Minimax Optimality of Score-based Diffusion Models:
Beyond the Density Lower Bound Assumptions

Kaihong Zhang [* 1] **Caitlyn H. Yin** [* 1] **Feng Liang** [1] **Jingbo Liu** [1]

Abstract

We study the asymptotic error of score-based diffusion model sampling in large-sample scenarios from a non-parametric statistics perspective. We show that a kernel-based score estimator achieves an optimal mean square error of $\widetilde{O}\left(n^{-1}t^{-\frac{d+2}{2}}(t^{\frac{d}{2}} \vee 1)\right)$ for the score function of $p_0 * \mathcal{N}(0, t\boldsymbol{I}_d)$, where n and d represent the sample size and the dimension, t is bounded above and below by polynomials of n, and p_0 is an arbitrary sub-Gaussian distribution. As a consequence, this yields an $\widetilde{O}\left(n^{-1/2}t^{-\frac{d}{4}}\right)$ upper bound for the total variation error of the distribution of the sample generated by the diffusion model under a mere sub-Gaussian assumption. If in addition, p_0 belongs to the nonparametric family of the β-Sobolev space with $\beta \leq 2$, by adopting an early stopping strategy, we obtain that the diffusion model is nearly (up to log factors) minimax optimal. This removes the crucial lower bound assumption on p_0 in previous proofs of the minimax optimality of the diffusion model for nonparametric families.

1. Introduction

Diffusion models have emerged as a powerful tool of generative models, demonstrating exceptional performance in a wide range of applications. Pioneering work by Ho et al. (2020) and Dhariwal & Nichol (2021) contributed to the development of image generation. Recent advances have demonstrated the effectiveness of diffusion models in a variety of domains, as exemplified by state-of-the-art results in image and text generation (Ramesh et al., 2022; Nichol et al., 2022), text-to-speech synthesis (Jeong et al., 2021; Popov et al., 2021; Huang et al., 2022), and molecular structure modeling (Xu et al., 2023; Hua et al., 2023). Their reach has extended to scientific fields such as neuroscience (Pinaya et al., 2022) and materials science (Manica et al., 2023).

Score-based generative modeling, a specific family of diffusion modeling, uses learned score functions (i.e., gradients of the log probability density functions) to transform white noise into the target data distribution via solving a stochastic differential equation. More specifically, the forward process converts samples drawn from a data distribution, denoted as p_0 (such as natural images), into complete noise, while the reverse process effectively reverts complete noise back into samples from p_0. Implementing the reverse process requires approximating the score function, a task typically accomplished through training neural networks using a score matching objective, as seen in prior work such as Hyvärinen & Dayan (2005), Vincent (2011) and Song & Ermon (2019). While the score-based generative model has demonstrated remarkable performance in numerous applications, there remain gaps in our theoretical understanding.

From a statistical perspective, score-based diffusion model can be framed as the estimation of an unknown distribution from random samples. Any inaccuracies in score estimation introduce errors into the subsequent diffusion process. This naturally leads to the following question:

Consider viewing the score-based diffusion model as an algorithm designed to generate samples from an unknown distribution, using information from finite training samples generated by the forward process. Under what conditions does this algorithm attain statistically optimal error rate for a given training sample size?

The total error in the sample distribution generated by a diffusion model can be approximately attributed to three components: the error stemming from approximating the true score, the discretization error incurred when discretizing stochastic differential equations, and the error arising from the convergence of the forward process (reflecting the deviation between the initial distribution of the reverse pro-

---
*Equal contribution [1]Department of Statistics, University of Illinois Urbana-Champaign, Champaign, IL, USA. Correspondence to: Kaihong Zhang <kaihong5@illinois.edu>, Caitlyn H. Yin <heqiyin2@illinois.edu>, Feng Liang <liangf@illinois.edu>, Jingbo Liu <jingbol@illinois.edu>.

cess and the standard Gaussian). Existing literature often employs Pinsker's inequality and Girsanov's theorem to bound the total variation distance between the true and approximate reverse processes, as seen in Chen et al. (2023b), Chen et al. (2023a) and Benton et al. (2023).

However, existing results are limited in the following aspects:

Uniform-in-time assumption on score estimation error. In Chen et al. (2023b) and other works such as Benton et al. (2023), it is assumed that the distribution is arbitrary, but there is a uniform (in time) i.e., uniformly over $[0, T]$ upper bound on the score estimation error. The uniform upper bound assumption fails to capture the observation in practice that the score error typically decreases in time, due to the smoothing effect of the Gaussian kernel (see our numerical experiment in Appendix B). Therefore, the bound in Chen et al. (2023b) is not sharp enough to yield the minimax optimal rate. They also did not discuss a natural question in large sample scenarios: the relationship between estimation error and sample size n.

Strong assumptions on the true data distribution. By making an assumption on the data distribution class, it is sometimes possible to resolve the above issue by characterizing the decay of the score error in time. For example, Oko et al. (2023) derived an estimation error bound for the Besov class of density function, showing that the diffusion model is minimax as long as the distribution satisfies a density lower bound on its support. The density lower bound greatly simplifies the proof of the score estimation error bound; however, it excludes natural distribution classes, such as multi-modal distributions or mixtures with well-separated components. In fact, under the compact support and the density lower bound assumption in Oko et al. (2023), we can deduce that the true density satisfies the log-Sobolev inequality (LSI) by the Holley-Stroock perturbation principle (Holley & Stroock, 1987). It is known that under LSI, running the Langevin dynamics is sufficient for achieving statistical efficiency (Koehler et al., 2022), while the diffusion model is expected to achieve efficiency for a wider class of distributions, by introducing a class of smoothed distributions. In this regard, the theory of Oko et al. (2023) has not uncovered the true strength of the diffusion model. This paper uses a different proof technique than Oko et al. (2023) and does not have this limitation.

Comparing the above two lines of work, it might seem that a strong distribution assumption is a price we need to pay for deriving sharp bounds for settings where the score error is not uniform in time (such as the nonparametric class). In this paper, we show that, perhaps surprisingly, we can derive a general sampling error bound only using a Sub-Gaussian assumption, which, when specialized to the nonparametric class (without the density lower bound assumption), is

nearly (up to polylog factors) minimax optimal.

1.1. Main Contributions

In this paper, we derive a time-dependent error bound for the score function $s_t(x) = \nabla \log p_t(x)$ (see Theorem 3.5). Utilizing the error bound for the score function, we are able to control the total variation distance between the true data distribution p_0 and the distribution of a sample from the diffusion model as described in Algorithm 1. Our main results only require the first or both of the following assumptions on the ground true data distribution p_0. We will make these statements rigorous in Section 3.2.

A1 The true data distribution p_0 is σ_0-Sub-Gaussian.

A2 The true data distribution p_0 belongs to the Sobolev class of density functions with the order of smoothness $\beta \leq 2$.

In particular, under **A1** and **A2**, we show that the error in the distribution of the sample generated by the diffusion model coincides with the classical minimax optimal convergence rate of density estimation in nonparametric statistics (Tsybakov, 2009). This rate cannot be improved, since sampling error upper bounds the density estimation error. We summarize our main result informally as follows:

Theorem 1.1 (Informal; see Theorem 3.5 and Theorem 3.8)**.** *Suppose that p_0 satisfies assumption **A1**, and $C > 0$ is arbitrary. There exists a score estimator $\hat{s}_t(x)$ $(t > t_0)$ such that for the early stopping time $t_0 = n^{-C}$, The distribution of sample from Algorithm 1 differs from p_{t_0} by at most $\mathrm{polylog}(n)n^{-1/2}t_0^{-\frac{d}{4}}$ in the total variation (TV) distance, where p_{t_0} denotes the distribution of the forward process at time t_0. Furthermore, if p_0 also fulfils **A2**, by choosing $t_0 = n^{-\frac{2}{2\beta+d}}$, the distribution of the output sample differs from p_0 by at most $\mathrm{polylog}(n)n^{-\frac{\beta}{2\beta+d}}$ in TV.*

In previous studies, several restrictive assumptions were made to construct minimax optimal score estimators for diffusion models. One of the key assumptions is that p_0 has compact support and whose density is bounded from below (Oko et al., 2023). As alluded to before, these assumptions are strong enough to guarantee LSI and hence fail to unveil the key advantage of the diffusion model over the Langevin dynamics (Koehler et al., 2022). It is not at all obvious that the density lower bound assumption can be removed without impairing the convergence rate; in fact, in some examples of nonparametric statistics, such as density estimation under Wasserstein error, the minimax rate can indeed get worse without the density lower bound assumption (Niles-Weed & Berthet, 2022).

In this work, we employ a truncated version of the score estimator similar to that of Zhang (1997) in the context of

empirical Bayes, but with more refined analysis to control the error in the regime of polynomially small t. More precisely, we note that $s_t(x) = \nabla p_t(x)/p_t(x)$, and estimate the numerator and the denominator by carefully constructed kernel density estimators. We define the truncated score estimator by

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

where $\hat{p}_t(x)$ is a kernel density estimator of original data distribution $p_t(x)$ [see Appendix C.1 for details about properties of the kernel needed and proof of its existence]. We set our score estimator $\hat{s}_t(x)$ to be zero when the kernel density estimator $\hat{p}_t(x)$ is less than ρ_n, where $\rho_n := \frac{1}{nt^{d/2}}$ is a parameter decaying at the same rate as MSE of $\hat{p}_t$. If $\hat{p}_t(x)$ is larger than ρ_n, the MSE for $\nabla\hat{p}_t(x)$ is proportional to $p_t(x)$ (Proposition 4.1), so that $p_t(x)\text{MSE}(\hat{s}_t(x)) \approx \text{MSE}(\nabla\hat{p}_t(x))/p_t(x)$ is unaffected by the lower density. If $\hat{p}_t(x)$ is smaller than ρ_n, there will be too few observations near x, and we cannot show that $\text{MSE}[\nabla\hat{p}_t(x)]$ is proportional to $p_t(x)$ via Bernstein's concentration; however, the sub-Gaussian assumption ensures that the contribution to the mean square score estimation error from this case is *exactly* bounded at the level of the minimax rate. See Section 3.3 for details on the convergence of our truncated score estimator.

To summarize, we provide a convergence guarantee for the diffusion model in the sense of total variation (TV) distance under mild assumptions on the true distribution p_0. Our main results Corollary 3.7 and Theorem 3.8, obtain a convergence rate of $n^{-\frac{\beta}{2\beta+d}}\text{polylog}(n)$ for sampling from p_0 using the score-based diffusion model, which matches the minimax optimal rate in the classical nonparametric estimation theory (Tsybakov, 2009) up to logarithmic factors. While the recent work of Oko et al. (2023) showed a similar minimax optimality result, they made the restrictive assumptions that the data distribution p_0 has bounded support and a density lower bounded, i.e., $p_0(x) > C$ for some constant C. Here we relax the compact support assumption to sub-Gaussianity and completely remove the density lower bound.

1.2. Prior Works

Regarding the theoretical convergence results of score-based generative models, previous works, such as Benton et al. (2023), Chen et al. (2023a), Chen et al. (2023b), and Lee et al. (2023), often assume an oracle score estimator with bounded error, without explicitly touching on the issue of the statistical error of the score estimator for finite samples. For example, in Chen et al. (2023b) it is assumed that their

exists some estimator s_θ satisfying a uniform in time bound

$$\mathbb{E}_{q_{t_k}(\mathbf{x})}[\|\nabla \log q_{T-t_k}(\mathbf{x}) - s_\theta(\mathbf{x}, T - t_k)\|^2] \le \varepsilon_{\text{score}}^2, \tag{2}$$

where t_k are the time discretization points, and showed that the resulting reverse diffusion process produces a sample with error roughly the order of $\varepsilon_{\text{score}}$. It is still not clear whether such s_θ exists, or if the uniform in time assumption (2) is valid, in specific settings. In contrast, our study incorporates the statistical error in score estimation, by applying a (refined) truncated score estimator similar to the one in Zhang (1997). This provides a unique perspective for analyzing score functions within the low-density regions.

Furthermore, the broader landscape of score-generative models relies on a variety of assumptions regarding the true distribution p_0, as discussed in prior works including Song et al. (2021b), Pidstrigach (2022), and De Bortoli (2022). These studies assumed the true distribution p_0 is supported on a low dimensional manifold $\mathcal{M} \subset \mathbb{R}^d$, with a smooth density relative to the manifold. De Bortoli (2022) specifically highlight that, without further assumptions, the generalization error exhibits exponential dependence on both the diameter of $\mathcal{M}$ and the reciprocal of the desired error margin. Further research in this domain, such as Block et al. (2022), explored the convergence of score estimation error and finite-sample bounds for sampling using Langevin diffusion.

Another line of work imposes the restrictive assumption of smoothness conditions such as a log-Sobolev Inequality (LSI) for the true distribution p_0, as discussed by Lee et al. (2022), Wibisono & Yang (2022). These studies provide insights into polynomial convergence guarantees for score function estimation in TV distance under L^2-accurate error. Building upon this, Chen et al. (2023b) and Lee et al. (2023) have extended the analysis by removing the LSI assumption, addressing the complexities of real-world data distributions, although they still require smoothness conditions and bounded support for the true distribution.

Closer to our line of work are the results of Oko et al. (2023), which provides a theoretical analysis of approximation and generalization abilities of score-based diffusion modeling under Besov function spaces and other specific assumptions, such as the true distribution being supported on a bounded domain and having a density lower bound. They gave an upper bound for the total generalization error of $n^{-\frac{2s}{d+2s}}\log^{18} n$ for the score network estimation, which matches the optimal convergence rate in the nonparametric function class setting up to logarithmic factors. Considering the neural network architecture's covering number, they minimized the empirical score-matching error.

After we finalized our manuscript, we noted a concurrent work by Wibisono et al. (2024) shares similarities with our

work. They constructed a regularized version of the score estimator

$$\hat{s}_h^\epsilon(x) := \frac{\nabla \hat{p}_h(x)}{\max(\hat{p}_h(x), \epsilon)},$$

where $\hat{p}_h(x)$ is the kernel density estimator with the Gaussian kernel. Using this estimator, they achieved an optimal rate of $\tilde{O}(n^{-\frac{2}{d+4}})$ for estimating the score function of an unknown probability distribution that is sub-Gaussian and has score function that is Lipschitz-continuous.

1.3. Organization

The rest of the paper is organized as follows. In Section 2, we establish the background of score-based generative models in our settings, detailing the necessary notations and assumptions. Section 3 presents our main results concerning the estimation error of the score function (over time t), and the bound for the TV distance between the data distribution and the distribution of sample from the diffusion model in Algorithm 1. Section 4 provides proof overviews for the theorems in Section 3. We provide some discussions of this paper in Section 5. Finally, we briefly summarize our main findings in Section 6.

2. Background

2.1. Forward and Backward Processes

Forward Process. Given the ground truth data distribution p_0, a forward process $(X_t)_{t\in[0,T]}$ is defined as the solution for the following Itô SDE:

$$dX_t = F(X_t, t)dt + g(t)dB_t, \quad X_0 \sim p_0, \quad (3)$$

where $F(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$ is a vector-valued function called the drift coefficient, $g(t) : \mathbb{R} \to \mathbb{R}$ is a scalar function called the diffusion coefficient, and $(B_t)_{t\in[0,T]}$ denotes a standard Brownian motion.

In this paper, we focus on the case of Brownian motion process,

$$dX_t = dB_t, \quad X_0 \sim p_0, \quad (4)$$

which is a special case of (3). In this example the drift term $F(\cdot, t) = 0$ and the diffusion coefficient is a constant $g(t) = 1$. This SDE has an explicit solution

$$X_t = X_0 + \sqrt{t}Z. \quad Z \sim \mathcal{N}(0, I_d) \perp\!\!\!\perp X_0. \quad (5)$$

Then we have

$$X_t | X_0 \sim \mathcal{N}(X_0, t\,I_d). \quad (6)$$

Note that if we perturb our original data X_0 with Gaussian noise B_T for a large enough time T, the marginal distribution of X_T will have only a weak dependence on X_0 and will be approximately Gaussian distributed. We hereafter

denote by $p_t(x)$ the probability density function of X_t in (5).

Reverse Process for Sample Generation. If we reverse the diffusion process (3) in time, we can generate new samples from p_0. Importantly, denotes $Y_t = X_{T-t}$, where $(X_t)_{t\in[0,T]}$ is a solution to the SDE in (3), then $(Y_t)_{t\in[0,T]}$ satisfies:

$$dY_t = \left[-F(Y_t, T-t) + g^2(t)\nabla \log p_{T-t}(Y_t)\right] dt + g(t)dB_t, \quad Y_0 = X_T \sim p_T, \quad (7)$$

where $(B_t)_{t\in[0,T]}$ is another independent standard Brownian motion. Here, $\nabla \log p_t(x)$ is called the *score function* for p_t. If we run the process (7) with $Y_0 \sim p_T$, then $Y_T \sim p_0$ and we obtain a sample from the true data distribution. Setting $F(\cdot, t) = 0$ and $g(t) = 1$, the corresponding reverse SDE for the Brownian motion process (4) is

$$dY_t = \nabla \log p_{T-t}(Y_t)dt + dB_t, \quad Y_0 = X_T \sim p_T. \quad (8)$$

To compute the reverse SDE, we need to construct an estimator $\hat{s}_t(x)$ for the score function $\nabla \log p_t(x)$ from the dataset $\{x_i\}_{i=1}^n$. The *score error* (SE) (Song et al., 2021c) for the estimator $\hat{s}_t(x)$ is defined as:

$$SE(\hat{s}) := \int_{t\in[0,T]} \mathbb{E}_{x\sim p_t}\left[\|\nabla \log p_t(x) - \hat{s}_t(x)\|_2^2\right] dt. \quad (9)$$

The expectation of score error over the dataset $\mathbb{E}_{\{x_i\}_{i=1}^n}[SE(\hat{s})]$ can be used to evaluate the performance of the proposed score estimator. Once we have an estimator of the score function, we can plug it into the reverse SDE and solve it to get a sample from the original data distribution $p_0(x)$.

2.2. Sampling Method

In order to sample from the unknown distribution p_0, in this paper we utilize the Brownian diffusion process $(X_t)_{t\in[0,T]}$ as defined in (4) and the corresponding backward process $(Y_t)_{t\in[0,T]}$ defined in (8). If we run the forward process $(X_t)_{t\in[0,T]}$ for a large enough time T, the distribution of X_T can be approximated as Gaussian. Suppose we obtain an estimator $\hat{s}_t$ for the score function $\nabla \log p_t(x)$ from the dataset $\{x_i\}_{i=1}^n$, we can run the backward SDE starting from Gaussian distribution with the unknown score function replaced by the score estimator:

$$d\widehat{Y}_t = \hat{s}_{T-t}(\widehat{Y}_t)dt + dB_t, \quad \widehat{Y}_0 \sim \mathcal{N}(0, TI_d). \quad (10)$$

To avoid learning the true score function, $\nabla \log p_0(x)$, which can lead to large score estimation errors due to low-density regions in p_0, we introduce an early stopping time t_0. Consequently, $\widehat{Y}_{T-t_0}$ is a sample generated from the model.

Algorithm 1 Brownian diffusion model

1: **Input:** The score estimator $\hat{s}_t$, early stopping time t_0, and a large enough time T.
2: Sample z from $\mathcal{N}(\mathbf{0}, T\mathbf{I}_d)$.
3: Solve the backward SDE

$$\mathrm{d}\mathbf{Y}_t = \hat{s}_{T-t}(\mathbf{Y}_t)\mathrm{d}t + \mathrm{d}\mathbf{B}_t$$

  with $\mathbf{Y}_0 = z$
4: **Output:** $\mathbf{Y}_{T-t_0}$, a sample generated from the model.

We summarize the diffusion model based on the Brownian motion (5) in Algorithm 1. Note that discretizing the process for solving reverse SDE is not included here. But in practice, we need to use a discrete-time approximation for the process of sampling (10). This problem has been discussed in many previous works. For example, Chen et al. (2023b) in their Theorem 2 addressed the L-Lipschitz regularity condition on $\nabla \log q_t$ for all t, based on the TV metric. Theorem 1 of Chen et al. (2023a) also regarded the L Lipschitz condition, to the Kullback–Leibler (KL) divergence metric, etc. A recent work Benton et al. (2023) controlled the discretization error more precisely of the reverse SDE, based on tools from stochastic localization initially developed in El Alaoui et al. (2022), Montanari (2023). Without the Lipschitz continuity condition, Oko et al. (2023) provided further insight regarding this issue, while considering other assumptions, including Besov spaces and density lower bounds.

2.3. Connections with Ornstein–Uhlenbeck Process

Many works on diffusion models utilize the Ornstein-Uhlenbeck (OU) process rather than the Brownian motion process described in (5). The OU process is a special case of the Itô stochastic differential equation in (3), with drift function $\mathbf{F}(\mathbf{X}_t, t) = -\mathbf{X}_t$ and diffusion function $g(t) = \sqrt{2}$:

$$\mathrm{d}\mathbf{X}_t = -\mathbf{X}_t\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t, \quad \mathbf{X}_0 \sim p_0. \quad (11)$$

The solution $(\mathbf{X}_t)_{t \in [0,T]}$ admits a closed form:

$$\mathbf{X}_t = e^{-t}\mathbf{X}_0 + \sqrt{1 - e^{-2t}}\mathbf{Z} \quad \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d) \perp\!\!\!\perp \mathbf{X}_0. \quad (12)$$

As stated previously in (5), the solution to the Brownian motion process $(\mathbf{Y}_t)_{t \in [0,T]}$ satisfies:

$$\mathbf{Y}_t = \mathbf{Y}_0 + \sqrt{t}\mathbf{Z} \quad \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d) \perp\!\!\!\perp \mathbf{Y}_0. \quad (13)$$

If we set $\mathbf{X}_0 = \mathbf{Y}_0$ and make the time transformation $t = e^{2s} - 1$, then it follows that $e^s \mathbf{X}_s \stackrel{\mathrm{d}}{=} \mathbf{Y}_t$. Therefore, the Ornstein-Uhlenbeck process can be viewed as a time transformation and scaled version of the Brownian motion process. For the analysis in this paper, we will utilize the Brownian motion process given in (5). The results can then be generalized to the Ornstein-Uhlenbeck process due to the equivalence of these two processes.

3. Main Results

In this section, we present our main results of score estimation error bound and the optimal convergence rate for diffusion models.

3.1. Notation

Here we outline notations and definitions used throughout the paper (see details in Appendix A). $p_0 : \mathbb{R}^d \to \mathbb{R}$ represents the true data distribution's density, with $p_t(x) := p_0 * \phi_t(x)$ (convolution with Gaussian density ϕ_t of $\mathcal{N}(0, t\mathbf{I}_d)$). The term "polylog" refers to $\mathrm{polylog}(n) := (\log n)^C$ for some constant C.

The Fourier transform of function f is given by:

$$\mathcal{F}[f](\omega) = \int_{\mathbb{R}^d} f(x)e^{-i\langle \omega, x\rangle}dx,$$

where $\langle \omega, x \rangle$ is the inner product of ω and x in $\mathbb{R}^d$. For multi-index α and a vector ω, $\omega^\alpha := \prod_{i=1}^d w_i^{\alpha_i}$. We use $TV(p, q)$ to denote the total variation distance between two probability distributions p and q. Similarly, $\mathrm{D}_{\mathrm{KL}}(p\|q)$ denotes KL divergence.

3.2. Assumptions

In this subsection, we will introduce the assumptions that are imposed on the true data distribution p_0 in our analysis. Roughly speaking, the data distribution is assumed to be sub-Gaussian and in the Sobolev class of densities.

Assumption 3.1. The true distribution p_0 is σ_0-Sub-Gaussian.

While the sub-Gaussian random variables in 1-dimension have been well-studied, it is still necessary to clarify the definition of a sub-Gaussian random vector in higher dimensions. A natural approach to define the Sub-Gaussian distribution in higher dimensions is through the projections onto lines.

Definition 3.2 (Sub-Gaussian random vectors, Vershynin (2018)). A random vector $X \in \mathbb{R}^n$ is said to be Sub-Gaussian if the one-dimensional marginals $\langle X, v\rangle$ are Sub-Gaussian random variables for all $v \in \mathbb{R}^n$. The Sub-Gaussian norm of X is defined as

$$\|X\|_{\psi_2} = \sup_{v \in S^{n-1}} \|\langle X, v\rangle\|_{\psi_2}, \quad (14)$$

where S^{n-1} is the unit sphere in $\mathbb{R}^d$ and the Orlicz norm of random variable Y with respect to the function $\psi_2(x) = e^{x^2} - 1$ is defined as:

$$\|Y\|_{\psi_2} = \inf\left\{C > 0 : \mathbb{E}\left[\psi_2\left(\frac{|Y|}{C}\right)\right] \leq 1\right\}. \quad (15)$$

We say a random vector $X \in \mathbb{R}^d$ is σ-Sub-Gaussian if $\sigma := \|X\|_{\psi_2} < \infty$.

Remark 3.3. The sub-Gaussian assumption is relatively mild. It is well known that a Log-Sobolev inequality (LSI) implies sub-gaussian decay of the tails (see Vershynin (2018)), but the reverse implication is not true (consider for example a distribution supported on the union of two disjoint intervals). The LSI is a common assumption on true data distribution in many prior works, such as Lee et al. (2022); Wibisono & Yang (2022). In the work of Oko et al. (2023), the true data distribution is assumed to have bounded support and density bounded below, which implies the LSI by the Holley-Stroock principle (Holley & Stroock, 1987). However, the assumption that the true density is bounded below is unrealistic in practice, as we would expect the true data distribution to have many low-density regions. This can lead to challenges in score estimation, as discussed in Song & Ermon (2019).

As we will see in Theorem 3.5, it suffices to derive an estimation error bound for the score function $s_t(x)$ only under Assumption 3.1. This in turn implies the minimax optimal rate for the nonparametric class under the additional Sobolev class assumption.

Assumption 3.4. The true distribution p_0 belongs to the Sobolev class of density (Tsybakov, 2009) with $\beta \leq 2$. Specifically, for $\beta, L \in \mathbb{R}_+$, the Sobolev class of density is defined as followed:

$$\mathcal{P}_\mathcal{S}(\beta, L) = \Big\{ p \in \mathcal{L}^1(\mathbb{R}^d) \,|\, p \geq 0, \int p = 1,$$

$$\forall \alpha \text{ with } \sum_{i=1}^d \alpha_i = \beta, \int |\omega^\alpha|^2 |\mathcal{F}[p](\omega)|^2 \, d\omega \leq (2\pi)^d L^2 \Big\}.$$

Note that instead of using the common definition of the Sobolev class, this definition utilizes the Fourier transformation of the density p, thereby allowing each α_i to take values not only as integers but also as positive real numbers.

3.3. Analysis of Score Estimation Error

As illustrated in Section 2, the performance of the score-based diffusion model of Algorithm 1 highly depends on the estimation accuracy of the score function $s_t(x) = \nabla \log p_t(x)$ at each time $t \geq t_0$. The following theorem provides an upper bound for score estimation error. Notably, the score estimation error decays as time increases, implying that adding sufficient noise to the samples makes the score learning procedure easier.

Theorem 3.5. *Denote the true score function* $s_t(x) := \nabla \log p_t(x)$. *Suppose that Assumption 3.1 holds. Then there exists a score estimator* $\hat{s}_t(x)$ *constructed from i.i.d samples* $\{x_i\}_{i=1}^n \sim p_0$ *such that for any constants* $T_1, T_2 > 0$, *let*

$t_0 := n^{-T_1}$ *and* $T := n^{T_2}$. *Then for any* $t_0 < t < T$,

$$\mathbb{E}\left[\int_x \|\hat{s}_t(x) - s_t(x)\|^2 p_t(x) \, dx \right]$$

$$\lesssim \text{polylog}(n) \, n^{-1} t^{-\frac{d+2}{2}} \left(t^{\frac{d}{2}} + \sigma_0^d \right), \quad (16)$$

where $\lesssim$ *hide the constant that does not depend on time* t *and sample size* n.

The formal proof can be found in Appendix D.1.

Remark 3.6. From this theorem, we can see that the use of the early stopping technique is crucial for ensuring that the analysis proceeds correctly. If we set $t_0 = 0$, the upper bound on the right-hand side in Theorem 3.5 would blow up, resulting in a very large score estimation error. This is because, without early stopping, the estimation error for the score function $\nabla \log p_0(x)$ at $t = 0$ would be excessively high due to the low-density regions in p_0. We can address this issue by introducing an early stopping time t_0. Properly setting the early stopping time t_0 is essential to achieve the desired minimax rate, as will be shown in Corollary 3.7 and Theorem 3.8.

After integrating over time in Theorem 3.5, we will get the following result, and the proof is provided in Appendix D.2.

Corollary 3.7. *Suppose that* p_0 *satisfies Assumption 3.1, and* $T_1, T_2 > 0$ *is arbitrary.*

1. *For* $t_0 = n^{-T_1}$ *and* $T = n^{T_2}$, *we have*

$$\int_{t=t_0}^T \mathbb{E}\left[\int_x \|\hat{s}_t(x) - s_t(x)\|^2 p_t(x) \, dx \right] dt$$

$$\lesssim \text{polylog}(n) \, n^{-1} t_0^{-d/2}. \quad (17)$$

2. *In particular, for any* $\beta > 0$, *let* $t_0 = n^{-\frac{2}{2\beta+d}}$ *and* $T = n^{T_2}$, *we have*

$$\int_{t=t_0}^T \mathbb{E}\left[\int_x \|\hat{s}_t(x) - s_t(x)\|^2 p_t(x) \, dx \right] dt$$

$$\lesssim \text{polylog}(n) \, n^{-\frac{2\beta}{2\beta+d}}. \quad (18)$$

In the first part of Corollary 3.7, The upper bound for the cumulative score error is given by $n^{-1} t_0^{-d/2}$, where t_0 is an early stopping time that can be freely chosen. Notably, Corollary 3.7 does not impose any assumptions on p_0 about nonparametric class and requires only Sub-Gaussianity. However, in the further analysis in Theorem 3.8 of the TV error of the diffusion model, if we further assume that p_0 is in the Sobolev class of density with smoothness parameter β (Assumption 3.4), setting $t_0 = n^{-\frac{2}{2\beta+d}}$ can adjust the error bound to $n^{-\frac{2\beta}{2\beta+d}}$. Remark that if we choose $t_0 \leq n^{-\frac{2}{2\beta+d}}$ (as in (Oko et al., 2023)), then the optimal rate may no longer be achieved in (18) without further assumptions such as the density lower bound.

3.4. Analysis of Estimation Error for Diffusion Model

In this section, we analyze the estimation error of the diffusion model in Algorithm 1, based on the score estimation error results from Section 3.3. The following theorem concerns the bound of TV distance between the true data distribution p_0 and the samples generated by Algorithm 1.

Theorem 3.8. *Suppose that Assumption 3.1 and Assumption 3.4 holds. Let $\widehat{\boldsymbol{Y}}_{T-t_0}$ be the output of the diffusion model in Algorithm 1 at time $T - t_0$. Let $t_0 = n^{-\frac{2}{2\beta+d}}$ and $T = n^{\frac{2\beta}{2\beta+d}}$, then there exists a score estimator $\hat{s}_t$ such that*

$$\mathbb{E}\left[\mathrm{TV}\left(\boldsymbol{X}_0, \widehat{\boldsymbol{Y}}_{T-t_0}\right)\right] \lesssim \mathrm{polylog}(n)\, n^{-\frac{\beta}{2\beta+d}}. \quad (19)$$

Remark 3.9. This sampling error rate coincides with the classical minimax rate in nonparametric density estimation (up to logarithmic factors) (Stone, 1980; 1982; Tsybakov, 2009), hence must be optimal. The sampling distribution $\widehat{\boldsymbol{Y}}_{T-t_0}$ is considered minimax optimal because the sampling error is lower-bounded by the estimation error. To see this, suppose there exists a sampler $\widehat{\boldsymbol{Z}}$ that produces samples with a smaller TV distance $\mathbb{E}\left[\mathrm{TV}\left(\boldsymbol{X}_0, \widehat{\boldsymbol{Z}}\right)\right]$ than the minimax rate. We could then generate many independent samples from $\widehat{\boldsymbol{Z}}$ to construct a density estimator for p_0 with an error smaller than the minimax rate, leading to a contradiction. Therefore, the sampling error cannot be smaller than the density estimation error.

Remark 3.10. In the proof of Theorem 3.8 (see Section 4.2), we only use the Sobolev class assumption (Assumption 3.4) when controlling the early stopping error $\mathrm{TV}\left(\boldsymbol{X}_0, \boldsymbol{X}_{t_0}\right)$. Therefore, under only the sub-Gaussian assumption, we can generalize this theorem to other nonparametric classes, provided we can control the early stopping error: $\mathrm{TV}\left(\boldsymbol{X}_0, \boldsymbol{X}_{t_0}\right)$. Using part 1 of Corollary 3.7 to control the error of score estimation, the sampling error of the diffusion model for a general sub-Gaussian p_0 can be bounded as follows:

$$\mathbb{E}\left[\mathrm{TV}\left(\boldsymbol{X}_0, \widehat{\boldsymbol{Y}}_{T-t_0}\right)\right]$$
$$\lesssim \inf_{t_0}\left(\mathrm{TV}\left(\boldsymbol{X}_0, \boldsymbol{X}_{t_0}\right) + \mathrm{polylog}(n)\, n^{-1/2} t_0^{-d/4}\right).$$

In Theorem 3.8, under the Sobolev class assumption, the desired rate is obtained by $t_0 = n^{-\frac{2}{2\beta+d}}$.

4. Proof Overview

We now provide the proof sketches for Theorem 3.5 and Theorem 3.8. Formal proofs are provided in Appendix D.1 and Appendix E, respectively.

4.1. Proof sketch of Theorem 3.5

Construction of the kernel score estimator. We first construct the score estimator for proving Theorem 3.5. The

score function associated with perturbed data distribution p_t is $s_t(x) = \nabla \log p_t(x) = \frac{\nabla p_t(x)}{p_t(x)}$. It is natural to first construct a density estimator $\hat{p}_t(x)$ for $p_t(x)$, then use the plug-in estimator $\hat{s}_t(x) = \frac{\nabla \hat{p}_t(x)}{\hat{p}_t(x)}$ to estimate the score function. Here, we will use the kernel density estimator (KDE) to estimate $p_t(x)$. Specifically, given samples $X_i \overset{\text{i.i.d.}}{\sim} p_0$ and $Z_{i,t} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, tI_d\right)$, by definition of the forward process in (4), $X_i^t = X_i + Z_{i,t} \overset{\text{i.i.d.}}{\sim} p_t$. Given a kernel K_d with order $\ell \geq 1$ (Definition 1.3 of Tsybakov (2009), also see Definition C.1) and with bounded support $[-1,1]^d$ (we can construct a kernel satisfying these properties using the Legendre polynomials, see Lemma C.2 for details), the KDE for $p_t(x)$ is defined as follows:

$$\hat{p}_t(x) = \frac{1}{nh^d} \sum_{i=1}^n K_d\left(\frac{x - X_i^t}{h}\right),$$

where h is the bandwidth to be specified later. Naively estimating the score by $\frac{\nabla \hat{p}_t(x)}{\hat{p}_t(x)}$ would lead to poor performance for x such that $\hat{p}_t(x)$ is small. To avoid this issue, we introduce the truncated version of the kernel score estimator: for any time t, let the threshold be $\rho_n := \frac{1}{nt^{d/2}}$. The truncated kernel score estimator is defined as follows:

$$\hat{s}_t(x) := \frac{\nabla \hat{p}_t(x)}{\hat{p}_t(x)} \mathbb{1}_{\{x:\, \hat{p}_t(x) \geq \rho_n\}}. \quad (20)$$

Mean Squared Error Analysis. The pointwise MSE for $\hat{p}_t(x)$ and $\nabla \hat{p}_t(x)$ are defined as follows:

$$\mathrm{MSE}(\hat{p}_t(x)) := \mathbb{E}\left[|\hat{p}_t(x) - p_t(x)|^2\right],$$
$$\mathrm{MSE}(\nabla \hat{p}_t(x)) := \mathbb{E}\left[\|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2\right].$$

We follow the common framework of bias-variance trade-off Tsybakov (2009) with slight modifications, which allows us to obtain a sharper bound that depends on the location x. The result of MSE bounds for $\hat{p}_t$ and $\nabla \hat{p}_t$ is summarized in the following proposition, the proof can be found in Appendix C.2.

Proposition 4.1 (See also Proposition C.3, C.4). *If we take the bandwidth $h = C\sqrt{\frac{t}{\log n}}$ for some constant C, and the order of kernel $\ell = \log n$, then*

$$\mathrm{MSE}(\hat{p}_t(x)) \lesssim \mathrm{polylog}(n)\left(\frac{p_t^*(x)}{nt^{\frac{d}{2}}} + (\log n)^{-\log n}\right),$$

$$\mathrm{MSE}(\nabla \hat{p}_t(x)) \lesssim \mathrm{polylog}(n)\left(\frac{p_t^*(x)}{nt^{\frac{d+2}{2}}} + (\log n)^{-\log n} t^{-1}\right).$$

where $p_t^*(x) := \sup_{\|\lambda\|_\infty < h} p_t(x+\lambda)$ is the local maximum of the density function p_t.

Remark 4.2. The bound for MSE depends on the point x, dimension d, which allows us to obtain a sharper bound in the low-density area of p_t.

With these results, we can proceed to derive the upper bound (16) in Theorem 3.5. The main challenge is to control the error in low-density areas. To address this, we consider the following three cases. In the first case where $p_t(x) > \rho_n \log^c n$ for some constant c, the density is relatively high. Using the concentration property of the KDE $\hat{p}_t(x)$ (see Lemma F.6), we find that with high probability, $p_t(x)$ and $\hat{p}_t(x)$ will approximately have the same order. As a result, $\hat{p}_t(x) > \rho_n$, which implies $\hat{s}_t(x) = \frac{\nabla \hat{p}_t(x)}{\hat{p}_t(x)}$. Then by triangle inequality (See lemma D.1),

$$\|\hat{s}_t(x) - s_t(x)\|^2 = \left\| \frac{\nabla \hat{p}_t(x)}{\hat{p}_t(x)} - \frac{\nabla p_t(x)}{p_t(x)} \right\|^2$$
$$\lesssim \frac{\|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2 + \|s_t(x)\|^2 |\hat{p}_t(x) - p_t(x)|^2}{(\hat{p}_t(x))^2}$$

Hence, the score error in (16) can be controlled using the MSE results for $\hat{p}_t(x)$ and $\nabla \hat{p}_t(x)$ from Proposition 4.1.

In case 2, we consider the lower density areas where $p_t(x) < \rho_n \log^{-c} n$. With high probability, $\hat{p}_t(x) < \rho_n$ and therefore, $\hat{s}_t(x) = 0$. The score error in (16) is now

$$\int_{p_t(x) < \rho_n \log^{-c} n} \|s_t(x)\|^2 p_t(x) dx. \quad (21)$$

This represents the expectation of the score function's second moment, restricted to the lower-density area. For any integer $m \geq 1$, Hölder's inequality yields,

$$(21) \leq (\mathbb{E}[\|s_t(X_t)\|^m])^{2/m} (\mathbb{P}[p(X_t) \leq \rho_n \log^{-c} n])^{1-\frac{2}{m}},$$

where $X_t \sim p_t$. The moment of score can be bounded using the result in Bobkov (2019) (See Lemma F.16), i.e., $\mathbb{E}[\|s_t(X_t)\|^m] \lesssim t^{-\frac{m}{2}}$. By Markov's inequality, the second term can be upper bounded in terms of the Rényi Entropy (Rényi, 1961) of X_t. Combining this with the Rényi entropy bound for the sub-Gaussian distribution we derive in Appendix F.6, we obtain the desired result for case 2.

Finally, in case 3 where $\rho_n \log^{-c} n < p_t(x) < \rho_n \log^c n$, we must consider both $\hat{s}_t(x) \neq 0$ and $\hat{s}_t(x) = 0$, since neither is guaranteed with high probability. However, since p_t has both lower bound and upper bound, we can thereby apply the methods from both Case 1 and Case 2 to bound this term. Combining three cases we can prove the result in Theorem 3.5.

4.2. Proof sketch of Theorem 3.8

Here we sketch the proof of Theorem 3.8. Full details of the proof of Theorem 3.8 can be found in Appendix E.

We first recall some notations defined in Section 2:

- The forward process in (4) is denoted by $(X_t)_{t \in [0,T]}$ and $X_t \sim p_t$;

- The backward process (8) is denoted by $(Y_t)_{t \in [0,T]}$ and by definition, $Y_t \sim p_{T-t}$;

- The process of Algorithm 1 is denoted by $(\widehat{Y}_t)_{t \in [0,T]}$.

Let $(\bar{Y}_t)_{t \in [0,T]}$ be $(\widehat{Y}_t)_{t \in [0,T]}$ replacing $\widehat{Y}_0 \sim \mathcal{N}(0, T I_d)$ by $\bar{Y}_0 \sim p_T$, i.e., $(\bar{Y}_t)_{t \in [0,T]}$ satisfies:

$$d\bar{Y}_t = \hat{s}_{T-t}(\bar{Y}_t)dt + dB_t, \quad \bar{Y}_0 \sim p_T.$$

Using triangle inequality of TV distance,

$$\mathbb{E}[\text{TV}(X_0, \widehat{Y}_{T-t_0})] \leq \text{TV}(X_0, X_{t_0})$$
$$+ \mathbb{E}[\text{TV}(Y_{T-t_0}, \bar{Y}_{T-t_0})] + \text{TV}(X_T, \mathcal{N}(0, T I_d)).$$

Therefore, it suffices to control three errors: the error from early stopping: $\text{TV}(X_0, X_{t_0})$, the error from score estimation: $\mathbb{E}[\text{TV}(Y_{T-t_0}, \bar{Y}_{T-t_0})]$, and the error from the initialization of backward process: $\text{TV}(X_T, \mathcal{N}(0, T I_d))$.

Controlling the error from early stopping. Using standard Fourier analysis and tail bound for Sub-Gaussian distribution, the error $\text{TV}(X_0, X_{t_0})$ can be controlled by the following theorem.

Theorem 4.3 (See also Theorem E.1). *Under Assumption 3.1 and Assumption 3.4, let $t_0 = n^{-\frac{2}{2\beta+d}}$ and $p_{t_0} = p_0 * \Phi_{t_0}$, where Φ_t is the density of Gaussian distribution $\mathcal{N}(0, t I_d)$ and $*$ denote the convolution operator, then there exists a constant C that depends on p_0, β, L and dimension d such that*

$$\text{TV}(p_0, p_{t_0}) \leq C \operatorname{polylog}(n) n^{-\frac{\beta}{2\beta+d}}.$$

The proof can be found in Appendix E.1. This theorem shows that for the density p_0 belonging to the Sobolev class of order $\beta \leq 2$ (this is the only part of the argument requiring $\beta \leq 2$), by setting the early stopping time t_0 exactly the same as that in Corollary 3.7, the TV distance between the true data distribution $X_0 \sim p_0$ and the perturbed distribution $X_0 + B_{t_0} \sim p_{t_0}$ will not exceed the minimax optimal rate of convergence $n^{-\frac{\beta}{2\beta+d}}$.

Girsanov's theorem for controlling the score error. The application of Girsanov's Theorem in analyzing the convergence of diffusion models has been explored in several studies, see Song et al. (2021a); Chen et al. (2023b); Oko et al. (2023). In our analysis, we replaced the unknown drift term of the process $(Y_t)_{t \in [0, T-t_0]}$ by our kernel score estimator $\hat{s}_t(x)$, resulting in the new process $(\bar{Y}_t)_{t \in [0, T-t_0]}$. By Pinsker's inequality and data processing inequality, $\text{TV}(Y_{T-t_0}, \bar{Y}_{T-t_0}) \lesssim \sqrt{D_{\text{KL}}(\mathbb{P}_Y \| \mathbb{P}_{\bar{Y}})}$, where $\mathbb{P}_Y$ and $\mathbb{P}_{\bar{Y}}$ are path measure for $(Y_t)_{t \in [0, T-t_0]}$ and $(\bar{Y}_t)_{t \in [0, T-t_0]}$, accordingly. Girsanov's theorem (detailed in Appendix E.2) shows that the KL divergence between the

path measures is bounded by the accumulated score error over time t, which can be bounded using Corollary 3.7, i.e.,

$$
\begin{aligned}
&\mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}\left(\mathbb{P}_{\boldsymbol{Y}} \| \mathbb{P}_{\bar{\boldsymbol{Y}}}\right)\right] \\
&\lesssim \int_{t_0}^{T} \mathbb{E}\left[\int_{x \in \mathbb{R}^d} \|\hat{s}_t(x) - \nabla \log p_t(x)\|^2 \, p_t(x) dx\right] dt \\
&\lesssim \operatorname{polylog}(n) n^{-\frac{2\beta}{2\beta+d}}.
\end{aligned}
$$

Error from Initialization. Similar to the Exponential convergence of the Ornstein–Ulhenbeck process (Bakry et al., 2014), we show that the Brownian diffusion process in (4) is close to Gaussian distribution in KL divergence for large enough time T:

$$
\mathrm{D}_{\mathrm{KL}}\left(\boldsymbol{X}_T \| \mathcal{N}(0, T\boldsymbol{I}_d)\right) \lesssim \frac{1}{T}.
$$

Consequently, by applying Pinsker's inequality, the initialization error $\mathrm{TV}\left(\boldsymbol{X}_T, \mathcal{N}(0, T\boldsymbol{I}_d)\right)$ has polynomial decay rate of $n^{-\frac{\beta}{2\beta+d}}$ for $T \geq n^{\frac{2\beta}{2\beta+d}}$.

5. Discussions

Discretization error. In the analysis of Section 3.4, we assume it is possible to solve the approximated backward SDE (10) in Algorithm 1. However, in practical scenarios, solving the backward SDE (10) directly is infeasible. Therefore, we need to apply numerical SDE solvers such as Euler-Maruyama discretization to approximate a solution.

Specifically, to discretize the SDE in (10), consider the grid points $0 = t_1 < \ldots < t_N = T - t_0$ and define the step size $\gamma_k = t_{k+1} - t_k$. We then consider the following process: Let $\hat{\boldsymbol{Y}}_0 \sim \mathcal{N}(0, T\boldsymbol{I}_d)$, and for t in each intervals $[t_k, t_{k+1}]$, define $\hat{\boldsymbol{Y}}_t$ via the following SDE:

$$
\mathrm{d}\hat{\boldsymbol{Y}}_t = \hat{s}_{T-t_k}(\hat{\boldsymbol{Y}}_{t_k})\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t, \text{ for } t \in [t_k, t_{k+1}]. \quad (22)
$$

This is equivalent to the following iterative process:

$$
\hat{\boldsymbol{Y}}_{t_{k+1}} = \hat{\boldsymbol{Y}}_{t_k} + \gamma_k \hat{s}_{T-t_k}(\hat{\boldsymbol{Y}}_{t_k}) + \sqrt{\gamma_k}\boldsymbol{z}_k
$$

for $k = 0, \ldots N - 1$, where $\boldsymbol{z}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$.

This approximation introduces an extra discretization error in the convergence analysis of the diffusion model in Theorem 3.8. According to Theorem 2.2 of Chen et al. (2023a), when the data distribution p_0 has finite second moment, which is satisfied under our Sub-Gaussian assumption, the discretization error can be bounded by

$$
d^2 \sum_{k=0}^{N-1} \frac{\gamma_{k+1}^2}{t_k^2}. \quad (23)
$$

Therefore, by taking a constant step size $t_k = t_0 + kh$, the discretization error in (23) scales as $h^2 \int_{t_0}^{T} \frac{1}{t^2} dt =$

$h^2\left(\frac{1}{t_0} - \frac{1}{T}\right)$. In our setting, t_0 and T are both polynomial functions of n, the discretization error becomes negligible if we set $h \asymp n^{-C}$ for a sufficiently large constant $C > 0$.

Since discretization is not the main focus of our paper, we refer the reader to Theorem 2.2 of Chen et al. (2023a) for a more detailed discussion.

Future works. In practice, instead of directly applying kernel density estimation, the unknown score function is often learned through empirical risk minimization of a certain class of neural networks using score matching. We conclude this paper with two possible avenues for further research. First, it is worth investigating whether a neural network-based score estimator can achieve a similar score estimation error upper bound as described in Theorem 3.5. Additionally, the estimation error of the diffusion model, as described in Theorem 3.8, has been proven for $\beta \leq 2$. Future research could explore extending this proof to cases where $\beta > 2$.

6. Conclusion

In this paper, we prove general upper bounds for the score estimation error for Gaussian mixture distributions, under a mild sub-Gaussian assumption for the unknown density, using a truncated version of the Kernel Density Estimator. Our score error bound has optimal dependence on the sample size and the variance component of the Gaussian mixture. By applying the Girsanov theorem and adopting early stopping, this implies sharp bounds on the sampling error in the diffusion model. As a consequence, for the Sobolev class of density functions, we demonstrate that the total variation sampling error of the diffusion model matches the minimax rate in nonparametric statistics (Stone, 1982; 1980; Tsybakov, 2009). This removes crucial assumptions in similar recent results on the minimax optimality of diffusion models, such as the density lower bound.

Acknowledgements

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Bakry, D., Gentil, I., Ledoux, M., et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Benton, J., De Bortoli, V., Doucet, A., and Deligiannidis, G. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

Berens, S. Conditional Rényi entropy. Master's thesis, Universiteit Leiden, 2013. URL https://www.universiteitleiden.nl/binaries/content/assets/science/mi/scripties/masterberens.pdf. Thesis (M.Sc.) – Universiteit Leiden.

Block, A., Mroueh, Y., and Rakhlin, A. Generative modeling with denoising auto-encoders and langevin sampling, 2022.

Bobkov, S. G. Moments of the scores. *IEEE Transactions on Information Theory*, 65(9):5294–5301, 2019.

Chen, H., Lee, H., and Lu, J. 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.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023b.

De Bortoli, V. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

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

El Alaoui, A., Montanari, A., and Sellke, M. Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 323–334. IEEE, 2022.

Folland, G. B. Higher-order derivatives and taylor's formula in several variables. *Preprint*, pp. 1–4, 2005.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Holley, R. and Stroock, D. Logarithmic Sobolev inequalities and stochastic Ising models. *J. Stat. Phys.*, 46(5/6), 1987.

Hua, C., Luan, S., Xu, M., Ying, Z., Fu, J., Ermon, S., and Precup, D. MUDiff: Unified diffusion for complete molecule generation. In *The Second Learning on Graphs Conference*, 2023.

Huang, R., Zhao, Z., Liu, H., Liu, J., Cui, C., and Ren, Y. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 2595–2605, 2022.

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Indritz, J. An inequality for hermite polynomials. *Proceedings of the American Mathematical Society*, 12(6): 981–983, 1961.

Jeong, M., Kim, H., Cheon, S. J., Choi, B. J., and Kim, N. S. Diff-tts: A denoising diffusion model for text-to-speech. In *Interspeech*, 2021. URL https://api.semanticscholar.org/CorpusID:233025015.

Koehler, F., Heckett, A., and Risteski, A. Statistical efficiency of score matching: The view from isoperimetry. In *The Eleventh International Conference on Learning Representations*, 2022.

Le Gall, J.-F. *Brownian motion, martingales, and stochastic calculus*. Springer, 2016.

Lee, H., Lu, J., and Tan, Y. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35: 22870–22882, 2022.

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

Manica, M., Born, J., Cadow, J., Christofidellis, D., Dave, A., Clarke, D., Teukam, Y. G. N., Giannone, G., Hoffman, S. C., Buchan, M., et al. Accelerating material design with the generative toolkit for scientific discovery. *npj Computational Materials*, 9(1):69, 2023.

Montanari, A. Sampling, diffusions, and stochastic localization. *arXiv preprint arXiv:2305.10690*, 2023.

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., and Chen, M. GLIDE: Towards photorealistic image generation and

editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 16784–16804. PMLR, 17–23 Jul 2022.

Niles-Weed, J. and Berthet, Q. Minimax estimation of smooth densities in wasserstein distance. *The Annals of Statistics*, 50(3):1519–1540, 2022.

Oko, K., Akiyama, S., and Suzuki, T. Diffusion models are minimax optimal distribution estimators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 26517–26582, 2023.

Pidstrigach, J. Score-based generative models detect manifolds. *Advances in Neural Information Processing Systems*, 35:35852–35865, 2022.

Pinaya, W. H., Tudosiu, P.-D., Dafflon, J., Da Costa, P. F., Fernandez, V., Nachev, P., Ourselin, S., and Cardoso, M. J. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pp. 117–126. Springer, 2022.

Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., and Kudinov, M. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pp. 8599–8608. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rényi, A. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pp. 547–562. University of California Press, 1961.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34: 1415–1428, 2021a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021c. URL https://openreview.net/forum?id=PxTIG12RRHS.

Stone, C. J. Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pp. 1348–1360, 1980.

Stone, C. J. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pp. 1040–1053, 1982.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer, 2009.

Van Erven, T. and Harremos, P. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Wibisono, A. and Yang, K. Y. Convergence in kl divergence of the inexact langevin algorithm with application to score-based generative models. *arXiv preprint arXiv:2211.01512*, 2022.

Wibisono, A., Wu, Y., and Yang, K. Y. Optimal score estimation via empirical bayes smoothing. *arXiv preprint arXiv:2402.07747*, 2024.

Xu, M., Powers, A. S., Dror, R. O., Ermon, S., and Leskovec, J. Geometric latent diffusion models for 3d molecule generation. In *International Conference on Machine Learning*, pp. 38592–38610. PMLR, 2023.

Zhang, C.-H. Empirical bayes and compound estimation of normal means. *Statistica Sinica*, 7(1):181–193, 1997.

A. Notation

Here is a more detailed summary of notations. $p_0 : \mathbb{R}^d \to \mathbb{R}$ denotes the density function of true data distribution. For any time t, we denote the probability distribution $p_t(x) := p_0 * \phi_t(x) = \int p_0(y)\phi_t(x - y)dy$, where $*$ represents the convolution operator, and ϕ_t is the density function of Gaussian distribution $\mathcal{N}(0, t\mathbf{I}_d)$. Let $s_t(x) := \nabla \log p_t(x)$ denote the (Stein) score function corresponding to the probability distribution $p_t(x)$ at time t.

The operations $\vee$ is defined as $a \vee b := \max\{a, b\}$. For any functions f and g, $f(n) \asymp g(n)$ implies that there exist two positive constants C_1 and C_2 such that $C_1 g(n) \leq f(n) \leq C_2 g(n)$ for all n, $f(n) = O(g(n))$ indicates that there exists a positive real number M and an integer n_0 such that $|f(n)| \leq Mg(n)$ for all $n \geq n_0$, and $f(n) = \Omega(g(n))$ implies there exists a positive constant c and an integer n_0 such that $f(n) \geq c \cdot g(n)$ for all $n \geq n_0$. The symbols $\lesssim$ and $\gtrsim$ denotes the corresponding inequality up to a constant C. In the notations used here and throughout this paper, the implicit constants (such as C_1, C_2, M and C here) may depend on p_0 but not on t and n, unless otherwise indicated. Throughout the proofs, the constant $C, C_1, C_2, \ldots$ are used to denote positive constants and their values may change from one line to another, but are independent of everything else. We define $\mathrm{poly}(f(n)) := f(n)^C$ for some positive constant C. For a set $\mathcal{A} \subset \mathbb{R}^d$, $|\mathcal{A}|$ denotes its Lebesgue measure.

For a square-integrable function $f : \mathbb{R}^d \to \mathbb{R}$, denote $\|f\|_{L^2} := \int_x f^2(x)dx$ and $\|f\|_\infty := \sup_x |f(x)|$. For a vector $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, $\|x\|_2 := \sqrt{\sum_{i=1}^d x_i^2}$ and $\|x\|_1 := \sum_{i=1}^n |x_i|$. The ℓ_∞ norm of a vector $x \in \mathbb{R}^d$, is defined as $\|x\|_\infty = \max(|x_1|, |x_2|, \ldots, |x_d|)$. The total variation (TV) distance between to probability distribution p and q is defined as $\mathrm{TV}(p, q) := \frac{1}{2} \int |p(x) - q(x)| \, dx$.

B. Numerical Results

For numerical experiments using real data, we firstly demonstrate the impracticality of the prior uniform-in-time assumption in a real-world example by using 28×28 MNIST dataset. In this experiment, we use U-Net shaped score-net (Ronneberger et al., 2015) as our estimator. Note that the finding in Figure 1(b) matches the conclusion in (Oko et al., 2023) regarding the convergence rate for neural net estimators. Actually, the decay rate we showed in Figure 1(b) follows the results we discussed in Theorem 3.5, i.e., the slope approximately matches the theoretical result where the index of t is $-\frac{3}{2}$.

The baselines include: (1) To represent the training procedure of the reverse process, we used score-based U-Net architecture (Song et al., 2021c) to learn the score function, and (2) we apply a specific forward process, i.e., the Brownian diffusion process, described in Section 2, Algorithm 1. Our training objective $\mathbf{s}_\theta(\mathbf{x}, t)$ is a continuous weighted combination of Fisher
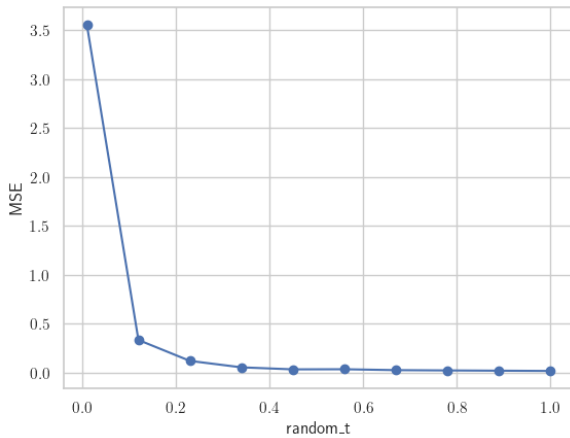
(a) 1-dimensional-MSE Convergence Rate
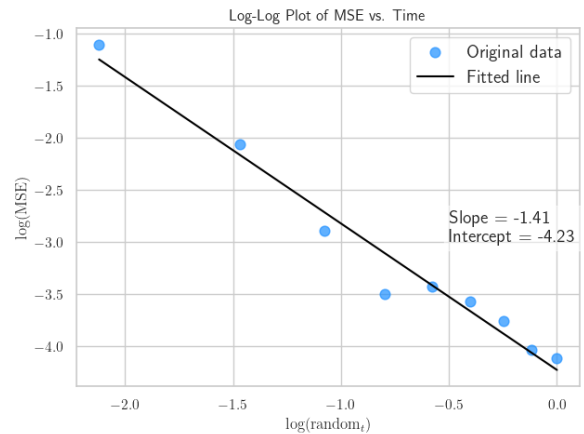
(b) Log-Log 1-dimensional-MSE Convergence Rate V.S. Time

Figure 1. Convergence Rates

divergences, given by

$$\mathbb{E}_{t \in \mathcal{U}(0,T)} \mathbb{E}_{p_t(\mathbf{x})} \left[\|\nabla_\mathbf{x} \log p_t(\mathbf{x}) - \mathbf{s}_\theta(\mathbf{x}, t)\|_2^2 \right]$$

where $\mathcal{U}(0,T)$ denotes a uniform distribution over the time interval $[T_1, 1]$, where $T_1 > 0$ is any constant. After training a convolutional neural network with a U-shaped architecture (known as U-Net), and applying the results to linear regression as illustrated in Figure 1(b), we observe that the slope closely approximates the main result of the mean square error as presented in Theorem 3.5, particularly when substituting $d = 1$ for relatively small values of t. Although the U-Net score estimator differs from the kernel estimator in the proof of Theorem 3.5, the $-\frac{3}{2}$ slope is the minimax rate that does not depend on the choice of the score estimator.

C. Construction and Properties of the Kernel Score Estimator

In this section, we will study the properties of the kernel score estimator we defined in (20), i.e.,

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

C.1. Construction of Kernel Function

In this section we construct kernel with order $\ell \geq 1$ that are bounded, compactly supported in $[-1, 1]^d$.

Definition C.1. (Tsybakov, 2009) Let $\ell \geq 1$ be an integer. A function $K_d : \mathbb{R}^d \to \mathbb{R}$ is a *kernel of order* ℓ if for any multi-index $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ with $|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_d \leq \ell$, the mappings

$$x \mapsto x^\alpha K(x)$$

are integrable and satisfy the following conditions:

$$\int_{\mathbb{R}^d} K(x)dx = 1,$$

$$\int_{\mathbb{R}^d} x^\alpha K(x)dx = 0, \quad \text{for all } 1 \leq |\alpha| \leq \ell - 1,$$

there exists an α with $|\alpha| = \ell$ such that $\int_{\mathbb{R}^d} x^\alpha K(x)dx \neq 0.$

Lemma C.2. *For any positive integer ℓ, there exists kernel of order ℓ supported on $[-1, 1]^d$, denoted as $K_d(x) := \prod_{i=1}^d K(x_i)$ such that for any $1 \leq i \leq d$ and $\alpha \leq \ell$,*

$$\int_{[-1,1]^d} K_d(x)dx = 1, \quad \|K_d\|_{L^2} = O(\ell^{\frac{3d}{2}}), \quad \|K_d\|_\infty = O(\ell^{\frac{5d}{2}}),$$

$$\int_{[-1,1]} |K(x_i)x_i^\alpha| \, dx_i = O\left(\ell^{\frac{5}{2}}/\alpha\right), \quad \int_{[-1,1]} |K'(x_i)x_i^\alpha| dx = O(\ell^{\frac{5}{2}}/\alpha)$$

and

$$\left\| \|\nabla K_d(\cdot)\|_2 \right\|_{L^2} = O(\sqrt{d}\ell^{\frac{3d}{2}}), \quad \|(\nabla K_d(\cdot))_i\|_\infty = O(\ell^{\frac{5d}{2}}).$$

Proof. We first consider the $d = 1$ case. We can construct such kernel using Legendre polynomials: Denote by $\{P_n(x)\}_{n=1}^\infty$ the Legendre polynomials. Define K by setting $K(-\infty) = 0$ and

$$K'(x) = \sum_{i=0}^{\ell+1} a_i P_i(x),$$

where $a_i \neq 0$ only if i is odd. Since P_i is an odd function for odd i, by this construction we automatically have $K(\pm 1) = 0$. Then from the orthogonality properties of Legendre polynomials

$$a_i = \frac{2i+1}{2} \int_{-1}^1 K'(x)P_i(x)dx.$$

13

To ensure that $K(x)$ is kernel of order ℓ, we need

$$\int_{-1}^{1} K'(x)dx = 0;$$

$$\int_{-1}^{1} K'(x)x dx = -\int_{-1}^{1} K(x)dx = -1;$$

$$\int_{-1}^{1} K'(x)x^k dx = -k\int_{-1}^{1} K(x)x^{k-1}dx = 0, \quad k = 2, \ldots, \ell + 1.$$

Note that these are $\ell + 2$ linear constraints, which uniquely determine $\{a_0, \ldots, a_{\ell+1}\}$. (For this we need that the matrix A, defined as $A := \{\int_{-1}^{1} x^i P_j(x)dx\}_{i,j=0,\ldots,\ell+1}$ which represents the coefficients of the system of linear equations, is invertible. This is true because $BA = \{\int_{-1}^{1} P_i(x)P_j(x)dx\}_{i,j=0,\ldots,\ell+1}$ for some matrix B.) Therefore,

$$a_i = -\frac{2i+1}{2}[x^i]P_i(x) = -\frac{2i+1}{2}2^i i\binom{\frac{i-1}{2}}{i},$$

where $[x^i]P_i(x)$ denotes the coefficients for x^i in the polynomials $P_i(x)$. Therefore,

$$|a_i| = (2i+1)\frac{((i)!!)^2}{i!}$$

$$\lesssim (2i+1)\frac{2i\left(\frac{i}{e}\right)^i}{\sqrt{2\pi i}\left(\frac{i}{e}\right)^i} \quad \left(i!! \sim \sqrt{2n}\left(\frac{i}{e}\right)^{i/2} \text{ and } i! \sim \sqrt{2\pi n}\left(\frac{i}{e}\right)^i\right)$$

$$\lesssim (2i+1)\sqrt{i}.$$

Therefore,

$$\|K'\|_{L^2}^2 = \int K'^2(x)dx = \sum_{i=0}^{\ell+1}\frac{2}{2i+1}a_i^2 \lesssim \sum_{i=0}^{\ell+1}(2i+1)i = O(\ell^3),$$

$$\|K'\|_\infty \le \frac{\ell+2}{2}\max|a_i| = O(\ell^{\frac{5}{2}}).$$

Last we consider $\int|K'(x)x^\alpha|dx$.

$$\int_{[-1,1]}|K'(x)x^\alpha|dx \le \|K'\|_\infty\int_{[-1,1]}|x^\alpha|dx \le O(\ell^{5/2})\frac{2}{\alpha+1} = O(\ell^{5/2}/\alpha).$$

Next we consider the bounds for kernel $K(x) = K(-1) + \int_{-1}^{x}K'(t)dt = \int_{-1}^{x}K'(t)dt$.

$$\|K\|_{L^2}^2 = \int K^2(x)dx$$

$$= \int\left(\int_{-1}^{x}K'(t)dt\right)^2 dx$$

$$\le \int\int_{-1}^{x}K'(t)^2 dt dx$$

$$\le 2\|K'\|^2$$

$$= O(\ell^3),$$

$$\|K\|_\infty = \sup_{x\in[-1,1]}\left|\int_{-1}^{x}K'(t)dt\right| \le 2\|K'\|_\infty = O(\ell^{\frac{5}{2}}).$$

Finally,

$$\int_{[-1,1]}|K(x_i)x_i^\alpha|\,dx_i \le \|K\|_\infty\int_{[-1,1]}|x^\alpha|\,dx = O(\ell^{5/2}/\alpha).$$

14

In general, for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, let

$$K_d(x) := \prod_{j=1}^{d} K(x_j), \quad (\nabla K_d(x))_j = K'(x_j) \prod_{i \neq j} K(x_i).$$

Using the above result in 1-dimension, we can directly find that

$$\|K_d\|_{L^2} = O(\ell^{\frac{3d}{2}}), \quad \|K_d\|_\infty = O(\ell^{\frac{5d}{2}})$$

and

$$\int_{[-1,1]^d} |K_d(x) x_j^\ell| \, dx \leq \|K\|_\infty^{d-1} \int_{[-1,1]} |K(x_j) x_j^\ell| \, dx_j = O\left(\ell^{\frac{5d-2}{2}}\right), \quad \text{for any } 1 \leq j \leq d.$$

Besides,

$$\begin{aligned}
\|\|\nabla K_d(\cdot)\|_2\|_{L^2}^2 &= \int \sum_{i=1}^{d} (\nabla K_d(x))_i^2 dx \\
&= \sum_{i=1}^{d} \int \left(K'(x_i) \prod_{j \neq i} K(x_j)\right)^2 dx \\
&\leq d\|K'\|_{L^2}^2 \|K\|_{L^2}^{2(d-1)} \\
&= O(d\ell^{3d}).
\end{aligned}$$

For any fixed i,

$$\|(\nabla K_d(\cdot))_i\|_\infty \leq \|K'\|_\infty \|K\|_\infty^{(d-1)} = O(\ell^{\frac{5d}{2}}).$$

$\square$

C.2. Mean Squared Error Analysis

Intuitively, if both the gradient estimator $\nabla \hat{p}_t(x)$ and the density estimator $\nabla \hat{p}_t(x)$ are close to their respective true values $\nabla p_t(x)$ and $p_t(x)$, the kernel score estimator $\hat{s}_t(x)$ defined in (20) should be also close to the true score function $s_t(x) = \frac{\nabla p_t(x)}{p_t(x)}$. In this section, we study the mean squared errors (MSE) of $\hat{p}_t(x)$ and $\nabla \hat{p}_t(x)$, and it will be shown in the proof of Theorem 3.5 that these mean squared errors are the dominant part of the score error.

The MSE for $\hat{p}_t(x)$ and $\nabla \hat{p}_t(x)$ are defined as follows:

$$\begin{aligned}
\mathrm{MSE}(\hat{p}_t(x)) &:= \mathbb{E}\left[|\hat{p}_t(x) - p_t(x)|^2\right], \\
\mathrm{MSE}(\nabla \hat{p}_t(x)) &:= \mathbb{E}\left[\|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2\right].
\end{aligned}$$

We follow the common framework of bias-variance trade-off Tsybakov (2009) with slight modifications, which allows us to obtain a sharper bound that depends on the location x.

Proposition C.3 (**MSE for estimator $\hat{p}_t$**). *If we take the bandwidth $h = \frac{\sqrt{t}}{D_n}$ for some constant D_n that depends on n, then there exists a constant C_2 that depend on p_0 and d, such that the MSE for $\hat{p}_t(x)$ satisfies*

$$\mathrm{MSE}(\hat{p}_t(x)) \leq C_2 \left(\frac{D_n^d \ell^{3d}}{n t^{\frac{d}{2}}} p_t^*(x) + \left(\frac{D_n}{de}\right)^{-2\ell} \ell^{-\ell + 5d - \frac{1}{2}}\right),$$

where ℓ is the order of kernel defined in Theorem C.2, $p_t^*(x) := \sup_{\|\lambda\|_\infty < h} p_t(x + \lambda)$. In particular, if we take $\ell = \Omega(\log n)$ and $D_n = C^* \sqrt{\log n}$, for some constant $C^* > de$, we have

$$\mathrm{MSE}(\hat{p}_t(x)) \leq C_2 \, \mathrm{polylog}(n) \left(\frac{p_t^*(x)}{n t^{\frac{d}{2}}} + (\log n)^{-\log n}\right).$$

15

Proof. By the Bias-Variance trade-off, we have

$$\text{MSE}(\hat{p}_t(x)) = \sigma^2(x) + b(x)^2,$$

where

$$\sigma^2(x) = \mathbb{E}\left[|\hat{p}_t(x) - \mathbb{E}[\hat{p}_t(x)]|^2\right], \quad b(x) = \mathbb{E}\left[\hat{p}_t(x)\right] - p_t(x).$$

We first analyze the variance term $\sigma^2(x)$. Denote

$$\eta_i(x) = K_d\left(\frac{x - X_i^t}{h}\right) - \mathbb{E}\left[K_d\left(\frac{x - X_i^t}{h}\right)\right],$$

where $X_i^t \sim p_t$ be the i-th observation at specific time point t. Then we have

$$\begin{aligned}
\mathbb{E}\left[\eta_i^2(x)\right] &\leq \mathbb{E}\left[K_d^2\left(\frac{x - X_i^t}{h}\right)\right] \\
&= \int_{\mathbb{R}^d} K_d^2\left(\frac{x - y}{h}\right) p_t(y)\, dy \\
&= h^d \int_{[-1,1]^d} K_d^2(z) p_t(x - hz) dz \\
&\leq h^d\, p_t^*(x) \int_{[-1,1]^d} K_d^2(z) dz,
\end{aligned}$$

where

$$p_t^*(x) := \sup_{\|\lambda\|_\infty \leq h} p_t(x + \lambda).$$

Therefore, since $\eta_i(x)$ are independent and $\mathbb{E}\eta_i(x) = 0$,

$$\sigma^2(x) = \mathbb{E}\left[\left(\frac{1}{nh^d}\sum_{i=1}^n \eta_i(x)\right)^2\right] = \frac{1}{nh^{2d}}\mathbb{E}\left[\eta_1^2(x)\right] \leq \frac{p_t^*(x)}{nh^d}\int K_d^2(z)dz \lesssim \frac{\ell^{3d}\, p_t^*(x)}{nh^d},$$

where in the last inequality we use the result in Lemma C.2. Taking $h = \frac{\sqrt{t}}{D_n}$, we get

$$\sigma^2(x) \lesssim \frac{\ell^{3d}\, D_n^d\, p_t^*(x)}{nt^{d/2}}. \tag{24}$$

Next we consider the bias $b(x)$:

$$\begin{aligned}
b(x) &= \frac{1}{nh^d}\mathbb{E}\left[\sum_{i=1}^n K_d\left(\frac{x - X_i^t}{h}\right)\right] - p_t(x) \\
&= \frac{1}{h^d}\int_{\mathbb{R}^d} K_d\left(\frac{x - z}{h}\right) p_t(z)dz - p_t(x) \\
&= \int_{\mathbb{R}^d} K_d(-u)\left[p_t(x + uh) - p_t(x)\right]du.
\end{aligned}$$

Using the Multivariate version of Taylor's Theorem (Folland, 2005), we have

$$p_t(x + uh) = p_t(x) + \sum_{|\alpha|=1} \frac{D^\alpha p_t(x)}{\alpha!}(uh)^\alpha + \cdots + \sum_{|\alpha|=\ell} \frac{D^\alpha p_t(x + \tau uh)}{\alpha!}(uh)^\alpha$$

for some $\tau \in [0, 1]$, where we use the multi-index notation,

$$|\alpha| = \alpha_1 + \cdots + \alpha_d = \ell, \quad u^\alpha = \prod_{i=1}^d u_i^{\alpha_i}, \quad \alpha! = \prod_{i=1}^d \alpha_i!.$$

Using the definition of kernel of order ℓ, $\int K_d(-u)du = 1$ and $\int u^\alpha K_d(u)du = 0$ for $|\alpha| \le \beta - 1$, we have

$$
\begin{aligned}
|b(x)| &= \left| \sum_{|\alpha|=\ell} \int_{\mathbb{R}^d} K_d(-u) \left(D^\alpha p_t(x + \tau u h) \right) \frac{(uh)^\alpha}{\alpha!} du \right| \\
&\le h^\ell \sum_{|\alpha|=\ell} \int_{\mathbb{R}^d} \left| K_d(-u) \left(D^\alpha p_t(x + \tau u h) \right) \frac{u^\alpha}{\alpha!} \right| du \\
&\le h^\ell \sum_{|\alpha|=\ell} \int_{\mathbb{R}^d} \left| K_d(-u) \frac{u^\alpha}{\alpha!} \right| du \left(\sup_x |D^\alpha p_t(x)| \right).
\end{aligned}
$$

By Proposition F.4, we have

$$
\sup_x |D^\alpha p_t(x)| \le C_1^d \|p_0\|_\infty \ell^{\frac{\ell}{2} + \frac{1}{4}} t^{-\ell/2}.
$$

Using the property of the kernel K_d in Lemma C.2,

$$
\begin{aligned}
\sum_{|\alpha|=\ell} \int_{\mathbb{R}^d} \left| K_d(-u) \frac{u^\alpha}{\alpha!} \right| du &= \sum_{|\alpha|=\ell} \prod_{i=1}^d \frac{1}{\alpha_i!} \int_{\mathbb{R}} |K(-u_i) u_i^{\alpha_i}| du_i \\
&\lesssim \sum_{|\alpha|=\ell} \prod_{i=1}^d \frac{1}{\alpha_i!} \frac{\ell^{5/2}}{\alpha_i} \\
&\le \ell^{5d/2} \sum_{|\alpha|=\ell} \frac{1}{\alpha!} \\
&= \ell^{5d/2} \frac{d^\ell}{\ell!}.
\end{aligned}
$$

Here in the last equality, we use the following multinomial theorem with $x_i = 1$ for all $i = 1, \ldots, d$:

$$
\left(\sum_{i=1}^d x_i \right)^\ell = \sum_{|\alpha|=\ell} \frac{\ell!}{\alpha!} \prod_{i=1}^d x_i^{\alpha_i}.
$$

By Stirling's approximation,

$$
\ell! > \sqrt{2\pi\ell}\, \ell^\ell e^{-\ell} e^{\frac{1}{12\ell+1}} > \ell^{\ell+\frac{1}{2}} e^{-\ell}.
$$

Thus,

$$
|b(x)| \lesssim C_1^d \|p_0\|_\infty h^\ell (de)^\ell \ell^{-\frac{\ell}{2} + \frac{5d}{2} - \frac{1}{4}} t^{-\frac{\ell}{2}}.
$$

Take the bandwidth as $h = \frac{\sqrt{t}}{D_n}$, we have:

$$
|b(x)| \lesssim C_1^d \|p_0\|_\infty \left(\frac{D_n}{de} \right)^{-\ell} \ell^{-\frac{\ell}{2} + \frac{5d}{2} - \frac{1}{4}}. \tag{25}
$$

Then the results follow by combining (24) and (25). $\qquad\square$

Proposition C.4 (**MSE for estimator** $\nabla \hat{p}_t$). *If we take the bandwidth $h = \frac{\sqrt{t}}{D_n}$ for some constant D_n that depends on n, then there exists a constant C_3, such that the MSE for $\nabla \hat{p}_t(x)$ satisfies*

$$
\mathrm{MSE}(\nabla \hat{p}_t(x)) \le C_3 \left(\frac{\ell^{\frac{3d}{2}} D_n^{d+2}}{n t^{\frac{d+2}{2}}} p_t^*(x) + \left(\frac{D_n}{de} \right)^{-2(\ell-1)} \ell^{-\ell+5d-\frac{1}{2}} t^{-1} \right),
$$

where ℓ is the order of kernel defined in Lemma C.2, $p_t^*(x) := \sup_{\|\lambda\|_\infty < h} p_t(x + \lambda)$. In particular, if we take $\ell = \log n$ and $D_n = C\sqrt{\log n}$ for some constant $C > de$, we have

$$
\mathrm{MSE}(\nabla \hat{p}_t(x)) \le C_3 \,\mathrm{polylog}(n) \left(\frac{p_t^*(x)}{n t^{\frac{d+2}{2}}} + (\log n)^{-\log n} t^{-1} \right).
$$

Proof. We follow the same procedure above in Proposition C.3. The Bias-Variance trade-off reads

$$\text{MSE}(\nabla \hat{p}_t(x)) = \sigma^2(x) + \|b(x)\|^2,$$

where

$$\sigma^2(x) = \mathbb{E}\left[\|\nabla \hat{p}_t(x) - \mathbb{E}\left[\nabla \hat{p}_t(x)\right]\|^2\right], \quad b(x) = \mathbb{E}\left[\nabla \hat{p}_t(x)\right] - \nabla p_t(x).$$

By definition, the kernel density estimation for the gradient $\nabla p_t(x)$ is:

$$\nabla \hat{p}_t(x) = \frac{1}{nh^{d+1}} \sum_{i=1}^{n} \nabla K_d \left(\frac{x - X_i^t}{h}\right),$$

where X_i^t stands for the $i-th$ sample at time t. We first analyze the variance term $\sigma^2(x)$. Denote

$$\eta_i(x) := \nabla K_d \left(\frac{x - X_i^t}{h}\right) - \mathbb{E}\left[\nabla K_d \left(\frac{x - X_i^t}{h}\right)\right],$$

Then, we get

$$
\begin{aligned}
\mathbb{E}\left[\|\eta_i(x)\|^2\right] &\leq \mathbb{E}\left[\left\|\nabla K_d \left(\frac{x - X_i^t}{h}\right)\right\|^2\right] \\
&= \int_{\mathbb{R}^d} \left\|\nabla K_d \left(\frac{x - y}{h}\right)\right\|^2 p_t(y) \, dy \\
&= h^d \int_{[-1,1]^d} \|\nabla K_d(u)\|^2 p_t(x - uh) du \\
&\leq h^d \sup_{\|\lambda\|_\infty < h} p_t(x + \lambda) \int_{[-1,1]^d} \|\nabla K_d(u)\|^2 \, du \\
&\lesssim h^d p_t^*(x) \, d\ell^{\frac{3d}{2}},
\end{aligned}
$$

where in the last inequality we use the result in Lemma C.2. Therefore, since $\eta_i(x)$ are independent and $\mathbb{E}\eta_i(x) = 0$, we have

$$\sigma^2(x) = \mathbb{E}\left[\left\|\frac{1}{nh^{d+1}} \sum_{i=1}^{n} \eta_i(x)\right\|^2\right] = \frac{1}{nh^{2d+2}} \mathbb{E}\left[\eta_1^2(x)\right] \lesssim \frac{d\ell^{\frac{3d}{2}}}{nh^{d+2}} p_t^*(x).$$

Then, if we take $h = \frac{\sqrt{t}}{D_n}$, then

$$\sigma^2(x) \lesssim \frac{d\ell^{\frac{3d}{2}} D_n^{d+2}}{nt^{\frac{d+2}{2}}} p_t^*(x). \tag{26}$$

Next, we consider the bias term $b(x)$,

$$
\begin{aligned}
b(x) &= \frac{1}{nh^{d+1}} \mathbb{E}\left[\sum_{i=1}^{n} \nabla K_d \left(\frac{x - X^i}{h}\right)\right] - \nabla p_t(x) \\
&= \frac{1}{h^{d+1}} \int_{\mathbb{R}^d} \nabla K_d \left(\frac{x - y}{h}\right) p(y) dy - \nabla p_t(x) \\
&= \frac{1}{h} \int_{\mathbb{R}^d} \nabla K_d (-u) \, p(x + uh) du - \nabla p_t(x).
\end{aligned}
$$

We consider the first coordinate of $b(x)$:

$$b_1(x) := \frac{1}{h} \int K'(-u_1) \prod_{j=2}^{d} K(-u_j) p_t(x_1 + u_1 h, \ldots, x_d + u_d h) du_1 \ldots du_d - \frac{\partial}{\partial x_1} p_t(x)$$

18

By Multivariate version of Taylor's Theorem Folland (2005), we have

$$p_t(x + uh) = \sum_{|\alpha| \le \ell-1} \frac{D^\alpha p(x)}{\alpha!}(uh)^\alpha + \sum_{|\alpha|=\ell} \frac{D^\alpha p(x + \tau uh)}{\alpha!}(uh)^\alpha, \quad \text{for some } \tau \in [0,1].$$

Using $K(1) = K(-1) = 0$, $\int K_d(-u)du = 1$, $\int u^\ell K_d(u)du = 0$ for $\ell \le \beta - 1$ and integration by parts, we have the first coordinate of $b(x)$:

$$|b_1(x)| = \frac{1}{h}\left|\int \sum_{|\alpha|=\ell} (\nabla K_d(-u))_1 D^\alpha p(x + \tau uh)\frac{(uh)^\alpha}{\alpha!}du\right|$$

$$\le h^{\ell-1} \sum_{|\alpha|=\ell} \int_{\mathbb{R}^d}\left|(\nabla K_d(-u))_1 (D^\alpha p_t(x + \tau uh))\frac{u^\alpha}{\alpha!}\right|du$$

$$\le h^{\ell-1} \sum_{|\alpha|=\ell} \int_{\mathbb{R}^d}\left|(\nabla K_d(-u))_1 \frac{u^\alpha}{\alpha!}\right|du \left(\sup_x |D^\alpha p_t(x)|\right).$$

By Lemma F.4,

$$\sup_x |D^\alpha p_t(x)| \le C_1^d \|p_0\|_\infty \ell^{\frac{\ell}{2}+\frac{1}{4}} t^{-\ell/2}.$$

Using the property of the kernel ∇K_d in Lemma C.2 and multinomial theorem, we have

$$\sum_{|\alpha|=\ell} \int_{\mathbb{R}^d}\left|(\nabla K_d(-u))_1 \frac{u^\alpha}{\alpha!}\right|du = \sum_{|\alpha|=\ell}\left[\left(\frac{1}{\alpha_1!}\int_{\mathbb{R}}|K'(-u_1)u_1^{\alpha_1}|\,du_1\right)\prod_{i=2}^d \frac{1}{\alpha_i!}\int_{\mathbb{R}}|K(-u_i)u_i^{\alpha_i}|\,du_i\right]$$

$$\lesssim \sum_{|\alpha|=\ell}\prod_{i=1}^d \frac{1}{\alpha_i!}\frac{\ell^{5/2}}{\alpha_i}$$

$$\le \ell^{5d/2} \sum_{|\alpha|=\ell} \frac{1}{\alpha!}$$

$$= \ell^{5d/2}\frac{d^\ell}{\ell!}.$$

By Stirling's approximation,

$$\ell! > \sqrt{2\pi\ell}\,\ell^\ell e^{-\ell} e^{\frac{1}{12\ell+1}} > \ell^{\ell+\frac{1}{2}}e^{-\ell}.$$

Thus,

$$|b_1(x)| \lesssim C_1^d \|p_0\|_\infty h^{\ell-1} (de)^\ell \ell^{-\frac{\ell}{2}+\frac{5d}{2}-\frac{1}{4}} t^{-\frac{\ell}{2}}.$$

Taking the bandwidth as $h = \frac{\sqrt{t}}{D_n}$, we have:

$$|b_1(x)| \lesssim C_1^d \|p_0\|_\infty de \left(\frac{D_n}{de}\right)^{-(\ell-1)} \ell^{-\frac{\ell}{2}+\frac{5d}{2}-\frac{1}{4}} t^{-\frac{1}{2}},$$

and then

$$\|b(x)\|^2 \le |b_1(x)|^2 \lesssim d^2 e C_1^{2d} \|p_0\|_\infty^2 \left(\frac{D_n}{de}\right)^{-2(\ell-1)} \ell^{-\ell+5d-\frac{1}{2}} t^{-1}. \tag{27}$$

Finally, The results follow by combining (26) and (27). □

D. Proof of Estimation Error for Score Function

D.1. Proof of Theorem 3.5

The proof of our main result begins with a discussion of the estimation of the target total error term. Using the kernel score estimator we constructed in Appendix C, i.e.,

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

the following proposition shows that the score error $\|\hat{s}_t(x) - s_t(x)\|^2$ can be decomposed into two components: the error of the gradient estimator $\nabla \hat{p}_t(x)$ and the error of the density estimator $\hat{p}_t(x)$. Therefore, it suffices to control the error of $\nabla \hat{p}_t(x)$ and $\hat{p}_t(x)$.

Lemma D.1. *For any $t > 0$, the score estimator defined in* (20) *satisfies*

$$\int_x \mathbb{E}\left[\|\hat{s}_t(x) - s_t(x)\|^2\right] p_t(x)dx$$
$$\leq 2\int_x \mathbb{E}\left[\frac{\|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2 + \|s_t(x)\|^2|\hat{p}_t(x) - p_t(x)|^2}{\hat{p}_t(x)^2}\mathbb{1}_{\{\hat{p}_t(x) > \rho_n\}}\right] p_t(x)dx$$
$$+ \int_x \mathbb{P}\left(\hat{p}_t(x) \leq \rho_n\right)\|s_t(x)\|^2 p_t(x)dx.$$

Proof. If x is such that $\hat{p}_t(x) > \rho_n$, we have the decomposition

$$\hat{s}_t(x) - s_t(x) = \nabla \log \hat{p}_t(x) - \nabla \log p_t(x)$$
$$= \frac{\nabla \hat{p}_t(x)}{\hat{p}_t(x)} - \frac{\nabla p_t(x)}{p_t(x)}$$
$$= \frac{\nabla \hat{p}_t(x)p_t(x) - \nabla p_t(x)p_t(x) + \nabla p_t(x)p_t(x) - \nabla p_t(x)\hat{p}_t(x)}{\hat{p}_t(x)p_t(x)}$$
$$= \frac{\nabla \hat{p}_t(x) - \nabla p_t(x) + s_t(x)(p_t(x) - \hat{p}_t(x))}{\hat{p}_t(x)}.$$

Therefore,

$$\int_x \mathbb{E}\left[\|\hat{s}_t(x) - s_t(x)\|^2 \mathbb{1}_{\hat{p}_t(x) > \rho_n}\right] p_t(x)dx$$
$$\leq 2\int_x \mathbb{E}\left[\frac{\|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2 + \|s_t(x)\|^2|\hat{p}_t(x) - p_t(x)|^2}{\hat{p}_t(x)^2}\mathbb{1}_{\hat{p}_t(x) > \rho_n}\right] p_t(x)dx.$$

On the other hand, if x is such that $\hat{p}_t(x) \leq \rho_n$, we have $\hat{s}_t(x) = 0$. Therefore,

$$\int_x \mathbb{E}\left[\|\hat{s}_t(x) - s_t(x)\|^2 \mathbb{1}_{\{\hat{p}_t(x) \leq \rho_n\}}\right] p_t(x)dx = \int_x \mathbb{P}\left(\hat{p}_t(x) \leq \rho_n\right)\|s_t(x)\|^2 p_t(x)dx.$$

$\square$

We further consider the score error in three cases, respectively. Recall that our score estimator is defined as

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

The key challenge is to control the error in low-density regions. We therefore consider the following three cases. Let $c > 4C_d + dT_2 + 2$ be a sufficiently large constant, where is a constant that only depends on the dimension d. In the first case $p_t(x) > \rho_n \log^c n$, the density is relatively high, so as we will see in the following proof, with high probability, $p_t(x)$ and $\hat{p}_t(x)$ will have the same order and as a result, $\hat{p}_t(x) > \rho_n$, which implies that the first term in Lemma D.1 will dominate. Whereas in case 2 we consider the lower density area $p_t(x) < \rho_n \log^{-c} n$, with high probability, $\hat{p}_t(x) < \rho_n$ and therefore, $\hat{s}_t(x) = 0$. Finally, in case 3 where $\rho_n \log^{-c} n < p_t(x) < \rho_n \log^c n$, we must consider both $\hat{s}_t(x) \neq 0$ and $\hat{s}_t(x) = 0$, since neither of them is guaranteed with high probability. In this case, we will apply the methods from both Case 1 and Case 2 to bound this term.

D.1.1. CASE 1: $p_t(x) > \rho_n \log^c n$

Lemma D.2. *For any time $n^{-T_1} < t < n^{T_2}$, and for any $0 < \varepsilon < 1$, taking $\rho_n = \frac{1}{n\, t^{d/2}}$, the score error on the set*

$$G_1 := \{x \colon p_t(x) > \rho_n \log^c n\}$$

can be upper bounded by

$$\mathbb{E}\left[\int_{G_1} \|\hat{s}_t(x) - s_t(x)\|^2 p_t(x)dx\right] \lesssim \operatorname{polylog}(n)\, n^{-1} t^{-\frac{d+2}{2}} (t^{\frac{d}{2}} + \sigma_0^d)\, g_1(n, \varepsilon),$$

where

$$g_1(n, \varepsilon) := \exp\left(\frac{1}{\varepsilon \log n}\right) (\varepsilon)^{-1} n^{C_1 \varepsilon},$$

where $C_1 = \frac{2+dT_2}{2}$.

Remark D.3. We will take $\varepsilon = \frac{\log \log n}{\log n}$ later in Section D.1.4 when Lemma D.2 is applied.

Proof of Lemma D.2. From Lemma D.1,

$$\int_{G_1} \mathbb{E}\left[\|\hat{s}_t(x) - s_t(x)\|^2\right] p_t(x)dx$$

$$\leq 2 \int_{G_1} \mathbb{E}\left[\frac{\|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2 + \|s_t(x)\|^2 |\hat{p}_t(x) - p_t(x)|^2}{\hat{p}_t(x)^2} \mathbb{1}_{\{\hat{p}_t(x) > \rho_n\}}\right] p_t(x)dx \tag{28}$$

$$+ \int_{G_1} \mathbb{P}\left(\hat{p}_t(x) \leq \rho_n\right) \|s_t(x)\|^2 p_t(x)dx \tag{29}$$

Denote

$$I_t(x) := \frac{\|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2 + \|s_t(x)\|^2 |\hat{p}_t(x) - p_t(x)|^2}{\hat{p}_t(x)^2}.$$

By Theorem F.6, define the random set

$$A := \left\{x \colon |\hat{p}_t(x) - p_t(x)| < C_4(\alpha) \operatorname{polylog}(n) \left(\sqrt{\frac{p_t^*(x)}{n t^{\frac{d}{2}}}} + \frac{1}{n t^{\frac{d}{2}}} + (\log n)^{-\frac{(\log n)}{2}}\right)\right\},$$

then $\{x \in A\}$ has probability greater than $1 - n^{-\alpha}$ for any $\alpha > 0$, where $p_t^*(x) := \sup_{\|\lambda\|_\infty < h} p_t(x + \lambda)$ and $C_4(\alpha)$ is some constant that only depends on p_0, d and α. Then we consider

$$I_t(x)\mathbb{1}_{\{\hat{p}(x) > \rho_n\}} = I_t(x)\mathbb{1}_{\{\hat{p}(x) > \rho_n\} \cap A} + I_t(x)\mathbb{1}_{\{\hat{p}(x) > \rho_n\} \cap A^c}. \tag{30}$$

Now we split the proof into several parts:

Part 1: The first term $I_t(x)\mathbb{1}_{\{\hat{p}(x) > \rho_n\} \cap A}$ **in** (30). Under the random set $\{x : \hat{p}(x) > \rho_n\} \cap A \cap G_1$, by Lemma D.4:

$$\hat{p}_t^2(x) \geq \left(p_t(x) - C_4(\alpha) \operatorname{polylog}(n) \left(\sqrt{\frac{p_t^*(x)}{n t^{\frac{d}{2}}}} + \frac{1}{n t^{\frac{d}{2}}} + (\log n)^{-\log n}\right)\right)^2 \geq \frac{1}{4} p_t^2(x).$$

Therefore, for any $x \in G_1$,

$$I_t(x)\mathbb{1}_{\{\hat{p}(x) > \rho_n\} \cap A} \leq \frac{\|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2 + \|s_t(x)\|^2 |\hat{p}_t(x) - p_t(x)|^2}{p_t(x)^2/4}.$$

Using Proposition C.3 and Proposition C.4:

$$\mathbb{E}\left[|\hat{p}_t(x) - p_t(x)|^2\right] \leq C_2 \operatorname{polylog}(n) \left(\frac{p_t^*(x)}{n t^{d/2}} + (\log n)^{-\log n}\right), \tag{31}$$

$$\mathbb{E}\left[\|\nabla\hat{p}_t(x) - \nabla p_t(x)\|^2\right] \le C_3 \operatorname{polylog}(n)\left(\frac{p_t^*(x)}{n\,t^{(d+2)/2}} + (\log n)^{-\log n}\,t^{-1}\right), \tag{32}$$

Therefore,

$$\int_{G_1} \mathbb{E}\left[I_t(x)\mathbb{1}_{\{\hat{p}(x) > \rho_n\} \cap A}\right] p_t(x)dx \lesssim \operatorname{polylog}(n)\left[\frac{1}{nt^{(d+2)/2}}\int_{G_1}\frac{p_t^*(x)}{p_t(x)}dx\right.$$
$$+ \frac{1}{nt^{d/2}}\int_{G_1}\frac{p_t^*(x)}{p_t(x)}\|s_t(x)\|^2 dx$$
$$\left.+ (\log n)^{-\log n}\int_{G_1}\left(\frac{t^{-1}}{p_t(x)} + \frac{\|s_t(x)\|^2}{p_t(x)}\right)dx\right]. \tag{33}$$

For the last term, we can use Lemma F.4, Lemma F.9 and the fact that $t > n^{-T_1}$, to obtain:

$$\int_{G_1}\left(\frac{t^{-1}}{p_t(x)} + \frac{\|s_t(x)\|^2}{p_t(x)}\right)dx \le |G_1|\left(t^{-1}\rho_n^{-1}\log^{-c}n + t^{-\frac{d}{2}}\rho_n^{-3}\log^{-3c}n\right) \lesssim \operatorname{poly}(n\log n).$$

Since $(\log n)^{-\log n} \ll \operatorname{poly}(n\log n)$ for sufficiently large n, the error for this term can be ignored. Using Lemma F.18, for any $0 < \varepsilon < 1$,

$$\frac{p_t^*(x)}{p_t^{1-\varepsilon}(x)} \le \frac{1}{\sqrt{1-\varepsilon}}\exp\left\{\frac{1-\varepsilon}{D_n^2\varepsilon}\right\},$$

we have

$$\int_{G_1}\frac{p_t^*(x)}{p_t(x)}dx = \int_{G_1}\frac{p_t^*(x)}{p_t^{1-\varepsilon}(x)}p_t^{-\varepsilon}(x)dx \le \frac{1}{\sqrt{1-\varepsilon}}\exp\left\{\frac{1-\varepsilon}{2D_n^2\varepsilon}\right\}(\log n)^{-c\varepsilon}\rho_n^{-\varepsilon}|G_1|,$$

and

$$\int_{G_1}\frac{p_t^*(x)}{p_t(x)}\|s_t(x)\|^2 dx \le \frac{1}{\sqrt{1-\varepsilon}}\exp\left\{\frac{1-\varepsilon}{D_n^2\varepsilon}\right\}(\log n)^{-c\varepsilon}\rho_n^{-\varepsilon}\int_{G_1}\|s_t(x)\|^2 dx.$$

Here, $|G_1|$ and $\int_{G_1}\|s_t(x)\|^2 dx$ can be upper bounded by Lemma F.9 and Lemma F.10, accordingly. Therefore, we have

$$\int_{G_1}\mathbb{E}\left[I_t(x)\mathbb{1}_{\{\hat{p}(x) > \rho_n\}\cap A}\right]p_t(x)dx$$
$$\lesssim \operatorname{polylog}(n)\frac{(t^{d/2} + \sigma_0^d)}{n\,t^{(d+2)/2}}\frac{1}{\sqrt{1-\varepsilon}}\exp\left\{\frac{1-\varepsilon}{D_n^2\varepsilon}\right\}(\log n)^{-c\varepsilon}\rho_n^{-\varepsilon}\left(1 + m(\log n)^{-\frac{c}{m}}\rho_n^{-\frac{1}{m}}\right)$$
$$\le \operatorname{polylog}(n)\frac{(t^{d/2} + \sigma_0^d)}{n\,t^{(d+2)/2}}\frac{1}{\sqrt{1-\varepsilon}}\exp\left\{\frac{1-\varepsilon}{D_n^2\varepsilon}\right\}\rho_n^{-\varepsilon}\left(1 + m\rho_n^{-\frac{1}{m}}\right)$$

Take $m \asymp \frac{1}{\varepsilon}$, and $D_n = C\sqrt{\log n}$, then $\frac{1}{\sqrt{1-\varepsilon}}\exp\left\{\frac{1-\varepsilon}{D_n^2\varepsilon}\right\} \lesssim \exp\left(\frac{1}{\varepsilon\log n}\right)$. Using $\rho_n = n^{-1}t^{-d/2} > n^{-1}n^{-dT_2/2}$, we then have

$$\rho_n^{-\varepsilon}\left(1 + m\rho_n^{-\frac{1}{m}}\right) = \rho_n^{-\varepsilon}\left(1 + \varepsilon^{-1}\rho_n^{-\varepsilon}\right)$$
$$\le \varepsilon^{-1}\rho_n^{-2\varepsilon}$$
$$\le \varepsilon^{-1}n^{C_1\varepsilon},$$

where $C_1 = \frac{2+dT_2}{2}$. Thus we can conclude that:

$$\int_{G_1}\mathbb{E}\left[I_t(x)\mathbb{1}_{\{\hat{p}(x) > \rho_n\}\cap A}\right]p_t(x)dx \lesssim \operatorname{polylog}(n)\,n^{-1}t^{-\frac{d+2}{2}}(t^{\frac{d}{2}} + \sigma_0^d)\exp\left(\frac{1}{\varepsilon\log n}\right)(\varepsilon)^{-1}n^{C_1\varepsilon}.$$

Part 2: The second term $I_t(x)\mathbb{1}_{\{\hat{p}(x) > \rho_n\}\cap A^c}$ **in** (30). The error bound in this part essentially uses the fact that A^c is small by concentration inequalities. Under the random set $\{x : \hat{p}(x) > \rho_n\} \cap A^c \cap G_1$, using $\hat{p}_t(x) > \rho_n$, $p_t(x) > \rho_n\log^c n$ and

Lemma F.11 of uniform upper bound for $\|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2$ and $\|\nabla p_t(x)\|^2 |\hat{p}_t(x) - p_t(x)|^2$,

$$
\begin{aligned}
I_t(x) \mathbb{1}_{\{\hat{p}(x) > \rho_n\} \cap A^c} &\leq \left(\frac{\|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2}{\rho_n^2} + \frac{\|\nabla p_t(x)\|^2 |\hat{p}_t(x) - p_t(x)|^2}{\rho_n^2 \, p_t^2(x)} \right) \mathbb{1}_{\{\hat{p}(x) > \rho_n\} \cap A^c} \\
&\lesssim \frac{t^{-(d+1)} \text{poly} \log(n)}{\rho_n^2} \left(1 + (\log n)^{-2c} \rho_n^{-2} \right) \mathbb{1}_{A^c}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\int_{G_1} \mathbb{E}\left[I_t(x) \mathbb{1}_{\{\hat{p}(x) > \rho_n\} \cap A^c} \right] p_t(x) dx &\lesssim \frac{t^{-(d+1)} \text{poly}(\log n)}{\rho_n^2} \left(1 + (\log n)^{-2c} \rho_n^{-2} \right) \int_{G_1} \mathbb{E}\left[\mathbb{1}_{A^c} \right] p_t(x) dx \\
&= \frac{t^{-(d+1)} \text{poly}(\log n)}{\rho_n^2} \left(1 + (\log n)^{-2c} \rho_n^{-2} \right) \int_{G_1} \mathbb{P}\left[x \in A^c \right] p_t(x) dx \\
&\leq \frac{t^{-(d+1)} \text{poly}(\log n)}{\rho_n^2} \left(1 + (\log n)^{-2c} \rho_n^{-2} \right) n^{-\alpha}, \quad (34)
\end{aligned}
$$

where in the last inequality we use the fact that $\mathbb{P}\left[x \in A^c \right] \leq n^{-\alpha}$. Since α can be arbitrarily large and $\frac{t^{-(d+1)} \text{poly} \log n}{\rho_n^2} \left(1 + (\log n)^{-2c} \rho_n^{-2} \right) = O(\text{poly}(n \log n))$, we can ignore the error of this term.

Part 3: The second term $\int_{G_1} \mathbb{P}\left(\hat{p}_t(x) \leq \rho_n \right) \|s_t(x)\|^2 p_t(x) dx$ **in** (29). Using Lemma D.5,

$$
\begin{aligned}
\int_{G_1} \mathbb{P}\left(\hat{p}_t(x) \leq \rho_n \right) \|s_t(x)\|^2 p_t(x) dx &\leq \int_{G_1} \mathbb{P}\left(x \in A^c \right) \|s_t(x)\|^2 p_t(x) dx \\
&\leq n^{-\alpha} \mathbb{E}_{X \sim p_t}\left[\|s_t(X)\|^2 \right] \\
&\lesssim n^{-\alpha} t^{-1}.
\end{aligned}
$$

Here in the last inequality, we use Lemma F.16 to bound the expectation of moment of the score function. Since α can be sufficiently large and $t^{-1} = O(\text{poly}(n))$, the error of this term can be negligible. $\qquad \square$

Lemma D.4. *Let A be the high probability set in Lemma F.6, i.e.,*

$$
A := \left\{ x \colon |\hat{p}_t(x) - p_t(x)| < C_4(\alpha) \, \text{polylog}(n) \left(\sqrt{\frac{p_t^*(x)}{nt^{\frac{d}{2}}}} + \frac{1}{nt^{\frac{d}{2}}} + (\log n)^{-\log n} \right) \right\}.
$$

For any $x \in A$, if x also satisfies: $\hat{p}_t(x) > \rho_n$ and $p_t(x) > \log^c n \, \rho_n$, then

$$
\hat{p}_t(x) \geq \frac{1}{2} p_t(x).
$$

sufficiently large n.

Proof. For any $x \in A$, since $p_t^*(x) \leq \|p\|_\infty$, we have

$$
\begin{aligned}
|\hat{p}_t(x) - p_t(x)| &\lesssim \text{polylog}(n) \left(\sqrt{\frac{1}{nt^{\frac{d}{2}}}} + \frac{1}{nt^{\frac{d}{2}}} + (\log n)^{-\log n} \right) \\
&= \text{polylog}(n) \left(\sqrt{\rho_n} + \rho_n + (\log n)^{-\log n} \right) \\
&\lesssim \text{polylog}(n) \left(\sqrt{\rho_n} + \rho_n \right)
\end{aligned}
$$

Therefore, under random set A and $G := \{x \colon p_t(x) > \log^c n \rho_n\}$,

$$
\begin{aligned}
\frac{|\hat{p}_t(x) - p_t(x)|}{p_t(x)} &\lesssim \text{polylog}(n) \frac{\sqrt{\rho_n} + \rho_n}{p_t(x)} \\
&\leq (\log^c n)^{-1} \text{polylog}(n) \left(\rho_n^{-1/2} + 1 \right) \\
&\lesssim \text{polylog}(n) \, n^{O(1)},
\end{aligned}
$$

23

where in the last inequality we use $\rho_n^{-1/2} = n^{1/2} t^{d/4} \leq n^{O(1)}$. Therefore

$$\frac{\hat{p}_t(x)}{p_t(x)} \geq 1 - \text{polylog}(n) \, n^{O(1)}.$$

When n is large, we can ensure that $\hat{p}_t(x) \geq \frac{1}{2} p_t(x)$. $\qquad\square$

Lemma D.5. *Let A be the high-probability random set defined in Lemma D.4. Define*

$$B := \{x : p_t(x) > \rho_n \log^c n\},$$

$$C := \{x : \hat{p}_t(x) < \rho_n\}.$$

Then $B \cap C \subseteq B \cap A^c$.

Proof. For $x \in B \cap C$,

$$|p_t(x) - \hat{p}_t(x)| > p_t(x) - \rho_n > \frac{1}{2} p_t(x),$$

when n is large. But by Lemma F.6, A implies $|p_t(x) - \hat{p}_t(x)| \lesssim (\log n)^{C_d} \sqrt{p_t^*(x)\rho_n}$ for some constant C_d that only depends on d. Using Lemma F.18, for any $\varepsilon > 0$, $A \cap B$ implies

$$
\begin{aligned}
|p_t(x) - \hat{p}_t(x)| &\lesssim (\log n)^{C_d} (\rho_n \log^c n)^{-\frac{\varepsilon}{2}} \sqrt{p_t(x)\rho_n} \\
&\leq (\log n)^{C_d} (\rho_n \log^c n)^{-\frac{\varepsilon}{2}} p_t(x) \log^{-\frac{c}{2}} n \\
&= (\log n)^{C_d} (n^{-1} t^{-d/2})^{-\frac{\varepsilon}{2}} (\log n)^{-\frac{c}{2}(\varepsilon+1)} p_t(x) \\
&\leq (\log n)^{C_d} n^{\frac{\varepsilon}{2}\left(1+\frac{dT_2}{2}\right)} (\log n)^{-\frac{c}{2}(\varepsilon+1)} p_t(x),
\end{aligned}
$$

where in the last inequality we use the fact that $t \leq n^{T_2}$. Therefore, by taking $\varepsilon = \frac{\log \log n}{\log n}$, we have $n^{\frac{\varepsilon}{2}\left(1+\frac{dT_2}{2}\right)} = (\log n)^{\frac{2+dT_2}{4}}$ and $(\log n)^{-\frac{c}{2}(\varepsilon+1)} \lesssim (\log n)^{-\frac{c}{2}}$, we have

$$(\log n)^{C_d} n^{\frac{\varepsilon}{2}\left(1+\frac{dT_2}{2}\right)} (\log n)^{-\frac{c}{2}(\varepsilon+1)} \leq (\log n)^{C_d + \frac{2+dT_2}{4} - \frac{c}{2}}.$$

Then since the constant c satisfies $\frac{c}{2} > 2C_d + \frac{2+dT_2}{2} > C_d + \frac{2+dT_2}{4}$, for large enough n we have $|p_t(x) - \hat{p}_t(x)| < \frac{1}{2} p_t(x)$ and $B \cap C \subseteq A^c$, as desired. $\qquad\square$

D.1.2. CASE 2: $p_t(x) < \rho_n \log^{-c} n$

Lemma D.6. *For any time $n^{-T_1} < t < n^{T_2}$, taking $\rho_n = \frac{1}{n \, t^{d/2}}$ and for any $0 < \varepsilon < 1$, the score error on the set*

$$G_2 := \{x : p_t(x) < \rho_n \log^{-c} n\}$$

can be upper bounded by

$$\mathbb{E}\left[\int_{G_2} \|\hat{s}_t(x) - s_t(x)\|^2 p_t(x) dx\right] \lesssim n^{-1} t^{-\frac{d+2}{2}} (t^{\frac{d}{2}} + \sigma_0^d) \, g_2(n, \varepsilon),$$

where

$$g_2(n, \varepsilon) := (\varepsilon)^{-(1+d/2)} n^{C_2 \varepsilon}$$

for some constant C_2 defined in the proof that only depends on T_1, T_2 and d (in turn, only depends on p_0).

Proof. Let A denote the high probability random set defined in Lemma F.6, then using Lemma D.7,

$$G_2 \cap A \subset \{x : \hat{p}_t(x) < \rho_n\}.$$

24

Then under $\{x \colon \hat{p}_t(x) < \rho_n\}$, by definition, $\hat{s}_t(x) = 0$. Therefore,

$$\mathbb{E}\left[\int_{G_2} \|\hat{s}_t(x) - s_t(x)\|^2 p_t(x) dx\right]$$

$$= \mathbb{E}\left[\int_{G_2} \|\hat{s}_t(x) - s_t(x)\|^2 \mathbb{1}_A(x) p_t(x) dx\right] + \mathbb{E}\left[\int_{G_2} \|\hat{s}_t(x) - s_t(x)\|^2 \mathbb{1}_{A^c}(x) p_t(x) dx\right]$$

$$\leq \mathbb{E}\left[\int_{G_2} \|\hat{s}_t(x) - s_t(x)\|^2 \mathbb{1}_{\{\hat{p}_t(x) < \rho_n\}}(x) p_t(x) dx\right] + \mathbb{E}\left[\int_{G_2} \|\hat{s}_t(x) - s_t(x)\|^2 \mathbb{1}_{A^c}(x) p_t(x) dx\right]$$

$$\lesssim \int_{G_2} \|s_t(x)\|^2 p_t(x) dx + \left(\mathbb{E}_{X \sim p_t, \{x_i\}_{i=1}^n}\left[\|\hat{s}_t(X) - s_t(X)\|^4\right]\right)^{1/2} \left(\int \mathbb{P}(x \in A^c) p_t(x) dx\right)^{1/2}$$

$$\leq \int_{G_2} \|s_t(x)\|^2 p_t(x) dx + \left(\mathbb{E}_{X \sim p_t, \{x_i\}_{i=1}^n}\left[\|\hat{s}_t(X) - s_t(X)\|^4\right]\right)^{1/2} n^{-\alpha/2},$$

where the last inequality follows by the Cauchy–Schwarz inequality. Therefore the error in the second term can be ignored since by Lemma F.12 the fourth moment of the score error is bounded by some $\mathrm{poly}(n^{-1}\log n)$ but $n^{-\alpha}$ can be arbitrarily small. For the first term, using Lemma F.17,

$$\int_{G_2} \|s_t(x)\|^2 p_t(x) dx \lesssim (\varepsilon)^{-(1+d/2)} t^{-1} \rho_n^{1-\varepsilon} \sigma_t^{d(1-\varepsilon)}.$$

Then the result follows by $\rho_n = \frac{1}{nt^{d/2}} = \mathrm{poly}(n)$ and $\sigma_t^d \lesssim t^{\frac{d}{2}} + \sigma_0^d \lesssim \mathrm{poly}(n)$, and the order of the polynomials only depends on T_1 and T_2.

Lemma D.7. *For any x satisfying $x \in A$ and $p_t(x) < \rho_n \log^{-c} n$, we have $\hat{p}_t(x) < \rho_n$.*

Proof. By definition of the random set A,

$$\hat{p}_t(x) \leq p_t(x) + (\log n)^{C_d}\left(\sqrt{p_t^*(x)\rho_n} + \rho_n + (\log n)^{-\frac{(\log n)}{2}}\right)$$

$$\lesssim \rho_n \log^{-c} n + (\log n)^{C_d}\sqrt{p_t^*(x)\rho_n}.$$

Using Lemma F.18, for any $\varepsilon > 0$,

$$\sqrt{p_t^*(x)\rho_n} \lesssim \sqrt{p_t^{1-\varepsilon}(x)\rho_n} \lesssim (\log n)^{-\frac{c(1-\varepsilon)}{2}} \rho_n^{1-\frac{\varepsilon}{2}}.$$

Using the fact that $t \leq n^{T_2}$,

$$\rho_n^{-\frac{\varepsilon}{2}} = (n^{-1}t^{-d/2})^{-\frac{\varepsilon}{2}} \leq n^{\frac{\varepsilon}{2}\left(1+\frac{dT_2}{2}\right)},$$

we have

$$\hat{p}_t(x) \lesssim \left(1 + (\log n)^{C_d} n^{\frac{\varepsilon}{2}\left(1+\frac{dT_2}{2}\right)} \log^{\frac{c(\varepsilon+1)}{2}} n\right) \rho_n \log^{-c} n.$$

Then we can take $\varepsilon = \frac{\log\log n}{\log n}$ such that $n^{\frac{\varepsilon}{2}\left(1+\frac{dT_2}{2}\right)} = (\log n)^{\frac{2+dT_2}{4}}$ and $\log^{\frac{c(\varepsilon+1)}{2}} n \leq \log^{\frac{c(1/2+1)}{2}} n = \log^{\frac{3c}{4}} n$. Therefore,

$$\left(1 + (\log n)^{C_d} n^{\frac{\varepsilon}{2}\left(1+\frac{dT_2}{2}\right)} \log^{\frac{c(\varepsilon+1)}{2}} n\right) \log^{-c} n < \log^{-c} n + (\log n)^{C_d + \frac{2+dT_2}{4} - \frac{c}{4}}$$

$$< 1,$$

for sufficiently large n, since $c > 4C_d + dT_2 + 2$. Therefore, this implies $\hat{p}_t(x) < \rho_n$. $\qquad\square$

$\square$

D.1.3. CASE 3: $\rho_n \log^{-c} n < p_t(x) < \rho_n \log^c n$

Lemma D.8. *For any time $n^{-T_1} < t < n^{T_2}$ and any $0 < \varepsilon < 1$, the score error on the set*

$$G_3 := \{x \colon \rho_n \log^{-c} n < p_t(x) < \rho_n \log^c n\}$$

can be upper bounded by

$$\mathbb{E}\left[\int_{G_3} \|\hat{s}_t(x) - s_t(x)\|^2 p_t(x) dx\right] \lesssim \text{polylog}(n)\, n^{-1}\, t^{-\frac{d+2}{2}}(t^{\frac{d}{2}} + \sigma_0^d)\, g_3(n, \varepsilon),$$

where

$$g_3(n, \varepsilon') := (\varepsilon)^{-(1+d/2)} \exp\left(\frac{1}{\varepsilon \log n}\right) n^{C_3 \varepsilon}$$

for some constant C_3 defined in the proof that only depends on T_1, T_2 and d.

Proof. Again, apply Lemma D.1,

$$\int_{G_3} \mathbb{E}\left[\|\hat{s}_t(x) - s_t(x)\|^2\right] p_t(x) dx$$

$$\leq 2\int_{G_3} \mathbb{E}\left[I_t(x)\mathbb{1}_{\{\hat{p}_t(x) > \rho_n\}}\right] p_t(x) dx + \int_{G_3} \mathbb{P}\left(\hat{p}_t(x) \leq \rho_n\right) \|s_t(x)\|^2 p_t(x) dx \qquad (35)$$

where

$$I_t(x) := \frac{\|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2 + \|s_t(x)\|^2 |\hat{p}_t(x) - p_t(x)|^2}{\hat{p}_t(x)^2}.$$

Now we split the proof into two parts:

Part 1: The first term $\int_{G_3} \mathbb{E}\left[I_t(x)\mathbb{1}_{\{\hat{p}_t(x) > \rho_n\}}\right] p_t(x) dx$ in (35).

First note that for any $x \in G_3$, we have $p_t(x) \log^{-c} n < \rho_n < p_t(x) \log^c n$. Therefore,

$$\hat{p}_t(x) \geq \rho_n \geq p_t(x) \log^{-c} n.$$

Then we have

$$I_t(x) \leq (\log^{2c} n)(p_t(x))^{-2} \left(\|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2 + \|s_t(x)\|^2 |\hat{p}_t(x) - p_t(x)|^2\right)$$

Using Proposition C.3 and Proposition C.4:

$$\mathbb{E}\left[|\hat{p}_t(x) - p_t(x)|^2\right] \leq C_2 \text{polylog}(n)\left(\frac{p_t^*(x)}{nt^{\frac{d}{2}}} + (\log n)^{-\log n}\right),$$

$$\mathbb{E}\left[\|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2\right] \leq C_3 \text{polylog}(n)\left(\frac{p_t^*(x)}{n\, t^{\frac{d+2}{2}}} + (\log n)^{-\log n}\, t^{-1}\right),$$

Then

$$\mathbb{E}\left[I_t(x)\right] \lesssim \text{polylog}(n)\, (p_t(x))^{-2}\left[\frac{p_t^*(x)}{n\, t^{\frac{d+2}{2}}} + \frac{p_t^*(x)\|s_t(x)\|^2}{n\, t^{1/2}} + (\log n)^{-\log n}\, (t^{-1} + \|s_t(x)\|^2)\right]$$

Integrate over x:

$$\int_{G_3} \mathbb{E}\left[I_t(x)\mathbb{1}_{\{\hat{p}(x) > \rho_n\}}\right] p_t(x) dx \lesssim \text{polylog}(n)\left[\frac{1}{nt^{\frac{d+2}{2}}}\int_{G_3}\frac{p_t^*(x)}{p_t(x)}dx\right.$$

$$+ \frac{1}{nt^{\frac{d}{2}}}\int_{G_3}\frac{p_t^*(x)}{p_t(x)}\|s_t(x)\|^2 dx$$

$$\left.+ (\log n)^{-\log n}\int_{G_3}\left(\frac{t^{-1}}{p_t(x)} + \frac{\|s_t(x)\|^2}{p_t(x)}\right) dx\right].$$

26

Notice that $G_3 \subset \{x\colon p_t(x) > \rho_n \log^{-c} n\} := G_3'$, which is similar to Case 1 (by replacing $\log^c n$ with $\log^{-c} n$). Therefore, using a similar approach in Lemma D.2 of Case 1, we know the term $(\log n)^{-\log n}$ is small than $\text{poly}(n \log n)$ for sufficiently large n, then the error for the last term can be ignored:

$$(\log n)^{-\log n} \int_{G_3} \left(\frac{t^{-1}}{p_t(x)} + \frac{\|s_t(x)\|^2}{p_t(x)} \right) dx \leq (\log n)^{-\log n} |G_3'| \left(t^{-1} \rho_n^{-1} \log^c n + t^{-\frac{d}{2}} \rho_n^3 \log^{3c} n \right)$$

$$\lesssim (\log n)^{-\log n} \text{poly}(n \log n)$$

$$\ll \text{poly}(n^{-1}).$$

Again, using the similar approach in Lemma D.2 in Case 1,

$$\frac{1}{nt^{(d+2)/2}} \int_{G_3} \frac{p_t^*(x)}{p_t(x)} dx + \frac{1}{n \, t^{d/2}} \int_{G_3} \frac{p_t^*(x)}{p_t(x)} \|s_t(x)\|^2 dx$$

$$\lesssim \frac{(t^{d/2} + \sigma_0^d)}{n \, t^{(d+2)/2}} \frac{1}{\sqrt{1-\varepsilon}} \exp\left\{ \frac{1-\varepsilon}{D_n^2 \varepsilon} \right\} (\log^c n)^\varepsilon \rho_n^{-\varepsilon} \left(1 + m (\log^c n)^{\frac{1}{m}} \rho_n^{-\frac{1}{m}} \right).$$

As before, by taking $m \asymp \frac{1}{\varepsilon}$, and $D_n = C\sqrt{\log n}$, then $\frac{1}{\sqrt{1-\varepsilon}} \exp\left\{ \frac{1-\varepsilon}{D_n^2 \varepsilon} \right\} \lesssim \exp\left(\frac{1}{\varepsilon \log n} \right)$. Since $\varepsilon < 1$, we know $(\log^c n)^\varepsilon < \log^c n$ and $(\log^c n)^{1/m} < \log^c n$. Using $\rho_n = n^{-1} t^{-d/2} > n^{-1} n^{-dT_2/2}$ we can conclude that

$$\int_{G_3} \mathbb{E}\left[I_t(x) \mathbb{1}_{\{\hat{p}(x) > \rho_n\}} \right] p_t(x) dx \lesssim \text{polylog}(n) \, n^{-1} \, t^{-\frac{d+2}{2}} (t^{\frac{d}{2}} + \sigma_0^d) \left(\frac{1}{\varepsilon} n^{(2+dT_2)\varepsilon} \exp\left(\frac{1}{\varepsilon \log n} \right) \right).$$

Part 2: The second term $\int_{G_3} \mathbb{P}\left(\hat{p}_t(x) \leq \rho_n \right) \|s_t(x)\|^2 p_t(x) dx$ **in** (35).

Notice that

$$\int_{G_3} \mathbb{P}\left(\hat{p}_t(x) \leq \rho_n \right) \|s_t(x)\|^2 p_t(x) dx \leq \int_{\{x \,:\, p_t(x) \leq \rho_n \log^c n\}} \|s_t(x)\|^2 p_t(x) dx.$$

Therefore, we can use a similar method of case 2 as in Lemma F.17, by replacing $\log^{-c} n$ with $\log^c n$:

$$\int_{\{x \,:\, p_t(x) \leq \rho_n \log^c n\}} \|s_t(x)\|^2 p_t(x) dx \leq \text{polylog}(n) \, (\varepsilon)^{-(1+d/2)} \, t^{-1} \rho_n^{1-\varepsilon} \sigma_t^{d(1-\varepsilon)}.$$

Since $\rho_n = \frac{1}{nt^{d/2}} = \text{poly}(n)$ and $\sigma_t^d \lesssim t^{\frac{d}{2}} + \sigma_0^d \lesssim \text{poly}(n)$, and the order of the polynomials C_2 (the same as that in case 2) only depends on T_1, T_2 and d. Then can conclude that

$$\int_{G_3} \mathbb{P}\left(\hat{p}_t(x) \leq \rho_n \right) \|s_t(x)\|^2 p_t(x) dx \leq \text{polylog}(n)(\varepsilon)^{-(1+d/2)} \, n^{-1} t^{-\frac{d+2}{2}} \left(\sigma_0^{\frac{d}{2}} + t^{\frac{d}{2}} \right) n^{C_2 \varepsilon}.$$

$\square$

D.1.4. COMBINING THE THREE CASES

Next, combining the three cases in Lemma D.2, Lemma D.6 and Lemma D.8 as discussed above,

$$\mathbb{E}\left[\int_x \|\hat{s}_t(x) - s_t(x)\|^2 p_t(x) \, dx \right] = \mathbb{E}\left[\int_{G_1 \cup G_2 \cup G_3} \|\hat{s}_t(x) - s_t(x)\|^2 p_t(x) \, dx \right]$$

$$\lesssim \text{polylog}(n) \, n^{-1} t^{-\frac{d+2}{2}} (t^{\frac{d}{2}} + \sigma_0^d) \, g(n, \varepsilon),$$

where

$$g(n, \varepsilon) := g_1(n, \varepsilon) + g_2(n, \varepsilon) + g_3(n, \varepsilon)$$

$$\leq \left(\exp\left(\frac{1}{\varepsilon \log n} \right) + (\varepsilon)^{-d/2} + (\varepsilon)^{-d/2} \exp\left(\frac{1}{\varepsilon \log n} \right) \right) (\varepsilon)^{-1} n^{(C_1 \vee C_2 \vee C_3)\varepsilon}.$$

Let $\varepsilon = \frac{\log \log n}{\log n}$, since $n^\varepsilon = n^{\frac{\log \log n}{\log n}} = \log n$, we have $g(n, \varepsilon) \leq C \text{polylog}(n)$ for some universal constant C. Therefore,

$$\mathbb{E}\left[\int_x \|\hat{s}_t(x) - s_t(x)\|^2 p_t(x) \, dx \right] \lesssim \text{polylog}(n) \, n^{-1} t^{-\frac{d+2}{2}} (t^{\frac{d}{2}} + \sigma_0^d).$$

D.2. Proof of Corollary 3.7

In Theorem 3.5, taking the integral with respect to time t from $t_0 = n^{-T_1}$ to $T = n^{T_2}$, we have:

$$\int_{t \in [t_0, T]} \int_x \mathbb{E}\|\hat{s}_t(x) - s_t(x)\|^2 p_t(x) dx dt \lesssim \text{polylog}(n) \, n^{-1} \int_{t_0}^{T} t^{-\frac{d+2}{2}} (t^{\frac{d}{2}} + \sigma_0^d) dt$$

$$= \text{polylog}(n) \, n^{-1} \left(\log(T) - \log(t_0) - \frac{2\sigma_0^d}{d} T^{-\frac{d}{2}} + \frac{2\sigma_0^d}{d} t_0^{-\frac{d}{2}} \right)$$

$$\leq \text{polylog}(n) \, n^{-1} \left(\log(T) - \log(t_0) + \frac{2\sigma_0^d}{d} t_0^{-\frac{d}{2}} \right)$$

$$\lesssim \text{polylog}(n) \, n^{-1} \, n^{\frac{d \, T_1}{2}}.$$

The second part of Corollary 3.7 simply follows by letting $T_1 = \frac{2}{2\beta + d}$ and noticing that $n^{-1} n^{\frac{d T_1}{2}} = n^{-\frac{2\beta}{2\beta + d}}$.

E. Proof of Theorem 3.8

E.1. Control of the error from early stopping

Theorem E.1. *Under Assumption 3.1 and Assumption 3.4, if $\beta \in [0, 2]$, $t_0 = n^{-\frac{2}{2\beta + d}}$ and $p_{t_0} = p_0 * \Phi_t$, where Φ_t is the density of Gaussian distribution in d-dimension, $\mathcal{N}(0, t\mathbf{I}_d)$ and $*$ denote the convolution operator, then there exists a constant C that depends on p_0, β, L and dimension d such that*

$$\text{TV}(p_0, p_{t_0}) \leq C \, \text{polylog}(n) \, n^{-\frac{\beta}{2\beta + d}}.$$

Proof. Firstly, the total variance (TV) distance between p_0 and p_t can be decomposed into two terms and by Jensen's inequality,

$$\text{TV}(p_0, p_t) = \int_{\mathbb{R}^d} |p_0(x) - p_t(x)| \, dx$$

$$= \int_{\|x\|_\infty < \log n} |p_0(x) - p_t(x)| \, dx + \int_{\|x\|_\infty > \log n} |p_0(x) - p_t(x)| \, dx$$

$$\leq (2 \log n)^{d/2} \sqrt{\int_{\mathbb{R}^d} |p_0(x) - p_t(x)|^2 \, dx} + \int_{\|x\|_\infty > \log n} |p_0(x) - p_t(x)| \, dx. \tag{36}$$

By the Sub-Gaussian tail bound in Lemma F.2, the error of the second term in (36) is negligible:

$$\int_{\|x\|_\infty > \log n} |p_0(x) - p_t(x)| dx \leq \int_{\|x\|_\infty > \log n} p_0(x) dx + \int_{\|x\|_\infty > \log n} p_t(x) dx$$

$$\leq 2d \exp\left(-\frac{(\log n)^2}{2\sigma_0^2} \right) + 2d \exp\left(-\frac{(\log n)^2}{2(\sigma_0 + \sqrt{t})^2} \right)$$

$$\leq 4dn^{-\frac{\log n}{\sigma_0^2 + 1}}$$

$$\ll n^{-\frac{\beta}{2\beta + d}}. \tag{37}$$

Next, we derive an upper bound for the first term $\int_{\mathbb{R}^d} |p_0(x) - p_t(x)|^2 \, dx$ in (36). By Plancherel's theorem,

$$\int |p_0(x) - p_t(x)|^2 \, dx = \frac{1}{(2\pi)^d} \int |\mathcal{F}[p_0](\omega) - \mathcal{F}[p_0](\omega)|^2 \, d\omega$$

$$= \frac{1}{(2\pi)^d} \int_{\|\omega\|_\infty \geq t^{-1/2}} |\mathcal{F}[p_0](\omega)|^2 \, |1 - \phi_t(\omega)|^2 \, d\omega \tag{38}$$

$$+ \frac{1}{(2\pi)^d} \int_{\|\omega\|_\infty < t^{-1/2}} |\mathcal{F}[p_0](\omega)|^2 \, |1 - \phi_t(\omega)|^2 \, d\omega. \tag{39}$$

Here $\phi_t(\omega) := \exp\left(-\frac{t^2 \|\omega\|^2}{2}\right)$ is the Fourier transformation (or Characteristic function) of Gaussian density. By the definition of the Sobolev class of density in Assumption 3.4, we have

$$
\begin{aligned}
d(2\pi)^d L^2 &\geq \int_{\mathbb{R}^d} \sum_{i=1}^{d} |\omega_i|^{2\beta} \left|\mathcal{F}[p_0](\omega)\right|^2 d\omega \\
&\geq \int_{\|\omega\|_\infty \geq t^{-1/2}} \sum_{i=1}^{d} |\omega_i|^{2\beta} \left|\mathcal{F}[p_0](\omega)\right|^2 d\omega \\
&\geq t^{-\beta} \int_{\|\omega\|_\infty \geq t^{-1/2}} \left|\mathcal{F}[p_0](\omega)\right|^2 d\omega
\end{aligned}
\tag{40}
$$

For the first term (38), use the fact that $|1 - \phi_t(\omega)| \leq 2$ and (40), substituting $t = n^{-\frac{2}{2\beta+d}}$, we have

$$
\frac{1}{(2\pi)^d} \int_{\|\omega\|_\infty t^{-1/2}} \left|\mathcal{F}[p_0](\omega)\right|^2 |1 - \phi_t(\omega)|^2 \, d\omega \leq dL^2 t^\beta = dL^2 \, n^{-\frac{2\beta}{2\beta+d}}.
\tag{41}
$$

For the second term in (39), use the fact that $|1 - \phi_t(\omega)|^2 = \left|1 - \exp\left(-\frac{\|\omega\|^2 t}{2}\right)\right|^2 \leq \frac{t^2 \|\omega\|^4}{4}$, we have

$$
\begin{aligned}
&\frac{1}{(2\pi)^d} \int_{\|\omega\|_\infty < t^{-1/2}} \left|\mathcal{F}[p_0](\omega)\right|^2 |1 - \phi_t(\omega)|^2 \, d\omega \\
&\leq \frac{1}{(2\pi)^d} \cdot \frac{t^2}{4} \int_{\|\omega\|_\infty < t^{-1/2}} \left|\mathcal{F}[p_0](\omega)\right|^2 \|\omega\|^4 d\omega \\
&= \frac{1}{(2\pi)^d} \cdot \frac{t^2}{4} \int_{\|\omega\|_\infty < t^{-1/2}} \left|\mathcal{F}[p_0](\omega)\right|^2 \|\omega\|^{2\beta} \cdot \|\omega\|^{4-2\beta} d\omega \\
&\leq \frac{1}{(2\pi)^d} \cdot \frac{t^2}{4} \sup_{\|\omega\|_\infty < t^{-1/2}} \|\omega\|^{4-2\beta} \int \left|\mathcal{F}[p_0](\omega)\right|^2 \|\omega\|^{2\beta} d\omega.
\end{aligned}
\tag{42}
$$

Since $\beta \leq 2$, we have

$$
\sup_{\|\omega\|_\infty < t^{-1/2}} \|\omega\|^{4-2\beta} = \sup_{\|\omega\|_\infty < t^{-1/2}} \left(\sum_{i=1}^{d} |\omega_i|^2\right)^{2-\beta} \leq \left(d \|\omega\|_\infty^2\right)^{2-\beta} \leq d^{2-\beta} t^{\beta-2}.
\tag{43}
$$

Besides, since $\|\omega\|^{2\beta} = \left(\sum_{i=1}^{d} |\omega_i|^2\right)^\beta \leq \left(d^{\beta-1} \vee 1\right) \sum_{i=1}^{d} |\omega_i|^{2\beta}$, and by the definition of Sobolev class of densities, we have

$$
\begin{aligned}
\int \left|\mathcal{F}[p_0](\omega)\right|^2 \|\omega\|^{2\beta} d\omega &\leq \left(d^{\beta-1} \vee 1\right) \sum_{i=1}^{d} \int \left|\mathcal{F}[p_0](\omega)\right|^2 |\omega_i|^{2\beta} d\omega \\
&\leq d \left(d^{\beta-1} \vee 1\right) (2\pi)^d L^2.
\end{aligned}
\tag{44}
$$

Therefore, combining (43) and (44),

$$
(42) \leq \frac{L^2 d^{3-\beta} \left(d^{\beta-1} \vee 1\right)}{4} t^\beta = \frac{L^2 d^{3-\beta} \left(d^{\beta-1} \vee 1\right)}{4} n^{-\frac{2\beta}{2\beta+d}}.
\tag{45}
$$

As a result of combining the above terms (41), (45) and (37), we are able to deduce that the total error between p_0 and p_t is $\mathrm{TV}(p_0, p_t) \lesssim \mathrm{polylog}(n) n^{-\frac{\beta}{2\beta+d}}$. $\qquad\square$

E.2. Girsanov's Theorem

We can translate the cumulative score error to the KL divergence error by using the following Girsanov Theorem.

Theorem E.2 (Girsanov Theorem, Le Gall (2016)). *Define the stochastic process $(\mathcal{L}_t)_{t \in [0,T]}$ by $\mathcal{L}_t := \int_0^t \langle b_s, dB_s \rangle$, where B is a $\mathbb{Q}$-Brownian motion. Assume that $E_{\mathbb{Q}} \int_0^T \|b_s\|^2 \, ds < \infty$. Then, L is a Q-martingale in $L^2(\mathbb{Q})$. Moreover, if*

$$E_{\mathbb{Q}} \mathcal{E}(\mathcal{L})_T = 1,$$

where

$$
\begin{aligned}
\mathcal{E}(\mathcal{L})_t &:= \exp\left(\mathcal{L}_t - \frac{1}{2}[\mathcal{L}]_t \right) \\
&= \exp\left(\int_0^t \langle b_s, dB_s \rangle - \frac{1}{2} \int_0^t \|b_s\|^2 \, ds \right).
\end{aligned}
$$

Then $\mathcal{E}(\mathcal{L})_T$ is also a $\mathbb{Q}$-martingale and the process

$$\tilde{B}_t := B_t - \int_0^t b_s \, ds$$

is a Brownian motion under the measure $\mathbb{P}$ defined by the Radon-Nikodym derivative:

$$d\mathbb{P} := \mathcal{E}(\mathcal{L})_T \, d\mathbb{Q}.$$

The Girsanov theorem can be used to convert a Brownian motion with drift into a standard Brownian motion by changing the measure. In this paper, we aim to derive the KL divergence between two stochastic processes, $(\boldsymbol{X}_t)_{t \in [0,T]}$ and $(\boldsymbol{Y}_t)_{t \in [0,T]}$. These processes have different drift terms. More specifically, $(\boldsymbol{X}_t)_{t \in [0,T]}$ and $(\boldsymbol{Y}_t)_{t \in [0,T]}$ are solutions to the following two SDEs:

$$d\boldsymbol{X}_t = b^{\boldsymbol{X}}(\boldsymbol{X}_t, t) \, dt + d\boldsymbol{B}_t, \quad \boldsymbol{X}_0 \sim p,$$
$$d\boldsymbol{Y}_t = b^{\boldsymbol{Y}}(\boldsymbol{Y}_t, t) \, dt + d\boldsymbol{B}_t; \quad \boldsymbol{Y}_0 \sim p.$$

Denote by $\mathbb{P}_{\boldsymbol{X}}$ and $\mathbb{P}_{\boldsymbol{Y}}$ the path measures of $(\boldsymbol{X}_t)_{t \in [0,T]}$ and $(\boldsymbol{Y}_t)_{t \in [0,T]}$, respectively. Since the initial distributions of these two processes are identical, the only difference for these two processes is their drift terms $b^{\boldsymbol{X}}$ and $b^{\boldsymbol{Y}}$. By Girsanov theorem, we can derive the Radon-Nikodym derivative for their path measure $\frac{d\mathbb{P}_{\boldsymbol{X}}}{d\mathbb{P}_{\boldsymbol{Y}}}$, which allows us to compute the KL divergence for these two processes:

$$D_{\mathrm{KL}}\left(\mathbb{P}_{\boldsymbol{X}} \,\|\, \mathbb{P}_{\boldsymbol{Y}} \right) = \mathbb{E}_{\mathbb{P}_{\boldsymbol{X}}} \left[\log \frac{d\mathbb{P}_{\boldsymbol{X}}}{d\mathbb{P}_{\boldsymbol{Y}}} \right].$$

The above discussion is summarized in the following lemma, as described in Chen et al. (2023b) and Oko et al. (2023).

Lemma E.3 (Chen et al. (2023b), Oko et al. (2023)). *Let p be any probability distribution and let $(\boldsymbol{X}_t)_{t \in [0,T]}$, $(\boldsymbol{Y}_t)_{t \in [0,T]}$ be solutions to the following two SDEs:*

$$d\boldsymbol{X}_t = b^{\boldsymbol{X}}(\boldsymbol{X}_t, t) \, dt + d\boldsymbol{B}_t, \quad \boldsymbol{X}_0 \sim p,$$

$$d\boldsymbol{Y}_t = b^{\boldsymbol{Y}}(\boldsymbol{Y}_t, t) \, dt + d\boldsymbol{B}_t; \quad \boldsymbol{Y}_0 \sim p.$$

We denote the distribution of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ as $p_t^{\boldsymbol{X}}$ and $p_t^{\boldsymbol{Y}}$ and the path measure of $(\boldsymbol{X}_t)_{t \in [0,T]}$ and $(\boldsymbol{Y}_t)_{t \in [0,T]}$ as $\mathbb{P}_{\boldsymbol{X}}$ and $\mathbb{P}_{\boldsymbol{Y}}$, respectively.

1. *Suppose the following Novikov's condition holds:*

$$\mathbb{E}_{\mathbb{P}_{\boldsymbol{X}}} \left[\exp\left(\int_0^T \|b^{\boldsymbol{X}}(\boldsymbol{X}_t, t) - b^{\boldsymbol{Y}}(\boldsymbol{X}_t, t)\|^2 \, dt \right) \right] < \infty. \tag{46}$$

Then the Radon-Nikodym derivative of $\mathbb{P}^{\boldsymbol{X}}$ with respect to $\mathbb{P}^{\boldsymbol{Y}}$ is

$$\frac{d\mathbb{P}_{\boldsymbol{X}}}{d\mathbb{P}_{\boldsymbol{Y}}}(\boldsymbol{Z}) = \exp\left(\frac{1}{2} \int_0^T \|b^{\boldsymbol{X}}(\boldsymbol{Z}_t, t) - b^{\boldsymbol{Y}}(\boldsymbol{Z}_t, t)\|^2 \, dt - \int_0^T \left(b^{\boldsymbol{X}}(\boldsymbol{Z}_t, t) - b^{\boldsymbol{Y}}(\boldsymbol{Z}_t, t) \right) \, d\boldsymbol{B}_t \right).$$

Therefore we have

$$D_{KL} \left(\mathbb{P}_{\boldsymbol{X}} \| \mathbb{P}_{\boldsymbol{Y}} \right) = \frac{1}{2} \, \mathbb{E}_{\mathbb{P}_{\boldsymbol{X}}} \left[\int_0^T \| b^{\boldsymbol{X}}(\boldsymbol{X}_t, t) - b^{\boldsymbol{Y}}(\boldsymbol{X}_t, t) \|^2 \, dt \right]$$

$$= \frac{1}{2} \int_x \int_0^T p_t^{\boldsymbol{X}}(x) \| b^{\boldsymbol{X}}(x, t) - b^{\boldsymbol{Y}}(x, t) \|^2 \, dt \, dx.$$

2. *If* $\int_x \int_0^T p_t^{\boldsymbol{X}}(x) \| b^{\boldsymbol{X}}(x, t) - b^{\boldsymbol{Y}}(x, t) \|^2 \, dt \, dx \leq C$ *holds for some constant* C, *then*

$$D_{KL} \left(\mathbb{P}_{\boldsymbol{X}} \| \mathbb{P}_{\boldsymbol{Y}} \right) \leq \frac{1}{2} \int_x \int_0^T p_t^{\boldsymbol{X}}(x) \| b^{\boldsymbol{X}}(x, t) - b^{\boldsymbol{Y}}(x, t) \|^2 \, dt \, dx,$$

even if Novikov's condition (46) *is not satisfied.*

E.3. Proof of Theorem 3.8

We first recall some notations of the stochastic processes we defined in Section 2. The forward process (4) is denoted by $(\boldsymbol{X}_t)_{t \in [0,T]}$ and $\boldsymbol{X}_t \sim p_t$; The backward process (8) is denoted by $(\boldsymbol{Y}_t)_{t \in [0,T]}$ and by definition, $\boldsymbol{Y}_t \sim p_{T-t}$; The process of Algorithm 1 is denoted by $(\hat{\boldsymbol{Y}}_t)_{t \in [0,T]}$. Let $(\bar{\boldsymbol{Y}}_t)_{t \in [0,T]}$ be $(\hat{\boldsymbol{Y}}_t)_{t \in [0,T]}$ replacing $\hat{\boldsymbol{Y}}_0 \sim \mathcal{N}(0, T\boldsymbol{I}_d)$ by $\bar{\boldsymbol{Y}}_0 \sim p_T$, i.e., $(\bar{\boldsymbol{Y}}_t)_{t \in [0,T]}$ satisfies:

$$d\bar{\boldsymbol{Y}}_t = \hat{s}_{T-t}(\bar{\boldsymbol{Y}}_t) dt + d\boldsymbol{B}_t, \quad \bar{\boldsymbol{Y}}_0 \sim p_T.$$

Now we start the proof. By the triangle inequality,

$$\mathbb{E} \left[\mathrm{TV} \left(\boldsymbol{X}_0, \hat{\boldsymbol{Y}}_{T-t_0} \right) \right] \leq \mathrm{TV} \left(\boldsymbol{X}_0, \boldsymbol{X}_{t_0} \right) + \mathrm{TV} \left(\boldsymbol{X}_{t_0}, \boldsymbol{Y}_{T-t_0} \right)$$

$$+ \mathbb{E} \left[\mathrm{TV} \left(\boldsymbol{Y}_{T-t_0}, \bar{\boldsymbol{Y}}_{T-t_0} \right) \right] + \mathbb{E} \left[\mathrm{TV} \left(\bar{\boldsymbol{Y}}_{T-t_0}, \hat{\boldsymbol{Y}}_{T-t_0} \right) \right].$$

By Theorem E.1, $\mathrm{TV} \left(\boldsymbol{X}_0, \boldsymbol{X}_{t_0} \right) \lesssim \mathrm{polylog}(n) \, n^{-\frac{\beta}{2\beta+d}}$. By definition of the backward process (8), $\mathrm{TV} \left(\boldsymbol{X}_{t_0}, \boldsymbol{Y}_{T-t_0} \right) = 0$. For the third term, by Pinsker's inequality and data-processing inequality,

$$\mathrm{TV} \left(\boldsymbol{Y}_{T-t_0}, \bar{\boldsymbol{Y}}_{T-t_0} \right) \lesssim \sqrt{D_{KL} \left(\boldsymbol{Y}_{T-t_0} \| \bar{\boldsymbol{Y}}_{T-t_0} \right)} \leq \sqrt{D_{KL} \left(\mathbb{P}_{\boldsymbol{Y}} \| \mathbb{P}_{\bar{\boldsymbol{Y}}} \right)},$$

where $\mathbb{P}_{\boldsymbol{Y}}$ and $\mathbb{P}_{\bar{\boldsymbol{Y}}}$ are the path measure for $(\boldsymbol{Y}_t)_{t \in [0, T-t_0]}$ and $(\bar{\boldsymbol{Y}}_t)_{t \in [0, T-t_0]}$. Then using the second part of Lemma E.3 by taking $b_t^{\boldsymbol{X}} = \nabla \log p_{T-t}(\boldsymbol{x})$ and $b_t^{\boldsymbol{Y}} = \hat{s}_{T-t}(\boldsymbol{x})$ (see Section E.2 for details), we have

$$D_{KL} \left(\mathbb{P}_{\boldsymbol{Y}} \| \mathbb{P}_{\bar{\boldsymbol{Y}}} \right) \leq \frac{1}{2} \int_0^{T-t_0} \mathbb{E}_{\boldsymbol{x} \sim p_t} \left[\| \hat{s}_{T-t}(\boldsymbol{x}) - \nabla \log p_{T-t}(\boldsymbol{x}) \|^2 \right] \, dt$$

$$= \frac{1}{2} \int_{t_0}^T \mathbb{E}_{\boldsymbol{x} \sim p_t} \left[\| \hat{s}_t(\boldsymbol{x}) - \nabla \log p_t(\boldsymbol{x}) \|^2 \right] \, dt.$$

Then $\mathbb{E} \left[D_{KL} \left(\mathbb{P}_{\boldsymbol{Y}} \| \mathbb{P}_{\bar{\boldsymbol{Y}}} \right) \right]$ can be bounded by Corollary 3.7. Since the only difference of $\bar{\boldsymbol{Y}}_{T-t_0}$ and $\hat{\boldsymbol{Y}}_{T-t_0}$ is their initial distribution, we have

$$\mathbb{E} \left[\mathrm{TV} \left(\bar{\boldsymbol{Y}}_{T-t_0}, \hat{\boldsymbol{Y}}_{T-t_0} \right) \right] \leq \mathrm{TV} \left(\boldsymbol{X}_T, \mathcal{N}(0, T\boldsymbol{I}_d) \right).$$

Again, by Pinsker's inequality,

$$\mathrm{TV} \left(\boldsymbol{X}_T, \mathcal{N}(0, T\boldsymbol{I}_d) \right) \leq \sqrt{D_{KL} \left(\boldsymbol{X}_T \| \mathcal{N}(0, T\boldsymbol{I}_d) \right)}.$$

Using Jensen's inequality,

$$D_{KL} \left(\boldsymbol{X}_T \| \mathcal{N}(0, T\boldsymbol{I}_d) \right) \leq \mathbb{E}_{\boldsymbol{x} \sim p_0} \left[D_{KL} \left(\boldsymbol{X}_T | \boldsymbol{X}_0 = \boldsymbol{x} \| \mathcal{N}(0, T\boldsymbol{I}_d) \right) \right]$$

$$= \frac{1}{2T} \mathbb{E}_{\boldsymbol{x} \sim p_0} \left[\boldsymbol{x}^T \boldsymbol{x} \right].$$

Here in the last equality, we use the fact that $\boldsymbol{X}_T | \boldsymbol{X}_0 = \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}, T\boldsymbol{I}_d)$ and the KL divergence between two Gaussian distributions. $\mathbb{E}_{\boldsymbol{x} \sim p_0} \left[\boldsymbol{x}^T \boldsymbol{x} \right]$ is finite since p_0 is Sub-Gaussian. Therefore by taking $T = n^{\frac{2\beta}{2\beta+d}}$

$$\mathbb{E} \left[\mathrm{TV} \left(\bar{\boldsymbol{Y}}_{T-t_0}, \hat{\boldsymbol{Y}}_{T-t_0} \right) \right] \lesssim n^{-\frac{\beta}{2\beta+d}}.$$

F. Auxiliary Results

F.1. Sub-Gaussian Properties of $p_t(x)$

We defined the Sub-Gaussian random vectors in Definition 3.2. Sub-Gaussian random vectors in higher dimensions also retain similar properties to those of 1-dimensional Sub-Gaussian random variables. The following lemma shows that the sum of two Sub-Gaussian random vectors is still Sub-Gaussian.

Lemma F.1. *Let $X, Y \in \mathbb{R}^d$ be independent Sub-Gaussian random vectors with Sub-Gaussian norms $\|X\|_{\psi_2}$ and $\|Y\|_{\psi_2}$ respectively. Then $X + Y$ is Sub-Gaussian and*

$$\|X + Y\|_{\psi_2} \leq \|X\|_{\psi_2} + \|Y\|_{\psi_2}.$$

Proof. We first prove that for $X, Y \in \mathbb{R}$, we have

$$\|X + Y\|_{\psi_2} \leq \|X\|_{\psi_2} + \|Y\|_{\psi_2}.$$

In fact, notice that the function $\psi_2(x) = e^{x^2} - 1$ is convex and increasing. Denote $\sigma_X := \|X\|_{\psi_2}$ and $\sigma_Y := \|Y\|_{\psi_2}$, then by triangle inequality and Jensen's inequality,

$$\begin{aligned}
\psi_2\left(\frac{|X + Y|}{\sigma_X + \sigma_Y}\right) &\leq \psi_2\left(\frac{|X|}{\sigma_X + \sigma_Y} + \frac{|Y|}{\sigma_X + \sigma_Y}\right) \\
&= \psi_2\left(\frac{\sigma_X}{\sigma_X + \sigma_Y}\frac{|X|}{\sigma_X} + \frac{\sigma_Y}{\sigma_X + \sigma_Y}\frac{|Y|}{\sigma_Y}\right) \\
&\leq \frac{\sigma_X}{\sigma_X + \sigma_Y}\psi_2\left(\frac{|X|}{\sigma_X}\right) + \frac{\sigma_Y}{\sigma_X + \sigma_Y}\psi_2\left(\frac{|Y|}{\sigma_Y}\right).
\end{aligned}$$

Take expectation of both sizes, and use the definition of Orlicz norm,

$$\begin{aligned}
\mathbb{E}\left[\psi_2\left(\frac{|X + Y|}{\sigma_X + \sigma_Y}\right)\right] &\leq \frac{\sigma_X}{\sigma_X + \sigma_Y}\mathbb{E}\left[\psi_2\left(\frac{|X|}{\sigma_X}\right)\right] + \frac{\sigma_Y}{\sigma_X + \sigma_Y}\mathbb{E}\left[\psi_2\left(\frac{|Y|}{\sigma_Y}\right)\right] \\
&\leq \frac{\sigma_X}{\sigma_X + \sigma_Y} + \frac{\sigma_Y}{\sigma_X + \sigma_Y} \\
&= 1.
\end{aligned}$$

Therefore,

$$\|X + Y\|_{\psi_2} \leq \sigma_X + \sigma_Y.$$

In general case where $X, Y \in \mathbb{R}^d$, for any $v \in S^{d-1}$,

$$\begin{aligned}
\|\langle X + Y, v\rangle\|_{\psi_2} &= \|\langle X, v\rangle + \langle Y, v\rangle\|_{\psi_2} \\
&\leq \|\langle X, v\rangle\|_{\psi_2} + \|\langle Y, v\rangle\|_{\psi_2} \\
&\leq \|X\|_{\psi_2} + \|Y\|_{\psi_2}.
\end{aligned}$$

The result follows from taking the supremum with respect to $v \in S^{d-1}$ of both sizes and the definition of Orlicz norm. $\square$

The next lemma provides a tail bound for a Sub-Gaussian random vector, analogous to the one-dimensional tail bound.

Lemma F.2. *Suppose $X \in \mathbb{R}^d$ is σ-Sub-Gaussian with $\mathbb{E}X = 0$, then for any $t \in \mathbb{R}$,*

$$\mathbb{P}\left(\|X\|_\infty > t\right) \leq 2d\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Proof. Let $v = (1, 0, \ldots, 0) \in \mathbb{R}^d$.

$$\mathbb{E}\exp\left(\frac{|X_1|^2}{\sigma^2}\right) = \mathbb{E}\exp\left(\frac{|\langle X, v\rangle|^2}{\sigma^2}\right) \leq 2.$$

Then X_1 is σ-Sub-Gaussian. Similarly, X_i is σ-Sub-Gaussian for all $i = 1, \ldots, d$. Therefore,

$$
\begin{aligned}
\mathbb{P}\left(\|X\|_\infty > t\right) &\leq \sum_{i=1}^d \mathbb{P}\left(|X_i| > t\right) \\
&\leq \sum_{i=1}^d 2\, e^{-\frac{t^2}{2\sigma^2}} \\
&= 2d\, e^{-\frac{t^2}{2\sigma^2}}.
\end{aligned}
$$

$\square$

Let $X_0 \sim p_0$, $Z_t \sim \mathcal{N}(0, t\boldsymbol{I}_d)$ and $X_t = X_0 + Z_t \sim p_t$. The following results of the Sub-Gaussian property of X_t can be directly seen from Lemma F.1.

Lemma F.3. *[**Sub-Gaussian property for** $p_t(x)$] Suppose p_0 is σ_0-Sub-Gaussian, and $p_t = p_0 * \phi_t$, where ϕ_t is the density of $\mathcal{N}(0, t\boldsymbol{I}_d)$ and $*$ denotes the convolution. Then p_t is also Sub-Gaussian and for $X_t \sim p_t$, $\|X_t\|_{\psi_2} \leq \sigma_0 + \sqrt{t}$.*

Proof. Let $Z \sim \mathcal{N}(0, t\boldsymbol{I}_d)$. For any $v \in S^{d-1}$, since $\|v\| = 1$, $\langle Z, v \rangle \sim \mathcal{N}(0, t)$. Using the fact that the Sub-Gaussian norm for a one-dimensional Gaussian random variable is equal to its square root of the variance, we have $\|\langle Z, v \rangle\|_{\psi_2} = \sqrt{t}$. Therefore,

$$
\|Z\|_{\psi_2} = \sup_{v \in S^{d-1}} \|\langle Z, v \rangle\|_{\psi_2} = \sqrt{t}.
$$

If $X \sim p_0$, then $X + Z \sim p_t$. By Lemma F.1,

$$
\|X + Z\|_{\psi_2} \leq \|X\|_{\psi_2} + \|Z\|_{\psi_2} = \sigma_0 + \sqrt{t}.
$$

$\square$

F.2. Bounds on $\|D^\alpha p_t(x)\|_\infty$

The following lemma provides an upper bound for any derivatives of p_t.

Lemma F.4. *Supposed $\|p_0\|_\infty := \sup_x p_0(x) < \infty$, and let $\alpha = (\alpha_1, \ldots, \alpha_d)$ with $\ell = \sum_{i=1}^d \alpha_i$, then*

$$
\sup_x |D^\alpha p_t(x)| \leq C_1^d \|p_0\|_\infty \ell^{\frac{\ell}{2} + \frac{1}{4}} t^{-\ell/2},
$$

for some universal constant C_1.

Proof. Denote $\phi_t(x) := \frac{1}{(2\pi t)^{d/2}} \exp\left\{-\frac{\|x\|^2}{2t}\right\}$ as the Gaussian density function with variance $t\boldsymbol{I}_d$.

$$
\begin{aligned}
|D^\alpha p_t(x)| &\leq \int_y |p(y) D^\alpha \phi_t(x - y)|\,dy \\
&\leq \|p_0\|_\infty \int_x |D^\alpha \phi_t(x)|\,dx.
\end{aligned}
$$

Next, we establish an upper bound for $\int_x |D^\alpha \phi_t(x)|\,dx = \int_x |\prod_{i=1}^d \frac{\partial^{\alpha_i}}{\partial x_i^{\alpha_i}} \phi_t(x)|\,dx$ using the Hermite polynomials. The Hermite polynomials of order n are defined as follows:

$$
H_n(x) := (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2}. \tag{47}
$$

Then we have:

$$
D^\alpha \phi_t(x) = \left(-\frac{1}{\sqrt{2t}}\right)^{|\alpha|} \prod_{i=1}^d H_{\alpha_i}\left(\frac{x_i}{\sqrt{2t}}\right) \phi_t(x).
$$

33

Using Lemma F.5 for Hermite polynomials, we have

$$|D^\alpha \phi_t(x)| \le C^d t^{-\ell/2} \left(\prod_{i=1}^d \alpha_i! \right)^{1/2} e^{\|x\|^2/4t} \phi_t(x),$$

Therefore, using $\prod_{i=1}^d \alpha_i! \le (\sum_{i=1}^d \alpha_i)! = \ell!$, we have,

$$\int_x |D^\alpha \phi_t(x)| dx \le C^d t^{-\ell/2} (\ell!)^{1/2} \int_x e^{\|x\|^2/4t} \phi_t(x) dx$$

$$= C^d t^{-\ell/2} (\ell!)^{1/2} \int_x \frac{1}{(2\pi t)^{d/2}} e^{-\|x\|^2/4t} dx$$

$$= C^d 2^{d/2} (\ell!)^{1/2} t^{-\ell/2}.$$

By Stirling's approximation,

$$n! \le \sqrt{2\pi n} \left(\frac{n}{e} \right)^n e^{\frac{1}{12n}},$$

we have

$$(\ell!)^{1/2} \le (2\pi)^{1/4} e^{\frac{1-12\ell^2}{24\ell}} \ell^{\frac{\ell}{2}+\frac{1}{4}} \le (2\pi)^{1/4} \ell^{\frac{\ell}{2}+\frac{1}{4}}.$$

Therefore,

$$\int_x |\phi_t^{(\ell)}(x)| dx \le C^d 2^{d/2} (2\pi)^{1/4} \ell^{\frac{\ell}{2}+\frac{1}{4}} t^{-\ell/2}.$$

$\square$

Lemma F.5. *Indritz (1961) The Hermite polynomials defined in* (47) *satisfies*

$$H_n(x) \le C \left(2^n n! \right)^{1/2} e^{x^2/2}$$

for some universal constant C.

F.3. Concentration Inequality for $\hat{p}_t$

Lemma F.6. *There exist a constant C_4 that only depends on p_0 and dimension d such that the kernel density estimator $\hat{p}_t(x)$ with the kernel of order ℓ we defined in Lemma C.2 and the choice of $h = \frac{\sqrt{t}}{D_n}$ satisfies*

$$|\hat{p}_t(x) - p_t(x)| < C_4 \left(\sqrt{\frac{D_n^d \ell^{3d} p_t^*(x) \log(1/\delta)}{nt^{\frac{d}{2}}}} + \frac{D_n^d \ell^{5d/2} \log(1/\delta)}{nt^{\frac{d}{2}}} + \left(\frac{D_n}{de} \right)^{-\ell} \ell^{-\frac{\ell}{2}+\frac{5d}{2}-\frac{1}{4}} \right),$$

with probability at least $1 - \delta$, where $p_t^*(x) := \sup_{\|\lambda\|_\infty < h} p_t(x+\lambda)$. Furthermore, by choosing $\ell = \log n$, $D_n = C\sqrt{\log n}$ and $\delta = n^{-\alpha}$ for any positive constant α, there exists a constant $C_5(\alpha)$ that depends on p_0, d and α, with probability at least $1 - n^{-\alpha}$,

$$|\hat{p}_t(x) - p_t(x)| < C_5(\alpha) \, \text{polylog}(n) \left(\sqrt{\frac{p_t^*(x)}{nt^{\frac{d}{2}}}} + \frac{1}{nt^{\frac{d}{2}}} + (\log n)^{-\log n} \right).$$

Proof. Take $Y_i = \frac{1}{h^d} K\left(\frac{x-X_i}{h} \right)$, then

$$\left| \frac{1}{n} \sum_{i=1}^n Y_i - \mathbb{E} Y \right| = |\hat{p}_t(x) - \mathbb{E}\, \hat{p}_t(x)|,$$

Using Lemma C.2, we have

$$|Y_i| \le \frac{1}{h^d} \|K_d\|_\infty \lesssim \frac{\ell^{\frac{5d}{2}}}{h},$$

$$\mathrm{Var}\,(Y_i) = \frac{1}{h^{2d}}\mathrm{Var}\left(K_d\left(\frac{x-X_i}{h}\right)\right) \le \frac{1}{h^d}\int K_d^2(u)du \sup_{\|\lambda\|_\infty < h} p_t(x+\lambda) \lesssim \frac{\ell^{3d}}{h^d}\,p_t^*(x),$$

where $p_t^*(x) := \sup_{\|\lambda\|_\infty < h} p_t(x+\lambda)$. Next we use the following Bernstein's inequality.

Proposition F.7 (Bernstein's inequality). *Suppose that Y_i are iid with mean μ, $\mathrm{Var}\,(Y_i) \le \sigma^2$ and $\mathbb{P}\,(Y_i \le M) = 1$. Then*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Y_i - \mu\right| > \varepsilon\right) \le 2\exp\left\{-\frac{n\varepsilon^2}{2\sigma^2 + 2M\varepsilon/3}\right\}.$$

Furthermore, with probability at least $1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^n Y_i - \mu\right| < \sqrt{\frac{2\sigma^2\log(1/\delta)}{n}} + \frac{2M\log(1/\delta)}{3n}.$$

By Bernstein's inequality,

$$\mathbb{P}\left(|\hat{p}_t(x) - \mathbb{E}\,\hat{p}_t(x)| < \sqrt{\frac{2\,\ell^{3d}\,p_t^*(x)\log(1/\delta)}{nh^d}} + \frac{2\ell^{5d/2}\log(1/\delta)}{3nh^d}\right) > 1 - \delta \tag{48}$$

By triangle inequality,

$$|\hat{p}_t(x) - p_t(x)| \le |\hat{p}_t(x) - \mathbb{E}\,\hat{p}_t(x)| + |\mathbb{E}\,\hat{p}_t(x) - p_t(x)| \tag{49}$$

Using the result of bias term in Proposition C.3, and by taking $h = \frac{\sqrt{t}}{D_n}$, we have

$$|\mathbb{E}\,\hat{p}_t(x) - p_t(x)| \le C_2\left(\frac{D_n}{de}\right)^{-\ell}\ell^{-\frac{\ell}{2}+\frac{5d}{2}-\frac{1}{4}}.$$

Therefore, there is a constant C_4 such that with probability at least $1 - n^{-\alpha}$,

$$|\hat{p}_t(x) - p_t(x)| < C_4\left(\sqrt{\frac{D_n^d\,\ell^{3d}\,p_t^*(x)\log(1/\delta)}{nt^{\frac{d}{2}}}} + \frac{D_n^d\ell^{5d/2}\log(1/\delta)}{nt^{\frac{d}{2}}} + \left(\frac{D_n}{de}\right)^{-\ell}\ell^{-\frac{\ell}{2}+\frac{5d}{2}-\frac{1}{4}}\right).$$

By choosing the order $\ell = \Omega(\log n)$, $D_n = C\sqrt{\log n}$ and $\delta = n^{-\alpha}$ for some positive constant α, we have with probability at least $1 - n^{-\alpha}$,

$$|\hat{p}_t(x) - p_t(x)| < C_5(\alpha)\,\mathrm{polylog}(n)\left(\sqrt{\frac{p_t^*(x)}{nt^{\frac{d}{2}}}} + \frac{1}{nt^{\frac{d}{2}}} + (\log n)^{-\log n}\right).$$

$\square$

F.4. Bounds on the Tail Density

In this section, we provide the proofs for the bound of $|G_1|$ and $\int_{G_1}\|s_t(x)\|^2 dx$ in the proofs in Lemma D.2, where $G = \{x\colon p_t(x) > \rho_n\log^c n\}$ and $\rho_n = \frac{1}{n\,t^{d/2}}$. To begin with, the following lemma shows that for sub-Gaussian density function p_t, for sufficiently large x satisfying $\|x\|_\infty \gtrsim \sqrt{\log n}$, the tail density can be controlled. Specifically, we show that $p_t(x)$ is bounded above by a polynomial decay in n^{-1}: $p_t(x) \lesssim \mathrm{poly}(n^{-1})$. This lemma is crucial for the proof of upper bounds on $|G_1|$ and $\int_{G_1}\|s_t(x)\|^2 dx$.

Lemma F.8. *For any $D > 0$ and $t > n^{-T_1}$, if $x \in \mathbb{R}^d$ satisfies $\|x\|_\infty \ge D\sigma_t\sqrt{\log n}$ and $n \ge \exp\left(\frac{\|p_0\|_\infty^2}{4D^2}\right)$, where $\sigma_t = \sigma_0 + \sqrt{t}$, then*

$$p_t(x) \le 2\sqrt{2}n^{-\frac{D}{2}+\frac{dT_1}{2}}$$

Proof. **Step 1**: Suppose $d = 1$. From Lemma F.3, $X_t \sim p_t(x)$ is Sub-Gaussian with Orlicz norm $\|X_t\|_{\psi_2} \leq \sigma_t = \sigma_0 + \sqrt{t}$. Then by the sub-Gaussian property in Definition 3.2,

$$\int e^{\frac{x^2}{\sigma_t^2}} p_t(x) dx \leq 2.$$

Consider two cases. If $t \geq 1$, fix x_0 that satisfies $|x_0| > D\sigma_t \sqrt{\log n}$. Define the following two sets:

$$\mathcal{A} := \left\{ x \colon |x - x_0| < \frac{D\sigma_t}{2} \sqrt{\log n} \right\},$$

$$\mathcal{B} := \left\{ x \colon |x - x_0| < \frac{p_t(x_0)\sqrt{t}}{4} \right\},$$

If $n > \exp\left(\frac{\|p_0\|_\infty^2}{4D^2} \right)$, then $\frac{D\sigma_t}{2} \sqrt{\log n} \geq \frac{D\sigma_t \|p_0\|_\infty}{4D} \geq \frac{p_t(x_0)\sqrt{t}}{4}$, where the last inequality is due to the fact that

$$p_t(x) = \int p(y)\phi_t(x - y) dy \leq \|p_0\|_\infty \qquad \text{for any } x.$$

Therefore $\mathcal{B} \subset \mathcal{A}$. Next, we use the Lipschitz property of p_t to control $p_t(x)$ for all $x \in \mathcal{B}$. Specifically, for any $x < x'$,

$$\begin{aligned}
|p_t(x) - p_t(x')| &\leq \int p_0(y)|\phi_t(x - y) - \phi_t(x' - y)| dy \\
&\leq |x - x'| \sup_x |\phi_t'(x)| \\
&\leq \frac{1}{t}|x - x'|,
\end{aligned} \tag{50}$$

where in the second inequality we use the mean value theorem, and in the third inequality we use the fact that

$$\sup_x |\phi_t'(x)| = \frac{1}{t\sqrt{2\pi t}} \sup_x \left\{ x \exp\left(-\frac{x^2}{2t} \right) \right\} = \frac{e^{-1/2}}{\sqrt{2\pi}} t^{-1}.$$

Then for any $y \in \mathcal{B}$, using (50), since $t > 1$, we have

$$\begin{aligned}
p_t(y) &\geq p_t(x_0) - \frac{1}{t}|x_0 - y| \\
&\geq p_t(x_0) - \frac{1}{\sqrt{t}}|x_0 - y| \\
&\geq p_t(x_0) - \frac{1}{\sqrt{t}}\frac{p_t(x_0)\sqrt{t}}{2} \\
&= \frac{1}{2}p_t(x_0).
\end{aligned} \tag{51}$$

Then using the definition of Sub-Gaussian and (51),

$$\begin{aligned}
2 &\geq \int e^{x^2/\sigma_t^2} p_t(x) \, dx \geq \inf_{x \in \mathcal{A}} e^{x^2/\sigma_t^2} \int_{\mathcal{A}} p_t(x) \, dx \\
&\geq e^{\frac{D^2}{4}\log n} \int_{\mathcal{B}} p_t(x) \, dx \\
&\geq n^{D^2/4} |\mathcal{B}| \frac{1}{2}p_t(x_0) \\
&= \frac{1}{4} n^{D^2/4} p_t^2(x_0)\sqrt{t}.
\end{aligned}$$

This implies that

$$p_t(x_0) \leq 2\sqrt{2}n^{-D/2}t^{-1/4} \leq 2\sqrt{2}n^{-D/2}.$$

If $n^{-\frac{2}{2\beta+d}} < t < 1$, define

$$\mathcal{C} := \left\{ x \colon |x - x_0| < \frac{p_t(x_0)t}{4} \right\}.$$

Similarly, if $n > \exp\left(\frac{\|p_0\|_\infty^2}{4D^2}\right)$, then $\frac{D\sigma_t}{2}\sqrt{\log n} \geq \frac{D\sigma_t\|p_0\|_\infty}{4D} \geq \frac{p_t(x_0)\sqrt{t}}{4} \geq \frac{p_t(x_0)\,t}{4}$ since $t < 1$. Therefore $\mathcal{C} \subset \mathcal{A}$. Then for any $y \in \mathcal{C}$, using (50),

$$
\begin{aligned}
p_t(y) &\geq p_t(x_0) - \frac{1}{t}|x_0 - y| \\
&\geq p_t(x_0) - \frac{1}{t}\frac{p_t(x_0)t}{2} \\
&= \frac{1}{2}p_t(x_0).
\end{aligned}
\tag{52}
$$

Then using the definition of Sub-Gaussian and (52),

$$
\begin{aligned}
2 \geq \int e^{x^2/\sigma_t^2} p_t(x)\,dx &\geq \inf_{x \in \mathcal{A}} e^{x^2/\sigma_t^2} \int_{\mathcal{A}} p_t(x)\,dx \\
&\geq e^{\frac{D^2}{4}\log n} \int_{\mathcal{C}} p_t(x)\,dx \\
&\geq n^{D^2/4}\,|\mathcal{C}|\,\frac{1}{2}p_t(x_0) \\
&= \frac{1}{4}\,n^{D^2/4}p_t^2(x_0)t.
\end{aligned}
$$

This implies that

$$p_t(x_0) \leq 2\sqrt{2}n^{-D/2}t^{-1/2} \leq 2\sqrt{2}n^{-\frac{D}{2}+\frac{T_1}{2}}.$$

Step 2: In general case $d \geq 1$, since $\|x\|_\infty \geq D\sigma_t\sqrt{\log n}$, without loss of generality, we suppose $|x_1| \geq D\sigma_t\sqrt{\log n}$. Then by definition of $p_t(x)$,

$$
\begin{aligned}
p_t(x) = \int_{y \in \mathbb{R}^d} p_0(y) \prod_{i=1}^{d} \phi_t(x_i - y_i)dy &\leq (2\pi t)^{-\frac{d-1}{2}} \int_{y \in \mathbb{R}} p_0(y)\phi_t(x_1 - y)dy \\
&= (2\pi t)^{-\frac{d-1}{2}} p_t(x_1).
\end{aligned}
$$

Using the 1-dimensional results, $p_t(x_1) \leq 2\sqrt{2}\,n^{-\frac{D}{2}+\frac{1}{2\beta+d}}\mathbb{1}_{t<1}$, and use the fact that $t > n^{-T_1}$, we have

$$p_t(x) \leq 2\sqrt{2}n^{-\frac{D}{2}+\frac{dT_1}{2}}.$$

$\square$

Lemma F.9. *Let* $G = \{x \colon p_t(x) > \rho_n \log^c n\}$ *where* $\rho_n = \frac{1}{n\,t^{d/2}}$ *and* $n^{-T_1} < t < n^{T_2}$,

$$|G| \leq D(t^{\frac{d}{2}} + \sigma_0^d)(\log n)^{\frac{d}{2}}$$

for some constant D only depending on d, T_1 and T_2.

Proof. For any $t \geq n^{-T_1}$ and constant $C > 0$. Let $D = 2C + 2dT_1$. Then by Lemma F.8, for any $n \geq \exp\left(\frac{\|p_0\|_\infty^2}{4D^2}\right)$,

$$p_t(x) \geq 2\sqrt{2}\,n^{-C}$$

implies

$$\|x\|_\infty \le D\,\sigma_t\,\sqrt{\log n} \le D\,(\sqrt{t}+\sigma_0)\,\sqrt{\log n}.$$

Since $\rho_n = n^{-1}t^{-d/2} \ge n^{-(1+\frac{dT_2}{2})}$, we can take $C = 1 + \frac{dT_2}{2}$ and conclude that

$$
\begin{aligned}
|G| &\le |\{x : \|x\|_\infty \le D\,(\sqrt{t}+\sigma_0)\,\sqrt{\log n}\}| \\
&\lesssim (\sqrt{t}+\sigma_0)^d (\log n)^{\frac{d}{2}} \\
&\le 2^{d-1}(t^{\frac{d}{2}} + \sigma_0^d)(\log n)^{\frac{d}{2}}.
\end{aligned}
$$

$\square$

Then using the bound for expectation of moment of score function $\mathbb{E}\left[\|s_t(X)\|^m\right]$ in Lemma F.16 and combining Theorem F.9, we can further provide an upper bound for

$$\int_G \|s_t(x)\|^2 dx. \tag{53}$$

Lemma F.10. *For any $m > 1$, define the set*

$$G := \{x : p_t(x) > \rho_n \, \log^c n\},$$

then

$$\int_G \|s_t(x)\|^2 dx \le C_6\, m\,(\log n)^{-\frac{c}{m}} \rho_n^{-\frac{1}{m}}\, t^{-1}\,(t^{\frac{d}{2}} + \sigma_0^d)\,(\log n)^{\frac{d}{2}},$$

for some constant C_6 depends on D_1 and p.

Proof. By Hölder's inequality,

$$
\begin{aligned}
\int_G \|s_t(x)\|^2 dx &= \int_G \|s_t(x)\|^2 \frac{1}{p_t(x)} p_t(x) dx \\
&\le \left(\int_G \|s_t(x)\|^{2m} p_t(x) dx\right)^{1/m} \left(\int_G p_t(x)^{-m'} p_t(x) dx\right)^{1/m'},
\end{aligned}
$$

where $m, m' > 1$ and satisfies $\frac{1}{m} + \frac{1}{m'} = 1$. By Lemma F.16,

$$\left(\int_G \|s_t(x)\|^{2m} p_t(x) dx\right)^{1/m} = \left(\mathbb{E}\left[\|s_t(X)\|^{2m}\right]\right)^{1/m} \le t^{-1}\,((2m-1)!!)^{1/m} \lesssim m\,t^{-1},$$

where in the last inequality we use Stirling's approximation. By Theorem F.9,

$$
\begin{aligned}
\left(\int_G p_t(x)^{-m'} p_t(x) dx\right)^{1/m'} &= \left(\int_G p_t(x)^{1-m'} dx\right)^{1/m'} \\
&\le \left(|G|(\log^c n)^{1-m'} \rho_n^{1-m'}\right)^{1/m'} \\
&\le \left(D_1(t^{\frac{d}{2}} + \sigma_0^d)(\log n)^{\frac{d}{2}}\right)^{1-1/m} (\log^c n)^{-\frac{1}{m}} \rho_n^{-\frac{1}{m}} \\
&\le D_1(t^{\frac{d}{2}} + \sigma_0^d)(\log n)^{\frac{d}{2}} (\log n)^{-\frac{c}{m}} \rho_n^{-\frac{1}{m}}.
\end{aligned}
$$

$\square$

38

F.5. Bounds on the Squared Error

The next proposition provides uniform upper bounds over $x \in \mathbb{R}$ for

$$\|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2$$

and

$$\|\nabla \hat{p}_t(x)\|^2 |\hat{p}_t(x) - p_t(x)|^2 .$$

Lemma F.11. *There exists constants C_5 depending on d and p_0 such that with probability 1,*

$$\sup_x \|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2 \leq C_5 \left(\|p_0\|_\infty^2 t^{-1} + \ell^{5d} h^{-(2d+2)} \right),$$

$$\sup_x \|\nabla p_t(x)\|^2 |\hat{p}_t(x) - p_t(x)|^2 \leq C_5 \left(\|p_0\|_\infty^2 t^{-1} (\ell^{5d} h^{-2d} + 2\|p_0\|_\infty^2) \right).$$

In particular, if $h = \frac{\sqrt{t}}{D_n}$, $D_n = C\sqrt{\log n}$ and $\ell = \log n$, then the two terms above are $O(t^{-2}\mathrm{poly}(\log n))$. Furthermore, if $\log \frac{1}{t} = O(\log n)$, then the two terms above are at most polynomial in n.

Proof. By Lemma F.4,

$$\|\nabla p_t(x)\|^2 = \sum_{i=1}^d \left| \frac{\partial}{\partial x_d} p_t(x) \right|^2 \leq d\, C_1^{2d} \|p\|_\infty^2 t^{-1}.$$

For the derivative of kernel density estimation,

$$\nabla \hat{p}_t(x) = \frac{1}{nh^{d+1}} \sum_{i=1}^n \nabla K_d \left(\frac{x - X_i^t}{h} \right),$$

By Lemma C.2,

$$\|\nabla \hat{p}_t(x)\|^2 \leq \frac{d\|\nabla K_d(\cdot)_1\|_\infty^2}{h^{2d+2}} \lesssim \frac{d\ell^{5d}}{h^{2d+2}}.$$

Therefore

$$\|\nabla \hat{p}_t(x) - \nabla p_t(x)\|^2 \leq 2d\, C_1^{2d} \|p\|_\infty^2 t^{-1} + 2C_2 \frac{d\ell^{5d}}{h^{2d+2}}.$$

Besides, since

$$\hat{p}_t(x) = \frac{1}{nh^d} \sum_{i=1}^n K_d \left(\frac{x - X_i^t}{h} \right),$$

we have

$$|\hat{p}_t(x) - p_t(x)|^2 \leq 2\|K\|_\infty^2 h^{-2s} + 2\|p\|_\infty^2 \lesssim \ell^{5d} h^{-2d} + 2\|p\|_\infty^2.$$

$\square$

Moreover, the fourth moments of the score error $\mathbb{E}_{x \sim p_t, \{x_i\}_{i=1}^n} \|\hat{s}_t(x) - s_t(x)\|^4$ can also be bounded.

Lemma F.12. *There exists some constants C_6 depending on d and p_0 such that*

$$\mathbb{E}_{X \sim p_t, \{x_i\}_{i=1}^n} \|\hat{s}_t(X) - s_t(X)\|^4 \leq C_6 \left(t^{-2} + \rho_n^{-4} h^{-4(d+1)} \right).$$

In particular, if $h = \frac{\sqrt{t}}{D_n}$, $D_n = \sqrt{\log n}$, $\ell = \log n$ and $\log \frac{1}{t} = O(\log n)$, then the right side is at most polynomial in n.

Proof.

$$\mathbb{E}_{X \sim p_t, \{x_i\}_{i=1}^n} \|\hat{s}_t(X) - s_t(X)\|^4 \lesssim \mathbb{E}_{X \sim p_t, \{x_i\}_{i=1}^n} \|\hat{s}_t(X)\|^4 + \mathbb{E}_{X \sim p_t} \|s_t(X)\|^4.$$

Lemma F.16 shows that the moment of score is bounded:

$$\mathbb{E}_{X \sim p_t} \|s_t(X)\|^4 \lesssim t^{-2} d^2.$$

Using the definition of $\hat{s}_t(x) = \frac{\nabla \hat{p}_t(x)}{\hat{p}_t(x)} \mathbb{1}_{\hat{p}_t(x) > \rho_n}$ and Lemma C.2,

$$\mathbb{E}_{X \sim p_t, \{x_i\}_{i=1}^n} \|\hat{s}_t(X)\|^4 \leq \rho_n^{-4} \mathbb{E}_{X \sim p_t, \{x_i\}_{i=1}^n} \|\nabla \hat{p}_t(X)\|^4$$

$$\leq \rho_n^{-4} \left\| \frac{1}{nh^{d+1}} \sum_{i=1}^d \nabla K_d \left(\frac{x - X_i^t}{h} \right) \right\|^4$$

$$\lesssim \rho_n^{-4} h^{-4(d+1)} \ell^{5d}.$$

$\square$

F.6. Control of the Score moments via the Rényi Entropies

F.6.1. RÉNYI ENTROPIES FOR SUB-GAUSSIAN RANDOM VECTOR

Definition F.13. The α-Rényi entropy for random vector X with distribution p is defined as

$$h_\alpha(X) := \frac{1}{1 - \alpha} \log \int (p(x))^\alpha \, dx.$$

The next proposition shows that sub-Gaussian random variables have finite Rényi entropy and that the maximum Rényi entropy is attainable by the Gaussian distribution.

Lemma F.14. *Suppose that $X \in \mathbb{R}$ is σ-Sub-Gaussian random variable. Let $\alpha \in (0, 1)$ and p denotes the density of X. Then*

$$h_\alpha(X) \leq \frac{1}{1 - \alpha} \log \left((2\pi\sigma^2)^{\frac{1-\alpha}{2}} (1 - \alpha)^{\frac{1-\alpha}{2}} \alpha^{-\frac{1}{2}} \right).$$

Proof. By the sub-Gaussian property in 1 dimension, X satisfies

$$\mathbb{E}[e^{X^2/\sigma^2}] \leq 2. \tag{54}$$

The map $p(\cdot) \mapsto \int p^\alpha dx$ is concave for $\alpha \in (0, 1)$, so the maximum of $h_\alpha(X)$ under the given Sub-Gaussian constraint (54) is achieved at the stationary point via calculus of variations. Namely, define the Lagrangian with Lagrange multipliers $\eta > 0$:

$$\mathcal{L}(p(x), \eta) = \int p(x)^\alpha \, dx + \eta \left(2 - \int p(x) \exp \left(\frac{x^2}{\sigma^2} \right) \, dx \right).$$

The functional derivative of $\mathcal{L}$ with respect to $p(x)$ is:

$$\frac{\delta \mathcal{L}}{\delta p(x)} = \alpha p(x)^{\alpha-1} - \eta \exp \left(\frac{x^2}{\sigma^2} \right) = 0.$$

Therefore the maximum of $h_\alpha(X)$ is attained by

$$\tilde{p}(x) = Z e^{-\frac{x^2}{(1-\alpha)\sigma^2}}$$

for some normalizing constant $Z > 0$. As p is a probability measure, we can verify that $Z = \frac{1}{\sqrt{\pi(1-\alpha)}\sigma}$. Therefore, for all σ-sub-Gaussian random variable X,

$$h_\alpha(X) \leq \frac{1}{1 - \alpha} \log \int (\tilde{p}(x))^\alpha \, dx = \frac{1}{1 - \alpha} \log \left((2\pi\sigma^2)^{\frac{1-\alpha}{2}} (1 - \alpha)^{\frac{1-\alpha}{2}} \alpha^{-\frac{1}{2}} \right).$$

$\square$

Next, we generalize this result to d-dimension σ-sub-Gaussian random vector X using the following property of Rényi entropy.

Lemma F.15. *Suppose that $X \in \mathbb{R}^d$ is σ-Sub-Gaussian random vector. Let $\alpha \in (0,1)$ and p denotes the density of X. Then for $\alpha \to 0$, we have*

$$h_\alpha(X) \le \log\left(C_d \sigma^d \alpha^{-\frac{d}{2}}\right)$$

for some constant C_d that only depends on the dimension d.

Proof. We define a transformation $T : \mathbb{R}^d \to S^{d-1} \times \mathbb{R}^+$ by $T(x) = (u, r)$, where $u = x/\|x\|$ and $r = \|x\|$. This transformation converts Cartesian coordinates to spherical coordinates, with u representing the direction and r the radius. The determinant of the Jacobian of this transformation is $|J_{T^{-1}}(r, u)| = r^{d-1}$. Therefore, by change of variable, the joint distribution of $(U, R) = (\frac{X}{\|X\|}, \|X\|)$ is given by

$$p_{U,R}(u, r) = p_X(T^{-1}(u, r))|J_{T^{-1}}(r, u)| = p_X(ur)r^{d-1}.$$

Next, we formulate the Rényi entropy of X in terms of the joint distribution of $(X/\|X\|, \|X\|)$:

$$
\begin{aligned}
h_\alpha(X) &= \log\left(\mathbb{E}\left[\left(\frac{1}{p_X(X)}\right)^{1-\alpha}\right]\right)^{\frac{1}{1-\alpha}} \\
&= \log\left(\mathbb{E}\left[\left(\frac{\|X\|^{d-1}}{p_{U,R}(X/\|X\|, \|X\|)}\right)^{1-\alpha}\right]\right)^{\frac{1}{1-\alpha}} \\
&\le \log\left(\mathbb{E}\left[\left(\frac{1}{p_{U,R}(X/\|X\|, \|X\|)}\right)^{p}\right]\right)^{\frac{1}{p}} + \log\left(\mathbb{E}\left[\left(\|X\|^{(d-1)}\right)^q\right]\right)^{\frac{1}{q}} \\
&= h_{1-p}\left(\frac{X}{\|X\|}, \|X\|\right) + \frac{1}{q}\log \mathbb{E}\left[\|X\|^{q(d-1)}\right],
\end{aligned}
$$

where in the inequality term we apply the Hölder's inequality with $\frac{1}{1-\alpha} = \frac{1}{p} + \frac{1}{q}$ for $0 < p < 1$ and $q > 1$. Combining Proposition 4.9 and Proposition 4.12 in Berens (2013), we obtain an Rényi Entropy upper bound for the join distribution $\left(\frac{X}{\|X\|}, \|X\|\right)$

$$h_{1-p}\left(\frac{X}{\|X\|}, \|X\|\right) \le h_{1-p}(\|X\|) + h_0\left(\frac{X}{\|X\|}\right).$$

The last term is bounded by a constant $\log C_d$ that only depends on the dimension because the direction vector $\frac{X}{\|X\|}$ has bounded supported on a unit sphere S^{d-1} (Van Erven & Harremos, 2014). Notice that $\|X\| \in \mathbb{R}$ is $\sqrt{d}\sigma$-sub-Gaussian since by Hölder's inequality,

$$\mathbb{E}\left[\exp\left(\frac{\|X\|^2}{d\sigma^2}\right)\right] \le \prod_{i=1}^{d}\left(\mathbb{E}\left[\exp\left(\frac{X_i^2}{\sigma^2}\right)\right]\right)^{\frac{1}{d}} \le 2.$$

Therefore, using the result of one dimension in Lemma F.14,

$$h_{1-p}(\|X\|) \le \frac{1}{p}\log\left((2\pi d\sigma^2)^{\frac{p}{2}} p^{\frac{p}{2}}(1-p)^{-\frac{1}{2}}\right).$$

The moments of $\|X\|$ can also be bounded since it is sub-Gaussian:

$$
\begin{aligned}
\mathbb{E}\left[\|X\|^{q(d-1)}\right] &= \int_0^\infty \mathbb{P}\left(\|X\|^{q(d-1)} > u\right) du \\
&= q(d-1) \int_0^\infty \mathbb{P}\left(\|X\| > t\right) t^{q(d-1)-1} dt \\
&\leq q(d-1) \int_0^\infty 2\exp\left(-\frac{t^2}{2d\sigma^2}\right) t^{q(d-1)-1} dt \\
&= q(d-1)(2d\sigma^2)^{\frac{q(d-1)}{2}} \Gamma\left(\frac{q(d-1)}{2}\right) \\
&\leq 3q(d-1)(2d\sigma^2)^{\frac{q(d-1)}{2}} \left(\frac{q(d-1)}{2}\right)^{\frac{q(d-1)}{2}}
\end{aligned}
$$

where in the last inequality we use $\Gamma(x) \leq 3x^x$ for all $x \geq 1/2$.

Combining these results and take $p = 1 - \frac{\alpha}{2}$ and $q = \frac{2}{\alpha} + \alpha - 3$, when α is small we have

$$
h_\alpha(X) \leq \log\left(\tilde{C}_d \sigma^d \alpha^{-\frac{d}{2}}\right)
$$

for some constant $\tilde{C}_d$ that only depends on the dimension d. $\qquad \square$

F.6.2. CONTROLLING THE EXPECTATION OF SCORE FUNCTION WITHIN LOWER DENSITY AREAS

For any random vector, we denote $I_m(X) := \mathbb{E}[\|s_X(X)\|^m]$, where $s_X(x) := \nabla \log p_X(x)$. For any random vector X, we can use the following proposition to control any order of moments of the score function $I_m(X)$.

Lemma F.16. *Let X and G be independent random vectors such that $X \sim p_0$, $G \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $Y = X + \sqrt{t}G$, then we have*

$$
\begin{aligned}
I_m(Y) &\leq \min\{I_m(X), I_m(\sqrt{t}G)\} \\
&\leq I_m(\sqrt{t}G) \\
&\leq t^{-\frac{m}{2}} d^{\frac{m}{2}} (m-1)!!
\end{aligned}
$$

Proof. The proof follows from a simple convexity argument of Corollary 3.2 in Bobkov (2019).
According to Bobkov (2019)[Corollary 3.2], for all independent random variables X, Y, $I_k(X+Y) \leq \min\{I_k(X), I_k(Y)\}$. We can generalize this result to higher dimensions. Here we have

$$
\begin{aligned}
I_m(X) &= \int \|s_X(x)\|^m p_X(x) dx \\
&= \int \frac{\|\nabla p(x)\|^m}{p^{m-1}(x)} dx
\end{aligned}
$$

where the items are vectors instead of scalars. In Bobkov (2019), follows the fact that the homogeneous function $R(u,v) = u^k/v^{k-1}$ is convex on the plane $u \in \mathbb{R}, v \in \mathbb{R}$. We need to extend this to the convexity of $(u,v) \longmapsto \frac{\|u\|^m}{v^{m-1}}$ where u is vector from $\mathbb{R}^d$ vector space. In fact, as in the constraints in classical convex optimization problems, we can check that the Hessian matrix for u and v are positive semi-definite:

$$
\begin{aligned}
\det\left(\nabla^2 \frac{\|u\|^m}{v^{m-1}}\right) &= \det\left(\begin{array}{cc} (m-2)(u \otimes u) + I & -(m-1)u^\top \\ -(m-1)u & m-1 \end{array}\right) \cdot \frac{\|u\|^{2m-3}}{v^{2m}} \cdot m^2 \\
&\geq 0.
\end{aligned}
$$

Therefore, the function $(u,v) \longmapsto \frac{\|u\|^m}{v^{m-1}}$ is still convex within the $u \in \mathbb{R}^d, v \in \mathbb{R}$.

Next, we calculate the moment of scores for Gaussian distribution $I_m(\sqrt{t}G)$. Let $\sqrt{t}G \sim \phi_t(x) := (2\pi t)^{-\frac{d}{2}} \exp\left(-\frac{\|x\|^2}{2t}\right)$ and $Z \sim \mathcal{N}(0,1)$,

$$
\begin{aligned}
I_m(\sqrt{t}G) &= \mathbb{E}_{X\sim\phi_t}\|\nabla \log \phi_t(X)\|^m \\
&= t^{-\frac{m}{2}} \mathbb{E}_{G\sim\phi_1}\|G\|^m \\
&\leq t^{-\frac{m}{2}} d^{\frac{m}{2}} \mathbb{E}|Z|^m \\
&\leq t^{-\frac{m}{2}} d^{\frac{m}{2}} (m-1)!!.
\end{aligned}
$$

$\square$

Lemma F.17. *Suppose that $X \sim p_t$ is σ_t^2-Sub-Gaussian, and $s_t(x) := \frac{\nabla p_t(x)}{p_t(x)}$ denoted the score function for $p_t(x)$. Then for any $0 < \varepsilon < 1$,*

$$
\int_{p_t(x)\leq\rho_n \log^{-c} n} \|s_t(x)\|^2 p_t(x)dx \lesssim \varepsilon^{-(1+\frac{d}{2})} t^{-1} \rho_n^{1-\varepsilon} \sigma^{d(1-\varepsilon)},
$$

where $\rho_n = n^{-1}t^{-d/2}$.

Proof. By Hölder's inequality,

$$
\begin{aligned}
\int_{p_t(x)\leq\rho_n \log^{-c} n} \|s_t(x)\|^2 p_t(x)dx &= \mathbb{E}[\|s_t(X)\|^2 \mathbb{1}_{\{p_t(x)\leq\rho_n \log^{-c} n\}}(X)] \\
&\leq (\mathbb{E}[\|s_t(X)\|^m])^{2/m} \left(\mathbb{P}[p(X) \leq \rho_n \log^{-c} n]\right)^{1-\frac{2}{m}}.
\end{aligned}
$$

The first term is the moments of score and can be upper bounded by Lemma F.16:

$$
\begin{aligned}
(\mathbb{E}[\|s_t(X)\|^m])^{2/m} &\leq (\mathbb{E}[\|s_t(X)\|^m])^{2/m} \\
&\leq t^{-1}d((m-1)!!)^{2/m} \\
&\lesssim m\,t^{-1},
\end{aligned}
$$

where in the last inequality we use Stirling's approximation. By Markov's inequality and the Upper bound for Rényi entropies in Lemma F.15,

$$
\begin{aligned}
\mathbb{P}[p(X) \leq \rho_n \log^{-c} n] &= \mathbb{P}\left[\frac{1}{p^{1-\alpha}(X)} \geq \frac{1}{\rho_n^{1-\alpha}(\log n)^{-(1-\alpha)c}}\right] \\
&\leq \rho_n^{1-\alpha}(\log n)^{-(1-\alpha)c} \exp((1-\alpha)h_\alpha(Y)) \\
&\leq C_d \rho_n^{1-\alpha} \sigma^{d(1-\alpha)} \alpha^{-\frac{d}{2}}.
\end{aligned}
$$

So combining the results we have

$$
\int_{p_t(x)\leq\rho_n \log^{-c} n} \|s_t(x)\|^2 p_t(x)dx \lesssim mt^{-1} \rho_n^{(1-\alpha)(1-2/m)} \sigma^{d(1-\alpha)(1-2/m)} \alpha^{-\frac{d(1-2/m)}{2}}.
$$

Finally, we take $\alpha = \varepsilon \downarrow 0$, $m = \varepsilon^{-1} \to \infty$ such that $(1-\alpha)(1-\frac{2}{m}) \asymp 1-\varepsilon$, then

$$
\int_{p_t(x)\leq\rho_n \log^{-c} n} \|s_t(x)\|^2 p_t(x)dx \lesssim \varepsilon^{-(1+\frac{d}{2})} t^{-1} \rho_n^{1-\varepsilon} \sigma^{d(1-\varepsilon)}.
$$

$\square$

F.7. Bounds on the Maximal Function

In the MSE analysis for $\hat{p}_t$ and $\nabla \hat{p}_t$ in Proposition C.3 and Proposition C.4, we introduce the local maximal function $p_t^*(x) := \sup_{\|\lambda\|_\infty < h} p_t(x + \lambda)$. The following lemma shows that $p_t^*(x)$ and $p_t(x)$ are uniformly close to each other for all x. Specifically, we demonstrate that if we take $h = \frac{\sqrt{t}}{D_n}$, then for any small ε, the ratio $\frac{p_t^*(x)}{p_t^{1-\varepsilon}(x)}$ is bounded by a constant that only depends on ε for all x and t.

Lemma F.18. *Define*

$$p_t^*(x) = \sup_{\|\lambda\|_\infty \leq h} p_t(x + \lambda).$$

If we take $h = \frac{\sqrt{t}}{D_n}$, for some D_n that depends on sample size n, then for any $t > 0$, $x \in \mathbb{R}^d$, and $\varepsilon \in (0, 1)$,

$$p_t(x) \leq p_t^*(x) \leq (1 - \varepsilon)^{-d/2} \exp\left\{ \frac{d(1 - \varepsilon)}{2D_n^2 \varepsilon} \right\} p_t^{1-\varepsilon}(x).$$

In particular, by taking $D_n = C\sqrt{\log n}$ for some constant C we have

$$p_t(x) \leq p_t^*(x) \leq (1 - \varepsilon)^{-d/2} \exp\left\{ \frac{d(1 - \varepsilon)}{2C^2 (\log n)\varepsilon} \right\} p_t^{1-\varepsilon}(x).$$

Remark F.19. If we take $\varepsilon \to 0$ such that $\epsilon \log n \to +\infty$, we have $p_t(x) \leq p_t^*(x) \lesssim p_t^{1-\varepsilon}(x)$.

Proof of Lemma F.18. The first inequality follows from the definition of $p_t^*(x)$. Next we prove the second inequality. By Jensen's inequality,

$$p_t^{1-\varepsilon}(x) = (p_0 * \phi_t)^{1-\varepsilon}(x) \geq p_0 * \phi_t^{1-\varepsilon}(x) \geq p_0 * \phi_{\frac{t}{1-\varepsilon}}(x),$$

where $\phi_t(x) = \exp\left(-\frac{\|x\|_2^2}{2t} \right)$. Then

$$\frac{p_t^*(x)}{p_t^{1-\varepsilon}(x)} \leq \sup_{\substack{x \in \mathbb{R}^d \\ \|\lambda\|_\infty \leq \frac{\sqrt{t}}{D_n}}} \frac{p_0 * \phi_t(x + \lambda)}{p_0 * \phi_{\frac{t}{1-\varepsilon}}(x)}$$

By changing of variable and Lemma F.20,

$$\frac{p_t^*(x)}{p_t^{1-\varepsilon}(x)} \leq \sup_{\substack{x \in \mathbb{R}^d \\ \|\lambda\|_\infty \leq D_n^{-1}}} \frac{\int p_0(x + \sqrt{t}\lambda - \sqrt{t}y)\phi_1(y)dy}{\int p_0(x - \sqrt{t}y)\phi_{\frac{1}{1-\varepsilon}}(y)dy}$$

$$= \sup_{\substack{x \in \mathbb{R}^d \\ \|\lambda\|_\infty \leq D_n^{-1}}} \frac{\int p_0(y)\phi_1\left(\lambda + \frac{x-y}{\sqrt{t}}\right)dy}{\int p_0(y)\phi_{\frac{1}{1-\varepsilon}}\left(\frac{x-y}{\sqrt{t}}\right)dy}$$

$$\leq \sup_{\substack{x \in \mathbb{R}^d \\ \|\lambda\|_\infty \leq D_n^{-1}}} \frac{\phi_1(\lambda + x)}{\phi_{\frac{1}{1-\varepsilon}}(x)}$$

$$= (1 - \varepsilon)^{-d/2} \sup_{\substack{x \in \mathbb{R}^d \\ \|\lambda\|_\infty \leq D_n^{-1}}} \exp\left\{ \frac{1}{2}\left[(1 - \varepsilon)\|x\|^2 - \|x + \lambda\|^2 \right] \right\}.$$

The exponent above is

$$\sup_{\substack{x \in \mathbb{R}^d \\ \|\lambda\|_\infty \leq D_n^{-1}}} \left((1 - \varepsilon)\|x\|^2 - \|x + \lambda\|^2 \right) = \sup_{\substack{x \in \mathbb{R}^d \\ \|\lambda\|_\infty \leq D_n^{-1}}} \left(-\varepsilon\|x + \lambda/\varepsilon\|^2 + (1/\varepsilon - 1)\|\lambda\|^2 \right)$$

$$= \sup_{\|\lambda\|_\infty \leq D_n^{-1}} (1/\varepsilon - 1)\|\lambda\|^2$$

$$= d(1/\varepsilon - 1)D_n^{-2}.$$

Therefore, denote $D_\varepsilon := \frac{p_t^*(x)}{p_t^{1-\varepsilon}(x)}$, then for all $x \in \mathbb{R}^d$

$$D_\varepsilon \leq (1-\varepsilon)^{-d/2} \exp\left\{ \frac{d(1-\varepsilon)}{2D_n^2 \varepsilon} \right\}.$$

$\square$

Lemma F.20. *Suppose that $f(x)$, $g(x)$ and $p(x)$ are probability density functions. Suppose that the density ratio of f and g is bounded, i.e.,*

$$\sup_x \frac{f(x)}{g(x)} < \infty.$$

Then

$$\frac{\int_x f(x)p(x)\,dx}{\int_x g(x)p(x)\,dx} \leq \sup_x \frac{f(x)}{g(x)}.$$

Proof. Denote

$$\lambda = \sup_x \frac{f(x)}{g(x)} < \infty.$$

Then

$$f(x) \leq \lambda g(x).$$

Integrate both sizes we have

$$\int f(x)p(x)\,dx \leq \lambda \int g(x)p(x)\,dx.$$

$\square$