REFLECTED DIFFUSIONS AND THEIR APPLICATIONS: FROM DATA-DRIVEN CONTROL TO GENERATIVE MODELS

ASBJØRN HOLK THOMSEN

PHD DISSERTATION

MARCH 2026

SOLIDUM PETIT IN PROFUNDIS · UNIVERSITAS ARHUSIENSIS ·

Reflected Diffusions and their Applications: From Data-Driven Control to Generative Models

Department of Mathematics
Aarhus University
Ny Munkegade 118
8000 Aarhus C
Denmark
E-mail: a.holk@math.au.dk

PhD dissertation by Asbjørn Holk Thomsen, written under supervision of Claudia Strauch, Heidelberg University, and Lukas Trottner, Stuttgart University.

Layout and typography by the author, typeset in LaTeX in the `memoir` class using a template borrowed with gratitude from Ragnhild Laursen.

AARHUS
UNIVERSITY

Abstract

Reflected stochastic differential equations (RSDEs) are a class of stochastic processes that, loosely speaking, behave like stochastic differential equations (SDEs) in the interior of some bounded domain $\Omega \subseteq \mathbb{R}^d$ while being reflected at the boundary $\partial\Omega$ of Ω, effectively confining them to the compact set $\overline{\Omega}$. Since SDEs are prevalent both in the modelling of real-world phenomena and in machine learning, it is natural to consider how such models and methods might be expanded through the use of RSDEs. In this dissertation, we consider some of the applications of RSDEs, where we first show that these provide novel ways of performing stochastic control, and later that they perform well in certain machine learning methods, where a bounded state space is natural.

In the first article, we consider how the reflection set Ω can provide a form of control over a stochastic process. In particular, we formulate an intuitive objective function based on the trajectory of an RSDE that expresses the goal of confining a particle to be near the origin without it hitting the boundary of the confining set too often. We then derive a closed form expression of this objective function using probabilistic analysis, and show that this cost function can be optimised both numerically and in a data-driven fashion.

In the second article, inspired by the use of SDEs in generative AI, we show that by replacing the underlying SDEs with RSDEs in these models, we obtain a more natural framework, since the data being generated is often itself bounded. Using spectral decomposition and rigorous neural network constructions, we show that the samples generated converge to the target distribution in total variation at a minimax optimal rate when assuming Sobolev smoothness of the target density.

Finally, in the third article, we continue in the framework of the second, but now considering a type of manifold hypothesis, where the target distribution is concentrated on an affine plane of much lower dimension than the ambient space. Using stochastic analysis, we derive an explicit solution to the involved RSDE, which we use to effectively approximate the target distribution and the affine plane on which it is supported, ultimately yielding near minimax optimal rates in Wasserstein-1-distance under assumptions of Sobolev smoothness.

Resumé

Reflekterede stokastiske differentialligninger (RSD'er) dækker over en familie af stokastiske processer, som løst sagt opfører sig som stokastiske differentialligninger (SD'er) inden i et begrænset domæne $\Omega \subseteq \mathbb{R}^d$, men som bliver reflekteret i randen $\partial\Omega$ af Ω, hvormed de bliver begrænset til den kompakte mængde $\overline{\Omega}$. Eftersom SD'er er meget udbredte i både modelleringen af fænomener fra den virkelige verden og i machine learning, er det naturligt at overveje, hvordan sådanne modeller kan udvides ved brug af RSD'er. I denne afhandling overvejer vi nogle af anvendelserne af RSD'er, hvor vi først viser, at disse tilbyder nye måder at udføre stokastisk kontrol, og senere at de præsterer godt i visse machine learning-metoder, hvor et begrænset tilstandsrum er naturligt.

I den første artikel betragter vi, hvordan reflektionsmængden Ω kan benyttes som en form for kontrol over en stokastisk proces. Specifikt formulerer vi en intuitiv målfunktion baseret på stien af en RSD, som udtrykker ønsket om at begrænse en partikel tæt på origo, men uden at den rammer randen af reflektionsmængden for ofte. Vi udleder derefter et lukket udtryk for denne målfunktion via sandsynlighedsteoretisk analyse, og viser at denne kan blive optimeret både numerisk og ved hjælp af observeret data.

I den anden artikel viser vi, inspireret af brugen af SD'er i generativ AI, at vi ved at erstatte de underliggende SD'er med RSD'er opnår en mere naturlig konstruktion, da det genererede data ofte selv er begrænset. Ved hjælp af spektralanalyse og dybdegående konstruktioner af neurale netværk viser vi, at det genererede data konvergerer mod det ønskede mål i total variation med en minimax-optimal rate, når vi antager at det underliggende mål har en Sobolev-glat tæthed.

Til sidst fortsætter vi i tredje artikel med den samme problemstilling, men hvor vi nu antager en slags mangfoldighedshypotese, hvor det ønskede mål er koncentreret på et affint plan af meget lavere dimension end det latente rum. Via stokastisk analyse udleder vi en eksplicit løsning til den involverede RSD, hvilket vi bruger til effektivt at approximere både det ønskede mål og dets støtte. Herved opnår vi under lignende glathedsantagelser nær minimax-optimale konvergensrater i Wasserstein-1-afstand.

Acknowledgements

First and foremost, I would like to extend by deepest gratitude to my supervisors Claudia Strauch and Lukas Trottner. Your friendly attitudes, words of encouragement and belief in me not just as a mathematician, but also as an independent researcher have been truly invaluable. I could not have asked for better supervision, and I hope that we get the chance to collaborate again in the future. To this, I would also like to thank my third co-author, Sören Christensen, who was a great help when I was just starting my PhD by not only being a great collaborator, but also by inviting me to Kiel for a research stay. I would also like to extend my heartfelt regards to the many great people I met during my extended research stay in Heidelberg, in particular Ricardo Blum, Timo Dörzbach, Bianca Neubert, Hans Reimann, Maximilian Siebel and Henning Stein. Your generosity in welcoming me to Heidelberg truly made me feel at home, and for that I am forever grateful. I wish you all the best and hope to see you again soon. I also want to thank my collaborators at the Interdisciplinary Nanoscience Center (iNANO) from whom I have learned how my research can actually be applied to real-world problems. I am of course also very grateful to the Villum Foundation for their financial support allowing for this project to exist in the first place.

I want to also thank my family – my father, my sister, Kis and Jørgen – for their unwavering support of me, not only for the past three years, but for my entire life. You have shaped me into the person that I am today, and I could not be more grateful for the encouragement I have received from you all. I would also like to thank all of my friends, without whom I would never have been able to finish this degree. From cake in MatKant with my friends from university to enjoying craft beers with my many dear friends and fellow volunteers from Fairbar, all of these great moments away from the research have been what has kept me sane, when the stress of deadlines have built up. Thank you all for filling my life with joy each day.

Lastly I want to thank my wonderful girlfriend Josephine. Over the past two years, you have reassured me when I had doubts in my abilities, you have put up with late days at work and you have helped me proof-read countless emails. But most importantly, you have been there to comfort me when I was at my lowest, and celebrated with me when I was at my highest. I hope we will have many more opportunities to do the latter soon.

Asbjørn Holk Thomsen
Aarhus, March 2026

Preface

This dissertation is the final product of a three year PhD programme at the Department of Mathematics, Aarhus University under the supervision of Claudia Strauch and Lukas Trottner. The main text consists of two main parts: first are two introductory chapters written for this dissertation, the first of which provides an introduction to the core concepts to appear in the remainder, and the second a brief and high level overview of the results and proof strategies used in the articles that follow. The second part consists of three articles in full, written in the course of PhD programme. They are:

Article A Sören Christensen, Asbjørn Holk Thomsen, and Lukas Trottner (2024)
Data-Driven Rules for Multidimensional Reflection Problems
SIAM/ASA Journal on Uncertainty Quantification 12:4, 1240-1272
DOI: 10.1137/23M1618570; arXiv:2311.06639 [math.OC]

Article B Asbjørn Holk, Claudia Strauch and Lukas Trottner (2024)
Statistical Guarantees for Denoising Reflected Diffusion Models
Accepted in the Journal of Machine Learning Research (JMLR)
arXiv:2411.01563 [math.ST]

Article C Asbjørn Holk, Claudia Strauch and Lukas Trottner (2026)
Reflected Diffusion Models Adapt to Low-dimensional Data
Preprint available
arXiv:2603.24495 [math.ST]

In accordance with GSNS rules and regulations, Article A and parts of an early draft of Article B were also used in the progress report for the qualifying examination.

All articles as they appear in this dissertation have been slightly modified for the sake of consistent notation and layout. The contents, however, have not been changed and are identical to how they appear either in their published form when applicable or the arXiv version.

* * *

We note here some common notation: in general, random variables, processes etc. are written with sans-serif font (e.g. X, B etc.) for sake of distinguishability. The d-dimensional Lebesgue measure is denoted as λ_d. For a metric space S, $x \in S$ and $r > 0$, the ball of radius r with centre x is denoted $B(x, r)$. The distance from x to a set $A \subseteq S$ is denoted as $\operatorname{dist}(x, A) = \inf_{y \in A} d(x, y)$. For general $x \in \mathbb{R}^d$, $|x|$ denotes the Euclidean norm, however for multi-indices $\beta \in \mathbb{N}_0^d$, $|\beta|$ refers to the 1-norm, i.e. $|\beta| = \beta_1 + \beta_2 + \ldots + \beta_d$. We write $a \lesssim_\theta b$ if there exists a constant $C_\theta > 0$ depending on θ such that $a \leq C_\theta b$ and $a \lesssim b$ if C_θ is a universal constant.

Contents

1 Introduction

The aim of this introductory chapter is to equip the reader with the necessary definitions and concepts to allow them to delve into the articles that follow, as well as tie these together into one contiguous study of the central topic – reflected diffusions and their applications. It is not aimed at domain experts. Rather, the target audience is those who have a curiosity in and keen understanding of mathematics, especially statistics and probability theory, at a level of a masters student or above. As such, many topics such as stochastic processes and stochastic and partial differential equations are assumed to be well-known and will not be explained in detail. A brief intuitive reminder, however, will be provided for each such topic when appropriate.

1.1 Reflected processes

The core concept of reflected processes is, in this author's opinion, quite intuitive. Suppose you have some stochastic process $\mathsf{X} = (\mathsf{X}_t)_{t \geq 0}$ defined on $\mathbb{R}^d$. Given some subset $\Omega \subseteq \mathbb{R}^d$, the associated *reflected* process $\mathsf{X}^\Omega = (\mathsf{X}_t^\Omega)_{t \geq 0}$ is then another stochastic process which in some sense behaves the same as X on the interior of Ω, but which is *reflected* at the boundary $\partial\Omega$ of Ω in the sense that it is pushed back into the interior, essentially constraining the process to only exist on $\overline{\Omega}$. In general, these can be rigorously defined via the associated Skorokhod problems and maps:

Definition 1.1: For any non-empty open set $\Omega \subseteq \mathbb{R}^d$ with reflection vector $v : \partial\Omega \to \mathbb{R}^d$, the solution to the associated *Skorokhod problem* given a càdlàg function $x : [0, T) \to \mathbb{R}^d$ with $T \in (0, \infty]$ is a pair of càdlàg functions $x^\Omega : [0, T) \to \overline{\Omega}$ and $\ell : [0, T) \to [0, \infty)$ such that $\ell(0) = 0$,

$$x^\Omega(t) = x(t) + \int_0^t v\big(x^\Omega(s)\big)\, \mathrm{d}\ell(s) \tag{1.1}$$

and

$$\ell(t) = \int_0^t \mathbf{1}_{\{x^\Omega(s) \in \partial\Omega\}}\, \mathrm{d}\ell(s).$$

The associated *Skorokhod map* is a mapping Γ^Ω such that $\Gamma^\Omega(x) = x^\Omega$.

Remark 1.2: In the above definition, the reflection vector $v : \partial\Omega \to \mathbb{R}^d$ is left intentionally vague, as it is highly situationally dependent. However, a common choice which we will be making in the articles that follow is choosing v as (a scaling of) the inward-pointing normal vector n whenever such a vector is uniquely defined. For instance, if $\partial\Omega$ is smooth enough in the sense that the signed distance

$$d_\Omega(x) = \begin{cases} \operatorname{dist}(x, \partial\Omega), & \text{if } x \in \Omega \\ -\operatorname{dist}(x, \partial\Omega), & \text{if } x \in \Omega^{\mathrm{c}} \end{cases} \tag{1.2}$$

is differentiable on $\partial\Omega$, then we may define $n = \nabla d_\Omega$.

With this, if X has almost surely càdlàg paths, we may then define X^Ω as

$$\mathsf{X}^\Omega(\omega) = \begin{cases} \Gamma^\Omega\big(\mathsf{X}(\omega)\big), & \text{if } t \mapsto \mathsf{X}_t(\omega) \text{ is càdlàg} \\ 0, & \text{otherwise.} \end{cases}$$

We remark that this construction already opens a wealth of options for X as all Lévy processes have an almost surely càdlàg version. However, since we will in this context only be working with processes that are pathwise continuous, we will not be exploring further generalisations.

Reflected stochastic differential equations

If instead of considering general processes X as our reflection candidates, we restrict ourselves to diffusion processes in the form of solutions to stochastic differential equations (SDEs) of the form

$$\mathrm{d}\mathsf{X}_t = b(\mathsf{X}_t)\,\mathrm{d}t + \sigma(\mathsf{X}_t)\,\mathrm{d}\mathsf{B}_t, \tag{1.3}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times m}$ are suitable functions and $(\mathsf{B}_t)_{t\geq 0}$ is a standard m-dimensional Brownian motion, the resulting process X^Ω now solves an associated *reflected stochastic differential equation* (RSDE). To better motivate their definition, note that we may re-write (1.1) as

$$\mathrm{d}x^\Omega(t) = \mathrm{d}x(t) + v\big(x^\Omega(t)\big)\,\mathrm{d}\ell(t).$$

Plugging in (1.3), we have

$$\mathrm{d}\mathsf{X}_t^\Omega = b(\mathsf{X}_t)\,\mathrm{d}t + \sigma(\mathsf{X}_t)\,\mathrm{d}\mathsf{B}_t + v(\mathsf{X}_t^\Omega)\,\mathrm{d}\mathsf{L}_t,$$

where $(\mathsf{L}_t)_{t\geq 0}$ is the now stochastic pathwise function from Definition 1.1. The process X^Ω defined here is special in that it is defined via an already existing solution to (1.3), i.e. $\mathsf{X}^\Omega = \Gamma^\Omega(\mathsf{X})$. Such processes are referred to as *strong solutions* to the associated RSDE. However, we are often only interested in the law of such processes, and to this, only *weak solutions* are necessary, i.e. solutions to the SDE

$$\mathrm{d}\mathsf{X}_t^\Omega = b(\mathsf{X}_t^\Omega)\,\mathrm{d}t + \sigma(\mathsf{X}_t^\Omega)\,\mathrm{d}\mathsf{B}_t + v(\mathsf{X}_t^\Omega)\,\mathrm{d}\mathsf{L}_t.$$

This is all summarised in the following definition:

Definition 1.3: Given a set $\Omega \subseteq \mathbb{R}^d$, a distribution μ on Ω and functions $b : \mathbb{R}^d \to \mathbb{R}^d$, $v : \partial\Omega \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times m}$, a solution to the RSDE

$$\mathrm{d}\mathsf{X}_t^\Omega = b(\mathsf{X}_t^\Omega)\,\mathrm{d}t + \sigma(\mathsf{X}_t^\Omega)\,\mathrm{d}\mathsf{B}_t + v(\mathsf{X}_t^\Omega)\,\mathrm{d}\mathsf{L}_t \tag{1.4}$$

is a collection $(\mathsf{X}_t^\Omega, \mathcal{F}_t, \mathsf{B}_t, \mathsf{L}_t)_{t\geq 0}$ satisfying the following:

(i) $(\mathcal{F}_t)_{t\geq0}$ is a filtration satisfying the natural conditions and $(\mathsf{B}_t)_{t\geq0}$ is an m-dimensional $(\mathcal{F}_t)$-Brownian Motion

(ii) $(\mathsf{L}_t)_{t\geq0}$ is a continuous, non-decreasing process satisfying $\mathsf{L}_0 \equiv 0$ and

$$\mathsf{L}_t = \int_0^t \mathbf{1}_{\{\mathsf{X}_t\in\partial D\}}\,\mathrm{d}\mathsf{L}_s$$

(iii) $(\mathsf{X}_t^\Omega)_{t\geq0}$ is a continuous and adapted process satisfying $\mathsf{X}_0^\Omega \sim \mu$, $\mathsf{X}_t^\Omega \in \overline{\Omega}$, and

$$\mathsf{X}_t^\Omega = \mathsf{X}_0^\Omega + \int_0^t b(\mathsf{X}_s^\Omega)\,\mathrm{d}s + \int_0^t \sigma(\mathsf{X}_s^\Omega)\,\mathrm{d}\mathsf{B}_s + \int_0^t v(\mathsf{X}_s^\Omega)\,\mathrm{d}\mathsf{L}_s$$

for all $t \geq 0$.

We say that a solution is *strong* if $\mathsf{X}^\Omega = \Gamma^\Omega(\mathsf{X})$ where Γ^Ω is the Skorokhod map and X is a strong solution to the associated non-reflected SDE

$$\mathrm{d}\mathsf{X}_t = b(\mathsf{X}_t)\,\mathrm{d}t + \sigma(\mathsf{X}_t)\,\mathrm{d}\mathsf{B}_t.$$

The process $\mathsf{L} = (\mathsf{L}_t)_{t\geq0}$ defined here is known as the *local time of* X^Ω *at* $\partial\Omega$, and is to be thought of as the amount of time X^Ω has spent at the boundary $\partial\Omega$. However, in general this process depends on the choice of reflection vector as one can simply choose for some $a > 0$ to instead use the reflection vector av resulting in a local time of $a^{-1}\mathsf{L}$. In the case where $d = 1$, this is different. Here, one can show that for any convex function φ and continuous semi-martingale X, there exists a process $\mathsf{A}^\varphi = (\mathsf{A}_t^\varphi)_{t\geq0}$ such that

$$\varphi(\mathsf{X}_t) = \varphi(\mathsf{X}_0) + \int_0^t \varphi'_-(\mathsf{X}_s)\,\mathrm{d}\mathsf{X}_s + \frac{1}{2}\mathsf{A}_t^\varphi,$$

where φ'_- is the left derivative of φ. Then, the local time at some point $a \in \mathbb{R}$ is defined as $\mathsf{L}^a = 2\mathsf{A}^{\varphi_a}$, where $\varphi_a(x) = |x - a|$. In fact, with this choice, the process A^φ can be derived explicitly via the so-called Itô–Tanaka formula which states that

$$\varphi(\mathsf{X}_t) = \varphi(\mathsf{X}_0) + \int_0^t \varphi'_-(\mathsf{X}_s)\,\mathrm{d}\mathsf{X}_s + \frac{1}{2}\int_{\mathbb{R}} \mathsf{L}_t^a\,\varphi''(\mathrm{d}a), \tag{1.5}$$

where φ'' is understood in a distributional sense. With this definition, one can also show the following alternative characterisation of L^a, if X has continuous paths:

$$\mathsf{L}^a_t = \lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_0^t \mathbf{1}_{(a-\varepsilon, a+\varepsilon)}(\mathsf{X}_s)\, \mathrm{d}\langle \mathsf{X} \rangle_s,$$

supporting the idea that L^a_t in some sense denotes how long X has been at a up to time t. This idea can be extended to higher dimensions where we may appropriately scale the reflection vector v such that it satisfies a similar occupation time formula:

$$\mathsf{L}_t = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \int_0^t \mathbf{1}_{(\partial\Omega)_\varepsilon}(\mathsf{X}^\Omega_s)\, \mathrm{d}s, \tag{1.6}$$

where $(\partial\Omega)_\varepsilon = \{x \in \mathbb{R}^d \mid \operatorname{dist}(x, \partial\Omega) < \varepsilon\}$ is the ε-fattening of $\partial\Omega$.

Applications

One might be rightly curious as to why one should study these processes and if they serve any purpose beyond academic curiosity. To this, we first consider the most straight-forward physical interpretation of RSDEs. Famously, Brownian motion is named after botanist Robert Brown who observed the erratic movement of pollen suspended in water, and the connection to physical particles is consistent in the mathematical vernacular. Diffusion processes are named as such since these mathematical objects describe the movement of individual particles suspended in a fluid when they *diffuse* in space. As such, the natural extension to reflected diffusions is that these describe the movement of particles when they diffuse in a *confined space* such as a room. With this notion, it is perhaps not surprising that as with the deep connection between SDEs and physical models involving partial differential equations (PDEs), so too is there one between RSDEs and PDEs with boundary conditions. In particular, if X solves (1.3) then the family of densities $(p_t)_{t\geq 0}$ of X solve the PDE

$$\begin{cases} \frac{\partial}{\partial t} u(x,t) = \mathcal{A} u(x,t), & (x,t) \in \mathbb{R}^d \times (0,\infty) \\ u(x,0) = p_0(x), & x \in \mathbb{R}^d \end{cases}$$

where $\mathcal{A} f = b^\top \nabla f + \frac{1}{2}\operatorname{tr}(\sigma\sigma^\top \nabla^2 f)$ is the generator of X. Conversely, if X^Ω solves (1.4) then its densities $(p^\Omega_t)_{t\geq 0}$ solve the same PDE with Neumann boundary conditions, i.e.

$$\begin{cases} \frac{\partial}{\partial t} u(x,t) = \mathcal{A} u(x,t), & (x,t) \in \Omega \times (0,\infty) \\ \frac{\partial}{\partial v} u(x,t) = 0, & (x,t) \in \partial\Omega \times (0,\infty) \\ u(x,0) = p^\Omega_0(x), & x \in \Omega. \end{cases}$$

That is, the above PDEs describe how the physical systems as a whole evolve over time, while the stochastic processes X and X^Ω describe how any single particle in that system behaves. As such, RSDEs hold the potential to further our understanding of more realistic physical models, much like SDEs have in the past through discoveries like the Feynman–Kac formula. Of course there are a wealth of other fields such as economics, biology, computer science etc. where SDEs have played an important role in modelling and simulating a myriad of phenomena, and here too an extension to RSDEs might very well also lead to more accurate models.

1.2 Approximation theory

Another reoccurring theme in the articles that follow is that of approximation rates and estimator optimality. To illustrate the importance of this, consider the following, purposefully ridiculous example: suppose you have i.i.d. data $(\mathsf{X}_n)_{n\in\mathbb{N}}$ with $\mathsf{X}_1 \sim N(\theta, 1)$, and you wish to construct for $n \in \mathbb{N}$ an estimator $\widehat{\theta}_n$ of θ based on the first n observations. The obvious choice here, of course, is the average $\widehat{\theta}_n^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \mathsf{X}_i$. However, let us also consider the alternative estimator $\widehat{\theta}_n^{(2)} = \frac{1}{\lceil \log n \rceil}\sum_{i=1}^{\lceil \log n \rceil} \mathsf{X}_i$. Both estimators are unbiased and converge both almost surely and in L^2 as $n \to \infty$. The difference is in the *rate of convergence*, in that no matter what rate with which $\widehat{\theta}_n^{(1)}$ converges to θ, we need exponentially more data for $\widehat{\theta}_n^{(2)}$ to achieve the same accuracy. This leads to the notion of *statistical efficiency* of an estimator, i.e. how quickly it converges to the object it is estimating. For parametric models, i.e. ones where given $\mathsf{X} \sim \mu_{\theta_0} \in \{\mu_\theta \mid \theta \in \Theta \subseteq \mathbb{R}^k\}$ we try to estimate θ_0, the maximum likelihood estimator (when available) is known to have many desirable properties. For instance, it is consistent and asymptotically efficient, meaning that it converges to the true parameter being estimated and that its variance is asymptotically minimal as per the Cramér–Rao lower bound, implying that its mean squared error (or L^2 distance) is asymptotically lower than any other estimator. With this, we might say that the maximum likelihood estimator is *asymptotically optimal in* L^2. However, while statisticians often impose a parametric model onto observed data, many real-world phenomena are much too complicated to conform to such models, and so we can benefit from eschewing such assumptions.

Non-parametric statistics and minimax optimality

In the world of non-parametric statistics, we no longer assume that the observed data follows a parametric distribution where we are then interested in these parameters. Instead, we strive to make estimations of certain properties of the underlying distribution μ (such as its density, correlation structure, modes etc.) in a way that is *model free* i.e. making no or as few as possible assumptions on μ. A classic example of this is the kernel density estimator: suppose μ has a continuous density f with respect to the Lebesgue measure and we observe $\mathsf{X}_1, \mathsf{X}_2, \dots, \mathsf{X}_n \overset{\text{i.i.d.}}{\sim} \mu$. Then, for some non-negative function K (the *kernel*) with $\int_{\mathbb{R}} K(x)\,\mathrm{d}x = 1$ and $h > 0$ we set

$$\widehat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - \mathsf{X}_i}{h}\Big).$$

One can show (see e.g. [124]) that if we assume that there exists some $\alpha, L > 0$ such that f is $\lfloor \alpha \rfloor$ times differentiable and $|f^{(\lfloor \alpha \rfloor)}(x) - f^{(\lfloor \alpha \rfloor)}(y)| \leq L|x-y|^{\alpha - \lfloor \alpha \rfloor}$ (here $\lfloor x \rfloor$ denotes the largest integer *strictly* less than x), then for a fitting choice of kernel K, there exists constants C_1, C_2 such that

$$\mathbb{E}\left[\left(\widehat{f}_n(x) - f(x)\right)^2\right] \leq \frac{C_1}{nh} + C_2^2 h^{2\alpha},$$

which is minimized at $h^* = \left(\frac{C_1}{2\alpha C_2^2}\right)^{\frac{1}{2\alpha+1}} n^{-\frac{1}{2\alpha+1}}$ yielding a convergence rate of $O(n^{-\frac{2\alpha}{2\alpha+1}})$. This pointwise convergence rate can be extended to an integrated risk, i.e. one can show that

$$\mathbb{E}\left[\int_{\mathbb{R}} \left(\widehat{f}_n(x) - f(x)\right)^2 \mathrm{d}x\right] = \mathbb{E}\left[\|\widehat{f}_n - f\|_{L^2}^2\right] \lesssim n^{-\frac{2\alpha}{2\alpha+1}}$$

Can we do better? As it turns out, we cannot. Indeed, one can show that (see e.g. [82]) for $p \geq 2$ and densities $f : \mathbb{R}^d \to \mathbb{R}$ in the more general Nikol'skii class

$$\mathcal{N}_p(\alpha, L) = \{f : \mathbb{R}^d \to \mathbb{R} \mid \partial^\beta f \text{ exists and } \|\partial^\beta f(\cdot + t) - \partial^\beta f\|_{L^p} \leq C|t|^{\alpha - |\beta|} \text{ for all } |\beta| \leq \alpha\},$$

we have

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{N}_p(\alpha, L)} \mathbb{E}\left[\|f - \widehat{f}\|_p\right] \gtrsim n^{-\frac{\alpha}{2\alpha+d}}.$$

Here, the infimum in front is taken over estimators $\widehat{f}$ of f based on n samples $\mathsf{X}_1, \mathsf{X}_2, \dots, \mathsf{X}_n \sim f$. In other words, we can never hope to achieve a better asymptotic error rate, as this is the smallest possible over this particular class of densities. Since this rate arises from the minimiser of the maximal error, we call this the *minimax rate*, and since the kernel density estimator above achieves this rate, we say that it is *minimax optimal.* Of course, we can choose a different risk measure or a different class of densities, and so we give a more general definition as follows:

Definition 1.4 (minimax risk and -optimality): Given a class $\mathcal{P}$ of probability measures on some space S, a mapping $\theta : \mathcal{P} \to \Theta$ and a distance function $\rho : \Theta^2 \to [0, \infty)$, let $\widehat{\Theta}_n$ denote the set of all estimators $\widehat{\theta}_n : S^n \to \Theta$ of θ. Then the *minimax risk* is given by

$$R_n = \inf_{\widehat{\theta}_n \in \widehat{\Theta}_n} \sup_{\mu \in \mathcal{P}} \mathbb{E}_\mu\left[\rho\left(\widehat{\theta}_n(\mathsf{X}_1, \mathsf{X}_2, \dots, \mathsf{X}_n), \theta(\mu)\right)\right].$$

If a family of estimators $\{\widehat{\theta}_n\}_{n \in \mathbb{N}}$ with $\widehat{\theta}_n \in \widehat{\Theta}_n$ satisfies

$$\sup_{\mu \in \mathcal{P}} \mathbb{E}_\mu\left[\rho\left(\widehat{\theta}_n(\mathsf{X}_1, \mathsf{X}_2, \dots, \mathsf{X}_n), \theta(\mu)\right)\right] \leq C R_n, \qquad \forall n \in \mathbb{N}$$

for some universal constant $C > 0$, we say that the family is *minimax optimal.*

In the above definition, we consider only discrete observations $(\mathsf{X}_n)_{n\in\mathbb{N}}$, however it can be straightforwardly modified to include continuous observations $(\mathsf{X}_t)_{t\geq 0}$ or any other index set. For the example above, we would have

$$\mathcal{P} = \{\mu \mid \mu(\mathrm{d}x) = f(x)\,\mathrm{d}x, f \in \mathcal{N}_p(\alpha, L)\},$$

the distance function ρ would be $\|\cdot\|_{L^p}$ and θ would be the function mapping a measure $\mu \in \mathcal{P}$ to its density f. We finish this section by listing some key settings and rates which will appear in the articles that follow:

Definition 1.5 (Total variation): Given a measurable space $(\Omega, \mathcal{F})$ and two probability measures μ and ν thereon, the *total variation distance* μ and ν is defined as

$$\mathrm{TV}(\mu, \nu) := \sup_{A\in\mathcal{F}} |\mu(A) - \nu(A)|.$$

One can show that if μ and ν admit densities f and g, respectively, then $\mathrm{TV}(\mu, \nu) = \frac{1}{2}\|f - g\|_{L^1}$

Beyond providing a very intuitive way of measuring distance between two distributions, total variation can also be nicely bounded via Pinsker's inequality which states that

$$\mathrm{TV}(\mu, \nu) \leq \sqrt{\frac{1}{2}\mathrm{KL}(\mu \,\|\, \nu)},$$

where KL denotes the Kullback–Leibler divergence. Together with Girsanov's theorem, this bounds the total variation between the laws of two (R)SDE's with differing drifts by the integrated L^2-distance between these drifts – this will become important in Articles B and C.

Definition 1.6 (Wasserstein distance): Given probability measures μ and ν on $\mathbb{R}^d$, let $\Pi(\mu, \nu)$ denote the set of all couplings of μ and ν, i.e.

$$\Pi(\mu, \nu) = \left\{\pi : \mathcal{B}(\mathbb{R}^d)^2 \to [0, 1] \mid \forall A, B \in \mathcal{B}(\mathbb{R}^d) : \pi(A, \mathbb{R}^d) = \mu(A), \pi(\mathbb{R}^d, B) = \nu(B)\right\}.$$

Then for $p \geq 1$ the *Wasserstein-p-distance* is given by

$$\mathcal{W}_p(\mu, \nu) := \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_\pi\left[|\mathsf{X} - \mathsf{Y}|^p\right]^{1/p}.$$

One drawback of total variation is that it cannot measure similarity of singular measures, even if they seem close. For instance, consider for $\sigma > 0$ the distributions $\mu_\sigma = N(0, \sigma^2)$ and $\nu = \delta_0$. Then $\mu_\sigma \xrightarrow{\mathrm{w}} \nu$ as $\sigma \to 0$, however regardless of σ, we have

$$\mathrm{TV}(\mu_\sigma, \nu) \geq |\mu_\sigma(\{0\}) - \nu(\{0\})| = 1.$$

In contrast, letting $\mathsf{Z} \sim N(0,1)$, we have for any $p \geq 1$

$$\mathcal{W}_p(\mu_\sigma, \nu) \leq \mathbb{E}\big[|\sigma\mathsf{Z} - 0|^p\big]^{1/p} = \sigma\mathbb{E}\big[|\mathsf{Z}|^p\big]^{1/p} \xrightarrow[\sigma\to 0]{} 0.$$

Definition 1.7 (Anisotropic Hölder space): For $\boldsymbol{\beta}, \boldsymbol{L} \in (0,\infty)^d$, the *anisotropic Hölder space* $\mathcal{H}_d(\boldsymbol{\beta}, \boldsymbol{L})$ is the set of all functions $g : \mathbb{R}^d \to \mathbb{R}$ satisfying

$$\|\partial_i^k g\|_\infty \leq L_i, \quad k = 1, \dots, \lfloor \beta_i \rfloor,$$

and

$$\|\partial_i^{\lfloor \beta_i \rfloor} g(\,\cdot + te_i) - \partial_i^{\lfloor \beta_i \rfloor} g\| \leq L_i |t|^{\beta_i - \lfloor \beta_i \rfloor}, \quad t \in \mathbb{R}.$$

Here, $\lfloor x \rfloor$ denotes the largest integer *strictly* less than x.

A canonical example of such a function is given by $x \mapsto \mathrm{e}^{-|x|^2} x^{\boldsymbol{\beta}} = \prod_{i=1}^d \mathrm{e}^{-x_i^2} x_i^{\beta_i}$. Note that any anisotropic Hölder smooth function is also automatically $\beta_{\min}$-Hölder smooth where $\beta_{\min} = \min\{\beta_1, \beta_2, \dots, \beta_d\}$. The benefit of considering this class rather than the usual (isotropic) Hölder-class is that we can attain convergence rates depending on the harmonic mean smoothness $\overline{\boldsymbol{\beta}+1} := (\frac{1}{d}\sum_{i=1}^d \frac{1}{\beta_i+1})^{-1}$ rather than the overall smoothness $\beta_{\min}$ which may be much lower. Indeed, it shown in [111] that for a certain subclass of diffusion processes X with stationary density $\rho \in \mathcal{H}_d(\boldsymbol{\beta}, \boldsymbol{L})$, the minimax-rate when observing $(\mathsf{X}_t)_{t\in[0,T]}$ wrt. $\|\cdot\|_\infty$ is $O\Big(\big(\frac{\log T}{T}\big)^{\frac{\overline{\boldsymbol{\beta}+1}}{\overline{\boldsymbol{\beta}+1}+d-2}}\Big)$ for $d \geq 3$.

Definition 1.8 (Sobolev space): For $\alpha \in \mathbb{N}$, $p \in [1,\infty]$ and $S \subseteq \mathbb{R}^d$, we define the *Sobolev space* $W^{\alpha,p}(S)$ as the set of functions $f : S \to \mathbb{R}$ with weak derivatives of all orders less than α in $L^p(S)$, i.e.

$$W^{\alpha,p}(S) := \big\{ f : S \to \mathbb{R} \mid \forall \beta \in \mathbb{N}_0, |\beta| \leq \alpha : \partial^\beta f \in L^p(S) \big\}.$$

We denote the case of $p = 2$ as $H^\alpha(S)$. We also let $H^\alpha(S)_0$ denote the subset of $H^\alpha(S)$ consisting of functions that vanish in a trace sense at the boundary ∂S of S, i.e.

$$H_0^\alpha(S) := \big\{ f \in H^\alpha(S) \mid \forall \beta \in \mathbb{N}_0, |\beta| \leq \alpha - 1 : \partial^\beta f|_{\partial S} = 0 \big\}$$

Sobolev spaces allow us to assume certain smoothness properties of target distributions without necessarily being smooth in the sense of classical derivatives. Moreover, the specific spaces $H^\alpha(S)$ have nice analytic properties in that these spaces equipped with the inner product

$$\langle f, g \rangle_{H^\alpha(S)} = \sum_{|\beta|\leq\alpha} \langle \partial^\beta f, \partial^\beta g \rangle_{L^2(S)}$$

are in fact Hilbert spaces. Finally, so long as the boundary ∂S of S is sufficiently smooth, $H^\alpha(S)$ can be extended and embedded in larger spaces in the sense that there exists a continuous extension $\mathcal{E} : H^\alpha(S) \to H^\alpha(\mathbb{R}^d)$ such that $\mathcal{E}f(x) = f(x)$ for $x \in S$, and we have $H^\alpha(S) \subseteq C^k(S)$ for $k < \alpha - \frac{d}{2}$. For the space $H_0^\alpha(S)$, the extension can be made explicit in that

$$\mathcal{E}f(x) = \begin{cases} f(x), & \text{if } x \in S \\ 0, & \text{otherwise.} \end{cases}$$

For distributions with densities in $H^\alpha(S)$, it follows from [133, Theorem 4] that the minimax-optimal rate in total variation is $O(n^{-\frac{\alpha}{2\alpha+d}})$ while [86, Theorem 3] states that for Wasserstein-p-distance it is $O(n^{-\frac{\alpha+1}{2\alpha+d}})$.

1.3 Neural networks

Neural networks are increasingly common as tools to approximate a wide range of functions. Although the term sometimes covers different architectures, in this introduction and the chapters that follow, we take neural networks to mean feed-forward neural networks, i.e. functions $\varphi : \mathbb{R}^d \to \mathbb{R}^m$ which consists of a series of affine transformations interspersed with activation functions that allow for non-linearity. We can specify any such network by its affine transformations and activation function as follows: let $A_i \in \mathbb{R}^{W_{i+1} \times W_i}$, $b_i \in \mathbb{R}^{W_i}$ for $i = 0, \dots, L$, where $\{W_i\}_{i=0}^{L+1} \subseteq \mathbb{N}$ are the widths of each layer of φ – in particular $W_0 = d$ and $W_{L+1} = m$. Also for $i = 0, \dots, L$, let $\sigma_i : \mathbb{R} \to \mathbb{R}$ be some non-linear activation function and set

$$\sigma_{b_i} x = \begin{bmatrix} \sigma_i(x_1 - b_{i,1}) & \sigma_i(x_2 - b_{i,2}) & \cdots & \sigma_i(x_{W_i} - b_{i,W_i}) \end{bmatrix}^\top, \qquad x \in \mathbb{R}^{W_i}.$$

Then we may write

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

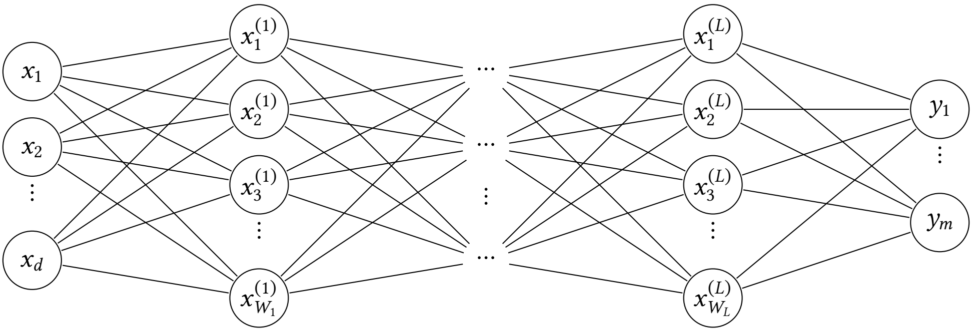

Figure 1.1: *Structure diagram of a neural network.*

In our setting, we will always take $\sigma_i = \sigma$ to be the *rectified linear unit* (ReLU) given by $\sigma(x) = x \vee 0$. It is not difficult to see that with this choice of activation function, every neural network is piecewise linear, since this is true of σ and this class is closed under composition. Such functions are dense in $C(K)$ wrt. $\|\cdot\|_\infty$ for any compact $K \subseteq \mathbb{R}^d$. That is, for any continuous function $f : K \to \mathbb{R}^m$ and $\varepsilon > 0$, there exists a piecewise linear function $\widehat{f} : K \to \mathbb{R}^m$ such that $\|f - \widehat{f}\|_\infty < \varepsilon$.

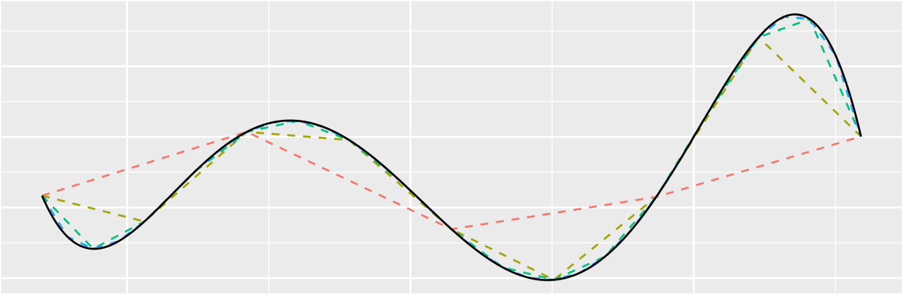

Figure 1.2: *Example of a function being approximated arbitrarily well by piecewise linear functions*

One could then hope that the same is true of neural networks, and luckily this is the case. Indeed, for a much wider range of activation functions, several so-called universal approximation theorems have been shown, see e.g. [59]. One strategy common in constructing neural networks targeting specific functions is to first approximate another class of functions which themselves are known to have good approximation properties. For instance, one might start by constructing an explicit neural network approximating the function $x \mapsto x^2$, and then since $xy = \frac{(x+y)^2-(x-y)^2}{4}$, this approximation easily extends to polynomials which often have known convergence rates through e.g. Taylor's theorem. Such methods are useful since explicit convergence rates allow us to evaluate how large a neural network must be to adequately approximate a given function. One way of parametrising neural networks of a certain size is given in [89] where for $L \in \mathbb{N}$, $W \in \mathbb{N}^{L+2}$, $S \in \mathbb{N}$ and $B > 0$ we let $\Phi(L, W, S, B)$ denote the set of neural networks ϕ satisfying:

- **Depth constraint**: The number of hidden layers in ϕ is no more than L
- **Width constraint**: The width the i'th layer is given by W_i for $i = 0, \dots, L+1$
- **Sparsity constraint**: The total number of non-zero entries in A_i, b_i is no more than S
- **Magnitude constraint**: All entries of A_i, b_i are numerically less than B

Often, as the networks grow large, it is inconvenient to keep track of the exact sizes of all networks, and we care only about their asymptotic sizes. This leads also to the class $\widetilde{\Phi}(L, W, S, B)$

where the above conditions only hold asymptotically up to a universal constant. This is summarised succinctly in the following definition:

Definition 1.9: $L \in \mathbb{N}$, $W \in \mathbb{N}^{L+2}$, $S \in \mathbb{N}$ and $B > 0$, we define the neural network class $\Phi(L, W, S, B)$ as

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

Furthermore, for $\widetilde{L}, \widetilde{W}, \widetilde{S}, \widetilde{B} > 0$, we define the asymptotic neural network class $\widetilde{\Phi}(\widetilde{L}, \widetilde{W}, \widetilde{S}, \widetilde{B})$ as

$$\widetilde{\Phi}(\widetilde{L}, \widetilde{W}, \widetilde{S}, \widetilde{B}) = \left\{ \Phi(L, W, S, B) \mid L \lesssim \widetilde{L}, \|W\|_\infty \lesssim \widetilde{W}, S \lesssim \widetilde{S}, B \lesssim \widetilde{B} \right\}.$$

2 Article overview and proof sketches

In this chapter, we will give a brief overview of the articles that follow, and the main results therein. In particular, we will give a high-level introduction and motivation to each of the topics discussed and as well as sketches of the main proofs and proof strategies used in each article. In the nature of the chapter, many details will be omitted, and we defer to the corresponding full articles for more rigorous explanations and derivations.

2.1 Article A: Data-driven rules for multidimensional reflection problems

We first motivate the central problem of this article by an (admittedly somewhat silly) example. Imagine that you are a goat herder and need to construct a fence to hold in your goats. Ideally, you want the goats to be near your farmhouse so that you will not need to walk as far every day to reach them, and there are certain areas you particularly want to keep the goats away from, e.g. a vegetable garden. However, you also want to minimise repair costs of the fence, and if you choose the fence perimeter too small, goats will more often bump into them, incurring damages. Thus you are faced with an optimisation problem: on one hand you want the fence to be as small as possible to avoid travelling too far, but not so small that the fence will be constantly damaged. How do you find the optimal fence perimeter?

We can model this using reflected processes as follows: suppose the position of any one goat is given by some stochastic process $\mathsf{X} = (\mathsf{X}_t)_{t\geq 0}$ on $\mathbb{R}^2$. Choosing some containment area $\Omega \subseteq \mathbb{R}^2$ with perimeter (i.e. fence) $\partial\Omega$, the movement of the constrained goats is then the reflected process $\mathsf{X}^\Omega = (\mathsf{X}^\Omega_t)_{t\geq 0}$ with associated local time $\mathsf{L}^\Omega = (\mathsf{L}^\Omega_t)_{t\geq 0}$. To model the cost of goats straying far away or into areas they are not allowed, choose some cost function $c : \mathbb{R}^2 \to [0, \infty)$ with its minimum at the origin (your farmhouse), such that $c(\mathsf{X}^\Omega_t)$ is the running cost of any goat at time t. To represent the cost of repairs, recall that L^Ω is to be thought of as the time the process X^Ω has spent at $\partial\Omega$, and hence the amount that any goat has bumped into the fence. Then, for some constant repair cost $\kappa > 0$, the total cost of the specific choice of fence at time $t > 0$ is given by

$$\mathsf{C}_t(\Omega) := \int_0^t c(\mathsf{X}^\Omega_s)\,\mathrm{d}s + \kappa \mathsf{L}^\Omega_t.$$

Assuming that all goats move independently of each other and of the time t, you invoke the law of large numbers to consider the long-term average cost, i.e.

$$J(\Omega) := \lim_{t\to\infty} \frac{1}{t}\mathbb{E}[\mathsf{C}_t(\Omega)] = \lim_{t\to\infty} \frac{1}{t}\mathbb{E}\bigg[\int_0^t c(\mathsf{X}^\Omega_s)\,\mathrm{d}s + \kappa \mathsf{L}^\Omega_t\bigg], \tag{2.1}$$

and the optimisation problem can now be exactly posed: find $\Omega^* \subseteq \mathbb{R}^2$ such that $J(\Omega^*) = \min_{\Omega \subseteq \mathbb{R}^2} J(\Omega)$.

An explicit objective function

It turns out that under certain assumptions, the objective function defined in (2.1) has an explicit expression. In particular, if we suppose now that X is a Langevin diffusion, i.e. a solution to the SDE

$$\mathrm{d}\mathsf{X}_t = -\nabla V(\mathsf{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathsf{B}_t$$

and X^Ω a solution to the associated normally reflected SDE

$$\mathrm{d}\mathsf{X}_t^\Omega = -\nabla V(\mathsf{X}_t^\Omega)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathsf{B}_t + n(\mathsf{X}_t^\Omega)\,\mathrm{d}\mathsf{L}_t^\Omega$$

for some function $V : \mathbb{R}^d \to \mathbb{R}$, then we have the following:

Theorem 2.1: If Ω is sufficiently smooth, then

$$J(\Omega) = \frac{1}{\int_\Omega \mathrm{e}^{-V(x)}\,\mathrm{d}x}\left(\int_\Omega c(x)\mathrm{e}^{-V(x)}\,\mathrm{d}x + \kappa \int_{\partial\Omega} \mathrm{e}^{-V(x)}\,\mathcal{H}^{d-1}(\mathrm{d}x)\right). \tag{2.2}$$

Here, the last integral is to be simply understood as the surface integral over $\partial\Omega$, formalised via the so-called *Hausdorff measure* $\mathcal{H}^{d-1}$. Langevin diffusions have a number of desirable qualities, first of which is that if the function e^{-V} is finitely integrable, that is if $\mathrm{e}^{-V} \in \mathcal{L}^1(\mathbb{R}^d)$, then X is ergodic with stationary distribution π where $\frac{\mathrm{d}\pi}{\mathrm{d}x} \propto \mathrm{e}^{-V}$. Secondly, this property carries over to the reflected process X^Ω very straightforwardly in that this is also ergodic with proportional density, but now restricted to $\overline{\Omega}$. That is, if X_∞ and X_∞^Ω are such that $\mathsf{X}_t \overset{\sim}{\to} \mathsf{X}_\infty$ and $\mathsf{X}_t^\Omega \overset{\sim}{\to} \mathsf{X}_\infty^\Omega$ as $t \to \infty$, then $\mathsf{X}_\infty^\Omega \sim (\mathsf{X}_\infty \mid \mathsf{X}_\infty \in \overline{\Omega})$. Combining these, we get the wonderfully intuitive interpretation of (2.2) that

$$J(\Omega) = \mathbb{E}\big[c(\mathsf{X}_\infty) \mid \mathsf{X}_\infty \in \overline{\Omega}\big] + \kappa\mathbb{P}\big(\mathsf{X}_\infty \in \partial\Omega \mid \mathsf{X}_\infty \in \overline{\Omega}\big).$$

Here, the last probability is a bit of abuse of notation in that the usual conditional probability above is often 0 as $\partial\Omega$ is of co-dimension 1. Instead, it should be interpreted in a limiting sense, i.e.

$$\mathbb{P}\big(\mathsf{X}_\infty \in \partial\Omega \mid \mathsf{X}_\infty \in \overline{\Omega}\big) := \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}\mathbb{P}\big(\mathrm{dist}(\mathsf{X}_\infty, \partial\Omega) \le \varepsilon \mid \mathsf{X}_\infty \in \overline{\Omega}\big).$$

The key to deriving (2.2) is using Itô's formula for a clever choice of function φ. First, note that for $f \in C^2(\overline{\Omega}, \mathbb{R})$ we have

$$\begin{aligned} &f(\mathsf{X}^\Omega_t) - f(\mathsf{X}^\Omega_0) \\ &= \int_0^t \nabla f(\mathsf{X}^\Omega_s) \cdot \mathrm{d}\mathsf{X}^\Omega_s + \frac{1}{2} \sum_{1 \le i,j \le d} \int_0^t \frac{\partial^2}{\partial x_i \partial x_j} f(\mathsf{X}^\Omega_s)\, \mathrm{d}\langle \mathsf{X}^{\Omega,(i)}, \mathsf{X}^{\Omega,(j)} \rangle_s \\ &= \int_0^t \nabla f(\mathsf{X}^\Omega_s) \cdot \mathrm{d}\mathsf{B}_t + \int_0^t \nabla f(\mathsf{X}^\Omega_s) \cdot n(\mathsf{X}^\Omega_s)\, \mathrm{d}\mathsf{L}^\Omega_s + \int_0^t \Delta f(\mathsf{X}^\Omega_s) - \nabla f(\mathsf{X}^\Omega_s) \cdot \nabla V(\mathsf{X}^\Omega_s)\, \mathrm{d}t. \end{aligned}$$

Now, since Ω is bounded and ∇f continuous, the first term is a martingale, and hence has expectation 0 under all initial measures. Furthermore, by the divergence theorem, we have

$$\begin{aligned} \int_{\overline{\Omega}} \big(\Delta f(x) - \nabla f(x) \cdot \nabla V(x)\big) \mathrm{e}^{-V(x)}\, \mathrm{d}x &= \int_{\overline{\Omega}} \Delta f(x) \mathrm{e}^{-V(x)} + \nabla f(x) \cdot \nabla \mathrm{e}^{-V(x)}\, \mathrm{d}x \\ &= - \int_{\partial\Omega} \big(\nabla f(x) \cdot n(x)\big) \mathrm{e}^{-V(x)}\, \mathcal{H}^{d-1}(\mathrm{d}x), \end{aligned}$$

where we recall that n is the *inward* pointing unit normal vector to $\partial\Omega$, which causes the sign-change from the usual formulation. In particular, if we assume that $\partial\Omega$ is C^2-smooth (meaning that it is the level-set of a C^2 function, i.e. $\partial\Omega = \{\Phi = 0\}$ for some $\Phi \in C^2(\mathbb{R}^d, \mathbb{R})$), the signed distance function introduced in (1.2) can be extended to a function $\varphi_\Omega \in C^2(\mathbb{R}^d, \mathbb{R})$ with $\nabla\varphi_\Omega(x) = n(x)$ for $x \in \partial\Omega$, and so

$$\begin{aligned} \frac{\mathbb{E}[\varphi_\Omega(\mathsf{X}^\Omega_t) - \varphi_\Omega(\mathsf{X}^\Omega_0)]}{t} &= \frac{1}{t} \mathbb{E}\bigg[\int_0^t \nabla\varphi_\Omega(\mathsf{X}^\Omega_s) \cdot n(\mathsf{X}^\Omega_s)\, \mathrm{d}\mathsf{L}^\Omega_s + \int_0^t \Delta\varphi_\Omega(\mathsf{X}^\Omega_s) - \nabla\varphi_\Omega(\mathsf{X}^\Omega_s) \cdot \nabla V(\mathsf{X}^\Omega_s)\, \mathrm{d}t \bigg] \\ &= \frac{1}{t} \mathbb{E}\bigg[\mathsf{L}^\Omega_t + \int_0^t \Delta\varphi_\Omega(\mathsf{X}^\Omega_s) - \nabla\varphi_\Omega(\mathsf{X}^\Omega_s) \cdot \nabla V(\mathsf{X}^\Omega_s)\, \mathrm{d}t \bigg], \end{aligned}$$

where we use that L^Ω is concentrated on the set $\{t \ge 0, \mathsf{X}^\Omega_t \in \partial\Omega\}$, and hence

$$\int_0^t \nabla\varphi_\Omega(\mathsf{X}^\Omega_s) \cdot n(\mathsf{X}^\Omega_s)\, \mathrm{d}\mathsf{L}^\Omega_s = \int_{\{0 \le s \le t, \mathsf{X}^\Omega_s \in \partial\Omega\}} n(\mathsf{X}^\Omega_s) \cdot n(\mathsf{X}^\Omega_s)\, \mathrm{d}\mathsf{L}^\Omega_s = \mathsf{L}^\Omega_t.$$

Rearranging and applying the ergodicity of X^Ω, we see that

$$\begin{aligned} \lim_{t\to\infty} \frac{1}{t} \mathbb{E}[\mathsf{L}^\Omega_t] &= \lim_{t\to\infty} \frac{1}{t} \mathbb{E}\bigg[\varphi_\Omega(\mathsf{X}^\Omega_t) - \varphi_\Omega(\mathsf{X}^\Omega_0) - \int_0^t \Delta\varphi_\Omega(\mathsf{X}^\Omega_s) - \nabla\varphi_\Omega(\mathsf{X}^\Omega_s) \cdot \nabla V(\mathsf{X}^\Omega_s)\, \mathrm{d}t \bigg] \\ &= - \int_{\overline{\Omega}} \big(\Delta\varphi_\Omega(x) - \nabla\varphi_\Omega(x) \cdot \nabla V(x)\big) \mathrm{e}^{-V(x)}\, \mathrm{d}x \\ &= \int_{\partial\Omega} \big(\nabla\varphi_\Omega(x) \cdot n(x)\big) \mathrm{e}^{-V(x)} \\ &= \int_{\partial\Omega} \mathrm{e}^{-V(x)}\, \mathcal{H}^{d-1}(\mathrm{d}x), \end{aligned}$$

which together with another application of the ergodicity of X^Ω yields (2.2).

Shape optimisation

Having now arrived at a deterministic expression for the objective function we have now transformed a somewhat vague problem into an explicit shape optimisation problem. There are a myriad of ways to approach such problems, however the one that we take in our case is to parametrise the problem, transforming it from a shape optimisation problem to simply a problem of optimising a function over $\mathbb{R}^n$ for some $n \in \mathbb{N}$. In particular, we first restrict ourselves to *strongly star-shaped sets* of the form

$$\Omega(r) = \{sq \mid q \in S^{d-1}, 0 < s < r(q)\} \quad \text{where} \quad r \in C^2\big(S^{d-1}, (0,\infty)\big),$$

and the idea is then that we can implicitly approximate $\Omega(r)$ by approximating r by piecewise multilinear interpolation. That is, we first fix $n \in \mathbb{N}$ points $q_1, q_2, \dots, q_n \in S^{d-1}$ and approximate $\Omega(r)$ by the star-shaped polytope spanned by the points $\{p_i\}_{i=1}^n = \{r(q_i)q_i\}_{i=1}^n$. Since the q_i's are fixed, each p_i is determined only by $r(q_i) \coloneqq r_i$, and so to any collection $r_1, r_2, \dots, r_n > 0$ we can assign a polytope $\widetilde{\Omega}(\boldsymbol{r})$, where $\boldsymbol{r} = \begin{bmatrix} r_1 & r_2 & \cdots & r_n \end{bmatrix}^\top$.

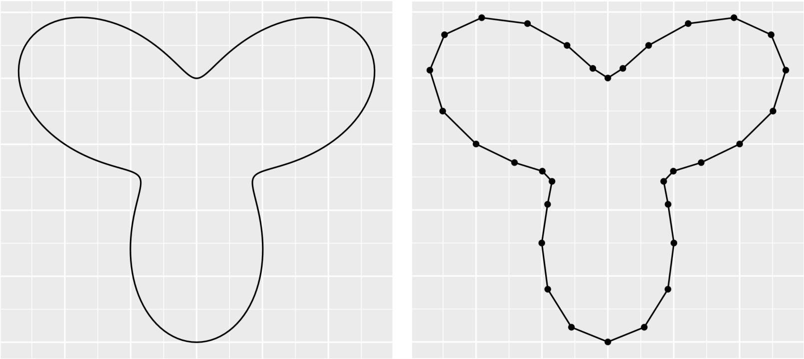

Figure 2.1: *Example of a star-shaped set* $\Omega(r)$ *with* $r(\theta) = 2 + \sin(3\theta)$ *and its approximation* $\widetilde{\Omega}(\boldsymbol{r})$.

The core assumption is then that

$$\min_{r \in C^2(S^{d-1},(0,\infty))} J\big(\Omega(r)\big) \approx \min_{\boldsymbol{r} \in (0,\infty)^n} \underbrace{J\big(\widetilde{\Omega}(\boldsymbol{r})\big)}_{\coloneqq \widetilde{J}(\boldsymbol{r})},$$

where the right hand side is a simple parametric problem which can be readily solved (at least approximately) using gradient-based methods. Parametrising the objective function in this way has the added benefit of making the integrals appearing in (2.2) (relatively) easy to calculate, as both $\widetilde{\Omega}(\boldsymbol{r})$ and its surface $\partial\widetilde{\Omega}(\boldsymbol{r})$ are now simply unions of d- and $(d-1)$-dimensional simplices, respectively, and so we need only know how to integrate over such sets. Especially the surface integral with respect to $\mathcal{H}^{d-1}$ becomes much more tractable, as we can simply employ an isomorphism Ψ to identify any facet F of $\partial\widetilde{\Omega}(\boldsymbol{r})$ with a subset $\Psi(F) \subseteq \mathbb{R}^{d-1}$ whence

$$\int_F g(x)\,\mathcal{H}^{d-1}(\mathrm{d}x) = \int_{\Psi(F)} g \circ \Psi(x)\,\mathrm{d}x.$$

Specifically, let S be one of the simplices making up $\widetilde{\Omega}(\boldsymbol{r})$, and let F be the facet of S opposite the origin. Then, there exists points $p_1, p_2, \dots, p_d$ (given as $p_j = r_{i_j} q_{i_j}$ for some subset $\{i_j\}_{i=1}^d \subseteq [n]$) such that

$$F = \left\{ \sum_{i=1}^d t_i p_i \mid t_i \geq 0, \sum_{i=1}^d t_i = 1 \right\} \quad \text{and} \quad S = \left\{ t_0 \sum_{i=1}^d t_i p_i \mid t_i \geq 0, \sum_{i=1}^d t_i = 1 \right\}.$$

Thus there is a natural bijection between F and the unit simplex $\{\boldsymbol{t} \in [0,1]^d \mid \mathbf{1}^\top \boldsymbol{t} = 1\}$, however to ease calculation and implementation details, we also compose this with the surjective map (bijective on the interior) from $(0,1)^{d-1}$ to the unit simplex, ultimately getting the bijection $\eta : (0,1)^{d-1} \to F^\circ$ given by

$$\eta(t_1, t_2, \dots, t_{d-1}) = (1-t_1)p_1 + t_1(1-t_2)p_2 + \cdots + \left(\prod_{i=1}^{d-2} t_i \right)(1-t_{d-1})p_{d-1} + \left(\prod_{i=1}^{d-1} t_i \right) p_d.$$

This of course also yields a mapping from $(0,1)^d$ to S given by $(s, \boldsymbol{t}) = s\eta(\boldsymbol{t})$. From here, finding the Jacobian and Gramian of these transforms is a fairly straight-forward exercise in calculus and linear algebra, and we arrive at the following:

Theorem 2.2: Let P denote the $d \times d$ matrix whose i'th column is p_i and P_{-1} the $d \times (d-1)$ matrix whise i'th column is $p_{i+1} - p_1$. Then we have the following for $g \in C(\mathbb{R}^d, \mathbb{R})$:

$$\int_S g(x)\,\mathrm{d}x = |P| \int_0^1 \int_{(0,1)^{d-1}} g\big(s\eta(\boldsymbol{t})\big)\,\psi(\boldsymbol{t}) r^{d-1} \mathrm{d}\boldsymbol{t}\,\mathrm{d}s$$

$$\int_F g(x)\,\mathcal{H}^{d-1}(\mathrm{d}x) = \sqrt{|P_{-1}^\top P_{-1}|} \int_{(0,1)^{d-1}} g\big(\eta(\boldsymbol{t})\big)\,\psi(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}$$

where $\psi(t_1, \dots, t_{d-1}) = \prod_{i=1}^{d-2} t_i^{d-1-i}$.

Note that the theorem as stated here differs from that in Article A – this is simply to avoid introducing too much superfluous notation at this stage, and the expressions are indeed equal. Using the same strategy we can also derive expressions for the derivatives of these integrals with respect to the vector lengths, and simple calculus yields an implementable expression for both $\widetilde{J}(\boldsymbol{r})$ and $\nabla\widetilde{J}(\boldsymbol{r})$, allowing for numerical optimisation using well-known gradient based methods.

Data-driven estimation

So far, we have assumed throughout that we know explicitly the dynamics of X, in particular the drift function $-\nabla V$ and hence the invariant density $\rho \propto \mathrm{e}^{-V}$. Of course, in real applications this is seldom the case, and we must first use observed data to infer such dynamics. To this, we let $\widehat{\rho}_T$ denote some estimator of ρ based on an observed (unreflected) trajectory $(\mathsf{X}_t)_{t\in[0,T]}$ for some $T > 0$. We also assume that the optimal reflection set Ω^* satisfies $B(0,\underline{\lambda}) \subseteq \Omega^* \subseteq B(0,\overline{\lambda})$ and $\mathcal{H}^{d-1}(\partial\Omega^*) \leq \Lambda$ for some constants $\underline{\lambda}, \overline{\lambda}, \Lambda$. This essentially means that the optimal set is neither too small, too large nor has a too irregular boundary. We then let $\rho_{\min} = \inf_{B(0,\overline{\lambda})} \rho > 0$ and set $\widehat{\rho}^*_T = \widehat{\rho}_T \vee \rho_{\min}$, and define the plug-in estimator of the objective function J as

$$\widehat{J}_T(\Omega) := \frac{1}{\int_\Omega \widehat{\rho}^*_T(x)\,\mathrm{d}x}\bigg(\int_\Omega c(x)\widehat{\rho}^*_T(x)\,\mathrm{d}x + \kappa\int_{\partial\Omega}\widehat{\rho}^*_T(x)\,\mathcal{H}^{d-1}(\mathrm{d}x)\bigg).$$

Assuming we can optimise this estimator, we then define our estimated optimal reflection set as $\widehat{\Omega}_T$ as a minimizer of $\widehat{J}_T$ over all sets satisfying the above assumptions. To show that $\widehat{\Omega}_T$ is indeed a good estimator of Ω^*, we first find that

$$\begin{aligned}|J(\Omega) - \widehat{J}_T(\Omega)| &= \frac{1}{\int_\Omega \widehat{\rho}^*_T(x)\,\mathrm{d}x}\bigg|J(\Omega)\int_\Omega \widehat{\rho}^*_T(x) - \rho(x)\,\mathrm{d}x \\ &\qquad + \int_\Omega c(x)\big(\rho(x) - \widehat{\rho}^*_T(x)\big)\,\mathrm{d}x + \kappa\int_{\partial\Omega}\rho(x) - \widehat{\rho}^*_T(x)\,\mathcal{H}(\mathrm{d}x)\bigg| \\ &\leq \frac{\int_{B(0,\overline{\lambda})} c(x)\,\mathrm{d}x + \kappa\Lambda}{\rho_{\min}\lambda_d(B(0,\underline{\lambda}))}\bigg(1 + \frac{\|\rho\|_\infty}{\rho_{\min}\lambda_d(B(0,\underline{\lambda}))}\bigg)\|\rho - \widehat{\rho}^*_T\|_\infty,\end{aligned}$$

where the constant in front is independent of Ω. Now, assuming $\mathsf{X}_0 \sim \mu$ for some measure $\mu \ll \pi$ with $\frac{\mathrm{d}\mu}{\mathrm{d}\pi} \in L^2(\pi)$, we find by Cauchy–Schwartz that

$$\begin{aligned}\mathbb{E}_\mu\Big[\sup_\Omega |J(\Omega) - \widehat{J}_T(\Omega)|\Big] &\leq C\mathbb{E}_\mu[\|\rho - \widehat{\rho}^*_T\|_\infty] \\ &\leq C\bigg\|\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\bigg\|_{L^2(\pi)}\mathbb{E}_\pi[\|\rho - \widehat{\rho}^*_T\|^2_\infty]^{1/2},\end{aligned}$$

and since Ω^* and $\widehat{\Omega}_T$ are minimizers of J and $\widehat{J}_T$, respectively, we have

$$|J(\Omega^*) - J(\widehat{\Omega}_T)| \leq J(\widehat{\Omega}_T) - J(\Omega^*) + \widehat{J}_T(\Omega^*) - \widehat{J}_T(\widehat{\Omega}_T) \leq 2\sup_\Omega |J(\Omega) - \widehat{J}_T(\Omega)|.$$

Combining this, we have that regardless of the initial distribution μ,

$$\mathbb{E}_\mu\big[|J(\Omega^*) - J(\widehat{\Omega}_T)|\big] \lesssim_\mu \mathbb{E}_\pi[\|\rho - \widehat{\rho}^*_T\|^2_\infty]^{1/2},$$

that is, we can approximate Ω^* (wrt. J) as precisely as we can approximate the invariant density ρ (wrt. $\|\cdot\|_\infty$) under π. In particular, assuming that $\rho \in \boldsymbol{\mathcal{H}(\beta, L)}$ cf. Definition 1.7 as well as some regularity conditions on X which we will omit here for brevity, following the method prescribed in [111] and employing a kernel density estimator as the approximation $\widehat{\rho}_T$, we arrive at the following theorem

Theorem 2.3: Assuming the above, we have for any measure $\mu \ll \pi$ with $\|\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\|_{L^2(\pi)} < \infty$

$$\mathbb{E}_\mu\Big[\big|J(\Omega^*) - J(\widehat{\Omega}_T)\big|\Big] \lesssim \begin{cases} \frac{\log T}{\sqrt{T}}, & \text{if } d = 2 \\ \left(\frac{\log T}{\sqrt{T}}\right)^{\frac{\overline{\boldsymbol{\beta}}+1}{2\overline{\boldsymbol{\beta}}+1+d-2}}, & \text{if } d \geq 3 \end{cases}$$

On one hand, this result implies that if we have access to another (unreflected) process, estimating the optimal reflection boundary in an online fashion, the regret decays according to the minimax rate. It also provides a bound for the regret faced if employing an explore-then-commit strategy. However, it is unclear how one would implement an exploration-exploitation strategy. In the one-dimensional setting, this is easy, as one can choose a fixed point $a \in \mathbb{R}$ that both the controlled uncontrolled process will hit almost surely. One can then “glue together” two exploration periods by starting and ending the exploitation periods at stopping times where the controlled process hits a.

This is not possible in higher dimensions, where points are of co-dimension at least 2 and hence will not be hit in general. In lieu of this, we describe for sake of completeness a scheme of considering exponentially increasing lengths of exploration and exploitation periods, yielding an expected regret rate of $O\big((\frac{(\log T)^2}{T})^{\frac{1}{3}}\big)$ for $d = 2$ and $O\big((\frac{\log T}{T})^{\frac{\overline{\boldsymbol{\beta}}+1}{3\overline{\boldsymbol{\beta}}+1}+d-2}\big)$ for $d \geq 3$. While suboptimal compared to the rate achieved in 2.3, the loss in rate is comparable to the one-dimensional case cf. [28].

2.2 Article B: Statistical guarantees for denoising reflected diffusion models

In this article, we take inspiration from the emerging field of generative AI. Here, given a set of samples, e.g. images, we want to generate new samples that in some sense resemble the originals. At a high level overview, one modern way to do this is by first adding noise to the samples in such a way that the noise can be removed. Then, when enough noise is added to the samples, they become near indistinguishable from pure noise, and so to generate new samples, we first generate new pure noise and run the de-noising procedure on this pure noise. To motivate how this is done mathematically, consider the simulation of a diffusion process $\mathsf{X} = (\mathsf{X}_t)_{t\geq 0}$ in Figure 2.3. Here, we note two important features. First, it seems that the process settles into the same distribution regardless of the initial point X_0. Secondly, looking at the process in reverse, that is going from right to left, it still seems to “behave” like a diffusion process, albeit with different dynamics forcing it into one specific point. These are the key properties that allow us to add

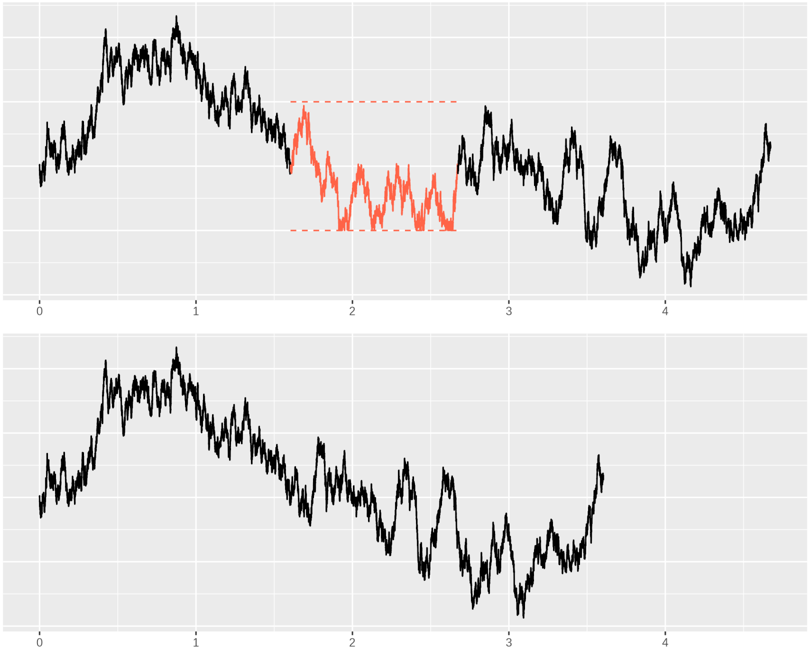

Figure 2.2: *Example of exploration periods being "glued together" to form a single trajectory*

noise to samples in such a way that the perturbed samples converge to pure noise and such that we may reverse the noising process.

To be exact, we assume that we are given samples $\mathsf{Y}_1, \mathsf{Y}_2, \dots, \mathsf{Y}_n \overset{\text{i.i.d.}}{\sim} \mu$ for some target measure μ on $\mathbb{R}^D$[1] – one concrete example is the CIFAR-10 dataset consisting of 32×32 colour images, where each $\mathsf{Y}_i \in \mathbb{R}^D$ with $D = 3 \cdot 32^2$ represents one image and μ the distribution of all such images on $\mathbb{R}^D$ – and our goal is then to generate new samples $\widehat{\mathsf{X}}$ such that the distribution of $\widehat{\mathsf{X}}$ in a fitting way resembles μ. In the method known as *de-noising diffusion models* (DDMs) we consider first the process $\mathsf{X} = (\mathsf{X}_t)_{t \geq 0}$ solving

$$\mathrm{d}\mathsf{X}_t = b(t, \mathsf{X}_t)\,\mathrm{d}t + \sigma(t, \mathsf{X}_t)\,\mathrm{d}\mathsf{B}_t, \qquad \mathsf{X}_0 \sim \mu$$

for suitable $b : [0, \infty) \times \mathbb{R}^D \to \mathbb{R}^D$ and $\sigma : [0, \infty) \times \mathbb{R}^D \to \mathbb{R}^{D \times D}$ of our choosing and a D-dimensional Brownian motion $\mathsf{B} = (\mathsf{B}_t)_{t \geq 0}$. This process, known as the *forward process*, will be our way of adding noise to the samples. As alluded to above, to remove noise, we then run this process in reverse. That is, for some fixed $\overline{T} > 0$ of our choosing, we consider the *backward*

---

[1] We denote here the dimension as a capital D to emphasise the potentially large dimensionality of the target distribution – this will become more important in the next article.

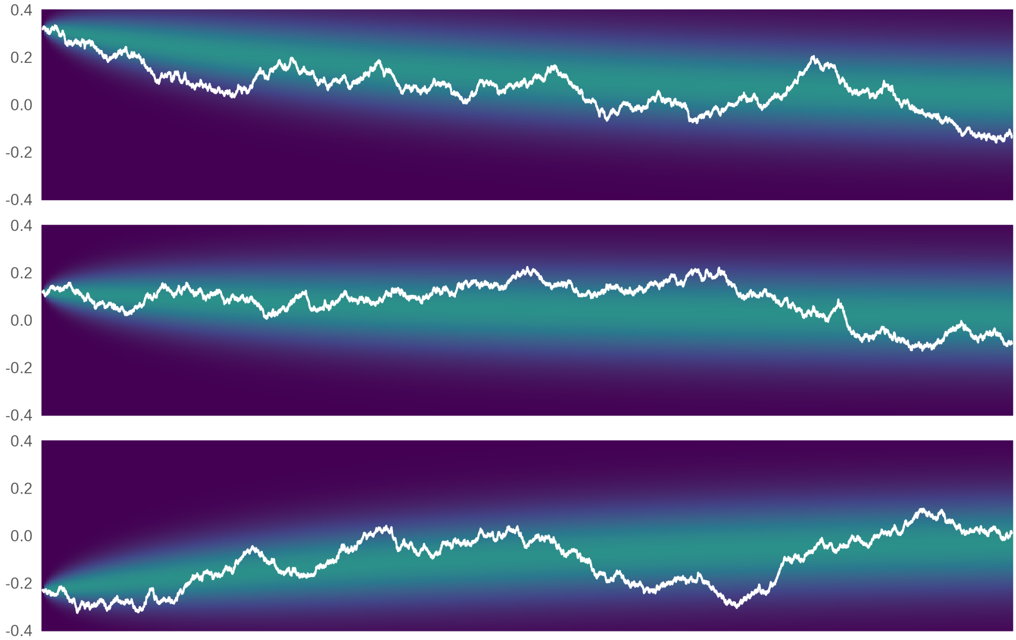

Figure 2.3: *Simulated paths of an Ornstein–Uhlenbeck process with three different initial points superimposed on a heat-map of their (scaled) densities.*

process $\overleftarrow{\mathsf{X}} = (\overleftarrow{\mathsf{X}}_t)_{t\in[0,\overline{T}]}$ given by $\overleftarrow{\mathsf{X}}_t = \mathsf{X}_{\overline{T}-t}$. Letting p_t denote the density of X_t for $t > 0$, it was then shown in [3] that under sufficient regularity conditions, there exists another Brownian motion $\overline{\mathsf{B}} = (\overline{\mathsf{B}}_t)_{t\geq 0}$ such that $\overleftarrow{\mathsf{X}}$ solves

$$\mathrm{d}\overleftarrow{\mathsf{X}}_t = -\overline{b}(\overline{T}-t, \overleftarrow{\mathsf{X}}_t)\,\mathrm{d}t + \sigma(\overline{T}-t, \overleftarrow{\mathsf{X}}_t)\,\mathrm{d}\overline{\mathsf{B}}_t, \qquad \overleftarrow{\mathsf{X}}_0 \sim p_{\overline{T}}$$

where

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

In particular, if $\sigma(t,x) = \gamma(t)I_D$ for some scalar function $\gamma : [0,\infty) \to [0,\infty)$, $\overline{b}$ simplifies significantly to

$$\overline{b}(t,x) = b(t,x) - \gamma(t)^2 \frac{\nabla p_t(x)}{p_t(x)} = b(t,x) - \gamma(t)^2 \nabla \log p_t(x)$$

At this point, if we knew p_t for $t \in (0,\overline{T}]$, we could simply use this process to generate our new sample $\widehat{\mathsf{X}}$. Unfortunately, p_t depends implicitly on the unknown distribution μ, and so we must approximate. In particular, given an estimator $s(x,t)$ of the *score function* $\nabla \log p_t(x)$ and assuming that $\mathsf{X}_t \overset{\sim}{\longrightarrow} \nu$ for some known distribution ν as $t \to \infty$, we consider instead the approximation $\widehat{\mathsf{X}}^s = (\widehat{\mathsf{X}}^s_t)_{t\in[0,\overline{T}]}$ solving

$$\mathrm{d}\widehat{\mathsf{X}}^s_t = -\big(b(\overline{T}-t,\widehat{\mathsf{X}}^s_t) - \gamma(\overline{T}-t)^2 s(\widehat{\mathsf{X}}^s_t,\overline{T}-t)\big)\,\mathrm{d}t + \gamma(\overline{T}-t)\,\mathrm{d}\overline{\mathsf{B}}_t, \qquad \widehat{\mathsf{X}}^s_0 \sim \nu.$$

Then, for some chosen $\underline{T} > 0$ and appropriate estimator s, we set $\widehat{\mathsf{X}} = \widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}$. This method is quite effective, having been shown to be nearly mini-max optimal as an estimator of μ, see e.g. [6, 89, 118]. However, it often suffers from an inherent limitation of SDEs such as the above, namely that the generated samples are necessarily unbounded. This is problematic when we have à priori knowledge that μ has a compact support $\Omega \subset \mathbb{R}^D$. For instance, in the setting of image generation, it is natural to consider an image as a vector in $[0,1]^D$ where each entry represents a pixel brightness value as a proportion. What should values outside of $[0,1]$ represent in this case? A common workaround used in practice is then that of thresholding, where samples outside of Ω are simply projected onto Ω as a final step. However, this pushes all of the probability mass outside of Ω onto the boundary $\partial\Omega$ leading to an overrepresentation of samples on $\partial\Omega$. A better solution would be to prevent the process $\widehat{\mathsf{X}}^s$ from ever leaving the support set Ω. This is the key motivation behind using reflected diffusions rather than unreflected ones, leading to the method of de-noising *reflected* diffusion models. Here, the procedure is much the same, except we consider as our forward model a reflected SDE. In particular, in our setting we consider for some compact, smooth and convex reflection set $\Omega \supseteq \operatorname{supp}\mu$ and chosen $f \in C^\infty(\overline{\Omega})$ with $f \geq f_{\min} > 0$ the forward process $\mathsf{X} = (\mathsf{X}_t)_{t\geq 0}$ solving

$$\mathrm{d}\mathsf{X}_t = \nabla f(\mathsf{X}_t)\,\mathrm{d}t + \sqrt{2f(\mathsf{X}_t)}\,\mathrm{d}\mathsf{B}_t + \nu(\mathsf{X}_t)\,\mathrm{d}\mathsf{L}_t, \qquad \mathsf{X}_0 \sim \mu,$$

where $\nu = fn$ is the scaled inward pointing normal vector ensuring that (1.6) holds. Like in the unreflected setting, it was shown in [18] that there exists another Brownian motion $\overline{\mathsf{B}} = (\overline{\mathsf{B}}_t)_{t\geq 0}$ such that the backward process $\overleftarrow{\mathsf{X}} = (\overleftarrow{\mathsf{X}}_t)_{t\in[0,\overline{T}]}$ given by $\overleftarrow{\mathsf{X}}_t = \mathsf{X}_{\overline{T}-t}$ solves

$$\mathrm{d}\overleftarrow{\mathsf{X}}_t = -\big(\nabla f(\overleftarrow{\mathsf{X}}_t) - 2f(\overleftarrow{\mathsf{X}}_t)\nabla\log p_{\overline{T}-t}(\overleftarrow{\mathsf{X}}_t)\big)\,\mathrm{d}t + \sqrt{2f(\overleftarrow{\mathsf{X}}_t)}\,\mathrm{d}\overline{\mathsf{B}}_t + \nu(\overleftarrow{\mathsf{X}}_t)\,\mathrm{d}\overline{\mathsf{L}}_t, \qquad \overleftarrow{\mathsf{X}}_0 \sim p_{\overline{T}},$$

where $\overline{\mathsf{L}}_t = \mathsf{L}_{\overline{T}} - \mathsf{L}_{\overline{T}-t}$ is the local time of $\overleftarrow{\mathsf{X}}_t$ at $\partial\Omega$. Since one can show that $\mathsf{X}_t \overset{\sim}{\longrightarrow} \mathrm{Unif}(\Omega)$ as $t \to \infty$, we then for some score approximation s take as our sample $\widehat{\mathsf{X}} = \widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}$, where

$$\mathrm{d}\widehat{\mathsf{X}}^s_t = -\big(\nabla f(\widehat{\mathsf{X}}^s_t) - 2f(\widehat{\mathsf{X}}^s_t)s(\widehat{\mathsf{X}}^s_t,\overline{T}-t)\big)\,\mathrm{d}t + \sqrt{2f(\widehat{\mathsf{X}}^s_t)}\,\mathrm{d}\overline{\mathsf{B}}_t + \nu(\widehat{\mathsf{X}}^s_t)\,\mathrm{d}\overline{\mathsf{L}}_t, \qquad \widehat{\mathsf{X}}^s_0 \sim \mathrm{Unif}(\Omega).$$

Beyond ensuring an explicit and easy-to-sample-from stationary distribution, this choice of forward model also provides us with a semi-explicit transition kernel through spectral decomposition. In particular, there exists orthonormal eigenpairs $(\ell_j, e_j)_{j\in\mathbb{N}_0}$ of the non-negative operator $-\nabla\cdot f\nabla$ such that the transition density q_t is given by

$$q_t(x,y) = \sum_{j\in\mathbb{N}_0} \mathrm{e}^{-t\ell_j} e_j(x)e_j(y). \tag{2.3}$$

Furthermore, these eigenpairs satisfy $0 = \ell_0 < \ell_1 \leq \ell_2 \leq \cdots$ with $\ell_j \asymp j^{\frac{2}{D}}$, $e_0 \equiv \lambda_D(\Omega)^{-\frac{D}{2}}$, $\|e_j\|_{H^k} \lesssim \ell_j^{k/2} \asymp j^{k/D}$ for $j \in \mathbb{N}$ and $\|e_j\|_\infty \lesssim j^\tau$ for $\tau < 1/2$; see [85] for details.

Error decomposition

Intuitively, there are three main sources of error in our estimate, namely the ones stemming from early stopping (i.e. considering $\widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}$ rather than $\widehat{\mathsf{X}}^s_{\overline{T}}$), from starting in the stationary distribution and from using an approximation s of the true score. Using total variation as our measure of error, this is reflected by the triangle inequality in that

$$\mathrm{TV}(\mu, \widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}) \leq \mathrm{TV}(\mu, \mathsf{X}_{\underline{T}}) + \mathrm{TV}(\mathsf{X}_{\underline{T}}, \widehat{\mathsf{X}}^{s_0}_{\overline{T}-\underline{T}}) + \mathrm{TV}(\widehat{\mathsf{X}}^{s_0}_{\overline{T}-\underline{T}}, \widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}), \tag{2.4}$$

where $s_0(x, t) = \nabla \log p_t(x)$ denotes the true score. The fact that the second term on the right hand side corresponds to the error from starting in the stationary distribution stems from the fact that $\mathsf{X}_{\underline{T}} = \overleftarrow{\mathsf{X}}_{\overline{T}-\underline{T}}$ and that the only difference between $\overleftarrow{\mathsf{X}}$ and $\widehat{\mathsf{X}}^{s_0}$ is in their initial distribution. To further analyse these error terms, we make the following assumption throughout on μ:

($\mathcal{H}0$) The target measure μ is absolutely continuous (wrt. the D-dimensional Lebesgue measure) with density $p_0 \in H^\alpha$ where $\alpha \in \mathbb{N} \cap (\frac{D}{2}, \infty)$ and $p_0 \geq p_{\min}$ for a constant $p_{\min} > 0$.

Since $\alpha > \frac{D}{2}$, we have by the Sobolev embedding theorem that p_0 is β-Hölder continuous for some $\beta \in (0, 1]$, i.e. there exists a constant $c_\beta > 0$ such that $|p_0(x) - p_0(y)| \leq c_\beta |x - y|^\beta$ for all $x, y \in \Omega$. Also, by symmetry of q_t, we have

$$p_t(x) = \int p_0(y) q_t(y, x)\,\mathrm{d}y = \int p_0(y) q_t(x, y)\,\mathrm{d}y = \mathbb{E}[p_0(\mathsf{X}_t) \mid \mathsf{X}_0 = x],$$

and so

$$|p_t(x) - p_0(x)| = |\mathbb{E}[p_0(\mathsf{X}_t) - p_0(\mathsf{X}_0) \mid \mathsf{X}_0 = x]| \leq c_\beta \mathbb{E}[|\mathsf{X}_t - \mathsf{X}_0|^\beta \mid \mathsf{X}_0 = x].$$

By our choice of forward model, we have by e.g. [29] that $q_t(x, y) \lesssim t^{-\frac{D}{2}} \mathrm{e}^{-\frac{|x-y|^2}{Ct}}$ for some $C > 0$, implying that

$$\mathbb{E}[|\mathsf{X}_t - \mathsf{X}_0|^\beta \mid \mathsf{X}_0 = x] = \int |x - y|^\beta q_t(x, y)\,\mathrm{d}y \lesssim t^{\frac{\beta}{2}}.$$

Since $\mathrm{TV}(\mu, \mathsf{X}_t) = \frac{1}{2}\|p_0 - p_t\|_{L^1}$, this yields the following lemma, controlling the first term of (2.4):

Lemma 2.4: Under ($\mathcal{H}0$) we have

$$\mathrm{TV}(\mu, \mathsf{X}_{\underline{T}}) \lesssim \underline{T}^{\frac{\beta}{2}}.$$

As for the second term of (2.4), we have that since $\widehat{\mathsf{X}}^{s_0}$ and $\overleftarrow{\mathsf{X}}$ follow the same Markovian dynamics, their total variation distance is decreasing in t, and hence

$$\mathrm{TV}(\mathsf{X}_{\underline{T}}, \widehat{\mathsf{X}}^{s_0}_{\overline{T}-\underline{T}}) = \mathrm{TV}(\overleftarrow{\mathsf{X}}_{\overline{T}-\underline{T}}, \widehat{\mathsf{X}}^{s_0}_{\overline{T}-\underline{T}}) \leq \mathrm{TV}(\overleftarrow{\mathsf{X}}_0, \widehat{\mathsf{X}}^{s_0}_0) = \mathrm{TV}(\mathsf{X}_{\overline{T}}, \mathrm{Unif}(\Omega)).$$

Also, noting that by the spectral decomposition (2.3) we have

$$p_t(x) = \sum_{j\in\mathbb{N}_0} \mathrm{e}^{-t\ell_j}\langle p_0, e_j\rangle_{L^2} e_j(x) = \frac{1}{\lambda_D(\Omega)} + \sum_{j\in\mathbb{N}} \mathrm{e}^{-t\ell_j}\langle p_0, e_j\rangle_{L^2} e_j(x),$$

it follows that

$$\Big\| p_t - \frac{1}{\lambda_D(\Omega)}\Big\|_{L^2}^2 = \Big\| \sum_{j\in\mathbb{N}} \mathrm{e}^{-t\ell_j}\langle p_0, e_j\rangle_{L^2} e_j\Big\|_{L^2}^2 = \sum_{j\in\mathbb{N}} \mathrm{e}^{-2t\ell_j}\langle p_0, e_j\rangle_{L^2}^2 \leq \|p_0\|_{L^2}^2 \mathrm{e}^{-2t\ell_1}.$$

Finally, since by the Cauchy–Schwarz inequality

$$\mathrm{TV}(\mathsf{X}_{\overline{T}}, \mathrm{Unif}(\Omega)) = \frac{1}{2}\Big\| p_{\overline{T}} - \frac{1}{\lambda_D(\Omega)}\Big\|_{L^1} \leq \frac{\sqrt{\lambda_D(\Omega)}}{2}\Big\| p_{\overline{T}} - \frac{1}{\lambda_D(\Omega)}\Big\|_{L^2},$$

we arrive at the next lemma bounding the second term of (2.4):

Lemma 2.5: Under $(\mathcal{H}0)$ we have

$$\mathrm{TV}(\mathsf{X}_{\underline{T}}, \widehat{\mathsf{X}}^{s_0}_{\overline{T}-\underline{T}}) \lesssim \mathrm{e}^{-\ell_1\overline{T}}.$$

Already we see that simply choosing $\underline{T} \asymp n^{-\frac{2\alpha}{\beta(2\alpha+D)}}$ and $\overline{T} = \frac{\alpha}{\ell_1(2\alpha+D)}\log n$ that the first two terms of (2.4) will be of order $O(n^{-\frac{\alpha}{2\alpha+D}})$, i.e. minimax optimal.

Score approximation

Controlling the third term of (2.4) is, however, a lot more involved. First, we must settle on a score approximation s, to which we note by Pinsker's inequality as well as Girsanov's theorem

$$\mathrm{TV}(\widehat{\mathsf{X}}^{s_0}_{\overline{T}-\underline{T}}, \widehat{\mathsf{X}}^{s}_{\overline{T}-\underline{T}}) \leq \sqrt{\frac{1}{2}\mathrm{KL}(\widehat{\mathsf{X}}^{s_0}_{\overline{T}-\underline{T}} \,\|\, \widehat{\mathsf{X}}^{s}_{\overline{T}-\underline{T}})} \leq \sqrt{\frac{\|f\|_\infty}{2}\int_{\underline{T}}^{\overline{T}} \mathbb{E}\big[|s(\mathsf{X}_t, t) - s_0(\mathsf{X}_t, t)|^2\big]\,\mathrm{d}t}. \tag{2.5}$$

As such, we must find an estimator which minimizes the L^2-distance to the true score under the unknown measure μ. This of course raises two main problems, as we do not have access to the measure μ whence we cannot take the expectation on the right hand side, but also we do not have access to the true score s_0 to even evaluate the difference at a single point. The key here is in the equivalence between this explicit score matching and the so-called *denoising score matching*, since we have

$$\int_{\underline{T}}^{\overline{T}} \mathbb{E}\left[|s(\mathsf{X}_t,t)-s_0(\mathsf{X}_t,t)|^2\right]\mathrm{d}t = \mathbb{E}\left[\int_{\underline{T}}^{\overline{T}} \mathbb{E}\left[|s(\mathsf{X}_t,t)-\nabla\log q_t(\mathsf{X}_0,\mathsf{X}_t)|^2 \mid \mathsf{X}_0\right]\mathrm{d}t\right]+C,$$

where

$$C=\int_{\underline{T}}^{\overline{T}} \mathbb{E}\left[|s_0(\mathsf{X}_t,t)|^2\right]\mathrm{d}t-\mathbb{E}\left[\int_{\underline{T}}^{\overline{T}} \mathbb{E}\left[|\nabla\log q_t(\mathsf{X}_0,\mathsf{X}_t)|^2 \mid \mathsf{X}_0\right]\mathrm{d}t\right],$$

is a constant independent of s. Thus, since q_t is determined only by our choice of forward model and is independent of μ, setting

$$L_s(x)=\int_{\underline{T}}^{\overline{T}} \mathbb{E}\left[|s(\mathsf{X}_t,t)-\nabla\log q_t(\mathsf{X}_0,\mathsf{X}_t)|^2 \mid \mathsf{X}_0=x\right]\mathrm{d}t,$$

minimising the L^2-distance is equivalent to minimising $\mathbb{E}[L_s(\mathsf{X}_0)]$, and by the law of large numbers, this is approximately given as

$$\mathbb{E}[L_s(\mathsf{X}_0)]\approx\frac{1}{n}\sum_{i=1}^{n}L_s(\mathsf{Y}_i).$$

This motivates our choice of score estimator as follows: given some hypothesis class of functions (later to be chosen as a certain class of neural networks) $\mathcal{S}$, we choose

$$\widehat{s}_n\in\operatorname*{argmin}_{s\in\mathcal{S}}\frac{1}{n}\sum_{i=1}^{n}L_s(\mathsf{Y}_i).$$

As such, the quality of our estimator is deeply tied to our choice of hypothesis class $\mathcal{S}$. On one hand, if we choose $\mathcal{S}$ "too small", we may not be able to capture the details of the true score s_0, yielding a high estimation error. On the other hand, if we choose $\mathcal{S}$ "too large", the minimiser $\widehat{s}_n$ might overfit the data $\mathsf{Y}_1,\mathsf{Y}_2,\dots,\mathsf{Y}_n$, yielding a high generalisation error. This is made explicit in [89]:

Theorem 2.6: Set $\mathcal{L}(\mathcal{S})=\{L_s\mid s\in\mathcal{S}\}$ and suppose that $\sup_{s\in\mathcal{S}\cup\{s_0\}}\|L_s\|_\infty\le C(\mathcal{L})$. Then, for appropriate $\delta>0$

$$\mathbb{E}\left[\int_{\underline{T}}^{\overline{T}} \mathbb{E}\left[|\widehat{s}_n(\mathsf{X}_t,t)-s_0(\mathsf{X}_t,t)|^2 \mid \mathsf{Y}_1,\mathsf{Y}_2,\dots,\mathsf{Y}_n\right]\mathrm{d}t\right]$$
$$\lesssim\inf_{s\in\mathcal{S}}\int_{\underline{T}}^{\overline{T}} \mathbb{E}\left[|s(\mathsf{X}_t,t)-s_0(\mathsf{X}_t,t)|^2\right]\mathrm{d}t+\frac{C(\mathcal{L})}{n}\log\mathcal{N}(\mathcal{L},\|\cdot\|_\infty,\delta)+\delta.$$

Here, $\mathcal{N}(\mathcal{L}, \|\cdot\|_\infty, \delta)$ denotes the covering number of $\mathcal{L}$ with respect to $\|\cdot\|_\infty$ and radius δ, i.e.

$$\mathcal{N}(\mathcal{L}, \|\cdot\|_\infty, \delta) = \inf\left\{ n \in \mathbb{N}_0 \mid \exists l_1, l_2, \dots, l_n \in \mathcal{L} \,:\, \sup_{l \in \mathcal{L}} \min_{i \in [n]} \|l - l_i\|_\infty \leq \delta \right\}.$$

By our choice of forward model, imposing some simple growth conditions on $s \in \mathcal{S}$, we can turn any covering of $\mathcal{S}$ into one of $\mathcal{L}(\mathcal{S})$ as well as turn bounds on $\mathcal{S}$ into ones on $\mathcal{L}$ – we will skip the details here and simply state the following lemma:

Lemma 2.7: If for all $t > 0$ we have $\sup_{s \in \mathcal{S}} \|s(\cdot, t)\|_\infty \leq \frac{C(\mathcal{S})}{\sqrt{t \wedge 1}}$ and $\underline{T} \leq 1 \leq \overline{T}$, then there exists a constant $c > 0$ such that

$$\mathcal{N}(\mathcal{L}, \|\cdot\|_\infty, \delta) \leq \mathcal{N}\left(\mathcal{S}, \|\cdot\|_\infty, \frac{\delta}{cC(\mathcal{S})\overline{T}}\right)$$

and

$$\sup_{s \in \mathcal{S} \cup \{s_0\}} \|L_s\|_\infty \lesssim (C(\mathcal{S})^2 \vee 1)(\log \underline{T}^{-1} + \overline{T}).$$

Neural network approximations

For our hypothesis class we will choose a restriction of the neural network class $\Phi(L, W, S, B)$ introduced in 1.9 that satisfies the growth constraint mentioned above, i.e. we set

$$\mathcal{S} = \mathcal{S}(L, W, S, B) := \left\{ \varphi \in \Phi(L, W, S, B) \mid |\varphi(x, t)| \leq \frac{C}{\sqrt{t \wedge 1}} \right\}$$

for some universal constant $C > 0$. This specific parametrisation enjoys not only very well-documented approximation properties, but also a clear bound on its covering number, in that we have cf. a straightforward modification of [89, Lemma C.2]

$$\log \mathcal{N}(\mathcal{S}, \|\cdot\|_\infty, \delta) \lesssim L(n)S(n) \log\left(\delta^{-1} L(n) \|W(n)\|_\infty B(n)\right).$$

Combining this, Lemma 2.7, Theorem 2.6 and (2.5) it follows that if we construct an estimator $s^* \in S(L(n), W(n), S(n), B(n))$ with

$$\int_{\underline{T}}^{\overline{T}} \mathbb{E}\left[|s^*(\mathrm{X}_t, t) - s_0(\mathrm{X}_t, t)|^2\right] \mathrm{d}t \lesssim n^{-\frac{2\alpha}{2\alpha + D}} (\log n)^3$$

and $|s^*(x, t)| \leq \frac{C}{\sqrt{t \wedge 1}}$, where

$$\begin{aligned} L(n) &\lesssim \log n \log\log n \\ \|W(n)\|_\infty &\lesssim n^{\frac{D}{2\alpha + D}} (\log n)^2 \\ S(n) &\lesssim n^{\frac{D}{2\alpha + D}} (\log n)^3 \\ B(n) &\lesssim n^{\frac{2\alpha}{\beta(2\alpha + D)}}, \end{aligned}$$

then

$$\mathbb{E}\big[\mathrm{TV}(\widehat{\mathsf{X}}^{s_0}_{\overline{T}-\underline{T}}, \widehat{\mathsf{X}}^{\widehat{s}_n}_{\overline{T}-\underline{T}})\big] \lesssim n^{-\frac{\alpha}{2\alpha+D}} (\log n)^3 (\log\log n)^{1/2}.$$

Here, the exact sizes of $L(n)$, $W(n)$ and $B(n)$ are largely irrelevant beyond that they should be polynomial in $\log n$, n and n, respectively, as they will then only contribute logarithmically to the covering number and hence the total variation. Likewise, the log-factor of $S(n)$ is also not important, and the main takeaway of the above limitations should be that $S(n)$ should be of order $O(n^{\frac{D}{2\alpha+D}})$ up to logarithmic factors. Presuming for now the existence of such an estimator s^*, this together with Lemmas 2.4 and 2.5 yields the following main theorem:

Theorem 2.8: Assuming $(\mathcal{H}0)$ and choosing $\underline{T} \asymp n^{-\frac{2\alpha}{\beta(2\alpha+D)}}$, $\overline{T} = \frac{\alpha}{\ell_1(2\alpha+D)} \log n$ and $\mathcal{S}$ as above, we have

$$\mathbb{E}\big[\mathrm{TV}(\mu, \widehat{\mathsf{X}}^{\widehat{s}_n}_{\overline{T}-\underline{T}})\big] \lesssim n^{-\frac{\alpha}{2\alpha+D}} (\log n)^3 (\log\log n)^{1/2}.$$

That is, up to small logarithmic factors this approximation is minimax-optimal. In order to actually construct such an estimator, we first note that by (2.3) we have

$$p_t(x) = \int_\Omega p_0(y) q_t(y,x)\,\mathrm{d}y = \sum_{j=0}^{\infty} \mathrm{e}^{-\ell_j t} \langle e_j, p_0\rangle_{L^2} e_j(x).$$

The first step to constructing s^* is then to truncate this sum by considering the function $h_N(x,t) = \sum_{j=0}^{N} \mathrm{e}^{-\ell_j t} \langle e_j, p_0\rangle_{L^2} e_j(x)$ and approximating instead $\nabla \log h_N$ for $N \in \mathbb{N}$. Some rather technical calculations (which we will skip here for the sake of brevity) bound the error in using this approximation, ultimately yielding that

$$\int_{\underline{T}}^{\overline{T}} \mathbb{E}\Big[\Big| s_0(\mathsf{X}_t, t) - \frac{\nabla h_N(\mathsf{X}_t, t)}{h_N(\mathsf{X}_t, t) \vee p_{\min}} \Big|^2\Big] \lesssim \log \underline{T}^{-1} N^{-\frac{2\alpha}{D}},$$

suggesting that we should aim to approximate h_N and ∇h_N at a rate of $N^{-\frac{2\alpha}{D}}$ for $N \asymp n^{\frac{D}{2\alpha+D}}$. Now, by [114] there exists neural networks $\varphi_j \in \widetilde{\Phi}(\log N, N, N\log N, N^{1/D})$ such that $\|\varphi_j - e_j\|_{L^2} \lesssim N^{-\frac{\alpha}{D}}$. An immediate idea is then to simply set[2]

$$\varphi_{h_N}(x,t) = \sum_{j=0}^{N} \mathrm{e}^{-\ell_j t} \langle e_j, p_0\rangle_{L^2} \varphi_j(x).$$

The problem with this approach is that the sparsity constraint of a sum of neural networks scales as the sum of the individual sparsity constraints, and hence the sparsity constraint here would be of order $N(N\log N) \gtrsim n^{\frac{2D}{2\alpha+D}} > n^{\frac{D}{2\alpha+D}}$. Instead we utilise the explicit and smooth dependence

---

[2] Here and in the following, we assume for sake of notational clarity that "elementary" functions such as exponentials, polynomials, division and multiplication are available in the construction of φ^*; in reality they should all be approximated by neural networks as in Article B.

on t in h_N to interpolate between fixed-time approximations of $x \mapsto h_N(x,t)$. That is, for some fixed time points $\{t_i\}_{i=1}^K$ we approximate for each $i \in [K]$ the function $x \mapsto h_N(x, t_i)$ by some φ_i using the above result, and then interpolate between them using polynomial interpolation, i.e. setting

$$\varphi_{h_N}(x,t) = \sum_{i=1}^K \varphi_i(x) p_i(t), \quad \text{where} \quad p_i(t) = \prod_{j \neq i} \frac{t - t_j}{t_i - t_j}.$$

Since the function $t \mapsto h_N(x,t)$ is entire as the finite sum of entire functions, its corresponding interpolating polynomial converges exponentially fast in K, provided the correct time points $\{t_i\}_{i=1}^K$. In particular, one can show that on any interval $[t_0, 4t_0]$, there exists time points $\{t_i\}_{i=1}^K$ such that

$$\forall t \in \left[t_0, 4t_0\right] : \left\| h_N(\cdot, t) - \sum_{i=1}^K h_N(\cdot, t_i) p_i(t) \right\|_{L^2}^2 \lesssim N 3^{-2K},$$

whence we need only take $K \asymp \log N \asymp \log n$. Finally, to extend this to all of $[\underline{T}, \overline{T}]$, we simply employ the above strategy on intervals of the form $[2^{m-1}\underline{T}, 2^{m+1}\underline{T}]$ for $m = 1, 2, \ldots, M$ such that $2^{M+1} \geq \overline{T}$ and combine all of these into one network. Since there are $M \asymp \log \frac{\overline{T}}{\underline{T}} \asymp \log n$ such networks, the final sparsity of the network ends up being of order $O(N(\log N)^3) = O(n^{\frac{D}{2\alpha+D}}(\log n)^3)$ as desired. The exact same method can be applied to approximate ∇h_N, and from there piecing these together into an approximation of $\frac{\nabla h_N}{h_N \vee p_{\min}}$ is fairly elementary, finally yielding the desired estimator s^* and proving Theorem 2.8.

2.3 Article C: Reflected diffusion models adapt to low-dimensional data

In the previous article, we showed that using de-noising reflected diffusion models to generate samples from a certain class of probability distributions is nearly minimax-optimal in total variation. However, as the dimensionality of these distributions can often be very large, even the minimax rate of $n^{-\frac{\alpha}{2\alpha+D}}$ can be very slow, unless we assume a large degree of smoothness α, which may be unrealistic. However, such high-dimensional data often exhibit lower-dimensional structures, which we can exploit. For instance, suppose that the observed data $\mathsf{Y}_1, \mathsf{Y}_2, \ldots, \mathsf{Y}_n \overset{\text{i.i.d.}}{\sim} \mu$ all lie in some space $M \subseteq \mathbb{R}^D$ homeomorphic to some $M^* \subseteq \mathbb{R}^d$ where $d \ll D$, i.e. there exists a bi-continuous bijection $f : M^* \to M$ such that $\mathsf{Y}_i = f(\mathsf{Y}_i^*)$. Then estimating μ is tantamount to estimating $\mu^* = \mu \circ f$, and since this distribution is d-dimensional we might hope for the much faster convergence rate of $n^{-\frac{\alpha}{2\alpha+d}}$. In this context, we say that D is the *latent dimension* while d is the *intrinsic dimension*. This type of supposition is known as a *manifold hypothesis*, and might explain how de-noising diffusion models perform so well even on extremely high-dimensional data. Of course, in practice we seldom have any knowledge of M, M^* or f, so simply transforming the data to its lower-dimensional representation is not an option. Instead, the hope is that whichever approximation technique we employ can simultaneously discover this low-dimensional structure and the distribution thereon. In this article, this is exactly what we demonstrate for a specific subclass of the de-noising reflected diffusion models which we introduced in the previous article. In particular, we now assume that the target measure μ is supported on a subset of an affine subspace of dimension d. That is, we assume that there exists orthonormal vectors $v_1, v_2, \dots, v_d$ and an offset vector $v_0 \in [0,1]^D$ such that μ is supported on some subset $M \subseteq (V + v_0) \cap [0,1]^D$, where $V = \operatorname{span}\{v_1, v_2, \dots, v_d\}$. We then choose $[0,1]^D$ as our reflection set Ω and consider as our forward model the simple reflected Brownian motion:

$$\mathrm{d}\mathsf{X}_t = \mathrm{d}\mathsf{B}_t + \frac{1}{2} n(\mathsf{X}_t)\,\mathrm{d}\mathsf{L}_t, \qquad \mathsf{X}_0 \sim \mu. \tag{2.6}$$

Note that this corresponds to the model considered in the previous article with $f \equiv \frac{1}{2}$, whence we also have that $\mathsf{X}_t \overset{\sim}{\to} \mathrm{Unif}[0,1]^D$ as $t \to \infty$ and so we consider for some score approximation s and $\overline{T} > 0$ the process $\widehat{\mathsf{X}}^s = (\widehat{\mathsf{X}}^s_t)_{t \in [0,\overline{T}]}$ solving

$$\mathrm{d}\widehat{\mathsf{X}}^s_t = s(\widehat{\mathsf{X}}^s_t, \overline{T} - t)\,\mathrm{d}t + \mathrm{d}\overline{\mathsf{B}}_t + \frac{1}{2} n(\widehat{\mathsf{X}}^s_t)\,\mathrm{d}\overline{\mathsf{L}}_t, \qquad \widehat{\mathsf{X}}^s_0 \sim \mathrm{Unif}[0,1]^D,$$

and take as our sample $\widehat{\mathsf{X}} = \widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}$.

Probabilistic analysis

In the previous article, we relied on a spectral decomposition to derive a semi-explicit representation of the transition density q_t and hence the true score $\nabla \log p_t$. While still available here, this representation offers no immediate options for dimensionality reduction, as the spacial components (i.e. the eigenfunctions e_i) are not known. With this simpler model, this is different, as we can explicitly construct a weak solution to the forward process (2.6). The key insight is that for a regular D-dimensional Brownian motion $(\mathsf{B}_t)_{t \geq 0}$, each component is a 1-dimensional Brownian motion, independent of the rest, while the simple rectangular geometry suggests that the reflection in one coordinate is independent of the others. As such, we may hope that each component of a D-dimensional reflected Brownian motion behaves like an independent 1-dimensional Brownian motion, which are easily constructed through the Itô–Tanaka formula. In particular, letting $\widehat{f} : \mathbb{R} \to [0,1]$ denote the 2-periodic function given by

$$\widehat{f}(x) = \begin{cases} x, & \text{if } x \in [0,1) \\ -x, & \text{if } x \in [-1,0), \end{cases}$$

a 1-dimensional reflected Brownian motion $\mathsf{W}^{[0,1]} = (\mathsf{W}^{[0,1]}_t)_{t \geq 0}$ can be constructed from a regular 1-dimensional Brownian motion $\mathsf{W} = (\mathsf{W}_t)_{t \geq 0}$ by simply setting $\mathsf{W}^{[0,1]}_t = \widehat{f}(\mathsf{W}_t)$. Indeed, since $\widehat{f}$ is the difference between two convex functions, we have by the Itô–Tanaka formula (1.5) that

$$\begin{aligned} \widehat{f}(\mathsf{W}_t) &= \widehat{f}(\mathsf{W}_0) + \int_0^t \widehat{f}'_-(\mathsf{W}_s)\,\mathrm{d}\mathsf{W}_s + \frac{1}{2} \int_{\mathbb{R}} \mathsf{L}^a_t \widehat{f}''(\mathrm{d}a) \\ &= \int_0^t (-1)^{\lfloor \mathsf{W}_s \rfloor}\,\mathrm{d}\mathsf{W}_s + \frac{1}{2} \sum_{k \in \mathbb{Z}} (-1)^k \mathsf{L}^k_t. \end{aligned}$$

Here, the quadratic variation of the first term is given by

$$\left\langle \int_0^t (-1)^{\lfloor \mathsf{W}_s \rfloor} \,\mathrm{d}\mathsf{W}_s \right\rangle = \int_0^t \left((-1)^{\lfloor \mathsf{W}_s \rfloor}\right)^2 \mathrm{d}\langle \mathsf{W}\rangle_s = \int_0^t 1 \,\mathrm{d}s = t,$$

and so it follows by Lévy's characterisation that this is in fact a Brownian motion. As for the remaining terms, consider the local time $\widehat{\mathsf{L}}_t^0$ of $\widehat{f}(\mathsf{W}_t)$ at 0, and note that $\widehat{f}(\mathsf{W}_s) \in (-\varepsilon, \varepsilon)$ if and only if $\mathsf{W}_s \in (2k - \varepsilon, 2k + \varepsilon)$ for some $k \in \mathbb{Z}$, whence

$$\begin{aligned} \widehat{\mathsf{L}}_t^0 &= \lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_0^t \mathbf{1}_{(-\varepsilon,\varepsilon)}\big(\widehat{f}(\mathsf{W}_s)\big) \,\mathrm{d}\langle \widehat{f}(\mathsf{W})\rangle_s \\ &= \lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_0^t \sum_{k\in\mathbb{Z}} \mathbf{1}_{(2k-\varepsilon,2k+\varepsilon)}(\mathsf{W}_s) \,\mathrm{d}s \\ &= \sum_{k\in\mathbb{Z}} \mathsf{L}_t^{2k}. \end{aligned}$$

The same analysis shows that $\widehat{\mathsf{L}}_t^1 = \sum_{k\in\mathbb{Z}} \mathsf{L}_t^{2k+1}$ and so it follows that

$$\widehat{f}(\mathsf{W}_t) = \widehat{\mathsf{W}}_t + \frac{1}{2}\left(\widehat{\mathsf{L}}_t^0 - \widehat{\mathsf{L}}_t^1\right),$$

where $\widehat{\mathsf{W}}_t := \int_0^t (-1)^{\lfloor \mathsf{W}_s \rfloor} \,\mathrm{d}\mathsf{W}_s$ is a Brownian motion, showing that $\widehat{f}(\mathsf{W}_t)$ indeed is a reflected Brownian motion. Applying this entry-wise to a regular D-dimensional Brownian motion $(\mathsf{B}_t)_{t\geq 0}$ and adding an offset of $\mathsf{Y} \sim \mu$ yields the desired process – the details are given in Lemma C.50 – and we arrive at the following lemma:

Lemma 2.9: Given a distribution μ on $[0,1]^D$, a solution $\mathsf{X} = (\mathsf{X}_t)_{t\geq 0}$ to (2.6) is given by

$$\mathsf{X}_t = f(\mathsf{B}_t + \mathsf{Y}),$$

where f is the entry-wise application of $\widehat{f}$, $\mathsf{B} = (\mathsf{B}_t)_{t\geq 0}$ is a D-dimensional Brownian motion independent of $\mathsf{Y} \sim \mu$.

With this, we can construct an explicit transition density, in that we note that for any $x, y \in [0,1]^D$, the process X travels from y to x if and only if the Brownian motion B travels from y to any point in $f^{-1}(\{x\})$. We can enumerate this preimage as follows: for $z \in \mathbb{Z}^D$, let $R_z(x)$ be given by

$$\big(R_z(x)\big)_i = \begin{cases} x_i, & \text{if } z_i \text{ even} \\ 1 - x_i, & \text{if } z_i \text{ odd,} \end{cases}$$

and note that $f^{-1}(\{x\}) = \{z + R_z(x) \mid z \in \mathbb{Z}^D\}$. Combining these facts, we have

$$\begin{aligned} q_t(y,x)\,\mathrm{d}x &= \mathbb{P}(\mathsf{X}_t \in \mathrm{d}x \mid \mathsf{Y} = y) \\ &= \mathbb{P}(f(\mathsf{B}_t + y) \in \mathrm{d}x) \\ &= \mathbb{P}(\mathsf{B}_t + y \in f^{-1}(\mathrm{d}x)) \\ &= \sum_{z\in\mathbb{Z}^D} \mathbb{P}(\mathsf{B}_t + y - z \in R_z(\mathrm{d}x)) \\ &= \sum_{z\in\mathbb{Z}^D} (2\pi t)^{-\frac{D}{2}} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}\,\mathrm{d}x. \end{aligned}$$

From here, applying the chain rule yields that

$$\nabla q_t(y,x) = -(2\pi t)^{-\frac{D}{2}} \sum_{z\in\mathbb{Z}^D} (-1)^z \frac{R_z(x)+z-y}{t} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}},$$

and hence

$$\nabla \log p_t(x) = -\frac{\sum_{z\in\mathbb{Z}^D} (-1)^z \int_{[0,1]^D} (R_z(x)+z-y) \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}\,\mu(\mathrm{d}y)}{t\sum_{z\in\mathbb{Z}^D} (-1)^z \int_{[0,1]^D} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}\,\mu(\mathrm{d}y)},$$

where we use the shorthand $(-1)^z = \operatorname{diag}\{(-1)^{z_i}\}$.

Score approximation and dimensionality reduction

As in the previous article, in order to reduce the error stemming from using an approximation of the score function, we must construct an explicit estimator with low L^2-distance to the true score under μ – we defer the details for now to the next subsection. However, to take advantage of the lower-dimensional structure of μ and its support, we need to re-write the score $\nabla \log p_t$ to also reflect this. In essence, the idea is to split the contributions to p_t into an on-support part stemming from μ, which is hard to approximate but lower-dimensional, and an off-support part stemming from the forward process, which is high-dimensional but (comparatively) easy to approximate. The key to doing this is the following two observations:

(1) Letting $A = \begin{bmatrix} v_1 & v_2 & \cdots & v_d \end{bmatrix}$, and assuming that μ has a density p_0 wrt. the d-dimensional Lebesgue measure on $V + v_0$, we have for any integrable function $g : M \to \mathbb{R}$

$$\int_M g(y)\,\mu(\mathrm{d}y) = \int_{M^*} g(Au + v_0) p_0(Au + v_0)\,\mathrm{d}u,$$

where $M^* = A^\top(M - v_0) \subseteq \mathbb{R}^d$

(2) Letting $P = AA^\top$ denote the orthogonal projection onto V, we have by the Pythagorean theorem that for $x \in \mathbb{R}^D$ and $u \in \mathbb{R}^d$,

$$\begin{aligned}|x - (Au + v_0)|^2 &= |P(x - v_0) - Au|^2 + |(I - P)(x - v_0)|^2\\ &= |A^\top(x - v_0) - u|^2 + |(I - P)(x - v_0)|^2,\end{aligned}$$

since A is an isometry.

Setting $x^\perp = (I - P)(x - v_0)$ and $x^* = A^\top(x - v_0)$ for $x \in \mathbb{R}^D$, these together yield that

$$\int_M \mathrm{e}^{-\frac{|x-y|^2}{2t}}\,\mu(\mathrm{d}y) = \mathrm{e}^{-\frac{|x^\perp|^2}{2t}} \int_{M^*} \mathrm{e}^{-\frac{|x^*-u|^2}{2t}}\, p_0(Au + v_0)\,\mathrm{d}u$$

and

$$\begin{aligned}\int_M \frac{x-y}{t}\mathrm{e}^{-\frac{|x-y|^2}{2t}}\,\mu(\mathrm{d}y) = \mathrm{e}^{-\frac{|x^\perp|^2}{2t}} \Bigg(&\frac{x^\perp}{t}\int_{M^*} \mathrm{e}^{-\frac{|x^*-u|^2}{2t}}\, p_0(Au + v_0)\,\mathrm{d}u\\ &+ A\int_{M^*} \frac{x^*-u}{t}\mathrm{e}^{-\frac{|x^*-u|^2}{2t}}\, p_0(Au + v_0)\,\mathrm{d}u\Bigg).\end{aligned}$$

As such, we see that all components of $\nabla \log p_t(x)$ including the unknown distribution μ rely on x only through its lower-dimensional projection x^*, and it is this that allows for dimensionality reduction. In particular, if we let $f_1 : \mathbb{R}^d \times (0, \infty) \to \mathbb{R}$ and $f_2 : \mathbb{R}^d \times (0, \infty) \to \mathbb{R}^d$ be given by

$$f_1(v, t) = \int_{M^*} \mathrm{e}^{-\frac{|v-u|^2}{2t}}\, p_0(Au + v_0)\,\mathrm{d}u$$

and

$$f_2(v, t) = \int_{M^*} \frac{v-u}{t}\mathrm{e}^{-\frac{|v-u|^2}{2t}}\, p_0(Au + v_0)\,\mathrm{d}u,$$

then we can re-write the score as

$$s_0(x, t) = -\frac{\sum_{z\in\mathbb{Z}^D}(-1)^z \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}}\left(\frac{x_z^\perp}{t} f_1(x_z^*, t) + A f_2(x_z^*, t)\right)}{\sum_{z\in\mathbb{Z}^D} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}} f_1(x_z^*, t)},$$

where $x_z = R_z(x) + z$ for $x \in [0, 1]^D$ and $z \in \mathbb{Z}^D$. Thus we need only approximate the functions f_1 and f_2 which are intrinsically d-dimensional in their spacial components. To do this, we make the following assumptions on μ and M:

($\mathcal{H}0$) The target measure μ, its density p_0 and its support $M \subseteq (V + v_0) \cap [0, 1]^D$ satisfy

(i) $p_0 \in H_0^\alpha(M)$ with $\alpha \in \mathbb{N} \cap (d/2, \infty)$, i.e. it is α-Sobolev smooth with derivatives vanishing at the boundary;

(ii) There exists constants $c_0 \geq d$ and $r_0 > 0$ such that $\mu(B(x,r) \cap M) \gtrsim (r \wedge r_0)^{c_0}$ and $\lambda_d(B(x,r) \cap M) \gtrsim (r \wedge r_0)^d$ for all $x \in M$;

(iii) There exists an $\varepsilon_M > 0$ and constants $0 < p_{\min} \leq p_{\max} < \infty$ such that for $x \in M$ with $\operatorname{dist}(x, \partial M) \geq \frac{\varepsilon}{2}$, $p_{\min} \leq p_0(x) \leq p_{\max}$;

(iv) When restricted to the set $\{x \in M \mid \operatorname{dist}(x, \partial M) \leq \varepsilon_M\}$, p_0 has continuous derivatives of order $\kappa := \frac{d(c_0-d)}{2} + d + 3\alpha + 2$;

(v) There exists a constant $\rho_{\min} > 0$ such that $\operatorname{dist}(M, \partial[0,1]^D) \geq \rho_{\min}$.

Here, condition (i) ensures that p_0 has overall smoothness α while the vanishing derivatives ensure that p_0 can be extended to all of $V + v_0$ smoothly by simply setting it to 0 outside of M. This in turn means that we have no uniform lower bound on p_0 on all of M as we did in the previous article – instead conditions (ii) and (iii) ensure that it is lower bounded on the part of M with minimal smoothness and does not decay too rapidly towards the boundary, while (iv) yields a higher degree of smoothness where p_0 is not lower bounded, allowing for better convergence rates. Finally, condition (v) ensures that with high probability, the forward process X_t behaves like an unreflected Brownian motion for small t, as there is a positive distance to the reflection barrier, which will become important later. With these assumptions, we can begin to approximate the true score function s_0. To start, we first truncate the sums in both numerator and denominator of s_0 – as these decay exponentially in $|z|$, we need only consider $\operatorname{Poly}(\log n)$ many terms to achieve an optimal rate. Next, we approximate for fixed times $t > 0$ the functions $f_1(\cdot, t)$ and $f_2(\cdot, t)$. Here we divide into two regimes: for "small" t, we utilise the increased smoothness of p_0 near the boundary to achieve better convergence rates where the denominator of s_0 is small, while for "large" t, we use the induced smoothness of forward process to achieve overall better convergence rates. We then, like in the previous article, employ polynomial interpolation to extend these fixed-time approximations to small time intervals, and finally piece these together to approximate s_0 on such intervals. Following this general strategy, we ultimately show the following:

Theorem 2.10: Under $(\mathcal{H}0)$ for any $\delta > 0$ and suitable $m \in \mathbb{N}$ and $\underline{t} > 0$, there exists a neural network

$$\varphi_{s_0} \in \begin{cases} \widetilde{\Phi}\left((\log m)^2 (\log\log m)^2, m(\log m)^{D+1}, m(\log m)^{D+2}, m^{\frac{\alpha}{d}} \underline{t}^{-1} \vee m^{\nu}\right), & \text{if } \underline{t} \leq \frac{1}{2} m^{-\frac{2-\delta}{d}} \\ \widetilde{\Phi}\left((\log m)^2 (\log\log m)^2, m'(\log m)^{D+1}, m'(\log m)^{D+2}, m'\right), & \text{if } \underline{t} > \frac{1}{2} m^{-\frac{2-\delta}{d}} \end{cases},$$

where $\nu = \frac{2d}{2\alpha - d} + \frac{1}{d}$ and $m' = \underline{t}^{-\frac{d}{2}} m^{\frac{\delta}{2}}$ satisfying

$$\int_{\underline{t}}^{2\underline{t}} \mathbb{E}[|s_0(X_t) - \varphi_{s_0}(X_t)|^2] \, \mathrm{d}t \lesssim \begin{cases} (\log m)^{d+2D+3} m^{-\frac{2\alpha}{d}}, & \text{if } \underline{t} \leq m^{-\frac{2-\delta}{d}} \\ (\log m)^{d+2D+3} m^{-\frac{2(\alpha+1)}{d}}, & \text{if } \underline{t} > m^{-\frac{2-\delta}{d}}. \end{cases}$$

Approximation in Wasserstein distance

With this manifold hypothesis in mind, we cannot directly use total variation as our error measure as in the previous article, as μ and the distribution of $\widehat{\mathsf{X}}$ are necessarily singular whence

$$\mathrm{TV}(\mu, \widehat{\mathsf{X}}) \geq |\mu(M) - \mathbb{P}(\widehat{\mathsf{X}} \in M)| = 1.$$

Instead, we choose to use the Wasserstein-1-distance, which does not suffer from this issue. To decompose the error like before, we introduce the process $\overline{\mathsf{X}}^s = (\overline{\mathsf{X}}_t^s)_{t \geq 0}$ solving

$$\mathrm{d}\overline{\mathsf{X}}_t^s = s(\overline{\mathsf{X}}_t^s, \overline{T} - t)\,\mathrm{d}t + \mathrm{d}\overline{\mathsf{B}}_t + \frac{1}{2} n(\overline{\mathsf{X}}_t^s)\,\mathrm{d}\overline{\mathsf{L}}_t, \qquad \overline{\mathsf{X}}_0^s \sim p_{\overline{T}}.$$

We then have

$$\mathcal{W}_1(\mu, \widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}) \leq \mathcal{W}_1(\mu, \mathsf{X}_{\underline{T}}) + \mathcal{W}_1(\mathsf{X}_{\underline{T}}, \overline{\mathsf{X}}^s_{\overline{T}-\underline{T}}) + \mathcal{W}_1(\overline{\mathsf{X}}^s_{\overline{T}-\underline{T}}, \widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}),$$

where the simple coupling of $\mathsf{Y} \sim \mu$ and $\mathsf{X}_t \sim f(\mathsf{B}_t + \mathsf{Y})$ cf. Lemma 2.9 yields

$$\mathcal{W}_1(\mu, \mathsf{X}_t) \leq \mathbb{E}[|f(\mathsf{B}_t + \mathsf{Y}) - \mathsf{Y}|] = \mathbb{E}[|f(\mathsf{B}_t + \mathsf{Y}) - f(\mathsf{Y})|] \leq \mathbb{E}[|\mathsf{B}_t|] \leq \sqrt{Dt},$$

since f is 1-Lipschitz. Furthermore, since $\mathcal{W}_1(\nu, \nu') \leq 2\sqrt{D}\mathrm{TV}(\nu, \nu')$ for all measures ν, ν' on $\mathbb{R}^D$, we can show as in the previous article that $\mathcal{W}_1(\overline{\mathsf{X}}^s_{\overline{T}-\underline{T}}, \widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}})$ decays exponentially in $\overline{T}$. In principle, then, we could progress as in the previous article to achieve a convergence rate of $n^{-\frac{\alpha}{2\alpha+d}}$ up to log-factors. For Wasserstein-1-distance this is, however, suboptimal as the minimax-risk is $n^{-\frac{\alpha+1}{2\alpha+d}}$, and so we must be more creative. The detailed argument is quite involved and is deferred to Article C, but the core idea is as follows:

(i) Split the time interval $[\underline{T}, \overline{T}]$ into $K \asymp \log n$ many subintervals $[t_{i-1}, t_i)$ with $t_{i+1}/t_i \leq 2$

(ii) On each subinterval consider a coupling of two backward processes started in the same point, but one using the true score s_0 and one using an approximation s

(iii) When t_i is small, we can only approximate the score at rate $n^{-\frac{\alpha}{2\alpha+d}}$, however since the interval is small, the coupling stays close with high probability

(iv) When t_i is large, the induced smoothness of the forward process allows us to estimate the score at a faster rate of $n^{-\frac{\alpha+1}{2\alpha+d}}$

Following this strategy, we show in particular that

$$\mathcal{W}_1(\mathsf{X}_{\underline{T}}, \overline{\mathsf{X}}^s_{\overline{T}-\underline{T}}) \lesssim \log n \sum_{i=1}^{K} \sqrt{t_i \wedge 1}\Big(\int_{t_{i-1}}^{t_i} \mathbb{E}\big[|s(\mathsf{X}_t, t) - s_0(\mathsf{X}_t, t)|^2\big]\,\mathrm{d}t\Big)^{\frac{1}{2}},$$

which combined with Theorem 2.10 shows for all $\delta > 0$ the existence of a function s^*, defined piecewise in time as a sequence of neural networks, satisfying

$$\mathcal{W}_1(\mathsf{X}_{\underline{T}}, \overline{\mathsf{X}}^{s^*}_{\overline{T}-\underline{T}}) \lesssim n^{-\frac{\alpha+1-\delta}{2\alpha+d}},$$

i.e. one that is arbitrarily close to minimax-optimal. Finally, since the network sizes in Theorem 2.10 also decay with t, this ensures that the generalisation error of using the estimator $\widehat{s}_n$ defined piecewise in the same fashion as in the previous article is also of order $n^{-\frac{\alpha+1-\delta}{2\alpha+d}}$. All in all, this yields the following final theorem:

Theorem 2.11: Assuming $(\mathcal{H}0)$ and choosing $\underline{T} \in \mathrm{Poly}(n^{-1})$ and $\overline{T} \asymp \log n$, for every δ there exists a family $\{\mathcal{S}_i\}_{i=1}^{K}$ of neural networks such that

$$\mathbb{E}\left[\mathcal{W}_1(\mu, \widehat{\mathsf{X}}^{\widehat{s}_n}_{\overline{T}-\underline{T}})\right] \lesssim n^{-\frac{\alpha+1-\delta}{2\alpha+d}}.$$

Article A

*Note: this is a **copy** of the article*

Sören Christensen, Asbjørn Holk Thomsen, and Lukas Trottner (2024)
Data-Driven Rules for Multidimensional Reflection Problems
SIAM/ASA Journal on Uncertainty Quantification 12:4, 1240-1272
DOI: 10.1137/23M1618570; arXiv:2311.06639 [math.OC]

The only changes made are to the notation and layout for sake of consistency with the rest of the dissertation.

Data-driven rules for multidimensional reflection problems

Sören Christensen, Asbjørn Holk Thomsen and Lukas Trottner

Abstract

Over the recent past data-driven algorithms for solving stochastic optimal control problems in face of model uncertainty have become an increasingly active area of research. However, for singular controls and underlying diffusion dynamics the analysis has so far been restricted to the scalar case. In this paper we fill this gap by studying a multivariate singular control problem for reversible diffusions with controls of reflection type. Our contributions are threefold. We first explicitly determine the long-run average costs as a domain-dependent functional, showing that the control problem can be equivalently characterized as a shape optimization problem. For given diffusion dynamics, assuming the optimal domain to be strongly star-shaped, we then propose a gradient descent algorithm based on polytope approximations to numerically determine a cost-minimizing domain. Finally, we investigate data-driven solutions when the diffusion dynamics are unknown to the controller. Using techniques from nonparametric statistics for stochastic processes, we construct an optimal domain estimator, whose static regret is bounded by the minimax optimal estimation rate of the unreflected process' invariant density. In the most challenging situation, when the dynamics must be learned simultaneously to controlling the process, we develop an episodic learning algorithm to overcome the emerging exploration-exploitation dilemma and show that given the static regret as a baseline, the loss in its sublinear regret per time unit is of natural order compared to the one-dimensional case.

Keywords: Stochastic singular control, reinforcement learning, exploration vs. exploitation, reversible diffusions, shape optimization, nonparametric statistics

MSC Codes: Primary 93E35, 68T05; secondary 49Q10, 60J60, 62M05

A.1 Introduction and problem formulation

Our underlying processes are Langevin diffusions X on $\mathbb{R}^d$ for $d \geq 2$, which is a well studied class of reversible diffusion processes with drift of potential form. For a C^2-function $V : \mathbb{R}^d \to \mathbb{R}$ and a d-dimensional Brownian motion B, X solves the SDE

$$\mathrm{d}\mathsf{X}_t = -\nabla V(\mathsf{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathsf{B}_t.$$

We consider a basic class of stochastic control problems with a clear interpretation: X is interpreted as the position of a particle, which we want to be close to a target state, 0, say. The distance is measured by a locally bounded function $f : \mathbb{R}^d \to [0, \infty)$ in such a way that $f(\mathsf{X}_t)$ stands for the costs associated with being away from the target state. The decision maker can now control the process by choosing a nonempty bounded domain (= open, connected set) $\Omega \subseteq \mathbb{R}^d$ of class C^2 and normally reflecting the process at $\partial\Omega$. We denote the resulting normally reflected processes by X^Ω, which is given as the solution to

$$\mathrm{d}\mathsf{X}^\Omega_t = -\nabla V(\mathsf{X}^\Omega_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathsf{B}_t + n(\mathsf{X}^\Omega_t)\,\mathrm{d}\mathsf{L}^\Omega_t,$$

where n is the unit inward normal vector of Ω at $x \in \partial\Omega$ and L^Ω denotes the local time of X^Ω on $\partial\Omega$, that is, a non-decreasing one-dimensional process with continuous paths that increases only when $\mathsf{X}^\Omega \in \partial\Omega$. We assume that controlling the process is associated with costs proportional to L^Ω (with proportionality factor $\kappa > 0$), so that for each Ω, the total costs associated with Ω until time t are

$$\int_0^t f(\mathsf{X}_s^\Omega)\mathrm{d}s + \kappa \mathsf{L}_t^\Omega .$$

Here, we consider a long-term-average criterion and our problem consists of minimizing

$$J(\Omega) := \liminf_{t\to\infty} \frac{1}{t}\mathbb{E}^x \left[\int_0^t f(\mathsf{X}_s^\Omega)\mathrm{d}s + \kappa \mathsf{L}_t^\Omega \right] \tag{A.7}$$

over all admissible Ω. The limit is independent of the initial value $x \in \Omega$ due to the ergodic nature of the reflected diffusion, which will be made precise in Theorem A.15. For known potential V of the underlying process, the problem we consider here is closely connected to singular stochastic control problems. We discuss this in more detail in Section A.1 below. The problem is a general formulation for trade-offs, where a decision maker wants to keep a stochastic system X as close as possible to a target state, but control incurs proportional costs. We discuss a number of applications in the following section. For the dimension $d = 1$, the setting considered here already includes essentially all such problems. In the multidimensional case $d \geq 2$, more problem-specific formulations are needed, especially with respect to restrictions on transaction costs and the type of diffusion, so that one often concentrates on generic formulations in order to understand general phenomena, see for example [62]. This is the approach we take in this paper.

Contributions

In this paper, we first address the question how the solution to the problem (A.7) can be meaningfully characterized when the potential V of the underlying process is known. Our main contribution here is Theorem A.15 which shows under minimal assumptions that $J(\Omega)$ is the L^1-limit of the average costs and is explicitly given by

$$J(\Omega) = \frac{1}{\int_\Omega \mathrm{e}^{-V(y)}\mathrm{d}y} \left(\int_\Omega f(y)\mathrm{e}^{-V(y)}\,\mathrm{d}y + \kappa \int_{\partial\Omega} \mathrm{e}^{-V(y)}\,\mathcal{H}^{d-1}(\,\mathrm{d}y) \right), \tag{A.8}$$

where $\mathcal{H}^{d-1}$ denotes the $(d-1)$-dimensional Hausdorff measure. The formula is interesting in that the control problem has been transformed into a shape optimization problem. Let us comment here on the corresponding control model with non-constant reflection costs, that is replacing (A.7) by

$$\widetilde{J}(\Omega) := \liminf_{t\to\infty} \frac{1}{t}\mathbb{E}^x \left[\int_0^t f(\mathsf{X}_s^\Omega)\mathrm{d}s + \int_0^t \kappa(\mathsf{X}_s^\Omega)\,\mathrm{d}\mathsf{L}_s^\Omega \right], \tag{A.9}$$

for some locally bounded function $\kappa : \mathbb{R}^d \to [0, \infty)$. Recent results from [73] that are based on solutions to appropriate Poisson equations with inhomogeneous boundary conditions, show that for $\kappa \in C^3$ and given stronger regularity assumptions on the boundary $\partial\Omega$ and the potential V than those that we impose for constant reflection costs, the L^1-limit is given by

$$\widetilde{J}(\Omega) = \frac{1}{\int_\Omega \mathrm{e}^{-V(y)} \mathrm{d}y} \left(\int_\Omega f(y) \mathrm{e}^{-V(y)} \, \mathrm{d}y + \int_{\partial\Omega} \kappa(y) \mathrm{e}^{-V(y)} \, \mathcal{H}^{d-1}(\mathrm{d}y) \right),$$

and the rates of convergence are the same as in Theorem A.15. Thus, given sufficient regularity assumptions, the results from this paper can be extended to non-constant reflection costs with straightforward modifications to the numerical and statistical results that we describe next.

In Section A.3 we show that the approach we present provides a suitable basis for addressing the problem in the context of model-based reinforcement learning. More precisely, we show that the problem can be solved when the drift function $b = -\nabla V$ is unknown to the decision maker, so that the control has to be purely data driven. Our approach is based on estimating the stationary density of the uncontrolled process ρ nonparametrically and specifying the exact rate with techniques from nonparametric statistics for diffusion processes (Theorem A.21). We use this via (A.8) to estimate the optimal boundary based on a path of the uncontrolled process and obtain that the resulting static regret has the same sublinear rate (Proposition A.23). In the more practically relevant situation of simultaneous optimization and data collection, we face an exploration vs. exploitation dilemma. The previous results together with an episodic learning approach lead to a proof of a sublinear regret rate (Theorem A.24 and Corollary A.25).

Finally, in Section A.4 we give a numerical approach to minimizing $J(\Omega)$ based on approximating Ω by polytopes and then applying gradient-based methods. To this end, we derive explicit formulas (Theorem A.26) under the assumption that Ω is star-shaped. We also discuss how this might be implemented in practice for any general $d \geq 2$ and give examples of the discussed optimization procedures, both where the potential V is known and when it is learned in accordance with Section A.3.

Related literature

Stochastic singular control problems are a class of stochastic control problems that have been extensively studied, see [47, 90] for textbook treatments. They arise in various applications such as inventory management in operations research [51], control of queueing networks [66], portfolio selection with transaction costs in finance [31], equity issuance in insurance mathematics [78], position control in engineering [83], or optimal harvesting in biology [1]. Underlying problems where the processes directly correspond to the mean-reverting-type processes used later in this paper include the management of exchange rates [44] and interest rates [130]. The actions of the agents affect the state and the costs in proportion to the size of the action. This structure implies that the optimal controls usually have the following general form: The space is divided into a region where action is required and a region Ω where no action is required. The optimal control then reflects the controlled process at the boundary of the no-action region to keep it inside. Therefore, the problem class is closely related to reflected SDEs [76, 94], as observed for a variate of examples, see e.g., [32, 52, 67, 105] (although this connection is difficult to be made precise in full generality, see the discussion in [13]). Thus, stochastic singular control can typically be reduced to finding the optimal no-action region Ω and the optimal reflection direction at the boundary of Ω. However, finding the solution is usually only possible explicitly in a few examples, typically with underlying one-dimensional diffusions, see the following section. Characterizing optimal controls becomes much harder when the problem has more than one dimension. Some characterizations of the optimal solutions in special multidimensional cases can be found in [30, 35, 41, 42, 68]. Thus, the problem (A.7) can be viewed as an optimization problem over a class of strategies typically relevant to general singular control problems. We restrict this, however, by the fact that only normal reflections are admissible. This assumption may be justifiable in some cases from the real world problem being modeled. However, we will also see below that in subclasses of problems, such as the radially symmetric case, this is not a restriction at all. In what follows, we will see that the assumption of normal reflection simplifies the problem to the point where a deeper analysis can be performed.

There is a fair amount of literature on the numerical treatment of singular control problems. The methods range from approaches based on a discretization with discrete Markov chains [70] over an approximation of the solution of the corresponding Hamilton-Jacobi-Bellman equation [69], approaches using linear programming [127] up to finite element approximations [126]. However, all methods have limited applicability, especially in higher dimensions, and care must be taken in the exact implementation. In this paper, we present a new approach that is structurally different from the existing ones.

As described above, the approach we have taken allows for data-driven control when the drift of the process is unknown. This question falls into the currently fast growing field of model based reinforcement learning (RL), where the agent does not know the system parameters and learns them by interacting with the environment and getting feedback. The agent chooses policies based on the current parameter estimation and tries to minimize the regret, which is the gap between the expected reward of the best policy and the actual reward achieved. Many discrete-time RL problems have been studied, where sublinear regret bounds have been obtained for different scenarios, such as bandit problems, tabular Markov decision problems, and linear quadratic (LQ) problems [33, 54, 91]. However, it is well known that the transition to high-frequency situations poses a structural problem for standard approaches such as Deep-Q learning [116]. Therefore, the concrete model must be included here. Most of the previous works only propose algorithms, and only a few analyze their regrets, mostly for LQ problems. In particular, [39] proved an asymptotic sublinear regret for regularized least-squares algorithms in an ergodic continuous-time LQ problem, but without giving the exact order of the bound. Recently, [11, 50] generalized the least-squares algorithms to finite-time horizon episodic settings and gave non-asymptotic regret bounds. These works assume a parametric structure of the problem. On the other hand, [84] considered propagator models and combined exploration and exploitation schemes to achieve sublinear regrets with high probability. Closest to this paper are the articles [27, 28], in which singular and impulse control problems with a nonparametric statistical diffusion structure are considered. However, these papers make heavy use of the one-dimensional structure of the underlying processes. This leads to the strategy already being described by one or two values, so only these need to be learned. One of the main contributions of this work is that we can give exact sublinear rates of regression even for the case where the optimal strategy is only infinite-dimensional parameterizable. To the best of our knowledge, this work is the first to provide such results in this context.

Notation We write λ_d for the Lebesgue measure on $\mathbb{R}^d$. Both the notations $\overline{B}$ and $\operatorname{cl} B$ will be used for the closure of a set $B \subset \mathbb{R}^d$. We write $a \lesssim_\theta b$ if $a \leq C(\theta)b$ for some constant $C(\theta)$ depending on a parameter θ, and $a \lesssim b$ if $a \leq Cb$ for some universal constant C.

A.2 Optimal reflection as a shape optimization problem for known potential V

As detailed above, for our approach it is central to find explicit expressions for $J(\Omega)$ for known drift. Before we get to that, we start here with a brief discussion of the one-dimensional special case, since it serves as the main motivation. The main observation for obtaining explicit solutions for all singular control problems for underlying linear diffusions is that the values $J(\Omega)$ can be found (semi-)explicitly in terms of speed measure and scale function. In this linear case, solving (A.7) therefore boils down to a standard optimization problem, see [2, 26] and the references therein. In the one dimensional ergodic case discussed here, this is realized by the fact that the stationary density of the reflected diffusion is just the conditional density of the uncontrolled diffusion. The key observation in this section is that this also holds in the multivariate case with underlying Langevin diffusions. This turns out to be key to establishing (A.8).

Ergodicity of the reflected Langevin diffusions

It is well-known that for an (uncontrolled) Langevin diffusion as introduced above, if e^{-V} is integrable, then X has a stationary density given by

$$\rho_{\mathbb{R}^d}(x) := \rho(x) := c^{-1}\exp(-V(x)), \text{ where } c = c_{\mathbb{R}^d}(V) = \int_{\mathbb{R}^d} \mathrm{e}^{-V(u)}\mathrm{d}u.$$

For general diffusions, knowing the distribution on the whole state space does not give any information about the stationary distributions of the corresponding diffusions with reflection in a subdomain Ω. A main observation for our approach is that for Langevin diffusions, this is different. Indeed, in this particular situation one can obtain the stationary distribution on Ω by conditioning:

Lemma A.12: For a bounded domain $\Omega \subseteq \mathbb{R}^d$ of class C^2, the density ρ_Ω given by

$$\rho_\Omega(x) = c_\Omega^{-1} \exp(-V(x)), \quad x \in D, \text{ where } c_\Omega = \int_\Omega \mathrm{e}^{-V(u)} \mathrm{d}u,$$

is a stationary density of the normally reflected process X^Ω.

This result is well-known and can, e.g., be proved with [61, Theorem 1], using that for the differential operator given by

$$Af(x) = -\langle \nabla V(x), \nabla f(x) \rangle + \Delta f(x), \quad f \in C^2(\mathbb{R}^d), x \in \mathbb{R}^d, \tag{A.10}$$

we have by the divergence theorem and the relation $\nabla \rho_\Omega = -\nabla V \rho_\Omega$,

$$\int_\Omega Af(x)\, \pi_\Omega(\,\mathrm{d}x) = -\int_{\partial\Omega} \langle \nabla f(x), n(x) \rangle \rho_\Omega(x)\, \mathcal{H}^{d-1}(\,\mathrm{d}x), \quad f \in C^2(\mathbb{R}^d), \tag{A.11}$$

where $\mathcal{H}^{d-1}$ denotes the $(d-1)$-dimensional Hausdorff measure. Since this property is essential for our statistical approach, let us briefly comment on generalizations thereof for diffusions with non-constant diffusion matrix. The conditioning property is more generally connected to reversibility of the non-constrained diffusion. Indeed, consider a general elliptic diffusion Y satisfying

$$\mathrm{d}\mathsf{Y}_t = b(\mathsf{Y}_t)\,\mathrm{d}t + \sigma(\mathsf{Y}_t)\,\mathrm{d}\mathsf{B}_t,$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ satisfies $\Sigma(x) := \sigma(x)\sigma(x)^\top \gtrsim \mathbb{I}$. If Y has a strictly positive invariant density $\rho^\mathsf{Y} \propto \exp(-\Psi)$ for some potential $\Psi : \mathbb{R}^d \to \mathbb{R}$, then Y is reversible (in the sense $(\mathsf{Y}_{T-t})_{t\in[0,T]} \overset{d}{=} (\mathsf{Y}_t)_{t\in[0,T]}$ for any $T > 0$ when $\mathsf{Y}_0 \sim \rho^\mathsf{Y}$) if, and only if, the detailed balance equation

$$b(x) = \frac{1}{2\rho^\mathsf{Y}(x)} \nabla \cdot (\Sigma(x)\rho^\mathsf{Y}(x)), \quad x \in \mathbb{R}^d,$$

is satisfied. In this case, if we replace the normal reflection direction $n(x)$ by co-normal reflection in direction $\nu(x) = \frac{1}{2}\Sigma(x)n(x)$ at $\partial\Omega$, then the co-normally reflected process Y^Ω on Ω has an invariant density obtained from conditioning, i.e., $\rho^\mathsf{Y}_\Omega = \frac{1}{\rho^\mathsf{Y}(\Omega)}\rho^\mathsf{Y}$ on Ω. In this sense, the results of this paper can potentially be extended to multivariate reversible diffusions with co-normal reflection controls (which, of course, requires Σ to be known).

It will be important for our purposes to have sufficiently fast convergence of the reflected diffusion to equilibrium. To this end, we will assume that $\Omega \subset \mathbb{R}^d$ is a bounded domain of class C^2 that is sufficiently nice to guarantee that the Markov process $(\mathsf{X}^\Omega, (\mathbb{P}^x)_{x\in\overline{\Omega}})$ has transition densities $(p_t^\Omega(x,y))_{t>0,x,y\in\overline{\Omega}}$—that is, $\mathbb{E}^x[f(\mathsf{X}_t^\Omega)] = \int_{\overline{\Omega}} p_t^\Omega(x,y) f(y)\,\mathrm{d}y$ for any bounded measurable function $f : \overline{\Omega} \to \mathbb{R}$ and $x \in \overline{\Omega}$—such that for any $t > 0$, p_t^Ω is continuous on $\overline{\Omega} \times \overline{\Omega}$ and we have the minorization property

$$\inf_{x,y\in\overline{\Omega}} p_1^\Omega(x,y) \geq \delta, \tag{A.12}$$

for some $\delta > 0$, where necessarily $\delta\lambda_d(\Omega) \in (0,1)$. Let us denote by $\boldsymbol{\Omega}$ the class of bounded domains $\Omega \subset \mathbb{R}^d$ of class $\mathcal{C}^2$ such that these assumptions hold for the process reflected in $\partial\Omega$. These can be verified under mild assumption on the boundary $\partial\Omega$ by general results on fundamental solutions of parabolic PDEs with Neumann boundary conditions. For instance, continuity of the transition densities and (A.12) are ensured whenever $\partial\Omega$ is the union of a finite number of hypersurfaces of class $\mathcal{C}^3$, cf. [60, p.166]. These assumptions now guarantee uniqueness of the invariant distribution and exponential ergodicity of the process.

Lemma A.13: Let $\Omega \in \boldsymbol{\Omega}$ with transition minorization as in (A.12). Then, the unique stationary distribution π_Ω of the normally reflected diffusion X^Ω is given by

$$\pi_\Omega(\mathrm{d}x) = \rho_\Omega(x)\,\mathrm{d}x, \quad x \in \overline{\Omega},$$

and X^Ω is uniformly ergodic, satisfying the bound

$$\|P_t^\Omega(x,\cdot) - \pi_\Omega\|_{\mathrm{TV}} \le \tfrac{2}{1-\delta\lambda_d(\Omega)}\mathrm{e}^{t\log(1-\delta\lambda_d(\Omega))}, \quad x \in \overline{\Omega}, \quad t > 0.$$

The short proof can be found in the supplement. As a consequence we have the following rate in the ergodic theorem, which is also proved in the supplement.

Corollary A.14: Let $\Omega \in \boldsymbol{\Omega}$. There exists a constant $C = C(\Omega) > 0$ such that for any $h \in L^\infty(\overline{\Omega})$ and $x \in \overline{\Omega}$ it holds

$$\frac{1}{t}\mathbb{E}^x\Big[\Big|\int_0^t (h(\mathsf{X}_s^\Omega) - \pi_\Omega(h))\,\mathrm{d}s\Big|\Big] \le \frac{C\|h\|_{L^\infty(\overline{\Omega})}}{\sqrt{t}}.$$

Solution of the ergodic control problem

Given the ergodicity assumptions from the previous subsection we can now fully characterize the ergodic average expected costs $J(\Omega)$ from (A.7) in terms of the invariant distribution of the reflected diffusion X^Ω. In fact, we will show more: we prove that the bias of the average costs vanishes linearly in time and that their stochastic fluctuation measured in terms of the L^1-deviation from $J(\Omega)$ vanishes at square-root rate.

Theorem A.15: Let $\Omega \in \boldsymbol{\Omega}$. Then, there exist constants $C(\Omega), C'(\Omega) > 0$ that depend on Ω but are independent of $x \in \overline{\Omega}$ and $t \ge 1$ such that

$$\mathbb{E}^x\Big[\Big|\frac{1}{t}\Big(\int_0^t f(\mathsf{X}_s^\Omega)\,\mathrm{d}s + \kappa\mathsf{L}_t^\Omega\Big) - \Big(\int_\Omega f(y)\rho_\Omega(y)\mathrm{d}y + \kappa\int_{\partial\Omega}\rho_\Omega(y)\,\mathcal{H}^{d-1}(\mathrm{d}y)\Big)\Big|\Big] \le \frac{C(\Omega)}{\sqrt{t}}, \tag{A.13}$$

and

$$\left|\mathbb{E}^x\left[\frac{1}{t}\left(\int_0^t f(\mathsf{X}_s^\Omega)\,\mathrm{d}s + \kappa \mathsf{L}_t^\Omega\right)\right] - \left(\int_\Omega f(y)\rho_\Omega(y)\mathrm{d}y + \kappa\int_{\partial\Omega}\rho_\Omega(y)\,\mathcal{H}^{d-1}(\mathrm{d}y)\right)\right| \le \frac{C'(\Omega)}{t}. \tag{A.14}$$

In particular,

$$J(\Omega) = \int_\Omega f(y)\rho_\Omega(y)\mathrm{d}y + \kappa\int_{\partial\Omega}\rho_\Omega(y)\,\mathcal{H}^{d-1}(\mathrm{d}y).$$

We will need the following basic result for the proof, which we include here for the lack of a precise reference.

Lemma A.16: Let $O \subset \mathbb{R}^d$ be a bounded open set of class C^k for some $k \ge 2$ and let n be the unit inward normal vector on ∂O. Then there exists a function $\varphi \in C^k(\mathbb{R}^d)$ such that $\nabla\varphi = n$ on ∂O.

This statement is a straightforward consequence of [65, Theorem 3], which shows that the signed distance function

$$s(x) := \begin{cases} d(x, \partial O), & x \in O, \\ -d(x, \partial O), & x \in \mathbb{R}^d \setminus O, \end{cases}$$

inherits the smoothness properties of the boundary in an open neighborhood of ∂O, and observing that $\nabla s = n$ on ∂O.

Proof of Theorem A.15. We only prove the claim on $L^1(\mathbb{P}^x)$-convergence at square root rate in (A.13), the statement on convergence in expectation at linear rate in (A.14) follows from similar, but easier considerations. Since Ω is of class C^2, by Lemma A.16 we may choose $\varphi \in C^2(\mathbb{R}^d)$ such that $\nabla\varphi(x) = n(x)$ for $x \in \partial\Omega$, and we find using (A.11)

$$\int_\Omega A\varphi(x)\,\pi_\Omega(\,\mathrm{d}x) = -\int_{\partial\Omega}\rho_\Omega(x)\,\mathcal{H}^{d-1}(\,\mathrm{d}x).$$

Now by Itô's formula for $t \ge 0$ almost surely,

$$\begin{aligned}\varphi(\mathsf{X}_t^\Omega) - \varphi(\mathsf{X}_0^\Omega) &= \int_0^t A\varphi(\mathsf{X}_s^\Omega)\,\mathrm{d}s + \int_0^t \langle\nabla\varphi(\mathsf{X}_s^\Omega), n(\mathsf{X}_s^\Omega)\rangle\,\mathrm{d}\mathsf{L}_s^\Omega + \sqrt{2}\int_0^t\langle\nabla\varphi(\mathsf{X}_s^\Omega),\,\mathrm{d}\mathsf{B}_s\rangle \\ &= \int_0^t A\varphi(\mathsf{X}_s^\Omega)\,\mathrm{d}s + \mathsf{L}_t^\Omega + \sqrt{2}\int_0^t\langle\nabla\varphi(\mathsf{X}_s^\Omega),\,\mathrm{d}\mathsf{B}_s\rangle,\end{aligned}$$

where we used

$$\begin{aligned}\int_0^t\langle\nabla\varphi(\mathsf{X}_s^\Omega), n(\mathsf{X}_s^\Omega)\rangle\,\mathrm{d}\mathsf{L}_s^\Omega = \int_{\{s\le t\,:\,\mathsf{X}_s^\Omega\in\partial D\}}\langle\nabla\varphi(\mathsf{X}_s^\Omega), n(\mathsf{X}_s^\Omega)\rangle\,\mathrm{d}\mathsf{L}_s^\Omega &= \int_{\{s\le t\,:\,\mathsf{X}_s^\Omega\in\partial D\}}\mathrm{d}\mathsf{L}_s^\Omega \\ &= \mathsf{L}_t^\Omega.\end{aligned}$$

Note also that since $\nabla\varphi$ is continuous, it is bounded on $\overline{\Omega}$, and hence $\int_0^\cdot \langle \nabla\varphi(\mathsf{X}_s^\Omega), \mathrm{d}\mathsf{B}_s\rangle$ is a L^2-martingale. Combining these, we find for any $x \in \overline{\Omega}$,

$$\begin{aligned}
&\mathbb{E}^x\Big[\Big|\frac{1}{t}\mathsf{L}_t^\Omega - \int_{\partial\Omega} \rho_\Omega(x)\,\mathcal{H}^{d-1}(\,\mathrm{d}x)\Big|\Big] \\
&= \mathbb{E}^x\Big[\Big|\frac{1}{t}\Big(\varphi(\mathsf{X}_t^\Omega) - \varphi(\mathsf{X}_0^\Omega) - \int_0^t (A\varphi(\mathsf{X}_s^\Omega) - \pi_\Omega(A\varphi))\,\mathrm{d}s - \sqrt{2}\int_0^t \langle\nabla\varphi(\mathsf{X}_s^\Omega), \mathrm{d}\mathsf{B}_s\rangle\Big)\Big|\Big] \\
&\leq 2\frac{\|\varphi\|_{L^\infty(\overline{\Omega})}}{t} + \mathbb{E}^x\Big[\frac{1}{t}\Big|\int_0^t (A\varphi(\mathsf{X}_s^\Omega) - \pi_\Omega(A\varphi))\,\mathrm{d}s\Big|\Big] + \frac{\sqrt{2}}{t}\mathbb{E}^x\Big[\Big|\int_0^t \langle\nabla\varphi(\mathsf{X}_s^\Omega), \mathrm{d}\mathsf{B}_s\rangle\Big|\Big] \\
&\lesssim \frac{\|\varphi\|_{L^\infty(\overline{\Omega})} + \|A\varphi\|_{L^\infty(\overline{\Omega})}}{\sqrt{t}} + \frac{1}{t}\mathbb{E}^x\Big[\Big|\int_0^t \langle\nabla\varphi(\mathsf{X}_s^\Omega), \mathrm{d}\mathsf{B}_s\rangle\Big|\Big] \\
&\lesssim \frac{\|\varphi\|_{L^\infty(\overline{\Omega})} + \|A\varphi\|_{L^\infty(\overline{\Omega})}}{\sqrt{t}} + \frac{1}{t}\Big(\mathbb{E}^x\Big[\int_0^t |\nabla\varphi(\mathsf{X}_s^\Omega)|^2\,\mathrm{d}s\Big]\Big)^{1/2} \\
&\lesssim \frac{\|\varphi\|_{L^\infty(\overline{\Omega})} + \|A\varphi\|_{L^\infty(\overline{\Omega})} + \||\nabla\varphi|\|_{L^\infty(\Omega)}}{\sqrt{t}}
\end{aligned}$$

where we used Corollary A.14 for the second inequality and Hölder inequality together with Itô-isometry for the third inequality. Using this, another application of Corollary A.14 for the continuous cost component yields

$$\mathbb{E}^x\Big[\Big|\frac{1}{t}\Big(\int_0^t f(\mathsf{X}_s^\Omega)\,\mathrm{d}s + \kappa\mathsf{L}_t^\Omega\Big) - \Big(\int_\Omega f(y)\rho_\Omega(y)\mathrm{d}y + \kappa\int_{\partial\Omega}\rho_\Omega(y)\,\mathcal{H}^{d-1}(\mathrm{d}y)\Big)\Big|\Big] \leq \frac{C(\Omega)}{\sqrt{t}},$$

as claimed. ☐

Remark A.17: Section 4 in [73] (see eqs. (4.23), (4.25), (4.35) and (4.37)) shows that if Ω has a C^4 boundary and $V \in C^3(\mathbb{R}^d)$, then an analogous result holds for the ergodic costs (A.9) in the model with non-constant reflection costs.

Putting pieces together, we obtain a formula for $J(\Omega)$ which is just based on the stationary density ρ of the uncontrolled process if the latter is ergodic.

Corollary A.18: For any $\Omega \in \mathbf{\Omega}$, it holds that

$$J(\Omega) = \frac{1}{\int_\Omega \mathrm{e}^{-V(y)}\mathrm{d}y}\left(\int_\Omega f(y)\mathrm{e}^{-V(y)}\mathrm{d}y + \kappa\int_{\partial\Omega}\mathrm{e}^{-V(y)}\,\mathcal{H}^{d-1}(\,\mathrm{d}y)\right).$$

If, moreover, $\exp(-V) \in L^1(\mathbb{R}^d)$, then

$$J(\Omega) = \frac{1}{\int_\Omega \rho(y)\mathrm{d}y}\left(\int_\Omega f(y)\rho(y)\mathrm{d}y + \kappa\int_{\partial\Omega}\rho(y)\,\mathcal{H}^{d-1}(\,\mathrm{d}y)\right).$$

We point out that this result is a multidimensional version of [2, Lemma 2.1].

A.3 Learning the optimal boundary

We now turn to the challenging situation, when the dynamics of the unconstrained Langevin diffusion are unknown, which makes it impossible to set any optimization algorithm utilizing the dynamics into motion, without feeding it information based on collected data first. As apparent from the explicit form of the cost functional given in Corollary A.18, a natural data-driven reflection procedure can be based on a plug-in approach, provided that we have an efficient estimator of (functionals of) the invariant density of the unconstrained Langevin diffusion at our disposal.

Adaptive nonparametric estimation of the invariant density

As in the scalar case discussed in [27], we employ a kernel estimator of the invariant density, whose sup-norm risk given appropriate conditions on the diffusion coefficients is well-understood in a general context by now. In the following, we will concentrate on a class of potentials V s.t. the process satisfies certain functional inequalities. This setting is quite natural given the reversible nature of Langevin diffusions and is studied in [111], where minimax optimal estimation rates for a Lepski type *adaptive* kernel estimator are established under *anisotropic* Hölder smoothness assumptions on the invariant density. To recall these results, some preliminary definitions are necessary.

Let π be the invariant distribution of X with density $\rho \propto \exp(-V)$ and for a probability measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ let $\mathbb{P}^\mu(\cdot) := \int_{\mathbb{R}^d} \mathbb{P}^x(\cdot)\,\mu(\mathrm{d}x)$, where $\mathbb{P}^x(\cdot) = \mathbb{P}(\cdot \mid \mathsf{X}_0 = x)$ for $x \in \mathbb{R}^d$, so that under $\mathbb{P}^\mu$, the process X is started according to μ. Let also L be the $L^2(\pi)$-generator of X with domain $\mathcal{D}(\mathrm{L})$. Since $-\mathrm{L}$ is self-adjoint and nonnegative on the Hilbert space $L^2(\pi)$ endowed with the inner product $\langle f, g\rangle_\pi := \int fg\,\mathrm{d}\pi$, we may define $\sqrt{-\mathrm{L}}$ via spectral calculus and note that for any $f \in \mathcal{D}(\sqrt{-\mathrm{L}})$, we have $\|\sqrt{-\mathrm{L}}f\|_\pi^2 = -\langle \mathrm{L}f, f\rangle_\pi$, which for $f \in C_c^2(\mathbb{R}^d)$ is equal to $\|\nabla f\|_\pi^2$. We treat diffusion models satisfying the following conditions.

Definition A.19: (PI) X satisfies a *Poincaré inequality* with constant C_{P} if, for any $f \in \mathcal{D}(\sqrt{-\mathrm{L}})$,

$$\mathrm{Var}_\pi(f) := \pi(f^2) - \pi(f)^2 \le C_{\mathrm{P}}\|\sqrt{-\mathrm{L}}f\|_\pi^2.$$

(NI) X satisfies a *Nash inequality* with constants $C_{\mathrm{N}}^1, C_{\mathrm{N}}^2$ if, for any $f \in \mathcal{D}(\sqrt{-\mathrm{L}})$,

$$\|f\|_\pi^{d+2} \le \left(C_{\mathrm{N}}^1\|f\|_\pi^2 + C_{\mathrm{N}}^2\|\sqrt{-\mathrm{L}}f\|_\pi^2\right)^{d/2}\|f\|_{L^1(\pi)}^2.$$

Denote by $\Sigma(C_{\mathrm{P}}, C_{\mathrm{N}}^1, C_{\mathrm{N}}^2)$ the class of potentials $V\colon \mathbb{R}^d \to \mathbb{R}^d$ s.t. the corresponding Langevin diffusion satisfies (PI) and (NI).

The combination of a Poincaré inequality with a Nash inequality is particularly attractive from a statistical point of view. A Poincaré inequality is equivalent to exponential ergodicity in $L^2(\pi)$.

More precisely, for any $f \in L^2(\pi)$ with $\pi(f) = 0$, $\|P_t f\|_\pi \lesssim \exp(-t/C_{\mathrm{P}})\|f\|_\pi$, which enforces a fast mixing behavior of the diffusion. A Nash inequality on the other hand is equivalent to *ultracontractivity* expressed through the heat kernel bound $\|P_t\|_{L^1(\pi)\to L^\infty(\pi)} \lesssim t^{-d/2}$ for $t \in (0,1]$, cf. [8, Theorem 6.3.1]. The key observation for the statistical approach is that the combination of both yields tight variance bounds for path integrals $\int_0^t f(\mathsf{X}_s)\,\mathrm{d}s$ of functionals $f \in L^2(\pi)$, which allows for efficient control of the stochastic fluctuations of kernel estimators.

To control the bias, we impose anisotropic Hölder regularity conditions on the invariant density.

Definition A.20: Let $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d) \in (0,\infty)^d$, $\boldsymbol{\mathcal{L}} = (\mathcal{L}_1, \dots, \mathcal{L}_d) \in (0,\infty)^d$. A function $g : \mathbb{R}^d \to \mathbb{R}$ is said to belong to the anisotropic Hölder class $\mathcal{H}_d(\boldsymbol{\beta}, \boldsymbol{\mathcal{L}})$ if, for all $i = 1, \dots, d$,

$$\|D_i^k g\|_\infty \le \mathcal{L}_i, \quad k = 1, \dots, \lfloor \beta_i \rfloor,$$
$$\|D_i^{\lfloor \beta_i \rfloor} g(\cdot + te_i) - D_i^{\lfloor \beta_i \rfloor} g(\cdot)\|_\infty \le \mathcal{L}_i |t|^{\beta_i - \lfloor \beta_i \rfloor}, \quad t \in \mathbb{R},$$

where $\lfloor \beta \rfloor$ denotes the largest integer strictly smaller than $\beta > 0$ and $\|\cdot\|_\infty$ denotes the *sup*-norm on $\mathbb{R}^d$. Denote by

$$\mathbb{H}_d(\boldsymbol{\beta}, \boldsymbol{\mathcal{L}}) = \mathbb{H}_d(\boldsymbol{\beta}, \boldsymbol{\mathcal{L}};\ C_\infty, C_{\mathrm{P}}, C_{\mathrm{N}}^1, C_{\mathrm{N}}^2),$$

the set of invariant densities $\rho_V \in \mathcal{H}_d(\boldsymbol{\beta} + \mathbf{1}, \boldsymbol{\mathcal{L}})$ s.t. $\|\rho_V\|_\infty \le C_\infty$ and $V \in \Sigma(C_{\mathrm{P}}, C_{\mathrm{N}}^1, C_{\mathrm{N}}^2)$.

Let $K : \mathbb{R} \to \mathbb{R}$ be a symmetric Lipschitz kernel function with $\operatorname{supp}(K) \subset [-1/2, 1/2]$ and $\int_{\mathbb{R}} K(x)\,\mathrm{d}x = 1$. We say that K is of order $\ell \in \mathbb{N}$ if $\int_{\mathbb{R}} x^m K(x)\,\mathrm{d}x = 0$ for any $m = 1, \dots, \ell$. For $h > 0$ we let $K_h(\cdot) := h^{-1}K(\cdot/h)$ and for $\boldsymbol{h} \in (0,\infty)^d$ we set

$$\mathbb{K}_{\boldsymbol{h}}(x) := \prod_{i=1}^d K_{h_i}(x_i), \quad x \in \mathbb{R}^d.$$

We now define the following kernel estimators given a continuous record $(\mathsf{X}_t)_{t\in[0,T]}$ s.t. $\rho_V \in \mathbb{H}_d(\boldsymbol{\beta}, \boldsymbol{L})$:

$$\widehat{\rho}_{\boldsymbol{h},T}(x) := \frac{1}{T}\int_0^T \mathbb{K}_{\boldsymbol{h}}(x - \mathsf{X}_s)\,\mathrm{d}s.$$

In order to efficiently estimate ρ_V via $\widehat{\rho}_{\boldsymbol{h},T}$, the bandwidth $\boldsymbol{h}$ has to be carefully chosen to achieve an optimal balance between bias and variance of the kernel estimator. If the Hölder smoothness parameter $\boldsymbol{\beta}$ is unknown, the bias cannot be evaluated directly, which poses the fundamental challenge to design a fully data-driven/adaptive bandwith selection procedure to obtain a rate-optimal but possibly random bandwidth $\widehat{\boldsymbol{h}}_T$.

In dimension $d = 2$ this problem is significantly simplified since then a tight variance bound of the kernel estimator only depends logarithmically on the bandwidth. Consequently, the smoothness independent and deterministic bandwith choice

$$\widehat{\boldsymbol{h}}_T \sim T^{-1/2}(1,1),$$

yields the optimal sup-norm estimation rate $\log T/\sqrt{T}$ given that the order of K is chosen large enough. In dimension $d \geq 3$, the situation is more involved and the bandwidth $\boldsymbol{h}_T = \widehat{\boldsymbol{h}}_T^{(\mathfrak{q})}$, depending on a parameter $\mathfrak{q} \geq 1$ that ensures $\mathsf{L}^{\mathfrak{q}}$-convergence at the minimax optimal rate, is chosen according Lepski type selection rule. The details are given in the supplement.

Finally, for $\mathfrak{q} \geq 1$, $d \geq 2$ and $\widehat{\boldsymbol{h}}_T = \widehat{\boldsymbol{h}}_T^{(\mathfrak{q})}$ given as above, we set

$$\widehat{\rho}_{\widehat{\boldsymbol{h}}_T,T}(x) = \frac{1}{T}\int_0^T \mathbb{K}_{\widehat{\boldsymbol{h}}_T}(x - \mathsf{X}_s)\,\mathrm{d}s, \quad x \in \mathbb{R}^d.$$

According to the discussion in [111, Section 3] on the two-dimensional case and [111, Theorem 3.4], we now have the following uniform sup-norm estimation result.

Theorem A.21: Suppose $\boldsymbol{\beta} \in (0,\mathfrak{b}]^d$ for some $\mathfrak{b} \in \mathbb{N} \cap [2,\infty)$ and let K have order $\mathfrak{b}+1$. Then, for any $\mathfrak{q} \geq 1$ and $\boldsymbol{\mathcal{L}} \in (0,\infty)^d$ it holds

$$\sup_{\rho_V \in \mathbb{H}_d(\boldsymbol{\beta},\boldsymbol{\mathcal{L}})} \left(\mathbb{E}^\pi\left[\left\|\widehat{\rho}_{\widehat{\boldsymbol{h}}_T,T} - \rho_V\right\|_\infty^{\mathfrak{q}}\right]\right)^{1/\mathfrak{q}} = \mathcal{O}(\Psi_{d,\boldsymbol{\beta}}(T)),$$

where for the harmonic mean smoothness $\overline{\boldsymbol{\beta}+\boldsymbol{1}} := (d^{-1}\sum_{i=1}^d \frac{1}{\beta_i+1})^{-1}$, the rate $\Psi_{d,\boldsymbol{\beta}}$ is specified by

$$\Psi_{d,\boldsymbol{\beta}}(T) := \begin{cases} \frac{\log T}{\sqrt{T}}, & d = 2, \\ \left(\frac{\log T}{T}\right)^{\frac{\overline{\boldsymbol{\beta}+1}}{2\overline{\boldsymbol{\beta}+1}+d-2}}, & d \geq 3. \end{cases}$$

Data-driven estimation of the optimal reflection boundary

In this section we consider a set of domains $\Theta \subset \Omega$ endowed with some topology such that the set of minimizers $\operatorname{argmin}_{\Omega\in\Theta} J(\Omega)$ is well-defined and let $\Omega^* \in \operatorname{argmin}_{\Omega\in\Theta} J(\Omega)$. For a Borel measurable function $h\colon \Theta \to \mathbb{R}$ and a Θ-valued random variable Z we write $\mathbb{E}[h(\mathsf{Z})] = \int_\Theta h(\Omega)\,\mathbb{P}(\mathsf{Z} \in \mathrm{d}D)$. Our data-driven procedure to determine reflection domains $\widehat{\Omega}$ whose average costs are close to the optimal ergodic costs $J(\Omega^*)$ uses the following assumption.

Assumption A.22: (i) For some constants $\underline{\lambda}, \overline{\lambda}, \Lambda$ it holds $B(0,\underline{\lambda}) \subset D^* \subset B(0,\overline{\lambda})$ and $\mathcal{H}^{d-1}(\partial D^*) \leq \Lambda$;

(ii) we are given information on a constant $\underline{\rho} > 0$ such that $\underline{\rho} \leq \inf_{B(0,\overline{\lambda})} \rho$.

Note that such a constant $\underline{\rho}$ always exists, because V is locally bounded by assumption and therefore $\rho \propto \mathrm{e}^{-V}$ is bounded away from 0 on every bounded set. Accordingly, we define the truncated invariant density estimator $\widehat{\rho}^*_{T,q}$ based on data $(\mathsf{X}_t)_{t\in[0,T]}$ of the uncontrolled diffusion process as

$$\widehat{\rho}^*_{T,q}(x) := \widehat{\rho}_{T,q}(x) \vee \underline{\rho}, \quad x \in \mathbb{R}^d,$$

with $\widehat{\rho}_{T,q}$ specified as the adaptive invariant density estimator from the previous subsection with bandwidth choice $\boldsymbol{h} = \widehat{\boldsymbol{h}}_T^{(q)}$ for $q \geq 1$. Moreover, we let $\Theta(\underline{\lambda}, \overline{\lambda}, \Lambda) \subset \Theta$ be the subfamily of reflection domains satisfying Assumption A.22.(i). Let

$$\widehat{J}_{T,q}(\Omega) := \frac{1}{\int_\Omega \widehat{\rho}^*_{T,q}(x)\,\mathrm{d}x}\Big(\int_\Omega f(y)\widehat{\rho}^*_{T,q}(y)\,\mathrm{d}y + \kappa \int_{\partial\Omega} \widehat{\rho}^*_{T,q}(y)\,\mathcal{H}^{d-1}(\,\mathrm{d}y)\Big), \quad D \in \Theta,$$

be the estimator of the asymptotic costs associated to the reflection domain Ω and define the reflection domain estimator

$$\widehat{\Omega}_{T,q} \in \operatorname*{argmin}_{\Omega\in\Theta(\underline{\lambda},\overline{\lambda},\Lambda)} \widehat{J}_{T,q}(\Omega).$$

Here, we must assume that the topological space $\Theta \subset \mathbf{\Omega}$ is sufficiently nice to allow a measurable choice of $\widehat{\Omega}_{T,q}$ considered as a random mapping into the Borel space $(\Theta, \mathcal{B}(\Theta))$ associated to Θ. We now have the following concentration result for the simple regret.

Proposition A.23: Suppose that $\mathsf{X}_0 \sim \mu$, where $\mu \ll \pi$ with $\|\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\|_{L^q(\pi)} < \infty$ for some $q \in (1, \infty]$. Then, for any $p \geq 1$, given the assumptions from Theorem A.21 we have the regret bound

$$\mathbb{E}^\mu\Big[\big|J\big(\Omega^*\big) - J\big(\widehat{\Omega}_{T,p\overline{q}}\big)\big|^p\Big]^{1/p} \leq C\|\tfrac{\mathrm{d}\mu}{\mathrm{d}\pi}\|^{1/p}_{L^q(\pi)}\Psi_{d,\beta}(T),$$

where $\overline{q} := q/(q-1)$ is the conjugate Hölder exponent of q and C depends on $\kappa, p, q, f, \|\rho\|_\infty$ and the constants from Assumption A.22.

Proof. For fixed $\Omega \in \Theta(\underline{\lambda}, \overline{\lambda}, \Lambda)$, write (in obvious notation)

$$J(\Omega) = A(\Omega)/B(\Omega), \quad \widehat{J}_{T,p\overline{q}}(\Omega) = \widehat{A}_T(\Omega)/\widehat{B}_T(\Omega).$$

Using

$$B(\Omega) \wedge \widehat{B}_T(\Omega) \geq \int_{B(0,\underline{\lambda})} \rho(x) \wedge \widehat{\rho}^*_{T,p\overline{q}}(x)\,\mathrm{d}x \geq \underline{\rho}\lambda_d(B(0,\underline{\lambda})) =: \varpi,$$

it follows ,

$$\begin{aligned}\big|J(\Omega) - \widehat{J}_{T,p\overline{q}}(\Omega)\big| &\leq \frac{A(\Omega)}{B(\Omega)\widehat{B}_T(\Omega)}\big|\widehat{B}_T(\Omega) - B(\Omega)\big| + \frac{|\widehat{A}_T(\Omega) - A(\Omega)|}{\widehat{B}_T(\Omega)} \\ &\leq \frac{\|f\|_{L^1(\pi)} + \kappa\|\rho\|_\infty\Lambda}{\varpi^2}\|\rho - \widehat{\rho}^*_{T,p\overline{q}}\|_{L^\infty(B(0,\overline{\lambda}))} \\ &\quad + \frac{\|f\|_{L^1(B(0,\overline{\lambda}))} + \kappa\Lambda}{\varpi}\|\rho - \widehat{\rho}^*_{T,p\overline{q}}\|_{L^\infty(B(0,\overline{\lambda}))}.\end{aligned}$$

Thus, for some constant $C = C(\kappa, \underline{\lambda}, \overline{\lambda}, \Lambda, \underline{\rho}, \|\rho\|_\infty, p, f)$,

$$\begin{aligned}\mathbb{E}^\mu\Big[\sup_{\Omega\in\Theta(\underline{\lambda},\overline{\lambda},\Lambda)} |J(\Omega) - \widehat{J}_{T,p\overline{q}}(\Omega)|^p\Big]^{1/p} &\leq C\mathbb{E}^\mu\big[\|\rho - \widehat{\rho}^*_{T,p\overline{q}}\|^p_{L^\infty(B(0,\overline{\lambda}))}\big]^{1/p} \\ &\leq C\|\tfrac{\mathrm{d}\mu}{\mathrm{d}\pi}\|^{1/p}_{L^q(\pi)}\mathbb{E}^\pi\big[\|\rho - \widehat{\rho}^*_{T,p\overline{q}}\|^{p\overline{q}}_{L^\infty(B(0,\overline{\lambda}))}\big]^{1/p\overline{q}} \\ &\leq C\|\tfrac{\mathrm{d}\mu}{\mathrm{d}\pi}\|^{1/p}_{L^q(\pi)}\mathbb{E}^\pi\big[\|\rho - \widehat{\rho}_{T,p\overline{q}}\|^{p\overline{q}}_{L^\infty(B(0,\overline{\lambda}))}\big]^{1/p\overline{q}} \\ &\lesssim_{p,q} C\|\tfrac{\mathrm{d}\mu}{\mathrm{d}\pi}\|^{1/p}_{L^q(\pi)}\Psi_{d,\boldsymbol{\beta}}(T),\end{aligned} \tag{A.15}$$

where for the second line we used Hölder inequality twice and for the third line used that $|\widehat{\rho}^*_{T,p\overline{q}}(x) - \rho(x)| \leq |\widehat{\rho}_{T,p\overline{q}}(x) - \rho(x)|$ for $x \in B(0, \overline{\lambda})$ since $\rho \geq \underline{\rho}$ on $B(0, \overline{\lambda})$. The final inequality follows from Theorem A.21. Finally,

$$\mathbb{E}^\mu\big[\big|J\big(\Omega^*\big) - J\big(\widehat{\Omega}_{T,p\overline{q}}\big)\big|^p\big]^{1/p} \leq 2\mathbb{E}^\mu\Big[\sup_{\Omega\in\Theta(\underline{\lambda},\overline{\lambda},\Lambda)} |J(\Omega) - \widehat{J}_{T,p\overline{q}}(\Omega)|^p\Big]^{1/p}, \tag{A.16}$$

where we used that since $\Omega^* \in \operatorname{argmin}_{\Omega\in\Theta(\underline{\lambda},\overline{\lambda},\Lambda)} J(\Omega)$ and $\widehat{\Omega}_{T,p\overline{q}} \in \operatorname{argmin}_{\Omega\in\Theta(\underline{\lambda},\overline{\lambda},\Lambda)} \widehat{J}_{T,p\overline{q}}(\Omega)$, we have

$$\begin{aligned}|J(\Omega^*) - J(\widehat{\Omega}_{T,p\overline{q}})| = J(\widehat{\Omega}_{T,p\overline{q}}) - J(\Omega^*) &\leq J(\widehat{\Omega}_{T,p\overline{q}}) - J(\Omega^*) + \widehat{J}_{T,p\overline{q}}(\Omega^*) - \widehat{J}_{T,p\overline{q}}(\widehat{\Omega}_{T,p\overline{q}}) \\ &\leq 2 \sup_{\Omega\in\Theta(\underline{\lambda},\overline{\lambda},\Lambda)} |J(\Omega) - \widehat{J}_{T,p\overline{q}}(\Omega)|.\end{aligned}$$

Combining (A.15) and (A.16) yields the claim. □

This result may be interpreted in two different ways. On the one hand, it shows for the generic situation, where the controller has access to a separate diffusion data sample and uses it in online estimation of an optimal reflection boundary, that the regret vanishes at the nonparametric estimation rate. On the other hand, it demonstrates that a simple explore-then-commit strategy, where we first estimate an optimal set for T time units and afterwards exploit by reflecting the process at the estimated boundaries, yields a regret bounded by $\Psi_{d,\beta}(T)$.

Episodic domain learning Aiming now at strategies with sublinear regret rates without any simplifying assumptions on the data-collection mechanism, we face a classical exploration-exploitation dilemma in light of the necessity to simultaneously control the process and estimate its dynamics over time. The key bound from Proposition A.23 will allow us to do so.

Our episodic learning algorithm separates the time-line into exploration and exploitation phases. Let T_i be the start of the i-th exploration period, where we let the diffusion run freely without reflection and S_i be the start of the i-th exploitation period, where we reflect the process according to an estimate of the optimal reflection boundary based on past observations of the exploration process. We always start with an exploration period, i.e. $T_1 = 0$, and then alternate between exploration and exploitation periods. We denote by $\tau_i = S_i - T_i$ the length of the i-th exploration period and by $\sigma_i = T_{i+1} - S_i$ the length of the i-th exploitation period.

Contrary to the scalar diffusion case in [28], the multivariate diffusion does not hit points, which makes it difficult to introduce an appropriate life-cycle decomposition of the exploration process that allows for elegant renewal theoretic arguments in the analysis. Instead we choose sequences $(a_i) \subset [1,\infty)^{\mathbb{N}}$, $(b_i) \subset [1,\infty)^{\mathbb{N}}$ and simply let $S_i = \inf\{t \geq T_i + a_i : \widetilde{\mathsf{X}}_t \in \mathrm{cl}(B(0,\underline{\lambda}))\}$ and $T_{i+1} = S_i + b_i$, where $(\widetilde{\mathsf{X}}_t)_{t\geq 0}$ denotes the process that is controlled according to the above strategy. This implies that the i-th exploitation length is deterministically given by $\sigma_i = b_i$ and given sufficiently strong recurrent behavior of the process the i-th exploration length is comparable to a_i in the sense that $\sup_i \mathbb{E}[\tau_i - a_i] < \infty$. Such recurrence properties will be enforced by the drift condition (A.17) below. Moreover, the strategy makes sure that at the start of an exploitation period we have $\widetilde{\mathsf{X}}_{S_i} \in \mathrm{cl}\, B(0,\underline{\lambda}) \subset \overline{\Omega}$ for any $\Omega \in \Theta(\underline{\lambda}, \overline{\lambda}, \Lambda)$. For the estimator of the reflection boundary $\widehat{\Omega}_i$ in the i-th exploitation period, we only take into account the observations from the last exploration period on the time interval $[T_i, T_i + a_i]$ by letting

$$\widehat{\rho}_i^*(x) := \widehat{\rho}_{i,2}(x) \vee \underline{\rho}, \quad x \in \mathbb{R}^d,$$

where $\widehat{\rho}_{i,2}$ is the adaptive invariant density estimator based on diffusion data $(\widetilde{\mathsf{X}}_t)_{t\in[T_i,T_i+a_i]}$ for the parameter choice $\mathfrak{q} = 2$ in the construction of the stochastic bandwidth, and then set

$$\widehat{\Omega}_i \in \operatorname*{argmin}_{\Omega\in\Theta(\underline{\lambda},\overline{\lambda},\Lambda)} \frac{1}{\int_\Omega \widehat{\rho}_i^*(y)\,\mathrm{d}y} \Big(\int_\Omega f(y)\widehat{\rho}_i^*(y)\,\mathrm{d}y + \kappa \int_{\partial\Omega} \widehat{\rho}_i^*(y)\,\mathcal{H}^{d-1}(\mathrm{d}y) \Big).$$

Let us remark here that our choice to take only observations from the previous exploration period into account is for technical reasons only because it allows us to exploit the simple regret bound from Proposition A.23 directly. In order to obtain an asymptotic sublinear regret, this will require increasing exploration lengths, see the doubling strategy in Corollary A.25. Alternative strategies that are based on the full exploration history and that may allow to avoid long single exploration periods would of course be desirable, but appear very challenging because of the lack of appropriate renewal structures in the multivariate case. The following should therefore be regarded as a verification result for the existence of a fully data-driven strategy that solves the exploration-exploitation tradeoff. The development of strategies with more desirable properties from a practical perspective will be left to future work.

Let $\widetilde{\mathbb{F}} = (\widetilde{\mathcal{F}}_t)_{t\geq 0}$ be the filtration generated by the controlled process $\widetilde{\mathsf{X}}$ and set

$$\widetilde{C}_{a,b} := \int_a^b f(\widetilde{\mathsf{X}}_s)\,\mathrm{d}s + \kappa(\widetilde{\mathsf{L}}_b - \widetilde{\mathsf{L}}_a),$$

as the costs on the time interval $[a, b]$ associated to $\widetilde{\mathsf{X}}$, where $\widetilde{\mathsf{L}}$ is the local time on the reflection boundaries during the exploitation phases and is set equal to zero during the exploration phases. We also let $C_{a,b}(x, D)$ be the costs on the time interval $[a, b]$ associated to a Langevin diffusion $Z^{x,\Omega}$ that is driven by a Brownian motion independent of $\widetilde{\mathbb{F}}$ and that is reflected in Ω and is started in x. Denote by $\tau(x, D)$ its first hitting time of $\mathrm{cl}(B(0, \underline{\lambda}))$. Furthermore, we set $n(T) := \min\{i \in \mathbb{N} : \sum_{j=1}^{i}(a_j + b_j) \geq T\}$ and note that

$$n(T) \geq \min\Big\{ i \in \mathbb{N} : \sum_{j=1}^{i}(\tau_j + \sigma_j) \geq T \Big\} = \min\{i \in \mathbb{N} : T_{i+1} \geq T\} \geq \min\{i \in \mathbb{N} : T_i \geq T\} - 1,$$

which in particular implies that $S_{n(T)+1} > T_{n(T)+1} \geq T$. For technical reasons, we assume that the potential V satisfies the following drift condition: for some constants $r, M > 0$,

$$\forall |x| \geq M : \langle \nabla V(x), x/|x| \rangle \geq r. \tag{A.17}$$

Due to the specific structure of the generator, it is well known, see e.g. [7], that the Langevin diffusion then satisfies a Poincaré inequality (PI) and that its generator has a Lyapunov function $G \geq 1$ that is locally bounded, is for some $a, R > 0$ given by $G(x) = \exp(a|x|)$ for all $|x| \geq R$, and satisfies $\pi(G) < \infty$. This implies, see [38, Theorem 5.2, Theorem 7.2], that X is G-uniformly ergodic in the sense that for some constant $b > 0$,

$$\sup_{|g|\leq G} |P_t g(x) - \pi(g)| \lesssim G(x)\exp(-bt), \quad x \in \mathbb{R}^d, \tag{A.18}$$

and that for any set B s.t. $\lambda_d(B) > 0$, it holds

$$\mathbb{E}^x\Big[\int_0^{\tau_B} G(\mathsf{X}_s)\,\mathrm{d}s \Big] \leq c(B)G(x), \quad x \in \mathbb{R}^d, \tag{A.19}$$

where $c(B)$ is a constant depending on B and τ_B is the first hitting time of B. We also need some assumptions on the set of viable reflection domains $\Theta(\underline{\lambda}, \overline{\lambda}, \Lambda)$ that allow sufficient uniform bounds in the following. More precisely, we assume the constants $C'(\Omega)$ appearing in Theorem A.15 to be uniformly bounded in Ω, that is,

$$\sup_{\Omega \in \Theta(\underline{\lambda}, \overline{\lambda}, \Lambda)} C'(\Omega) < \infty. \tag{A.20}$$

Note that this assumption boils down to uniform lower bounds on the transition densities, cf. Lemma A.13, and uniform bounds on the maxima of the functions φ_Ω and their partial derivatives described in Lemma A.16, i.e., certain uniform regularity assumptions on the boundaries.

Moreover, we also require a uniform upper bound on the transition densities in the form

$$\sup_{\Omega \in \Theta(\underline{\lambda}, \overline{\lambda}, \Lambda)} \sup_{t \geq 1, (x,y) \in \overline{\Omega}^2} p_t^{\Omega}(x,y) < \infty. \tag{A.21}$$

From the Gaussian upper bound on $p_t^{\Omega}(x,y)$ given in [92, Corollary 6.15] we know that there exists a constant $c(\Omega)$ such that for a.e. $(x,y) \in \overline{\Omega}^2$ we have $\sup_{t \geq 1} p_t^{\Omega}(x,y) \leq c(\Omega)$. By continuity of $(x,y) \mapsto p_t^{\Omega}(x,y)$ for any $t > 0$ it therefore follows that $\sup_{t \geq 1, (x,y) \in \overline{\Omega}^2} p_t^{\Omega}(x,y) < \infty$ for any $\Omega \in \Theta(\underline{\lambda}, \overline{\lambda}, \Lambda)$. Verifying the uniform bounds (A.20) and (A.21) is highly problem specific and is a difficult task with the tools available in the literature when $\Theta(\underline{\lambda}, \overline{\lambda}, \Lambda)$ is infinite. Still, these assumptions are not unreasonable provided that appropriate uniform regularity conditions on the boundaries are in force. With this technical preparation we can prove our final theorem.

Theorem A.24: Suppose that the non-reflected Langevin diffusion satisfies (A.17) and (NI) and that its invariant density satisfies $\rho_V \in \mathcal{H}_d(\boldsymbol{\beta} + \mathbf{1}, \boldsymbol{\mathcal{L}})$. Assume also that $f \in L^2(\pi)$, $f \lesssim G$, that the initial distribution of the first exploration phase μ satisfies $\mu \ll \pi$ and $\mathrm{d}\mu/\mathrm{d}\pi \in L^2(\pi)$ and that (A.20) and (A.21) hold. Then, the average regret per time unit is bounded by

$$\frac{1}{T}\mathbb{E}[\widetilde{C}_{0,T}] - J(\Omega^*) \leq \frac{C}{T}\Big(\sum_{i=1}^{n(T)} a_i + \sum_{i=1}^{n(T)} b_i \Psi_{d,\beta}(a_i)\Big),$$

where C depends on $\kappa, f, \|\rho\|_\infty$ and the constants from (A.20), (A.21) and Assumption A.22.

Proof. Without loss of generality, let $T \geq 1$. Throughout the proof we write $a \lesssim b$ if $a \leq cb$ for some constant c that may depend on $\kappa, f, \|\rho\|_\infty$ and the constants from (A.20), (A.21) and Assumption A.22. Using that the costs are nonnegative, we have

$$\begin{aligned}
\mathbb{E}[\widetilde{C}_{0,T}] &= \mathbb{E}\Big[\sum_{T_i \leq T} \widetilde{C}_{T_i, S_i \wedge T}\Big] + \mathbb{E}\Big[\sum_{S_i \leq T} \widetilde{C}_{S_i, T_{i+1} \wedge T}\Big] \leq \sum_{i=1}^{n(T)} \mathbb{E}[\widetilde{C}_{T_i, S_i}] + \sum_{i=1}^{n(T)} \mathbb{E}[\widetilde{C}_{S_i \wedge T, T_{i+1} \wedge T}] \\
&\leq \sum_{i=1}^{n(T)} \Big(\mathbb{E}\big[C_{0,a_i}(\widetilde{\mathsf{X}}_{T_i}, \mathbb{R}^d)\big] + \mathbb{E}\big[C_{0,\tau(\widetilde{\mathsf{X}}_{T_i+a_i}, \mathbb{R}^d)}(\widetilde{\mathsf{X}}_{T_i+a_i}, \mathbb{R}^d)\big]\Big) + \sum_{i=1}^{n(T)-1} \mathbb{E}\big[C_{0,b_i}(\widetilde{\mathsf{X}}_{S_i}, \widehat{\Omega}_i)\big] \\
&\qquad + \mathbb{E}\big[C_{0,(T - S_{n(T)}) \vee 1}(\widetilde{\mathsf{X}}_{S_{n(T)}}, \widehat{\Omega}_{n(T)})\big].
\end{aligned} \tag{A.22}$$

We start by bounding the second term, associated to the exploitation periods until the $(n(T)-1)$-th episode. By conditioning on $\widetilde{\mathcal{F}}_{S_i}$ we see that

$$\begin{aligned}
&\Big|\mathbb{E}\big[C_{0,b_i}(\widetilde{\mathsf{X}}_{S_i}, \widehat{\Omega}_i)\big] - b_i \mathbb{E}[J(\widehat{\Omega}_i)]\Big| \\
&\leq b_i \int_{\Theta(\underline{\lambda}, \overline{\lambda}, \Lambda)} \int_{\mathrm{cl}\, B(0, \underline{\lambda})} \Big|\mathbb{E}^x\Big[\frac{1}{b_i}\int_0^{b_i} f(\mathsf{X}_s^{\Omega})\,\mathrm{d}s + \kappa \mathsf{L}_{b_i}^{\Omega} - J(\Omega)\Big]\Big| \mathbb{P}(\widetilde{\mathsf{X}}_{S_i} \in \mathrm{d}x, \widehat{\Omega}_i \in \mathrm{d}D) \\
&\leq \sup_{\Omega \in \Theta(\underline{\lambda}, \overline{\lambda}, \Lambda)} C'(\Omega) < \infty,
\end{aligned} \tag{A.23}$$

where the last two lines follow from Theorem A.15 and (A.20). Observe now that on the previous data collection interval $[T_i, T_i + a_i]$ the process $\widetilde{\mathsf{X}}$ is equal in law to the Langevin diffusion with potential V started according to the law $\mathbb{P}_{\widetilde{\mathsf{X}}_{T_i}}$. For $i = 1$, the latter has, by assumption on μ, a Radon–Nikodym derivative w.r.t. the invariant distribution π that lies in $L^2(\pi)$. Extend the transition densities p_t^Ω from $\overline{\Omega}^2$ to $\mathbb{R}^d \times \mathbb{R}^d$ by setting $p_t^\Omega(x, y) = 0$ for $x, y \notin \overline{\Omega}$. For $i \geq 2$, we then observe that for any $y \in \operatorname{cl}(B(0, \overline{\lambda}))$ we have

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}(\widetilde{\mathsf{X}}_{T_i} \in \cdot)}{\mathrm{d}\lambda_d}(y) &= \int_{\Theta(\underline{\lambda},\overline{\lambda},\Lambda)} \int_{\operatorname{cl} B(0,\underline{\lambda})} p_{b_i}^\Omega(x, y)\, \mathbb{P}(\widetilde{\mathsf{X}}_{S_{i-1}} \in \mathrm{d}x, \widehat{\Omega}_i \in \mathrm{d}D) \\ &\leq \sup_{\Omega\in\Theta(\underline{\lambda},\overline{\lambda},\Lambda)} \sup_{t\geq 1,(x,y)\in\overline{\Omega}^2} p_t^\Omega(x, y) < \infty,\end{aligned}$$

where we used (A.21). Since the Lebesgue density of π is bounded away from zero on $B(0, \overline{\lambda})$, this now implies

$$\sup_{i\in\mathbb{N}} \Big\| \frac{\mathrm{d}\mathbb{P}(\widetilde{\mathsf{X}}_{T_i}\in\cdot)}{\mathrm{d}\pi} \Big\|_{L^2(\pi)} < \infty. \tag{A.24}$$

Proposition A.23 therefore yields that

$$\mathbb{E}\big[|J(\Omega^*) - J(\widehat{\Omega}_i)|\big] \lesssim \Psi_{d,\beta}(a_i).$$

In summary, the above estimates and the triangle inequality therefore yield

$$\sum_{i=1}^{n(T)-1} \mathbb{E}\big[C_{0,b_i}(\widetilde{\mathsf{X}}_{S_i}, \widehat{\Omega}_i)\big] \lesssim n(T) + \sum_{i=1}^{n(T)-1} b_i(\Psi_{d,\beta}(a_i) + J(\Omega^*)) \tag{A.25}$$

We turn to the last summand in (A.22). Similarly to (A.23) we see that

$$\begin{aligned}&\Big|\mathbb{E}\big[C_{0,(T-S_{n(T)})\vee 1}(\widetilde{\mathsf{X}}_{S_{n(T)}}, \widehat{\Omega}_{n(T)})\big] - \mathbb{E}\big[((T - S_{n(T)}) \vee 1)J(\widehat{\Omega}_{n(T)})\big]\Big| \\ &\leq \sup_{\Omega\in\Theta(\underline{\lambda},\overline{\lambda},\Lambda)} C'(\Omega) < \infty,\end{aligned} \tag{A.26}$$

and using (A.24) and Proposition A.23 we have

$$\begin{aligned}\mathbb{E}\big[((T - S_{n(T)}) \vee 1)|J(\Omega^*) - J(\widehat{\Omega}_{n(T)})|\big] &\leq b_{n(T)}\mathbb{E}\big[|J(\Omega^*) - J(\widehat{\Omega}_{n(T)})|\big] \\ &\lesssim b_{n(T)}\Psi_{d,\beta}(a_{n(T)}).\end{aligned} \tag{A.27}$$

Combining (A.25), (A.26) and (A.27) we finally arrive at

$$\begin{aligned}
&\sum_{i=1}^{n(T)-1} \mathbb{E}\big[C_{0,b_i}(\widetilde{\mathsf{X}}_{S_i}, \widehat{\Omega}_i)\big] + \mathbb{E}\big[C_{0,(T-S_{n(T)})\vee 1}(\widetilde{\mathsf{X}}_{S_{n(T)}}, \widehat{\Omega}_{n(T)})\big] \\
&\leq n(T) + \sum_{i=1}^{n(T)} b_i \Psi_{d,\beta}(a_{n(T)}) + J(\Omega^*)\Big(\sum_{i=1}^{n(T)-1} b_i + \mathbb{E}[(T - S_{n(T)}) \vee 1]\Big) \qquad (A.28)\\
&\leq n(T) + TJ(\Omega^*) + \sum_{i=1}^{n(T)} b_i \Psi_{d,\beta}(a_i).
\end{aligned}$$

Let us now treat the exploration periods. Recall that since the unreflected diffusion satisfies a Poincaré inequality, we have exponentially fast convergence of its semigroup in $L^2(\pi)$. Since moreover $f \in L^2(\pi)$, the combined statements of Theorem 3.1 and Corollary 3.2 in [19] imply that for $\widetilde{f} := f - \mu(f)$ and $g := -\int_0^\infty P_s \widetilde{f}\, \mathrm{d}s$, we have $g \in \mathcal{D}(\mathrm{L})$ and

$$\Big\| \int_0^t (f - \pi(f))(\mathsf{X}_s)\,\mathrm{d}s \Big\|^2_{L^2(\mathbb{P}^\pi)} \leq Ct \| \sqrt{-\mathrm{L}} g \|^2_\pi,$$

for some constant C that is independent of $t \geq 0$. Using (A.24) and Hölder inequality yields

$$\begin{aligned}
\Big|\mathbb{E}\big[C_{0,a_i}(\widetilde{\mathsf{X}}_{T_i}, \mathbb{R}^d)\big] - a_i \pi(f)\Big| &= \Big|\mathbb{E}\Big[\mathbb{E}^{\widetilde{\mathsf{X}}_{T_i}}\Big[\int_0^{a_i} (f - \pi(f))(\mathsf{X}_s)\,\mathrm{d}s\Big]\Big]\Big| \\
&\leq \int_{\mathbb{R}^d} \mathbb{E}^x\Big[\Big| \int_0^{a_i} (f - \pi(f))(\mathsf{X}_s)\,\mathrm{d}s\Big|\Big] \frac{\mathrm{d}\mathbb{P}(\widetilde{\mathsf{X}}_{T_i} \in \cdot)}{\mathrm{d}\pi}(x)\,\pi(\mathrm{d}x) \\
&\leq \Big\| \frac{\mathrm{d}\mathbb{P}(\widetilde{\mathsf{X}}_{T_i} \in \cdot)}{\mathrm{d}\pi} \Big\|_{L^2(\pi)} \Big\| \int_0^{a_i} (f - \pi(f))(\mathsf{X}_s)\,\mathrm{d}s \Big\|_{L^2(\mathbb{P}^\pi)} \\
&\leq \widetilde{C} \sqrt{a_i}.
\end{aligned}$$

It therefore follows by the triangle inequality that

$$\sum_{i=1}^{n(T)} \mathbb{E}\big[C_{0,a_i}(\widetilde{\mathsf{X}}_{T_i}, \mathbb{R}^d)\big] \lesssim \sum_{i=1}^{n(T)} a_i. \qquad (A.29)$$

Furthermore, using (A.19) and the assumption $f \lesssim G$, we can write

$$\begin{aligned}
\mathbb{E}\big[C_{0,\tau(\widetilde{\mathsf{X}}_{T_i+a_i},\mathbb{R}^d)}(\widetilde{\mathsf{X}}_{T_i+a_i}, \mathbb{R}^d)\big] &= \int_{\mathbb{R}^d} \mathbb{E}^x\Big[\int_0^{\tau_{\mathrm{cl}(B(0,\underline{\lambda}))}} f(\mathsf{X}_s)\,\mathrm{d}s\Big]\,\mathbb{P}(\widetilde{\mathsf{X}}_{T_i+a_i} \in \mathrm{d}x) \\
&\lesssim \mathbb{E}\big[G(\widetilde{\mathsf{X}}_{T_i+a_i})\big].
\end{aligned}$$

Now since $\widetilde{\mathsf{X}}_{T_i+a_i}$ has the same law as X_{a_i} with X started according to the law $\mathbb{P}_{\widetilde{\mathsf{X}}_{T_i}}$ which, by construction, is supported on $\mathrm{cl}(B(0,\overline{\lambda}))$, it follows from (A.18) that

$$\begin{aligned}\mathbb{E}\big[G(\widetilde{\mathsf{X}}_{T_i+a_i})\big] &\leq \sup_{x\in\mathrm{cl}(B(0,\overline{\lambda}))} \mathbb{E}^x[G(\mathsf{X}_{a_i})]\\ &\leq \pi(G) + \sup_{x\in\mathrm{cl}(B(0,\overline{\lambda}))} |P_{a_i}G(x) - \pi(G)| \leq \pi(G) + C\|G\|_{L^\infty(\mathrm{cl}(B(0,\overline{\lambda})))} < \infty,\end{aligned}$$

for some constant independent of $i \in \mathbb{N}$. This allows us to conclude that

$$\sum_{i=1}^{n(T)} \mathbb{E}\big[C_{0,\tau(\widetilde{\mathsf{X}}_{T_i+a_i},\mathbb{R}^d)}(\widetilde{\mathsf{X}}_{T_i+a_i},\mathbb{R}^d)\big] \leq \big(\pi(G) + C\|G\|_{L^\infty(\mathrm{cl}(B(0,\overline{\lambda})))}\big)n(T). \tag{A.30}$$

Taking together the bounds (A.28), (A.29) and (A.30) proves the assertion. □

As a consequence, we obtain the following explicit rates for a strategy that doubles exploration times and chooses subsequent exploitation times inverse proportionally to the nonparametric estimation rate.

Corollary A.25: Let $a_i = 2^i$ and $b_i = a_i/\Psi_{d,\boldsymbol{\beta}}(a_i)$ for $i \in \mathbb{N}$. For the corresponding reflection strategy, given the assumptions of Theorem A.24 it holds

$$\frac{1}{T}\mathbb{E}\big[\widetilde{C}_{0,T}\big] - J(\Omega^*) \lesssim \begin{cases}\Big(\frac{(\log T)^2}{T}\Big)^{\frac{1}{3}}, & d = 2,\\ \Big(\frac{\log T}{T}\Big)^{\frac{\overline{\boldsymbol{\beta}+\mathbf{1}}}{3\overline{\boldsymbol{\beta}+\mathbf{1}}+d-2}}, & d \geq 3.\end{cases}$$

The rate loss of the strategy's regret per time unit relative to the static regret from Proposition A.23 provides a natural analogue to the regret bounds from [28, Theorem 2.5] in the one-dimensional case, even though the construction of the strategies substantially differs. Let us also remark that *doubling tricks* in the strategy design that make sure that the number of episodes at time T is of order $\log T$, are commonly encountered in algorithms with verifiable optimal regret rates for undiscounted reinforcement learning problems. For instance, the popular UCRL2 algorithm in [5] recomputes policies as soon as the occurrence of a state-action pair has doubled. Such strategies have the drawback of choosing suboptimal policies for arbitrarily long periods of time (here, not reflecting at all), see [17]. Recently, [14] have proposed the *regret of exploration* as an appropriate measure to capture such inefficiencies. It is an interesting and challenging question for future work to adapt reflection strategies based on the nonparametric approach advocated in this paper to such finer-grained performance measures.

A.4 Numerical optimization

In this section, we present results showing how the objective function $J(\Omega)$ can not only be evaluated but also optimized numerically when limiting the set of eligible reflection sets $\boldsymbol{\Omega}$. Let us remark that the computational costs of the methods presented grow very quickly in d (see, e.g., Remark A.27). This, together with the degrading efficiency of kernel density estimators, cf. Theorem A.21, suggests that what follows is mostly appropriate for low to moderate dimensional problems. Corollary A.18 shows that, for known $\kappa > 0, V$ and f, our problem for known dynamics boils down to minimizing the functional

$$J : D \mapsto \frac{1}{\int_\Omega \mathrm{e}^{-V(x)} \mathrm{d}x} \left(\int_\Omega f(x) \mathrm{e}^{-V(x)} \, \mathrm{d}x + \kappa \int_{\partial\Omega} \mathrm{e}^{-V(x)} \, \mathcal{H}^{d-1}(\mathrm{d}x) \right)$$

over $\boldsymbol{\Omega}$. We are therefore faced with a shape optimization problem. For a general overview on shape optimization we refer to [34]. To approach our particular problem numerically, we restrict ourselves to bounded domains that are *strongly starshaped at 0.* Specifically, we assume that for any Ω the boundary is given by

$$\partial D = \{ r(q) q \ : \ q \in S^{d-1} \},$$

for some suitably smooth radial function $r : S^{d-1} \to (0, \infty)$ on the d-sphere S^{d-1}. Rather than optimizing over all such functions, we discretize the problem by considering $N \in \mathbb{N}$ points placed uniformly (in a suitable sense, see Remark A.27) on the sphere, say $\{q_i\}_{i=1}^N \subseteq S^{d-1}$, and then approximating any star-shaped set Ω by the polytope $\widetilde{\Omega}$ with vertices $\{p_i\}_{i=1}^N := \{r(q_i) q_i\}_{i=1}^N$. The surface $\partial\widetilde{\Omega}$ of the polytope $\widetilde{\Omega}$ can be triangulated into some number $N_F \in \mathbb{N}$ of facets, each spanned by d of the vertices of $\widetilde{\Omega}$, say, $p_{i_1}, p_{i_2}, \dots, p_{i_d}$ for some suitable index set $\{i_1, i_2, \dots, i_d\}$. Let $\mathcal{I}$ denote the set of such index sets, and for $I \in \mathcal{I}$, let F_I denote the facet spanned by $\{p_i \mid i \in I\}$ and S_I denote the d-simplex spanned by $\{0\} \cup \{p_i \mid i \in I\}$. It then follows that $\widetilde{\Omega} = \bigcup_{I \in \mathcal{I}} S_I$ while $\partial\widetilde{\Omega} = \bigcup_{I \in \mathcal{I}} F_I$, and so an approximation of $J(\Omega)$ is given by

$$J(\Omega) \approx J(\widetilde{\Omega}) = \frac{1}{\sum_{I \in \mathcal{I}} \int_{S_I} \mathrm{e}^{-V(x)} \mathrm{d}x} \sum_{I \in \mathcal{I}} \left(\int_{S_I} f(x) \mathrm{e}^{-V(x)} \, \mathrm{d}x + \kappa \int_{F_I} \mathrm{e}^{-V(x)} \, \mathcal{H}^{d-1}(\mathrm{d}x) \right). \tag{A.31}$$

Note that this approximation depends on r (and hence on Ω) only through the N lengths $\{r_i\}_{i=1}^N := \{r(q_i)\}_{i=1}^N$, and hence we may consider J as a function of N variables, $J(\boldsymbol{r}) := J(\widetilde{\Omega})$. Gradient based optimization schemes can now be used, since

$$\frac{\partial J(\boldsymbol{r})}{\partial r_i} = \frac{1}{C} \sum_{I \in \mathcal{I}_i} \left(\tfrac{\partial}{\partial r_i} \Big(\int_{S_I} f(x) \mathrm{e}^{-V(x)} \, \mathrm{d}x + \kappa \int_{F_I} \mathrm{e}^{-V(x)} \, \mathcal{H}^{d-1}(\mathrm{d}x) \Big) - J(\boldsymbol{r}) \tfrac{\partial}{\partial r_i} \Big(\int_{S_I} \mathrm{e}^{-V(x)} \, \mathrm{d}x \Big) \right), \tag{A.32}$$

where $C = \sum_{I \in \mathcal{I}} \int_{S_I} \mathrm{e}^{-V(x)} \, \mathrm{d}x$, and $\mathcal{I}_i = \{ I \in \mathcal{I} \ : \ i \in I \}$. To evaluate these expressions, we use the following theorem.

Theorem A.26: Let $q_1, \dots, q_d \in S^{d-1}$ and $r_1, \dots, r_d > 0$ be given and let $p_i = r_i q_i$ for $i = 1, \dots, d$. Denote by S the simplex in $\mathbb{R}^d$ spanned by the origin and the points $p_1, \dots, p_d$ and by F the facet of S opposite the origin. Finally, let Q denote the $d \times d$ matrix whose i'th column is q_i and $\mathrm{r}_{\mathrm{rat}} \in \mathbb{R}^d$ the vector given by $(r_{\mathrm{rat}})_j = \prod_{k \neq j} r_k$. Then we have the following for $g \in C(\mathbb{R}^d, \mathbb{R})$:

$$\int_S g(x)\,\mathrm{d}x = \Big(\prod_{k=1}^{d} r_k\Big)|Q| \int_0^1 \int_{(0,1)^{d-1}} g\big(r\eta(\boldsymbol{t})\big)\,\psi(\boldsymbol{t}) r^{d-1} \mathrm{d}\boldsymbol{t}\,\mathrm{d}r \tag{A.33}$$

$$\int_F g(x)\,\mathcal{H}^{d-1}(\mathrm{d}x) = \|\mathrm{adj}(Q^\top) r_{\mathrm{rat}}\| \int_{(0,1)^{d-1}} g\big(\eta(\boldsymbol{t})\big)\,\psi(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t} \tag{A.34}$$

where $\psi(t_1, \dots, t_{d-1}) = \prod_{i=1}^{d-2} t_i^{d-1-i}$, and

$$\eta(t_1, \dots, t_{d-1}) = (1 - t_1)p_1 + t_1(1 - t_2)p_2 + \cdots + \Big(\prod_{i=1}^{d-2} t_i\Big)(1 - t_{d-1})p_{d-1} + \Big(\prod_{i=1}^{d-1} t_i\Big)p_d.$$

Furthermore, for $i = 1, \dots, d$, we have

$$\frac{\partial}{\partial r_i} \int_S g(x)\,\mathrm{d}x = \Big(\prod_{k \neq i} r_k\Big)|Q| \int_{(0,1)^{d-1}} g\big(\eta_i(\boldsymbol{t})\big)\,\widehat{\psi}(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}, \tag{A.35}$$

where $\widehat{\psi}(t_1, \dots, t_{d-1}) = (1 - t_1)\psi(t_1, \dots, t_{d-1})$, and η_i denotes η after swapping p_1 and p_i. Finally, if also $g \in C^1(\mathbb{R}^d, \mathbb{R})$,

$$\begin{aligned} \frac{\partial}{\partial r_i} \int_F g(x)\,\mathcal{H}^{d-1}(\mathrm{d}x) = {} & \frac{r_{\mathrm{rat}}^\top \,\mathrm{adj}(QQ^\top) r_{\mathrm{rat}}^{(i)}}{\|\mathrm{adj}(Q^\top) r_{\mathrm{rat}}\|} \int_{(0,1)^{d-1}} g\big(\eta(\boldsymbol{t})\big)\,\psi(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t} \\ & + \|\mathrm{adj}(Q^\top) r_{\mathrm{rat}}\| \int_{(0,1)^{d-1}} \big\langle \nabla g\big(\eta_i(\boldsymbol{t})\big), q_i \big\rangle\, \widehat{\psi}(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}, \end{aligned} \tag{A.36}$$

where $(r_{\mathrm{rat}}^{(i)})_j = \prod_{k \neq i,j} r_k$ for $j \neq i$ and $(r_{\mathrm{rat}}^{(i)})_i = 0$.

The proof is given in Appendix A.A.

Remark A.27: While one can always use Theorem A.26 to evaluate $J(\widetilde{D})$ for any chosen set of vertices $\{q_i\}_{i=1}^N \subseteq S^{d-1}$, the approximation (A.31) might suffer if these are chosen poorly. An intuitive approach chooses vertices “uniformly” on the sphere, although better domain-dependent choices may be possible. Here, “uniformly” is understood in the sense that the distances between neighbouring vertices (i.e. vertices sharing an edge) should be approximately equal. Clearly, for $d = 2$, this is possible to do exactly, however for $d > 2$ there are only finitely many N for which this is possible (corresponding e.g. to the Platonic solids if $d = 3$), and so we must settle for approximations. There are several different optimality criteria, and their solutions depend on d, making it nigh impossible to implement for general d. Therefore, we propose the following method which, while generally suboptimal in the goal of minimizing variation in inter-point distances, generalizes easily to any $d \geq 2$. First, split the $d-1$-dimensional hypercube $[-1,1]^{d-1}$ evenly into sub-hypercubes of length $2/(K-1)$ for some $K \geq 2$. Then, each of these sub-hypercubes is triangulated into $(d-1)!$ simplices using the standard triangulation scheme. Finally, this construction is rotated appropriately $2d$ times to cover the surface of $[-1,1]^d$, and the result is projected onto S^{d-1} simply by normalizing. This results in a total of $N = K^d - (K-2)^d$ vertices and $S = 2d!(K-1)^{d-1}$ simplices to integrate over, showing that the curse of dimensionality affects this problem immensely. As such, for large dimensions, one might prefer other methods such as Monte Carlo integration for evaluating $J(\Omega)$, see [73]. However, such methods do not give access to gradients and it is unclear how to construct meaningful derivative-free optimization methods over shape domains for the given problem. Taking also account of the various necessary optimization steps and multi-threading that come on top of the cost function evaluations, our method is therefore feasible for small to moderate dimensions with the above implementations. In particular, the summands in (A.31) and (A.32) can be computed independently of each other and so if one were to implement these in practice, it would be preferable to use a programming language that can utilize multi-threading efficiently, e.g. by using a GPU, to speed up optimizations.

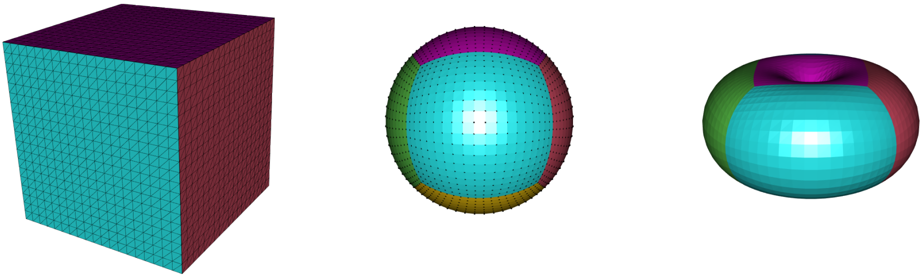

Figure A.4: *Illustration of our proposed method to generate points on S^{d-1}, here with $d = 3$. Left: Initial point generation and triangulation of the cube $[-1,1]^d$. Middle: Generated points on S^{d-1}. Right: Approximation polytope $\widetilde{\Omega}$ of Ω.*

When the dynamics of the process are known, that is we have access to the potential V, the theorem allows us to find explicit expressions for the right hand side of (A.31) and (A.32) given N points on the sphere and therefore to numerically optimize within the space of polytope approximations to star-shaped sets via gradient-based methods. For the two-dimensional case, strikingly simple expressions can be derived from this and, given sufficient regularity of the star-shaped domain, the approximation rate in (A.31) is at most of order $1/N$.

Corollary A.28: Let $r \in C^2([0, 2\pi], [r_{\min}, r_{\max}])$ with $0 < r_{\min} \le r_{\max} < \infty$ be some periodic radial function, and denote by $\Omega \subseteq \mathbb{R}^2$ the set with $\partial D = r([0, 2\pi])$. For any $N \in \mathbb{N}$, let $r_i = r(\frac{2i\pi}{N})$, $q_i = (\cos\frac{2i\pi}{N}, \sin\frac{2i\pi}{N})$ and $p_i = r_i q_i$ and denote by $\widetilde{\Omega} \subseteq \mathbb{R}^2$ the simplex with vertices $\{p_i\}_{i=1}^N$. Finally, for $i = 1, \dots, N$ and $t \in (0, 1)$, let $\eta_i^{\pm}(t) = p_i + t(p_{i\pm 1} - p_i)$, where we identify $p_0 = p_N$ and $p_{N+1} = p_1$. Then, there exists a constant $K \ge 0$ such that

$$|J(\Omega) - J(\widetilde{\Omega})| \le \frac{K}{N},$$

where for $\widetilde{\rho} := \exp(-V)$ we have the explicit representations

$$J(\widetilde{\Omega}) = \frac{1}{C} \sum_{i=1}^{N} \Big(\sin\frac{2\pi}{N} r_i r_{i+1} \int_0^1 \int_0^1 (f\widetilde{\rho})\big(r\eta_i^{+}(t)\big) r \mathrm{d}r \mathrm{d}t + \kappa \|p_{i+1} - p_i\| \int_0^1 \widetilde{\rho}\big(\eta_i^{+}(t)\big) \mathrm{d}t \Big),$$

and

$$\frac{\partial J(\widetilde{\Omega})}{\partial r_i} = \frac{1}{C} \int_0^1 \big(\psi^{+}(t)\widetilde{\rho}\big(\eta_i^{+}(t)\big) + \psi^{-}(t)\widetilde{\rho}\big(\eta_i^{-}(t)\big)\big) \,\mathrm{d}t,$$

where

$$\psi^{\pm}(t) = \big(\sin\tfrac{2\pi}{N} r_{i\pm 1}\big(f\big(\eta_i^{\pm}(t)\big) - J(\widetilde{\Omega})\big) - \kappa\|p_{i\pm 1} - p_i\| \big\langle \nabla V\big(\eta_i^{\pm}(t)\big), q_i \big\rangle\big)(1 - t) + \tfrac{\kappa(r_i - \cos\frac{2\pi}{N} r_{i\pm 1})}{\|p_i - p_{i\pm 1}\|},$$

and $C = \sum_{i=1}^N \sin\frac{2\pi}{N} r_i r_{i+1} \int_0^1 \int_0^1 \widetilde{\rho}\big(r\eta_i^{+}(t)\big) r \,\mathrm{d}r \,\mathrm{d}t$.

The proof is deferred to the supplement. This now implies for $d = 2$ that for suitable admissible domain families $\Theta \supset \Omega$, the infimum of $\Omega \mapsto J(\Omega)$ over Θ is well approximated by the infimum over the polytope approximations.

Corollary A.29: Let $d = 2$ and let Θ be the family of domains Ω that are strongly star-shaped at 0 and are identified by C^2 periodic radial functions $r_\Omega : [0, 2\pi] \to (0, \infty)$ such that for some global constants $\underline{\lambda}, \overline{\lambda}, \Lambda$,

$$\underline{\lambda} \le r_\Omega \le \overline{\lambda} \quad \text{and} \quad \max_{\theta \in [0, 2\pi]} (|r_\Omega'(\theta)| + |r_\Omega''(\theta)|) \le \Lambda.$$

Then, letting $\widetilde{\Omega}_N$ be the polytope approximation of Ω from Corollary A.28, it holds

$$\big|\inf_{\Omega \in \Theta} J(\widetilde{\Omega}_N) - \inf_{\Omega \in \Theta} J(\Omega)\big| \lesssim_{\underline{\lambda}, \overline{\lambda}, \Lambda} 1/N.$$

The proof is deferred to the supplement.

Remark A.30: While a detailed analysis of the approximation error appears cumbersome in higher dimensions, the approximation of $J(\Omega)$ by $J(\widetilde{\Omega})$ generally depends mainly on the approximation quality of r by $\widetilde{r}$ (where the latter denotes the radial function associated with $\widetilde{\Omega}$), in particular $\|r-\widetilde{r}\|_\infty$ and $\|\|\nabla r-\nabla\widetilde{r}\|\|_\infty$. Assuming $r \in C^2$, then by Lipschitz continuity, this is again proportionally bounded by the $d-1$-dimensional volume of the largest facet of the initial approximation of S^{d-1}, i.e. the polytope spanned by $q_1, \dots, q_N$.

Numerical experiments and simulations

In what follows, we test the methods described above on a variety of choices of the potential V, the cost function f and the reflection cost κ. To this end, we wish to simulate the associated reflected SDEs X^Ω when Ω is a polytope, so let us first discuss how this is carried out in practice. We utilize a standard Euler–Maruyama-type scheme with some alterations to account for reflection, cf. [103]. In particular, we initialize $\mathsf{X}_0 = 0$ and $\mathsf{L}_0^\Omega = 0$ and then for some $\Delta > 0$ and $n \in \mathbb{N}_0$ we first generate a proposal point

$$\mathsf{X}^{\mathrm{prop}}_{(n+1)\Delta} = \mathsf{X}_{n\Delta} - \nabla V(\mathsf{X}_{n\Delta}) \cdot \Delta + \sqrt{2}\xi_{n+1},$$

where $\xi_i \sim N_d(0, \Delta I_d)$. Then, letting P_Ω denote the projection operator onto Ω, we set $\mathsf{X}_{(n+1)\Delta} = P_\Omega(\mathsf{X}^{\mathrm{prop}}_{(n+1)\Delta})$ and $\mathsf{L}^\Omega_{(n+1)\Delta} = \mathsf{L}^\Omega_{n\Delta} + \|\mathsf{X}_{(n+1)\Delta} - \mathsf{X}^{\mathrm{prop}}_{(n+1)\Delta}\|$. When Ω is a polytope, P_Ω can be constructed explicitly as follows: first, clearly if $x \in \Omega$, we have $P_\Omega(x) = x$. Otherwise, let $F_1, F_2, \dots, F_S$ be the facets of Ω and let $p_j^{(i)}$ be the j'th vertex of F_i. Then the projection $P_{F_i}(x)$ of $x \in D^{\mathrm{c}}$ onto F_i can be found by solving the quadratic program

$$\min_{w\in\mathbb{R}^d} \|x - \sum w_j p_j^{(i)}\|^2, \quad \text{subject to} \quad \sum w_i = 1, w_i \geq 0,$$

to which standard numerical solvers are available. Then, choosing $i^*(x) \in \operatorname{argmin}\|x - P_{F_i}(x)\|$, we set $P_\Omega(x) = P_{F_{i^*(x)}}(x)$. While, in theory, $i^*(x)$ may not be unique, we remark that the set for which there exists several candidates is of dimension at most $d-1$ and hence $i^*(\mathsf{X}^{\mathrm{prop}}_{(n+1)\Delta})$ is unique with probability 1.

In practice, a number of steps can be taken to make the simulation more efficient. Firstly, when checking if $x \in \mathbb{R}^d$ is in Ω, a simple first test is to see if $\|x\| \leq r_{\min}$, where $r_{\min} := \sup\{r \geq 0 \mid B(0,r) \subseteq D\}$. If not, there exists unique vertices $p_1(x), p_2(x), \dots, p_d(x)$ such that these vertices span a facet, say $F_{c(x)}$, of Ω and such that x is in their conical hull, i.e. $x = \sum w_i p_i(x)$ where $w_i \geq 0$, and we have $x \in \Omega$ if and only if $\sum w_i \leq 1$. If the step size Δ is small, we have most likely that either $c(\mathsf{X}^{\mathrm{prop}}_{(n+1)\Delta}) = c(\mathsf{X}_{n\Delta})$, or that $F_{c(\mathsf{X}^{\mathrm{prop}}_{(n+1)\Delta})}$ is a neighbouring facet of $F_{c(\mathsf{X}_{n\Delta})}$. As such, keeping track of $c(\mathsf{X}_{n\Delta})$ enables us to quickly find $c(x_{(n+1)\Delta})$ with high probability and hence check if $\mathsf{X}^{\mathrm{prop}}_{(n+1)\Delta} \in \Omega$. In a similar fashion, we have most likely that $c(x) = i^*(x)$ or that $F_{i^*(x)}$ is a neighbouring facet of $F_{c(x)}$, thence, if we accept a very small probability of error, we can drastically reduce computation time by only considering the neighbours of $F_{c(x)}$ and $c(x)$ itself when computing $i^*(x)$.

While implementations ideally should make use of multi-threading, the following examples are supposed to be proofs of concept, so that it was sufficient for our purposes to implement the algorithms in R, a high-level language incompatible with multi-threading. We therefore only discuss examples in dimensions $d = 2$ and $d = 3$, where visualization is also much more straight-forward.

In the simple case of a pure Brownian motion, corresponding to $\mathrm{e}^{-V} \equiv 1$, and given radially symmetric costs $f = \|\cdot\|$, one expects the optimal reflection boundary to be a sphere centered at 0. Optimizing the corresponding cost functional over the space of such balls only, gives the optimization problem a parametric structure that can be easily solved analytically to reveal the optimal ball to be $\Omega^* = B(0, r^*)$, where $r^* = \sqrt{(d+1)\kappa}$. It is now interesting to test our method with regard to two questions: does the numerical optimization over the more general class of star-shaped domains support our intuition by identifying a ball as the optimal reflection domain, and if so, do we obtain a good approximation of the optimal radius $r^* = \sqrt{(d+1)\kappa}$ as well? The result for different choices of κ is visualized in Figure A.5, giving an affirmative answer to both questions. Curiously, it also appears in Figure A.5 that convergence towards an optimum is actually faster for $d = 3$ than for $d = 2$ in this particular example. A possible explanation for this is that while for $d = 2$, each vertex of the polygon approximation is connected to only 2 neighbours, for $d = 3$ each vertex is connected to 6. This results in each variable having a larger impact on the objective function, which may lead to the gradient being a better estimate for which direction is optimal for the objective.

Our method is also well-equipped for handling more challenging non-symmetric situations, where it is hard to make an educated guess on the optimal shape. As such, we also test the method on Ornstein–Uhlenbeck processes with strong correlation, in particular the processes

$$\mathrm{d}\mathsf{X}_t^{\Omega} = -M\mathsf{X}_t^{\Omega}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathsf{B}_t + n(\mathsf{X}_t^{\Omega})\,\mathrm{d}\mathsf{L}_t^{\Omega}, \qquad M = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}^{-1},$$

as well as a skewed cost-function, namely $f(x, y) = \sqrt{x^2 + 5y^2}$. The found approximately optimal shapes can be seen in Figure A.6. Here, for each shape we take $N = 50$, $\kappa = 1$ and use the BFGS algorithm to find the optimal shape with starting values $r_i = 1$ for $i = 1, \dots, 50$. Figure A.6 also shows simulations of the above reflected processes in these approximately optimal shapes to asses the convergence of the realized costs towards the theoretical objective function. We simulate the relevant processes with time-steps of 10^{-4} until time $T = 100$ (but plot only until $T = 10$ for visual clarity). The average realized cost in comparison to the expected average costs are given in Table A.1. Repeating the experiment for Brownian motion, but with $d = 3$ (here, we take the skew cost function as $f(x, y, z) = \sqrt{x^2 + 5y^2 + z^2})$ reveals similar optimal shapes, as seen in figure A.7.

For a more exotic example, we now test the effect of κ on the optimal shape for different cost functions f. In particular, in the following we let $\mu_1 = (0, -2)$, $\mu_2 = (2\cos\frac{\pi}{6}, 1)$ and $\mu_3 = (-2\cos\frac{\pi}{6}, 1)$ and consider as ρ the mixture distribution of three Gaussians with these points as

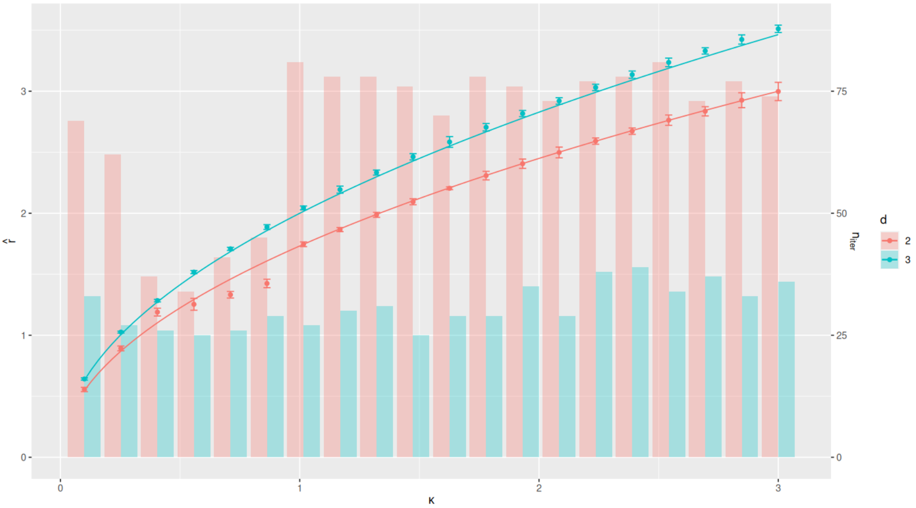

Figure A.5: *For each value of* κ*, we use the BFGS algorithm (using the built-in R implementation* `optim`*) to find an approximate optimal shape. To not bias the results towards a ball, we initialize the algorithm with* $r_i = 1 + \frac{1}{2}U_i$*, where* $U_i \sim \mathrm{Unif}[-1,1]$ *for* $i = 1, \ldots, N$ *(*$N \approx 200$*). Once the approximate optimal values* $\widehat{r}_1, \widehat{r}_2, \ldots, \widehat{r}_N$ *are found, we plot the mean of these along with error bars with height of their standard deviation. The standard deviations all being small indicate that the optimal shapes are approximately ball-shaped and their means give an approximation of the optimal radius. For reference we draw a curve of the theoretical optimal values* $r^* = \sqrt{(d+1)\kappa}$*. Finally, we also add a bar-plot illustrating the number of iterations of the BFGS algorithm were needed to compute the shapes.*

	Brownian motion	Ornstein–Uhlenbeck
norm cost function	2.22 (2.31)	1.18 (1.15)
skewed cost function	2.83 (2.91)	1.66 (1.74)

Table A.1: *Average realized costs vs. expected average long term costs (in brackets)*

means, i.e.

$$\rho(x) \propto \exp\Big(-\frac{\|x-\mu_1\|^2}{2}\Big) + \exp\Big(-\frac{\|x-\mu_2\|^2}{2}\Big) + \exp\Big(-\frac{\|x-\mu_3\|^2}{2}\Big).$$

We then consider for a variety of values of κ the two cost functions $f = \|\cdot\|$ and $f(x) = \min\{\|x-\mu_1\|, \|x-\mu_2\|, \|x-\mu_3\|\}$ and plot the results in figure A.8. Note that the second cost function does *not* satisfy $\operatorname{argmin} f = 0$, so while it does not fit exactly into the general framework of this article, we include it here since it illustrates the interplay between ρ, κ and f fairly well. As intuitively expected, for the norm cost function it appears optimal to reflect near each of the three modes of ρ, revealing especially for larger κ a triangular-like shape. For the other cost function, it appears

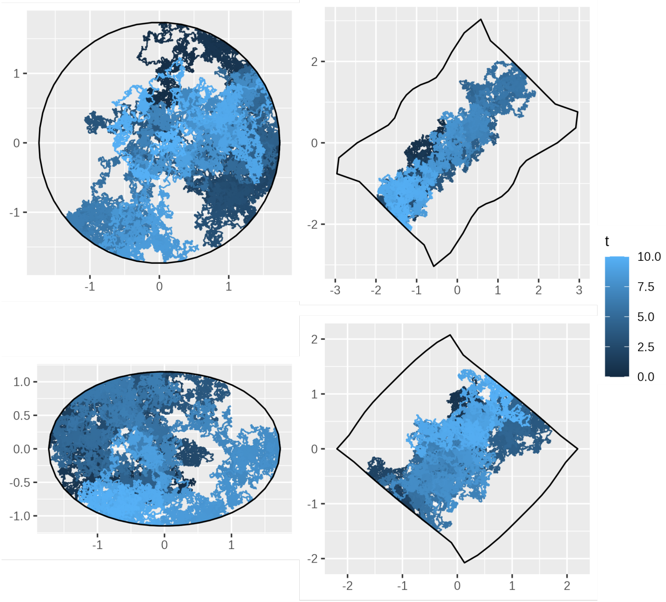

Figure A.6: *Simulated optimal shapes and corresponding path realizations of reflected processes. Top left: Brownian motion with norm cost. Top right: Ornstein–Uhlenbeck process with norm cost. Bottom left: Brownian motion with skewed cost. Bottom right: Ornstein–Uhlenbeck process with skewed cost.*

that essentially a merger of three balls around the three modes is optimal. This might, however, be an artefact of our searching only through shapes that are star-shaped at 0. It might in fact be the case that $J(\Omega)$ has three distinct minima centred around the three modes of ρ.

Finally, we wish to numerically assess the effectiveness of the methods discussed in Section A.3, i.e. when the density ρ is unknown. In particular, we consider the Ornstein–Uhlenbeck process X in $\mathbb{R}^2$ governed by

$$\mathrm{d}\mathsf{X}_t = -\frac{\mathsf{X}_t}{10}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathsf{B}_t,$$

and simulate data from this model until time T_{end} for increasing values of T_{end}, corresponding to increasing periods of exploration. For each simulation, we then estimate the invariant density ρ (here a normal density) via a time-discretized version of the kernel density estimator as well as its gradient $\nabla\rho$ via the gradient of the kernel estimator. In particular, for simulated observations $\mathsf{X}_{n\Delta}$, $n = 0, \ldots, T_{\mathrm{end}}/\Delta$ (here we take $\Delta = 10^{-3}$) we use the following estimates:

$$\widehat{\rho}(x) := \frac{1}{T_{\mathrm{end}}/\Delta} \sum_{n=0}^{T_{\mathrm{end}}/\Delta} \mathbb{K}_{\boldsymbol{h}}(x - \mathsf{X}_{n\Delta}), \quad \text{and} \quad \widehat{\nabla\rho}(x) := \frac{1}{T_{\mathrm{end}}/\Delta} \sum_{n=0}^{T_{\mathrm{end}}/\Delta} \nabla\mathbb{K}_{\boldsymbol{h}}(x - \mathsf{X}_{n\Delta}).$$

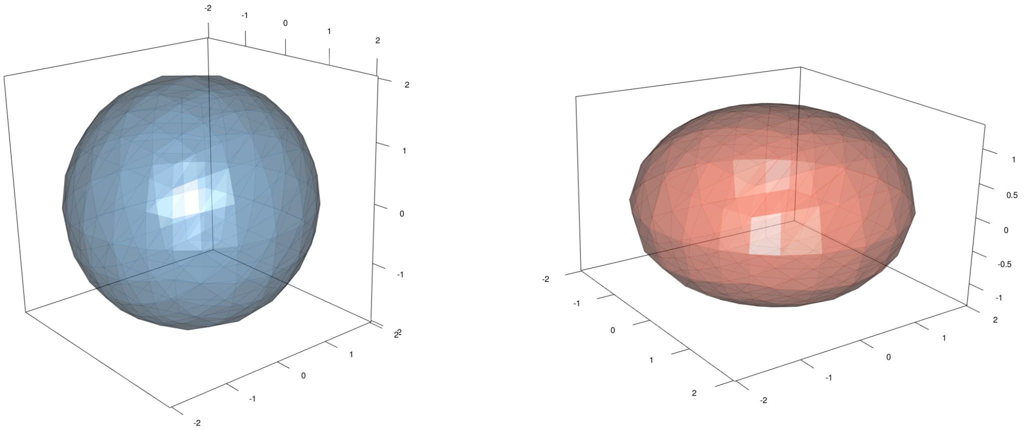

Figure A.7: *Optimal shapes for Brownian motion with reflection cost* $\kappa = 1$ *and cost function* $f = \|\cdot\|$ *(left) and* $f(x, y, z) = \sqrt{x^2 + 5y^2 + z^2}$ *(right).*

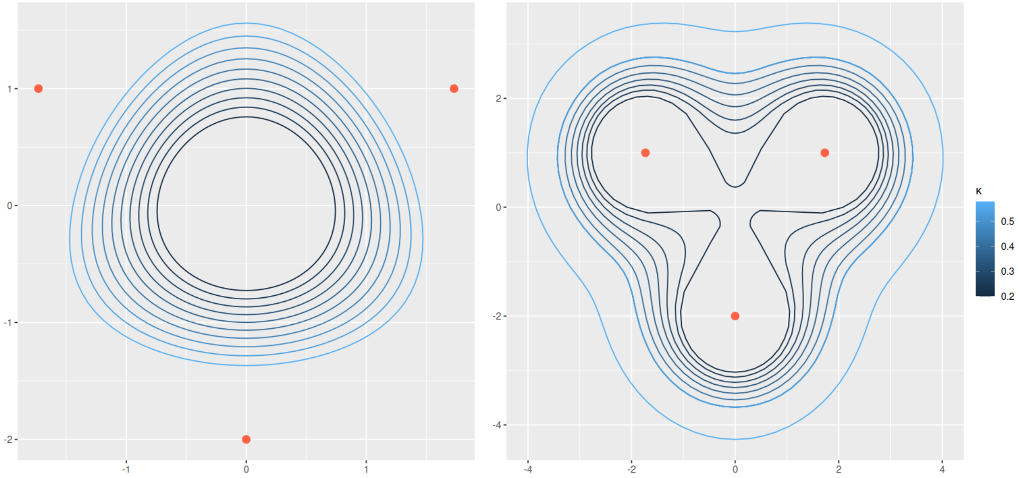

Figure A.8: *For each* κ*, we plot the optimal shape found by the BFGS algorithm, where* ρ *is a mixture of three Gaussians with means at the points marked in red. Left: Norm cost function,* $f = \|\cdot\|$*. Right: Cost function* $f(x) = \min\{\|x - \mu_1\|, \|x - \mu_2\|, \|x - \mu_3\|\}$*.*

Here, for numerical simplicity, we take K to be a standard normal density and $h_1 = h_2 = \frac{1}{\sqrt{T_{\text{end}}}}$. Finally, setting $f = \|\cdot\|$ and $\kappa = 1$, we use Theorem A.26 along with (A.31) and (A.32) to obtain a data-driven estimate of $J(\Omega)$ and $\nabla J(\Omega)$, which we then optimize as before with the BFGS algorithm. The results are shown in Figure A.9.

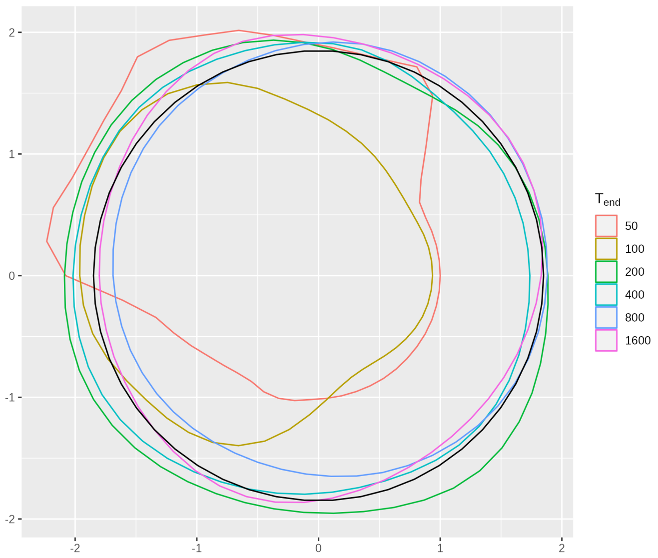

Figure A.9: *Estimates of the optimal shape (black) using kernel estimates after increasing periods of exploration. Notably, after only* $T = 150$*, the estimated optimal shape has an associated cost only* 0.61% *higher than the true optimum.*

A.A Proof of Theorem A.26

Proof. We first note that η is simply repeated linear interpolation. That is, if $L(t; x, y) := x + t(y - x)$ for $t \in (0, 1)$ and $x, y \in \mathbb{R}^d$, we have

$$\eta(t_1, \dots, t_{d-1}) = L\Big(t_1; p_1, L\big(t_2; p_2, L(t_3; p_3 \dots)\big)\Big).$$

As such, for any point $x \in \operatorname{int} F$, there exists a unique $\boldsymbol{t} \in (0, 1)^d$ such that $\eta(\boldsymbol{t}) = x$, and similarly for any $y \in \operatorname{int} S$, a unique $r \in (0, 1)$ such that $r\eta(\boldsymbol{t}) = y$. As such, to verify (A.33) and (A.34), we only need to show that the functions

$$(r, \boldsymbol{t}) \mapsto \Big(\prod_{k=1}^{d} r_k\Big)|Q|\psi(\boldsymbol{t})r^{d-1}, \qquad \text{and} \qquad \boldsymbol{t} \mapsto \|\mathrm{adj}(Q^\top)r_{\mathrm{rat}}\|^2 \psi(\boldsymbol{t}),$$

denote the Jacobian and Gramian of $(r, \boldsymbol{t}) \mapsto r\eta(\boldsymbol{t})$ and η, respectively. To this end, we introduce the following notation: let $t_d \equiv 0$, $T_1 = 1$ and set $T_i = \prod_{j=1}^{i-1} t_j$ for $i = 2, \dots, d$. Then we may write $\eta(t_1, \dots, t_{d-1}) = \sum_{i=1}^{d} T_i(1 - t_i)p_i$, whereby

$$\frac{\partial}{\partial t_i} r\eta(t_1, \dots, t_{d-1}) = r\Big(\frac{\eta(t_1, \dots, t_{d-1}) - \sum_{j=1}^{i-1} T_j(1 - t_j)p_j - T_i p_i}{t_i}\Big), \quad i = 1, \dots, d-1,$$

while of course $\frac{\partial}{\partial r} r\eta = \eta$. From this it follows by determinant properties that the Jacobian of $(r, \boldsymbol{t}) \mapsto r\eta(\boldsymbol{t})$ is given by

$$|\mathcal{J}(r, \boldsymbol{t})| = r^{d-1}\Big(\prod_{i=1}^{d-1} \frac{1}{t_i}\Big)\Big(\prod_{i=1}^{d} T_i\Big)|P| = |P|\psi(\boldsymbol{t})r^{d-1},$$

where P is the matrix whose i'th column is p_i. Noting then that $P = \mathrm{diag}\{r_1, \dots, r_d\}Q$ whereby $|P| = (\prod_{k=1}^{d} r_k)|Q|$ shows (A.33). To show (A.34), we note that since F lies in a $d-1$-dimensional hyperplane, say H, we may embed it in $\mathbb{R}^{d-1}$ by an isometry $\Psi : H \to \mathbb{R}^{d-1}$ with $\Psi(p_1) = 0$. Specifically, Ψ can be constructed as $x \mapsto \Psi'\big(A(x - p_1)\big)$, where A is the rotation matrix such that the d'th coordinate of Ax is 0 for all $x \in H$, and $\Psi' : \mathbb{R}^d \to \mathbb{R}^{d-1}$ simply discards the last coordinate. Then, integrating a function g over F with respect to $\mathcal{H}^{d-1}$ is equivalent to integrating $g \circ \Psi^{-1}$ over $\Psi(F)$ with respect to the $d-1$-dimensional Lebesgue measure. Now, since $\Psi(F)$ by the above construction of Ψ is a simplex in $\mathbb{R}^{d-1}$ consisting of the origin and $d-1$ other points, say $p'_1, \dots, p'_{d-1}$, it follows by the above,

$$\int_{\Psi(F)} g \circ \Psi^{-1}(x)\mathrm{d}x = |P'| \int_0^1 \int_{(0,1)^{d-2}} g \circ \Psi^{-1}\big(r\eta'(\boldsymbol{t})\big)\psi'(\boldsymbol{t})r^{d-1}\, \mathrm{d}\boldsymbol{t}\, \mathrm{d}r,$$

where η' similarly is linear interpolation between $p'_1, \dots, p'_{d-1}$, ψ' is the $d-1$-dimensional equivalent of ψ and P' is the $(d-1)\times(d-1)$ matrix whose i'th column is p'_i. By some elementary substitutions and renaming of variables, we may write in an abuse of notation $\Psi^{-1}\big(r\eta'(\boldsymbol{t})\big) = \eta(\boldsymbol{t})$ and $\psi'(\boldsymbol{t})r^{d-1} = \psi(\boldsymbol{t})$. Finally, to find $|P'|$, we see

$$|P'| = (d-1)!\lambda_{d-1}(\Psi(F)) = (d-1)!\lambda_{d-1}(F),$$

where we use that Ψ is an isometry and hence preserves volumes. This scaled $d-1$ dimensional volume is given by the square root of a Gram determinant, i.e. the determinant of the $(d-1) \times (d-1)$ matrix G whose i, j'th entry is $\langle p_{i+1} - p_1, p_{j+1} - p_1\rangle$. Such matrices enjoy the property that

$$|G| = \left|\begin{bmatrix} P \\ \mathbf{1}^\top \end{bmatrix}^\top \begin{bmatrix} P \\ \mathbf{1}^\top \end{bmatrix}\right| - |P^\top P| = \left|\begin{bmatrix} P \\ \mathbf{1}^\top \end{bmatrix}^\top \begin{bmatrix} P \\ \mathbf{1}^\top \end{bmatrix}\right| - |P|^2,$$

where $\mathbf{1} \in \mathbb{R}^d$ is the vector of all 1's. Furthermore, the Cauchy–Binet formula yields that

$$\left| \begin{bmatrix} P \\ \mathbf{1}^\top \end{bmatrix}^\top \begin{bmatrix} P \\ \mathbf{1}^\top \end{bmatrix} \right| = \sum_{k=1}^{d+1} |P^{(k)}|^2,$$

where $P^{(k)}$ is the matrix $\begin{bmatrix} P^\top & \mathbf{1} \end{bmatrix}^\top$ with the k'th row removed. Thus, we have $|P'| = \sqrt{\sum_{k=1}^d |P^{(k)}|^2}$, since the $d+1$'st term of the above sum corresponds exactly to $|P|^2$. Now, by the recursive formula for determinants, the determinant of $P^{(k)}$ is given as the sum of cofactors along the d'th row (since this row is all 1's). But removing the d'th row and j'th column from $P^{(k)}$ is exactly the same as removing the k'th row and j'th column from P, whereby

$$|P^{(k)}| = \sum_{j=1}^d (-1)^{k+j} |P^{(k,j)}| = \sum_{j=1}^d (-1)^{k+j} \Big(\prod_{l \neq j} r_l \Big) |Q^{(k,j)}| = (\operatorname{adj} Q^\top)_k^\top r_{\mathrm{rat}},$$

where $P^{(k,j)}$ denotes P after removing the k'th row and j'th column, and similarly for Q. From this, it follows clearly that $|P'| = \|\operatorname{adj}(Q^\top) r_{\mathrm{rat}}\|$, as desired.

To show (A.35), let $i \in \{1, \dots, d\}$ be fixed and consider now for some small h the simplex S_h with a vertex at the origin and at the points $p_1, \dots, p_{i-1}, \frac{r_i+h}{r_i} p_i, p_{i+1}, \dots, p_d$. Since either $S \subseteq S_h$ or $S_h \subseteq S$, the symmetric difference $S \vartriangle S_h$ is another simplex with vertices at $p_1, \dots, p_d$ and $\frac{r_i+h}{r_i} p_i$. Assume without loss of generality that $S \subseteq S_h$. Shifting the coordinate system so that p_i lies at the origin, we get a simplex with a vertex at the origin and at the points $p_1 - p_i, p_2 - p_i, \dots, \frac{h}{r_i} p_i, \dots, p_d - p_i$. Note that by properties of the determinant, we get

$$\left| \begin{bmatrix} p_1 - p_i & p_2 - p_i & \cdots & \frac{h}{r_i} p_i & \cdots & p_d - p_i \end{bmatrix} \right| = \frac{h}{r_i} |P|.$$

Using this and (A.33), we find that

$$\begin{aligned} \int_{S_h} g(x) \mathrm{d}x - \int_S g(x) \mathrm{d}x &= \int_{S_h \setminus S} g(x) \mathrm{d}x \\ &= h \Big(\prod_{k \neq i} r_k \Big) |Q| \int_0^1 \int_{(0,1)^{d-1}} g\big(r \eta_{i,h}(\boldsymbol{t}) + p_i\big) r^{d-1} \psi(\boldsymbol{t}) \, \mathrm{d}\boldsymbol{t} \, \mathrm{d}r, \end{aligned}$$

where, using the same notation as earlier,

$$\eta_{i,h}(t_1, \dots, t_{d-1}) = T_i(1 - t_i) \frac{h}{r_i} p_i + \sum_{j \neq i} T_j (1 - t_j)(p_j - p_i).$$

Dividing by h and letting $h \to 0$ (implicitly using dominated convergence and the continuity of g), we thus find

$$\frac{\partial}{\partial r_i} \int_S g(x) \mathrm{d}x = \Big(\prod_{k \neq i} r_k \Big) |Q| \int_0^1 \int_{(0,1)^{d-1}} g\big(r \eta_{i,0}(\boldsymbol{t}) + p_i\big) r^{d-1} \psi(\boldsymbol{t}) \, \mathrm{d}\boldsymbol{t} \, \mathrm{d}r. \tag{A.37}$$

Noting that

$$\sum_{j\neq i} T_j(1-t_j) = \sum_{j\neq i}^{d-1}(T_j - T_{j+1}) + T_d = T_1 - (T_i - T_{i+1}) = 1 - T_i(1-t_i),$$

we find

$$\begin{aligned} r\eta_{i,0}(t_1,\dots,t_{d-1}) + p_i &= r\Big(\sum_{j\neq i} T_j(1-t_j)(p_j - p_i)\Big) + p_i \\ &= r\Big(\sum_{j\neq i} T_j(1-t_j)p_j - \big(1 - T_i(1-t_i)\big)p_i\Big) + p_i \\ &= r\eta(t_1,\dots,t_{d-1}) + (1-r)p_i, \end{aligned}$$

which together with (A.37) yields

$$\frac{\partial}{\partial r_i}\int_S g(x)\mathrm{d}x = \Big(\prod_{k\neq i} r_k\Big)|Q|\int_0^1\int_{(0,1)^{d-1}} g\big(r\eta(\boldsymbol{t}) + (1-r)p_i\big)r^{d-1}\psi(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}\,\mathrm{d}r.$$

At this point we remark that changing the ordering of the points $p_1,\dots,p_d$ yields the same simplex, and so $J(\Omega)$ is independent of the ordering. Hence, swapping p_i and p_1 and changing η to η_i has no influence on $J(\Omega)$, and going through the above calculations again with this change (effectively setting $i = 1$), we arrive at the same result. Thus, changing η to η_i in the above, we see

$$r\eta_i(t_1,\dots,t_{d-1}) + (1-r)p_i = (1-rt_1)p_i + rt_1(1-t_2)p_2 + \cdots + rt_1\Big(\prod_{j=2}^{d-1} t_j\Big)p_d = \eta_i(rt_1,\dots,t_{d-1}).$$

Thus, making the substitution $u = rt_1$ in the above integral, we get

$$\begin{aligned} \int_0^1\int_{(0,1)^{d-1}} g\big(r\eta_i(\boldsymbol{t}) + (1-r)p_i\big)r^{d-1}\psi(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}\,\mathrm{d}r &= \int_0^1\int_{(0,1)^{d-2}}\int_0^r g\big(\eta_i(u,\boldsymbol{t})\big)\,\psi(u,\boldsymbol{t})\mathrm{d}u\,\mathrm{d}\boldsymbol{t}\,\mathrm{d}r \\ &= \int_{(0,1)^{d-2}}\int_0^1 g\big(\eta_i(u,\boldsymbol{t})\big)\,\psi(u,\boldsymbol{t})(1-u)\,\mathrm{d}u\,\mathrm{d}\boldsymbol{t} \\ &= \int_{(0,1)^{d-1}} g\big(\eta_i(\boldsymbol{t})\big)\,\widehat{\psi}(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}, \end{aligned}$$

which shows (A.35). Finally, (A.36) follows from the fact that

$$\begin{aligned} \frac{\partial}{\partial r_i}\|\mathrm{adj}(Q^\top)r_{\mathrm{rat}}\| = \frac{\partial}{\partial r_i}\sqrt{\|\mathrm{adj}(Q^\top)r_{\mathrm{rat}}\|^2} &= \frac{(\mathrm{adj}(Q^\top)r_{\mathrm{rat}})^\top(\frac{\partial}{\partial r_i}\,\mathrm{adj}(Q^\top)r_{\mathrm{rat}})}{\|\mathrm{adj}(Q^\top)r_{\mathrm{rat}}\|} \\ &= \frac{r_{\mathrm{rat}}^\top\,\mathrm{adj}(QQ^\top)r_{\mathrm{rat}}^{(i)}}{\|\mathrm{adj}(Q^\top)r_{\mathrm{rat}}\|}, \end{aligned}$$

and that $\frac{\partial}{\partial r_i}\eta_i(t_1,\dots,t_{d-1}) = (1-t_1)q_i$. ☐

Acknowledgements

The authors would like to thank two anonymous referees for their careful reading and valuable feedback.

Article B

*Note: this is a **copy** of the article*

Asbjørn Holk, Claudia Strauch and Lukas Trottner (2024)
Statistical Guarantees for Denoising Reflected Diffusion Models
Accepted in the Journal of Machine Learning Research (JMLR)
arXiv:2411.01563 [math.ST]

The only changes made are to the notation and layout for sake of consistency with the rest of the dissertation.

Statistical guarantees for denoising reflected diffusion models

Asbjørn Holk, Claudia Strauch and Lukas Trottner

Abstract

In recent years, denoising diffusion models have become a crucial area of research due to their abundance in the rapidly expanding field of generative AI. While recent statistical advances have delivered explanations for the generation ability of idealised denoising diffusion models for high-dimensional target data, implementations introduce thresholding procedures for the generating process to overcome issues arising from the unbounded state space of such models. This mismatch between theoretical design and implementation of diffusion models has been addressed empirically by using a *reflected* diffusion process as the driver of noise instead. In this paper, we study statistical guarantees of these denoising reflected diffusion models. In particular, under Sobolev smoothness assumptions, we establish rates of convergence in total variation which, up to a polylogarithmic factor, match the minimax lower bound. Our main contributions include the statistical analysis of this novel class of denoising reflected diffusion models and a refined score approximation method in both time and space, leveraging spectral decomposition and rigorous neural network analysis.

Keywords: diffusion models, score-based generative models, reflected diffusions, minimax optimality, deep neural networks

B.1 Introduction

Deep generative models (DGMs) are a broad class of models that train deep neural networks to generate synthetic samples from a target distribution representing a given training data set. Examples that have shown tremendous empirical success in the last decade include the classes of Generative Adversarial Networks (GANs) [4, 49, 88, 136], Variational Autoencoders [64] and normalising flows [93, 100, 115]. Most recently, dynamic generative methods, which work by learning the reverse dynamics of a forward noising process that gradually evolves the original data distribution into a simple and easy-to-sample-from distribution, have gained much traction in the machine learning community. The most prominent class of such models are Denoising Diffusion Models (DDMs) in discrete and continuous time [55, 104, 106, 107], which add Gaussian noise to the data and for which the backward dynamics is determined by the gradient of the log likelihood of the forward marginals, also known as the score. The score is intractable as it depends on the unknown initial data distribution and therefore needs to be learned, which is why these models are also referred to as score-based generative models.

These models have been adapted and improved for enhanced efficiency and performance in numerous applications. However, the statistical theory for standard DGMs is still under active development with many basic questions left unanswered. The statistical theory for GANs is perhaps the most mature among DGMs, with significant efforts in the recent past [20, 23, 75, 97, 102, 109, 110, 119] to build a minimax theory that, among other things, draws on deeper connections to optimal transport theory [24, 128].

Related work Starting only recently with the seminal paper [89], the study of statistical convergence guarantees for standard denoising diffusion models has experienced rapid growth in the past two years. For denoising diffusion models in $\mathbb{R}^D$ that transform data into noise via a forward Ornstein–Uhlenbeck process, [89] prove total variation and Wasserstein-1 rates that match the minimax lower bound over Besov classes up to log factors for the total variation and arbitrarily small polynomial loss for the Wasserstein-1 distance. Their assumptions on the initial distribution with density p_0 can be summarised by three key components:

(i) p_0 is compactly supported on the D-dimensional hypercube $[-1, 1]^D$;

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

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

These assumptions provide a natural starting point from a statistical perspective, since they allow one to easily control certain quantities in the approximation analysis that would otherwise obscure the central mathematical ideas. However, these assumptions have practical limitations since, in many empirical applications, data distributions tend to be multimodal and extremely high-dimensional. Consequently, the resulting optimal rates, derived in terms of the ambient dimension D, often do not align with the empirical success of diffusion models in generating high-quality samples in such complex scenarios. This gap has motivated recent studies to extend these results to cases where data distributions are constrained to lower-dimensional structures, aligning with the widely held *manifold hypothesis*, which suggests that popular high-dimensional training data sets are supported on lower-dimensional manifolds. [22], [118] and [6] have expanded the analysis to such structured settings, refining the rates in Wasserstein-1 distance and providing theoretical foundations for understanding the adaptivity of DDMs to low-dimensional structures.

Our contribution Our work offers a different type of extension: we provide statistical convergence guarantees for *denoising reflected diffusion models* (DRDMs). This class of generative models was introduced and empirically implemented in [80], motivated by the observation that, in practice, the implementation of the generation process of standard DDMs relies significantly on the incorporation of thresholding procedures to prevent the backward process, which has unbounded state space, from exiting the permitted range of target samples (say, e.g., tensors containing bounded RGB values of pixels in an image). Since such thresholding procedures have no theoretical justification in the design of unconstrained diffusion models, [80] suggest using reflected diffusion processes on bounded domains Ω as the forward noising process and demonstrate competitive performance to state-of-the-art models with comparable training and inference times. Specifically, [80] implement a time-changed reflected Brownian motion as forward model, which yields a uniform ergodic distribution on the bounded domain $\overline{\Omega}$ that is used as initialisation of the backward generative process with learned score obtained by denoising score matching, see below. We also refer to [45] and [46] for related work on constrained diffusion models.

To help the reader quickly identify the main technical novelties and to place our results in the context of the existing literature, let us briefly summarise core contributions of this paper:

- We derive an explicit upper bound on the expected total variation distance between the data distribution p_0 and the distribution induced by the DRDM with estimated score, with a resulting convergence rate which matches the well-known minimax lower bounds over Sobolev classes up to logarithmic factors.
- We derive a spectral characterisation of the time-dependent score function for reflected diffusions generated by self-adjoint operators with Neumann boundary conditions, yielding an explicit representation amenable to statistical analysis.
- We construct a calibrated neural network approximation scheme for these spectral score representations, combining truncation of eigenexpansions with space-time interpolation to balance approximation accuracy and model complexity.
- We analyse how the semi-explicit nature of transition densities in the reflected setting impacts score estimation and show how the resulting approximation and estimation errors can be controlled despite the absence of closed-form Gaussian expressions.

Having outlined these contributions, we now describe the construction and analysis of the reflected denoising dynamics in more detail. Time-reversal of the forward noising process motivates the formulation of the backward generative model, which under suitable regularity assumptions is again given by a reflected diffusion whose drift is determined by the score function

$$s_0(x,t) := \nabla \log p_t(x), \quad p_t(x) := \int_\Omega q_t(y,x)\,\mathbb{P}(\mathsf{X}_0 \in \mathrm{d}y),$$

where $q_t(x,y)$ are the transition densities of the reflected forward diffusion and $\mathbb{P}(\mathsf{X}_0 \in \cdot)$ is the data distribution. Since the latter is unknown, the score must be learned from a given i.i.d. data sample $(\mathsf{X}_{0,i})_{i=1,\dots,n}$, which, similarly to the unconstrained diffusion case, is implemented by minimising an empirical version of the denoising score matching loss

$$\int_{\underline{T}}^{\overline{T}} \int_{\Omega^2} |s(y,t) - \nabla_y \log q_t(x,y)|^2 q_t(x,y)\,\mathbb{P}(\mathsf{X}_0 \in \mathrm{d}x)\,\mathrm{d}y\,\mathrm{d}t,$$

in a class of feedforward ReLU neural networks $\mathcal{S} \ni s$, where $\overline{T} - \underline{T} \in (0, \overline{T}]$ is the stopping time of the backward reflected diffusion. By the equivalence of denoising and explicit score matching [129], it therefore becomes crucial to precisely calibrate the neural network class in terms of the number of observations n, the dimension D and the smoothness α of the data to balance the explicit score approximation error

$$\min_{s\in\mathcal{S}} \int_{\underline{T}}^{\overline{T}} \int_\Omega |s(x,t) - \nabla \log p_t(x)|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t$$

and the complexity of the class $\mathcal{S}$, measured in terms of its covering number w.r.t. the supremum norm.

While the approximation analysis for standard diffusion models with Ornstein–Uhlenbeck forward noising process [6, 22, 89, 118] can rely on the explicitly known Gaussian transition densities $q_t(x, y)$, this situation dramatically changes in the reflected model, where the transition densities are at best semi-explicit and need to be numerically approximated. We choose a forward reflected diffusion with self-adjoint weighted Laplacian $\nabla \cdot f\nabla$ subject to a Neumann boundary condition as a generator, which provides us with a space-time spectral decomposition of the transition densities of the form

$$q_t(x, y) = \sum_{j=0}^{\infty} \mathrm{e}^{-t\ell_j} e_j(x)e_j(y), \quad t > 0, x, y \in \Omega, \tag{B.38}$$

for eigenpairs $(\ell_j, e_j)_{j\in\mathbb{N}_0}$ of $-\nabla \cdot f\nabla$ with known growth behaviour.

The statistical estimation of the conductivity f based on such reflected diffusion data has been recently studied for low- and high-frequency observations in [85] and [56], respectively. In the low-frequency statistical setting, the computational challenges arising from the semi-explicit nature of (B.38) have been subsequently analysed in [48]. Let us note that, for the specific choice of $f \equiv 1/2$, the generator is given by $\frac{1}{2}\Delta$ and thus yields a reflected Brownian motion as a forward model. Up to a time change, we therefore cover the reflected diffusion models implemented in [80]. Independently of our choice for the heat conductivity $f : \mathbb{R}^D \to [f_{\min}, \infty) \subset (0, \infty)$, the forward noising process has a uniform ergodic distribution from which we sample to initialise the backward generating process. The conductivity $f(x)$ controls the speed at which the forward process diffuses around x, with larger values pushing the process faster towards equilibrium, but also increasing the oscillation of the eigenfunctions e_j around x. This implies a tradeoff between the convergence speed of the forward model and the approximation difficulty of the backward model. For some potential $V : \mathbb{R}^D \to \mathbb{R}$ we could also consider the generator $\mathcal{A}_V := \mathrm{e}^V \nabla \cdot \mathrm{e}^{-V} f\nabla$ with Neumann boundary conditions instead, resulting in a normally reflected forward diffusion with stationary distribution $\mu_V \propto \mathrm{e}^{-V}$ and a spectral decomposition as in (B.38) with respect to the dominating measure $\mu_V(y)\,\mathrm{d}y$, see [85]. For the particular choice of $f = 1/2$, this yields a reflected (overdamped) Langevin diffusion process. Our general approximation approach, which is described below, could be adapted to this scenario. However, in the context of generative modelling, we are primarily interested in a stationary forward distribution that is easy to sample from. As long as the domain Ω is sufficiently nice, this is clearly satisfied for the uniform ergodic distribution in the model considered here, but it is generally difficult for space-dependent choices of V, which require sampling from $\mu_V \propto \mathrm{e}^{-V}$ via MCMC methods, which are expensive in the dimension D. From a practical perspective, it is therefore quite natural to consider reflected diffusion models with uniform stationary distribution.

To get a grasp on the neural network approximation of the score obtained from (B.38), we impose assumptions on the data distribution $\mathbb{P}(\mathsf{X}_0 \in \cdot)$, which in the technical setting of [85] provide a natural analogue to the three essential conditions assumed for the data distribution in the statistical analysis of unconstrained diffusion models alluded to above. More precisely, let

$H_c^\alpha(\Omega)$ be the class of compactly supported Sobolev functions of order α on Ω, where we extend any function $\varphi \in H_c^\alpha(\Omega)$ to $\overline{\Omega}$ by setting $\varphi|_{\partial\Omega} = 0$. We assume that

($\mathcal{H}0$) there exist a nonnegative function $\widetilde{p}_0 \in H_c^\alpha(\Omega)$, for $\alpha \in \mathbb{N} \cap (D/2, \infty)$, and $p_{\min} > 0$ such that $\mathbb{P}(\mathsf{X}_0 \in \cdot)$ has a density $p_0 : \overline{\Omega} \to [p_{\min}, \infty)$ given by $p_0 = \widetilde{p}_0 + p_{\min}$.

This simplifying assumption provides a strictly positive lower bound $p_{\min} > 0$ on the data distribution, while at the same time avoiding boundary issues associated with reflection thanks to $p_0 \in H_c^\alpha(\Omega)/\mathbb{R}$ and guaranteeing Hölder continuity of p_0 by the Sobolev smoothness condition $\alpha > D/2$.

These conditions are directly comparable to the assumptions made on the data density p_0 in [89]:

- in their unconstrained model, they allow the noising and denoising processes to leave the support $[-1,1]^D$ and assume a lower bound $p_0|_{[-1,1]^D} \geq p_{\min} > 0$. We enforce the support constraint by restricting the generative model to $\overline{\Omega} = \operatorname{supp} p_0$ and therefore assume $p_0 = p_0|_{\overline{\Omega}} \geq p_{\min} > 0$.
- [89] assumes Besov smoothness $p_0 \in B_{p,q}^\alpha$ for $\alpha > (1/p - 1/2) \vee 0$, with the additional restriction that p_0 is arbitrarily smooth near the boundary. Since $H^\alpha = B_{2,2}^\alpha$, our Sobolev assumption on p_0 is more restrictive, and our requirement that $\alpha > D/2$ implies continuity of p_0. For the particular case $p = q = 2$, however, [89] impose no restrictions on the smoothness parameter $\alpha > 0$. We enforce arbitrary smoothness close to the boundary in a specific way by modelling p_0 as the shift of a compactly supported function on the open domain Ω.

Given the preceding discussion, our focus in this paper is therefore *not* to optimise our initial data assumptions with regard to what is observed in practice for popular data sets. We rather opt for a set of conditions that shares common principles with previous studies for standard DDMs, but allows us to highlight the central mathematical ideas for score approximation in the reflected setting and to develop a fairly compact approximation theory that we regard as a versatile starting point for future statistical investigations.

Given ($\mathcal{H}0$), the score can be written as

$$s(x,t) = \nabla \log p_t(x) = \frac{\sum_{j=0}^\infty \mathrm{e}^{-t\ell_j} \langle p_0, e_j \rangle_{L^2} \nabla e_j(x)}{\sum_{j=0}^\infty \mathrm{e}^{-t\ell_j} \langle p_0, e_j \rangle_{L^2} e_j(x)}, \quad x \in \Omega, t > 0.$$

Our neural network approximation strategy is then broken down into the following steps:

1. truncate the series representation of $p_t(x)$ at $j = N$ to obtain an approximation $h_N(x,t)$ and corresponding truncated score $\nabla \log h_N(x,t)$, where N needs to be appropriately chosen depending on n, α and 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 [114];
3. approximate the space-time function $h_N(x, t)$ by constructing a neural network approximation of the time interpolation of the neural networks from Step 2., where the interpolation degree is adapted to the parameters N, α and D.

All steps require a careful calibration of the approximation parameters and accordingly the neural network sizes, which is made possible by a precise analysis of the different levels of numerical errors induced by our stepwise approach. These techniques may be of independent interest since they are applicable to generic spectral decompositions of semigroups associated to self-adjoint Markov generators. In particular, this is potentially relevant for extending the statistical analysis to the unifying class of *denoising Markov models* that has been recently introduced in [12], see also [99]: following our previous discussion of generalised reflected forward models with self-adjoint generator $\mathcal{A}_V = \mathrm{e}^V \nabla \cdot f \mathrm{e}^{-V} \nabla$ and associated spectral decomposition of the transition density $q_t(x, y)$ relative to the invariant density $\mu_V \propto \mathrm{e}^{-V}$, for a general self-adjoint Markov generator $\mathcal{A}$ with invariant distribution μ and eigendecomposition $(\ell_j, e_j)_{j \in \mathbb{N}_0}$ in $L^2(\mu)$, the unknown forward marginals that are needed to express the backward generator in denoising Markov models are decomposed as

$$p_t(x) = \sum_{j=0}^{\infty} \mathrm{e}^{-t\ell_j} \langle p_0, e_j \rangle_{L^2(\mu)} e_j(x).$$

Thus, under appropriate assumptions on p_0 and controls on the eigenvalues and eigenfunctions, as well as with detailed knowledge of the invariant distribution μ, a neural-network approximation of the space-time function $p_t(x)$ and functionals thereof appears plausible, using the general strategy introduced in this paper.

This strategy allows us to determine an explicit calibration of the neural network class $\mathcal{S}$, the forward terminal time $\overline{T}$ and the backward early stopping time $\overline{T} - \underline{T}$, such that for the empirical denoising score loss minimiser $\widehat{s}_n$ in this class of neural networks, we obtain the convergence rate

$$\mathbb{E}\big[\mathrm{TV}(p_0, \overleftarrow{p}^{\widehat{s}_n}_{\overline{T}-\underline{T}})\big] \lesssim n^{-\frac{\alpha}{2\alpha+D}} (\log n)^3 (\log\log n)^{1/2}.$$

Here, $\overleftarrow{p}^{\widehat{s}_n}_t$ denotes the density at time t of the backward generating reflected diffusion that has drift term determined by $\widehat{s}_n$ and is started in the uniform distribution on Ω. This establishes an expected total variation convergence rate (up to small log-factors) for denoising reflected diffusion models that matches the well-known minimax lower bound over Sobolev classes [133].

Organisation of the paper The paper is structured as follows. In Section B.2, we provide the necessary technical background on denoising diffusion models. Here, we briefly outline the fundamentals underlying standard unconstrained DDMs, before providing a precise mathematical framework for denoising reflected diffusion models. Section B.3 introduces the exact specification of the score estimator via denoising score matching and the associated generative model. We then present our main result, Theorem B.31, with the remainder of the paper dedicated to its proof. The proof preparation proceeds along three sections. In Section B.3, we present the basic error decomposition of the expected total variation risk and establish bounds on the first two sources of error, arising from early stopping and the initiation of the backwards generating process in the invariant uniform distribution on $\overline{\Omega}$. The score matching error as the last component of the error decomposition is discussed in Section B.3, before in Section B.3, we construct the neural network approximation of the score that allows us to optimally bound the score matching error. Section B.3 is then dedicated to proving Theorem B.31, based on the results of the previous subsections.

In the concluding Section B.4, we discuss our findings and highlight promising directions for future research, motivated by the insights and limitations of this paper.

Notation For an open set Ω and $\alpha \in \mathbb{N}$, we denote by $H^{\alpha}(\Omega)$ the Sobolev space of functions having weak partial derivatives up to order α in $L^2(\Omega)$. We denote the inner product and corresponding norm on $L^2(\Omega)$ by $\langle \cdot, \cdot \rangle_{L^2(\Omega)}$ and $\|\cdot\|_{L^2(\Omega)}$, or simply $\langle \cdot, \cdot \rangle_{L^2}$ and $\|\cdot\|_{L^2}$ if the domain Ω is fixed and there is no room for confusion. $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^D$ for any $D \in \mathbb{N}$, and for a function $f : \mathcal{X} \to \mathbb{R}^D$ for some space $\mathcal{X}$, $\|f\|_{\mathcal{X}} = \sup_{x \in \mathcal{X}} |f(x)|$ is its supremum norm. $C^{\alpha}(\Omega)$ denotes the space of functions with continuous partial derivatives of order α in $\overline{\Omega}$, and we let $C^{\infty}(\Omega) = \bigcap_{s \in \mathbb{N}} C^{\alpha}(\Omega)$. For $\beta \in (0, 1]$, $C^{0,\beta}(\Omega)$ is the space of β-Hölder continuous functions on Ω. For $1 \le p, q \le \infty$ and $p_{\min} > 0$, $\mathcal{B}^{\alpha}_{p,q}(\Omega)$ denote the usual Besov spaces on Ω, cf. [123] for details.

B.2 Theoretical background

Before introducing our framework for DRDMs, we first recall the basic ideas of unconstrained diffusion models.

Based on observing a finite number of samples corresponding to an *unknown* distribution p_0 on $\mathbb{R}^D$, DDMs provide an iterative generative algorithm to create new samples that approximately match the target distribution p_0. The general idea is to find a stochastic process that perturbs p_0 to a new distribution p_T in such a way 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. In an idealised setting where we have access to the exact specifications of the backward dynamics, samples from p_0 can be generated exactly by first sampling from p_T and then running the backward process. However, these dynamics must naturally adapt to the information contained in the unknown p_0, so the true backward dynamics must be estimated from the data.

In the framework of DDMs, this perturbation is done via an SDE, i.e., for some fixed time $T > 0$ and suitable drift $b : [0, T] \times \mathbb{R}^D \to \mathbb{R}^D$ and diffusion coefficient $\sigma : [0, T] \times \mathbb{R}^D \to \mathbb{R}^{D \times D}$, we consider the forward model

$$\mathrm{d}\mathsf{X}_t = b(t, \mathsf{X}_t)\,\mathrm{d}t + \sigma(t, \mathsf{X}_t)\,\mathrm{d}\mathsf{B}_t, \quad t \in [0, T], \mathsf{X}_0 \sim p_0,$$

where $\mathsf{B} = (\mathsf{B}_t)_{t\in[0,T]}$ is a standard D-dimensional Brownian motion. Under sufficient regularity conditions [3, 53], the forward model has a solution $\mathsf{X} = (\mathsf{X}_t)_{t\in[0,T]}$ with marginal densities $(p_t)_{t\in[0,T]}$ such that the time-reversed process $\overleftarrow{\mathsf{X}}_t = \mathsf{X}_{T-t}$, $t \in [0, T]$, solves

$$\mathrm{d}\overleftarrow{\mathsf{X}}_t = -\overline{b}(T-t, \overleftarrow{\mathsf{X}}_t)\,\mathrm{d}t + \sigma(T-t, \overleftarrow{\mathsf{X}}_t)\,\mathrm{d}\overline{\mathsf{B}}_t, \quad t \in [0, T], \overleftarrow{\mathsf{X}}_0 \sim p_T, \tag{B.39}$$

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

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

Thus, the time-reversed process solves a time-inhomogeneous SDE, with drift $-\overline{b}(T-\cdot, \cdot)$ and diffusion coefficient $\sigma(T-\cdot, \cdot)$.

In many practical implementations as well as in the statistical studies [6, 22, 89, 118], the diffusion coefficient is set to $\sigma(t, x) = \gamma(t)\mathbb{I}_D$ for some scalar function γ, which implies that the forward model is given by a (possibly time-inhomogeneous) Ornstein–Uhlenbeck process with explicit transition densities and the backward drift becomes

$$\overline{b}(t, x) = b(t, x) - \gamma^2(t)\nabla \log p_t(x),$$

where $\nabla \log p_t$ is referred to as the *score* of the forward model.

Substituting this into (B.39), we obtain the dynamics

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

of the backward process. Then, when $t \to T$, the density of $\overleftarrow{\mathsf{X}}_t$ approaches p_0, so that simulating the reverse process generates new data samples corresponding to the target p_0.

While we are free to choose the coefficients of our forward process (i.e., b and σ), the score function $\nabla \log p_t$ depends on p_0 and hence needs to be estimated from the data, which is referred to as *score matching*.

Reflected generative diffusion models RGDMs follow the same generative principle, but constrain both forward and backward dynamics to a bounded domain. To avoid some technicalities, let us assume that $\Omega \subseteq \mathbb{R}^D$ is an open, connected and bounded set with C^∞ boundary $\partial\Omega$. We consider the reflected time-homogeneous forward model

$$\mathrm{d}\mathsf{X}_t = b(\mathsf{X}_t)\,\mathrm{d}t + \sigma(\mathsf{X}_t)\,\mathrm{d}\mathsf{B}_t + \nu(\mathsf{X}_t)\,\mathrm{d}\mathsf{L}_t, \quad \mathsf{X}_0 \in \overline{\Omega},$$

with smooth and bounded coefficients $b : \overline{\Omega} \to \mathbb{R}^D$, $\sigma : \overline{\Omega} \to \mathbb{R}^{D\times D}$ and conormal reflection determined by

$$\nu(x) := \frac{1}{2}a(x)n(x) = \frac{1}{2}\sum_{i=1}^{D}\langle \sigma_{\cdot,i}(x), n(x)\rangle \sigma_{\cdot,i}(x) > 0, \quad x \in \partial\Omega,$$

where $a := \sigma\sigma^\top$ is assumed to be uniformly elliptic.

Here, n is the inward unit normal vector at the boundary $\partial\Omega$, and the process $(\mathsf{L}_t)_{t\geq 0}$ is the local time at $\partial\Omega$, which is a nondecreasing continuous process of bounded variation that increases only when the solution X hits the boundary, i.e., $\mathsf{L}_t = \int_0^t \mathbf{1}_{\partial\Omega}(\mathsf{X}_s)\,\mathrm{d}\mathsf{L}_s$ almost surely. Because of the conormal reflection, the boundary local time can equivalently be characterised by

$$\mathsf{L}_t = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^t \mathbf{1}_{(\partial\Omega)_\varepsilon}(\mathsf{X}_s)\,\mathrm{d}s, \tag{B.40}$$

where $(\partial\Omega)_\varepsilon := \{x \in \mathbb{R}^D : \operatorname{dist}(x, \partial\Omega) \leq \varepsilon\}$, and the limit holds both in L^2 and almost surely, uniformly on $[0, T]$, cf. Cattiaux [18, Proposition 1.3].

The boundary reflection process $L_t := \int_0^t \varphi(\mathsf{X}_s)\,\mathrm{d}\mathsf{L}_s$ reflects X in a conormal direction whenever it hits the boundary $\partial\Omega$, thus constraining the state space of the diffusion to the compact set $\overline{\Omega}$.

The process X is a time-homogeneous Markov process with transition semigroup $(Q_t)_{t\geq 0}$ determined by transition densities $(q_t)_{t\geq 0}$, given as the fundamental solutions of the PDE with conormal Neumann boundary condition,

$$\begin{cases} \frac{\partial}{\partial t}u(x,t) = \mathcal{A}u(x,t), & (x,t) \in \Omega \times (0,T], \\ \frac{\partial}{\partial \nu}u(x,t) = 0, & (x,t) \in \partial\Omega \times (0,T], \end{cases}$$

where $\mathcal{A}$ is the second-order differential operator given by

$$\mathcal{A} = \sum_{i=1}^{D} b_i(x)\partial_{x_i} + \frac{1}{2}\sum_{i,j=1}^{D} a_{i,j}(x)\partial_{x_i}\partial_{x_j}.$$

We denote the density of the forward process at time t by p_t, i.e.,

$$p_t(x)\,\mathrm{d}x = \mathbb{P}(\mathsf{X}_t \in \mathrm{d}x \mid \mathsf{X}_0 \sim p_0) = \int_{\overline{\Omega}} p_0(y)\,Q_t(y, \mathrm{d}x)\,\mathrm{d}y = \int_{\overline{\Omega}} p_0(y)q_t(y,x)\,\mathrm{d}y\,\mathrm{d}x, \quad x \in \overline{\Omega}.$$

Similarly to the unconstrained model, Cattiaux [18, Theorem 2.5] shows that time-reversion of the reflected diffusion process yields a time-inhomogeneous reflected diffusion process, whose drift is reminiscent of the unconstrained model.

More precisely, letting $\overleftarrow{\mathsf{X}}^T = (\overleftarrow{\mathsf{X}}_t)_{t\in[0,T]}$, $\overleftarrow{\mathsf{X}}_t = \mathsf{X}_{T-t}$ and $\overline{\mathsf{L}}^T = (\overline{\mathsf{L}}_t)_{t\in[0,T]}$, $\overline{\mathsf{L}}_t = \mathsf{L}_T - \mathsf{L}_{T-t}$, $t \in [0,T]$, there exists a Brownian motion $\overline{\mathsf{B}}^T = (\overline{\mathsf{B}}_t)_{t\in[0,T]}$ w.r.t. an enlargement of the natural filtration generated by $\overleftarrow{\mathsf{X}}^T$ such that $\overleftarrow{\mathsf{X}}^T$ solves

$$\mathrm{d}\overleftarrow{\mathsf{X}}_t = -\overline{b}(\overleftarrow{\mathsf{X}}_t)\,\mathrm{d}t + \sigma(\overleftarrow{\mathsf{X}}_t)\,\mathrm{d}\overline{\mathsf{B}_t} + \nu(\overleftarrow{\mathsf{X}}_t)\,\mathrm{d}\overline{\mathsf{L}_t}, \quad \overleftarrow{\mathsf{X}}_0 \sim p_T, \tag{B.41}$$

on $[0, T)$, where $\overline{b} : [0, T] \times \overline{\Omega} \to \mathbb{R}^D$ is given by

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

Note that, by definition, $\overline{\mathsf{L}}^T$ is nondecreasing, has bounded variation and satisfies

$$\overline{\mathsf{L}}_t = \int_0^t \mathbf{1}_{\partial\Omega}(\overleftarrow{\mathsf{X}}_s)\, \mathrm{d}\overline{\mathsf{L}}_s$$

for $t \in [0, T]$, i.e., $\overline{\mathsf{L}}^T$ is the local time at the boundary of the backward process $\overleftarrow{\mathsf{X}}^T$.

A fundamental requirement for diffusion generative modeling is a precise understanding of the limiting behaviour of the forward process to evaluate the required run time of the forward process for the backward initialisation to be a sufficiently good approximation of the true terminal forward distribution p_T. To this end, we choose $b = \nabla f$ and $\sigma = \sqrt{2f}\mathbb{I}_{d\times d}$ for some potential $f : \mathbb{R}^D \to [f_{\min}, \infty) \subset (0, \infty)$. To avoid technicalities, we assume that $f \in C^\infty(\overline{\Omega})$ and that the smooth and bounded domain Ω is also convex. This enables us to remain within the technical framework of [85], which provides useful technical results required for our analysis. With our choice of coefficients, the time-homogeneous forward dynamics are described by the divergence form L^2-generator

$$\mathcal{A} = \nabla \cdot f\nabla = \langle \nabla f, \nabla \cdot \rangle + f\Delta \tag{B.42}$$

with core

$$H^1_\nu(\Omega) := \{\varphi \in H^1(\Omega) : \partial\varphi/\partial\nu = 0 \text{ on } \partial\Omega\}$$

for $\nu = fn$, corresponding to the constrained SDE

$$\mathrm{d}\mathsf{X}_t = \nabla f(\mathsf{X}_t)\,\mathrm{d}t + \sqrt{2f(\mathsf{X}_t)}\,\mathrm{d}\mathsf{B}_t + \nu(\mathsf{X}_t)\,\mathrm{d}\mathsf{L}_t. \tag{B.43}$$

Note that the ellipticity condition $f \geq f_{\min} > 0$ implies that the conormal Neumann boundary condition $\langle \nu(x), \nabla\varphi(x)\rangle = \frac{\partial\varphi}{\partial\nu}(x) = 0$ for $x \in \partial\Omega$ is equivalent to a normal Neumann boundary condition, i.e., $H^1_\nu(\Omega) = H^1_n(\Omega)$.

Thus, both the reflected forward and backward SDEs exhibit normal reflection at the boundary, and (B.43) induces a space-dependent scaling of local time that ensures that the occupation limit (B.40) holds true. Moreover, the specific choice $f \equiv 1/2$ yields a normally reflected Brownian motion.

By the divergence theorem, it follows that the invariant distribution of the forward Markov process X is the easy-to-sample-from uniform distribution on $\overline{\Omega}$, i.e., $\mu = \frac{\lambda_D|_{\overline{\Omega}}}{\lambda_D(\overline{\Omega})}$. Furthermore, there exist orthonormal eigenpairs $(\ell_j, e_j)_{j\geq 0}$ of the nonnegative operator $-\nabla \cdot f\nabla$ satisfying $0 = \ell_0 < \ell_1 \leq \ell_2 \leq \cdots$ and obeying the Weyl asymptotics $\ell_j \asymp j^{2/D}$ and $e_0 = \frac{1}{\lambda_D(\Omega)^{1/2}}\mathbf{1}$ and $(e_j)_{j\geq 1} \subset H^1_\nu(\Omega) \cap L^2_0(\Omega)$ such that

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

See Nickl [85, Section 3] for a detailed discussion of these properties.

Because $f \in C^\infty(\overline{\Omega})$, Nickl [85, Corollary 1] yields the bounds

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

This implies the smoothing property that, for any bounded initial density p_0,

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

for arbitrary $\tau > 1/2$. By the Sobolev imbedding theorem, we therefore have $p_t \in C^\infty(\Omega)$, regardless of the smoothness properties of p_0, and we can identify the weak derivatives of p_t with its classical derivatives. The SDE (B.41) governing the backward dynamics becomes

$$\mathrm{d}\overleftarrow{\mathsf{X}}_t = \big(\nabla f(\overleftarrow{\mathsf{X}}_t) + 2f(\overleftarrow{\mathsf{X}}_t)\nabla \log p_{T-t}(\overleftarrow{\mathsf{X}}_t)\big)\,\mathrm{d}t + \sqrt{2f(\overleftarrow{\mathsf{X}}_t)}\,\mathrm{d}\overline{\mathsf{B}}_t + \nu(\overleftarrow{\mathsf{X}}_t)\,\mathrm{d}\overline{\mathsf{L}}_t,$$

with initialisation $\overleftarrow{\mathsf{X}}_0 \sim p_T$. Thus, even though the diffusion coefficient is state-dependent, the particular interplay between drift and diffusion coefficient ensures that the backward drift is fully determined by the forward drift and the score $(x, t) \mapsto \nabla \log p_t(x)$, as in the unconstrained Ornstein–Uhlenbeck forward model with state-independent diffusion coefficient.

By the spectral decomposition of the transition densities, the score is explicitly given by

$$\nabla \log p_t(x) = \frac{\sum_{j\geq 0} \mathrm{e}^{-t\ell_j} \langle p_0, e_j\rangle_{L^2} \nabla e_j(x)}{\sum_{j\geq 0} \mathrm{e}^{-t\ell_j} \langle p_0, e_j\rangle_{L^2} e_j(x)}, \quad x \in \Omega, t > 0, \tag{B.44}$$

which will be instrumental in analysing the score approximation properties of neural networks underlying the algorithm described in the next section.

Neural network classes We will construct an estimator of the score via minimising the denoising score matching error in an appropriate class of neural networks. For doing so, we introduce parameterised classes of neural networks with ReLU (Rectified Linear Unit) activation function.

In particular, for any $b, x \in \mathbb{R}^m$, let

$$\sigma_b(x) = \begin{bmatrix} \sigma(x_1 - b_1) \\ \sigma(x_2 - b_2) \\ \vdots \\ \sigma(x_m - b_m) \end{bmatrix}, \quad \sigma(y) = y \vee 0,$$

and denote for $L \in \mathbb{N}$, $W \in \mathbb{N}^{L+2}$, $S \in \mathbb{N}$ and $B > 0$ by $\Phi(L, W, S, B)$ the class of neural networks with depth (i.e., number of hidden layers) L, layer widths (including input and output layers) W, sparsity constraint S, and norm constraint B. We thus 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 $i = 0, \dots, L$ (to ease notation, we always set $b_0 = 0$), 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. In an abuse of notation, we denote σ_0 simply by σ. This can be written succinctly as

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

B.3 Generative modelling with reflected diffusions

Denote the true score by $s_0(x, t) := \nabla \log p_t(x)$, and assume we are given samples $(\mathsf{X}_{0,i})_{i \in [n]} \overset{\text{i.i.d.}}{\sim} p_0$. For a hypothesis class $\mathcal{S}$ of neural networks with ReLU activation function, to be exactly calibrated later, and $s \in \mathcal{S} \cup \{s_0\}$, we define

$$\begin{aligned} L_s(x) &:= \int_{\underline{T}}^{\overline{T}} \int_{\Omega} |s(y, t) - \nabla_y \log q_t(x, y)|^2 q_t(x, y) \, \mathrm{d}y \, \mathrm{d}t \\ &= \mathbb{E}\Big[\int_{\underline{T}}^{\overline{T}} |s(\mathsf{X}_t, t) - \nabla_y \log q_t(x, \mathsf{X}_t)|^2 \mid \mathsf{X}_0 = x \Big], \end{aligned} \tag{B.45}$$

where $\overline{T}$ is the terminal runtime of the reflected forward process and $\underline{T} \in (0, \overline{T})$ is such that we run the reflected backward process, which is initialised with distribution $\mathcal{U}(\Omega)$, until $\overline{T} - \underline{T}$. We then denote the empirical score matching loss associated to s by

$$\widehat{L}_{s,n} := \frac{1}{n} \sum_{i=1}^{n} L_s(\mathsf{X}_{0,i}),$$

and we define the empirical score minimiser by

$$\widehat{s}_n := \operatorname*{argmin}_{s \in \mathcal{S}} \widehat{L}_{s,n}. \tag{B.46}$$

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

$$\begin{aligned} \mathrm{d}\overline{\mathsf{X}}_t^s &= \big(\nabla f(\overline{\mathsf{X}}_t^s) + 2 f(\overline{\mathsf{X}}_t^s) s(\overline{\mathsf{X}}^s, t)\big) \, \mathrm{d}t + \sqrt{2 f(\overline{\mathsf{X}}_t^s)} \, \mathrm{d}\overline{\mathsf{B}}_t + \nu(\overline{\mathsf{X}}_t^s) \, \mathrm{d}\overline{\mathsf{L}}_t, \quad t \in [0, \overline{T} - \underline{T}], \\ \overline{\mathsf{X}}_0^s &\sim \mathcal{U}(\overline{\Omega}), \end{aligned} \tag{B.47}$$

for some Brownian motion $(\overline{\mathsf{B}}_t)_{t \in [0, \overline{T} - \underline{T}]}$ and local time $(\overline{\mathsf{L}}_t)_{t \in [0, \overline{T} - \underline{T}]}$ at the boundary $\partial\Omega$, and we denote its density at time t by $\overleftarrow{p}_t^s$. Here, the initialisation $\mathsf{X}_0^{\widehat{s}_n} \sim \mathcal{U}(\overline{\Omega})$ and the Brownian motion $\overline{\mathsf{B}}$ are chosen independently of the data $(\mathsf{X}_{0,i})_{i=1,\dots,n}$. Then, $(\overleftarrow{p}_t^{\widehat{s}_n})_{t \in [0, \overline{T})}$ are the densities of the backward process driven by the score estimate $\widehat{s}_n$. In particular, $\overleftarrow{p}_{\overline{T} - \underline{T}}^{\widehat{s}_n}$ is the density of the generated data obtained from stopping the backward process early at time $\overline{T} - \underline{T}$. Assessing the quality of the generated samples therefore boils down to analysing the distance between the distribution induced by p_0 and the (random) distribution induced by $\overleftarrow{p}^{\widehat{s}_n}_{\overline{T}-\underline{T}}$.

In this paper, we use the total variation distance as the divergence measure. For two probability measures $\mathbf{P}, \mathbf{Q}$ on $\overline{\Omega}$ with Lebesgue densities p, q, the total variation distance is denoted by

$$\mathrm{TV}(p, q) \equiv \mathrm{TV}(\mathbf{P}, \mathbf{Q}) := \sup_{A \in \mathcal{B}(\overline{\Omega})} |\mathbf{P}(A) - \mathbf{Q}(A)|.$$

Our main result is the following.

Theorem B.31: Assume that $p_0 = \widetilde{p}_0 + p_{\min}$ for some nonnegative $\widetilde{p}_0 \in H^\alpha_c(\Omega)$ and $p_{\min} > 0$, where $\alpha \in \mathbb{N} \cap (D/2, \infty)$. Let

$$\underline{T} \asymp n^{-\frac{2\alpha}{\beta(2\alpha+D)}} \quad \text{and} \quad \overline{T} = \frac{\alpha}{\ell_1(2\alpha + D)} \log n,$$

where $\beta = 1$ if $\alpha > D/2 + 1$, $\beta = \alpha - D/2$ if $\alpha \in (D/2, D/2 + 1)$ and for $\alpha = D/2 + 1$, β can be arbitrarily chosen in $(0, 1)$.

Then, there exists a class of neural networks

$$\mathcal{S} = \left\{ s \in \Phi\big(L(n), W(n), S(n), B(n)\big) \ : \ \|s(\cdot, t)\|_\Omega \le C(t^{-1/2} \vee 1)\, \forall t \in [\underline{T}, \overline{T}] \right\},$$

for some global constant $C > 0$, with network sizes

$$\begin{aligned} L(n) &\lesssim \log n \log\log n, \\ \|W(n)\|_\infty &\lesssim n^{\frac{D}{2\alpha+D}} (\log n)^2, \\ S(n) &\lesssim n^{\frac{D}{2\alpha+D}} (\log n)^3, \quad \text{and} \\ B(n) &\lesssim n^{\frac{2\alpha}{\beta(2\alpha+D)}}, \end{aligned}$$

such that, for $\widehat{s}_n$ given by (B.46), it holds for n large enough that

$$\mathbb{E}\big[\mathrm{TV}(p_0, \overleftarrow{p}^{\widehat{s}_n}_{\overline{T}-\underline{T}})\big] \lesssim n^{-\frac{\alpha}{2\alpha+D}} (\log n)^3 (\log\log n)^{1/2}. \tag{B.48}$$

The proof of this theorem is prepared in the following sections. There, we will always work under the assumption on p_0 from the theorem, that is, we assume

($\mathcal{H}0$) there exist a nonnegative function $\widetilde{p}_0 \in H^\alpha_c(\Omega)$, for $\alpha \in \mathbb{N} \cap (D/2, \infty)$, and $p_{\min} > 0$ such that $p_0 = \widetilde{p}_0 + p_{\min}$,

without further comment. Note that since $\alpha > D/2$, the Sobolev imbedding theorem yields the continuous imbedding $H^\alpha_c(\Omega) \hookrightarrow C^{0,\beta}(\Omega)$, where β is defined depending on α and D as stated in

Theorem B.31. Thus, we may consider p_0 as a continuous function such that there exist some Hölder constants $\beta \in (0,1]$, $c_\beta \in (0,\infty)$, fixed throughout the rest of the paper, such that

$$|p_0(x) - p_0(y)| \leq c_\beta |x-y|^\beta, \quad x, y \in \Omega. \tag{B.49}$$

Error decomposition

Let $\overline{p}_t^s$ be the density at time t of the solution to the SDE (B.47) with initial condition $\overline{\mathsf{X}}_0^s \sim p_{\overline{T}}$ (independent of the data $(\mathsf{X}_{0,i})_{i=1,\dots,n}$) replacing $\overline{\mathsf{X}}_0^s \sim \mathcal{U}(\overline{\Omega})$. In particular, we then notice that $\overline{p}_t^{s_0} = p_{\overline{T}-t}$ and therefore $\overline{p}^{s_0}_{\overline{T}-\underline{T}} = p_{\underline{T}}$. As for unconstrained diffusion models, cf. [89], the triangle inequality for the total variation distance then implies the error decomposition bound

$$\begin{aligned}\mathbb{E}\big[\operatorname{TV}(p_0, \overleftarrow{p}^{\widehat{s}_n}_{\overline{T}-\underline{T}})\big] \leq \operatorname{TV}(p_0, p_{\underline{T}}) + \operatorname{TV}(\mathbb{P}(\mathsf{X}_{\overline{T}} \in \cdot \mid \mathsf{X}_0 \sim p_0), \mathcal{U}(\overline{\Omega}))\\ + \mathbb{E}\big[\operatorname{TV}(\overline{p}^{s_0}_{\overline{T}-\underline{T}}, \overline{p}^{\widehat{s}_n}_{\overline{T}-\underline{T}})\big].\end{aligned} \tag{B.50}$$

The generalisation error therefore splits into three separate error contributions:

1. the first term represents the error induced by stopping early the backward process initialised by the true forward terminal density $p_{\overline{T}}$ at time $\overline{T} - \underline{T}$;
2. the second term is the error associated to starting the backward process in its stationary distribution instead of $p_{\overline{T}}$;
3. the third term quantifies the error coming from running the backward process with the drift determined by the estimated score $\widehat{s}_n$ instead of the true score s_0.

We start with controlling the first two terms before treating the most challenging score approximation error. The early stopping contribution to the error decomposition is controlled via small time heat kernel bounds for the transition densities in the following lemma.

Lemma B.32: There exists a constant C depending only on f, D, Ω, β and c_β such that

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

Proof. Fix $0 < t \leq 1$. We need to show that $\frac{1}{2}\|p_t - p_0\|_{L^1} \leq Ct^{\beta/2}$. To do so, note that the reversibility of X implies the symmetry of its transition densities. Hence,

$$p_t(x) = \int p_0(y) q_t(y,x)\,\mathrm{d}y = \int p_0(y) q_t(x,y)\,\mathrm{d}y = \mathbb{E}[p_0(\mathsf{X}_t) \mid \mathsf{X}_0 = x], \qquad x \in \Omega.$$

Using the Hölder continuity stated in (B.49), we therefore obtain

$$|p_t(x) - p_0(x)| = |\mathbb{E}[p_0(\mathsf{X}_t) - p_0(\mathsf{X}_0) \mid \mathsf{X}_0 = x]| \leq c_\beta \mathbb{E}[|\mathsf{X}_t - \mathsf{X}_0|^\beta \mid \mathsf{X}_0 = x], \qquad x \in \Omega.$$

Furthermore, by Davies [29, Corollary 3.2.9], we have the small time Gaussian heat kernel bound

$$q_t(x,y) \le C_0 \frac{1}{t^{D/2}} \exp\Big(-C_1 \frac{|x-y|^2}{t}\Big), \quad x, y \in \Omega,$$

where $C_0, C_1 \ge 0$ are constants depending only on f, D and Ω. Thus, for $x \in \Omega$,

$$\mathbb{E}[|\mathsf{X}_t - \mathsf{X}_0|^\beta \mid \mathsf{X}_0 = x] = \int_D q_t(x,y)|x-y|^\beta \, \mathrm{d}y \le C_0 \frac{1}{t^{D/2}} \int_\Omega |x-y|^\beta \exp\Big(-C_1 \frac{|x-y|^2}{t}\Big) \mathrm{d}y.$$

Next, note that

$$\begin{aligned}\int_\Omega |x-y|^\beta \exp\Big(-C_1 \frac{|x-y|^2}{t}\Big) \mathrm{d}y &\le \int_{\mathbb{R}^D} |y|^\beta \exp\Big(-C_1 \frac{|y|^2}{t}\Big) \mathrm{d}y \\ &= \kappa_D \int_0^\infty r^{\beta+D-1} \exp\Big(-C_1 \frac{r^2}{t}\Big) \mathrm{d}r \\ &= \frac{\kappa_D}{2} t^{\frac{\beta+D}{2}} C_1^{-\frac{\beta+d}{2}} \int_0^\infty u^{\frac{\beta+D}{2}-1} \mathrm{e}^{-u} \, \mathrm{d}u \\ &= C_2 t^{\frac{\beta+D}{2}},\end{aligned}$$

where κ_D is the surface area of S^{D-1} and $C_2 = \frac{\kappa_D}{2} C_1^{-\frac{\beta+D}{2}} \Gamma(\frac{\beta+D}{2})$. Finally, since Ω is bounded, the result follows by setting $C = c_\beta C_0 C_2 \lambda_D(\Omega)/2$. □

The error contribution from starting the backward process uniformly on Ω is controlled in terms of the spectral gap ℓ_1 of $\mathcal{A}$ defined in (B.42), which can be lower bounded by $\ell_1 \ge f_{\min}/C_{\mathrm{P}}(\Omega)$, where $C_{\mathrm{P}}(\Omega)$ is the Poincaré constant of the domain Ω.

Lemma B.33: It holds that

$$\mathrm{TV}(\mathbb{P}(\mathsf{X}_{\overline{T}} \in \cdot \mid \mathsf{X}_0 \sim p_0), \mathcal{U}(\overline{\Omega})) \le \frac{\sqrt{\lambda_D(\Omega)}}{2} \|p_0\|_{L^2} \mathrm{e}^{-\ell_1 \overline{T}}, \quad \overline{T} > 0.$$

Proof. Let $t > 0$.

The Cauchy–Schwarz inequality implies that

$$\begin{aligned}\mathrm{TV}(\mathbb{P}(\mathsf{X}_t \in \cdot \mid \mathsf{X}_0 \sim p_0), \mathcal{U}(\overline{\Omega})) &= \frac{1}{2} \int_\Omega \big|p_t(x) - \tfrac{1}{\lambda_D(\Omega)}\big| \, \mathrm{d}x \\ &\le \frac{\sqrt{\lambda_D(\Omega)}}{2} \Big(\int_\Omega \big|p_t(x) - \tfrac{1}{\lambda_D(\Omega)}\big|^2 \, \mathrm{d}x \Big)^{1/2}.\end{aligned}$$

By the spectral decomposition of $q_t(y,x)$, it follows for $a_k := \langle p_0, e_k\rangle_{L^2}$ that

$$\begin{aligned}\int_\Omega \Big|p_t(x) - \tfrac{1}{\lambda_D(\Omega)}\Big|^2 \,\mathrm{d}x &= \int_\Omega \Big|\sum_{k\geq 1} \mathrm{e}^{-\ell_k t} e_k(x) a_k\Big|^2 \,\mathrm{d}x \\ &= \sum_{k\geq 1} a_k^2 \mathrm{e}^{-2\ell_k t} \\ &\leq \|p_0\|_{L^2}^2 \mathrm{e}^{-2\ell_1 t},\end{aligned}$$

where we used $\ell_0 = 0$, $e_0 \equiv \lambda_D(\Omega)^{-1/2}$ and $\lambda_D(\Omega)^{1/2} a_0 = \langle p_0, 1\rangle_{L^2} = 1$ for the first line and orthonormality of $(e_k)_{k\geq 0}$ for the second one. This yields the claim. ☐

We now move on to the treatment of score approximation error in the next section.

Score matching error

In this section, $\mathcal{S}$ denotes a generic neural network class that shall be exactly calibrated at the end of the section for our score approximation purposes. Recall that the score estimator $\widehat{s} = \widehat{s}_n$ is defined according to (B.46).

By Girsanov's theorem, cf. Theorem B.44, and Pinsker's inequality, the third term in the decomposition (B.50) is controlled by

$$\begin{aligned}&\Big(\frac{1}{2}\mathbb{E}\Big[\int_{\underline{T}}^{\overline{T}} \int_\Omega f(x)|\widehat{s}(x,t) - \nabla \log p_t(x)|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t\Big]\Big)^{1/2} \\ &\quad\asymp \Big(\mathbb{E}\Big[\int_{\underline{T}}^{\overline{T}} \int_\Omega |\widehat{s}(x,t) - \nabla \log p_t(x)|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t\Big]\Big)^{1/2},\end{aligned} \tag{B.51}$$

where we used that $0 < f_{\min} \leq f(x) \leq \|f\|_{\overline{\Omega}} < \infty$ for all $x \in \Omega$.

The key to bounding this term is the equivalence between explicit and denoising score matching, i.e.,

$$\begin{aligned}&\int_{\underline{T}}^{\overline{T}} \int_\Omega |s(y,t) - \nabla \log p_t(y)|^2 p_t(y)\,\mathrm{d}y\,\mathrm{d}t \\ &\quad= \int_{\underline{T}}^{\overline{T}} \int_{\Omega^2} |s(y,t) - \nabla_y \log q_t(x,y)|^2 q_t(x,y) p_0(x)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + C \\ &\quad= \mathbb{E}[L_s(\mathsf{X}_0)] + C,\end{aligned} \tag{B.52}$$

where

$$C = \int_{\underline{T}}^{\overline{T}} \int_\Omega |\nabla \log p_t(y)|^2 p_t(y)\,\mathrm{d}y\,\mathrm{d}t - \int_{\underline{T}}^{\overline{T}} \int_{\Omega^2} |\nabla_y \log q_t(x,y)|^2 q_t(x,y) p_0(x)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t \leq 0$$

is a constant that is independent of s. Note that (B.52) is valid in our reflected diffusion model by the same arguments as in [129], see also the proof of Lemma C.3 in [89].

Using (B.52), the generalisation loss (B.51) can be bounded in terms of the minimal score approximation error over the class $\mathcal{S}$ and the complexity of the induced function class $\mathcal{L} := \{L_s : s \in \mathcal{S}\}$ for a desired precision level δ.

Theorem B.34: Suppose that $\sup_{s\in\mathcal{S}\cup\{s_0\}}\|L_s\|_\Omega \leq C(\mathcal{L}) < \infty$. Then, for any $\delta > 0$ such that $\mathcal{N}(\mathcal{L}, \|\cdot\|_\Omega, \delta) \geq 3$, it holds that

$$\begin{aligned}&\mathbb{E}\Big[\int_{\underline{T}}^{\overline{T}}\int_\Omega |\widehat{s}(x,t) - \nabla\log p_t(x)|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t\Big]\\ &\quad\leq 2\inf_{s\in\mathcal{S}}\int_{\underline{T}}^{\overline{T}}\int_\Omega |s(x,t) - \nabla\log p_t(x)|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t + 2\frac{C(\mathcal{L})}{n}\Big(\frac{145}{9}\log\mathcal{N}(\mathcal{L},\|\cdot\|_\Omega,\delta) + 160\Big)\\ &\quad\quad + 5\delta.\end{aligned}$$

The proof of this theorem is in principle the same as that of Theorem C.4 in [89], but let us emphasize that larger numeric constants appear in our statement. Apart from a few simple typos in [89], the main reason for this is a small gap in the proof of Oko, Akiyama, and Suzuki [89, Theorem C.4], which has recently been pointed out in [132], and which requires fixing. More precisely, [89] claim that the excess loss satisfies

$$\mathbb{E}\big[(L_s(\mathsf{X}_0) - L_{s_0}(\mathsf{X}_0))^2\big] \leq C(\mathcal{L})\mathbb{E}[L_s(\mathsf{X}_0) - L_{s_0}(\mathsf{X}_0)],$$

which is unjustified because $L_s(x) - L_{s_0}(x)$ is not necessarily non-negative pointwise. This is not dramatic however, since we can show that instead

$$\mathbb{E}\big[(L_s(\mathsf{X}_0) - L_{s_0}(\mathsf{X}_0))^2\big] \leq 4C(\mathcal{L})\mathbb{E}[L_s(\mathsf{X}_0) - L_{s_0}(\mathsf{X}_0)], \tag{B.53}$$

holds, i.e., the bound from [89] is true up to a universal multiplicative constant that does not matter in an essential way for the remainder of the proof.

Using the terminology of [10], (B.53) shows that the denoising score matching excess loss satisfies a *Bernstein condition*, which for a variety of problems in the empirical risk minimisation literature has been identified as a crucial ingredient to obtain minimax optimal convergence rates for empirical risk minimisers. For reasons of reproducibility in other modelling contexts, we prove (B.53) in Appendix B.B in a general Markovian framework.

To deal with the stochastic error in the upper bound, it is essential to control both the uniform loss upper bound $C(\mathcal{L})$ and the covering number $\mathcal{N}(\mathcal{L}, \|\cdot\|_\Omega, \delta)$. The size of the networks from the generic class $\mathcal{S}$ which is required to have a sufficient bound on the approximation error

$$\inf_{s\in\mathcal{S}}\int_{\underline{T}}^{\overline{T}}\int_\Omega |s(x,t) - \nabla\log p_t(x)|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t$$

translates directly into covering number bounds on $\mathcal{N}(\mathcal{L}, \|\cdot\|_\Omega, \delta)$, as the following result shows.

Lemma B.35: If for any $t > 0$, $\sup_{s\in\mathcal{S}}\|s(\cdot,t)\|_\Omega \le C(\mathcal{S})(t^{-1/2}\vee 1)$, for some finite constant $C(\mathcal{S})$, then there exists a constant c, depending only on D, f and Ω, such that, for any $\delta > 0$,

$$\mathcal{N}(\mathcal{L},\|\cdot\|_\Omega,\delta) \le \mathcal{N}\big(\mathcal{S},\|\cdot\|_{\Omega\times[\underline{T},\overline{T}]},\tfrac{\delta}{cC(\mathcal{S})\overline{T}}\big).$$

Proof. For $\widetilde{\delta} > 0$, let $s_1,\dots,s_N : \mathbb{R}^D \to \mathbb{R}$ be a $\widetilde{\delta}$-net for $\mathcal{S}$ w.r.t. $\|\cdot\|_{\Omega\times[\underline{T},\overline{T}]}$. Denote $h_s(x;y,t) := s(y,t) - \nabla_y \log q_t(x,y)$ for $s\in\mathcal{S}$. Let $s\in\mathcal{S}$, and choose $s'\in\{s_1,\dots,s_N\}$ such that $\|s-s'\|_{\Omega\times[\underline{T},\overline{T}]} \le \widetilde{\delta}$. Using definition (B.45) and the inequality $\big||h_s|-|h_{s'}|\big|(x;y,t) \le |s-s'|(y,t)$, we obtain

$$\begin{aligned}
|L_s - L_{s'}|(x) &\le \int_{\underline{T}}^{\overline{T}}\int_\Omega \big||h_s|^2-|h_{s'}|^2\big|(x;y,t)q_t(x,y)\,\mathrm{d}y\,\mathrm{d}t \\
&= \int_{\underline{T}}^{\overline{T}}\int_\Omega \big||h_s|-|h_{s'}|\big|(x;y,t)\left(|h_s|+|h_{s'}|\right)(x;y,t)q_t(x,y)\,\mathrm{d}y\,\mathrm{d}t \\
&\lesssim \|s-s'\|_{\Omega\times[\underline{T},\overline{T}]}\int_{\underline{T}}^{\overline{T}}\int_\Omega \left(|h_s|+|h_{s'}|\right)(x;y,t)q_t(x,y)\,\mathrm{d}y\,\mathrm{d}t \qquad \text{(B.54)}\\
&\lesssim \widetilde{\delta}\Big(\int_{\underline{T}}^{\overline{T}}\sup_{s\in\mathcal{S}}\sup_{z\in\Omega}|s(z,t)|\,\mathrm{d}t + \int_{\underline{T}}^{\overline{T}}\int_\Omega |\nabla_y \log q_t(x,y)|q_t(x,y)\,\mathrm{d}y\,\mathrm{d}t\Big) \\
&\lesssim \widetilde{\delta}\Big(C(\mathcal{S})\overline{T} + \int_{\underline{T}}^{\overline{T}}\int_\Omega |\nabla_y \log q_t(x,y)|q_t(x,y)\,\mathrm{d}y\,\mathrm{d}t\Big).
\end{aligned}$$

Using Ouhabaz [92, Theorem 6.19] and symmetry of q_t, we obtain for some constants $C,\gamma > 0$ only depending on D, f and Ω,

$$\begin{aligned}
\int_\Omega |\nabla_y \log q_t(x,y)|q_t(x,y)\,\mathrm{d}y &= \int_\Omega |\nabla_y q_t(x,y)|\,\mathrm{d}y \\
&= \int_\Omega |\nabla_y q_t(y,x)|\,\mathrm{d}y \\
&\le Ct^{-1/2}\mathrm{e}^{\gamma t}.
\end{aligned}$$

This shows that

$$\int_{\underline{T}}^{1}\int_\Omega |\nabla_y \log q_t(x,y)|q_t(x,y)\,\mathrm{d}y\,\mathrm{d}t \lesssim 1. \tag{B.55}$$

Furthermore, as in the proof of Nickl [85, Proposition 3], we have

$$\sup_{(x,y)\in\Omega^2}|\nabla_y q_t(x,y)| \lesssim \sum_{j\ge 1} j^{\tau+1/D}\mathrm{e}^{-ctj^{2/D}}, \quad t > 0,$$

for some constants $c > 0$, $\tau > 1/2$, showing also that

$$\int_{1}^{\overline{T}}\int_\Omega |\nabla \log q_t(x,y)|q_t(x,y)\,\mathrm{d}y\,\mathrm{d}t \lesssim 1. \tag{B.56}$$

Plugging (B.55) and (B.56) into (B.54), it follows that $\{L_{s_i} : i \in [N]\}$ is an $\widetilde{\delta} c C(\mathcal{S})\overline{T}$-covering of $\mathcal{L}$ w.r.t. $\|\cdot\|_\Omega$, where c is some constant depending only on f, D and Ω. This implies the claimed result. ☐

For the uniform loss upper bound $C(\mathcal{L})$ we again need to deal with the challenge of not having access to a simple analytic expression for the transition densities $q_t(x, y)$. While upper and lower heat kernel bounds are available for q_t and its gradient under specific assumptions on the domain, these are not sufficient to yield appropriate pointwise estimates for the log-gradient $\nabla_y q_t(x, y)$. Instead, we exploit that for our purposes it suffices to have (time)-integrated bounds. The basic idea is best illustrated for the particular case of a constant diffusivity $f \equiv 1$. Then, the generator of the forward process is just the Neumann Laplacian Δ on $\overline{\Omega}$ which implies that for any $t > 0$ and $x, y \in \Omega$

$$\Delta_y q_t(x, y) = \partial_t q_t(x, y),$$

because $q_t(x, y)$ is a symmetric fundamental solution to the Neumann heat equation. Furthermore,

$$\Delta_y \log q_t(x, y) = \frac{\Delta_y q_t(x, y)}{q_t(x, y)} - |\nabla_y \log q_t(x, y)|^2,$$

which together with the above establishes the fundamental relation

$$|\nabla_y \log q_t(x, y)|^2 - \partial_t \log q_t(x, y) = -\Delta_y \log q_t(x, y),$$

between spatial and temporal log-gradients. Based on this observation, the famous Li–Yau estimate [74] establishes that for $f \equiv 1$ it holds that

$$|\nabla_y \log q_t(x, y)|^2 - \partial_t \log q_t(x, y) \leq \frac{d}{2t}.$$

This result could be directly used for $f \equiv 1$ to establish the bound in the following lemma, but if f is not constant the situation becomes a bit more tricky. While we may always interpret $\Delta_f = \nabla \cdot f \nabla$ as a weighted Neumann Laplacian on the manifold $\overline{\Omega}$ equipped with a Riemannian metric induced by the Riemannian tensor associated to f, corresponding Li–Yau type estimates from the literature [9, 98] require the validation of certain curvature conditions on $\nabla \cdot f \nabla$, which may be hard to check for specific choices of f and Ω. To circumvent this problem, we follow a more elementary approach, which is however still based on the type of reasoning outlined above.

Lemma B.36: Assume that $\underline{T} \leq 1$ and $\overline{T} \geq 1$. If for any $t > 0$, $\sup_{s \in \mathcal{S}} \|s(\cdot, t)\|_\Omega \leq C(\mathcal{S})(t^{-1/2} \vee 1)$, then

$$\sup_{s \in \mathcal{S} \cup \{s_0\}} \|L_s\|_\Omega \lesssim (C(\mathcal{S})^2 \vee 1)(|\log \underline{T}| + \overline{T}).$$

Proof. Let $s \in \mathcal{S} \cup \{s_0\}$ and $x \in \Omega$ be arbitrarily chosen. By the assumption on the uniform temporal growth of s, and using that $\int_\Omega q_t(x,y)\,\mathrm{d}y = 1$

$$
\begin{aligned}
L_s(x) &\leq 2\Big(\int_{\underline{T}}^{\overline{T}} \int_\Omega \big(|s(y,t)|^2 + |\nabla_y \log q_t(x,y)|^2\big) q_t(x,y)\,\mathrm{d}y\Big) \\
&\lesssim C(\mathcal{S})^2 \int_{\underline{T}}^{\overline{T}} t^{-1} \vee 1 \,\mathrm{d}t + \int_{\underline{T}}^{\overline{T}} \int_\Omega |\nabla_y \log q_t(x,y)|^2 q_t(x,y)\,\mathrm{d}y \\
&\leq C(\mathcal{S})^2 (|\log \underline{T}| + \overline{T}) + \int_{\underline{T}}^{\overline{T}} \int_\Omega |\nabla_y \log q_t(x,y)|^2 q_t(x,y)\,\mathrm{d}y.
\end{aligned}
$$

To bound the remaining integral involving the log-gradient of the transition density, we first observe that $\nabla_y \cdot f\nabla_y \log q_t(x,y)$ satisfies

$$
\begin{aligned}
\nabla_y \cdot f\nabla_y \log q_t(x,y) &= \nabla_y \cdot f(y) \frac{\nabla_y q_t(x,y)}{q_t(x,y)} \\
&= \frac{\nabla_y \cdot f\nabla_y q_t(x,y)}{q_t(x,y)} - f(y)\frac{|\nabla_y q_t(x,y)|^2}{q_t(x,y)^2}.
\end{aligned}
$$

Since $q_t(x,y) = q_t(y,x)$ is a fundamental solution to the elliptic PDE $(\nabla \cdot f\nabla - \partial_t)u(y,t) = 0$ on Ω with Neumann boundary conditions, we can write

$$
\begin{aligned}
\int_\Omega \int_{\underline{T}}^{\overline{T}} \frac{\nabla_y \cdot f\nabla_y q_t(x,y)}{q_t(x,y)} q_t(x,y)\,\mathrm{d}t\,\mathrm{d}y &= -\int_\Omega \int_{\underline{T}}^{\overline{T}} \underbrace{\frac{\partial_t q_t(x,y)}{q_t(x,y)}}_{=\partial_t \log q_t(x,y)} q_t(x,y)\,\mathrm{d}y\,\mathrm{d}t \\
&= -\int_\Omega \int_{\underline{T}}^{\overline{T}} \partial_t q_t(x,y)\,\mathrm{d}t\,\mathrm{d}y = 0.
\end{aligned}
$$

Combining these two observations with $0 < f_{\min} \leq f(y) \leq \|f\|_{\overline{\Omega}} < \infty$ for all $y \in \overline{\Omega}$, and denoting the generator by $\Delta_f = \nabla \cdot f\nabla$, we obtain

$$
\begin{aligned}
\int_{\underline{T}}^{\overline{T}} \int_\Omega \frac{|\nabla_y q_t(x,y)|^2}{q_t(x,y)^2} q_t(x,y)\,\mathrm{d}y\,\mathrm{d}t &\asymp -\int_{\underline{T}}^{\overline{T}} \int_\Omega (\nabla_y \cdot f(y)\nabla_y \log q_t(x,y)) q_t(x,y)\,\mathrm{d}y\,\mathrm{d}t \\
&= -\int_{\underline{T}}^{\overline{T}} \int_\Omega ((\Delta_f + \partial_t)\log q_t(x,y)) q_t(x,y)\,\mathrm{d}y\,\mathrm{d}t \\
&= -\mathbb{E}^x\Big[\int_{\underline{T}}^{\overline{T}} (\Delta_f + \partial_t) \log q_t(x, \mathsf{X}_t)\,\mathrm{d}t\Big] \\
&= \mathbb{E}^x[\log q_{\underline{T}}(x, \mathsf{X}_{\underline{T}})] - \mathbb{E}^x[\log q_{\overline{T}}(x, \mathsf{X}_{\overline{T}})].
\end{aligned}
$$

Noting that $q_t(x,\cdot)$ satisfies the Neumann boundary condition at $\partial\Omega$, the last equality is a consequence of Itô's formula, by which

$$\begin{aligned}
&\log q_{\overline{T}}(x, \mathsf{X}_{\overline{T}}) - \log q_{\underline{T}}(x, \mathsf{X}_{\underline{T}}) \\
&= \int_{\underline{T}}^{\overline{T}} (\partial_t + \Delta_f) \log q_t(x, \mathsf{X}_t)\,\mathrm{d}t + \int_{\underline{T}}^{\overline{T}} \sqrt{2f(\mathsf{X}_t)}\langle \nabla \log q_t(x, \mathsf{X}_t),\, \mathrm{d}\mathsf{B}_t\rangle \\
&\quad + \int_{\underline{T}}^{\overline{T}} \underbrace{\langle \nabla_y \log q_t(x, \mathsf{X}_t),\, \nu(\mathsf{X}_t)\rangle}_{=0}\, \mathrm{d}\mathsf{L}_t,
\end{aligned}$$

where the expectation of the stochastic integral is zero because $(y,t) \mapsto |\sqrt{f(y)}\nabla_y \log q_t(x,y)|$ is bounded on $\overline{\Omega} \times [\underline{T}, \overline{T}]$, which follows from the lower bound

$$\inf_{t\geq t_0, x, y\in\overline{\Omega}} q_t(x,y) > 0, \tag{B.57}$$

for any $t_0 > 0$, see, e.g., Itô [60, p.166], and the space-time smoothness of $(t,x,y) \mapsto q_t(x,y)$ on the compact set $\overline{\Omega} \times [\underline{T}, \overline{T}]$. We finish the proof by noting that for some constant C, the upper heat kernel bound

$$q_t(x,y) \lesssim (t^{-D/2} \vee 1) \exp\Big(- C\frac{|x-y|^2}{t}\Big), \qquad t > 0, x, y \in \overline{\Omega},$$

from Davies [29, Theorem 3.29] yields

$$\sup_{x,y\in\overline{\Omega}} \log q_{\underline{T}}(x,y) \lesssim \frac{D}{2} \log \underline{T}^{-1},$$

and, moreover, (B.57) implies $\inf_{x,y\in\overline{\Omega}} q_{\overline{T}}(x,y) \geq c$, for some constant c not depending on $\underline{T}, \overline{T}$ since $\overline{T} \geq 1$. Putting these bounds together, we conclude that

$$\mathbb{E}^x[\log q_{\underline{T}}(x, \mathsf{X}_{\underline{T}})] - \mathbb{E}^x[\log q_{\overline{T}}(x, \mathsf{X}_{\overline{T}})] \lesssim 1 + \log \underline{T}^{-1},$$

and therefore also

$$\int_{\underline{T}}^{\overline{T}} \int_{\mathcal{X}} \frac{|\nabla_y q_t(x,y)|^2}{q_t(x,y)^2} q_t(x,y)\,\mathrm{d}y\,\mathrm{d}t \lesssim 1 + \log \underline{T}^{-1}.$$

□

Bounding the approximation error

Before going into details, we first outline here our general strategy for bounding the approximation error:

1. approximate the true score by truncation of the spectral decomposition h_N, i.e., approximate $\nabla_x \log p_t(x)$ by $\nabla_x \log h_N(x, t)$;
2. approximate h_N and $\nabla_x h_N$ by neural networks on $[\underline{T}, \overline{T}]$ via
    (a) dividing $[\underline{T}, \overline{T}]$ into sub-intervals of increasing length, totalling a number of intervals on the order of $\log N$;
    (b) on each sub-interval, fix a number of time-points $\{t_i\}$, also on the order of $\log N$, and at each of these, make an approximation of $h_N(t_i)$ and $\nabla_x h_N(t_i)$ using existing results;
    (c) extend these discrete approximations to the entire sub-interval using polynomial interpolation;
    (d) combine approximations on each sub-interval into one final approximation via a partition of unity;
3. combine the first two steps, along with general results on neural networks, to achieve a neural network approximation of $\nabla_x \log p_t$.

At this point, we comment on how and why our approximation strategy differs from those in, for example, [89]. The authors there assume a Gaussian transition kernel with a density that is comparatively easy to approximate using neural networks. The difficulty then lies in approximating the initial density p_0 and the convolution of the two. By contrast, in our spectral composition, the influence of p_0 enters the function only as the weights $\langle p_0, e_j \rangle_{L^2}$. The growth of these weights encodes the smoothness of p_0 and thereby influences the truncation parameter N but they need no approximation by neural networks. Thus, the difficulty lies instead in approximating the eigenfunctions e_j themselves. One might think that one could use existing results, e.g. from [101, 114], to approximate each e_j and the individual time components $t \mapsto \mathrm{e}^{-t\ell_j}$ separately, and then sum them together. However, using these existing results, the sparsity constraint of each summand would be of order at least N, and thus the sparsity constraint of the network as a whole would be at least N^2. This is problematic, since, according to Oko, Akiyama, and Suzuki [89, Lemma C.2], the sparsity constraint enters exponentially in the covering number of the associated class and, consequently, by Theorem B.34, linearly in the generalisation error. This is why we employ a different strategy: we approximate the entire sum at fixed time points and use polynomials to interpolate between them over time. This ultimately gives us a sparsity constraint of order $N\,\mathrm{Poly}(\log N)$. Following this general strategy, we can prove the following score approximation result.

Theorem B.37: 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 $\varphi_s \in \Phi(L(n), W(n), S(n), B(n))$ satisfying

$$\int_{\underline{T}}^{\overline{T}} \int_{\Omega} \left|\varphi_s(x,t) - \nabla_x \log p_t(x)\right|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t \lesssim n^{-\frac{2\alpha}{2\alpha+D}} (\log n)^2 (\overline{T} + \log(\underline{T}^{-1})). \tag{B.58}$$

The size of the network is evaluated as

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

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 Ω such that $|\varphi_s(x,t)| \leq \frac{C}{\sqrt{t}}$ for all $t \in [\underline{T}, \overline{T}]$ and $x \in \Omega$.

As alluded to above, the idea is to break up the score approximation error by using the spectral score representation (B.44) and to reduce this to the problem of approximating $\nabla \log h_N(x,t) = \nabla h_N(x,t)/h_N(x,t)$, where, for $N \in \mathbb{N}$ and $t \in [0,T]$, the truncated series $h_N(t) = h_N(\cdot, t)$ is given by

$$h_N(t) := \sum_{j=0}^{N} \mathrm{e}^{-t\ell_j} \langle p_0, e_j \rangle_{L^2} e_j. \tag{B.59}$$

It will then be crucial to choose the cutoff value N of the right order in terms of n to balance the tradeoff between the error incurred through the truncation procedure and the increased approximation quality of $\nabla \log h_N$ in terms of neural networks for smaller N. Our analysis will demonstrate that for the desired approximation accuracy $n^{-\alpha/(2\alpha+D)}$, the choice $N \asymp n^{D/(2\alpha+D)}$ is appropriate.

Step 1: Bounding the truncation loss We start by considering the approximation properties of h_N and $\nabla_x h_N$. To this end, let us introduce the homogeneous Sobolev space of order α that is induced by the eigendecomposition of $-\nabla \cdot f \nabla$ via

$$\bar{H}^\alpha(\Omega) := \left\{ \phi \in L_0^2(\Omega) : \|\phi\|_{\bar{H}^\alpha}^2 := \sum_{j \geq 1} \ell_j^\alpha \langle \phi, e_j \rangle_{L^2}^2 < \infty \right\}.$$

Lemma B.38: It holds that

$$\int_{\underline{T}}^{\overline{T}} \|p_t - h_N(t)\|_{H^1}^2 \,\mathrm{d}t \lesssim \|p_0 - \tfrac{1}{\lambda_D(\Omega)}\|_{H^\alpha}^2 N^{-2\alpha/D}.$$

Proof. Arguing as in Nickl [85, Proposition 3], we see that $p_t, h_N(t) \in H^k(\Omega)$ for any $k \in \mathbb{N}$.

Since $p_t - h_N(t) \in L_0^2$, Nickl [85, Proposition 2] shows that $p_t - h_N(t) \in \bar{H}^1(\Omega)$ and $\|p_t - h_N(t)\|_{\bar{H}^1} \asymp \|p_t - h_N(t)\|_{H^1}$. Thus,

$$\begin{aligned}\int_{\underline{T}}^{\overline{T}} \|p_t - h_N(t)\|_{H^1}^2 \,\mathrm{d}t &\asymp \int_{\underline{T}}^{\overline{T}} \|p_t - h_N(t)\|_{\bar{H}^1}^2 \,\mathrm{d}t \\ &= \sum_{j \geq N+1} \int_{\underline{T}}^{\overline{T}} \ell_j \mathrm{e}^{-2\ell_j t} \,\mathrm{d}t \,\langle p_0, e_j\rangle_{L^2}^2 \\ &\leq \sum_{j \geq N+1} \langle p_0, e_j\rangle_{L^2}^2.\end{aligned}$$

As $p_0 \in H_c^\alpha(\Omega)/\mathbb{R}$, it holds $p_0 - \frac{1}{\lambda_D(\Omega)} \in H_c^\alpha(\Omega)/\mathbb{R} \cap L_0^2(\Omega)$.

Furthermore, by Nickl [85, Proposition 2], we have $H_c^\alpha(\Omega)/\mathbb{R} \cap L_0^2 \subset \bar{H}^\alpha(\Omega)$ and $\|\phi\|_{H^\alpha} \asymp \|\phi\|_{\bar{H}^\alpha}$ for $\phi \in \bar{H}^\alpha(\Omega)$, implying that $\|p_0 - \frac{1}{\lambda_D(\Omega)}\|_{H^\alpha} \asymp \|p_0 - \frac{1}{\lambda_D(\Omega)}\|_{\bar{H}^\alpha}$.

Using this and $\ell_j \asymp j^{2/D}$, it follows

$$\begin{aligned}\sum_{j \geq N+1} \langle p_0, e_j\rangle_{L^2}^2 \leq \frac{\sum_{j=1}^\infty \langle p_0, e_j\rangle_{L^2}^2 j^{2\alpha/D}}{(N+1)^{2\alpha/D}} &= \frac{\sum_{j=1}^\infty \langle p_0 - \frac{1}{\lambda_D(\Omega)}, e_j\rangle_{L^2}^2 j^{2\alpha/D}}{(N+1)^{2\alpha/D}} \\ &\lesssim \|p_0 - \tfrac{1}{\lambda_D(\Omega)}\|_{H^\alpha}^2 N^{-2\alpha/D}.\end{aligned} \tag{B.60}$$

□

This result allows us to study the approximation quality of (a truncated version of) $\nabla \log h_N$.

Proposition B.39: It holds that

$$\int_{\underline{T}}^{\overline{T}} \int_\Omega \left|\nabla_x \log p_t(x) - \frac{\nabla_x h_N(x,t)}{h_N(x,t) \vee p_{\min}}\right|^2 p_t(x) \,\mathrm{d}x \,\mathrm{d}t \lesssim C(p_0)(1 + \ell_1^{-1}) p_{\min}^{-4} \log(\underline{T}^{-1}) N^{-2\alpha/D},$$

where

$$C(p_0) = \|p_0\|_\infty \|p_0 - \tfrac{1}{\lambda_D(\Omega)}\|_{H^\alpha}^2 (1 + \|p_0\|_{H^\alpha}^2).$$

Proof. Symmetry of the transition densities implies that for any $t \geq 0, x \in \Omega$,

$$p_t(x) = \int_\Omega p_0(y) q_t(y,x) \,\mathrm{d}y = \int_\Omega p_0(y) q_t(x,y) \,\mathrm{d}y \leq \|p_0\|_\infty \int_\Omega q_t(x,y) \,\mathrm{d}y = \|p_0\|_\infty,$$

and, similarly,

$$p_t(x) \geq p_{\min}.$$

It follows that

$$\begin{aligned}
&\int_{\underline{T}}^{\overline{T}} \int_{\Omega} \left| \nabla_x \log p_t(x) - \frac{\nabla_x h_N(x,t)}{h_N(x,t) \vee p_{\min}} \right|^2 p_t(x) \,\mathrm{d}x \,\mathrm{d}t \\
&\quad \leq \|p_0\|_\infty \int_{\underline{T}}^{\overline{T}} \left\| \nabla_x \log p_t - \frac{\nabla_x h_N(t)}{h_N(t) \vee p_{\min}} \right\|_{L^2}^2 \mathrm{d}t \\
&\quad \leq 2\|p_0\|_\infty \Big(\int_{\underline{T}}^{\overline{T}} \left\| \frac{\nabla_x (p_t - h_N(t))}{h_N(t) \vee p_{\min}} \right\|_{L^2}^2 \mathrm{d}t + \int_{\underline{T}}^{\overline{T}} \left\| \nabla_x p_t \frac{p_t - (h_N(t) \vee p_{\min})}{p_t (h_N(t) \vee p_{\min})} \right\|_{L^2}^2 \mathrm{d}t \Big) \\
&\quad \leq 2\|p_0\|_\infty \Big(\frac{1}{p_{\min}^2} \int_{\underline{T}}^{\overline{T}} \left\| \nabla_x (p_t - h_N(t)) \right\|_{L^2}^2 \mathrm{d}t + \frac{1}{p_{\min}^4} \int_{\underline{T}}^{\overline{T}} \left\| \nabla_x p_t (p_t - (h_N(t) \vee p_{\min})) \right\|_{L^2}^2 \mathrm{d}t \Big).
\end{aligned} \tag{B.61}$$

By Lemma B.38, we obtain for the first term

$$\begin{aligned}
\frac{1}{p_{\min}^2} \int_{\underline{T}}^{\overline{T}} \left\| \nabla_x (p_t - h_N(t)) \right\|_{L^2}^2 \mathrm{d}t &\leq p_{\min}^{-2} \int_{\underline{T}}^{\overline{T}} \|p_t - h_N(t)\|_{H^1}^2 \,\mathrm{d}t \\
&\lesssim \|p_0 - \tfrac{1}{\lambda_D(\Omega)}\|_{H^\alpha}^2 p_{\min}^{-2} N^{-2\alpha/D}.
\end{aligned} \tag{B.62}$$

To bound the second term, note first that

$$\begin{aligned}
\|p_t - \tfrac{1}{\lambda_D(\Omega)}\|_{\bar{H}^{\alpha+1}}^2 = \sum_{j=1}^{\infty} \mathrm{e}^{-2\ell_j t} \ell_j^{\alpha+1} \langle p_0, e_j \rangle_{L^2}^2 \leq \frac{1}{2t} \sum_{j \geq 1} \ell_j^{\alpha} \langle p_0, e_j \rangle_{L^2}^2 &= \frac{2}{t} \|p_0 - \tfrac{1}{\lambda_D(\Omega)}\|_{\bar{H}^\alpha}^2 \\
&\asymp \frac{2}{t} \|p_0 - \tfrac{1}{\lambda_D(\Omega)}\|_{H^\alpha}^2 < \infty, \quad t > 0.
\end{aligned}$$

Therefore, for any $i \in [d]$,

$$\begin{aligned}
\|\partial_{x_i} p_t\|_{H^\alpha} = \|\partial_{x_i} (p_t - \tfrac{1}{\lambda_D(\Omega)})\|_{H^\alpha} \leq \|p_t - \tfrac{1}{\lambda_D(\Omega)}\|_{H^{\alpha+1}} &\asymp \|p_t - \tfrac{1}{\lambda_D(\Omega)}\|_{\bar{H}^{\alpha+1}} \\
&\lesssim \frac{1}{\sqrt{t}} \|p_0 - \tfrac{1}{\lambda_D(\Omega)}\|_{H^\alpha}.
\end{aligned}$$

Since $\alpha > D/2$, the Sobolev imbedding theorem yields

$$\sup_{i \in [d]} \|\partial_{x_i} p_t\|_\infty \lesssim t^{-1/2} \|p_0 - \tfrac{1}{\lambda_D(\Omega)}\|_{H^\alpha}.$$

Consequently, for N as above,

$$\begin{aligned}\int_{\underline{T}}^{\overline{T}} \big\|\nabla_x p_t(p_t - (h_N(t) \vee p_{\min}))\big\|_{L^2}^2 \, \mathrm{d}t &\lesssim \|p_0 - \tfrac{1}{\lambda_D(\Omega)}\|_{H^\alpha}^2 \int_{\underline{T}}^{\overline{T}} \frac{1}{t} \|p_t - (h_N(t) \vee p_{\min})\|_{L^2}^2 \, \mathrm{d}t \\ &\leq \|p_0 - \tfrac{1}{\lambda_D(\Omega)}\|_{H^\alpha}^2 \int_{\underline{T}}^{\overline{T}} \frac{1}{t} \|p_t - h_N(t)\|_{L^2}^2 \, \mathrm{d}t \\ &= \|p_0 - \tfrac{1}{\lambda_D(\Omega)}\|_{H^\alpha}^2 \int_{\underline{T}}^{\overline{T}} \frac{1}{t} \mathrm{e}^{-2\ell_1 t} \, \mathrm{d}t \sum_{j \geq N+1} \langle p_0, e_j \rangle_{L^2}^2 \\ &\lesssim \|p_0 - \tfrac{1}{\lambda_D(\Omega)}\|_{H^\alpha}^4 N^{-2\alpha/D} \Big(\int_{\underline{T}}^{1} \frac{1}{t} \, \mathrm{d}t + \int_{1}^{\overline{T}} \mathrm{e}^{-\ell_1 t} \, \mathrm{d}t \Big) \\ &\leq \|p_0 - \tfrac{1}{\lambda_D(\Omega)}\|_{H^\alpha}^4 (1 + 1/\ell_1) \log(\underline{T}^{-1}) N^{-2\alpha/D}, \end{aligned} \tag{B.63}$$

where we used (B.60) and that $|p_t(x) - (h_N(x,t) \vee p_{\min})| \leq |p_t(x) - h_N(x,t)|$ since $p_t(x) \geq p_{\min}$.

Combining (B.61), (B.62) and (B.63) yields the assertion. ☐

Step 2: Approximation of truncated score by neural networks Given the previous result, it remains to show that (the truncated version of) $\nabla_x \log h_N$ with h_N as defined in (B.59) can be well approximated by a neural network whose size is quantified in terms of N.

To this end, we will make repeated use of the following fundamental observations about compositions and parallelisations of the type of ReLU neural networks introduced in Section B.2 that we are dealing with.

First, we observe that the concatenation of such neural networks is itself simply another neural network. In particular, if $\varphi_1 \in \Phi(L_1, W_1, S_1, B_1)$ and $\varphi_2 \in \Phi(L_2, W_2, S_2, B_2)$ are such that $W_{2,L_2+2} = W_{1,1}$, then $\varphi_1 \circ \varphi_2 \in \Phi(L_1 + L_2 + 1, W_{1,2}^{\mathrm{cat}}, S_1 + S_2, B_1 \vee B_2)$, where

$$W_{1,2}^{\mathrm{cat}} = \begin{bmatrix} W_{2,1} & W_{2,2} & \cdots & W_{2,L_2+1} & W_{1,1} & W_{1,2} & \cdots & W_{1,L_1+2} \end{bmatrix}^\top.$$

In general, if for some $k \in \mathbb{N}$, $\varphi_i \in \Phi(L_i, W_i, S_i, B_i)$ for $i = 1, \ldots, k$, then

$$\varphi_1 \circ \varphi_2 \circ \cdots \circ \varphi_k \in \Phi\Big(\sum_{i=1}^{k} L_i + k, W_{[k]}^{\mathrm{cat}}, \sum_{i=1}^{k} S_i, \max_{i \in \{1,\ldots,k\}} B_i \Big),$$

where $W_{[k]}^{\mathrm{cat}}$ is defined recursively as above. Similarly, if φ_1, φ_2 are as before but with $L_1 = L_2 = L$, we can parallelise the two into one network $\varphi_{1,2}^{\mathrm{par}} \in \Phi(L, W_{1,2}^{\mathrm{par}}, S_1 + S_2, B_1 \vee B_2)$ such that $\varphi_{1,2}^{\mathrm{par}}(x, y) = \begin{bmatrix} \varphi_1(x) & \varphi_2(y) \end{bmatrix}^\top$ for $x \in \mathbb{R}^{W_{1,1}}$ and $y \in \mathbb{R}^{W_{2,1}}$. The simplest way to construct this is using block matrices, i.e., by

$$\varphi_{1,2}^{\mathrm{par}} = \begin{bmatrix} A_{1,L} & 0 \\ 0 & A_{2,L} \end{bmatrix} \sigma_{\begin{bmatrix} b_{1,L} \\ b_{2,L} \end{bmatrix}} \begin{bmatrix} A_{1,L-1} & 0 \\ 0 & A_{2,L-1} \end{bmatrix} \sigma_{\begin{bmatrix} b_{1,L-1} \\ b_{2,L-1} \end{bmatrix}} \cdots \begin{bmatrix} A_{1,1} & 0 \\ 0 & A_{2,1} \end{bmatrix} \sigma_{\begin{bmatrix} b_{1,1} \\ b_{2,1} \end{bmatrix}} \begin{bmatrix} A_{1,0} & 0 \\ 0 & A_{2,0} \end{bmatrix},$$

which would mean $W_{1,2}^{\text{par}} = W_1 + W_2$.

However, if φ_1 and φ_2 share some inputs (i.e., if the first, say, $m \in \mathbb{N}$ entries of x, y are the same), then the rightmost matrix in the above may be altered to

$$\begin{bmatrix} A_{1,0} & 0 \\ 0 & A_{2,0} \end{bmatrix} \begin{bmatrix} I_m & 0 & 0 \\ 0 & I_{W_{1,1}-m} & 0 \\ I_m & 0 & 0 \\ 0 & 0 & I_{W_{2,1}-m} \end{bmatrix},$$

whereby $(W_{1,2}^{\text{par}})_1 = W_{1,1} + W_{2,1} - m$ instead. Again, this can of course naturally be generalised to k networks of equal depth, where we then have

$$\varphi_{[k]}^{\text{par}} \in \Phi\Big(L, W_{[k]}^{\text{par}}, \sum_{i=1}^{k} S_k, \max_{i\in\{1,\dots,k\}} B_i\Big), \quad (W_{[k]}^{\text{par}})_j = \sum_{i=1}^{k} W_{i,j} \text{ for } j > 1.$$

Finally, note that multiplying the network $\varphi_{[k]}^{\text{par}}$ with the vector $[1 \ \cdots \ 1]$ from the left sums the entries of $\varphi_{[k]}^{\text{par}}$ without changing the size of the network substantially, whence

$$\sum_{i=1}^{k} \varphi_i \in \Phi\Big(L, W_{[k]}^{\text{sum}}, k + \sum_{i=1}^{k} S_k, 1 \vee \max_{i\in\{1,\dots,k\}} B_i\Big), \quad (W_{[k]}^{\text{sum}})_j = \sum_{i=1}^{k} W_{i,j} \text{ for } 1 < j < k.$$

For larger and more complicated neural networks, their exact sizes are often unavailable, and we only have access to their asymptotic sizes. Due to this, we also introduce the following class of neural networks that eases network size analysis in the proofs that follow,

$$\widetilde{\Phi}(\widetilde{L}, \widetilde{W}, \widetilde{S}, \widetilde{B}) := \Big\{ \varphi \in \Phi(L, W, S, B) \, : \, L \lesssim \widetilde{L}, \|W\|_\infty \lesssim \widetilde{W}, S \lesssim \widetilde{S} \text{ and } B \lesssim \widetilde{B} \Big\}.$$

With this notation, we have for arbitrary networks $\varphi_i \in \widetilde{\Phi}(L_i, W_i, S_i, B_i)$, $i = 1, 2$, that $\varphi_1 \circ \varphi_2 \in \widetilde{\Phi}(L_1 + L_2, W_1 \vee W_2, S_1 + S_2, B_1 \vee B_2)$ and $\varphi_{1,2}^{\text{par}} \in \widetilde{\Phi}(L_1 \vee L_2, W_1 + W_2, S_1 + S_2, B_1 \vee B_2)$.

With this class established, we can begin to approximate $\nabla_x \log h_N$ with a suitable network. We do this by approaching the problem in smaller pieces, first noting that, for $t \geq 0$ and $x \in \Omega$,

$$\nabla_x \log h_N(x, t) = \frac{\sum_{j=1}^{N} \mathrm{e}^{-t\ell_j} \langle p_0, e_j \rangle_{L^2} \nabla_x e_j(x)}{\frac{1}{\lambda_D(\Omega)} + \sum_{j=1}^{N} \mathrm{e}^{-t\ell_j} \langle p_0, e_j \rangle_{L^2} e_j(x)}.$$

Thus, in order to approximate $\nabla_x \log h_N$, we need to be able to approximate products and quotients of functions. Such approximation results already exist in the literature, see, e.g., [89, 101, 121, 134], but here we give slightly stronger versions with optimised neural network sizes, which lead to improved convergence rates (in terms of log factors). More detailed statements and their proofs are given in Appendix B.C, but for our purposes the following will suffice.

Lemma B.40: For $m \in \mathbb{N}$ and $C \geq 1$, there exist neural networks $\varphi_m^{\mathrm{mult}} \in \widetilde{\Phi}(m, 1, m, C)$ and $\varphi_m^{\mathrm{mult},D} \in \widetilde{\Phi}(m, D, Dm, C)$ satisfying

$$|\varphi_m^{\mathrm{mult}}(x, y) - xy| \leq C2^{-m}, \quad x \in [0, 1], y \in [-C, C],$$

and

$$|\varphi_m^{\mathrm{mult},D}(x, y) - xy| \leq \sqrt{D}C2^{-m}, \quad x \in [0, 1], y \in [-C, C]^D.$$

These also satisfy $\varphi_m^{\mathrm{mult}}(x, 0) = \varphi_m^{\mathrm{mult}}(0, y) = 0$.

Lemma B.41: For $m, \underline{k}, \overline{k} \in \mathbb{N}$, there exists a neural network

$$\varphi_m^{\mathrm{rec}} \in \widetilde{\Phi}((k + m)\log(k + m), k, (k + m)\log(k + m), 2^k),$$

where $k = \underline{k} + \overline{k}$, satisfying

$$|\varphi_m^{\mathrm{rec}}(x) - x^{-1}| \leq 2^{-m}, \qquad x \in [2^{-\underline{k}}, 2^{\overline{k}}].$$

Lemma B.42: For each $m \in \mathbb{N}$, there exists a neural network

$$\varphi^{\mathrm{cap}} \in \widetilde{\Phi}(m \log m, m, m \log m, 2^{m/2})$$

satisfying $\varphi^{\mathrm{cap}}(t) \asymp \frac{1}{\sqrt{t}}$ for all $t \in [2^{-m}, 1]$.

Lemma B.43: Let $0 < \underline{T} < \overline{T}$ with $\underline{T} \in \mathrm{Poly}(N^{-1})$ be given, and let f denote either h_N or $\partial_{x_i} h_N$, for some $i \in \{1, \dots, d\}$ and $N \in \mathbb{N}$ sufficiently large. Then, there exists a neural network $\varphi_f \in \Phi(L(N), W(N), S(N), B(N))$ satisfying

$$\forall t \in [\underline{T}, \overline{T}] \, : \, \|\varphi_f(\,\cdot\,, t) - f(\,\cdot\,, t)\|_{L^2}^2 \lesssim \begin{cases} N^{-\frac{2\alpha}{D}}(\log N)^2, & \text{if } f = h_N \\ \varepsilon(t)N^{-\frac{2\alpha}{D}}(\log N)^2, & \text{if } f = \partial_{x_i} h_N, \end{cases}$$

where

$$\int_{\underline{T}}^{\overline{T}} \varepsilon(t)\,\mathrm{d}t \lesssim \overline{T} + \log(\underline{T}^{-1})$$

and whose network size is evaluated as

$$\begin{aligned} L(N) &= \log N \log\log N, \\ W(N) &= MN\log N, \\ S(N) &= MN(\log N)^2, \quad \text{and} \\ B(N) &= \frac{N^{\frac{2\alpha+D}{2D}}}{\log N} \vee \frac{1}{\underline{T}}, \end{aligned}$$

where $M \in O(\log \frac{\overline{T}}{\underline{T}})$. Furthermore, there exists a constant $C < \infty$ depending only on p_0 and D such that $\sup_{x\in\Omega} |\varphi_{\partial_{x_i} h_N}(x,t)| \leq C(1 \vee \frac{1}{\sqrt{t}})$ for all $t \in [\underline{T}, \overline{T}]$.

Proof. We first construct a network φ_f with the desired error rate and specify its size at the end. To this end, suppose that there exist neural networks $\varphi_f^{(1)}, \dots, \varphi_f^{(M)}$, where $M = \lfloor \log_2 \frac{\overline{T}}{\underline{T}} \rfloor$ such that

$$\|\varphi_f^{(m)}(\cdot,t) - f(\cdot,t)\|_{L^2}^2 \lesssim \begin{cases} N^{-\frac{2\alpha}{D}}(\log N)^2, & \text{if } f = h_N, \\ \big(\frac{1}{2^{m-1}\underline{T}} \vee 1\big) N^{-\frac{2\alpha}{D}}(\log N)^2, & \text{if } f = \partial_{x_i} h_N, \end{cases}$$

for $m = 1, \dots, M$ and $t \in [2^{m-1}\underline{T}, 2^{m+1}\underline{T}]$. Then, consider the partition of unity $\{\pi_m\}_{m=1}^M$ given by

$$\pi_m(t) = 0 \vee \left(\frac{t - 2^{m-1}\underline{T}}{2^{m-1}\underline{T}} \wedge \frac{2^{m+1}\underline{T} - t}{2^m \underline{T}} \right) = \begin{cases} \frac{t}{2^{m-1}\underline{T}} - 1, & \text{if } t \in [2^{m-1}\underline{T}, 2^m\underline{T}] \\ 2 - \frac{t}{2^m\underline{T}}, & \text{if } t \in [2^m\underline{T}, 2^{m+1}\underline{T}] \\ 0, & \text{otherwise} \end{cases},$$

for $m = 2, \dots, M-1$, while $\pi_1(t) = 0 \vee \Big(1 \wedge \frac{4\underline{T}-t}{2\underline{T}}\Big)$ and $\pi_M(t) = 0 \vee \Big(1 \wedge \frac{t - 2^{M-1}\underline{T}}{2^{M-1}\underline{T}}\Big)$. Since for $a, b \in \mathbb{R}$, $a \vee b = a + \sigma(b-a)$ and $a \wedge b = a - \sigma(a-b)$, each π_m is representable as a neural network. We then claim that φ_f, defined as

$$\varphi_f(x,t) = \sum_{m=1}^{M} \varphi_{m_1}^{\text{mult}}\big(\pi_m(t), \varphi_f^{(m)}(x,t)\big), \qquad m_1 = \left\lceil \frac{\alpha}{D} \log_2 N \right\rceil,$$

yields the desired network. Indeed, we first note that since at most two of the π_m's are non-zero for any $t \in [\underline{T}, \overline{T}]$ and $\varphi_m^{\text{mult}}(0,y) = 0$ for all $y \in \mathbb{R}$, we have for $m = 2, \dots, M$ and $t \in [2^{m-1}\underline{T}, 2^m\underline{T}]$ that

$$\begin{aligned} \|\varphi_f(\cdot,t) - f(\cdot,t)\|_{L^2} &= \|\varphi_{m_1}^{\text{mult}}\big(\pi_{m-1}(t), \varphi_f^{(m-1)}(\cdot,t)\big) + \varphi_{m_1}^{\text{mult}}\big(\pi_m(t), \varphi_f^{(m)}(\cdot,t)\big) - f(\cdot,t)\|_{L^2} \\ &\lesssim 2^{-m_1} + \|\pi_{m-1}(t)\varphi_f^{(m-1)}(\cdot,t) + \pi_m(t)\varphi_f^{(m)}(\cdot,t) - f(\cdot,t)\|_{L^2} \\ &\lesssim \begin{cases} 2^{-m_1} + N^{-\frac{\alpha}{D}}\log N, & \text{if } f = h_N \\ 2^{-m_1} + \big(\frac{1}{\sqrt{2^{m-2}\underline{T}}} \vee 1\big) N^{-\frac{\alpha}{D}}\log N, & \text{if } f = \partial_{x_i} h_N, \end{cases} \end{aligned}$$

where in the last inequality we used

$$\begin{aligned}
&\|\pi_{m-1}(t)\varphi_f^{(m-1)}(\cdot,t) + \pi_m(t)\varphi_f^{(m)}(\cdot,t) - f(\cdot,t)\|_{L^2} \\
&\quad= \|\pi_{m-1}(t)\big(\varphi_f^{(m-1)}(\cdot,t) - f(\cdot,t)\big) + \pi_m(t)\big(\varphi_f^{(m)}(\cdot,t) - f(\cdot,t)\big)\|_{L^2} \\
&\quad\le \pi_{m-1}(t)\|\varphi_f^{(m-1)}(\cdot,t) - f(\cdot,t)\|_{L^2} + \pi_m(t)\|\varphi_f^{(m)}(\cdot,t) - f(\cdot,t)\|_{L^2} \\
&\quad\lesssim \begin{cases} N^{-\frac{\alpha}{D}}\log N, & \text{if } f = h_N \\ \big(\frac{1}{\sqrt{2^{m-2}\underline{T}}} \vee 1\big)N^{-\frac{\alpha}{D}}\log N, & \text{if } f = \partial_{x_i}h_N, \end{cases}.
\end{aligned}$$

Setting

$$\varepsilon(t) = \sum_{m=1}^{M+1}\Big(\frac{1}{2^{m-2}\underline{T}} \vee 1\Big)\mathbf{1}_{[2^{m-1}\underline{T},2^m\underline{T}]}(t),$$

and by choice of m_1, squaring both sides of the inequality yields the desired error rate. A similar but simpler analysis shows that this also holds for $t \in [\underline{T}, 2\underline{T}]$ and $t \in [2^M\underline{T}, 2^{M+1}\underline{T}]$, whence the inequality holds for all $t \in [\underline{T}, 2^{M+1}\underline{T}] \supseteq [\underline{T}, \overline{T}]$. Furthermore, we have

$$\begin{aligned}
\int_{\underline{T}}^{\overline{T}} \varepsilon(t)\,\mathrm{d}t &\le \sum_{m=1}^{M+1}\Big(\frac{1}{2^{m-2}\underline{T}} \vee 1\Big)(2^m\underline{T} - 2^{m-1}\underline{T}) \\
&= \sum_{m=1}^{\lfloor\log_2(\underline{T}^{-1})\rfloor+2} \frac{2^{m-1}\underline{T}}{2^{m-2}\underline{T}} + \sum_{m=\lfloor\log_2(\underline{T}^{-1})\rfloor+3}^{M+1} 2^{m-1}\underline{T} \\
&= 2\big(\big\lfloor\log_2(\underline{T}^{-1})\big\rfloor + 2\big) + \underline{T}\Big(\sum_{m=0}^{M}2^m - \sum_{m=0}^{\lfloor\log_2(\underline{T}^{-1})\rfloor+2} 2^m\Big) \\
&= 2\big(\big\lfloor\log_2(\underline{T}^{-1})\big\rfloor + 2\big) + 2\underline{T}\Big(2^M - 2^{\lfloor\log_2(\underline{T}^{-1})\rfloor}\Big) \\
&\lesssim \overline{T} + \log(\underline{T}^{-1})
\end{aligned}$$

as claimed.

As such, we only need to construct the networks $\varphi_f^{(m)}$ for all $m \in \{1,\dots,M\}$, so let some such m be fixed. Then, let $a_m = 3\cdot 2^{m-2}\underline{T}$ and $b_m = 5\cdot 2^{m-2}\underline{T}$, and set $f_m(x,t) = f(x, a_m t + b_m)$ such that $f_m(x,[-1,1]) = f(x,[2^{m-1}\underline{T}, 2^{m+1}\underline{T}])$ for all $x \in \Omega$. As in the proof of Lemma B.38, we see that for each fixed $t \in [\underline{T}, \overline{T}]$, $h_N(\cdot,t) \in H^{s+1}(\Omega)$, whence $f(\cdot,t) \in H^\alpha(\Omega) = B^\alpha_{2,2}(\Omega)$. Furthermore, $\|f(\cdot,t)/(1 \vee \|f(\cdot,t)\|_{B^\alpha_{2,2}(\Omega)})\|_{B^\alpha_{2,2}(\Omega)} \le 1$. Thus, since D is bounded, a slight modification of Suzuki [114, Proposition 1] using the Sobolev extension theorem yields the existence of a neural network $\widetilde{\varphi}_{f,t} \in \widetilde{\Phi}(\log N, N, N\log N, N^{1/D})$ satisfying $\|\widetilde{\varphi}_{f,t} - f(\cdot,t)/(1 \vee \|f(\cdot,t)\|_{B^\alpha_{2,2}(\Omega)})\|_{L^2(\Omega)} \lesssim N^{-\frac{\alpha}{D}}$. Then, setting $\varphi_{f,t} = (1 \vee \|f(\cdot,t)\|_{B^\alpha_{2,2}(\Omega)})\widetilde{\varphi}_{f,t}$, we have

$$\begin{aligned}
\|\varphi_{f,t} - f(\cdot,t)\|_{L^2(\Omega)} &= (1 \vee \|f(\cdot,t)\|_{B^\alpha_{2,2}(\Omega)})\Big\|\widetilde{\varphi}_{f,t} - \frac{f(\cdot,t)}{(1 \vee \|f(\cdot,t)\|_{B^\alpha_{2,2}(\Omega)})}\Big\|_{L^2(\Omega)} \\
&\lesssim (1 \vee \|f(\cdot,t)\|_{B^\alpha_{2,2}(\Omega)})N^{-\frac{\alpha}{D}}.
\end{aligned}$$

Noting that

$$\|f(\cdot,t)\|_{B^{\alpha}_{2,2}(\Omega)} \asymp \|f(\cdot,t)\|_{H^{\alpha}(\Omega)} \lesssim \begin{cases} \|p_0\|_{H^{\alpha}}, & \text{if } f = h_N, \\ \frac{1}{\sqrt{t}}\|p_0\|_{H^{\alpha}}, & \text{if } f = \partial_{x_i} h_N, \end{cases}$$

we thus have that $\varphi_{f,t} \in \widetilde{\Phi}(\log N, N, N\log N, N^{1/D} \vee \frac{1}{\sqrt{\underline{T}}})$, and

$$\|\varphi_{f,t} - f(\cdot,t)\|_{L^2(\Omega)} \lesssim \begin{cases} N^{-\frac{\alpha}{D}}, & \text{if } f = h_N, \\ \left(\frac{1}{\sqrt{t}} \vee 1\right) N^{-\frac{\alpha}{D}}, & \text{if } f = \partial_{x_i} h_N. \end{cases}$$

Then, for each $t \in [-1,1]$, let $\varphi_{f_m,t} = \varphi_{f,a_m t + b_m}$ such that $\varphi_{f_m,t}$ is an approximation of $f_m(t,\cdot)$. We will then approximate f_m by polynomial interpolation in time and by $\varphi_{f_m,t}$ in space. There are then three main sources of error: the error in approximating a polynomial with a neural network, the error from polynomial interpolation, and finally the error from approximating $f_m(\cdot,t)$ by $\varphi_{f_m,t}$. To separate these sources of error, we now let $\{t_i\}_{i=0}^{k}$ be the first $k+1$ Chebyshev nodes on $[-1,1]$ for some k to be determined later, i.e. $t_i = \cos\frac{i\pi}{k}$. Then, for $i = 0,\dots,k$, let $p_i(t) = \prod_{j\neq i}(t - t_j)$ and set $c_i = \frac{1}{p_i(t_i)}$. Furthermore, set

$$\begin{aligned} \varphi_{f_m}(x,t) &= \sum_{i=0}^{k} c_i \varphi^{\text{mult}}_{m_2}\left(\varphi_{p_i}(t), \varphi_{f_m,t_i}(x)\right), \\ \psi_m(x,t) &= \sum_{i=0}^{k} c_i p_i(t) \varphi_{f_m,t_i}(x), \\ P_m(x,t) &= \sum_{i=0}^{k} c_i p_i(t) f_m(x,t_i), \end{aligned}$$

where φ_{p_i} is a neural network approximation of p_i satisfying $|\varphi_{p_i}(t) - p_i(t)| \lesssim k2^{-m_3}$ to be constructed later. We then have that

$$\begin{aligned} \|\varphi_{f_m}(\cdot,t) - f_m(\cdot,t)\|_{L^2} \leq \|\varphi_{f_m}(\cdot,t) - \psi_m(\cdot,t)\|_{L^2} + \|\psi_m(\cdot,t) - P_m(\cdot,t)\|_{L^2} \\ + \|P_m(\cdot,t) - f_m(\cdot,t)\|_{L^2}. \end{aligned} \tag{B.64}$$

By Trefethen [122, Theorem 5.2], it holds that $|c_i| \leq \frac{2^{k-1}}{k}$, and so the first term of (B.64) is upper bounded by

$$\begin{aligned}\|\varphi_{f_m}(\,\cdot\,,t)-\psi_m(\,\cdot\,,t)\|_{L^2} &\leq \sum_{i=0}^{k}|c_i|\|\varphi^{\mathrm{mult}}_{m_2}\big(\varphi_{p_i}(t),\varphi_{f_m,t_i}\big)-p_i(t)\varphi_{f_m,t_i}\|_{L^2}\\ &\lesssim \sum_{i=0}^{k}|c_i|(2^{-m_2}+k2^{-m_3}\|\varphi_{f_m,t_i}\|_{L^2})\\ &\lesssim \sum_{i=0}^{k}|c_i|(2^{-m_2}+k2^{-m_3}(\underline{T}^{-\frac{1}{2}}N^{-\frac{\alpha}{D}}+\|f_m(\,\cdot\,,t_i)\|_{L^2}))\\ &\leq 2^{k-1}(2^{-m_2}+k2^{-m_3}(\underline{T}^{-\frac{1}{2}}N^{-\frac{\alpha}{D}}+\|p_0\|_{H^1})),\end{aligned}$$

and choosing $m_2=\lceil\frac{\alpha}{D}\log_2 N+k\rceil$ and $m_3=m_2+\lceil\log_2(k+\underline{T}^{-\frac{1}{2}})\rceil$ bounds this term by $N^{-\alpha/D}$. For the second term of (B.64), it is a well-known property of Chebyshev nodes that $|p_i(t)c_i|\leq 2$, whence

$$\begin{aligned}\|\psi_m(\,\cdot\,,t)-P_m(\,\cdot\,,t)\|_{L^2} &\leq \sum_{i=0}^{k}|c_i p_i(t)|\|\varphi_{f_m,t_i}-f_m(\,\cdot\,,t_i)\|_{L^2}\\ &\lesssim \begin{cases} kN^{-\frac{\alpha}{D}}, & \text{if } f=h_N\\ k\big(\frac{1}{\sqrt{2^{m-1}\underline{T}}}\vee 1\big)N^{-\frac{\alpha}{D}}, & \text{if } f=\partial_{x_i}h_N,\end{cases}\end{aligned}$$

where the last inequality in the case of $f=\partial_{x_i}h_N$ follows from the fact that for all $t\in[-1,1]$

$$\begin{aligned}\|\varphi_{f_m,t}-f_m(\,\cdot\,,t)\|_{L^2} &= \|\varphi_{f,a_m t+b_m}-f(\,\cdot\,,a_m t+b_m)\|_{L^2}\\ &\lesssim \Big(\frac{1}{\sqrt{a_m t+b_m}}\vee 1\Big)N^{-\frac{\alpha}{D}}\\ &\leq \Big(\frac{1}{\sqrt{2^{m-1}\underline{T}}}\vee 1\Big)N^{-\frac{\alpha}{D}}.\end{aligned}$$

Finally, for the third term of (B.64), we note that for each $x\in\Omega$, the function $t\mapsto f_m(x,t)$ is entire on $\mathbb{C}$ as an affine combination of exponentials. It then follows from Trefethen [122, Theorem 8.2] that, for $\rho>1$,

$$|f_m(x,t)-P_m(x,t)|\leq\frac{4M_{m,\rho}(x)\rho^{-k}}{\rho-1},\qquad t\in[-1,1],$$

where

$$M_{m,\rho}(x)=\max_{z\in\partial E_\rho}|f_m(x,z)|,\quad\text{and}\quad \partial E_\rho=\left\{\frac{z+z^{-1}}{2}\mid |z|=\rho\right\}.$$

For $z\in\partial E_\rho$, we have, letting $\widetilde{e}_n$ denote either e_n or $\partial_{x_i}e_n$, depending on f,

$$f_m(x,z)=f(x,a_m z+b_m)=\sum_{n=0}^{N}\mathrm{e}^{-\ell_n(a_m\Re z+b_m)}\langle p_0,e_n\rangle\widetilde{e}_n(x)\mathrm{e}^{-\mathrm{i}\ell_n a_m\Im z}=\langle\boldsymbol{r}(x,\Re z),\boldsymbol{\theta}(\Im z)\rangle,$$

where $(\boldsymbol{r}(x,y))_n = \mathrm{e}^{-\ell_n(a_m y + b_m)}\langle p_0, e_n\rangle \widetilde{e}_n(x)$ and $\boldsymbol{\theta}(y)_n = \mathrm{e}^{-\mathrm{i}\ell_n a_m y}$, so $|\boldsymbol{\theta}(y)| = N+1$. Thus, by Cauchy–Schwarz inequality,

$$
\begin{aligned}
M_{m,\rho}(x) &\leq \left((N+1) \max_{z \in \partial E_\rho} \sum_{n=0}^{N} \mathrm{e}^{-2\ell_n(a_m \Re z + b_m)} \langle p_0, e_n\rangle^2 \widetilde{e}_n(x)^2 \right)^{1/2} \\
&= \left((N+1) \sum_{n=0}^{N} \mathrm{e}^{-2\ell_n(a_m(\frac{-\rho-\rho^{-1}}{2}) + b_m)} \langle p_0, e_n\rangle^2 \widetilde{e}_n(x)^2 \right)^{1/2}.
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
\|M_{m,\rho}\|_{L^2}^2 &\leq (N+1) \sum_{n=0}^{N} \mathrm{e}^{-2\ell_n(a_m(\frac{-\rho-\rho^{-1}}{2}) + b_m)} \langle p_0, e_n\rangle^2 \|\widetilde{e}_n\|_{L^2}^2 \\
&\leq (N+1) \sum_{n=0}^{N} \mathrm{e}^{-2\ell_n(a_m(\frac{-\rho-\rho^{-1}}{2}) + b_m)} \langle p_0, e_n\rangle^2 \|e_n\|_{H^1}^2 \\
&\lesssim (N+1) \sum_{n=0}^{N} \mathrm{e}^{-2\ell_n(a_m(\frac{-\rho-\rho^{-1}}{2}) + b_m)} \ell_n \langle p_0, e_n\rangle^2,
\end{aligned}
$$

and if $a_m(\frac{-\rho-\rho^{-1}}{2}) + b_m = 0$, the right hand side is bounded by $(N+1)\|p_0\|_{\widetilde{H}^1}^2$. This is exactly the case when $\rho = 3$, in which case we have

$$
\|f_m(\,\cdot\,,t) - P_m(\,\cdot\,,t)\|_{L^2}^2 \lesssim N 3^{-2k},
$$

and setting $k = \lceil (\frac{\alpha}{D} + \frac{1}{2}) \log N \rceil$ yields an approximation φ_{f_m} of f_m with the correct error rate. Setting $\varphi_f^{(m)}(x,t) = \varphi_{f_m}(\frac{t-b_m}{a_m}, x)$ for $t \in [2^{m-2}\underline{T}, 2^{m+2}\underline{T}]$ then gives the desired network. We thus only need to construct the network φ_{p_i} as detailed above. To this end, note that the neural network

$$
\varphi_{p_i}^{(0)} :\ t \mapsto \begin{bmatrix} I_k & -I_k & 0 \\ 0 & 0 & I_{2^{\lceil \log_2 k \rceil} - k} \end{bmatrix} \sigma_{\begin{bmatrix} \boldsymbol{t}_i \\ -\boldsymbol{t}_i \\ -1 \end{bmatrix}} \begin{bmatrix} I_k \\ -I_k \\ 0 \end{bmatrix} t, \qquad \boldsymbol{t}_i = \begin{bmatrix} t_0 & \cdots & t_{i-1} & t_{i+1} & \cdots & t_k \end{bmatrix}^\top,
$$

maps t to the vector

$$
\begin{bmatrix} t - t_0 & \cdots & t - t_{i-1} & t - t_{i+1} & \cdots & t - t_k & 1 & \cdots & 1 \end{bmatrix}^\top \in [-2,2]^{2^{\lceil \log_2 k \rceil}},
$$

regardless of the signs of its entries. Without altering the size of $\varphi_{p_i}^{(0)}$, we can swap the entries of $\varphi_{p_i}^{(0)}(t)$ such that adjacent entries correspond to opposing t_j's, e.g., such that the first two entries are $t - t_0$ and $t - t_k$ rather than $t - t_0$ and $t - t_1$ and so on. This ensures that the products of adjacent entries stay uniformly bounded rather than some products growing and some shrinking, better bounding the size of the network. A slight modification of (the proof of) Lemma B.47 then yields a network $\overline{\varphi}_{m_3}^{\text{mult}}$ of the same asymptotic size as in Lemma B.47 such that $|\varphi_{m_3}^{\text{mult}}(x,y)-xy| \leq 2^{-m_3}$ for all $x, y \in [-2,2]$. Let $\varphi_{p_i}^{(j)}$ be a parallelisation of $2^{\lceil \log_2 k \rceil - j}$ copies of this network, and set

$$\varphi_{p_i} = \varphi_{p_i}^{(\lceil \log_2 k \rceil)} \circ \cdots \circ \varphi_{p_i}^{(1)} \circ \varphi_{p_i}^{(0)}.$$

To ensure that φ_{p_i} satisfies $|\varphi_{p_i} - p_i| \lesssim k2^{-m_3}$, fix $t \in [-1,1]$, and for notation, set $\xi_n^{(j)} = (\varphi_{p_i}^{(j)} \circ \cdots \circ \varphi_{p_i}^{(1)} \circ \varphi_{p_i}^{(0)}(t))_n$ and $y_l = (\varphi_{p_i}^{(0)}(t))_l$. We then claim that

$$\left|\xi_n^{(j)} - \prod_{l=(n-1)2^j+1}^{n2^j} y_l\right| \leq (2^j - 1)2^{-m_3}.$$

This is true by construction for $j = 1$, so assume this holds for some $j \geq 1$. We then have

$$\begin{aligned}
\left|\xi_n^{(j+1)} - \prod_{l=(n-1)2^{j+1}+1}^{n2^{j+1}} y_l\right| &= \left|\varphi_{m_3}^{\text{mult}}(\xi_{2n-1}^{(j)}, \xi_{2n}^{(j)}) - \prod_{l=(n-1)2^{j+1}+1}^{n2^{j+1}} y_l\right| \\
&\leq 2^{-m_3} + \left|\xi_{2n-1}^{(j)}\xi_{2n}^{(j)} - \left(\prod_{l=(2n-2)2^j+1}^{(2n-1)2^j} y_l\right)\left(\prod_{l=(2n-1)2^j+1}^{2n2^j} y_l\right)\right| \\
&\leq 2^{-m_3}\left(1 + \left(|\xi_{2n-1}^{(j)}| + \prod_{l=(2n-1)2^j+1}^{2n2^j} |y_l|\right)(2^j - 1)\right).
\end{aligned}$$

By our previous rearranging of entries in $\varphi_{p_i}^{(0)}(t)$, we have that $|\xi_{2n-1}^{(j)}| + \prod_{l=(2n-1)2^j+1}^{2n2^j} |y_l| \leq 2$ for all $j \geq 1$, and the claim follows. This then implies that $|\varphi_{p_i}(t) - p_i(t)| \lesssim k2^{-m_3}$ for all $t \in [-1,1]$ as desired.

We now shift from analysing the error of the network to its size instead. First, it is apparent from their construction that $\varphi_{p_i}^{(0)} \in \widetilde{\Phi}(1,k,k,1)$, while it also holds that

$$\varphi_{p_i}^{(j)} \in \widetilde{\Phi}(m_3, 2^{\lceil \log_2 k \rceil-(j-1)}, 2^{\lceil \log_2 k \rceil - j}m_3, 1).$$

Hence, since $\sum_{j=1}^{\lceil \log_2 k \rceil} 2^{\lceil \log_2 k \rceil - j} = \sum_{j=0}^{\lceil \log_2 k \rceil - 1} 2^j = 2^{\lceil \log_2 k \rceil} - 1$,

$$\begin{aligned}
\varphi_{p_i} &\in \widetilde{\Phi}\Big(\lceil \log_2 k \rceil m_3, k, (2^{\lceil \log_2 k \rceil} - 1)m_3 + k, 1\Big) \\
&= \widetilde{\Phi}(\log N \log\log N, \log N, (\log N)^2, 1).
\end{aligned}$$

Parallelising this with φ_{f_m,t_i}, we thus get a network in $\widetilde{\Phi}(\log N \log\log N, N, N\log N, N^{\frac{1}{D}} \vee \frac{1}{\sqrt{T}})$. Since this dominates the size of $\varphi_{m_2}^{\text{mult}}$ and since $|c_i| \leq \frac{2^{k-1}}{k} \leq \frac{N^{\frac{2\alpha+D}{2D}}}{\frac{2\alpha+D}{d}\log N}$, it follows that each term of φ_{f_m} is in $\widetilde{\Phi}(L_M, W_M, S_M, B_M)$, where

$$\begin{aligned} L_M &= \log N \log\log N, \\ W_M &= N \log N, \\ S_M &= N(\log N)^2, \quad \text{and} \\ B_M &= \frac{N^{\frac{2\alpha+D}{2D}}}{\log N} \vee \frac{1}{\sqrt{\underline{T}}}. \end{aligned}$$

Next, by construction of the π_m's, we have that $\pi_m \in \widetilde{\Phi}(1, 1, 1, \frac{1}{\underline{T}})$, where parallelising $\varphi_f^{(m)}$ with π_m does not change the asymptotic size of the network. Since the size of $\varphi_{m_1}^{\text{mult}}$ is also dominated by that of φ_{f_m}, it follows that each term of φ_f is included in $\widetilde{\Phi}(L_M, W_M, S_M, B_M)$ as well. Parallelising all of these and summing them yields

$$\begin{aligned} L(N) &= \log N \log\log N, \\ W(N) &= MN \log N, \\ S(N) &= MN(\log N)^2, \quad \text{and} \\ B(N) &= \frac{N^{\frac{2\alpha+D}{2D}}}{\log N} \vee \frac{1}{\underline{T}}. \end{aligned}$$

Finally, since $\alpha > \frac{D}{2}$, we have by similar calculations as those in the proof of Lemma B.39 that for some $\widetilde{C} < \infty$ depending only on D and p_0, it holds that $\sup_{x\in\Omega} |\partial_{x_i} h_N(x, t)| \leq \widetilde{C}\frac{1}{\sqrt{t}}$. Letting φ^{cap} be as in Lemma B.42 (with $m = \log_2(\underline{T}^{-1}) \asymp \log N$), it follows that also $\sup_{x\in\Omega} |\partial_{x_i} h_N(x, t)| \leq C\varphi^{\text{cap}}(t)$ for all $t \in [\underline{T}, \overline{T}]$, whence we get a no worse approximation by replacing $\varphi_{\partial_{x_i} h_N}$ with $\varphi_{\partial_{x_i} h_N} \wedge (C\varphi^{\text{cap}})$ and this network has the desired bound. Furthermore, since

$$\varphi^{\text{cap}} \in \widetilde{\Phi}(\log N \log\log N, \log N, \log N \log\log N, 1/\sqrt{\underline{T}}),$$

and this is dominated by the network size of $\varphi_{\partial_{x_i} h_N}$, taking this minimum does not alter the size of the network. This finishes the proof. ☐

Step 3: Putting things together With the essential preparations from Step 1 and 2, we can now finally prove Theorem B.37.

Proof of Theorem B.37. Let $\varphi_{h_N}, \varphi_{\partial_{x_i} h_N}, i \in [D]$ and $N \in \mathbb{N}$, be as in Lemma B.43. Parallelising the latter of these yields a network $\varphi_{\nabla_x h_N} \in \widetilde{\Phi}(L(N), W(N), S(N), B(N))$ approximating $\nabla_x h_N$. We then claim that φ_s defined as

$$\varphi_s(x, t) = \varphi_m^{\text{mult},D}(\varphi_m^{\text{rec}}(\varphi_{h_N}(x, t) \vee p_{\min}), \varphi_{\nabla_x h_N}(x, t)), \qquad m = \left\lceil \frac{\alpha}{D} \log_2 N \right\rceil,$$

has the desired properties. Indeed, for the size of the network, notice that $\varphi_{h_N} \vee p_{\min}$ is bounded above (by $\|p_0\|_\infty$) and below, and hence $\varphi_m^{\mathrm{rec}} \circ (\varphi_{h_N} \vee p_{\min})$ is bounded above by $2p_{\min}^{-1}$ for N large enough, while the entries of $\varphi_{\nabla_x h_N}$ are all bounded numerically by $\frac{C}{\sqrt{\underline{T}}}$ some $C < \infty$ since $\alpha > \frac{D}{2}$. Thus, $\varphi_m^{\mathrm{rec}} \in \widetilde{\Phi}(\log N \log\log N, \log N \log\log N, \log N \log\log N, 1)$ and $\varphi_m^{\mathrm{mult},d} \in \widetilde{\Phi}(\log N, \log N, \log N, \frac{1}{\sqrt{\underline{T}}})$, whereby φ_s has the same asymptotic size as $\varphi_{\nabla_x h_N}$. Note also that $|\varphi_s(x,t)| \le \frac{4Cp_{\min}^{-1}}{\sqrt{t}} \vee 1$ for all $x \in \Omega$, $t \in [\underline{T}, \overline{T}]$ and N large enough. So all that remains is to show that φ_s, as defined above, satisfies (B.58). By Proposition B.39 and the triangle inequality, this is equivalent to verifying

$$\int_{\underline{T}}^{\overline{T}} \int_\Omega \left| \varphi_s(x,t) - \frac{\nabla_x h_N(x,t)}{h_N(x,t) \vee p_{\min}} \right|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t \lesssim N^{-\frac{2\alpha}{D}} (\log N)^2 (\overline{T} + \log(\underline{T}^{-1})).$$

which follows if we can show that $\| \varphi_s(\,\cdot\,,t) - \frac{\nabla_x h_N(\cdot,t)}{h_N(\cdot,t)\vee p_{\min}} \|_{L^2}^2 \lesssim \varepsilon(t) N^{-\frac{2\alpha}{D}} (\log N)^2$ for each $t \in [\underline{T}, \overline{T}]$, where $\varepsilon(t)$ is as in Lemma B.43. For doing so, first note that

$$\begin{aligned} \left\| \varphi_s(\,\cdot\,,t) - \frac{\nabla_x h_N(\cdot,t)}{h_N(\cdot,t) \vee p_{\min}} \right\|_{L^2} &\le \| \varphi_s(\,\cdot\,,t) - \varphi_m^{\mathrm{rec}}(\varphi_{h_N}(\,\cdot\,,t) \vee p_{\min}) \varphi_{\nabla_x h_N}(\,\cdot\,,t) \|_{L^2} \\ &\quad + \left\| \left(\varphi_m^{\mathrm{rec}}(\varphi_{h_N}(\,\cdot\,,t) \vee p_{\min}) - \frac{1}{\varphi_{h_N}(\,\cdot\,,t) \vee p_{\min}} \right) |\varphi_{\nabla_x h_N}(\,\cdot\,,t)| \right\|_{L^2} \\ &\quad + \left\| \frac{\varphi_{\nabla_x h_N}(\cdot,t)}{\varphi_{h_N}(\,\cdot\,,t) \vee p_{\min}} - \frac{\nabla_x h_N(\cdot,t)}{h_N(\cdot,t) \vee p_{\min}} \right\|_{L^2}. \end{aligned}$$

The first term is simply evaluated as

$$\| \varphi_s(\,\cdot\,,t) - \varphi_m^{\mathrm{rec}}(\varphi_{h_N}(\,\cdot\,,t) \vee p_{\min}) \varphi_{\nabla_x h_N}(\,\cdot\,,t) \|_{L^2} \le 2^{-m} \Big(4\sqrt{D} p_{\min}^{-1} \Big\| p_0 - \frac{1}{\lambda_D(\Omega)} \Big\|_{H^\alpha} \lambda_D(\Omega) \Big) \lesssim N^{-\frac{\alpha}{D}},$$

while for the second term, we have

$$\begin{aligned} \left\| \left(\varphi_m^{\mathrm{rec}}(\varphi_{h_N}(\,\cdot\,,t) \vee p_{\min}) - \frac{1}{\varphi_{h_N}(\,\cdot\,,t) \vee p_{\min}} \right) |\varphi_{\nabla_x h_N}(\,\cdot\,,t)| \right\|_{L^2} &\le 2^{-m} \| \varphi_{\nabla_x h_N}(\,\cdot\,,t) \|_{L^2} \\ &\lesssim 2^{-m} \Big(\sqrt{D\varepsilon(t)} N^{-\frac{\alpha}{D}} \log N + \|p_0\|_{H^1} \Big) \\ &\lesssim \sqrt{\varepsilon(t)} N^{-\frac{\alpha}{D}}. \end{aligned}$$

For the final term, one obtains, similarly to the proof of Proposition B.39,

$$\begin{aligned} \left\| \frac{\varphi_{\nabla_x h_N}(\cdot,t)}{\varphi_{h_N}(\,\cdot\,,t) \vee p_{\min}} - \frac{\nabla_x h_N(\cdot,t)}{h_N(\cdot,t) \vee p_{\min}} \right\|_{L^2} &\le p_{\min}^{-1} \| \varphi_{\nabla_x h_N}(\,\cdot\,,t) - \nabla_x h_N(\cdot,t) \|_{L^2} \\ &\qquad + p_{\min}^{-2} \| \nabla_x h_N(\cdot,t)(h_N(\cdot,t) - \varphi_{h_N}(\,\cdot\,,t)) \|_{L^2} \\ &\lesssim p_{\min}^{-1} \sqrt{D\varepsilon(t)} N^{-\frac{\alpha}{D}} \log N + p_{\min}^{-2} \|p_0\|_{H^1} \sqrt{\varepsilon(t)} N^{-\frac{\alpha}{D}} \log N \\ &\lesssim \sqrt{\varepsilon(t)} N^{-\frac{\alpha}{D}} \log N. \end{aligned}$$

Finally, setting $N = n^{\frac{D}{2\alpha+D}}$ yields both the desired network size and error rate, which finishes the proof. $\square$

Proof of the main result

With the previous preparations, we can now prove our main result on the generative error, Theorem B.31. To this end, we use the general error decomposition (B.50) in combination with Lemma B.32 and Lemma B.33 to control the early stopping and ergodic error contributions, as well as with the approximation result Theorem B.37 that allows us to obtain an optimised upper bound on the empirical score loss via Theorem B.34.

Proof of Theorem B.31. Choose $\delta = n^{-2\alpha/(2\alpha+D)}$ and $N = n^{D/(2\alpha+D)}$. By the choices for $\underline{T}, \overline{T}$ and N, Theorem B.37 implies that there exists a family of neural networks $\mathcal{S}$ with the specified size constraints, such that for some $s \in \mathcal{S}$ we have

$$\int_{\underline{T}}^{\overline{T}} \int_\Omega \left|s(x,t) - \nabla_x \log p_t(x)\right|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t \lesssim n^{-\frac{2\alpha}{2\alpha+D}} (\log n)^3.$$

With these network size constraints, we get for $C(\mathcal{S}) := C$ and c from Lemma B.35, by using a straightforward modification of Oko, Akiyama, and Suzuki [89, Lemma C.2], that

$$\begin{aligned}\log \mathcal{N}(\mathcal{S}, \|\cdot\|_{\Omega\times[\underline{T},\overline{T}]}, \delta/(cC\overline{T})) &\lesssim S(n)L(n)\log\left(\delta^{-1}\overline{T}^2 L(n)\|W(n)\|_\infty B(n)\right)\\ &\lesssim n^{\frac{D}{2\alpha+D}}(\log n)^5 \log\log n.\end{aligned}$$

Lemma B.35 therefore implies that

$$\log \mathcal{N}(\mathcal{L}, \|\cdot\|_\Omega, \delta) \lesssim n^{\frac{D}{2\alpha+D}}(\log n)^5 \log\log n$$

as well. By Lemma B.36 and the choices of $\overline{T}, \underline{T}$, we can choose $C(\mathcal{L}) \lesssim \log n$ so that

$$\frac{C(\mathcal{L})}{n}\log \mathcal{N}(\mathcal{L}, \|\cdot\|_\Omega, \delta) \lesssim n^{\frac{-2\alpha}{2\alpha+D}}(\log n)^6 \log\log n.$$

Using Theorem B.37, it follows from the above and Theorem B.34 by our choice of N that

$$\begin{aligned}&\mathbb{E}\Big[\int_{\underline{T}}^{\overline{T}} \int_\Omega |\widehat{s}_n(x,t) - \nabla \log p_t(x)|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t\Big]\\ &\le 2\inf_{s\in\mathcal{S}} \int_{\underline{T}}^{\overline{T}} \int_\Omega |s(x,t) - \nabla \log p_t(x)|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t + 2\frac{C(\mathcal{L})}{n}\Big(\frac{37}{9}\log \mathcal{N}(\mathcal{L}, \|\cdot\|_\Omega, \delta) + 32\Big) + 3\delta\\ &\lesssim n^{-\frac{2\alpha}{2\alpha+D}}(\log n)^3 + n^{-\frac{2\alpha}{2\alpha+D}}(\log n)^6 \log\log n + n^{-\frac{2\alpha}{2\alpha+D}}\\ &\lesssim n^{-\frac{2\alpha}{2\alpha+D}}(\log n)^6 \log\log n.\end{aligned}$$

Thus,

$$
\begin{aligned}
\mathbb{E}\Big[\operatorname{TV}(\overline{p}^{s_0}_{\overline{T}-\underline{T}}, \overline{p}^{\widehat{s}_n}_{\overline{T}-\underline{T}})\Big] &\leq \sqrt{\frac{1}{2}\mathbb{E}\Big[\operatorname{KL}(\overline{p}^{s_0}_{\overline{T}-\underline{T}} \,\|\, \overline{p}^{\widehat{s}_n}_{\overline{T}-\underline{T}})\Big]} \\
&= \sqrt{\mathbb{E}\Big[\int_{\underline{T}}^{\overline{T}} \int_{\overline{\Omega}} f(x)|\widehat{s}_n(x,t) - \nabla \log p_t(x)|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t\Big]} \\
&\leq \sqrt{\|f\|_{\overline{\Omega}}\mathbb{E}\Big[\int_{\underline{T}}^{\overline{T}} \int_{\overline{\Omega}} |\widehat{s}_n(x,t) - \nabla \log p_t(x)|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t\Big]} \\
&\lesssim n^{-\frac{\alpha}{2\alpha+D}} (\log n)^3 (\log\log n)^{1/2},
\end{aligned}
$$

where we used Pinsker's inequality and Jensen's inequality together with concavity of $x \mapsto \sqrt{x}$ for the first line, while the second line follows from Theorem B.44 (Girsanov), together with independence of the driving Brownian motion $\overline{\mathsf{B}}$ and the initialisation $\overline{\mathsf{X}}_0 \sim p_T$ from the data $(\mathsf{X}_{0,i})_{i=1,\dots,n}$. The proof is concluded by applying the error decomposition (B.50) and using that, by Lemma B.32 and Lemma B.33, we have

$$
\operatorname{TV}(p_0, p_{\underline{T}}) + \operatorname{TV}(\mathbb{P}(\mathsf{X}_{\overline{T}} \in \cdot \mid \mathsf{X}_0 \sim p_0), \mathcal{U}(\overline{\Omega})) \lesssim \underline{T}^{\beta/2} + \mathrm{e}^{-\ell_1 \overline{T}} \lesssim n^{-\frac{\alpha}{2\alpha+D}},
$$

by our choices of $\underline{T}, \overline{T}$. ☐

B.4 Discussion

This paper presents a rigorous investigation of the non-standard class of denoising reflected diffusion models (DRDMs), focusing on statistical convergence guarantees and approximation within constrained settings, as it is relevant for scenarios involving bounded state spaces. This is a first step in extending the statistical analysis of standard denoising diffusion models to the generalised class of *denoising Markov models* proposed in [12].

A key result of our analysis is the derivation of convergence rates in total variation that match the minimax lower bound on Sobolev classes up to a logarithmic factor. More precisely, from Yang and Barron [133, Theorem 4], see also Oko, Akiyama, and Suzuki [89, Proposition 5.2] for the verification of their assumptions for α-smooth Sobolev functions on $[0,1]^D$, we have that

$$
\inf_{\widehat{p}_n} \sup_{p_0 \in H^\alpha(\Omega)} \mathbb{E}_{p_0}[\operatorname{TV}(p_0, \widehat{p}_n)] \asymp n^{-\frac{\alpha}{2\alpha+D}},
$$

where the infimum ranges over all estimators $\widehat{p}_n$ based on n i.i.d. data points having density p_0 under $\mathbb{P}_{p_0}$. Our upper bound stated in (B.48) therefore establishes that DRDMs attain the minimax optimal rate of convergence up to log-factors for specific Sobolev densities. Note that our logarithmic loss is comparatively small relative to that of unconstrained DDMs in [89], where it is of order $(\log n)^8$.

However, convergence rates (even optimal ones) expressed in terms of the ambient dimension d fall short of capturing the empirical success of DDMs. This gap is related to the *manifold hypothesis*, a prominent idea that real-world high-dimensional data often reside on lower-dimensional manifolds, to which well-trained generative models are believed to adapt. Developing a theoretical underpinning for this hypothesis has therefore been one of the central goals in statistical theory for generative models. In the pioneering paper [89], the authors also take a first step towards investigating statistical convergence guarantees of DDMs for data distributed on such lower-dimensional structures by extending their analysis to initial distributions supported on a lower-dimensional hyperplane, where they obtain the almost minimax optimal rate $n^{-\frac{\alpha+1}{2\alpha+d}+\varepsilon}$ in the Wasserstein-1 distance in terms of the sample size n and the subspace dimension d. Related statistical results under linear subspace assumptions are given in [22]. In the recent work [118], Tang and Yang significantly extend this result by establishing (up to log factors) the minimax convergence rate $C(D)n^{-\frac{\alpha+1}{2\alpha+d}}$ in Wasserstein-1 distance for distributions p_0 such that

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

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

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

Note that these three conditions mirror the three assumptions from [89] mentioned in the introductory Section B.1 in the manifold setting. The multiplicative factor $C(D)$ in [118]'s convergence rate is of order $D^{\alpha+D/2}$ and thus potentially very large for high ambient dimension D. Most recently, [6] show that this multiplicative factor can be significantly reduced to the order $\sqrt{D}$ by appropriately choosing the neural network class for score approximation. Extending the DRDM framework to support data on submanifolds (and thus improving the convergence rate) presents additional mathematical challenges. For example, enforcing a reflecting boundary for data supported on lower-dimensional submanifolds would require non-trivial modifications to the spectral score representation and a revised analysis of the associated Sobolev bounds. Such adjustments are beyond the scope of this study, yet our current work may provide a foundational approach for future research in the reflected diffusion context.

The assumptions on p_0 in our model (see ($\mathcal{H}0$)) play a key role in controlling the approximation error in our DRDM framework under bounded domain constraints. Such assumptions, although somewhat limiting in a practical context, are comparable to those made in [89] for the total variation convergence analysis of unconstrained models while maintaining compatibility with the spectral methods that we employ for our statistical analysis. In this context, our more stringent smoothness assumption $\alpha > D/2$ implies Hölder continuity of the data density p_0, but allows us to avoid some technical difficulties arising due to the less explicit analytical nature of DRDMs compared to standard DDMs.

How to remove the lower bound assumption $p_0|_{\operatorname{supp} p_0} \geq p_{\min}$ that is present in all the works discussed above is a highly relevant and conceptually challenging question. Recent works by [112, 135], and [132] prove minimax optimal rates for particular unconstrained diffusion models without lower bound assumptions for sub-Gaussian densities p_0 and push-forward distributions on the ambient space $\mathbb{R}^D$ of the form $g_\# \mathcal{U}[0,1]^d$ for Hölder-continuous g and $d \leq D$, respectively. [135] use a more classical kernel estimation and truncation strategy instead of neural network approximations. [112] exploit deeper results on the space-time regularity of the score function of an Ornstein–Uhlenbeck process for direct approximation with tanh activation function, thereby avoiding the need to approximate p_t and ∇p_t separately. Finally [132] exploit the structure of the score induced by their data assumption $p_0 \sim g_\# \mathcal{U}[0,1]^d$ and the Gaussian Ornstein–Uhlenbeck forward transition densities to construct their ReLU neural network based approximation class in a very specific way. These approaches do not translate directly to our non-Gaussian reflected setting and we leave the statistical study of reflected diffusion models without lower bound assumptions to future work.

In general, our analytical approach in DRDMs differs significantly from that in unconstrained diffusion models, as the bounded domain prevents the explicit Gaussian transition densities commonly used for error control. The semi-explicit nature of these densities in the reflected setting means that, rather than relying on straightforward Gaussian approximations, we implement general spectral decompositions and Sobolev-based bounds informed by the Sobolev smoothness of p_0. The score approximation here presents additional technical challenges, which we address through an innovative polynomial-time interpolation procedure that proves crucial to achieving feasible convergence rates. This technique introduces new and effective methods for controlling error contributions in generative models with bounded state spaces and, as outlined in the introduction, may also provide a versatile tool for statistical analysis of generalised denoising Markov models [12, 99] beyond the scope of this paper.

Finally, it should be noted that our analysis does not address sampling issues, in particular those arising from simulating the backward reflected process with an estimated drift. While this aspect is important in practical implementations, our current focus on theoretical convergence guarantees serves to isolate and address the approximation errors inherent to the reflection-based model setup itself. Future studies could incorporate error analysis for sampling methodologies specific to reflected processes to extend the results presented here.

B.A Girsanov's theorem for reflected diffusions

The following result is a version of Girsanov's theorem for reflected diffusions, which is a correction of Theorem 7.1 and Theorem A.6 in [80], where the influence of the diffusion coefficient on the KL divergence has been overlooked.

Theorem B.44: Let $(\mathsf{X}_t)_{t\in[0,T]}$ and $(\widetilde{\mathsf{X}}_t)_{t\in[0,T]}$ be solutions of the normally reflected SDEs

$$\begin{aligned} \mathrm{d}\mathsf{X}_t &= b(t,\mathsf{X}_t)\,\mathrm{d}t + \sigma(t,\mathsf{X}_t)\,\mathrm{d}\mathsf{B}_t + \nu(\mathsf{X}_t)\,\mathrm{d}\mathsf{L}_t, \\ \mathrm{d}\widetilde{\mathsf{X}}_t &= \widetilde{b}(t,\widetilde{\mathsf{X}}_t)\,\mathrm{d}t + \sigma(t,\widetilde{\mathsf{X}}_t)\,\mathrm{d}\widetilde{\mathsf{B}}_t + \nu(\widetilde{\mathsf{X}}_t)\,\mathrm{d}\widetilde{\mathsf{L}}_t, \quad \widetilde{\mathsf{X}}_0 \overset{d}{=} \mathsf{X}_0, \end{aligned} \tag{B.65}$$

where $b, \widetilde{b}$ are bounded on $[0,T]\times\overline{\Omega}$ and $\sigma(t,\cdot)$ is bounded, Lipschitz and uniformly elliptic in the sense $a(t,\cdot) := \sigma(t,\cdot)\sigma(t,\cdot)^\top \succeq \underline{\lambda}\mathbb{I}$ for some $\underline{\lambda} > 0$ and all $t \in [0,T]$. Denote by $\mathbb{P}_{\mathsf{X}^T}$ and $\mathbb{P}_{\widetilde{\mathsf{X}}^T}$ their respective path measures on $C([0,T],\overline{\Omega})$. Then, for $L_t = \int_0^t \nu(\mathsf{X}_s)\,\mathrm{d}\mathsf{L}_s$,

$$\begin{aligned}
&\log \frac{\mathrm{d}\mathbb{P}_{\widetilde{\mathsf{X}}^{\mathrm{T}}}}{\mathrm{d}\mathbb{P}_{\mathsf{X}^{\mathrm{T}}}}(\mathsf{X}^T) \\
&= \int_0^T (\widetilde{b}(t,\mathsf{X}_t) - b(t,\mathsf{X}_t))^\top a^{-1}(t,\mathsf{X}_t)\,\mathrm{d}(\mathsf{X}_t - L_t) \\
&\qquad - \frac{1}{2}\int_0^T (\widetilde{b}(t,\mathsf{X}_t) - b(t,\mathsf{X}_t))^\top a^{-1}(t,\mathsf{X}_t)(\widetilde{b}(t,\mathsf{X}_t) + b(t,\mathsf{X}_t))\,\mathrm{d}t,
\end{aligned}$$

a.s., and

$$\mathrm{KL}(\mathbb{P}_{\mathsf{X}^T}\,\|\,\mathbb{P}_{\widetilde{\mathsf{X}}^T}) = \frac{1}{2}\mathbb{E}\Big[\int_0^T \|\sigma^{-1}(t,\mathsf{X}_t)(\widetilde{b}(t,\mathsf{X}_t) - b(t,\mathsf{X}_t))\|^2\,\mathrm{d}t\Big].$$

Proof. Let

$$\begin{aligned}
\mathsf{Z}_T := \exp\Big(&\int_0^T (\widetilde{b}(t,\mathsf{X}_t) - b(t,\mathsf{X}_t))^\top(\sigma^{-1}(t,\mathsf{X}_t))^\top\,\mathrm{d}\mathsf{B}_t \\
&- \frac{1}{2}\int_0^T \|\sigma^{-1}(t,\mathsf{X}_t)(\widetilde{b}(t,\mathsf{X}_t) - b(t,\mathsf{X}_t))\|^2\,\mathrm{d}t\Big),
\end{aligned}$$

and define $\mathrm{d}\mathbb{Q}_{\mathbb{T}} := \mathsf{Z}_T\,\mathrm{d}\mathbb{P}$. Since $\overline{\beta} := \sup_{t\in[0,T],x\in\overline{\Omega}}\|\beta(t,x)\| < \infty$ for $\beta \in \{b, \widetilde{b}\}$, we have

$$\sup_{t\in[0,T],x\in\overline{\Omega}} \|\sigma^{-1}(t,x)\beta(t,x)\| \leq \frac{\overline{\beta}}{\underline{\lambda}} < \infty.$$

Thus, Novikov's condition is satisfied, making $\mathbb{Q}_T$ a probability measure equivalent to $\mathbb{P}$, and Girsanov's theorem implies that

$$\widetilde{\mathsf{B}}_t := \mathsf{B}_t - \int_0^t \sigma^{-1}(s,\mathsf{X}_s)(\widetilde{b}(s,\mathsf{X}_s) - b(s,\mathsf{X}_s))\,\mathrm{d}s, \quad t \in [0,T],$$

is a $\mathbb{Q}_T$-Brownian motion, see Karatzas and Shreve [63, Chapter 3, Theorem 5.1, Corollary 5.13]. Thus,

$$\mathsf{X}_t = \mathsf{X}_0 + \int_0^t \widetilde{b}(s,\mathsf{X}_s)\,\mathrm{d}s + \int_0^t \sigma(s,\mathsf{X}_s)\,\mathrm{d}\widetilde{\mathsf{B}}_s + \int_0^t \nu(\mathsf{X}_s)\,\mathrm{d}\mathsf{L}_s, \quad t \in [0,T],$$

where $\mathsf{L}_t = \int_0^t \mathbf{1}_{\{\mathsf{X}_s\in\partial\Omega\}}\,\mathrm{d}\mathsf{L}_s$, for any $t \in [0,T]$, $\mathbb{P}$-a.s. and hence also $\mathbb{Q}_T$-a.s. Since $\mathsf{X}_0 \overset{d}{=} \widetilde{\mathsf{X}}_0$, it follows that X^T is a weak solution to the reflected SDE (B.65) and analogously to Karatzas and Shreve [63, Chapter 5, Proposition 3.10], it holds that under the given assumptions on $\widetilde{b}, \sigma$, weak solutions to (B.65) are unique in law. Thus, X^T under $\mathbb{Q}_T$ has the same law as $\widetilde{\mathsf{X}}^T$ under $\mathbb{P}$, such that, upon noting that

$$\begin{aligned}\log \mathsf{Z}_T = \int_0^T &(\widetilde{b}(t,\mathsf{X}_t) - b(t,\mathsf{X}_t))^\top a^{-1}(t,\mathsf{X}_t)\,\mathrm{d}(\mathsf{X}_t - L_t) \\ &- \frac{1}{2}\int_0^T (\widetilde{b}(t,\mathsf{X}_t) - b(t,\mathsf{X}_t))^\top a^{-1}(t,\mathsf{X}_t)(\widetilde{b}(t,\mathsf{X}_t) + b(t,\mathsf{X}_t))\,\mathrm{d}t,\end{aligned}$$

the first assertion follows. This yields immediately

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\mathsf{X}^T} \,\|\, \mathbb{P}_{\widetilde{\mathsf{X}}^T}) = -\mathbb{E}\Big[\log \frac{\mathrm{d}\mathbb{P}_{\widetilde{\mathsf{X}}^\mathrm{T}}}{\mathrm{d}\mathbb{P}_{\mathsf{X}^\mathrm{T}}}(\mathsf{X}^T)\Big] &= -\mathbb{E}[\log \mathsf{Z}_T] \\ &= \frac{1}{2}\mathbb{E}\Big[\int_0^T \|\sigma^{-1}(t,\mathsf{X}_t)(\widetilde{b}(t,\mathsf{X}_t) - b(t,\mathsf{X}_t))\|^2\,\mathrm{d}t\Big].\end{aligned}$$

□

B.B Bernstein condition for the denoising score matching excess loss and generalisation error

The goal of this section is to prove (B.53) and thereby justify the generalisation error decomposition from Theorem B.34 in a generalised Markovian framework that in particular includes our reflected forward model, as well as the unconstrained Ornstein–Uhlenbeck models used in other works.

Let $\mathsf{Y}_1, \dots, \mathsf{Y}_n \overset{\text{i.i.d}}{\sim} \mu$ be a collection of random variables with state space $\mathcal{X} \subset \mathbb{R}^d$. Let also $(\mathsf{X}_t)_{t \geq 0}$ be a Markov process with state space $\mathcal{X}$ and transition densities $(q_t)_{t\geq 0}$ w.r.t. some reference measure ν, that is, $\mathbb{P}^x(\mathsf{X}_t \in \mathrm{d}y) = q_t(x,y)\,\nu(\mathrm{d}y)$, where $\mathbb{P}^x = \mathbb{P}(\cdot \mid \mathsf{X}_0 = x)$. Denote by p_t the density of X_t started in μ at time $t > 0$, i.e., $p_t(y)\nu(\mathrm{d}y) = \mathbb{P}^\mu(\mathsf{X}_t \in \mathrm{d}y) = \int_{\mathcal{X}} q_t(x,y)\,\mu(\mathrm{d}x)\,\nu(\mathrm{d}y)$. We assume that the *score* s_0 of this Markov process given by $s_0(x,t) = \nabla \log p_t(x)$ as well as $\nabla \log q_t(x,y)$ are well defined for all $t > 0, x, y \in \mathcal{X}$, and define for some measurable candidate function s the *denoising score matching loss* over an interval $[\underline{T}, \overline{T}] \subset (0,\infty)$ by

$$\begin{aligned}L_s(x) &= \int_{\underline{T}}^{\overline{T}} \int_{\mathcal{X}} |s(y,t) - \nabla \log q_t(x,y)|^2 q_t(x,y)\,\nu(\mathrm{d}y)\,\mathrm{d}t \\ &= \mathbb{E}^x\Big[\int_{\underline{T}}^{\overline{T}} |s(\mathsf{X}_t,t) - \nabla \log q_t(x,\mathsf{X}_t)|^2\,\mathrm{d}t\Big],\end{aligned}$$

which gives

$$\mathbb{E}^\mu[L_s(\mathsf{X}_0)] = \int_{\underline{T}}^{\overline{T}} \iint_{\mathcal{X}\times\mathcal{X}} |s(y,t) - \nabla \log q_t(x,y)\|^2 q_t(x,y)\,\nu(\mathrm{d}y)\,\mu(\mathrm{d}x)\,\mathrm{d}t.$$

Then, given sufficient integrability properties that allow interchanging the order of differentiation and integration, a few simple lines of calculation establish the equivalence between denoising and *explicit* score matching [129], which is the first equality in

$$\begin{aligned}\mathbb{E}^{\mu}[L_s(\mathsf{X}_0)] &= \int_{\underline{T}}^{\overline{T}} \int_{\mathcal{X}} |s(y,t) - \nabla \log p_t(y)|^2 p_t(y)\, \nu(\,\mathrm{d}y)\, \mathrm{d}t + C_{\underline{T},\overline{T}} \\ &= \int_{\underline{T}}^{\overline{T}} \int_{\mathcal{X}} |s(y,t) - s_0(y,t)|^2 p_t(y)\, \nu(\,\mathrm{d}y)\, \mathrm{d}t + C_{\underline{T},\overline{T}},\end{aligned}$$

where $C_{\underline{T},\overline{T}} = \mathbb{E}^{\mu}[L_{s_0}(\mathsf{X}_0)]$ is independent of $\mathcal{S}$. In particular, we obtain

$$\int_{\underline{T}}^{\overline{T}} \int_{\mathcal{X}} |s(y,t) - \nabla \log p_t(y)|^2 p_t(y)\, \nu(\,\mathrm{d}y)\, \mathrm{d}t = \mathbb{E}^{\mu}[L_s(\mathsf{X}_0)] - \mathbb{E}^{\mu}[L_{s_0}(\mathsf{X}_0)]. \tag{B.66}$$

As discussed before, a natural and tractable choice for the score s_0 given the data $Y_1, \ldots, Y_n$ is therefore given by the empirical risk minimiser

$$\hat{s} \in \operatorname*{argmin}_{s \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^{n} L_s(Y_i),$$

where $\mathcal{S}$ is some appropriate class of candidate functions. The following result shows that the excess risk $L_s - L_{s_0}$ satisfies a Bernstein condition w.r.t. μ, which is essential for the denoising score loss to work well as an empirical risk minimisation objective.

Lemma B.45: Suppose that $\sup_{s \in \mathcal{S} \cup \{s_0\}} \|L_s\|_{\mathcal{X}} \leq C(\mathcal{L}) < \infty$. Then,

$$\mathbb{E}^{\mu}\left[(L_s(\mathsf{X}_0) - L_{s_0}(\mathsf{X}_0))^2\right] \leq 4C(\mathcal{L})\mathbb{E}^{\mu}[L_s(\mathsf{X}_0) - L_{s_0}(\mathsf{X}_0)].$$

Proof. Using the elementary equality $|a|^2 - |b|^2 = \langle a+b, a-b \rangle$ for vectors a, b, and the Cauchy–

Schwarz inequality several times, we find for any $x \in \mathcal{X}$,

$$\begin{aligned}
&|L_s(x) - L_{s_0}(x)| \\
&\leq \int_{\underline{T}}^{\overline{T}} \int_{\mathcal{X}} |\langle s(y,t) - s_0(y,t), s(y,t) + s_0(y,t) - 2\nabla \log q_t(x,y)\rangle| q_t(x,y)\, \nu(\mathrm{d}y)\, \mathrm{d}t \\
&\leq \int_{\underline{T}}^{\overline{T}} \int_{\mathcal{X}} |s(y,t) - s_0(y,t)||s(y,t) + s_0(y,t) - 2\nabla \log q_t(x,y)| q_t(x,y)\, \nu(\mathrm{d}y)\, \mathrm{d}t \\
&\leq \int_{\underline{T}}^{\overline{T}} \Big(\int_{\mathcal{X}} |s(y,t) - s_0(y,t)|^2 q_t(x,y)\, \nu(\mathrm{d}y) \Big)^{1/2} \\
&\qquad \times \Big(\int_{\mathcal{X}} |s(y,t) + s_0(y,t) - 2\nabla \log q_t(x,y)|^2 q_t(x,y)\, \nu(\mathrm{d}y) \Big)^{1/2} \mathrm{d}t \\
&\leq \Big(\int_{\underline{T}}^{\overline{T}} \int_{\mathcal{X}} |s(y,t) - s_0(y,t)|^2 q_t(x,y)\, \nu(\mathrm{d}y)\, \mathrm{d}t \Big)^{1/2} \\
&\qquad \times \Big(\int_{\underline{T}}^{\overline{T}} \int_{\mathcal{X}} |s(y,t) + s_0(y,t) - 2\nabla \log q_t(x,y)|^2 q_t(x,y)\, \nu(\mathrm{d}y)\, \mathrm{d}t \Big)^{1/2} \\
&\leq \sqrt{2(L_s(x) + L_{s_0}(x))} \Big(\int_{\underline{T}}^{\overline{T}} \int_{\mathcal{X}} |s(y,t) - s_0(y,t)|^2 q_t(x,y)\, \nu(\mathrm{d}y)\, \mathrm{d}t \Big)^{1/2} \\
&\leq 2\sqrt{C(\mathcal{L})} \Big(\int_{\underline{T}}^{\overline{T}} \int_{\mathcal{X}} |s(y,t) - s_0(y,t)|^2 q_t(x,y)\, \nu(\mathrm{d}y)\, \mathrm{d}t \Big)^{1/2}
\end{aligned}$$

Thus, using (B.66),

$$\begin{aligned}
\mathbb{E}^{\mu}\big[(L_s(\mathsf{X}_0) - L_{s_0}(\mathsf{X}_0))^2\big] &\leq 4C(\mathcal{L}) \int_{\mathcal{X}} \int_{\underline{T}}^{\overline{T}} \int_{\mathcal{X}} |s(y,t) - s_0(y,t)|^2 q_t(x,y)\, \nu(\mathrm{d}y)\, \mathrm{d}t\, \mu(\mathrm{d}x) \\
&= 4C(\mathcal{L}) \int_{\underline{T}}^{\overline{T}} \int_{\mathcal{X}} |s(y,t) - s_0(y,t)|^2 p_t(y)\, \nu(\mathrm{d}y)\, \mathrm{d}t \\
&= 4C(\mathcal{L}) \mathbb{E}^{\mu}[L_s(\mathsf{X}_0) - L_{s_0}(\mathsf{X}_0)].
\end{aligned}$$

□

This result renders the reasoning in Oko, Akiyama, and Suzuki [89, Theorem C.4] valid up to a multiplicative factor that only results in a minor change of constants in their generalisation error upper bound. For completeness we state a corrected version of Oko, Akiyama, and Suzuki [89, Theorem C.4] in the general context of this section with corrected constants.

Theorem B.46: Suppose that $\sup_{s\in\mathcal{S}}\|L_s\|_{\mathcal{X}} \le C(\mathcal{L})$, where $C(\mathcal{L}) < \infty$. Then, for any $\delta > 0$ such that $\mathcal{N}(\mathcal{L}, \|\cdot\|_{\mathcal{X}}, \delta) \ge 3$, it holds that

$$\begin{aligned}
&\mathbb{E}\Big[\int_{\underline{T}}^{\overline{T}}\int_{\mathcal{X}}|\widehat{s}(x,t) - \nabla\log p_t(x)|^2 p_t(x)\, \nu(\,\mathrm{d}x)\,\mathrm{d}t\Big] \\
&\quad \le 2\inf_{s\in\mathcal{S}}\int_{\underline{T}}^{\overline{T}}\int_{\mathcal{X}}|s(x,t) - \nabla\log p_t(x)|^2 p_t(x)\, \nu(\,\mathrm{d}x)\,\mathrm{d}t + \frac{2C(\mathcal{L})}{n}\Big(\frac{145}{9}\log\mathcal{N}(\mathcal{L}, \|\cdot\|_{\mathcal{X}}, \delta) + 160\Big) \\
&\qquad + 5\delta.
\end{aligned}$$

B.C Technical results on neural network approximations

We first consider neural network approximations of products.

Lemma B.47: For any $C \ge 1$ and $m, d \in \mathbb{N}$, there exist neural networks $\varphi_m^{\mathrm{mult}} \in \Phi(L, W, S, B)$ and $\varphi_m^{\mathrm{mult},D} \in \Phi(L, W_D, D\cdot S, B)$ satisfying

$$|\varphi_m^{\mathrm{mult}}(x,y) - xy| \le C2^{-m}, \quad x \in [0,1], y \in [-C, C].$$

and

$$|\varphi_m^{\mathrm{mult},D}(x,y) - xy| \le \sqrt{D}C2^{-m}, \quad x \in [0,1], y \in [-C, C]^D.$$

Furthermore, $\varphi^{\mathrm{mult}}(0,y) = \varphi^{\mathrm{mult}}(x,0) = 0$. The sizes of the networks are evaluated as $L = m + 8$, $S = 58 + 16m$, $B = C$ and

$$W = \begin{bmatrix} 2 & 3 & 3 & \underbrace{12 \ \cdots \ 12}_{m+2 \text{ times}} & 2 & 2 & 1\end{bmatrix}^\top, \quad (W_D)_i = \begin{cases} D+1, & \text{if } i = 1 \\ D\cdot W_i & \text{otherwise.}\end{cases}$$

In particular, we have $L, S \lesssim m$, $\|W\|_\infty \lesssim 1$, $\|W_D\|_\infty \lesssim D$ and $B \lesssim C$.

Proof. We proceed as in the proof of Oko, Akiyama, and Suzuki [89, Lemma F.6], but adapted to this specific setting. Thus, let $m \in \mathbb{N}$ and $C \ge 0$ be fixed. Then, by Schmidt-Hieber [101, Lemma A.2], there exists a neural network $\overline{\varphi}_m^{\mathrm{mult}} \in \Phi(m+4, W_0, 24+16m, 1)$, where

$$W_0 = \begin{bmatrix} 2 & \underbrace{6 \ \cdots \ 6}_{m+2 \text{ times}} & 1\end{bmatrix}^\top,$$

satisfying $|\overline{\varphi}_m^{\mathrm{mult}}(x, y) - xy| \le 2^{-m}$ and $\overline{\varphi}_m^{\mathrm{mult}}(x, 0) = \overline{\varphi}_m^{\mathrm{mult}}(0, y) = 0$ for all $x, y \in [0, 1]$. Thus, for $y \in [-1, 1]$, we have $|\mathrm{sgn}(y)\overline{\varphi}_m^{\mathrm{mult}}(x, |y|) - xy| \le 2^{-m}$. Note then that

$$\begin{aligned}\mathrm{sgn}(y)\overline{\varphi}_m^{\mathrm{mult}}(x, |y|) &= \sigma_R\big(\overline{\varphi}_m^{\mathrm{mult}}(x, \sigma_R(y))\big) - \sigma_R\big(\overline{\varphi}_m^{\mathrm{mult}}(x, \sigma_R(-y))\big) \\ &= \overline{\varphi}_m^{\mathrm{mult},3} \circ \overline{\varphi}_m^{\mathrm{mult},2} \circ \overline{\varphi}_m^{\mathrm{mult},1}(x, y),\end{aligned}$$

where

$$\begin{aligned}\overline{\varphi}_m^{\mathrm{mult},1} &= I_3\sigma \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} \in \Phi(1, \begin{bmatrix} 2 & 3 & 3 \end{bmatrix}^\top, 6, 1), \\ \overline{\varphi}_m^{\mathrm{mult},3} &= \begin{bmatrix} 1 & -1 \end{bmatrix} \sigma I_2 \in \Phi(1, \begin{bmatrix} 2 & 2 & 1 \end{bmatrix}^\top, 4, 1),\end{aligned}$$

and $\overline{\varphi}_m^{\mathrm{mult},2}(x, y, z) = \begin{bmatrix} \overline{\varphi}_m^{\mathrm{mult}}(x, y) & \overline{\varphi}_m^{\mathrm{mult}}(x, z) \end{bmatrix}^\top$. Thus, $\overline{\varphi}_m^{\mathrm{mult},2} \in \Phi(m + 4, W_1, 48 + 32m, 1)$ where W_1 is the parallelization of W_0 with itself, and hence $\mathrm{sgn}(y)\overline{\varphi}_m^{\mathrm{mult}}(x, |y|) := \widetilde{\varphi}_m^{\mathrm{mult}}(x, y) \in \Phi(m + 8, W, 58 + 16m, 1)$, where W is as specified in the statement of the lemma. Finally, if $y \in [-C, C]$, we have similarly $|C\widetilde{\varphi}_m^{\mathrm{mult}}(x, y/C) - xy| \le C2^{-m}$, and since these are simply linear transformations (i.e., neural networks of depth 0), it follows that $C\widetilde{\varphi}_m^{\mathrm{mult}}(x, y/C) := \varphi_m^{\mathrm{mult}}(x, y) \in \Phi(m+8, W, 58+16m, C)$. Parallelization of $\varphi_m^{\mathrm{mult}}$ a total of D times yields a network $\varphi_m^{\mathrm{mult},D} \in \Phi(m + 8, W_D, D(58 + 16m), C)$ with W_D as specified above, which satisfies

$$\max_{i\in\{1,\dots,D\}} |(\varphi_m^{\mathrm{mult},D}(x, y))_i - xy_i| \le C2^{-m}, \quad x \in [0, 1], y \in [-C, C]^D,$$

implying the desired result since $|y| \le \sqrt{D}\|y\|_\infty$ for $y \in \mathbb{R}^d$. ☐

Next, we consider the reciprocals. Note that by not relying on the comparatively slow converging Taylor series of the reciprocal, we achieve a substantially smaller network size than Oko, Akiyama, and Suzuki [89, Lemma F.7], who achieve a network size of $\widetilde{\Phi}(m^2, m^3, m^4, 2^{2m})$ (assuming, as they do, that $m = \underline{k} = \overline{k}$).

Lemma B.48: For any $\underline{k}, \overline{k}, m \in \mathbb{N}$ with $m > \overline{k}$, there exists a neural network $\varphi_m^{\mathrm{rec}} \in \Phi(L, W, S, B)$ satisfying

$$|\varphi_m^{\mathrm{rec}}(x) - \frac{1}{x}| \le 2^{-m}, \quad x \in [2^{-\underline{k}}, 2^{\overline{k}}].$$

The size of φ_m^{rec} is evaluated as

$$\begin{aligned}L &= \big(4(\underline{k} + \overline{k}) + 2m + 17\big)\lceil\log_2(\underline{k} + m + 2)\rceil + 2\lceil\log_2(\underline{k} + \overline{k})\rceil + 1, \\ S &= \big(260 + 68(\underline{k} + \overline{k} + m)\big)\lceil\log_2(\underline{k} + m + 2)\rceil + 8(\underline{k} + \overline{k}),\end{aligned}$$

$B = 2^{2(\underline{k}+\overline{k})}$ and W is the concatenation of $\lceil \log_2(\underline{k}+m+2)\rceil$ copies of W_{iter} and W_{init}, where

$$W_{\text{iter}} = \Big[2 \quad \underbrace{7 \quad \cdots \quad 7}_{2(\underline{k}+\overline{k})+m+3 \text{ times}} \quad 2 \quad \underbrace{7 \quad \cdots \quad 7}_{2(\underline{k}+\overline{k})+m+3 \text{ times}} \quad 2\Big]^\top,$$

and

$$W_{\text{init}} = \Big[1 \quad \underline{k}+\overline{k} \quad 2^{\lceil \log_2(\underline{k}+\overline{k})\rceil} \quad 2^{\lceil \log_2(\underline{k}+\overline{k})\rceil-1} \quad 2^{\lceil \log_2(\underline{k}+\overline{k})\rceil-1} \quad \cdots \quad 2 \quad 2 \quad 1\Big]^\top.$$

In particular, we have $L, S \lesssim (k+m)(\log(k+m))$, $\|W\|_\infty \lesssim k$, where $k = \underline{k}+\overline{k}$.

Proof. Let $\underline{k}, \overline{k}, m \in \mathbb{N}$ with $m > \overline{k}$ be fixed. We approximate the reciprocal by Newton–Raphson iterations, adapted to neural networks. In a usual Newton–Raphson scheme approximating x^{-1} for some $x > 0$, we would take x_0 to be some initial approximation and set $x_n = x_{n-1}(2 - xx_{n-1})$ for $n \in \mathbb{N}$. This, however, involves two multiplications by non-constants in each iteration, and as such is not directly accessible by neural networks. To overcome this, suppose that we have access to a neural network φ_m^{iter} such that $|\varphi_m^{\text{iter}}(x,y) - x(2-xy)| \le 2^{-(m+1)}$ as well as a neural network φ^{init} such that $|\varphi^{\text{init}} - x^{-1}| \le \frac{1}{2}x^{-1}$. Then, setting $\widetilde{x}_0 = \varphi^{\text{init}}(x)$ and $\widetilde{x}_n = \varphi_m^{\text{iter}}(\widetilde{x}_{n-1}, x)$ for $n \in \mathbb{N}$, we claim that $|\widetilde{x}_{n_0} - x^{-1}| \le 2^{-m}$ where $n_0 = \lceil \log_2(\underline{k}+m+2)\rceil$. To show this, we first find by the recursive definition of $\widetilde{x}_n$ that

$$|\widetilde{x}_{n+1} - x^{-1}| = |\varphi_m^{\text{iter}}(\widetilde{x}_n, x) - x^{-1}| \le |\widetilde{x}_n(2 - x\widetilde{x}_n) - x^{-1}| + 2^{-(m+1)} = x(\widetilde{x}_n - x^{-1})^2 + 2^{-(m+1)}.$$

Hence, setting $e_n(x) := xe_{n-1}^2(x) + 2^{-(m+1)}$ with $e_0(x) := \frac{1}{2}x^{-1}$, we find that $|\widetilde{x}_n - x^{-1}| \le e_n(x)$ for all $n \in \mathbb{N}$. To show that $e_{n_0}(x) \le 2^{-m}$ on $[2^{-\underline{k}}, 2^{\overline{k}}]$, we introduce $d_n(x) := e_n(x) - \frac{1}{x}2^{-2^n}$ for $n \in \mathbb{N}$ and $x \in [2^{-\underline{k}}, 2^{\overline{k}}]$. Since $\frac{1}{x}2^{-2^n} = x\left(\frac{1}{x}2^{-2^{n-1}}\right)^2$, d_n satisfies the recursion $d_n(x) = xd_{n-1}^2(x) + 2^{-2^n+1}d_{n-1}(x) + 2^{-(m+1)}$ with $d_0 \equiv 0$. Furthermore, we claim that d_n is non-decreasing and convex. Clearly, this is true for d_0, and by the recursion above we have

$$d_n'(x) = \frac{\mathrm{d}}{\mathrm{d}x}[xd_{n-1}^2(x) + 2^{-2^n+1}d_{n-1}(x)] = d_{n-1}^2(x) + 2xd_{n-1}'(x)d_{n-1}(x) + 2^{-2^n+1}d_{n-1}'(x).$$

By induction, this is non-negative and non-decreasing, so d_n is non-decreasing and convex as claimed. Since $e_n(x) = \frac{1}{x}2^{-2^n} + d_n(x)$, it follows that e_n is also convex, and we thus only need to check that $e_{n_0}(2^{-\underline{k}}) \vee e_{n_0}(2^{\overline{k}}) \le 2^{-m}$. To this end, it is clear that for $n \ge 2$, we have $d_n(2^{-\underline{k}}) \ge d_{n+1}(2^{-\underline{k}})$, and some tedious but straightforward calculations show that

$$d_3(2^{-\underline{k}}) = \left(1 + 2^{-\underline{k}}2^{-(m+1)}\left(2^{-k}2^{-(m+1)} + \frac{17}{16}\right)^2 + \frac{1}{256}\left(2^{-k}2^{-(m+1)} + \frac{17}{16}\right)\right)2^{-(m+1)}.$$

A very rough estimate yields from this that $d_n(2^{-k}) \le \frac{3}{2}2^{-(m+1)} = 2^{-m} - 2^{-(m+2)}$ for $n \ge 3$. Now, since $\underline{k} + m + 2 > 4$, we have $n_0 \ge 3$, and so

$$e_{n_0}(2^{-\underline{k}}) = 2^{\underline{k}}2^{-2^n} + d_{n_0}(2^{-\underline{k}}) \le 2^{\underline{k}}2^{-(\underline{k}+m+2)} + 2^{-m} - 2^{-(m+2)} = 2^{-m}.$$

As for $e_{n_0}(2^{\overline{k}})$, since $\overline{k} \leq m-1$, the same convexity argument implies that it suffices to check that $e_{n_0}(2^{m-1}) \leq 2^{-m}$. But we clearly have that $e_0(2^{m-1}) = 2^{-m}$, and since

$$2^{m-1}(2^{-m})^2 + 2^{-(m+1)} = 2 \cdot 2^{-(m+1)} = 2^{-m},$$

it follows that $e_n(2^{m-1}) = 2^{-m}$ for all $n \in \mathbb{N}$. Now, let us construct this network. For notation, let $k = \underline{k} + \overline{k}$. We begin by constructing φ^{init} as a piecewise linear function. In particular, for $i = 0, \ldots, k$, let $p_i = 2^{i-\underline{k}}$, and for $i = 1, \ldots, k$ let ℓ_i be the linear interpolation between (p_{i-1}, p_{i-1}^{-1}) and (p_i, p_i^{-1}), i.e.,

$$\ell_i(x) = \frac{p_i^{-1} - p_{i-1}^{-1}}{p_i - p_{i-1}} x - \left(\frac{p_i^{-1} - p_{i-1}^{-1}}{p_i - p_{i-1}} p_i - \frac{1}{p_i} \right) = \frac{1}{p_i}\left(3 - \frac{2}{p_i} x\right), \quad x \in [p_{i-1}, p_i].$$

Note that by convexity of the reciprocal function, we have $\ell_i(x) \geq x^{-1}$ for all $x \in [p_{i-1}, p_i]$, and hence $E_i(x) := |\ell_i(x) - \frac{1}{x}| = \frac{1}{p_i}\left(3 - \frac{2}{p_i}x\right) - \frac{1}{x}$. We then have that $E_i'(x) = \frac{1}{x^2} - \frac{2}{p_i^2}$, and since E_i has exactly one extremal point and this is a maximum, we have that the error is maximized when $x = \frac{p_i}{\sqrt{2}}$ and that the corresponding error is $E_i(\frac{p_i}{\sqrt{2}}) = \frac{3-2\sqrt{2}}{p_i} \leq \frac{1}{2x}$. Thus, since ℓ_i is merely an affine transformation and hence clearly representable as a neural network, ℓ_i satisfies the properties of φ^{init} on $[p_{i-1}, p_i]$. Also, if $x \in [p_{i-1}, p_i]$, then $\ell_j(x) < x^{-1}$ for all $j \neq i$, and so $\varphi^{\text{init}}(x) := \max_{j=1,\ldots,k} \ell_j(x)$ works. Since $a \vee b = a + \sigma(b-a)$ for $a, b \geq 0$, the function $(y_1, \ldots, y_k) \mapsto \max_{j=1,\ldots,k} y_j$ is realized by a network $\varphi^{\max} \in \Phi(2n_k, W_{\max}, 5k, 1)$, where $n_k = \lceil \log_2(k) \rceil$ and

$$W_{\max} = \begin{bmatrix} k & 2^{n_k} & 2^{n_k-1} & 2^{n_k-1} & \cdots 2 & 2 & 1 \end{bmatrix}^\top.$$

Thus, by parallelizing, we have $\varphi^{\text{init}} \in \Phi(2(n_k+1), \begin{bmatrix} 1 & k & W_{\max}^\top \end{bmatrix}^\top, 8k, 2^{2k})$. Next, we construct φ_m^{iter}. Similarly to the proof of Lemma B.47, a small modification of Schmidt-Hieber [101, Lemma A.2] yields a network $\varphi_l^{\text{mult}} \in \Phi(l+4, W_{\text{mult}}, 24+16l, 2^k)$ with

$$W_{\text{mult}} = \begin{bmatrix} 2 & \underbrace{6 \quad \cdots \quad 6}_{l+2 \text{ times}} & 1 \end{bmatrix}^\top,$$

satisfying $|\varphi_l^{\text{mult}}(x,y) - xy| \leq 2^{k+1-l}$ for all $x \in [0,2]$ and $y \in [0, 2^k]$. Since we always have $x \wedge \widetilde{x}_n \leq 2$ and $|2 - x\widetilde{x}_n| \leq 2$, setting $\varphi_m^{\text{iter}}(x,y) = \varphi_l^{\text{mult}}(x, 2 - \varphi_l^{\text{mult}}(x,y))$ with $l \geq 2k + m + 3$ works. Parallelizing with the identity, we can now simply chain these together without altering the size of the network other than increasing the width of each layer by 1. That is, setting $\widetilde{\varphi}^{\text{init}}(x) := \begin{bmatrix} \varphi^{\text{init}}(x) & x \end{bmatrix}^\top$ and $\widetilde{\varphi}_m^{\text{iter}}(x,y) := \begin{bmatrix} \varphi_m^{\text{iter}}(x,y) & x \end{bmatrix}^\top$,

$$\varphi_m^{\text{rec}}(x) := \varphi_m^{\text{iter}} \circ \underbrace{\widetilde{\varphi}_m^{\text{iter}} \circ \cdots \circ \widetilde{\varphi}_m^{\text{iter}}}_{n_0-1 \text{ times}} \circ \widetilde{\varphi}^{\text{init}}(x),$$

yields the desired network. □

Proof of Lemma B.42. We first approximate $t \mapsto \sqrt{t}$ by a piecewise linear function. To this end, let $t_i = 2^{-m} + i/m$ for $i = 0, \dots, m$ such that $t_m \geq 1$. Next, for $j = 1, \dots, m$ and $t \in [t_{j-1}, t_j]$, set

$$\ell_j(t) = m(\sqrt{t_j} - \sqrt{t_{j-1}})(t - t_{j-1}) + \sqrt{t_{j-1}},$$

such that ℓ_j is the linear interpolation between the points $(t_{j-1}, \sqrt{t_{j-1}})$ and $(t_j, \sqrt{t_j})$. Since $t \mapsto \sqrt{t}$ is concave, it follows that, for $t \in [t_{i-1}, t_i]$, we have $\ell_i(t) \leq \sqrt{t}$ while $\ell_j(t) \geq \sqrt{t}$ for all $j \neq i$. Hence, setting $\ell(t) = \min_{j \in [m]} \ell_j(t)$ yields the desired piecewise linear approximation, and since both affine functions and minima are exactly representable as neural networks, we find that $\ell \in \widetilde{\Phi}(\log m, m, m, \sqrt{m})$. We claim that setting $\varphi^{\mathrm{cap}} = \varphi_1^{\mathrm{rec}} \circ \ell$ yields the desired network. Indeed, for the size of the network, note that since $\ell(t) \in [2^{-m/2}, 1]$ for all $t \in [2^{-m}, 1]$, we have $\varphi_1^{\mathrm{rec}} \in \widetilde{\Phi}(m \log m, m, m \log m, 2^{m/2})$, and since this dominates the size of ℓ, it follows that φ^{cap} is of the same size as φ_1^{rec}.

As for the claim that $\varphi^{\mathrm{cap}}(t) \asymp \frac{1}{\sqrt{t}}$, first note that

$$\varphi^{\mathrm{cap}}(t) = \frac{1}{\sqrt{t}} + \left(\varphi_1^{\mathrm{rec}} \circ \ell(t) - \frac{1}{\ell(t)}\right) + \left(\frac{1}{\ell(t)} - \frac{1}{\sqrt{t}}\right),$$

where the second term is in $[-1/2, 1/2]$, whence

$$\frac{1}{2\sqrt{t}} \leq \frac{1}{\sqrt{t}} + \left(\varphi_1^{\mathrm{rec}} \circ \ell(t) - \frac{1}{\ell(t)}\right) \leq \frac{3}{2\sqrt{t}}.$$

For the last term, note that since $\ell(t_i) = \sqrt{t_i}$ for $i = 0, \dots, m$, we have

$$\left|\frac{1}{\ell(t)} - \frac{1}{\sqrt{t}}\right| \leq \frac{1}{2^{m/2}} - \frac{1}{\sqrt{2^m + m}} < \frac{1}{4},$$

whereby

$$\frac{1}{4\sqrt{t}} \leq \varphi^{\mathrm{cap}}(t) \leq \frac{7}{4\sqrt{t}},$$

as desired. ☐

Article C

*Note: this is a **copy** of the article*

Asbjørn Holk, Claudia Strauch and Lukas Trottner (2026)
Reflected Diffusion Models Adapt to Low-dimensional Data
Preprint available
arXiv:2603.24495 [math.ST]

The only changes made are to the notation and layout for sake of consistency with the rest of the dissertation.

Reflected diffusion models adapt to low-dimensional data

Asbjørn Holk, Claudia Strauch and Lukas Trottner

Abstract

While the mathematical foundations of score-based generative models are increasingly well understood for unconstrained Euclidean spaces, many practical applications involve data restricted to bounded domains. This paper provides a statistical analysis of reflected diffusion models on the hypercube $[0,1]^D$ for target distributions supported on d-dimensional linear subspaces. A primary challenge in this setting is the absence of Gaussian transition kernels and translation invariance, both of which play a central role in standard theory in $\mathbb{R}^D$. By employing an easily implementable infinite series expansion of the transition densities, we develop analytic tools to bound the score function and its approximation by sparse ReLU networks. For target densities with Sobolev smoothness α, we establish a convergence rate in the 1-Wasserstein distance of order $n^{-\frac{\alpha+1-\delta}{2\alpha+d}}$ for arbitrarily small $\delta > 0$, demonstrating that the generative algorithm fully adapts to the intrinsic dimension d. These results confirm that the presence of reflecting boundaries does not degrade the fundamental statistical efficiency of the diffusion paradigm, matching the almost optimal rates known for unconstrained settings.

C.1 Introduction

Deep generative models constitute a broad and rapidly evolving class of methods for learning complex data distributions from samples, with score-based diffusion models [107] emerging as a particularly powerful dynamic paradigm in recent years. Motivated by the fact that many modern data sets are intrinsically low-dimensional yet embedded in high-dimensional bounded ambient spaces, we study the statistical performance diffusion-based generative modelling for probability measures supported on lower-dimensional manifolds within a bounded domain.

The study of the statistical performance of diffusion models has become a central avenue of research in statistics for machine learning. Several papers [6, 21, 22, 37, 40, 72, 89, 97, 118, 135] consider the convergence of such algorithms in the standard setting of an Ornstein–Uhlenbeck (OU in the following) noising model under different regularity and structural assumptions on the target distribution as well as different score approximation classes such as neural networks with or without sparsity assumptions and kernel-type estimators. We provide more details on existing results for unconstrained models and how they relate to our work in the discussion in Section C.4, but for now focus on the class of reflected diffusion models that we consider here.

Such generative models were first introduced in [45, 80] motivated by the fact that practical implementations of the generative backward process often rely on thresholding procedures to enforce geometric constraints on the data, even though the forward OU training model ignores such constraints. To overcome such theoretical discrepancies [45, 80] suggest to use a *reflected* diffusion process as driver of noise instead. This allows to follow the same time-reversal rationale underlying unconstrained diffusion models since the time-reversal of a reflected diffusion is again a reflected diffusion with adjusted drift incorporating the information on the target data provided by the time evolution of the *score*, that is the log-gradient of the forward marginals given initialisation in the data distribution.

First statistical guarantees for such models were given in [58] under the assumption that the data distribution has full support on the bounded reflection domain with Sobolev density bounded away from zero. The goal of this paper is to extend this analysis to singular target distributions supported on a lower-dimensional manifold $M \subset [0,1]^D$ and to demonstrate that the convergence rate of reflected diffusion models adapt to the intrinsic dimension of the data. In doing so, we provide the first rigorous statistical analysis of diffusion-based constrained generative models on bounded domains that explicitly accounts for low-dimensional data structures. This is particularly important from both an applied and a theoretical perspective in light of the so-called *manifold hypothesis* [43, 77, 81]. It postulates that image or text data (and many others) although being extremely high-dimensional have common structural features that make them supported on (unions of) much lower dimensional manifolds. Empirical evidence of this has, e.g., been provided by [15, 96, 108], which makes the manifold hypothesis a reasonable explanation for the tremendous success of deep generative models, provided that their adaptivity to intrinsic lower-dimensional geometric structures can be theoretically verified.

As a natural starting point, we focus on the simplified case where M lies in a linear subspace of $\mathbb{R}^D$. This agrees with the route taken by first studies on adaptivity of unconstrained diffusion models to lower-dimensional data [22, 89] and provides an important foundation for further investigations into more complex manifold structures. Our main contributions can be summarised as follows:

- While existing literature on manifold adaptation relies fundamentally on the Gaussian transition kernels of unconstrained OU processes, we develop analytic tools to control the score function associated with *reflected Brownian motion on the hypercube.* Compared to [58] we do not work with an eigenfunction expansion of the the score, but use the simple geometry of the hypercube together with symmetries of reflected Brownian motion to expand the transition densities as an infinite mixture of restricted Gaussian densities. This allows us to provide precise bounds on the spatial growth of the score and its singular behaviour as $t \searrow 0$, effectively *decoupling the boundary effects of the domain* from the concentration of the measure μ around the low-dimensional subspace M.

- We prove that, despite the analytic complexities introduced by the reflecting boundaries and the resulting non-Gaussian transition densities, denoising score matching via sparse ReLU networks achieves the required approximation rates. In particular, we demonstrate that the *complexity of the estimator depends only on the intrinsic dimension d* of the linear subspace supporting the data, rather than the ambient dimension D.

- We derive an *upper bound for the* 1*-Wasserstein distance* between the target distribution and the law of the generated samples. The established rate of $O(n^{-(\alpha+1-\delta)/(2\alpha+d)})$ confirms that imposing hard physical constraints via reflection on $[0,1]^D$ does *not* degrade the fundamental statistical efficiency of the diffusion model, matching the almost optimal rates known for unconstrained dynamics.

In the following, we summarise the generative and statistical estimation procedure, introduce and discuss our assumptions on the target distribution and provide an informal version of our main result on 1-Wasserstein convergence rates of reflected diffusion models.

Forward reflected diffusion Let μ be a target probability distribution on $\mathbb{R}^D$, concentrated on a compact d-dimensional manifold $M \subset [0,1]^D$, where possibly $d \ll D$. Given an i.i.d. sample of data with distribution μ, our aim is to generate approximate samples for μ in a two-step procedure via a time-reversal mechanism for reflected diffusions.

As a first step, we perturb μ by adding isotropic noise through a reflected Brownian motion on the hypercube. Specifically, we consider the reflected SDE

$$\mathrm{d}\mathsf{X}_t = \mathrm{d}\mathsf{B}_t + n(\mathsf{X}_t)\,\mathrm{d}\mathsf{L}_t, \quad \mathsf{X}_0 \sim \mu, \tag{C.67}$$

where $(\mathsf{B}_t)_{t\geq 0}$ is a standard D-dimensional Brownian motion, $n(x)$ denotes an inward-pointing normal vector at $x \in \partial[0,1]^D$, and $(\mathsf{L}_t)_{t\geq 0}$ is the local time of $(\mathsf{X}_t)_{t\geq 0}$ at $\partial[0,1]^D$, i.e., a one-dimensional, continuous and non-decreasing process satisfying $\mathsf{L}_t = \int_0^t \mathbf{1}_{\{\mathsf{X}_s\in\partial[0,1]^D\}}\,\mathrm{d}\mathsf{L}_s$ and $\int_0^t |n(\mathsf{X}_s)|\,\mathrm{d}\mathsf{L}_s < \infty$ almost surely. The presence of the stochastic forcing term $n(\mathsf{X}_t)\,\mathrm{d}\mathsf{L}_t$ in the dynamics prevents the process from escaping the unit cube by normally reflecting it back into the interior when it hits the boundary. Note that for boundary points $x \in \partial[0,1]^d$ where two or more faces of the cube intersect, the direction of $n(x)$ is not uniquely defined. If for $x \in \partial[0,1]^D$, we let $I(x), J(x) \subseteq [D]$ denote the indices of x for which $x_i = 0$ and $x_j = 1$, respectively, we specify $n(x) = \sum_{i\in I(x)} e_i - \sum_{j\in J(x)} e_j$, where e_i is the i-th standard unit vector in $\mathbb{R}^D$. In particular, $n(x)$ is the unique inward pointing normal vector on smooth parts of the cube boundary, where faces do not intersect. The particular choice on the non-smooth part of the boundary $E := \{x \mid |I(x)| + |J(x)| > 1\}$ is without consequences, since the reflected Brownian motion will never hit E almost surely when started in $[0,1]^D \setminus E$, cf. [131, Theorem 1.1] and we will assume without further mention that $\operatorname{supp}(\mu) \cap E = \emptyset$.

Existence and pathwise uniqueness of strong solutions for general reflected diffusions in bounded convex domains has been shown in [117] under mild conditions on the coefficients, which are satisfied for the Brownian case considered here. In particular, because of the simple geometry of $[0,1]^D$ and the normal reflection direction, the i-th coordinate X^i of the strong solution of (C.67) is a strong solution to the one-dimensional reflected SDE

$$\mathrm{d}\mathsf{X}^i_t = \mathrm{d}\mathsf{B}^i_t - \operatorname{sgn}(\mathsf{X}^i_t)\,\mathrm{d}\mathsf{L}^i_t,$$

where L^i is the local time at $\{0,1\}$ and $\operatorname{sgn}(x) = -1$ for $x \leq 0$ and $\operatorname{sgn}(x) = 1$ for $x > 0$. Thus, conditional on the initialisation X_0, the components of X are independent reflected Brownian motions on $[0,1]$ and $\mathsf{L} = \sum_{i=1}^D \mathsf{L}^i$ almost surely. The boundary local times L^i at the faces are characterised via the occupation limit

$$\mathsf{L}^i_t = \lim_{\varepsilon\downarrow 0} \frac{1}{2\varepsilon}\int_0^t \mathbf{1}_{[0,\varepsilon]\cup[1-\varepsilon,1]}(\mathsf{X}^i_s)\,\mathrm{d}s,$$

which hold both almost surely and in L^2, uniformly on relatively compact sets in t, see [16, Theorem 2.6] in a more general context. Consequently,

$$\mathsf{L}_t = \lim_{\varepsilon\downarrow 0}\frac{1}{2\varepsilon}\int_0^t \sum_{i=1}^{D} \mathbf{1}_{[0,\varepsilon]\cup[1-\varepsilon,1]}(\mathsf{X}_s^i)\,\mathrm{d}s, \tag{C.68}$$

uniformly in L^2 and almost surely on relatively compact sets of t. In the following, we let p_t denote the density of X_t wrt Lebesgue measure on $\mathbb{R}^D$.

Time reversal and generative sampling Fix a terminal time $\overline{T} > 0$, and define the time-reversed process $\overleftarrow{\mathsf{X}}_t := \mathsf{X}_{\overline{T}-t}$ for $t \in [0, \overline{T}]$. Then, there exists a Brownian motion $(\overline{\mathsf{B}}_t)_{t\geq 0}$ such that $\overleftarrow{\mathsf{X}}$ satisfies the reflected SDE

$$\mathrm{d}\overleftarrow{\mathsf{X}}_t = \nabla\log p_{\overline{T}-t}(\overleftarrow{\mathsf{X}}_t)\,\mathrm{d}t + \mathrm{d}\overline{\mathsf{B}}_t + n(\overleftarrow{\mathsf{X}}_t)\,\mathrm{d}\overline{\mathsf{L}}_t, \qquad \overleftarrow{\mathsf{X}}_0 \sim p_{\overline{T}}, \tag{C.69}$$

where $\overline{\mathsf{L}}_t := \mathsf{L}_{\overline{T}} - \mathsf{L}_{\overline{T}-t}$ is the local time of $\overleftarrow{\mathsf{X}}$ at the boundary $\partial[0,1]^D$. This result is proved in [18, Theorem 2.5] for more general reflected diffusions on smooth domains, while [45, Theorem 3.2] give an instructive probabilistic proof inspired by the non-reflected case [53] for reflected Brownian motion on precompact, *smooth* convex domains. Their proof relies on the boundary occupation limit characterisation of the local time, which in our case is provided by (C.68), and sufficient smoothness properties of the transition densities of X_t, which in our model can be verified based on the representation given in Lemma C.51. Thus, even though the unit cube $[0,1]^D$ is non-smooth, its simple geometry allows us to provide the technical tools needed to follow the proof of [45, Theorem 2.3] to verify (C.69).

If the score function $s_0(x,t) := \nabla\log p_t(x)$ were known, then simulating (C.69) would yield exact samples from μ at time $\overline{T}$. Since p_t and s_0 depend on the unknown target distribution, they must be approximated from data.

Approximate backward diffusion. Given an approximation $s(x,t)$ of the score function, we instead consider the reflected SDE

$$\mathrm{d}\overline{\mathsf{X}}^s_t = s(\overline{\mathsf{X}}^s_t, \overline{T}-t)\,\mathrm{d}t + \mathrm{d}\overline{\mathsf{B}}_t + n(\overline{\mathsf{X}}^s_t)\,\mathrm{d}\overline{\mathsf{L}}_t, \qquad \overline{\mathsf{X}}^s_0 \sim \mathcal{U}([0,1]^D). \tag{C.70}$$

Several structural features of the proposed framework motivate the use of reflected diffusions on the hypercube. First, the state space $[0,1]^D$ is natural in many applications, including image and signal generation, and reflection provides a principled mechanism to ensure that generated samples remain within prescribed bounds. Second, the forward process (C.67) has zero drift and unit diffusion, which allows for an explicit representation of its transition kernel (see Lemma C.50) and substantially simplifies the probabilistic analysis. Third, the uniform distribution on $[0,1]^D$ is invariant for the forward reflected Brownian motion, yielding a simple and practically convenient initialisation for the backward dynamics. For numerical stability, we do not run the backward dynamics all the way to time $\overline{T}$, but instead output the sample $\overline{\mathsf{X}}^s_{\overline{T}-\underline{T}}$ for some small $\underline{T} > 0$. This introduces three distinct sources of error:

(i) the truncation error due to early stopping at time $\underline{T}$;

(ii) the initialisation error from starting at stationarity rather than $p_{\overline{T}}$;

(iii) the approximation error from using s instead of the true score s_0.

Error metric. Our goal is to quantify the discrepancy between μ and the law of the generated samples. Since the algorithm produces a random probability measure as the terminal law of a stochastic process, a natural notion of error is provided by Wasserstein distances. More specifically, our error criterion is the 1-Wasserstein distance, for probability measures ν_1, ν_2 on $[0,1]^D$ defined by

$$\mathcal{W}_1(\nu_1, \nu_2) := \inf_{\pi \in \Pi(\nu_1, \nu_2)} \int_{[0,1]^D \times [0,1]^D} |x - y| \, \pi(\mathrm{d}x, \, \mathrm{d}y),$$

where $\Pi(\nu_1, \nu_2)$ denotes the set of all couplings of ν_1 and ν_2. Unlike divergences based on densities, $\mathcal{W}_1$ remains meaningful when ν_1 or ν_2 are supported on lower-dimensional sets, and it admits a natural interpretation in terms of couplings of stochastic processes, making it well suited for diffusion-based generative models.

Score estimation via denoising score matching. To estimate the score function, we discretise the time interval $[\underline{T}, \overline{T}]$ into $K \in \mathbb{N}$ subintervals $\{[t_{i-1}, t_i]\}_{i=1}^K$, where $K \asymp \log n$ and $t_i = \underline{T} c^i$ for some $c \in (1, 2]$, $t_K = \overline{T}$. On each subinterval, we approximate the map $(x, t) \mapsto \nabla \log p_t(x)$ separately. The construction of the score estimator is based on the classical equivalence between the explicit score matching loss

$$\int_{t_{i-1}}^{t_i} \mathbb{E}\left[|s(\mathrm{X}_t, t) - \nabla \log p_t(\mathrm{X}_t)|^2\right] \mathrm{d}t$$

and the denoising score matching loss

$$\int_{t_{i-1}}^{t_i} \mathbb{E}\left[|s(\mathrm{X}_t, t) - \nabla \log q_t(\mathrm{X}_0, \mathrm{X}_t)|^2\right] \mathrm{d}t,$$

where $q_t(x, \cdot)$ denotes the transition density of the forward reflected diffusion at time t, started from x. More precisely, for any measurable function s, one has

$$\int_{t_{i-1}}^{t_i} \mathbb{E}\left[|s(\mathrm{X}_t, t) - \nabla \log p_t(\mathrm{X}_t)|^2\right] \mathrm{d}t = \mathbb{E}\left[\int_{t_{i-1}}^{t_i} |s(\mathrm{X}_t, t) - \nabla \log q_t(\mathrm{X}_0, \mathrm{X}_t)|^2 \, \mathrm{d}t\right] + C_i,$$

where $C_i = -\mathbb{E}\left[\int_{t_{i-1}}^{t_i} |\nabla \log p_t(\mathrm{X}_t) - \nabla \log q_t(\mathrm{X}_0, \mathrm{X}_t)|^2 \, \mathrm{d}t\right]$ is a constant independent of s. Accordingly, for a given approximation class $\mathcal{S}_i$ and $i \in [K]$, we define

$$L_s^{(i)}(x) := \int_{t_{i-1}}^{t_i} \mathbb{E}\left[|s(\mathrm{X}_t, t) - \nabla \log q_t(x, \mathrm{X}_t)|^2 \mid \mathrm{X}_0 = x\right] \mathrm{d}t.$$

Minimising the explicit score matching loss over $\mathcal{S}_i$ is then equivalent to minimising $\mathbb{E}[L_s^{(i)}(\mathsf{X}_0)]$ over the same class. Given i.i.d. samples $\mathsf{Y}_1, \dots, \mathsf{Y}_n \sim \mu$, a natural estimator of the score on the interval $[t_{i-1}, t_i]$ is obtained by minimising the empirical denoising score matching loss

$$\widehat{\mathsf{L}}_s^{(i)} := \frac{1}{n} \sum_{j=1}^{n} L_s^{(i)}(\mathsf{Y}_j).$$

For a collection of sparse ReLU neural network classes $\{\mathcal{S}_i\}_{i=1}^K$, we thus define the overall score estimator as the piecewise function

$$\widehat{s}_n(x, t) = \sum_{i=1}^{K} \widehat{s}_n^{(i)}(x, t)\, \mathbf{1}_{[t_{i-1}, t_i)}(t), \qquad \text{where } \widehat{s}_n^{(i)} \in \operatorname*{argmin}_{s \in \mathcal{S}_i} \widehat{L}_s^{(i)}.$$

Conditional on $\widehat{s}_n$ we then simulate the reflected SDE (C.70) with $s = \widehat{s}_n$ until time $\overline{T} - \underline{T}$ and use $\overline{\mathsf{X}}^{\widehat{s}_n}_{\overline{T}-\underline{T}}$ as an approximate sample for the target distribution μ.

Assumptions and main result The probabilistic results developed in Section C.2 apply to general target measures supported on smooth submanifolds. For the statistical analysis of score estimation and for deriving explicit convergence rates, however, we restrict attention to a setting in which the geometry of the support is sufficiently simple to permit sharp approximation bounds for neural networks. Specifically, we introduce the following assumptions about μ and its support M:

($\mathcal{H}1$) There exist orthonormal vectors $v_1, \dots, v_d \in \mathbb{R}^D$ with $d \leq D$ and a shift $v_0 \in [0,1]^D$ such that M is connected with non-empty interior, has a Lipschitz boundary and is a closed subset of $(V + v_0) \cap [0,1]^D$ where $V = \operatorname{Span}(v_1, v_2, \dots, v_d)$.

Moreover, there exist constants $c_0 \geq d, r_0 > 0$ such that, for all $x \in M$ and all $r > 0$, we have $\mu\big(B(x,r) \cap M\big) \gtrsim (r \wedge r_0)^{c_0}$ and $\lambda_d(B(x,r) \cap M) \gtrsim (r \wedge r_0)^d$, where λ_d denotes the restriction of the d-dimensional Lebesgue measure to $(V + v_0) \cap [0,1]^D$.

($\mathcal{H}2$) The target distribution μ admits a density p_0 wrt to the λ_d such that

(i) $p_0 \in H_0^\alpha(M)$ with $\alpha \in \mathbb{N} \cap (d/2, \infty)$, i.e., the density has Sobolev smoothness α on M and the weak derivatives up to order $\alpha - 1$ vanish at the boundary in the trace sense;

(ii) p_0 is bounded and bounded away from zero on an interior region of M, i.e., there exist constants $0 < p_{\min} \leq p_{\max} < \infty$ and $\varepsilon_M > 0$ satisfying $p_0(x) \leq p_{\max}$ for all $x \in M$ and $p_0(x) \geq p_{\min}$ for all $x \in M_{-\varepsilon_M/2} := M \setminus (\partial M)_{\varepsilon_M/2}$, where $(\partial M)_{\varepsilon_M/2}$ denotes the $\varepsilon_M/2$-fattening of ∂M.

Existence of $r_0 > 0$ such that $\lambda_d(B(x,r) \cap M) \gtrsim (r \wedge r_0)^d$ for all $r > 0$ and $x \in M$ is guaranteed if M is β-smooth for some $\beta \geq 2$ and has positive reach $\tau > 0$; see, e.g., [36, Lemma 20]. In this setting, existence of $c_0 \geq d$ such that $\mu\big(B(x,r) \cap M\big) \gtrsim (r \wedge r_0)^{c_0}$ is then further guaranteed if the target density p_0 decays polynomially towards ∂M, i.e., if $p_0(x) \gtrsim \operatorname{dist}(x, \partial M)^{c_0 - d}$ for x sufficiently close to the boundary. A typical construction of densities $p_0 \in H_0^\alpha(M)$ would model $p_0(x) = c \operatorname{dist}(x, \partial M)^{c_0 - d}$ in a neighbourhood of ∂M with $c_0 \geq \alpha + d$. A variation of such an assumption on controlled decay at the boundary has also been used in [113] and allows us to avoid rather artificial strict lower boundedness assumptions on the target density. A visualisation of our support assumption is given in Figure C.13. For small times t, the forward density p_t concentrates sharply around M, and the score $\nabla \log p_t(x)$ grows rapidly as x moves away from M. Accurate score estimation in this regime is statistically delicate, particularly near the boundary of M. To avoid technical complications associated with boundary singularities, we introduce the following auxiliary regularity and geometric conditions.

($\mathcal{H}3$) When restricted to an area near the boundary, the target density p_0 is sufficiently smooth. Specifically, there exists $\varepsilon_M > 0$ such that the restriction of p_0 to a neighbourhood of the boundary, $p_0|_{(\partial M)_{\varepsilon_M} \cap M} \in C^\kappa((\partial M)_{\varepsilon_M} \cap M, \mathbb{R})$, where $\kappa := \frac{d(c_0 - d)}{2} + d + 3\alpha + 2$ and $(\partial M)_{\varepsilon_M}$ denotes the ε_M-fattening of ∂M.

($\mathcal{H}4$) M does not intersect $\partial[0,1]^D$, i.e., there exists $\rho_{\min} > 0$ such that

$$\operatorname{dist}(M, \partial[0,1]^D) := \inf_{x \in M, y \in \partial[0,1]^D} |x - y| \geq \rho_{\min}.$$

When imposing both ($\mathcal{H}2$) and ($\mathcal{H}3$), we assume that the values of ε_M coincide. We note that ($\mathcal{H}3$) is comparable to assumptions made in related work on statistical estimation rates of unconstrained diffusion models, see, e.g., [89, Assumption 6.3], [71, Assumption (**B**)]. If, e.g., $p_0(x) = c \operatorname{dist}(x, \partial M)^{c_0 - d}$ close to the boundary as in the discussion above, this assumption is always satisfied provided ∂M is sufficiently smooth. Assumption ($\mathcal{H}4$) can always be enforced by rescaling the data and undo this scaling for the generated output. Generally, for any of our results, we will precisely state which (if any) of the above assumptions are needed.

With this setup, we can state an informal version of our main theorem; the precise statement is given in Theorem C.59.

Theorem C.49: Assume ($\mathcal{H}1$)–($\mathcal{H}4$). For any $\delta > 0$, choose $\underline{T} \in \operatorname{Poly}(n^{-1})$ and $\overline{T} \asymp \log n$. Then there exists a family $\{\mathcal{S}_i\}_{i=1}^K$ of sparse ReLU neural network classes such that the reflected diffusion generative algorithm driven by the empirical denoising score matching estimator $\widehat{s}_n$ satisfies

$$\mathbb{E}\Big[\mathcal{W}_1\big(\mu, \mathcal{L}(\overline{\mathsf{X}}^{\widehat{s}_n}_{\overline{T}-\underline{T}})\big)\Big] = O\big(n^{-\frac{\alpha+1-\delta}{2\alpha+d}}\big),$$

where $\mathcal{L}(\overline{\mathsf{X}}^{\widehat{s}_n}_{\overline{T}-\underline{T}})$ denotes the law of the output $\overline{\mathsf{X}}^{\widehat{s}_n}_{\overline{T}-\underline{T}}$ conditional on the data.

Organisation of the paper The remainder of the paper is organised as follows. In Section C.2, we develop the probabilistic foundations of our model, including the construction of the forward reflected diffusion and the derivation of the explicit score representation. We also establish crucial bounds on the growth and regularity of the score function in Lemma C.53. Section C.3 is dedicated to the statistical analysis and the proof of our main result. We present the error decomposition for the 1-Wasserstein distance, specify the sparse ReLU network classes used for estimation, and combine these results to prove Theorem C.59. Finally, Section C.4 places our findings in the context of recent minimax results and discusses extensions to general manifolds and discretisation errors. All technical proofs omitted from the main part and auxiliary results are collected in the Appendix.

C.2 Probabilistic analysis

Our statistical analysis relies on a detailed understanding of distributional and path properties of normally reflected Brownian motions in a hypercube, which we develop in this section. Generally, strong solutions to reflected SDEs can be constructed via the so-called Skorokhod map. In the present setting of a normally reflected Brownian motion in the hypercube, this is a mapping $\Gamma : C([0,\infty);\mathbb{R}^D) \to C([0,\infty);[0,1]^D)$ such that the strong solution of (C.67) may be written as $\mathsf{X}_\cdot = \Gamma(\mathsf{Y} + \mathsf{B}_\cdot)$. Although existence and uniqueness of the Skorokhod map are well known, its explicit characterisation is generally intractable. For the purposes of simulation and for analysing distributional properties of the forward process, however, it is sufficient to work with weak solutions. By pathwise uniqueness for reflected diffusions in convex domains, cf. [95, Theorem 2.5.1], weak solutions are also unique in law (the classical Yamada–Watanabe argument assuming pathwise uniqueness for unconstrained SDEs, cf. [63, Proposition 5.3.20], extends to the reflected setting). We therefore begin by constructing a simple weak solution of (C.67), which is particularly convenient for training and theoretical analysis for the following reasons:

(i) simulating the weak solution is as easy as simulating a Brownian motion.

(ii) the weak solution yields a simple, interpretable and numerically simple to approximate series representation of the transition densities $q_t(x,y)$, see Lemma C.51. Together with property (i), this yields a simple recipe that does not require full forward path simulations to numerically approximate the empirical denoising score matching loss by using Algorithm 1.

(iii) the transition density formula from Lemma C.51 is perfectly suited to capture the influence of the intrinsic dimensionality of the data support on theoretical approximation properties of the score $\nabla \log p_t(x)$, which eventually translates to faster estimation rates.

All proofs of the lemmata stated in this section are deferred to Appendix C.A.

Lemma C.50: Let $\widehat{f} : \mathbb{R} \to [0,1]$ be the 2-periodic function defined by

$$\widehat{f}(x) := \begin{cases} x, & \text{if } x \in [0,1), \\ -x, & \text{if } x \in [-1,0), \end{cases}$$

and extended periodically to all of $\mathbb{R}$. Define $f : \mathbb{R}^D \to [0,1]^D$ by applying $\widehat{f}$ component-wise, that is, $f(x) := (\widehat{f}(x_i))_{i=1,\dots,D}$. If $\mathsf{Y} \sim \nu$ for some probability measure ν on $[0,1]^D$ and $(\mathsf{B}_t)_{t\geq 0}$ is a D-dimensional Brownian motion independent of Y, then the process $(\mathsf{X}_t)_{t\geq 0}$ defined by $\mathsf{X}_t := f(\mathsf{B}_t + \mathsf{Y})$ is a weak solution to the reflected SDE

$$\mathrm{d}\mathsf{X}_t = \mathrm{d}\mathsf{W}_t + n(\mathsf{X}_t)\,\mathrm{d}\mathsf{L}_t, \quad t \geq 0,$$

with initial distribution $\mathsf{X}_0 \sim \nu$ and Brownian motion W.

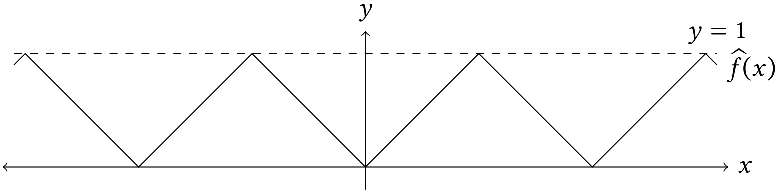

Figure C.10: *Graph of the function $\widehat{f}$ from Lemma C.50. The function reflects the identity between the lines $y = 0$ and $y = 1$; applied component-wise, f therefore essentially reflects the identity at the boundary $\partial[0,1]^D$.*

Using the explicit construction from Lemma C.50, we can derive a closed-form expression for the density p_t of the forward process $(\mathsf{X}_t)_{t\geq 0}$ and, consequently, for the associated score function s_0.

Lemma C.51: Let $(\mathsf{X}_t)_{t\geq 0}$ be a solution to (C.67) and define the reflection operator R_z component-wise by

$$\big(R_z(x)\big)_i := \begin{cases} x_i, & \text{if } z_i \text{ is even}, \\ 1 - x_i, & \text{if } z_i \text{ is odd}, \end{cases}$$

where $i \in \mathbb{N}, z_i \in \mathbb{Z}, x_i \in [0,1]$. Then, for all $x, y \in [0,1]^D$ and $t > 0$, the transition density of the reflected Brownian motion in $[0,1]^D$ is given by

$$q_t(y,x) = (2\pi t)^{-D/2} \sum_{z\in\mathbb{Z}^D} \exp\left(-\frac{|R_z(x) + z - y|^2}{2t}\right),$$

and, for p_t denoting the density of X_t wrt. the Lebesgue measure on $\mathbb{R}^D$,

$$p_t(x) = (2\pi t)^{-D/2} \sum_{z\in\mathbb{Z}^D} \int_{[0,1]^D} \exp\left(-\frac{|R_z(x)+z-y|^2}{2t}\right) \mu(\mathrm{d}y), \qquad x \in [0,1]^D.$$

In particular, the score function $s_0(x,t) = \nabla \log p_t(x)$ admits the explicit representation

$$s_0(x,t) = -\frac{\sum_{z\in\mathbb{Z}^D} (-1)^z \int_{[0,1]^D} (R_z(x)+z-y) \exp\left(-\frac{|R_z(x)+z-y|^2}{2t}\right) \mu(\mathrm{d}y)}{t \sum_{z\in\mathbb{Z}^D} \int_{[0,1]^D} \exp\left(-\frac{|R_z(x)+z-y|^2}{2t}\right) \mu(\mathrm{d}y)}, \tag{C.71}$$

where we use the shorthand $(-1)^z := \operatorname{diag}\big(((-1)^{z_i})_{i=1}^D\big)$.

Combining Lemma C.50 and Lemma C.51 yields the simple numerical algorithm 1 to approximate the empirical denoising score matching loss, which is necessary for implementation of the score estimation procedure.

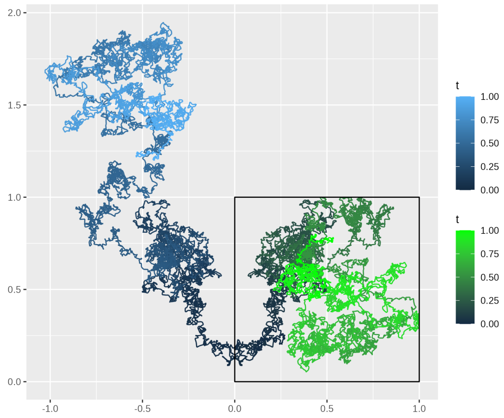

Figure C.11: *Simulation of a reflected Brownian motion (green) along with the non-reflected version (blue) that is used for its construction using Lemma C.50.*

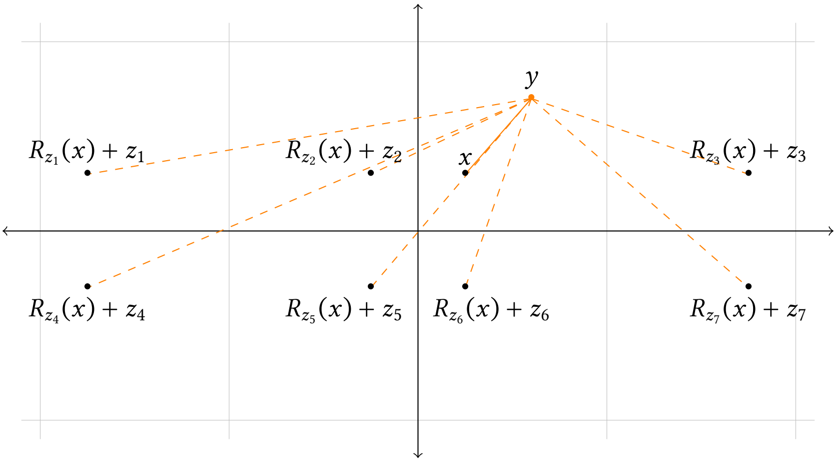

Figure C.12: *The points $R_{z_i}(x)+z_i$ are all mapped back to x under the function f from Lemma C.50. Consequently, the reflected process X_t moves from y to x if and only if the Brownian motion B_t moves from y to x* or *to any of the points $R_{z_i}(x)+z_i$.*

Algorithm 1 Numerical approximation of empirical denoising score matching loss $\widehat{L}_s^{(i)}$

```
**Input:** data $\{\mathsf{Y}_j\}_{j=1}^n \overset{iid}{\sim} \nu$, $N \in \mathbb{N}$, time interval index $i \in [K]$, $s \in \mathcal{S}_i$
set $\widetilde{L}_s^{(i)} = 0$
**for** $k = 1$ **to** $N$ **do**
    draw $y = \mathsf{Y}_{j_k}$ uniformly from $\{\mathsf{Y}_j\}_{j=1}^n$
    draw independently $t \sim \mathcal{U}([t_{i-1}, t_i])$ and $\mathsf{B}_t \sim \mathcal{N}(0, tI_D)$
    set $x_t = f(\mathsf{Y}_{j_k} + \mathsf{B}_t)$
    choose $K \in \mathbb{N}$ and set
        $$\widetilde{\nabla \log q_t}(y, x_t) = \frac{\sum_{z\in\mathbb{Z}^D, |z|\le K} (-1)^z (R_z(x_t) + z - y) \exp\left(-\frac{|R_z(x_t)+z-y|^2}{2t}\right)}{t \sum_{z\in\mathbb{Z}^D, |z|\le K} \exp\left(-\frac{|R_z(x_t)+z-y|^2}{2t}\right)}$$
    set $\widetilde{L}_s^{(i)} \leftarrow \widetilde{L}_s^{(i)} + \frac{1}{N}|s(t, x_t) - \widetilde{\nabla \log q_t}(y, x_t)|^2$
**end for**
**Output:** $\widetilde{L}_s^{(i)}$
```

Remark C.52: (i) choosing the cutoff parameter K dependent on the initialisation y and the drawn time t is important since it should be proportional to the number of reflections along the path $y \to x_t$ (decreasing in t and the distance of y to $\partial[0,1]^D$). See also the discussion in [80] on implementation of the model, where Gaussian approximations are made for small t and spectral decompositions of the transition density are exploited for large t approximations.

(ii) in practice, one may simulate $(y_k, t_k, x_{t_k})_{k=1}^N$ only once and use these for Monte–Carlo approximation of $\widehat{L}_s^{(i)}$ for the updated approximators s in every optimisation step.

These results now allow us to give a precise analysis of the growth of the score $\nabla \log p_t(x)$ in t depending on the distance of x to the data manifold M as well as the path behaviour of the reflected Brownian motion for small times. These properties will play a crucial role for constructing efficient neural network score approximators and for proving almost optimal rates in 1-Wasserstein distance.

Lemma C.53: Fix $t > 0$ and let $M_{\rho,t} = \{x \in [0,1]^D \,:\, \operatorname{dist}(x, M) \leq \sqrt{t(D+2\rho)}\}$ denote the $\sqrt{t(D+2\rho)}$-fattening of M for some $\rho > 1$. Then, under $(\mathcal{H}1)$ the following hold:

(a) $|\nabla \log p_t(x)| \lesssim \frac{1}{t \wedge \sqrt{t}}$ for $x \in [0,1]^D$;

(b) $\mathbb{E}[|\nabla \log p_t(\mathsf{X}_t)|^2 \mathbf{1}_{M_{\rho,t}^c}(\mathsf{X}_t)] \lesssim \frac{1}{t^2 \wedge 1} \mathrm{e}^{-\rho}$;

(c) if $t \leq 1/2$, there exists a universal constant C such that

$$\mathbb{P}\Big(\forall s \in [t,1] \,:\, |\mathsf{X}_s - \mathsf{X}_0| \leq C\sqrt{D}\sqrt{s}\big(\log(1 + \log t^{-1}) + y\big)\Big) \geq 1 - 4D\mathrm{e}^{-2y^2}, \quad y > 0;$$

(d) $p_t(x) \gtrsim t^{\frac{c_0 - D}{2}} \mathrm{e}^{-\rho}$ for $x \in M_{\rho,t}$;

(e) for $t \in (0,1]$,

  (i) $\forall x, y \in [0,1]^D \,:\, |\nabla_x \log q_t(y, x)| \lesssim \frac{|x-y|}{t} + \frac{1}{\sqrt{t}}$;

  (ii) $\forall x \in [0,1]^D \,:\, |\nabla \log p_t(x)| = |\mathbb{E}[\nabla_2 \log q_t(\mathsf{X}_0, \mathsf{X}_t) \mid \mathsf{X}_t = x]| \lesssim \frac{1}{t}\mathbb{E}[|\mathsf{X}_t - \mathsf{X}_0| \mid \mathsf{X}_t = x] + \frac{1}{\sqrt{t}}$;

  (iii) $\forall t \in (0,1], x \in M_{\rho,t} :\ |\nabla \log p_t(x)| \lesssim \frac{\sqrt{\rho + \log t^{-1}}}{\sqrt{t}}$.

Note that part (e).(ii) together with the upper bound $|\mathsf{X}_t - \mathsf{X}_0| \leq \sqrt{D}$ immediately implies the bound from part (a) for $t \in (0,1]$. However, our combinatorial proof technique for (a) can be translated directly to truncated versions of the score representation given in (C.71), which will form the basis of our neural network approximation strategy in the next section. Conversely, the proof of (e).(ii) relies on the denoising score representation $\nabla \log p_t(x) = \mathbb{E}[\nabla_2 \log q_t(\mathsf{X}_0, \mathsf{X}_t) \mid \mathsf{X}_t = x]$, which has no probabilistic analogue for the truncated score representation.

C.3 Wasserstein convergence rate

This section establishes quantitative convergence guarantees in the 1-Wasserstein distance for the reflected diffusion generative scheme driven by an estimated score. Our analysis builds on a careful decomposition of the approximation error and combines statistical bounds for score estimation with probabilistic stability estimates for reflected stochastic dynamics. In our earlier work [58], we derived upper bounds of order $n^{-\alpha/(2\alpha+D)}$ (up to polylogarithmic factors) for the total variation distance under Sobolev smoothness $\alpha > D/2$, expressed in terms of the ambient dimension D. In the present bounded-domain setting, such bounds immediately imply corresponding guarantees in the 1-Wasserstein distance. However, classical results from nonparametric density estimation in the i.i.d. setting (see, for instance, Theorem 2 in [87]) suggest that these rates are suboptimal. Indeed, [89] obtained an improved upper bound of order $n^{-(\alpha+1-\delta)/(2\alpha+D)}$, for arbitrary $\delta > 0$, by exploiting a refined multiscale analysis of the reverse diffusion. While our overall strategy is inspired by the approach of [89], their arguments cannot be transferred verbatim to the present setting. In particular, the pathwise stability estimates in [89] rely heavily on properties of OU processes on $\mathbb{R}^D$, whereas our model is governed by reflected diffusions on a compact domain with boundary. As a consequence, we must develop and invoke genuinely new probabilistic tools, including precise growth and regularity bounds for reflected Brownian paths and their associated scores, as established in Lemma C.53. At the same time, the compactness of the state space allows us to simplify several technical aspects of the construction and to avoid certain localisation arguments that are necessary in the unbounded setting.

Error decomposition

As outlined in the introduction, the overall approximation error decomposes into three distinct contributions: the error due to early stopping of the backward dynamics, the error incurred by approximating the score function, and the error arising from initialising the dynamics with the uniform distribution on $[0,1]^D$ rather than with the target distribution. To disentangle these effects, we introduce, for a given score approximation s, an auxiliary reflected diffusion $(\widehat{\mathsf{X}}^s_t)_{t\in[0,\overline{T}]}$ defined as the solution to

$$\mathrm{d}\widehat{\mathsf{X}}^s_t = s(\widehat{\mathsf{X}}^s_t, \overline{T}-t)\,\mathrm{d}t + \mathrm{d}\mathsf{B}_t + n(\widehat{\mathsf{X}}^s_t)\,\mathrm{d}\mathsf{L}_t, \qquad t \in [0,\overline{T}],$$

with initial condition $\widehat{\mathsf{X}}^s_0 \sim p_{\overline{T}}$. The triangle inequality for the 1-Wasserstein metric $\mathcal{W}_1$ yields for the process defined in (C.70) the error decomposition

$$\mathcal{W}_1\big(\mu, \overline{\mathsf{X}}^s_{\overline{T}-\underline{T}}\big) \le \mathcal{W}_1(\mu, \mathsf{X}_{\underline{T}}) + \mathcal{W}_1\big(\mathsf{X}_{\underline{T}}, \widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}\big) + \mathcal{W}_1\big(\widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}, \overline{\mathsf{X}}^s_{\overline{T}-\underline{T}}\big). \tag{C.72}$$

Since $\mathsf{X}_{\underline{T}} \sim \widehat{\mathsf{X}}^{s_0}_{\overline{T}-\underline{T}}$, we may rewrite

$$\mathcal{W}_1\big(\mathsf{X}_{\underline{T}}, \widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}\big) = \mathcal{W}_1\big(\widehat{\mathsf{X}}^{s_0}_{\overline{T}-\underline{T}}, \widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}\big),$$

where the two reflected processes $\widehat{\mathsf{X}}^s$ and $\widehat{\mathsf{X}}^{s_0}$ are initialised in the same distribution $p_{\overline{T}}$, but have different drifts s and s_0, respectively. Likewise, $\widehat{\mathsf{X}}^s$ and $\overline{\mathsf{X}}^s$ share the same drift, but are started in different initial distributions $p_{\overline{T}}$ and $\mathcal{U}[0,1]^D$, respectively. Consequently, the three terms on the right-hand side of (C.72) correspond, respectively, to the error due to early stopping, the error caused by approximating the score function, and the error introduced by initialising the dynamics with the uniform distribution. We begin by controlling the first and third term, which admit comparatively elementary bounds.

Lemma C.54: Let μ be an arbitrary probability distribution on $[0,1]^D$, and let $(\mathsf{X}_t)_{t\geq 0}$ be a solution to (C.67) with initial condition $\mathsf{X}_0 \sim \mu$. Then, for all $t \geq 0$,

$$\mathcal{W}_1(\mu, \mathsf{X}_t) \leq \sqrt{D\,t}.$$

Proof. Let $\mathsf{Y} \sim \mu$, and define $\mathsf{X}_t = f(\mathsf{B}_t + \mathsf{Y})$, where f is as in Lemma C.50, and $(\mathsf{B}_t)_{t\geq 0}$ is a Brownian motion independent of Y. By Lemma C.50, the process $(\mathsf{X}_t)_{t\geq 0}$ indeed solves (C.67). Since f is 1-Lipschitz, we obtain

$$\mathcal{W}_1(\mu, \mathsf{X}_t) \leq \mathbb{E}\big[|f(\mathsf{B}_t + \mathsf{Y}) - \mathsf{Y}|\big] = \mathbb{E}\big[|f(\mathsf{B}_t + \mathsf{Y}) - f(\mathsf{Y})|\big] \leq \mathbb{E}[|\mathsf{B}_t|] \leq \sqrt{D\,t},$$

where the final inequality follows from the Cauchy–Schwarz inequality. ☐

We next address the error arising from initialising the backward dynamics in the stationary distribution rather than in $p_{\overline{T}}$. This can be bounded using uniform ergodicity of reflected Brownian motions in bounded convex domains, which is analysed in detail in [79].

Lemma C.55: Let $0 < \underline{T} \leq \overline{T}$, and suppose that s is such that (C.70) admits a unique strong solution on $[0, \overline{T} - \underline{T}]$ for any initial distribution. Let $(\overline{\mathsf{X}}^s_t)_{t\in[0,\overline{T}]}$ and $(\widehat{\mathsf{X}}^s_t)_{t\in[0,\overline{T}]}$ denote such solutions, with $\overline{\mathsf{X}}^s_0 \sim \mathcal{U}[0,1]^D$ and $\widehat{\mathsf{X}}^s_0 \sim p_{\overline{T}}$, respectively. Then,

$$\mathcal{W}_1\big(\widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}, \overline{\mathsf{X}}^s_{\overline{T}-\underline{T}}\big) \leq \frac{8\sqrt{D}}{\pi} \exp\Big(-\frac{\pi^2 \overline{T}}{2D}\Big).$$

Remark C.56: For existence and uniqueness of strong solutions, it suffices that for each $t \in [0, \overline{T} - \underline{T}]$ the map $x \mapsto s(x,t)$ is Lipschitz continuous with a Lipschitz constant independent of t, cf. [117]. This condition is satisfied by all neural network score approximations s considered in this paper.

Proof of Lemma C.55. We begin by recalling that, for any two probability measures ν, ν' on $[0,1]^D$,

$$\mathcal{W}_1(\nu, \nu') \leq 2\,\mathrm{diam}([0,1]^D)\ \mathrm{TV}(\nu, \nu') = 2\sqrt{D}\ \mathrm{TV}(\nu, \nu'). \tag{C.73}$$

Let $(Q_{0,t})_{t\geq 0}$ denote the transition kernels of the (possibly time-inhomogeneous) SDE (C.70), and for any probability measure ν define $\nu Q_{0,t}(\mathrm{d}x) := \int_{[0,1]^D} Q_{0,t}(y, \mathrm{d}x)\,\nu(\mathrm{d}y)$. Writing $\mu_{\overline{T}}(\mathrm{d}x) =$

$p_{\overline{T}}(x)\,\mathrm{d}x$ and letting ρ denote the uniform distribution on $([0,1]^D, \mathcal{B}([0,1]^D))$, we have $\widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}} \sim \mu_{\overline{T}} Q_{0,\overline{T}-\underline{T}}$, while $\overline{\mathsf{X}}^s_{\overline{T}-\underline{T}} \sim \rho Q_{0,\overline{T}-\underline{T}}$. Since $(Q_{0,t})_{t\geq 0}$ is a contraction semigroup, it follows that

$$\begin{aligned}
\mathcal{W}_1\big(\widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}, \overline{\mathsf{X}}^s_{\overline{T}-\underline{T}}\big) &\leq 2\sqrt{D}\ \mathrm{TV}\big(\mu_{\overline{T}} Q_{0,\overline{T}-\underline{T}}, \rho Q_{0,\overline{T}-\underline{T}}\big) \\
&\leq 2\sqrt{D}\ \mathrm{TV}(\mu_{\overline{T}}, \rho) \\
&= 2\sqrt{D} \sup_{A\in\mathcal{B}([0,1]^D)} \Big| \int_{[0,1]^D} \int_A (p_{\overline{T}}(x,y) - 1)\,\mathrm{d}y\, \mu(\mathrm{d}x) \Big| \\
&\leq 2\sqrt{D} \sup_{x\in[0,1]^D} \sup_{A\in\mathcal{B}([0,1]^D)} \Big| \int_A (p_{\overline{T}}(x,y) - 1)\,\mathrm{d}y \Big| \\
&= 2\sqrt{D} \sup_{x\in[0,1]^D} \mathrm{TV}(q_{\overline{T}}(x,\cdot), \rho).
\end{aligned}$$

The result now follows from [79, Theorem 4], which states that

$$\sup_{x\in[0,1]^D} \mathrm{TV}(q_{\overline{T}}(x,\cdot), \rho) \leq \frac{4}{\pi} \exp\Big(- \frac{\pi^2 \overline{T}}{2D} \Big).$$

□

We now turn to the second term in (C.72) and follow the general strategy from [89] to control it. We start by decomposing the time interval $[0, \overline{T} - \underline{T}]$ into a sequence of geometrically shrinking sub-intervals and introduce, on each such sub-interval, an auxiliary process in which the true score is only partially replaced by its approximation. This multilevel construction allows us to localise the score approximation error in time and to derive sharper bounds on the resulting Wasserstein distance. Fix a constant $c \in (1,2]$, and choose $K \in \mathbb{N}$ such that $\underline{T} c^K = \overline{T}$. Define the intermediate times $t_i := \underline{T} c^i$, $i = 0, \dots, K$. For each $i \in \{0, \dots, K\}$ and a given score approximation s, let $\mathsf{Y}^{(i)} = (\mathsf{Y}^{(i)}_t)_{t\in[0,\overline{T}-\underline{T}]}$ denote the solution to the reflected SDE

$$\begin{aligned}
\mathrm{d}\mathsf{Y}^{(i)}_t &= \nabla \log p_{\overline{T}-t}\big(\mathsf{Y}^{(i)}_t\big)\,\mathrm{d}t + \mathrm{d}\mathsf{B}_t + n\big(\mathsf{Y}^{(i)}_t\big)\,\mathrm{d}\mathsf{L}_t, \quad && t \in [0, \overline{T} - t_i), \\
\mathrm{d}\mathsf{Y}^{(i)}_t &= s\big(\mathsf{Y}^{(i)}_t, \overline{T} - t\big)\,\mathrm{d}t + \mathrm{d}\mathsf{B}_t + n\big(\mathsf{Y}^{(i)}_t\big)\,\mathrm{d}\mathsf{L}_t, && t \in [\overline{T} - t_i, \overline{T} - \underline{T}],
\end{aligned}$$

with initial condition $\mathsf{Y}^{(i)}_0 \sim p_{\overline{T}}$. Thus, the process $\mathsf{Y}^{(i)}$ follows the exact reverse-time dynamics driven by the true score up to time $\overline{T} - t_i$, and subsequently evolves according to the approximate score s. By construction, we have the distributional identities

$$\mathsf{X}_{\underline{T}} \sim \widehat{\mathsf{X}}^{s_0}_{\overline{T}-\underline{T}} \sim \mathsf{Y}^{(0)}_{\overline{T}-\underline{T}}, \qquad \widehat{\mathsf{X}}^{s}_{\overline{T}-\underline{T}} \sim \mathsf{Y}^{(K)}_{\overline{T}-\underline{T}}.$$

Applying the triangle inequality for the 1-Wasserstein distance therefore yields

$$\mathcal{W}_1\big(\mathsf{X}_{\underline{T}}, \widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}\big) \leq \sum_{i=1}^K \mathcal{W}_1\big(\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \mathsf{Y}^{(i)}_{\overline{T}-\underline{T}}\big).$$

The following proposition provides a bound on each of the incremental Wasserstein distances, which essentially improves by a factor of $((t_i \wedge 1)\rho)^{1/2}$ the rough upper bound that can be derived from combining a total variation bound and Girsanov's theorem, provided that the score approximation satisfies $|s(x,t)| \leq C\sqrt{\rho/t}$ for $t \leq 1$. This growth control is motivated by Lemma C.53, whose combined conclusion tells us that for any $p \in \mathbb{N}$, with probability at least $1 - 1/n^p$, the true score satisfies

$$\forall t \geq \underline{T} \,:\, |s_0(t, \mathsf{X}_t)| = |\nabla \log p_t(\mathsf{X}_t)| \lesssim t^{-1/2}\sqrt{\log(1 + \log \underline{T}^{-1}) + \sqrt{p \log n}}.$$

The improved Wasserstein bound allows to compensate the higher difficulty of score approximation for small times t caused by its increasing irregularity as $t \to 0$ and thereby obtain faster Wasserstein convergence rates. The detailed proof is postponed to Appendix C.B.

Proposition C.57: Assume ($\mathcal{H}4$). Let s be a score approximation satisfying $|s(x,t)| \leq C\sqrt{\rho/t}$ for all $x \in [0,1]^D$, $t \in (0,1]$, for some constants $C > 0$ and $\rho > 1$. Assume moreover that $t_1 \leq 1$ and that $\log\big(1 + \log(t_1^{-1})\big) \lesssim \sqrt{\rho}$. Then, for any $i = 1, \dots, K$, the corresponding processes $\mathsf{Y}^{(i-1)}$ and $\mathsf{Y}^{(i)}$ satisfy

$$\mathcal{W}_1\big(\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \mathsf{Y}^{(i)}_{\overline{T}-\underline{T}}\big) \leq \mathfrak{C}\left(\mathrm{e}^{-\rho} + \left((t_i \wedge 1)\rho \int_{t_{i-1}}^{t_i} \mathbb{E}\big[|s(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2\big]\,\mathrm{d}t\right)^{\frac{1}{2}}\right),$$

for some constant $\mathfrak{C} > 0$ independent of i. In particular,

$$\mathcal{W}_1\big(\mathsf{X}_{\underline{T}}, \widehat{\mathsf{X}}^s_{\overline{T}-\underline{T}}\big) \leq \mathfrak{C}\left(K\mathrm{e}^{-\rho} + \sqrt{\rho}\sum_{i=1}^{K}\sqrt{t_i \wedge 1}\left(\int_{t_{i-1}}^{t_i} \mathbb{E}\big[|s(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2\big]\,\mathrm{d}t\right)^{\frac{1}{2}}\right)$$

.

Recall that, for given $\underline{T}, \overline{T}$ and $t_i = \underline{T}c^i$, $i = 0, \dots, K$, as above such that $c^K = \overline{T}$, we estimate the score on on $[t_{i-1}, t_i)$ by minimising the empirical denoising score loss via

$$\widehat{s}^{(i)}_n \in \underset{s \in \mathcal{S}_i}{\operatorname{argmin}} \frac{1}{n}\sum_{k=1}^{n} L^{(i)}_s(\mathsf{Y}_k), \tag{C.74}$$

where $\mathsf{Y}_1, \dots, \mathsf{Y}_n \overset{iid}{\sim} \mu$ is our given data,

$$L^{(i)}_s(x) = \mathbb{E}\Big[\int_{t_{i-1}}^{t_i} |s(\mathsf{X}_t, t) - \nabla \log q_t(x, \mathsf{X}_t)|^2\,\mathrm{d}t\Big],$$

and $\mathcal{S}_i$ is an approximating class of neural networks that needs to be chosen. The full score estimator is then obtained by concatenating these minimisers across time,

$$\widehat{s}_n(x,t) = \sum_{i=1}^{K} \widehat{s}^{(i)}_n(x,t)\mathbf{1}_{[t_{i-1}, t_i)}(t), \tag{C.75}$$

which reflects the multiscale structure of the diffusion and allows the approximation complexity to adapt to the effective noise level at time t. By the equivalence of explicit and denoising score matching, $\widehat{s}_n^{(i)}$ therefore serves as an empirical risk minimiser for the true score s_0 on $[t_{i-1}, t_i)$. For a given approximation class $\mathcal{S}_i$, the L^2 estimation error

$$\mathbb{E}\Big[\int_{t_{i-1}}^{t_i} |\widehat{s}_n^{(i)}(\mathrm{X}_t, t) - \nabla \log q_t(x, \mathrm{X}_t)|^2 \,\mathrm{d}t\Big]$$

therefore naturally splits into the conflicting effects of an approximation error

$$\min_{s\in\mathcal{S}_i} \mathbb{E}\Big[\int_{t_{i-1}}^{t_i} |s(t, \mathrm{X}_t) - \nabla \log q_t(x, \mathrm{X}_t)|^2 \,\mathrm{d}t\Big],$$

which decreases with larger networks sizes that increase the expressivity of the approximation class, and a complexity term that increases with the size of the network class.

Score approximation

In order to optimally balance these two effects, given a desired target accuracy, it is necessary to make parsimonious choices regarding the network sizes. To specify this, we now introduce the class of sparsity-constrained neural networks with ReLU activation function that we use for score approximation. For $b, x \in \mathbb{R}^m$, define

$$\sigma_b(x) = \begin{bmatrix} \sigma(x_1 - b_1) \\ \sigma(x_2 - b_2) \\ \vdots \\ \sigma(x_m - b_m) \end{bmatrix}, \quad \sigma(y) = y \vee 0,$$

and for $L \in \mathbb{N}$, $W \in \mathbb{N}^{L+2}$, $S \in \mathbb{N}$ and $B > 0$ denote by $\Phi(L, W, S, B)$ the class of neural networks with depth (i.e., number of hidden layers) L, layer widths (including input and output layers) W, sparsity constraint S, and norm constraint B. We thus 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 $i = 0, \dots, L$ (to ease notation, we always set $b_0 = 0$), 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. In an abuse of notation, we denote σ_0 simply by σ. This can be written succinctly as

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

For larger and more complicated neural networks, their exact sizes are often unavailable, and we only have access to their asymptotic sizes. Due to this, we also introduce the following class of neural networks that eases network size analysis in the proofs that follow:

$$\widetilde{\Phi}(\widetilde{L}, \widetilde{W}, \widetilde{S}, \widetilde{B}) := \Big\{ \varphi \in \Phi(L, W, S, B) \,:\, L \lesssim \widetilde{L}, \|W\|_\infty \lesssim \widetilde{W}, S \lesssim \widetilde{S} \text{ and } B \lesssim \widetilde{B} \Big\}.$$

With this notation, we have for arbitrary networks $\varphi_i \in \widetilde{\Phi}(L_i, W_i, S_i, \mathsf{B}_i)$, that

$$\begin{aligned} \varphi_1 \circ \varphi_2 &\in \widetilde{\Phi}(L_1 + L_2, W_1 \vee W_2, S_1 + S_2, \mathsf{B}_1 \vee \mathsf{B}_2) \qquad \text{and} \\ \begin{bmatrix} \varphi_1 \\ \varphi_2 \end{bmatrix} &\in \widetilde{\Phi}(L_1 \vee L_2, W_1 + W_2, S_1 + S_2, \mathsf{B}_1 \vee \mathsf{B}_2). \end{aligned}$$

In particular, since $\varphi_1 + \varphi_2 = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \varphi_1 & \varphi_2 \end{bmatrix}^\top$, we have also

$$\sum_{i=1}^k \varphi_i \in \widetilde{\Phi}\Big(\max\{L_i\}, \sum_{i=1}^k W_i, \sum_{i=1}^k S_i, \max\{\mathsf{B}_i\} \Big).$$

Some basic neural network approximation results that we shall frequently use in our analysis are given in Appendix C.C. Our main approximation result is the following.

Theorem C.58: Under assumptions ($\mathcal{H}1$)–($\mathcal{H}4$), for any $\delta > 0$, large enough $m \in \mathbb{N}$ and $\underline{t} > 0$ with $m^{-\frac{2\alpha+2}{2\alpha+d}} \lesssim \underline{t} \lesssim \log m$, there exists a neural network

$$\varphi_{s_0} \in \begin{cases} \widetilde{\Phi}\big((\log m)^2(\log\log m)^2, m(\log m)^{D+1}, m(\log m)^{D+2}, m^{\frac{\alpha}{d}}\underline{t}^{-1} \vee m^{\nu}\big), & \text{if } \underline{t} \le \frac{1}{2}m^{-\frac{2-\delta}{d}} \\ \widetilde{\Phi}\big((\log m)^2(\log\log m)^2, m'(\log m)^{D+1}, m'(\log m)^{D+2}, m'\big), & \text{if } \underline{t} > \frac{1}{2}m^{-\frac{2-\delta}{d}} \end{cases},$$

where $\nu = \frac{2d}{2\alpha - d} + \frac{1}{d}$ and $m' = \underline{t}^{-\frac{d}{2}} m^{\frac{\delta}{2}}$ satisfying

$$\int_{\underline{t}}^{2\underline{t}} \mathbb{E}[|s_0(\mathsf{X}_t) - \varphi_{s_0}(\mathsf{X}_t)|^2] \,\mathrm{d}t \lesssim \begin{cases} (\log m)^{d+2D+3} m^{-\frac{2\alpha}{d}}, & \text{if } \underline{t} \le m^{-\frac{2-\delta}{d}} \\ (\log m)^{d+2D+3} m^{-\frac{2(\alpha+1)}{d}}, & \text{if } \underline{t} > m^{-\frac{2-\delta}{d}}. \end{cases}$$

Moreover, this network can be chosen such that $|s_0(x, t)| \lesssim \frac{\sqrt{\log m}}{\sqrt{\underline{t} \wedge 1}}$ for all $x \in [0, 1]^D$ and $t \in [\underline{t}, 2\underline{t}]$.

The proof is technically involved and proceeds through several stages. We provide a high-level overview of the argument here, while all details are deferred to Section C.B. The approximation strategy exploits the explicit score representation established in Lemma C.51, together with the general neural network approximation framework for space-time functions developed in [58].

Recall from Assumptions ($\mathcal{H}1$) and ($\mathcal{H}2$) that the target distribution μ is supported on a closed subset M of a d-dimensional affine subspace $(V + v_0) \cap [0, 1]^D$ with non-empty interior, where $V = \operatorname{Span}(v_1, \dots, v_d)$, $v_0 \in [0, 1]^D$, and $v_1, \dots, v_d \in \mathbb{R}^D$ are vectors in the D-dimensional ambient space. Let $A := (v_1, \dots, v_d) \in \mathbb{R}^{D \times d}$ and $P := AA^\top$, so that P is the orthogonal projection onto V and any $x \in V$ can be written as $x = Au$ for some $u \in \mathbb{R}^d$. Then, for any integrable function $g : M \to \mathbb{R}$,

$$\int_M g \,\mathrm{d}\mu = \int_{M^*} g(Au + v_0) p_0(Au + v_0) \,\mathrm{d}u,$$

where $M^* := A^\top(M - v_0) = \{u \in \mathbb{R}^d \,:\, Au + v_0 \in M\}$. Moreover, for any $x \in \mathbb{R}^D$ and $u \in M^*$,

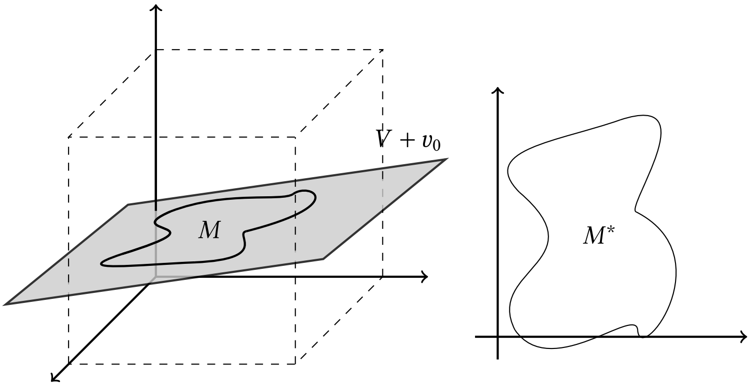

Figure C.13: *Example of a domain $M \subseteq \mathbb{R}^3$ and its lower-dimensional representation $M^* \subseteq \mathbb{R}^2$.*

$$\begin{aligned} x - (Au + v_0) &= \big((v_0 + P(x - v_0)) - (Au + v_0)\big) + \big(x - (v_0 + P(x - v_0))\big) \\ &= \big(P(x - v_0) - Au\big) + \big((I - P)(x - v_0)\big), \end{aligned}$$

where the first term lies in V and the second in $V^\perp$. By the Pythagorean theorem,

$$|x - (Au + v_0)|^2 = |P(x - v_0) - Au|^2 + |(I - P)(x - v_0)|^2.$$

Consequently, for any $x \in \mathbb{R}^D$,

$$\begin{aligned} \int_M \mathrm{e}^{-\frac{|x-y|^2}{2t}} \,\mu(\mathrm{d}y) &= \int_{M^*} \mathrm{e}^{-\frac{|x-(Au+v_0)|^2}{2t}} p_0(Au + v_0) \,\mathrm{d}u \\ &= \mathrm{e}^{-\frac{|(I-P)(x-v_0)|^2}{2t}} \int_{M^*} \mathrm{e}^{-\frac{|P(x-v_0)-Au|^2}{2t}} p_0(Au + v_0) \,\mathrm{d}u \\ &= \mathrm{e}^{-\frac{|(I-P)(x-v_0)|^2}{2t}} \int_{M^*} \mathrm{e}^{-\frac{|A^\top(x-v_0)-u|^2}{2t}} p_0(Au + v_0) \,\mathrm{d}u. \end{aligned}$$

Here we used that $|Au| = |u|$ for all $u \in \mathbb{R}^d$. A similar decomposition yields

$$\begin{aligned} \int_M \frac{x - y}{t} \mathrm{e}^{-\frac{|x-y|^2}{2t}} \,\mu(\mathrm{d}y) = \mathrm{e}^{-\frac{|(I-P)(x-v_0)|^2}{2t}} \Bigg(&\frac{(I - P)(x - v_0)}{t} \int_{M^*} \mathrm{e}^{-\frac{|A^\top(x-v_0)-u|^2}{2t}} p_0(Au + v_0) \,\mathrm{d}u \\ &+ A \int_{M^*} \frac{A^\top(x - v_0) - u}{t} \mathrm{e}^{-\frac{|A^\top(x-v_0)-u|^2}{2t}} p_0(Au + v_0) \,\mathrm{d}u \Bigg). \end{aligned}$$

In view of Lemma C.51, which gives

$$s_0(x,t) = -\frac{\sum_{z\in\mathbb{Z}^D}(-1)^z \int_{[0,1]^D}(R_z(x)+z-y)\exp\left(-\frac{|R_z(x)+z-y|^2}{2t}\right)\mu(\mathrm{d}y)}{t\sum_{z\in\mathbb{Z}^D}\int_{[0,1]^D}\exp\left(-\frac{|R_z(x)+z-y|^2}{2t}\right)\mu(\mathrm{d}y)}, \quad (x,t)\in[0,1]^D\times(0,\infty),$$

all dependence on μ enters through the lower-dimensional projection $A^\top(x-v_0)$. Accordingly, a substantial part of the approximation task reduces to approximating the functions

$$f_1 : (u,t) \mapsto \int_{M^*} \mathrm{e}^{-\frac{|u-v|^2}{2t}} p_0(Av+v_0)\,\mathrm{d}v, \qquad f_2 : (u,t) \mapsto \int_{M^*} \frac{u-v}{t}\mathrm{e}^{-\frac{|u-v|^2}{2t}} p_0(Av+v_0)\,\mathrm{d}v, \quad \text{(C.76)}$$

defined on $\mathbb{R}^d\times(0,\infty)$. This dimensional reduction allows us to derive error bounds that depend on the intrinsic dimension d rather than the ambient dimension D. A comparable mechanism appears in [89], where the Gaussian transition densities of the OU forward process give rise to analogous expressions. In contrast, the transition densities in the present model are given by infinite series of Gaussian densities restricted to $[0,1]^D$, which introduces substantial additional technical difficulties. These are addressed using an approximation strategy adapted from [58], where spectral representations of the forward density and its gradient were analysed.

The proof proceeds along the following steps:

1. For fixed $\underline{t} > 0$ and $t \in [\underline{t}, 2\underline{t}]$, we truncate the series representation of the forward density by

   $$p_t^K(x) = (2\pi t)^{-\frac{D}{2}} \sum_{\substack{z\in\mathbb{Z}^D \\ \|z\|_\infty \le \sqrt{2\underline{t}(D+2K)}}} \int_{[0,1]^D} \exp\left(-\frac{|R_z(x)+z-y|^2}{2t}\right)\mu(\mathrm{d}y)$$

   and define the corresponding truncated score

   $$s_0^K(x,t) := \frac{\nabla p_t^K(x)}{p_t^K(x)} \quad \text{(C.77)}$$

   Lemma C.60 establishes for $t\in[\underline{t},2\underline{t}]$ an exponential convergence of s_0^K to s_0 in L^2 as $K\to\infty$ allowing us to restrict attention to truncation levels of order $\log n$. Since the truncated sums can be approximated termwise by neural networks, combining these approximations increases the network depth only by a logarithmic factor, leading to a negligible additional error.

2. For fixed $\delta > 0$, we split the time domain into small- and large-time regimes according to $t \lesseqgtr m^{-(2-\delta)/d}$. Using Lemma C.69, we approximate $f_1(\cdot,t)$ and $f_2(\cdot,t)$ in (C.76) for fixed t; see Lemmas C.61 and C.62. In the small-time regime, Assumption ($\mathcal{H}3$) yields improved approximation rates of order $m^{-\kappa/d}$ outside M^*, compared to $m^{-\alpha/d}$ on M^* (up to logarithmic factors), while in the large-time regime the regularising effect of the forward diffusion leads to rates of order $m^{-(\kappa+1)/d}$.

3. We extend these fixed-time approximations to short time intervals using polynomial interpolation in t, as shown in Lemma C.63.

4. Finally, we combine the resulting constructions with the general neural network approximation results from Appendix C.C to prove Theorem C.58, which provides L^2 approximation bounds for the score in both time regimes.

Main result

With these preparations, we are now in a position to give a concise proof our main result.

Theorem C.59: Assume $(\mathcal{H}1)$–$(\mathcal{H}3)$, and set

$$\underline{T} = D^{-1} n^{-\frac{2(\alpha+1)}{2\alpha+d}} \quad \text{and} \quad \overline{T} = \frac{8}{\pi^2} \log\left(\frac{8D^{\frac{3}{2}}}{\pi} n^{\frac{\alpha+1}{2\alpha+d}} \right).$$

Then, for any $\delta > 0$ and $n \in \mathbb{N}$ large enough, there exist neural network classes

$$\mathcal{S}_i = \left\{ \varphi \in \Phi(L, W_i, S_i, B) : \ |\varphi(x,t)| \lesssim \frac{\sqrt{\log n}}{\sqrt{t_i \wedge 1}} \right\},$$

where

$$\begin{aligned} L &\lesssim \log n \log\log n, \\ \|W_i\|_\infty &\lesssim \left(n^{\frac{d}{2\alpha+d}} \wedge \left[(t_i \wedge 1)^{-\frac{d}{2}} n^{\frac{\delta d}{2\alpha+d}} \right] \right) (\log n)^{D+1}, \\ S_i &\lesssim \left(n^{\frac{d}{2\alpha+d}} \wedge \left[(t_i \wedge 1)^{-\frac{d}{2}} n^{\frac{\delta d}{2\alpha+d}} \right] \right) (\log n)^{D+2}, \\ B &\lesssim n^{\frac{4(\alpha+1)+d(c_0-d+2)}{2(2\alpha+d)}} \vee n^{\frac{2d^2}{4\alpha^2-d^2}+\frac{1}{2\alpha+d}}, \end{aligned}$$

such that the reflected diffusion generative algorithm associated to the empirical denoising score matching loss minimiser $\widehat{s}_n$ defined via (C.74) and (C.75) satisfies

$$\mathbb{E}\left[\mathcal{W}_1(\mu, \overline{\mathsf{X}}^{\widehat{s}_n}_{\overline{T}-\underline{T}}) \right] \lesssim n^{-\frac{\alpha+1-\delta}{2\alpha+d}}.$$

Proof. First, recalling the decomposition (C.72), it follows that

$$\mathbb{E}\left[\mathcal{W}_1(\mu, \overline{\mathsf{X}}^{\widehat{s}_n}_{\overline{T}-\underline{T}}) \right] \leq \mathcal{W}_1(\mu, \mathsf{X}_{\underline{T}}) + \mathbb{E}[\mathcal{W}_1(\widehat{\mathsf{X}}^{\widehat{s}_n}_{\overline{T}-\underline{T}}, \overline{\mathsf{X}}^{\widehat{s}_n}_{\overline{T}-\underline{T}})] + \mathbb{E}[\mathcal{W}_1(\mathsf{X}_{\underline{T}}, \widehat{\mathsf{X}}^{\widehat{s}_n}_{\overline{T}-\underline{T}})].$$

Here, it immediately follows by Lemmas C.54 and C.55 that the first two terms are each bounded by $n^{-\frac{\alpha+1}{2\alpha+d}}$, and so we now focus on bounding $\mathbb{E}[\mathcal{W}_1(\mathsf{X}_{\underline{T}}, \widehat{\mathsf{X}}^{\widehat{s}_n}_{\overline{T}-\underline{T}})]$. By Lemma C.57, and its preceding discussion, we have for $\rho > 0$ that

$$\begin{aligned}\mathbb{E}[\mathcal{W}_1(\mathsf{X}_{\underline{T}}, \widehat{\mathsf{X}}^{\widehat{s}_n}_{\overline{T}-\underline{T}})] &\lesssim K\mathrm{e}^{-\rho} + \sum_{i=1}^{K} \mathbb{E}\left[\left((t_i \wedge 1)\rho \int_{t_{i-1}}^{t_i} \mathbb{E}\left[|\widehat{s}_n(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2 \mid \widehat{s}_n\right] \mathrm{d}t\right)^{\frac{1}{2}}\right] \\ &\leq K\mathrm{e}^{-\rho} + \sum_{i=1}^{K} \left((t_i \wedge 1)\rho\, \mathbb{E}\left[\int_{t_{i-1}}^{t_i} \mathbb{E}\left[|\widehat{s}_n(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2 \mid \widehat{s}_n\right] \mathrm{d}t\right]\right)^{\frac{1}{2}},\end{aligned}$$

where in the last inequality we use Jensen's inequality. Recall here that $t_i = \underline{T}c^i$, where $c \in (1, 2]$ and $K \in \mathbb{N}$ are chosen such that $t_K = \overline{T}$, i.e. $K = \log_c \frac{\overline{T}}{\underline{T}} \asymp \log n$. Setting $\rho = \frac{\alpha+1}{2\alpha+d} \log n$, we thus have

$$K\mathrm{e}^{-\rho} \asymp n^{-\frac{\alpha+1}{2\alpha+d}} \log n \lesssim n^{-\frac{\alpha+1-\delta}{2\alpha+d}}.$$

Now, to further analyse each term, we fix $i \in [K]$ and introduce the induced function class $\mathcal{L}^{(i)} = \{L_s^{(i)} \mid s \in \mathcal{S}_i\}$. We then have by [58, Theorem 3.4, Theorem B.2] if $\sup_{s \in \mathcal{S}_i \cup \{s_0\}} \|L_s^{(i)}\|_\infty \leq C(\mathcal{L}^{(i)}) < \infty$ that for suitable $\Delta > 0$,

$$\begin{aligned}&\mathbb{E}\left[\int_{t_{i-1}}^{t_i} \mathbb{E}\left[|\widehat{s}_n(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2 \mid \widehat{s}_n\right] \mathrm{d}t\right] \\ &\quad\leq 2 \inf_{s \in \mathcal{S}_i} \int_{t_{i-1}}^{t_i} \mathbb{E}\left[|s(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2\right] \mathrm{d}t + 2\frac{C(\mathcal{L}^{(i)})}{n}\left(\frac{145}{9} \log \mathcal{N}(\mathcal{L}^{(i)}, \|\cdot\|_\infty, \Delta) + 160\right) + 5\Delta.\end{aligned}$$

Following the proof of [58, Lemma 3.5], we also have that if for $t \in [t_{i-1}, t_i)$, $\sup_{s \in \mathcal{S}_i} \|s(\cdot, t)\|_\infty \leq \frac{C(\mathcal{S}_i)}{\sqrt{t \wedge 1}}$ for some $C(\mathcal{S}_i) < \infty$ and

$$\int_{t_{i-1}}^{t_i} \int_{[0,1]^D} |\nabla \log q_t(x, y)| q_t(x, y)\, \mathrm{d}y\, \mathrm{d}t \lesssim \sqrt{t_i}, \tag{C.78}$$

then there exists a constant c_0 such that

$$\mathcal{N}(\mathcal{L}^{(i)}, \|\cdot\|_\infty, \delta) \leq \mathcal{N}\left(\mathcal{S}_i, \|\cdot\|_\infty, \frac{\Delta}{c_0 C(\mathcal{S}_i)(t_i + \sqrt{t_i})}\right).$$

To this last condition, we have

$$\begin{aligned}\int_{[0,1]^D} |\nabla \log q_t(x, y)| q_t(x, y)\, \mathrm{d}y &= \int_{[0,1]^D} |\nabla q_t(x, y)|\, \mathrm{d}y \\ &\leq \frac{1}{t} \sum_{z \in \mathbb{Z}^D} \int_{[0,1]^D} (2\pi t)^{-\frac{D}{2}} |R_z(x) + z - y| \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}\, \mathrm{d}y \\ &= \frac{1}{t} \int_{\mathbb{R}^D} (2\pi t)^{-\frac{D}{2}} |y| \mathrm{e}^{-\frac{|y|^2}{2t}}\, \mathrm{d}y \\ &\leq \frac{\sqrt{D}}{\sqrt{t}},\end{aligned}$$

implying (C.78). Furthermore, using the Li–Yau bound from [74, Theorem 1.1], which gives $|\nabla \log q_t(x,y)|^2 \lesssim \frac{D}{t} + \partial_t \log q_t(x,y)$, it follows that for $s \in \mathcal{S}_i$

$$\begin{aligned}
\mathsf{L}_s(x) &= \int_{t_{i-1}}^{t_i} \int_{[0,1]^D} |s(y,t) - \nabla \log q_t(x,y)|^2 q_t(x,y) \, \mathrm{d}y \, \mathrm{d}t \\
&\leq 2 \int_{t_{i-1}}^{t_i} \int_{[0,1]^D} \big(|s(y,t)|^2 + |\nabla \log q_t(x,y)|^2 \big) q_t(x,y) \, \mathrm{d}y \, \mathrm{d}t \\
&\lesssim 2 \int_{t_{i-1}}^{t_i} \int_{[0,1]^D} \Big(\frac{C(\mathcal{S}_i)^2 + D}{t} + \partial_t \log q_t(x,y) \Big) q_t(x,y) \, \mathrm{d}y \, \mathrm{d}t \\
&= 2\big(C(\mathcal{S}_i)^2 + D\big) \log c,
\end{aligned}$$

where we used that

$$\begin{aligned}
\int_{t_{i-1}}^{t_i} \int_{[0,1]^D} \partial_t \log q_t(x,y) q_t(x,y) \, \mathrm{d}y \, \mathrm{d}t &= \int_{[0,1]^D} \int_{t_{i-1}}^{t_i} \partial_t q_t(x,y) \, \mathrm{d}t \, \mathrm{d}y \\
&= \int_{[0,1]^D} q_{t_i}(x,y) - q_{t_{i-1}}(x,y) \, \mathrm{d}y = 0.
\end{aligned}$$

Thus, it follows that $C(\mathcal{L}^{(i)}) \lesssim (C(\mathcal{S}_i)^2 + D) \log c \lesssim C(\mathcal{S}_i)^2$, and hence by the above

$$\begin{aligned}
&\mathbb{E}\Big[\int_{t_{i-1}}^{t_i} \mathbb{E}\big[|\widehat{s}_n(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2 \mid \widehat{s}_n \big] \, \mathrm{d}t \Big] \\
&\quad \lesssim \inf_{s \in \mathcal{S}_i} \int_{t_{i-1}}^{t_i} \mathbb{E}\big[|s(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2 \big] \, \mathrm{d}t + \frac{C(\mathcal{S}_i)^2 \log \mathcal{N}(\mathcal{S}_i, \|\cdot\|_\infty, \frac{\Delta}{c_0 C(\mathcal{S}_i)(t_i + \sqrt{t_i})})}{n} + \Delta.
\end{aligned}$$

Here, by a small modification of [89, Lemma C.2], if $\mathcal{S}_i \subseteq \Phi(L, W, S, B)$, then

$$\log \mathcal{N}\Big(\mathcal{S}_i, \|\cdot\|_\infty, \frac{\Delta}{c_0 C(\mathcal{S}_i)(t_i + \sqrt{t_i})} \Big) \lesssim LS \log(c_0 C(\mathcal{S}_i)(t_i + \sqrt{t_i}) \Delta^{-1} L \|W\|_\infty (B \vee 1)).$$

Next, setting $m = \lceil n^{\frac{d}{2\alpha+d}} \rceil$, if $t_i \leq n^{-\frac{2-\delta}{2\alpha+d}}$, we have by Theorem C.58 that there exists a neural network $\varphi_{s_0}^{(i)} \in \Phi(L, W, S, B)$ with

$$\begin{aligned}
L &\lesssim (\log m)^2 (\log\log m)^2 && \lesssim \mathrm{Poly}(\log n) \\
\|W\|_\infty &\lesssim m (\log m)^{D+1} && \lesssim \mathrm{Poly}(n) \\
S &\lesssim m (\log m)^{D+2} && \lesssim n^{\frac{d}{2\alpha+d}} \,\mathrm{Poly}(\log n) \\
B &\lesssim m^{\frac{2(\alpha+1)}{d} + \frac{c_0 - d}{2} + 1} \vee m^{\nu} && \lesssim \mathrm{Poly}(n),
\end{aligned}$$

satisfying $|\varphi_{s_0}^{(i)}| \lesssim \frac{\sqrt{\log n}}{\sqrt{t_i \wedge 1}} \leq \frac{\sqrt{\log n}}{\sqrt{t \wedge 1}}$ for $t \in [t_{i-1}, t_i)$ and

$$\int_{t_{i-1}}^{t_i} \mathbb{E}\big[|\varphi_{s_0}^{(i)}(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2 \big] \, \mathrm{d}t \lesssim (\log m)^{2+2D+3} m^{-\frac{2\alpha}{d}} \leq n^{-\frac{2\alpha}{2\alpha+d}} \,\mathrm{Poly}(\log n)$$

Thus, setting $\mathcal{S}_i = \{\varphi \in \Phi(L, W, S, B) \mid |\varphi| \lesssim \frac{\sqrt{\log n}}{\sqrt{t_{i-1}}}\}$ and $\Delta = n^{-\frac{2\alpha}{2\alpha+d}}$, it follows that

$$\mathbb{E}\left[\int_{t_{i-1}}^{t_i} \mathbb{E}\left[|\widehat{s}_n(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2 \mid \widehat{s}_n\right] \mathrm{d}t\right] \lesssim n^{-\frac{2\alpha}{2\alpha+d}} \operatorname{Poly}(\log n).$$

Conversely, if $t_i > n^{-\frac{2-\delta}{2\alpha+d}}$, the same theorem yields a network $\varphi_{s_0}^{(i)} \in \Phi(L, W_i, S_i, B)$ with

$$\begin{aligned}
L &\lesssim (\log m)^2 (\log\log m)^2 && \lesssim \operatorname{Poly}(\log n)\\
\|W_i\|_\infty &\lesssim (t_i \wedge 1)^{-\frac{d}{2}} m^{\frac{\delta}{2}} (\log m)^{D+1} && \lesssim \operatorname{Poly}(n)\\
S_i &\lesssim (t_i \wedge 1)^{-\frac{d}{2}} m^{\frac{\delta}{2}} (\log m)^{D+2} && \lesssim (t_i \wedge 1)^{-\frac{d}{2}} n^{\frac{\delta d}{2(2\alpha+d)}} \operatorname{Poly}(\log n)\\
\mathsf{B}_i &\lesssim m^{\frac{2(\alpha+1)}{d} + \frac{c_0 - d}{2} + 1} && \lesssim \operatorname{Poly}(n),
\end{aligned}$$

satisfying

$$\int_{t_{i-1}}^{t_i} \mathbb{E}\left[|\varphi_{s_0}^{(i)}(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2\right] \mathrm{d}t \lesssim (\log m)^{2+2D+3} m^{-\frac{2(\alpha+1)}{d}} \leq n^{-\frac{2(\alpha+1)}{2\alpha+d}} \operatorname{Poly}(\log n).$$

Thus, setting $\Delta = n^{-\frac{2(\alpha+1)}{2\alpha+d}}$, we have now

$$\mathbb{E}\left[\int_{t_{i-1}}^{t_i} \mathbb{E}\left[|\widehat{s}_n(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2 \mid \widehat{s}_n\right] \mathrm{d}t\right] \lesssim \left(n^{-\frac{2(\alpha+1)}{2\alpha+d}} + \frac{(t_i \wedge 1)^{-\frac{d}{2}} n^{\frac{\delta d}{2(2\alpha+d)}}}{n}\right) \operatorname{Poly}(\log n).$$

Now, letting $K^* = \max\{i \in [K] \,:\, t_i \leq n^{-\frac{2-\delta}{2\alpha+d}}\}$, we have

$$\begin{aligned}
&\sum_{i=1}^{K} \left((t_i \wedge 1)\rho\, \mathbb{E}\left[\int_{t_{i-1}}^{t_i} \mathbb{E}\left[|\widehat{s}_n(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2 \mid \widehat{s}_n\right] \mathrm{d}t\right]\right)^{\frac{1}{2}}\\
&\quad\lesssim \log n \left[\sum_{i=1}^{K^*} \left(t_i\, \mathbb{E}\left[\int_{t_{i-1}}^{t_i} \mathbb{E}\left[|\widehat{s}_n(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2 \mid \widehat{s}_n\right] \mathrm{d}t\right]\right)^{\frac{1}{2}}\right.\\
&\qquad\qquad \left.\sum_{i=K^*+1}^{K} \left((t_i \wedge 1)\, \mathbb{E}\left[\int_{t_{i-1}}^{t_i} \mathbb{E}\left[|\widehat{s}_n(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2 \mid \widehat{s}_n\right] \mathrm{d}t\right]\right)^{\frac{1}{2}}\right].
\end{aligned}$$

Here, the first sum is easily bounded by

$$\sum_{i=1}^{K^*} \left(t_i\, \mathbb{E}\left[\int_{t_{i-1}}^{t_i} \mathbb{E}\left[|\widehat{s}_n(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2 \mid \widehat{s}_n\right] \mathrm{d}t\right]\right)^{\frac{1}{2}} \lesssim n^{-\frac{1-\delta/2}{2\alpha+d}} n^{-\frac{\alpha}{2\alpha+d}} \operatorname{Poly}(\log n) \lesssim n^{-\frac{\alpha+1-\delta}{2\alpha+d}},$$

where we use that $K^* \leq K \lesssim \log n$. As for the second sum, we have

$$\begin{aligned}
&\sum_{i=K^*+1}^{K} \left((t_i \wedge 1) \mathbb{E}\left[\int_{t_{i-1}}^{t_i} \mathbb{E}\left[|\widehat{s}_n(\mathsf{X}_t, t) - \nabla \log p_t(\mathsf{X}_t)|^2 \mid \widehat{s}_n \right] \mathrm{d}t \right] \right)^{\frac{1}{2}} \\
&\lesssim \mathrm{Poly}(\log n) \left(K n^{-\frac{\alpha+1}{2\alpha+d}} + \frac{n^{\frac{\delta d}{4(2\alpha+d)}}}{\sqrt{n}} \sum_{i=K^*+1}^{K} (t_i \wedge 1)^{\frac{2-d}{4}} \right) \\
&\lesssim \mathrm{Poly}(\log n) \left(n^{-\frac{\alpha+1}{2\alpha+d}} + t_{K^*+1}^{\frac{2-d}{4}} \frac{n^{\frac{\delta d}{4(2\alpha+d)}}}{\sqrt{n}} \right) \\
&\lesssim \mathrm{Poly}(\log n) \left(n^{-\frac{\alpha+1}{2\alpha+d}} + n^{\frac{-(2-d)(2-\delta)+\delta d-2(2\alpha+d)}{4(2\alpha+d)}} \right) \\
&\lesssim n^{-\frac{\alpha+1-\delta}{2\alpha+d}}.
\end{aligned}$$

Thus, it follows that

$$\mathbb{E}[\mathcal{W}_1(\mathsf{X}_{\underline{T}}, \widehat{\mathsf{X}}^{\widehat{s}_n}_{\overline{T}-\underline{T}})] \lesssim n^{-\frac{\alpha+1-\delta}{2\alpha+d}},$$

as desired. ☐

C.4 Discussion

We conclude by placing our results in the context of the existing statistical theory of diffusion-based generative models and by clarifying the scope and limitations of the present analysis.

Minimax optimal estimation We do *not* aim to establish results on minimax optimality, as such questions, while important, lie beyond the scope of the present work. The first near-optimal convergence rates for diffusion models under the 1-Wasserstein metric were obtained by [89] who consider target distributions on $\mathbb{R}^D$ that admit an α-smooth, compactly supported density wrt Lebesgue measure, and analyse classical Ornstein–Uhlenbeck (OU) dynamics. Under the assumption of perfect SDE simulation and working without the manifold hypothesis, they show that diffusion models achieve the rate $n^{-\frac{\alpha+1-\delta}{2\alpha+D}}$, for arbitrary $\delta > 0$, which is near minimax-optimal in the ambient dimension D and consistent with results established in the classical i.i.d. density estimation setting by [87]; see in particular Theorem 1 therein for densities bounded away from zero.

In the context of diffusion-based generative models with non-singular target distributions, [113] establish an upper bound of order $n^{-\frac{\alpha+1}{2\alpha+D}}$, up to polylogarithmic factors, for the 1-Wasserstein distance between the true distribution and the law induced by the generative model. Their analysis uses empirical score matching over suitably chosen neural network classes with tanh-activation function and holds uniformly over classes of Hölder-α smooth densities that are bounded below on any compact subset of the interior of the support and exhibit controlled decay near its boundary. The avoidance of an arbitrarily small polynomial inefficiency in the rate comes here at the price of having to consider polynomially many (in terms of the data) separate time intervals in the score estimation procedure as compared to the logarithmic dependence in our and previous work on 1-Wasserstein estimation rates [6, 89, 118]. This theoretical understanding has been further advanced by [21], who provide sharp finite-sample error bounds measured in the p-Wasserstein distance for arbitrary $p \geq 1$. Notably, they relax typical compact-support or smooth density assumptions, requiring only finite-moment conditions on the target distribution.

While these works constitute significant progress in the statistical theory of generative modelling, their analyses fundamentally rely on the explicit Gaussian structure of the transition kernels associated with unconstrained OU or standard Brownian dynamics on $\mathbb{R}^D$. In particular, [113] and [21] exploit a control on the temporal score regularity, which is heavily tied to the specific analytical smoothing properties of the Gaussian transition kernels. Our reflected diffusion framework on $[0,1]^D$ inherently breaks these structural properties. The corresponding transition densities are instead governed by an infinite series expansion of restricted Gaussian densities (or alternatively, by a spectral decomposition as in [58]), which leads to a fundamentally different and analytically more complex score regularity, in particular due to the influence of boundary reflections, which become increasingly notable for larger t. Extending the minimax optimality proofs from the unconstrained setting to bounded domains with reflections would require entirely new analytic techniques to rigorously control these boundary effects and is consequently beyond the scope of this paper.

Furthermore, a strictly minimax-optimal convergence rate, completely free of extraneous logarithmic terms, has recently been derived by [37] in the unconstrained variance-exploding setting. For target distributions with Hölder smoothness $\alpha > 0$, they establish rates wrt the score matching loss, which allows them to prove that the generated distribution achieves the minimax-optimal rate in terms of the expected squared total variation distance and, for $\alpha \geq 1$, the 1-Wasserstein distance. However, these bounds are achieved by departing from neural network approximations and employing kernel-based score estimators instead. While this constitutes an important theoretical contribution to the understanding of the fundamental statistical limits of diffusion models, it diverges from common algorithmic practice, where gradient descent methods for score matching exploit the flexibility and inductive biases of expressive neural architectures. Our analysis maintains this connection by explicitly studying neural network-based score estimators. Moreover, from an analytical perspective, translating kernel-based techniques to reflected diffusions on bounded domains would introduce severe boundary biases, necessitating the construction of highly specialised boundary-correction kernels. Given the practical prevalence of neural networks in generative modelling and these domain-specific analytic challenges, a theoretical investigation of kernel-based score matching is conceptually distinct from our objectives and therefore falls outside the scope of the present work.

Extension to general manifolds In this work, we restrict our geometric setup to data supported on a linear subspace $d \ll D$ intersecting the hypercube. While simpler than general non-linear manifolds, this setting already poses significant mathematical challenges due to the aforementioned absence of Gaussian transition kernels. In unconstrained OU settings, this Gaussian structure allowed [6, 118] to extend convergence rates to unknown compact d-dimensional manifolds, achieving bounds that scale with the intrinsic dimension d. Independently of any assumptions on the target distribution, the reflected diffusion case requires managing complex transition densities (cf. Lemma C.51), making the bounding of the score function and its estimation considerably more involved. Let us emphasize here that all of the results given in Section C.2 do not rely on any specific geometric assumptions on the support of μ but only partially on the assumption that p_0 has controlled decay at the boundary. For curved manifolds without boundary such that the density wrt the volume measure is bounded away from zero as in [6, 118] all statements therefore remain true if the lower bound in Lemma C.53.(d) is replaced by $t^{-d/2}\mathrm{e}^{-\rho}$. In particular, parts of Lemma C.53 provide a natural analogue to the crucial technical Lemma C.1 in [118] and can therefore serve as a natural starting point for score approximation for general manifold data via considering linear sub-problems associated to local chart parametrisations of the data density.

Discretisation and sampling errors A further extension of the present analysis involves the relaxation of the assumption of exact simulation for the backward dynamics. In practice, sampling from the generative model requires the discretisation of the backward SDE. For standard unconstrained processes, this is typically achieved via the Euler–Maruyama scheme, whose error properties have been extensively quantified in recent literature.

For reflected diffusions, however, the discretisation is more involved, as simulated paths must be strictly constrained to the domain $[0,1]^D$. This necessitates the use of alternative schemes, such as projected or penalised Euler–Maruyama methods, which explicitly account for the local time at the boundary. The numerical analysis of these methods requires bounding the error in the presence of reflecting barriers, where the discretisation error is coupled with the approximation of the score function near the boundary. Establishing a comprehensive end-to-end bound that incorporates these discretisation effects is a significant objective in numerical stochastic analysis. This constitutes a natural direction for future research, as also pointed out in the survey article [120].

Acknowledgements We thank Iskander Azangulov and Judith Rousseau for helpful discussions during the *MFO Mini-Workshop: Statistical Challenges for Deep Generative Models* and Simon Bienewald for helpful remarks on Wasserstein generalisation bounds.

C.A Proofs for Section C.2

Proof of Lemma C.50. Since $f(\mathsf{Y}) = \mathsf{Y}$ and $\mathsf{B}_0 = 0$ a.s., we have $\mathsf{X}_0 \sim \nu$. Fix $i \in [D]$ and define the one-dimensional process $\mathsf{Z}^{(i)}_t := \mathsf{B}^{(i)}_t + \mathsf{Y}^{(i)}$, $t \geq 0$. For $x \in \mathbb{R}$, let $\mathsf{L}^{(i)}_{t,x}$ denote the local time of $\mathsf{Z}^{(i)}$ at level x, that is, the unique process satisfying

$$\mathsf{L}^{(i)}_{t,x} = \lim_{\varepsilon \searrow 0} \frac{1}{2\varepsilon} \int_0^t \mathbf{1}_{(x-\varepsilon,x+\varepsilon)}(\mathsf{Z}^{(i)}_s)\,\mathrm{d}s.$$

Since $\widehat{f}$ is the difference of two convex functions, the Itô–Tanaka formula implies that $\mathsf{X}_t^{(i)} := \widehat{f}(\mathsf{Z}_t^{(i)})$ is a continuous semimartingale satisfying

$$\mathsf{X}_t^{(i)} = \mathsf{Y}^{(i)} + \int_0^t \widehat{f}'_-(\mathsf{Z}_s^{(i)})\,\mathrm{d}\mathsf{Z}_s^{(i)} + \frac{1}{2}\int_{\mathbb{R}} \mathsf{L}_{t,x}^{(i)}\,\widehat{f}''(\mathrm{d}x),$$

where $\widehat{f}'_-$ denotes the left derivative of $\widehat{f}$ and $\widehat{f}''$ its distributional second derivative. Define $\mathsf{W}_t^{(i)} := \int_0^t \widehat{f}'_-(\mathsf{Z}_s^{(i)})\,\mathrm{d}\mathsf{Z}_s^{(i)}$. As $\widehat{f}'_-(x) \in \{-1, 1\}$ for all $x \in \mathbb{R}$, we obtain

$$\langle \mathsf{W}^{(i)}\rangle_t = \int_0^t \widehat{f}'_-(\mathsf{Z}_s^{(i)})^2\,\mathrm{d}\langle \mathsf{B}^{(i)}\rangle_s = \int_0^t 1\,\mathrm{d}s = t,$$

and hence $(\mathsf{W}_t^{(i)})_{t\geq 0}$ is a standard Brownian motion by Lévy's characterisation. Next, define

$$\mathsf{L}_t^{(i),0} := \sum_{k\in\mathbb{Z}} \mathsf{L}_{t,2k}^{(i)}, \quad \mathsf{L}_t^{(i),1} := \sum_{k\in\mathbb{Z}} \mathsf{L}_{t,2k+1}^{(i)}.$$

Then, $\int_{\mathbb{R}} \mathsf{L}_{t,x}^{(i)}\,\widehat{f}''(\mathrm{d}x) = \mathsf{L}_t^{(i),0} - \mathsf{L}_t^{(i),1}$. Let $\mathsf{T}_k := \inf\{t \geq 0 : |\mathsf{Z}_t^{(i)}| \geq k\}$, $k \in \mathbb{Z}$, and denote by $L^{\mathsf{X}^{(i)}}$ the local time of $\mathsf{X}^{(i)}$ at 0. For all $t \geq 0$ and $n \in \mathbb{N}$, we have a.s.

$$\begin{aligned}
\mathsf{L}_{t\wedge\mathsf{T}_{2n}}^{\mathsf{X}^{(i)}} = \lim_{\varepsilon\searrow 0}\frac{1}{2\varepsilon}\int_0^{t\wedge\mathsf{T}_{2n}} \mathbf{1}_{(-\varepsilon,\varepsilon)}(\mathsf{X}_s^{(i)})\,\mathrm{d}s &= \lim_{\varepsilon\searrow 0}\frac{1}{2\varepsilon}\sum_{k\in\mathbb{Z},|k|\leq n}\int_0^{t\wedge\mathsf{T}_{2n}} \mathbf{1}_{(2k-\varepsilon,2k+\varepsilon)}(\mathsf{Z}_s^{(i)})\,\mathrm{d}s \\
&= \sum_{k\in\mathbb{Z},\,|k|\leq n}\lim_{\varepsilon\searrow 0}\frac{1}{2\varepsilon}\int_0^{t\wedge\mathsf{T}_{2n}} \mathbf{1}_{(2k-\varepsilon,2k+\varepsilon)}(\mathsf{Z}_s^{(i)})\,\mathrm{d}s \\
&= \sum_{k\in\mathbb{Z},\,|k|\leq n} \mathsf{L}_{t\wedge\mathsf{T}_{2n},2k}^{(i)},
\end{aligned}$$

where we used that $\mathsf{X}_s^{(i)} = 0$ if and only if $\mathsf{Z}_s^{(i)} = 2k$ for some $k \in \mathbb{Z}$. By monotone convergence,

$$\mathsf{L}_t^{\mathsf{X}^{(i)}} = \sum_{k\in\mathbb{Z}} \mathsf{L}_{t,2k}^{(i)} = \mathsf{L}_t^{(i),0}, \quad \text{for all } t \geq 0 \text{ a.s.}$$

An analogous argument shows that $\mathsf{L}_t^{(i),1}$ is the local time of $\mathsf{X}^{(i)}$ at 1. Consequently, $\mathsf{X}^{(i)}$ satisfies the one-dimensional reflected SDE

$$\mathrm{d}\mathsf{X}_t{}^{(i)} = \mathrm{d}\mathsf{W}_t^{(i)} + \frac{1}{2}\big(\mathrm{d}\mathsf{L}_t^{(i),0} - \mathrm{d}\mathsf{L}_t^{(i),1}\big).$$

Finally, define $\mathsf{L}_t := \frac{1}{2}\sum_{i=1}^{D}\big(\mathsf{L}_t^{(i),0} + \mathsf{L}_t^{(i),1}\big)$ and $\mathsf{W}_t := (\mathsf{W}_t^{(1)}, \dots, \mathsf{W}_t^{(D)})$. Then,

$$\begin{aligned}
\mathsf{X}_t &= \mathsf{X}_0 + \mathsf{W}_t + \frac{1}{2}\sum_{i=1}^{D} e_i\Big(\mathsf{L}_t^{(i),0} - \mathsf{L}_t^{(i),1}\Big) \\
&= \mathsf{X}_0 + \mathsf{W}_t + \sum_{i=1}^{D}\int_{s\in[0,t]\,:\,\{\mathsf{X}_s^{(i)}\in\{0,1\}\}} n(\mathsf{X}_t)\,\mathrm{d}\mathsf{L}_s^{(i),\mathsf{X}_s^{(i)}} \\
&= \mathsf{X}_0 + \mathsf{W}_t + \int_{s\in[0,t]\,:\,\bigcup_{i=1}^{D}\{\mathsf{X}_s^{(i)}\in\{0,1\}\}} n(\mathsf{X}_s)\,\mathrm{d}\Big(\frac{1}{2}\sum_{j=1}^{D}\mathsf{L}_s^{(j),\mathsf{X}_s^{(j)}}\Big) \\
&= \mathsf{X}_0 + \mathsf{W}_t + \int_{s\in[0,t]\,:\,\{\mathsf{X}_s\in\partial[0,1]^D\}} n(\mathsf{X}_s)\,\mathrm{d}\mathsf{L}_s = \mathsf{X}_0 + \mathsf{W}_t + \int_0^t n(\mathsf{X}_s)\,\mathrm{d}\mathsf{L}_s.
\end{aligned}$$

Moreover, for $i, j \in [D]$,

$$\langle \mathsf{W}^{(i)}, \mathsf{W}^{(j)}\rangle_t = \int_0^t \widehat{f'_{-}}(\mathsf{Z}_s^{(i)})\widehat{f'_{-}}(\mathsf{Z}_s^{(j)})\,\mathrm{d}\langle \mathsf{B}^{(i)}, \mathsf{B}^{(j)}\rangle_s = \begin{cases} t, & i = j, \\ 0, & i \neq j, \end{cases}$$

so that $(\mathsf{W}_t)_{t\geq 0}$ is a D-dimensional Brownian motion. Since $(\mathsf{L}_t)_{t\geq 0}$ is a local time of $(\mathsf{X}_t)_{t\geq 0}$, this completes the proof. □

Proof of Lemma C.51. Let f be defined as in Lemma C.50. By that lemma and by uniqueness in law of weak solutions to (C.67), the associated transition density q_t satisfies

$$q_t(y,x)\ \mathrm{d}x = \mathbb{P}(\mathsf{X}_t \in\ \mathrm{d}x \mid \mathsf{X}_0 = y) = \mathbb{P}\big(f(\mathsf{B}_t + y) \in\ \mathrm{d}x\big).$$

For any $z \in \mathbb{Z}^D$ and $x \in [0,1]^D + z$, we have $f(x) = R_z(x-z)$. Since the collection $\big([0,1]^D + z\big)_{z\in\mathbb{Z}^D}$ forms a partition of $\mathbb{R}^D$ up to Lebesgue null sets, it follows that for $x, y \in [0,1]^D$,

$$\begin{aligned}
q_t(y,x)\ \mathrm{d}x &= \sum_{z\in\mathbb{Z}^D}\mathbb{P}\big(f(\mathsf{B}_t + y) \in\ \mathrm{d}x,\ \mathsf{B}_t + y \in [0,1]^D + z\big) \\
&= \sum_{z\in\mathbb{Z}^D}\mathbb{P}\big(R_z(\mathsf{B}_t + y - z) \in\ \mathrm{d}x,\ \mathsf{B}_t + y \in [0,1]^D + z\big) \\
&= \sum_{z\in\mathbb{Z}^D}\mathbb{P}\big(\mathsf{B}_t \in R_z(\mathrm{d}x) + z - y\big) = (2\pi t)^{-D/2}\sum_{z\in\mathbb{Z}^D}\exp\left(-\frac{|R_z(x) + z - y|^2}{2t}\right)\ \mathrm{d}x.
\end{aligned}$$

In the third equality, we used that, for each $z \in \mathbb{Z}^D$, the mapping R_z is an involution on $[0,1]^D$, and that $R_z(\mathrm{d}x) + z - y \subseteq [0,1]^D + z - y$ for all $x \in [0,1]^D$. The fourth equality follows from the transformation theorem and the fact that the determinant of Jacobian $J_{R_z}(x) = (-1)^z$ has unit absolute value. Applying monotone convergence and integrating $q_t(y,x)$ against $\mu(\mathrm{d}y)$ yields

$$p_t(x) = (2\pi t)^{-D/2}\sum_{z\in\mathbb{Z}^D}\int_{[0,1]^D}\exp\left(-\frac{|R_z(x) + z - y|^2}{2t}\right)\ \mu(\mathrm{d}y),$$

as claimed. To compute the score, note that for $z \in \mathbb{Z}^D$ and $x, y \in [0,1]^D$, the chain rule gives

$$\nabla \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} = -\frac{\nabla |R_z(x)+z-y|^2}{2t} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} = -(-1)^z \frac{R_z(x)+z-y}{t} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}.$$

Hence, dominated convergence again yields

$$s_0(x,t) = \frac{\nabla p_t(x)}{p_t(x)} = -\frac{\sum_{z\in\mathbb{Z}^D} (-1)^z \int_{[0,1]^D} (R_z(x)+z-y) \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} \,\mu(\mathrm{d}y)}{t \sum_{z\in\mathbb{Z}^D} \int_{[0,1]^D} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} \,\mu(\mathrm{d}y)},$$

as desired. ☐

Proof of Lemma C.53. To show (a), notice that, by the triangle inequality,

$$t|s_0(x,t)| \leq \frac{\sum_{z\in\mathbb{Z}^D} \int_{[0,1]^D} |R_z(x)+z-y| \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} \,\mu(\mathrm{d}y)}{\sum_{z\in\mathbb{Z}^D} \int_{[0,1]^D} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} \,\mu(\mathrm{d}y)},$$

where, since the numerator is convergent, we must have for every $\varepsilon > 0$ and $x, y \in [0,1]^D$ that there exists a $K > 0$ such that

$$\sum_{\substack{z\in\mathbb{Z}^D \\ |R_z(x)+z-y|>K}} |R_z(x)+z-y| \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} < \varepsilon.$$

In particular, there exists a $K(x,y,t) > 0$ such that

$$\sum_{\substack{z\in\mathbb{Z}^D \\ |R_z(x)+z-y|>K(x,y,t)}} |R_z(x)+z-y| \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} \leq \sum_{z\in\mathbb{Z}^D} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}.$$

Thus, if we can show that $K(x,y,t) \leq K(t)$, independent of x and y, integrating both sides yields

$$\begin{aligned}\sum_{z\in\mathbb{Z}^D} \int_{[0,1]^D} |R_z(x)+z-y| \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} \,\mu(\mathrm{d}y) &= \int_{[0,1]^D} \Bigg(\sum_{\substack{z\in\mathbb{Z}^D \\ |R_z(x)+z-y|\leq K(t)}} |R_z(x)+z-y| \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} \\ &\qquad + \sum_{\substack{z\in\mathbb{Z}^D \\ |R_z(x)+z-y|>K(t)}} |R_z(x)+z-y| \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} \Bigg) \mu(\mathrm{d}y) \\ &\leq \big(K(t)+1\big) \int_{[0,1]^D} \sum_{z\in\mathbb{Z}^D} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} \,\mu(\mathrm{d}y),\end{aligned}$$

which then implies

$$|s_0(x,t)| \leq \frac{K(t)+1}{t}. \tag{C.79}$$

To do this, we first note that, for each $x, y \in [0,1]^D$ and $z \in \mathbb{Z}^D$, the point $R_z(x) + z - y$ lies in $[-1,1]^D + z$, and that the function $|u|\mathrm{e}^{-\frac{|u|^2}{2t}}$ is decreasing in $|u|$ whenever $|u| > \sqrt{t}$. Thus, when $|z| > 2\sqrt{D} + \sqrt{t}$, we have the rough estimate of

$$|R_z(x) + z - y|\mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} \leq \int_{[-2,2]^D} |z+u|\mathrm{e}^{-\frac{|z+u|^2}{2t}}\,\mathrm{d}u.$$

Now, notice that the set $\{z + [-2,2]^D \mid z \in \mathbb{Z}^D\}$ constitutes a covering of $\mathbb{R}^D$, where each hypercube $z + [0,1]^D$ is covered a total of 4^D times. Thus, we have for each integrable $g : \mathbb{R}^D \to \mathbb{R}$,

$$\sum_{z \in \mathbb{Z}^D} \int_{[-2,2]^D} g(u+z)\,\mathrm{d}u = 4^D \int_{\mathbb{R}^D} g(u)\,\mathrm{d}u.$$

Similarly, the set $\{z + [-2,2]^D \mid z \in \mathbb{Z}^D, |z| > K\}$ is a covering of a subset of $\mathbb{R}^D$ containing $\{u \in \mathbb{R}^D \mid |u| > K - 2\sqrt{D}\}$, and each hypercube $z + [0,1]^D$ is covered at most 4^D times, whence

$$\sum_{\substack{z \in \mathbb{Z}^D \\ |z| > K}} \int_{[-2,2]^D} g(u+z)\,\mathrm{d}u \leq 4^D \int_{\{|u| > K - 2\sqrt{D}\}} g(u)\,\mathrm{d}u.$$

Applying this to the above for some $K > 2\sqrt{D} + \sqrt{t}$, we get

$$\begin{aligned}
\sum_{\substack{z \in \mathbb{Z}^D \\ |z| > K}} |R_z(x) + z - y|\mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} &\leq \sum_{\substack{z \in \mathbb{Z}^D \\ |z| > K}} \int_{[-2,2]^D} |u+z|\mathrm{e}^{-\frac{|u+z|^2}{2t}}\,\mathrm{d}u \\
&\leq 4^D \int_{\{|u| > K - 2\sqrt{D}\}} |u|\mathrm{e}^{-\frac{|u|^2}{2t}}\,\mathrm{d}u \\
&= (32\pi t)^{\frac{D}{2}} \mathbb{E}[|\mathsf{B}_t| \mathbf{1}_{\{|\mathsf{B}_t| > K - 2\sqrt{D}\}}].
\end{aligned}$$

Setting $K(t) = 2\sqrt{D} + \sqrt{t(D + \frac{D}{t})} \lesssim 1 + \sqrt{t} \vee 1$, it follows by Lemma C.71.(b) that

$$\mathbb{E}[|\mathsf{B}_t| \mathbf{1}_{\{|\mathsf{B}_t| > K(t) - 2\sqrt{D}\}}] \lesssim t^{-\frac{D}{2}} \mathrm{e}^{-\frac{D}{2t}},$$

whence

$$\sum_{\substack{z \in \mathbb{Z}^D \\ |z| > K(t)}} |R_z(x) + z - y|\mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} \lesssim \mathrm{e}^{-\frac{D}{2t}} \leq \mathrm{e}^{-\frac{|x-y|^2}{2t}} \leq \sum_{z \in \mathbb{Z}^D} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}.$$

By the previous discussion discussion leading to (C.79), this proves part (a).

To prove (b), we first note that (a) gives

$$\mathbb{E}[|\nabla \log p_t(\mathsf{X}_t)|^2 \mathbf{1}_{M^{\mathrm{c}}_{\rho,t}}(\mathsf{X}_t)] \lesssim \frac{1}{t^2 \wedge 1} \mathbb{P}(\mathsf{X}_t \in M^{\mathrm{c}}_{\rho,t}).$$

To control the right hand side, let $f, \widehat{f}$ be as in Lemma C.50 such that $\mathsf{X}_t \sim f(\mathsf{B}_t + \mathsf{Y})$. For all $x \in \mathbb{R}$ and $y \in [0,1]$, we have $|\widehat{f}(x+y) - y| \leq |x|$, implying that $|f(\mathsf{B}_t + \mathsf{Y}) - \mathsf{Y}| \leq |\mathsf{B}_t|$, and hence that if $|\mathsf{B}_t| \leq \sqrt{t(D+2\rho)}$, then also $\operatorname{dist}(f(\mathsf{B}_t + \mathsf{Y}), M) \leq \sqrt{t(D+2\rho)}$. Using this and Lemma C.71.(a), we have

$$\mathbb{P}(\mathsf{X}_t \in M^{\mathrm{c}}_{\rho,t}) \leq \mathbb{P}(|\mathsf{B}_t| > \sqrt{t(D+2\rho)}) \lesssim \mathrm{e}^{-\rho},$$

showing (b).

We continue with the proof of (c). Using the same reasoning as above, we find for $\mathsf{Z}_u := \mathsf{B}^{(1)}_{\mathrm{e}^{-u}} / \sqrt{\mathrm{e}^{-u}}$ and $z > 0$ that

$$\mathbb{P}\big(\exists s \in [t,1] : |\mathsf{X}_s - \mathsf{X}_0| > \sqrt{s}z\big) \leq \mathbb{P}\Big(\sup_{s\in[t,1]} |\mathsf{B}_s/\sqrt{s}| > z\Big) \leq D\mathbb{P}\Big(\sup_{u\in[0,\log t^{-1}]} |\mathsf{Z}_u| > z/\sqrt{D}\Big). \quad \text{(C.80)}$$

Note that $(\mathsf{Z}_u)_{u\in[0,\log t^{-1}]}$ is a Gaussian process with canonical distance d_{Z} given by

$$d_{\mathsf{Z}}(u,v)^2 := \mathbb{E}\big[|\mathsf{Z}_u - \mathsf{Z}_v|^2\big] = 2\big(1 - \tfrac{\mathrm{e}^{-(u\vee v)}}{\mathrm{e}^{-(u+v)/2}}\big) = 2\big(1 - \mathrm{e}^{-|u-v|/2}\big).$$

It is readily verified that for $\varepsilon \in (0, \sqrt{2})$ the covering number $N([0,\log t^{-1}], d_{\mathsf{Z}}, \varepsilon)$ is bounded by

$$N([0,\log t^{-1}], d_{\mathsf{Z}}, \varepsilon) \leq 1 + \frac{\log t^{-1}}{-2\log(1-\varepsilon^2/2)} \leq 1 + \frac{\log t^{-1}}{\varepsilon^2}$$

and thus we obtain for the entropy integral

$$\begin{aligned}\int_0^\infty \log N([0,\log t^{-1}], d_{\mathsf{Z}}, \varepsilon)\,\mathrm{d}\varepsilon &\leq \int_0^{\sqrt{2}} \log\Big(1 + \frac{\log t^{-1}}{\varepsilon^2}\Big)\,\mathrm{d}\varepsilon \\ &\leq \sqrt{2}\log(2+\log t^{-1}) + 2\int_0^{\sqrt{2}} \log(1/\varepsilon)\,\mathrm{d}\varepsilon \\ &\leq \sqrt{2}\log(2+\log t^{-1}) + 2 \\ &\leq C_1 \log(1+\log t^{-1}),\end{aligned}$$

for some universal constant C_1. Moreover, we find for the diameter

$$\Delta_{d_{\mathsf{Z}}}([0,\log t^{-1}]) := \sup_{s,s'\in[0,\log t^{-1}]} d_{\mathsf{Z}}(s,s') \leq \sqrt{2}.$$

Thus, Dudley's entropy concentration bound for suprema of Gaussian processes, cf. e.g., [125, Remark 8.1.6], yields for any $y > 0$ and some constant $C_1 > 0$ that

$$\sup_{u\in[0,\log t^{-1}]} |\mathsf{Z}_u - \mathsf{Z}_0| \lesssim \int_0^\infty \log N([0,\log t^{-1}], d_{\mathsf{Z}}, \varepsilon)\,\mathrm{d}\varepsilon + \Delta_{d_{\mathsf{Z}}}([0,\log t^{-1}])u \leq C_1\log(1+\log t^{-1}) + \sqrt{2}y,$$

with probability larger than $1-2\mathrm{e}^{-y^2/2}$. Since Z_0 is standard normal, we therefore conclude that there exists a universal constant $C \geq 1$ such that

$$\sup_{u\in[0,\log t^{-1}]} |\mathsf{Z}_u| \leq C(\log(1+\log t^{-1})+y),$$

with probability larger than $1-4\mathrm{e}^{-2y^2}$. Consequently, it follows from (C.80) that

$$\mathbb{P}\Big(\forall s \in [t,1] \,:\, |\mathsf{X}_s - \mathsf{X}_0| \leq C\sqrt{Ds}\big(\log(1+\log t^{-1})+y\big)\Big) \geq 1-4D\mathrm{e}^{-2y^2},$$

which proves (c).

To show (d), let $x \in M_{\rho,t}$ be given and choose some $y_0 \in \overline{M}$ with $|y_0 - x| \leq \sqrt{t(D+2\rho)}$. Then, set $y_1 = y_0 + \frac{\sqrt{t}}{|y_0-x|}(y_0 - x)$ such that y_1 lies on the line containing x and y_0, but a distance of $\sqrt{t}$ further away from x than y_0. Then, $B(y_0, \sqrt{t}) \subseteq B(y_1, 2\sqrt{t})$ while $|x-y| \leq |x-y_1| + |y_1 - y| \leq \sqrt{t(D+3+2\rho)}$ for all $y \in B(y_0, \sqrt{t})$. Thus for all such y we have $\mathrm{e}^{\frac{|x-y|^2}{2t}} \gtrsim \mathrm{e}^{-\rho}$, and it follows by Assumption $(\mathcal{H}1)$ that

$$\int_{[0,1]^D} \mathrm{e}^{-\frac{|x-y|^2}{2t}}\,\mu(\mathrm{d}y) \geq \mathrm{e}^{-\frac{D+3+2\rho}{2}}\,\mu\big(B(y_0,\sqrt{t})\big) \gtrsim t^{\frac{c_0}{2}}\,\mathrm{e}^{-\rho}.$$

This implies that for such x, the same is true of $(2\pi t)^{\frac{D}{2}} p_t(x)$, showing (d).

Finally, we prove part (e). We first show part (i), that is $|\nabla_x \log q_t(y,x)| \lesssim \frac{|x-y|}{t} + \frac{1}{\sqrt{t}}$ for all $x, y \in [0,1]^D$. To this end, for such x, y let $Z_1(x,y) = \{z \in \mathbb{Z}^D \,:\, |R_z(x) + z - y| \leq \sqrt{2}|x-y|\}$ and $Z_2 = \mathbb{Z}^D \setminus Z_1$. Then,

$$\begin{aligned}
|\nabla_x \log q_t(y,x)| &\leq \frac{\sum_{z\in\mathbb{Z}^D} \frac{|R_z(x)+z-y|}{t}\mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}}{\sum_{z\in\mathbb{Z}^D} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}} \\
&= \frac{\sum_{z\in Z_1(x,y)} \frac{|R_z(x)+z-y|}{t}\mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}}{\sum_{z\in\mathbb{Z}^D} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}} + \frac{\sum_{z\in Z_2(x,y)} \frac{|R_z(x)+z-y|}{t}\mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}}{\sum_{z\in\mathbb{Z}^D} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}} \\
&\leq \frac{\sqrt{2}|x-y|}{t} + \frac{\sum_{z\in Z_2(x,y)} \frac{|R_z(x)+z-y|}{t}\mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}}{\sum_{z\in\mathbb{Z}^D} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}}.
\end{aligned}$$

Now, note that for $z \in Z_2(x,y)$ we have $|R_z(x)+z-y|^2 - |x-y|^2 > \frac{1}{2}|R_z(x)+z-y|^2$, whence

$$\begin{aligned}
\frac{\sum_{z\in Z_2(x,y)} \frac{|R_z(x)+z-y|}{t}\mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}}{\sum_{z\in\mathbb{Z}^D} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}}} &\leq \sum_{z\in Z_2(x,y)} \frac{|R_z(x)+z-y|}{t}\mathrm{e}^{\frac{|x-y|^2-|R_z(x)+z-y|^2}{2t}} \\
&\leq 2\sum_{z\in\mathbb{Z}^D} \frac{|R_z(x)+z-y|}{2t}\mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{4t}}.
\end{aligned}$$

Now, to evaluate this final sum, let $f_t(x) = \frac{|x|}{t} \mathrm{e}^{-\frac{|x|^2}{2t}}$, and note that for all $z \in \mathbb{Z}^D$, we have

$$\#\big\{ z' \in \mathbb{Z}^D \,:\, R_{z'}(x) + z' - y \in [0,1]^D + z \big\} \leq 2^D,$$

whereby

$$\sum_{z \in \mathbb{Z}^D} f_{2t}(R_z(x) + z - y) \leq 2^D \sum_{z \in \mathbb{Z}^D} \sup_{x \in [0,1]^D + z} f_{2t}(x)$$

Since $f_t(x) \leq f_t(\sqrt{t}\frac{x}{|x|}) \leq \frac{1}{\sqrt{t}}$, it follows that for the 2^D terms where $[0,1]^D + z \subseteq [-1,1]^D$, we have $\sup_{x \in [0,1]^D + z} f_{2t}(x) \leq \frac{1}{\sqrt{2t}}$, while for all others, it is the point in $[0,1]^D + z$ closest to the origin since f_t is decreasing in $|x|$ for $|x| > \sqrt{t}$. The set of all such points is merely $\mathbb{Z}^D \setminus \{0\}$ with points near the axes being repeated a maximum of 2^D times, whence

$$\sum_{z \in \mathbb{Z}^D} \frac{|R_z(x) + z - y|}{2t} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{4t}} \leq 4^D \Big(\frac{1}{\sqrt{2t}} + \sum_{z \in \mathbb{Z}^D} \frac{|z|}{2t} \mathrm{e}^{-\frac{|z|^2}{4t}} \Big). \tag{C.81}$$

Since

$$\sum_{z \in \mathbb{Z}^D} \frac{|z|}{2t} \mathrm{e}^{-\frac{|z|^2}{4t}} \asymp \int_{\mathbb{R}^D} \frac{|x|}{2t} \mathrm{e}^{-\frac{|x|^2}{4t}} \,\mathrm{d}x = \frac{(2\pi t)^{\frac{D}{2}}}{2t} \mathbb{E}\big[|\mathsf{B}_{2t}|\big] \leq \frac{(2\pi t)^{\frac{D}{2}}}{\sqrt{2t}},$$

it follows from (C.81) that

$$\sum_{z \in \mathbb{Z}^D} \frac{|R_z(x) + z - y|}{2t} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{4t}} \lesssim \frac{1}{\sqrt{t}},$$

and hence $|\nabla_x \log q_t(y,x)| \lesssim \frac{|x-y|}{t} + \frac{1}{\sqrt{t}}$ as claimed. By the score matching identity

$$\nabla \log p_t(x) = \int \nabla_x \log q_t(y,x) \frac{q_t(y,x)\, \mu(\mathrm{d}y)}{p_t(x)} = \mathbb{E}[\nabla_2 \log q_t(\mathsf{X}_0, \mathsf{X}_t) \mid \mathsf{X}_t = x],$$

it follows that

$$|\nabla \log p_t(x)| \leq \mathbb{E}[|\nabla_2 \log q_t(\mathsf{X}_0, \mathsf{X}_t)| \mid \mathsf{X}_t = x] \lesssim \frac{1}{t} \mathbb{E}[|\mathsf{X}_0 - \mathsf{X}_t| \mid \mathsf{X}_t = x] + \frac{1}{\sqrt{t}},$$

proving part (e).(ii). Let now $t < 1$. We have

$$\begin{aligned} \mathbb{E}[|\mathsf{X}_0 - \mathsf{X}_t| \mid \mathsf{X}_t = x] &= \frac{\int_M |x-y| q_t(y,x)\, \mu(\mathrm{d}y)}{p_t(x)} \\ &= \frac{\int_{M \cap B(x,\rho_t)} |x-y| q_t(y,x)\, \mu(\mathrm{d}y)}{p_t(x)} + \frac{\int_{M \cap B(x,\rho_t)^{\mathrm{c}}} |x-y| q_t(y,x)\, \mu(\mathrm{d}y)}{p_t(x)}, \end{aligned}$$

where $\rho_t = \sqrt{3t(\rho + \frac{c_0+1}{2} \log t^{-1})}$. Clearly, for the first term we have

$$\frac{\int_{M \cap B(x,\rho_t)} |x-y| q_t(y,x)\, \mu(\mathrm{d}y)}{p_t(x)} \leq \frac{\rho_t \int_M q_t(y,x)\, \mu(\mathrm{d}y)}{p_t(x)} = \rho_t \lesssim \sqrt{t(\rho + \log t^{-1})},$$

while for the second term, we have by (d) that $p_t(x) \gtrsim t^{\frac{c_0-D}{2}}\mathrm{e}^{-\rho}$ for $x \in M_{\rho,t}$, and since the diameter of $[0,1]^D$ is $\sqrt{D}$, we get

$$\frac{\int_{M\cap B(x,\rho_t)^{\mathrm{c}}} |x-y|q_t(y,x)\,\mu(\mathrm{d}y)}{p_t(x)} \leq \sqrt{D}\frac{\int_{M\cap B(x,\rho_t)^{\mathrm{c}}} q_t(y,x)\,\mu(\mathrm{d}y)}{p_t(x)} \lesssim t^{\frac{D-c_0}{2}}\mathrm{e}^{\rho}\int_{M\cap B(x,\rho_t)^{\mathrm{c}}} q_t(y,x)\,\mu(\mathrm{d}y).$$

Now, by [74, Theorem 3.2] we have $q_t(y,x) \lesssim t^{-\frac{D}{2}}\mathrm{e}^{-\frac{|x-y|^2}{3t}}$, whence

$$t^{\frac{D-c_0}{2}}\mathrm{e}^{\rho}\int_{M\cap B(x,\rho_t)^{\mathrm{c}}} q_t(y,x)\,\mu(\mathrm{d}y) \lesssim t^{-\frac{c_0}{2}}\mathrm{e}^{\rho-\frac{\rho_t^2}{3t}} = \sqrt{t}.$$

Thus, using part (e).(ii), we find

$$|\nabla \log p_t(x)| \lesssim \frac{1}{t}\mathbb{E}[|\mathsf{X}_0 - \mathsf{X}_t| \mid \mathsf{X}_t = x] + \frac{1}{\sqrt{t}} \lesssim \frac{\sqrt{t(\rho+\log t^{-1})}+\sqrt{t}}{t} + \frac{1}{\sqrt{t}} \lesssim \frac{\sqrt{\rho+\log t^{-1}}}{\sqrt{t}},$$

showing (e).(iii). □

C.B Remaining proofs for Section C.3

Proof of Proposition C.57. Combining (C.73) with Pinsker's inequality and Girsanov's theorem for reflected diffusions, cf. [58, Theorem A.1], we obtain

$$\begin{aligned}\mathcal{W}_1\big(\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \mathsf{Y}^{(i)}_{\overline{T}-\underline{T}}\big) &\leq 2\sqrt{D}\,\mathrm{TV}\big(\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \mathsf{Y}^{(i)}_{\overline{T}-\underline{T}}\big)\\ &\leq \sqrt{2D\ \mathrm{KL}(\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}} \,\|\, \mathsf{Y}^{(i)}_{\overline{T}-\underline{T}})} = \sqrt{D\int_{t_{i-1}}^{t_i} \mathbb{E}\left[\left|s(\mathsf{X}_t,t) - \nabla\log p_t(\mathsf{X}_t)\right|^2\right]\mathrm{d}t}.\end{aligned}$$

This bound already yields the claim whenever $t_i \gtrsim \rho^{-1}$. Hence, in the remainder of the proof we may and do assume that $t_i \leq 1/(C_1\rho)$, for a constant $C_1 > 0$ to be chosen later. Let $\mathbb{Q}^{(i)}$ denote the law of the full path $\mathsf{Y}^{(i)}$ on $C([0,\overline{T}-\underline{T}],[0,1]^D)$. By the Kantorovich–Rubinstein duality,

$$\mathcal{W}_1\big(\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \mathsf{Y}^{(i)}_{\overline{T}-\underline{T}}\big) = \sup_{\|f\|_{\mathrm{Lip}}\leq 1}\left|\int f\big(p(\overline{T}-\underline{T})\big)\,(\mathbb{Q}^{(i-1)} - \mathbb{Q}^{(i)})(\mathrm{d}p)\right|.$$

Since the processes $\mathsf{Y}^{(i-1)}$ and $\mathsf{Y}^{(i)}$ coincide on $[0,\overline{T}-t_i)$, their marginals at time $\overline{T}-t_i$ agree, and thus

$$\sup_{\|f\|_{\mathrm{Lip}}\leq 1}\left|\int f\big(p(\overline{T}-t_i)\big)\,(\mathbb{Q}^{(i-1)} - \mathbb{Q}^{(i)})(\mathrm{d}p)\right| = \mathcal{W}_1\big(\mathsf{Y}^{(i-1)}_{\overline{T}-t_i}, \mathsf{Y}^{(i)}_{\overline{T}-t_i}\big) = 0.$$

Subtracting this null term yields

$$\mathcal{W}_1\big(\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \mathsf{Y}^{(i)}_{\overline{T}-\underline{T}}\big) = \sup_{\|f\|_{\mathrm{Lip}}\leq 1}\left|\int \big[f(p(\overline{T}-\underline{T})) - f(p(\overline{T}-t_i))\big](\mathbb{Q}^{(i-1)} - \mathbb{Q}^{(i)})(\mathrm{d}p)\right|.$$

Fix $C_2 > 0$, and introduce the event $A_i = \big\{ |p(\overline{T}-\underline{T}) - p(\overline{T}-t_i)| \le C_2\sqrt{t_i\rho} \big\} \subset C([\overline{T}-\underline{T}],[0,1]^D)$. Splitting the integral over A_i and A_i^{c}, and denoting by $|\nu|$ the total variation of a signed measure ν, we obtain

$$\begin{aligned}
\sup_{\|f\|_{\mathrm{Lip}}\le 1} &\left| \int \big[f(p(\overline{T}-\underline{T})) - f(p(\overline{T}-t_i))\big](\mathbb{Q}^{(i-1)} - \mathbb{Q}^{(i)})(\mathrm{d}p) \right| \\
&\le C_2\sqrt{t_i\rho} \int |\mathbb{Q}^{(i-1)} - \mathbb{Q}^{(i)}|(\mathrm{d}p) + \sqrt{D}\big(\mathbb{Q}^{(i-1)}(A_i^{\mathsf{c}}) + \mathbb{Q}^{(i)}(A_i^{\mathsf{c}})\big).
\end{aligned}$$

The first term is bounded using Pinsker's inequality as

$$C_2\sqrt{t_i\rho}\,\mathrm{TV}\big(\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}}, \mathsf{Y}^{(i)}_{\overline{T}-\underline{T}}\big) \lesssim \sqrt{t_i\rho}\left(\int_{t_{i-1}}^{t_i} \mathbb{E}\left[\big|s(\mathsf{X}_t,t) - \nabla\log p_t(\mathsf{X}_t)\big|^2\right] \mathrm{d}t\right)^{1/2}.$$

Hence, it remains to control the tail probabilities

$$\mathbb{P}\big(|\mathsf{Y}^{(i)}_{\overline{T}-\underline{T}} - \mathsf{Y}^{(i)}_{\overline{T}-t_i}| > C_2\sqrt{t_i\rho}\big), \qquad \mathbb{P}\big(|\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}} - \mathsf{Y}^{(i-1)}_{\overline{T}-t_i}| > C_2\sqrt{t_i\rho}\big).$$

To this end, let $\mathsf{Z}^{(i)}$ denote a solution to the SDE

$$\mathrm{d}\mathsf{Z}^{(i)}_t = s\big(\mathsf{Z}^{(i)}_t, \overline{T}-t_i-t\big)\,\mathrm{d}t + \mathrm{d}\mathsf{B}_{\overline{T}-t_i+t}, \qquad \mathsf{Z}^{(i)}_0 = \mathsf{Y}^{(i)}_{\overline{T}-t_i},$$

and define $\tau := \inf\{t \ge 0 : \mathsf{Z}^{(i)}_t \in \partial[0,1]^D\}$. By construction, we have $\mathsf{Z}^{(i)}_t = \mathsf{Y}^{(i)}_{\overline{T}-t_i+t}$ for all $t \in [0, \tau \wedge (t_i - \underline{T})]$. Consequently,

$$\mathbb{P}\big(|\mathsf{Y}^{(i)}_{\overline{T}-\underline{T}} - \mathsf{Y}^{(i)}_{\overline{T}-t_i}| > C_2\sqrt{t_i\rho}\big) \le 1 - \mathbb{P}\big(|\mathsf{Z}^{(i)}_{t_i-\underline{T}} - \mathsf{Z}^{(i)}_0| \le C_2\sqrt{t_i\rho},\ \tau \ge t_i - \underline{T}\big).$$

We now introduce the events

$$A_1 := \big\{\forall t \in [t_i, 1] : |\mathsf{X}_t - \mathsf{X}_0| \le C_3\sqrt{t\rho}\big\}, \quad A_2 := \Big\{\sup_{t\in[0,t_i-\underline{T}]} \big|\mathsf{B}_{\overline{T}-t_i+t} - \mathsf{B}_{\overline{T}-t_i}\big| \le C_4\sqrt{t_i\rho}\Big\},$$

where the constants $C_3, C_4 > 0$ will be specified below. On the event $A_1 \cap A_2$, we use that $\mathsf{Z}^{(i)}_0 = \mathsf{Y}^{(i)}_{\overline{T}-t_i} = \mathsf{X}_{t_i}$ and obtain $\mathrm{dist}(\mathsf{Z}^{(i)}_0, M) \le |\mathsf{X}_{t_i} - \mathsf{X}_0| \le C_3\sqrt{t_i\rho}$. Moreover, using the assumed bound $|s(x,t)| \le C\sqrt{\rho/t}$, we have for all $t \in [0, t_i - \underline{T}]$,

$$\begin{aligned}
|\mathsf{Z}^{(i)}_t - \mathsf{Z}^{(i)}_0| &\le \int_0^t |s(\mathsf{Z}^{(i)}_u, \overline{T}-t_i-u)|\,\mathrm{d}u + \big|\mathsf{B}_{\overline{T}-\underline{T}} - \mathsf{B}_{\overline{T}-t_i}\big| \\
&\le C\sqrt{\rho}\int_0^t \frac{1}{\sqrt{\overline{T}-t_i-u}}\,\mathrm{d}u + C_4\sqrt{t_i\rho} \\
&\le (2C + C_4)\sqrt{t_i\rho}
\end{aligned}$$

on $A_1 \cap A_2$. Combining the previous two bounds yields

$$\operatorname{dist}(\mathsf{Z}_t^{(i)}, M) \le \operatorname{dist}(\mathsf{Z}_0^{(i)}, M) + |\mathsf{Z}_t^{(i)} - \mathsf{Z}_0^{(i)}| \le (C_3 + 2C + C_4)\sqrt{t_i \rho}$$

on $A_1 \cap A_2$. Since $t_i \le (C_1\rho)^{-1}$ by assumption, we further obtain that on $A_1 \cap A_2$,

$$\operatorname{dist}(\mathsf{Z}_t^{(i)}, \partial[0,1]^D) \ge \rho_{\min} - \operatorname{dist}(\mathsf{Z}_t^{(i)}, M) \ge \rho_{\min} - \frac{C_3 + 2C + C_4}{\sqrt{C_1}}.$$

Choosing $C_1 \ge 1 \vee \left(\frac{2(C_3+2C+C_4)}{\rho_{\min}}\right)^2$ ensures that $\operatorname{dist}(\mathsf{Z}_t^{(i)}, \partial[0,1]^D) > 0$ for all $t \in [0, t_i - \underline{T}]$, and hence $\tau \ge t_i - \underline{T}$ on $A_1 \cap A_2$. On this event, we therefore have

$$|\mathsf{Z}_{t_i-\underline{T}}^{(i)} - \mathsf{Z}_0^{(i)}| \le C_{2,1}\sqrt{t_i\rho}, \quad C_{2,1} := 2C + C_4.$$

It follows that

$$\mathbb{P}\big(|\mathsf{Y}_{\overline{T}-\underline{T}}^{(i)} - \mathsf{Y}_{\overline{T}-t_i}^{(i)}| > C_{2,1}\sqrt{t_i\rho}\big) \le \mathbb{P}(A_1^{\mathsf{c}}) + \mathbb{P}(A_2^{\mathsf{c}}).$$

We now bound the two probabilities on the right-hand side. Choosing

$$C_3 = \widetilde{C}\sqrt{D}\,\frac{\log\left(1+\log(t_1^{-1})\right)+\sqrt{\rho}}{\sqrt{\rho}} \lesssim 1,$$

where $\widetilde{C}$ is the universal constant from Lemma C.53.(c), that lemma yields $\mathbb{P}(A_1^{\mathsf{c}}) \lesssim \mathrm{e}^{-\rho}$. Similarly, choosing $C_4 = \frac{\sqrt{D+4\rho}}{\sqrt{\rho}} \lesssim 1$, it follows by Lemma C.71.(a) that

$$\begin{aligned}
\mathbb{P}(A_2^{\mathsf{c}}) &= \mathbb{P}\Big(\sup_{t\in[0,t_i-\underline{T}]} |\mathsf{B}_{\overline{T}-t_i+t} - \mathsf{B}_{\overline{T}-t_i}| > \sqrt{t_i(D+4\rho)}\Big) \\
&= \mathbb{P}\Big(\sup_{t\in[0,t_i-\underline{T}]} |\mathsf{B}_t| > \sqrt{t_i(D+4\rho)}\Big) \\
&\le 2\mathbb{P}\big(|\mathsf{B}_{t_i-\underline{T}}| > \sqrt{t_i(D+4\rho)}\big) \\
&\lesssim \mathrm{e}^{-\rho}.
\end{aligned}$$

For the first inequality we used the following classical argument for Brownian motion: let $t, a > 0$ and $\tau_a := \inf\{s \ge 0 : |\mathsf{B}_s| = a\}$. Then

$$\begin{aligned}
\mathbb{P}\Big(\sup_{s\le t} |\mathsf{B}_s| > a\Big) = \mathbb{P}(\tau_a \le t) &= \mathbb{P}(\tau_a \le t, |\mathsf{B}_t| > a) + \mathbb{P}(\tau_a \le t, |\mathsf{B}_t| \le a) \\
&\le \mathbb{P}(|\mathsf{B}_t| > a) + \mathbb{P}(\tau_a \le t, |\mathsf{B}_t| \le a),
\end{aligned}$$

and by the tower rule and strong Markov property,

$$\mathbb{P}(\tau_a \le t, |\mathsf{B}_t| \le a) = \mathbb{E}\big[\mathbf{1}_{\{\tau_a \le t\}}\mathbb{P}^{\mathsf{B}_{\tau_a}}(|\mathsf{B}_{t-\tau_a}| \le a)\big] \le \frac{1}{2}\mathbb{P}(\tau_a \le t), \tag{C.82}$$

where we used that for any $s \geq 0$ and $x \in \mathbb{R}^D$ with $|x| = a$ we have

$$\mathbb{P}^x(|\mathsf{B}_s| \leq a) = \mathbb{P}^0(|\mathsf{B}_s - x| \leq a) \leq \mathbb{P}(\mathsf{B}_s \cdot x \geq 0) = \frac{1}{2},$$

because $\mathsf{B}_s \cdot x$ is a centered Gaussian random variable. Inserting this into (C.82) and rearranging yields

$$\mathbb{P}\Big(\sup_{s \leq t} |\mathsf{B}_s| > a\Big) \leq 2\mathbb{P}(|\mathsf{B}_t| \geq a).$$

Next, since $|\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}} - \mathsf{Y}^{(i)}_{\overline{T}-t_i}| \leq |\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}} - \mathsf{Y}^{(i-1)}_{\overline{T}-t_{i-1}}| + |\mathsf{Y}^{(i-1)}_{\overline{T}-t_{i-1}} - \mathsf{Y}^{(i-1)}_{\overline{T}-t_i}|$, it follows that

$$\mathbb{P}(|\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}} - \mathsf{Y}^{(i-1)}_{\overline{T}-t_i}| > C_2\sqrt{t_i\rho}) \leq \mathbb{P}(|\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}} - \mathsf{Y}^{(i-1)}_{\overline{T}-t_{i-1}}| > C_2\sqrt{t_i\rho}) + \mathbb{P}(|\mathsf{Y}^{(i-1)}_{\overline{T}-t_{i-1}} - \mathsf{Y}^{(i-1)}_{\overline{T}-t_i}| > C_2\sqrt{t_i\rho}).$$

Here, since $t_{i-1} \leq t_i$, we have

$$\mathbb{P}(|\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}} - \mathsf{Y}^{(i-1)}_{\overline{T}-t_{i-1}}| > C_2\sqrt{t_i\rho}) \leq \mathbb{P}(|\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}} - \mathsf{Y}^{(i-1)}_{\overline{T}-t_{i-1}}| > C_2\sqrt{t_{i-1}\rho}) \lesssim \mathrm{e}^{-\rho}$$

by the exact same argument as above. Meanwhile, since $\mathsf{Y}^{(i-1)}_{\overline{T}-t} = \mathsf{X}_t$ for $t \in [t_{i-1}, t_i]$, we have

$$|\mathsf{Y}^{(i-1)}_{\overline{T}-t_{i-1}} - \mathsf{Y}^{(i-1)}_{\overline{T}-t_i}| = |\mathsf{X}_{t_{i-1}} - \mathsf{X}_{t_i}| \leq |\mathsf{X}_{t_{i-1}} - \mathsf{X}_0| + |\mathsf{X}_{t_i} - \mathsf{X}_0|,$$

and so once again by Lemma C.53.(c)

$$\begin{aligned}
\mathbb{P}(|\mathsf{Y}^{(i-1)}_{\overline{T}-t_{i-1}} - \mathsf{Y}^{(i-1)}_{\overline{T}-t_i}| \leq C_2\sqrt{t_i\rho}) &\geq \mathbb{P}\Big(|\mathsf{X}_{t_{i-1}} - \mathsf{X}_0| \leq \frac{C_2\sqrt{t_i\rho}}{2}, |\mathsf{X}_{t_i} - \mathsf{X}_0| \leq \frac{C_2\sqrt{t_i\rho}}{2}\Big) \\
&\geq \mathbb{P}\Big(\forall t \in [t_{i-1}, 1] \,:\, |\mathsf{X}_t - \mathsf{X}_0| \leq \frac{C_2\sqrt{t\rho}}{2}\Big) \\
&\gtrsim 1 - \mathrm{e}^{-\rho},
\end{aligned}$$

where $C_2 = (2C + C_4) \vee (2\widetilde{C}\sqrt{D}\frac{\log(1+\log t_{i-1}^{-1}) + \sqrt{\rho}}{\sqrt{\rho}}) \geq C_{2,1}$. In summary, we conclude that

$$\begin{aligned}
&\mathbb{P}(|\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}} - \mathsf{Y}^{(i-1)}_{\overline{T}-t_i}| > C_2\sqrt{t_i\rho}) + \mathbb{P}(|\mathsf{Y}^{(i)}_{\overline{T}-\underline{T}} - \mathsf{Y}^{(i)}_{\overline{T}-t_i}| > C_2\sqrt{t_i\rho}) \\
&\leq \mathbb{P}(|\mathsf{Y}^{(i-1)}_{\overline{T}-\underline{T}} - \mathsf{Y}^{(i-1)}_{\overline{T}-t_i}| > C_2\sqrt{t_i\rho}) + \mathbb{P}(|\mathsf{Y}^{(i)}_{\overline{T}-\underline{T}} - \mathsf{Y}^{(i)}_{\overline{T}-t_i}| > C_{2,1}\sqrt{t_i\rho}) \lesssim \mathrm{e}^{-\rho},
\end{aligned}$$

which finishes the proof. ☐

Proof of the score approximation accuracy

We follow the approximation programme outlined in Section C.3.

Step 1: score truncation error

Lemma C.60: Fix $\underline{t} > 0$, let $K \in \mathbb{N}_0$ be given and let s_0^K be given by (C.77). Then we have for $t \in [\underline{t}, 2\underline{t}]$

$$\mathbb{E}[|s_0(\mathsf{X}_t, t) - s_0^K(\mathsf{X}_t, t)|^2] \lesssim \frac{1}{t^2 \wedge 1} K^{\frac{D}{2}} \mathrm{e}^{-K}.$$

Proof. For notation, set $K_{\underline{t}} = \sqrt{2\underline{t}(D + 2K)}$. We first have by the triangle inequality that

$$\begin{aligned} \mathbb{E}[|s_0(\mathsf{X}_t, t) - s_0^K(\mathsf{X}_t, t)|^2] &= \mathbb{E}\Big[\Big|\frac{1}{p_t(\mathsf{X}_t)}\Big(s_0^K(\mathsf{X}_t, t)\big(p_t(\mathsf{X}_t) - p_t^K(\mathsf{X}_t)\big) - \nabla\big(p_t(\mathsf{X}_t) - p_t^K(\mathsf{X}_t)\big)\Big)\Big|^2\Big] \\ &\lesssim \mathbb{E}\Big[|s_0^K(\mathsf{X}_t, t)|^2 \frac{\big(p_t(\mathsf{X}_t) - p_t^K(\mathsf{X}_t)\big)^2}{p_t(\mathsf{X}_t)^2}\Big] + \mathbb{E}\Big[\frac{|\nabla(p_t(\mathsf{X}_t) - p_t^K(\mathsf{X}_t))|^2}{p_t(\mathsf{X}_t)^2}\Big]. \end{aligned} \tag{C.83}$$

We first concentrate on the first term. From (the proof of) Lemma C.53.(a), it follows that also $|s_0^K(x, t)| \lesssim \frac{1}{t \wedge 1}$, whence

$$\begin{aligned} \mathbb{E}\Big[|s_0^K(\mathsf{X}_t, t)|^2 \frac{\big(p_t(\mathsf{X}_t) - p_t^K(\mathsf{X}_t)\big)^2}{p_t(\mathsf{X}_t)^2}\Big] &\lesssim \frac{1}{t^2 \wedge 1} \mathbb{E}\Big[\frac{\big(p_t(\mathsf{X}_t) - p_t^K(\mathsf{X}_t)\big)^2}{p_t(\mathsf{X}_t)^2}\Big] \\ &= \frac{1}{t^2 \wedge 1} \int_{[0,1]^D} \frac{\big(p_t(x) - p_t^K(x)\big)^2}{p_t(x)} \,\mathrm{d}x \\ &\leq \frac{1}{t^2 \wedge 1} \int_{[0,1]^D} p_t(x) - p_t^K(x) \,\mathrm{d}x, \end{aligned}$$

where in the last inequality we use that the terms of $p_t(x)$ are non-negative, whence $p_t(x) - p_t^K(x) \leq p_t(x)$. Next, notice that for all $z \in \mathbb{Z}^D$, we have $R_z([0,1]^D) = [0,1]^D$, whence $\{R_z([0,1]^D) + z \mid z \in \mathbb{Z}^D\}$ is a disjoint (apart from a measure 0 set) partition of $\mathbb{R}^D$, and similarly $\{R_z([0,1]^D) + z \mid z \in \mathbb{Z}^D, \|z\|_\infty > K_{\underline{t}}\}$ is a partition of the set $\mathbb{R}^D \setminus [-\lfloor K_{\underline{t}} \rfloor, \lfloor K_{\underline{t}} \rfloor + 1]^D$, whence for any integrable g,

$$\sum_{\substack{z \in \mathbb{Z}^D \\ \|z\|_\infty > K_{\underline{t}}}} \int_{[0,1]^D} g(R_z(x) + z) \,\mathrm{d}x = \int_{\mathbb{R}^D \setminus [-\lfloor K_{\underline{t}} \rfloor, \lfloor K_{\underline{t}} \rfloor + 1]^D} g(x) \,\mathrm{d}x.$$

In particular, by Fubini–Tonelli's theorem

$$\begin{aligned} \int_{[0,1]^D} p_t(x) - p_t^K(x) \,\mathrm{d}x &= (2\pi t)^{-\frac{D}{2}} \sum_{\substack{z \in \mathbb{Z}^D \\ \|z\|_\infty > K_{\underline{t}}}} \int_{[0,1]^D} \int_{[0,1]^D} \mathrm{e}^{-\frac{|R_z(x) + z - y|^2}{2t}} \,\mu(\mathrm{d}y) \,\mathrm{d}x \\ &= \int_{[0,1]^D} \Big(\int_{\mathbb{R}^D \setminus [-\lfloor K_{\underline{t}} \rfloor, \lfloor K_{\underline{t}} \rfloor + 1]^D} (2\pi t)^{-\frac{D}{2}} \mathrm{e}^{-\frac{|x-y|^2}{2t}} \,\mathrm{d}x \Big) \,\mu(\mathrm{d}y). \end{aligned}$$

Now, for each $y \in M$ and $x \in \mathbb{R}^D \setminus [-\lfloor K_{\underline{t}} \rfloor, \lfloor K_{\underline{t}} \rfloor + 1]^D$, we necessarily have that $|x-y| \geq \|x-y\|_\infty \geq \lfloor K_{\underline{t}} \rfloor$, whence Lemma C.71.(a) yields

$$\begin{aligned}
\int_{\mathbb{R}^D \setminus [-\lfloor K_{\underline{t}} \rfloor, \lfloor K_{\underline{t}} \rfloor + 1]^D} (2\pi t)^{-\frac{D}{2}} \mathrm{e}^{-\frac{|x-y|^2}{2t}} \, \mathrm{d}x &\leq \int_{\{|x-y| \geq \lfloor K_{\underline{t}} \rfloor\}} (2\pi t)^{-\frac{D}{2}} \mathrm{e}^{-\frac{|x-y|^2}{2t}} \, \mathrm{d}x \\
&= \mathbb{P}(|\mathsf{B}_t| \geq \lfloor K_{\underline{t}} \rfloor) \\
&= \mathbb{P}(|\mathsf{B}_t| \in [\lfloor K_{\underline{t}} \rfloor, K_{\underline{t}})) + \mathbb{P}(|\mathsf{B}_t| \geq K_{\underline{t}}) \\
&\lesssim \mathbb{P}(|\mathsf{B}_t| \in [\lfloor K_{\underline{t}} \rfloor, K_{\underline{t}})) + K^{\frac{D}{2}} \mathrm{e}^{-K},
\end{aligned}$$

since $t \leq 2\underline{t}$. Meanwhile, we have

$$\begin{aligned}
\mathbb{P}(|\mathsf{B}_t| \in [\lfloor K_{\underline{t}} \rfloor, K_{\underline{t}})) \propto t^{-\frac{D}{2}} &\int_{\lfloor K_{\underline{t}} \rfloor}^{K_{\underline{t}}} r^{D-1} \mathrm{e}^{-\frac{r^2}{2t}} \, \mathrm{d}r \\
&\lesssim t^{-\frac{D}{2}} K_{\underline{t}}^D \mathrm{e}^{-\frac{K_{\underline{t}}^2}{2t}} \\
&\lesssim K^{\frac{D}{2}} \mathrm{e}^{-K},
\end{aligned}$$

and hence

$$\int_{\mathbb{R}^D \setminus [-\lfloor K_{\underline{t}} \rfloor, \lfloor K_{\underline{t}} \rfloor + 1]^D} (2\pi t)^{-\frac{D}{2}} \mathrm{e}^{-\frac{|x-y|^2}{2t}} \, \mathrm{d}x \lesssim K^{\frac{D}{2}} \mathrm{e}^{-K}.$$

By the above, this implies

$$\mathbb{E}\Big[|s_0^K(\mathsf{X}_t, t)|^2 \frac{\big(p_t(\mathsf{X}_t) - p_t^K(\mathsf{X}_t)\big)^2}{p_t(\mathsf{X}_t)^2} \Big] \lesssim \frac{1}{t^2 \wedge 1} K^{\frac{D}{2}} \mathrm{e}^{-K}.$$

As for the second term of (C.83), we again note that by following the proof of Lemma C.53.(a), we have $\frac{|\nabla(p_t(x) - p_t^K(x))|}{p_t(x)} \lesssim \frac{1}{t \wedge 1}$, whence we have similar to before

$$\mathbb{E}\Big[\frac{|\nabla(p_t(\mathsf{X}_t) - p_t^K(\mathsf{X}_t))|^2}{p_t(\mathsf{X}_t)^2} \Big] \lesssim \frac{1}{t \wedge 1} \int_{[0,1]^D} \big|\nabla\big(p_t(x) - p_t^K(x)\big)\big| \, \mathrm{d}x.$$

Furthermore, we have by the triangle inequality and similar calculations as before

$$\begin{aligned}
\int_{[0,1]^D} \big|\nabla\big(p_t(x) - p_t^K(x)\big)\big| \, \mathrm{d}x &\leq (2\pi t)^{-\frac{D}{2}} \sum_{\substack{z \in \mathbb{Z}^D \\ \|z\|_\infty > K_{\underline{t}}}} \int_{[0,1]^D} \int_{[0,1]^D} \frac{|R_z(x) + z - y|}{t} \mathrm{e}^{-\frac{|R_z(x)+z-y|^2}{2t}} \, \mu(\mathrm{d}y) \, \mathrm{d}x \\
&= \int_{[0,1]^D} \Big(\int_{\mathbb{R}^D \setminus [-K_{\underline{t}}, K_{\underline{t}}+1]^D} (2\pi t)^{-\frac{D}{2}} \frac{|x-y|}{t} \mathrm{e}^{-\frac{|x-y|^2}{2t}} \, \mathrm{d}x \Big) \, \mu(\mathrm{d}y),
\end{aligned}$$

where we have by Lemma C.71.(b)

$$\begin{aligned}
\int_{\mathbb{R}^D \setminus [-K_{\underline{t}}, K_{\underline{t}}+1]^D} (2\pi t)^{-\frac{D}{2}} \frac{|x-y|}{t} \mathrm{e}^{-\frac{|x-y|^2}{2t}} \, \mathrm{d}x &\leq \frac{1}{t} \mathbb{E}[|\mathsf{B}_t| \mathbf{1}_{\{|\mathsf{B}_t|_\infty \geq K_{\underline{t}}\}}] \\
&\lesssim \frac{1}{\sqrt{t}} K^{\frac{D}{2}} \mathrm{e}^{-K}.
\end{aligned}$$

Inserting this into the above, we have

$$\mathbb{E}\Big[\frac{|\nabla(p_t(\mathsf{X}_t)-p_t^K(\mathsf{X}_t))|^2}{p_t(\mathsf{X}_t)^2}\Big] \lesssim \frac{1}{t^2\wedge 1}K^{\frac{D}{2}}\mathrm{e}^{-K},$$

which combined with the above and (C.83) yields the result. □

Step 2: approximation of $f_1(\cdot,t), f_2(\cdot,t)$ for fixed t We start with the small time regime.

Lemma C.61: Under assumptions $(\mathcal{H}1)$–$(\mathcal{H}3)$, for large enough $m\in\mathbb{N}$ and fixed $t\le m^{-\frac{2-\delta}{d}}$ there exist neural networks

$$\varphi_{1,t}\in\widetilde{\Phi}(\log m, m, m\log m, m^{\nu}) \quad \text{and} \quad \varphi_{2,t}\in\widetilde{\Phi}(\log m, m, m\log m, t^{-\frac{1}{2}}\vee m^{\nu}),$$

where $\nu=\frac{2d}{2\alpha-d}+\frac{1}{d}$ such that for $u\in\mathbb{R}^d$,

$$|(2\pi t)^{-\frac{d}{2}}f_1(u,t)-\varphi_{1,t}(u)|\lesssim\begin{cases} m^{-\frac{\alpha}{d}}, & \text{if } u\in M^*_{-\varepsilon_M/2}\\ (\log m)^{\frac{d}{2}}m^{-\frac{\kappa}{d}}, & \text{if } u\notin M^*_{-\varepsilon_M/2}\end{cases}$$

and

$$|(2\pi t)^{-\frac{d}{2}}f_2(u,t)-\varphi_{2,t}(u)|\lesssim\begin{cases} \frac{1}{\sqrt{t}}m^{-\frac{\alpha}{d}}, & \text{if } u\in M^*_{-\varepsilon_M/2}\\ \frac{1}{\sqrt{t}}(\log m)^{\frac{d+1}{2}}m^{-\frac{\kappa}{d}}, & \text{if } u\notin M^*_{-\varepsilon_M/2}\end{cases}$$

where f_1, f_2 are as in (C.76).

Proof. To construct the neural networks $\varphi_{i,t}$, we first construct separate networks $\varphi_{i,t}^{(1)}$ and $\varphi_{i,t}^{(2)}$ corresponding to the cases where either $u\in M^*_{-\varepsilon_M/2}$ or $u\notin M^*_{-\varepsilon_M/2}$, in the latter case utilizing the increased smoothness near the boundary ∂M of M per assumption $(\mathcal{H}3)$, which we then stitch together into one network. To this end, we first establish the following common notation: let $q(u)=p_0(Au+v_0)$ and $n_t(u)=(2\pi t)^{-\frac{d}{2}}\mathrm{e}^{-\frac{|u|^2}{2t}}$, such that $(2\pi t)^{-\frac{d}{2}}f_1(u)=n_t * q(u)$, while $(2\pi t)^{-\frac{d}{2}}f_2(u)=\nabla n_t * q(u)$. Suppose then that $\mathrm{dist}(u,M^*)>\varepsilon_M/2$. Since $t\le m^{-\frac{2-\delta}{d}}$, we have $t(d+2\frac{\kappa}{d}\log m)\to 0$ for $m\to\infty$, so we may assume that m is large enough that $\sqrt{t(d+2\frac{\kappa}{d}\log m)}\le\varepsilon_M/2$, and it follows by Lemma C.71 that

$$|n_t * q(u)|\le p_{\max}\mathbb{P}\left(|\mathsf{B}_t^*|>\sqrt{t\Big(d+2\frac{\kappa}{d}\log m\Big)}\right)\lesssim(\log m)^{\frac{d}{2}}m^{-\frac{\kappa}{d}}$$

and

$$|\nabla n_t * q(u)|\le\frac{p_{\max}}{t}\mathbb{E}\big[|\mathsf{B}_t^*|\mathbf{1}_{\{|\mathsf{B}_t^*|>\sqrt{t(d+2\frac{\kappa}{d}\log m)}\}}\big]\lesssim\frac{1}{\sqrt{t}}(\log m)^{\frac{d+1}{2}}m^{-\frac{\kappa}{d}},$$

where $(\mathsf{B}_t^*)_{t\geq 0}$ is a d-dimensional Brownian motion. Thus for $u \notin M^*_{-\varepsilon_M/2}$, it suffices to approximate $n_t * q$ and $\nabla n_t * q$ on the set $(\partial M^*)_{3\varepsilon_M/4} \supset (\partial M^*)_{\varepsilon_M/2}$. Now since both $M^*_{-\varepsilon_M/2}$ and $(\partial M^*)_{3\varepsilon_M/4}$ are compact and have Lipschitz boundary, it follows by Lemma C.69 that there exists neural networks $\widetilde{\varphi}_{1,t}^{(i)} \in \widetilde{\Phi}(\log m, m, m\log m, C_{i,1} \vee m^\nu)$ and $\widetilde{\varphi}_{2,t}^{(i),j} \in \widetilde{\Phi}(\log m, m, m\log m, C_{i,2} \vee m^\nu)$ for $i = 1, 2$ and $j \in [d]$ such that

$$|\varphi_{1,t}^{(i)}(u) - n_t * q(u)| \lesssim \begin{cases} \|n_t * q\|_{H^\alpha(M^*_{-\varepsilon_M/2})} m^{-\frac{\alpha}{d}}, & \text{if } u \in M^*_{-\varepsilon_M/2}, i = 1 \\ \|n_t * q\|_{H^\kappa((\partial M^*)_{3\varepsilon_M/4})} m^{-\frac{\kappa}{d}}, & \text{if } u \in (\partial M^*)_{3\varepsilon/4}, i = 2 \end{cases}$$

and

$$|\varphi_{2,t}^{(i),j}(u) - \partial_j n_t * q(u)| \lesssim \begin{cases} \|\partial_j n_t * q\|_{H^\alpha(M^*_{-\varepsilon_M/2})} m^{-\frac{\alpha}{d}}, & \text{if } u \in M^*_{-\varepsilon_M/2}, i = 1 \\ \|\partial_j n_t * q\|_{H^\kappa((\partial M^*)_{3\varepsilon_M/4})} m^{-\frac{\kappa}{d}}, & \text{if } u \in (\partial M^*)_{3\varepsilon/4}, i = 2, \end{cases}$$

and where

$$C_{i,j} = \begin{cases} \|n_t * q\|_{H^\alpha(M^*_{-\varepsilon_M/2})}, & \text{if } (i,j) = (1,1) \\ \|n_t * q\|_{H^\kappa((\partial M^*)_{3\varepsilon_M/4})}, & \text{if } (i,j) = (2,1) \\ \|\partial_j n_t * q\|_{H^\alpha(M^*_{-\varepsilon_M/2})}, & \text{if } (i,j) = (1,2) \\ \|\partial_j n_t * q\|_{H^\kappa((\partial M^*)_{3\varepsilon_M/4})}, & \text{if } (i,j) = (2,2). \end{cases}$$

In order to bound $\|n_t * q\|_{H^\alpha(M^*_{-\varepsilon_M/2})}$, first note that by Assumption $(\mathcal{H}2)$ we may extend q to a global α-Sobolev function on $\mathbb{R}^d$ with compact support by simply setting $q(u) = 0$ for $u \notin M^*$, i.e. we can assume $q \in H_c^\alpha(\mathbb{R}^d)$. It then follows by Young's convolution inequality that

$$\begin{aligned} \|n_t * q\|^2_{H^\alpha(\mathbb{R}^d)} &= \sum_{|\beta|\leq\alpha} \|\partial^\beta(n_t * q)\|^2_{L^2(\mathbb{R}^d)} = \sum_{|\beta|\leq\alpha} \|n_t * (\partial^\beta q)\|^2_{L^2(\mathbb{R}^d)} \\ &\leq \|n_t\|^2_{L^1(\mathbb{R}^d)} \sum_{|\beta|\leq\alpha} \|\partial^\beta q\|^2_{L^2(\mathbb{R}^d)} = \|q\|^2_{H^\alpha(\mathbb{R}^d)}, \end{aligned} \tag{C.84}$$

and hence also $\|n_t * q\|_{H^\alpha(M^*_{-\varepsilon_M/2})} \lesssim 1$. Similarly,

$$\|\partial_j n_t * q\|_{H^\alpha(M^*_{-\varepsilon_M/2})} \lesssim \|\partial_j n_t\|_{L^1(\mathbb{R}^d)} = \int_{\mathbb{R}^d} (2\pi t)^{-\frac{d}{2}} \frac{|u_j|}{t} \mathrm{e}^{-\frac{|u|^2}{2t}} \,\mathrm{d}u = \sqrt{\frac{2}{\pi t}} \int_0^\infty \frac{r}{t} \mathrm{e}^{-\frac{r^2}{2t}} \,\mathrm{d}r = \sqrt{\frac{2}{\pi t}}, \tag{C.85}$$

and setting $(\varphi_{2,t}^{(1)})_j = \varphi_{2,t}^{(1),j}$ yields the desired networks on $M^*_{-\varepsilon_M/2}$.

Next, to bound $\|n_t * q\|_{H^\kappa((\partial M^*)_{3\varepsilon_M/4})}$, fix $u \in (\partial M^*)_{3\varepsilon/4}$ and $\beta \in \mathbb{N}_0^d$ with $|\beta| \leq \kappa$. Then we have

$$(\partial^\beta n_t) * q(u) = \int_{B(u,\varepsilon_M/4)} \partial^\beta n_t(u - v) q(v) \,\mathrm{d}v + \int_{B(u,\varepsilon_M/4)^\mathrm{c}} \partial^\beta n_t(u - v) q(v) \,\mathrm{d}v := I_1(u) + I_2(u),$$

where we note that in the first integral, $u - v \in (\partial M^*)_{\varepsilon_M}$ for all $v \in B(u, \varepsilon_M/4)$, whence q here is in C^κ. Thus we have by integration by parts for $i \in [d]$

$$\begin{aligned}\int_{B(u,\varepsilon_M/4)} \partial_{u_i} n_t(u-v)q(v)\,\mathrm{d}v &= -\int_{B(u,\varepsilon_M/4)} \partial_{v_i} n_t(u-v)q(v)\,\mathrm{d}v \\ &= \int_{B(u,\varepsilon_M/4)} n_t(u-v)\partial_{v_i} q(v)\,\mathrm{d}v \\ &\qquad - \frac{4}{\varepsilon_M}\int_{\partial B(u,\varepsilon_M/4)} n_t(u-v)q(v)(v_i - u_i)\,\mathcal{H}^{d-1}(\mathrm{d}v),\end{aligned}$$

where we use that $\frac{4(v_i-u_i)}{\varepsilon_M}$ is the outward pointing normal vector of $B(u, \varepsilon_M/4)$ at $v \in \partial B(u, \varepsilon_M/4)$. Repeating this and setting $\beta = \beta^{(1)} + \beta^{(2)} + \ldots + \beta^{(|\beta|)}$ with $|\beta^{(i)}| = 1$, we have

$$I_1(u) = \int_{B(u,\varepsilon/4)} n_t(u-v)\partial^\beta q(v)\,\mathrm{d}v - \frac{4}{\varepsilon_M}\sum_{i=1}^{|\beta|} B_i(u),$$

where

$$B_i(u) = \int_{\partial B(u,\varepsilon_M/4)} \partial^{\beta - \sum_{j=1}^{i}\beta^{(j)}} n_t(u-v)\partial^{\sum_{j=1}^{i-1}\beta^{(j)}} q(v)(v-u)^{\beta^{(i)}}\,\mathcal{H}^{d-1}(\mathrm{d}v).$$

Clearly we have

$$\left|\int_{B(u,\varepsilon/4)} n_t(u-v)\partial^\beta q(v)\,\mathrm{d}v\right| \le \sup_{|\beta|\le\kappa}\ \sup_{u\in(\partial M^*)_{\varepsilon_M}} |\partial^\beta q(u)|$$

independently of β, while

$$|B_i(u)| \le \sup_{|\beta|\le\kappa}\|\partial^\beta q\|_\infty \int_{\partial B(u,\varepsilon_M/4)} \left|\partial^{\beta-\sum_{j=1}^{i}\beta^{(j)}} n_t(u-v)(v-u)^{\beta^{(i)}}\right|\mathcal{H}^{d-1}(\mathrm{d}v).$$

Here, the integrand is of the form $\mathrm{Poly}(t^{-1})\mathrm{Poly}((u-v))\mathrm{e}^{-\frac{|u-v|^2}{2t}} \lesssim t^{-(d+|\beta|)}\mathrm{e}^{-\frac{\varepsilon_M^2}{32t}}$, and is hence uniformly bounded by $\sup_{t>0} t^{-(d+\kappa)}\mathrm{e}^{-\frac{\varepsilon_M^2}{32t}} < \infty$, ultimately implying that $|I_1(u)| \lesssim 1$. The same is true of the integrand in $I_2(u)$, implying that also $|I_2(u)| \lesssim 1$. Putting things together, we have

$$\begin{aligned}\|n_t * q\|_{H^\kappa((\partial M^*)_{3\varepsilon_M/4})} &= \sum_{|\beta|\le\kappa} \|\partial^\beta(n_t * q)\|_{L^2((\partial M^*)_{3\varepsilon_M/4})} \\ &= \sum_{|\beta|\le\kappa} \|(\partial^\beta n_t) * q\|_{L^2((\partial M^*)_{3\varepsilon_M/4})} \\ &\lesssim \sup_{|\beta|\le\kappa}\ \sup_{u\in(\partial M^*)_{3\varepsilon_M/4}}\ \sup_{t>0} |(\partial^\beta n_t) * q(u)| \\ &\lesssim 1.\end{aligned}$$

A similar analysis can be used to bound $\|\partial_j n_t * q\|_{H^\kappa((\partial M^*)_{3\varepsilon_M/4})}$, the only difference being that we can no longer move all derivatives from n_t to q using integration by parts as this would require q to be locally $C^{\kappa+1}$. In particular, following the same notation as before, we find that

$$|I_1(u)| \lesssim \left| \int_{B(u,\varepsilon_M/4)} \partial_j n_t(u-v)\partial^\beta q(v)\,\mathrm{d}v \right| + 1 \lesssim \|\partial_j n_t\|_{L^1(\mathbb{R}^d)} \lesssim \frac{1}{\sqrt{t}},$$

implying in the same way as before that $\|\partial_j n_t * q\|_{H^\kappa((\partial M^*)_{3\varepsilon_M/4})} \lesssim \frac{1}{\sqrt{t}}$. Once again setting $(\varphi_{2,t}^{(2)})_j = \varphi_{2,t}^{(2),j}$ yields the desired networks on $(\partial M^*)_{\varepsilon_M/2}$. Now, on the overlap $M^*_{-\varepsilon_M/2} \cap (\partial M^*)_{3\varepsilon_M/4}$, both $\varphi_{1,t}^{(i)}$ approximate $n_t * q$ at the desired rate, and hence the same is true of any convex combination of the two. In particular, if we let $\varphi_{\mathbf{1}_{M^*_{-3\varepsilon_M/4}}}$ and $\varphi_{\mathbf{1}_{M^*_{\varepsilon_M/4}}}$ be as in Lemma C.68 with $\varepsilon = \varepsilon_M/4$, it follows that

$$\widetilde{\varphi}_{1,t} := \varphi_{\mathbf{1}_{M^*_{\varepsilon_M/4}}} \left(\varphi_{1,t}^{(1)} \varphi_{\mathbf{1}_{M^*_{-3\varepsilon_M/4}}} + \varphi_{1,t}^{(2)} \left(1 - \varphi_{\mathbf{1}_{M^*_{-3\varepsilon_M/4}}}\right)\right)$$

has the desired error rate for all $u \in \mathbb{R}^d$. Setting

$$\varphi_{1,t} := \varphi_\ell^{\mathrm{mult}}\left(\varphi_{\mathbf{1}_{M^*_{\varepsilon_M/4}}}, \varphi_\ell^{\mathrm{mult}}\left(\varphi_{\mathbf{1}_{M^*_{-3\varepsilon_M/4}}}, \varphi_{1,t}^{(1)}\right) + \varphi_\ell^{\mathrm{mult}}\left(1 - \varphi_{\mathbf{1}_{M^*_{-3\varepsilon_M/4}}}, \varphi_{1,t}^{(2)}\right)\right)$$

with $\ell = \lceil \frac{\kappa}{d} \log m \rceil$ yields the desired network, as the error and network size stemming from the multiplication networks are negligible compared to the rest. The exact same method can be used to construct $\varphi_{2,t}$, finishing the proof. ☐

Having approximated f_1 and f_2 for fixed small t, we now use the induced smoothness of the forward process to approximate these for fixed large t.

Lemma C.62: Under assumptions $(\mathcal{H}1)$ and $(\mathcal{H}2)$, for $\delta > 0$, large enough $m \in \mathbb{N}$ and fixed $t > 0$ with $\frac{1}{2}m^{-\frac{2-\delta}{d}} < t \lesssim \log m$, there exists neural networks

$$\varphi_{1,t}, \varphi_{2,t} \in \widetilde{\Phi}(\log m', m', m' \log m', m'),$$

where $m' = (t \wedge 1)^{-\frac{d}{2}} m^{\frac{\delta}{2}}$ such that for $u \in \mathbb{R}^d$ and $i = 1, 2$

$$|(2\pi t)^{-\frac{d}{2}} f_i(u,t) - \varphi_{i,t}(u)| \lesssim m^{-\frac{\kappa+1}{d}}$$

where f_1 and f_2 are as in (C.76).

Proof. We start by bounding the Sobolev norm of Gaussian densities. As in the previous proof, let $n_t(u) = (2\pi t)^{-\frac{d}{2}} \mathrm{e}^{-\frac{|u|^2}{2t}}$ denote the density of $N(0, tI_d)$. Also, for $s \in \mathbb{R}$, let $\psi(s) = \mathrm{e}^{-s^2}$ and $\eta_t(s) = \frac{s}{\sqrt{2t}}$ such that $n_t(u) = (2\pi t)^{-\frac{d}{2}} \prod_{i=1}^d \psi \circ \eta_t(u_i)$. Thus, for $\beta \in \mathbb{N}_0^d$,

$$\partial^\beta n_t(u) = (2\pi t)^{-\frac{d}{2}} \prod_{i=1}^d \left(\frac{\mathrm{d}^{\beta_i}}{\mathrm{d}s^{\beta_i}} \psi \circ \eta_t \right)(u_i) = (2\pi t)^{-\frac{d}{2}} \prod_{i=1}^d (2t)^{-\frac{\beta_i}{2}} \left(\frac{\mathrm{d}^{\beta_i}}{\mathrm{d}s^{\beta_i}} \psi \right)\left(\eta_t(u_i)\right).$$

Also, for any function $g \in L^2(\mathbb{R})$,

$$\|g \circ \eta_t\|_{L^2}^2 = \int_{\mathbb{R}} g\big(\eta_t(s)\big)^2 \,\mathrm{d}s = \sqrt{2t} \int_{\mathbb{R}} g(r)^2 \,\mathrm{d}r = \sqrt{2t}\|g\|_{L^2}^2.$$

Combining these, we have that

$$\begin{aligned}
\|\partial^\beta n_t\|_{L^2}^2 &= \int_{\mathbb{R}^d} (2\pi t)^{-d} \prod_{i=1}^{d} (2t)^{-\beta_i} \Big(\frac{\mathrm{d}^{\beta_i}}{\mathrm{d}s^{\beta_i}} \psi \Big) \big(\eta_t(u_i)\big)^2 \,\mathrm{d}u \\
&= \pi^{-d} (2t)^{-(d+|\beta|)} \prod_{i=1}^{d} \Big\| \Big(\frac{\mathrm{d}^{\beta_i}}{\mathrm{d}s^{\beta_i}} \psi \Big) \circ \eta_t \Big\|_{L^2}^2 \\
&= \pi^{-d} (2t)^{-\frac{d+2|\beta|}{2}} \prod_{i=1}^{d} \Big\| \frac{\mathrm{d}^{\beta_i}}{\mathrm{d}s^{\beta_i}} \psi \Big\|_{L^2}^2,
\end{aligned}$$

implying that $\|\partial^\beta n_t\|_{L^2}^2 = t^{-\frac{d+2|\beta|}{2}} \|\partial^\beta n_1\|_{L^2}^2$, and hence for $\gamma \in \mathbb{N}_0$

$$\|n_t\|_{H^\gamma} = \sqrt{\sum_{|\beta| \le \gamma} \|\partial^\beta n_t\|_{L^2}^2} \le (t \wedge 1)^{-\frac{d+2\gamma}{4}} \|n_1\|_{H^\gamma}. \tag{C.86}$$

Next, in order to actually approximate $n_t * q$ and $\nabla n_t * q$, we first restrict the set on which to approximate these in order to apply Lemma C.69. To this end, let c^*, r^* be as in the previous proof and set $\rho_{t,\gamma} = \sqrt{t(d + 2\frac{\gamma}{d} \log m)}$ for $\gamma \ge 0$, and note that for $u \in \mathbb{R}^d$ with $\|u - c^*\|_\infty > r^* + \rho_{t,\gamma}$, we have $\operatorname{dist}(u, M^*) > \rho_{t,\gamma}$ and hence by Lemma C.71 we have

$$n_t * q(u) \lesssim \Big(\frac{\gamma}{d} \log m \Big)^{\frac{d}{2}} m^{-\frac{\gamma}{d}} \quad \text{and} \quad |\nabla n_t * q(u)| \lesssim \frac{1}{\sqrt{t}} \Big(\frac{\gamma}{d} \log m \Big)^{\frac{d+1}{2}} m^{-\frac{\gamma}{d}},$$

whence we need only approximate $n_t * q$ and $\nabla n_t * q$ on $[-(r^* + \rho_{t,\gamma}), r^* + \rho_{t,\gamma}]^d + c^*$. As such, let $\widetilde{\varphi}_1, \widetilde{\varphi}_{2,j} \in \widetilde{\Phi}(\gamma^2 \log m', \gamma^2 m', \gamma^4 m' \log m', (m')^\nu)$ where $\nu = \frac{2d}{2\gamma - d} + \frac{1}{d}$ be such that

$$\begin{aligned}
|\widetilde{\varphi}_1(u) - n_t * q(u)| &\lesssim (1 + \rho_{t,\gamma})^{\gamma - \frac{d}{2}} \|n_t * q\|_{H^\gamma} (m')^{-\frac{\gamma}{d}} \qquad \text{and} \\
|\widetilde{\varphi}_{2,j}(u) - \partial_j n_t * q(u)| &\lesssim (1 + \rho_{t,\gamma})^{\gamma - \frac{d}{2}} \|\partial_j n_t * q\|_{H^\gamma} (m')^{-\frac{\gamma}{d}}
\end{aligned}$$

for all u with $\|u - c^*\|_\infty \le r^* + \rho_{t,\gamma} + 1$ in accordance with Lemma C.69. Then, letting $\ell = \lceil \frac{\gamma}{d} \log_2 m \rceil \asymp \gamma \log m$ and $\varphi_{\rho_{t,\gamma}} = (1 \wedge (r^* + \rho_{t,\gamma} + 1 - \|u - c^*\|_\infty)) \vee 0$, set

$$\varphi_{1,t}(u) = \varphi_\ell^{\mathrm{mult}}(\varphi_{\rho_{t,\gamma}}(u), \widetilde{\varphi}_1(u)) \quad \text{and} \quad \varphi_{2,t}(u) = \varphi_\ell^{\mathrm{mult},d}(\varphi_{\rho_{t,\gamma}}(u), \widetilde{\varphi}_2(u)),$$

where $(\widetilde{\varphi}_2)_j = \widetilde{\varphi}_{2,j}$. Once again, as the sizes of the multiplication networks and $\varphi_{\rho_{t,\gamma}}$ are negligible compared to those of $\widetilde{\varphi}_i$, it follows that also $\varphi_{1,t}, \varphi_{2,t} \in \widetilde{\Phi}(\gamma^2 \log m', \gamma^2 m', \gamma^4 m' \log m', (m')^\nu)$, while

$$|\varphi_{1,t}(u) - n_t * q(u)| \lesssim \Big(\Big(\frac{\gamma}{d}\log m\Big)^{\frac{d}{2}} \vee \big((1+\rho_{t,\gamma})^{\gamma-\frac{d}{2}}\|n_t * q\|_{H^\gamma}\big)\Big)(m')^{-\frac{\gamma}{d}} \quad \text{and}$$
$$|\varphi_{2,t}(u) - \nabla n_t * q(u)| \lesssim \Big(\frac{1}{\sqrt{t}}\Big(\frac{\gamma}{d}\log m\Big)^{\frac{d+1}{2}} \vee \big((1+\rho_{t,\gamma})^{\gamma-\frac{d}{2}}\|n_t * q\|_{H^{\gamma+1}}\big)\Big)(m')^{-\frac{\gamma}{d}},$$

where we use that $\|\partial_j n_t * q\|_{H^\gamma} \leq \|n_t * q\|_{H^{\gamma+1}}$. Since $t \lesssim \log m$, it follows that

$$|\varphi_{1,t}(u) - n_t * q(u)| \lesssim \mathrm{Poly}(\log m)\|n_t * q\|_{H^\gamma}(m')^{-\frac{\gamma}{d}},$$

where as in (C.84) we have by Young's inequality and (C.86) that

$$\|n_t * q\|_{H^\gamma} \leq \|q\|_{L^1}\|n_t\|_{H^\gamma} \lesssim (t \wedge 1)^{-\frac{d+2\gamma}{4}}.$$

Inserting the definition of m' and using the assumption that $t > \frac{1}{2}m^{-\frac{2-\delta}{d}}$ we see

$$\begin{aligned}\|n_t * q\|_{H^\gamma}(m')^{-\frac{\gamma}{d}} &\lesssim (t \wedge 1)^{-\frac{d+2\gamma}{4}}\big((t \wedge 1)^{-\frac{d}{2}}m^{\frac{\delta}{2}}\big)^{-\frac{\gamma}{d}} \\ &= (t \wedge 1)^{-\frac{d}{4}}m^{-\frac{\delta\gamma}{2d}} \\ &\lesssim m^{\frac{2-\delta}{4}-\frac{\gamma\delta}{2d}}.\end{aligned}$$

Similarly, we have

$$|\varphi_{2,t}(u) - \nabla n_t * q(u)| \lesssim \mathrm{Poly}(\log m)m^{\frac{2-\delta}{4}\frac{d+2}{d}-\frac{\gamma\delta}{2d}}.$$

Setting $\gamma = \lceil\frac{2d}{\delta}(\frac{2-\delta}{4}\frac{d+2}{d} + \frac{\kappa+1}{d})\rceil$, it follows that

$$|\varphi_{2,t}(u) - \nabla n_t * q(u)| \lesssim \big(\mathrm{Poly}(\log m)m^{-\frac{\delta}{2d}}\big)m^{-\frac{\kappa+1}{d}} \lesssim m^{-\frac{\kappa+1}{d}},$$

and hence also $|\varphi_{1,t}(u) - n_t * q(u)| \lesssim m^{-\frac{\kappa+1}{d}}$ as desired. Notice also that since $\kappa \geq 2d$, we have $\gamma > 5d$ and hence $\nu = \frac{2d}{2\gamma-d} + \frac{1}{d} \leq 1$. ☐

Step 3: extend fixed time approximations to time intervals

Lemma C.63: Under assumptions $(\mathcal{H}1)$–$(\mathcal{H}3)$, for $\delta > 0$, large enough $m \in \mathbb{N}$ and $\underline{t} > 0$ with $m^{-\frac{2\alpha+2}{2\alpha+d}} \lesssim \underline{t} \lesssim \log m$ there exists neural networks

$$\varphi_1, \varphi_2 \in \begin{cases}\widetilde{\Phi}(\log m \log\log m, m\log m, m\log^2 m, m^\nu \vee \underline{t}^{-1}), & \text{if } \underline{t} \leq \frac{1}{2}m^{-\frac{2-\delta}{d}} \\ \widetilde{\Phi}(\log m \log\log m, m'\log m, m'\log^2 m, m'), & \text{if } \underline{t} > \frac{1}{2}m^{-\frac{2-\delta}{d}}\end{cases}$$

where $\nu = \frac{2d}{2\alpha - d} + \frac{1}{d}$ and $m' = (\underline{t} \wedge 1)^{-\frac{d}{2}} m^{\frac{\delta}{2}}$ such that for $u \in \mathbb{R}^d$ and $t \in [\underline{t}, 2\underline{t}]$,

$$|(2\pi t)^{-\frac{d}{2}} f_1(u,t) - \varphi_1(u,t)| \lesssim \begin{cases} (\log m) m^{-\frac{\alpha}{d}}, & \text{if } \underline{t} \leq \frac{1}{2} m^{-\frac{2-\delta}{d}}, u \in M^*_{-\varepsilon_M/2} \\ (\log m)^{\frac{d+2}{2}} m^{-\frac{\kappa}{d}}, & \text{if } \underline{t} \leq \frac{1}{2} m^{-\frac{2-\delta}{d}}, u \notin M^*_{-\varepsilon_M/2} \\ (\log m) m^{-\frac{\kappa+1}{d}}, & \text{if } \underline{t} > \frac{1}{2} m^{-\frac{2-\delta}{d}} \end{cases}$$

and

$$|(2\pi t)^{-\frac{d}{2}} f_2(u,t) - \varphi_2(u,t)| \lesssim \begin{cases} \frac{1}{\sqrt{\underline{t} \wedge 1}} (\log m) m^{-\frac{\alpha}{d}}, & \text{if } \underline{t} \leq \frac{1}{2} m^{-\frac{2-\delta}{d}}, u \in M^*_{-\varepsilon_M/2} \\ \frac{1}{\sqrt{\underline{t} \wedge 1}} (\log m)^{\frac{d+3}{2}} m^{-\frac{\kappa}{d}}, & \text{if } \underline{t} \leq \frac{1}{2} m^{-\frac{2-\delta}{d}}, u \notin M^*_{-\varepsilon_M/2} \\ \frac{1}{\sqrt{\underline{t} \wedge 1}} (\log m) m^{-\frac{\kappa+1}{d}}, & \text{if } \underline{t} > \frac{1}{2} m^{-\frac{2-\delta}{d}} \end{cases}$$

where f_1, f_2 are as in (C.76).

Proof. We start by constructing networks with the desired approximation rates and consider their sizes at the end. To this end, for notation, let

$$\varepsilon_1(u,t) = \begin{cases} m^{-\frac{\alpha}{d}}, & \text{if } t \leq \frac{1}{2} m^{-\frac{2-\delta}{d}}, u \in M^*_{-\varepsilon_M/2} \\ (\log m)^{\frac{d}{2}} m^{-\frac{\kappa}{d}}, & \text{if } t \leq \frac{1}{2} m^{-\frac{2-\delta}{d}}, u \notin M^*_{-\varepsilon_M/2} \\ m^{-\frac{\kappa+1}{d}}, & \text{if } t > \frac{1}{2} m^{-\frac{2-\delta}{d}} \end{cases}$$

and

$$\varepsilon_2(u,t) = \begin{cases} \frac{1}{\sqrt{\underline{t} \wedge 1}} m^{-\frac{\alpha}{d}}, & \text{if } t \leq \frac{1}{2} m^{-\frac{2-\delta}{d}}, u \in M^*_{-\varepsilon_M/2} \\ \frac{1}{\sqrt{\underline{t} \wedge 1}} (\log m)^{\frac{d+1}{2}} m^{-\frac{\kappa}{d}}, & \text{if } t \leq \frac{1}{2} m^{-\frac{2-\delta}{d}}, u \notin M^*_{-\varepsilon_M/2} \\ \frac{1}{\sqrt{\underline{t} \wedge 1}} m^{-\frac{\kappa+1}{d}}, & \text{if } t > \frac{1}{2} m^{-\frac{2-\delta}{d}} \end{cases}$$

and for $t > 0$ let $\varphi_{i,t}$ denote either the networks in Lemma C.61 if $t \leq \frac{1}{2} m^{-\frac{2-\delta}{d}}$ or those in Lemma C.62 if $t > \frac{1}{2} m^{-\frac{2-\delta}{d}}$. In either case, we have $|(2\pi t)^{-\frac{d}{2}} f_i(u,t) - \varphi_{i,t}| \lesssim \varepsilon_i(u,t)$. Also, as in the previous proofs, let $q(u) = p_0(Au + v_0)$ and $n_t(u) = (2\pi t)^{-\frac{d}{2}} \mathrm{e}^{-\frac{|u|^2}{2t}}$, such that $(2\pi t)^{-\frac{d}{2}} f_1 = n_t * q$ and $(2\pi t)^{-\frac{d}{2}} f_2 = \nabla n_t * q$. The idea of the proof is, as in the proof of [58, Lemma 3.13], to use polynomial interpolation in time between φ_{1,t_i} and φ_{2,t_i} for appropriate time points $\{t_i\}$. Since the time dependence of both $n_t * q$ and $\nabla n_t * q$ are well-behaved, for any fixed $u \in \mathbb{R}^d$, the functions $t \mapsto n_t * q(u)$ and $t \mapsto \nabla n_t * q(u)$ can be efficiently approximated by polynomial interpolation, and this property carries over to the neural network approximations, as we will show.

To this end, we first center our time interval, as this makes analysis easier, so let $a = \frac{1}{2}\underline{t}$ and $b = \frac{3}{2}\underline{t}$, and set $n^*_t(u) = n_{at+b}(u)$ for $t \in (-1,1)$ such that $n^*_{(-1,1)}(u) = n_{(\underline{t},\overline{t})}(u)$. Then, for some $k \in \mathbb{N}$ to be determined later, let $\{t_i\}_{i=0}^k = \{\cos \frac{i\pi}{k}\}_{i=0}^k$ be the first $k+1$ Chebyshev nodes on

$(-1,1)$. Then, for $i = 0, \dots, k$, let $p_i(t) = \prod_{j \neq i}(t - t_j)$ and set $c_i = \frac{1}{p_i(t_i)}$. Furthermore, set

$$\begin{aligned}
\varphi_1^*(u,t) &= \sum_{i=0}^{k} c_i \varphi_\ell^{\mathrm{mult}}\big(\varphi_\ell^{p_i}(t), \varphi_{1,t_i}^*(u)\big), \\
\psi(u,t) &= \sum_{i=0}^{k} c_i p_i(t) \varphi_{1,t_i}^*(u), \qquad \text{and} \\
P(u,t) &= \sum_{i=0}^{k} c_i p_i(t) n_{t_i}^* * q(u).
\end{aligned}$$

Here, $\varphi_{1,t}^* = \varphi_{1,at+b}$, while $\varphi_\ell^{\mathrm{mult}}$ is as in Lemma C.64 and $\varphi_\ell^{p_i}$ is a neural network approximations of p_i. In particular, we can construct $\varphi_\ell^{p_i} \in \widetilde{\Phi}(\ell \log k, k, k\ell, 1)$ such that

$$|\varphi_\ell^{p_i}(t) - p_i(t)| \lesssim k 2^{-\ell}, \qquad \forall t \in [-1,1].$$

We defer this construction to (the proof of) [58, Lemma 3.13]. We then find by the triangle inequality that

$$|\varphi_1^*(u,t) - n_t^* * q(u)| \leq |\varphi_1^*(u,t) - \psi(u,t)| + |\psi(u,t) - P(u,t)| + |P(u,t) - n_t^* * q(u)|, \tag{C.87}$$

and so setting $\varphi_1(u,t) = \varphi_1^*(u, \frac{2}{t}t - 3)$, the error analysis is completed if we can show that each of the above terms can be bounded by $\varepsilon_1(u,t) \log m$. Recalling from the proof of Lemma C.61 that $\|\varphi_{1,t}\|_\infty \leq p_{\max}$, we find that

$$\begin{aligned}
\big|\varphi_\ell^{\mathrm{mult}}\big(\varphi_\ell^{p_i}(t), \varphi_{1,t_i}^*(u)\big) - p_i(t)\varphi_{1,t_i}^*(u)\big| &\leq \big|\varphi_\ell^{\mathrm{mult}}\big(\varphi_\ell^{p_i}(t), \varphi_{1,t_i}^*(u)\big) - \varphi_\ell^{p_i}(t)\varphi_{1,t_i}^*(u)\big| \\
&\quad + p_{\max}|\varphi_\ell^{p_i}(t) - p_i(t)| \\
&\lesssim k 2^{-\ell}.
\end{aligned}$$

Furthermore, by [122, Theorem 5.2], it holds that $|c_i| \leq \frac{2^{k-1}}{k}$, and so the first term of (C.87) is upper bounded by

$$|\varphi_1^*(u,t) - \psi(u,t)| \leq \sum_{i=0}^{k} |c_i| |\varphi_{\ell_1}^{\mathrm{mult}}\big(\varphi_{\ell_2}^{p_i}(t), \varphi_{1,t_i}^*(u)\big) - p_i(t)\varphi_{1,t_i}^*(u)| \lesssim k 2^{k-\ell}$$

and choosing $\ell = \lceil k + \log_2 k + \frac{\kappa+1}{d} \log_2 m \rceil \asymp k + \log m$ bounds this term by $m^{-\frac{\kappa+1}{d}} \leq \varepsilon_1(u,t)$. For the second term of (C.87), it can be shown that $|p_i(t) c_i| \lesssim 1$ (see Appendix), whence

$$|\psi(u,t) - P(u,t)| \leq \sum_{i=0}^{k} |c_i p_i(t)| |\varphi_{1,t_i}^*(u) - n_{t_i}^* * q(u)| \lesssim k \varepsilon_1(u,t).$$

Finally, for the third term of (C.87), we start by showing that for each fixed $u \in \mathbb{R}^d$, the function $t \mapsto n_t * q(u)$ is analytically extendable to $\mathbb{C}^+ := \{w \in \mathbb{C} \mid \Re w > 0\}$. To this end, we first see that $t \mapsto n_t(u)$ is analytic on $\mathbb{C}^+$ for all $u \in \mathbb{R}^d$ as the composition of an analytic function with a rational function with a pole at 0. Thus, for each $w_0 \in \mathbb{C}^+$, there exists an open neighbourhood D_0 of w_0 and integrable functions $\{a_n\}_{n \in \mathbb{N}_0}$ such that

$$n_w(u) = \sum_{n=0}^{\infty} a_n(u)(w - w_0)^n, \qquad \forall w \in D_0,$$

where this sum converges uniformly and absolutely on D_0. Since q is a probability density, it then follows by dominated convergence that for $w \in D_0$

$$n_w * q(u) = \int_{M^*} \Big(\sum_{n=0}^{\infty} a_n(u - v)(w - w_0)^n \Big) q(v) \,\mathrm{d}v = \sum_{n=0}^{\infty} \Big(\int_{M^*} a_n(u - v) q(v) \,\mathrm{d}v \Big)(w - w_0)^n,$$

showing that $n_t * q$ is analytic as claimed. It then follows by [122, Theorem 8.2] that for $\rho > 1$ satisfying $b - a(\frac{\rho + \rho^{-1}}{2}) > 0$, we have

$$|P(u, t) - n_t^* * q(u)| \leq \frac{4R_\rho(u)\rho^{-k}}{\rho - 1}, \qquad \forall t \in (-1, 1),$$

where

$$R_\rho(u) = \max_{w \in \partial E_\rho} |n_w^*(u)|, \quad \text{and} \quad \partial E_\rho = \left\{ \frac{w + w^{-1}}{2} \mid |w| = \rho \right\}.$$

We claim that $\rho = 2$ works for our purposes. Indeed, we have $b - a(\frac{2 + 2^{-1}}{2}) = \frac{7}{8}\underline{t} > 0$, and since one readily checks that

$$\min_{w \in \partial E_2} \frac{a \Re w + b}{|aw + b|^2} = \frac{1}{\frac{5}{4}a + b} = \frac{8}{17\underline{t}},$$

we find that for $u \in \mathbb{R}^d$ and $w \in \partial E_2$

$$|n_w^*(u)| = \left|(2\pi w)^{-\frac{d}{2}} \mathrm{e}^{-\frac{|u|^2}{2(aw+b)}}\right| = (2\pi|w|)^{-\frac{d}{2}} \mathrm{e}^{-\frac{|u|^2}{2} \cdot \frac{a\Re w + b}{|aw+b|^2}} \leq \Big(\frac{14\pi}{8}\underline{t}\Big)^{-\frac{d}{2}} \mathrm{e}^{-\frac{4|u|^2}{17\underline{t}}},$$

whence

$$\begin{aligned} R_2(u) &\leq \int_{M^*} \Big(\frac{7\pi}{4}\underline{t}\Big)^{-\frac{d}{2}} \mathrm{e}^{-\frac{4|u-v|^2}{17\underline{t}}} q(v) \,\mathrm{d}v \\ &= \Big(\frac{17}{7}\Big)^{\frac{d}{2}} \int_{M^*} \Big(2\pi\Big(\frac{17}{8}\underline{t}\Big)\Big)^{-\frac{d}{2}} \mathrm{e}^{-\frac{|u-v|^2}{2(\frac{17}{8}\underline{t})}} q(v) \,\mathrm{d}v \\ &= \Big(\frac{17}{7}\Big)^{\frac{d}{2}} n_{\frac{17}{8}\underline{t}} * q(u). \end{aligned}$$

Now by Young's convolution inequality, we find that

$$\|R_2\|_{L^\infty} \leq \Big(\frac{17}{7}\Big)^{\frac{d}{2}} \|n_{\frac{17}{8}\underline{t}}\|_{L^1} \|q\|_{L^\infty} \leq \Big(\frac{17}{7}\Big)^{\frac{d}{2}} p_{\max},$$

whereby

$$|P(u,t) - n_t^* * q(u)| \lesssim 2^{-k},$$

and choosing $k = \lceil \frac{\kappa+1}{d} \log m \rceil \asymp \log m$ ensures that this is bounded by $m^{-\frac{\kappa+1}{d}} \leq \varepsilon_1(u,t)$.

The exact same strategy can be used to approximate $\nabla n_t * q$, i.e. if we set in a recycling of notation

$$\begin{aligned} \varphi_2^*(u,t) &= \sum_{i=0}^{k} c_i \varphi_\ell^{\mathrm{mult},d}\big(\varphi_\ell^{p_i}(t), \varphi_{2,t_i}^*(u)\big), \\ \psi(u,t) &= \sum_{i=0}^{k} c_i p_i(t) \varphi_{2,t_i}^*(u), \qquad \text{and} \\ P(u,t) &= \sum_{i=0}^{k} c_i p_i(t) \nabla n_{t_i}^* * q(u), \end{aligned}$$

then we once again have by the triangle inequality

$$|\varphi_2^*(u,t) - \nabla n_t^* * q(u)| \leq |\varphi_2^*(u,t) - \psi(u,t)| + |\psi(u,t) - P(u,t)| + |P(u,t) - \nabla n_t^* * q(u)|. \tag{C.88}$$

Here, using the exact same approach as before, we can bound the first two terms by $\varepsilon_2(u,t)$. As for the third term, noting that $|P(u,t) - \nabla n_t * q(u)|^2 = \sum_{j=1}^{d} |P_j(u,t) - \partial_j n_t * q(u)|^2$, we can use the same method as before to bound each of these summands. In particular, following the same steps as above and defining $R_{2,j}$ as above for the j'th summand, we would find that

$$\|R_{2,j}\|_{L^\infty} \leq \left(\frac{17}{7}\right)^{\frac{d}{2}} \|\partial_j n_{\frac{17}{8}\underline{t}}\|_{L^1} \|q\|_{L^\infty},$$

and thus by (C.85),

$$|P(u,t) - \nabla n_t * q(u)|^2 \lesssim 2^{-2k} \sum_{j=1}^{d} \|\partial_j n_{\frac{17}{8}\underline{t}}\|_{L^1}^2 \lesssim \frac{1}{\underline{t}} m^{-\frac{2(\kappa+1)}{d}} \leq \varepsilon_2(u,t)^2.$$

As for the sizes of the networks, we first recall that if $\underline{t} \leq \frac{1}{2} m^{-\frac{2-\delta}{d}}$, we have for each i that $\varphi_{1,t_i} \in \widetilde{\Phi}(\log m, m, m \log m, m^\nu \vee (\underline{t} \log m)^{-\frac{1}{2}})$, while $\varphi_\ell^{p_i} \in \widetilde{\Phi}(\log m \log\log m, \log m, \log^2 m, 1)$, whereby each summand in the definition of φ_1^* is in $\widetilde{\Phi}(\log m \log\log m, m, m \log m, m^\nu \vee (\underline{t} \log m)^{-\frac{1}{2}})$, and since there are $k \asymp \log m$ such terms, we have $\varphi_1 \in \widetilde{\Phi}(\log m \log\log m, m \log m, m \log^2 m, m^\nu \vee \underline{t}^{-1})$. Similarly, if $\underline{t} > \frac{1}{2} m^{-\frac{2-\delta}{d}}$, we have $\varphi_{1,t_i} \in \widetilde{\Phi}(\log m', m', m' \log m', m')$, and hence by the same argumentation $\varphi_1 \in \widetilde{\Phi}(\log m \log\log m, m', m' \log m', m')$. Similar analysis shows that the same is true of φ_2, finishing the proof. ☐

Step 4: Putting things together

With all of the above we are now in a position to prove Theorem C.58.

Proof of Theorem C.58. First, by Lemmas C.53.(b) and C.60 and the triangle inequality we need only approximate $s_0^K \mathbf{1}_{M_{\rho,\underline{t}}}$ with

$$\rho = K = \frac{2(\alpha+1)}{d} \log m + \log \underline{t}^{-1} \lesssim \log m.$$

To this end, we let for sake of notation $x^* = A^\top(x - v_0) \in \mathbb{R}^d$ and $x^\perp = (I - P)(x - v_0) \in \mathbb{R}^D$ for $x \in \mathbb{R}^D$ such that x^* is the local coordinates of the projection Px of x onto M, while $x^\perp$ is the perpendicular component of x with respect to M. Recalling then that

$$\int_M \mathrm{e}^{-\frac{|x-y|^2}{2t}} \, \mu(\mathrm{d}y) = \mathrm{e}^{-\frac{|x^\perp|^2}{2t}} \int_{M^*} \mathrm{e}^{-\frac{|x^*-u|}{2t}} \, p_0(Au + v_0) \, \mathrm{d}u = \mathrm{e}^{-\frac{|x^\perp|^2}{2t}} f_1(x^*, t),$$

and

$$\begin{aligned} \int_M \frac{x-y}{t} \mathrm{e}^{-\frac{|x-y|^2}{2t}} \, \mu(\mathrm{d}y) &= \mathrm{e}^{-\frac{|x^\perp|^2}{2t}} \left(\frac{x^\perp}{t} \int_{M^*} \mathrm{e}^{-\frac{|x^*-u|^2}{2t}} \, p_0(Au + v_0) \, \mathrm{d}u \right. \\ &\qquad \left. + A \int_{M^*} \frac{x^* - u}{t} \mathrm{e}^{-\frac{|x^*-u|^2}{2t}} \, p_0(Au + v_0) \, \mathrm{d}u \right) \\ &= \mathrm{e}^{-\frac{|x^\perp|^2}{2t}} \left(\frac{x^\perp}{t} f_1(x^*, t) + A f_2(x^*, t) \right), \end{aligned}$$

we let

$$h_1(x,t) = (2\pi t)^{-\frac{d}{2}} \mathrm{e}^{-\frac{|x^\perp|^2}{2t}} f_1(x^*, t) \quad \text{and} \quad h_2(x,t) = (2\pi t)^{-\frac{d}{2}} \mathrm{e}^{-\frac{|x^\perp|^2}{2t}} \big(\frac{x^\perp}{t} f_1(x^*, t) + A f_2(x^*, t) \big)$$

such that

$$s_0^K(x,t) = \frac{-\sum_{\substack{z \in \mathbb{Z}^D \\ \|z\|_\infty \le K_{\underline{t}}}} (-1)^z h_2(R_z(x) + z, t)}{\sum_{\substack{z \in \mathbb{Z}^D \\ \|z\|_\infty \le K_{\underline{t}}}} h_1(R_z(x) + z, t)},$$

where $K_{\underline{t}} = \sqrt{2\underline{t}(D + 2K)}$. Furthermore, let

$$\widehat{h}_1(x,t) = \mathrm{e}^{-\frac{|x^\perp|^2}{2t}} \varphi_1(x^*, t) \quad \text{and} \quad \widehat{h}_2(x,t) = \mathrm{e}^{-\frac{|x^\perp|^2}{2t}} \big(\frac{x^\perp}{t} \varphi_1(x^*, t) + A \varphi_2(x^*, t) \big),$$

where φ_1, φ_2 are as in Lemma C.63. By (the proof of) Lemma C.53.(d) there exists a constant $c > 0$ such that for all $(x,t) \in M_{\rho,\underline{t}} \times [\underline{t}, 2\underline{t}]$,

$$p_t^K(x) \ge c t^{\frac{c_0 - D}{2}} \mathrm{e}^{-\rho} = c t^{\frac{c_0-D}{2}+1} m^{-\frac{2(\alpha+1)}{d}}, \tag{C.89}$$

and we set

$$\widehat{s}_0^K(x,t) = \frac{-\sum_{\substack{z \in \mathbb{Z}^D \\ \|z\|_\infty \le K_{\underline{t}}}} (-1)^z \widehat{h}_2(R_z(x) + z, t)}{\Big(\sum_{\substack{z \in \mathbb{Z}^D \\ \|z\|_\infty \le K_{\underline{t}}}} \widehat{h}_1(R_z(x) + z, t) \Big) \vee c \underline{t}^{\frac{c_0-d}{2}+1} m^{-\frac{2(\alpha+1)}{d}}} =: \frac{\widehat{\nabla p_t}^K(x)}{\widehat{p}_t^K(x)}.$$

To clearly separate our sources of error, we first show that $\widehat{s}_0^K$ is a good approximation of s_0^K on $M_{\rho,\underline{t}} \times [\underline{t}, 2\underline{t}]$ and only then approximate $\widehat{s}_0^K$ by a neural network. To this end, we have first

$$\begin{aligned} |s_0^K(x,t) - \widehat{s}_0^K(x,t)| &= \left|\frac{\widehat{\nabla p}_t^K(x)}{\widehat{p}_t^K(x)} - \frac{(2\pi t)^{\frac{D-d}{2}}\nabla p_t^K(x)}{(2\pi t)^{\frac{D-d}{2}} p_t^K(x)}\right| \\ &\le \frac{1}{p_t^K(x)\widehat{p}_t^K(x)}\Big(|\nabla p_t^K(x)|\big|(2\pi t)^{\frac{D-d}{2}} p_t^K(x) - \widehat{p}_t^K(x)\big| \\ &\qquad\qquad + p_t^K(x)\big|(2\pi t)^{\frac{D-d}{2}}\nabla p_t^K(x) - \widehat{\nabla p}_t^K(x)\big|\Big) \\ &= \frac{1}{\widehat{p}_t^K(x)}\Big(|s_0^K(x,t)|\big|(2\pi t)^{\frac{D-d}{2}} p_t^K(x) - \widehat{p}_t^K(x)\big| + \big|(2\pi t)^{\frac{D-d}{2}}\nabla p_t^K(x) - \widehat{\nabla p}_t^K(x)\big|\Big). \end{aligned}$$

Then by (C.89),

$$|(2\pi t)^{\frac{D-d}{2}} p_t^K(x) - \widehat{p}_t^K(x)| \lesssim \Big|(2\pi t)^{\frac{D-d}{2}} p_t^K(x) - \sum_{\substack{z\in\mathbb{Z}^D \\ \|z\|_\infty \le K_{\underline{t}}}} \widehat{h}_1(R_z(x)+z,t)\Big|.$$

Next, we further split this error into four parts to analyse separately. In particular, let $S_1 = \{x \in \mathbb{R}^D \mid x^* \in M^*_{-\varepsilon_M/2}\}$ and $S_2 = \mathbb{R}^D \setminus S_1$. Furthermore, for $x \in [0,1]^D$, and $z \in \mathbb{Z}^D$ let $x_z = R_z(x) + z$ and set $Z_i^K(x) = \{z \in \mathbb{Z}^D \mid \|z\|_\infty \le K_{\underline{t}}, x_z \in S_i\}$. Repeated use of the triangle inequality (along with the fact that A and $(-1)^z$ are orthogonal matrices) then yields that the distance $|s_0^K(x,t) - \widehat{s}_0^K(x,t)|$ is upper bounded by the sum

$$\frac{1}{\widehat{p}_t^K(x)}|s_0^K(x,t)| \sum_{z\in Z_1^K(x)} e^{-\frac{|x_z^\perp|^2}{2t}} |(2\pi t)^{-\frac{d}{2}} f_1(x_z^*,t) - \varphi_1(x_z^*,t)| \tag{C.90}$$

$$+ \frac{1}{\widehat{p}_t^K(x)}|s_0^K(x,t)| \sum_{z\in Z_2^K(x)} e^{-\frac{|x_z^\perp|^2}{2t}} |(2\pi t)^{-\frac{d}{2}} f_1(x_z^*,t) - \varphi_1(x_z^*,t)| \tag{C.91}$$

$$+ \frac{1}{\widehat{p}_t^K(x)} \sum_{z\in Z_1^K(x)} e^{-\frac{|x_z^\perp|^2}{2t}} \Big(\frac{|x_z^\perp|}{t}|(2\pi t)^{-\frac{d}{2}} f_1(x_z^*,t) - \varphi_1(x_z^*,t)| + |(2\pi t)^{-\frac{d}{2}} f_2(x_z^*,t) - \varphi_2(x_z^*,t)|\Big) \tag{C.92}$$

$$+ \frac{1}{\widehat{p}_t^K(x)} \sum_{z\in Z_2^K(x)} e^{-\frac{|x_z^\perp|^2}{2t}} \Big(\frac{|x_z^\perp|}{t}|(2\pi t)^{-\frac{d}{2}} f_1(x_z^*,t) - \varphi_1(x_z^*,t)| + |(2\pi t)^{-\frac{d}{2}} f_2(x_z^*,t) - \varphi_2(x_z^*,t)|\Big). \tag{C.93}$$

We will analyse each of these terms separately. To ease notation, let $\varepsilon_{\underline{t}}$ denote either $m^{-\frac{1}{d}}$ if $\underline{t} > m^{-\frac{2-\delta}{d}}$ and 1 otherwise.

Term (C.90): For $z \in Z_1^K(x)$, we have $|(2\pi t)^{-\frac{d}{2}} f_1(x_z^*,t) - \varphi_1(x_z^*,t)| \lesssim m^{-\frac{\alpha}{d}} \varepsilon_{\underline{t}} \log m$ by Lemma C.63, while also

$$f_1(x_z^*,t) \ge \int_{B(x_z^*,\sqrt{t})\cap M^*_{-\varepsilon_M/2}} e^{-\frac{|x_z^*-u|^2}{2t}} p_0(Au+v_0)\,\mathrm{d}u \ge e^{-\frac{1}{2}} p_{\min}\lambda_d(B(x_z^*,\sqrt{t})\cap M^*_{-\varepsilon_M/2}) \gtrsim (t\wedge r_0^2)^{\frac{d}{2}}$$

by assumption ($\mathcal{H}1$). This implies in particular that

$$\begin{aligned}
\widehat{p}_t^K(x) &\geq \sum_{z\in Z_1^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}} \varphi_1(x_z^*, t) \\
&\geq \sum_{z\in Z_1^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}} \left((2\pi t)^{-\frac{d}{2}} f_1(x_z^*, t) - |(2\pi t)^{-\frac{d}{2}} f_1(x_z^*, t) - \varphi_1(x_z^*, t)|\right) \\
&\gtrsim \sum_{z\in Z_1^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}} \left((1 \wedge t^{-\frac{d}{2}} r_0^d) - m^{-\frac{\alpha}{d}} \log m\right) \\
&\gtrsim (\log m)^{-\frac{d}{2}} \sum_{z\in Z_1^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}},
\end{aligned}$$

where we use that $t^{-1} \geq (2\underline{t})^{-1} \gtrsim (\log m)^{-1}$. Combining these, we have

$$\frac{1}{\widehat{p}_t^K(x)} |s_0^K(x,t)| \sum_{z\in Z_1^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}} |(2\pi t)^{-\frac{d}{2}} f_1(x_z^*, t) - \varphi_1(x_z^*, t)| \lesssim |s_0^K(x,t)| m^{-\frac{\alpha}{d}} \varepsilon_{\underline{t}} (\log m)^{\frac{d+2}{2}}.$$

Term (C.91): For $z \in Z_2^K(x)$, we can no longer lower bound $\widehat{p}_t^K(x)$ as we did above. Instead, we have by definition that $\widehat{p}_t^K(x) \geq \underline{t}^{\frac{c_0-d}{2}+1} m^{-\frac{2(\alpha+1)}{d}}$. Furthermore, using Lemma C.63,

$$|(2\pi t)^{-\frac{d}{2}} f_1(x_z^*, t) - \varphi_1(x_z^*, t)| \lesssim (\log m)^{\frac{d+2}{2}} m^{-\frac{\kappa}{d}} \varepsilon_{\underline{t}} \lesssim \underline{t}^{\frac{c_0-d}{2}+1} (\log m)^{\frac{d+2}{2}} m^{-\frac{3\alpha+2}{d}} \varepsilon_{\underline{t}},$$

where we used that $\underline{t} \gtrsim m^{-\frac{2\alpha+2}{2\alpha+d}} \geq m^{-1}$. Thus it follows that

$$\begin{aligned}
&\frac{1}{\widehat{p}_t^K(x)} |s_0^K(x,t)| \sum_{z\in Z_2^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}} |(2\pi t)^{-\frac{d}{2}} f_1(x_z^*, t) - \varphi_1(x_z^*, t)| \\
&\quad\lesssim \frac{m^{\frac{2(\alpha+1)}{d}}}{\underline{t}^{\frac{c_0-d}{2}+1}} |s_0^K(x,t)| \sum_{z\in Z_2^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}} \underline{t}^{\frac{c_0-d}{2}+1} (\log m)^{\frac{d+2}{2}} m^{-\frac{3\alpha+2}{d}} \varepsilon_{\underline{t}} \\
&\quad= |s_0^K(x,t)| (\log m)^{\frac{d+2}{2}} m^{-\frac{\alpha}{d}} \varepsilon_{\underline{t}} \sum_{z\in Z_2^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}} \\
&\quad\lesssim |s_0^K(x,t)| (\log m)^{\frac{d+2D+2}{2}} m^{-\frac{\alpha}{d}} \varepsilon_{\underline{t}},
\end{aligned}$$

where in the last step we use that

$$\# Z_2^K(x) \leq \#\{z \in \mathbb{Z}^D \mid \|z\|_\infty \leq K_{\underline{t}}\} \leq (2K_{\underline{t}} + 1)^D \lesssim (\log m)^D.$$

Term (C.92): Here again using Lemma C.63 we have $|(2\pi t)^{-\frac{d}{2}} f_1(x_z^*, t) - \varphi_1(x_z^*, t)| \lesssim m^{-\frac{\alpha}{d}} \varepsilon_{\underline{t}} \log m$ and $|(2\pi t)^{-\frac{d}{2}} f_2(x_z^*, t) - \varphi_2(x_z^*, t)| \lesssim \frac{1}{\sqrt{\underline{t}\wedge 1}} m^{-\frac{\alpha}{d}} \varepsilon_{\underline{t}} \log m$, and so by the exact same reasoning as in case (C.90),

$$\frac{1}{\widehat{p}_t^K(x)} \sum_{z\in Z_1^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}} |(2\pi t)^{-\frac{d}{2}} f_2(x_z^*, t) - \varphi_2(x_z^*, t)| \lesssim \frac{1}{\sqrt{\underline{t} \wedge 1}} m^{-\frac{\alpha}{d}} \varepsilon_{\underline{t}} (\log m)^{\frac{d+2}{2}}.$$

Also, like in case (C.90), we have

$$\frac{1}{\widehat{p}_t^K(x)} \sum_{z\in Z_1^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}} \frac{|x_z^\perp|}{t} |(2\pi t)^{-\frac{d}{2}} f_1(x_z^*, t) - \varphi_1(x_z^*, t)| \lesssim (\log m)^{\frac{d+2}{2}} m^{-\frac{\alpha}{d}} \varepsilon_{\underline{t}} \frac{\sum_{z\in Z_1^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}} \frac{|x_z^\perp|}{t}}{\sum_{z\in Z_1^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}}},$$

and to bound this, we first note that for all $z \in \mathbb{Z}^D$ with $\|z\|_\infty \leq K_{\underline{t}}$, we have $x_z \in [-K_{\underline{t}}, K_{\underline{t}}]^D$, whence $|x_z^\perp| \leq 2\sqrt{D} K_{\underline{t}} \lesssim \sqrt{t} \log m$ and so

$$(\log m)^{\frac{d+2}{2}} m^{-\frac{\alpha}{d}} \varepsilon_{\underline{t}} \frac{\sum_{z\in Z_1^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}} \frac{|x_z^\perp|}{t}}{\sum_{z\in Z_1^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}}} \lesssim \frac{1}{\sqrt{\underline{t} \wedge 1}} (\log m)^{\frac{d+4}{2}} m^{-\frac{\alpha}{d}} \varepsilon_{\underline{t}}.$$

Term (C.93): Repeating the arguments used in cases (C.91) and (C.92), we obtain

$$\begin{aligned} &\frac{1}{\widehat{p}_t^K(x)} \sum_{z\in Z_2^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}} |(2\pi t)^{-\frac{d}{2}} f_2(x_z^*, t) - \varphi_2(x_z^*, t)| \\ &\quad \lesssim \frac{1}{\sqrt{\underline{t} \wedge 1}} (\log m)^{\frac{d+3}{2}} m^{-\frac{\alpha}{d}} \varepsilon_{\underline{t}} \sum_{z\in Z_2^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}} \\ &\quad \lesssim \frac{1}{\sqrt{\underline{t} \wedge 1}} (\log m)^{\frac{d+2D+3}{2}} m^{-\frac{\alpha}{d}} \varepsilon_{\underline{t}}, \end{aligned}$$

and, since $\mathrm{e}^{-\frac{r}{2t}} \frac{r}{t} \leq \frac{1}{\sqrt{t}}$ for all $r > 0$,

$$\frac{1}{\widehat{p}_t^K(x)} \sum_{z\in Z_2^K(x)} \mathrm{e}^{-\frac{|x_z^\perp|^2}{2t}} \frac{|x_z^\perp|}{t} |(2\pi t)^{-\frac{d}{2}} f_1(x_z^*, t) - \varphi_1(x_z^*, t)| \lesssim \frac{1}{\sqrt{\underline{t} \wedge 1}} (\log m)^{\frac{d+2D+3}{2}} m^{-\frac{\alpha}{d}} \varepsilon_{\underline{t}}.$$

Combining all of these different cases, we see that for $x \in M_{\rho,\underline{t}}$

$$|s_0^K(x, t) - \widehat{s}_0^K(x, t)| \lesssim \Big(|s_0^K(x, t)| + \frac{1}{\sqrt{\underline{t} \wedge 1}}\Big) (\log m)^{\frac{d+2D+3}{2}} m^{-\frac{\alpha}{d}} \varepsilon_{\underline{t}}, \tag{C.94}$$

whence

$$\int_{\underline{t}}^{2\underline{t}} \mathbb{E}\big[|s_0^K(\mathsf{X}_t, t) - \widehat{s}_0^K(\mathsf{X}_t, t)|^2 \mathbf{1}_{M_{\rho,t}}(\mathsf{X}_t)\big] \,\mathrm{d}t \lesssim \Big(\int_{\underline{t}}^{2\underline{t}} \mathbb{E}[|s_0^K(\mathsf{X}_t, t)|^2] \,\mathrm{d}t + \frac{\underline{t}}{\underline{t} \wedge 1}\Big) (\log m)^{d+2D+3} m^{-\frac{2\alpha}{d}} \varepsilon_{\underline{t}}.$$

Now, to estimate the remaining integral, we first have by Lemma C.60 and our choice of K that

$$\begin{aligned} \int_{\underline{t}}^{2\underline{t}} \mathbb{E}[|s_0^K(\mathsf{X}_t, t)|^2] \,\mathrm{d}t &\leq 2 \int_{\underline{t}}^{2\underline{t}} \big(\mathbb{E}[|s_0(\mathsf{X}_t, t)|^2] + \mathbb{E}[|s_0(\mathsf{X}_t, t) - s_0^K(\mathsf{X}_t, t)|^2]\big) \,\mathrm{d}t \\ &\lesssim \int_{\underline{t}}^{2\underline{t}} \mathbb{E}[|s_0(\mathsf{X}_t, t)|^2] \,\mathrm{d}t + (\log m)^{\frac{D}{2}} m^{-\frac{2(\alpha+1)}{d}}. \end{aligned}$$

Furthermore, since $(x,t) \mapsto p_t(x)$ is a positive solution of the heat equation $\partial_t u(t,x) = \frac{1}{2}\Delta u(t,x)$ with Neumann boundary conditions on $M \times (0,\infty)$ for the compact, convex set $M = [0,1]^D$, the Li–Yau bound [74, Theorem 1.1] yields that

$$|s_0(x,t)|^2 = |\nabla \log p_t(x)|^2 \lesssim \partial_t \log p_t(x) + \frac{D}{t}, \quad (x,t) \in [0,1]^d \times (0,\infty).$$

Thus,

$$\int_{\underline{t}}^{2\underline{t}} \mathbb{E}[|s_0(\mathsf{X}_t,t)|^2]\,\mathrm{d}t \lesssim \int_{\underline{t}}^{2\underline{t}} \mathbb{E}\Big[\Big(\frac{\partial_t p_t(\mathsf{X}_t)}{p_t(\mathsf{X}_t)}\Big] + \frac{D}{t}\Big)\,\mathrm{d}t = D\int_{\underline{t}}^{2\underline{t}} \frac{1}{t}\,\mathrm{d}t = D\log 2,$$

where we use that by Fubini–Tonelli's theorem

$$\int_{\underline{t}}^{2\underline{t}} \mathbb{E}\Big[\frac{\partial_t p_t(\mathsf{X}_t)}{p_t(\mathsf{X}_t)}\Big]\,\mathrm{d}t = \int_{\underline{t}}^{2\underline{t}} \int_{[0,1]^D} \partial_t p_t(x)\,\mathrm{d}x\,\mathrm{d}t = \int_{[0,1]^D} \int_{\underline{t}}^{2\underline{t}} \partial_t p_t(x)\,\mathrm{d}t\,\mathrm{d}x = 0.$$

Inserting this into the above, we have that

$$\int_{\underline{t}}^{2\underline{t}} \mathbb{E}\big[|s_0^K(\mathsf{X}_t,t) - \widehat{s}_0^K(\mathsf{X}_t,t)|^2 \mathbf{1}_{M_{\rho,t}}(\mathsf{X}_t)\big]\,\mathrm{d}t \lesssim (\log m)^{d+2D+3} m^{-\frac{2\alpha}{d}}\,\varepsilon_{\underline{t}}^2, \tag{C.95}$$

as desired. With this established, all that is left is to approximate $\widehat{s}_0^K \mathbf{1}_{M_{\rho,\underline{t}}}$ by a neural network φ_{s_0}. To this end, letting $\varphi_\ell^{\mathrm{exp}}$, $\varphi_\ell^{\mathrm{mult}}$, $\varphi_\ell^{\mathrm{rec}}$ and $\varphi_\ell^{\mathrm{norm}}$ be as in Lemmas C.66, C.64, C.65 and C.67 and setting

$$\widehat{\varphi}_\ell^{\mathrm{exp}}(x,t) = \varphi_\ell^{\mathrm{exp}}\Big(\frac{1}{2}\varphi_\ell^{\mathrm{mult}}\big(\varphi_\ell^{\mathrm{rec}}(t), \varphi_\ell^{\mathrm{norm}}(x)\big)\Big),$$

we have

$$\begin{aligned}
\Big|\mathrm{e}^{-\frac{|x|^2}{2t}} - \widehat{\varphi}_\ell^{\mathrm{exp}}(x,t)\Big| &\leq 2^{-\ell} + \Big|\mathrm{e}^{-\frac{|x|^2}{2t}} - \mathrm{e}^{-\frac{1}{2}\varphi_\ell^{\mathrm{mult}}(\varphi_\ell^{\mathrm{rec}}(t),\varphi_\ell^{\mathrm{norm}}(x))}\Big| \\
&\lesssim 2^{-\ell} + \Big|\frac{|x|^2}{t} - \varphi_\ell^{\mathrm{mult}}\big(\varphi_\ell^{\mathrm{rec}}(t), \varphi_\ell^{\mathrm{norm}}(x)\big)\Big| \\
&\lesssim K_{\underline{t}} 2^{-\ell} + \Big|\frac{|x|^2}{t} - \varphi_\ell^{\mathrm{rec}}(t)\varphi_\ell^{\mathrm{norm}}(x)\Big| \\
&\lesssim K_{\underline{t}} 2^{-\ell} + K_{\underline{t}}^2\Big|\frac{1}{t} - \varphi_\ell^{\mathrm{rec}}(t)\Big| + \frac{1}{\underline{t}}\Big||x|^2 - \varphi_\ell^{\mathrm{norm}}(x)\Big| \\
&\lesssim \frac{K_{\underline{t}}^2}{\underline{t}} 2^{-\ell}
\end{aligned}$$

for all $x \in \mathbb{R}^D$ with $\|x\|_\infty \leq K_{\underline{t}}$ and all $t \in [\underline{t}, 2\underline{t}]$. Next, let

$$\begin{aligned}
\varphi_{h_1}(x,t) &= \varphi_\ell^{\mathrm{mult}}\big(\widehat{\varphi}_\ell^{\mathrm{exp}}(x^\perp,t), \varphi_1(x^*,t)\big), \\
\widetilde{\varphi}_1(x,t) &= \varphi_\ell^{\mathrm{mult},D}\big(\varphi_1(x^*,t), \varphi_\ell^{\mathrm{mult},D}(\varphi_\ell^{\mathrm{rec}}(t), x^\perp)\big), \quad \text{and} \\
\varphi_{h_2}(x,t) &= \varphi_\ell^{\mathrm{mult},D}\big(\widehat{\varphi}_\ell^{\mathrm{exp}}(x^\perp,t), \widetilde{\varphi}_1(x,t) + A\varphi_2(x^*,t)\big),
\end{aligned}$$

and it follows that

$$|\widehat{h}_1(x,t) - \varphi_{h_1}(x,t)| \lesssim |\varphi_1(x^*,t)|\Big(1 + \frac{K_{\underline{t}}^2}{\underline{t}}\Big)2^{-\ell} \lesssim (2\pi\underline{t})^{-\frac{d}{2}}\frac{K_{\underline{t}}^2}{\underline{t}}2^{-\ell},$$

where we once again use that we can assume $|\varphi_1| \le (2\pi\underline{t})^{-\frac{d}{2}}|f_1| \le (2\pi\underline{t})^{-\frac{d}{2}}$ by Lemma C.70. Similarly,

$$\Big|\widetilde{\varphi}_1(x,t) - \frac{x^{\perp}}{t}\varphi_1(x,t)\Big| \lesssim K_{\underline{t}}(2\pi\underline{t})^{-\frac{d}{2}}\Big(1 + \frac{1}{\underline{t}}\Big)2^{-\ell},$$

and so, since we may once again assume $|\widetilde{\varphi}_1 + A\varphi_2| \lesssim (2\pi\underline{t})^{-\frac{d}{2}}\frac{K_{\underline{t}}}{\underline{t}}$,

$$|\widehat{h}_2(x,t) - \varphi_{h_2}(x,t)| \lesssim (2\pi\underline{t})^{-\frac{d}{2}}\frac{K_{\underline{t}}^2}{\underline{t}}2^{-\ell}.$$

Setting $\varphi_{p_t^K}(x,t) = \sum_{\substack{z\in\mathbb{Z}^D\\ \|z\|_\infty\le K_{\underline{t}}}}\varphi_{h_1}(x_z,t)$ and $\varphi_{\nabla p_t^K}(x,t) = -\sum_{\substack{z\in\mathbb{Z}^D\\ \|z\|_\infty\le K_{\underline{t}}}}(-1)^z\varphi_{h_2}(x_z,t)$, it thus follows that

$$|\varphi_{p_t^K}(x,t) - \widehat{p}_t^K(x,t)|, |\varphi_{\nabla p_t^K}(x,t) - \widehat{\nabla p_t}^K(x,t)| \lesssim (2\pi\underline{t})^{-\frac{d}{2}}\frac{K_{\underline{t}}^{D+2}}{\underline{t}}2^{-\ell}.$$

Finally, let

$$\varphi_{s_0}(x,t) = \varphi_\ell^{\mathrm{mult},D}\Big(\varphi_\ell^{\mathrm{rec}}(\varphi_{p_t^K}(x,t) \vee \underline{t}^{\frac{c_0-d}{2}+1}m^{-\frac{2(\alpha+1)}{d}}), \varphi_{\nabla p_t^K}(x,t)\Big),$$

and we have for $x \in M_{\rho,\underline{t}}$ that

$$\begin{aligned}
&|\varphi_{s_0}(x,t) - \widehat{s}_0^K(x,t)|\\
&\quad\le |\varphi_{s_0}(x,t) - \varphi_\ell^{\mathrm{rec}}(\varphi_{p_t^K}(x,t) \vee c\underline{t}^{\frac{c_0-d}{2}+1}m^{-\frac{2(\alpha+1)}{d}})\varphi_{\nabla p_t^K}(x,t)|\\
&\qquad + |\varphi_{\nabla p_t^K}(x,t)|\Big|\varphi_\ell^{\mathrm{rec}}(\varphi_{p_t^K}(x,t) \vee c\underline{t}^{\frac{c_0-d}{2}+1}m^{-\frac{2(\alpha+1)}{d}}) - \frac{1}{\varphi_{p_t^K}(x,t) \vee c\underline{t}^{\frac{c_0-d}{2}+1}m^{-\frac{2(\alpha+1)}{d}}}\Big|\\
&\qquad + \Big|\frac{\varphi_{\nabla p_t^K}(x,t)}{\varphi_{p_t^K}(x,t) \vee c\underline{t}^{\frac{c_0-d}{2}+1}m^{-\frac{2(\alpha+1)}{d}}} - \frac{\widehat{\nabla p_t}^K(x)}{\widehat{p}_t^K(x)}\Big|.
\end{aligned}$$

Here, the first two terms together are bounded by $|\varphi_{\nabla p_t^K}(x,t)|(c\underline{t}^{-\frac{c_0-d}{2}-1}m^{\frac{2(\alpha+1)}{d}}+1)2^{-\ell}$, where again $|\varphi_{\nabla p_t^K}(x,t)| \lesssim (2\pi\underline{t})^{-\frac{d}{2}}\frac{K_{\underline{t}}^D}{\underline{t}}$ by Lemma C.70. For the third term, we have

$$\begin{aligned}
\Big|\frac{\varphi_{\nabla p_t^K}(x,t)}{\varphi_{p_t^K}(x,t) \vee c\underline{t}^{\frac{c_0-d}{2}+1}m^{-\frac{2(\alpha+1)}{d}}} - \frac{\widehat{\nabla p_t}^K(x)}{\widehat{p}_t^K(x,t)}\Big| &\le \frac{m^{\frac{2(\alpha+1)}{d}}}{c\underline{t}^{\frac{c_0-d}{2}+1}}\Big(|\widehat{s}_0^K(x,t)||\widehat{p}_t^K(x,t) - \varphi_{p_t^K}(x,t)|\\
&\qquad + |\widehat{\nabla p_t}^K(x,t) - \varphi_{\nabla p_t^K}(x,t)|\Big)\\
&\lesssim \frac{K_{\underline{t}}^{D+2}m^{\frac{2(\alpha+1)}{d}}}{\underline{t}^{\frac{c_0}{2}+2}}(|\widehat{s}_0^K(x,t)| + 1)2^{-\ell}.
\end{aligned}$$

Here, since $|s_0^K(x,t)| \lesssim \frac{1}{t}$ by (the proof of) Lemma C.53.(a) and $|s_0^K(x,t) - \widehat{s}_0^K(x,t)| \lesssim \frac{1}{t}$ by (C.94), we have also that $|\widehat{s}_0^K(x,t)| \lesssim \frac{1}{t}$ for $x \in M_{\rho,\underline{t}}$, and hence

$$\Big|\frac{\varphi_{\nabla p_t^K}(x,t)}{\varphi_{p_t^K}(x,t) \vee c\underline{t}^{\frac{c_0-d}{2}+1} m^{-\frac{2(\alpha+1)}{d}}} - \frac{\widehat{\nabla p_t}^K(x)}{\widehat{p}_t^K(x,t)}\Big| \lesssim \frac{K_{\underline{t}}^{D+2} m^{\frac{2(\alpha+1)}{d}}}{\underline{t}^{\frac{c_0}{2}+3}} 2^{-\ell},$$

whereby all in all

$$|\varphi_{s_0}(x,t) - \widehat{s}_0(x,t)| \lesssim \frac{K_{\underline{t}}^{D+2} m^{\frac{2(\alpha+1)}{d}}}{\underline{t}^{\frac{c_0}{2}+3}} 2^{-\ell} \lesssim (\log m)^{D+2} m^{\frac{2(\alpha+1)}{d}+\frac{c_0}{2}+3} 2^{-\ell}.$$

Thus, setting $\ell = \lceil (\frac{3(\alpha+1)}{d} + \frac{c_0}{2} + 3)\log_2 m + (D+2)\log_2 \log m \rceil \lesssim \log m$ ensures that this is bounded by $m^{-\frac{\alpha+1}{d}}$. As for $x \notin M_{\rho,t}$, we can once again assume by Lemmas C.70 and C.53.(e) that $|\varphi_{s_0}(x,t)| \lesssim \frac{\sqrt{\rho + \log \underline{t}^{-1}}}{\sqrt{\underline{t} \wedge 1}} \lesssim \frac{\sqrt{\log m}}{\sqrt{\underline{t} \wedge 1}}$ (since this is true of $s_0^K \mathbf{1}_{M_{\rho,t}}$), and so it follows by Lemma C.71.(a)

$$\begin{aligned}
\mathbb{E}[|\varphi_{s_0}(\mathsf{X}_t,t) - \widehat{s}_0^K(\mathsf{X}_t,t)\mathbf{1}_{M_{\rho,t}}(\mathsf{X}_t)|^2] &\lesssim m^{-\frac{2(\alpha+1)}{d}} \mathbb{P}(\mathsf{X}_t \in M_{\rho,t}) + \frac{\log m}{\underline{t} \wedge 1} \mathbb{P}(\mathsf{X}_t \notin M_{\rho,t}) \\
&\lesssim m^{-\frac{2(\alpha+1)}{d}} + \frac{\log m}{\underline{t} \wedge 1} \rho^{\frac{D}{2}} \mathrm{e}^{-\rho} \\
&\lesssim (\log m)^{\frac{D+2}{2}} m^{-\frac{2(\alpha+1)}{d}}.
\end{aligned}$$

Finally, combining this, (C.95) and Lemmas C.53.(b) and C.60 along with repeated use of the triangle inequality, we have that

$$\begin{aligned}
\int_{\underline{t}}^{2\underline{t}} \mathbb{E}[|\varphi_{s_0}(\mathsf{X}_t,t) - s_0(\mathsf{X}_t,t)|^2]\,\mathrm{d}t &\lesssim \int_{\underline{t}}^{2\underline{t}} \mathbb{E}[|\varphi_{s_0}(\mathsf{X}_t,t) - \widehat{s}_0^K(\mathsf{X}_t,t)\mathbf{1}_{M_{\rho,t}}(\mathsf{X}_t)|^2]\,\mathrm{d}t \\
&\quad + \int_{\underline{t}}^{2\underline{t}} \mathbb{E}\left[|\widehat{s}_0^K(\mathsf{X}_t,t) - s_0^K(\mathsf{X}_t,t)|^2 \mathbf{1}_{M_{\rho,t}}(\mathsf{X}_t)\right]\mathrm{d}t \\
&\quad + \int_{\underline{t}}^{2\underline{t}} \mathbb{E}\left[|s_0^K(\mathsf{X}_t,t) - s_0(\mathsf{X}_t,t)|^2 \mathbf{1}_{M_{\rho,t}}(\mathsf{X}_t)\right]\mathrm{d}t \\
&\quad + \int_{\underline{t}}^{2\underline{t}} \mathbb{E}[s_0(\mathsf{X}_t,t)\mathbf{1}_{M_{\rho,t}}(\mathsf{X}_t) - s_0(\mathsf{X}_t,t)|^2]\,\mathrm{d}t \\
&\lesssim (\log m)^{d+2D+3} m^{-\frac{2\alpha}{d}} \varepsilon_{\underline{t}}^2,
\end{aligned}$$

as desired. As for the size of the network, we first have that for our choice of ℓ (and recalling that all sizes being multiplied or divided are bounded by $\underline{t}^{\frac{d-c_0}{2}-1} m^{\frac{2(\alpha+1)}{d}} \lesssim m^{\frac{2(\alpha+1)}{d}+\frac{c_0-d}{2}+1}$)

$$\begin{aligned}
\varphi_\ell^{\mathrm{exp}} &\in \widetilde{\Phi}((\log m)^2(\log\log m)^2, \log m \log\log m, (\log m)^3(\log\log m)^3, 1)\\
\varphi_\ell^{\mathrm{mult}} &\in \widetilde{\Phi}(\log m, 1, \log m, m^{\frac{2(\alpha+1)}{d}+\frac{c_0-d}{2}+1})\\
\varphi_\ell^{\mathrm{rec}} &\in \widetilde{\Phi}(\log m \log\log m, \log m, \log m \log\log m, m^{\frac{2(\alpha+1)}{d}+\frac{c_0-d}{2}+1})\\
\varphi_\ell^{\mathrm{norm}} &\in \widetilde{\Phi}(\log m, 1, \log m, \log m),
\end{aligned}$$

whereby $\widehat{\varphi}_\ell^{\mathrm{exp}} \in \widetilde{\Phi}((\log m)^2(\log\log m)^2, \log m \log\log m, (\log m)^3(\log\log m)^3, m^{\frac{2(\alpha+1)}{d}+\frac{c_0-d}{2}+1})$. Since the size of $\varphi_\ell^{\mathrm{mult}}$ is comparably negligible to those of φ_1 and $\widehat{\varphi}_\ell^{\mathrm{exp}}$, this implies that

$$\varphi_{h_1} \in \begin{cases} \widetilde{\Phi}\big((\log m)^2(\log\log m)^2, m\log m, m\log^2 m, m^{\frac{2(\alpha+1)}{d}+\frac{c_0-d}{2}+1} \vee m^\nu\big), & \text{if } \underline{t} \le \frac{1}{2}m^{-\frac{2-\delta}{d}}\\ \widetilde{\Phi}\big((\log m)^2(\log\log m)^2, m'\log m, m'(\log m)^2, m^{\frac{2(\alpha+1)}{d}+\frac{c_0-d}{2}+1}\big), & \text{if } \underline{t} > \frac{1}{2}m^{-\frac{2-\delta}{d}} \end{cases}$$

and similar analysis shows that the same is true of φ_{h_2}. Finally, summing $\lceil K_{\underline{t}} \rceil^D \lesssim (\log m)^D$ copies of these yields that

$$\begin{bmatrix} \varphi_{p_t^K} \\ \varphi_{\nabla p_t^K} \end{bmatrix} \in \begin{cases} \widetilde{\Phi}\big((\log m)^2(\log\log m)^2, m(\log m)^{D+1}, m(\log m)^{D+2}, m^{\frac{2(\alpha+1)}{d}+\frac{c_0-d}{2}+1} \vee m^\nu\big), & \text{if } \underline{t} \le \frac{1}{2}m^{-\frac{2-\delta}{d}}\\ \widetilde{\Phi}\big((\log m)^2(\log\log m)^2, m'(\log m)^{D+1}, m'(\log m)^{D+2}, m^{\frac{2(\alpha+1)}{d}+\frac{c_0-d}{2}+1}\big), & \text{if } \underline{t} > \frac{1}{2}m^{-\frac{2-\delta}{d}} \end{cases}$$

and since the remaining networks are once again negligible to this, we have also

$$\varphi_{s_0} \in \begin{cases} \widetilde{\Phi}\big((\log m)^2(\log\log m)^2, m(\log m)^{D+1}, m(\log m)^{D+2}, m^{\frac{2(\alpha+1)}{d}+\frac{c_0-d}{2}+1} \vee m^\nu\big), & \text{if } \underline{t} \le \frac{1}{2}m^{-\frac{2-\delta}{d}}\\ \widetilde{\Phi}\big((\log m)^2(\log\log m)^2, m'(\log m)^{D+1}, m'(\log m)^{D+2}, m^{\frac{2(\alpha+1)}{d}+\frac{c_0-d}{2}+1}\big), & \text{if } \underline{t} > \frac{1}{2}m^{-\frac{2-\delta}{d}} \end{cases},$$

as desired. □

C.C Basic neural network approximation results

Lemma C.64 ([58] Lemma 3.10): For $m \in \mathbb{N}$ and $C \ge 1$, there exist neural networks $\varphi_m^{\mathrm{mult}} \in \widetilde{\Phi}(m, 1, m, C)$ and $\varphi_m^{\mathrm{mult},d} \in \widetilde{\Phi}(m, d, dm, C)$ satisfying

$$|\varphi_m^{\mathrm{mult}}(x, y) - xy| \le C2^{-m}, \quad x \in [0,1], y \in [-C, C],$$

and

$$|\varphi_m^{\mathrm{mult},d}(x, y) - xy| \le \sqrt{d}C2^{-m}, \quad x \in [0,1], y \in [-C, C]^d.$$

These also satisfy $\varphi_m^{\mathrm{mult}}(x, 0) = \varphi_m^{\mathrm{mult}}(0, y) = 0$.

Lemma C.65 ([58] Lemma 3.11): For $m, \underline{k}, \overline{k} \in \mathbb{N}$, there exists a neural network

$$\varphi_m^{\mathrm{rec}} \in \widetilde{\Phi}((k+m)\log(k+m), k, (k+m)\log(k+m), 2^k),$$

where $k = \underline{k} + \overline{k}$, satisfying

$$|\varphi_m^{\mathrm{rec}}(x) - x^{-1}| \le 2^{-m}, \qquad x \in [2^{-\underline{k}}, 2^{\overline{k}}].$$

Lemma C.66: For $m \in \mathbb{N}$ there exists a neural network

$$\varphi_m^{\exp} \in \widetilde{\Phi}(m^2 \log^2 m, m \log m, m^3 \log^3 m, 1)$$

satisfying

$$|\varphi_m^{\exp}(x) - \mathrm{e}^{-x}| \le 2^{-m}, \qquad x \ge 0.$$

Proof. First note that for $x \ge m \log 2 := K$, we have $\mathrm{e}^{-x} \le 2^{-m}$, and so we need find an approximation $\varphi_m^{\exp}$ satisfying $|\varphi_m^{\exp}(x) - \mathrm{e}^{-x}| \le 2^{-m}$ for $x \in [0, K]$ with $|\varphi_m^{\exp}(x)| \le 2^{-m}$ for $x \ge K$. To this end, note then that $\mathrm{e}^{-x} = \left(\mathrm{e}^{-\frac{x}{\lceil K \rceil}}\right)^{\lceil K \rceil}$, whereby we need only approximate e^{-x} and x^n for $x \in [0,1]$ and $n \in \mathbb{N}$. To this end, let $\varphi_{k_1}^{\mathrm{mult}}$ be as in Lemma C.64 with $C = 1$ for some $k_1 \in \mathbb{N}$ to be determined later and set $\varphi_{k_1}^{\mathrm{pow},n}(x) = \varphi_{k_1}^{\mathrm{mult}}(x, \varphi_{k_1})^{\mathrm{pow},n-1}$ for $n \ge 3$ with $\varphi_{k_1}^{\mathrm{pow},2}(x) = \varphi_{k_1}^{\mathrm{mult}}(x, x)$. Then, $\varphi_{k_1}^{\mathrm{pow},n} \in \widetilde{\Phi}(nk_1, 1, nk_1, 1)$, and we have

$$|\varphi_{k_1}^{\mathrm{pow},n}(x) - x^n| \le |\varphi_{k_1}^{\mathrm{pow},n}(x) - x\varphi_{k_1}^{\mathrm{pow},n-1}(x)| + x|\varphi_{k_1}^{\mathrm{pow},n-1}(x) - x^{n-1}| \le 2^{-k_1} + |\varphi_{k_1}^{\mathrm{pow},n-1}(x) - x^{n-1}|,$$

from which it follows by elementary induction that $|\varphi_{k_1}^{\mathrm{pow},n}(x) - x^n| \lesssim n2^{-k_1}$. Next, for some $k_2 \in \mathbb{N}$, also to be determined later, let

$$\widetilde{\varphi}_{k_1,k_2}^{\exp}(x) = 1 - x + \sum_{k=2}^{k_2} \frac{(-1)^k \varphi_{k_1}^{\mathrm{pow},k}(x)}{k!},$$

such that $\widetilde{\varphi}_{k_1,k_2}^{\exp} \in \widetilde{\Phi}(k_1 k_2, k_2, k_1 k_2^2, 1)$ and

$$|\widetilde{\varphi}_{k_1,k_2}^{\exp}(x) - \mathrm{e}^{-x}| \le \left|\mathrm{e}^{-x} - \sum_{k=0}^{k_2} \frac{(-x)^k}{k!}\right| + \sum_{k=2}^{k_2} \frac{|\varphi_{k_1}^{\mathrm{pow},k}(x) - x^k|}{k!} \lesssim \frac{1}{k_2!} + 2^{-k_1} \sum_{k=2}^{k_2} \frac{k-1}{k!} \le \frac{1}{k_2!} + 2^{-k_1}$$

for all $x \in [0,1]$. Setting $\overline{\varphi}_m^{\exp}(x) = \varphi_k^{\mathrm{pow},\lceil K \rceil} \circ \widetilde{\varphi}_{k,k}^{\exp}\left(\frac{x}{\lceil K \rceil}\right)$ with $k \ge m \log m \vee 4$ (ensuring $\frac{1}{k!} \le 2^{-k}$)

for $x \in [0, K]$, we then see that $\overline{\varphi}^{\text{exp}}_m \in \widetilde{\Phi}(m^2 \log^2 m, m \log m, m^3 \log^3 m, 1)$

$$\begin{aligned}
\left|\overline{\varphi}^{\text{exp}}_m(x) - \mathrm{e}^{-x}\right| &\le \left|\overline{\varphi}^{\text{exp}}_m(x) - \widetilde{\varphi}^{\text{exp}}_{k,k}\Big(\frac{x}{\lceil K\rceil}\Big)^{\lceil K\rceil}\right| + \left|\widetilde{\varphi}^{\text{exp}}_{k,k}\Big(\frac{x}{\lceil K\rceil}\Big)^{\lceil K\rceil} - \Big(\mathrm{e}^{-\frac{x}{\lceil K\rceil}}\Big)^{\lceil K\rceil}\right| \\
&\lesssim \lceil K\rceil\Big(2^{-k} + \left|\widetilde{\varphi}^{\text{exp}}_{k,k}\Big(\frac{x}{\lceil K\rceil}\Big) - \mathrm{e}^{-\frac{x}{\lceil K\rceil}}\right|\Big) \\
&\lesssim m2^{-k} \\
&\le 2^{-m}.
\end{aligned}$$

Finally, the function

$$p(x) = \begin{cases} 1, & \text{if } x \le K-1 \\ K-x, & \text{if } K-1 \le x \le K \\ 0, & \text{if } x \ge K \end{cases}$$

is exactly representable as a neural network, and setting $\varphi^{\text{exp}}_m(x) = \varphi^{\text{mult}}_m(p(x), \overline{\varphi}^{\text{exp}}_m(x))$ ensures that $|\varphi^{\text{exp}}_m(x)| \le 2^{-m}$ for $x \ge K$ without altering the asymptotic size of the network. ☐

Lemma C.67: For $m \in \mathbb{N}$ and $K \in \mathbb{N}_0$ there exists a neural network $\varphi^{\text{norm}}_m \in \widetilde{\Phi}(m, D, Dm, K)$ satisfying

$$\left|\varphi^{\text{norm}}_m(x) - |x|^2\right| \le DK^2 2^{-m}, \qquad \|x\|_\infty \le K.$$

Proof. A small modification of Lemma C.64 yields a network $\widetilde{\varphi}^{\text{mult}}_m \in \widetilde{\varphi}(m, 1, m, K)$ with

$$|\widetilde{\varphi}^{\text{mult}}_m(x, y) - xy| \le K^2 2^{-m}, \qquad x, y \in [-K, K].$$

Setting $\varphi^{\text{norm}}_m(x) = \sum_{i=1}^{D} \widetilde{\varphi}^{\text{mult}}_m(x_i, x_i)$, we have that $\varphi^{\text{norm}}_m \in \widetilde{\Phi}(m, D, Dm, K)$, and for all $x \in [-K, K]^D$, we have

$$\left|\varphi^{\text{norm}}_m(x) - |x|^2\right| \le \sum_{i=1}^{D} |\widetilde{\varphi}^{\text{mult}}_m(x_i, x_i) - x_i^2| \le DK^2 2^{-m}.$$

☐

Lemma C.68: For every compact set $K \subseteq \mathbb{R}^d$ with diameter $R > 0$ and every $\varepsilon > 0$, there exists a neural network $\varphi_{\mathbf{1}_K} \in \widetilde{\Phi}\big(\log \frac{R}{\varepsilon}, (\frac{R}{\varepsilon})^d, (\frac{R}{\varepsilon})^d, \frac{1}{\varepsilon}\big)$ satisfying $\varphi_{\mathbf{1}_K}(x) \in [0, 1]$ for all $x \in \mathbb{R}^d$, $\varphi_{\mathbf{1}_K}(K) = 1$ and $\varphi_{\mathbf{1}_K}(K_\varepsilon^{\mathrm{c}}) = 0$, where K_ε is the ε-fattening of K.

Proof. First, for $r > 0$, let $K_r^\infty = \{x \in \mathbb{R}^d : \exists y \in K \text{ s.t. } \|x - y\|_\infty < r\}$ denote the r-fattening of K with respect to $\|\cdot\|_\infty$. Since $|x| \leq \sqrt{d}\|x\|_\infty$, we then have that $K_\varepsilon^{\mathrm{c}} \subseteq (K_{\varepsilon'}^\infty)^{\mathrm{c}}$, where $\varepsilon' = \frac{\varepsilon}{\sqrt{d}}$, and so we need only find a neural network $\varphi_{\mathbf{1}_K}$ satisfying $\varphi_{\mathbf{1}_K}(x) \in [0,1]$ with $\varphi_{\mathbf{1}_K}(K) = 1$ and $\varphi_{\mathbf{1}_K}((K_{\varepsilon'}^\infty)^{\mathrm{c}}) = 0$. The reason for working with $\|\cdot\|_\infty$ rather than $|\cdot|$ is that while $|x|$ needs to be approximated by neural networks, $\|x\|_\infty$ is itself exactly representable as a neural network, as $a \vee b = a + \sigma(b - a)$ and $\|x\|_\infty = (-x_1 \vee x_1) \vee (-x_2 \vee x_2) \vee \ldots \vee (-x_d \vee x_d)$. Now, let $y_1, y_2, \ldots, y_N$ be a minimal $\frac{\varepsilon'}{4}$-covering of K with respect to $\|\cdot\|_\infty$, and set $\varphi^{\mathrm{dist}}(x) = \min_{i \in [N]} \|x - y_i\|_\infty$. Since also $a \wedge b = b - \sigma(b - a)$, φ^{dist} is also representable as a neural network. In particular, by using a divide and conquer strategy, we have that $\varphi^{\mathrm{dist}} \in \widetilde{\Phi}(\log N, N, N, 1)$, and we see that φ^{dist} satisfies $\varphi^{\mathrm{dist}}(x) > \frac{3\varepsilon'}{4}$ for $x \notin K_{\varepsilon'}^\infty$, while $\varphi^{\mathrm{dist}}(x) < \frac{\varepsilon'}{4}$ for $x \in K$. Lastly, set

$$\varphi_{\mathbf{1}_K}(x) = \left(1 \wedge \left(\frac{3}{2} - \frac{2\varphi^{\mathrm{dist}}(x)}{\varepsilon'}\right)\right) \vee 0,$$

and we see that $\varphi_{\mathbf{1}_K}$ satisfies our criteria and that $\varphi_{\mathbf{1}_K} \in \widetilde{\Phi}(\log N, N, N, \frac{1}{\varepsilon'})$. Finally, noting that since K is of diameter R and hence contained in $[0,R]^d + y_0$ for some $y_0 \in \mathbb{R}^d$, its covering number is less than that of $[0,R]^D$, i.e.

$$N = N\Big(K, \|\cdot\|_\infty, \frac{\varepsilon'}{4}\Big) \leq N\Big([0,R]^d, \|\cdot\|_\infty, \frac{\varepsilon'}{4}\Big) \leq \left\lceil \frac{4\sqrt{d}R}{\varepsilon} \right\rceil^d \lesssim \Big(\frac{R}{\varepsilon}\Big)^d,$$

whence $\varphi_{\mathbf{1}_K} \in \widetilde{\Phi}\big(\log \frac{R}{\varepsilon}, (\frac{R}{\varepsilon})^d, (\frac{R}{\varepsilon})^d, \frac{1}{\varepsilon}\big)$ as desired. ☐

Lemma C.69 (Proposition 1 in [114]): Let S be a Lipschitz domain with $S \subseteq [-K,K]^d$ for some $K \geq 1$ and let $g : S \to \mathbb{R}$ have Sobolev smoothness γ for some $\gamma > \frac{d}{2}$, i.e. $\|g\|_{H^\gamma} < \infty$. Then, for large enough $m \in \mathbb{N}$, there exists a neural network $\varphi_g \in \widetilde{\Phi}(\gamma^2 \log m, \gamma^2 m, \gamma^4 m \log m, m^\nu)$ where $\nu = \frac{d}{\gamma - \frac{d}{2}} + \frac{1}{d}$, satisfying

$$|g(u) - \varphi_g(u)| \lesssim K^{\gamma - \frac{d}{2}} \|g\|_{H^\gamma} m^{-\frac{\gamma}{d}}, \qquad u \in [-K,K]^d$$

Proof. First, since S is Lipschitz, we may extend g to a function $\widehat{g} : [-K,K]^d \to \mathbb{R}$, also with Sobolev smoothness γ. To avoid the cumbersome notation, we simply assume without loss of generality that $S = [-K,K]^d$. Then, let $\eta_K(u) = Ku$ and set $\widetilde{g} = \frac{g \circ \eta_K}{\|g \circ \eta_K\|_{B_{2,2}^\gamma}}$, where $\|\cdot\|_{B_{2,2}^\gamma}$ denotes the norm associated with the Besov space $B_{2,2}^\gamma$. Since $H^\gamma \cong B_{2,2}^\gamma$, we have $\widetilde{g} \in B_{2,2}^\gamma$ and $\|\widetilde{g}\|_{B_{2,2}^\gamma} = 1$, whence it follows by [114, Proposition 1] that there exists a neural network $\widetilde{\varphi}_g \in \widetilde{\Phi}(\gamma^2 \log m, \gamma^2 m, \gamma^4 m \log m, m^\nu)$ with

$$|\widetilde{\varphi}_g(u) - \widetilde{g}(u)| \lesssim m^{-\frac{\gamma}{d}}, \qquad u \in [-1,1]^d.$$

Now, letting $\varphi_g(u) = \|g \circ \eta_K\|_{B^\gamma_{2,2}} \widetilde{\varphi}_g(\frac{u}{K})$, it follows that for any $u \in [-K, K]^d$, we have

$$|\varphi_g(u) - g(u)| = \|g \circ \eta_K\|_{B^\gamma_{2,2}} \left|\widetilde{\varphi}_g\left(\frac{u}{K}\right) - \widetilde{g}\left(\frac{u}{K}\right)\right| \lesssim \|g \circ \eta_K\|_{B^\gamma_{2,2}} m^{-\frac{\gamma}{d}}.$$

To bound $\|g \circ \eta_K\|_{B^\gamma_{2,2}}$, we first note that $\|g \circ \eta_K\|_{B^\gamma_{2,2}} \asymp \|g \circ \eta_K\|_{H^\gamma}$, and that for $\beta \in \mathbb{N}_0^d$ we have $\partial^\beta(g \circ \eta_K) = K^{|\beta|}(\partial^\beta g) \circ \eta_K$, while

$$\|(\partial^\beta g) \circ \eta_K\|_{L^2}^2 = \int_{\mathbb{R}^d} |\partial^\beta g(Ku)|^2 \,\mathrm{d}u = K^{-d} \int_{\mathbb{R}^d} |\partial^\beta g(v)|^2 \,\mathrm{d}v = K^{-d}\|\partial^\beta g\|_{L^2}^2.$$

Combining these, we have

$$\|g \circ \eta_K\|_{H^\gamma} = \sqrt{\sum_{|\beta| \le \gamma} \|\partial^\beta (g \circ \eta_K)\|_{L^2}^2} = \sqrt{\sum_{|\beta| \le \gamma} K^{2|\beta| - d} \|\partial^\beta g\|_{L^2}^2} \le K^{\gamma - \frac{d}{2}} \|g\|_{H^\gamma},$$

as desired. ☐

Lemma C.70: Let $g : E \to \mathbb{R}^k$ be a function on some subset $E \subset \mathbb{R}^{W_0}$ such that $|g(s)| \le C$ for all $s \in E$ and some constant $C > 0$. Then, for all $\widetilde{\varphi} \in \widetilde{\Phi}(L, W, S, B)$, where $W_{L+1} = k$, there exists a neural network $\varphi \in \widetilde{\Phi}(L, W, S, C \vee B)$ satisfying

$$|g(s) - \varphi(s)| \le |g(s) - \widetilde{\varphi}(s)|,$$

and $|\varphi(s)| \le \sqrt{k}C$ for all $s \in E$.

Proof. First, note that for all $s \in E$ we have $\|g(s)\|_\infty \le |g(s)| \le C$, so setting

$$\varphi(s) = \begin{bmatrix} \varphi_1(s) \\ \vdots \\ \varphi_k(s) \end{bmatrix} = \begin{bmatrix} \widetilde{\varphi}_1(s) \wedge C \vee (-C) \\ \vdots \\ \widetilde{\varphi}_k(s) \wedge C \vee (-C) \end{bmatrix},$$

we have immediately that $|\varphi_i(s) - g_i(s)| \le |\widetilde{\varphi}_i(s) - g_i(s)|$ and hence $|\varphi(s) - g(s)| \le |\widetilde{\varphi}(s) - g(s)|$, while $|\varphi(s)| \le \sqrt{k}\|\varphi(s)\|_\infty \le kC$ for all $s \in E$. Finally, noting that

$$\widetilde{\varphi}_i(s) \wedge C \vee (-C) = \widetilde{\varphi}_i(s) - \sigma(\widetilde{\varphi}_i(s) - C) + \sigma\big(\sigma(\widetilde{\varphi}_i(s) - C) - \widetilde{\varphi}_i(s) - C\big),$$

it follows that $\varphi \in \widetilde{\Phi}(L, W, S, C \vee B)$. ☐

C.D Auxiliary technical results

Lemma C.71: Let $(\mathsf{W}_t)_{t\geq 0}$ be a k-dimensional Brownian motion for some $k \in \mathbb{N}$, and let $\rho > 1$. Then, the following bounds hold:

(a) $\mathbb{P}\big(|\mathsf{W}_t| > \sqrt{t(k+2\rho)}\big) \lesssim \rho^{\frac{k}{2}} \mathrm{e}^{-\rho}$;

(b) $\mathbb{E}\big[|\mathsf{W}_t| \mathbf{1}_{\{|\mathsf{W}_t| > \sqrt{t(k+2\rho)}\}}\big] \lesssim \sqrt{t}\, \rho^{\frac{k+1}{2}} \mathrm{e}^{-\rho}$.

Proof. Let $\mathsf{Z}_k \sim \mathcal{N}(0, I_k)$. For any $\lambda \in (0, \frac{1}{2})$, Markov's inequality yields

$$\mathbb{P}\big(|\mathsf{Z}_k| > \sqrt{k+2\rho}\big) = \mathbb{P}\Big(\mathrm{e}^{\lambda|\mathsf{Z}_k|^2} > \mathrm{e}^{\lambda(k+2\rho)}\Big) \leq \mathbb{E}\big[\mathrm{e}^{\lambda|\mathsf{Z}_k|^2}\big] \mathrm{e}^{-\lambda(k+2\rho)} = M_{\chi^2_k}(\lambda)\, \mathrm{e}^{-\lambda(k+2\rho)},$$

where $M_{\chi^2_k}(\lambda) = (1-2\lambda)^{-k/2}$ denotes the moment generating function of the χ^2_k distribution. Define $\psi(\lambda) := (1-2\lambda)^{-\frac{k}{2}} \mathrm{e}^{-\lambda(k+2\rho)}$. A direct computation shows that

$$\psi'(\lambda) = 2\big(\lambda(k+2\rho) - \rho\big)(1-2\lambda)^{-\frac{k}{2}-1} \mathrm{e}^{-\lambda(k+2\rho)},$$

which is negative for $\lambda < \frac{\rho}{k+2\rho}$ and positive for $\lambda > \frac{\rho}{k+2\rho}$. Hence, ψ attains its minimum at $\lambda = \frac{\rho}{k+2\rho}$, and we obtain

$$\mathbb{P}\big(|\mathsf{Z}_k| > \sqrt{k+2\rho}\big) \leq \psi\Big(\frac{\rho}{k+2\rho}\Big) = \Big(1 - \frac{2\rho}{k+2\rho}\Big)^{-\frac{k}{2}} \mathrm{e}^{-\rho} = \Big(1 + \frac{2\rho}{k}\Big)^{\frac{k}{2}} \mathrm{e}^{-\rho} \lesssim \rho^{\frac{k}{2}} \mathrm{e}^{-\rho}.$$

Since $\mathsf{W}_t \overset{d}{=} \sqrt{t}\, \mathsf{Z}_k$, this proves (a). We next consider the truncated first moment. Using polar coordinates, we compute

$$\begin{aligned} \mathbb{E}\big[|\mathsf{Z}_k| \mathbf{1}_{\{|\mathsf{Z}_k| > \sqrt{k+2\rho}\}}\big] &= (2\pi)^{-\frac{k}{2}} \int_{\{|x| > \sqrt{k+2\rho}\}} |x| \mathrm{e}^{-\frac{|x|^2}{2}} \,\mathrm{d}x = \frac{2^{-\frac{k}{2}+1}}{\Gamma(\frac{k}{2})} \int_{\sqrt{k+2\rho}}^{\infty} r^k \mathrm{e}^{-\frac{r^2}{2}} \,\mathrm{d}r \\ &= \frac{\sqrt{2}}{\Gamma(\frac{k}{2})} \int_{\frac{k+2\rho}{2}}^{\infty} u^{\frac{k+1}{2}-1} \mathrm{e}^{-u} \,\mathrm{d}u = \sqrt{2}\, \frac{\Gamma\big(\frac{k+1}{2}, \frac{k+2\rho}{2}\big)}{\Gamma(\frac{k}{2})}, \end{aligned}$$

where $\Gamma(s, x)$ denotes the Gamma function. Moreover,

$$\mathbb{P}\big(|\mathsf{Z}_k| > \sqrt{k+2\rho}\big) = \frac{\Gamma\big(\frac{k}{2}, \frac{k+2\rho}{2}\big)}{\Gamma(\frac{k}{2})}.$$

Combining this with part (a), we find

$$\mathbb{E}\big[|\mathsf{Z}_k| \mathbf{1}_{\{|\mathsf{Z}_k| > \sqrt{k+2\rho}\}}\big] \lesssim \frac{\Gamma\big(\frac{k+1}{2}, \frac{k+2\rho}{2}\big)}{\Gamma\big(\frac{k}{2}, \frac{k+2\rho}{2}\big)} \rho^{\frac{k}{2}} \mathrm{e}^{-\rho}.$$

Using the asymptotic relation $\Gamma(s,x) \sim x^{s-1}\mathrm{e}^{-x}$ as $x \to \infty$, we obtain

$$\frac{\Gamma(s+\frac{1}{2},x)}{\Gamma(s,x)}x^{-\frac{1}{2}} = \frac{\Gamma(s+\frac{1}{2},x)x^{1-s-\frac{1}{2}}\mathrm{e}^{x}}{\Gamma(s,x)x^{1-s}\mathrm{e}^{x}} \to 1 \qquad \text{as } x \to \infty,$$

and hence

$$\mathbb{E}\big[|\mathsf{Z}_k|\mathbf{1}_{\{|\mathsf{Z}_k|>\sqrt{k+2\rho}\}}\big] \lesssim \frac{\Gamma(\frac{k+1}{2},\frac{k+2\rho}{2})}{\Gamma(\frac{k}{2},\frac{k+2\rho}{2})}\rho^{\frac{k}{2}}\mathrm{e}^{-\rho} \asymp \sqrt{\frac{k+2\rho}{2}}\,\rho^{\frac{k}{2}}\mathrm{e}^{-\rho} \lesssim \rho^{\frac{k+1}{2}}\mathrm{e}^{-\rho}.$$

Finally, since $|\mathsf{W}_t| \overset{d}{=} \sqrt{t}\,|\mathsf{Z}_k|$, the last display already yields (b). ☐

Lemma C.72: Let $k \in \mathbb{N}$ be given and set $t_j = \cos(\frac{j\pi}{k})$ for $j = 0,\dots,k$ and let $p_i(t) = \prod_{j\neq i}(t-t_j)$ for $i = 0,\dots,k$. Then, $|\frac{p_i(t)}{p_i(t_i)}| \leq 2$ for all i and $t \in (-1,1)$.

Proof. Let $p(t) = \prod_{j=0}^{k}(t-t_j)$ and $\widehat{p}(t) = \prod_{j=1}^{k-1}(t-t_j)$. By comparing roots, we see that $\widehat{p}$ is simply a re-scaling of the $k-1$'st Chebyshev polynomial of the second kind U_{k-1} given by $U_{k-1}(\cos\theta)\sin\theta = \sin k\theta$. In particular, since by L'Hôpital we have

$$U_{k-1}(1) = \lim_{\theta\to 0} U_{k-1}(\cos\theta) = \lim_{\theta\to 0}\frac{\sin k\theta}{\sin\theta} = \lim_{\theta\to 0}\frac{k\cos k\theta}{\cos\theta} = k,$$

and by [122] that $\widehat{p}(1) = \frac{p_0(t_0)}{2} = \frac{k}{2^{k-1}}$, we have $\widehat{p}(t) = \frac{U_{k-1}(t)}{2^{k-1}}$. In particular,

$$p(\cos\theta) = (\cos\theta-1)(\cos\theta+1)\widehat{p}(\cos\theta) = -\sin^2(\theta)\frac{U_{k-1}(\cos\theta)}{2^{k-1}} = \frac{-\sin\theta\sin k\theta}{2^{k-1}},$$

and so for $\theta \neq \frac{i\pi}{k}$,

$$p_i(\cos\theta) = -\frac{\sin\theta\sin k\theta}{2^{k-1}(\cos\theta - \cos\frac{i\pi}{k})},$$

while [122] again yields that

$$p_i(t_i) = \begin{cases} (-1)^i\frac{k}{2^{k-1}}, & \text{if } i \notin \{0,k\} \\ (-1)^i\frac{k}{2^{k-2}}, & \text{if } i \in \{0,k\}. \end{cases}$$

Thus, for $\theta \neq \frac{i\pi}{k}$,

$$\Big|\frac{p_i(\cos\theta)}{p_i(t_i)}\Big| \leq \frac{|\sin\theta\sin k\theta|}{k|\cos\theta-\cos\frac{i\pi}{k}|}.$$

From the above it follows that $|\frac{p_i(\cos\theta)}{p_i(t_i)}| = |\frac{p_{k-i}(\cos(\pi-\theta))}{p_{k-i}(t_{k-i})}|$, and so we may assume going forward that $\theta < \frac{i\pi}{k}$. Next, since $|\cos\theta - \cos\frac{i\pi}{k}| = 2|\sin(\frac{\theta}{2}+\frac{i\pi}{2k})\sin(\frac{\theta}{2}-\frac{i\pi}{2k})|$, it follows that

$$\left|\frac{p_i(t)}{p_i(t_i)}\right| \leq \frac{1}{2k}\sup_{\theta\in(0,\frac{i\pi}{k})}\left|\frac{\sin\theta}{\sin(\frac{\theta}{2}+\frac{i\pi}{2k})}\right| \sup_{\theta\in(0,\frac{i\pi}{k})}\left|\frac{\sin k\theta}{\sin(\frac{\theta}{2}-\frac{i\pi}{2k})}\right|.$$

Now, let $g(a,\theta) = \frac{\sin\theta}{\sin\frac{\theta+a}{2}}$ and note that $g(a,\theta) \geq 0$ for $\theta \leq a \leq \pi$, and we have

$$\sup_{\theta\in(0,\frac{i\pi}{k})}\left|\frac{\sin\theta}{\sin(\frac{\theta}{2}+\frac{i\pi}{2k})}\right| = \sup_{\theta\in(0,\frac{i\pi}{k})} g\left(\frac{i\pi}{k},\theta\right) \leq \sup_{a\in(0,\pi)}\sup_{\theta\in(0,a)} g(a,\theta) = \sup_{\theta\in(0,\pi)}\sup_{a\in(\pi-\theta,\pi)} g(a,\theta).$$

Since for fixed $\theta \in (0,\pi)$ and $a \in (\pi-\theta,\pi)$ we have

$$\frac{\mathrm{d}}{\mathrm{d}a}g(a,\theta) = -\frac{1}{2}g(a,\theta)\cot\frac{\theta+a}{2} \geq 0,$$

it follows that $\sup_{a\in(\pi-\theta,\pi)} g(a,\theta) = g(\pi,\theta)$. Therefore, since $g(\pi,\theta) = 2\sin\frac{\theta}{2}$, we have

$$\sup_{\theta\in(0,\frac{i\pi}{k})}\left|\frac{\sin\theta}{\sin(\frac{\theta}{2}+\frac{i\pi}{2k})}\right| \leq 2.$$

Next, we have that

$$\sup_{\theta\in(0,\frac{i\pi}{k})}\left|\frac{\sin k\theta}{\sin(\frac{\theta}{2}-\frac{i\pi}{2k})}\right| \leq \sup_{\theta\in(0,\pi)}\left|\frac{\sin k\theta}{\sin(\frac{\theta}{2}-\frac{i\pi}{2k})}\right| = \sup_{\theta\in(0,\pi)}\left|\frac{\sin(k(\theta-\frac{i\pi}{k}))}{\sin(\frac{\theta}{2}-\frac{i\pi}{2k})}\right| = \sup_{\theta\in(0,\pi)}\left|\frac{\sin k\theta}{\sin\frac{\theta}{2}}\right| = 2k.$$

Plugging both of these into the estimate above, we find that $|\frac{p_i(t)}{p_i(t_i)}| \leq 2$ as desired. □

Bibliography

[1] Luis H. R. Alvarez and Larry A. Shepp. "Optimal harvesting of stochastically fluctuating populations". In: *J. Math. Biol.* 37.2 (1998), pp. 155–177. doi: `10.1007/s002850050124`.

[2] Luis HR Alvarez. *A Class of Solvable Stationary Singular Stochastic Control Problems.* 2018. arXiv: `1803.03464`.

[3] Brian D. O. Anderson. "Reverse-time diffusion equation models". In: *Stochastic Process. Appl.* 12.3 (1982), pp. 313–326. doi: `10.1016/0304-4149(82)90051-5`.

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 214–223.

[5] Peter Auer, Thomas Jaksch, and Ronald Ortner. "Near-optimal Regret Bounds for Reinforcement Learning". In: *Advances in Neural Information Processing Systems.* Vol. 21. 2008.

[6] Iskander Azangulov, George Deligiannidis, and Judith Rousseau. *Convergence of Diffusion Models Under the Manifold Hypothesis in High-Dimensions.* 2024. arXiv: `2409.18804 [stat.ML]`.

[7] Dominique Bakry, Patrick Cattiaux, and Arnaud Guillin. "Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré". In: *J. Funct. Anal.* 254.3 (2008), pp. 727–759. doi: `10.1016/j.jfa.2007.11.002`.

[8] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators.* Vol. 348. Grundlehren der mathematischen Wissenschaften. Springer, Cham, 2014, pp. xx+552. doi: `10.1007/978-3-319-00227-9`.

[9] Dominique Bakry and Zhongmin M. Qian. "Harnack inequalities on a manifold with positive or negative Ricci curvature". In: *Rev. Mat. Iberoamericana* 15.1 (1999), pp. 143–179. doi: `10.4171/RMI/253`.

[10] Peter L. Bartlett and Shahar Mendelson. "Empirical minimization". In: *Probab. Theory Related Fields* 135.3 (2006), pp. 311–334. doi: `10.1007/s00440-005-0462-3`.

[11] Matteo Basei, Xin Guo, Anran Hu, and Yufei Zhang. "Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon". In: *Journal of Machine Learning Research* 23.178 (2022), pp. 1–34.

[12] Joe Benton, Yuyang Shi, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. "From denoising diffusions to denoising Markov models". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 86.2 (2024), pp. 286–301. doi: `10.1093/jrsssb/qkae005`.

[13] Frederik Boetius and Michael Kohlmann. “Connections between optimal stopping and singular stochastic control”. In: *Stochastic Process. Appl.* 77.2 (1998), pp. 253–281. DOI: `10.1016/S0304-4149(98)00049-0`.

[14] Victor Boone and Bruno Gaujal. “The Regret of Exploration and the Control of Bad Episodes in Reinforcement Learning”. In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. PMLR, 2023, pp. 2824–2856.

[15] Bradley CA Brown, Anthony L. Caterini, Brendan Leigh Ross, Jesse C Cresswell, and Gabriel Loaiza-Ganem. “Verifying the Union of Manifolds Hypothesis for Image Data”. In: *International Conference on Learning Representations*. 2023.

[16] Krzysztof Burdzy, Zhen-Qing Chen, and John Sylvester. “The heat equation and reflected Brownian motion in time-dependent domains”. In: *Ann. Probab.* 32.1B (2004), pp. 775–804. DOI: `10.1214/aop/1079021464`.

[17] Apostolos N. Burnetas and Michael N. Katehakis. “Optimal adaptive policies for Markov decision processes”. In: *Math. Oper. Res.* 22.1 (1997), pp. 222–255. DOI: `10.1287/moor.22.1.222`.

[18] Patrick Cattiaux. “Time reversal of diffusion processes with a boundary condition”. In: *Stochastic Process. Appl.* 28.2 (1988), pp. 275–292. DOI: `10.1016/0304-4149(88)90101-9`.

[19] Patrick Cattiaux, Djalil Chafaï, and Arnaud Guillin. “Central limit theorems for additive functionals of ergodic Markov diffusions processes”. In: *ALEA Lat. Am. J. Probab. Math. Stat.* 9.2 (2012), pp. 337–382.

[20] Minwoo Chae. *Rates of convergence for nonparametric estimation of singular distributions using generative adversarial networks*. 2022. arXiv: `2202.02890 [math.ST]`.

[21] Saptarshi Chakraborty, Quentin Berthet, and Peter L. Bartlett. *Generalization Properties of Score-matching Diffusion Models for Intrinsically Low-dimensional Data*. 2026. arXiv: `2603.03700 [stat.ML]`.

[22] Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. “Score Approximation, Estimation and Distribution Recovery of Diffusion Models on Low-Dimensional Data”. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 4672–4712.

[23] Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. *Distribution Approximation and Statistical Estimation Guarantees of Generative Adversarial Networks*. 2022. arXiv: `2002.03938 [cs.LG]`.

[24] Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical Optimal Transport.* Vol. 2364. Lecture Notes in Mathematics. École d'Été de Probabilités de Saint-Flour XLIX – 2019. Springer, Cham, 2025, pp. xiv+260. DOI: `10.1007/978-3-031-85160-5`. arXiv: `2407.18163 [math.ST]`.

[25] Sören Christensen, Asbjø rn Holk Thomsen, and Lukas Trottner. "Data-driven rules for multidimensional reflection problems". In: *SIAM/ASA J. Uncertain. Quantif.* 12.4 (2024), pp. 1240–1272. DOI: `10.1137/23M1618570`.

[26] Sören Christensen, Ernesto Mordecki, and Facundo Oliú Eguren. *Two sided ergodic singular control and mean field game for diffusions.* 2023. arXiv: `2306.09263`.

[27] Sören Christensen and Claudia Strauch. "Nonparametric learning for impulse control problems—Exploration vs. exploitation". In: *Ann. Appl. Prob.* 33.2 (2023), pp. 1569 –1587. DOI: `10.1214/22-AAP1849`.

[28] Sören Christensen, Claudia Strauch, and Lukas Trottner. "Learning to reflect: A unifying approach for data-driven stochastic control strategies". In: *Bernoulli* 30.3 (2024), pp. 2074–2101. DOI: `10.3150/23-BEJ1665`.

[29] E. B. Davies. *Heat kernels and spectral theory.* Vol. 92. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1989, pp. x+197. DOI: `10.1017/CBO9780511566158`.

[30] M. H. A. Davis and M. Zervos. "A pair of explicitly solvable singular stochastic control problems". In: *Appl. Math. Optim.* 38.3 (1998), pp. 327–352. DOI: `10.1007/s002459900094`.

[31] Mark HA Davis and Andrew R Norman. "Portfolio selection with transaction costs". In: *Mathematics of operations research* 15.4 (1990), pp. 676–713.

[32] Tiziano De Angelis, Giorgio Ferrari, and John Moriarty. "A solvable two-dimensional degenerate singular stochastic control problem with nonconvex costs". In: *Math. Oper. Res.* 44.2 (2019), pp. 512–531. DOI: `10.1287/moor.2018.0934`.

[33] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. "On the sample complexity of the linear quadratic regulator". In: *Found. Comput. Math.* 20.4 (2020), pp. 633–679. DOI: `10.1007/s10208-019-09426-y`.

[34] M. C. Delfour and J.-P. Zolésio. *Shapes and geometries.* Second. Vol. 22. Advances in Design and Control. Metrics, analysis, differential calculus, and optimization. SIAM, Philadelphia, PA, 2011, pp. xxiv+622. DOI: `10.1137/1.9780898719826`.

[35] Jodi Dianetti and Giorgio Ferrari. "Multidimensional singular control and related Skorokhod problem: Sufficient conditions for the characterization of optimal controls". In: *Stochastic Process. Appl.* 162 (2023), pp. 547–592. DOI: `10.1016/j.spa.2023.05.006`.

[36] Vincent Divol. “Measure estimation on manifolds: an optimal transport approach”. In: *Probab. Theory Related Fields* 183.1-2 (2022), pp. 581–647. DOI: `10.1007/s00440-022-01118-z`.

[37] Zehao Dou, Subhodh Kotekal, Zhehao Xu, and Harrison H. Zhou. *From optimal score matching to optimal sampling*. 2024. arXiv: `2409.07032 [stat.ML]`.

[38] D. Down, S. P. Meyn, and R. L. Tweedie. “Exponential and uniform ergodicity of Markov processes”. In: *Ann. Probab.* 23.4 (1995), pp. 1671–1691.

[39] T. E. Duncan, L. Guo, and B. Pasik-Duncan. “Adaptive continuous-time linear quadratic Gaussian control”. In: *IEEE Trans. Automat. Control* 44.9 (1999), pp. 1653–1662. DOI: `10.1109/9.788532`.

[40] Jianqing Fan, Yihong Gu, and Ximing Li. *Optimal estimation of a factorizable density using diffusion models with ReLU neural networks*. 2025. arXiv: `2510.03994 [math.ST]`.

[41] Salvatore Federico, Giorgio Ferrari, and Patrick Schuhmann. “A singular stochastic control problem with interconnected dynamics”. In: *SIAM J. Control Optim.* 58.5 (2020), pp. 2821–2853. DOI: `10.1137/19M1296288`.

[42] Salvatore Federico, Giorgio Ferrari, and Patrick Schuhmann. “Singular control of the drift of a Brownian system”. In: *Appl. Math. Optim.* 84 (2021), S561–S590. DOI: `10.1007/s00245-021-09779-3`.

[43] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. “Testing the manifold hypothesis”. In: *J. Amer. Math. Soc.* 29.4 (2016), pp. 983–1049. DOI: `10.1090/jams/852`.

[44] Giorgio Ferrari and Tiziano Vargiolu. “On the singular control of exchange rates”. In: *Annals of Operations Research* 292.2 (2020), pp. 795–832.

[45] Nic Fishman, Leo Klarner, Valentin De Bortoli, Emile Mathieu, and Michael John Hutchinson. “Diffusion Models for Constrained Domains”. In: *Transactions on Machine Learning Research* (2023). arXiv: `2304.05364 [cs.LG]`.

[46] Nic Fishman, Leo Klarner, Emile Mathieu, Michael Hutchinson, and Valentin De Bortoli. “Metropolis Sampling for Constrained Diffusion Models”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023, pp. 62296–62331.

[47] Wendell H. Fleming and H. Mete Soner. *Controlled Markov processes and viscosity solutions*. 2nd ed. Vol. 25. Stochastic Modelling and Applied Probability. Springer, New York, 2006, pp. xviii+429.

[48] Matteo Giordano and Sven Wang. “Statistical algorithms for low-frequency diffusion data: a PDE approach”. In: *Ann. Statist.* 53.3 (2025), pp. 1150–1175. DOI: `10.1214/25-aos2496`. arXiv: `2405.01372 [stat.ME]`.

[49] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[50] Xin Guo, Anran Hu, and Yufei Zhang. *Reinforcement learning for linear-convex models with jumps via stability analysis of feedback controls.* 2021. arXiv: `2104.09311`.

[51] Xin Guo, Philip Kaminsky, Pascal Tomecek, and Ming Yuen. "Optimal spot market inventory strategies in the presence of cost and price risk". In: *Mathematical Methods of Operations Research* 73 (2011), pp. 109–137.

[52] Xin Guo and Pascal Tomecek. "A class of singular control problems and the smooth fit principle". In: *SIAM J. Control Optim.* 47.6 (2008), pp. 3076–3099. DOI: `10.1137/070685336`.

[53] U. G. Haussmann and É. Pardoux. "Time reversal of diffusions". In: *Ann. Probab.* 14.4 (1986), pp. 1188–1205.

[54] Jiafan He, Dongruo Zhou, and Quanquan Gu. "Logarithmic regret for reinforcement learning with linear function approximation". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 4171–4180.

[55] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.

[56] Marc Hoffmann and Kolyan Ray. "Nonparametric Bayesian estimation in a multidimensional diffusion model with high frequency data". In: *Probab. Theory Relat. Fields* (2024). DOI: `10.1007/s00440-024-01317-w`. arXiv: `2211.12267 [math.ST]`.

[57] Asbjørn Holk, Claudia Strauch, and Lukas Trottner. *Reflected diffusion models adapt to low-dimensional data.* 2026. arXiv: `2603.24495 [math.ST]`.

[58] Asbjørn Holk, Claudia Strauch, and Lukas Trottner. *Statistical guarantees for denoising reflected diffusion models.* 2024. arXiv: `2411.01563 [math.ST]`.

[59] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5 (1989), pp. 359–366. DOI: `https://doi.org/10.1016/0893-6080(89)90020-8`.

[60] Seizô Itô. *Diffusion equations.* Vol. 114. Translations of Mathematical Monographs. American Mathematical Society, Providence, RI, 1992, pp. x+225. DOI: `10.1090/mmono/114`.

[61] Weining Kang and Kavita Ramanan. "Characterization of stationary distributions of reflected diffusions". In: *Ann. Appl. Probab.* 24.4 (2014), pp. 1329–1374. DOI: `10.1214/13-AAP947`.

[62] Ioannis Karatzas. "A class of singular stochastic control problems". In: *Advances in Applied Probability* 15.2 (1983), pp. 225–254.

[63] Ioannis Karatzas and Steven E. Shreve. *Brownian motion and stochastic calculus*. Second. Vol. 113. Graduate Texts in Mathematics. Springer-Verlag, New York, 1991, pp. xxiv+470. DOI: `10.1007/978-1-4612-0949-2`.

[64] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *2nd International Conference on Learning Representations*. Ed. by Yoshua Bengio and Yann LeCun. 2014.

[65] Steven G. Krantz and Harold R. Parks. "Distance to C^k hypersurfaces". In: *J. Differential Equations* 40.1 (1981), pp. 116–120. DOI: `10.1016/0022-0396(81)90013-9`.

[66] Elena V Krichagina and Michael I Taksar. "Diffusion approximation for GI/G/1 controlled queues". In: *Queueing systems* 12 (1992), pp. 333–367.

[67] Lukasz Kruk. "Optimal policies for n-dimensional singular stochastic control problems. I. The Skorokhod problem". In: *SIAM J. Control Optim.* 38.5 (2000), pp. 1603–1622. DOI: `10.1137/S0363012998347535`.

[68] Lukasz Kruk. "Optimal policies for n-dimensional singular stochastic control problems. II. The radially symmetric case. Ergodic control". In: *SIAM J. Control Optim.* 39.2 (2000), pp. 635–659. DOI: `10.1137/S0363012998347547`.

[69] Sunil Kumar and Kumar Muthuraman. "A numerical method for solving singular stochastic control problems". In: *Oper. Res.* 52.4 (2004), pp. 563–582. DOI: `10.1287/opre.1030.0107`.

[70] Harold J Kushner. "Numerical methods for stochastic control problems in continuous time". In: *SIAM J. Control Optim.* 28.5 (1990), pp. 999–1048.

[71] Hyeok Kyu Kwon, Dongha Kim, Ilsang Ohn, and Minwoo Chae. *Nonparametric estimation of a factorizable density using diffusion models*. 2025. arXiv: `2501.01783 [math.ST]`.

[72] Hyeok Kyu Kwon, Dongha Kim, Ilsang Ohn, and Minwoo Chae. "Nonparametric Estimation of a Factorizable Density using Diffusion Models". In: *J. Mach. Learn. Res.* 27.22 (2026), pp. 1–125.

[73] Benedict Leimkuhler, Akash Sharma, and Michael V. Tretyakov. "Simplest random walk for approximating Robin boundary value problems and ergodic limits of reflected diffusions". In: *Ann. Appl. Probab.* 33.3 (2023), pp. 1904–1960. DOI: `10.1214/22-aap1856`.

[74] Peter Li and Shing-Tung Yau. "On the parabolic kernel of the Schrödinger operator". In: *Acta Math.* 156.3-4 (1986), pp. 153–201. DOI: `10.1007/BF02399203`.

[75] Tengyuan Liang. "How well generative adversarial networks learn distributions". In: *J. Mach. Learn. Res.* 22.228 (2021), Paper No. 228, 41.

[76] P.-L. Lions and A.-S. Sznitman. "Stochastic differential equations with reflecting boundary conditions". In: *Comm. Pure Appl. Math.* 37.4 (1984), pp. 511–537. DOI: `10.1002/cpa.3160370408`.

[77] Gabriel Loaiza-Ganem, Brendan Leigh Ross, Rasa Hosseinzadeh, Anthony L. Caterini, and Jesse C. Cresswell. "Deep Generative Models through the Lens of the Manifold Hypothesis: A Survey and New Connections". In: *Transactions on Machine Learning Research* (2024).

[78] Arne Løkka and Mihail Zervos. "Optimal dividend and issuance of equity policies in the presence of proportional costs". In: *Insurance: Mathematics and Economics* 42.3 (2008), pp. 954–961.

[79] Jackson Loper. "Uniform ergodicity for Brownian motion in a bounded convex set". In: *J. Theoret. Probab.* 33.1 (2020), pp. 22–35. DOI: `10.1007/s10959-018-0848-7`.

[80] Aaron Lou and Stefano Ermon. "Reflected Diffusion Models". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 22675–22701.

[81] Yunqian Ma and Yun Fu, eds. *Manifold learning theory and applications*. CRC Press, Boca Raton, FL, 2012, pp. xxiv+290.

[82] Daniel McDonald. "Minimax Density Estimation for Growing Dimension". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by Aarti Singh and Jerry Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, 2017, pp. 194–203.

[83] José-Luis Menaldi and Michael I Taksar. "Optimal correction problem of a multidimensional stochastic system". In: *Automatica* 25.2 (1989), pp. 223–232.

[84] Eyal Neuman and Yufei Zhang. *Statistical Learning with Sublinear Regret of Propagator Models*. 2023. arXiv: `2301.05157`.

[85] Richard Nickl. "Consistent inference for diffusions from low frequency measurements". In: *Ann. Statist.* 52.2 (2024), pp. 519–549. DOI: `10.1214/24-aos2357`. arXiv: `2210.13008 [math.ST]`.

[86] Jonathan Niles-Weed and Quentin Berthet. *Minimax estimation of smooth densities in Wasserstein distance*. 2020. arXiv: `1902.01778 [math.ST]`.

[87] Jonathan Niles-Weed and Quentin Berthet. "Minimax estimation of smooth densities in Wasserstein distance". In: *Ann. Statist.* 50.3 (2022), pp. 1519–1540. DOI: `10.1214/21-aos2161`.

[88] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. "f-GAN: training generative neural samplers using variational divergence minimization". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 271–279.

[89] Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. "Diffusion Models are Minimax Optimal Distribution Estimators". In: *International Conference on Machine Learning*. 2023.

[90] Bernt Øksendal and Agnès Sulem. *Applied stochastic control of jump diffusions*. 3rd ed. Universitext. Springer, Cham, 2019, pp. xvi+436. DOI: `10.1007/978-3-030-02781-0`.

[91] Ian Osband and Benjamin Van Roy. “Model-based reinforcement learning and the eluder dimension”. In: *Advances in Neural Information Processing Systems* 27 (2014).

[92] El Maati Ouhabaz. *Analysis of heat equations on domains*. Vol. 31. London Mathematical Society Monographs Series. Princeton University Press, Princeton, NJ, 2005, pp. xiv+284.

[93] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. “Normalizing Flows for Probabilistic Modeling and Inference”. In: *Journal of Machine Learning Research* 22.57 (2021), pp. 1–64.

[94] Andrey Pilipenko. *An introduction to stochastic differential equations with reflection.* Vol. 1. Universitätsverlag Potsdam, 2014.

[95] Andrey Pilipenko. *An introduction to stochastic differential equations with reflection.* Vol. 1. Universitätsverlag Potsdam, 2014.

[96] Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. “The Intrinsic Dimension of Images and Its Impact on Learning”. In: *International Conference on Learning Representations*. 2021.

[97] Nikita Puchkin, Sergey Samsonov, Denis Belomestny, Eric Moulines, and Alexey Naumov. “Rates of convergence for density estimation with generative adversarial networks”. In: *Journal of Machine Learning Research* 25.29 (2024), pp. 1–47.

[98] Zhong Min Qian. “Gradient estimates and heat kernel estimate”. In: *Proc. Roy. Soc. Edinburgh Sect. A* 125.5 (1995), pp. 975–990. DOI: `10.1017/S0308210500022599`.

[99] Yinuo Ren, Grant M. Rotskoff, and Lexing Ying. *A Unified Approach to Analysis and Design of Denoising Markov Models*. arXiv preprint, arXiv:2504.01938. 2025. arXiv: `2504.01938 [cs.LG]`.

[100] Danilo Rezende and Shakir Mohamed. “Variational Inference with Normalizing Flows”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, 2015, pp. 1530–1538.

[101] Johannes Schmidt-Hieber. “Nonparametric regression using deep neural networks with ReLU activation function”. In: *Ann. Statist.* 48.4 (2020), pp. 1875–1897. DOI: `10.1214/19-AOS1875`.

[102] Nicolas Schreuder, Victor-Emmanuel Brunel, and Arnak Dalalyan. “Statistical guarantees for generative models without domination”. In: *Proceedings of the 32nd International Conference on Algorithmic Learning Theory.* Ed. by Vitaly Feldman, Katrina Ligett, and Sivan Sabato. Vol. 132. Proceedings of Machine Learning Research. PMLR, 2021, pp. 1051–1071.

[103] Leszek Słomiński. “On approximation of solutions of multidimensional SDEs with reflecting boundary conditions”. In: *Stochastic Process. Appl.* 50.2 (1994), pp. 197–219. DOI: `10.1016/0304-4149(94)90118-X`.

[104] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. “Deep Unsupervised Learning using Nonequilibrium Thermodynamics”. In: *Proceedings of the 32nd International Conference on Machine Learning.* Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, 2015, pp. 2256–2265.

[105] H. Mete Soner and Steven E. Shreve. “Regularity of the value function for a two-dimensional singular stochastic control problem”. In: *SIAM J. Control Optim.* 27.4 (1989), pp. 876–907. DOI: `10.1137/0327047`.

[106] Yang Song and Stefano Ermon. “Generative Modeling by Estimating Gradients of the Data Distribution”. In: *Advances in Neural Information Processing Systems.* Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[107] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. “Score-Based Generative Modeling through Stochastic Differential Equations”. In: *International Conference on Learning Representations.* 2021.

[108] Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. “Diffusion Models Encode the Intrinsic Dimension of Data Manifolds”. In: *Proceedings of the 41st International Conference on Machine Learning.* Vol. 235. Proceedings of Machine Learning Research. PMLR, 2024, pp. 46412–46440.

[109] Arthur Stéphanovitch, Eddie Aamari, and Clément Levrard. “Wasserstein generative adversarial networks are minimax optimal distribution estimators”. In: *Ann. Statist.* 52.5 (2024), pp. 2167–2193. DOI: `10.1214/24-aos2430`.

[110] Arthur Stéphanovitch, Ugo Tanielian, Benoît Cadre, Nicolas Klutchnikoff, and Gérard Biau. “Optimal 1-Wasserstein distance for WGANs”. In: *Bernoulli* 30.4 (2024), pp. 2955–2978. DOI: `10.3150/23-bej1701`.

[111] Claudia Strauch. “Adaptive invariant density estimation for ergodic diffusions over anisotropic classes”. In: *Ann. Statist.* 46.6B (2018), pp. 3451–3480. DOI: `10.1214/17-AOS1664`.

[112] Arthur Stéphanovitch, Eddie Aamari, and Clément Levrard. *Generalization bounds for score-based generative models: a synthetic proof.* arXiv preprint, arXiv:2507.04794. 2025.

[113] Arthur Stéphanovitch, Eddie Aamari, and Clément Levrard. *Generalization bounds for score-based generative models: a synthetic proof.* 2025. arXiv: `2507.04794 [math.ST]`.

[114] Taiji Suzuki. "Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality". In: *International Conference on Learning Representations.* 2019.

[115] E. G. Tabak and Cristina V. Turner. "A family of nonparametric density estimation algorithms". In: *Comm. Pure Appl. Math.* 66.2 (2013), pp. 145–164. DOI: `10.1002/cpa.21423`.

[116] Corentin Tallec, Léonard Blier, and Yann Ollivier. "Making deep q-learning methods robust to time discretization". In: *International Conference on Machine Learning.* PMLR. 2019, pp. 6096–6104.

[117] Hiroshi Tanaka. "Stochastic differential equations with reflecting boundary condition in convex regions". In: *Hiroshima Math. J.* 9.1 (1979), pp. 163–177.

[118] Rong Tang and Yun Yang. "Adaptivity of Diffusion Models to Manifold Structures". In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics.* Ed. by Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li. Vol. 238. Proceedings of Machine Learning Research. PMLR, 2024, pp. 1648–1656.

[119] Rong Tang and Yun Yang. "Minimax rate of distribution estimation on unknown submanifolds under adversarial losses". In: *Ann. Statist.* 51.3 (2023), pp. 1282–1308. DOI: `10.1214/23-aos2291`.

[120] Wenpin Tang and Hanyang Zhao. "Score-based diffusion models via stochastic differential equations". In: *Stat. Surv.* 19 (2025), pp. 28–64. DOI: `10.1214/25-ss152`.

[121] Matus Telgarsky. "Neural networks and rational functions". In: *Proceedings of the 34th International Conference on Machine Learning.* Proceedings of Machine Learning Research. PMLR, 2017, pp. 3387–3393.

[122] Lloyd N. Trefethen. *Approximation theory and approximation practice.* Extended. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, [2020] ©2020, pp. xi+363. DOI: `10.1137/1.9781611975949`.

[123] Hans Triebel. *Theory of function spaces.* Modern Birkhäuser Classics. Birkhäuser/Springer Basel AG, Basel, 2010, p. 285.

[124] Alexandre B Tsybakov. *Introduction to Nonparametric Estimation.* Springer series in statistics. Dordrecht: Springer, 2009. DOI: `10.1007/b13794`.

[125] Roman Vershynin. *High-dimensional probability.* Vol. 47. Cambridge Series in Statistical and Probabilistic Mathematics. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, 2018, pp. xiv+284. DOI: `10.1017/9781108231596`.

[126] Martin G. Vieten and Richard H. Stockbridge. “Convergence of finite element methods for singular stochastic control”. In: *SIAM J. Control Optim.* 56.6 (2018), pp. 4336–4364. DOI: `10.1137/17M1155119`.

[127] Martin G. Vieten and Richard H. Stockbridge. “On the solution structure of infinite-dimensional linear problems stemming from singular stochastic control problems”. In: *SIAM J. Control Optim.* 58.6 (2020), pp. 3363–3388. DOI: `10.1137/19M1297415`.

[128] Cédric Villani. *Optimal transport.* Vol. 338. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Old and new. Springer-Verlag, Berlin, 2009, pp. xxii+973. DOI: `10.1007/978-3-540-71050-9`.

[129] Pascal Vincent. “A connection between score matching and denoising autoencoders”. In: *Neural Comput.* 23.7 (2011), pp. 1661–1674. DOI: `10.1162/NECO\_a\_00142`.

[130] Ananda P. N. Weerasinghe. “Stationary stochastic control for Itô processes”. In: *Adv. in Appl. Probab.* 34.1 (2002), pp. 128–140. DOI: `10.1239/aap/1019160953`.

[131] R. J. Williams. “Reflected Brownian motion with skew symmetric data in a polyhedral domain”. In: *Probab. Theory Related Fields* 75.4 (1987), pp. 459–485. DOI: `10.1007/BF00320328`.

[132] Konstantin Yakovlev and Nikita Puchkin. “Generalization error bound for denoising score matching under relaxed manifold assumption”. In: *Proceedings of 38th Conference on Learning Theory.* Vol. 291. Proceedings of Machine Learning Research. PMLR, 2025, pp. 5824–5891. arXiv: `2502.13662 [cs.LG]`.

[133] Yuhong Yang and Andrew Barron. “Information-theoretic determination of minimax rates of convergence”. In: *Ann. Statist.* 27.5 (1999), pp. 1564–1599. DOI: `10.1214/aos/1017939142`.

[134] Dmitry Yarotsky. “Error bounds for approximations with deep ReLU networks”. In: *Neural Netw.* 94 (2017), pp. 103–114. DOI: `https://doi.org/10.1016/j.neunet.2017.07.002`.

[135] Kaihong Zhang, Caitlyn H. Yin, Feng Liang, and Jingbo Liu. “Minimax optimality of score-based diffusion models: beyond the density lower bound assumptions”. In: *Proceedings of the 41st International Conference on Machine Learning.* Proceedings of Machine Learning Research. Vienna, Austria: PMLR, 2024.

[136] Junbo Zhao, Michael Mathieu, and Yann LeCun. “Energy-based Generative Adversarial Networks”. In: *International Conference on Learning Representations.* 2017.