DIFFUSION MODELS WITH HEAVY-TAILED TARGETS: SCORE ESTIMATION AND SAMPLING GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

Score-based diffusion models have emerged as a powerful framework for generative modeling, with score estimation as a central statistical bottleneck. However, existing theoretical guarantees largely focus on light-tailed distributions or impose restrictive assumptions such as compact support, limiting their applicability to heavy-tailed data encountered in practice. In this work, we study standard score-based diffusion models for heavy-tailed distributions in a Sobolev class with smoothness $\beta > 0$, covering both exponential and polynomial tail decay characterized by a tail parameter γ. Based on kernel density estimation, we derive sharp minimax rates for score estimation, revealing a qualitative distinction between the two regimes: under exponential tails, the estimation rate matches the light-tailed benchmark up to polylogarithmic factors, whereas under polynomial tails it explicitly depends on the tail parameter γ. We further establish the convergence guarantees for the associated continuous-time reverse diffusion dynamics. For exponential tails and $\beta \in (0, 2]$, the generated distribution converges in total variation at the rate $n^{-\beta/(2\beta+d)}$ up to logarithmic factors, matching the light-tailed benchmark. For polynomial tails, we obtain a γ-dependent total variation bound, while leaving its optimality open. These results characterize the statistical limits of score estimation and the resulting sampling accuracy for heavy-tailed targets, extending diffusion theory beyond the light-tailed setting.

1 INTRODUCTION

Diffusion models have become a central tool in modern generative modeling, with impressive performance in image generation (Dhariwal & Nichol, 2021; Ho et al., 2020; Ramesh et al., 2022; Song et al., 2020a), natural language processing (Popov et al., 2021), and computational biology (Anand & Achim, 2022; Xu et al., 2022). A commonly used formulation is the score-based diffusion model (SDM), which is implemented through stochastic differential equations (Song et al., 2020b). At a high level, SDMs consist of a forward process that gradually perturbs data drawn from the target distribution into noise, and a reverse process that aims to reconstruct the target distribution from noise. The reverse dynamics depend on the unknown score function, which specifies the local direction of denoising at each noise level. Consequently, score estimation controls the accuracy of the reverse diffusion process and is widely regarded as the core statistical bottleneck in SDMs.

Recently, a substantial line of research has investigated how accurately the score function can be learned in diffusion models. Early work Block et al. (2020) establishes score estimation guarantees in ℓ_2-norm, though the bounds depend on Rademacher complexities of unknown hypothesis classes. Later, Gupta et al. (2024) shows exponentially improved sample complexity using sufficiently expressive neural network architectures. More recently, Cole & Lu (2024) proves that for sub-Gaussian targets, the score can be locally approximated by a neural network with uniformly bounded parameters, mitigating certain high-dimensional issues, while Han et al. (2024) derives explicit sample complexity guarantees via neural tangent kernel tools. Complementing these learnability results, Chewi et al. (2025) reduces DDPM (Ho et al., 2020) score estimation to classical distribution learning and establishes computational lower bounds, including cryptographic hardness for learning scores of general Gaussian mixture models.

Other works have explored score estimation under additional structural assumptions on the target distribution. For instance, Gatmiry et al. (2024) and Chen et al. (2024) show that when the data

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

follow a Gaussian mixture distribution, the score function can be learned efficiently using piecewise low-degree polynomial approximations. Another line of research considers data supported on low-dimensional structures. In this vein, Chen et al. (2023b) shows that neural networks can accurately approximate the score when the data lie in an unknown linear subspace, while Azangulov et al. (2024) establishes accurate score approximation for data supported on a bounded low-dimensional manifold via suitably designed neural networks. More recently, Yakovlev & Puchkin (2025) and Yakovlev et al. (2025) study denoising and implicit score matching under relaxed manifold assumptions and derive nonasymptotic approximation and generalization bounds with rates governed by the intrinsic dimension.

A separate line of work develops minimax convergence rates for diffusion models, characterizing the best achievable sampling error as a function of the sample size and ambient dimension. However, existing analyses rely on strong regularity assumptions on the target distribution, such as sub-Gaussian tails (Zhang et al., 2024; Stéphanovitch et al., 2025; Cai & Li, 2025; Dou et al., 2024; Fu et al., 2024), explicit lower bounds on the density (Oko et al., 2023; Fan et al., 2025), or compact support (Stéphanovitch et al., 2025; Dou et al., 2024; Tang & Yang, 2024). Such conditions are restrictive and often fail for heavy-tailed distributions arising in practice, for example, in finance (Harvey, 2013), signal and image processing (Achim et al., 2003; Briassouli et al., 2005), and various fields involving measurement errors with non-negligible tail mass (Adler et al., 1998). As a result, current minimax theories do not apply to heavy-tailed targets and leave open fundamental questions about the behavior and limits of standard SDMs in this setting[1].

In this work, we address this gap by studying score estimation and sampling for SDMs with heavy-tailed targets under the conventional Gaussian forward process. This choice is natural because Gaussian noising is the canonical baseline in SDMs and yields the explicit smoothing representation. At the same time, Gaussian convolution regularizes the density without washing out heavy-tail effects, making it possible to characterize when the standard pipeline remains statistically robust and when tail behavior creates intrinsic difficulties. Since the score function drives the reverse dynamics, its estimation complexity is central to sampling guarantees. We therefore study the following questions:

a) Under the standard score-based diffusion pipeline with Gaussian noising, what are the minimax optimal rates for estimating the score when the underlying data distribution is heavy-tailed?

b) How do these score estimation rates translate into bounds on the sampling error between the generated and target distributions?

We provide explicit answers to both questions. Specifically, we consider densities in a Sobolev class with smoothness parameter $\beta > 0$ and analyze two representative tail regimes: (i) exponential tail decay, which includes the sub-Gaussian case, and (ii) polynomial tail decay, corresponding to heavier-tailed distributions. Both regimes are indexed by a tail parameter γ.

For each regime, we construct kernel-based score estimators and show that they attain the corresponding minimax optimal rates for score estimation. We further analyze the associated continuous reverse diffusion process and derive convergence guarantees for the resulting sampling distribution.

Towards that, we make the following contributions.

- We analyze the Gaussian forward diffusion and derive a nearly minimax optimal rate for score estimation under polynomial tails. Specifically, when the target density has polynomial tails with index γ, we show in Theorem 1 that a kernel-based score estimator achieves the rate $n^{-\frac{\gamma+1}{d+\gamma+1}}\, t^{-1-\frac{d}{2}\frac{\gamma+1}{d+\gamma+1}}$, matching the minimax lower bound in Theorem 9 up to polylogarithmic factors over its stated time window and for the fixed compatible class.
- We further study the associated continuous reverse diffusion driven by the learned score. In Theorem 4, we obtain, to our knowledge, the first total variation sampling error upper bound in the polynomial tail regime, of order $\mathcal{O}\Big(\operatorname{polylog}(n) n^{-\frac{2\beta(\gamma+1)}{4\beta(d+\gamma+1)+d(d+2(\gamma+1))}}\Big)$.
- For exponentially decaying targets, we obtain sharp score-estimation rates and sampling guarantees matching the light-tailed benchmark for $\beta \in (0, 2]$. Under the Gaussian forward

[1] Due to space constraints, we defer detailed comparisons with closely related papers and additional theoretical background to Appendix A.

process, Theorem 5 shows that score estimation achieves the rate $n^{-1}t^{-1-d/2}$, matching the minimax lower bound in Theorem 10 up to logarithmic factors over its stated time window and for the fixed compatible class. Moreover, Theorem 8 shows that, for $\beta \in (0, 2]$, the continuous-time reverse diffusion driven by the learned score attains the rate $n^{-\beta/(2\beta+d)}$ in total variation, matching known guarantees for sub-Gaussian targets up to polylogarithmic factors.

Taken together, these results provide a unified statistical perspective on score estimation and the induced sampling error for standard SDMs under heavy-tailed targets. They reveal a sharp dichotomy between exponential and polynomial decay: with exponential tails, the conventional Gaussian diffusion framework remains statistically robust and achieves rates consistent with the sub-Gaussian regime. In contrast, under polynomial tails, the minimax score estimation rate depends explicitly on the tail index, and this tail-dependent penalty propagates to our sampling upper bound, indicating that heavier tails make the problem statistically harder.

Notation. Let $L^1(\mathbb{R}^d)$ be the integrable functions on the d-dimensional Euclidean space $\mathbb{R}^d$. For $p \geqslant 1$, define $\|f\|_{L^p} := \left(\int_{\mathbb{R}^d} |f(x)|^p \, \mathrm{d}x \right)^{1/p}$, and $\|f\|_{L^p(q)} := \left(\int_{\mathbb{R}^d} |f(x)|^p q(x) \, \mathrm{d}x \right)^{1/p}$, where $q : \mathbb{R}^d \to [0, \infty)$ is measurable. Given a random vector X, $\mathcal{L}(X)$ denotes its law and $\|X\|_{\mathbb{L}_2} := (\mathbb{E}\|X\|^2)^{1/2}$ (Euclidean norm $\|\cdot\|$). Write $\mathrm{polylog}(n) := (\log n)^C$ for some $C > 0$. For $f, g : \mathbb{R}^d \to \mathbb{R}$, $(f * g)(x) := \int_{\mathbb{R}^d} f(x - y)g(y) \, \mathrm{d}y$, and the Fourier transform for f is defined via $\mathcal{F}[f](\omega) := \int_{\mathbb{R}^d} f(x) \, e^{-i\langle \omega, x \rangle} \, \mathrm{d}x$. $\mathrm{TV}(p, q)$ and $\mathrm{KL}(p\|q)$ denote total variation and Kullback-Leibler divergence between probability measures p and q, respectively.

2 Preliminaries

In this section, we introduce the score-based diffusion framework considered in this work, state the tail decay and regularity assumptions, and present the proposed score estimator.

2.1 Score-Based Diffusion Models

Framework. We consider the forward process

$$\mathrm{d}X_t = f(X_t, t) \, \mathrm{d}t + g(X_t, t) \, \mathrm{d}B_t \,,$$

where the initial point $X_0 \sim p_0$ follows the target distribution, and B_t denotes the standard d–dimensional Brownian motion. Here, the drift $f : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d$ and the function $g : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^{d \times d}$ are diffusion parameters (see Tang & Zhao (2024) for common choices). For clarity, we take $f(X_t, t) = 0$ and $g(X_t, t) = I_d$, so the forward process reduces to a standard Brownian motion

$$\mathrm{d}X_t = \mathrm{d}B_t \,. \tag{1}$$

The solution $(X_t)_{t \in [0,T]}$ admits a closed-form conditional distribution given X_0, namely

$$X_t = X_0 + \sqrt{t} Z, \quad Z \sim \mathcal{N}(0, I_d) \,.$$

Consequently, for sufficiently large T, Gaussian perturbation makes p_T approximately Gaussian when p_0 is light-tailed. When p_0 is heavy-tailed, however, Gaussian convolution smooths the distribution without eliminating its heavy tails, so the approximation is accurate mainly in the high-probability region while tail deviations persist. Then, diffusion models generate new data by reversing the SDE (1), which leads to the following backward SDE

$$\mathrm{d}X_t^{\leftarrow} = \nabla \log p_{T-t}(X_t^{\leftarrow}) \, \mathrm{d}t + \mathrm{d}W_t \,, \tag{2}$$

where $X_0^{\leftarrow} \sim p_T$, and the term $\nabla \log p_t$, referred to as the *score function* for p_t. Additionally, W_t denotes another standard Brownian motion independent of B_t. Under mild conditions, when initialized at $X_0^{\leftarrow} \sim p_T$, the backward process $\{X_t^{\leftarrow}\}_{0 \leqslant t \leqslant T}$ has the same distribution as the forward process $\{X_{T-t}\}_{0 \leqslant t \leqslant T}$ (Anderson, 1982; Cattiaux et al., 2023). As a result, running the reverse diffusion $X_t^{\leftarrow}$ from $t = 0$ to T will generate a sample from the target p_0. Since p_T is unknown, we approximate it by $\hat{p}_T = \mathcal{N}(0, TI_d)$ [2]. Therefore, we derive a reverse diffusion process defined via

$$\mathrm{d}Y_t = \nabla \log p_{T-t}(Y_t) \, \mathrm{d}t + \mathrm{d}W_t, \quad Y_0 \sim \hat{p}_T \,.$$

[2]We note that, even when p_0 is heavy-tailed, initializing with $\hat{p}_T = \mathcal{N}(0, TI_d)$ is accurate in the bulk for large T. Any remaining discrepancy in the far tails can be viewed as a terminal initialization error.

Score matching. Another challenge in working with (2) is that $\nabla \log p_t$ is unknown since p_t is intractable. In practice, it is approximated from data by learning a surrogate s_θ via the score-matching objective

$$\underset{\theta \in \Theta}{\text{minimize}} \quad \mathbb{E}[\|s_\theta(t, X_t) - \nabla \log p_t(X_t)\|^2] \,,$$

where $\{s_\theta : \theta \in \Theta\}$ is a sufficiently expressive function class. This objective can be minimized using either parametric estimators based on neural networks or nonparametric methods such as kernel score estimators. While neural networks dominate practical implementations, our goal is to characterize the fundamental statistical limits of score estimation rather than optimize a particular parameterization. Kernel-based estimators offer a natural nonparametric framework for this purpose due to their analytical tractability. We therefore adopt a kernel-based score estimator in our analysis.

Substituting the learned score estimate $\hat{s}$ into the backward process (2) yields the practical continuous-time backward SDE

$$\mathrm{d}\hat{Y}_t = \hat{s}_{T-t}(\hat{Y}_t)\,\mathrm{d}t + \mathrm{d}W_t \,, \quad \hat{Y}_0 \sim \hat{p}_T \tag{3}$$

Our analysis focuses on score estimation and on bounding the total variation distance between the distribution of the generated sample $\hat{Y}_t$ and the target p_0. The discretization of this continuous reverse SDE is beyond the scope of this work.

2.2 Tail Decay and Regularity Conditions

To analyze score estimation and induced sampling error in the heavy-tailed regime, we assume p_0 has either exponential or polynomial tail decay and belongs to a Sobolev smoothness class, as detailed below.

Assumption 1 (Exponential Tail Decay)**.** *The density p_0 satisfies the following decay condition for constants $c_1, C_1 > 0$*

$$p_0(x) \leqslant C_1 \exp(-c_1 \|x\|^\gamma), \quad \gamma \in (0, \infty).$$

The parameter γ controls tail decay. When $\gamma = 2$, p_0 is dominated by a Gaussian-type tail with all moments finite. The case $\gamma = 1$ corresponds to Laplace-type distributions, which retain exponential decay but have heavier tails than the Gaussian. For $\gamma \in (0, 1)$, the decay is slower than exponential, leading to stretched-exponential tails.

Assumption 2 (Polynomial Tail Decay)**.** *The target distribution p_0 satisfies the following polynomial decay condition: there exists a constant $C_2 > 0$ such that*

$$p_0(x) \leqslant \frac{C_2}{(1 + \|x\|^2)^{(1+\gamma+d)/2}}, \quad \gamma \in (0, \infty).$$

The polynomial decay condition covers many practical heavy-tailed targets, including multivariate Student's t distributions, Pareto family, and mixtures with heavy-tailed components. This envelope implies, for $R \geqslant 1$, $\mathbb{P}(\|X\| > R) \lesssim R^{-(1+\gamma)}$ for $X \sim p_0$, so γ directly controls tail heaviness, following the standard density-to-tail convention in regular variation (de Haan & Resnick, 1987; Bobkov & Ledoux, 2009). In this regime, variance or higher moments may be infinite, making it broader and more challenging than exponential tails. Beyond tail behavior, we impose mild Sobolev smoothness (Tsybakov, 2009) on the target p_0.

Assumption 3 (Sobolev Regularity)**.** *The true distribution p_0 belongs to the Sobolev class $\mathcal{P}_\mathcal{S}(\beta, L)$ with smoothness parameter $\beta > 0$. Specifically, for $\beta, L \in \mathbb{R}_+$,*

$$\mathcal{P}_\mathcal{S}(\beta, L) = \Big\{ p \in L^1(\mathbb{R}^d) \;\Big|\; p \geqslant 0, \int_{\mathbb{R}^d} p(x)\,\mathrm{d}x = 1, \forall \alpha \in [0,\infty)^d \text{ with } \sum_{i=1}^d \alpha_i = \beta,$$
$$\text{where } |\omega^\alpha| := \prod_{j=1}^d |\omega_j|^{\alpha_j}, \int_{\mathbb{R}^d} |\omega^\alpha|^2 \, |\mathcal{F}[p](\omega)|^2 \,\mathrm{d}\omega \leqslant (2\pi)^d L^2 \Big\}.$$

This condition is equivalent, up to constants depending only on d and β, to the usual radial Fourier characterization of Sobolev smoothness (see Appendix C.1). Such integrated smoothness is standard

in nonparametric minimax theory and is well suited to L_2 analyses of kernel estimators (Zhang et al., 2024). In contrast, many theoretical works on SDMs assume Hölder regularity (Stéphanovitch et al., 2025; Fan et al., 2025; Cai & Li, 2025; Dou et al., 2024), which provides pointwise control of derivatives. More broadly, Oko et al. (2023) adopts Besov smoothness, a more general assumption that includes Sobolev and Hölder classes as special cases.

2.3 Kernel-Based Score Estimation

To obtain a statistically tractable estimator of the score $s_t(x) = \nabla \log p_t(x)$, we use a kernel-based nonparametric construction. Given n i.i.d. samples $\{X^{(i)}\}_{i=1}^n$ from p_0, we estimate the marginal density $p_t = \varphi_t * p_0$ using the Gaussian kernel φ_t and form a plug-in estimator of $\nabla \log p_t$, with an additional threshold to stabilize the estimate in low-density regions.

Gaussian KDE and score estimator. We define the Gaussian kernel density estimator (KDE) of p_t and its gradient by

$$\hat{p}_t(x) = \frac{1}{n}\sum_{i=1}^{n} \varphi_t\Big(x - X^{(i)}\Big), \qquad \nabla\hat{p}_t(x) = \frac{1}{nt}\sum_{i=1}^{n} \Big(X^{(i)} - x\Big)\, \varphi_t\Big(x - X^{(i)}\Big),$$

where $\varphi_t(x) = \frac{1}{(2\pi t)^{d/2}} \exp\big(-\|x\|^2/2t\big)$ is the Gaussian transition density. In contrast to standard KDE, the bandwidth is not a tuning parameter here; instead, the diffusion time t naturally serves as the bandwidth.

We then define the thresholded score estimator

$$\hat{s}_t(x) := \frac{\nabla\hat{p}_t(x)}{\hat{p}_t(x)} \mathbb{1}_{\{\hat{p}_t(x) \geqslant \rho_n\}}(x), \qquad \rho_n := \frac{\log n}{n(2\pi t)^{d/2}},$$

so that the denominator is bounded away from zero on the region where the ratio is evaluated.

3 Main Results: Convergence Analysis

In this section, we establish theoretical guarantees for the proposed score estimators and the resulting sampling procedure under the standard diffusion framework described above, assuming the heavy-tailed target density p_0. We first consider the polynomial-tail setting and then turn to the exponential-tail setting.

Due to space constraints, we provide proof sketches for the sampling error bounds in Theorems 4 and 8 in Appendix B, with the full proofs deferred to Appendices D and E. These sketches highlight the key ideas behind the corresponding score estimation bounds, the main technical challenges, and the ways in which our setting and arguments differ from prior work.

3.1 Polynomial Decay

In this part, we first analyze the score estimation error, and then derive a sampling error bound for the distribution generated by the reverse SDE (3) driven by the proposed score estimator, assuming p_0 has polynomially decaying tails.

We now state our main statistical guarantee for score estimation at diffusion time $t \in (0, T]$. The following theorem provides a mean squared error bound for the proposed estimator $\hat{s}_t$.

Theorem 1. *Suppose that p_0 satisfies Assumption 2 and 3. Let $\kappa = (\gamma+1)/(d+\gamma+1)$ and $\eta = d\kappa/2$. Then, for any $t \in (0, T]$ and $n \geqslant 3$,*

$$\mathbb{E}\big[\|\hat{s}_t - s_t\|^2_{L^2(p_t)}\big] \lesssim (\log n)^{1+\kappa} n^{-\kappa} \begin{cases} t^{-1-\eta}, & 0 < t \leqslant 1, \\ t^{-1}, & t \geqslant 1. \end{cases}$$

The implicit constant depends only on d, γ, and the polynomial-tail envelope constant.

This MSE bound degrades as t decreases, reflecting the fact that score estimation becomes statistically ill-conditioned at very small noise levels. For downstream sampling guarantees, however, the relevant

quantity is not the pointwise error at a fixed t, but its accumulation over time along the diffusion trajectory. We therefore introduce a cutoff $t_0 \in (0, T)$ and bound the integrated error over $[t_0, T]$ in the next theorem.

Theorem 2. *Suppose that p_0 satisfies Assumption 2 and 3. Then, for any $0 < t_0 < T$ and $n \geqslant 3$,*

$$\int_{t_0}^{T} \mathbb{E}\left[\|\hat{s}_t - s_t\|^2_{L^2(p_t)}\right] \mathrm{d}t \lesssim (\log n)^{1+\kappa} n^{-\kappa} \left[\frac{(t_0 \wedge 1)^{-\eta} - (T \wedge 1)^{-\eta}}{\eta} + \log \frac{T \vee 1}{t_0 \vee 1}\right],$$

In particular, if $0 < t_0 \leqslant 1 \leqslant T$,

$$\int_{t_0}^{T} \mathbb{E}\left[\|\hat{s}_t - s_t\|^2_{L^2(p_t)}\right] \mathrm{d}t \lesssim (\log n)^{1+\kappa} n^{-\kappa} \big(t_0^{-\eta} + \log T\big).$$

Existing score estimation analyses for SDMs typically assume light tails (e.g., sub-Gaussian or compact support). Our bounds interpolate between heavy and light tails via the tail index γ: as γ increases, the polynomial tail condition becomes more restrictive, and the tail dependent penalty disappears since

$$\lim_{\gamma \to \infty} \frac{\gamma + 1}{d + \gamma + 1} = 1.$$

Consequently, the sample size dependence in Theorem 2 approaches n^{-1}, matching known sub-Gaussian case (Zhang et al., 2024).

We also note that the dependence on t_0 is unavoidable: as $t \to 0$, the marginal p_t becomes increasingly irregular in the tails, so the plug-in score estimator $\hat{s}_t$ suffers variance inflation because few samples fall in low-density regions. We therefore treat $t_0 > 0$ as a regularization cutoff at the cost of an early-stopping bias: since the reverse SDE is only run back to time t_0, its endpoint targets $p_{t_0} = p_0 * \varphi_{t_0}$ rather than p_0. The next theorem quantifies this smoothing bias in total variation distance.

Theorem 3. *Under Assumptions 2 and 3 with Sobolev smoothness $\beta \in (0, 2]$, and define $p_{t_0} = p_0 * \varphi_{t_0}$. Then, the early stopping bias is bounded in total variation as*

$$\mathrm{TV}(p_0, p_{t_0}) \lesssim t_0^{\frac{\beta(\gamma+1)}{d+2(\gamma+1)}}.$$

This shows that early stopping at time t_0 incurs a controllable approximation error $\mathrm{TV}(p_0, p_{t_0})$, which vanishes as $t_0 \to 0$ at a rate governed by the Sobolev smoothness β and the tail index γ. Combined with Theorem 2, this yields a standard bias–variance tradeoff: smaller t_0 reduces bias but increases the score estimation error via $t_0^{-\frac{d(\gamma+1)}{2(d+\gamma+1)}}$. This tradeoff motivates selecting t_0 to balance the two terms in the final sampling bound as follows.

Theorem 4. *Under the assumptions stated in Theorem 3, set*

$$t_0 \;=\; n^{-\frac{2(d+2(\gamma+1))}{4\beta(d+\gamma+1)+d(d+2(\gamma+1))}} \quad \textit{and} \quad T = n^{\frac{4\beta(\gamma+1)}{4\beta(d+\gamma+1)+d(d+2(\gamma+1))}}.$$

Then, there exists a score estimator $(\hat{s}_t)_{t\in[0,T]}$ such that the sample $\hat{Y}_{T-t_0}$ from the corresponding reverse time diffusion process (3) *satisfies*

$$\mathbb{E}\left[\mathrm{TV}\left(p_0, \mathcal{L}(\hat{Y}_{T-t_0})\right)\right] \;\lesssim\; \mathrm{polylog}(n)\, n^{-\frac{2\beta(\gamma+1)}{4\beta(d+\gamma+1)+d(d+2(\gamma+1))}}.$$

As $\gamma \to \infty$, the tails become progressively lighter and the tail-dependent penalty vanishes. In this limit, the total variation rate reduces to $n^{-\beta/(2\beta+d)}$, recovering the sub-Gaussian case studied in Zhang et al. (2024) and the classical minimax rate for nonparametric density estimation (Tsybakov, 2009).

Alternative low-noise strategy. In Appendix G, we study a decoupled kernel score estimator under stronger regularity and score-moment assumptions. Its score MSE bound is uniform as $t \downarrow 0$, and combining it with the Gaussian-KDE estimator yields a cumulative score-error bound for a hybrid construction.

3.2 EXPONENTIAL DECAY

In this section, we study score estimation and sampling under exponentially decaying targets. As in the polynomial-tail regime, we first derive a pointwise mean-squared error bound. The effective tail exponent after Gaussian convolution is $\min\{\gamma, 2\}$.

Theorem 5. *Suppose Assumption 1 holds. For any $t \in (0, T]$ and $n \geqslant 3$,*

$$\mathbb{E}\big[\|\hat{s}_t - s_t\|^2_{L^2(p_t)}\big] \lesssim \frac{(\log n)^{2+d/\min\{\gamma,2\}}}{nt(t \wedge 1)^{d/2}}.$$

Equivalently, the time dependence is $t^{-1-d/2}$ for $t \leq 1$ and t^{-1} for $t \geq 1$. The implicit constant depends only on d, γ, c_1, and C_1.

Comparing this with Theorem 1, the most striking difference is the dependence on the sample size n. Under polynomial tails, the rate $n^{-\frac{\gamma+1}{d+\gamma+1}}$ degrades as tails get heavier (smaller γ), whereas under exponential tails it improves to the near-parametric rate n^{-1} up to logarithmic factors regardless of the value of γ. This suggests that the standard diffusion pipeline is statistically robust across exponentially decaying targets. The next theorem bounds the cumulative score estimation error.

Theorem 6. *Suppose that p_0 satisfies Assumption 1 and 3, for any $T > 0$ and any $0 < t_0 < T$ with $n \geqslant 3$,*

$$\int_{t_0}^{T} \mathbb{E}\big[\|\hat{s}_t - s_t\|^2_{L^2(p_t)}\big]\,\mathrm{d}t \lesssim \frac{(\log n)^{2+d/\min\{\gamma,2\}}}{n}\left[\frac{2}{d}\big[(t_0 \wedge 1)^{-d/2} - (T \wedge 1)^{-d/2}\big] + \log\frac{T \vee 1}{t_0 \vee 1}\right].$$

In particular, if $t_0 \leqslant 1 \leqslant T$, the bracketed term is $\mathcal{O}(t_0^{-d/2} + \log T)$.

Under Assumption 1, the target p_0 has stretched exponential tails, which includes the sub-Gaussian case but also allows more general decay through γ. Theorem 6 shows that this additional generality does not deteriorate the statistical rate: for polynomially bounded horizons T, the cumulative score error over $[t_0, T]$ still scales as $\mathrm{polylog}(n)n^{-1}$ with an explicit small-noise dependence $t_0^{-d/2}$. In particular, this bound recovers the rate in the sub-Gaussian case (Zhang et al., 2024, Corollary 3.7).

The bound in Theorem 6 worsens as $t_0 \to 0$, reflecting the ill-conditioning of the low-noise regime. We therefore stop the reverse dynamics at time t_0 and target $p_{t_0} = p_0 * \varphi_{t_0}$, incurring an early-stopping bias $\mathrm{TV}(p_0, p_{t_0})$. The next theorem bounds this bias.

Theorem 7. *Suppose that p_0 satisfies Assumption 1 and Assumption 3 with $\beta \in (0, 2]$. Then, for $0 < t_0 \leqslant 1$, the early stopping bias satisfies*

$$\mathrm{TV}(p_0, p_{t_0}) \lesssim t_0^{\beta/2} \cdot (\log(e/t_0))^{\frac{d}{2\min\{\gamma,2\}}}.$$

Building on this result and the score estimation guarantees in Theorem 6, we establish a total variation bound between p_0 and the distribution of samples produced by the reverse time diffusion dynamics (3).

Theorem 8. *Under the assumptions stated in Theorem 7. Set $t_0 = n^{-\frac{2}{2\beta+d}}$ and $T = n^{\frac{2\beta}{2\beta+d}}$. Then, there exists a score estimator $(\hat{s}_t)_{t\in[0,T]}$ such that the sample $\hat{Y}_{T-t_0}$ from the corresponding reverse time diffusion process* (3) *satisfies*

$$\mathbb{E}\big[\mathrm{TV}\big(p_0, \mathcal{L}(\hat{Y}_{T-t_0})\big)\big] \;\lesssim\; \mathrm{polylog}(n)\, n^{-\frac{\beta}{2\beta+d}}.$$

The choices of t_0 and T align with prior analyses for sub-Gaussian targets and lead to the same rate $n^{-\beta/(2\beta+d)}$ (Zhang et al., 2024, Theorem 3.8). Allowing p_0 to have stretched exponential tails does not degrade convergence relative to the sub-Gaussian or compact support setting, highlighting the robustness of the standard diffusion pipeline across exponentially decaying targets.

Comparing our analysis in the polynomial and exponential regimes reveals a qualitative dichotomy. In the polynomial case, the tail index γ appears explicitly through the factor $n^{-(\gamma+1)/(d+\gamma+1)}$ in Theorem 2. In contrast, in the exponential case, the rate improves to n^{-1} up to logarithmic factors in Theorem 6, and in the corresponding sampling guarantee (Theorem 8) the dependence on γ disappears

from the leading term. The intuition comes from the geometry of where most of the probability mass lives. For polynomially decaying tails, capturing most of the mass requires enlarging the region of interest at a *polynomial* rate in n. Consequently, the size of the domain that must be covered grows rapidly, and the available samples are spread thinly over a large domain. For exponentially decaying tails, the region containing almost all the mass expands only *logarithmically* with n, so its volume grows much more slowly. In this case, the estimation error scales like n^{-1}, and the tail index γ only enters through these logarithmic terms, which explains its apparent disappearance from the main convergence rate.

4 Main Results: Minimax Lower Bound for Score Estimation

To complement our upper bounds, we prove matching minimax lower bounds for score estimation in the heavy-tailed regime. The key challenge is to construct a least favorable family of hypotheses (candidate data distributions) that are hard to distinguish yet yield well-separated scores. We first state the minimax lower bounds for both polynomial and exponential tails, and then describe the least favorable family used in the construction. Since the two regimes follow the same overall strategy, with the polynomial case being more technically involved, we focus on the polynomial-tail setting when presenting the construction. Full proofs are deferred to Appendix H.

4.1 Minimax Lower Bound under Polynomial Tail Regime

Theorem 9. *Let $\mathcal{H}_\gamma$ denote the class of heavy-tailed distributions satisfying Assumption 2 with tail index γ, together with the fixed parameters and polynomial baseline compatibility specified in Appendix H. For all sufficiently large n and all diffusion times satisfying*

$$n^{-\frac{2(d+2\gamma+2)}{d(d+2(\gamma+1))+2\beta(d+\gamma+1)}} \lesssim t \leq t_{\mathrm{P}},$$

where $t_{\mathrm{P}} \in (0,1)$ is a fixed constant chosen as in Appendix H.1, the minimax risk satisfies

$$\inf_{\hat{s}_t} \sup_{p_0 \in \mathcal{H}_\gamma \cap \mathcal{P}_{\mathcal{S}}(\beta,L)} \mathbb{E}\left[\int_{\mathbb{R}^d} \|\hat{s}_t(x) - s_t(x)\|^2 p_t(x)\mathrm{d}x\right] \gtrsim n^{-\frac{\gamma+1}{d+\gamma+1}} t^{-1-\frac{d(\gamma+1)}{2(d+\gamma+1)}}$$

This lower bound matches the rate in Theorem 1 up to logarithmic factors over the stated range of diffusion times, establishing its minimax optimality. In particular, the γ-dependence is intrinsic and cannot be improved uniformly over the class.

4.2 Minimax Lower Bound under Exponential Tail Regime

Theorem 10. *Let $\mathcal{H}'_\gamma$ denote the class of heavy-tailed distributions satisfying Assumption 1, with the fixed parameters and exponential baseline compatibility specified in Appendix H. For all sufficiently large n and all diffusion times satisfying*

$$A_{\mathrm{E}} n^{-\frac{2}{d+2\beta}} \leqslant t \leqslant t_{\mathrm{E}}\,,$$

where $A_{\mathrm{E}} > 0$ and $t_{\mathrm{E}} \in (0,1)$ are fixed constants chosen as in Appendix H.2,

$$\inf_{\hat{s}_t} \sup_{p_0 \in \mathcal{H}'_\gamma \cap \mathcal{P}_{\mathcal{S}}(\beta,L)} \mathbb{E}\left[\int_{\mathbb{R}^d} \|\hat{s}_t(x) - s_t(x)\|^2 p_t(x)\mathrm{d}x\right] \gtrsim n^{-1}\, t^{-1-d/2}.$$

This lower bound matches Theorem 5 up to logarithmic factors over the stated range of diffusion times, thereby establishing the minimax optimality of the score estimation rate.

4.3 Construction of Least Favorable Distributions

The proof of the lower bounds is based on Fano's method, applied to a carefully constructed finite family of target densities. In the polynomial-tail regime considered in Section 4.1, the construction differs from classical nonparametric lower-bound arguments on a fixed compact domain, such as $[-1,1]^d$ studied in Dou et al. (2024): to reflect the polynomial decay of the target density, we instead place localized perturbations in the tail region.

Perturbation domain in the tail. We place localized perturbations on a hypercube grid inside $[R, 2R]^d$, where R will be chosen as a function of n and t. This tail placement has two roles: the baseline density in this region is small (of order $\|x\|^{-(d+\gamma+1)}$), which keeps the KL divergence between hypotheses small and makes them hard to distinguish, while the perturbations can still be tuned to ensure the resulting score functions are well separated. The achievable separation reflects a trade-off between tail decay and perturbation magnitude, and depends on γ.

Perturbed hypotheses. Let $q_0(x) \propto (1 + \|x\|^2)^{-(d+\gamma+1)/2}$. Fix a grid $\{x_\mathbf{i}\}_{\mathbf{i}\in\mathcal{I}} \subset [R, 2R]^d$ with spacing h, where $\mathcal{I}$ denotes the corresponding index set. For a binary vector $b = (b_\mathbf{i})_{\mathbf{i}\in\mathcal{I}} \in \{0, 1\}^{|\mathcal{I}|}$, define the perturbed density

$$q_b(x) = q_0(x) + \epsilon \sum_{\mathbf{i}\in\mathcal{I}} b_\mathbf{i}\, \omega(x - x_\mathbf{i}),$$

where ω is a smooth bump function supported at scale h. We then select a finite subset $\mathcal{B} \subseteq \{0, 1\}^{|\mathcal{I}|}$ and consider the hypothesis family $\{q_b\}_{b\in\mathcal{B}}$.

Coupling of parameters. The lower bound follows by balancing the tail radius R, grid scale h, and amplitude ϵ to satisfy the KL and separation conditions. In particular, we choose

$$R \asymp n^{\frac{1}{d+\gamma+1}}\, t^{\frac{d}{2(d+\gamma+1)}},$$

linking sample size and diffusion time to the geometry of polynomial tails.

5 Discussion

In this work, we characterize minimax limits for score estimation and the resulting sampling error in the canonical Gaussian SDM framework with heavy-tailed targets. We show that Gaussian noising remains statistically robust for exponentially decaying targets, while this robustness breaks down under polynomial tails, where principled remedies remain open. One possible direction is to modify the forward process by adopting heavy-tailed noise, such as stable or Lévy-type perturbations (Yoon et al., 2023; Shariatian et al., 2025). However, this moves beyond the Gaussian diffusion family and introduces substantially different analytic challenges, so extending minimax theory to this regime is nontrivial and beyond the scope of the present work. Another promising approach is to consider tempered or data-dependent noise schedules that avoid placing excessive mass in regions that dominate the statistical error, and we leave a systematic investigation of such remedies for future work.

Furthermore, our sampling analysis focuses on the continuous-time reverse SDE, while practical diffusion models rely on discretized dynamics. Extending our results to the discretized setting remains an important direction for future work.

Finally, although our score estimation rates under polynomial tails are minimax optimal, the optimality of the resulting sampling rates remains unclear. The main difficulty arises in the low-noise regime: our minimax lower bound for score estimation is valid only over a suitable range of diffusion times t and does not cover the regime where t is very close to zero. As a result, in the total variation analysis, we rely on the standard early-stopping argument rather than a sharp characterization of the near-zero-time behavior. Under stronger regularity and score-moment assumptions, Appendix G provides a uniform low-noise score bound and an associated hybrid construction. These auxiliary bounds do not resolve minimax sampling optimality over the original polynomial-tail class. Therefore, determining whether the conversion from score estimation error to total variation sampling error is sharp would require matching sampling lower bounds or a sharper propagation analysis, which we leave for future work.

AI Assistance Disclosure

In this work, we used generative AI tools only to assist with editing and polishing the writing, including improving the clarity, readability, and presentation of the manuscript. We have not used generative AI tools for research ideation, development of theoretical models, formulation or verification of mathematical claims, design or execution of experiments, generation of synthetic data, or drafting sections of the paper.

All AI-assisted edits were carefully reviewed by the authors for accuracy and consistency with the intended meaning. We take responsibility for the final content of this work, including all text, claims, and results presented in the manuscript.

REFERENCES

Alin Achim, Panagiotis Tsakalides, and Anastasios Bezerianos. Sar image denoising via bayesian wavelet shrinkage based on heavy-tailed modeling. *IEEE Transactions on Geoscience and Remote sensing*, 41(8):1773–1784, 2003.

Robert Adler, Raya Feldman, and Murad Taqqu. *A practical guide to heavy tails: statistical techniques and applications*. Springer Science & Business Media, 1998.

Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022.

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Vahan Arsenyan, Elen Vardanyan, and Arnak Dalalyan. Assessing the quality of denoising diffusion models in wasserstein distance: Noisy score and optimal bounds. *arXiv preprint arXiv:2506.09681*, 2025.

Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*, 2024.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

Sergey G. Bobkov and Michel Ledoux. Weighted Poincaré-type inequalities for Cauchy and other convex measures. *The Annals of Probability*, 37(2):403–427, 2009. doi: 10.1214/08-AOP407.

Alexia Briassouli, Panagiotis Tsakalides, and Athanasios Stouraitis. Hidden messages in heavy-tails: Dct-domain watermark detection using alpha-stable models. *IEEE Transactions on Multimedia*, 7 (4):700–715, 2005.

Changxiao Cai and Gen Li. Minimax optimality of the probability flow ode for diffusion models. *arXiv preprint arXiv:2503.09583*, 2025.

Patrick Cattiaux, Giovanni Conforti, Ivan Gentil, and Christian Léonard. Time reversal of diffusion processes under a finite entropy condition. In *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, volume 59, pp. 1844–1881. Institut Henri Poincaré, 2023.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pp. 4735–4763. PMLR, 2023a.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pp. 4672–4712. PMLR, 2023b.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023c.

Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*, 2024.

Sinho Chewi, Alkis Kalavasis, Anay Mehrotra, and Omar Montasser. Ddpm score matching and distribution learning. *arXiv preprint arXiv:2504.05161*, 2025.

Frank Cole and Yulong Lu. Score-based generative models break the curse of dimensionality in learning a family of sub-gaussian probability distributions. *arXiv preprint arXiv:2402.08082*, 2024.

Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. Kl convergence guarantees for score diffusion models under minimal data assumptions. *SIAM Journal on Mathematics of Data Science*, 7(1):86–109, 2025.

Paula Cordero-Encinar, O Deniz Akyildiz, and Andrew B Duncan. Non-asymptotic analysis of diffusion annealed langevin monte carlo for generative modelling. *arXiv preprint arXiv:2502.09306*, 2025.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.

Laurens de Haan and Sidney I. Resnick. On regular variation of probability densities. *Stochastic Processes and their Applications*, 25:83–93, 1987. doi: 10.1016/0304-4149(87)90191-8.

Jacob Deasy, Nikola Simidjievski, and Pietro Liò. Heavy-tailed denoising score matching. *arXiv preprint arXiv:2112.09788*, 2021.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Eleonora Di Nezza, Giampiero Palatucci, and Enrico Valdinoci. Hitchhiker's guide to the fractional Sobolev spaces. *Bulletin des Sciences Mathématiques*, 136(5):521–573, 2012. doi: 10.1016/j.bulsci.2011.12.004.

Zehao Dou, Subhodh Kotekal, Zhehao Xu, and Harrison H Zhou. From optimal score matching to optimal sampling. *arXiv preprint arXiv:2409.07032*, 2024.

Jianqing Fan, Yihong Gu, and Ximing Li. Optimal estimation of a factorizable density using diffusion models with relu neural networks. *arXiv preprint arXiv:2510.03994*, 2025.

Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *arXiv preprint arXiv:2403.11968*, 2024.

Xuefeng Gao, Hoang M Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of machine learning research*, 26(43):1–54, 2025.

Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of gaussians using diffusion models. *arXiv preprint arXiv:2404.18869*, 2024.

Shivam Gupta, Aditya Parulekar, Eric Price, and Zhiyang Xun. Improved sample complexity bounds for diffusion model training. *Advances in Neural Information Processing Systems*, 37:40976–41012, 2024.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Neural network-based score estimation in diffusion models: Optimization and generalization. *arXiv preprint arXiv:2401.15604*, 2024.

Andrew C Harvey. *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*, volume 52. Cambridge University Press, 2013.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pp. 946–985. PMLR, 2023.

Eliya Nachmani, Robin San Roman, and Lior Wolf. Denoising diffusion gamma models. *arXiv preprint arXiv:2110.05948*, 2021.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pp. 26517–26582. PMLR, 2023.

Kushagra Pandey, Jaideep Pathak, Yilun Xu, Stephan Mandt, Michael Pritchard, Arash Vahdat, and Morteza Mardani. Heavy-tailed diffusion models. *arXiv preprint arXiv:2410.14171*, 2024.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pp. 8599–8608. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Dario Shariatian, Umut Simsekli, and Alain Oliviero Durmus. Heavy-tailed diffusion with denoising levy probabilistic models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

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

Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In *International Conference on Artificial Intelligence and Statistics*, pp. 1648–1656. PMLR, 2024.

Wenpin Tang and Hanyang Zhao. Score-based diffusion models via stochastic differential equations–a technical tutorial. *arXiv preprint arXiv:2402.07487*, 2024.

Hans Triebel. *Theory of Function Spaces*, volume 78 of *Monographs in Mathematics*. Birkhäuser, Basel, 1983. doi: 10.1007/978-3-0346-0416-1.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*, volume 11. Springer, 2009.

Alexander Yu. Veretennikov. On strong solutions and explicit formulas for solutions of stochastic integral equations. *Mathematics of the USSR-Sbornik*, 39(3):387–403, 1981. doi: 10.1070/SM1981v039n03ABEH001522.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

Konstantin Yakovlev and Nikita Puchkin. Generalization error bound for denoising score matching under relaxed manifold assumption. *arXiv preprint arXiv:2502.13662*, 2025.

Konstantin Yakovlev, Anna Markovich, and Nikita Puchkin. Implicit score matching meets denoising score matching: improved rates of convergence and log-density hessian estimation. *arXiv preprint arXiv:2512.24378*, 2025.

Eun Bi Yoon, Keehun Park, Sungwoong Kim, and Sungbin Lim. Score-based generative models with lévy processes. *Advances in Neural Information Processing Systems*, 36:40694–40707, 2023.

Yifeng Yu and Lu Yu. Advancing wasserstein convergence analysis of score-based models: Insights from discretization and second-order acceleration. *Advances in Neural Information Processing Systems*, 38:138411–138465, 2026.

Kaihong Zhang, Caitlyn H. Yin, Feng Liang, and Jingbo Liu. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. *Proceedings of Machine Learning Research*, 235:60134–60178, 2024. ISSN 2640-3498.

A Additional Related Works

A.1 Comparison with Closely Related Works

We compare our work with a closely related line of work that studies KDE-based score estimation in diffusion models.

Comparison with Zhang et al. (2024). Our upper-bound analysis is related to theirs in that both use KDE-based score estimators under Gaussian noising. The main differences lie in the kernel construction and the treatment of tail behavior. First, Zhang et al. (2024) use a high-order KDE whose kernel order depends on n, whereas we exploit the identity $p_t = \varphi_t * p_0$ and use the Gaussian transition kernel itself; consequently, the smoothing scale is determined directly by the diffusion time t rather than by an additional bandwidth choice. Second, our heavy-tailed setting requires different arguments for controlling low-density regions after Gaussian smoothing.

The sub-Gaussian/Rényi-entropy approach used in their analysis is not available in our setting, and our exponential-tail assumption already includes stretched-exponential targets with $0 < \gamma < 2$, which lie outside the sub-Gaussian regime. Instead, we control the low-density contribution directly through the assumed tail decay and establish new bulk–tail estimates for the smoothed density. In the polynomial-tail regime, these estimates yield rates that depend explicitly on the tail index γ.

Comparison with Cai & Li (2025). Cai & Li (2025) is related to ours in that both establish statistical guarantees for score estimation and sampling in diffusion-based generative models. The main differences concern the sampling dynamics and the regularity conditions required by the analysis. First, Cai & Li (2025) study probability-flow ODE samplers, whereas we analyze the standard reverse SDE. Because deterministic ODE samplers cannot be handled directly using the usual Girsanov argument, their analysis requires a smooth regularized score estimator that controls not only the L^2 score estimation error but also the corresponding mean Jacobian error. By contrast, our reverse-SDE analysis does not require Jacobian control. Instead, the main statistical challenge is to control the score estimator in low-density regions under heavy-tailed data distributions. Second, their theory assumes sub-Gaussian target distributions with Hölder-smooth densities, while our framework allows both exponential and polynomial tail decay. In the polynomial-tail regime, this yields tail-dependent score estimation and sampling rates that are outside the scope of their sub-Gaussian setting.

Comparison with Dou et al. (2024). Our lower-bound analysis is related to theirs in that both rely on Fano-type arguments to study minimax score estimation after Gaussian noising. The main differences concern the perturbation construction and the role of tail behavior. First, Dou et al. (2024) consider compactly supported Hölder densities on a fixed interval, and construct perturbations within the support. This setting avoids the need to control unbounded tail regions, allowing their analysis to focus on sharp score estimation rates across different noise regimes, especially in the low-noise regime. Second, a key technical challenge in their work is to handle the discontinuity of the density at the boundary of the compact support and to show that this boundary effect does not deteriorate the score estimation rate. By contrast, our heavy-tailed setting has no compact-support boundary, but it requires new arguments to control the contribution of low-density tail regions after Gaussian smoothing. As a result, both our lower-bound construction and upper-bound analysis explicitly reflect the tail decay of the target distribution, leading, in the polynomial-tail regime, to rates that depend on the tail index γ.

A.2 Broader Related Works

Convergence Analysis for Diffusion Models. Motivated by the empirical success of score-based diffusion models, a growing literature has examined their convergence properties. Most existing analyses assume access to accurate score estimates and focus on the convergence behavior of the resulting sampling dynamics. These works typically rely on bounded second-moment assumptions on the target distribution (e.g., (Cordero-Encinar et al., 2025; Chen et al., 2023c; Yu & Yu, 2026; Gao et al., 2025; Conforti et al., 2025; Chen et al., 2023a)) or impose compact support conditions (e.g., (De Bortoli, 2022; Lee et al., 2023; Azangulov et al., 2024; Arsenyan et al., 2025)). While these studies offer valuable insights into the stability, discretization error, and convergence behavior of diffusion-based samplers, they treat the score function as known and therefore bypass the fundamental statistical challenge of learning the score from finite data.

Diffusion Models Beyond Gaussian Noise. A separate thread of work considers SDMs that replace Gaussian perturbations in the forward process with non-Gaussian noise, such as generalized Gaussian distribution (Deasy et al., 2021), gamma distribution (Nachmani et al., 2021), t-distributions (Pandey et al., 2024) or α-stable Lévy processes (Yoon et al., 2023; Shariatian et al., 2025). These approaches share a common motivation: to enhance robustness in settings with imbalanced or heavy-tailed data and to broaden the applicability of diffusion models beyond the Gaussian framework. However, theoretical development in this direction has been limited. Unlike the Gaussian case, non-Gaussian perturbations typically do not yield closed-form transition densities or tractable reverse SDEs, making it difficult to analyze the resulting dynamics or characterize sampling error. As a consequence, most of these works provide no explicit convergence guarantees, nor do they study the statistical difficulty of estimating the corresponding score functions.

B Proof Sketch

In this section, we sketch the proofs of Theorems 4 and 8. We first treat polynomial tails, outlining the sampling analysis and the roles of score estimation error, early-stopping bias, and initialization. We then summarize the key ideas for the associated score estimation bounds, highlighting the main technical obstacles and how our setting and arguments depart from prior work. We then turn to exponential tails, which follow the same template, and highlight only the differences from the polynomial case.

B.1 Proof Sketch for the Sampling Error in the Polynomial Decay Regime (Theorem 4)

The sampling proof combines the score estimation and early-stopping bounds via a triangle inequality decomposition. All stochastic arguments are conditional on the training sample $\mathcal{D}_n$, after which we will take expectations over $\mathcal{D}_n$. We have

$$\mathrm{TV}\big(p_0, \mathcal{L}(\hat{Y}_{T-t_0})\big) \leqslant \mathrm{TV}(p_0, p_{t_0}) + \mathrm{TV}\big(p_{t_0}, \mathcal{L}(\hat{Y}_{T-t_0})\big),$$

which separates the early stopping bias and the reverse process error. Next, let $\tilde{Y}$ denote the backward process with the same learned score but initialized from p_T. Then,

$$\mathrm{TV}\big(p_{t_0}, \mathcal{L}(\hat{Y}_{T-t_0})\big) \leqslant \mathrm{TV}\big(p_{t_0}, \mathcal{L}(\tilde{Y}_{T-t_0})\big) + \mathrm{TV}\big(\mathcal{L}(\tilde{Y}_{T-t_0}), \mathcal{L}(\hat{Y}_{T-t_0})\big).$$

The first term corresponds to the score estimation error. For the second term, TV contraction under the backward kernel yields

$$\mathrm{TV}\big(\mathcal{L}(\tilde{Y}_{T-t_0}), \mathcal{L}(\hat{Y}_{T-t_0})\big) \leqslant \mathrm{TV}\big(p_T, \mathcal{N}(0, TI_d)\big),$$

which we interpret as the initialization error. Combining these bounds gives the following decomposition.

$$\mathrm{TV}(p_0, \mathcal{L}(\hat{Y}_{T-t_0})) \leqslant \underbrace{\mathrm{TV}(p_0, p_{t_0})}_{\text{Early Stopping Bias}} + \underbrace{\mathrm{TV}(p_{t_0}, \mathcal{L}(\tilde{Y}_{T-t_0}))}_{\text{Score Matching Error}} + \underbrace{\mathrm{TV}(p_T, \mathcal{N}(0, TI_d))}_{\text{Initialization Error}}.$$

Score matching error (Theorems 1 and 2). By Pinsker's inequality and Girsanov's theorem with a localization argument, the sampling error is controlled by the square root of the cumulative score estimation error.

$$\mathbb{E}\left[\mathrm{TV}\big(\mathcal{L}(\tilde{Y}_{T-t_0}), p_{t_0}\big)\right] \lesssim \left(\int_{t_0}^{T} \mathbb{E}\Big[\int_{\mathbb{R}^d} \|\hat{s}_t(x) - s_t(x)\|^2 p_t(x)\,\mathrm{d}x\Big]\mathrm{d}t\right)^{1/2}.$$

Combining this with the refined cumulative bound in Theorem 2, and using that the chosen horizon satisfies $T = n^a$ so the $\log T$ term is lower order, we obtain

$$\mathbb{E}\left[\mathrm{TV}\big(\mathcal{L}(\tilde{Y}_{T-t_0}), p_{t_0}\big)\right] \lesssim \mathrm{polylog}(n)\, n^{-\frac{\gamma+1}{2(d+\gamma+1)}}\, t_0^{-\frac{d(\gamma+1)}{4(d+\gamma+1)}}.$$

Early stopping bias (Theorem 3). We early stop the reverse process at t_0 and target $p_{t_0} = p_0 * \varphi_{t_0}$. Theorem 3 bounds the resulting total variation bias

$$\mathrm{TV}(p_0, p_{t_0}) \lesssim t_0^{\frac{\beta(\gamma+1)}{d+2(\gamma+1)}}.$$

Initialization error (Lemma 19). The discrepancy between the distribution $p_T = p_0 * \varphi_T$ and the Gaussian prior $\mathcal{N}(0, TI_d)$ is controlled by Lemma 19, which yields $\mathrm{TV}(p_T, \varphi_T) \lesssim T^{-1/2}$. Thus, the initialization error can be made negligible by choosing T sufficiently large.

Balancing the early stopping bias and score estimation error determines the optimal t_0 and yields the stated rate.

B.1.1 Proof Sketch of Theorem 1

Our main approach is to decompose the error using a thresholding argument. We control the weighted mean integrated squared error (MISE) of the score estimator $\hat{s}_t(x)$ by splitting the domain according to (i) the magnitude of the true density and (ii) the thresholding applied to the estimator.

First, we split $\mathbb{R}^d$ into a bulk region $\mathcal{G}_1 := \{x \in \mathbb{R}^d : p_t(x) > c_\alpha \rho_n\}$ and a tail region $\mathcal{G}_2 := \{x \in \mathbb{R}^d : p_t(x) \leqslant c_\alpha \rho_n\}$, where ρ_n is the given truncation threshold and $c_\alpha > 0$ is a constant to be specified. Second, We then exploit the estimator's thresholding: by construction, $\hat{s}_t(x) = 0$ whenever $\hat{p}_t(x) < \rho_n$.

The link between p_t and $\hat{p}_t$ is handled via the following pointwise concentration bound.

Lemma 11. *For any fixed $\alpha > 0$ and $n \geqslant 3$, there exists a constant $C_\alpha > 0$, depending only on α, such that for every fixed $x \in \mathbb{R}^d$,*

$$\mathbb{P}\left(\left|\hat{p}_t(x) - p_t(x)\right| < C_\alpha\left(\frac{\log n}{n(2\pi t)^{d/2}} + \sqrt{\frac{p_t(x)\log n}{n(2\pi t)^{d/2}}}\right)\right) \geqslant 1 - n^{-\alpha},$$

where $C_\alpha = \max\{\sqrt{8(\alpha+1)},\, 16(\alpha+1)/3\}$.

Let $A_\alpha(x)$ denote the high-probability event on which Lemma 11 holds and set the function $\Delta s_t(x) := \hat{s}_t(x) - s_t(x)$. Using this event together with the bulk and tail partition, we decompose the error as

$$\begin{aligned}
&\mathbb{E}\Big[\int_{\mathbb{R}^d} \|\Delta s_t(x)\|^2 p_t(x)\,\mathrm{d}x\Big] \\
&= \underbrace{\int_{\mathcal{G}_1} \mathbb{E}[\|\Delta s_t(x)\|^2\, \mathbb{1}_{\{\omega \in A_\alpha(x)\}}]\, p_t(x)\,\mathrm{d}x}_{\mathrm{I}} + \underbrace{\int_{\mathcal{G}_2} \mathbb{E}[\|\Delta s_t(x)\|^2\, \mathbb{1}_{\{\omega \in A_\alpha(x)\}}]\, p_t(x)\,\mathrm{d}x}_{\mathrm{II}} \\
&\quad + \underbrace{\int_{\mathbb{R}^d} \mathbb{E}[\|\Delta s_t(x)\|^2\, \mathbb{1}_{\{\omega \in A_\alpha(x)^c\}}]\, p_t(x)\,\mathrm{d}x}_{\mathrm{III}}.
\end{aligned}$$

Within terms I and II, we further split into the active region $\hat{p}_t(x) \geqslant \rho_n$ (score estimated) and the truncated region $\hat{p}_t(x) < \rho_n$ (score set to zero).

Step 1: Lower bounds on the denominator. The main difficulty is the unstable denominator in

$$\hat{s}_t(x) = \frac{\nabla \hat{p}_t(x)}{\hat{p}_t(x)}\, \mathbb{1}_{\{\hat{p}_t(x) \geqslant \rho_n\}}.$$

We control it by lower bounding $\hat{p}_t(x)$ when $x \in \mathcal{G}_1$ and $\omega \in A_\alpha(x)$, Lemma 11 implies $\hat{p}_t(x) \asymp p_t(x)$, hence $\hat{p}_t(x) \geqslant \frac{1}{2} p_t(x)$, while elsewhere on the active regions we have $\hat{p}_t(x) \geqslant \rho_n$.

Step 2: Reduction to pointwise MSE bounds. With the denominator controlled, the score error reduces to terms involving the errors of $\hat{p}_t$ and $\nabla \hat{p}_t$. These are bounded by pointwise MSE estimates that scale with the true density $p_t(x)$.

Proposition 12. *The mean squared error of $\hat{p}_t(x)$ satisfies* $\mathrm{MSE}(\hat{p}_t(x)) \leqslant \frac{p_t(x)}{n(2\pi t)^{d/2}}$.

Proposition 13. *The mean squared error of $\nabla \hat{p}_t(x)$ satisfies* $\mathrm{MSE}(\nabla \hat{p}_t(x)) \leqslant \frac{p_t(x)}{nt(2\pi t)^{d/2}}$.

Step 3: Integration using bulk and tail lemmas. Plugging these MSE bounds into terms I and II yields integrals over $\mathcal{G}_1$ and $\mathcal{G}_2$ weighted by $p_t(x)$ and $s_t(x)$. We control them using two lemmas that characterize the geometry of the bulk and tail region under heavy tails.

Lemma 14 (Also Lemma 16)**.** *Under Assumption 2, we have*

$$|\mathcal{G}_1| \lesssim \left(\frac{C_\gamma(t)}{\rho_n}\right)^{\frac{d}{1+\gamma+d}}, \text{ and } \int_{\mathcal{G}_1} \|s_t(x)\|^2 \,\mathrm{d}x \lesssim \big(C_\gamma(t)\big)^{\frac{d}{(1+\gamma+d)m'}} \, m\, t^{-1} \rho_n^{-\frac{1}{m}-\frac{d}{(1+\gamma+d)m'}}.$$

Lemma 15 (Also Lemma 17)**.** *Under Assumption 2, it holds that*

$$\int_{\mathcal{G}_2} p_t(x)\,\mathrm{d}x \lesssim \big(C_\gamma(t)\big)^{\frac{d}{d+\gamma+1}} \rho_n^{\frac{\gamma+1}{d+\gamma+1}}, \text{ and } \int_{\mathcal{G}_2} \|s_t(x)\|^2 p_t(x)\,\mathrm{d}x \lesssim \varepsilon^{-1} t^{-1} \big(C_\gamma(t)\big)^{\frac{d(1-\varepsilon)}{d+\gamma+1}} \rho_n^{\frac{(\gamma+1)(1-\varepsilon)}{d+\gamma+1}}.$$

Here $m, m' > 1$ satisfying $1/m + 1/m' = 1$, $C_\gamma(t)$ depends on γ, t, and d, whose explicit form is given in the appendix. Combining the bounds for terms I–III, the chosen ρ_n both (i) guarantees $\hat{p}_t \geqslant \frac{1}{2} p_t$ on $\mathcal{G}_1$ (Lemma 11) and (ii) balances the contributions of terms I and II, thereby yielding the rate in Theorem 1.

Remark 1 (Origin of the heavy-tail rate) We emphasize that the rate $n^{-\frac{\gamma+1}{d+\gamma+1}}$ in Theorem 1 is not an artifact of a loose analysis. It arises from two intrinsic geometric features of polynomial tails that alter the bias–variance trade-off relative to the sub-Gaussian case. **i) Bulk expansion:** In light-tailed (sub-Gaussian) settings, the stable high-density region $\mathcal{G}_1$ grows only logarithmically with n. Under polynomial tails, $|\mathcal{G}_1|$ expands polynomially as $\rho_n^{-d/(d+\gamma+1)}$ (Lemma 14), reducing the effective sample density and inflating the bulk variance. **ii) Tail mass:** Under sub-Gaussian tails, the mass of the low-density region $\mathcal{G}_2$ is bounded by ρ_n times a polylogarithmic factor in n at fixed diffusion time. With polynomial tails, $\mathcal{G}_2$ retains non-negligible mass of order $\rho_n^{(\gamma+1)/(d+\gamma+1)}$ (Lemma 15), forcing a conservative ρ_n and tightening the variance–bias trade-off.

Remark 2 (Challenges in the low-noise regime) Our analysis also clarifies why a naive kernel-based score estimator becomes unstable as $t \to 0$ on unbounded domains. Under both exponential and polynomial tails, samples are extremely sparse in the far tails, and when the bandwidth $\sqrt{t}$ is too small relative to this sparsity, the estimator averages over almost no observations, causing its variance to blow up. This issue is largely absent under compact support, where the main difficulty is boundary bias rather than large empty regions. The key difference is how much probability mass these regions carry: with exponential tails the mass decays rapidly, so their contribution to the L_2-error is negligible and rates can match compact support case (Dou et al., 2024), whereas with polynomial tails the tail mass decays slowly, so far tail variance dominates and degrades the convergence rate.

B.2 Proof Sketch for the Sampling Error in the Exponential Decay Regime (Theorem 8)

The proof follows the same template as the polynomial tail case, but exponential tails yield sharper geometric control, so the score matching estimator avoids the heavy-tail penalty. We start from the same decomposition.

$$\mathrm{TV}(p_0, \mathcal{L}(\hat{Y}_{T-t_0})) \leqslant \underbrace{\mathrm{TV}(p_0, p_{t_0})}_{\text{Early Stopping Bias}} + \underbrace{\mathrm{TV}(p_{t_0}, \mathcal{L}(\tilde{Y}_{T-t_0}))}_{\text{Score Matching Error}} + \underbrace{\mathrm{TV}(p_T, \mathcal{N}(0, TI_d))}_{\text{Initialization Error}}.$$

Score Matching Error (Theorems 5 and 6). As before, Pinsker's inequality and Girsanov's theorem bound the sampling error by the square root of the cumulative score MSE. In the exponential regime, geometry is sharper: the volume $|\mathcal{G}_1|$ grows only logarithmically in n at fixed t (Lemma 21) and the low-density mass is bounded by the threshold times the effective volume (Lemma 22), which yields a faster cumulative rate when plugged into the score MSE bound, for the polynomial horizon and cutoff used below.

$$\mathbb{E}\left[\mathrm{TV}\big(\mathcal{L}(\tilde{Y}_{T-t_0}), p_{t_0}\big)\right] \lesssim \mathrm{polylog}(n)\, n^{-1/2}\, t_0^{-d/4}.$$

Early stopping bias (Theorem 7). Early stopping at t_0 again targets $p_{t_0} = p_0 * \varphi_{t_0}$. Under exponential tails, truncation is negligible, and the resulting bias is governed by Sobolev regularity up

to a logarithmic factor

$$\mathrm{TV}(p_0, p_{t_0}) \lesssim t_0^{\beta/2} (\log(e/t_0))^{d/(2\min\{\gamma,2\})}.$$

Initialization error (Lemma 19). The initialization mismatch between $p_T = p_0 * \varphi_T$ and the Gaussian prior satisfies $\mathrm{TV}(p_T, \varphi_T) \lesssim T^{-1/2}$.

The final rate follows by balancing the early stopping bias with the score matching error.

C Additional Details for the Preliminaries

C.1 Equivalent Fourier Characterizations of Sobolev Smoothness

Assumption 3 imposes a Fourier-domain fractional Sobolev condition. It is equivalent, up to constants depending only on d and β, to the radial bound $\int_{\mathbb{R}^d} \|\omega\|^{2\beta} |\mathcal{F}[p](\omega)|^2 \, \mathrm{d}\omega \lesssim L^2$: the inequality $|\omega^\alpha|^2 \leqslant \|\omega\|^{2\beta}$ holds whenever $\sum_j \alpha_j = \beta$, while finitely many choices such as $\alpha = \beta e_j$ recover the radial weight by norm equivalence (Triebel, 1983; Di Nezza et al., 2012).

C.2 Properties of the Gaussian KDE

Additional Notation. We denote by $\mathbb{B}_r(x) := \{y \in \mathbb{R}^d : \|y - x\| \leqslant r\}$ the Euclidean ball of radius r centered at x, and simply write $\mathbb{B}_r$ for $\mathbb{B}_r(0)$.

Before we proceed with the proof, we first derive the unbiasedness of the proposed estimator of p_t and ∇p_t, which is central to our error analysis and will be used repeatedly throughout the heavy-tailed setting we consider.

Unbiasedness. With expectation taken over the training sample, $\hat{p}_t$ and $\nabla \hat{p}_t$ are unbiased estimators of p_t and ∇p_t, respectively. In particular,

$$\mathbb{E}[\hat{p}_t(x)] = \int_{\mathbb{R}^d} \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{\|x-y\|^2}{2t}\right) p_0(y) \, \mathrm{d}y = (\varphi_t * p_0)(x) = p_t(x)$$

and

$$\mathbb{E}\left[\nabla \hat{p}_t(x)\right] = \int_{\mathbb{R}^d} \nabla_x \varphi_t(x-y) \, p_0(y) \, \mathrm{d}y = \nabla_x \int_{\mathbb{R}^d} \varphi_t(x-y) \, p_0(y) \, \mathrm{d}y = \nabla p_t(x).$$

Proof of Proposition 12. By the definition of MSE and the construction of the $\hat{p}_t$, we have

$$\mathrm{MSE}(\hat{p}_t(x)) \leqslant \mathrm{Var}(\hat{p}_t(x)) \leqslant \frac{1}{n} \mathrm{Var}(\varphi_t(X^{(i)} - x)) \, .$$

Let $Z_0 \sim p_0$ be an independent copy of $\{X^{(i)}\}_{i=1}^n$. We then have

$$\begin{aligned}
\frac{1}{n} \mathrm{Var}(\varphi_t(X^{(i)} - x)) &\leqslant \frac{1}{n} \mathbb{E}[\varphi_t^2(Z_0 - x)] \\
&= \frac{1}{n} \int_{\mathbb{R}^d} \frac{1}{(2\pi t)^d} \exp\left(-\frac{\|x-y\|_2^2}{t}\right) p_0(y) \, \mathrm{d}y \\
&\leqslant \frac{1}{n(2\pi t)^{d/2}} \int_{\mathbb{R}^d} \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{\|x-y\|_2^2}{2t}\right) p_0(y) \, \mathrm{d}y \\
&= \frac{p_t(x)}{n(2\pi t)^{d/2}} \, .
\end{aligned}$$

Combining this with the previous display gives the desired result. □

Proof of Proposition 13. Let $Z_0 \sim p_0$ be an independent copy of $\{X^{(i)}\}_{i=1}^n$. Since $\nabla \hat{p}_t(x)$ is unbiased, we have

$$\mathrm{MSE}(\nabla \hat{p}_t(x)) = \mathrm{Var}(\nabla \hat{p}_t(x))$$

$$
\begin{aligned}
&= \frac{1}{nt^2}\mathrm{Var}((X^{(i)}-x)\varphi_t(X^{(i)}-x)) \\
&\leqslant \frac{1}{nt^2}\mathbb{E}\left[\|Z_0-x\|^2\varphi_t^2(Z_0-x)\right] \\
&= \frac{1}{nt^2}\int_{\mathbb{R}^d}\frac{\|y-x\|^2}{(2\pi t)^d}\exp\left(-\frac{\|x-y\|_2^2}{t}\right)p_0(y)\mathrm{d}y \\
&\leqslant \frac{1}{nt(2\pi t)^{d/2}}\int_{\mathbb{R}^d}\frac{1}{(2\pi t)^{d/2}}\exp\left(-\frac{\|x-y\|_2^2}{2t}\right)p_0(y)\mathrm{d}y \\
&= \frac{p_t(x)}{nt(2\pi t)^{d/2}}
\end{aligned}
$$

as desired. □

D Proof of Section 3.1

For each fixed x, we define the event

$$
A_\alpha(x) = \left\{\omega : |\hat{p}_t(x)-p_t(x)| < C_\alpha\left(\frac{\log n}{n(2\pi t)^{d/2}} + \sqrt{\frac{p_t(x)\log n}{n(2\pi t)^{d/2}}}\right)\right\}.
$$

By Lemma 11, it holds that $\mathbb{P}(A_\alpha(x)^c) < n^{-\alpha}$ for each fixed x. We further decompose the space into the bulk region $\mathcal{G}_1$ and the tail region $\mathcal{G}_2$ through

$$
\mathcal{G}_1 := \{x : p_t(x) > c_\alpha\rho_n\}, \qquad \mathcal{G}_2 := \{x : p_t(x) \leqslant c_\alpha\rho_n\},
$$

where c_α is chosen to satisfy $c_\alpha > 2$ and $c_\alpha > 2C_\alpha(1+\sqrt{c_\alpha})$. Then for any $x \in \mathcal{G}_1$ and $\omega \in A_\alpha(x)$,

$$
\begin{aligned}
\hat{p}_t(x) &> p_t(x) - C_\alpha\big(\rho_n + \sqrt{p_t(x)\rho_n}\big) \\
&> p_t(x) - C_\alpha\left(\frac{p_t(x)}{c_\alpha} + \frac{p_t(x)}{\sqrt{c_\alpha}}\right) > \Big(1 - C_\alpha(c_\alpha^{-1}+c_\alpha^{-1/2})\Big)p_t(x) > \frac{1}{2}p_t(x),
\end{aligned}
$$

and consequently $\hat{p}_t(x) > \frac{1}{2}p_t(x) > \frac{c_\alpha}{2}\rho_n > \rho_n$. This shows that in the bulk region $\mathcal{G}_1$, the density estimate is bounded away from zero (on event $A_\alpha(x)$), whereas $\mathcal{G}_2$ isolates the low-density region where thresholding is needed for stability.

D.1 Proof of Theorem 1

Proof. The proof relies on partitioning the integration domain into a high-density region $\mathcal{G}_1$, where the kernel estimator concentrates well, and a low-density tail region $\mathcal{G}_2$, where the estimator is truncated. We divide the L^2 score matching error into three parts

$$
\begin{aligned}
&\mathbb{E}\left[\int_{\mathbb{R}^d}\|\hat{s}_t(x)-s_t(x)\|^2 p_t(x)\,\mathrm{d}x\right] \\
&= \int_{\mathbb{R}^d}\mathbb{E}\left[\|\hat{s}_t(x)-s_t(x)\|^2\right]p_t(x)\,\mathrm{d}x \\
&= \underbrace{\int_{\mathcal{G}_1}\mathbb{E}\left[\|\hat{s}_t(x)-s_t(x)\|^2\mathbb{1}\{\omega\in A_\alpha(x)\}\right]p_t(x)\,\mathrm{d}x}_{\mathrm{I}} + \underbrace{\int_{\mathcal{G}_2}\mathbb{E}\left[\|\hat{s}_t(x)-s_t(x)\|^2\mathbb{1}\{\omega\in A_\alpha(x)\}\right]p_t(x)\,\mathrm{d}x}_{\mathrm{II}} \\
&\quad + \underbrace{\int_{\mathbb{R}^d}\mathbb{E}\left[\|\hat{s}_t(x)-s_t(x)\|^2\mathbb{1}\{\omega\in A_\alpha(x)^c\}\right]p_t(x)\,\mathrm{d}x}_{\mathrm{III}}.
\end{aligned}
$$

Step 1. Note that when $x \in \mathcal{G}_1$ and $\omega \in A_\alpha(x)$, $\hat{p}_t(x) > \max\{\rho_n, \frac{1}{2}p_t(x)\}$, then

$$
\mathrm{I} \leqslant \int_{\mathcal{G}_1}\mathbb{E}[\|\hat{s}_t(x)-s_t(x)\|^2\mathbb{1}\{\hat{p}_t(x) > \max(\rho_n, \frac{1}{2}p_t(x))\}]p_t(x)\,\mathrm{d}x \tag{4}
$$

$$\leqslant \int_{\mathcal{G}_1} 2\mathbb{E}\left[\frac{\|\nabla\hat{p}_t(x)-\nabla p_t(x)\|^2+\|s_t(x)\|^2|\hat{p}_t(x)-p_t(x)|^2}{\hat{p}_t(x)^2}\mathbb{1}\{\hat{p}_t(x)>\frac{1}{2}p_t(x)\}\right]p_t(x)\,\mathrm{d}x$$
$$\leqslant 8\int_{\mathcal{G}_1}\mathbb{E}\left[\|\nabla\hat{p}_t(x)-\nabla p_t(x)\|^2+\|s_t(x)\|^2|\hat{p}_t(x)-p_t(x)|^2\right]\frac{1}{p_t(x)}\,\mathrm{d}x$$
$$\leqslant 8\int_{\mathcal{G}_1}\left(\mathrm{MSE}(\nabla\hat{p}_t(x))+\mathrm{MSE}(\hat{p}_t(x))\cdot\|s_t(x)\|^2\right)\cdot\frac{1}{p_t(x)}\,\mathrm{d}x$$

By Proposition 12 and 13, we have

$$\mathrm{I}\leqslant 8|\mathcal{G}_1|\cdot\frac{1}{nt(2\pi t)^{d/2}}+8\int_{\mathcal{G}_1}\|s_t(x)\|^2\,\mathrm{d}x\cdot\frac{1}{n(2\pi t)^{d/2}} \tag{5}$$

Lemma 16. *Under Assumption 2, we have*

$$|\mathcal{G}_1|\lesssim\left(\frac{C_\gamma(t)}{\rho_n}\right)^{\frac{d}{1+\gamma+d}},$$

and

$$\int_{\mathcal{G}_1}\|s_t(x)\|^2\mathrm{d}x\lesssim C_\gamma(t)^{\frac{d}{(1+\gamma+d)m'}}mt^{-1}\rho_n^{-\frac{1}{m}-\frac{d}{(1+\gamma+d)m'}},$$

where $C_\gamma(t)=C_*\max\{1,t^{(\gamma+1)/2}\}$ is the global envelope constant in Lemma 23; $C_*\geqslant 1$ is fixed independently of n and t. Here $m>1$ and $m'=m/(m-1)$.

By taking $m=\frac{1}{\varepsilon}$ in Lemma 16, we obtain

$$\int_{\mathcal{G}_1}\|s_t(x)\|^2\mathrm{d}x\lesssim C_\gamma(t)^{\frac{d(1-\varepsilon)}{1+\gamma+d}}\varepsilon^{-1}t^{-1}\rho_n^{-\varepsilon-\frac{d(1-\varepsilon)}{1+\gamma+d}}.$$

We now simplify the bound obtained by substituting this estimate into (5). Set $D=d+\gamma+1$, $\kappa=(\gamma+1)/D$, and

$$A_t:=C_\gamma(t)^{d/D}\rho_n^\kappa.$$

Since $[n(2\pi t)^{d/2}]^{-1}=\rho_n/\log n$, the bulk bound is

$$\mathrm{I}\lesssim\frac{C_\gamma(t)^{d/D}\rho_n^{-d/D}+\varepsilon^{-1}C_\gamma(t)^{d(1-\varepsilon)/D}\rho_n^{-\varepsilon-d(1-\varepsilon)/D}}{nt(2\pi t)^{d/2}}$$
$$=\frac{A_t+\varepsilon^{-1}A_t^{1-\varepsilon}}{t\log n}.$$

Step 2. Recall the definition of $\mathcal{G}_2$ and $\hat{s}_t(x)$, we have

$$\mathrm{II}\leqslant\int_{\mathcal{G}_2}\mathbb{E}\left[\|s_t(x)\|^2\mathbb{1}\{\hat{p}_t(x)<\rho_n\}\right]p_t(x)\,\mathrm{d}x+\int_{\mathcal{G}_2}\mathbb{E}\left[\|\hat{s}_t(x)-s_t(x)\|^2\mathbb{1}\{\hat{p}_t(x)\geqslant\rho_n\}\right]p_t(x)\,\mathrm{d}x.$$

Lemma 17. *Under Assumption 2, it holds that*

$$\int_{\mathcal{G}_2}p_t(x)\,\mathrm{d}x\lesssim C_\gamma(t)^{\frac{d}{d+\gamma+1}}\rho_n^{\frac{\gamma+1}{d+\gamma+1}},$$

and

$$\int_{\mathcal{G}_2}\|s_t(x)\|^2p_t(x)\,\mathrm{d}x\lesssim\varepsilon^{-1}t^{-1}C_\gamma(t)^{\frac{d(1-\varepsilon)}{d+\gamma+1}}\rho_n^{\frac{(\gamma+1)(1-\varepsilon)}{d+\gamma+1}}.$$

Applying Lemma 17, the first term of II can be bounded by

$$\int_{\mathcal{G}_2}\mathbb{E}\left[\|s_t(x)\|^2\mathbb{1}\{\hat{p}_t(x)<\rho_n\}\right]p_t(x)\,\mathrm{d}x\leqslant\int_{\mathcal{G}_2}\|s_t(x)\|^2p_t(x)\,\mathrm{d}x\lesssim\varepsilon^{-1}t^{-1}C_\gamma(t)^{\frac{d(1-\varepsilon)}{d+\gamma+1}}\rho_n^{\frac{(\gamma+1)(1-\varepsilon)}{(d+\gamma+1)}}.$$

For the second term of II, by Proposition 12 and Proposition 13, and following the similar approach in (4), we have

$$\begin{aligned}
&\int_{\mathcal{G}_2} \mathbb{E}\left[\|\hat{s}_t(x) - s_t(x)\|^2 \mathbb{1}\{\hat{p}_t(x) \geqslant \rho_n\}\right] p_t(x)\,\mathrm{d}x \\
&\leqslant \int_{\mathcal{G}_2} 2\mathbb{E}\left[\frac{\|\nabla\hat{p}_t(x) - \nabla p_t(x)\|^2 + \|s_t(x)\|^2|\hat{p}_t(x) - p_t(x)|^2}{\hat{p}_t(x)^2}\mathbb{1}\{\hat{p}_t(x) \geqslant \rho_n\}\right] p_t(x)\,\mathrm{d}x \\
&\leqslant 2\rho_n^{-2}\int_{\mathcal{G}_2}\left(\mathrm{MSE}(\nabla\hat{p}_t(x)) + \mathrm{MSE}(\hat{p}_t(x))\cdot\|s_t(x)\|^2\right)p_t(x)\,\mathrm{d}x \\
&\leqslant 2\rho_n^{-2}\left(\int_{\mathcal{G}_2}\frac{p_t(x)}{nt(2\pi t)^{d/2}}p_t(x)\,\mathrm{d}x + \int_{\mathcal{G}_2}\frac{p_t(x)}{n(2\pi t)^{d/2}}\cdot\|s_t(x)\|^2 p_t(x)\,\mathrm{d}x\right) \\
&\leqslant 2\rho_n^{-1}\left(\frac{c_\alpha}{nt(2\pi t)^{d/2}}\int_{\mathcal{G}_2}p_t(x)\,\mathrm{d}x + \frac{c_\alpha}{n(2\pi t)^{d/2}}\int_{\mathcal{G}_2}\|s_t(x)\|^2 p_t(x)\,\mathrm{d}x\right).
\end{aligned}$$

Applying Lemma 17 again, it then follows that

$$\begin{aligned}
&\int_{\mathcal{G}_2} \mathbb{E}\left[\|\hat{s}_t(x) - s_t(x)\|^2 \mathbb{1}\{\hat{p}_t(x) \geqslant \rho_n\}\right] p_t(x)\,\mathrm{d}x \\
&\lesssim 2\rho_n^{-1}\left(\frac{c_\alpha}{nt(2\pi t)^{d/2}}C_\gamma(t)^{\frac{d}{d+\gamma+1}}\rho_n^{\frac{\gamma+1}{d+\gamma+1}} + \frac{c_\alpha}{n(2\pi t)^{d/2}}\varepsilon^{-1}t^{-1}C_\gamma(t)^{\frac{d(1-\varepsilon)}{d+\gamma+1}}\rho_n^{\frac{(\gamma+1)(1-\varepsilon)}{d+\gamma+1}}\right) \\
&= 2c_\alpha(t\log n)^{-1}\left(C_\gamma(t)^{\frac{d}{d+\gamma+1}}\rho_n^{\frac{\gamma+1}{d+\gamma+1}} + \varepsilon^{-1}C_\gamma(t)^{\frac{d(1-\varepsilon)}{d+\gamma+1}}\rho_n^{\frac{(\gamma+1)(1-\varepsilon)}{d+\gamma+1}}\right)
\end{aligned}$$

Step 3. By Hölder's inequality, we obtain

$$\begin{aligned}
\mathrm{III} &\leqslant \left(\int_{\mathbb{R}^d}\mathbb{E}\left[\|\hat{s}_t(x) - s_t(x)\|^4\right]p_t(x)\,\mathrm{d}x\right)^{1/2}\left(\int_{\mathbb{R}^d}\mathbb{E}\left[\mathbb{1}\{\omega\in A_\alpha(x)^c\}\right]p_t(x)\,\mathrm{d}x\right)^{1/2} \\
&\lesssim n^{-\alpha/2}\left(\int_{\mathbb{R}^d}\|s_t(x)\|^4 p_t(x)\,\mathrm{d}x + \int_{\mathbb{R}^d}\mathbb{E}\left[\|\hat{s}_t(x)\|^4\right]p_t(x)\,\mathrm{d}x\right)^{1/2}.
\end{aligned}$$

Lemma 18 (Lemma F.16 in Zhang et al. (2024))**.** *Let X and G be independent d-dimensional random vectors such that $X \sim p_0, G \sim \mathcal{N}(0, I_d)$ and $Y = X + \sqrt{t}G$. Then,*

$$\mathbb{E}[\|\nabla \log p_Y(Y)\|^m] \leqslant t^{-\frac{m}{2}} d^{\frac{m}{2}} (m-1)!!\,.$$

By Lemma 18, the first term is bounded by

$$\int_{\mathbb{R}^d}\|s_t(x)\|^4 p_t(x)\,\mathrm{d}x \leqslant 3t^{-2}d^2\,,$$

and the second term can be bounded by

$$\begin{aligned}
\int_{\mathbb{R}^d}\mathbb{E}\|\hat{s}_t(x)\|^4 p_t(x)\,\mathrm{d}x &= \int_{\mathbb{R}^d}\mathbb{E}\left[\|\hat{s}_t(x)\|^4\mathbb{1}\{\hat{p}_t(x) > \rho_n\}\right]p_t(x)\,\mathrm{d}x \\
&\leqslant \rho_n^{-4}\int_{\mathbb{R}^d}\mathbb{E}\left[\|\nabla\hat{p}_t(x)\|^4\right]p_t(x)\,\mathrm{d}x\,.
\end{aligned}$$

Recall the definition of $\nabla p_t(x)$, it holds that

By convexity of $z \mapsto \|z\|^4$ and the identity $\varphi_t(z)^4 = (2\pi t)^{-2d}\exp(-2\|z\|^2/t)$, we have, uniformly in x,

$$\begin{aligned}
\mathbb{E}\left[\|\nabla\hat{p}_t(x)\|^4\right] &= \frac{1}{(nt)^4}\mathbb{E}_{X^{(i)}\sim p_0}\left[\left\|\sum_{i=1}^n (X^{(i)} - x)\varphi_t(X^{(i)} - x)\right\|^4\right] \\
&\leqslant \frac{1}{t^4}\mathbb{E}_{X\sim p_0}\left[\|(X-x)\varphi_t(X-x)\|^4\right] \\
&\leqslant \frac{1}{t^4(2\pi t)^{2d}}\sup_{r\geqslant 0} r^4\exp\left(-2r^2/t\right)
\end{aligned}$$

$$= \frac{1}{e^2(2\pi)^{2d}t^{2d+2}}\,.$$

Here $r = \|X - x\|$, and $r^4 e^{-2r^2/t}$ attains its maximum $t^2 e^{-2}$ at $r = \sqrt{t}$. The Gaussian normalization contributes the factor t^{-2d}.

Using the fourth-moment bound above and $\int p_t = 1$, we obtain

$$\int_{\mathbb{R}^d} \mathbb{E}\|\hat{s}_t(x)\|^4 p_t(x)\,\mathrm{d}x \leqslant \frac{\rho_n^{-4}}{e^2(2\pi)^{2d}t^{2d+2}} = \frac{n^4}{e^2(\log n)^4 t^2},$$

where the equality uses $\rho_n = (\log n)/[n(2\pi t)^{d/2}]$. Consequently, for every $t > 0$ and $n \geqslant 3$,

$$\begin{aligned}\mathrm{III} &\lesssim n^{-\alpha/2}t^{-1}\left(3d^2 + \frac{n^4}{e^2(\log n)^4}\right)^{1/2} \\ &\lesssim_d t^{-1}\left(n^{-\alpha/2} + \frac{n^{2-\alpha/2}}{(\log n)^2}\right).\end{aligned}$$

Fixing $\alpha = 8$ in the pointwise concentration event (and choosing the corresponding fixed c_α) therefore gives

$$\mathrm{III} \leqslant C_d n^{-2}t^{-1}, \qquad t > 0, \tag{6}$$

for a constant C_d independent of n and t. Writing $\kappa = (\gamma+1)/(d+\gamma+1) \in (0,1)$, for $0 < t \leqslant 1$ we have

$$n^{-2}t^{-1} \leqslant n^{-\kappa}t^{-1-d\kappa/2}, \qquad n^{-2}t^{-1} \leqslant n^{-1}t^{-1-d/2}.$$

Thus the bad-event term is below the small-time leading bounds in both tail regimes. For $t \geqslant 1$, it is also below the respective large-time comparison terms $n^{-\kappa}t^{-1}$ and $n^{-1}t^{-1}$. We combine this estimate with the bulk and tail bounds in (7) for polynomial tails and in Appendix E for exponential tails.

The two bounds in Step 2, expressed using the same $A_t = C_\gamma(t)^{d/D}\rho_n^\kappa$, give

$$\mathrm{II} \lesssim \frac{\varepsilon^{-1}A_t^{1-\varepsilon}}{t} + \frac{A_t + \varepsilon^{-1}A_t^{1-\varepsilon}}{t\log n}.$$

Combining this with Step 1 and (6), for $n \geqslant 3$ we obtain

$$\mathbb{E}\big[\|\hat{s}_t - s_t\|^2_{L^2(p_t)}\big] \lesssim \frac{1}{t}\left[\frac{A_t + \varepsilon^{-1}A_t^{1-\varepsilon}}{\log n} + \varepsilon^{-1}A_t^{1-\varepsilon} + n^{-2}\right]. \tag{7}$$

Choose ε by setting

$$m = \lceil \log n \rceil \geqslant 2, \qquad \varepsilon = 1/m, \qquad \eta = d\kappa/2.$$

The score moment bound used in Hölder's inequality has a constant independent of m after extracting the factor m: indeed, $\big(\mathbb{E}_{p_t}\|s_t\|^{2m}\big)^{1/m} \leqslant dt^{-1}[(2m-1)!!]^{1/m} \leqslant 2dmt^{-1}$. For the global envelope $C_\gamma(t) = C_* \max\{1, t^{(\gamma+1)/2}\}$, define

$$B_n := C_*^{d/D}\left(\frac{\log n}{n(2\pi)^{d/2}}\right)^\kappa.$$

Then the dependence on t is exactly

$$A_t = \begin{cases} B_n t^{-\eta}, & 0 < t \leqslant 1, \\ B_n, & t \geqslant 1, \end{cases} \qquad \frac{d(\gamma+1)}{2D} - \frac{d\kappa}{2} = 0.$$

In particular $A_t \geqslant B_n \gtrsim n^{-\kappa}$, so $A_t^{-1/m} \lesssim n^{\kappa/m} \leqslant e^\kappa$ uniformly in $t > 0$ and $n \geqslant 3$. Consequently $A_t^{1-\varepsilon} \lesssim A_t$, $m \lesssim \log n$, and $n^{-2} \lesssim A_t$. Substitution into (7) yields

$$\mathbb{E}\big[\|\hat{s}_t - s_t\|^2_{L^2(p_t)}\big] \lesssim \frac{(\log n)A_t}{t} \lesssim (\log n)^{1+\kappa}n^{-\kappa}\begin{cases} t^{-1-\eta}, & 0 < t \leqslant 1, \\ t^{-1}, & t \geqslant 1. \end{cases}$$

This proves Theorem 1 uniformly for all $t > 0$. □

D.2 Proof of Theorem 2

Proof. Let $\kappa = (\gamma+1)/(d+\gamma+1)$ and $\eta = d\kappa/2$. For any $0 < t_0 < T$, integrating the bound in Theorem 1 over the intersections of $[t_0, T]$ with $(0,1]$ and $[1,\infty)$ gives

$$\begin{aligned}\int_{t_0}^{T} \mathbb{E}\big[\|\hat{s}_t - s_t\|_{L^2(p_t)}^2\big]\,\mathrm{d}t &\lesssim (\log n)^{1+\kappa} n^{-\kappa} \int_{t_0}^{T} \frac{\mathrm{d}t}{t(t\wedge 1)^{\eta}} \\ &= (\log n)^{1+\kappa} n^{-\kappa} H_\eta(t_0, T).\end{aligned}$$

Explicitly,

$$H_\eta(t_0,T) = \begin{cases} (t_0^{-\eta} - T^{-\eta})/\eta, & 0 < t_0 < T \leqslant 1, \\ (t_0^{-\eta} - 1)/\eta + \log T, & 0 < t_0 < 1 < T, \\ \log(T/t_0), & 1 \leqslant t_0 < T. \end{cases}$$

In particular, the contribution from times greater than 1 is logarithmic. The bad-event contribution satisfies, by (6),

$$\int_{t_0}^{T} \mathrm{III}(t)\,\mathrm{d}t \leqslant C_d n^{-2} \log(T/t_0), \qquad 0 < t_0 < T. \tag{8}$$

Since $H_\eta(t_0,T) \geqslant \log(T/t_0)$ and $n^{-2} \leqslant n^{-\kappa}$, this term is absorbed in the cumulative bound for every $0 < t_0 < T$. The sampling specialization $t_0 \leqslant 1 \leqslant T$ follows immediately. □

D.3 Proof of Theorem 3

Proof. It suffices to consider $0 < t_0 \leqslant 1$, since the case $t_0 \geqslant 1$ follows from $\mathrm{TV} \leqslant 1$. Firstly, the total variation distance between p_0 and p_{t_0} can be decomposed into a central part and a tail part by introducing a truncation radius $R > 0$. By Jensen's inequality, it holds that

$$\begin{aligned}\mathrm{TV}(p_0, p_{t_0}) &= \frac{1}{2}\int_{\mathbb{R}^d} |p_0(x) - p_{t_0}(x)|\,\mathrm{d}x \\ &= \frac{1}{2}\int_{\|x\|\leqslant R} |p_0(x) - p_{t_0}(x)|\,\mathrm{d}x + \frac{1}{2}\int_{\|x\|>R} |p_0(x) - p_{t_0}(x)|\,\mathrm{d}x \\ &\leqslant \frac{1}{2}\sqrt{\mathrm{Vol}(\mathbb{B}_R)}\sqrt{\int_{\mathbb{R}^d} |p_0(x) - p_{t_0}(x)|^2\,\mathrm{d}x} + \frac{1}{2}\int_{\|x\|>R} (p_0(x) + p_{t_0}(x))\,\mathrm{d}x \\ &\lesssim R^{d/2}\|p_0 - p_{t_0}\|_{L^2} + \int_{\|x\|>R} p_0(x)\,\mathrm{d}x + \int_{\|x\|>R} p_{t_0}(x)\,\mathrm{d}x\,.\end{aligned} \tag{9}$$

For the second term, we have

$$\int_{\|x\|>R} p_0(x)\,\mathrm{d}x \leqslant C_0 \int_R^\infty r^{-(d+\gamma+1)} r^{d-1}\,\mathrm{d}r = C_0 \int_R^\infty r^{-\gamma-2}\,\mathrm{d}r \lesssim C_0 R^{-(\gamma+1)}\,.$$

For the second tail term, recall that $X_{t_0} = X_0 + \sqrt{t_0}Z$, where $X_0 \sim p_0$ and $Z \sim \mathcal{N}(0, I_d)$. For sufficiently large R (relative to $\sqrt{t_0}$), the heavy tail of X_0 dominates the Gaussian noise. Specifically,

$$\begin{aligned}\mathbb{P}(\|X_{t_0}\| > R) &\leqslant \mathbb{P}(\|X_0\| > R/2) + \mathbb{P}(\|\sqrt{t_0}Z\| > R/2) \\ &\leqslant C_0 (R/2)^{-(\gamma+1)} + 2^{d/2}\exp\left(-\frac{R^2}{16 t_0}\right)\,.\end{aligned}$$

Since polynomial decay dominates exponential decay for large R, we have:

$$\text{Tail Error} \lesssim R^{-(\gamma+1)}\,. \tag{10}$$

The following step follows Plancherel's theorem,

$$\begin{aligned}\|p_0 - p_{t_0}\|_{L^2}^2 &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} |\mathcal{F}[p_0](\omega) - \mathcal{F}[p_{t_0}](\omega)|^2\,\mathrm{d}\omega \\ &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} |\mathcal{F}[p_0](\omega)|^2 |1 - \phi_{t_0}(\omega)|^2\,\mathrm{d}\omega\,,\end{aligned}$$

where $\phi_{t_0}(\omega) = \exp(-\frac{t_0\|\omega\|^2}{2})$. We split the integral into low and high frequencies at threshold $K = t_0^{-1/2}$.

Note that $|1 - \phi_{t_0}(\omega)| \leqslant 2$, it then follows that

$$\begin{aligned}\int_{\|\omega\|\geqslant t_0^{-1/2}} |\mathcal{F}[p_0]|^2|1-\phi_{t_0}|^2 \,\mathrm{d}\omega &\leqslant 4\int_{\|\omega\|\geqslant t_0^{-1/2}} |\mathcal{F}[p_0]|^2 \frac{\|\omega\|^{2\beta}}{\|\omega\|^{2\beta}}\,\mathrm{d}\omega \\ &\leqslant 4(t_0^{-1/2})^{-2\beta}\int_{\mathbb{R}^d} \|\omega\|^{2\beta}|\mathcal{F}[p_0]|^2\,\mathrm{d}\omega \\ &\lesssim t_0^{\beta} L^2\,.\end{aligned}$$

Since $|1 - e^{-x}|^2 \leqslant x^2$ for $x \geqslant 0$, we then have $|1 - \phi_{t_0}(\omega)|^2 \leqslant (\frac{t_0\|\omega\|^2}{2})^2 = \frac{t_0^2\|\omega\|^4}{4}$. Thus,

$$\begin{aligned}\int_{\|\omega\|<t_0^{-1/2}} |\mathcal{F}[p_0]|^2|1-\phi_{t_0}|^2\,\mathrm{d}\omega &\lesssim t_0^2\int_{\|\omega\|<t_0^{-1/2}} |\mathcal{F}[p_0]|^2\|\omega\|^4\,\mathrm{d}\omega \\ &= t_0^2\int_{\|\omega\|<t_0^{-1/2}} |\mathcal{F}[p_0]|^2\|\omega\|^{2\beta}\|\omega\|^{4-2\beta}\,\mathrm{d}\omega\,.\end{aligned}$$

When $\beta \leqslant 2$, it holds that $4 - 2\beta \geqslant 0$. Thus, $\|\omega\|^{4-2\beta} \leqslant (t_0^{-1/2})^{4-2\beta} = t_0^{\beta-2}$, and

$$\int_{\|\omega\|<t_0^{-1/2}} |\mathcal{F}[p_0]|^2|1-\phi_{t_0}|^2\,\mathrm{d}\omega \lesssim t_0^2\cdot t_0^{\beta-2}\int_{\mathbb{R}^d}\|\omega\|^{2\beta}|\mathcal{F}[p_0]|^2 \lesssim t_0^{\beta}L^2\,.$$

Combining high and low frequencies, we obtain

$$\|p_0 - p_{t_0}\|_{L^2} \lesssim \sqrt{t_0^{\beta}} = t_0^{\beta/2}\,. \tag{11}$$

Step 3: Optimal Radius Selection. Plugging (10) and (11) back into the display (9) yields

$$\mathrm{TV}(p_0, p_{t_0}) \lesssim R^{d/2}t_0^{\beta/2} + R^{-(\gamma+1)}\,.$$

To optimize this bound, we balance the two contributions by choosing R such that

$$R^{\frac{d}{2}+\gamma+1} \asymp t_0^{-\beta/2}\,,$$

which yields the choice

$$R \asymp t_0^{-\frac{\beta}{d+2(\gamma+1)}}\,.$$

Plugging this optimal R back into the previous display gives

$$\mathrm{TV}(p_0, p_{t_0}) \lesssim \left(t_0^{-\frac{\beta}{d+2(\gamma+1)}}\right)^{-(\gamma+1)} = t_0^{\frac{\beta(\gamma+1)}{d+2(\gamma+1)}}$$

as desired. □

D.4 Proof of Theorem 4

Proof. Condition on the observed sample $\mathcal{D}_n = D$ throughout the stochastic-analysis part of the argument. For fixed D, the thresholded estimator $\hat{s}_t^D$ is a bounded Borel-measurable drift on $[t_0, T] \times \mathbb{R}^d$: each summand in $\nabla\hat{p}_t^D(x)$ is uniformly bounded in x, and the threshold condition $\hat{p}_t^D(x) \geqslant \rho_n$ on the active region keeps the ratio bounded. Hence the learned reverse SDE admits a unique strong solution by Veretennikov's theorem for bounded Borel drifts (Veretennikov, 1981). Since $p_t = p_0 * \varphi_t$ is smooth and strictly positive for every $t > 0$, the Brownian noising process admits the usual time-reversal description on $[t_0, T]$. We then apply the stopped-process version of Girsanov's theorem conditionally on D, as in (Chen et al., 2023c, Section 5.2) and (Zhang et al., 2024, Lemma E.3(2)); the required energy condition is finite because $\hat{s}_t^D$ is bounded and the exact score has finite second moment on $[t_0, T]$. More explicitly, let τ_M be the first time at which the accumulated squared drift difference reaches M. Girsanov's theorem applied up to τ_M gives the stopped path-space KL bound. Finite integrated score error implies $\tau_M \uparrow T - t_0$ almost surely as $M \to \infty$; lower

semicontinuity of relative entropy and monotone convergence of the stopped drift-error integral then give the full conditional bound. Averaging over D, followed by data processing, Pinsker's inequality, and Jensen's inequality, yields the asserted expected total-variation bound.

Recall that the backward process $X^{\leftarrow}$ is approximated by $\hat{Y}_t$. Let $(\tilde{Y}_t)_{t\in[0,T]}$ be $(\hat{Y}_t)_{t\in[0,T]}$ replacing $\hat{Y}_0 \sim \mathcal{N}(0, TI_d)$ by $\tilde{Y}_0 \sim p_T$, that is, $(\tilde{Y}_t)_{t\in[0,T]}$ satisfies:

$$\mathrm{d}\tilde{Y}_t = \hat{s}_{T-t}(\tilde{Y}_t)\,\mathrm{d}t + \mathrm{d}W_t, \quad \tilde{Y}_0 \sim p_T\,.$$

By the triangle inequality, it holds that

$$\begin{aligned}&\mathbb{E}\left[\mathrm{TV}(\mathcal{L}(X_0), \mathcal{L}(\hat{Y}_{T-t_0}))\right]\\ &\leqslant \mathrm{TV}(\mathcal{L}(X_0), \mathcal{L}(X^{\leftarrow}_{T-t_0})) + \mathbb{E}[\mathrm{TV}(\mathcal{L}(X^{\leftarrow}_{T-t_0}), \mathcal{L}(\tilde{Y}_{T-t_0}))] + \mathbb{E}[\mathrm{TV}(\mathcal{L}(\tilde{Y}_{T-t_0}), \mathcal{L}(\hat{Y}_{T-t_0}))]\,.\end{aligned}$$

By Theorem 3, $\mathrm{TV}(\mathcal{L}(X_0), \mathcal{L}(X^{\leftarrow}_{T-t_0})) \lesssim n^{-\frac{2\beta(\gamma+1)}{4\beta(d+\gamma+1)+d(d+2(\gamma+1))}}$. For the second term, by Pinsker's inequality and data-processing inequality, we obtain

$$\mathrm{TV}(\mathcal{L}(X^{\leftarrow}_{T-t_0}), \mathcal{L}(\tilde{Y}_{T-t_0})) \lesssim \sqrt{\mathrm{KL}(\mathcal{L}(X^{\leftarrow}_{T-t_0}) \| \mathcal{L}(\tilde{Y}_{T-t_0}))} \leqslant \sqrt{\mathrm{KL}(\mathbb{P}^D_{X^{\leftarrow}} \| \mathbb{P}^D_{\tilde{Y}})}$$

where $\mathbb{P}^D_{X^{\leftarrow}}$ and $\mathbb{P}^D_{\tilde{Y}}$ are the conditional path laws of $(X^{\leftarrow}_t)_{t\in[0,T-t_0]}$ and $(\tilde{Y}_t)_{t\in[0,T-t_0]}$. Then, we have

$$\begin{aligned}\mathrm{KL}(\mathbb{P}^D_{X^{\leftarrow}} \| \mathbb{P}^D_{\tilde{Y}}) &\leqslant \frac{1}{2}\int_0^{T-t_0}\int_{\mathbb{R}^d} \|\hat{s}^D_{T-t}(x) - s_{T-t}(x)\|^2 p_{T-t}(x)\,\mathrm{d}x\,\mathrm{d}t\\ &= \frac{1}{2}\int_{t_0}^{T}\int_{\mathbb{R}^d} \|\hat{s}_t(x) - s_t(x)\|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t\end{aligned}$$

Then $\mathbb{E}[\mathrm{KL}(\mathbb{P}^D_{X^{\leftarrow}} \| \mathbb{P}^D_{\tilde{Y}})]$ can be bounded by Theorem 2, and by Jensen's inequality, we obtain

$$\mathbb{E}[\mathrm{TV}(\mathcal{L}(X^{\leftarrow}_{T-t_0}), \mathcal{L}(\tilde{Y}_{T-t_0}))] \lesssim \mathbb{E}\left[\sqrt{\mathrm{KL}(\mathbb{P}^D_{X^{\leftarrow}} \| \mathbb{P}^D_{\tilde{Y}})}\right] \leqslant \sqrt{\mathbb{E}[\mathrm{KL}(\mathbb{P}^D_{X^{\leftarrow}} \| \mathbb{P}^D_{\tilde{Y}})]}\,.$$

It follows that

$$\mathbb{E}[\mathrm{TV}(\mathcal{L}(X^{\leftarrow}_{T-t_0}), \mathcal{L}(\tilde{Y}_{T-t_0}))] \lesssim (\log n)^{(1+\kappa)/2} n^{-\kappa/2} H_\eta(t_0, T)^{1/2}\,.$$

Here $\kappa = (\gamma+1)/(d+\gamma+1)$ and $\eta = d\kappa/2$. For the chosen cutoff $t_0 = n^{-b}$ with fixed $b > 0$ and horizon $T = n^{\frac{4\beta(\gamma+1)}{4\beta(d+\gamma+1)+d(d+2(\gamma+1))}}$, we have $H_\eta(t_0, T) \lesssim t_0^{-\eta} + \log T$, with the $\log T$ term lower order. Therefore

$$\mathbb{E}[\mathrm{TV}(\mathcal{L}(X^{\leftarrow}_{T-t_0}), \mathcal{L}(\tilde{Y}_{T-t_0}))] \lesssim \mathrm{polylog}(n) n^{-\frac{2\beta(\gamma+1)}{4\beta(d+\gamma+1)+d(d+2(\gamma+1))}}\,.$$

Under the same conditional Pinsker–Girsanov argument, the additive bad-event term in (8) contributes at most

$$C_d n^{-1}\sqrt{\log(T/t_0)}$$

to the resulting upper bound on expected total variation, by $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$. For the polynomial choices of T and t_0^{-1} in the sampling theorems, this is $\mathcal{O}(n^{-1}\sqrt{\log n})$, lower order than both stated sampling rates.

Since the only difference of $\tilde{Y}_{T-t_0}$ and $\hat{Y}_{T-t_0}$ is their initial distribution, we have

$$\mathbb{E}[\mathrm{TV}(\mathcal{L}(\tilde{Y}_{T-t_0}), \mathcal{L}(\hat{Y}_{T-t_0}))] \leqslant \mathrm{TV}(\mathcal{L}(X_T), \mathcal{N}(0, TI_d)) \lesssim \frac{1}{\sqrt{T}}\,.$$

The last inequality follows from Lemma 19. By taking $T = n^{\frac{4\beta(\gamma+1)}{4\beta(d+\gamma+1)+d(d+2(\gamma+1))}}$, we reach at

$$\mathbb{E}[\mathrm{TV}(\mathcal{L}(\tilde{Y}_{T-t_0}), \mathcal{L}(\hat{Y}_{T-t_0}))] \lesssim n^{-\frac{2\beta(\gamma+1)}{4\beta(d+\gamma+1)+d(d+2(\gamma+1))}}\,.$$

This completes the proof. □

Lemma 19. *Let p_0 be a probability density on $\mathbb{R}^d$ with finite first moment $M_1 := \int_{\mathbb{R}^d} \|z\| \, p_0(z) \, \mathrm{d}z < \infty$. Let φ_T denote the density of $\mathcal{N}(0, TI_d)$, and define $p_T := p_0 * \varphi_T$. Then, it holds that*

$$\mathrm{TV}(p_T, \varphi_T) \leqslant \frac{M_1}{2\sqrt{T}}.$$

E Proof of Section 3.2

Proposition 12, 13 and Lemma 11 are independent of the behavior of tail decay, so they remain valid without change. For the global convergence theorem for score estimation, we revisit the proof step by step: only Lemmas 16, 17, and 23 require modification, after which the desired result follows immediately.

Lemma 20 (Modification of Lemma 23)**.** *Under Assumption 1, there exist constants $C_{\exp}, c_{\exp} > 0$, depending only on d, γ, c_1, C_1, such that for all $t > 0$ and $x \in \mathbb{R}^d$,*

$$p_t(x) \leqslant C_{\exp}(1+t)^{-d/2} \exp\left[-c_{\exp}\left(\frac{\|x\|}{\sqrt{1+t}}\right)^{\min\{\gamma,2\}}\right].$$

E.1 Proof of Theorem 5

Proof. The proof follows the same threshold-based decomposition strategy as the proof of Theorem 1. We decompose the weighted MISE into contributions from the high-density bulk region $\mathcal{G}_1$ and the low-density tail region $\mathcal{G}_2$.

Recall the decomposition from (4) and (5). On the bulk region $\mathcal{G}_1$, the error includes a gradient-variance term which scales as $|\mathcal{G}_1| \cdot \frac{1}{nt(2\pi t)^{d/2}}$. Under Assumption 1 (exponential decay), the following lemma adapting Lemma 16 shows that the size of $\mathcal{G}_1$ grows only logarithmically in n, with an explicit time-dependent volume factor.

Lemma 21 (Modification of Lemma 16)**.** *Under Assumption 1, for the threshold $\rho_n = \frac{\log n}{n(2\pi t)^{d/2}}$ with $n \geqslant 3$, the volume of the high-density region $\mathcal{G}_1 = \{x : p_t(x) \geqslant c_\alpha \rho_n\}$ satisfies*

$$|\mathcal{G}_1| \lesssim V_t, \qquad \int_{\mathcal{G}_1} \|s_t(x)\|^2 \mathrm{d}x \lesssim \frac{m}{t} \rho_n^{-1/m} V_t^{1-1/m},$$

where $V_t := (1+t)^{d/2} (\log n)^{d/\min\{\gamma,2\}}$ and $m > 1$; the implicit constants are uniform in n, t, m.

Consequently, the integrated error over the bulk satisfies

$$\mathrm{I} \lesssim \mathrm{polylog}(n) \frac{(1+t)^{d/2}}{nt^{1+d/2}}.$$

For the tail region $\mathcal{G}_2$, the error is controlled by the probability mass of the tail.

Lemma 22 (Modification of Lemma 17)**.** *Under Assumption 1, for the threshold $\rho_n = \frac{\log n}{n(2\pi t)^{d/2}}$ with $n \geqslant 3$, the probability mass of the tail region $\mathcal{G}_2 = \{x : p_t(x) < c_\alpha \rho_n\}$ satisfies*

$$\int_{\mathcal{G}_2} p_t \lesssim \min\{1, \rho_n V_t\}, \qquad \int_{\mathcal{G}_2} \|s_t\|^2 p_t \lesssim \frac{m}{t} (\rho_n V_t)^{1-1/m},$$

with V_t as above and $m > 1$. No assumption $\rho_n < 1$ is needed.

Set $B_t = \rho_n V_t$ and $m = \lceil \log n \rceil$. Substitution of the two lemmas into the same I/II decomposition, together with $\mathrm{III} \lesssim n^{-2}/t$ from (6), gives

$$\mathbb{E}\|\hat{s}_t - s_t\|^2_{L^2(p_t)} \lesssim \frac{1}{t}\left[\frac{B_t + mB_t^{1-1/m}}{\log n} + mB_t^{1-1/m} + n^{-2}\right].$$

Since $B_t \gtrsim n^{-1}$, $B_t^{-1/m} \lesssim 1$ uniformly in $t > 0$. Hence

$$\mathbb{E}\|\hat{s}_t - s_t\|^2_{L^2(p_t)} \lesssim \frac{(\log n) B_t}{t} \asymp \frac{(\log n)^{2+d/\min\{\gamma,2\}}}{n} \frac{(1+t)^{d/2}}{t^{1+d/2}} \lesssim \frac{(\log n)^{2+d/\min\{\gamma,2\}}}{nt(t \wedge 1)^{d/2}}.$$

The low-density mass is controlled by the threshold times V_t. This proves the all-time bound. □

E.2 Proof of Theorem 6

Proof. Integrating the all-time bound in Theorem 5, for any $0 < t_0 < T$ we obtain

$$\int_{t_0}^{T} \mathbb{E}\|\hat{s}_t - s_t\|^2_{L^2(p_t)}\,\mathrm{d}t \lesssim \frac{(\log n)^{2+d/\min\{\gamma,2\}}}{n} \int_{t_0}^{T} \frac{\mathrm{d}t}{t(t\wedge 1)^{d/2}}$$
$$= \frac{(\log n)^{2+d/\min\{\gamma,2\}}}{n} \left[\frac{2}{d}\big[(t_0\wedge 1)^{-d/2} - (T\wedge 1)^{-d/2}\big] + \log\frac{T\vee 1}{t_0\vee 1}\right].$$

The integral covers all three positions of $[t_0, T]$ relative to 1. The bad-event contribution $C_d n^{-2}\log(T/t_0)$ is absorbed in this bound because the time integral above is at least $\log(T/t_0)$.

This yields the desired result. □

E.3 Proof of Theorem 7

Proof. The proof proceeds analogously to that of Theorem 3, utilizing a truncation radius R to decompose the total variation distance.

$$\mathrm{TV}(p_0, p_{t_0}) \lesssim \frac{1}{2}R^{d/2}\|p_0 - p_{t_0}\|_{L^2} + \frac{1}{2}\int_{\|x\|>R}(p_0(x) + p_{t_0}(x))\mathrm{d}x.$$

For the smoothing error in the central region, the Sobolev assumption yields the same bound as in the polynomial case (derived via Plancherel's theorem)

$$\|p_0 - p_{t_0}\|_{L^2} \lesssim t_0^{\beta/2}.$$

For $0 < t_0 \leqslant 1$, the envelope in Lemma 20 gives $\int_{\|x\|>R}(p_0 + p_{t_0}) \lesssim e^{-cR^{\min\{\gamma,2\}}}$ for $R \geqslant 1$. Choose $R = A(\log(e/t_0))^{1/\min\{\gamma,2\}}$ with fixed A sufficiently large. The tail is then $\mathcal{O}(t_0^{\beta/2})$, and the central-region estimate yields

$$\mathrm{TV}(p_0, p_{t_0}) \lesssim t_0^{\beta/2}(\log(e/t_0))^{d/(2\min\{\gamma,2\})}.$$

□

E.4 Proof of Theorem 8

Proof. The proof proceeds by decomposing the total variation error into three components—early stopping bias, score matching error, and initialization error following the same conditional well-posedness and stopped-process Girsanov argument as in the proof of Theorem 4.

$$\mathbb{E}[\mathrm{TV}(p_0, \mathcal{L}(\hat{Y}_{T-t_0}))] \leqslant \underbrace{\mathrm{TV}(p_0, p_{t_0})}_{\mathrm{I}} + \underbrace{\mathbb{E}[\mathrm{TV}(p_{t_0}, \mathcal{L}(\tilde{Y}_{T-t_0}))]}_{\mathrm{II}} + \underbrace{\mathrm{TV}(p_T, \mathcal{N}(0, TI_d))}_{\mathrm{III}}.$$

For the early stopping bias I, Theorem 7 establishes that under exponential tails, the error is dominated by the Sobolev smoothing term $t_0^{\beta/2}$, up to the factor $(\log(e/t_0))^{d/(2\min\{\gamma,2\})}$.

For the score matching error II, applying Pinsker's inequality and the cumulative MSE bound from Theorem 6, with $0 < t_0 \leqslant 1 \leqslant T$, yields

$$\mathbb{E}\big[\mathrm{TV}\big(p_{t_0}, \mathcal{L}(\tilde{Y}_{T-t_0})\big)\big] \lesssim (\log n)^{1+d/(2\min\{\gamma,2\})} n^{-1/2}\big(t_0^{-d/2} + \log T\big)^{1/2}.$$

For $t_0 = n^{-2/(2\beta+d)}$ and $T = n^{2\beta/(2\beta+d)}$, this time range holds and $\log T$ is lower order than $t_0^{-d/2}$. Thus the score contribution remains $\mathrm{polylog}(n) n^{-1/2} t_0^{-d/4}$.

For the initialization error III, Lemma 19 ensures a decay of $T^{-1/2}$, which vanishes for large T. Balancing the bias $t_0^{\beta/2}$ against the score error $n^{-1/2}t_0^{-d/4}$ up to logarithmic factors yields the rate-balancing stopping time $t_0 \asymp n^{-\frac{2}{2\beta+d}}$ and the stated convergence rate. □

F Proof of Auxiliary Lemma for Upper Bounds

F.1 Proof of Lemmas for Polynomial Decay

Proof of Lemma 11. Note that $\hat{p}_t$ is an unbiased estimator, and Proposition 12 derives the upper bound of $\mathrm{MSE}(\hat{p}_t)$,

$$\mathrm{MSE}(\hat{p}_t(x)) \leqslant \frac{p_t(x)}{n(2\pi t)^{d/2}}\,.$$

It also holds that

$$\max_{1\leqslant i\leqslant n}\frac{1}{n}\left(\varphi_t(X^{(i)}-x)-\mathbb{E}[\varphi_t(X^{(i)}-x)]\right)\leqslant\frac{2}{n}\|\varphi_t(x)\|_\infty=\frac{2}{n(2\pi t)^{d/2}}\,.$$

By Bernstein's Inequality, we have

$$\mathbb{P}\left(|\hat{p}_t(x)-p_t(x)|\geqslant\varepsilon\right)\leqslant 2\exp\left(-\frac{\varepsilon^2}{2p_t(x)/(n(2\pi t)^{d/2})+4\varepsilon/(3n(2\pi t)^{d/2})}\right)\,.$$

Then, with probability at least $1-\delta$, it holds that

$$|\hat{p}_t(x)-p_t(x)|<\max\left\{\sqrt{\frac{4p_t(x)\log(2/\delta)}{n(2\pi t)^{d/2}}},\frac{8\log(2/\delta)}{3n(2\pi t)^{d/2}}\right\}\,.$$

Let $\delta=n^{-\alpha}$, we obtain that

$$\mathbb{P}\left(|\hat{p}_t(x)-p_t(x)|<\max\left\{\sqrt{\frac{4p_t(x)\log(2n^\alpha)}{n(2\pi t)^{d/2}}},\frac{8\log(2n^\alpha)}{3n(2\pi t)^{d/2}}\right\}\right)\geqslant 1-n^{-\alpha}$$

as desired. For $n\geqslant 2$, $\log(2n^\alpha)\leqslant(\alpha+1)\log n$, let $C_\alpha=\max\{\sqrt{8(\alpha+1)},16(\alpha+1)/3\}$ be the uniform constant, then we reach the desired result in Lemma 11. □

To proceed to proof of Lemma 16 and 17, we need the following lemma

Lemma 23. *Under Assumption 2, there exists $C_*\geqslant 1$, depending only on d, γ, and C_2, such that for every $x\in\mathbb{R}^d$ and $t>0$,*

$$p_t(x)\leqslant C_\gamma(t)(1+\|x\|^2)^{-(1+d+\gamma)/2},\qquad C_\gamma(t)=C_*\max\{1,t^{(1+\gamma)/2}\}.$$

Proof of Lemma 23. First suppose $\|x\|>1$ and let $R=\|x\|/2$. We decompose the density p_t as

$$p_t(x)=(p_0*\varphi_t)(x)=\underbrace{\int_{\|y\|\leqslant R}p_0(y)\varphi_t(x-y)\,\mathrm{d}y}_{\mathrm{I}}+\underbrace{\int_{\|y\|>R}p_0(y)\varphi_t(x-y)\,\mathrm{d}y}_{\mathrm{II}}\,.$$

The first term on the right side can be bounded as follows:

$$\begin{aligned}\mathrm{I}&\leqslant\int_{\|y\|\leqslant R}p_0(y)\sup_{\|z\|\leqslant R}\varphi_t(x-z)\,\mathrm{d}y\\&=\int_{\|y\|\leqslant R}p_0(y)\frac{1}{(2\pi t)^{d/2}}\exp\left(-\frac{\|x\|^2}{8t}\right)\mathrm{d}y\\&\leqslant\frac{1}{(2\pi t)^{d/2}}\exp\left(-\frac{\|x\|^2}{8t}\right).\end{aligned}$$

Denote

$$F(t,r):=\frac{1}{(2\pi t)^{d/2}}\exp(-r/(8t))(1+r)^{(1+\gamma+d)/2},\qquad r=\|x\|^2.$$

For fixed $r\geqslant 1$, the maximum point of $F(t,r)$ over $t>0$ is $t(r)=r/(4d)$. It follows that

$$\sup_{0<t\leqslant 1,\,r\geqslant 1}F(t,r)=\max\left\{\sup_{1\leqslant r\leqslant 4d}\left(\frac{2d}{\pi}\right)^{d/2}e^{-d/2}(1+r)^{\frac{1+\gamma+d}{2}}r^{-d/2},\right.$$

$$\sup_{r>4d}(2\pi)^{-d/2}e^{-r/8}(1+r)^{\frac{1+\gamma+d}{2}}\Bigg\} =: C_1(\gamma,d) < \infty.$$

The first supremum is over a compact interval, and the exponential factor makes the second finite. Thus $C_1(\gamma,d)$ is independent of t. For fixed $t>1$, the maximum point of $F(t,r)$ over $r \geqslant 1$ is $r(t) = 4t(1+\gamma+d)-1$, and hence

$$\begin{aligned}\sup_{r\geqslant 1} F(t,r) &= \frac{t^{(1+\gamma)/2}}{(2\pi)^{d/2}}\exp\left(-\frac{4t(1+\gamma+d)-1}{8t}\right)[4(1+\gamma+d)]^{(1+\gamma+d)/2}\\ &\leqslant \frac{t^{(1+\gamma)/2}}{(2\pi)^{d/2}}e^{-\frac{1+\gamma+d}{4}}[4(1+\gamma+d)]^{\frac{1+\gamma+d}{2}} =: \widetilde{C}_2(\gamma,d)t^{(1+\gamma)/2}.\end{aligned}$$

Therefore, $\mathrm{I} \leqslant C_1(\gamma,d)(1+\|x\|^2)^{-(1+\gamma+d)/2}$ when $0<t\leqslant 1$, and $\mathrm{I} \leqslant \widetilde{C}_2(\gamma,d)t^{(1+\gamma)/2}(1+\|x\|^2)^{-(1+\gamma+d)/2}$ when $t>1$. For the second term, Assumption 2 gives

$$\begin{aligned}\mathrm{II} &\leqslant \int_{\|y\|>R}\frac{C_2}{(1+\|y\|^2)^{(1+\gamma+d)/2}}\varphi_t(x-y)\,\mathrm{d}y\\ &\leqslant \frac{2^{1+\gamma+d}C_2}{(1+\|x\|^2)^{(1+\gamma+d)/2}}.\end{aligned}$$

Combining these two terms, for $\|x\|>1$ we obtain

$$p_t(x)(1+\|x\|^2)^{(1+\gamma+d)/2} \leqslant \begin{cases} C_1(\gamma,d)+2^{1+\gamma+d}C_2, & 0<t\leqslant 1,\\ \widetilde{C}_2(\gamma,d)t^{(1+\gamma)/2}+2^{1+\gamma+d}C_2, & t>1.\end{cases}$$

For $\|x\|\leqslant 1$, convolution preserves the upper bound $\|p_t\|_\infty \leqslant \|p_0\|_\infty \leqslant C_2$, so

$$p_t(x)(1+\|x\|^2)^{(1+\gamma+d)/2} \leqslant 2^{(1+\gamma+d)/2}C_2.$$

Taking

$$C_* := \max\left\{1,\ C_1(\gamma,d)+2^{1+\gamma+d}C_2,\ \widetilde{C}_2(\gamma,d)+2^{1+\gamma+d}C_2,\ 2^{(1+\gamma+d)/2}C_2\right\},$$

we conclude that $p_t(x) \leqslant C_\gamma(t)(1+\|x\|^2)^{-(1+\gamma+d)/2}$ for all $x\in\mathbb{R}^d$ and $t>0$, where

$$C_\gamma(t) = \begin{cases} C_*, & 0<t\leqslant 1,\\ C_* t^{(1+\gamma)/2}, & t>1.\end{cases}$$

□

Proof of Lemma 16. By Lemma 23, we have

$$p_t(x) \leqslant \frac{C_\gamma(t)}{(1+\|x\|^2)^{(1+\gamma+d)/2}}.$$

Since $\rho_n = \frac{\log n}{n(2\pi t)^{d/2}}$, then $p_t(x) > c_\alpha\rho_n$ implies that

$$\|x\| \leqslant (C_\gamma(t)/(c_\alpha\rho_n))^{\frac{1}{1+\gamma+d}} \lesssim \left(\frac{C_\gamma(t)}{\rho_n}\right)^{\frac{1}{1+\gamma+d}}.$$

Therefore, we obtain

$$|\mathcal{G}_1| \leqslant \left|\left\{x : \|x\| \leqslant \left(\frac{C_\gamma(t)}{\rho_n}\right)^{\frac{1}{1+\gamma+d}}\right\}\right| \lesssim \left(\frac{C_\gamma(t)}{\rho_n}\right)^{\frac{d}{1+\gamma+d}}.$$

By Hölder's inequality, it holds that

$$\begin{aligned}\int_{\mathcal{G}_1}\|s_t(x)\|^2\mathrm{d}x &= \int_{\mathcal{G}_1}\|s_t(x)\|^2\frac{1}{p_t(x)}p_t(x)\,\mathrm{d}x\\ &\leqslant \left(\int_{\mathcal{G}_1}\|s_t(x)\|^{2m}p_t(x)\,\mathrm{d}x\right)^{1/m}\left(\int_{\mathcal{G}_1}p_t(x)^{-m'}p_t(x)\,\mathrm{d}x\right)^{1/m'},\end{aligned}$$

where $m, m' > 1$ and satisfies $\frac{1}{m} + \frac{1}{m'} = 1$. By Lemma 18, it holds that

$$\left(\int_{\mathcal{G}_1} \|s_t(x)\|^{2m} p_t(x)\,\mathrm{d}x\right)^{1/m} \lesssim mt^{-1}\,,$$

Invoking the bound of $|\mathcal{G}_1|$, we have

$$\begin{aligned}\left(\int_{\mathcal{G}_1} p_t(x)^{-m'} p_t(x)\mathrm{d}x\right)^{1/m'} &= \left(\int_{\mathcal{G}_1} p_t(x)^{1-m'}\mathrm{d}x\right)^{1/m'} \\ &\leqslant [|\mathcal{G}_1|(c_\alpha\rho_n)^{1-m'}]^{1/m'} \\ &\lesssim C_\gamma(t)^{\frac{d}{(1+\gamma+d)m'}}\rho_n^{-\frac{1}{m}-\frac{d}{(1+\gamma+d)m'}}\,.\end{aligned}$$

Combining these two terms, we have

$$\int_{\mathcal{G}_1} \|s_t(x)\|^2\mathrm{d}x \lesssim C_\gamma(t)^{\frac{d}{(1+\gamma+d)m'}} mt^{-1}\rho_n^{-\frac{1}{m}-\frac{d}{(1+\gamma+d)m'}}\,.$$

□

Proof of Lemma 17. Note that

$$\int_{\mathcal{G}_2} p_t(x)\,\mathrm{d}x = \mathbb{P}_{X\sim p_t}(p_t(X) \leqslant c_\alpha\rho_n)\,.$$

By Hölder's inequality and Lemma 18, it holds that

$$\begin{aligned}\int_{\mathcal{G}_2} \|s_t(x)\|^2 p_t(x)\,\mathrm{d}x &= \mathbb{E}_{X\sim p_t}\left[\|s_t(X)\|^2 \mathbb{1}\{p_t(X) \leqslant c_\alpha\rho_n\}\right] \\ &\leqslant \mathbb{E}_{X\sim p_t}[\|s_t(X)\|^{2m}]^{\frac{1}{m}}\mathbb{E}\left[\mathbb{1}\{p_t(X) \leqslant c_\alpha\rho_n\}\right]^{1-\frac{1}{m}} \\ &\lesssim mt^{-1}\mathbb{P}(p_t(X) \leqslant c_\alpha\rho_n)^{1-\frac{1}{m}}\,.\end{aligned}$$

Let $R_n = \left(\frac{C_\gamma(t)}{c_\alpha\rho_n}\right)^{\frac{1}{d+\gamma+1}}$, by Lemma 23, it holds that

$$\{x : \|x\| \geqslant R_n\} \subset \{x : p_t(x) \leqslant c_\alpha\rho_n\}\,.$$

Then, we can decompose the probability

$$\begin{aligned}\mathbb{P}(p_t(X) \leqslant c_\alpha\rho_n) &= \mathbb{P}(p_t(X) \leqslant c_\alpha\rho_n, \|X\| < R_n) + \mathbb{P}(p_t(X) \leqslant c_\alpha\rho_n, \|X\| \geqslant R_n) \\ &\leqslant \mathbb{P}(p_t(X) \leqslant c_\alpha\rho_n, \|X\| < R_n) + \mathbb{P}(\|X\| \geqslant R_n) \\ &\leqslant \int_{\|x\|<R_n} p_t(x)\mathbb{1}\{p_t(x) \leqslant c_\alpha\rho_n\}\,\mathrm{d}x + \int_{\|x\|\geqslant R_n} \frac{C_\gamma(t)}{(1+\|x\|^2)^{(d+\gamma+1)/2}}\mathrm{d}x \\ &\leqslant c_\alpha\rho_n \mathrm{Vol}(\mathbb{B}_{R_n}) + C_\gamma(t)\cdot\frac{2\pi^{d/2}}{\Gamma(d/2)}\int_{R_n}^\infty \rho^{-\gamma-2}\mathrm{d}\rho \qquad (12) \\ &\leqslant c_\alpha\rho_n\cdot\frac{\pi^{d/2}}{\Gamma(d/2+1)}R_n^d + C_\gamma(t)\cdot\frac{2\pi^{d/2}}{\Gamma(d/2)}R_n^{-\gamma-1} \\ &= c_\alpha^{\frac{\gamma+1}{d+\gamma+1}} C_\gamma(t)^{\frac{d}{d+\gamma+1}}\left(\frac{\pi^{d/2}}{\Gamma(d/2+1)} + \frac{2\pi^{d/2}}{\Gamma(d/2)}\right)\cdot\rho_n^{\frac{\gamma+1}{d+\gamma+1}}\,.\end{aligned}$$

Therefore, we obtain

$$\int_{\mathcal{G}_2} p_t(x)\,\mathrm{d}x \lesssim C_\gamma(t)^{\frac{d}{d+\gamma+1}}\rho_n^{\frac{\gamma+1}{d+\gamma+1}}\,,$$

and

$$\int_{\mathcal{G}_2} \|s_t(x)\|^2 p_t(x)\,\mathrm{d}x \lesssim \varepsilon^{-1}t^{-1}C_\gamma(t)^{\frac{d(1-\varepsilon)}{d+\gamma+1}}\rho_n^{\frac{(\gamma+1)(1-\varepsilon)}{(d+\gamma+1)}}$$

as desired. □

Proof of Lemma 19. We start with the definition of TV distance

$$\mathrm{TV}(p_T, \varphi_T) = \frac{1}{2} \int_{\mathbb{R}^d} |p_T(x) - \varphi_T(x)| \mathrm{d}x.$$

Substituting the convolution representation of p_T into the total variation distance and applying the triangle inequality yields

$$\begin{aligned}
|p_T(x) - \varphi_T(x)| &= \left| \int_{\mathbb{R}^d} p_0(z) \varphi_T(x - z) \mathrm{d}z - \varphi_T(x) \right| \\
&= \left| \int_{\mathbb{R}^d} p_0(z) \varphi_T(x - z) \mathrm{d}z - \int_{\mathbb{R}^d} p_0(z) \varphi_T(x) \mathrm{d}z \right| \\
&= \left| \int_{\mathbb{R}^d} p_0(z) \left(\varphi_T(x - z) - \varphi_T(x) \right) \mathrm{d}z \right| \\
&\leqslant \int_{\mathbb{R}^d} p_0(z) \left| \varphi_T(x - z) - \varphi_T(x) \right| \mathrm{d}z.
\end{aligned}$$

Integrating both sides over $\mathbb{R}^d$ with respect to x and then interchanging the order of integration gives

$$\int_{\mathbb{R}^d} |p_T(x) - \varphi_T(x)| \mathrm{d}x \leqslant \int_{\mathbb{R}^d} p_0(z) \left(\int_{\mathbb{R}^d} \left| \varphi_T(x - z) - \varphi_T(x) \right| \mathrm{d}x \right) \mathrm{d}z.$$

By the definition of TV distance, it holds that

$$\int_{\mathbb{R}^d} |\varphi_T(x - z) - \varphi_T(x)| \mathrm{d}x = 2 \, \mathrm{TV}(\varphi_T(\cdot - z) \| \varphi_T).$$

Substituting this into the previous inequality provides us with

$$\int_{\mathbb{R}^d} |p_T(x) - \varphi_T(x)| \mathrm{d}x \leqslant 2 \int_{\mathbb{R}^d} p_0(z) \, \mathrm{TV}(\varphi_T(\cdot - z), \varphi_T) \mathrm{d}z.$$

Dividing both sides by the factor 2 gives

$$\mathrm{TV}(p_T, \varphi_T) \leqslant \int_{\mathbb{R}^d} p_0(z) \, \mathrm{TV}(\varphi_T(\cdot - z), \varphi_T) \mathrm{d}z.$$

By Pinsker's inequality, we have

$$\mathrm{TV}(\varphi_T(\cdot - z), \varphi_T) \leqslant \sqrt{\frac{1}{2} \, \mathrm{KL}(\mathcal{N}(z, TI_d), \mathcal{N}(0, TI_d))} = \frac{\|z\|}{2\sqrt{T}}.$$

Therefore,

$$\begin{aligned}
\mathrm{TV}(p_T, \varphi_T) &\leqslant \int_{\mathbb{R}^d} p_0(z) \cdot \frac{\|z\|}{2\sqrt{T}} \mathrm{d}z \\
&= \frac{1}{2\sqrt{T}} \int_{\mathbb{R}^d} \|z\| p_0(z) \mathrm{d}z \\
&= \frac{M_1}{2\sqrt{T}}.
\end{aligned}$$

Since $M_1 < \infty$ by assumption, this completes the proof. □

F.2 Proof of Lemmas for Exponential Decay

Proof of Lemma 20.

$$p_t(x) = \int_{\mathbb{R}^d} p_0(y) \varphi_t(x - y) \, \mathrm{d}y.$$

To analyze the decay behavior for large $\|x\|$, we split the integration domain $\mathbb{R}^d$ into two regions. Let $R = \|x\|/2$. We define the regions $D_1 = \{y : \|x - y\| \geqslant R\}$ and $D_2 = \{y : \|x - y\| < R\}$.

We first consider the integral over the first region D_1. In this area, the distance between x and y satisfies $\|x-y\| \geqslant \|x\|/2$. Consequently, the Gaussian kernel $\varphi_t(x-y)$ decays rapidly. We can derive the upper bound for the kernel

$$\varphi_t(x-y) = \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{\|x-y\|^2}{2t}\right) \leqslant \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{\|x\|^2}{8t}\right).$$

Since $\int_{D_1} p_0(y)\mathrm{d}y \leqslant 1$, we then have

$$I_1 = \int_{D_1} p_0(y)\varphi_t(x-y)\,\mathrm{d}y \leqslant \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{\|x\|^2}{8t}\right).$$

Next, we turn to the second region D_2. In this area, the noise is small, satisfying $\|x-y\| < \|x\|/2$. By the reverse triangle inequality, the magnitude of y is bounded from below

$$\|y\| = \|x-(x-y)\| \geqslant \|x\| - \|x-y\| > \frac{\|x\|}{2}.$$

Using Assumption 1 on the initial density p_0, for any $y \in D_2$, we have

$$p_0(y) \leqslant C_1 \exp(-c_1\|y\|^\gamma) \leqslant C_1 \exp\left(-c_1\frac{\|x\|^\gamma}{2^\gamma}\right).$$

Similar to the previous step, we bound the contribution from this region by taking the supremum of p_0 over D_2 outside the integral. Since $\int_{D_2} \varphi_t(x-y)\mathrm{d}y \leqslant 1$, it holds that

$$I_2 = \int_{D_2} p_0(y)\varphi_t(x-y)\,\mathrm{d}y \leqslant C_1 \exp\left(-2^{-\gamma}c_1\|x\|^\gamma\right).$$

For $0 < t \leqslant 1$, we sharpen the bound on I_1 using $p_0 \leqslant C_1$. With $Z \sim \mathcal{N}(0, I_d)$,

$$\begin{aligned} I_1 &\leqslant C_1 \int_{\|x-y\|\geqslant R} \varphi_t(x-y)\,\mathrm{d}y \\ &= C_1\mathbb{P}(\|\sqrt{t}Z\| \geqslant R) \leqslant C_{\exp}e^{-c_{\exp}R^2/t}. \end{aligned}$$

Together with the bound on I_2 above, this gives

$$p_t(x) = I_1 + I_2 \leqslant C_{\exp}e^{-c_{\exp}\|x\|^{\min\{\gamma,2\}}}, \qquad 0 < t \leqslant 1,$$

after adjusting constants, including for $\|x\| \leqslant 1$.

For $t \geqslant 1$, we instead retain the bound on I_1 and sharpen the bound on I_2. Since $D_2 \subset \{y : \|y\| > R\}$,

$$\begin{aligned} I_2 &\leqslant (2\pi t)^{-d/2} \int_{\|y\|>R} p_0(y)\,\mathrm{d}y \\ &\leqslant C_{\exp}t^{-d/2}e^{-c_{\exp}R^{\min\{\gamma,2\}}}, \end{aligned}$$

where the last inequality follows by integrating the exponential envelope in Assumption 1 and adjusting constants. Combining the bounds on I_1 and I_2 yields

$$p_t(x) \leqslant C_{\exp}t^{-d/2}\exp\left[-c_{\exp}\left(\frac{\|x\|}{\sqrt{t}}\right)^{\min\{\gamma,2\}}\right], \qquad t \geqslant 1.$$

Here $C_{\exp}, c_{\exp} > 0$ depend only on d, γ, c_1, C_1 and may be adjusted between bounds. Since $\sqrt{1+t} \asymp 1$ for $t \leqslant 1$ and $\sqrt{1+t} \asymp \sqrt{t}$ for $t \geqslant 1$, choosing a common prefactor and decay constant gives the claimed bound for all $t > 0$ and $x \in \mathbb{R}^d$. □

Proof of Lemmas 21 and 22. Because $\rho_n(1+t)^{d/2} \gtrsim (\log n)/n$, the scaled envelope implies $\mathcal{G}_1 \subset \{\|x\| \leqslant C\sqrt{1+t}(\log n)^{1/\min\{\gamma,2\}}\}$ (or $\mathcal{G}_1$ is empty), so $|\mathcal{G}_1| \lesssim V_t$. Hölder's inequality and the uniform score moment bound $(\mathbb{E}_{p_t}\|s_t\|^{2m})^{1/m} \lesssim m/t$ give $\int_{\mathcal{G}_1}\|s_t\|^2 \lesssim (m/t)\rho_n^{-1/m}V_t^{1-1/m}$.

For the low-density region, choose $R = A\sqrt{1+t}(\log n)^{1/\min\{\gamma,2\}}$ with a fixed sufficiently large A. Integration of the scaled envelope gives $\mathbb{P}(\|X_t\| > R) \lesssim n^{-2}$. Hence

$$\int_{\mathcal{G}_2} p_t \leqslant c_\alpha\rho_n \operatorname{Vol}(\mathbb{B}_R) + \mathbb{P}(\|X_t\| > R) \lesssim \rho_n V_t + n^{-2} \lesssim \rho_n V_t.$$

The mass is also at most 1. A second application of Hölder gives $\int_{\mathcal{G}_2}\|s_t\|^2 p_t \lesssim (m/t)(\rho_n V_t)^{1-1/m}$. All bounds are uniform in $t > 0$ and $n \geqslant 3$, including $\rho_n \geqslant 1$. □

G Alternative Low-Noise Strategy for Polynomially Decaying Targets

As discussed in Section B.1.1, in the low-noise regime (small diffusion time t), the naive estimator based on the intrinsic diffusion scale $\sqrt{t}$ suffers from a variance blow-up in heavy-tailed regions due to sample sparsity. To control score estimation at small noise levels, we introduce a decoupled kernel score estimator whose MSE bound remains uniform as $t \downarrow 0$. Throughout this section, we assume that p_0 satisfies Assumption 3 with $\beta > 2$, and for pointwise error control, we further impose a Hölder continuity condition on p_0.

Assumption 4 (Hölder Regularity)**.** *The target density p_0 is bounded and belongs to the Hölder class $\Sigma(\beta, L)$ with $\beta > 2$. Let $\ell = \lfloor \beta \rfloor$ denote the integer part of β. Assume p_0 is ℓ-times continuously differentiable, and its ℓ-th order partial derivatives are Hölder continuous*

$$|D^\alpha p_0(x) - D^\alpha p_0(y)| \leqslant L\|x - y\|^{\beta-\ell}, \quad \forall x, y \in \mathbb{R}^d, \, |\alpha| = \ell,$$

where D^α denotes the mixed partial derivative associated with the multi-index α, and $L > 0$ is a universal constant independent of x and y.

We note that the Hölder constant L is preserved uniformly for all t, since

$$\begin{aligned}
|D^\alpha p_t(x) - D^\alpha p_t(y)| &= |D^\alpha (p_0 * \varphi_t)(x) - D^\alpha (p_0 * \varphi_t)(y)| \\
&\leqslant \int_{\mathbb{R}^d} |(D^\alpha p_0)(x - z) - (D^\alpha p_0)(y - z)| \, \varphi_t(z) \mathrm{d}z \\
&\leqslant \int_{\mathbb{R}^d} L\|x - y\|^{\beta-\ell} \varphi_t(z) \mathrm{d}z \\
&= L\|x - y\|^{\beta-\ell}.
\end{aligned}$$

Score estimator construction. Let $K \in C_c^\infty(\mathbb{R}^d)$ be a compactly supported kernel of order $\ell = \lfloor \beta \rfloor$, supported on $[-1, 1]^d$, satisfying $\int_{\mathbb{R}^d} K(u) \, \mathrm{d}u = 1$, $\int_{\mathbb{R}^d} u^\alpha K(u) \, \mathrm{d}u = 0$ for all multi-indices α with $1 \leqslant |\alpha| \leqslant \ell$, and $K, \nabla K \in L^2(\mathbb{R}^d)$.

The kernel density estimator for p_0 with bandwidth $h > 0$ is defined via

$$\hat{f}_h(y) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{y - X_i}{h}\right), \quad \text{where } X_i \overset{\text{i.i.d.}}{\sim} p_0.$$

The density estimator is defined via convolution with the Gaussian kernel φ_t

$$\hat{p}_t(x) := (\hat{f}_h * \varphi_t)(x) = \int_{\mathbb{R}^d} \hat{f}_h(y) \varphi_t(x - y) \, \mathrm{d}y.$$

We stabilize the score estimator using a unified threshold $\rho_n = C_\rho h^\beta$, where $C_\rho > 0$ is chosen sufficiently large. Specifically,

$$\hat{s}_t(x) := \frac{\nabla \hat{p}_t(x)}{\hat{p}_t(x)} \mathbb{1}_{\{\hat{p}_t(x) \geqslant \rho_n\}}.$$

We first establish global MISE bounds using Fourier analysis, then derive pointwise error bounds for the proposed score estimator.

Proposition 24 (Global MISE Bounds)**.** *Suppose Assumption 3 holds. The moment conditions and compact support of the kernel K specified above imply*

$$|\mathcal{F}[K](v) - 1| \leqslant C_K \min(\|v\|^\beta, 1), \qquad v \in \mathbb{R}^d,$$

for some C_K depending on K, d, and β. Then, for any $t > 0$, the decoupled estimator satisfies

$$\begin{aligned}
\mathbb{E}\left[\|\nabla \hat{p}_t - \nabla p_t\|_{L^2(\mathbb{R}^d)}^2\right] &\lesssim L^2 h^{2(\beta-1)} + \frac{\|\nabla K\|_{L^2}^2}{nh^{d+2}} \\
\mathbb{E}\left[\|\hat{p}_t - p_t\|_{L^2(\mathbb{R}^d)}^2\right] &\lesssim L^2 h^{2\beta} + \frac{\|K\|_{L^2}^2}{nh^d},
\end{aligned}$$

In particular, these bounds are uniform in t. The implicit constants depend only on d, β, and K.

Proof. We invoke Plancherel's theorem under the Fourier transform convention

$$\mathcal{F}[g](\omega) = \int_{\mathbb{R}^d} g(x) e^{-i\omega \cdot x} \mathrm{d}x$$

. We then decompose the MISE into the integrated squared bias and the integrated variance. We focus on $\nabla \hat{p}_t$. The density case follows by the same argument.

We begin with the integrated squared bias. Observe that

$$\mathbb{E}[\nabla \hat{p}_t] = \nabla (p_0 * K_h) * \varphi_t \,,$$

so the pointwise bias can be written as

$$\mathrm{Bias}(x) = \nabla \left((p_0 * K_h) - p_0 \right) * \varphi_t(x) \,.$$

Taking Fourier transforms gives

$$\mathcal{F}[\mathrm{Bias}](\omega) = i\omega \left(\mathcal{F}[K](h\omega) - 1 \right) \mathcal{F}[p_0](\omega) e^{-t\|\omega\|^2/2}.$$

By Plancherel's theorem, the squared L^2-norm of the bias is therefore

$$\|\mathrm{Bias}(\nabla \hat{p}_t)\|_{L^2}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \|\omega\|^2 |\mathcal{F}[K](h\omega) - 1|^2 |\mathcal{F}[p_0](\omega)|^2 e^{-t\|\omega\|^2} \, \mathrm{d}\omega.$$

For $\beta > 2$, the Fourier approximation property implies

$$\|\omega\|^2 |\mathcal{F}[K](h\omega) - 1|^2 \leqslant C_K^2 h^{2(\beta-1)} \|\omega\|^{2\beta}.$$

Indeed, this follows separately on $\|h\omega\| \leqslant 1$ and $\|h\omega\| > 1$. Thus

$$\begin{aligned} \|\,\mathrm{Bias}(\nabla \hat{p}_t)\|_{L^2}^2 &\leqslant \frac{C_K^2 h^{2(\beta-1)}}{(2\pi)^d} \int_{\mathbb{R}^d} \|\omega\|^{2\beta} |\mathcal{F}[p_0](\omega)|^2 \, \mathrm{d}\omega \\ &\lesssim L^2 h^{2(\beta-1)}. \end{aligned}$$

Here we used $e^{-t\|\omega\|^2} \leqslant 1$ and the Sobolev norm equivalence in Appendix C.1. The same argument without the gradient multiplier gives $\|\,\mathrm{Bias}(\hat{p}_t)\|_{L^2}^2 \lesssim L^2 h^{2\beta}$.

Next, we analyze the integrated variance. Define $\Psi_{h,t} = \nabla K_h * \varphi_t$, then the integrated variance of $\nabla \hat{p}_t$ is bounded by $\frac{1}{n} \|\Psi_{h,t}\|_{L^2}^2$. By Young's convolution inequality and $\|\varphi_t\|_{L^1} = 1$, it holds that

$$\|\Psi_{h,t}\|_{L^2} \leqslant \|\nabla K_h\|_{L^2} \|\varphi_t\|_{L^1} = \|\nabla K_h\|_{L^2}.$$

By the scaling of the kernel gradient, it holds that

$$\nabla K_h(x) = h^{-(d+1)} \nabla K(x/h) \,,$$

and therefore

$$\|\nabla K_h\|_{L^2}^2 = \int_{\mathbb{R}^d} h^{-2(d+1)} \|\nabla K(x/h)\|^2 \mathrm{d}x = h^{-(d+2)} \|\nabla K\|_{L^2}^2.$$

Substituting this identity yields the integrated variance term $\frac{\|\nabla K\|_{L^2}^2}{n h^{d+2}}$.

□

Proposition 25 (Uniform Pointwise Bias)**.** *Under Assumption 4, for any $t > 0$ and $x \in \mathbb{R}^d$, the pointwise bias of the decoupled estimators satisfies:*

$$\begin{aligned} |\mathbb{E}[\hat{p}_t(x)] - p_t(x)| &\leqslant C_{B,0} h^{\beta}, \\ \|\mathbb{E}[\nabla \hat{p}_t(x)] - \nabla p_t(x)\| &\leqslant C_{B,1} h^{\beta-1}, \end{aligned}$$

where $C_{B,0} = L \cdot \frac{(2d)^\ell}{\ell!} \int_{\mathbb{R}^d} \|u\|^\beta |K(u)| \mathrm{d}u$ and $C_{B,1} = L\sqrt{d} \cdot \frac{(2d)^{\ell-1}}{(\ell-1)!} \int_{\mathbb{R}^d} \|u\|^{\beta-1} |K(u)| \mathrm{d}u$, depending only on L, d, ℓ and the kernel.

Proof. We begin by studying the bias of the density estimator $\hat{p}_t$. The bias admits the representation

$$\text{Bias}(\hat{p}_t)(x) = ((p_0 * K_h) - p_0) * \varphi_t(x).$$

Since φ_t is a probability density, convolution with φ_t is an L^∞-contraction. Consequently,

$$\sup_x |\text{Bias}(\hat{p}_t)(x)| \leqslant \sup_y |(p_0 * K_h)(y) - p_0(y)|\,.$$

For any fixed y, expanding $p_0(y - hu)$ around y via the multivariate Taylor formula with an integral remainder gives

$$\begin{aligned} p_0(y - hu) = p_0(y) &+ \sum_{1\leqslant|\alpha|\leqslant\ell} \frac{(-h)^{|\alpha|}u^\alpha}{\alpha!} D^\alpha p_0(y) \\ &+ (-h)^\ell \sum_{|\alpha|=\ell} \frac{\ell}{\alpha!} \int_0^1 (1-s)^{\ell-1}[D^\alpha p_0(y - shu) - D^\alpha p_0(y)]\mathrm{d}s\, u^\alpha. \end{aligned}$$

The polynomial terms vanish after integration because the kernel has the required vanishing moments. Under Assumption 4, the remainder term satisfies

$$|R_\ell(y, -hu)| \leqslant L \cdot \frac{(2d)^\ell}{\ell!} \|hu\|^\beta.$$

Consequently,

$$|(p_0 * K_h)(y) - p_0(y)| \leqslant \int_{\mathbb{R}^d} |K(u)| \cdot L \cdot \frac{(2d)^\ell}{\ell!} h^\beta \|u\|^\beta \mathrm{d}u =: C_{B,0} h^\beta.$$

Next, we analyze the bias of the gradient estimator $\nabla\hat{p}_t$. Its bias is

$$\text{Bias}(\nabla\hat{p}_t)(x) = (\nabla(p_0 * K_h) - \nabla p_0) * \varphi_t(x).$$

Using the identity $\nabla(p_0 * K_h) = (\nabla p_0) * K_h$, we can repeat the same Taylor expansion argument with ∇p_0 in place of p_0. Apply the Taylor argument to each component of ∇p_0 with smoothness $\beta - 1$ and the same kernel K. Combining the componentwise bounds in the Euclidean norm gives the stated $C_{B,1}$.

□

Proposition 26 (Uniform Pointwise Variance)**.** *For any $t > 0$ and $x \in \mathbb{R}^d$, the pointwise variance of the decoupled estimators is bounded by the local density $p_t^*(x) = \sup_{\|y-x\|_\infty \leqslant h} p_t(y)$ via*

$$\text{Var}(\hat{p}_t(x)) \leqslant \frac{\|K\|_{L^2}^2\, p_t^*(x)}{nh^d}\,, \quad \text{Var}(\nabla\hat{p}_t(x)) \leqslant \frac{\|\nabla K\|_{L^2}^2\, p_t^*(x)}{nh^{d+2}}\,,$$

where the constants are independent of t; $p_t^*(x)$ retains its dependence on t.

Proof. We first bound the variance of the density estimator $\hat{p}_t$. Note that

$$\hat{p}_t(x) = \frac{1}{n}\sum_{i=1}^n Z_i, \qquad Z_i = (K_h * \varphi_t)(x - X_i)\,.$$

Since Z_i are i.i.d, we have

$$\text{Var}(\hat{p}_t(x)) \leqslant \frac{1}{n}\mathbb{E}[Z_1^2]\,.$$

By Jensen's inequality applied to convolution with the probability density φ_t

$$(K_h * \varphi_t)^2(z) \leqslant (K_h^2 * \varphi_t)(z).$$

which implies

$$\mathbb{E}[Z_1^2] \leqslant (K_h^2 * p_t)(x)\,.$$

If K is compactly supported, then $K_h^2(u) = h^{-2d}K^2(u/h)$ has support in $[-h,h]^d$, and a change of variables yields

$$(K_h^2 * p_t)(x) = \frac{1}{h^d}\int_{\mathbb{R}^d} K^2(u)p_t(x-hu)\mathrm{d}u \leqslant \frac{p_t^*(x)}{h^d}\int_{\mathbb{R}^d} K^2(u)\mathrm{d}u = \frac{\|K\|_{L^2}^2 p_t^*(x)}{h^d}.$$

where $p_t^*(x) := \sup_{\|u\|_\infty \leqslant 1} p_t(x-hu)$. Dividing by n gives the stated variance bound for $\hat{p}_t(x)$.

Then, we bound the variance of the gradient estimator $\nabla\hat{p}_t$. Note that

$$\nabla\hat{p}_t(x) = \frac{1}{n}\sum_{i=1}^n W_i, \qquad W_i = (\nabla K_h * \varphi_t)(x - X_i)\,.$$

Using the scaling relation

$$\nabla K_h(u) = h^{-(d+1)}\nabla K(u/h)\,,$$

we see that the gradient introduces an extra factor h^{-1} compared with K_h. Applying the same Jensen inequality and the same compact support argument as in the density case yields an analogous bound, with $\|K\|_{L^2}^2$ replaced by $\|\nabla K\|_{L^2}^2$ and h^d replaced by h^{d+2}.

□

Lemma 27 (High-Probability Concentration)**.** *Suppose Assumptions 2 and 4 hold. Let $h \asymp n^{-1/(2\beta+d)}$ and $\rho_n = C_\rho h^\beta$, where $C_\rho > 2C_{B,0}$ is fixed. For $\alpha > 0$, $n \geqslant 2$, and fixed $t > 0, x \in \mathbb{R}^d$, define*

$$A_\alpha(x) = \left\{ |\hat{p}_t(x) - p_t(x)| \leqslant C_{B,0}h^\beta + C_\alpha\left(\sqrt{\frac{p_t^*(x)\log n}{nh^d}} + \frac{\log n}{nh^d}\right)\right\},$$

where

$$C_\alpha = \max\left\{\sqrt{8(\alpha+1)\|K\|_{L^2}^2}, \frac{16(\alpha+1)\|K\|_\infty}{3}\right\}.$$

Then $\mathbb{P}(A_\alpha(x)^c) \leqslant n^{-\alpha}$. For sufficiently large n, uniformly in $t > 0$ and $x \in G_1 := \{p_t \geqslant 2\rho_n\}$, on $A_\alpha(x)$ we have $|\hat{p}_t(x) - p_t(x)| \leqslant p_t(x)/2$. The probability statement is pointwise in x, not a simultaneous event over $\mathbb{R}^d$.

Proof. Set $Z_i = (K_h * \varphi_t)(x - X_i)$. By the preceding variance bound,

$$|Z_i - \mathbb{E}Z_i| \leqslant 2h^{-d}\|K\|_\infty, \qquad \mathrm{Var}(Z_i) \leqslant h^{-d}\|K\|_{L^2}^2 p_t^*(x).$$

Bernstein's inequality gives, with probability at least $1 - n^{-\alpha}$,

$$|\hat{p}_t(x) - \mathbb{E}\hat{p}_t(x)| \leqslant \sqrt{\frac{2\|K\|_{L^2}^2 p_t^*(x)\log(2n^\alpha)}{nh^d}} + \frac{4\|K\|_\infty\log(2n^\alpha)}{3nh^d}.$$

Since $\log(2n^\alpha) \leqslant (\alpha+1)\log n$, the bias bound proves the probability assertion.

For the denominator comparison, boundedness of p_0 and its global Hölder regularity imply $\|\nabla^2 p_0\|_\infty \lesssim C_2 + L$, where the norm is the supremum of the Hessian operator norm. Indeed, on each unit ball its degree-ℓ Taylor polynomial is bounded by $\|p_0\|_\infty$ plus the uniformly bounded remainder; norm equivalence for fixed-degree polynomials controls its second-order coefficients uniformly in the center. Convolution therefore bounds $\|\nabla^2 p_t\|_\infty$ uniformly in t. Nonnegativity and Taylor's inequality imply

$$\|\nabla p_t(x)\|^2 \leqslant 2\|\nabla^2 p_t\|_\infty p_t(x), \qquad p_t(x+v) \leqslant 2p_t(x) + \|\nabla^2 p_t\|_\infty\|v\|^2.$$

Consequently,

$$p_t^*(x) \leqslant 2p_t(x) + d\|\nabla^2 p_0\|_\infty h^2.$$

On G_1, using $(nh^d)^{-1} \asymp h^{2\beta}$ yields

$$\frac{\sqrt{p_t^*(x)\log n/(nh^d)} + \log n/(nh^d)}{p_t(x)} \lesssim h^{\beta/2}\sqrt{\log n} + h\sqrt{\log n} + h^\beta\log n = o(1).$$

The bias divided by $p_t(x)$ is at most $C_{B,0}/(2C_\rho) < 1/4$. Thus the total relative error on $A_\alpha(x)$ is at most $1/2$ for sufficiently large n, uniformly in $x \in G_1$ and $t > 0$. □

Assumption 5 (Score Regularity)**.** *The initial distribution p_0 satisfies*

$$\mathbb{E}_{p_0}\|s_0\|^{2k} \leqslant M_{2k}$$

for some fixed $k > 1$ and $M_{2k} < \infty$.

Remark. This assumption is used only for the alternative low-noise analysis, not for the main Gaussian-KDE bounds. The constants in the auxiliary bounds may depend on k and M_{2k}.

Lemma 28 (Monotonicity of Score Moments)**.** *Suppose the initial score function $s_0 = \nabla \log p_0$ satisfies $\mathbb{E}_{p_0}[\|s_0(x)\|^{2k}] < \infty$ for some $k \geqslant 1$. Then, for any $t > 0$, the moments of the diffused score s_t are non-increasing with respect to t*

$$\mathbb{E}_{p_t}[\|s_t(x)\|^{2k}] \leqslant \mathbb{E}_{p_0}[\|s_0(x)\|^{2k}].$$

Proof. We utilize the interpretation of the diffused score as a conditional expectation (often referred to as Tweedie's formula). The score function can be expressed as

$$s_t(x) = \nabla \log p_t(x) = \frac{\nabla p_t(x)}{p_t(x)} = \frac{\int_{\mathbb{R}^d} \nabla p_0(y)\varphi_t(x-y)\,\mathrm{d}y}{p_t(x)}.$$

Using the identity $\nabla p_0(y) = s_0(y)p_0(y)$, we have

$$s_t(x) = \int_{\mathbb{R}^d} s_0(y)\frac{p_0(y)\varphi_t(x-y)}{p_t(x)}\,\mathrm{d}y.$$

Notice that $w(y|x) := \frac{p_0(y)\varphi_t(x-y)}{p_t(x)}$ corresponds exactly to the posterior density $p(y|x)$ (the conditional distribution of the clean data y given the noisy observation x). Thus, $s_t(x)$ is the posterior expectation of the initial score

$$s_t(x) = \mathbb{E}_{Y|X=x}[s_0(Y)].$$

Consider the convex function $\Phi(u) = \|u\|^{2k}$ (for $k \geqslant 1$). By Jensen's inequality for conditional expectations, it holds that

$$\|s_t(x)\|^{2k} = \left\|\mathbb{E}_{Y|X=x}[s_0(Y)]\right\|^{2k} \leqslant \mathbb{E}_{Y|X=x}\left[\|s_0(Y)\|^{2k}\right].$$

Now, we compute the global expectation with respect to the marginal density $p_t(x)$

$$\begin{aligned}
\mathbb{E}_{p_t}[\|s_t(x)\|^{2k}] &= \int_{\mathbb{R}^d} \|s_t(x)\|^{2k} p_t(x)\,\mathrm{d}x \\
&\leqslant \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \|s_0(y)\|^{2k} \frac{p_0(y)\varphi_t(x-y)}{p_t(x)}\,\mathrm{d}y\right) p_t(x)\,\mathrm{d}x \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|s_0(y)\|^{2k} p_0(y)\varphi_t(x-y)\,\mathrm{d}y\,\mathrm{d}x.
\end{aligned}$$

By Fubini's theorem, we have

$$\mathbb{E}_{p_t}[\|s_t(x)\|^{2k}] \leqslant \int_{\mathbb{R}^d} \|s_0(y)\|^{2k} p_0(y) \left(\int_{\mathbb{R}^d} \varphi_t(x-y)\,\mathrm{d}x\right)\mathrm{d}y.$$

Since φ_t is a probability density, and $\int \varphi_t(x-y)\,\mathrm{d}x = 1$, it follows that

$$\mathbb{E}_{p_t}[\|s_t(x)\|^{2k}] \leqslant \int_{\mathbb{R}^d} \|s_0(y)\|^{2k} p_0(y)\,\mathrm{d}y = \mathbb{E}_{p_0}[\|s_0(x)\|^{2k}]$$

as desired. Applying the same argument to p_u and an additional Gaussian convolution of variance $t-u$ proves monotonicity for every $0 \leqslant u < t$.

□

We are now ready to establish the MSE for the proposed score estimator.

Theorem 29 (MSE for $\hat{s}_t$)**.** *Suppose Assumptions 2, 3 with $\beta > 2$, 4 and 5 with $k > 1$ hold. Set $h \asymp n^{-\frac{1}{2\beta+d}}$ and $\rho_n = C_\rho h^\beta \asymp n^{-\frac{\beta}{2\beta+d}}$ with $C_\rho > 2C_{B,0}$. For all sufficiently large n and any $0 < t \leqslant 1$, the score estimation error satisfies*

$$\mathbb{E}\left[\int_{\mathbb{R}^d} \|\hat{s}_t(x) - s_t(x)\|^2 p_t(x)\,\mathrm{d}x\right] \lesssim n^{-\frac{\beta-2}{2\beta+d}} + n^{-\frac{\beta(\gamma+1)}{k'(2\beta+d)(d+\gamma+1)}},$$

where $\frac{1}{k} + \frac{1}{k'} = 1$. The implicit constant is independent of n, t.

Proof. Fix $\alpha = 4$ and write $A(x) = A_4(x)$. We decompose the score risk into its truncated and active parts, separating the contribution on $A(x)^c$:

$$\mathcal{R} = \underbrace{\mathbb{E}\int \|s_t(x)\|^2 p_t(x) \mathbb{1}_{\{\hat{p}_t(x) < \rho_n\}} \mathbb{1}_{A(x)}\,\mathrm{d}x}_{E^A_{\mathrm{trunc}}}$$
$$+ \underbrace{\mathbb{E}\int \left\| \frac{\nabla \hat{p}_t(x)}{\hat{p}_t(x)} - s_t(x) \right\|^2 p_t(x) \mathbb{1}_{\{\hat{p}_t(x) \geqslant \rho_n\}} \mathbb{1}_{A(x)}\,\mathrm{d}x}_{E^A_{\mathrm{est}}} + \mathcal{R}_{\mathrm{bad}}.$$

All integrals without a specified domain are over $\mathbb{R}^d$.

Tail truncation. Let $G_1 = \{p_t \geqslant 2\rho_n\}$ and $G_2 = \{p_t < 2\rho_n\}$. Lemma 27 shows that truncation cannot occur on $G_1 \cap A(x)$. For $0 < t \leqslant 1$, the tail envelope constant in Lemma 23 is uniformly bounded. The argument of (12), with threshold $2\rho_n$, therefore gives

$$\int_{G_2} p_t(x)\,\mathrm{d}x \lesssim \rho_n^{(\gamma+1)/(d+\gamma+1)}.$$

By Hölder's inequality and Lemma 28,

$$E^A_{\mathrm{trunc}} \leqslant \int_{G_2} \|s_t(x)\|^2 p_t(x)\,\mathrm{d}x$$
$$\leqslant M_{2k}^{1/k} \left(\int_{G_2} p_t(x)\,\mathrm{d}x\right)^{1/k'} \lesssim h^{\beta(\gamma+1)/[k'(d+\gamma+1)]}.$$

Estimation on the active region. Set

$$T_1 = \frac{\|\nabla \hat{p}_t - \nabla p_t\|^2}{\hat{p}_t^2} p_t, \qquad T_2 = \frac{|\hat{p}_t - p_t|^2}{\hat{p}_t^2} \|s_t\|^2 p_t.$$

Then E^A_{est} is at most twice the integral of $T_1 + T_2$ on the active region and $A(x)$. On this region, $p_t/\hat{p}_t \leqslant 2$: on G_1 this follows from concentration, and on G_2 it follows from $p_t < 2\rho_n \leqslant 2\hat{p}_t$. In particular,

$$\frac{p_t}{\hat{p}_t^2} \leqslant \frac{2}{\rho_n}, \qquad 0 \leqslant \frac{|\hat{p}_t - p_t|^2}{\hat{p}_t^2} \leqslant 1.$$

The gradient MISE bound thus yields

$$\mathbb{E}\int_{\{\hat{p}_t \geqslant \rho_n\}} T_1 \mathbb{1}_{A(x)}\,\mathrm{d}x \lesssim \rho_n^{-1}\left(h^{2\beta-2} + \frac{1}{nh^{d+2}}\right)$$
$$\lesssim h^{\beta-2}.$$

For T_2, apply Hölder's inequality on the product of the sample probability space and $p_t(x)\,\mathrm{d}x$:

$$\mathbb{E}\int_{\{\hat{p}_t \geqslant \rho_n\}} T_2 \mathbb{1}_{A(x)}\,\mathrm{d}x \leqslant M_{2k}^{1/k}\left[\mathbb{E}\int_{\{\hat{p}_t \geqslant \rho_n\}} \left(\frac{|\hat{p}_t - p_t|^2}{\hat{p}_t^2}\right)^{k'} p_t\, \mathbb{1}_{A(x)}\,\mathrm{d}x\right]^{1/k'}$$
$$\lesssim \left[\rho_n^{-1}\mathbb{E}\|\hat{p}_t - p_t\|_{L^2}^2\right]^{1/k'}$$

$$\lesssim h^{\beta/k'}.$$

The second step uses the relative-error bound by 1 on this region and $p_t/\hat{p}_t^2 \leqslant 2/\rho_n$.

Bad event. Smoothness of the fixed compactly supported kernel gives the deterministic bound, uniformly in $t > 0$,

$$\|\hat{s}_t\|_\infty \leqslant \rho_n^{-1}\|\nabla K_h * \varphi_t\|_\infty \leqslant \rho_n^{-1}h^{-d-1}\|\nabla K\|_\infty.$$

Together with $\mathbb{P}(A(x)^c) \leqslant n^{-4}$ and $\int \|s_t\|^2 p_t \leqslant M_{2k}^{1/k}$, this gives

$$\begin{aligned}\mathcal{R}_{\mathrm{bad}} &\lesssim n^{-4}\left(\rho_n^{-2}h^{-2d-2} + M_{2k}^{1/k}\right)\\ &\lesssim h^{6\beta+2d-2} + h^{8\beta+4d} \lesssim h^{\beta-2}.\end{aligned}$$

Here $n^{-1} \asymp h^{2\beta+d}$ and $\rho_n = C_\rho h^\beta$.

Combining the three contributions, and using $h \leqslant 1$ for large n and $(\gamma+1)/(d+\gamma+1) < 1$, we obtain

$$\begin{aligned}\mathcal{R} &\lesssim h^{\beta-2} + h^{\beta/k'} + h^{\beta(\gamma+1)/[k'(d+\gamma+1)]}\\ &\lesssim n^{-(\beta-2)/(2\beta+d)} + n^{-\beta(\gamma+1)/[k'(2\beta+d)(d+\gamma+1)]}.\end{aligned}$$

All bounds are uniform over $0 < t \leqslant 1$. □

Remark. Theorem 29 gives a uniform low-noise upper bound under additional regularity and score-moment assumptions. Define

$$\kappa_{\mathrm{grad}} = \frac{\beta-2}{2\beta+d}, \qquad \kappa_{\mathrm{tail}} = \frac{\beta(\gamma+1)}{k'(2\beta+d)(d+\gamma+1)}, \qquad \kappa_{\mathrm{dec}} = \min\{\kappa_{\mathrm{grad}}, \kappa_{\mathrm{tail}}\}.$$

Thus the integrated-in-space score MSE is $O(n^{-\kappa_{\mathrm{dec}}})$ uniformly on $0 < t \leqslant 1$; this is an upper bound, not an optimality claim. For a hybrid rule, let $\kappa_1 = (\gamma+1)/(d+\gamma+1)$. Balancing the two available MSE bounds at the level of powers of n gives

$$t_{\mathrm{switch}} \asymp n^{-2(\kappa_1-\kappa_{\mathrm{dec}})/(2+d\kappa_1)}.$$

Using the decoupled estimator below this time and the Gaussian KDE estimator above it yields

$$\begin{aligned}\int_0^1 \mathbb{E}\|\hat{s}_t^{\mathrm{hybrid}} - s_t\|_{L^2(p_t)}^2 \,\mathrm{d}t &\lesssim n^{-\kappa_{\mathrm{dec}}}t_{\mathrm{switch}} + (\log n)^{1+\kappa_1}n^{-\kappa_1}t_{\mathrm{switch}}^{-d\kappa_1/2}\\ &\lesssim \mathrm{polylog}(n)n^{-\kappa_{\mathrm{dec}}-2(\kappa_1-\kappa_{\mathrm{dec}})/(2+d\kappa_1)}.\end{aligned}$$

The time integral is an improper integral at zero. There is no general ordering between this switching time and the early-stopping time of Theorem 4. Because this section assumes $\beta > 2$, that theorem's Gaussian-smoothing benchmark uses smoothness 2. These upper bounds do not establish suboptimality of the alternative estimator or minimax optimality of a sampling strategy.

H Proof of Section 4

In this part, we present the proof of the minimax lower bounds for score estimation.

Baseline densities and compatibility for the lower bounds. Set $D = d + \gamma + 1$ and use the normalized baselines

$$q_{\mathrm{P}}(x) = c_{\mathrm{P}}(1+\|x\|^2)^{-D/2}, \qquad q_{\mathrm{E}}(x) = Z^{-1}e^{-2c_1(1+\|x\|^2)^{\gamma/2}}.$$

Here c_{P} and Z are normalizing constants. Both baselines are smooth, strictly positive, and have finite radial Sobolev seminorm $S_\beta(q) := [(2\pi)^{-d}\int \|\xi\|^{2\beta}|\widehat{q}(\xi)|^2\mathrm{d}\xi]^{1/2}$ for every fixed $\beta > 0$. For the respective lower bound, assume that its baseline q_* satisfies $q_*(x) \leqslant (1-\tau)E(x)$ and $S_\beta(q_*) \leqslant L-\ell$ for fixed $\tau \in (0,1)$ and $\ell > 0$, where E is the corresponding class envelope. Sufficient conditions are $C_2 > c_{\mathrm{P}}$ (polynomial) or $C_1 > Z^{-1}$ (exponential), together with $L > S_\beta(q_*)$. Choosing the perturbations to satisfy $|q_b - q_*| \leqslant \eta q_*$ with $\eta < \min\{1/2, \tau/(1-\tau)\}$ and $S_\beta(q_b - q_*) \leqslant \ell$ then keeps every hypothesis in the same fixed class and Sobolev ball. All lower-bound constants and time-window endpoints may depend on these fixed parameters and margins, but not on n, t or the selected density.

Fractional-Sobolev almost orthogonality. We first record the calculation used in both lower bounds. Let $\psi \in C_c^\infty(\mathbb{R}^d)$, let $h \in (0,1]$, and let $x_i = x_0 + hk_i$ for distinct $k_i \in \mathbb{Z}^d$. For

$$\Delta_b(x) = \epsilon \sum_{i\in\mathcal{I}} b_i \psi\left(\frac{x - x_i}{h}\right), \qquad |b_i| \leqslant 1,$$

we have, for every real $\beta > 0$,

$$\|\Delta_b\|_{\dot{H}^\beta}^2 \leqslant C_{\psi,\beta,d}\epsilon^2 |\mathcal{I}| h^{d-2\beta}.$$

Indeed,

$$\widehat{\Delta_b}(\xi) = \epsilon h^d \widehat{\psi}(h\xi) \sum_{i\in\mathcal{I}} b_i e^{-ix_i\cdot\xi}.$$

After the change of variables $\eta = h\xi$ and removal of the common phase $e^{-ix_0\cdot\xi}$,

$$\|\Delta_b\|_{\dot{H}^\beta}^2 = \epsilon^2 h^{d-2\beta} \int_{\mathbb{R}^d} W(\eta) \left|\sum_{i\in\mathcal{I}} b_i e^{-ik_i\cdot\eta}\right|^2 \mathrm{d}\eta, \qquad W(\eta) := \|\eta\|^{2\beta} |\widehat{\psi}(\eta)|^2.$$

Since $\widehat{\psi}$ is rapidly decaying, the periodization $W_{\mathrm{per}}(\eta) := \sum_{\ell\in\mathbb{Z}^d} W(\eta + 2\pi\ell)$ is bounded. Orthogonality of the lattice exponentials on $[0, 2\pi]^d$ therefore gives

$$\begin{aligned}\int_{\mathbb{R}^d} W(\eta) \left|\sum_i b_i e^{-ik_i\cdot\eta}\right|^2 \mathrm{d}\eta &\leqslant \|W_{\mathrm{per}}\|_\infty \int_{[0,2\pi]^d} \left|\sum_i b_i e^{-ik_i\cdot\eta}\right|^2 \mathrm{d}\eta \\ &\leqslant C_{\psi,\beta,d}|\mathcal{I}|.\end{aligned}$$

Moreover, because $h \leqslant 1$, the same upper bound controls the L^2 term and hence the inhomogeneous H^β norm. In the polynomial construction $|\mathcal{I}| \asymp R^d h^{-d}$, so the bound becomes $\|\Delta_b\|_{\dot{H}^\beta}^2 \lesssim \epsilon^2 R^d h^{-2\beta}$; in the exponential construction $|\mathcal{I}| \asymp h^{-d}$, it becomes $\|\Delta_b\|_{\dot{H}^\beta}^2 \lesssim \epsilon^2 h^{-2\beta}$.

H.1 Proof of Theorem 9

Choose a fixed $\psi \in C_c^\infty([-1/2, 1/2]^d)$ satisfying $\int_{[-1/2,1/2]^d} \psi = 0$ and $\nabla\psi(0) = v_0 \neq 0, L_\psi = \sup_{x\in[-1/2,1/2]^d} \|\nabla^2\psi\|$. Set

$$h := \Lambda\sqrt{t}, \quad \Lambda \geqslant \frac{4L_\psi\sqrt{d}}{\|\nabla\psi(0)\|},$$

and construct perturbations based on ψ and h via

$$\omega(x) := \psi(x/h).$$

Then, ω is supported on $[-h/2, h/2]^d$ and $\int_{[-h/2,h/2]^d} \omega = 0$. Choose

$$q_0 = C_\gamma(1 + \|x\|^2)^{-(d+\gamma+1)/2}, \quad q_b = q_0 + \epsilon \sum_{\mathbf{i}\in\mathcal{I}} b_{\mathbf{i}} \omega(x - x_{\mathbf{i}}).$$

Then, with $q_0 = q_{\mathrm{P}}$ and the fixed-class constraints verified below, $q_0, q_b \in \mathcal{H}_\gamma \cap \mathcal{P}_{\mathcal{S}}(\beta, L)$. Choose the grid centers so that $x_{\mathbf{i}} + [-h/2, h/2]^d \subset [R, 2R]^d$; this still gives $|\mathcal{I}| \asymp (R/h)^d$ when R/h is large. We work with $R \geqslant 2$ and $0 < t \leqslant \Lambda^{-2}$, so $h \leqslant 1$, and enlarge the fixed Λ below if needed. Let

$$f_0 = \varphi_t * q_0, \quad f_b = \varphi_t * q_b.$$

Writing $D = d + \gamma + 1$ and $s_0 = \nabla \log f_0$, the Student-type baseline satisfies $f_0(x) \asymp R^{-D}$ and $\|s_0(x)\| \leqslant C_s/R$ uniformly on $[R, 2R]^d$ for $R \geqslant 2$ and $t \leqslant 1$. These follow by convolving the polynomial bounds on q_0 and ∇q_0 with the Gaussian; all constants below are independent of n, t, R.

Separation of scores. Note that

$$\int_{\mathbb{R}^d} \|s_b(x) - s_{b'}(x)\|^2 f_0(x)\mathrm{d}x$$

$$= \int_{\mathbb{R}^d} \left\| \frac{\nabla f_b}{f_b} - \frac{\nabla f_{b'}}{f_{b'}} \right\|^2 f_0(x)\mathrm{d}x$$
$$= \int_{\mathbb{R}^d} \left\| \frac{\nabla f_b \cdot (f_{b'} - f_b) + f_b \cdot (\nabla f_b - \nabla f_{b'})}{f_b(x) f_{b'}(x)} \right\|^2 f_0(x)\mathrm{d}x\,.$$

Denote $\Delta f_{b,b'} = f_b - f_{b'}, \Delta\nabla f_{b,b'} = \nabla f_b - \nabla f_{b'}$, then we have

$$\begin{aligned}
\nabla f_b \cdot \Delta f_{b',b} &= \nabla f_b \cdot \epsilon \sum_{\mathbf{i}\in\mathcal{I}} (b'_{\mathbf{i}} - b_{\mathbf{i}})(\varphi_t * \omega)(x - x_{\mathbf{i}}) \\
&= \epsilon \nabla f_0 \cdot \sum_{\mathbf{i}\in\mathcal{I}} (b'_{\mathbf{i}} - b_{\mathbf{i}})(\varphi_t * \omega)(x - x_{\mathbf{i}}) \\
&\quad + \epsilon^2 \left[\sum_{\mathbf{i}\in\mathcal{I}} b_{\mathbf{i}}(\varphi_t * \nabla\omega)(x - x_{\mathbf{i}})\right] \left[\sum_{\mathbf{i}\in\mathcal{I}} (b'_{\mathbf{i}} - b_{\mathbf{i}})(\varphi_t * \omega)(x - x_{\mathbf{i}})\right] \\
f_b \cdot \Delta\nabla f_{b,b'} &= \epsilon f_b \cdot \sum_{\mathbf{i}\in\mathcal{I}} (b_{\mathbf{i}} - b'_{\mathbf{i}})(\varphi_t * \nabla\omega)(x - x_{\mathbf{i}}) \\
&= \epsilon f_0 \cdot \sum_{\mathbf{i}\in\mathcal{I}} (b_{\mathbf{i}} - b'_{\mathbf{i}})(\varphi_t * \nabla\omega)(x - x_{\mathbf{i}}) \\
&\quad + \epsilon^2 \left[\sum_{\mathbf{i}\in\mathcal{I}} b_{\mathbf{i}}(\varphi_t * \omega)(x - x_{\mathbf{i}})\right] \left[\sum_{\mathbf{i}\in\mathcal{I}} (b_{\mathbf{i}} - b'_{\mathbf{i}})(\varphi_t * \nabla\omega)(x - x_{\mathbf{i}})\right]
\end{aligned}$$

Let $U_{\mathbf{i}}(x) = (\varphi_t * \omega)(x - x_{\mathbf{i}})$, $r = \delta h$, and $Q_{\mathbf{i}} = x_{\mathbf{i}} + [-r, r]^d$. Restricting to cores with $b_{\mathbf{i}} \neq b'_{\mathbf{i}}$, the uniform second-order control given below justifies

$$\|s_b - s_{b'}\|^2_{L^2(f_0)} \gtrsim \epsilon^2 \sum_{\mathbf{i}: b_{\mathbf{i}} \neq b'_{\mathbf{i}}} \int_{Q_{\mathbf{i}}} \frac{1}{f_0(x)} \left\| (b_{\mathbf{i}} - b'_{\mathbf{i}})[s_0(x)U_{\mathbf{i}}(x) - \nabla U_{\mathbf{i}}(x)] + \xi_{\mathbf{i}}(x) \right\|^2 \mathrm{d}x.$$

The implicit constant accounts for the quadratic terms and the comparable denominators, where

$$\xi_{\mathbf{i}}(x) = \sum_{\mathbf{j}\neq\mathbf{i}\in\mathcal{I}} (b_{\mathbf{j}} - b'_{\mathbf{j}}) \left[\frac{\nabla f_0}{f_0} \cdot (\varphi_t * \omega)(x - x_{\mathbf{j}}) - (\varphi_t * \nabla\omega)(x - x_{\mathbf{j}})\right].$$

Note that when x is away from $[-h/2, h/2]^d$, we can derive an upper bound for $\varphi_t * \omega$ using the d-dimensional kernel $\varphi_t(z) = (2\pi t)^{-d/2} \exp(-\|z\|^2/(2t))$:

$$\begin{aligned}
|(\varphi_t * \omega)(x)| &= \left| \int_{\mathbb{R}^d} \varphi_t(x - y)\omega(y)\mathrm{d}y \right| \\
&\leqslant \int_{[-h/2,h/2]^d} \varphi_t(x - y)|\omega(y)|\mathrm{d}y \\
&\leqslant h^d \sup_{y\in[-h/2,h/2]^d} \varphi_t(x - y) \cdot \|\omega\|_\infty \\
&\leqslant \sup_{y\in[-h/2,h/2]^d} \frac{h^d}{(2\pi t)^{d/2}} \exp\left(-\frac{\|x - y\|^2_\infty}{2t}\right) \cdot \|\omega\|_\infty\,.
\end{aligned}$$

Likewise, since $\|\nabla\omega\|_\infty = h^{-1}\|\nabla\psi\|_\infty$, with $a(x) = \mathrm{dist}_\infty(x, [-h/2, h/2]^d)$,

$$\|(\varphi_t * \nabla\omega)(x)\| \leqslant \frac{h^{d-1}}{(2\pi t)^{d/2}} \|\nabla\psi\|_\infty e^{-a(x)^2/(2t)}.$$

Fix $0 < \delta \leqslant 1/4$. For $x \in Q_{\mathbf{i}}$ and $\|x_{\mathbf{j}} - x_{\mathbf{i}}\|_\infty = kh$ with $k \geqslant 1$, the distance to $x_{\mathbf{j}} + [-h/2, h/2]^d$ is at least $(k - 1/2 - \delta)h$. Thus

$$\|\xi_{\mathbf{i}}(x)\| \leqslant \left(\frac{C_s}{R}\|\psi\|_\infty + \frac{1}{h}\|\nabla\psi\|_\infty\right) E_{d,\delta}(\Lambda),$$
$$E_{d,\delta}(\Lambda) := \frac{\Lambda^d}{(2\pi)^{d/2}} \sum_{k=1}^{\infty} [(2k + 1)^d - (2k - 1)^d] e^{-\Lambda^2(k-1/2-\delta)^2/2}$$

$$\leqslant C_d \Lambda^d e^{-\Lambda^2/32} \longrightarrow 0 \qquad (\Lambda \to \infty).$$

Here the finite grid is bounded by the full lattice shells, $k - 1/2 - \delta \geqslant k/4$, and the shell counts are $O_d(k^{d-1})$. The Gaussian exponential dominates the kernel and support-volume prefactor $h^d t^{-d/2} = \Lambda^d$ as Λ increases. The same shell estimate controls $\sum_{\mathbf{j}\neq\mathbf{i}} |U_{\mathbf{j}}|$ and $\sum_{\mathbf{j}\neq\mathbf{i}} \|\nabla U_{\mathbf{j}}\|$ with factors $\|\psi\|_\infty$ and $h^{-1}\|\nabla\psi\|_\infty$, respectively. We also note that

$$\begin{aligned}&\left\| \frac{\nabla f_0}{f_0} \cdot (\varphi_t * \omega)(x - x_{\mathbf{i}}) - (\varphi_t * \nabla\omega)(x - x_{\mathbf{i}}) \right\| \\ &\geqslant \|(\varphi_t * \nabla\omega)(x - x_{\mathbf{i}})\| - \left\| \frac{\nabla f_0}{f_0} (\varphi_t * \omega)(x - x_{\mathbf{i}}) \right\|\end{aligned}$$

We handle the two terms on the right-hand side of this display separately. Let

$$\delta < \min\{1/4, \|v_0\|/(4\sqrt{d}L_\psi)\},$$

it follows that

$$\|\nabla\omega(u) - \nabla\omega(0)\| \leqslant \sup_{x\in[-h/2,h/2]^d} \|\nabla^2\omega(x)\| \cdot \|u\| \leqslant L_\psi \sqrt{d}\, r/h^2 < \frac{\|v_0\|}{4h},$$

then $\|\nabla\omega(u)\| > \frac{3}{4}\|\nabla\omega(0)\|$ for $u = x - x_{\mathbf{i}} \in [-r, r]^d$. Denote

$$g_t(u) := (\varphi_t * \nabla\omega)(u) = \int_{\mathbb{R}^d} \varphi_t(u - y)\nabla\omega(y)\mathrm{d}y\,.$$

It holds that

$$\begin{aligned}\|g_t(u) - \nabla\omega(u)\| &= \left\| \int_{\mathbb{R}^d} \varphi_t(u - y)(\nabla\omega(y) - \nabla\omega(u))\mathrm{d}y \right\| \\ &\leqslant \int_{\mathbb{R}^d} \varphi_t(u - y) \frac{L_\psi}{h^2} \|u - y\| \mathrm{d}y \\ &= \frac{L_\psi\sqrt{t}}{h^2} \mathbb{E}[\|Z\|] \\ &\leqslant \frac{L_\psi\sqrt{dt}}{h^2},\end{aligned}$$

where $Z \sim \mathcal{N}(0, I_d)$, and the last inequality follows from Jensen inequality. Therefore, we have

$$\|g_t(u)\| \geqslant \|\nabla\omega(u)\| - \|g_t(u) - \nabla\omega(u)\| \geqslant \frac{3\|v_0\|}{4h} - \frac{L_\psi\sqrt{d}}{\Lambda h} \geqslant \frac{\|v_0\|}{2h}$$

and

$$\left\| \frac{\nabla f_0}{f_0} \cdot (\varphi_t * \omega)(x - x_{\mathbf{i}}) \right\| \leqslant \frac{C_s}{R}\|\omega\|_\infty = \frac{C_s}{R}\|\psi\|_\infty\,.$$

Let $R > 4C_s h\|\psi\|_\infty/\|v_0\|$, it then follows that

$$\left\| \frac{\nabla f_0}{f_0} \cdot (\varphi_t * \omega)(x - x_{\mathbf{i}}) - (\varphi_t * \nabla\omega)(x - x_{\mathbf{i}}) \right\| \geqslant \frac{\|v_0\|}{4h}\,.$$

Choose the fixed Λ large enough that $E_{d,\delta}(\Lambda)(\|v_0\|/4 + \|\nabla\psi\|_\infty) \leqslant \|v_0\|/8$. The preceding estimates then give

$$\|(b_{\mathbf{i}} - b'_{\mathbf{i}})(s_0 U_{\mathbf{i}} - \nabla U_{\mathbf{i}}) + \xi_{\mathbf{i}}\| \geqslant \frac{\|v_0\|}{8h}, \qquad x \in Q_{\mathbf{i}}, \quad b_{\mathbf{i}} \neq b'_{\mathbf{i}}.$$

To control the quadratic remainder, set $S_b = \sum_{\mathbf{j}} b_{\mathbf{j}} U_{\mathbf{j}}$, $\Delta S = S_b - S_{b'}$, $L = \nabla\Delta S - s_0\Delta S$, and $Q = S_{b'}\nabla\Delta S - \Delta S\nabla S_{b'}$. The exact score numerator is $\epsilon f_0 L + \epsilon^2 Q$. Disjoint bump supports and Gaussian convolution imply $\|S_b\|_\infty \lesssim 1$ and $\|\nabla S_b\|_\infty \lesssim h^{-1}$, so on each differing core

$$\|\epsilon^2 Q\| \leqslant C(\epsilon R^D)\|\epsilon f_0 L\| \leqslant \tfrac{1}{2}\|\epsilon f_0 L\|,$$

provided $\epsilon R^D = c_R$ is a sufficiently small fixed constant. The same choice ensures $f_b \asymp f_{b'} \asymp f_0$, which justifies the lower bound in terms of the linear contribution above. Integrating over the disjoint differing cores, each of volume $(2r)^d$, now yields

$$\begin{aligned}\|s_b - s_{b'}\|^2_{L^2(f_0)} &\gtrsim \epsilon^2 R^D (2\delta h)^d h^{-2}\, d_H(b, b') \\ &\asymp \epsilon^2 R^D h^{d-2}\, d_H(b, b') \\ &\asymp \epsilon^2 R^D t^{d/2-1}\, d_H(b, b').\end{aligned}$$

Here δ and Λ are fixed before n, t are chosen; their powers affect constants only. Therefore, we have

$$\int_{\mathbb{R}^d} \|s_b(x) - s_{b'}(x)\|^2 f_0(x)\mathrm{d}x \gtrsim \epsilon^2 R^{d+\gamma+1} t^{d/2-1} d_H(b, b')\,.$$

where $d_H(b, b') = \sum_{\mathbf{i}\in\mathcal{I}} \mathbf{1}\{b_{\mathbf{i}} \neq b'_{\mathbf{i}}\}$ is the Hamming distance. By the Gilbert–Varshanov bound, there exists an exponentially large packing $\mathcal{B}' \subset \mathcal{B}$, whose minimum Hamming distance is linear in the dimension, namely, $|\mathcal{B}'| \geqslant \exp(c_0(R/h)^d)$ and $\min_{b\neq b'\in\mathcal{B}'} d_H(b, b') \geqslant c_0(R/h)^d$ for some universal constant c_0. Thus, we obtain

$$\|s_b - s_{b'}\|^2_{L^2(f_0)} \gtrsim \epsilon^2 R^{2d+\gamma+1} t^{-1}\,.$$

Bounding the KL radius. The observations have joint law $q_b^{\otimes n}$. Since the perturbations have disjoint supports, $\int \omega = 0$, and $q_0(x) \asymp R^{-(d+\gamma+1)}$ on their supports, we have

$$\begin{aligned}\mathrm{KL}(q_b\|q_0) \leqslant \chi^2(q_b\|q_0) &= \int_{\mathbb{R}^d} \frac{(q_b(x) - q_0(x))^2}{q_0(x)}\mathrm{d}x \\ &\lesssim \epsilon^2 R^{d+\gamma+1}|\mathcal{I}|h^d.\end{aligned}$$

Using $|\mathcal{I}| \asymp R^d h^{-d}$, $h = \Lambda\sqrt{t}$, and $\epsilon = \frac{c_\epsilon}{nh^d}$, choose

$$R = \Big(\frac{c_R}{\epsilon}\Big)^{\frac{1}{d+\gamma+1}}$$

with $c_\epsilon, c_R > 0$ sufficiently small. Then

$$\frac{n\,\mathrm{KL}(q_b\|q_0)}{|\mathcal{I}|} \lesssim n\epsilon^2 R^{d+\gamma+1} h^d = c_\epsilon c_R.$$

Since the Gilbert–Varshamov packing satisfies $\log|\mathcal{B}'| \gtrsim |\mathcal{I}|$, this gives

$$\max_{b\in\mathcal{B}'} \mathrm{KL}(q_b^{\otimes n}\|q_0^{\otimes n}) \leqslant c\log|\mathcal{B}'|$$

after shrinking c_ϵ, c_R if needed.

Fano method. Applying Fano's inequality to the observations with joint law $q_b^{\otimes n}$ therefore yields a lower bound controlled by the score separation above. Moreover, the smallness of c_R ensures $(1-\eta)q_0 \leqslant q_b \leqslant (1+\eta)q_0$ on the perturbation region, and convolution with φ_t preserves the same comparison for f_b and f_0. Hence $L^2(f_b)$ and $L^2(f_0)$ are equivalent up to constants, so

$$\begin{aligned}\inf_{\hat{s}} \sup_{b\in\mathcal{B}'} \mathbb{E}_{q_b^{\otimes n}} \|\hat{s} - s_b\|^2_{L^2(f_b)} &\gtrsim \min_{b\neq b'\in\mathcal{B}'} \|s_b - s_{b'}\|^2_{L^2(f_0)} \\ &\gtrsim \epsilon^2 R^{2d+\gamma+1} h^{-2} \\ &\gtrsim n^{-\frac{\gamma+1}{d+\gamma+1}} t^{-1-\frac{d(\gamma+1)}{2(d+\gamma+1)}}.\end{aligned}$$

Since $h = \Lambda\sqrt{t}$ with fixed Λ, the kernel normalization and core-volume factors contribute only fixed constants to this rate. Together with the construction constraints verified below, this proves the lower bound in Theorem 9.

To ensure the perturbed densities q_b define a valid worst-case construction, we must verify three constraints: the heavy-tailed decay, the Sobolev regularity, and the geometric consistency of the local perturbations. Recall our parameter choices: $h \asymp t^{1/2}$, $\epsilon \asymp n^{-1}t^{-d/2}$, and $R \asymp \epsilon^{-\frac{1}{d+\gamma+1}}$.

Heavy-tailedness ($q_b \in \mathcal{H}_\gamma$): The perturbed density must remain non-negative and satisfy the polynomial tail decay. On the perturbation supports, $q_0(x) \asymp R^{-(d+\gamma+1)}$, so choosing c_R sufficiently small ensures the perturbation is dominated by the base density:

$$\epsilon\|\omega\|_\infty \lesssim \inf_{x\in[R,2R]^d} q_0(x) \asymp R^{-(d+\gamma+1)}.$$

Choose c_R so that $|q_b - q_0| \leqslant \eta q_0$ with the fixed η in the baseline compatibility condition. Since $\int \omega = 0$ and $(1+\eta)(1-\tau) \leqslant 1$, every q_b is a density in the same $\mathcal{H}_\gamma$, with no enlargement of C_2 as n, t vary. Gaussian convolution preserves $(1 \pm \eta)$-comparability.

Sobolev Regularity ($q_b \in \mathcal{P}_\mathcal{S}(\beta, L)$):

Let $\Delta_b(x) = \epsilon \sum_{\mathbf{i}} b_{\mathbf{i}}\omega(x - x_{\mathbf{i}})$. Since fractional Sobolev norms are nonlocal, we verify this constraint in Fourier space. Using the Fourier estimate established at the beginning of this appendix, we obtain

$$\int_{\mathbb{R}^d} \|\xi\|^{2\beta} |\widehat{\Delta_b}(\xi)|^2 \mathrm{d}\xi \lesssim \epsilon^2 |\mathcal{I}| h^{d-2\beta} \asymp \epsilon^2 R^d h^{-2\beta}.$$

The same bound controls the inhomogeneous Sobolev norm when $h \leqslant 1$. In the normalized seminorm S_β, choose the fixed construction constants so that $S_\beta(\Delta_b) \leqslant \ell$. Then $S_\beta(q_b) \leqslant S_\beta(q_0) + \ell \leqslant L$, which implies membership in the stated Sobolev ball since $|\xi^\alpha| \leqslant \|\xi\|^\beta$. Thus the condition $\epsilon^2 R^d h^{-2\beta} \lesssim 1$, equivalently $R^{d/2}\epsilon t^{-\beta/2} \lesssim 1$, has a fixed right-hand budget depending on ℓ, which determines the constant in the lower endpoint for t below.

Substituting $\epsilon \asymp n^{-1}t^{-d/2}$ and $R \asymp (nt^{d/2})^{\frac{1}{d+\gamma+1}}$, this imposes a lower bound on t

$$R^{d/2}(n^{-1}t^{-d/2})t^{-\beta/2} \lesssim 1\,,$$

which implies

$$(nt^{d/2})^{\frac{d}{2(d+\gamma+1)}} n^{-1} t^{-(d+\beta)/2} \lesssim 1.$$

Rearranging terms yields $t \gtrsim n^{-\frac{2(d+2\gamma+2)}{d^2+2d\gamma+2d+2\beta(d+\gamma+1)}}$.

Geometric Consistency ($h \lesssim R$): The wavelet support size h must be smaller than the tail location R. This requires $t^{1/2} \lesssim (nt^{d/2})^{\frac{1}{d+\gamma+1}}$. Rearranging gives $t^{\gamma+1} \lesssim n^2$, which is trivially satisfied in the regime of interest where $n \to \infty$ and $t \to 0$. The constants required for separation are compatible with this window: if $A = d(d+2\gamma+2) + 2\beta(d+\gamma+1)$ and $t \gtrsim n^{-2(d+2\gamma+2)/A}$, then $R \gtrsim n^{2\beta/A} \to \infty$. Fix $\Lambda > 1$ sufficiently large to satisfy the preceding separation conditions, and set $t_\mathrm{P} := \Lambda^{-2}$. Then $t \leqslant t_\mathrm{P}$ gives $h \leqslant 1$, hence also $R/h \to \infty$ and $R > 4C_s h\|\psi\|_\infty/\|v_0\|$ for sufficiently large n. The amplitude choice $\epsilon R^D = c_R$ is precisely the uniform smallness condition used for the quadratic terms.

H.2 Proof of Theorem 10

The construction of least favorable hypotheses in the exponential tail regime follows the same overall fashion as in the polynomial case. A key difference, however, makes the analysis considerably simpler: **in the exponential regime the optimal rate is governed by the bulk rather than the tail.**

In the polynomial regime, the minimax rate is driven by sample scarcity in an expanding tail region, which forces us to place perturbations at a large radius $R_n \asymp n^{\frac{1}{d+\gamma+1}} t^{\frac{d}{2(d+\gamma+1)}}$. For exponentially decaying targets, the effective support grows only logarithmically, so it is enough to introduce perturbations on a fixed compact set (or a very slowly growing region) to obtain the parametric rate n^{-1} up to the diffusion time dependence.

Hypothesis construction. Use the smooth baseline $q_0 = q_\mathrm{E}$ defined at the start of this appendix, under its fixed tail-envelope and Sobolev margins. This density is smooth at the origin for every $\gamma > 0$ and belongs to the Schwartz class, so all fixed-order Sobolev seminorms are finite. Place an h-spaced lattice of centers with entire support cubes $x_{\mathbf{i}} + [-h/2, h/2]^d \subset K := [0,1]^d$. For sufficiently small fixed $h_0 > 0$ and $h \leqslant h_0$, its size is $M := |\mathcal{I}| \asymp h^{-d}$. Define

$$q_b(x) = q_0(x) + \epsilon \sum_{\mathbf{i}\in\mathcal{I}} b_{\mathbf{i}}\omega(x - x_{\mathbf{i}}),$$

where the wavelet ω is the same as in the proof of Theorem 9, and the grid spacing is set to $h = \Lambda\sqrt{t}$. Let

$$f_0 = \varphi_t * q_0, \quad f_b = \varphi_t * q_b\,.$$

Scaling. Because the perturbations are supported on a bounded region where $q_0(x) \asymp 1$, the weights $1/f_0(x)$ and $1/q_0(x)$ in the separation and KL bounds, respectively, do not introduce any additional polynomial factor (in contrast to the polynomial tail case). The key scalings become

1. **Separation of scores.** Uniformly for $x \in K$ and $0 < t \leqslant 1$, we have $c_K \leqslant f_0(x) \leqslant C_K$ and $\|s_0(x)\| \leqslant B_K$: the lower bound follows from positivity of q_0 on a fixed neighborhood of K and a fixed Gaussian small-ball probability, while $\|\nabla f_0\|_\infty \leqslant \|\nabla q_0\|_\infty$. Write $U_{\mathbf{i}} = \varphi_t * \omega(\cdot - x_{\mathbf{i}})$ and $Q_{\mathbf{i}} = x_{\mathbf{i}} + [-\delta h, \delta h]^d$. Choose $\delta \leqslant 1/4$ and Λ as in the Gaussian-bump estimates for the polynomial-tail construction, then decrease h_0 so that $B_K h_0 \|\psi\|_\infty \leqslant \|v_0\|/4$. On $Q_{\mathbf{i}}$,

$$\|\nabla U_{\mathbf{i}} - s_0 U_{\mathbf{i}}\| \geqslant \frac{\|v_0\|}{4h},$$
$$\sum_{\mathbf{j}\neq\mathbf{i}} \|\nabla U_{\mathbf{j}} - s_0 U_{\mathbf{j}}\| \leqslant \frac{E_{d,\delta}(\Lambda)}{h}(\|\nabla\psi\|_\infty + B_K h_0\|\psi\|_\infty) \leqslant \frac{\|v_0\|}{8h}.$$

   The last inequality holds after fixing Λ sufficiently large, since $E_{d,\delta}(\Lambda) \leqslant C_d \Lambda^d e^{-\Lambda^2/32}$. Set $S_b = \sum b_{\mathbf{j}} U_{\mathbf{j}}$, $\Delta S = S_b - S_{b'}$, $L_b = \nabla\Delta S - s_0 \Delta S$, and $Q_b = S_{b'}\nabla\Delta S - \Delta S \nabla S_{b'}$. The exact score difference has numerator $\epsilon f_0 L_b + \epsilon^2 Q_b$ and denominator $f_b f_{b'}$. Disjoint supports give $\|S_b\|_\infty \lesssim 1$ and $\|\nabla S_b\|_\infty \lesssim h^{-1}$. Thus on every differing core, $\|L_b\| \geqslant \|v_0\|/(8h)$ and

$$\|\epsilon^2 Q_b\| \leqslant C\epsilon\|\epsilon f_0 L_b\| \leqslant \tfrac{1}{2}\|\epsilon f_0 L_b\|$$

   for sufficiently small fixed amplitude. With $f_b \asymp f_0$, integration over the disjoint cores yields

$$\|s_b - s_{b'}\|^2_{L^2(f_0)} \gtrsim \epsilon^2 h^{d-2} d_H(b, b').$$

   Choose a binary packing $\mathcal{B}$ with $d_H(b, b') \gtrsim M$ and $\log|\mathcal{B}| \gtrsim M$; make h_0 small enough that M exceeds the fixed packing threshold. For distinct packed hypotheses the separation is therefore $\gtrsim \epsilon^2 h^{-2} \asymp \epsilon^2 t^{-1}$. All these bounds hold on $0 < t \leqslant t_{\mathrm{E}} := (h_0/\Lambda)^2$, with $h_0 \leqslant 1$.

2. **KL divergence.** Since the observations are i.i.d. samples from q_b, we bound $\mathrm{KL}(q_b\|q_0)$. The perturbations are supported in a fixed bounded region on which $q_0(x) \asymp 1$ and have disjoint supports, so

$$\mathrm{KL}(q_b\|q_0) \leqslant \chi^2(q_b\|q_0) \lesssim \epsilon^2|\mathcal{I}|h^d \asymp \epsilon^2,$$

   where $M := |\mathcal{I}| \asymp h^{-d} \asymp t^{-d/2}$.

Fano argument. Let $\mathcal{B}$ be a packing set with $\log|\mathcal{B}| \asymp M \asymp t^{-d/2}$. Fano's inequality applies provided

$$\max_{b\in\mathcal{B}} \mathrm{KL}(q_b^{\otimes n}\|q_0^{\otimes n}) \lesssim \log|\mathcal{B}|.$$

The KL bound above shows that this condition holds when $n\epsilon^2 \lesssim t^{-d/2}$. More precisely, take $\epsilon = c_\epsilon (nh^d)^{-1/2}$ with a sufficiently small fixed $c_\epsilon > 0$; then $n\,\mathrm{KL}(q_b\|q_0) \leqslant c\log|\mathcal{B}|$ for a sufficiently small numerical c. In rate notation,

$$\epsilon \asymp n^{-1/2}\, t^{-d/4},$$

so that $n\,\mathrm{KL}(q_b\|q_0) \lesssim t^{-d/2} \asymp \log|\mathcal{B}|$. For sufficiently small perturbation constants, $(1-\eta)q_0 \leqslant q_b \leqslant (1+\eta)q_0$ on the perturbation region. Gaussian convolution preserves this comparison, giving $(1-\eta)f_0 \leqslant f_b \leqslant (1+\eta)f_0$ and hence equivalence of the $L^2(f_b)$ and $L^2(f_0)$ losses. Plugging the chosen amplitude into the separation bound therefore yields

$$\text{Minimax risk} \gtrsim \epsilon^2\, t^{-1} \asymp n^{-1}\, t^{-(1+d/2)}.$$

Verification of constraints. It remains to check that the construction defines valid hypotheses.

- **Positivity and fixed tail envelope**. Since $q_0 \geqslant c > 0$ on K and the bump supports are disjoint, $|q_b - q_0| \leqslant \epsilon\|\psi\|_\infty \leqslant \eta q_0$ on K for sufficiently small amplitude, and $q_b = q_0$ outside K. Together with $\int \omega = 0$ and the fixed baseline margin, this proves positivity, normalization, and membership in the same $\mathcal{H}'_\gamma$. In the time window $t \geqslant A_{\mathrm{E}} n^{-2/(d+2\beta)}$, the amplitude is $O(n^{-\beta/(d+2\beta)}) \to 0$ uniformly, so the condition and the quadratic-term bound both hold for all sufficiently large n.
- **Sobolev regularity.**
  The Fourier almost-orthogonality calculation used in the polynomial-tail construction gives $S_\beta(\Delta_b)^2 \lesssim \epsilon^2 |\mathcal{I}| h^{d-2\beta} \asymp \epsilon^2 h^{-2\beta}$. Require $S_\beta(\Delta_b) \leqslant \ell$, so that $S_\beta(q_b) \leqslant S_\beta(q_0) + \ell \leqslant L$. With $\epsilon = c_\epsilon (nh^d)^{-1/2}$, it suffices that $C c_\epsilon^2/(nh^{d+2\beta}) \leqslant \ell^2$. Choose a fixed $A_{\mathrm{E}} > 0$ sufficiently large that this inequality holds whenever $t \geqslant A_{\mathrm{E}} n^{-2/(d+2\beta)}$, using $h = \Lambda\sqrt{t}$. Thus, in rate notation (with a fixed budget depending on ℓ), the constraint becomes

  $$n^{-1/2} t^{-d/4} \cdot t^{-\beta/2} \lesssim 1\,,$$

  which is equivalent to

  $$t \gtrsim n^{-\frac{2}{d+2\beta}}\,.$$

  This gives the lower time endpoint in Theorem 10; the upper endpoint t_{E} is required by the compact-grid separation.

Conclusion. The resulting lower bound is $n^{-1}t^{-(1+d/2)}$, matching the upper bound in Theorem 5 up to logarithmic factors on $A_{\mathrm{E}} n^{-2/(d+2\beta)} \leqslant t \leqslant t_{\mathrm{E}}$ for the fixed compatible class. This shows that, under exponential tails, the statistical difficulty is dominated by the parametric scaling n^{-1} together with the ill-conditioning of the diffusion kernel through $t^{-(1+d/2)}$, without any additional penalty from tail sparsity.