Statistical Convergence of Spherical First Hitting Diffusion Models

Simon Bienewald
Institute for Stochastics and Applications
University of Stuttgart
Stuttgart, Germany
simon.bienewald@isa.uni-stuttgart.de

Lukas Trottner
Institute for Stochastics and Applications
University of Stuttgart
Stuttgart, Germany
lukas.trottner@isa.uni-stuttgart.de

Abstract

Denoising diffusion models have evolved into a state-of-the-art method for tasks in various fields, such as denoising and generation of images, text generation, or generation of synthetic data for training of other machine learning models. First hitting diffusion models (FHDM) are a particular class of denoising diffusion models with *random* adaptive generation time tailored to generate data on a known manifold. Building on the conditioning framework of Doob's h-transform these models leverage the given information on the target data manifold to demonstrate strong performance across tasks while offering distinct features such as time-homogeneous dynamics of the generating process and a reduced average simulation time. Even though the theoretical investigation of standard forward-backward diffusion models has attracted much attention in the recent past, the statistical convergence properties of FHDMs are not yet understood. In this work, we show that, up to logarithmic factors, FHDMs achieve the minimax optimal convergence rate in total variation for spherically supported Sobolev smooth data distributions. Our result holds under the idealized assumption that the model can be trained on the exact denoising score matching loss and sampled from without discretisation or Monte-Carlo error. In particular, this is the first statistical optimality result for denoising diffusion modelling with random generation time.

1 Introduction

Generative modelling has rapidly gained in importance in recent years, with applications including the generation of hyperrealistic images and videos, text and also synthetic data for training of large machine learning models. Many state of the art generative models are variants of score-based denoising diffusion models (DDMs) [27], which train neural networks to learn the drift of a stochastic diffusion process that is initialised in some easy-to-sample-from distribution and terminates in the targeted data distribution. The most common procedure simulates forward SDEs with Gaussian transition kernels initialised in the training data until approximate convergence to a known Gaussian distribution to create a training data set of paths. Based on a denoising score matching procedure the *score* of the forward process (the $\log$-gradient of the forward marginal densities), which characterises the dynamics of the reverse (generating) process, is then learned. New samples are then generated as the terminal value of the approximated reverse SDE initialised in the approximate terminal distribution of the forward process. A drawback of denoising diffusion models compared to one-step sampling procedures such as GANs or variational autoencoders is the numerically expensive and highly sensitive generative procedure that requires many iterative sampling steps to approximate the reverse SDE on a fine discretised time-grid. Moreover, because of the time-inhomogeneous nature of the reverse process, the algorithm does not adapt to the noise level along generated paths since it relies on a fixed deterministic simulation time that needs to be sufficiently large for the forward

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

process to terminate in the desired Gaussian distribution for backward initialisation. To address such issues when some prior geometric knowledge on the data support is given, [36] introduced First Hitting Diffusion Models (FHDMs). While FHDMs, similarly to DDMs, use an iterative SDE based sampling procedure, they do not rely on a time-reversal mechanism but aim to condition a simple SDE, such as a Brownian motion, to terminate in the target distribution Π^* at the *random* first hitting time of the known data manifold. This is achieved by learning the log-gradient of a Doob h-transform [8] (the analogue of the score in DDMs) that is fitted to Π^* via a denoising score matching procedure that involves drawing first hitting bridges to the given data at the training stage. For instance, FHDMs have been shown in [36] to produce high quality point clouds in $\mathbb{R}^3$, climate data on the sphere, unweighted graphs, and image segmentations. Importantly, the authors demonstrate that leveraging the geometric information on the data in the design of the model leads to better generation performance compared to conventional DDMs, while the average (here *random* first-hitting) generation time can be significantly decreased.

Our contribution To the best of our knowledge, no theoretical analysis for FHDMs is available that can explain their impressive algorithmic performance. In this work, we therefore provide the first statistical convergence analysis of FHDMs for the particular case of data supported on the unit sphere $\partial B \subset \mathbb{R}^d$ in dimension $d \geq 3$. Specifically, given i.i.d. data $\{X_1, \dots, X_n\}$ with α-Sobolev smooth density ($\alpha \in \mathbb{N} \cap ((d-1)/2, \infty)$) on the sphere that is uniformly bounded away from zero, we are interested in the total variation convergence rate of FHDMs in terms of the size n, dimension d and smoothness α of the data. Approximating the score $\nabla \log h$ with sparse ReLU-neural network functions, our main result, Theorem 3.1, demonstrates that the output distribution $\widehat{\Pi}_n$ of the learned model with an appropriate choice of n-dependent network sizes, converges at rate

$$\mathbb{E}[\mathrm{TV}(\Pi^*, \widehat{\Pi}_n)] \lesssim n^{-\frac{\alpha}{2\alpha+d-1}} (\log n)^{7/2}.$$

The rate matches, up to small log-factors, the minimax optimal nonparametric rate for the $(d-1)$-dimensional data distributions under consideration. This is also the first statistical optimality result on generative modelling with adaptive random generation time.

Related work Statistical theory for neural network based DDMs has made tremendous advances in recent years since the seminal work of [24], who were the first to prove statistical optimality in total variation of DDMs for Besov-smooth and lower bounded data densities on the cube $[-1, 1]^d$. They also show that DDMs can implicitly adapt to the data geometry by proving almost optimal 1-Wasserstein convergence rates for data supported in a lower-dimensional linear subspace, see also [6]. The latter result was further refined in [31] and [4] to almost optimal Wasserstein rates for general low-dimensional manifold structures, thus providing statistical evidence for the ability of DDMs to overcome the curse of dimensionality. For other classes of data distributions [34, 11, 17] also demonstrate that the convergence rate of standard DDMs only depends on different notions of intrinsic dimensionality d^*. Notably, the results of [34] do not require lower-boundedness of the data density, which has also been demonstrated for full-dimensional supports in [37] for kernel-based score approximations and more recently in [28] for neural network estimators. While all of the works above consider DDMs in their standard formulation with Gaussian transitions, some recent progress has also been made for statistical convergence of iterative generative models, which in a similar spirit to FHDMs, are designed to explicitly enforce geometric constraints on the data distribution such as [14, 13] for *reflected* diffusion models [21] and [33] for discrete diffusion models [22].

2 Setting

We first introduce the technical h-transform setup of [36] and provide some additional mathematical details that are needed for the statistical analysis. Some standard notation that we frequently use is summarised in Appendix K. Let Π^* be the target distribution, supported on the boundary of the (open) unit ball $B \subset \mathbb{R}^d$, where $d \in \mathbb{N} \cap [3, \infty)$. Let also W be a Brownian motion started in $z \in \overline{B}$ under $\mathbb{P}^z$ and Z be the Brownian motion absorbed in ∂B, that is

$$\mathrm{d}Z_t = \mathbf{1}_B(Z_t)\,\mathrm{d}W_t, \quad Z_0 = W_0 = z \in \overline{B}, \tag{1}$$

with law $\mathbb{Q}^z$. The canonical process ξ on the path space $(C([0,\infty); \mathbb{R}^d), \mathcal{G}_\infty, \mathbb{G}, \mathbb{Q}^z)$ is then a Brownian motion started in z and stopped in $\tau(\xi) := \inf\{t \geq 0 : \xi_t \in \partial B\}$, where $\mathbb{G} = (\mathcal{G}_t)_{t\geq 0}$ is the $(\mathbb{Q}^z)_{z\in\overline{B}}$ completion of the natural filtration of ξ and $\mathcal{G}_\infty := \sigma(\bigcup_{t\geq 0} \mathcal{G}_t)$.

The idea of [36] is to transform this process into a process Z^h using Doob's h-transform to achieve $Z^h_{\tau^h} \sim \Pi^*$, where $\tau^h := \inf\{t \geq 0 : Z^h_t \in \partial B\} = \tau(Z^h)$ is the first hitting time of the unit sphere. To this end, we abbreviate $\tau := \tau(W) = \tau(Z)$ for the first hitting time of ∂B (or, equivalently, the first exit time of B) by the Brownian motion W, and define $\mathbb{Q}_{\partial B}(z, \mathrm{d}x) := \mathbb{P}^z(W_\tau \in \mathrm{d}x)$, $z \in B$, $x \in \partial B$, to be the Poisson kernel of the Brownian motion. The Poisson kernel has a density wrt the $(d-1)$-dimensional Hausdorff measure σ, which is given by [12]

$$q(x \mid z) := \frac{\mathrm{d}\mathbb{Q}_{\partial B}(z, \cdot)}{\mathrm{d}\sigma}(x) = \frac{1}{\sigma(\partial B)} \frac{\|x\|^2 - \|z\|^2}{\|x\| \, \|z - x\|^d} = \frac{1}{\sigma(\partial B)} \frac{1 - \|z\|^2}{\|z - x\|^d}, \quad z \in B, x \in \partial B. \tag{2}$$

We assume that the target measure is absolutely continuous wrt the Poisson kernel with a fixed initialisation $z_0 \in B$ and let $\pi^* := \mathrm{d}\Pi^* / \, \mathrm{d}\mathbb{Q}_{\partial B}(z_0, \cdot)$ be its density. We then choose the function h that induces the desired h-transform as

$$h(z) := \mathbb{E}^z[\pi^*(Z_\tau)] = \begin{cases} \int_{\partial B} \pi^*(x) \, \mathbb{Q}_{\partial B}(z, \mathrm{d}x) = \int_{\partial B} \pi^*(x) q(x \mid z) \, \sigma(\mathrm{d}x), & z \in B \\ \pi^*(z), & z \in \partial B. \end{cases}$$

Furthermore, we will assume that $\sup_{y \in \partial B} \pi^*(y) < \infty$ and $\inf_{y \in \partial B} \pi^*(y) > 0$, which implies that h is bounded and that $h > 0$ on $\overline{B}$. The function h is harmonic for Z in the sense that for any $z \in \overline{B}$ and $t \geq 0$ it holds that $\mathbb{E}^z[h(Z_t)] = h(z)$, which follows straightforwardly from the Markov property of Z and τ being a terminal $\mathbb{F}$-stopping time, where $\mathbb{F}$ is the $(\mathbb{P}^z)_{z \in \overline{B}}$ completion of the natural filtration of Z. In particular, this yields that $(h(Z_t))_{t \geq 0}$ is an $\mathbb{F}$-martingale under any $\mathbb{P}^z$. This allows us to use Doob's h-transform technique [8, Chapter 11] to define Markov transition kernels $P^h_t(z, A) := \frac{1}{h(z)} \mathbb{E}^z[h(Z_t) \mathbf{1}_A(Z_t)]$ for $z \in \overline{B}, A \in \mathcal{B}(\overline{B})$, with the associated Markov process Z^h being again of diffusion type with a continuous version solving the absorbed SDE

$$\mathrm{d}Z^h_t = \mathbf{1}_B(Z^h_t)(\nabla \log h(Z^h_t) \, \mathrm{d}t + \mathrm{d}W_t), \quad Z^h_0 = Z_0. \tag{3}$$

In particular, $Z^h_0 = z$, $\mathbb{P}^z$-a.s. for all $z \in \overline{B}$. In analogy to denoising diffusion models, we call the additional term $\nabla \log h$ arising in (3) the *score*. We denote by $\mathbb{Q}^{h,z}$ the law of Z^h under $\mathbb{P}^z$ and set $\mathbb{Q}^{\Pi^*} \equiv \mathbb{Q}^h \equiv \mathbb{Q}^{h,z_0}$ to indicate that Z^h has terminal distribution Π^* when started in z_0. [8, Theorem 11.9] now yields for any $\mathbb{G}$-stopping time T that

$$\mathbb{Q}^{h,z}(\Lambda) = \frac{1}{h(z)} \int_\Lambda h(\xi_T) \, \mathrm{d}\mathbb{Q}^z, \quad \Lambda \in \mathcal{G}_T, z \in \overline{B}. \tag{4}$$

Noting that $h(z_0) = 1$, this yields $\mathbb{Q}^{\Pi^*}(\Lambda) = \int_\Lambda h(\xi_T) \, \mathrm{d}\mathbb{Q}^{z_0}$ and since $\xi_t = \xi_\tau$ and $\tau < \infty$, $\mathbb{Q}^{z_0}$-a.s., we obtain from this that generally

$$\mathbb{Q}^{\Pi^*}(\Lambda) = \int_\Lambda \pi^*(\xi_\tau) \, \mathrm{d}\mathbb{Q}^{z_0}, \quad \Lambda \in \mathcal{G}_\infty. \tag{5}$$

A *formal* calculation via disintegration then gives

$$\begin{aligned} \mathbb{Q}^{\Pi^*}(\mathrm{d}\omega) = \pi^*(\xi_\tau(\omega)) \, \mathbb{Q}^{z_0}(\mathrm{d}\omega) &= \int_{\partial B} \frac{\mathrm{d}\Pi^*}{\mathrm{d}\mathbb{Q}^{z_0}(\xi_\tau \in \cdot)}(x) \, \mathbb{Q}^{z_0}(\xi_\tau \in \mathrm{d}x) \, \mathbb{Q}^{z_0}(\mathrm{d}\omega \mid \xi_\tau = x) \\ &= \int_{\partial B} \mathbb{Q}^{z_0}(\mathrm{d}\omega \mid \xi_\tau = x) \, \Pi^*(\mathrm{d}x) = \int_{\partial B} \mathbb{Q}^{q(x \mid \cdot)}(\mathrm{d}\omega) \, \Pi^*(\mathrm{d}x), \end{aligned} \tag{6}$$

where $\mathbb{Q}^{q(x \mid \cdot)}$ is the law of the $q(x \mid \cdot)$-transform of the Brownian motion W killed on first hitting ∂B, started in z_0; see also [36, equation (8)]. For $Z^x_0 = z_0$, this transform is realised by the SDE

$$\mathrm{d}Z^x_t = \nabla_2 \log q(x \mid Z^x_t) \, \mathrm{d}t + \mathrm{d}W_t, \quad Z^x_0 = z_0, \, t < \tau^x := \inf\{t \geq 0 : Z^x \in \partial B\}, \tag{7}$$

where $\nabla_2 \log q(x \mid z)$ points with increasing force towards x as $z \to \partial B$, thus conditioning the SDE to hit the unit sphere in x just before being killed. Making (6) precise is technically delicate because of the pole of $q(x \mid \cdot)$ at x. We will not need or attempt to formally prove this statement thanks to an early stopping procedure that we introduce for the generation algorithm and the statistical analysis. For our purposes it therefore suffices that (7) is a well-behaved SDE up to the first hitting-time of the boundary of any slightly smaller ball $B_{1-\varepsilon}$ for some $\varepsilon \in (0, 1)$. This motivates a natural sampling procedure for the h-transform (3): first sample x from Π^* on ∂B and then simulate a first

hitting bridge Z^x according to (7). This procedure is of central importance for score-based generative modelling since the score $\nabla \log h$ is implicitly defined via the unknown data distribution Π^* and is thus inaccessible in practice. The score therefore needs to be approximated by a learnable function s based on paths of Z^h, which given the data set $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} \Pi^*$ are generated as above but sampling from the empirical data distribution instead of Π^*. An approximated process Z^s is then given by

$$\mathrm{d}Z_t^s = \mathbf{1}_B(Z_t^s)(s(Z_t^s)\,\mathrm{d}t + \mathrm{d}W_t), \quad Z_0^s = z_0 \in B. \tag{8}$$

We now have the following formula for the distribution of Z^h at its terminal time that further motivates the choice of h and yields a denoising score representation of $\nabla \log h$, proved in Appendix B.

Lemma 2.1. *For any $z \in B$, $x \in \partial B$ it holds that $\mathbb{P}^z(Z^h_{\tau^h} \in \mathrm{d}x) = \frac{\pi^*(x)}{h(z)}\,\mathbb{Q}_{\partial B}(z, \mathrm{d}x)$ and in particular $\mathbb{P}^{z_0}(Z^h_{\tau^h} \in \mathrm{d}x) = \Pi^*(\mathrm{d}x)$. Moreover, the score $\nabla \log h$ is given by $\nabla \log h(z) = \mathbb{E}^z\big[\nabla_2 \log q(Z^h_{\tau^h} \mid Z^h_0)\big]$ for any $z \in B$.*

Assumptions on the target distribution Π^* We define the space $L^2(\partial B)$ of square-integrable functions on ∂B as the set of functions that are integrable with respect to the surface measure σ. On a subspace of this space, we consider the *Laplace–Beltrami operator* $\Delta_{\partial B}$ [29], which is a generalisation of the Laplace operator on manifolds. $\Delta_{\partial B}$ is a self-adjoint operator wrt the $L^2(\partial B)$-scalar product on the sphere, with eigenvalues $l(l+d-2)$, $l \in \mathbb{N}$, and orthonormal eigenfunctions Y_{lm}, $m = 1, \dots, M_l$, where M_l denotes the dimension of the eigenspace to the lth eigenvalue. The Y_{lm} are the so-called *spherical harmonics*, which are well-known from harmonic analysis and form a complete basis of $L^2(\partial B)$. The natural domain of $\Delta_{\partial B}$ is given by $\mathrm{Dom}(\Delta_{\partial B}) \coloneqq \{u \in L^2(\partial B) : \mathbb{R}^d \ni z \mapsto u(z/\|z\|) \text{ is twice weakly differentiable}\}$, and the space $H^\alpha(\partial B)$ of Sobolev functions on ∂B with smoothness $\alpha \in \mathbb{N}$ is defined by

$$H^\alpha(\partial B) \coloneqq \left\{u \in \mathrm{Dom}(\Delta_{\partial B}) : \|u\|_{H^\alpha(\partial B)} < \infty\right\},$$

with the spherical Sobolev norm

$$\|u\|^2_{H^\alpha(\partial B)} \coloneqq \|u\|^2_{L^2(\partial B)} + \langle u, (-\Delta_{\partial B})^\alpha u\rangle_{L^2(\partial B)} = \sum_{l=0}^{\infty}\sum_{m=1}^{M_l}[1 + (l(l+d-2))^\alpha]\,|\langle u, Y_{lm}\rangle|^2.$$

We impose the following assumptions on the target distribution for the remainder of the paper.

- ($\mathcal{H}1$) $\pi^* \in H^\alpha(\partial B)$ for $\alpha \in \mathbb{N}$ and $\alpha > (d-1)/2 \geq 1$.
- ($\mathcal{H}2$) π^* is uniformly bounded away from zero, that is, $\pi^* \geq \pi_{\min}$ for a constant $\pi_{\min} > 0$.

According to the Sobolev embedding theorem for manifolds [2, Theorem 2.20], Assumption ($\mathcal{H}1$) implies that π^* is continuous and thus bounded from above as well. Assumption ($\mathcal{H}2$) is typical in the literature on statistical convergence rates of diffusion models [24, 31, 4, 14] and simplifies the approximation analysis of the score function $\nabla \log h = \nabla h/h$ since it implies that $h(z) \geq \pi_{\min} > 0$.

2.1 Generation and estimation strategy

As for forward-backward diffusion models with fixed time horizon, the main idea to approximate the unknown score $\nabla \log h$ is to use Girsanov's theorem [18, Theorem 5.22] to first express the KL divergence between the path measures induced by the approximating drift s and the "true" drift $\nabla \log h$ in terms of an L^2-loss along the paths $t \mapsto Z_t^h$ and to obtain from this an implementable L^2-loss via an equivalent *denoising score matching* objective. However, $\|\nabla \log h\|$ is not guaranteed to be bounded in a neighbourhood of ∂B, making the applicability of Girsanov's theorem and finiteness of the explicit score matching loss a delicate issue. To circumvent such problems, we do not simulate Z^h until its terminal time τ^h, but stop a little early at first hitting of the $(1-\varepsilon)$-sphere $\partial B_{1-\varepsilon}$ for small $\varepsilon > 0$. To generate a data distribution with the correct support, the simulated value of the early stopped process on $\partial B_{1-\varepsilon}$ is simply projected onto ∂B. Let us therefore introduce $\tau_{1-\varepsilon}(\omega) \coloneqq \inf\{t \geq 0 : \omega_t \in \partial B_{1-\varepsilon}\}$ for $\omega \in \mathbb{R}^{[0,\infty)}$ and set $\tau^h_{1-\varepsilon} \coloneqq \tau_{1-\varepsilon}(Z^h)$, $\tau^s_{1-\varepsilon} \coloneqq \tau_{1-\varepsilon}(Z^s)$ as the first hitting times of $\partial B_{1-\varepsilon}$ by Z^h and Z^s, respectively. If not said otherwise, we always assume that $\varepsilon < (1 - \|z_0\|)/2$ so that $\partial B_{1-\varepsilon}$ is well separated from the initialisation z_0. Furthermore, let $\mathbb{Q}^h_{\partial B_{1-\varepsilon}}$ and $\mathbb{P}^s_{\partial B_{1-\varepsilon}}$ denote the laws of the random variables $Z^h_{\tau^h_{1-\varepsilon}}$ and $Z^s_{\tau^s_{1-\varepsilon}}$ when the processes

are started in z_0, where Z^s solves the absorbed SDE (8) for some locally Lipschitz approximating function s. Let also

$$P_{\partial B}\colon B \setminus \{0\} \to \partial B, \quad x \mapsto \frac{x}{\|x\|},$$

be the projection onto the unit sphere. We now decompose the total variation distance between our target distribution Π^* and the distribution of the projection of the simulated $Z^s_{\tau^s_{1-\varepsilon}}$ according to

$$\begin{aligned}\mathrm{TV}(\Pi^*, P_{\partial B}\sharp\mathbb{P}^s_{\partial B_{1-\varepsilon}}) &\le \mathrm{TV}(\Pi^*, P_{\partial B}\sharp\mathbb{Q}^h_{\partial B_{1-\varepsilon}}) + \mathrm{TV}(P_{\partial B}\sharp\mathbb{Q}^h_{\partial B_{1-\varepsilon}}, P_{\partial B}\sharp\mathbb{P}^s_{\partial B_{1-\varepsilon}}) \\ &= \underbrace{\mathrm{TV}(\Pi^*, P_{\partial B}\sharp\mathbb{Q}^h_{\partial B_{1-\varepsilon}})}_{\text{early stopping error}} + \underbrace{\mathrm{TV}(\mathbb{Q}^h_{\partial B_{1-\varepsilon}}, \mathbb{P}^s_{\partial B_{1-\varepsilon}})}_{\text{generation error on } \partial B_{1-\varepsilon}}, \end{aligned} \tag{9}$$

where we used that the restriction of $P_{\partial B}$ is a bijection between $\partial B_{1-\varepsilon}$ and ∂B. Projecting $Z^s_{\tau^s_{1-\varepsilon}}$ onto ∂B is necessary to obtain meaningful total variation bounds, since $\mathrm{TV}(\Pi^*, \mathbb{P}^s_{\partial B_{1-\varepsilon}}) = 1$ by disjoint supports. To minimise the rhs of (9) for fixed $\varepsilon > 0$ and a given approximation class $\mathcal{S} \ni s$, we need to optimise $\mathrm{TV}(\mathbb{Q}^h_{\partial B_{1-\varepsilon}}, \mathbb{P}^s_{\partial B_{1-\varepsilon}})$ or a suitable upper bound thereof. By Pinsker's inequality, we have

$$\mathrm{TV}(\mathbb{Q}^{\Pi^*}_{\partial B_{1-\varepsilon}}, \mathbb{P}^s_{\partial B_{1-\varepsilon}})^2 \le \frac{1}{2}\mathbf{KL}(\mathbb{Q}^{\Pi^*}_\varepsilon \,\|\, \mathbb{P}^s_\varepsilon),$$

with the KL-divergence given by the following result, whose proof can be found in Appendix B.

Proposition 2.2. *Suppose that s is locally Lipschitz. Then, $\mathbb{Q}^h_\varepsilon \approx \mathbb{P}^s_\varepsilon$ and*

$$\mathrm{KL}(\mathbb{Q}^h_\varepsilon \,\|\, \mathbb{P}^s_\varepsilon) = \frac{1}{2}\mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|s(Z^h_t) - \nabla \log h(Z^h_t)\|^2 \,\mathrm{d}t\Big]. \tag{10}$$

Since the explicit score matching loss on the rhs of (10) depends explicitly on the unknown score $\nabla \log h$, we next derive a denoising score matching identity that allows us to replace $\nabla \log h$ by the Poisson kernel $q(x \mid \cdot)$ with observations $x \in \partial B$ for estimation purposes.

Proposition 2.3. *Let $z \in B$, $\varepsilon \in (0,1)$ and suppose that s is bounded on $B_{1-\varepsilon}$. For any $\varepsilon > 0$ and $\mathbb{F}$-stopping times $\underline{\tau} \le \overline{\tau} \le \tau^h_{1-\varepsilon}$, it holds that*

$$\mathbb{E}^z\Big[\int_{\underline{\tau}}^{\overline{\tau}} \|\nabla \log h(Z^h_t) - s(Z^h_t)\|^2 \,\mathrm{d}t\Big] = \mathbb{E}^z\Big[\int_{\underline{\tau}}^{\overline{\tau}} \|\nabla_2 \log q(Z^h_{\tau^h} \mid Z^h_t) - s(Z^h_t)\|^2 \,\mathrm{d}t\Big] + C, \tag{11}$$

where $C = -\mathbb{E}^z[\int_{\underline{\tau}}^{\overline{\tau}} \|\nabla_2 \log q(Z^h_{\tau^h} \mid Z^h_t) - \nabla \log h(Z^h_t)\|^2 \,\mathrm{d}t]$ is independent of s. Furthermore, for Z^x denoting the $q(x \mid \cdot)$-transform of Z given by (7) and $\tau^x_r \coloneqq \tau_r(Z^x)$ for $r \in [0,1]$, we have

$$\begin{aligned}&\mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|\nabla_2 \log q(Z^h_{\tau^h} \mid Z^h_t) - s(Z^h_t)\|^2 \,\mathrm{d}t\Big] \\ &\quad = \int_{\partial B} \mathbb{E}^{z_0}\Big[\int_0^{\tau^x_{1-\varepsilon}} \|\nabla_2 \log q(x \mid Z^x_t) - s(Z^x_t)\|^2 \,\mathrm{d}t\Big]\, \Pi^*(\mathrm{d}x).\end{aligned} \tag{12}$$

The proof is given in Appendix B. Given the data sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \Pi^*$, this result suggests to approximate $\nabla \log h$ via the empirical risk minimiser $\widehat{s}$ given by

$$\widehat{s} \coloneqq \operatorname*{arg\,min}_{s \in \mathcal{S}} \frac{1}{n}\sum_{i=1}^n L_s(X_i) \tag{13}$$

for the denoising score loss function

$$L_s(x) = L^\varepsilon_s(x) \coloneqq \mathbb{E}^{z_0}\Big[\int_0^{\tau^x_{1-\varepsilon}} \|s(Z^x_t) - \nabla_2 \log q(x \mid Z^x_t)\|^2 \,\mathrm{d}t\Big], \quad x \in \partial B. \tag{14}$$

For the concrete neural network approximation class $\mathcal{S}$ that we introduce next, existence of a measurable selection of the minimizer $\widehat{s}$ is guaranteed, see Appendix J for details. Let us also remark that an approximate minimizer up to a $\delta = 1/n$ error margin would suffice for our main result to hold. As approximation class $\mathcal{S}$ we choose sparse neural networks with ReLU activation

function, which we now briefly introduce. For some $m \in \mathbb{N}$ and $v, b \in \mathbb{R}^m$, define the shifted ReLU $\sigma^b(v) \coloneqq (\mathrm{ReLU}(v_i - b_i))_{i=1,\dots,m}^\top$. We then define the set of ReLU neural networks with number of layers $\mathsf{L} \in \mathbb{N}$, maximal width $\mathsf{W} \in \mathbb{R}_+$, sparsity constraint $\mathsf{S} \in \mathbb{N}$ and norm constraint $\mathsf{B} \in \mathbb{R}_+$, as

$$\begin{aligned}&\mathrm{NN}(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})\\ &\coloneqq \begin{array}{l}\Big\{A^{\mathsf{L}}\sigma^{b^{\mathsf{L}}} \cdots A^1\sigma^{b^1}A^0 : W \in \mathbb{N}^{\mathsf{L}+2}, A^i \in \mathbb{R}^{W_{i+1}\times W_i}, b^i \in \mathbb{R}^{W_{i+1}}, \|W\|_\infty \le \mathsf{W},\\ \sum_{i=1}^{\mathsf{L}}(\|A^i\|_0 + \|b^i\|_0) + \|A^0\|_0 \le \mathsf{S}, \quad \max_{i=1}^{\mathsf{L}}(\|A^i\|_\infty \vee \|b^i\|_\infty) \vee \|A^0\|_\infty \le \mathsf{B}\Big\}.\end{array}\end{aligned} \tag{15}$$

We summarise some (by now) standard results on sparse neural network approximations from the literature in Appendix I. Based on such neural networks, we define our approximation class $\mathcal{S}$ by

$$\mathcal{S} = \left\{\phi \in \mathrm{NN}(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B}) : \|\phi(z)\| \le \frac{12(d+2)}{1-\|z\|}\right\}, \tag{16}$$

which is motivated by the following score bound.

Lemma 2.4. *For any $z \in B$ it holds that $\|\nabla \log h(z)\| \le \frac{d+2}{1-\|z\|}$.*

Imposing uniform growth restrictions on the neural networks from the approximation class $\mathcal{S}$ is also a typical feature in statistical analysis of score-based generative models with deterministic time-horizon [24, 31, 4, 34, 14, 13]. In our case, the distance to the sphere as our target manifold can be regarded as an intrinsic time-scale of the algorithm since it reflects the average time left until termination. With this preparation we can now state the full generative Algorithm 1.

Algorithm 1 Generative algorithm

Input: data $\{X_1, \dots, X_n\} \overset{\text{iid}}{\sim} \Pi^*$

• choose an early stopping parameter ε and network class parameters $\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B}$ depending on the number n of samples

• determine the empirical denosing score loss minimiser $\widehat{s}$ according to (13) with $\mathcal{S}$ as in (16) with the chosen network parameters

• for a Brownian motion W independent of the data $\{X_1, \dots, X_n\}$, simulate the SDE $Z^{\widehat{s}}$

$$\mathrm{d}Z_t^{\widehat{s}} = \widehat{s}(Z_t^{\widehat{s}})\,\mathrm{d}t + \mathrm{d}W_t, \quad Z_0^{\widehat{s}} = z_0,$$

until its first hitting time $\widehat{\tau}_{1-\varepsilon}$ of the $(1-\varepsilon)$-sphere $\partial B_{1-\varepsilon}$.

Output: projected value $P_{\partial B} Z^{\widehat{s}}_{\widehat{\tau}_{1-\varepsilon}}$ as new (approximate) sample for Π^*.

The generative algorithm uses the simplifying assumption that we can evaluate $L_s(x)$ and that for a score estimator $\widehat{s}$ we can exactly simulate the corresponding absorbed SDE. In practice, $L_s(x)$ needs to be numerically approximated based on discretely sampled paths of Z^x, which for $z_0 = 0$ can be efficiently implemented via the simple rotation algorithm given in [36, Proposition 2.11].

3 Main result

Our main result is the following.

Theorem 3.1. *Let $\varepsilon = n^{-\alpha/(\beta(2\alpha+d-1))}$ for $\beta = (\alpha - (d-1)/2) \wedge 1$. Then for Π^* satisfying assumptions ($\mathcal{H}1$) and ($\mathcal{H}2$), there exist neural network size parameters of order*

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

such that Algorithm 1 with $\mathcal{S}$ chosen as in (16) *produces an output $P_{\partial B} Z^{\widehat{s}}_{\widehat{\tau}_{1-\varepsilon}}$ such that*

$$\mathbb{E}[\mathrm{TV}(\Pi^*, P_{\partial B}\sharp\mathbb{P}^{\widehat{s}}_{\partial B_{1-\varepsilon}})] \lesssim n^{-\alpha/(2\alpha+d-1)} (\log n)^{7/2},$$

where $\mathrm{TV}(\Pi^*, P_{\partial B}\sharp\mathbb{P}^{\widehat{s}}_{\partial B_{1-\varepsilon}}) \coloneqq \mathrm{TV}(\Pi^*, P_{\partial B}\sharp\mathbb{P}^{s}_{\partial B_{1-\varepsilon}})|_{s=\widehat{s}}$ is the total variation distance between Π^* and the law of $P_{\partial B} Z^{\widehat{s}}_{\widehat{\tau}_{1-\varepsilon}}$ given the estimator $\widehat{s}$.

By the following proposition, Theorem 3.1 demonstrates that up to log-factors, the convergence rate of FHDMs matches the minimax lower bound for total variation density estimation for spherical data. The proof of the lower bound is deferred to Appendix C.

Proposition 3.2. *Let $\mathcal{B}(H^\alpha(\partial B), L) \coloneqq \{\pi \in H^\alpha(\partial B) : \|\pi\|_{H^\alpha(\partial B)} \leq L, \int_{\partial B} \pi \, \mathrm{d}\sigma = 1\}$ be the space of α-Sobolev probability densities on the sphere with Sobolev norm bounded by $L \in (0, \infty)$. Then, it holds that*

$$\inf_{\widehat{\rho}} \sup_{\pi \in \mathcal{B}(H^\alpha(\partial B), L)} \mathbb{E}_\pi[\mathrm{TV}(\pi, \widehat{\rho})] \gtrsim n^{-\alpha/(2\alpha+d-1)},$$

where the infimum is taken over all random probability densities $\widehat{\rho}$ on ∂B that are measurable wrt the data $\{X_1, \dots, X_n\} \overset{iid}{\sim} \pi$ under $\mathbb{P}_\pi$.

Note that, up to the choice of L, it is irrelevant whether the functions in $\mathcal{B}(H^\alpha(\partial B), L)$ are chosen as densities wrt σ or wrt the harmonic measure $q(\cdot \mid z_0) \, \mathrm{d}\sigma$: for $z_0 \notin \partial B$ the function $q(\cdot \mid z_0)$ is arbitrarily smooth and bounded from below, so $\pi \in H^\alpha(\partial B)$ if and only if $\pi / q(\cdot \mid z_0) \in H^\alpha(\partial B)$; furthermore, the total variation distance is invariant to the choice of dominating measure.

The proof strategy of Theorem 3.1 follows a similar pattern to previous statistical work on diffusion models. However, the technical details are of a fundamentally different nature and impose new challenges because of the stochastic termination criterion, which places the analytic framework quite naturally into potential theory for Markov processes [5, 8]. We give a short breakdown of the proof below before giving detailed intermediate results in the next section.

In a first step, we control the early stopping error in (9), where we show that the total variation distance between the target distribution Π^* and the projected distribution of the h-transform stopped in $\partial B_{1-\varepsilon}$ is at most of order $\varepsilon^{\beta/2}$, where β is the minimally guaranteed Hölder smoothness of the target density π^*. This allows us to focus on the second error component in (9), which is the generation error on the slightly smaller sphere $\partial B_{1-\varepsilon}$. By Proposition 2.2 and Pinsker's inequality, this error is characterised by the expected explicit score matching loss (10) of the empirical risk minimiser $\widehat{s}$. Given the equivalence of explicit and denoising score matching loss (Proposition 2.3) this can be decomposed into an approximation error and an error influenced by the complexity of the class $\mathcal{L} = \{L_s : s \in \mathcal{S}\}$ of denoising score losses, cf. Theorem 4.2.(i), which needs to be balanced via an appropriate selection of the network parameters to achieve the optimal rate.

The complexity term is determined by the covering number of $\mathcal{L} = \{L_s : s \in \mathcal{S}\}$ built from $\mathcal{S}$ and the size of the uniform control $C(\mathcal{L}) = \sup_{s \in \mathcal{S}} \|L_s\|_\infty$. These components are controlled in terms of the covering number of the neural network class $\mathcal{S}$ and a term $\mathcal{O}(\log \varepsilon^{-1})$, respectively, exploiting the growth rate of the neural networks and the score $\nabla \log h$ near ∂B. The approximation error analysis of the neural network class is carried out in three main steps: First, the score is approximated on the $(d-1)$-dimensional coordinate space, yielding convergence rates depending on the manifold dimension $(d-1)$ instead of the ambient dimension d. Then, the coordinate map $\varphi_{\partial B} \colon \partial B \to \mathbb{R}^{d-1}$ is approximated, which is cheap in terms of neural network size due to the arbitrary smoothness of $\varphi_{\partial B}$. Finally, the constructed neural networks are concatenated to approximate the score on the ambient space. This yields a parsimonious choice for sufficient network sizes that allows (up to log factors) approximation at rate $\mathcal{O}(n^{-\frac{\alpha}{2\alpha+d-1}})$, while keeping the complexity term at the same order.

4 Central results for the proof of Theorem 3.1

4.1 Early-stopping bound

We start with the error contribution by stopping the generation procedure early on $\partial B_{1-\varepsilon}$ and then projecting onto the sphere ∂B. This is described by the first term in (9) and can be bounded as follows in terms of the early stopping parameter ε and the minimally guaranteed Hölder smoothness β of the data density π^*, which is proved in Appendix D

Proposition 4.1. *Let $\beta \coloneqq (\alpha - (d-1)/2) \wedge 1$. Then, for a constant C independent of ε,*

$$\mathrm{TV}(\Pi^*, P_{\partial B} \sharp \mathbb{Q}^h_{\partial B_{1-\varepsilon}}) \leq C \varepsilon^{\beta/2}.$$

4.2 Generalisation error

We now treat in detail the second error component in (9), which for $s = \widehat{s}$ encodes the total variation error between the output of our generation procedure given in Algorithm 1 and the output generated by the true h-transform Z^h stopped early in $\partial B_{1-\varepsilon}$. With Pinsker's inequality and Proposition 2.2, this can be controlled in expectation via the following generalisation bound of the learned score $\widehat{s}$.

Theorem 4.2. *Let $\delta > 0$, L_s, $\mathcal{S}$, $\widehat{s}$ be as in Algorithm 1 and $C(\mathcal{L}) \coloneqq \sup_{s\in\mathcal{S}} \|L_s\|_{L^\infty(\partial B)}$. Denote by $\mathcal{N}(\mathcal{A}, \|\cdot\|, \delta)$ the δ-covering number of a function class $\mathcal{A}$ with respect to a norm $\|\cdot\|$.*

(i) If $\mathcal{N}(\mathcal{L}, \|\cdot\|_{L^\infty(\partial B)}, \delta) \geq 3$, then

$$\mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|\widehat{s}(Z^h_t) - \nabla \log h(Z^h_t)\|^2 \,\mathrm{d}t\Big] \lesssim \inf_{s\in\mathcal{S}} \mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|s(Z^h_t) - \nabla \log h(Z^h_t)\|^2 \,\mathrm{d}t\Big] + C(\mathcal{L})\frac{\log \mathcal{N}(\mathcal{L}, \|\cdot\|_{L^\infty(\partial B)}, \delta)}{n} + \delta. \tag{17}$$

(ii) For a constant c independent of ε and $\mathcal{S}$ we have $\mathcal{N}(\mathcal{L}, \|\cdot\|_{L^\infty(\partial B)}, \delta) \leq \mathcal{N}\big(\mathcal{S}, \|\cdot\|_{L^\infty(B_{1-\varepsilon})}, c\delta/\sqrt{C(\mathcal{L})}\big)$.

(iii) It holds that $C(\mathcal{L}) \lesssim \log \varepsilon^{-1} + 1$.

While the proof of part (i) follows along the lines of [24, Theorem 4.3] with the correction pointed out in [14, Appendix B] based on Proposition 2.3, we emphasize that the uniform score loss bound in (iii) is considerably more involved than the corresponding result in [24]. In fact, proving (iii) for standard fixed-time diffusion models follows from a two-line calculation using the $1/t$ growth rate of the score, which, integrated from $\underline{T}$ to $\overline{T}$, directly gives a $\mathcal{O}(\log \underline{T}^{-1})$ bound. In our setting such a direct analytic approach is not possible due to the random integration time interval, which therefore requires martingale and h-transform techniques. The proof of Theorem 4.2 is given in Appendix E.

4.3 Approximation error

The above results allow us to bound the complexity term fully in terms of ε and the covering number of the neural network class $\mathcal{S}$. To control the latter, we need an efficient neural network approximator for the score given the target rate $n^{-\alpha/(2\alpha+d-1)}$ of the approximation error

$$\inf_{s\in\mathcal{S}} \mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|s(Z^h_t) - \nabla \log h(Z^h_t)\|^2 \,\mathrm{d}t\Big] = \inf_{s\in\mathcal{S}} \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}} \|s(W_t) - \nabla \log h(W_t)\|^2 h(W_t) \,\mathrm{d}t\Big]. \tag{18}$$

This is achieved by the following main result of this section.

Theorem 4.3. *For any $N \in \mathbb{N} \cap [3, \infty)$ and $\varepsilon > 0$ small enough, there exists a neural network $s \in \mathrm{NN}(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})$ with $\mathsf{L} \lesssim \log N \log\log N + \log^2 \varepsilon^{-1}$, $\mathsf{W} \lesssim N \log^2 N$, $\mathsf{S} \lesssim N \log^3 N + \log^2 \varepsilon^{-1}$, and $\mathsf{B} \lesssim \mathrm{Poly}(N) \vee \varepsilon^{-4}$, such that*

$$\mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|s(Z^h_t) - \nabla \log h(Z^h_t)\|^2 \,\mathrm{d}t\Big] \lesssim N^{-2\alpha/(d-1)} \log N \log \varepsilon^{-1}.$$

The proof along with the proofs of intermediate approximation results given below can be found in Appendix F. A core difference for the proof of Theorem 4.3 as compared to the approximation analysis in fixed time denoising diffusion models is that the space-time analysis of (18) cannot be disentangled since the stopping time $\tau^h_{1-\varepsilon}$ depends on the whole path of Z^h. Instead, we use that the path integral of stopped Brownian motion can be expressed in terms of its potential measure, that is a spatial Lebesgue integral weighted by the *Green kernel* $G_{1-\varepsilon}$ of the domain $B_{1-\varepsilon}$, cf. (50) for its explicit formula, which measures the time spent by the stopped Brownian motion in a Borel set $A \subset B_{1-\varepsilon}$ via the occupation formula $\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}} \mathbf{1}_A(W_t) \,\mathrm{d}t\Big] = \int_A G_{1-\varepsilon}(z_0, z) \,\mathrm{d}z$. We split the spatial integral into two sub-domains: a fixed interior ball, where the Green kernel is singular, but the integrand is uniformly bounded, and the remaining annulus, where the Green kernel $G_{1-\varepsilon}(z, z_0)$ decreases linearly to 0 with the distance of z to the boundary, cf. Lemma H.3, thus limiting the average time spent by the Brownian motion in a critical environment of the sphere before absorption.

Proposition 4.4. *Let $\|z_0\| < R < 1 - \varepsilon$ be some fixed radius. Then,*

$$\mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|s(Z^h_t) - \nabla \log h(Z^h_t)\|^2 \,\mathrm{d}t\Big] = \int_{B_{1-\varepsilon}} G_{1-\varepsilon}(z, z_0)\|s(z) - \nabla \log h(z)\|^2 \, h(z)\,\mathrm{d}z$$
$$\lesssim \Big\|(s - \nabla \log h)\sqrt{G_{1-\varepsilon}(z_0, \cdot)}\Big\|^2_{L^2(B_{1-\varepsilon}/B_R)} + \|s - \nabla \log h\|^2_{L^\infty(B_R)}. \tag{19}$$

L^∞ approximation on B_R The approximation of the score on B_R is straightforward since $\nabla \log h|_{B_R}$ is smooth with supremum norm of the derivatives depending only on the fixed value R, which we choose independently of ε and n. Thus, we can apply approximation results from [30] to $\nabla \log h$ to obtain the following result, proved in Appendix F.

Proposition 4.5. *For $N \in \mathbb{N}$, there exists a neural network $\underline{s} \in \mathrm{NN}(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})$ with $\mathsf{L} \lesssim \log N$, $\mathsf{W} \lesssim N$, $\mathsf{S} \lesssim N \log N$, and $\mathsf{B} = \mathrm{Poly}(N)$, such that*

$$\|\nabla \log h - \underline{s}\|_{L^\infty(B_R)} \lesssim N^{-\alpha/(d-1)}.$$

L^2 approximation on $B_{1-\varepsilon} \setminus B_R$ The approximation of the score on $B_{1-\varepsilon}$ requires a more careful construction of a neural network (NN), since the Sobolev norm of $\nabla \log h$ diverges as ε decreases to zero. Our approximation strategy is inspired by the approach from [14] for neural network approximation (NNA) of space-time functions, where in our setting the radial coordinate plays the rôle of the time variable. The approximation strategy can be summarised as follows:

1. We start with approximating h and ∇h by functions h_N and ∇h_N obtained from truncating orthogonal series representations of h, h_N wrt the spherical harmonic basis.
2. For fixed radii r_i we then construct spherical NNAs of $h_N(r_i, \cdot)$ and $\nabla h_N(r_i, \cdot)$ on the coordinate space, exploiting the induced smoothness from the target π^*.
3. The NNAs for fixed radii are then combined via polynomial interpolation in r to obtain NNAs of $(r, x) \mapsto h_N(r, x), \nabla h_N(r, x)$, which are in turn approximated by NNs. We show that it suffices to choose the number of radii of order $\log N$ for the desired convergence rate.
4. The above results in NNAs $s_{h_N \circ \varphi_\pm^{-1}}$ and $s_{\nabla h_N \circ \varphi_\pm^{-1}}$ in spherical coordinates for stereographical coordinate maps $\varphi_\pm$. We concatenate these with NNAs of $\varphi_\pm$ to get NNAs of h_N and ∇h_N in cartesian space, which are then combined into a NNA of $\nabla \log h \approx \nabla h_N / h_N$.

The function h has a unique spherical representation $\widetilde{h}\colon (0,1) \times \partial B \to \mathbb{R}$ specified by $h(z) = \widetilde{h}(\|z\|, z/\|z\|)$, $z \in B \setminus \{0\}$. To not overburden notation, we identify $h = \widetilde{h}$. The same convention will be used for ∇h and for their approximations h_N and ∇h_N, which we now introduce. In order to approximate h by a polynomial in the radial coordinate, we first note that the Poisson kernel can be represented in terms of the spherical harmonics Y_{lm} by $q(y \mid z) = \sum_{l=0}^\infty \|z\|^l \sum_{m=1}^{M_l} Y_{lm}(y) Y_{lm}(z/\|z\|)$ for $y \in \partial B, z \in B \setminus \{0\}$. A straightforward calculation shows that the right hand side is indeed harmonic and that for any $f \in L^2(\partial B)$, $\langle q(\cdot \mid x), f\rangle_{L^2(\partial B)} = f(x)$ for all $x \in \partial B$, which is also known as the reproducing property of the Poisson kernel [3]. As a consequence, h is given by

$$h(r, x) = \sum_{l=0}^\infty r^l \sum_{m=1}^{M_l} Y_{lm}(x) \underbrace{\int_{\partial B} Y_{lm}(y)\pi^*(y)\,\sigma(\mathrm{d}y)}_{=:a_{lm}},$$

whence a natural approximation h_N is obtained by truncation of the representing series, that is, $h_N(r, x) := \sum_{l=0}^N \sum_{m=1}^{M_l} a_{lm} r^l Y_{lm}(x)$. For $\nabla_{\partial B}$ the covariant derivative on the sphere, the gradient of h_N is then given by

$$\nabla h_N(r, x) = \sum_{l=1}^N \sum_{m=1}^{M_l} a_{lm}\big(l r^{l-1} Y_{lm}(x) x + r^{l-1} \nabla_{\partial B} Y_{lm}(x)\big). \tag{20}$$

The approximation rates of h_N and ∇h_N are given in the following lemma.

Lemma 4.6. *It holds*

(i) $\|\nabla h_N(h-h_N)\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|^2_{L^2(B_{1-\varepsilon}\setminus B_R)} \lesssim \log\varepsilon^{-1}(1-\varepsilon)^{2N+d+1}N^{-2\alpha}\|\pi^*\|^4_{H^\alpha(\partial B)}$,

(ii) $\|\nabla h - \nabla h_N\|^2_{L^2(B_{1-\varepsilon})} \lesssim (1-\varepsilon)^{2N+d}N^{-2\alpha+1}\|\pi^*\|^2_{H^\alpha(\partial B)}$.

What mainly remains is the approximation of the simpler functions h_N and ∇h_N by NNs designed to achieve an L^2-approximation rate of order $\mathcal{O}(N^{-\alpha/(d-1)})$ up to logarithmic factors in N and ε^{-1} while requiring at most $\operatorname{Poly}\log(N)$ network layers and $N\operatorname{Poly}\log(N)$ non-zero parameters in order to control the complexity of the NN class in Theorem 4.2.(i).

Proposition 4.7. *There exist NNs s_{h_N} and $s_{\nabla h_N}$ in the NN class $\mathcal{S}(\mathsf{L},\mathsf{W},\mathsf{S},\mathsf{B})$ with $\mathsf{L} \lesssim \log N\log\log N + \log^2\varepsilon^{-1}$, $\mathsf{W}\lesssim N\log^2 N$, $\mathsf{S}\lesssim N\log^3 N+\log^2\varepsilon^{-1}$, and $\mathsf{B}\lesssim N^{1/(d-1)}\vee\varepsilon^{-4}$, which achieve*

$$\|s_{\nabla h_N}(h_N - s_{h_N})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_{1-\varepsilon}/B_R)} \lesssim N^{-\alpha/(d-1)}\log N\log\varepsilon^{-1},$$
$$\|(\nabla h_N - s_{\nabla h_N})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_{1-\varepsilon}/B_R)} \lesssim N^{-\alpha/(d-1)}\log N\log\varepsilon^{-1}.$$

Combining the two previous results with Proposition 4.5 gives us the main ingredients for achieving the overall approximation rate of the score in Theorem 4.3. Detailed derivations are given in Appendix F. Putting together Theorems 4.2 and Theorem 4.3 with the choices $N = n^{(d-1)/(2\alpha+d-1)}$ and $\delta = n^{-2\alpha/(2\alpha+d-1)}$ now easily yields Theorem 3.1. The full proof is given in Appendix G.

5 Conclusion

In this paper, we gave the first statistical optimality results for iterative generative models with random termination criterion by showing that FHDMs achieve the minimax optimal convergence rate $\mathcal{O}(n^{-\alpha/(2\alpha+d-1)})$ up to logarithmic factors for spherical data with α-Sobolev smooth and uniformly lower bounded density π^*, where $\alpha \in ((d-1)/2,\infty)\cap\mathbb{N}$. Given the stochastic termination criterion and the time-homogeneous nature of the algorithm, both the probabilistic as well as the score approximation analysis imposed several new challenges compared to the analysis of denoising diffusion models with deterministic time horizon, which we solved based on arguments rooted in general Markov processes theory and spherical Riemannian geometry.

To finish the paper, let us comment on limitations that arise from our setting and the made assumptions. The restriction on α implies that the density must be Hölder smooth and it is a technically challenging question how to extend this to a non-smooth setting, where we have no access to Sobolev embeddings, similarly to [24] and their general Besov smoothness framework in the context of DDMs. The lower boundedness assumption on π^*, albeit a common assumption in the field of statistics for generative models, also imposes a notable restriction on the generality of the statistical setting and is highly non-trivial to overcome.

Extending the convergence analysis of FHDMs to target manifolds $\mathcal{M}$ other than the sphere is demanding. On the one hand, the general score matching results from section 2 and the general risk decomposition extend without any problems, where a suitable early stopping criterion becomes the first hitting time of an ε-environment of $\mathcal{M}$. Moreover, [16] shows that the Poisson kernel for more general smooth domains is asymptotically comparable to the Poisson kernel for the sphere near the boundary, which allows to extend some technical arguments in this paper that are based on the explicit form of the Poisson kernel. On the other hand, our score approximation strategy critically depends on the geometry and smoothness of $\mathcal{M}$, with the availability of a suitable expansion of the Poisson kernel being a central feature of our analysis. It should, however, also be noted that FHDMs are specifically designed for data supports admitting an easily implementable Poisson kernel such that highly complicated or even unknown domains require a different ansatz for random adaptive termination rules, see e.g. [7]. Finally, we emphasise that we have not considered discrete sampling effects for both the training as well as the generation procedure as discussed in Section 2.1, which, given the stochastic termination criterion, is an interesting question from a numerical point of view for future work.

Acknowledgments and Disclosures

The authors received no third-party funding or third-party support for this work. There are no competing financial or personal interests to declare.

References

[1] Charalambos D. Aliprantis and Kim C. Border. *Infinite dimensional analysis*. 3rd ed. A hitchhiker's guide. Springer, Berlin, 2006, pp. xxii+703.

[2] Thierry Aubin. *Some nonlinear problems in Riemannian geometry*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 1998, pp. xviii+395. URL: https://doi.org/10.1007/978-3-662-13006-3.

[3] Sheldon Axler, Paul Bourdon, and Wade Ramey. *Harmonic function theory*. 2nd ed. Vol. 137. Graduate Texts in Mathematics. Springer-Verlag, New York, 2001, pp. xii+259. URL: https://doi.org/10.1007/978-1-4757-8137-3.

[4] Iskander Azangulov, George Deligiannidis, and Judith Rousseau. *Convergence of Diffusion Models Under the Manifold Hypothesis in High-Dimensions*. 2025. arXiv: 2409.18804 [stat.ML].

[5] R. M. Blumenthal and R. K. Getoor. *Markov processes and potential theory*. Vol. Vol. 29. Pure and Applied Mathematics. Academic Press, New York-London, 1968, pp. x+313.

[6] 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*. Vol. 202. 2023, pp. 4672–4712. URL: https://proceedings.mlr.press/v202/chen23o.html.

[7] Sören Christensen, Jan Kallsen, Claudia Strauch, and Lukas Trottner. "Beyond Fixed Horizons: A Theoretical Framework for Adaptive Denoising Diffusions". In: *Trans. Mach. Learn. Res.* (2026). URL: https://openreview.net/forum?id=xowTpMf9fM.

[8] Kai Lai Chung and John B. Walsh. *Markov processes, Brownian motion, and time symmetry*. Second. Vol. 249. Grundlehren der mathematischen Wissenschaften. Springer, New York, 2005, pp. xii+431. URL: https://doi.org/10.1007/0-387-28696-9.

[9] Kai Lai Chung and Zhong Xin Zhao. *From Brownian motion to Schrödinger's equation*. Vol. 312. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, 1995, pp. xii+287. URL: https://doi.org/10.1007/978-3-642-57856-4.

[10] Dennis Elbrächter, Dmytro Perekrestenko, Philipp Grohs, and Helmut Bölcskei. "Deep neural network approximation theory". In: *IEEE Trans. Inform. Theory* 67.5 (2021), pp. 2581–2623. URL: https://doi.org/10.1109/TIT.2021.3062161.

[11] 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]. URL: https://arxiv.org/abs/2510.03994.

[12] David Gilbarg and Neil S. Trudinger. *Elliptic partial differential equations of second order*. Classics in Mathematics. Reprint of the 1998 edition. Springer-Verlag, Berlin, 2001, pp. xiv+517.

[13] Asbjørn Holk, Claudia Strauch, and Lukas Trottner. *Reflected diffusion models adapt to low-dimensional data*. 2026. arXiv: 2603.24495 [math.ST]. URL: https://arxiv.org/abs/2603.24495.

[14] Asbjørn Holk, Claudia Strauch, and Lukas Trottner. "Statistical guarantees for denoising reflected diffusion models". In: *Journal of Machine Learning Research* 27.116 (2026), pp. 1–50. URL: http://jmlr.org/papers/v27/25-1588.html.

[15] K. K. J. Kinateder, Patrick McDonald, and David Miller. "Exit time moments, boundary value problems, and the geometry of domains in Euclidean space". In: *Probab. Theory Related Fields* 111.4 (1998), pp. 469–487. URL: https://doi.org/10.1007/s004400050174.

[16] Steven G. Krantz. "Calculation and estimation of the Poisson kernel". In: *J. Math. Anal. Appl.* 302.1 (2005), pp. 143–148. URL: https://doi.org/10.1016/j.jmaa.2004.08.010.

[17] 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. URL: http://jmlr.org/papers/v27/25-0121.html.

[18] Jean-François Le Gall. *Brownian Motion, Martingales, and Stochastic Calculus*. First. Graduate Texts in Mathematics. Springer Cham, 2016, pp. xxiii+273. URL: https://doi.org/10.1007/978-3-319-31089-3.

[19] John M. Lee. *Introduction to smooth manifolds*. Second. Vol. 218. Graduate Texts in Mathematics. Springer, New York, 2013, pp. 28–29. URL: https://doi.org/10.1007/978-1-4419-9982-5.

[20] John M. Lee. *Riemannian manifolds*. Vol. 176. Graduate Texts in Mathematics. An introduction to curvature. Springer-Verlag, New York, 1997, pp. 35–36. URL: https://doi.org/10.1007/b98852.

[21] Aaron Lou and Stefano Ermon. "Reflected Diffusion Models". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. 2023, pp. 22675–22701. URL: https://proceedings.mlr.press/v202/lou23a.html.

[22] Aaron Lou, Chenlin Meng, and Stefano Ermon. "Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution". In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. 2024, pp. 32819–32848. URL: https://proceedings.mlr.press/v235/lou24a.html.

[23] Jan-Henrik Metsch. *Parabolic and Elliptic Schauder Theory on Manifolds for a Fourth-Order Problem with a First- and a Third-Order Boundary Condition*. 2023. arXiv: 2304.04184 [math.AP].

[24] Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. "Diffusion Models are Minimax Optimal Distribution Estimators". In: *International Conference on Machine Learning*. 2023. URL: https://openreview.net/forum?id=ORyo7fxcIA.

[25] L. C. G. Rogers and David Williams. *Diffusions, Markov processes, and martingales. Vol. 2*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000, pp. xiv+480. URL: https://doi.org/10.1017/CBO9781107590120.

[26] R. T. Seeley. "Spherical harmonics". In: *Amer. Math. Monthly* 73.4 (1966), pp. 115–121. URL: https://doi.org/10.2307/2313760.

[27] 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. URL: https://openreview.net/forum?id=PxTIG12RRHS.

[28] Arthur Stéphanovitch, Eddie Aamari, and Clément Levrard. *Generalization bounds for score-based generative models: a synthetic proof*. 2025. arXiv: 2507.04794 [math.ST].

[29] M.A. Šubin. *Pseudodifferential Operators and Spectral Theory*. Soviet Mathematics Series. Springer-Verlag, 1987, pp. 167–169.

[30] 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. URL: https://openreview.net/forum?id=H1ebTsActm.

[31] Rong Tang and Yun Yang. "Adaptivity of Diffusion Models to Manifold Structures". In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. Vol. 238. Proceedings of Machine Learning Research. 2024, pp. 1648–1656. URL: https://proceedings.mlr.press/v238/tang24a.html.

[32] Lloyd N. Trefethen. *Approximation theory and approximation practice*. Extended. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2020, pp. xi+363. URL: https://doi.org/10.1137/1.9781611975949.

[33] Shintaro Wakasugi and Taiji Suzuki. "State Size Independent Statistical Error Bound for Discrete Diffusion Models". In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2026. URL: https://openreview.net/forum?id=GCwJd0JH5k.

[34] Konstantin Yakovlev and Nikita Puchkin. "Generalization error bound for denoising score matching under relaxed manifold assumption". In: *Proceedings of Thirty Eighth Conference on Learning Theory*. Vol. 291. 2025, pp. 5824–5891. URL: https://proceedings.mlr.press/v291/yakovlev25a.html.

[35] Yuhong Yang and Andrew Barron. "Information-theoretic determination of minimax rates of convergence". In: *Ann. Statist.* 27.5 (1999), pp. 1564–1599. URL: https://doi.org/10.1214/aos/1017939142.

[36] Mao Ye, Lemeng Wu, and Qiang Liu. "First Hitting Diffusion Models for Generating Manifold, Graph and Categorical Data". In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 27280–27292. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/ae87d80f5a0f3ee5c5643448f9599d1b-Paper-Conference.pdf.

[37] 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*. 2024.

A Technical results on hitting times and h-transforms

Lemma A.1. *Let Z be the strong solution of*

$$\mathrm{d}Z_t = b(Z_t)\,\mathrm{d}t + \sigma(Z_t)\,\mathrm{d}W_t, \quad Z_0 \sim \eta,$$

where $b\colon \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma\colon \mathbb{R}^d \to \mathbb{R}^{d\times n}$, $d \le n$, are Lipschitz functions, $\sigma\sigma^\top$ is uniformly positive definite, ν is concentrated on B and W is a n-dimensional Brownian motion independent of Z_0. Denote by τ the first exit time of Z out of B. Then

$$\mathbb{E}[\tau^k] < \infty \quad \textit{for all } k \in \mathbb{N}. \tag{21}$$

Proof. According to [15], the function

$$u_k(x) = \mathbb{E}^x[\tau^k \mid Z_0 = x], \quad x \in B$$

is the solution of the iterated PDE with $u_0 = 1$:

$$\begin{cases} Au_k = -ku_{k-1} & \text{on } B, \\ u_k = 0 & \text{on } \partial B, \end{cases}$$

where A is the infinitesimal generator of Z. We show that solutions in $C^2(B)$ exist for all $k \in \mathbb{N}$ by induction. Since both u_0 and the diffusion coefficients are in $C^\gamma(B)$ for any $\gamma < 1$, the solution $u_1 \in C^{2+\gamma}(\overline{B})$ exists by [12, Corollary 6.9.]. Thus, for fixed $k \in \mathbb{N}$, $u_{k-1} \in C^\gamma(\overline{B})$ and so $u_k \in C^{2+\gamma}(\overline{B})$ exists. Using that $\overline{B}$ is compact and u_k is continuous, we get $\sup_{x\in\overline{B}} \mathbb{E}[\tau^k \mid Z_0 = x] < \infty$ and finally

$$\mathbb{E}[\tau^k] = \int_{\overline{B}} \mathbb{E}[\tau^k \mid Z_0 = x]\,\eta(\mathrm{d}x) < \infty.$$

□

Lemma A.2. *For any $z \in B_{1-\varepsilon}$ it holds that $\mathbb{E}^z[\|W_\tau - W_{\tau_{1-\varepsilon}}\|] \le \sqrt{\varepsilon(2-\varepsilon)}$.*

Proof. By Jensen's inequality and the strong Markov property of Brownian motion [8, Chapter 2.3, Theorem 3], it follows that

$$\begin{aligned} \mathbb{E}^z[\|W_\tau - W_{\tau_{1-\varepsilon}}\|] &\le \left(\mathbb{E}^z[\|W_\tau - W_{\tau_{1-\varepsilon}}\|^2]\right)^{\frac{1}{2}} \\ &= \mathbb{E}^z\left[\mathbb{E}^{W_{\tau_{1-\varepsilon}}}[\|W_\tau - W_0\|^2]\right]^{1/2} \end{aligned} \tag{22}$$

where we used for the second line that $\tau_{1-\varepsilon} \le \tau$ under $\mathbb{P}^z$ since $z \in B_{1-\varepsilon}$. For any $x \in B$ we have $\mathbb{E}^x[\|W_0\|^2 - d \times 0] = \mathbb{E}^x[\|W_\tau\|^2 - d\tau]$ by the optional stopping theorem and thus

$$\|x\|^2 = \mathbb{E}^x[\|W_\tau\|^2 - d\tau] = 1 - d\,\mathbb{E}^x[\tau].$$

Similarly, optional stopping for the mean zero martingale $\|W_t - W_0\|^2 - dt$ under $\mathbb{P}^x$ yields

$$\mathbb{E}^x[\|W_\tau - W_0\|^2 - d\tau] = 0.$$

Combining the above gives $\mathbb{E}^x[\|W_\tau - W_0\|^2] = 1 - \|x\|^2$, which inserted into (22) finally yields

$$\mathbb{E}^z[\|W_\tau - W_{\tau_{1-\varepsilon}}\|] \le \mathbb{E}^z\left[1 - \|W_{\tau_{1-\varepsilon}}\|^2\right]^{1/2} = \left[1 - (1-\varepsilon)^2\right]^{1/2} = \sqrt{\varepsilon(2-\varepsilon)}$$

□

Lemma A.3 (L^2-bound on the score)**.** *Let ν be some probability measure on the unit sphere, $\overline{h}$ a function of the form*

$$\overline{h}(z) = \int_{\partial B} q(x \mid z)\,\nu(\mathrm{d}x), \quad z \in B,$$

$Z^{\overline{h}}$ be the corresponding $\overline{h}$-transform of some Brownian motion absorbed in ∂B, and let $\mathbb{F}$-stopping times $\underline{\tau} \le \overline{\tau} \le \tau^{\overline{h}}_{1-\varepsilon}$ be given. Then, for any $z_0 \in B_{1-\varepsilon}$, it holds

$$\mathbb{E}^{z_0}\left[\int_{\underline{\tau}}^{\overline{\tau}} \|\nabla \log \overline{h}(Z^{\overline{h}}_t)\|^2\,\mathrm{d}t\right] = 2\mathbb{E}^{z_0}[\log \overline{h}(Z^{\overline{h}}_{\overline{\tau}})] - 2\mathbb{E}^{z_0}[\log \overline{h}(Z^{\overline{h}}_{\underline{\tau}})].$$

In particular,

$$\mathbb{E}^{z_0}\left[\int_0^{\tau^{\overline{h}}_{1-\varepsilon}} \|\nabla \log \overline{h}(Z^{\overline{h}}_t)\|^2\,\mathrm{d}t\right] \lesssim d\log\varepsilon^{-1} + 1.$$

Proof. First, we note that

$$\|\nabla \log \overline{h}(z)\|^2 = \frac{\Delta \overline{h}(z)}{\overline{h}(z)} - \Delta \log \overline{h}(z), \quad z \in B.$$

Since $q(x \mid \cdot)$ is bounded and smooth on $\overline{B}_{1-\varepsilon}$ for all $x \in \partial B$, derivatives can be pulled inside the integral and thus $\overline{h}$ is smooth on $\overline{B}_{1-\varepsilon}$ as well. This yields

$$\Delta \overline{h} = \int_{\partial B} \Delta q(x \mid \cdot)\, \nu(\mathrm{d}x) = 0,$$

since the Poisson kernel is harmonic in B. Thus, we get

$$\mathbb{E}^{z_0}\left[\int_{\underline{\tau}}^{\overline{\tau}} \|\nabla \log \overline{h}(Z_t^{\overline{h}})\|^2 \,\mathrm{d}t\right] = -\mathbb{E}^{z_0}\left[\int_{\underline{\tau}}^{\overline{\tau}} \Delta \log h(Z_t^{\overline{h}}) \,\mathrm{d}t\right].$$

By the strong mean value property of harmonic functions, $\overline{h} > 0$ in $\overline{B}_{1-\varepsilon}$ (otherwise ν would need to be the zero measure), so $\log \overline{h}$ is bounded and smooth in $\overline{B}_{1-\varepsilon}$. Thus, we can use Itô's formula for the continuous semimartingale $(Z^{\overline{h}}_{t\wedge\tau^{\overline{h}}_{1-\varepsilon}})_{t\geq 0}$ to obtain

$$\begin{aligned}
&\mathbb{E}^{z_0}\left[\int_{\underline{\tau}}^{\overline{\tau}} \Delta \log \overline{h}(Z_t^{\overline{h}}) \,\mathrm{d}t\right] \\
&= 2\mathbb{E}^{z_0}\left[\log \overline{h}(Z_{\overline{\tau}}^{\overline{h}}) - \log \overline{h}(Z_{\underline{\tau}}^{\overline{h}}) - \int_{\underline{\tau}}^{\overline{\tau}} \nabla \log \overline{h}(Z_t^{\overline{h}}) \cdot \mathrm{d}Z_t^{\overline{h}}\right] \\
&= 2\Big(\mathbb{E}^{z_0}[\log \overline{h}(Z_{\overline{\tau}}^{\overline{h}})] - \mathbb{E}^{z_0}[\log \overline{h}(Z_{\underline{\tau}}^{\overline{h}})] - \mathbb{E}^{z_0}\Big[\int_{\underline{\tau}}^{\overline{\tau}} \|\nabla \log \overline{h}(Z_t^{\overline{h}})\|^2 \,\mathrm{d}t\Big] - \mathbb{E}^{z_0}\Big[\int_{\underline{\tau}}^{\overline{\tau}} \nabla \log \overline{h}(Z_t^{\overline{h}}) \cdot \mathrm{d}W_t\Big]\Big) \\
&= 2\Big(\mathbb{E}^{z_0}[\log \overline{h}(Z_{\overline{\tau}}^{\overline{h}})] - \mathbb{E}^{z_0}[\log \overline{h}(Z_{\underline{\tau}}^{\overline{h}})] - \mathbb{E}^{z_0}\Big[\int_{\underline{\tau}}^{\overline{\tau}} \|\nabla \log \overline{h}(Z_t^{\overline{h}})\|^2 \,\mathrm{d}t\Big]\Big)
\end{aligned}$$

where the last line follows from optional stopping[1] for the martingale $M = \int_0^{\cdot} \nabla \log \overline{h}(Z^{\overline{h}}_{s\wedge\tau^{\overline{h}}_{1-\varepsilon}}) \cdot \mathrm{d}W_s$ and the fact that on $\{t \leq \overline{\tau}\} \subset \{t \leq \tau^{\overline{h}}_{1-\varepsilon}\}$ it holds that $M_t = \int_0^t \nabla \log \overline{h}(Z_s^{\overline{h}}) \cdot \mathrm{d}W_s$. Finally, we can solve for the term of interest and get

$$\mathbb{E}\left[\int_{\underline{\tau}}^{\overline{\tau}} \|\nabla \log \overline{h}(Z_t^{\overline{h}})\|^2 \,\mathrm{d}t\right] = 2\mathbb{E}[\log \overline{h}(Z_{\overline{\tau}}^{\overline{h}})] - 2\mathbb{E}[\log \overline{h}(Z_{\underline{\tau}}^{\overline{h}})].$$

For the special case $\overline{\tau} = \tau^{\overline{h}}_{1-\varepsilon}$ and $\underline{\tau} = 0$, it follows

$$\begin{aligned}
\mathbb{E}^{z_0}\left[\int_0^{\tau^{\overline{h}}_{1-\varepsilon}} \|\nabla \log \overline{h}(Z_t^{\overline{h}})\|^2 \,\mathrm{d}t\right] &= 2\mathbb{E}^{z_0}[\log \overline{h}(Z^{\overline{h}}_{\tau^{\overline{h}}_{1-\varepsilon}})] - 2\mathbb{E}^{z_0}[\log \overline{h}(Z_0^{\overline{h}})] \\
&\lesssim 2d \log \varepsilon^{-1} + 2 \log \overline{h}(z_0),
\end{aligned}$$

where we used

$$\begin{aligned}
\overline{h}(z) = \int_{\partial B} q(x \mid z)\, \nu(\mathrm{d}x) = \frac{1 - \|z\|^2}{\sigma(\partial B)} \int_{\partial B} \frac{1}{\|x - z\|^d}\, \nu(\mathrm{d}x) &\lesssim \int_{\partial B} \frac{1}{(1 - \|z\|)^d} \nu(\mathrm{d}x) \\
&= (1 - \|z\|)^{-d} = \varepsilon^{-d}
\end{aligned}$$

for $z \in \partial B_{1-\varepsilon}$. □

Lemma A.4. *Let $\tau^x_{1-\varepsilon} = \tau_{1-\varepsilon}(Z^x)$ for $x \in \partial B$ and τ be the first hitting of the unit sphere by a Brownian motion. Let $z_0 \in B_{1-\varepsilon}$ and $\varepsilon < (1 - \|z_0\|/2)$. Then, for any $x \in \partial B$*

$$\mathbb{E}^{z_0}[\tau^x_{1-\varepsilon}] \leq C(\mathbb{E}^0[\tau] + \mathbb{E}^{z_0}[\tau]) < \infty,$$

for a constant C depending on z_0 but not on ε and x.

[1] optional stopping is applicable here since h and $\nabla \log h$ are bounded on $\overline{B}_{1-\varepsilon}$ and hence $\mathbb{E}^{z_0}[\langle M \rangle_{\overline{\tau}}] \lesssim \mathbb{E}^{z_0}[\overline{\tau}] \leq \mathbb{E}[\tau^{\overline{h}}_{1-\varepsilon}] < \infty$ by Lemma A.1.

Proof. Let $R = 5\|z_0\|/4$ be a fixed radius (though the proof does not rely on this particular choice) such that $z_0 \in B_R$ and $B_R \cap \partial B_{1-\varepsilon} = \varnothing$ by assumption. Since Z^x is a $q(x \mid \cdot)$-transform of the stopped Brownian motion, we have

$$\begin{aligned}
\mathbb{E}^{z_0}[\tau^x_{1-\varepsilon}] &= \frac{1}{q(x \mid z_0)} \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}} q(x \mid W_t)\,\mathrm{d}t\Big] \\
&= \frac{1}{q(x \mid z_0)} \int_{B_{1-\varepsilon}} G_{1-\varepsilon}(z_0, z) q(x \mid z)\,\mathrm{d}z \\
&\lesssim \frac{1}{q(x \mid z_0)} \int_{B_R} G_{1-\varepsilon}(z, z_0)\,\mathrm{d}z + \frac{1}{q(x \mid z_0)} \int_{B_{1-\varepsilon}\setminus B_R} G_{1-\varepsilon}(z, z_0) q(x \mid z)\,\mathrm{d}z \\
&\lesssim \frac{1}{q(x \mid z_0)} \int_{B_{1-\varepsilon}} G_{1-\varepsilon}(z, z_0)\,\mathrm{d}z + \frac{1}{q(x \mid z_0)} \int_{B_{1-\varepsilon}\setminus B_R} G_{1-\varepsilon}(z, z_0) q(x \mid z)\,\mathrm{d}z. \quad (23)
\end{aligned}$$

For the first term we obtain by definition of the Green kernel $G_{1-\varepsilon}$ of the Brownian motion killed in $\partial B_{1-\varepsilon}$,

$$\frac{1}{q(x \mid z_0)} \int_{B_{1-\varepsilon}} G_{1-\varepsilon}(z, z_0)\,\mathrm{d}z = \frac{1}{q(x \mid z_0)} \mathbb{E}^{z_0}[\tau_{1-\varepsilon}] = \frac{(1-\varepsilon)^2 - \|z_0\|^2}{d\, q(x \mid z_0)} \le \frac{(1+\|z_0\|)^d}{d(1-\|z_0\|^2)} < \infty.$$

The second term in (23) can be related to the expected hitting time of a Brownian motion started in the origin with a change of measure from $G_{1-\varepsilon}(\cdot, z_0)$ to $G_{1-\varepsilon}(\cdot, 0)$. This is possible, due to the 3G-Theorem [9, Theorem 6.5], which gives

$$\frac{G_1(x,y)G_1(y,z)}{G_1(x,z)} \lesssim \frac{\|x-z\|^{d-2}}{\|x-y\|^{d-2}\,\|y-z\|^{d-2}}.$$

Together with the fact that $G_1(z/(1-\varepsilon), z_0/(1-\varepsilon)) = (1-\varepsilon)^{d-2} G_{1-\varepsilon}(z, z_0)$, it implies

$$\frac{G_{1-\varepsilon}(z,z_0)}{G_{1-\varepsilon}(z,0)} = \frac{G_1(z/(1-\varepsilon), z_0/(1-\varepsilon))}{G_1(z/(1-\varepsilon), 0)} \lesssim \frac{1}{G_1(z_0/(1-\varepsilon), 0)} \frac{(1-\varepsilon)^{d-2}\|z\|^{d-2}}{\|z-z_0\|^{d-2}\,\|z_0\|^{d-2}}$$

and the right hand side is uniformly bounded for $\varepsilon < (1-\|z_0\|)/2$ and $z \in B_{1-\varepsilon} \setminus B_R$.

Thus,

$$\begin{aligned}
\frac{1}{q(x \mid z_0)} \int_{B_{1-\varepsilon}\setminus B_R} G_{1-\varepsilon}(z, z_0) q(x \mid z)\,\mathrm{d}z &\lesssim \frac{1}{q(x \mid z_0)} \int_{B_{1-\varepsilon}\setminus B_R} G_{1-\varepsilon}(z, 0) q(x \mid z)\,\mathrm{d}z \\
&\le \frac{(1+\|z_0\|)^d}{1-\|z_0\|^2} \mathbb{E}^0[\tau^x_{1-\varepsilon}].
\end{aligned}$$

Under $\mathbb{P}^0$, it holds that $W_{\tau_{1-\varepsilon}}$ and $\tau_{1-\varepsilon}$ are independent, see [25, Chapter 4, Theorem 39.6]. Thus,

$$\mathbb{E}^0[\tau^x_{1-\varepsilon}] = \frac{1}{q(x \mid 0)} \mathbb{E}^0[\tau_{1-\varepsilon} q(x \mid W_{\tau_{1-\varepsilon}})] = \mathbb{E}^0[\tau_{1-\varepsilon}]\,\mathbb{E}^0[q(x \mid W_{\tau_{1-\varepsilon}})] = \mathbb{E}^0[\tau_{1-\varepsilon}] \le \mathbb{E}^0[\tau] < \infty,$$

where we used that $(q(x \mid W_{t\wedge\tau}))_{t\ge 0}$ is a martingale, because $q(x \mid \cdot)$ is harmonic for the stopped Brownian motion, and thus by optional stopping and dominated convergence

$$\mathbb{E}^0[q(x \mid W_{\tau_{1-\varepsilon}})] = \lim_{n\to\infty} \mathbb{E}^0[q(x \mid W_{n\wedge\tau_{1-\varepsilon}})] = \lim_{n\to\infty} \mathbb{E}^0[q(x \mid W_0)] = q(x \mid 0) = 1.$$

Combining the previous bounds yields the claim. □

Lemma A.5. *It holds*

$$\mathbb{E}^{z_0}\left[\int_0^{\tau_{1-\varepsilon}} \frac{1}{(1-\|W_t\|)^2}\,\mathrm{d}t\right] \lesssim \log \varepsilon^{-1} + 1.$$

Proof. First, we use the binomial formula $1 - r = (1-r^2)/(1+r) \ge (1-r^2)/2$ to get

$$\mathbb{E}^{z_0}\left[\int_0^{\tau_{1-\varepsilon}} \frac{1}{(1-\|W_t\|)^2}\,\mathrm{d}t\right] \le 4\mathbb{E}^{z_0}\left[\int_0^{\tau_{1-\varepsilon}} \frac{1}{(1-\|W_t\|^2)^2}\,\mathrm{d}t\right],$$

which renders the integrand differentiable on the whole domain. Then, we define $u \in C^2(B_{1-\varepsilon})$ as the solution of the Dirichlet boundary problem

$$\begin{cases} \frac{1}{2}\Delta u(z) = \frac{1}{(1-\|z\|^2)^2}, & z \in B_{1-\varepsilon}, \\ u(z) = 0, & z \in \partial B_{1-\varepsilon}. \end{cases}$$

Then, by Itô's formula, it holds

$$u(z_0) = \mathbb{E}^{z_0}\left[\int_0^{\tau_{1-\varepsilon}} \frac{1}{(1-\|W_t\|^2)^2}\,\mathrm{d}t\right].$$

Now, it is easy to see that u is isotropic, i.e., there is a function $f\colon [0,1-\varepsilon] \to \mathbb{R}$ such that $u(z) = f(\|z\|)$, which means that the above PDE reduces to the ODE

$$f''(r) + \frac{d-1}{r} f'(r) = r^{1-d}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^{d-1}f'(r)\right) = -\frac{2}{(1-r)^2}, \qquad r \in (0, 1-\varepsilon).$$

This equation has the explicit solution

$$\begin{aligned} f(\|z_0\|) &= \int_{\|z_0\|}^{1-\varepsilon} v^{1-d}\left(\int_0^v \frac{r^{d-1}}{(1-r)^2}\,\mathrm{d}r\right)\mathrm{d}v \\ &\le \int_{\|z_0\|}^{1-\varepsilon} v^{1-d} v^{d-1}\left(\int_0^v \frac{1}{(1-r)^2}\,\mathrm{d}r\right)\mathrm{d}v \\ &= \int_{\|z_0\|}^{1-\varepsilon} \Big(\frac{1}{1-v} - 1\Big)\,\mathrm{d}v \\ &= \log \varepsilon^{-1} + c. \end{aligned}$$

□

B Remaining proofs for Section 2

Proof of Lemma 2.1. The family $(q_x)_{x\in\partial B} \coloneqq (q(x \mid \cdot))_{x\in\partial B}$ is a family of harmonic functions for the absorbed Brownian motion Z and we may express h as

$$h(z) = \int_{\partial B} q_x(z)\,\nu(dx),$$

for the measure $\nu(\mathrm{d}x) \coloneqq \pi^*(x)\,\sigma(\mathrm{d}x)$. The formula

$$\mathbb{P}^z(Z^h_{\tau^h} \in \mathrm{d}x) = \frac{\pi^*(x)}{h(z)}\,\mathbb{Q}_{\partial B}(z, \mathrm{d}x) \tag{24}$$

is therefore a consequence of [8, Proposition 11.10] and the argument used in the proof of [8, Theorem 13.39]. Consequently,

$$\begin{aligned} \nabla \log h(z) = \int \nabla_z \log q(x \mid z)\frac{\pi^*(x) q(x \mid z)}{h(z)}\,\sigma(\mathrm{d}x) &= \int \nabla_z \log q(x \mid z)\,\mathbb{P}^z(Z^h_{\tau^h} \in \mathrm{d}x) \\ &= \mathbb{E}^z\big[\nabla_2 \log q(Z^h_{\tau^h} \mid Z^h_0)\big], \end{aligned}$$

where we used (24) for the second equality. □

Proof of Proposition 2.2. To ease notation, define $b \coloneqq s - \nabla \log h$. First, we apply Girsanov's theorem [18, Theorem 5.22] to obtain for any $T > 0$ and $A \in \mathcal{G}_T$,

$$\mathbb{P}^s_\varepsilon(A) = \int_A \mathcal{E}\left(\int_0^{T\wedge\tau_{1-\varepsilon}} b(\xi_t)\,\mathrm{d}w_t\right)\mathrm{d}\mathbb{Q}^h_\varepsilon.$$

where w is a Brownian motion under $\mathbb{Q}_\varepsilon^h$ such that $\mathrm{d}\xi_t = s(\xi_t)\,\mathrm{d}t + \mathrm{d}w_t$ on $[\![0, \tau_{1-\varepsilon}]\!]$, $\mathbb{Q}_\varepsilon^h$-a.s. Let us note here that the stochastic integral is well-defined, since b is bounded on $\overline{B}_{1-\varepsilon}$ and $T \wedge \tau_{1-\varepsilon} < \infty$ a.s. Girsanov's theorem can be applied since the following Novikov condition is fulfilled:

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}_\varepsilon^h}\left[\exp\left(\frac{1}{2}\int_0^{T\wedge\tau_{1-\varepsilon}} \|b(\xi_t)\|^2\,\mathrm{d}t\right)\right] &\le \mathbb{E}_{\mathbb{Q}_\varepsilon^h}\left[\exp\left(\frac{1}{2}\sup_{z\in\overline{B}_{1-\varepsilon}}\|b(z)\|^2\,T\wedge\tau_{1-\varepsilon}\right)\right]\\ &\le \mathrm{e}^{\sup_{z\in\overline{B}_{1-\varepsilon}}\|b(z)\|^2 T} < \infty.\end{aligned}$$

The process $(\int_0^{T\wedge\tau_{1-\varepsilon}} b(\xi_t)\,\mathrm{d}w_t)_{T\ge 0}$ is a uniformly integrable martingale wrt $(\mathcal{G}_t)_{t\ge 0}$ thanks to Lemma A.1 (note here that the drift b of Z^h is Lipschitz on the compact set $\overline{B_{1-\varepsilon}}$ since s is locally Lipschitz, $(x,z)\mapsto q(x\mid z)$ is smooth on $\overline{B_{1-\varepsilon}}$ and $h \ge \pi_{\min} > 0$), which implies

$$\mathbb{E}_{\mathbb{Q}_\varepsilon^h}\left[\left\langle \int_0^{\cdot} b(\xi_t)\,\mathrm{d}w_t\right\rangle_{\tau_{1-\varepsilon}}\right] = \mathbb{E}_{\mathbb{Q}_\varepsilon^h}\left[\int_0^{\tau_{1-\varepsilon}}\|b(\xi_t)\|^2\,\mathrm{d}t\right] \lesssim \mathbb{E}_{\mathbb{Q}_\varepsilon^h}[\tau_{1-\varepsilon}] = \mathbb{E}^{z_0}[\tau_{1-\varepsilon}^h] < \infty. \quad (25)$$

Since $\tau_{1-\varepsilon} < \infty$, $\mathbb{Q}_\varepsilon^h$-a.s., for $\mathbb{Q}_\varepsilon^h$-a.e. ω it holds $\tau_{1-\varepsilon}(\omega)\wedge T = \tau_{1-\varepsilon}(\omega)$ for T large enough and we get

$$\lim_{T\to\infty}\mathcal{E}\left(\int_0^{T\wedge\tau_{1-\varepsilon}} b(\xi_t)\,\mathrm{d}w_t\right) = \mathcal{E}\left(\int_0^{\tau_{1-\varepsilon}} b(\xi_t)\,\mathrm{d}w_t\right) \quad \mathbb{Q}_\varepsilon^h\text{-a.s..}$$

By Fatou's lemma, it follows for $A\in\mathcal{G}_t$, $t>0$,

$$\int_A \mathcal{E}\left(\int_0^{\tau_{1-\varepsilon}} b(\xi_t)\,\mathrm{d}w_t\right)\mathrm{d}\mathbb{Q}_\varepsilon^h \le \liminf_{T\to\infty}\int_A \mathcal{E}\left(\int_0^{T\wedge\tau_{1-\varepsilon}} b(\xi_t)\,\mathrm{d}w_t\right)\mathrm{d}\mathbb{Q}_\varepsilon^h = \mathbb{P}_\varepsilon^s(A). \quad (26)$$

We extend (26) to all $A\in\mathcal{G}_\infty$: define

$$\mu(A) \coloneqq \mathbb{P}_\varepsilon^s(A) - \int_A \mathcal{E}\left(\int_0^{\tau_{1-\varepsilon}} b(\xi_t)\,\mathrm{d}w_t\right)\mathrm{d}\mathbb{Q}_\varepsilon^h \quad \text{for all } A\in\bigcup_{t\ge 0}\mathcal{G}_t.$$

Due to (26), μ is a pre-measure on $\bigcup_{t\ge0}\mathcal{G}_t$, which can be extended to a measure on $\mathcal{G}_\infty = \sigma(\cup_{t\ge0}\mathcal{G}_t)$ using Carathéodory's theorem. Then, we have

$$\int_\cdot \mathcal{E}\left(\int_0^{\tau_{1-\varepsilon}} b(\xi_t)\,\mathrm{d}w_t\right)\mathrm{d}\mathbb{Q}_\varepsilon^h + \mu = \mathbb{P}_\varepsilon^s$$

on $\bigcup_{t\ge0}\mathcal{G}_t$ by construction, and therefore on $\mathcal{G}_\infty$. Since μ is non-negative, we get (26) for all $A\in\mathcal{G}_\infty$.

Consequently, using also that $\mathcal{E}(\int_0^{\tau_{1-\varepsilon}} b(\xi_t)\,\mathrm{d}w_t) > 0$, $\mathbb{Q}_\varepsilon^h$-a.s., it follows that $\mathbb{P}_\varepsilon^s(A) = 0$ implies $\mathbb{Q}_\varepsilon^h(A) = 0$ for all $A\in\mathcal{G}_\infty$, and we get $\mathbb{Q}_\varepsilon^h \ll \mathbb{P}_\varepsilon^s$ on $\mathcal{G}_\infty$. With symmetric arguments, the above can be repeated with $\mathbb{Q}_\varepsilon^h$ and $\mathbb{P}_\varepsilon^s$ interchanged, yielding $\mathbb{P}_\varepsilon^s \ll \mathbb{Q}_\varepsilon^h$ as well. Therefore, according to [18, Proposition 5.20], the process of conditional Radon–Nikodym densities is uniformly integrable and

$$\frac{\mathrm{d}\mathbb{P}_\varepsilon^s}{\mathrm{d}\mathbb{Q}_\varepsilon^h} = \lim_{T\to\infty}\mathbb{E}_{\mathbb{Q}_\varepsilon^h}\left[\frac{\mathrm{d}\mathbb{P}_\varepsilon^s}{\mathrm{d}\mathbb{Q}_\varepsilon^h}\middle|\mathcal{G}_T\right] = \lim_{T\to\infty}\mathcal{E}\left(\int_0^{T\wedge\tau_{1-\varepsilon}} b(\xi_t)\,\mathrm{d}w_t\right) = \mathcal{E}\left(\int_0^{\tau_{1-\varepsilon}} b(\xi_t)\,\mathrm{d}w_t\right).$$

Thus, we conclude

$$\begin{aligned}\mathrm{KL}(\mathbb{Q}_\varepsilon^h \,\|\, \mathbb{P}_\varepsilon^s) &= \mathbb{E}_{\mathbb{Q}_\varepsilon^h}\left[\log\left(\frac{\mathrm{d}\mathbb{Q}_\varepsilon^h}{\mathrm{d}\mathbb{P}_\varepsilon^s}\right)\right]\\ &= \underbrace{-\,\mathbb{E}^{z_0}\left[\int_0^{\tau_{1-\varepsilon}^h} b(Z_t^h)\,\mathrm{d}W_t\right]}_{=0 \text{ (u.i. martingale due to (25))}} + \frac{1}{2}\mathbb{E}^{z_0}\left[\int_0^{\tau_{1-\varepsilon}^h}||b(Z_t^h)||^2\,\mathrm{d}t\right]\\ &= \frac{1}{2}\mathbb{E}^{z_0}\left[\int_0^{\tau_{1-\varepsilon}^h}\|b(Z_t^h)\|^2\,\mathrm{d}t\right].\end{aligned}$$

□

Proof of Proposition 2.3. Using the score representation from Lemma 2.1, we obtain

$$\begin{aligned}\mathbb{E}^z\Big[\int_{\underline{\tau}}^{\overline{\tau}} \langle \nabla \log h(Z_t^h), s(Z_t^h)\rangle \,\mathrm{d}t\Big] &= \mathbb{E}^z\Big[\int_{\underline{\tau}}^{\overline{\tau}} \big\langle \mathbb{E}^{Z_t^h}\big[\nabla_2 \log q(Z_{\tau^h}^h \mid Z_0^h)\big], s(Z_t^h)\big\rangle \,\mathrm{d}t\Big] \\ &= \mathbb{E}^z\Big[\int_0^{\infty} \mathbb{E}\big[\mathbf{1}_{\{\overline{\tau}>t\geq\underline{\tau}\}}\big\langle \nabla_2 \log q(Z_{\tau^h}^h \mid Z_t^h), s(Z_t^h)\big\rangle \,\big|\, \mathcal{F}_t\big] \,\mathrm{d}t\Big] \\ &= \mathbb{E}^z\Big[\int_{\underline{\tau}}^{\overline{\tau}} \big\langle \nabla_2 \log q(Z_{\tau^h}^h \mid Z_t^h), s(Z_t^h)\big\rangle \,\mathrm{d}t\Big].\end{aligned}$$

For the second line we used the Markov property and the facts that $\{\overline{\tau} > t \geq \underline{\tau}\} \in \mathcal{F}_t$ and that since τ^h is a terminal time we have $Z_{\tau^h}^h \circ \theta_t = Z_{t+\tau^h\circ\theta_t}^h = Z_{\tau^h}^h$ on $\{\overline{\tau} > t\} \subset \{\tau^h > t\}$ for the shift operators $(\theta_t)_{t\geq 0}$ of Z^h. The last line is then a consequence of the tower property of the conditional expectation and Fubini. This shows that

$$\begin{aligned}&\mathbb{E}^z\Big[\int_{\underline{\tau}}^{\overline{\tau}} \|\nabla \log h(Z_t^h) - s(Z_t^h)\|^2 \,\mathrm{d}t\Big] \\ &= \mathbb{E}^z\Big[\int_{\underline{\tau}}^{\overline{\tau}} \|s(Z_t^h)\|^2 \,\mathrm{d}t\Big] - 2\mathbb{E}^z\Big[\int_{\underline{\tau}}^{\overline{\tau}} \langle \nabla \log h(Z_t^h), s(Z_t^h)\rangle \,\mathrm{d}t\Big] + C \\ &= \mathbb{E}^z\Big[\int_{\underline{\tau}}^{\overline{\tau}} \|s(Z_t^h)\|^2 \,\mathrm{d}t\Big] - 2\mathbb{E}^z\Big[\int_{\underline{\tau}}^{\overline{\tau}} \big\langle \nabla_2 \log q(Z_{\tau^h}^h \mid Z_t^h), s(Z_t^h)\big\rangle \,\mathrm{d}t\Big] + C \\ &= \mathbb{E}^z\Big[\int_{\underline{\tau}}^{\overline{\tau}} \|\nabla_2 \log q(Z_{\tau^h}^h \mid Z_t^h) - s(Z_t^h)\|^2 \,\mathrm{d}t\Big] + C',\end{aligned}$$

where the constants C, C' are independent of s and C' is given by

$$\begin{aligned}C' &\coloneqq \mathbb{E}^z\Big[\int_{\underline{\tau}}^{\overline{\tau}} \|\nabla \log h(Z_t^h)\|^2 \,\mathrm{d}t\Big] - \mathbb{E}^z\Big[\int_{\underline{\tau}}^{\overline{\tau}} \|\nabla_2 \log q(Z_{\tau^h}^h \mid Z_t^h)\|^2 \,\mathrm{d}t\Big] \\ &= -\mathbb{E}^z\Big[\int_{\underline{\tau}}^{\overline{\tau}} \|\nabla_2 \log q(Z_{\tau^h}^h \mid Z_t^h) - \nabla \log h(Z_t^h)\|^2 \,\mathrm{d}t\Big],\end{aligned}$$

where the second line follows from plugging in $s = \nabla \log h$ in (11). Finally, we prove (12). By the Markov property of W it $\mathbb{P}^{z_0}$-a.s. holds

$$\begin{aligned}&\mathbb{E}^{z_0}\Big[\|s(W_t) - \nabla \log q(W_\tau \mid W_t)\|^2 \pi^*(W_\tau) \mid \mathcal{F}_t\Big]\mathbf{1}_{\{t<\tau\}} \\ &= \mathbb{E}^{W_t}\Big[\|s(W_0) - \nabla \log q(W_\tau \mid W_0)\|^2 \pi^*(W_\tau)\Big]\mathbf{1}_{\{t<\tau\}} \\ &= \mathbb{E}^{z}\Big[\|s(z) - \nabla \log q(W_\tau \mid z)\|^2 \pi^*(W_\tau)\Big]\mathbf{1}_{\{t<\tau\}}\Big|_{z=W_t} \\ &= \int_{\partial B} \|s(z) - \nabla \log q(x \mid z)\|^2 \pi^*(x) q(x \mid z)\,\sigma(\mathrm{d}x)\,\mathbf{1}_{\{t<\tau\}}\Big|_{z=W_t} \\ &= \int_{\partial B} \|s(W_t) - \nabla \log q(x \mid W_t)\|^2 \pi^*(x) q(x \mid W_t)\,\sigma(\mathrm{d}x)\,\mathbf{1}_{\{t<\tau\}}.\end{aligned}$$

Using this together with (5), the Markov property and Fubini's theorem it follows that

$$\begin{aligned}&\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^h} \|s(Z_t^h) - \nabla_2 \log q(Z_{\tau^h}^h \mid Z_t^h)\|^2 \,\mathrm{d}t\Big] \\ &= \mathbb{E}^{z_0}\Big[\int_0^{\infty} \mathbb{E}^{z_0}\Big[\|s(W_t) - \nabla \log q(W_\tau \mid W_t)\|^2 \pi^*(W_\tau) \mid \mathcal{F}_t\Big]\mathbf{1}_{\{t<\tau_{1-\varepsilon}\}} \,\mathrm{d}t\Big] \\ &= \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}} \int_{\partial B} \|s(W_t) - \nabla \log q(x \mid W_t)\|^2 \pi^*(x) q(x \mid W_t)\,\sigma(\mathrm{d}x)\,\mathrm{d}t\Big] \\ &= \int_{\partial B} \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}} \|s(W_t) - \nabla \log q(x \mid W_t)\|^2 q(x \mid W_t)\,\mathrm{d}t\Big]\,\pi^*(x)\,\sigma(\mathrm{d}x) \\ &= \int_{\partial B} \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}} \|s(W_t) - \nabla \log q(x \mid W_t)\|^2 \,\mathrm{d}t\, q(x \mid W_{\tau_{1-\varepsilon}})\Big]\,\pi^*(x)\,\sigma(\mathrm{d}x),\end{aligned}$$

where for the last line we used that $q(x \mid \cdot)$ is bounded on $B_{1-\varepsilon}$ and harmonic for the killed Brownian motion, which implies that $(q(x \mid W_{t\wedge\tau_{1-\varepsilon}}))_{t\geq 0}$ is a bounded martingale and hence $q(x \mid W_t)\mathbf{1}_{\{t<\tau_{1-\varepsilon}\}} = \mathbb{E}^{z_0}[q(x \mid W_{\tau_{1-\varepsilon}}) \mid \mathcal{F}_t]\mathbf{1}_{\{t<\tau_{1-\varepsilon}\}}$, $\mathbb{P}^{z_0}$-a.s. by the Markov property and the optional stopping theorem. By [8, Theorem 11.9] and the fact that $\tau^x_{1-\varepsilon} < \tau^x$ a.s., the expectation on the rhs can be expressed in terms of the $q(x \mid \cdot)$-transform Z^x, yielding

$$\begin{aligned}
&\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}} \|s(W_t) - \nabla \log q(x \mid W_t)\|^2 \,\mathrm{d}t\, q(x \mid W_{\tau_{1-\varepsilon}})\Big] \\
&\quad = \mathbb{E}^{z_0}\Big[\int_0^{\tau^x_{1-\varepsilon}} \|s(Z^x_t) - \nabla \log q(x \mid Z^x_t)\|^2 \,\mathrm{d}t\Big] q(x \mid z_0),
\end{aligned}$$

and thus by the above,

$$\begin{aligned}
&\mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|s(Z^h_t) - \nabla_2 \log q(Z^h_{\tau^h} \mid Z^h_t)\|^2 \,\mathrm{d}t\Big] \\
&= \int_{\partial B} \mathbb{E}^{z_0}\Big[\int_0^{\tau^x_{1-\varepsilon}} \|s(Z^x_t) - \nabla \log q(x \mid Z^x_t)\|^2 q(x \mid Z^x_t)\,\mathrm{d}t\Big]\, \pi^*(x) q(x \mid z_0)\, \sigma(\mathrm{d}x) \\
&= \int_{\partial B} \mathbb{E}^{z_0}\Big[\int_0^{\tau^x_{1-\varepsilon}} \|s(Z^x_t) - \nabla \log q(x \mid Z^x_t)\|^2 q(x \mid Z^x_t)\,\mathrm{d}t\Big]\, \Pi^*(\mathrm{d}x)
\end{aligned}$$

as claimed. □

Proof of Lemma 2.4. According to Lemma 2.1, we have $\nabla \log h(z) = \mathbb{E}[\nabla_z \log q(Z^h_{\tau^h} \mid z) \mid Z^h_0 = z]$ and for $x \in \partial B$ it holds

$$\nabla_z \log q(x \mid z) = -\frac{2z}{1-\|z\|^2} - d\frac{z-x}{\|z-x\|^2}.$$

Combining both results, we get

$$\begin{aligned}
\|\nabla \log h(z)\| &= \big\|\mathbb{E}[\nabla_z \log q(Z^h_{\tau^h} \mid z) \mid Z^h_0 = z]\big\| \\
&= \Big\| -\frac{2z}{1-\|z\|^2} - d\mathbb{E}\Big[\frac{Z^h_{\tau^h} - z}{\|Z^h_{\tau^h} - z\|^2} \,\Big|\, Z^h_0 = z\Big]\Big\| \\
&\leq 2\frac{\|z\|}{1-\|z\|^2} + d\mathbb{E}\Big[\frac{1}{\|Z^h_{\tau^h} - z\|} \,\Big|\, Z^h_0 = z\Big] \\
&\leq \frac{2}{(1-\|z\|)(1+\|z\|)} + \frac{d}{1-\|z\|} \\
&\leq \frac{d+2}{1-\|z\|},
\end{aligned}$$

where we used the reverse triangle inequality in the fourth line. □

C Proof of the minimax lower bound

Proof of Proposition 3.2. The sphere can be split into an upper and a lower half $B_\pm$, which can then be separately parameterised in stereographic coordinates via bijective mappings $\varphi_{\partial B,\pm}\colon B_\pm \to B_1^{(d-1)}$, cf. Section F for details. Let $\Pi^*(\mathrm{d}x) = \pi(x)\,\sigma(\mathrm{d}x)$ and $\widehat{\mu}(\mathrm{d}x) = \widehat{\rho}(x)\,\sigma(\mathrm{d}x)$ and consider a density π concentrated on B_+. Then,

$$\begin{aligned}
\|\Pi^* - \widehat{\mu}\|_{\mathrm{TV}} &= \|(\Pi^* - \widehat{\mu})|_{B_+}\|_{\mathrm{TV}} + \|(\Pi^* - \widehat{\mu})|_{B_-}\|_{\mathrm{TV}} \\
&= \|(\Pi^* - \widehat{\mu}|_{B_+}) \circ \varphi^{-1}_{\partial B,+}\|_{\mathrm{TV}} + \|\widehat{\mu}|_{B_-} \circ \varphi^{-1}_{\partial B,-}\|_{\mathrm{TV}},
\end{aligned}$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation norm on ∂B in the first and on $B_1^{(d-1)}$ in the second line, and therefore

$$\begin{aligned}\|\Pi^* - \widehat{\mu}\|_{\mathrm{TV}} &\geq \|(\Pi^* - \widehat{\mu}|_{B_+}) \circ \varphi_{\partial B,+}^{-1}\|_{\mathrm{TV}} \\ &= \int_{B_1^{(d-1)}} \left|\pi \circ \varphi_{\partial B,+}^{-1} |J_{\varphi_{\partial B,+}^{-1}}|(x) - \widehat{\rho} \circ \varphi_{\partial B,+}^{-1} |J_{\varphi_{\partial B,+}^{-1}}|(x)\right| \mathrm{d}x \\ &=: \int_{B_1^{(d-1)}} |\overline{\pi}(x) - \widetilde{\rho}(x)| \, \mathrm{d}x,\end{aligned}$$

where $|J_{\varphi_{\partial B,+}^{-1}}|$ denotes the determinant of the Jacobian of $\varphi_{\partial B,+}^{-1}$ and $\overline{\pi} = \pi \circ \varphi_{\partial B,+}^{-1} |J_{\varphi_{\partial B,+}^{-1}}|$ is a probability density on $B_1^{(d-1)}$ such that $\overline{\pi} \in \mathcal{B}(H^\alpha(B_1^{(d-1)}), C)$ for some $C > 0$, which follows from Lemma H.1 and the fact that $|J_{\varphi_{\partial B,+}^{-1}}|$ is uniformly bounded on $B_1^{(d-1)}$. Note also that above $\widetilde{\rho}$ can be considered as a measurable function wrt the data $\{Y_i\}_{i=1}^n = \{\varphi_{\partial B,+}(X_i)\}_{i=1}^n \overset{\text{iid}}{\sim} \overline{\pi}$ since $\{X_i\}_{i=1}^n$ may be interpreted as iid B_+-valued random variables under $\mathbb{P}_\pi$ for π concentrated on B_+ as above. Thus,

$$\begin{aligned}\inf_{\widehat{\rho}} \sup_{\pi \in \mathcal{B}(H^\alpha(\partial B), L)} \mathbb{E}_\pi[\mathrm{TV}(\pi, \widehat{\rho})] &\geq \inf_{\widehat{\rho}} \sup_{\pi \in \mathcal{B}(H^\alpha(\partial B), L), \mathrm{supp}\pi \subset B_+} \mathbb{E}_\pi[\mathrm{TV}(\pi, \widehat{\rho})] \\ &\geq \inf_{\widetilde{\rho}} \sup_{\overline{\pi} \in \mathcal{B}(H^\alpha(B_1^{(d-1)}), C)} \mathbb{E}_{\overline{\pi}}\big[\|\overline{\pi} - \widetilde{\rho}\|_{L^1(B_1^{(d-1)})}\big].\end{aligned}$$

where the last infimum is taken over all random functions (not necessarily Lebesgue probability densities) $\widetilde{\rho}$ on $B_1^{(d-1)}$ that are measurable with respect to data $\{Y_1, \ldots, Y_n\}$ such that $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \overline{\pi}$ under $\mathbb{P}_{\overline{\pi}}$. By [35, Theorem 4], whose assumptions can be verified as in the proof of [24, Proposition D.4] with $p = q = 2$ and $s \in \mathbb{N}$ due to $B_{2,2}^s(D) = H^s(D)$ for Lipschitz domains D, the latter minimax risk is lower bounded by

$$\inf_{\widetilde{\rho}} \sup_{\overline{\pi} \in \mathcal{B}(H^\alpha(B_1^{(d-1)}), C)} \mathbb{E}_{\overline{\pi}}\big[\|\overline{\pi} - \widetilde{\rho}\|_{L^1(B_1^{(d-1)})}\big] \gtrsim n^{-\frac{\alpha}{2\alpha + d - 1}},$$

which finishes the proof. □

D Proof of the early stopping bound

Proof of Proposition 4.1. The proof uses Scheffé's lemma to relate the total variation distance to the L^1-distance of the densities wrt some reference measure, which we choose as the distribution $\mathbb{P}^{z_0}(W_\tau \in \cdot)$. While the density of Π^* is π^* by definition, the density of $P_{\partial B} \sharp \mathbb{Q}^h_{\partial B_{1-\varepsilon}}$ is not directly obvious. For any $x \in B \setminus \{z_0\}$ let $q(x \mid z_0)$ be defined as in the first equality of (2). For $A \in \mathcal{B}$ it holds

$$\begin{aligned}P_{\partial B} \sharp \mathbb{Q}^h_{\partial B_{1-\varepsilon}}(A) &= \mathbb{P}^{z_0}\Big(\tfrac{Z^h_{\tau^h_{1-\varepsilon}}}{1-\varepsilon} \in A\Big) \\ &= \mathbb{E}^{z_0}\left[\mathbf{1}_A\left(\frac{W_{\tau_{1-\varepsilon}}}{1-\varepsilon}\right) h(W_{\tau_{1-\varepsilon}})\right] \\ &= \frac{1}{\sigma(\partial B_{1-\varepsilon})} \int_{\partial B_{1-\varepsilon}} \mathbf{1}_A\left(\frac{y}{1-\varepsilon}\right) h(y) q(y \mid z_0) \, \sigma(\mathrm{d}y) \\ &= \frac{1}{\sigma(\partial B)(1-\varepsilon)^{d-1}} \int_{\partial B} \mathbf{1}_A(x) h((1-\varepsilon)x) q((1-\varepsilon)x \mid z_0)(1-\varepsilon)^{d-1} \, \sigma(\mathrm{d}x) \\ &= \int_{\partial B} \mathbf{1}_A(x) h((1-\varepsilon)x) \frac{q((1-\varepsilon)x \mid z_0)}{q(x \mid z_0)} \, \mathbb{P}^{z_0}(W_\tau \in \mathrm{d}x),\end{aligned}$$

which shows that for $x \in \partial B$

$$\frac{\mathrm{d}P_{\partial B} \sharp \mathbb{Q}^h_{\partial B_{1-\varepsilon}}}{\mathrm{d}\mathbb{P}^{z_0}(W_\tau \in \cdot)}(x) = h((1-\varepsilon)x) \frac{q((1-\varepsilon)x \mid z_0)}{q(x \mid z_0)}.$$

Thus, using $h(y) = \mathbb{E}^y[\pi^*(W_\tau)]$ and the fact that $(W_{\tau_{1-\varepsilon}}, W_{\tau_{1-\varepsilon}}/(1-\varepsilon)) \overset{d}{=} (W_\tau(1-\varepsilon), W_\tau)$ unter $\mathbb{P}^0$, we obtain

$$
\begin{aligned}
&\mathrm{TV}(\Pi^*, P_{\partial B}\sharp\mathbb{Q}^h_{\partial B_{1-\varepsilon}}) \\
&\lesssim \mathbb{E}^{z_0}\left[\left|\pi^*(W_\tau) - h((1-\varepsilon)W_\tau)\frac{q((1-\varepsilon)W_\tau \mid z_0)}{q(W_\tau \mid z_0)}\right|\right] \\
&= \mathbb{E}^0\left[\left|\pi^*(W_\tau) - h((1-\varepsilon)W_\tau)\frac{q((1-\varepsilon)W_\tau \mid z_0)}{q(W_\tau \mid z_0)}\right| q(W_\tau \mid z_0)\right]\sigma(\partial B) \\
&\asymp \mathbb{E}^0\left[\left|\pi^*\left(\frac{W_{\tau_{1-\varepsilon}}}{1-\varepsilon}\right) q\left(\frac{W_{\tau_{1-\varepsilon}}}{1-\varepsilon}\,\middle|\, z_0\right) - h(W_{\tau_{1-\varepsilon}})q(W_{\tau_{1-\varepsilon}} \mid z_0)\right|\right] \\
&= \mathbb{E}^0\left[\left|\pi^*\left(\frac{W_{\tau_{1-\varepsilon}}}{1-\varepsilon}\right) q\left(\frac{W_{\tau_{1-\varepsilon}}}{1-\varepsilon}\,\middle|\, z_0\right) - \mathbb{E}^{W_{\tau_{1-\varepsilon}}}\left[\pi^*(W_\tau)\right]q(W_{\tau_{1-\varepsilon}} \mid z_0)\right|\right] \\
&= \mathbb{E}^0\left[\left|\mathbb{E}^0\left[\pi^*\left(\frac{W_{\tau_{1-\varepsilon}}}{1-\varepsilon}\right) q\left(\frac{W_{\tau_{1-\varepsilon}}}{1-\varepsilon}\,\middle|\, z_0\right) - \pi^*(W_\tau)q(W_{\tau_{1-\varepsilon}} \mid z_0)\,\middle|\, \mathcal{F}_{\tau_{1-\varepsilon}}\right]\right|\right] \\
&\leq \mathbb{E}^0\left[\left|\pi^*\left(\frac{W_{\tau_{1-\varepsilon}}}{1-\varepsilon}\right) q\left(\frac{W_{\tau_{1-\varepsilon}}}{1-\varepsilon}\,\middle|\, z_0\right) - \pi^*(W_\tau)q(W_{\tau_{1-\varepsilon}} \mid z_0)\right|\right] \\
&\leq \mathbb{E}^0\left[q(W_{\tau_{1-\varepsilon}} \mid z_0)\left|\pi^*\left(\frac{W_{\tau_{1-\varepsilon}}}{1-\varepsilon}\right) - \pi^*(W_\tau)\right|\right] \\
&\quad + \mathbb{E}^0\left[\pi^*\left(\frac{W_{\tau_{1-\varepsilon}}}{1-\varepsilon}\right)\left|q\left(\frac{W_{\tau_{1-\varepsilon}}}{1-\varepsilon}\,\middle|\, z_0\right) - q(W_{\tau_{1-\varepsilon}} \mid z_0)\right|\right].
\end{aligned}
\tag{27}
$$

To bound the first term in (27), we use that for any $x \in \partial B_{1-\varepsilon}$

$$
\begin{aligned}
q(x \mid z_0) = \frac{1}{\sigma(\partial B)}\frac{(1-\varepsilon)^2 - \|z_0\|^2}{(1-\varepsilon)\|z_0 - x\|^d} &\leq \frac{2}{\sigma(\partial B)(1+\|z_0\|)}\frac{1-\|z_0\|^2}{((1-\varepsilon)-\|z_0\|)^d} \\
&\leq \frac{2^{d+1}}{\sigma(\partial B)(1-\|z_0\|)^{d-1}}
\end{aligned}
\tag{28}
$$

since by assumption $1 - \varepsilon > (1 + \|z_0\|)/2$. Furthermore, according to the Sobolev embedding theorem on manifolds [2, Theorem 2.20], π^* is β-Hölder smooth in the sense

$$
\|\pi^*(x) - \pi^*(y)\| \lesssim d_{\partial B}(x,y)^\beta, \qquad x, y \in \partial B,
$$

where $d_{\partial B}$ is the distance measure on the sphere

$$
d_{\partial B}(x,y) \coloneqq \inf\left\{\int_0^1 \|\tfrac{\mathrm{d}}{\mathrm{d}t}\gamma(t)\|\,\mathrm{d}t \,\middle|\, \gamma\colon [0,1] \to \partial B \text{ differentiable}, \gamma(0) = x, \gamma(1) = y\right\}.
$$

On the sphere it thus holds $d_{\partial B}(x,y) \leq \pi\|x - y\|$ and therefore

$$
\begin{aligned}
\left\|\pi^*\left(\frac{W_{\tau_{1-\varepsilon}}}{1-\varepsilon}\right) - \pi^*(W_\tau)\right\| &\lesssim \left\|\frac{W_{\tau_{1-\varepsilon}}}{1-\varepsilon} - W_\tau\right\|^\beta \\
&\leq \frac{\|W_{\tau_{1-\varepsilon}} - W_\tau\|^\beta}{1-\varepsilon} + \|W_\tau\|^\beta\left(\frac{1}{1-\varepsilon} - 1\right)^\beta \\
&\lesssim \|W_{\tau_{1-\varepsilon}} - W_\tau\|^\beta + \varepsilon^\beta.
\end{aligned}
\tag{29}
$$

For the second term in (27), we simply note that based on (2), extending $q(\cdot \mid z_0)$ to the annulus $\Lambda_{z_0} \coloneqq \{x \in \mathbb{R} : \|x\| \in (1 - \|z_0\|/2, 1 + \|z_0\|/2)\}$ by

$$
q(x \mid z_0) = \frac{1}{\sigma(\partial B)}\frac{\|x\|^2 - \|z_0\|^2}{\|x\|\|z_0 - x\|^d}, \quad x \in \Lambda_{z_0},
$$

gives a smooth function on Λ_{z_0}, which is thus in particular Lipschitz and hence Hölder continuous for any $\gamma \in (0,1]$. Choosing $\gamma = \beta$ therefore yields

$$
\left|q(W_{\tau_{1-\varepsilon}}/(1-\varepsilon) \mid z_0) - q_{1-\varepsilon}(W_{\tau_{1-\varepsilon}} \mid z_0)\right| \lesssim \|W_{\tau_{1-\varepsilon}}/(1-\varepsilon) - W_{\tau_{1-\varepsilon}}\|^\beta = \varepsilon^\beta. \tag{30}
$$

Inserting (28)–(30) in (27) yields

$$\begin{aligned}\mathrm{TV}(\Pi^*, P_{\partial B}\sharp\mathbb{Q}^h_{\partial B_{1-\varepsilon}}) &\lesssim 4\sigma(B)^{-1}\left(\mathbb{E}^0[\|W_\tau - W_{\tau_{1-\varepsilon}}\|^\beta] + \varepsilon^\beta\right) + \|\pi^*\|_{L^\infty(\partial B)}\varepsilon^\beta \\ &\lesssim \mathbb{E}^0[\|W_\tau - W_{\tau_{1-\varepsilon}}\|]^\beta + \varepsilon^\beta \\ &\lesssim (\varepsilon(2-\varepsilon))^{\beta/2} + \varepsilon^\beta \\ &\lesssim \varepsilon^{\beta/2},\end{aligned}$$

where we used Jensen's inequality for the second and Lemma A.2 for the third line. □

E Proofs for Section 4.2

Proof of Theorem 4.2.(i). The only difference to the proof [24, Lemma C.4] lies in the first few lines due to the different definition of L_s. Recall that the driving Brownian motion W of the generative process Z^h is chosen independently of the data $X_1, \dots, X_n$, making Z^h and $\{X_1, \dots, X_n\}$ independent. By conditioning and using the first part of Proposition 2.3 we then obtain

$$\begin{aligned}&\mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|\widehat{s}(Z^h_t) - \nabla \log h(Z^h_t)\|^2 \,\mathrm{d}t\Big] \\ &\mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|\widehat{s}(Z^h_t) - \nabla \log h(Z^h_t)\|^2 \,\mathrm{d}t\Big] - \mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|\nabla \log h(Z^h_t) - \nabla \log h(Z^h_t)\|^2 \,\mathrm{d}t\Big] \\ &= \mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|\widehat{s}(Z^h_t) - \nabla \log q(Z^h_{\tau^h} \mid Z^h_t)\|^2 \,\mathrm{d}t\Big] + C \\ &\quad - \mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|\nabla \log h(Z^h_t) - \nabla \log q(Z^h_{\tau^h} \mid Z^h_t)\|^2 \,\mathrm{d}t\Big] - C \\ &= \mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|\widehat{s}(Z^h_t) - \nabla \log q(Z^h_{\tau^h} \mid Z^h_t)\|^2 \,\mathrm{d}t\Big] \\ &\qquad - \mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|\nabla \log h(Z^h_t) - \nabla \log q(Z^h_{\tau^h} \mid Z^h_t)\|^2 \,\mathrm{d}t\Big].\end{aligned}$$

We now use (12) from Proposition 2.3 to find

$$\begin{aligned}&\mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|\widehat{s}(Z^h_t) - \nabla \log q(Z^h_{\tau^h} \mid Z^h_t)\|^2 \,\mathrm{d}t\Big] - \mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}} \|\nabla \log h(Z^h_t) - \nabla \log q(Z^h_{\tau^h} \mid Z^h_t)\|^2 \,\mathrm{d}t\Big] \\ &= \int_{\partial B} \mathbb{E}^{z_0}\Big[\int_0^{\tau^x_{1-\varepsilon}} \|\widehat{s}(Z^x_t) - \nabla_2 \log q(x \mid Z^x_t)\|^2 - \int_0^{\tau^x_{1-\varepsilon}} \|\nabla \log h(Z^x_t) - \nabla_2 \log q(x \mid Z^x_t)\|^2 \,\mathrm{d}t\Big] \,\Pi^*(\mathrm{d}x) \\ &= \mathbb{E}^{z_0}\left[\frac{1}{n}\sum_{i=1}^n (L_{\widehat{s}}(X_i) - L_{\nabla \log h}(X_i))\right],\end{aligned}$$

where the last line uses Fubini's theorem and $X_i \sim \Pi^*$ for all $i = 1, \dots, n$. From this point on, the proof is completely analogous to [24, Lemma C.4], where we further note that essentially repeating the calculation from [14, Lemma B.1] based on Proposition 2.3 gives the Bernstein inequality $\mathbb{E}^{z_0}[(L_s(X_1) - L_{\nabla \log h}(X_1))^2] \le 4C(\mathcal{L})\mathbb{E}^{z_0}[L_s(X_1) - L_{\nabla \log h}(X_1)]$, which is needed to justify a critical step in [24, Lemma C.4]; see [14, Section B] for details. □

Proof of Theorem 4.2.(ii). Let $s_1, \dots, s_m \in \mathcal{S}$ be a $\widetilde{\delta}$-net for $\mathcal{S}$ wrt $\|\cdot\|_{L^\infty(B_{1-\varepsilon})}$. Then, for any $s \in \mathcal{S}$, choose $j \in \mathbb{N}$ s.t. $\|s - s_j\|_{L^\infty(B_{1-\varepsilon})} \le \widetilde{\delta}$. For better readability, we define for any $x \in \partial B$

$$b(z) \coloneqq s(z) - \nabla \log q(x \mid z), \qquad b_j(z) \coloneqq s_j(z) - \nabla \log q(x \mid z).$$

Then,

$$
\begin{aligned}
|L_s(x) - L_{s_j}(x)| &= \Big|\mathbb{E}^{z_0}[\int_0^{\tau_{1-\varepsilon}^x} (\|b(Z_t^x)\|^2 - \|b_j(Z_t^x)\|^2)\,\mathrm{d}t]\Big| \\
&\le \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x} \big|\|b(Z_t^x)\| - \|b_j(Z_t^x)\|\big| \left(\|b(Z_t^x)\| + \|b_j(Z_t^x)\|\right)\mathrm{d}t\Big] \\
&\le \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x} \|b(Z_t^x) - b_j(Z_t^x)\| \left(\|b(Z_t^x)\| + \|b_j(Z_t^x)\|\right)\mathrm{d}t\Big] \\
&= \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x} \underbrace{\|s(Z_t^x) - s_j(Z_t^x)\|}_{\le \widetilde{\delta}} \left(\|b(Z_t^x)\| + \|b_j(Z_t^x)\|\right)\mathrm{d}t\Big] \\
&\le 2\widetilde{\delta} \sup_{s\in\mathcal{S}} \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x} \|s(Z_t^x) - \nabla \log q(x \mid Z_t^x)\|\,\mathrm{d}t\Big] \\
&\le 2\widetilde{\delta}\sqrt{\mathbb{E}^{z_0}[\tau_{1-\varepsilon}^x]} \sup_{s\in\mathcal{S}} \Big(\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x} \|s(Z_t^x) - \nabla \log q(x \mid Z_t^x)\|^2\,\mathrm{d}t\Big]\Big)^{1/2} \\
&\lesssim 2\widetilde{\delta}\sqrt{C(\mathcal{L})},
\end{aligned}
$$

where we used the Cauchy–Schwarz inequality in the penultimate line and Lemma A.4 to uniformly bound $\mathbb{E}^{z_0}[\tau_{1-\varepsilon}^x]$. Taking the supremum over all x, we get

$$\|L_s - L_{s_j}\|_{L^\infty(\partial B)} \lesssim \widetilde{\delta}\sqrt{C(\mathcal{L})}.$$

By choosing $\widetilde{\delta} \lesssim \delta/\sqrt{C(\mathcal{L})}$, we see that the functions $L_{s_1}, \dots, L_{s_m}$ form a δ-net for $\mathcal{L}$ wrt $\|\cdot\|_{L^\infty(\partial B)}$. □

Proof of Theorem 4.2.(iii). First, we have the elementary bound

$$L_s(x) \le 2\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x} \|s(Z_t^x)\|^2\,\mathrm{d}t\Big] + 2\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x} \|\nabla \log q(x \mid Z_t^x)\|^2\,\mathrm{d}t\Big]. \tag{31}$$

Up to a multiplicative constant, the second term is bounded by $\log \varepsilon^{-1}$ according to Lemma A.3 with $\overline{h} = q(x \mid \cdot)$. For the first term, we note that

$$\|s(z)\| \lesssim (1 - \|z\|)^{-1} \le \frac{2\|z\|}{1 - \|z\|^2} + 1 = \|\nabla \log f(z)\| + 1, \tag{32}$$

with $f(z) \coloneqq 1 - \|z\|^2$ for $z \in B$, because of

$$(1 - \|z\|)\left(\frac{2\|z\|}{1 - \|z\|^2} + 1\right) = \frac{2\|z\|}{1 + \|z\|} + 1 - \|z\| \ge 2\|z\| + 1 - \|z\| \ge 1.$$

The advantage of this rewriting is that f is smooth on the open ball B (as compared to $(1 - \|z\|)^{-1}$) allowing the use of Iô's formula and one can get rid of the squared norm on the right-hand side of (31) thanks to $\|\nabla \log f\|^2 = \Delta f/f - \Delta \log f$, $f \ge 0$, and

$$\Delta f(z) = -2 < 0.$$

Therefore, we have

$$
\begin{aligned}
\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x} \|s(Z_t^x)\|^2\,\mathrm{d}t\Big] &\lesssim \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x} \|\nabla \log f(Z_t^x)\|^2\,\mathrm{d}t\Big] + \mathbb{E}^{z_0}[\tau_{1-\varepsilon}^x] \\
&\le \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x} -\Delta \log f(Z_t^x)\,\mathrm{d}t\Big] + \mathbb{E}^{z_0}[\tau_{1-\varepsilon}^x] \\
&= \frac{1}{q(x \mid z_0)}\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}} -\Delta \log f(W_t) q(x \mid W_t)\,\mathrm{d}t\Big] + \mathbb{E}^{z_0}[\tau_{1-\varepsilon}^x].
\end{aligned} \tag{33}
$$

The second term is bounded by a constant that is independent of ε by Lemma A.4. For the first term, the idea is to use Itô's formula to eliminate the time-integral, for which we need the Laplacian to act on the product $q(x \mid \cdot) \log f$. This can be achieved using the vector identity $q(x \mid \cdot)\Delta \log f = \Delta(q(x \mid \cdot) \log f) - \Delta_z q(x \mid \cdot) \log f - 2\nabla \log f \cdot \nabla_z q(x \mid \cdot)$. Moreover, the second term vanishes, since $q(x \mid \cdot)$ is harmonic. Inserting this identity in (33) yields

$$\begin{aligned}
\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x} \|\nabla \log f(Z_t^x)\|^2 \,\mathrm{d}t\Big] &\leq \frac{1}{q(x \mid z_0)}\bigg(- \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}} \Delta(\log f(W_t) q(x \mid W_t)) \,\mathrm{d}t\Big] \\
&\qquad + 2\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}} \nabla \log f(W_t) \cdot \nabla_z q(x \mid W_t) \,\mathrm{d}t\Big]\bigg) \\
&= -\frac{1}{q(x \mid z_0)} \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}} \Delta(\log f(W_t) q(x \mid W_t)) \,\mathrm{d}t\Big] \\
&\qquad + 2\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x} \nabla \log f(Z_t^x) \cdot \nabla_z \log q(x \mid Z_t^x) \,\mathrm{d}t\Big].
\end{aligned} \tag{34}$$

For the first term, we apply Itô's formula:

$$\begin{aligned}
&\frac{1}{q(x \mid z_0)} \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}} \Delta(\log f(W_t) q(x \mid W_t)) \,\mathrm{d}t\Big] \\
&= \frac{2}{q(x \mid z_0)}\bigg(\mathbb{E}^{z_0}[\log f(W_{\tau_{1-\varepsilon}}) q(x \mid W_{\tau_{1-\varepsilon}})] \\
&\qquad\qquad - \mathbb{E}^{z_0}[\log f(W_0) q(x \mid W_0)] \\
&\qquad\qquad - \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}} \nabla(\log f(W_t) q(x \mid W_t)) \cdot \mathrm{d}W_t\Big]\bigg) \\
&= 2\mathbb{E}^{z_0}[\log f(Z^x_{\tau^x_{1-\varepsilon}})] - 2\mathbb{E}^{z_0}[\log f(Z_0^x)] \\
&\quad - \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}} \nabla(\log f(W_t) q(x \mid W_t)) \cdot \mathrm{d}W_t\Big] \\
&= 2\log \frac{1 - (1-\varepsilon)^2}{1 - \|z_0\|^2} \\
&\quad - \mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}} \nabla(\log f(W_t) q(x \mid W_t)) \cdot \mathrm{d}W_t\Big],
\end{aligned}$$

The stochastic integral term vanishes, since $\nabla(q(x \mid \cdot) \log f)$ is bounded in $B_{1-\varepsilon}$ and so is the expectation of its quadratic variation. For the second term in (34) we use the Cauchy–Schwarz inequality and therefore obtain with the above

$$\begin{aligned}
&\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x} \|\nabla \log f(Z_t^x)\|^2 \,\mathrm{d}t\Big] \\
&\leq 2\log \frac{1 - \|z_0\|^2}{1 - (1-\varepsilon)^2} \\
&\quad + 2\left(\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x} \|\nabla \log f(Z_t^x)\|^2 \,\mathrm{d}t\Big]\right)^{1/2} \left(\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x} \|\nabla_z \log q(x \mid Z_t^x)\|^2 \,\mathrm{d}t\Big]\right)^{1/2}
\end{aligned}$$

This is now a quadratic inequality for the square root of the integral of interest, which implies

$$\begin{aligned}
&\left(\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x}\|\nabla \log f(Z_t^x)\|^2\,\mathrm{d}t\Big]\right)^{1/2}\\
&\leq \left(\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x}\|\nabla_z \log q(x\mid Z_t^x)\|^2\,\mathrm{d}t\Big]\right)^{1/2}\\
&\quad+\left(\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x}\|\nabla_z \log q(x\mid Z_t^x)\|^2\,\mathrm{d}t\Big]+2\log\frac{1-\|z_0\|^2}{1-(1-\varepsilon)^2}\right)^{1/2}\\
&\lesssim \left(\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}^x}\|\nabla_z \log q(x\mid Z_t^x)\|^2\,\mathrm{d}t\Big]\right)^{1/2}+(1+\log\varepsilon^{-1})^{1/2}\\
&\lesssim (1+\log\varepsilon^{-1})^{1/2},
\end{aligned}$$

where in the last step we used Lemma A.3. Squaring both sides and inserting in (31) finally yields the claim. □

F Proofs for Section 4.3

We start by proving the representation of the explicit score loss decomposition in Proposition 4.4.

Proof of Proposition 4.4. By writing $(P_t^{1-\varepsilon})_{t\geq 0}$ for the semigroup of the Brownian motion killed on first exit of $B_{1-\varepsilon}$, then the potential measure $U^{1-\varepsilon}$ for the killed Brownian Motion started in z_0 is given by $U^{1-\varepsilon}(z_0,\mathrm{d}z)=\int_0^\infty P_t^{1-\varepsilon}(z_0,\mathrm{d}z)\,\mathrm{d}t = G_{1-\varepsilon}(z_0,z)\,\mathrm{d}z$ and

$$\begin{aligned}
&\mathbb{E}^{z_0}\Big[\int_0^{\tau_{1-\varepsilon}}\|s(W_t)-\nabla\log h(W_t)\|^2 h(W_t)\,\mathrm{d}t\Big]\\
&=\int_0^\infty\int_{B_{1-\varepsilon}}\|s(z)-\nabla\log h(z)\|^2 h(z) P_t^{1-\varepsilon}(z_0,\mathrm{d}z)\,\mathrm{d}t\\
&=\int_{B_{1-\varepsilon}}\|s(z)-\nabla\log h(z)\|^2 h(z) U^{1-\varepsilon}(z_0,\mathrm{d}z)\\
&=\int_{B_{1-\varepsilon}} G_{1-\varepsilon}(z_0,z)\|s(z)-\nabla\log h(z)\|^2 h(z)\,\mathrm{d}z,
\end{aligned}\tag{35}$$

which yields the first statement about the representation of the explicit score loss as a weighted L^2 norm with respect to the Lebesgue density $G(z_0,\cdot)h$. The function h can be bounded by $\|\pi^*\|_{L^\infty(\partial B)}$ once again. The Green kernel, however, diverges as $z\to z_0$ with rate $\mathcal{O}(\|\cdot-z_0\|^{-d+2})$. Therefore, we split the ball into a smaller ball B_R, which contains z_0 and an annulus $B_{1-\varepsilon}/B_R$ on which $G_{1-\varepsilon}$ is bounded. This yields the estimate

$$\begin{aligned}
&\int_{B_{1-\varepsilon}} G_{1-\varepsilon}(z,z_0)\|s(z)-\nabla\log h(z)\|^2 h(z)\,\mathrm{d}z\\
&\leq \|\pi^*\|_{L^\infty(\partial B)}\left\|(s-\nabla\log h)\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\right\|^2_{L^2(B_{1-\varepsilon}/B_R)}\\
&\quad+\|s-\nabla\log h\|^2_{L^\infty(B_R)}\int_{B_R} G_{1-\varepsilon}(z,z_0)\,\mathrm{d}z.
\end{aligned}$$

The remaining integral over the Green function can be evaluated by applying (35) in reverse direction and replacing the integrand $\|s-\nabla\log h\|^2 h$ by 1:

$$\int_{B_R} G_{1-\varepsilon}(z,z_0)\,\mathrm{d}z\leq\int_{B_{1-\varepsilon}} G_{1-\varepsilon}(z,z_0)\,\mathrm{d}z=\mathbb{E}^{z_0}[\tau_{1-\varepsilon}]=\frac{(1-\varepsilon)^2-\|z_0\|^2}{d}<\infty.$$

□

Proof of Proposition 4.5. By shifting the origin to $-\mathbf{1} \in \mathbb{R}^d$ and rescaling the space with a factor of 2, Lemma I.9 can be applied to functions on $[-1,1]^d$ as well. Since the ball B_R has a smooth boundary, the Sobolev extension theorem allows to extend the components of $\nabla \log h$ on B_R to functions $u_i \in H^\gamma(\mathbb{R}^d)$, $i = 1, \dots, d$, with $\|u_i\|_{H^\gamma(\mathbb{R}^d)} \lesssim \|(\nabla \log h)_i\|_{H^\gamma(B_R)}$ for any $\gamma \geq 1$. Choosing $\gamma \coloneqq \frac{d}{d-1}\alpha$, which is valid for $\alpha > (d-1)/2$, and parallelising the network approximations for each component, then yields the desired network based on Lemma I.9. □

The remainder of this section is devoted to the proof of Theorem 4.3. To this end, our goal is to first motivate and derive the central technical results Lemma 4.6 and Proposition 4.7 based on the approximation strategy outlined in Section 4.3, and then put things together for the proof of Theorem 4.3. In our analysis, we will make repeated use of the following uniform bounds on the gradient ∇h_N and the Hessian $\nabla^2 h_N$.

Lemma F.1. *For $r \in (0,1)$, it holds*

(i) $\|\nabla h_N(r,\cdot)\|_{L^\infty(\partial B)} \lesssim \frac{1}{1-r^2}\|\pi^*\|_{H^{(d-1)/2}(\partial B)}$,

(ii) $\|\nabla^2 h_N(r,\cdot)\|_{L^\infty(\partial B)} \coloneqq \sup_{i,j}\|\partial_i\partial_j h_N(r,\cdot)\|_{L^\infty(\partial B)} \lesssim \frac{r}{(1-r^2)^2}\|\pi^*\|_{H^{(d-1)/2}(\partial B)}$.

Proof. Let $x \in \partial B$. Then, using [26, Theorem 4]

$$\begin{aligned}
\|\nabla h_N(r,x)\| &\leq \sum_{l=1}^N \sum_{m=1}^{M_l} |a_{lm}| \big(l r^{l-1} |Y_{lm}(x)| \|x\| + r^{l-1} \|\nabla_{\partial B} Y_{lm}(x)\| \big) \\
&\lesssim \sum_{l=1}^N \sum_{m=1}^{M_l} |a_{lm}| r^{l-1} \big(l\, l^{d/2-1} + l^{d/2} \big) \\
&\leq \left(\sum_{l=1}^N l r^{2(l-1)} \right)^{1/2} \left(\sum_{l=1}^N \sum_{m=1}^{M_l} |a_{lm}|^2 l^{d-1} \right)^{1/2} \\
&\leq \frac{1}{1-r^2} \|\pi^*\|_{H^{(d-1)/2}(\partial B)},
\end{aligned}$$

which proves (i). For the Hessian of h_N, we have

$$\begin{aligned}
\nabla^2 h_N(r,x) = \sum_{l=1}^N \sum_{m=1}^{M_l} a_{lm} r^{l-2} \Big[& l\big(\mathbb{I} + (l-2)xx^\top\big) Y_{lm}(x) + l\big(x \nabla_{\partial B} Y_{lm}(x)^\top + \nabla_{\partial B} Y_{lm}(x) x^\top \big) \\
& + \nabla^2_{\partial B} Y_{lm}(x) \Big].
\end{aligned}$$

By using [26, Theorem 4] once again, it follows similarly

$$\begin{aligned}
\|\nabla^2 h_N(r,\cdot)\|_{L^\infty(\partial B)} &\leq \sum_{l=1}^N \sum_{m=1}^{M_l} |a_{lm}| r^{l-2} \big[l(l-1) \|Y_{lm}\|_{L^\infty(\partial B)} \\
&\qquad + 2l \|\nabla_{\partial B} Y_{lm}\|_{L^\infty(\partial B)} + \|\nabla^2_{\partial B} Y_{lm}\|_{L^\infty(\partial B)} \big] \\
&\lesssim \sum_{l=1}^N \sum_{m=1}^{M_l} |a_{lm}| r^{l-2} \big(l^2 l^{d/2-1} + l\, l^{d/2} + l^{d/2+1} \big) \\
&\leq \left(\sum_{l=1}^N l^3 r^{2(l-2)} \right)^{1/2} \left(\sum_{l=1}^N \sum_{m=1}^{M_l} |a_{lm}|^2 l^{d-1} \right)^{1/2} \\
&\leq \left(\frac{1}{r^2} + 8 + 6r^2 \sum_{l=3}^N l(l-1)(l-2) r^{2(l-3)} \right)^{1/2} \|\pi^*\|_{H^{(d-1)/2}(\partial B)} \\
&\leq \left(\frac{1}{R^2} + 8 + 6r^2 \frac{6}{(1-r^2)^4} \right)^{1/2} \|\pi^*\|_{H^{(d-1)/2}(\partial B)} \\
&\lesssim \frac{r}{(1-r^2)^2} \|\pi^*\|_{H^{(d-1)/2}(\partial B)},
\end{aligned}$$

which proves (ii). □

Defining the full network as

$$\overline{s}(z) \coloneqq \phi_{\mathrm{mult}}(s_{\nabla h_N}(z), \phi_{\mathrm{rec}}(s_{h_N}(z) \vee \pi^*_{\min}), \qquad z \in B_{1-\varepsilon} \setminus B_R, \tag{36}$$

with a multiplication network ϕ_{mult} as in Lemma I.5, a network ϕ_{rec} approximating $z \mapsto 1/z$ as in Lemma I.7, and approximating networks s_{h_N} and $s_{\nabla h_N}$ for h_N and ∇s_{h_N} defined later, the total approximation error on $B_{1-\varepsilon} \setminus B_R$ can be bounded by

$$\begin{aligned}
&\|(\overline{s} - \nabla \log h)\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_{1-\varepsilon}\setminus B_R)} \\
&\le \|(\overline{s} - s_{\nabla h_N}\phi_{\mathrm{rec}} \circ (s_{h_N} \vee \pi^*_{\min}))\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_{1-\varepsilon}\setminus B_R)} \\
&\quad + \left\| s_{\nabla h_N}\Big(\phi_{\mathrm{rec}} \circ (s_{h_N} \vee \pi^*_{\min}) - \frac{1}{s_{h_N} \vee \pi^*_{\min}}\Big)\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\right\|_{L^2(B_{1-\varepsilon}\setminus B_R)} \\
&\quad + \left\| \Big(\frac{s_{\nabla h_N}}{s_{h_N} \vee \pi^*_{\min}} - \nabla \log h\Big)\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\right\|_{L^2(B_{1-\varepsilon}\setminus B_R)}.
\end{aligned} \tag{37}$$

The first term is controlled by

$$\begin{aligned}
&\|(\overline{s} - s_{\nabla h_N}\phi_{\mathrm{rec}} \circ (s_{h_N} \vee \pi^*_{\min}))\underbrace{\sqrt{G_{1-\varepsilon}(z_0,\cdot)}}_{\lesssim 1}\|_{L^2(B_{1-\varepsilon}\setminus B_R)} \\
&\lesssim 2^{-l_1}\|s_{\nabla h_N}\|_{L^\infty(B_{1-\varepsilon})}\,\|\phi_{\mathrm{rec}} \circ (s_{h_N} \vee \pi^*_{\min})\|_{L^\infty(B_{1-\varepsilon})} \\
&\lesssim 2^{-l_1}\varepsilon^{-1}\left(\left\|\phi_{\mathrm{rec}} \circ (s_{h_N} \vee \pi^*_{\min}) - \frac{1}{s_{h_N} \vee \pi^*_{\min}}\right\|_{L^\infty(B_{1-\varepsilon})} + \left\|\frac{1}{s_{h_N} \vee \pi^*_{\min}}\right\|_{L^\infty(B_{1-\varepsilon})}\right) \\
&\lesssim 2^{-l_1}\varepsilon^{-1}(2^{-l_2} + \pi_{\min}^{-1}),
\end{aligned}$$

while the second term is bounded by

$$\begin{aligned}
&\left\| s_{\nabla h_N}\Big(\phi_{\mathrm{rec}} \circ (s_{h_N} \vee \pi^*_{\min}) - \frac{1}{s_{h_N} \vee \pi^*_{\min}}\Big)\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\right\|_{L^2(B_{1-\varepsilon}\setminus B_R)} \\
&\lesssim \left\|\phi_{\mathrm{rec}} \circ (s_{h_N} \vee \pi^*_{\min}) - \frac{1}{s_{h_N} \vee \pi^*_{\min}}\right\|_{L^\infty(B_{1-\varepsilon})} \int_{B_{1-\varepsilon}} \frac{1}{(1-\|z\|)^2} G_{1-\varepsilon}(z_0, z)\,\mathrm{d}z \\
&\lesssim 2^{-l_2}\log \varepsilon^{-1},
\end{aligned}$$

using Lemma A.5. The third term is the most challenging one and can be further decomposed as

$$\begin{aligned}
&\left\| \Big(\frac{s_{\nabla h_N}}{s_{h_N} \vee \pi^*_{\min}} - \nabla \log h\Big)\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\right\|_{L^2(B_{1-\varepsilon}\setminus B_R)} \\
&= \left\| \Big(\frac{\nabla h}{h} - \frac{\nabla h_N}{h_N \vee \pi^*_{\min}}\Big)\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\right\|_{L^2(B_{1-\varepsilon}\setminus B_R)} \\
&\quad + \left\| \Big(\frac{\nabla h_N}{h_N \vee \pi^*_{\min}} - \frac{s_{\nabla h_N}}{s_{h_N} \vee \pi^*_{\min}}\Big)\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\right\|_{L^2(B_{1-\varepsilon}\setminus B_R)} \\
&\le \left\| \Big(\frac{\nabla h - \nabla h_N}{h}\Big)\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\right\|_{L^2(B_{1-\varepsilon}\setminus B_R)} \\
&\quad + \left\| \nabla h_N\left(\frac{1}{h} - \frac{1}{h_N \vee \pi^*_{\min}}\right)\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\right\|_{L^2(B_{1-\varepsilon}\setminus B_R)} \\
&\quad + \left\| \frac{\nabla h_N - s_{\nabla h_N}}{h_N \vee \pi^*_{\min}}\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\right\|_{L^2(B_{1-\varepsilon}\setminus B_R)} \\
&\quad + \left\| s_{\nabla h_N}\left(\frac{1}{h_N \vee \pi^*_{\min}} - \frac{1}{s_{h_N} \vee \pi^*_{\min}}\right)\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\right\|_{L^2(B_{1-\varepsilon}\setminus B_R)}
\end{aligned}$$

$$
\begin{aligned}
&\lesssim \pi_{\min}^{-1}\|\nabla h - \nabla h_N\|_{L^2(B_{1-\varepsilon})} + \pi_{\min}^{-2}\|\nabla h_N \underbrace{|h - h_N \vee \pi^*_{\min}|}_{\le |h-h_N|}\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_{1-\varepsilon}\setminus B_R)} \\
&+ \pi_{\min}^{-1}\|(\nabla h_N - s_{\nabla h_N})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_{1-\varepsilon}\setminus B_R)} \\
&+ \pi_{\min}^{-2}\|s_{\nabla h_N}\underbrace{|h_N \vee \pi^*_{\min} - s_{h_N} \vee \pi^*_{\min}|}_{\le |h_N - s_{h_N}|}\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_{1-\varepsilon}\setminus B_R)},
\end{aligned} \tag{38}
$$

where we used $h = \mathbb{E}^{\cdot}[\pi^*(W_\tau)] \ge \pi_{\min} \coloneqq \inf_{x\in\partial B}\pi^*(x) > 0$. The first two terms in (38) are bounded by Lemma 4.6, which we now prove.

Proof of Lemma 4.6. For part (i), we first expand the left hand side as

$$
\begin{aligned}
&\|\nabla h_N(h-h_N)\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|^2_{L^2(B_{1-\varepsilon}\setminus B_R)} \\
&\le \int_R^{1-\varepsilon} \|\nabla h_N(r,\cdot)\|^2_{L^\infty(\partial B)}\|(h(r,\cdot)-h_N(r,\cdot))\underbrace{\sqrt{G_{1-\varepsilon}(z_0,(r,\cdot))}}_{\lesssim\sqrt{1-r}}\|^2_{L^2(\partial B)} r^{d-1}\,\mathrm{d}r.
\end{aligned} \tag{39}
$$

According to Lemma H.3, the Green kernel is bounded by $1-r$ up to constants. The L^2-distance between $h(r,\cdot)$ and $h_N(r,\cdot)$ evaluates as

$$
\begin{aligned}
\|h(r,\cdot)-h_N(r,\cdot)\|^2_{L^2(\partial B)} &= \sum_{l=N+1}^{\infty}\sum_{m=1}^{M_l} r^{2l}a_{lm}^2 \\
&\lesssim (1-\varepsilon)^{2(N+1)}N^{-2\alpha}\sum_{l=0}^{\infty}\sum_{m=1}^{M_l}(l(l+d-2))^{\alpha}|\langle\pi^*,Y_{lm}\rangle|^2 \\
&\lesssim (1-\varepsilon)^{2(N+1)}N^{-2\alpha}\|\pi^*\|^2_{H^\alpha(\partial B)}.
\end{aligned} \tag{40}
$$

Inserting (40) and Lemma F.1(i) in (39), and using $(1-r^2)^{-1}\le(1-r)^{-1}$, one gets

$$
\begin{aligned}
\|\nabla h_N(h-h_N)\|^2_{L^2(B_{1-\varepsilon})} &\lesssim (1-\varepsilon)^{2(N+1)}N^{-2\alpha}\|\pi^*\|^2_{H^{(d-1)/2}(\partial B)}\|\pi^*\|^2_{H^\alpha(\partial B)}\int_0^{1-\varepsilon}\frac{r^{d-1}}{1-r}\,\mathrm{d}r \\
&\lesssim (1-\varepsilon)^{2N+d+1}N^{-2\alpha}\log\varepsilon^{-1}\|\pi^*\|^4_{H^\alpha(\partial B)},
\end{aligned}
$$

which proves (i). For part (ii) we apply a similar idea to the expansion (20):

$$
\begin{aligned}
&\|\nabla h(r,\cdot)-\nabla h_N(r,\cdot)\|^2_{L^2(\partial B)} \\
&= \sum_{l,k=N+1}^{\infty}\sum_{m=1}^{M_l}\sum_{m'=1}^{M_k} r^{l+k-2}a_{lm}a_{km'}\Big(l^2\underbrace{\int_{\partial B}Y_{lm}(x)Y_{km'}(x)\overbrace{\|x\|^2}^{=1}\,\sigma(\mathrm{d}x)}_{=\delta_{lk}\delta_{mm'}} \\
&\qquad + 2l\int_{\partial B}Y_{km'}(x)x\cdot\nabla_{\partial B}Y_{lm}(x)\,\sigma(\mathrm{d}x) \\
&\qquad + \langle\nabla_{\partial B}Y_{lm}\cdot\nabla_{\partial B}Y_{km'}\rangle\Big).
\end{aligned}
$$

The second term vanishes, since the covariant derivative $\nabla_{\partial B}Y_{lm}(x)$ is tangential to the manifold and thus orthogonal to the normal vector x. The third term can be evaluated using integration by parts on the sphere:

$$
\langle\nabla_{\partial B}Y_{lm}\cdot\nabla_{\partial B}Y_{km'}\rangle = \int_{\partial B}\nabla_{\partial B}Y_{lm}\cdot\nabla_{\partial B}Y_{km'}\,\mathrm{d}\sigma = -\int_{\partial B}Y_{km'}\Delta_{\partial B}Y_{lm}\,\mathrm{d}\sigma = l(l+d-2)\delta_{lk}\delta_{km'}.
$$

Thus, it follows

$$
\|\nabla h(r,\cdot)-\nabla h_N(r,\cdot)\|^2_{L^2(\partial B)} \le \sum_{l=N+1}^{\infty}\sum_{m=1}^{M_l} r^{2(l-1)}l(l+d-2)|a_{lm}|^2,
$$

and

$$\begin{aligned}\|\nabla h - \nabla h_N\|^2_{L^2(B_{1-\varepsilon})} &= \int_0^{1-\varepsilon} \|\nabla h(r,\cdot) - \nabla h_N(r,\cdot)\|^2_{L^2(\partial B)} r^{d-1}\,\mathrm{d}r \\ &\leq \sum_{l=N+1}^{\infty}\sum_{m=1}^{M_l} l \underbrace{\frac{(l+d-2)}{2l+d-2}}_{\leq 1}(1-\varepsilon)^{2l+d-2}|a_{lm}|^2 \\ &\leq (1-\varepsilon)^{2N+d} N^{-2(\alpha-1/2)} \sum_{l=0}^{\infty}\sum_{m=1}^{M_l} (l(l+d-2))^{\alpha}|a_{lm}|^2 \\ &\lesssim (1-\varepsilon)^{2N+d} N^{-2\alpha+1} \|\pi^*\|^2_{H^\alpha(\partial B)},\end{aligned}$$

which proves (ii). □

For the two remaining terms in (38), we turn to the approximation of h_N and ∇h_N. To this end, it will be necessary to separate the radial from the spherical direction. Thus, we define the stereographic projection maps (see, e.g., [19] and [20]) given by

$$\varphi_+(x) \coloneqq (r,\theta_1,\dots,\theta_{d-1}) = \left(\sqrt{\sum_{i=1}^d x_i^2}, -\frac{x_1}{1+x_d},\dots,-\frac{x_{d-1}}{1+x_d}\right), \qquad x_d \in (-1,1],$$

$$\varphi_-(x) \coloneqq (r,\theta_1,\dots,\theta_{d-1}) = \left(\sqrt{\sum_{i=1}^d x_i^2}, \frac{x_1}{1-x_d},\dots,\frac{x_{d-1}}{1-x_d}\right), \qquad x_d \in [-1,1).$$

These two coordinate maps map the northern (for φ_+) and the southern (for φ_-) half of the unit ball to $[0,1]\times B_1^{(d-1)}$, where $B_R^{(d-1)} \subset \mathbb{R}^{d-1}$ denotes the ball of radius R in $(d-1)$-dimensional space. Thus, for fixed radius r, their inverse maps parametrise the northern and the southern hemisphere by taking vectors from the $(d-1)$-dimensional unit sphere as input, which are given explicitly by

$$\begin{aligned}\varphi_+^{-1}(r,\theta) &\coloneqq \left(\frac{2r^2\theta_1}{\|\theta\|^2+r^2},\dots,\frac{2r^2\theta_{d-1}}{\|\theta\|^2+r^2}, r\frac{\|\theta\|^2-r^2}{\|\theta\|^2+r^2}\right), \\ \varphi_-^{-1}(r,\theta) &\coloneqq \left(-\frac{2r^2\theta_1}{\|\theta\|^2+r^2},\dots,-\frac{2r^2\theta_{d-1}}{\|\theta\|^2+r^2}, -r\frac{\|\theta\|^2-r^2}{\|\theta\|^2+r^2}\right),\end{aligned}$$

The stereographic coordinate system has the advantage, compared to standard spherical coordinates, that it does not involve trigonometric functions and the determinant of its metric, which appears in surface integrals, is strictly bounded from below. This will come in handy, when approximating φ_+ and φ_- with neural networks and for changing from the curvy-linear surface measure on ∂B to the Lebesgue measure on $B_1^{(d-1)}$. In the following, since the analysis of the northern and southern half of the ball is completely analogous, we will sometimes write $\varphi_\pm$ as a place holder for either φ_+ or φ_-.

We will at some point consider functions at a fixed radius r and approximate their composition with $\varphi_\pm^{-1}(r,\cdot)$. For better readability, we define $\varphi_{\partial B,\pm}(x) \coloneqq (\varphi_\pm(x)_2,\dots,\varphi_\pm(x)_d)$ for $x \in \partial B$, which means that $\varphi_{\partial B,\pm}^{-1} = \varphi_\pm^{-1}(r,\cdot)/r = \varphi_\pm^{-1}(1,\cdot)$.

Our approximation strategy of h_N and ∇h_N is based on the following steps:

1. Approximate $h_N(r_i,\cdot)\circ\varphi_{\partial B,\pm}^{-1}$ and $\nabla h_N(r_i,\cdot)\circ\varphi_{\partial B,\pm}^{-1}$ uniformly on the coordinate space for the northern and southern half separately with neural networks $s_{h_N\circ\varphi_\pm^{-1},r_i}$ and $s_{\nabla h_N\circ\varphi_\pm^{-1},r_i}$ for fixed r_i.
2. Interpolate $s_{h_N\circ\varphi_\pm^{-1},r_i}$ and $s_{\nabla h_N\circ\varphi_\pm^{-1},r_i}$ with $\mathcal{O}(\log N)$ Chebyshev polynomials and nodes r_i on shells $[1-2^{m-1}\varepsilon, 1-2^{m+1}\varepsilon]\times\partial B$, $m=1,\dots,M$ and $M=\mathcal{O}(\log N)$.
3. Approximate Chebyshev polynomials by neural networks (analogously to [14]).

4. Approximate $h_N \circ \varphi_\pm^{-1}$ and $\nabla h_N \circ \varphi_\pm^{-1}$ in L^2 by combining the approximations on the shells with a partition of unity of neural networks (analogously to [14]) to networks $s_{h_N \circ \varphi_\pm^{-1}}$ and $s_{\nabla h_N \circ \varphi_\pm^{-1}}$.
5. Approximate φ_+ and φ_- by neural networks s_{φ_+} and s_{φ_-}.
6. Approximate h_N with a partition of unity (p_+, p_-) via $s_{h_N} = \phi_{\text{mult}} \circ (p_+, s_{h_N \circ \varphi_+^{-1}} \circ s_{\varphi_+}) + \phi_{\text{mult}} \circ (p_-, s_{h_S \circ \varphi_-^{-1}} \circ s_{\varphi_-})$ and similarly for ∇h_N.

The first four steps are summarised in the following proposition.

Proposition F.2. *There exist neural networks $s_{h_N \circ \varphi_\pm^{-1}}, s_{\nabla h_N \circ \varphi_\pm^{-1}} \in \mathcal{S}(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})$ with*

$$\mathsf{L} \lesssim \log N \log\log N, \qquad \mathsf{W} \lesssim N \log^2 N \qquad \mathsf{S} \lesssim N \log^3 N, \qquad \mathsf{B} \lesssim N^{1/(d-1)} \vee \varepsilon^{-1},$$

and for $r \in [R, 1-\varepsilon]$

$$\|h_N(r,\cdot) \circ \varphi_{\partial B,\pm}^{-1} - s_{h_N \circ \varphi_\pm^{-1}}(r,\cdot)\|_{L^\infty([-1,1]^{d-1})} \lesssim N^{-\alpha/(d-1)} \log N,$$
$$\|\nabla h_N(r,\cdot) \circ \varphi_{\partial B,\pm}^{-1} - s_{\nabla h_N \circ \varphi_\pm^{-1}}(r,\cdot)\|_{L^\infty([-1,1]^{d-1})} \lesssim \frac{1}{1-r} N^{-\alpha/(d-1)} \log N.$$

Proof. We follow the step-by-step approach outlined above. Since the procedure is identical for both halves of the ball, we just write $\varphi_{\partial B} \coloneqq \varphi_{\partial B,\pm}$.

Step 1: Approximation at fixed radius For fixed radius $r \in [R, 1-\varepsilon]$, we aim to apply Lemma I.9, which means that we have to bound the Sobolev norm of $h_N(r,\cdot) \circ \varphi_{\partial B}^{-1}$ and $\nabla h_N(r,\cdot) \circ \varphi_{\partial B}^{-1}$. Using Lemma H.1, we get

$$\begin{aligned}
\|h_N(r,\cdot) \circ \varphi_{\partial B}^{-1}\|_{H^\alpha([-1,1]^{d-1})}^2 &\lesssim \|h_N(r,\cdot)\|_{H^\alpha(\partial B)}^2 \\
&= \|h_N(r,\cdot)\|_{L^2(\partial B)}^2 + \sum_{l=0}^{N} (l(l+d-2))^\alpha \sum_{m=1}^{M_l} |\langle h_N(r,\cdot), Y_{lm}\rangle|^2 \\
&= \sum_{l=0}^{N} r^{2l} \sum_{m=1}^{M_l} |a_{lm}|^2 + \sum_{l=0}^{N} r^{2l} (l(l+d-2))^\alpha \sum_{m=1}^{M_l} |a_{lm}|^2 \\
&\leq \|\pi^*\|_{H^\alpha(\partial B)}^2,
\end{aligned}$$

where in the last step we used $r \leq 1$. Furthermore, using additionally Lemma H.2 for the second line, we obtain

$$\begin{aligned}
&\|\nabla h_N(r,\cdot) \circ \varphi_{\partial B}^{-1}\|_{H^\alpha([-1,1]^{d-1})}^2 \\
&\lesssim \|\nabla h_N(r,\cdot)\|_{H^\alpha(\partial B)}^2 \\
&= \sum_{l,k=1}^{N} r^{l+k-2} \Big\langle \sum_{m=1}^{M_l} (l Y_{lm} \cdot + \nabla_{\partial B} Y_{lm}), \sum_{m'=1}^{M_k} (k Y_{km'} \cdot + \nabla_{\partial B} Y_{km'}) \Big\rangle_{H^\alpha(\partial B)}^2 \\
&= \sum_{l=1}^{N} r^{2(l-1)} \left(l(l+d-1)(1 + ((l-1)(l+d-3))^\alpha)\right) |a_{lm}|^2 \\
&\lesssim \sum_{l=0}^{N} l^2 r^{2l} (1 + (l(l+d-2))^\alpha)) |a_{lm}|^2 \\
&\leq \sup_{l \in \mathbb{N}} l^2 r^{2l} \ \|\pi^*\|_{H^\alpha(\partial B)}^2 .
\end{aligned}$$

The supremum of the sequence $(l^2 r^{2l})_{l \in \mathbb{N}}$ is obviously bounded by the supremum of its extension to the non-negative real line, which can be computed by setting its derivative to zero, which yields

$$\sup_{l \in \mathbb{N}} l^2 r^{2l} \leq \log(r)^{-2} \mathrm{e}^{-2} \lesssim (1-r)^{-2}.$$

Thus, Lemma I.9 provides neural networks $\phi_{h_N\circ\varphi^{-1},r}$ and $\phi_{\nabla h_N\circ\varphi^{-1},r}$ with size

$$\mathsf{L}\lesssim\log N,\qquad \mathsf{W}\lesssim N,\qquad S\lesssim(\mathsf{L}-1)N,\qquad \mathsf{B}=\mathrm{Poly}(N),$$

such that

$$\|h_N(r,\cdot)\circ\varphi_{\partial B}^{-1}-\phi_{h_N\circ\varphi^{-1},r}\|_{L^\infty([-1,1]^{d-1})}\lesssim N^{-\alpha/(d-1)},$$
$$\|\nabla h_N(r,\cdot)\circ\varphi_{\partial B}^{-1}-\phi_{\nabla h_N\circ\varphi^{-1},r}\|_{L^\infty([-1,1]^{d-1})}\lesssim\frac{1}{1-r}N^{-\alpha/(d-1)}.$$

Step 2: Approximation on $[1-2^{m+1}\varepsilon,1-2^{m-1}\varepsilon]\times\partial B$ via Chebyshev interpolation We now use step 1 to approximate h_N and ∇h_N on $[b_m-a_m,b_m+a_m]\times\partial B$ with $a_m=3\times2^{m-2}\varepsilon$ and $b_m=1-5\times2^{m-2}\varepsilon$ for $m=1,\dots,M$, where $M\coloneqq\lceil\log_2((1-R)/\varepsilon)+1\rceil$ in order to cover the entire annulus $B_{1-\varepsilon}\setminus B_R$. Since the procedure is the same for both h_N and ∇h_N, we denote the rescaled functions $(r,x)\mapsto h_N(a_mr+b_m,x)$ and $(r,x)\mapsto\nabla h_N(a_mr+b_m,x)$, $r\in[-1,1]$, $x\in\partial B$, by f_m. This can be approximated by the Chebyshev polynomials

$$\psi_{f_m}(r,x)=\sum_{i=1}^{k}c_ip_i(r)f_m(r_i,x),\quad r\in[-1,1],\,x\in\partial B,\tag{41}$$

with $r_i=\cos(i\pi/k)$, $p_i(r)=\prod_{i\neq j}^{k}(r-r_i)$, $c_i=1/p_i(r_i)$ and $k\in\mathbb{N}$ to be determined later as well. Since f_m is an entire function on $\mathbb{C}$, it holds that [32, Theorem 8.2]

$$|f_m(r,x)-\psi_{f_m}(r,x)|\leq\frac{4M_{m,\rho}(x)\rho^{-k}}{\rho-1},\tag{42}$$

for some $\rho>1$ and

$$M_{m,\rho}(x)\coloneqq\max_{z\in\partial E_\rho}|f_m(z,x)|,\qquad\partial E_\rho\coloneqq\left\{\frac{z+z^{-1}}{2}:z\in\mathbb{C},|z|=\rho\right\}.$$

This function is indeed bounded for all $\rho>1$ on $[b_m-a_m,b_m+a_m]\times\partial B$. To see this, note that for $y=(z+z^{-1})/2\in E_\rho$ with $z=\rho\mathrm{e}^{\mathrm{i}\varphi}$

$$|y|^2=\frac{|z|^2+|z|^{-2}+2\mathrm{Re}(z/\bar z)}{4}=\frac{\rho^2+\rho^{-2}+2\cos(2\varphi)}{4}\leq\left(\frac{\rho+\rho^{-1}}{2}\right)^2,$$
$$\mathrm{Re}(y)=\frac{\mathrm{Re}(z)+\mathrm{Re}(1/z)}{2}=\frac{\rho+\rho^{-1}}{2}\cos(\varphi)\leq\frac{\rho+\rho^{-1}}{2}.$$

Thus,

$$\begin{aligned}|h_N(y,x)|&\leq\sum_{l=0}^{N}|a_my+b_m|^l|Y_{lm}(x)||a_{lm}|\\&=\sum_{l=0}^{N}(a_m^2|y|^2+2a_mb_m\mathrm{Re}(y)+b_m^2)^{l/2}|Y_{lm}(x)||a_{lm}|\\&\leq\sum_{l=0}^{N}\left[a_m^2\left(\frac{\rho+\rho^{-1}}{2}\right)^2+2a_mb_m\frac{\rho+\rho^{-1}}{2}+b_m^2\right]^{l/2}|Y_{lm}(x)||a_{lm}|\\&=\sum_{l=0}^{N}\left(a_m\frac{\rho+\rho^{-1}}{2}+b_m\right)^l|Y_{lm}(x)||a_{lm}|,\end{aligned}$$

and similarly

$$\begin{aligned}\|\nabla h_N(y,x)\|_\infty&\leq\sum_{l=1}^{N}|a_my+b_m|^{l-1}|a_{lm}|(l|Y_{lm}(x)|\underbrace{\|x\|_\infty}_{\leq1}+\|\nabla_{\partial B}Y_{lm}(x)\|_\infty)\\&\leq\sum_{l=1}^{N}\left(a_m\frac{\rho+\rho^{-1}}{2}+b_m\right)^{l-1}|a_{lm}|(l|Y_{lm}(x)|+\|\nabla_{\partial B}Y_{lm}(x)\|_\infty).\end{aligned}$$

For the right-hand sides to be bounded for $N \to \infty$, one needs $a_m(\rho + \rho^{-1})/2 + b_m \leq 1$ for all $m \in \mathbb{N}$, which holds for $\rho \in (1, 3]$. Fix such ρ in the following. Theorem 4 in [26] provides the bounds $|Y_{lm}(x)| \lesssim l^{d/2-1}$ and $\|\nabla_{\partial B} Y_{lm}(x)\| \lesssim l^{d/2}$. This results in

$$\max_{y \in E_\rho} |h_N(y,x)| \lesssim \sum_{l=0}^{N} l^{d/2-1} |a_{lm}| \leq \sqrt{N} \sqrt{\sum_{l=0}^{N} l^{d-2} |a_{lm}|^2} \leq \sqrt{N} \|\pi^*\|_{H^{d/2-1}(\partial B)},$$

$$\max_{y \in E_\rho} \|\nabla h_N(y,x)\| \lesssim \sum_{l=1}^{N} l^{d/2} |a_{lm}| \leq \sqrt{\sum_{l=1}^{N} l} \sqrt{\sum_{l=1}^{N} l^{d-1} |a_{lm}|^2} \leq \sqrt{\frac{N(N+1)}{2}} \|\pi^*\|_{H^{(d-1)/2}(\partial B)},$$

using Hölder's inequality in both cases, which means that

$$\|M_{m,\rho}\|_{L^\infty(\partial B)} \lesssim N \|\pi^*\|_{H^{(d-1)/2}(\partial B)}$$

and for (42)

$$\|f_m(r,\cdot) - \psi_{f_m}(r,\cdot)\|_{L^\infty(\partial B)} \lesssim N \|\pi^*\|_{H^{(d-1)/2}(\partial B)} \rho^{-k}.$$

Choosing $k \coloneqq \lceil (\alpha/(d-1) + 1) \log_\rho N \rceil$ yields the desired $N^{-\alpha/(d-1)}$ rate of convergence.

Now, since f_m can be approximated by a logarithmic number of Chebyshev polynomials with the appropriate rate, ψ_m can be approximated summand-by-summand with a polynomial rate in N, while keeping a total network size of $\mathcal{O}(\log N)$. Thus, we define the neural network approximation for ψ_m as

$$\phi_m(r,x) = \sum_{i=1}^{k} c_i \, \phi_{\text{mult}}(\phi_{p_i}(r), \phi_{f, a_m r_i + b_m}(x)),$$

where ϕ_{p_i} is the neural network approximation of p_i constructed in step 3 and ϕ_{mult} as in Lemma I.5. Repeating the arguments from [14, Lemma 3.13.], where we replace $\underline{T}$ by ε^2, it follows that

$$\|\phi_m(r,\cdot) - \psi_{f_m}(r,\cdot)\|_{L^\infty(\partial B)} \lesssim \begin{cases} N^{-\alpha/(d-1)} \log N, & \text{if } f_m = h_N \\ \frac{1}{2^{m-1}\varepsilon} N^{-\alpha/(d-1)} \log N, & \text{if } f_m = \nabla h_N, \end{cases}$$

with a network size

$$\mathsf{L} \lesssim \log N \log\log N, \qquad \mathsf{W} \lesssim N, \qquad \mathsf{S} \lesssim N \log N, \qquad \mathsf{B} \lesssim N^{1/(d-1)} \vee \varepsilon^{-1}.$$

Step 3: Approximation of Chebyshev polynomials The approximation of the Chebyshev polynomials p_i with neural networks $\phi_{p_i} \in \mathrm{NN}(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})$ has been shown in [14, Lemma 3.13], which have sizes

$$\mathsf{L} \lesssim \log N \log\log N, \qquad \mathsf{W} \lesssim \log N$$
$$\mathsf{S} \lesssim \log^2 N, \qquad \mathsf{B} = const.$$

Step 4: Combining neural networks on shells with partition of unity From the three previous steps, we get neural networks $\phi_1, \ldots, \phi_M$ such that

$$\|\phi_m(r,\cdot) - f_m(r,\cdot)\|_{L^\infty(\partial B)} \lesssim \begin{cases} N^{-\alpha/(d-1)} \log N, & \text{if } f = h_N \\ \frac{1}{2^{m-1}\varepsilon} N^{-\alpha/(d-1)} \log N, & \text{if } f = \nabla h_N, \end{cases}$$

for $r \in [1 - 2^{m+1}\varepsilon, 1 - 2^{m-1}\varepsilon]$. We define a partition of unity $(\mathsf{p}_m)_{m=1,\ldots,M}$ in radial direction via

$$\mathsf{p}_1(r) \coloneqq 0 \vee \left(1 \wedge \frac{1 - 2^{m+1}\varepsilon - r}{2^{m-1}\varepsilon}\right),$$

$$\mathsf{p}_m(r) \coloneqq 0 \vee \left(\frac{r - (1 - 2^{m-1}\varepsilon)}{2^{m-1}\varepsilon} \wedge \frac{1 - 2^{m+1}\varepsilon - r}{2^{m-1}\varepsilon}\right), \quad m = 2, \ldots, M-1,$$

$$\mathsf{p}_M(r) \coloneqq 0 \vee \left(\frac{r - (1 - 2^{m-1}\varepsilon)}{2^{m-1}\varepsilon} \wedge 1\right).$$

These functions can be represented exactly by two-layer neural networks due to Lemma I.4 (expressing max and min by neural networks). Thus, we define the overall network for f, being either h_N or $\partial_i h_N$, $i = 1, \ldots, d$, as

$$s_{f \circ \varphi^{-1}}(r,x) \coloneqq \sum_{m=1}^{M} \phi_{\text{mult}}(\phi_m(r,x), \mathsf{p}_m(r)),$$

where ϕ_{mult} is again given by Lemma I.5.

Step 5: Putting things together Using the approximation properties of ϕ_{mult} and the results from the previous steps, we can now control the approximation error of $s_{f_m \circ \varphi^{-1}}$: for $r \in [R, 1-\varepsilon]$, there exists $m \in \{1, \dots, M-1\}$ such that $r \in [1-2^m\varepsilon, 1-2^{m-1}\varepsilon]$ and

$$\begin{aligned}
&\|f \circ \varphi^{-1}(r,\cdot) - s_{f\circ\varphi^{-1}}(r,\cdot)\|_{L^\infty(\partial B)} \\
&= \|\phi_{\mathrm{mult}}(\phi_m(r,\cdot), \mathsf{p}_m(r)) + \phi_{\mathrm{mult}}(\phi_{m+1}(r,\cdot), \mathsf{p}_{m+1}(r)) - f \circ \varphi^{-1}(r,\cdot)\|_{L^\infty(\partial B)} \\
&\le 2 \times 2^{-l_1}\|\phi_m(r,\cdot)\|_{L^\infty(\partial B)} + \|\mathsf{p}_m(r)(\phi_m(r,\cdot) - f_m \circ \varphi^{-1}(r,\cdot))\|_{L^\infty(\partial B)} \\
&\quad + \|\mathsf{p}_{m+1}(r)(\phi_{m+1}(r,\cdot) - f_{m+1} \circ \varphi^{-1}(r,\cdot))\|_{L^\infty(\partial B)} \\
&\lesssim 2^{-l_1+1}\varepsilon^{-1} + \|\phi_m(r,\cdot) - f_m \circ \varphi^{-1}(r,\cdot)\|_{L^\infty(\partial B)} \\
&\quad + \|\phi_{m+1}(r,\cdot) - f_{m+1} \circ \varphi^{-1}(r,\cdot)\|_{L^\infty(\partial B)} \\
&\lesssim \varepsilon^{-1}2^{-l_1} + \begin{cases} N^{-\alpha/(d-1)}\log N, & \text{if } f = h_N \\ \frac{1}{2^{m-2}\varepsilon} N^{-\alpha/(d-1)}\log N, & \text{if } f = \nabla h_N. \end{cases}
\end{aligned}$$

Choosing $l_1 \coloneqq \lceil \log_2(\varepsilon^{-1}) + \log_2(\alpha/(d-1)) \rceil$ and noting that

$$\frac{1}{2^{m-2}\varepsilon} = \frac{4}{1-(1-2^m\varepsilon)} \le \frac{4}{1-r},$$

we finally get the desired approximation error bound.

For the size of $s_{f\circ\varphi^{-1}} \in \mathcal{S}(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})$, it suffices to note that it is a sum of compositions of ϕ_{mult} with a parallelisation of ϕ_m with p_m. By Lemma I.1 (composition of neural networks), Lemma I.2 (parallelisation of neural networks), Lemma I.3 (sum of neural networks), Lemma I.5 (approximation of multiplication) and the size of ϕ_m determined previously, we get

$$\mathsf{L} \lesssim \log N \log\log N, \quad W \lesssim MN\log N, \quad S \lesssim MN(\log N)^2, \quad B \lesssim N^{1/(d-1)} \vee \varepsilon^{-1},$$

with $M \lesssim \log N$. □

Now, all that is left is to approximate the coordinate maps $\varphi_\pm$.

Lemma F.3. *For any $\gamma > 0$, there exist neural networks $s_{\varphi_+} \in \mathcal{S}(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})$ and $s_{\varphi_-} \in \mathcal{S}(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})$ with*

$$\mathsf{L} \lesssim \log^2\gamma^{-1}, \qquad \mathsf{W} = \mathrm{const}, \qquad \mathsf{S} \lesssim \log^2\gamma^{-1}, \qquad \mathsf{B} \lesssim \gamma^{-1},$$

such that

$$\|\varphi_\pm - s_{\varphi_\pm}\|_{L^\infty(B_\pm \setminus B_R)} \le \gamma,$$

with

$$B_\pm \coloneqq \{z \in B_{1-\varepsilon} : \pm z_d \ge -1/2\}.$$

Moreover, $R \le s_{\varphi_\pm}(z)_1 \le 1-\varepsilon$ for $z \in B_\pm / B_R$.

Proof. The construction focuses mainly on the approximation of the radial coordinate r, since the angular coordinates are simply quotients of linear functions and the reciprocal function on $[1,2]$ can be efficiently approximated using Lemma I.7. For the latter, we can define

$$s_{\varphi_\pm}(x)_i \coloneqq \phi_{\mathrm{mult}}(\mp x_i, \phi_{\mathrm{rec}}(1 \pm x_d))$$

for $i = 2, \dots, d$, with ϕ_{mult} given by Lemma I.5 and with size

$$\mathsf{L} \lesssim \log\gamma^{-1} + \log\gamma^{-1}\,\log\log\gamma^{-1}, \quad \mathsf{W} = 1, \quad S \lesssim \log\gamma^{-1} + \log\gamma^{-1}\,\log\log\gamma^{-1}, \quad B = \mathrm{const},$$

for a maximal error of size $|\varphi_\pm(x)_i - s_{\varphi_\pm}(x)_i| \le \gamma$ for $x \in B_\pm \setminus B_R$.

For the approximation of the radial coordinate, we need to approximate $y \mapsto y^2$ and $y \mapsto \sqrt{y}$. The former is a special case of Lemma I.5: there exists a neural network $\phi_\square$ of size $\mathsf{L} \lesssim \log\gamma^{-1}$, $\mathsf{W} = 1$, $\mathsf{S} \lesssim \log\gamma^{-1}$ and $\mathsf{B} = \mathrm{const}$, such that

$$|\phi_\square(x) - x^2| \le \gamma.$$

For the square root, we note that on $[R, 1-\varepsilon]$ it has the Taylor expansion

$$\sqrt{y} = 1 - \sum_{k=1}^{\infty} \underbrace{\frac{(2k-1)!!}{2^k k!}}_{=:a_k} (1-y)^k =: f(y),$$

where

$$(2k-1)!! \coloneqq (2k-1)(2k-3)\cdots 3 \times 1.$$

We now follow a two-step procedure that first approximates the square root function by its partial Taylor sums f_K, $K \in \mathbb{N}$ and in a second step approximate the partial sums by a neural network, i.e.,

$$|\sqrt{y} - s_{\varphi_\pm}(y)| \leq |f(y) - f_K(y)| + |f_K(y) - s_{\varphi_\pm}(y)|. \tag{43}$$

The convergence rate of the series of partial sums is evaluated straightforwardly:

$$\begin{aligned}
|f(y) - f_K(y)| &= \sum_{k=K+1}^{\infty} a_k (1-y)^k \\
&= (1-y)^K \sum_{k=1}^{\infty} a_{k+K} (1-y)^k \\
&\leq (1-R)^K \sum_{k=1}^{\infty} a_k (1-R)^k \\
&= (1-R)^K (1-\sqrt{R}) \leq (1-R)^K.
\end{aligned}$$

Thus, it suffices to choose $K = \log\gamma / \log(1-R)$ to bound the first term in (43) by γ.

For the second term, we can invoke Lemma I.6 to get a neural network $\phi_{\mathrm{sqrt}} \in \mathcal{S}(K(\log\gamma^{-1} + \log K), 9, K(\log\gamma^{-1} + \log K), 1)$ such that

$$|\phi_{\mathrm{sqrt}}(y) - f_K(y)| \leq \gamma.$$

The network approximating the radial coordinate can then be defined as

$$s_{\varphi_\pm}(x)_1 \coloneqq \phi_{\mathrm{sqrt}} \left(1 \wedge \sum_{i=1}^{d} \phi_{\square}(x_i) \vee 0 \right),$$

whose size, according to Lemma I.3, can be chosen of order

$$\begin{aligned}
&\mathsf{L} \lesssim K(\log\gamma^{-1} + \log K) + 2 + \log\gamma^{-1} \lesssim \log^2\gamma^{-1}, && \mathsf{W} \leq 9d, \\
&\mathsf{S} \lesssim K(\log\gamma^{-1} + \log K) + 2 + d + d\,\log\gamma^{-1} \lesssim \log^2\gamma^{-1}, && \mathsf{B} = \mathrm{const}.
\end{aligned}$$

The approximation error of s_φ is then finally evaluated by

$$\begin{aligned}
|r(x) - s_{\varphi_\pm}(x)_1| &\leq \|\sqrt{\cdot} - \phi_{\mathrm{sqrt}}\|_{L^\infty([R,1-\varepsilon])} + \sum_{i=1}^{d} |\phi_{\square}(x_i) - x_i^2| \\
&\leq \gamma + d\gamma,
\end{aligned}$$

where in the first inequality we used the approximation error of ϕ_{sqrt} and in the second one the approximation error for the square function. Setting $(d+1)\gamma \mapsto \gamma$ yields the claim. □

Finally, the neural networks $s_{h_N \circ \varphi_\pm^{-1}}$ and $s_{\nabla h_N \circ \varphi_\pm^{-1}}$ can be concatenated with $s_{\varphi_\pm}$ to approximate the score on the cartesian space. This is the content of Proposition 4.7 and we are now in a position to prove it.

Proof of Proposition 4.7. Some parts of the proof follow the same reasoning for both h_N and $\partial_i h_N$ for any $i = 1, \ldots, d$, which we commonly denote as f. We define the neural networks as

$$\begin{aligned}
s_f(z) \coloneqq (&\phi^{\mathrm{cap}}(s_{\varphi_+}(z)_1) \wedge \phi_{\mathrm{mult}}(p_+(z_d), s_{f\circ\varphi_+^{-1}}(s_{\varphi_+}(z))) \\
&+ \phi_{\mathrm{mult}}(p_-(z_d), s_{f\circ\varphi_-^{-1}}(s_{\varphi_-}(z)))) \vee (-\phi^{\mathrm{cap}}(s_{\varphi_+}(z)_1)),
\end{aligned}$$

with ϕ^{cap} as in Lemma I.8, ϕ_{mult} as in Lemma I.5, $s_{\varphi_\pm}$ as in Lemma F.3 with $\gamma = \varepsilon^4$, and

$$p_\pm(z_d) \coloneqq 1 \wedge (\pm z_d + 1/2) \vee 0.$$

Bounding the values of s_f with ϕ^{cap} ensures that

$$\begin{aligned}
|s_f(z)| \leq 6\frac{d+2}{1-s_{\varphi_+}(z)_1} &\leq 6\frac{d+2}{1-\|z\|} + 6(d+2)\left|\frac{1}{1-s_{\varphi_+}(z)_1} - \frac{1}{1-\|z\|}\right| \\
&= 6\frac{d+2}{1-\|z\|} + 6\frac{|\|z\| - s_{\varphi_+}(z)_1|}{(1-s_{\varphi_+}(z)_1)(1-\|z\|)} \\
&\leq \frac{6(d+2)(1+\gamma\varepsilon^{-1})}{1-\|z\|} \\
&\leq \frac{12(d+2)}{1-\|z\|},
\end{aligned}$$

showing that $s_f \in \mathcal{S}(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})$ with hyperparameters determined at the end of the proof. Concerning the approximation error, since $|f(z)| \leq (d+2)/(1-\|z\|)$ and for ε small enough

$$\phi^{\mathrm{cap}}(s_{\varphi_+}(z)_1) \geq \frac{2(d+2)}{1-s_{\varphi_+}(z)_1} \geq \frac{2(d+2)(1-\varepsilon^{-1}\gamma)}{1-\|z\|} \geq \frac{d+2}{1-\|z\|},$$

cutting off a neural network with ϕ^{cap} does not alter the approximation of f, i.e.,

$$\begin{aligned}
|f(z) - s_f(z)| &\leq |f(z) - \phi_{\mathrm{mult}}(p_+(z_d), s_{f\circ\varphi_+^{-1}}(s_{\varphi_+}(z))) - \phi_{\mathrm{mult}}(p_-(z_d), s_{f\circ\varphi_-^{-1}}(s_{\varphi_-}(z)))| \\
&\leq 2 \times 2^{-l_1}\|s_{f\circ\varphi_\pm^{-1}} \circ s_{\varphi_\pm}\|_{L^\infty(B_{1-\varepsilon})} \\
&\quad + |f(z) - p_+(z_d)s_{f\circ\varphi_+^{-1}}(s_{\varphi_+}(z)) - p_-(z_d)s_{f\circ\varphi_-^{-1}}(s_{\varphi_-}(z))| \\
&\leq 2^{-l_1+1}\varepsilon^{-1} + \underbrace{p_+(z_d)}_{\leq \mathbf{1}_{B_+}}|f(z) - s_{f\circ\varphi_+^{-1}}(s_{\varphi_+}(z))| + \underbrace{p_-(z_d)}_{\leq \mathbf{1}_{B_-}}|f(z) - s_{f\circ\varphi_-^{-1}}(s_{\varphi_-}(z))|.
\end{aligned}$$

By choosing $l_1 \coloneqq \lceil \log \varepsilon^{-1} + \alpha/(d-1)\log N \rceil$, the first term achieves the desired convergence rate. Thus, without loss of generality, it suffices to check the statement for $s_f \coloneqq s_{f\circ\varphi_\pm^{-1}} \circ s_{\varphi_\pm}$ on the sliced balls $B_\pm$. We decompose the errors as

$$\begin{aligned}
&\|s_{\nabla h_N}(h_N - s_{h_N})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_\pm/B_R)} \\
&\leq \|s_{\nabla h_N}(h_N - h_N\circ\varphi_\pm^{-1}\circ s_{\varphi_\pm})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_\pm/B_R)} \\
&\quad + \|s_{\nabla h_N}(h_N\circ\varphi_\pm^{-1}\circ s_{\varphi_\pm} - s_{h_N\circ\varphi_\pm^{-1}}\circ s_{\varphi_\pm})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_\pm/B_R)},
\end{aligned}$$

and

$$\begin{aligned}
&\|(\nabla h_N - s_{\nabla h_N})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_\pm/B_R)} \\
&\leq \|(\nabla h_N - \nabla h_N\circ\varphi_\pm^{-1}\circ s_{\varphi_\pm})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_\pm/B_R)} \\
&\quad + \|(\nabla h_N\circ\varphi_\pm^{-1}\circ s_{\varphi_\pm} - s_{\nabla h_N\circ\varphi_\pm^{-1}}\circ s_{\varphi_\pm})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_\pm/B_R)}.
\end{aligned}$$

For the first terms, we use that $h_N\circ\varphi_\pm^{-1}$ and $\nabla h_N\circ\varphi_\pm^{-1}$ are differentiable on $[R, 1-\varepsilon]\times B_2^{(d-1)}$, and thus locally Lipschitz, to get

$$\begin{aligned}
&\|(\nabla h_N - \nabla h_N\circ\varphi_\pm^{-1}\circ s_{\varphi_\pm})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_\pm/B_R)} \\
&\leq \|\nabla^2 h_N\circ\varphi_\pm^{-1}\|_{L^\infty([R,1-\varepsilon]\times B_2^{(d-1)})}\|(\varphi_\pm - s_{\varphi_\pm})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_\pm\setminus B_R)} \\
&\lesssim \varepsilon^{-2}\gamma,
\end{aligned}$$

and

$$\begin{aligned}
&\|s_{\nabla h_N}(h_N - h_N\circ\varphi_\pm^{-1}\circ s_{\varphi_\pm})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_\pm/B_R)} \\
&\leq \|s_{\nabla h_N}\nabla h_N\circ\varphi_\pm^{-1}\|_{L^\infty([R,1-\varepsilon]\times B_2^{(d-1)})}\|(\varphi_\pm - s_{\varphi_\pm})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_\pm\setminus B_R)} \\
&\lesssim \varepsilon^{-2}\gamma,
\end{aligned}$$

using Lemma F.1 for the bounds on ∇h_N and $\nabla^2 h_N$ and Lemma F.3 in both cases. For the second term of the approximation of h_N, we directly get from Proposition F.2 and Lemma A.5

$$\begin{aligned}
&\|s_{\nabla h_N}(h \circ \varphi_\pm^{-1} \circ s_{\varphi_\pm} - s_{h\circ\varphi_\pm^{-1}} \circ s_{\varphi_\pm})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|^2_{L^2(B_\pm/B_R)} \\
&\leq \int_{B_\pm\setminus B_R} \|s_{\nabla h_N}\|_{L^\infty} \|h(\varphi_\pm^{-1}(s_{\varphi_\pm}(z))) - s_{h\circ\varphi_\pm^{-1}}(s_{\varphi_\pm}(z))\|^2 G_{1-\varepsilon}(z_0,z)\,\mathrm{d}z \\
&\lesssim N^{-2\alpha/(d-1)} \log N \int_{B_\pm\setminus B_R} \frac{1}{(1-\|z\|)^2} G_{1-\varepsilon}(z_0,z)\,\mathrm{d}z \\
&\lesssim N^{-2\alpha/(d-1)} \log N \log \varepsilon^{-1}.
\end{aligned}$$

For the second term of the approximation of ∇h_N, we first apply the same idea:

$$\begin{aligned}
&\|(\nabla h \circ \varphi_\pm^{-1} \circ s_{\varphi_\pm} - s_{\nabla h\circ\varphi_\pm^{-1}} \circ s_{\varphi_\pm})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|^2_{L^2(B_\pm/B_R)} \\
&= \int_{B_\pm\setminus B_R} \|\nabla h(\varphi_\pm^{-1}(s_{\varphi_\pm}(z))) - s_{\nabla h\circ\varphi_\pm^{-1}}(s_{\varphi_\pm}(z))\|^2 G_{1-\varepsilon}(z_0,z)\,\mathrm{d}z \\
&\lesssim N^{-2\alpha/(d-1)} \log N \int_{B_\pm\setminus B_R} \frac{1}{(1-s_{\varphi_\pm}(z)_1)^2} G_{1-\varepsilon}(z_0,z)\,\mathrm{d}z.
\end{aligned}$$

Since $s_{\varphi_\pm}(z)_1$ is an approximation of $\|z\|$, the remaining integral can be controlled again by $\log\varepsilon^{-1}$ plus the approximation error of the network:

$$\begin{aligned}
&\int_{B_\pm\setminus B_R} \frac{1}{(1-s_{\varphi_\pm}(z)_1)^2} G_{1-\varepsilon}(z_0,z)\,\mathrm{d}z \\
&= \int_R^{1-\varepsilon}\int_{\partial B} \frac{1}{(1-s_{\varphi_\pm}(\varphi_\pm^{-1}(r,x))_1)^2} G_{1-\varepsilon}(z_0,(r,x))\,\mathrm{d}\sigma(x) r^{d-1}\,\mathrm{d}r \\
&\leq \int_R^{1-\varepsilon}\int_{\partial B} \frac{1}{(1-r)^2} G_{1-\varepsilon}(z_0,(r,x))\,\mathrm{d}\sigma(x) r^{d-1}\,\mathrm{d}r \\
&\quad + \int_R^{1-\varepsilon}\int_{\partial B} \left|\frac{1}{(1-r)^2} - \frac{1}{(1-s_{\varphi_\pm}(\varphi_\pm^{-1}(r,x))_1)^2}\right| G_{1-\varepsilon}(z_0,(r,x))\,\mathrm{d}\sigma(x) r^{d-1}\,\mathrm{d}r \\
&\lesssim \log\varepsilon^{-1} + \varepsilon^{-4}\int_R^{1-\varepsilon}\int_{\partial B} |(1-s_{\varphi_\pm}(\varphi_\pm^{-1}(r,x))_1)^2 - (1-r)^2|\,\mathrm{d}\sigma(x) r^{d-1}\,\mathrm{d}r \\
&= \log\varepsilon^{-1} + \varepsilon^{-4}\int_R^{1-\varepsilon}\int_{\partial B} |s_{\varphi_\pm}(\varphi_\pm^{-1}(r,x))_1 - r|\,\underbrace{|s_{\varphi_\pm}(\varphi_\pm^{-1}(r,x))_1 + r + 2|}_{\leq 4}\,\mathrm{d}\sigma(x) r^{d-1}\,\mathrm{d}r \\
&\lesssim \log\varepsilon^{-1} + \varepsilon^{-4}\int_{B_\pm\setminus B_R} |s_{\varphi_\pm}(z)_1 - \|z\||\,\mathrm{d}z \\
&\lesssim \log\varepsilon^{-1} + \varepsilon^{-4}\gamma,
\end{aligned}$$

using Lemma F.3 in the last step once again. Choosing $\gamma = \varepsilon^4$ yields the desired approximation error.

The network size evaluates as follows: s_f is a sum of compositions of ϕ_{mult} with $p_\pm$, $s_{f\circ\varphi_\pm^{-1}}$ and $s_{\varphi_\pm}$. For ϕ_{mult}, we choose $l_1 = \lceil \alpha/(d-1)\log_2 N\rceil + 1$, thus the network belongs to the class $\mathcal{S}(\log N, 1, \log N, 1)$ up to constants. Combining the sizes of $s_{f\circ\varphi_\pm^{-1}}$ and $s_{\varphi_\pm}$ from Proposition F.2 and Lemma F.3 using Lemma I.1 (composition of neural networks), we get that $s_{f\circ\varphi_\pm^{-1}} \circ s_{\varphi_\pm} \in \mathcal{S}(\mathsf{L}_\pm, \mathsf{W}_\pm, \mathsf{S}_\pm, \mathsf{B}_\pm)$ with

$$\begin{aligned}
&\mathsf{L}_\pm \lesssim \log N \log\log N + \log^2\varepsilon^{-1}, && \mathsf{W}_\pm \lesssim (N\log^2 N)\vee 1 = N\log^2 N, \\
&\mathsf{S}_\pm \lesssim N\log^3 N + \log^2\varepsilon^{-1}, && \mathsf{B}_\pm \lesssim N^{1/(d-1)} \vee \varepsilon^{-1} \vee \varepsilon^{-4} = N^{1/(d-1)} \vee \varepsilon^{-4}.
\end{aligned}$$

Parallelizing $s_{f\circ\varphi_\pm^{-1}} \circ s_{\varphi_\pm}$ with $p_\pm$ does not alter the asymptotic size of the network, since the latter has constant size, thus, using Lemma I.1 once again and the size of ϕ_{mult}, the total size of

$\phi_{\mathrm{mult}} \circ (p_\pm, s_{f\circ\varphi_\pm^{-1}} \circ s_{\varphi_\pm})$ evaluates as

$$\begin{aligned}
\mathsf{L} &\lesssim \log N + \log N \log\log N + \log^2 \varepsilon^{-1} \leq 2 \log N \log\log N + \log^2 \varepsilon^{-1},\\
\mathsf{W} &\lesssim (N \log^2 N) \vee 1 = N \log^2 N,\\
\mathsf{S} &\lesssim \log N + N \log^3 N + \log^2 \varepsilon^{-1} \leq 2N \log^3 N + \log^2 \varepsilon^{-1},\\
\mathsf{B} &\lesssim N^{1/(d-1)} \vee \varepsilon^{-4} \vee 1 = N^{1/(d-1)} \vee \varepsilon^{-4}.
\end{aligned}$$

Similarly, the composition and parallelization with the remaining networks $\phi^{\mathrm{cap}} \circ s_{\varphi_+}(\cdot)_1$, max and min does not affect the size of s_f, since their sizes are dominated by that of $\phi_{\mathrm{mult}} \circ (p_\pm, s_{f\circ\varphi_\pm^{-1}} \circ s_{\varphi_\pm})$. □

Finally, Theorem 4.3 can be proven.

Proof of Theorem 4.3. Let $R, \widetilde{R}$ be fixed radii such that $\|z_0\| < R < \widetilde{R} < 1 - \varepsilon$. We define the overall network s as

$$s(z) \coloneqq \phi_{\mathrm{mult}}(\overline{p}(s_{r^2}(z)), \overline{s}(z)) + \phi_{\mathrm{mult}}(\underline{p}(s_{r^2}(z)), \underline{s}(z)),$$

with $\underline{s}$ given by Proposition 4.5, $\overline{s}$ defined as in (36), s_{r^2} the approximation of the squared radial coordinate, ϕ_{mult} as in Lemma I.5, and with the neural networks

$$\overline{p}(r) \coloneqq 0 \vee \left(\frac{r - R^2}{\widetilde{R}^2 - R^2}\right) \wedge 1, \qquad \underline{p}(r) \coloneqq 0 \vee \left(\frac{\widetilde{R}^2 - r}{\widetilde{R}^2 - R^2}\right) \wedge 1,$$

which act as a partition of unity. The network s_{r^2} is constructed as

$$s_{r^2}(z) \coloneqq \sum_{i=1}^{d} s_\square(z_i), \quad z \in [-1, 1]^d,$$

with $s_\square$ the neural network approximation of $x \mapsto x^2$ given in Lemma I.5. According to the same lemma and Lemma I.3, $s_{r^2} \in \mathrm{NN}(m + 1, d, dm, C)$ for some $C > 0$ with

$$|s_{r^2}(z) - r^2| \leq dC2^{-m}.$$

We choose $m \coloneqq \lceil \log(dC\varepsilon^{-1}) \rceil$. Then, by Lemma I.5 and Proposition 4.4 (with the separation at $\widetilde{R}$ for $\underline{s}$),

$$\begin{aligned}
&\inf_{s\in\mathcal{S}} \mathbb{E}^{z_0}\left[\int_0^{\tau^h_{1-\varepsilon}} \|s(Z_t^h) - \nabla \log h(Z_t^h)\|^2 \,\mathrm{d}t\right]\\
&\lesssim 2^{-2l_1}\|\overline{s}\|^2_{L^\infty(B_{1-\varepsilon})} + 2^{-2l_1}\|\underline{s}\|^2_{L^\infty(B_{1-\varepsilon})} + \Big\| \underbrace{\overline{p} \circ s_{r^2}}_{\leq \mathbf{1}_{B_{1-2\varepsilon}\setminus B_{R-\varepsilon}}} (\overline{s} - \nabla \log h)\sqrt{G_{1-\varepsilon}(z_0, \cdot)}\Big\|^2_{L^2(B_{1-\varepsilon})}\\
&\qquad + \Big\| \underbrace{\underline{p} \circ s_{r^2}}_{\leq \mathbf{1}_{B_{\widetilde{R}+\varepsilon}}} (\underline{s} - \nabla \log h)\sqrt{G_{1-\varepsilon}(z_0, \cdot)}\Big\|^2_{L^2(B_{1-\varepsilon})}\\
&\lesssim 2^{-2l_1}\varepsilon^{-2} + \Big\|(\overline{s} - \nabla \log h)\sqrt{G_{1-\varepsilon}(z_0, \cdot)}\Big\|^2_{L^2(B_{1-2\varepsilon}\setminus B_{R+\varepsilon})} + \|\underline{s} - \nabla \log h\|^2_{L^\infty(B_{\widetilde{R}+\varepsilon})}. \qquad (44)
\end{aligned}$$

For ε small enough, it still holds $R - \varepsilon > \|z_0\|$ and $\widetilde{R} + \varepsilon < 1 - 2\varepsilon$, whence we can replace $R - \varepsilon \to R$, $\widetilde{R} + \varepsilon \to \widetilde{R}$ and $2\varepsilon \to \varepsilon$ without loss of generality in the following.

The last term in (44) is directly bounded by $N^{-2\alpha/(d-1)}$ using Proposition 4.5, while the first term is bounded by the same rate with the choice $l_1 \coloneqq \lceil -\log\varepsilon + \alpha/(d-1)\log N \rceil$. The first term can be

decomposed according to (37) and the subsequent discussion, yielding

$$\begin{aligned}\left\|(\overline{s}-\nabla\log h)\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\right\|^2_{L^2(B_{1-\varepsilon}\setminus B_R)} &\lesssim 2^{-2l_1}\varepsilon^{-2}(2^{-l_2}+\pi_{\min}^{-1})^2 + 2^{-2l_2}\log^2\varepsilon^{-1}\\ &\quad+\pi_{\min}^{-1}\|\nabla h-\nabla h_N\|_{L^2(B_{1-\varepsilon})}\\ &\quad+\pi_{\min}^{-2}\|\nabla h_N(h-h_N)\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_{1-\varepsilon}\setminus B_R)}\\ &\quad+\pi_{\min}^{-1}\|(\nabla h_N-s_{\nabla h_N})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_{1-\varepsilon}\setminus B_R)}\\ &\quad+\pi_{\min}^{-2}\|s_{\nabla h_N}(h_N-s_{h_N})\sqrt{G_{1-\varepsilon}(z_0,\cdot)}\|_{L^2(B_{1-\varepsilon}\setminus B_R)}.\end{aligned}$$

The previous choice for l_1 and $l_2 \coloneqq \lceil \log\log\varepsilon^{-1} + \alpha/(d-1)\log N\rceil$ yield the rate $\mathcal{O}(N^{-2\alpha/(d-1)})$ for the first two terms. The bounds on the four remaining terms are the result of Lemma 4.6 and Proposition 4.7, yielding the final claim

$$\begin{aligned}\mathbb{E}^{z_0}\Big[\int_0^{\tau^h_{1-\varepsilon}}\|s(Z_t^h)-\nabla\log h(Z_t^h)\|^2\,\mathrm{d}t\Big] &\lesssim N^{-2\alpha/(d-1)}+N^{-2\alpha/(d-1)}+N^{-2\alpha/(d-1)}\\ &\quad+\log\varepsilon^{-1}(1-\varepsilon)^{2N+d-1}N^{-2\alpha}+(1-\varepsilon)^{2N+d}N^{-2\alpha+1}\\ &\quad+N^{-2\alpha/(d-1)}\log N\log\varepsilon^{-1}+N^{-2\alpha/(d-1)}\\ &\lesssim \log\varepsilon^{-1}(1-\varepsilon)^{2N+d-1}N^{-2\alpha}+(1-\varepsilon)^{2N+d}N^{-2\alpha+1}\\ &\quad+N^{-2\alpha/(d-1)}\log N\log\varepsilon^{-1}\\ &\lesssim N^{-2\alpha/(d-1)}\log N\log\varepsilon^{-1}.\end{aligned}$$

For the last step we used that it holds

$$N^{-2\alpha} = N^{-2\alpha/(d-1)}$$

and for the exponent of N of the second summand

$$\begin{aligned}&\frac{2\alpha-1}{\alpha/(d-1)} = 2(d-1)-\frac{d-1}{\alpha} > 2d-4\geq 2,\\ \Longrightarrow\ &N^{-2\alpha+1}\leq N^{-2\alpha/(d-1)},\end{aligned}$$

due to $\alpha > (d-1)/2$ and $d\geq 3$.

The size of s is straightforwardly evaluated with Proposition 4.7, the sizes of s_{r^2}, $\overline{p}$, $\underline{p}$ and Lemma I.5. The sizes of $\overline{p}\circ s_{r^2}$ and $\underline{p}\circ s_{r^2}$ evaluate as

$$\begin{aligned}\mathsf{L}&\lesssim\log\varepsilon^{-1}, & \mathsf{W}&\lesssim 1,\\ \mathsf{S}&\lesssim\log\varepsilon^{-1}, & \mathsf{B}&\lesssim 1.\end{aligned}$$

Thus, the size of the parallelization $(\overline{p}\circ s_{r^2},\overline{s})$ is dominated by the size of $\overline{s}$. The size of $\underline{s}$ is given in Proposition 4.5, such that the parallelization $(\underline{p}\circ s_{r^2},\underline{s})$ has size

$$\begin{aligned}\mathsf{L}&\lesssim\log N+\log\varepsilon^{-1}, & \mathsf{W}&\lesssim N,\\ \mathsf{S}&\lesssim N\log N+\log\varepsilon^{-1}, & \mathsf{B}&\lesssim \mathrm{Poly}(N).\end{aligned}$$

By the choice of l_1, the size of ϕ_{mult} is of the same order as the ones of $\overline{p}\circ s_{r^2}$ and $\underline{p}\circ s_{r^2}$, which are dominated by the size of $(\overline{p}\circ s_{r^2},\overline{s})$ and $(\underline{p}\circ s_{r^2},\underline{s})$. Thus, $\phi_{\mathrm{mult}}(\overline{p}\circ s_{r^2},\overline{s})$ has the same asymptotic size as $\overline{s}$ and $\phi_{\mathrm{mult}}(\underline{p}\circ s_{r^2},\underline{s})$ has the same size as $(\underline{p}\circ s_{r^2},\underline{s})$ derived above. Finally, the sum of $\phi_{\mathrm{mult}}(\overline{p}\circ s_{r^2},\overline{s})$ and $\phi_{\mathrm{mult}}(\underline{p}\circ s_{r^2},\underline{s})$ giving the total network size evaluates as

$$\begin{aligned}\mathsf{L}&\lesssim(\log N\log\log N+\log^2\varepsilon^{-1})\vee(\log N+\log\varepsilon^{-1})\\ &\lesssim\log N\log\log N+\log^2\varepsilon^{-1},\\ \mathsf{S}&\lesssim N\log^3 N+\log^2\varepsilon^{-1}+N\log N+\log\varepsilon^{-1}\\ &\lesssim N\log^3 N+\log^2\varepsilon^{-1},\end{aligned}\qquad\begin{aligned}\mathsf{W}&\lesssim N\log^2 N+N\\ &\lesssim N\log^2 N,\\ \mathsf{B}&\lesssim(N^{1/(d-1)}\vee\varepsilon^{-4})\vee\mathrm{Poly}(N)\\ &\lesssim\mathrm{Poly}(N)\vee\varepsilon^{-4}.\end{aligned}$$

□

G Proof of the main result

Proof of Theorem 3.1. Decomposing the risk as in (9) we get

$$\begin{aligned}\mathbb{E}^{z_0}[\mathrm{TV}(\Pi^*, P_{\partial B}\sharp\widehat{\mathbb{P}}^{\widehat{s}}_{\partial B_{1-\varepsilon}})] &\leq \mathrm{TV}(\Pi^*, P_{\partial B}\sharp\mathbb{Q}^h_{\partial B_{1-\varepsilon}}) + \mathbb{E}[\mathrm{TV}(\mathbb{Q}^h_{\partial B_{1-\varepsilon}}, P^{\widehat{s}}_{\partial B_{1-\varepsilon}})] \\ &\lesssim \varepsilon^{\beta/2} + \mathbb{E}[\mathrm{TV}(\mathbb{Q}^h_{\partial B_{1-\varepsilon}}, P^{\widehat{s}}_{\partial B_{1-\varepsilon}})] \\ &\lesssim \varepsilon^{\beta/2} + \left(\inf_{s\in\mathcal{S}} \mathbb{E}^{z_0}\Big[\int_{\underline{\tau}^h}^{\overline{\tau}^h} \|s(Z^h_t) - \nabla \log h(Z^h_t)\|^2 \,\mathrm{d}t\Big] \right. \\ &\quad \left. + C(\mathcal{L})\frac{\log \mathcal{N}(\mathcal{L}, \|\cdot\|_\infty, \delta)}{n} + \delta \right)^{1/2}, \end{aligned} \tag{45}$$

where we used Proposition 4.1 for the second line and afterwards used Pinsker's inequality combined with Proposition 2.2 and Theorem 4.2.(i). This shows that we must choose $\varepsilon = n^{-2\alpha/(\beta(2\alpha+d-1))}$ and $\delta \coloneqq n^{-2\alpha/(2\alpha+d-1)}$ for the desired convergence rate. For the complexity term, we invoke Theorem 4.2.(iii) to bound $C(\mathcal{L})$ by $\log \varepsilon^{-1}$ (up to constants) as well as Theorem 4.2.(ii) to bound the covering number $\mathcal{N}(\mathcal{L}, \|\cdot\|_\infty, \delta)$ by the covering number $\mathcal{N}(\mathcal{S}, \|\cdot\|_\infty, \delta/\sqrt{C(\mathcal{L})})$, where we choose $\mathcal{S}$ such that it contains the approximation network from Theorem 4.3, thus yielding the order of the network parameters from the statement of the theorem. Since $\mathcal{S}$ is a network class with depth L, width W, sparsity S and maximal parameter value B, its covering number can be bounded with [24, Lemma C.2]. For the given order of the network parameters, this yields

$$\log \mathcal{N}(\mathcal{L}, \|\cdot\|_\infty, \delta) \lesssim \log \mathcal{N}(\mathcal{S}, \|\cdot\|_\infty, c\delta/\sqrt{C(\mathcal{L})}) \lesssim \mathsf{LS}\log(c^{-1}\delta^{-1}\sqrt{C(\mathcal{L})}\mathsf{LWB}). \tag{46}$$

By choosing $N = n^{(d-1)/(2\alpha+d-1)}$ and ε, δ as above, the bound (46) reduces to $n^{(d-1)/(2\alpha+d-1)}\log^6 n$. Combining this with the approximation result Theorem 4.3, we find that (45) is bounded by

$$\begin{aligned}\mathbb{E}[\mathrm{TV}(\Pi^*, P_{\partial B}\sharp\widehat{\mathbb{P}}^{\widehat{s}}_{\partial B_{1-\varepsilon}})] &\lesssim \varepsilon^{\beta/2} + \left(N^{-\frac{2\alpha}{d-1}}\log N \log \varepsilon^{-1} + \log\varepsilon^{-1}\frac{n^{\frac{d-1}{2\alpha+d-1}}\log^6 n}{n} + \delta \right)^{1/2} \\ &\lesssim n^{-\frac{\alpha}{2\alpha+d-1}} + \left(n^{-\frac{2\alpha}{2\alpha+d-1}}(\log n)^2 + n^{-\frac{2\alpha}{2\alpha+d-1}}(\log n)^7 + n^{-\frac{2\alpha}{2\alpha+d-1}} \right)^{1/2} \\ &\lesssim n^{-\frac{\alpha}{2\alpha+d-1}} + n^{-\frac{\alpha}{2\alpha+d-1}}\log n + n^{-\frac{\alpha}{2\alpha+d-1}}(\log n)^{7/2} \\ &\lesssim n^{-\frac{\alpha}{2\alpha+d-1}}(\log n)^{7/2}. \end{aligned}$$

□

H Auxiliary lemmas

Lemma H.1. *Let $\varphi_{\partial B,\pm}$ denote the stereographic coordinate map for the $(d-1)$-dimensional sphere defined in Section F (either northern or southern hemisphere) and $k \in \mathbb{N}_0$. Then, for $u \in H^k(\partial B)$ and any $R > 0$*

$$\|u \circ \varphi^{-1}_{\partial B,\pm}\|_{H^k(B_R^{(d-1)})} \lesssim \|u\|_{H^k(\partial B)},$$

where the hidden constant diverges as $R \to \infty$.

Proof. The main idea relies essentially on [23, Lemma 2.3.1], which uses the elliptic regularity estimate of solutions of the Dirichlet/Neumann problem on a manifold. Let $\varphi_{\partial B} \in \{\varphi_{\partial B,+}, \varphi_{\partial B,-}\}$ and denote by $g = (g^{ij})_{i,j=1,\dots,d-1}$ the metric tensor induced by $\varphi_{\partial B}$. Its determinant is given by ([20] eq. (3.10))

$$|g(\theta)| \coloneqq |\det g(\theta)| = \left(\frac{2}{1+\|\theta\|^2}\right)^{2d}, \quad \theta \in \mathbb{R}^{d-1}.$$

We note that $|g|$ is that is bounded from above by $(2/(1-R))^d$) and from below by 1. Moreover, the (weak) Laplace–Beltrami operator can be expressed in coordinate space as

$$\Delta_{\partial B} = \frac{1}{\sqrt{|g|}} \sum_{i,j=1}^{d} \partial_j (g^{ij} \sqrt{|g|} \partial_i).$$

Now, for any $v \in W_0^{1,2}(B_R^{(d-1)})$ and some open set $U \supset \overline{B_R^{(d-1)}}$, we have the integration by parts formula

$$\begin{aligned}
\int_U (-\Delta_{\partial B})(u \circ \varphi_{\partial B}^{-1}) v \,\mathrm{d}\lambda &= -\sum_{i,j=1}^{d} \int_U \partial_j (g^{ij} \sqrt{|g|} \partial_i (u \circ \varphi_{\partial B}^{-1})) \frac{v}{\sqrt{|g|}} \,\mathrm{d}\lambda \\
&= \sum_{i,j=1}^{d} \int_U g^{ij} \partial_i (u \circ \varphi_{\partial B}^{-1}) \partial_j v - g^{ij} \frac{\sqrt{|g|}}{2\sqrt{|g|^3}} \partial_j |g| \partial_i (u \circ \varphi_{\partial B}^{-1}) v \,\mathrm{d}\lambda \\
&= \sum_{i,j=1}^{d} \int_U g^{ij} \partial_i (u \circ \varphi_{\partial B}^{-1}) \partial_j v - g^{ij} \underbrace{\frac{1}{2|g|} \partial_j |g|}_{=:\Gamma_{ij}} \partial_i (u \circ \varphi_{\partial B}^{-1}) v \,\mathrm{d}\lambda,
\end{aligned}$$

where the Γ_{ij} are the so-called *Christoffel symbols*. This is now a weak formulation for the PDE

$$\sum_{i,j=1}^{d} \partial_j (g^{ij} \partial_i - g^{ij} \Gamma_{ij} \partial_i)(u \circ \varphi_{\partial B}^{-1}) = (-\Delta_{\partial B}) u \circ \varphi_{\partial B}^{-1},$$

with some boundary condition that is not important here. Since the metric tensor is always positive definite by definition, the differential operator in the above PDE is elliptic and we can apply the elliptic interior regularity estimate [12, Theorem 8.20]

$$\|u \circ \varphi_{\partial B}^{-1}\|_{H^k(B_R^{(d-1)})} \lesssim \|(-\Delta_{\partial B}) u \circ \varphi_{\partial B}^{-1}\|_{H^{k-2}(B_R^{(d-1)})} + \|u \circ \varphi_{\partial B}^{-1}\|_{L^2(B_R^{(d-1)})}. \tag{47}$$

And since $\Delta_{\partial B} u \circ \varphi_{\partial B}^{-1} \in H^{k-2}(B_R^{(d-1)})$, we can apply the same argument iteratively to get

$$\begin{aligned}
&\|u \circ \varphi_{\partial B}^{-1}\|_{H^k(B_R^{(d-1)})} \\
&\lesssim \begin{cases} \sum_{l=0}^{k/2} \|(-\Delta_{\partial B})^l u \circ \varphi_{\partial B}^{-1}\|_{L^2(B_R^{(d-1)})}, & k \text{ even} \\ \sum_{l=0}^{(k-3)/2} \|(-\Delta_{\partial B})^l u \circ \varphi_{\partial B}^{-1}\|_{L^2(B_R^{(d-1)})} + \|(-\Delta_{\partial B})^{(k-1)/2} u \circ \varphi_{\partial B}^{-1}\|_{H^1(B_R^{(d-1)})}, & k \text{ odd} \end{cases} \\
&\lesssim \begin{cases} \|(-\Delta_{\partial B})^{k/2} u \circ \varphi_{\partial B}^{-1}\|_{L^2(B_R^{(d-1)})} + \|u \circ \varphi_{\partial B}^{-1}\|_{L^2(B_R^{(d-1)})}, & k \text{ even} \\ \|(-\Delta_{\partial B})^{(k-1)/2} u \circ \varphi_{\partial B}^{-1}\|_{H^1(B_R^{(d-1)})} + \|(-\Delta_{\partial B})^{k/2} u \circ \varphi_{\partial B}^{-1}\|_{L^2(B_R^{(d-1)})} & \\ \qquad + \|u \circ \varphi_{\partial B}^{-1}\|_{L^2(B_R^{(d-1)})}, & k \text{ odd.} \end{cases}
\end{aligned} \tag{48}$$

In the case of k being even, it suffices to use that $1 \leq \sqrt{|g|}$ on $B_R^{(d-1)}$ to get

$$\begin{aligned}
\|u \circ \varphi_{\partial B}^{-1}\|_{H^k(B_R^{(d-1)})} &\lesssim \|(-\Delta_{\partial B})^{k/2} u \circ \varphi_{\partial B}^{-1} |g|^{1/4}\|_{L^2(B_R^{(d-1)})} + \|u \circ \varphi_{\partial B}^{-1} |g|^{1/4}\|_{L^2(B_R^{(d-1)})} \\
&\leq \|(-\Delta_{\partial B})^{k/2} u \circ \varphi_{\partial B}^{-1} |g|^{1/4}\|_{L^2(\mathbb{R}^{d-1})} + \|u \circ \varphi_{\partial B}^{-1} |g|^{1/4}\|_{L^2(\mathbb{R}^{d-1})} \\
&\leq \|u\|_{H^k(\partial B)},
\end{aligned}$$

which was the claim. For odd k, however, there is one derivative left to convert into half a Laplace–Beltrami operator. This follows essentially by $1 \leq \sqrt{|g|}$ once again, the fact that g is uniformly

positive definite on $B_R^{(d-1)}$ and the integration by parts formula for manifolds:

$$\begin{aligned}
\|(-\Delta_{\partial B})^{(k-1)/2} u \circ \varphi_{\partial B}^{-1}\|^2_{H^1(B_R^{(d-1)})} &= \int_{B_R^{(d-1)}} \|\nabla(-\Delta_{\partial B})^{(k-1)/2} u \circ \varphi_{\partial B}^{-1}\|^2 \,\mathrm{d}\lambda \\
&\lesssim \int_{\mathbb{R}^{d-1}} \|g \nabla(-\Delta_{\partial B})^{(k-1)/2} u \circ \varphi_{\partial B}^{-1}\|^2 \sqrt{|g|} \,\mathrm{d}\lambda \\
&= \int_{\partial B} \Big(\nabla_{\partial B}(-\Delta_{\partial B})^{(k-1)/2} u\Big) \Big(\nabla_{\partial B}(-\Delta_{\partial B})^{(k-1)/2} u\Big) \,\mathrm{d}\sigma \\
&= \int_{\partial B} (-\Delta_{\partial B})^{(k-1)/2} u (-\Delta_{\partial B})^{(k+1)/2} u \,\mathrm{d}\sigma \\
&= \|(-\Delta_{\partial B})^{k/2} u\|_{L^2(\partial B)}.
\end{aligned}$$

Inserting the derived bound into (48) and using one more time that $\sqrt{|g|} \geq 1$ to the remaining two terms yields the claim for odd k as well. □

Lemma H.2. *Let $l \in \mathbb{N}$, $m \in \mathbb{N} \cap [1, M_l]$ and let p be a function of the form*

$$p_l(z) \coloneqq \|z\|^l \sum_{m=1}^{M_l} a_{lm} Y_{lm}(z/\|z\|), \qquad z \in \mathbb{R}^d.$$

Then,

$$\langle \nabla p_l, \nabla p_k \rangle_{H^\alpha(\partial B)} = \delta_{lk} l(l+d-1)(1 + ((l-1)(l+d-3))^\alpha) \sum_{m=1}^{M_l} |a_{lm}|^2.$$

Proof. First, we note that

$$\langle \nabla p_l, \nabla p_k \rangle_{H^\alpha(\partial B)} = \langle \nabla p_l, \nabla p_k \rangle_{L^2(\partial B)} + \langle \nabla p_l, \Delta_{\partial B}^\alpha \nabla p_k \rangle_{L^2(\partial B)}.$$

The first term can be evaluated straightforwardly, using $\nabla p_l(x) = \sum_{m=1}^{M_l} a_{lm}(l Y_{lm}(x) x + \nabla_{\partial B} Y_{lm}(x))$:

$$\begin{aligned}
&\langle \nabla p_l, \nabla p_k \rangle_{L^2(\partial B)} \\
&= \int_{\partial B} \nabla p_l(x) \cdot \nabla p_k(x) \,\sigma(\mathrm{d}x) \\
&= \sum_{m=1}^{M_l} \sum_{m'=1}^{M_k} a_{lm} a_{km'} \int_{\partial B} lk Y_{lm}(x) Y_{km'}(x) \|x\|^2 + l \underbrace{\nabla_{\partial B} Y_{lm}(x) \cdot x}_{=0} + k \underbrace{\nabla_{\partial B} Y_{km'}(x) \cdot x}_{=0} \\
&\qquad\qquad + \nabla_{\partial B} Y_{lm}(x) \cdot \nabla_{\partial B} Y_{km'}(x) \,\sigma(\mathrm{d}x) \\
&= \sum_{m=1}^{M_l} \sum_{m'=1}^{M_k} a_{lm} a_{km'} \left(l^2 \delta_{lk} \delta_{mm'} - \int_{\partial B} Y_{lm}(x) \Delta_{\partial B} Y_{km'}(x) \,\sigma(\mathrm{d}x) \right) \\
&= \delta_{lk} \sum_{m=1}^{M_l} |a_{lm}|^2 (l^2 + l(l+d-2)).
\end{aligned} \tag{49}$$

For the second term, the idea consists in relating the Laplace–Beltrami operator of ∇p_l on the sphere to its regular Laplacian on the ambient space, since the latter vanishes:

$$\begin{aligned}
\Delta p_l(z) &= \partial_r^2 p_l(r,x) + \frac{d-1}{r} \partial_r p_l(r,x) + \frac{1}{r^2} \Delta_{\partial B} p_l(r,x) \\
&= \sum_{m=1}^{M_l} a_{lm} [l(l-1) r^{l-2} Y_{lm}(x) + (d-1) l r^{l-2} Y_{lm}(x) + r^{l-2} \Delta_{\partial B} Y_{lm}(x)] \\
&= r^{l-2} [l(l+d-2) - l(l+d-2)] \sum_{m=1}^{M_l} a_{lm} Y_{lm}(x) = 0, \\
\implies \Delta \nabla p_l &= \nabla \Delta p_l = 0.
\end{aligned}$$

Writing the left hand side in spherical coordinates again, we can solve for $\Delta_{\partial B}\nabla p_l$ to get

$$\begin{aligned}
\Delta_{\partial B}\nabla p_l(x) &= -r^2\left(\partial_r^2 + \frac{d-1}{r}\partial_r\right)\nabla p_l(r,x)|_{r=1} \\
&= -r^2\sum_{m=1}^{M_l} a_{lm}\big[lY_{lm}(x)x((l-1)(l-2)r^{l-3} + (d-1)(l-2)r^{l-3}) \\
&\quad + \nabla_{\partial B}Y_{lm}(x)((l-1)(l-2)r^{l-3} + (d-2)(l-1)r^{l-3}\big]|_{r=1} \\
&= -(l-1)(l+d-3)r^{l-1}\sum_{m=1}^{M_l} a_{lm}\big[lY_{lm}(x)x + \nabla_{\partial B}Y_{lm}(x)\big]|_{r=1} \\
&= -(l-1)(l+d-3)\nabla p_l(x).
\end{aligned}$$

In other words, the components of the regular gradient $\partial_{x_i}p_l$ are eigenfunctions of the Laplace–Beltrami operator for the eigenvalue $(l-1)(l+d-3)$. In particular, $\partial_{x_i}p_l$ and $\partial_{x_i}p_k$ for $k \neq l$ are orthogonal, which is consistent with (49). Thus,

$$\begin{aligned}
\langle\nabla p_l, \Delta_{\partial B}^{\alpha}\nabla p_k\rangle_{L^2(\partial B)} &= ((l-1)(l+d-3))^{\alpha}\langle\nabla p_l, \nabla p_k\rangle_{L^2(\partial B)} \\
&= ((l-1)(l+d-3))^{\alpha} l(l+d-1)\delta_{lk},
\end{aligned}$$

which concludes the proof. □

Lemma H.3. *Let $R \in (0,1)$ be such that $R > \|z_0\|$ and $\varepsilon \leq 1-R$. Then there exists a constant $C > 0$ independent of ε such that for $R \leq \|z\| \leq 1$*

$$G_{1-\varepsilon}(z, z_0) \leq C(1 - \|z\|).$$

Proof. Since $d \geq 3$, the Green function $G_{1-\varepsilon}$ is given by [12]

$$G_{1-\varepsilon}(z, z_0) = \frac{1}{d(d-2)\sigma(\partial B)}\left(\frac{1}{\|z - z_0\|^{d-2}} - \frac{1}{\|z - z_0^*\|^{d-2}}\left(\frac{1-\varepsilon}{\|z_0\|}\right)^{d-2}\right), \tag{50}$$

for $z \in \mathbb{R}^d \setminus \{z_0, z_0^*\}$ with $z_0^* \coloneqq (1-\varepsilon)^2 z_0/\|z_0\|^2$. Away from z_0 and z_0^*, the Green kernel is arbitrarily smooth, thence it is bounded and differentiable with bounded derivative on B/B_R. As a function of the radial coordinate $r = \|z\|$, it is differentiable as well with $\partial_r G_{1-\varepsilon}(z, z_0) = \nabla G_{1-\varepsilon}(z, z_0) \cdot z/\|z\|$. By the (one-dimensional) fundamental theorem of calculus, we therefore get for any $x \coloneqq z/\|z\| \in \partial B$

$$\begin{aligned}
G_{1-\varepsilon}(z, z_0) &= \underbrace{G_{1-\varepsilon}((1-\varepsilon)x, z_0)}_{=0} - \int_{\|z\|}^{1-\varepsilon} \partial_r G_{1-\varepsilon}(rx, z_0)\,\mathrm{d}r \\
&\leq \sup_{y \in B/B_R} \|\nabla G_{1-\varepsilon}(y, z_0)\|\,(1 - \varepsilon - \|z\|) \\
&\leq \underbrace{\sup_{y \in B/B_R} \|\nabla G_{1-\varepsilon}(y, z_0)\|}_{=:C_\varepsilon}\,(1 - \|z\|).
\end{aligned}$$

For $\varepsilon \leq 1-R$, the map $\varepsilon \mapsto \|\nabla G_{1-\varepsilon}(y, z_0)\|$ is uniformly bounded as well, meaning

$$C \coloneqq \sup_{\varepsilon \leq 1-R}\sup_{y \in B\setminus B_R} \|\nabla G_{1-\varepsilon}(y, z_0)\| < \infty.$$

□

I Basic results in neural network approximation theory

Here we summarise some important known results on approximation properties of neural networks from the class $\mathrm{NN}(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})$ introduced in (15).

Lemma I.1 (composition of neural networks)**.** *Let $\phi_1 \in \mathrm{NN}(\mathsf{L}_1, \mathsf{W}_1, \mathsf{S}_1, \mathsf{B}_1)$ and $\phi_2 \in \mathrm{NN}(\mathsf{L}_2, \mathsf{W}_2, \mathsf{S}_2, \mathsf{B}_2)$. Then its composition $\phi = \phi_1 \circ \phi_2$ has size*

$$\mathsf{L} \leq \mathsf{L}_1 + \mathsf{L}_2 + 1, \qquad \mathsf{W} = \mathsf{W}_1 \vee \mathsf{W}_2, \qquad \mathsf{S} = \mathsf{S}_1 + \mathsf{S}_2, \qquad \mathsf{B} = \mathsf{B}_1 \vee \mathsf{B}_2.$$

Lemma I.2 (parallelisation of neural networks, [24] Lemma F.3)**.** *Let $\phi_1 \in \mathrm{NN}(\mathsf{L}_1, \mathsf{W}_1, \mathsf{S}_1, \mathsf{B}_1)$ and $\phi_2 \in \mathrm{NN}(\mathsf{L}_2, \mathsf{W}_2, \mathsf{S}_2, \mathsf{B}_2)$. Then the parallelisation $\phi = (\phi_1, \phi_2)$ has size*

$$\mathsf{L} \leq \mathsf{L}_1 \vee \mathsf{L}_2, \qquad \mathsf{W} = 2(\mathsf{W}_1 + \mathsf{W}_2), \qquad \mathsf{S} = 2(\mathsf{S}_1 + \mathsf{L}\mathsf{W}_1 + \mathsf{S}_2 + \mathsf{L}\mathsf{W}_2), \qquad \mathsf{B} = \mathsf{B}_1 \vee \mathsf{B}_2.$$

Lemma I.3 (sum of neural networks, [24] Lemma F.3)**.** *Let $\phi_i \in \mathrm{NN}(\mathsf{L}_\mathsf{i}, \mathsf{W}_\mathsf{i}, \mathsf{S}_\mathsf{i}, \mathsf{B}_\mathsf{i})$, $i = 1, \ldots, m$. Then the sum $\phi = \sum_{i=1}^m \phi_i$ is a neural network with size*

$$\mathsf{L} \leq \max_{i=1,\ldots,m} L_i + 1, \quad \mathsf{W} = 4\sum_{i=1}^m \mathsf{W}_i, \quad \mathsf{S} = 4\sum_{i=1}^m (\mathsf{S}_\mathsf{i} + \mathsf{L}\mathsf{W}_\mathsf{i}) + 2m, \quad \mathsf{B} = \max_{i=1,\ldots,m} \mathsf{B}_i.$$

Lemma I.4 (minimum and maximum function)**.** *The minimum and maximum function can be expressed as ReLU networks in the class $\mathrm{NN}(1, 3, 7, 1)$:*

$$\begin{aligned} \max(x, y) &= y + \mathrm{ReLU}(x - y), \\ &= \mathrm{ReLU}(y) - \mathrm{ReLU}(-y) \\ &\quad + \mathrm{ReLU}(x - y) \end{aligned} \qquad \min(x, y) = x - \mathrm{ReLU}(x - y), \qquad x, y \in \mathbb{R}.$$

Lemma I.5 (approximation of multiplication function, [14] Lemma C.1)**.** *For $l_1 \in \mathbb{N}$ and $C \geq 1$, there exists a neural network $\phi_{\mathrm{mult}} \in \mathrm{NN}(l_1, 1, l_1, C)$ s.t.*

$$|\phi_m^{\mathrm{mult}}(x, y) - xy| \leq C2^{-l_1}, \qquad x \in [0, 1], \quad y \in [-C, C].$$

Lemma I.6 (approximation of monomials, [10] Proposition III.5)**.** *For $\delta > 0$ and $l \in \mathbb{N}_0$, there exists $\phi_{\mathrm{mon}}^{(l)} \in \mathrm{NN}(l(\log \delta^{-1} + \log l), 9, l(\log \delta^{-1} + \log l), 1)$ s.t.*

$$|\phi_{\mathrm{mon}}^{(l)}(r) - r^l| \leq \delta, \qquad r \in [0, 1].$$

Lemma I.7 (approximation of reciprocal function, [14] Lemma C.2)**.** *For $l_2, \underline{k}, \overline{k} \in \mathbb{N}$, there exists a neural network $\phi_{\mathrm{rec}} \in \mathrm{NN}((k + l_2)\log(k + l_2), k, (k + l_2)\log(k + l_2), 2^k)$ with $k = \underline{k} + \overline{k}$ such that*

$$|\phi_{\mathrm{rec}}(x) - x^{-1}| \leq 2^{l_2}, \qquad x \in [2^{-\underline{k}}, 2^{\overline{k}}].$$

As an almost direct consequence of the previous lemma, one can get the following result, which will be needed at the end of the proof on the score approximation.

Lemma I.8 (growth control of neural network)**.** *For $m \in \mathbb{N}$, there exists a neural network $\phi^{\mathrm{cap}} \in \mathrm{NN}((m + 1)\log(m + 1), m + 1, (m + 1)\log(m + 1), 2^{m+1})$ such that*

$$\frac{2(d + 2)}{1 - r} \leq \phi^{\mathrm{cap}}(r) \leq \frac{6(d + 2)}{1 - r}, \qquad r \in [0, 1 - 2^{-m}].$$

Proof. The proof is a simpler version of [14, Lemma 3.12]. Invoking Lemma I.7 with $l_2 = 1$, $\underline{k} = m$, and $\overline{k} = 0$, there exists a neural network $\widetilde{\phi}^{\mathrm{cap}}$ such that

$$\begin{aligned} |\phi_{\mathrm{rec}}(x) - x^{-1}| \leq \frac{1}{2} \leq \frac{1}{2x}, \quad x \in [2^{-m}, 1] \\ \Rightarrow \frac{1}{2x} = \frac{1}{x} - \frac{1}{2x} \leq \widetilde{\phi}^{\mathrm{cap}}(x) \leq \frac{1}{x} + \frac{1}{2x} = \frac{3}{2x}. \end{aligned}$$

Multiplying with $4(d + 2)$, setting $r = 1 - x$, and $\phi^{\mathrm{cap}}(r) \coloneqq 4(d + 2)\widetilde{\phi}^{\mathrm{cap}}(1 - r)$ yields the claim. □

Lemma I.9 ([30] Proposition 1, special case)**.** *Let $\gamma > d/2$ and $f \in H^\gamma([0, 1]^d)$ with $\|f\|_{H^\gamma([0,1]^d)} \leq 1$, $N \in \mathbb{N}$ sufficiently large. Then there exists $\phi_N \in \mathrm{NN}(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})$ with*

$$\mathsf{L} \lesssim \log N, \qquad \mathsf{W} \lesssim N, \qquad \mathsf{S} \lesssim N \log N, \qquad \mathsf{B} = \mathrm{Poly}(N),$$

such that

$$\|f - \phi_N\|_{L^\infty([0,1]^d)} \lesssim N^{-\gamma/d}.$$

J Existence and measurability of the ERM

In this section, we clarify the existence of a measurable minimiser of the empirical risk defined in (13) for the particular neural network class $\mathcal{S}$ under consideration.

Let $X_1, \dots, X_n \sim \Pi^*$ be some random sample. For fixed $\omega \in \Omega$ the existence of a minimiser follows from a reparametrisation of the considered neural network class that allows to reformulate the minimisation problem as a minimisation of the continuous function $\theta \mapsto \widehat{L}(\theta, \omega) = n^{-1} \sum_{i=1}^n L_{\phi_\theta}\big(X_i(\omega)\big)$ over a non-empty compact set $\Theta \subset \mathbb{R}^D$, where D is the total number of network parameters (note here that we consider networks with the same architecture and that continuity in θ follows from continuity of the neural networks with respect to the parametrisation) and ϕ_θ are networks in $\mathcal{S}$.

To see this, it is essential to note that the parameter B ensures that Θ can be written as the intersection of the compact hypercube $[-\mathsf{B}, \mathsf{B}]^D$ with closed sets Θ_1 and Θ_2 that originate from the sparsity constraint and growth restriction (which, in fact, is only necessary to impose for $z \in \overline{B_{1-\varepsilon}}$), respectively. Here, Θ_1 is the finite union of at most S-dimensional (closed) hyperplanes in $\mathbb{R}^D$ and

$$\Theta_2 = \big\{\theta : \sup_{z \in \overline{B_{1-\varepsilon}}} \big(c(1 - \|z\|)^{-1} - \|\phi_\theta(z)\|\big)\big\}$$

is the pre-image of the closed set $[0, \infty)$ under the continuous map

$$\theta \mapsto (c(1 - \|z\|)^{-1} - \|\phi_\theta(z)\|),$$

which is therefore closed as well.

Concerning the measurability, continuity of $\theta \mapsto \widehat{L}(\theta, \omega)$ for every $\omega \in \Omega$ guarantees that the non-empty set of empirical risk minimisers $\arg\min_{\theta \in \Theta} \widehat{L}(\theta, \omega)$ is closed. Furthermore, additionally to the continuity of $\theta \mapsto \widehat{L}(\theta, \omega)$, the mapping $(\theta, \omega) \mapsto \widehat{L}(\theta, \omega)$ is bi-measurable and thus a Carathéodory function. The Kuratowski–Ryll-Nardzewski measurable selection theorem, cf. [1, Theorem 18.13], therefore guarantees the existence of a measurable selection $\omega \mapsto \theta(\omega) \in \arg\min_{\theta \in \Theta} \widehat{L}(\theta, \omega)$ and thus a measurable selection

$$\omega \mapsto \widehat{s}(\omega) \in \arg\min_{s \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^n L_s(X_i(\omega)),$$

where $\widehat{s} = \phi_{\widehat{\theta}}$.

K Some standard notation used in the paper

For $\alpha \in \mathbb{N}$ and any open domain $D \in \mathbb{R}^d$, we let $H^\alpha(D)$ be the Sobolev spaces of α-times weakly differentiable functions with square-integrable weak derivatives. We denote by $\|\cdot\|_{L^p(D)}$, $p \in [1, \infty]$ the usual functional L^p-norms and let $\langle \cdot, \cdot \rangle$ be the standard scalar product in $L^2(D)$. We write $\|\cdot\|$ for the Euclidean and $\|\cdot\|_\infty$ for the maximum norm, and let $\|\cdot\|_0$ be the number of non-zero values. Moreover, for $f \in H^\alpha(D)$, $\partial_i f$ denotes the weak derivative of f wrt the ith component. $\mathrm{TV}(\mu, \nu) \coloneqq \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|$ denotes the total variation distance between two probability measures μ, ν on a measurable space $(\Omega, \mathcal{F})$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main contribution is stated both in the abstract and, in more detail, in the introduction. The claims are a reformulation of the main theorem presented in Section 2.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our result only holds under certain assumptions (in other words: the convergence guarantee is limited to settings where the assumptions are met), which are clearly stated in the introduction and in the results section together with the reason why they are needed. A critical discussion of associated theoretical limitations that are grounds for future work is given in the conclsuion section. Questions regarding limitations to the performance of the model are not within the scope of our work, being a theoretic investigation under rather general assumptions on the data distribution.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: General assumptions, such as the ambient dimension and properties of the data distribution, are stated on the introduction and just before the main theorem. Additional assumptions needed for particular theorems and lemmata are always stated together with the statement, while detailed proofs are deferred to the supplementary material in order to meet the page limit.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: Our work does not involve specific data and code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Most guidelines regard experimental procedures, handling of data, models and code, which do not apply to our research. Potential impacts on society are discussed in the following question.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our contribution is a theoretical convergence guarantee for First-Hitting diffusion models and we thus see no societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not release any data or models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The class of generative models that we investigate in this paper was first introduced by Ye et. al, which we cite in our work.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not provide any assets that need documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.