Optimal score estimation via empirical Bayes smoothing

Andre Wibisono ANDRE.WIBISONO@YALE.EDU
Department of Computer Science, Yale University

Yihong Wu YIHONG.WU@YALE.EDU
Department of Statistics and Data Science, Yale University

Kaylee Yingxi Yang YINGXI.YANG@YALE.EDU
Department of Statistics and Data Science, Yale University

Editors: Shipra Agrawal and Aaron Roth

Abstract

We study the problem of estimating the score function of an unknown probability distribution ρ^* from n independent and identically distributed observations in d dimensions. Assuming that ρ^* is subgaussian and has a Lipschitz-continuous score function s^*, we establish the optimal rate of $\tilde{\Theta}(n^{-\frac{2}{d+4}})$ for this estimation problem under the loss function $\|\hat{s} - s^*\|^2_{L^2(\rho^*)}$ that is commonly used in the score matching literature, highlighting the curse of dimensionality where sample complexity for accurate score estimation grows exponentially with the dimension d. Leveraging key insights in empirical Bayes theory as well as a new convergence rate of smoothed empirical distribution in Hellinger distance, we show that a regularized score estimator based on a Gaussian kernel attains this rate, shown optimal by a matching minimax lower bound. We also discuss extensions to estimating β-Hölder continuous scores with $\beta \leq 1$, as well as the implication of our theory on the sample complexity of score-based generative models.

Keywords: Score estimation, kernel density estimation, empirical Bayes, Hellinger distance, minimax risk

1. Introduction

Sampling from a probability distribution is a fundamental algorithmic task in many applications; for example, in Bayesian statistics, we draw samples from the posterior distribution to perform approximate inference. The *score function* of a distribution, which is defined as the derivative of the logarithm of the density of the distribution, encodes rich information about the distribution. In particular, if we have access to the score function of a distribution, then we can sample from it by running any first-order sampling algorithm such as the Langevin dynamics or the Hamiltonian Monte Carlo. Recent results have shown mixing time guarantees for such algorithms under structural assumptions on the target distribution, such as log-concavity or isoperimetry; see for example Dalalyan (2017a); Durmus et al. (2019); Bou-Rabee et al. (2020). More recently, an alternative method for sampling known as the "Score-based Generative Models" (SGMs) have been proposed, which operates via following the reverse diffusion process from a standard distribution such as the standard Gaussian to the target distribution; see for example Song and Ermon (2019); Song et al. (2020b); Ho et al. (2020). Implementing SGMs as an algorithm requires approximating the score function of the target distribution along the forward diffusion process; this can be done for example via score matching (Hyvärinen and Dayan, 2005), and in practice this is typically trained via neural networks. A recent wave of theoretical results has shown that assuming we have access to a good sequence of score estimators along the forward process with provable error guarantees, then algorithms derived from SGMs have good mixing time guarantees, with the same or even better iteration complexity as classical algorithms such as based on the Langevin dynamics, but without requiring assumptions such as isoperimetry, log-concavity, or even smoothness on the target distribution; see for example Lee et al. (2022, 2023); Chen et al. (2023c); Benton et al. (2023).

Motivated by the wide applications of the score function, in this paper we study the problem of estimating the score function of a probability distribution from independent samples. We assume the target distribution has full support on $\mathbb{R}^d$, is subgaussian, and has a Lipschitz-continuous score function. The Lipschitzness of the score function is a common assumption in sampling literature, including for analyzing classical Langevin-based algorithms (Dalalyan, 2017a,b; Cheng and Bartlett, 2018; Durmus and Moulines, 2019; Dalalyan and Karagulyan, 2019; Durmus et al., 2019; Vempala and Wibisono, 2019) and the newly developed SGMs (Block et al., 2020; De Bortoli et al., 2021; Lee et al., 2022; Yang and Wibisono, 2022; Chen et al., 2023a; Lee et al., 2023; Chen et al., 2023c). Let $\mathcal{P}_{\alpha,L}$ denote the class of probability distributions on $\mathbb{R}^d$ that are α-subgaussian with L-Lipschitz score functions; here we assume $\alpha^2 L \geq 1$ to ensure that $\mathcal{P}_{\alpha,L}$ is not empty. Let ρ^* be a probability density in $\mathcal{P}_{\alpha,L}$. The **score function** of ρ^* is the vector field $s^* \colon \mathbb{R}^d \to \mathbb{R}^d$ defined by

$$s^*(x) = \nabla \log \rho^*(x)$$

for all $x \in \mathbb{R}^d$, where ∇ is the gradient with respect to x. Observing samples $\mathbf{X}_n = (X_1, \dots, X_n)$ drawn i.i.d. (independently and identically distributed) from ρ^*, our goal is to learn a **score estimator** $\hat{s}(\cdot) := \hat{s}(\cdot; \mathbf{X}_n)$ that uses the samples to approximate the true score function s^*. We measure the score estimation error using the following loss:

$$\ell(\hat{s}, \rho^*) := \|\hat{s} - s^*\|_{\rho^*}^2 = \int_{\mathbb{R}^d} \|\hat{s}(x) - s^*(x)\|^2 \rho^*(x) dx. \tag{1}$$

There are a number of reasons why this is a meaningful loss function for score estimation.

- The loss function (1) is the relevant error metric in the application of score matching as assumed in the recent works in SGMs. For example, Chen et al. (2023c) showed that the sampling error of a popular type of SGM known as Denoising Diffusion Probabilistic Modeling (Ho et al., 2020) can be bounded up to the score matching loss (1) and discretization error. We will revisit this in Section 3 and discuss the implication of our results on SGMs.

- If the estimator $\hat{s}$ is *proper*, i.e. it is the score function of a valid density $\hat{\rho}$, then (1) equals the *relative Fisher information* (or *Fisher distance*) (Villani, 2003, Eq. (9.25)) between ρ^* and $\hat{\rho}$. In this work, however, we do not limit the scope to proper score estimators.

- In the special case when ρ^* is a Gaussian mixture, the loss function (1) is precisely the *regret*, the central quantity in the theory of *empirical Bayes* (EB), that measures the excess risk of a data-driven procedure over the Bayesian oracle risk. Although our model class is far richer than Gaussian mixtures, this connection with empirical Bayes is crucial for developing our score estimator; see Section 1.1.

- Finally, it is also necessary to consider a squared loss weighted by the true density ρ^* as the score cannot be estimated well in low-density regions. Indeed, for the unweighted squared loss $\|\hat{s}-s^*\|_2^2=\int_{\mathbb{R}^d}\|\hat{s}(x)-s^*(x)\|^2dx$, it is easy to show that the minimax score estimation error is infinite.[1]

The minimax risk of score estimation over the density class $\mathcal{P}_{\alpha,L}$ under the loss function (1) with sample size n is defined as

$$\mathcal{R}_n(\mathcal{P}_{\alpha,L}) \coloneqq \inf_{\hat{s}} \sup_{\rho^*\in\mathcal{P}_{\alpha,L}} \mathbb{E}\ell(\hat{s},\rho^*) \tag{2}$$

where the expectation is over $\mathbf{X}_n=(X_1,\ldots,X_n)\sim(\rho^*)^{\otimes n}$ and the infimum is taken over all estimators $\hat{s}$ that is measurable with respect to $\mathbf{X}_n$. Taking a step toward understanding the theoretical aspects of score estimation, we summarize the major contributions of this paper as follows:

1. We study a regularized score estimator based on the KDE (kernel density estimator) using Gaussian kernel. We analyze the performance of this estimator $\hat{s}$ (see (6)) under the loss function (1) and establish an upper bound on the minimax risk as follows:

$$\sup_{\rho^*\in\mathcal{P}_{\alpha,L}} \mathbb{E}\ell(\hat{s},\rho^*) \lesssim n^{-\frac{2}{d+4}}\,\mathrm{polylog}(n). \tag{3}$$

When ρ^* is such that the score function s^* is (L,β)-Hölder continuous for some $0<\beta\leq 1$, using the same estimator, we extend the upper bound to the following:

$$\mathbb{E}\ell(\hat{s},\rho^*) \lesssim n^{-\frac{2\beta}{d+2\beta+2}}\mathrm{polylog}(n).$$

2. We prove a matching minimax lower bound:

$$\mathcal{R}_n(\mathcal{P}_{\alpha,L}) \gtrsim n^{-\frac{2}{d+4}} \tag{4}$$

thereby showing that the optimal rate of score estimation is $n^{-\frac{2}{d+4}}$ up to logarithmic factors. The proof adapts the standard approach for establishing minimax lower bounds in nonparametric density estimation using Fano's lemma with modifications made for scores. Comparing the lower bound (4) with the upper bound (3), we observe the typical "curse of dimensionality" which suggests that to achieve a specified level of accuracy in score estimation, the sample complexity must increase exponentially with dimension.

3. We discuss some implications of our results in the context of SGMs. In particular, we propose a regularized score estimator along the Ornstein-Uhlenbeck (OU) process targeting standard Gaussian, which is the usual forward process in SGMs. For estimating the score function at time t along the forward process, by using intermediate results in the proof of the upper bound (3), we derive an error bound of $\tilde{O}\left(\eta^{-d/2}(tn)^{-1}\right)$ in the weighted squared loss, where η is the step size in the SGM algorithm.

[1] For example, the densities $\rho_0=\mathcal{N}(0,1)$ and $\rho_1=\mathcal{N}(\mu_n,1)$ are statistically indistinguishable with sample size n for $\mu_n=2^{-n}$, but their score functions s_0 and s_1 differ by a constant, and hence $\|s_0-s_1\|_2=\infty$.

1.1. Main idea

Let us discuss the algorithm that attains the optimal rate (3) and the broad strokes of its analysis in connection to the empirical Bayes theory. Let $\hat{\rho} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ denote the empirical distribution of the sample and its Gaussian smoothed version:

$$\hat{\rho}_h = \hat{\rho} * \mathcal{N}(0, hI_d) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{N}(X_i, hI_d). \tag{5}$$

for some bandwidth parameter $h > 0$. This Gaussian smoothing is an algorithmic device that allows us to tap into the powerful machinery of empirical Bayes. Instead of applying the score of $\hat{\rho}_h$, we consider its regularized version:

$$\hat{s}_h^{\varepsilon}(x) := \frac{\nabla \hat{\rho}_h(x)}{\max(\hat{\rho}_h(x), \varepsilon)} \tag{6}$$

for some regularization parameter $\varepsilon > 0$. A deep result in empirical Bayes due to Jiang and Zhang (2009) (see also Saha and Guntuboyina (2020) for extensions to multiple dimensions) is that the error between regularized scores of two Gaussian mixtures is upper bounded within logarithmic factors by the squared Hellinger distance between the two mixtures. Applying this result to our setting and bounding the likelihood ratio of $\frac{\rho^*}{\rho_h^*}$ using the score smoothness, we obtain

$$\|\hat{s}_h^{\varepsilon} - s_h^{*\varepsilon}\|_{\rho_h^*}^2 \lesssim \|\hat{s}_h^{\varepsilon} - s_h^{*\varepsilon}\|_{\rho^*}^2 \lesssim \frac{1}{h} H^2(\hat{\rho}_h, \rho_h^*) \cdot \mathrm{polylog}\left(\frac{1}{H^2(\hat{\rho}_h, \rho_h^*)}, \frac{1}{\varepsilon}\right). \tag{7}$$

Here $\rho_h^* = \rho^* * \mathcal{N}(0, hI_d)$ is the smoothed version of the true density, $s_h^{*\varepsilon} = \frac{\nabla \rho_h^*}{\max(\rho_h^*, \varepsilon)}$ is its regularized score, and the squared Hellinger distance between two densities p and q is

$$H^2(p, q) := \int_{\mathbb{R}^d} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx.$$

Next, we bound the Hellinger distance between the smoothed empirical distribution and the population:

$$\mathbb{E}[H^2(\hat{\rho}_h, \rho_h^*)] \lesssim \frac{1}{nh^{d/2}} \cdot \mathrm{polylog}(n). \tag{8}$$

Crucially relying on the smoothness of the score of ρ^*, this result seems *not* obtainable from the literature on smooth empirical distribution based only on the subgaussianity of ρ^* (Goldfeld et al., 2020; Block et al., 2022). (In fact, we suspect whether (8) is true without the score smoothness. See Lemma 9 and the surrounding discussion in Appendix A.1.1 for details.)

Finally, we control the error due to smoothing and regularization:

$$\|s_h^{*\varepsilon} - s^*\|_{\rho^*}^2 \lesssim \frac{\varepsilon}{h}\mathrm{polylog}\left(\frac{1}{\varepsilon}, \frac{1}{h}\right) + h. \tag{9}$$

Since the regularization parameter only contributes logarithmically in (7), we may choose it rather aggressively as $\varepsilon = n^{-2}$. Balancing the main terms of $\frac{1}{nh^{1+d/2}}$ and h leading to the optimal choice of bandwidth parameter $h = n^{-\frac{2}{d+4}}$ and the optimal rate (3).

It is helpful to clarify the difference between classical empirical Bayes and the present paper. In EB denoising, one observes i.i.d. samples $Y_1, \dots, Y_n \sim \rho^*_h = \rho^* * \mathcal{N}(0, hI_d)$, where the variance parameter h is fixed by the problem and the prior ρ^* is an *arbitrary* distribution with only tail assumptions (e.g. subgaussian). Therein, the goal is to compete with the oracle who knows the prior ρ^* and computes the Bayes estimator of $X_i \sim \rho^*$ given the noisy observation Y_i. Thanks to the Tweedie's formula Efron (2011)

$$\mathbb{E}[X_i \mid Y_i = y] = y + hs^*_h(y), \tag{10}$$

this is equivalent to estimating the score of s^*_h. Given an approximate score $\tilde{s}$, the regret of the approximate Bayes denoiser $\tilde{X}(y) = y + h\tilde{s}(y)$, i.e., the excess risk over the Bayesian oracle applied to a fresh observation, is given by the score matching loss (1), namely $h^2\ell(\tilde{s}, \rho^*_h)$.

A popular method in EB (the so-called g-modeling approach (Efron, 2014)) is to first compute an estimate $\tilde{\rho}$ of the distribution ρ^* based on Y_i's using deconvolution techniques, such as nonparametric maximum likelihood (NPMLE) (Jiang and Zhang, 2009; Saha and Guntuboyina, 2020), then bound the density estimation error $\tilde{\rho}_h = \tilde{\rho} * \mathcal{N}(0, hI_d)$ in Hellinger distance, and finally the regret of the regularized score of $\tilde{\rho}_h$ using tools such as (7). In our setting, since we have access to samples X_i's drawn from ρ^* before convolution, we can directly apply the smoothed empirical distribution with an optimized bandwidth parameter h.

1.2. Related work

Empirical Bayes. As mentioned earlier, for Gaussian mixtures the score matching loss (1) and the empirical Bayes regret is equivalent. This connection can be made more precise. Consider the nonparametric class of Gaussian mixtures $\pi * \mathcal{N}(0, I_d)$, where the mixing distribution π is supported on a ball of radius $r = O(1)$. In view of (10), this is a subset of our model class $\mathcal{P}_{\alpha,L}$ for some α, L depending on r. On this subclass, the best score estimation error is at most $O(\frac{(\log n)^5}{n})$ (Jiang and Zhang, 2009) in one dimension (see Saha and Guntuboyina (2020) for extensions to d dimensions), achieved by the NPMLE. A different approach is carried out in Li et al. (2005) based on KDE that applies (different) polynomial kernels of logarithmic degree to estimate the density and the derivative leads to similar results with worse logarithmic factors. In terms of negative results, for $d = 1$, a lower bound $\Omega\left(\frac{(\log n)^2}{(\log\log n)^2 n}\right)$ is shown in Polyanskiy and Wu (2021). Compared with these near-parametric rates $\tilde{\Theta}(n^{-1})$, the nonparametric rate $\tilde{\Theta}(n^{-\frac{2}{d+4}})$ in (3) is much slower, as the class $\mathcal{P}_{\alpha,L}$ we consider is much richer than Gaussian mixtures.

Density Estimation. We can view score estimation as a density estimation problem under a different measurement of the estimation error. In score estimation, if $\hat{s}$ is a proper estimator, i.e., $\hat{s} = \nabla \log \hat{\rho}$ for some distribution $\hat{\rho}$, then the loss function (1) is the relative Fisher information:

$$\mathsf{FI}(\rho^* \,\|\, \hat{\rho}) := \|\nabla \log \rho^* - \nabla \log \hat{\rho}\|^2_{\rho^*}, \tag{11}$$

which depends on both the density itself and its first-order gradient information. In classical density estimation, common choices of the loss function include the squared $L^2(\mathbb{R}^d)$ loss $\|\rho^* - \hat{\rho}\|_2^2 := \int_{\mathbb{R}^d} (\rho^*(x) - \hat{\rho}(x))^2 dx$ and the squared Hellinger distance $H^2(\rho^*, \hat{\rho})$. There is a rich literature on density estimation of nonparametric Gaussian mixtures. In one dimension, the optimal $L^2(\mathbb{R})$ error is known to be $\Theta((\log n)^{1/2}/n)$; for the squared Hellinger, the lower bound is $\Omega(\log n/n)$ (Ibragimov, 2001; Kim, 2014), and the upper bound is $O((\log n)^2/n)$ achieved by the NPMLE (Jiang and

Zhang, 2009). In d dimensions, the optimal $L^2(\mathbb{R}^d)$ error is $\Theta((\log n)^{d/2}/n)$ (Kim and Guntuboyina, 2022); for the squared Hellinger, the lower bound is $\Omega((\log n)^d/n)$ (Kim and Guntuboyina, 2022), and the upper bound is $O((\log n)^{d+1}/n)$ which is established for the NPMLE (Saha and Guntuboyina, 2020).

Score Estimation. While many methods have been proposed for estimating the score function, theoretical results are scarce. One approach involves density estimation techniques, such as kernel density estimation or neural network-based methods, followed by differentiation of their logarithm; see for example Scott (1992); Papamakarios et al. (2017). Another approach estimates the unnormalized log-density and differentiates this estimate; this is effective since the score function does not depend on the normalizing constant. A popular method in this area is called score matching (Hyvärinen and Dayan, 2005); this method proceeds by minimizing the relative Fisher information between the data distribution and the learned model distribution, which is equivalent to the loss function (1) in the case that the score estimator is proper. It has been shown that the minimization of the score matching loss (1) is a consistent estimator assuming the global minimum is found by the optimization algorithm used in the estimation (Hyvärinen and Dayan, 2005). The work of Sutherland et al. (2018) uses the Nyström approximation to speed up the score matching procedure to learn an exponential family density model with the natural parameter in a reproducing kernel Hilbert space, which may be infinite-dimensional, as introduced in Sriperumbudur et al. (2017). The work of Zhou et al. (2020) studies nonparametric score estimation via kernel ridge regression and proved the sample complexity of the resulting score estimator under some assumptions, including that the underlying score function can be written as the image of an integral operator in a reproducing kernel Hilbert space. The work of Saremi et al. (2018) trained a neural network to minimize the score matching objective and output the energy–unnormalized log-density. For a review of modern approaches to energy-based model training, see for example Song and Kingma (2021). Besides parameter estimation in unnormalized models, one can also train a neural network to directly output the score by minimizing the score matching objective (Song et al., 2020a).

Score-based Generative Models. Recent advancements in SGMs have focused on convergence analyses of the algorithms assuming access to an accurate score estimator. Initial studies either hinged on structural assumptions on the data distribution such as a log-Sobolev inequality (Lee et al., 2022; Yang and Wibisono, 2022) or strong assumptions on score estimation error such as L^∞ error (De Bortoli et al., 2021), or they led to bounds that exponentially increased with the problem parameters (De Bortoli, 2022; Block et al., 2020). Subsequent studies have achieved polynomial convergence rates under less restrictive assumptions, including that the data distribution has a finite second moment and the scores along the forward process are Lipschitz (Chen et al., 2023a; Lee et al., 2023; Chen et al., 2023c). More recent results including Benton et al. (2023) have established polynomial convergence guarantees under a minimal assumption of a finite second moment of the data distribution. Parallel to these developments, significant efforts have been directed towards the problem of score estimation in SGMs, including the following. The work of Chen et al. (2023b) studied the score estimation using neural networks and derived a finite-sample bound for a specifically chosen network architecture and parameters, with the assumption that the data lies in a low-dimensional linear subspace. The work of Oko et al. (2023) bounded the estimation error when using a neural network and showed that diffusion models are nearly minimax-optimal estimators in the total variation and in the Wasserstein distance of order one, assuming the target density belongs to the Besov space. The work of Scarvelis et al. (2023) proposed to smooth the closed-form score from empirical distribution to obtain an SGM that can generate samples without training. The work of Cui et al. (2023) obtained an error rate of $\Theta(1/n)$ for SGM when the target distribution is a mixture of two Gaussians and using a two-layer neural network for learning the score function. The work of Cole and Lu (2024) showed that the score function can be approximated efficiently via neural networks when the target distribution is subgaussian and has a log-relative density with respect to the Gaussian measure which is a Barron function, i.e. can be approximated efficiently by neural networks. The work of Li et al. (2024) studied SGM with score estimator from empirical kernel density estimator, similar to our work; they showed the sample complexity when the target distribution is either a standard Gaussian or has bounded support, and discussed the issue of memorization of training samples. A concurrent work by Zhang et al. (2024) shows that when the target distribution belongs to the β-Sobolev space with $\beta \leq 2$, the diffusion model with a kernel-based score estimator is minimax optimal up to logarithmic factors.

1.3. Notations and definitions

We review the necessary notations and definitions. Let $\mathcal{P}(\mathbb{R}^d)$ denote the space of probability distributions on $\mathbb{R}^d$. For distributions $\rho, \nu \in \mathcal{P}(\mathbb{R}^d)$ which are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, for convenience we also write their probability density functions as $\rho\colon \mathbb{R}^d \to \mathbb{R}$ and $\nu\colon \mathbb{R}^d \to \mathbb{R}$. Recall the total variation (TV) distance between ρ and ν is

$$\mathrm{TV}(\rho, \nu) = \sup_{A \subseteq \mathbb{R}^d} |\rho(A) - \nu(A)| = \frac{1}{2} \int_{\mathbb{R}^d} |\rho(x) - \nu(x)| \, dx.$$

For a function $f\colon \mathbb{R}^d \to \mathbb{R}$ and a probability distribution ρ on $\mathbb{R}^d$, the squared $L^2(\rho)$-norm of f is

$$\|f\|_\rho^2 := \mathbb{E}_\rho[f^2] = \int_{\mathbb{R}^d} f(x)^2 \, \rho(x) \, dx.$$

We define $L^2(\rho)$ to be the space of functions $f\colon \mathbb{R}^d \to \mathbb{R}$ for which $\|f\|_\rho^2 < \infty$. Similarly, given a vector field $s\colon \mathbb{R}^d \to \mathbb{R}^d$, the squared $L^2(\rho)$-norm of s is

$$\|s\|_\rho^2 := \mathbb{E}_\rho\left[\|s\|^2\right] = \int_{\mathbb{R}^d} \|s(x)\|^2 \, \rho(x) \, dx.$$

We say a probability distribution ρ on $\mathbb{R}^d$ is **α-subgaussian** for some $0 < \alpha < \infty$ if for all $\theta \in \mathbb{R}^d$:

$$\mathbb{E}_\rho \exp(\theta^\top (X - \mathbb{E}_\rho X)) \leq \exp\left(\frac{\alpha^2 \|\theta\|^2}{2}\right).$$

We say a random variable $X \sim \rho$ is subgaussian if its distribution ρ is subgaussian.

We use $a = O(b)$ or $b = \Omega(a)$ to indicate that $a \leq Cb$ for a universal constant $C > 0$. We use $a = \Theta(b)$ to indicate that $C_1 b \leq a \leq C_2 b$ for $C_2 > C_1 > 0$. And $\tilde{O}(\cdot)$ hides logarithmic factors.

2. Main results

2.1. Score estimator via Empirical Bayes smoothing

Suppose we are given a sample of n i.i.d. observations $\mathbf{X}_n = (X_1, \dots, X_n)$ from an unknown distribution $\rho^* \in \mathcal{P}_{\alpha,L}$. Our goal is to estimate the score $s^* = \nabla \log \rho^*$.

For $h > 0$, let $\hat{\rho}_h$ be the smoothed empirical distribution which is a mixture of Gaussians:

$$\hat{\rho}_h = \frac{1}{n}\sum_{i=1}^{n} \mathcal{N}(X_i, hI_d).$$

We propose the following regularized KDE score estimator $\hat{s}_h^{\varepsilon}(\cdot) = \hat{s}_h^{\varepsilon}(\cdot; \mathbf{X}_n)$:

$$\hat{s}_h^{\varepsilon}(x) := \frac{\nabla \hat{\rho}_h(x)}{\max(\hat{\rho}_h(x), \varepsilon)} \tag{12}$$

for some bandwidth parameter $h > 0$ and regularization parameter $\varepsilon > 0$.

We measure the accuracy of the score estimator $\hat{s}_h^{\varepsilon}$ in the expected square $L^2(\rho^*)$-norm as in (1), and establish the following error bound on the order of $\tilde{O}(n^{-\frac{2}{d+4}})$. Here the expectation is taken over the sample $\mathbf{X}_n \sim (\rho^*)^{\otimes n}$.

Theorem 1 *Let $d \geq 1$ be fixed, and suppose we have $X_1, \ldots, X_n$ drawn i.i.d. from some $\rho^* \in \mathcal{P}_{\alpha,L}$. Setting*

$$\varepsilon = n^{-2} \qquad \text{and} \qquad h = \left(\frac{d^3(\alpha^2 \log n)^{d/2}}{L^2 n}\right)^{\frac{2}{d+4}},$$

for sufficiently large n, the score estimator (12) satisfies

$$\mathbb{E}\ell(\hat{s}_h^{\varepsilon}, \rho^*) \leq C d\alpha^2 L^2 \left(\log n\right)^{\frac{d}{d+4}} n^{-\frac{2}{d+4}}$$

where $\ell(\cdot,\cdot)$ is defined in (1), and $C > 0$ is a universal constant.

Proof We provide the main argument for proving Theorem 1, deferring some of the lemmas to the appendix. We define:

$$\rho_h^* = \rho^* * \mathcal{N}(0, hI_d), \qquad s_h^* = \nabla \log \rho_h^*, \qquad \text{and} \qquad s_h^{*\varepsilon} = \frac{\nabla \rho_h^*}{\max(\rho_h^*, \varepsilon)}. \tag{13}$$

Since s^* is L-Lipschitz, we can show that the density ratio of ρ^* to ρ_h^* is bounded from above everywhere by a constant: In Lemma 10 in Appendix A.1, we show that for all $x \in \mathbb{R}^d$,

$$\frac{\rho^*(x)}{\rho_h^*(x)} \leq \exp\left(dLh/2\right).$$

Then by a change of measure from ρ^* to ρ_h^*, we get

$$\mathbb{E}\ell(\hat{s}_h^{\varepsilon}, \rho^*) = \mathbb{E}\|\hat{s}_h^{\varepsilon} - s^*\|_{\rho^*}^2 \leq \exp\left(dLh/2\right)\mathbb{E}\|\hat{s}_h^{\varepsilon} - s^*\|_{\rho_h^*}^2. \tag{14}$$

We can decompose the last factor on the right-hand side above as follows:

$$\mathbb{E}\|\hat{s}_h^{\varepsilon} - s^*\|_{\rho_h^*}^2 \leq \; 3\mathbb{E}\|\hat{s}_h^{\varepsilon} - s_h^{*\varepsilon}\|_{\rho_h^*}^2 \; + 3\|s_h^{*\varepsilon} - s_h^*\|_{\rho_h^*}^2 + 3\|s_h^* - s^*\|_{\rho_h^*}^2. \tag{15}$$

We now bound each of the three terms above separately.

First term: The first term $\mathbb{E}\|\hat{s}_h^\varepsilon - s_h^{*\varepsilon}\|_{\rho_h^*}^2$ concerns the distance between the regularized score functions of $\hat{\rho}_h$ and of ρ_h^*, which we can bound as follows: If $\varepsilon \in (0, (2\pi h)^{-d/2}e^{-1/2}]$ and $\alpha^2 n^{-2/d}\log n \lesssim h \leq 1/(4L)$, then by Lemma 12 in Appendix A.2,

$$\mathbb{E}\|\hat{s}_h^\varepsilon - s_h^{*\varepsilon}\|_{\rho_h^*}^2 \leq \frac{Cd\,(C_{h,d,\alpha}+d)}{nh}\left[\left(\log\frac{1}{\varepsilon(2\pi h)^{d/2}}\right)^3 + \log\frac{n}{C_{h,d,\alpha}+d}\right]. \tag{16}$$

where $C>0$ is a universal constant and $C_{h,d,\alpha} \coloneqq \left(\frac{\alpha^2\log n}{h}\right)^{d/2}$. The proof of Lemma 12 proceeds by first bounding $\|\hat{s}_h^\varepsilon - s_h^{*\varepsilon}\|_{\rho_h^*}^2$ in terms of the squared Hellinger distance between $\hat{\rho}_h$ and ρ_h^* using the result of Saha and Guntuboyina (2020), extending (Jiang and Zhang, 2009, Theorem 3) to d dimensions. This crucially uses the regularization in the score estimates. Another crucial component of the proof involves establishing a bound for the expected Hellinger distance between Gaussian-smoothed empirical distribution to the population, which gives us the desired bound of $\frac{1}{nh^{d/2}}$ and thereby achieving the optimal rate in Theorem 1. We formally present this result as Lemma 9 in Appendix A.1.1.

Second term: The second term $\|s_h^{*\varepsilon} - s_h^*\|_{\rho_h^*}^2$ is the error induced by the regularization, which we can bound as follows: If $0 \leq \varepsilon \leq (2\pi h)^{-d/2}/e$ and $h \leq \alpha^2$, then

$$\|s_h^{*\varepsilon} - s_h^*\|_{\rho_h^*}^2 \leq \frac{2\varepsilon}{h}(64\alpha^2\log n)^{d/2}\log\frac{1}{\varepsilon(2\pi h)^{d/2}} + \frac{2d^{3/2}}{hn^2}. \tag{17}$$

We state the result formally as Lemma 13 in Appendix A.3.

Third term: The third term $\|s_h^* - s^*\|_{\rho_h^*}^2$ is a bias term that we can bound using the Lipschitzness of the score function. Concretely, by Lemma 16 in Appendix A.5, we have: If $h < 1/(4L)$,

$$\|s_h^* - s^*\|_{\rho_h^*}^2 \leq L^2 hd. \tag{18}$$

Combining the bounds. Combining the bounds (16)–(18), we have

$$\begin{aligned}\mathbb{E}\|\hat{s}_h^\varepsilon - s^*\|_{\rho_h^*}^2 \leq{}& \frac{Cd\,(C_{h,d,\alpha}+d)}{hn}\left(\log\frac{1}{\varepsilon(2\pi h)^{d/2}}\right)^3 + \frac{Cd\,(C_{h,d,\alpha}+d)}{hn}\log\frac{n}{C_{h,d,\alpha}+d}\\ &+\frac{6\varepsilon}{h}(64\alpha^2\log n)^{d/2}\log\frac{1}{\varepsilon(2\pi h)^{d/2}} + \frac{6d^{3/2}}{hn^2} + 3L^2hd\end{aligned}$$

where $C>0$ is a universal constant. By choosing $\varepsilon = n^{-2}$, the first term dominates the second and third terms for $n = \Omega(e^d)$. It follows that

$$\mathbb{E}\|\hat{s}_h^\varepsilon - s^*\|_{\rho_h^*}^2 \leq \frac{Cd^4(\alpha^2\log n)^{d/2}}{nh^{d/2+1}}\left(\log\frac{n}{h}\right)^3 + 3L^2hd \tag{19}$$

where $C>0$ is a different universal constant. Finally, we optimize the bound (19) over h. By choosing $h = \left(\frac{d^3(\alpha^2\log n)^{d/2}}{L^2 n}\right)^{\frac{2}{d+4}}$, the two terms in (19) are in the same order, and hence we obtain the desired bound by the change of measure in (14):

$$\mathbb{E}\|\hat{s}_h^\varepsilon - s^*\|_{\rho^*}^2 \leq Cd\alpha^2L^2\,(\log n)^{\frac{d}{d+4}}\,n^{-\frac{2}{d+4}}$$

for a universal constant $C > 0$. ■

The following result generalizes Theorem 1 to score function s^* that is (L, β)-Hölder-continuous with $0 < \beta \leq 1$, satisfying $\|s^*(x_1) - s^*(x_2)\| \leq L\|x_1 - x_2\|^\beta$ for any $x_1, x_2 \in \mathbb{R}^d$. The proof uses a more general argument based on the score bound of Gaussian mixture in (Polyanskiy and Wu, 2016, Proposition 2) in place of the more delicate argument (Lemma 14) based on log-concavity that requires the Lipschitzness of the score. For readability, we provide the proof of Theorem 2 separately in Appendix B.

Theorem 2 *Fix $d \geq 1$. Suppose we have $X_1, \cdots, X_n$ drawn i.i.d. from some ρ^* which is a α-subgaussian distribution on $\mathbb{R}^d$ with (L, β)-Hölder continuous score function for some $0 < \beta \leq 1$. Setting*

$$\varepsilon = n^{-2} \qquad \textit{and} \qquad h = \left(\frac{d^{4-\beta}(\alpha^2 \log n)^{d/2}}{L^2 n}\right)^{\frac{2}{d+2\beta+2}}$$

for sufficiently large n, the score estimator (12) satisfies

$$\mathbb{E}\ell(\hat{s}_h^\varepsilon, \rho^*) \leq C d^\beta L^2 \alpha^{2\beta} (\log n)^{\frac{d\beta}{d+2\beta+2}} n^{-\frac{2\beta}{d+2\beta+2}}$$

for a universal constant $C > 0$.

2.2. Minimax lower bound

The following minimax lower bound, matching the upper bound in Theorem 1 up to logarithmic factors, is proved in Appendix C. The same argument is expected to extend straightforwardly to yield a matching lower bound for Theorem 2 for the case of β-Hölder continuous scores.

Theorem 3 *For any $d \geq 1$ and $\alpha > 0$, there exist constants $c = c(d, \alpha)$ and $L = L(d, \alpha)$ such that*

$$\inf_{\hat{s}} \sup_{\rho^* \in \mathcal{P}_{\alpha,L}} \mathbb{E}\ell(s, \rho^*) \geq c(d, \alpha)\, n^{-\frac{2}{d+4}}. \tag{20}$$

The above convergence rate can be interpreted by drawing analogy with classic results on estimating smooth densities in the Hölder class. It is well-known (Stone, 1982, 1983) that, for m-smooth densities supported on a d-dimensional hypercube, the optimal rate (in mean squared L^2-error) of estimating the rth derivative is

$$n^{-\frac{2(m-r)}{2m+d}}. \tag{21}$$

It is conceivable that estimating the score function (derivative of log density) is at least as hard as estimating the derivative of the density itself. Since the Lipschitz assumption of the score translates to twice differentiability of the density, we see that Theorem 3 corresponds to (21) with $m = 2$ and $r = 1$.

Furthermore, optimal estimation of m-smooth densities in squared error is attained by KDE with kernel chosen depending on the smoothness parameter m and the optimal bandwidth $n^{-\frac{1}{d+2m}}$ (Stone, 1983). For $m = 2$ this curiously coincides with the bandwidth choice $\sqrt{h}$ of the Gaussian kernel in Theorem 1; for $m = \beta + 1$, this coincides with Theorem 2. By classical results in density estimation (Tsybakov, 2009), for densities with smoothness parameter $m \leq 2$, positive kernels are optimal; however, for higher smoothness $m > 2$, kernels with negative parts must be used. For this reason, the methodology in the current paper based on Gaussian kernel may cease to be optimal if the score function is smoother than Lipschitz. Determining the optimal rate of estimating β-Hölder scores with $\beta > 1$ remains an open question.

The optimal rate $n^{-\frac{2}{d+4}}$ exhibits the typical "curse of dimensionality", suggesting that the optimal sample complexity reaching a given accuracy of score estimation must grow exponentially with the dimension. Rigorously establishing this is an interesting question that requires proving a sharper non-asymptotic lower bound for score estimation that applies to d growing with n. We note that for estimating the density itself this was successfully carried out in McDonald (2017).

3. Application to SGM

In this section, we discuss some implications of our results from Section 2.1 in the context of SGMs. We derive a finite-sample error bound for a KDE score estimator along the forward process, and then we plug in our result to existing guarantees for SGMs to deduce a final sample complexity result. We first provide a brief review of a specific type of SGM called *Denoising Diffusion Probabilistic Modeling (DDPM)*; we refer the reader to Ho et al. (2020) for more details.

Suppose our target distribution is ν on $\mathbb{R}^d$. In DDPM, we start with an Ornstein-Uhlenbeck (OU) process targeting $\gamma = \mathcal{N}(0, I_d)$:

$$dX_t = -X_t\,dt + \sqrt{2}dW_t, \qquad X_0 \sim \nu_0 = \nu, \tag{22}$$

where W_t is the standard Brownian motion in $\mathbb{R}^d$. Let $\nu_t = \mathrm{Law}(X_t)$ be the distribution of X_t along the OU flow, and let $s_t = \nabla \log \nu_t$ be the score function of ν_t [2]. We run the OU process (22) until time $T > 0$, and then we simulate the backward (time-reversed) process, which can be described by the following stochastic differential equation:

$$d\tilde{Y}_t = (\tilde{Y}_t + 2s_{T-t}(\tilde{Y}_t))dt + \sqrt{2}dW_t, \qquad \tilde{Y}_0 \sim \mu_0 = \nu_T. \tag{23}$$

By construction, if we start the backward process of (23) from $\tilde{Y}_0 \sim \mu_0 = \nu_T$, then we will have $\tilde{Y}_t \sim \mu_t = \nu_{T-t}$ for $0 \leq t \leq T$, and thus $\tilde{Y}_T \sim \mu_T = \nu$ is a sample from the target distribution. However, in practice, we do not know the score functions $(s_t)_{0<t\leq T}$. Typically, we only assume we have independent samples from ν, which we use to construct score estimators $(\hat{s}_t)_{0<t\leq T}$. Then in the algorithm, we start the backward process (23) from $\tilde{Y}_0 \sim \gamma$ (the target distribution of forward process (22) which is close to ν_T for large T), and we simulate the backward process in discrete time with score estimators $(\hat{s}_t)_{0<t\leq T}$ that we learn from samples. Let $\eta > 0$ be the step size, and $K = \frac{T}{\eta}$ so $T = K\eta$; we assume $K \in \mathbb{N}$. In each step, the DDPM algorithm performs the following update:

$$\textbf{(DDPM)} \qquad y_{k+1} = e^{\eta} y_k + 2(e^{\eta} - 1)\hat{s}_{\eta(K-k)}(y_k) + \sqrt{e^{2\eta} - 1}z_k \tag{24}$$

[2] We note a change in the notation of s_t. Previously, s_t denotes the score function of the Gaussian-smoothed target distribution with t being the smoothing variance. In this section, s_t is the score function of ν_t along the OU process starting from target distribution ν.

where $\hat{s}_{\eta k}$ is an approximation to the score function of $\nu_{\eta k}$, and $z_k \sim \mathcal{N}(0, I_d)$ is an independent Gaussian random variable.

We will apply the convergence result from Chen et al. (2023c), who showed that the DDPM algorithm (24) returns a sample that is close in TV distance to ν up to the score estimation error and discretization error under some assumptions, including an error bound on the score estimation error. We note that in Chen et al. (2023c) the score estimator is learned from (and hence depends on) the sample, and their result is stated for a fixed score estimator. An inspection of their proof shows that the guarantees hold in expectation (with respect to the random sample), which we state below as Proposition 4.

Proposition 4 *Assume:*

1. *For all $t \geq 0$, the score $s_t = \nabla \log \nu_t$ is L-Lipschitz for some $1 \leq L < \infty$;*
2. *The target distribution ν has a finite second moment $m_2^2 := \mathbb{E}_\nu[\|X\|^2] < \infty$;*
3. *For $k = 1, \cdots, K$, the score estimator $\hat{s}_{k\eta}(\cdot) = \hat{s}_{k\eta}(\cdot; \mathbf{X}_n)$, which depends on the sample $\mathbf{X}_n$, has expected score estimation error $\mathbb{E}_{\mathbf{X}_n}\|s_{k\eta} - \hat{s}_{k\eta}\|^2_{\nu_{k\eta}} \leq \epsilon^2_{\text{score}}$; and*
4. *The step size $\eta = T/K$ satisfies $\eta \lesssim 1/L$.*

At each time $t = k\eta$, let ρ_t be the law of the iterate y_k of the DDPM algorithm (24) *conditioned on the sample $\mathbf{X}_n$. Then it holds that:*

$$\mathbb{E}_{\mathbf{X}_n}[\mathrm{TV}(\rho_T, \nu)] \lesssim e^{-T}\sqrt{\mathrm{KL}(\nu\|\gamma)} + \left(L\sqrt{d\eta} + Lm_2\eta + \epsilon_{\text{score}}\right)\sqrt{T}. \tag{25}$$

3.1. Score estimation along the OU flow

Given i.i.d. observations $X^{(1)}, \cdots, X^{(n)}$ from ν, we need to estimate the score functions along the OU process, i.e. $s_t = \nabla \log \nu_t$ for any $t > 0$. Recall that following the OU flow (22), $X_t \sim \nu_t = \mathrm{Law}(e^{-t}X_0) * \mathcal{N}(0, \tau(t)I_d)$ where $\tau(t) = 1 - e^{-2t}$. When we start the OU flow with a finite set of observations $\{X^{(i)}\}_{i=1}^n$, i.e. the initial distribution of the flow is the empirical distribution $\hat{\nu}_0 = \frac{1}{n}\sum_{i=1}^n \delta_{X^{(i)}}$, the perturbed distribution at time t is a Gaussian mixture

$$\hat{\nu}_t = \frac{1}{n}\sum_{i=1}^n \mathcal{N}(e^{-t}X^{(i)}, \tau(t)I_d)$$

Its score function, $\nabla \log \hat{\nu}_t$, can be easily expressed in a closed-form. Using this closed-form score function for the sampling allows for a sampler without training. This may seem appealing, but using this score function in the backward SDE (23) will convert the noise to the empirical distribution $\hat{\nu}_0$, which means the model will memorize the training set and cannot generate novel samples. Therefore, we propose to use the regularized score function of $\hat{\nu}_t$:

$$\hat{s}_t^\varepsilon = \frac{\nabla \hat{\nu}_t}{\max(\hat{\nu}_t, \varepsilon)} \tag{26}$$

for some $\varepsilon > 0$. The induced error in the closed-form score will enable the model to generalize. Furthermore, we can analyze its performance by appealing to similar techniques used in the proof of Theorem 1. We state the result as follows and provide the proof in Appendix D.

Theorem 5 *Fix $d \geq 1$. Assume ν is α-subgaussian and has an L-Lipschitz score. Choose $\varepsilon = n^{-2}$. If the step size of DDPM (24) satisfies*

$$\frac{1}{2}\log\left(1+\frac{\alpha^2 \log n}{n^{2/d}}\right) \lesssim \eta \leq \frac{1}{2}\log\left(1+\frac{1}{4L-1}\right),$$

then at time $t \geq \eta$, the squared $L^2(\nu_t)$ error of the estimator (26) satisfies

$$\mathbb{E}\ell(\hat{s}_t^\varepsilon, \nu_t) = \mathbb{E}\|\hat{s}_t^\varepsilon - s_t\|_{\nu_t}^2 \lesssim \frac{1}{n}\frac{d}{(1-e^{-2t})}\left(\left(\frac{\alpha^2\log n}{e^{2\eta}-1}\right)^{d/2} + d\right)\left(\log\frac{n}{(2\pi(1-e^{-2t}))^{d/4}}\right)^3 \tag{27}$$

where $\lesssim$ hides absolute constant, and the expectation is taken over the i.i.d. sample $X^{(1)}, \cdots, X^{(n)}$ from ν.

Note for small t and large n, the right-hand side above is $\tilde{O}(\eta^{-d/2}(tn)^{-1})$. The bound decreases in t; this is because ν_t converges to the standard Gaussian as $t \to \infty$. In fact, our method shows that to reach $\mathbb{E}\ell(\hat{s}_t^\varepsilon, \nu_t) \leq \epsilon_{\text{score}}^2$ for *all* $t \geq \eta$, it suffices to have $\tilde{O}\left(\frac{d\alpha^d}{\eta^{d/2+1}\epsilon_{\text{score}}^2}\right)$ samples. This is not obvious because, despite that both $\hat{\nu}_t$ and ν_t move closer to the same Gaussian as t increases, it is unclear whether the score estimation error $\|\hat{s}_t^\varepsilon - s_t\|_{\nu_t}^2$ is monotonically decreasing in t.[3] Nevertheless, empirical Bayes techniques (recall (7)) allow us to control $\ell(\hat{\nu}_t, \nu_t)$ in terms of the Hellinger distance $H^2(\hat{\nu}_t, \nu_t)$ which satisfies data processing inequality (Polyanskiy and Wu, 2024, Theorem 7.4) and hence decreasing in t. Furthermore, a simple application of Markov inequality shows that on a high probability event (with respect to $H^2(\hat{\nu}_\eta, \nu_\eta)$), the preceding bound on $\ell(\hat{s}_t^\varepsilon, \nu_t)$ holds simultaneously for all $t \geq \eta$.

Combining Theorem 5 with the previous convergence result for DDPM (Proposition 4), we obtain the following sample complexity guarantee for DDPM driven by the regularized score estimator (26):

Corollary 6 *Suppose the assumptions in Proposition 4 and Theorem 5 hold. In order to have $\mathbb{E}\mathrm{TV}(\rho_T, \nu) \leq \epsilon_{\mathrm{TV}}$, it suffices to run DDPM (24) with $\hat{s}_t = \hat{s}_t^\varepsilon$ in (26) for $T \asymp \log(\mathrm{KL}(\nu\|\gamma)/\epsilon_{\mathrm{TV}})$ and $\eta \asymp \frac{\epsilon_{\mathrm{TV}}^2}{L^2 d}$, and have*

$$n = \tilde{O}\left(\frac{d^{d/2+2}\alpha^d L^{d+2}}{\epsilon_{\mathrm{TV}}^{d+4}}\right)$$

samples for score estimation.

4. Discussion

In this paper, we study the score estimation for subgaussian densities in d dimensions with Lipschitz-continuous score functions. Under the score matching loss (1), we establish a minimax lower bound at the rate of $n^{-\frac{2}{d+4}}$ using Fano's lemma. Applying techniques from empirical Bayes and smoothed

[3] Even in the absence of the regularization parameter ϵ, in which case $\|\hat{s}_t - s_t\|_{\nu_t}^2$ coincides with the relative Fisher information $\mathsf{FI}(\hat{\nu}_t \,\|\, \nu_t)$ in (11), it is still unclear whether this is monotone in t because FI is convex in the first argument but not in the second. In comparison, $H^2(\hat{\nu}_t, \nu_t) = H^2(\hat{\nu} * \mathcal{N}(0, (e^{2t}-1)I_d), \nu * \mathcal{N}(0, (e^{2t}-1)I_d))$ is decreasing in t.

empirical distribution as well as new insights enabled by the Lipschitzness of the true score, a regularized KDE score estimator using a Gaussian kernel with optimized bandwidth is shown to achieve this optimal rate up to logarithmic factors. The convergence rate $n^{-\frac{2}{d+4}}$ suggests that an exponential increase in sample complexity is unavoidable as the dimension d grows.

Within the SGM framework, particularly considering an OU flow as the forward process, if we start with n independent observations, the score function along the OU flow has a closed-form (as the score of a Gaussian mixture). To improve generalization, we explicitly introduce a regularization term and analyze the performance of this estimator, leading to a sample complexity of $\tilde{O}\left(\frac{d\alpha^d}{\eta^{d/2+1}\epsilon_{\text{score}}^2}\right)$ for score matching up to error ϵ_{score}, and a sample complexity of $\tilde{O}\left(\frac{d^{d/2+2}\alpha^d L^{d+2}}{\epsilon_{\mathrm{TV}}^{d+4}}\right)$ for SGMs using this score estimator to reach a sampling error ϵ_{TV}. The exponential dependence of sample complexity on the dimension d is fundamental to the nonparametric distribution class $\mathcal{P}_{\alpha,L}$ we consider, which only assumes subgaussianity and score smoothness. There is a need to seek a meaningful distribution class whose sample complexity for score estimation has a milder dependency on d.

Acknowledgments

Y. Wu is grateful to Yandi Shen for helpful discussion on Lemma 13 and to Sinho Chewi for discussion that motivated Lemma 18. Y. Wu is supported in part by the NSF Grant CCF-1900507 and an Alfred Sloan fellowship.

References

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising autoencoders and Langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

Adam Block, Zeyu Jia, Yury Polyanskiy, and Alexander Rakhlin. Rate of convergence of the smoothed empirical Wasserstein distance. *arXiv preprint arXiv:2205.02128*, 2022.

Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *Annals of Applied Probability*, 30(3):1209–1250, 2020.

Herm Jan Brascamp and Elliott H Lieb. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of functional analysis*, 22(4):366–389, 1976.

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

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

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

Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Algorithmic Learning Theory*, pages 186–211. PMLR, 2018.

Frank Cole and Yulong Lu. Score-based generative models break the curse of dimensionality in learning a family of sub-gaussian probability distributions. In *International Conference on Learning Representations*, 2024.

Hugo Cui, Florent Krzakala, Eric Vanden-Eijnden, and Lenka Zdeborová. Analysis of learning a flow-based generative model from limited sample complexity. *arXiv preprint arXiv:2310.03575*, 2023.

Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689. PMLR, 2017a.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):651–676, 2017b.

Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

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

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger Bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, 2021.

Alain Durmus and Éric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854 – 2882, 2019.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Bradley Efron. Two modeling strategies for empirical Bayes estimation. *Statistical Science*, 29(2): 285, 2014.

Ziv Goldfeld, Kristjan Greenewald, Jonathan Niles-Weed, and Yury Polyanskiy. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66(7):4368–4391, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

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

Ildar Ibragimov. Estimation of analytic functions. *Lecture Notes-Monograph Series*, pages 359–383, 2001.

Zeyu Jia, Yury Polyanskiy, and Yihong Wu. Entropic characterization of optimal rates for learning gaussian mixtures. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4296–4335. PMLR, 12–15 Jul 2023.

Wenhua Jiang and Cun-Hui Zhang. General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647 – 1684, 2009.

Arlene K.H. Kim. Minimax bounds for estimation of normal mixtures. *Bernoulli*, 20(4):1802 – 1818, 2014.

Arlene K.H. Kim and Adityanand Guntuboyina. Minimax bounds for estimating multivariate Gaussian location mixtures. *Electronic Journal of Statistics*, 16(1):1461–1484, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems*, 2022.

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

Jianjun Li, Shanti S Gupta, and Friedrich Liese. Convergence rates of empirical Bayes estimation in exponential family. *Journal of statistical planning and inference*, 131(1):101–115, 2005.

Sixu Li, Shi Chen, and Qin Li. A good score does not lead to a good generative model. *arXiv preprint arXiv:2401.04856*, 2024.

Daniel McDonald. Minimax density estimation for growing dimension. In *Artificial Intelligence and Statistics*, 2017.

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

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in neural information processing systems*, 2017.

Yury Polyanskiy and Yihong Wu. Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Transactions on Information Theory*, 62(7):3992–4002, 2016.

Yury Polyanskiy and Yihong Wu. Sharp regret bounds for empirical Bayes and compound decision problems. *arXiv preprint arXiv:2109.03943*, 2021.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2024. Draft: http://www.stat.yale.edu/~yw562/teaching/itbook-export.pdf.

Sujayam Saha and Adityanand Guntuboyina. On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. *The Annals of Statistics*, 48(2):738–762, 2020.

Saeed Saremi, Arash Mehrjou, Bernhard Schölkopf, and Aapo Hyvärinen. Deep energy estimator networks. In *Advances in Neural Information Processing Systems*, 2018.

Christopher Scarvelis, Haitz Sáez de Ocáriz Borde, and Justin Solomon. Closed-form diffusion models. *arXiv preprint arXiv:2310.12395*, 2023.

David W Scott. *Multivariate Density Estimation*, chapter 6, pages 125–193. John Wiley & Sons, Ltd, 1992. ISBN 9780470316849.

Yandi Shen and Yihong Wu. Empirical bayes estimation: When does g-modeling beat f-modeling in theory (and in practice)? *arXiv preprint arXiv:2211.12692*, 2022.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pages 574–584, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020b.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyv, Revant Kumar, et al. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59, 2017.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

Charles J Stone. Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. In *Recent Advances in Statistics*, pages 393–406. Elsevier, 1983.

Danica J Sutherland, Heiko Strathmann, Michael Arbel, and Arthur Gretton. Efficient and principled score estimation with nyström kernel exponential families. In *International Conference on Artificial Intelligence and Statistics*, 2018.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY, 2009.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in neural information processing systems*, 2019.

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2003.

Kaylee Yingxi Yang and Andre Wibisono. Convergence of the Inexact Langevin Algorithm and Score-based Generative Models in KL Divergence. *arXiv preprint arXiv:2211.01512*, 2022.

Kaihong Zhang, Heqi Yin, Feng Liang, and Jingbo Liu. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. *arXiv preprint arXiv:2402.15602*, 2024.

Yuhao Zhou, Jiaxin Shi, and Jun Zhu. Nonparametric score estimators. In *International Conference on Machine Learning*, pages 11513–11522. PMLR, 2020.

Appendix A. Details for proof of Theorem 1

A.1. Preliminary results

We present preliminary results which we will use. We recall the following result from Saha and Guntuboyina (2020).

Proposition 7 ((Saha and Guntuboyina, 2020, Theorem E.1)) *Let ρ_0 and ν_0 be two distributions on $\mathbb{R}^d$. Let $\rho_1 = \rho_0 * \mathcal{N}(0, I_d)$ and $\nu_1 = \nu_0 * \mathcal{N}(0, I_d)$. For $\varepsilon > 0$, let $s^\varepsilon_{\rho_1}$ and $s^\varepsilon_{\nu_1}$ denote the regularized score functions of ρ_1 and ν_1 respectively. If $\varepsilon \le (2\pi)^{-d/2}e^{-1/2}$, then*

$$\|s^\varepsilon_{\rho_1} - s^\varepsilon_{\nu_1}\|^2_{\rho_1} \le Cd \max\left\{\left(\log \frac{(2\pi)^{-d/2}}{\varepsilon}\right)^3, |\log H(\rho_1,\nu_1)|\right\} H^2(\rho_1,\nu_1)$$

where C is a universal positive constant.

Via a rescaling argument, we have the following generalization.

Lemma 8 *Let ρ and ν be two distributions on $\mathbb{R}^d$. Let $h > 0$, $\rho_h = \rho * \mathcal{N}(0, hI_d)$ and $\nu_h = \nu * \mathcal{N}(0, hI_d)$. For any $\varepsilon > 0$, let $s^\varepsilon_{\rho_h}$ and $s^\varepsilon_{\nu_h}$ be the regularized score functions of ρ_h and ν_h respectively. If $0 < \varepsilon \le (2\pi h)^{-d/2}e^{-1/2}$, then*

$$\|s^\varepsilon_{\rho_h} - s^\varepsilon_{\nu_h}\|^2_{\rho_h} \le \frac{Cd}{h} \max\left\{\left(\log \frac{(2\pi h)^{-d/2}}{\varepsilon}\right)^3, |\log H(\rho_h,\nu_h)|\right\} H^2(\rho_h,\nu_h),$$

where C is a universal positive constant.

Proof Let $X \sim \rho$ and $Y \sim \nu$, so $X_h = X + \mathcal{N}(0, hI) \sim \rho_h$ and $Y_h = Y + \mathcal{N}(0, hI) \sim \nu_h$. We can also write X_h and Y_h as

$$X_h = \sqrt{h}X' \qquad \text{where } X' = \frac{X}{\sqrt{h}} + \mathcal{N}(0, I_d) \sim \rho'$$
$$Y_h = \sqrt{h}Y' \qquad \text{where } Y' = \frac{Y}{\sqrt{h}} + \mathcal{N}(0, I_d) \sim \nu'.$$

Note that $\rho' = \mathrm{Law}(X/\sqrt{h}) * \mathcal{N}(0, I_d)$ and $\nu' = \mathrm{Law}(Y/\sqrt{h}) * \mathcal{N}(0, I_d)$. It follows from Proposition 7 that if $0 < \varepsilon' \le (2\pi)^{-d/2}e^{-1/2}$, then

$$\|s_{\rho'}^{\varepsilon'} - s_{\nu'}^{\varepsilon'}\|_{\rho'}^2 \le Cd \max\left\{\left(\log \frac{(2\pi)^{-d/2}}{\varepsilon'}\right)^3, \left|\log H(\rho', \nu')\right|\right\} H^2(\rho', \nu')$$

for some positive constant C. By the relation of ρ_h and ρ', we have $\rho_h(x) = h^{-d/2}\rho'(\frac{x}{\sqrt{h}})$. Thus we have the following relation of the score functions of ρ_h and ρ',

$$s_{\rho_h}(x) = \frac{\nabla \rho_h(x)}{\rho_h(x)} = \frac{h^{-d/2}h^{-1/2}\nabla\rho'\left(\frac{x}{\sqrt{h}}\right)}{h^{-d/2}\rho'\left(\frac{x}{\sqrt{h}}\right)} = \frac{1}{\sqrt{h}} s_{\rho'}\left(\frac{x}{\sqrt{h}}\right)$$

and similarly for ν_h and ν', $s_{\nu_h}(y) = h^{-1/2}s_{\nu'}\left(\frac{y}{\sqrt{h}}\right)$. The same holds for the regularized score functions,

$$s_{\rho_h}^{\varepsilon}(x) = \frac{1}{\sqrt{h}} s_{\rho'}^{\varepsilon'}\left(\frac{x}{\sqrt{h}}\right) \qquad \text{and} \qquad s_{\nu_h}^{\varepsilon}(y) = \frac{1}{\sqrt{h}} s_{\nu'}^{\varepsilon'}\left(\frac{y}{\sqrt{h}}\right)$$

where $\varepsilon' = h^{d/2}\varepsilon$. Therefore, if $0 < \varepsilon \le (2\pi h)^{-d/2}e^{-1/2}$, i.e. $0 < \varepsilon' \le (2\pi)^{-d/2}e^{-1/2}$, then

$$\begin{aligned} \|s_{\rho_h}^{\varepsilon} - s_{\nu_h}^{\varepsilon}\|_{\rho_h}^2 &= \int \rho_h(\tilde{x})\, \|s_{\rho_h}^{\varepsilon}(\tilde{x}) - s_{\nu_h}^{\varepsilon}(\tilde{x})\|^2 d\tilde{x} \\ &= \frac{1}{h}\int \rho'(x)\, \|s_{\rho'}^{\varepsilon'}(x) - s_{\nu'}^{\varepsilon'}(x)\|^2 dx \qquad \text{(by letting } \tilde{x} = \sqrt{h}x\text{)} \\ &\overset{(a)}{\le} \frac{Cd}{h} \max\left\{\left(\log \frac{(2\pi h)^{-d/2}}{\varepsilon}\right)^3, \left|\log H(\rho', \nu')\right|\right\} H^2(\rho', \nu'). \end{aligned}$$

where (a) uses Proposition 7 and $\varepsilon' = h^{d/2}\varepsilon$. By the scale-invariance of the Hellinger distance,

$$H^2(\rho_h, \nu_h) = H^2(\rho', \nu').$$

Therefore, we obtain the desired result

$$\|s_{\rho_h}^{\varepsilon} - s_{\nu_h}^{\varepsilon}\|_{\rho_h}^2 \le \frac{Cd}{h} \max\left\{\left(\log \frac{(2\pi h)^{-d/2}}{\varepsilon}\right)^3, |\log H(\rho_h, \nu_h)|\right\} H^2(\rho_h, \nu_h).$$

■

A.1.1. HELLINGER CONVERGENCE RATE OF SMOOTHED EMPIRICAL DISTRIBUTION

Another crucial ingredient of the proof is bounding the Hellinger distance between Gaussian-smoothed empirical distribution to the population.

Lemma 9 *Let $d \geq 1$, $h > 0$ and $\alpha > 0$. Let $\rho^* \in \mathcal{P}_{\alpha,L}$, which is an α-subgaussian measure on $\mathbb{R}^d$ with an L-Lipschitz score s^*. Let $\rho_h^* = \rho^* * \mathcal{N}(0, hI_d)$. Let $\hat{\rho}$ be the empirical measure of an i.i.d. sample of size n drawn from ρ^* and $\hat{\rho}_h = \hat{\rho} * \mathcal{N}(0, hI_d)$. Assume that $h \leq \frac{1}{4L}$ and $h \leq \alpha^2$. Then*

$$\mathbb{E}H^2(\hat{\rho}_h, \rho_h^*) \leq \frac{1}{n}\left(\frac{C\alpha^2 \log n}{h}\right)^{d/2} + \frac{4d}{n}, \tag{28}$$

where C is some universal constant.

We note that convergence rate of smoothed empirical distribution has been well-studied in the literature (see, e.g., Goldfeld et al. (2020); Block et al. (2022)) under various metrics including different types of f-divergences and transportation distances, some of which exhibit a rich spectrum of behavior depending on the relationship between the smoothing parameter and the subgaussian parameter of the population. For example, in the setting of Lemma 9, (Goldfeld et al., 2020, Proposition 2) shows that

$$\mathbb{E}\mathrm{TV}(\rho_h^*, \hat{\rho}_h) \leq \left(\frac{1}{\sqrt{2}} + \frac{\alpha}{\sqrt{h}}\right)^{d/2} e^{\frac{3d}{16}} \frac{1}{\sqrt{n}}, \tag{29}$$

which holds for any α-subgaussian ρ^* without smoothness conditions on the score. Using the inequality $H^2/2 \leq \mathrm{TV} \leq H$ (Polyanskiy and Wu, 2024, Sec. 7.3), (29) implies that $\mathbb{E}H^2(\hat{\rho}_h, \rho_h^*) \lesssim \frac{1}{\sqrt{nh^{d/2}}}$ up to constant depending on d and α. Since the inequality $H^2 \leq \mathrm{TV}$ cannot be improved in general,[4] this falls short of the desired bound of $\frac{1}{nh^{d/2}}$ in (28) and hence the optimal rate of score estimation in Theorem 1. Another option is to apply the smoothed KL bound in (Block et al., 2022, Theorem 3) and the fact that $H^2 \leq \mathrm{KL}$, leading to $\mathbb{E}\mathrm{KL}(\hat{\rho}_h \| \rho_h^*) \leq \frac{C(\log n)^d}{n}$; unfortunately, examining the proof of this result shows that the constant C is exponential in $1/h$. In fact, our proof of Lemma 9 (notably, Lemma 10 below) crucially relies on the Lipschitzness of the score and directly deals with the Hellinger distance using a truncated second moment calculation. It is unclear whether Lemma 9 holds for all subgaussian distributions without smooth scores.

To show Lemma 9, we start with an intermediate result.

Lemma 10 *Let p be a density on $\mathbb{R}^d$ whose score $s = \nabla \log p$ is L-Lipschitz. Let $p_h = p * \varphi_h$ where φ_h is the density of $\mathcal{N}(0, hI_d)$. Then for all $y \in \mathbb{R}^d$,*

$$\frac{p}{p_h}(y) \leq \exp(dLh/2).$$

[4] Note that we do have the special structure that ρ_h^* are $\hat{\rho}_h$ are both Gaussian mixtures. The recent work Jia et al. (2023) shows that for Gaussian mixtures H^2 and KL are comparable. Whether H and TV are comparable is posed as an open problem in Jia et al. (2023).

Proof Let $Z \sim \mathcal{N}(0, I_d)$. Then $p_h(y) = \mathbb{E}[p(y - \sqrt{h}Z)]$. We have

$$\begin{aligned}
\log \frac{p}{p_h}(y) = & \ \log p(y) - \log \mathbb{E}[p(y - \sqrt{h}Z)] \\
\overset{(a)}{\leq} & \ \mathbb{E}[\log p(y) - \log p(y - \sqrt{h}Z)] \\
= & \ \mathbb{E} \int_{-\sqrt{h}}^{0} \langle -Z, s(y - uZ) \rangle du \\
\overset{(b)}{=} & \ \mathbb{E} \int_{-\sqrt{h}}^{0} \langle -Z, s(y - uZ) - s(y) \rangle du \\
\overset{(c)}{\leq} & \ \mathbb{E} \int_{-\sqrt{h}}^{0} L|u| \|Z\|^2 du = dLh/2
\end{aligned}$$

where (a) applies Jensen's inequality to the convex function $x \mapsto -\log x$; (b) is because $\mathbb{E}[Z] = 0$; (c) applies Cauchy-Schwarz and the Lipschitzness of s. ■

Remark 11 *As we shall see below, the second moment calculation in bounding the expected square Hellinger distance requires controlling a quantity of the form $\int_{\mathbb{R}^d} dy \frac{p}{p_h}(y)$ as $\mathrm{poly}(1/h)$. The hope is that the convolution p_h has a slightly heavier tail than that of p such that the ratio $\frac{p}{p_h}(y)$ decays like a Gaussian of variance approximately h; this is exactly the case when p is Gaussian. While we cannot prove this in general, Lemma 10 bounds the density ratio $\frac{p}{p_h}(y)$ by a constant independent of y. As such, additional truncation will need to be introduced before passing to the second moment, which we do below.*

On the other hand, with more assumptions on the density p, it is possible to control $\int_{\mathbb{R}^d} dy \, \frac{p}{p_h}(y)$ directly. For example, if p is not only log-smooth but also strongly log-concave, then by analyzing the heat equation satisfied by p_h, one can show that $\int_{\mathbb{R}^d} dy \, \frac{p}{p_h}(y) = O(h^{-d/2})$.

Proof [Proof of Lemma 9] Let $B \subset \mathbb{R}^d$. Note that for any distributions P and Q with densities p and q on $\mathbb{R}^d$,

$$H^2(p, q) = \int_{\mathbb{R}^d} (\sqrt{p} - \sqrt{q})^2 \leq \int_B (\sqrt{p} - \sqrt{q})^2 + P(B^c) + Q(B^c) \leq \int_B \frac{(p - q)^2}{q} + P(B^c) + Q(B^c).$$

Thus, applying $\mathbb{E}\hat{\rho}_h = \rho_h^*$, we obtain

$$\mathbb{E}H^2(\hat{\rho}_h, \rho_h^*) \leq \int_B dy \frac{\mathbb{E}[(\rho_h^*(y) - \hat{\rho}_h(y))^2]}{\rho_h^*(y)} + 2 \int_{B^c} \rho_h^*. \tag{30}$$

Recall that $\varphi_h(x) = \frac{1}{(2\pi h)^{d/2}} e^{-\|x\|^2/(2h)}$ is the density of $\mathcal{N}(0, hI_d)$. Let X_i be i.i.d. as ρ^*. Note that for each y,

$$\hat{\rho}_h(y) = \frac{1}{n} \sum_{i=1}^{n} \varphi_h(y - X_i).$$

Thus $\mathbb{E}[\hat{\rho}_h(y)] = \rho_h^*(y)$ and $\mathrm{Var}[\hat{\rho}_h(y)] = \frac{1}{n}\mathrm{Var}(\varphi_h(y - X_1))$. Note that $\varphi_h(x)^2 = (4\pi h)^{-d/2} \varphi_{h/2}(x)$. So

$$\mathbb{E}[\varphi_h(y - X_1))^2] = (4\pi h)^{-d/2} \mathbb{E}[\varphi_{h/2}(y - X_1)] = (4\pi h)^{-d/2} \rho_{h/2}^*(y)$$

and $\mathrm{Var}(\varphi_h(y - X_1)) = (4\pi h)^{-d/2}\rho^*_{h/2}(y) - (\rho^*_h(y))^2$.

Combining the above with (30), we get

$$\mathbb{E}H^2(\hat{\rho}_h, \rho^*_h) \le \frac{(4\pi h)^{-d/2}}{n} \int_B dy \frac{\rho^*_{h/2}(y)}{\rho^*_h(y)} + 2\int_{B^c} \rho^*_h. \tag{31}$$

Next, choose $B = \mu + [-a, a]^d$ where μ is the mean of ρ^*, $a = \sqrt{C_0 \log n}$ for some C_0 to be specified. Since ρ^* is α-subgaussian, ρ_h is $\sqrt{\alpha^2 + h}$-subgaussian with the same mean μ. Assuming $h \le \alpha^2$, by union bound

$$\int_{B^c} \rho^*_h \le 2d \exp\left(-\frac{a^2}{4\alpha^2}\right) \le \frac{2d}{n} \tag{32}$$

upon choosing $C_0 = 4\alpha^2$ and hence $a = 2\alpha\sqrt{\log n}$.

For the first term in (31), since $h < \frac{1}{L}$, Lemma 14 in Appendix A.4 below implies that the score of $\rho^*_{h/2}$ is $2L$-Lipschitz. Applying Lemma 10 to $p = \rho^*_{h/2}$ and $t = h/2$ yields

$$\int_B dy \frac{\rho^*_{h/2}(y)}{\rho^*_h(y)} \le \mathrm{vol}(B)\exp(dLh/4) = (16\alpha^2 \log n)^{d/2} \exp(dLh/4). \tag{33}$$

Combining everything we get

$$\mathbb{E}H^2(\hat{\rho}_h, \rho^*_h) \le \frac{1}{n}\left(\frac{C\alpha^2 \log n}{h}\right)^{d/2} + \frac{4d}{n}.$$

■

A.2. Bounding the empirical Bayes regret

Lemma 12 *Assume ρ^* is α-subgaussian and has an L-Lipschitz score s^*. Let $0 < \varepsilon \le (2\pi h)^{-d/2}e^{-1/2}$, and assume*

$$\frac{\alpha^2 \log n}{n^{2/d}} \lesssim h \le \frac{1}{4L}.$$

Then

$$\mathbb{E}\|\hat{s}^\varepsilon_h - s^{*\varepsilon}_h\|^2_{\rho^*_h} \le \frac{Cd\,(C_{h,d,\alpha} + d)}{nh}\left[\left(\log\frac{(2\pi h)^{-d/2}}{\varepsilon}\right)^3 + \log\frac{n}{C_{h,d,\alpha} + d}\right].$$

where $C > 0$ is a universal constant and $C_{h,d,\alpha} = \left(\frac{\alpha^2 \log n}{h}\right)^{d/2}$.

Proof First, by Lemma 8, we relate the quantity $\|\hat{s}^\varepsilon_h - s^{*\varepsilon}_h\|^2_{\rho^*_h}$ in terms of the squared Hellinger distance between $\hat{\rho}_h$ and ρ^*_h conditional on the samples:

$$\|\hat{s}^\varepsilon_h - s^{*\varepsilon}_h\|^2_{\rho^*_h} \le \frac{Cd}{h} \max\left\{\left(\log\frac{(2\pi h)^{-d/2}}{\varepsilon}\right)^3, |\log H(\rho^*_h, \hat{\rho}_h)|\right\} H^2(\rho^*_h, \hat{\rho}_h)$$

where $C > 0$ is a universal constant. Note that $|\log H| \leq \log \frac{2}{H}$ since $0 \leq H \leq \sqrt{2}$. Taking expectation over $\mathbf{X}_n$, and using the simple bound $\max(a,b) \leq a+b$ for $a, b \geq 0$, we have:

$$\mathbb{E}\|\hat{s}_h^{\varepsilon} - s_h^{*\varepsilon}\|_{\rho_h^*}^2 \leq \frac{Cd}{h}\left(\left(\log \frac{(2\pi h)^{-d/2}}{\varepsilon}\right)^3 \mathbb{E}H^2(\rho_h^*, \hat{\rho}_h) + \frac{1}{2}\mathbb{E}\left[H^2(\rho_h^*, \hat{\rho}_h)\, \log \frac{4}{H^2(\rho_h^*, \hat{\rho}_h)}\right]\right). \tag{34}$$

By Lemma 9, we have that: If $h \leq \frac{1}{4L}$

$$\mathbb{E}H^2(\rho_h^*, \hat{\rho}_h) \leq \frac{1}{n}\left(\frac{C\alpha^2 \log n}{h}\right)^{d/2} + \frac{4d}{n}. \tag{35}$$

On the other hand, since $x \mapsto x \log \frac{4}{x}$ is concave, by Jensen's inequality

$$\frac{1}{2}\mathbb{E}\left[H^2(\rho_h^*, \hat{\rho}_h) \log \frac{4}{H^2(\rho_h^*, \hat{\rho}_h)}\right] \leq \frac{1}{2}\mathbb{E}H^2(\rho_h^*, \hat{\rho}_h) \log \frac{4}{\mathbb{E}H^2(\rho_h^*, \hat{\rho}_h)}.$$

Recall that $x \mapsto x \log \frac{4}{x}$ is increasing in $(0, 4/e)$. Let

$$C_{h,d,\alpha} \triangleq \left(\frac{\alpha^2 \log n}{h}\right)^{d/2}.$$

If $\frac{C_{h,d,\alpha}}{n} \leq 4/e$, which can be satisfied when $h \gtrsim \alpha^2 n^{-2/d} \log n$, then

$$\frac{1}{2}\mathbb{E}H^2(\rho_h^*, \hat{\rho}_h) \log \frac{4}{\mathbb{E}H^2(\rho_h^*, \hat{\rho}_h)} \leq \frac{C_{h,d,\alpha} + d}{2n} \log \frac{n}{C_{h,d,\alpha} + d}. \tag{36}$$

Therefore, combining (34), (35) and (36), we obtain the desired result

$$\mathbb{E}\|\hat{s}_h^{\varepsilon} - s_h^{*\varepsilon}\|_{\rho_h^*}^2 \leq \frac{Cd\,(C_{h,d,\alpha} + d)}{nh}\left[\left(\log \frac{(2\pi h)^{-d/2}}{\varepsilon}\right)^3 + \log \frac{n}{C_{h,d,\alpha} + d}\right].$$

■

A.3. Bounding the regularization error

The following result bounds the error introduced by the regularization parameter ε. Similar results appeared before in (Jiang and Zhang, 2009, Theorem 3) for 1 dimension and (Saha and Guntuboyina, 2020, Lemma 4.3) for d dimensions, the latter of which is not convenient to apply. Instead, we provide a self-contained improved version following the simple approach in (Shen and Wu, 2022, Sec. 5.2).

Lemma 13 *Let ρ^* be α-subgaussian. Assume that $0 \leq \varepsilon \leq (2\pi h)^{-d/2}/e$ and $h \leq \alpha^2$. Then*

$$\|s_h^{*\varepsilon} - s_h^*\|_{\rho_h^*}^2 \leq \frac{2\varepsilon}{h}(64\alpha^2 \log n)^{d/2} \log \frac{1}{\varepsilon(2\pi h)^{d/2}} + \frac{2d^{3/2}}{hn^2}.$$

Proof We first control the size of the score s_h^*. Let $U \sim \rho_*$ and $X = U + \sqrt{h}Z \sim \rho_h^*$, where $Z \sim \mathcal{N}(0, I_d)$. Recall Tweedie's formula (10), namely

$$s_h^*(x) = \nabla \log \rho_h^*(x) = \frac{1}{h}(\mathbb{E}[U|X = x] - x). \tag{37}$$

Following Jiang and Zhang (2009); Saha and Guntuboyina (2020), applying Jensen's inequality yields

$$\begin{aligned} \|s_h^*(x)\|^2 &\leq \frac{1}{h^2}\mathbb{E}[\|X - U\|^2|X = x] \\ &\leq \frac{2}{h}\log \mathbb{E}[\exp(\|X - U\|^2/(2h))|X = x] = \frac{2}{h}\log\frac{1}{(2\pi h)^{d/2}\rho_h^*(x)} \end{aligned} \tag{38}$$

where the last equality is because the conditional density of U given $X = x$ is $\frac{\exp\left(-\|x-u\|^2/(2h)\right)\rho^*(u)}{(2\pi h)^{d/2}\rho_h^*(x)}$.

Next, as in the proof Lemma 9, set $B = \mu + [-a, a]^d$, where $\mu = \mathbb{E}[U] = \mathbb{E}[X]$ and $a = \sqrt{64\alpha^2 \log n}$. By the same subgaussian tail bound as in (32), we have

$$\mathbb{P}\left[X \notin B\right] \leq \frac{2d}{n^4}. \tag{39}$$

Now we are ready to bound $\|s_h^{*\varepsilon} - s_h^*\|^2_{\rho_h^*}$: Recall from (13) that $s_h^{*\varepsilon} = \frac{\nabla \rho_h^*}{\max(\rho_h^*, \varepsilon)}$. Then

$$\begin{aligned} \|s_h^{*\varepsilon} - s_h^*\|^2_{\rho_h^*} \leq& \mathbb{E}[\|s_h^*(X)\|^2 \mathbf{1}\left\{\rho_h^*(X) \leq \varepsilon\right\}] \\ \leq& \underbrace{\mathbb{E}[\|s_h^*(X)\|^2 \mathbf{1}\left\{\rho_h^*(X) \leq \varepsilon\right\}\mathbf{1}\left\{X \in B\right\}]}_{\text{(I)}} + \underbrace{\mathbb{E}[\|s_h^*(X)\|^2 \mathbf{1}\left\{X \notin B\right\}]}_{\text{(II)}}. \end{aligned}$$

For the first term, applying (38) we get

$$\begin{aligned} \text{(I)} =& \int_B dx\, \rho_h^*(x)\|s_h^*(x)\|^2 \mathbf{1}\left\{\rho_h^*(x) \leq \varepsilon\right\} \\ \leq& \frac{2}{h}\int_B dx\, \rho_h^*(x) \log\frac{1}{(2\pi h)^{d/2}\rho_h^*(x)}\mathbf{1}\left\{\rho_h^*(x) \leq \varepsilon\right\} \\ \leq& (2a)^d \frac{2\varepsilon}{h}\log\frac{1}{\varepsilon(2\pi h)^{d/2}} \end{aligned}$$

where the last inequality follows form the fact that $t \log \frac{1}{t}$ is increasing on $t \in (0, 1/e)$ and the assumption that $\varepsilon(2\pi h)^{d/2} < 1/e$.

For the second term, applying (37)

$$\begin{aligned} \text{(II)} =& \frac{1}{h}\mathbb{E}[\|\mathbb{E}[Z|X]\|^2\, \mathbf{1}\left\{X \notin B\right\}] \\ \leq& \frac{1}{h}\sqrt{\mathbb{E}[\|Z\|^4]\, \mathbb{P}\left[X \notin B\right]} \\ \leq& \frac{2d^{3/2}}{hn^2}. \end{aligned}$$

where the first inequality applies Jensen's inequality and Cauchy-Schwarz inequality, and the second applies Jensen's inequality, $\mathbb{E}[\|Z\|^4] = 2d + d^2$ and (39). ■

A.4. Lipschitzness of the Gaussian mixture score

Lemma 14 *Assume ρ_0 is L-log-smooth. For $t \in (0, \frac{1}{2L}]$, $\rho_t = \rho_0 * \mathcal{N}(0, tI_d)$ is $2L$-log-smooth.*

Proof For $X_0 \sim \rho_0$, let $X_t = X_0 + \sqrt{t}Z$ where $Z \sim \mathcal{N}(0, I_d)$ is independent, so $X_t \sim \rho_t$. For $t > 0$, let ρ_{0t} denote the joint distribution of (X_0, X_t). Let $\rho_{0|t}(\cdot \mid y)$ denote the conditional density of X_0 given $X_t = y$. Similarly, let $\rho_{t|0}(\cdot \mid x)$ denote the conditional density of X_t given $X_0 = x$, and note $\rho_{t|0}(\cdot \mid x) = \mathcal{N}(x, t\,I_d)$ by definition.

By Tweedie's formula (10),

$$s_t(y) = \nabla \log \rho_t(y) = \frac{\mathbb{E}_{\rho_{0|t}}[X \mid y] - y}{t}. \tag{40}$$

We then derive $\nabla^2 \log \rho_t(y)$: Noting that

$$\begin{aligned}
\nabla_y\, \rho_{0|t}(x \mid y) &= \nabla\, \frac{\rho_{t|0}(y \mid x)\rho_0(x)}{\rho_t(y)} \\
&= \frac{\nabla\, \rho_{t|0}(y \mid x)\rho_0(x)}{\rho_t(y)} - \frac{\rho_{t|0}(y \mid x)\rho_0(x)\nabla\rho_t(y)}{\rho_t^2(y)} \\
&= \rho_{0|t}(x \mid y)\frac{x-y}{t} - \rho_{0|t}(x \mid y)\nabla \log \rho_t(y),
\end{aligned}$$

we obtain the gradient of the posterior mean

$$\begin{aligned}
\nabla \mathbb{E}_{\rho_{0|t}}[X \mid y] &= \int \nabla \rho_{0|t}(x \mid y)x^\top dx \\
&= \mathbb{E}_{\rho_{0|t}}\left[\frac{(X-y)X^\top}{t} - \nabla \log \rho_t(y)X^\top \mid y\right] \\
&\overset{(40)}{=} \frac{\mathbb{E}_{\rho_{0|t}}[XX^\top \mid y]}{t} - \frac{\mathbb{E}_{\rho_{0|t}}[X \mid y]\,\mathbb{E}_{\rho_{0|t}}[X \mid y]^\top}{t} \\
&= \frac{\mathrm{Cov}_{\rho_{0|t}}[X \mid y]}{t}.
\end{aligned}$$

It follows that

$$-\nabla^2 \log \rho_t(y) = \frac{I_d}{t} - \frac{\mathrm{Cov}_{\rho_{0|t}}[X \mid y]}{t^2}. \tag{41}$$

We now bound the covariance term. Suppose $\rho_0 \propto e^{-f}$. Recall that $\rho_{0|t}(x \mid y) \propto e^{-f(x) - \frac{1}{2t}\|y-x\|^2}$, thus

$$-\nabla_x^2 \log \rho_{0|t}(x \mid y) = \nabla_x^2 \left(f(x) + \frac{1}{2t}\|y-x\|^2\right) = \nabla^2 f(x) + \frac{1}{t}I_d,$$

(note the derivative above is with respect to x). Since $\nabla^2 f(x) \preceq LI_d$, we have

$$-\nabla_x^2 \log \rho_{0|t}(x \mid y) \preceq \left(L + \frac{1}{t}\right) I_d.$$

This implies (see Lemma 15 below): For any $y \in \mathbb{R}^d$

$$\mathrm{Cov}_{\rho_{0|t}}[X \mid y] \succeq \frac{1}{L + 1/t} I_d.$$

Therefore, we obtain an upper bound of the Hessian matrix (41):

$$-\nabla^2 \log \rho_t(y) \preceq \left(\frac{1}{t} - \frac{1}{t(tL+1)}\right) I_d = \frac{L}{tL+1} I_d. \tag{42}$$

To get a lower bound, we note that since $\nabla^2 f(x) \succeq -LI_d$ for any $x \in \mathbb{R}^d$,

$$-\nabla_x^2 \log \rho_{0|t}(x \mid y) \succeq (-L + \frac{1}{t}) I_d \succeq 0.$$

So for $t < \frac{1}{L}$, $\rho_{0|t}(\cdot \mid y)$ is $(\frac{1}{t} - L)$-strongly log-concave, which implies

$$\mathrm{Cov}_{\rho_{0|t}}[X \mid y] \preceq \frac{1}{1/t - L} I_d.$$

Therefore, for any $y \in \mathbb{R}^d$

$$-\nabla^2 \log \rho_t(y) \succeq \left(\frac{1}{t} - \frac{1}{t(1-tL)}\right) I_d = -\frac{L}{1-tL} I_d. \tag{43}$$

Combining (42) and (43) gives

$$-\frac{L}{1-tL} I_d \preceq -\nabla^2 \log \rho_t(y) \preceq \frac{L}{1+tL} I_d.$$

For $0 \le t < \frac{1}{L}$, $\frac{L}{1-tL} \ge \frac{L}{1+tL}$. Therefore, ρ_t is $\frac{L}{1-tL}$-log-smooth. If $t \le \frac{1}{2L}$, then we have $\frac{L}{1-tL} \le 2L$, so we conclude ρ_t is $2L$-log-smooth for $0 \le t \le \frac{1}{2L}$. ■

Lemma 15 (Brascamp and Lieb (1976)) *Suppose a density ρ on $\mathbb{R}^d$ satisfies $-\nabla^2 \log \rho(x) \preceq LI_d$ for all $x \in \mathbb{R}^d$. Then*

$$\mathrm{Cov}_\rho(X) \succeq \frac{1}{L} I_d.$$

This is a classical result; here we provide an alternate proof based on Fisher information calculation.

Proof Let $\nu = \mathcal{N}(m, C)$ be a Gaussian with the same mean $m = \mathbb{E}_\rho[X]$ and covariance $C = \mathrm{Cov}_\rho(X)$ as ρ. Note $-\nabla \log \nu(x) = C^{-1}(x-m)$. We can compute the relative Fisher information matrix of ρ with respect to ν to be:

$$\begin{aligned}
\tilde{J}_\nu(\rho) &:= \mathbb{E}_\rho\left[\left(\nabla \log \frac{\rho}{\nu}\right)\left(\nabla \log \frac{\rho}{\nu}\right)^\top\right] \\
&= \mathbb{E}_\rho\left[(\nabla \log \rho)(\nabla \log \rho)^\top\right] + \mathbb{E}_\rho\left[(\nabla \log \rho)\left(C^{-1}(x-m)\right)^\top\right] \\
&\quad + \mathbb{E}_\rho\left[\left(C^{-1}(x-m)\right)(\nabla \log \rho)^\top\right] + \mathbb{E}_\rho\left[\left(C^{-1}(x-m)\right)\left(C^{-1}(x-m)\right)^\top\right] \\
&= \mathbb{E}_\rho[-\nabla^2 \log \rho] - C^{-1} - C^{-1} + C^{-1}CC^{-1} \\
&\preceq LI_d - C^{-1}
\end{aligned}$$

where the third equality above holds by integration by parts, and the last inequality holds by L-log-smoothness of ρ. Since $\tilde{J}_\nu(\rho) \succeq 0$, this implies $C^{-1} \preceq LI_d$ or equivalently $C \succeq \frac{1}{L} I_d$, as desired.
■

A.5. Bounding the score error of the Gaussian smoothing

Lemma 16 *Assume that s^* is (L,β)-Hölder continuous for $0<\beta\le 1$: For any $x_1,x_2\in\mathbb{R}^d$*

$$\|s^*(x_1)-s^*(x_2)\|\le L\|x_1-x_2\|^\beta.$$

Then

$$\|s_h^*-s^*\|_{\rho_h^*}^2\le L^2(hd)^\beta.$$

Proof For $X\sim\rho^*$, let $Y=X+\sqrt{h}Z$ where $Z\sim\mathcal{N}(0,I_d)$. Then $Y\sim\rho_h^*$ and $\rho_h^*(y)=\mathbb{E}\rho^*(y-\sqrt{h}Z)$. Moreover,

$$s_h^*(y)=\nabla\log\mathbb{E}\rho^*(y-\sqrt{h}Z)=\frac{\mathbb{E}\left[s^*(y-\sqrt{h}Z)\rho^*(y-\sqrt{h}Z)\right]}{\mathbb{E}\rho^*(y-\sqrt{h}Z)}=\mathbb{E}\left[s^*(X)\mid Y=y\right]. \quad (44)$$

Therefore, for any $y\in\mathbb{R}^d$

$$\|s_h^*(y)-s^*(y)\|^2\le\mathbb{E}\left[\|s^*(X)-s^*(y)\|^2\mid Y=y\right]\le L^2\mathbb{E}\left[\|X-y\|^{2\beta}\mid Y=y\right].$$

So $\|s_h^*-s^*\|_{\rho_h^*}^2\le L^2h^\beta\mathbb{E}[\|Z\|^{2\beta}]\le L^2(hd)^\beta$. ■

Remark 17 *A natural question is whether the score smoothing error can be improved if the true score has higher smoothness than Lipschitz (e.g. β-Hölder for $\beta>1$, which is well-studied in nonparametric statistics (Tsybakov, 2009).) However, Lemma 16 as stated cannot be improved. For an example, consider $\rho^*=\mathcal{N}(0,I_d)$ whose score is $s^*(x)=-x$. Then $\|s_h^*-s^*\|_{\rho_h^*}^2=\Theta(h)$.*

Appendix B. Extensions to Hölder continuous scores

In this appendix we prove Theorem 2 on the estimation of β-Hölder continuous score functions with $0<\beta\le 1$. The proof follows the same program of proving Theorem 1, except that the key Lemma 10 bounding the likelihood ratio of Gaussian convolutions, which relies on Lipschitzness of the score function, needs to be extended. More specifically, Lemma 14, which shows that the score function remains Lipschitz after convolving with sufficiently small Gaussian noise, applies the strong log-concavity of the posterior and the Brascamp-Lieb inequality. While it may be difficult to extend Lemma 14 to less smooth scores, it turns out that we can circumvent the score smoothness of Gaussian convolution in extending Lemma 10. The following result bounds the score difference between the smoothed and the original distributions by applying the score bound in Polyanskiy and Wu (2016).

Lemma 18 *Let s denote the score of p. Suppose s is (L,β)-Hölder continuous for some $0<\beta\le 1$. Let s_h be the score of $p_h=p*\mathcal{N}(0,hI_d)$. Then for any $y\in\mathbb{R}^d$ and $h>0$,*

$$\|s_h(y)-s(y)\|\le 4L(\|y-\mu\|+A)^\beta$$

where $A=\mathbb{E}_{X\sim p}[\|X-\mu\|]$ and $\mu=\mathbb{E}_{X\sim p}[X]$.

Proof Let $Y = X + \sqrt{h}Z$, where $Z \sim N(0, I_d)$ and $X \sim p$ are independent. Recall from (44) that $s_h(y) = \mathbb{E}[s(X) \,|\, Y = y] = \mathbb{E}[s(y - \sqrt{h}Z) \,|\, Y = y]$. Then

$$\begin{aligned}
\|s_h(y) - s(y)\| &\leq \mathbb{E}[\|s(y - \sqrt{h}Z)) - s(y)\| \,|\, Y = y] \\
&\leq L\, \mathbb{E}[\|\sqrt{h}Z\|^\beta \,|\, Y = y] \\
&= L\, \mathbb{E}[\|y - X\|^\beta \,|\, Y = y] \\
&\overset{(a)}{\leq} L\, (\mathbb{E}[\|y - X\| \,|\, Y = y])^\beta \\
&= L\, (\mathbb{E}[\|(y - \mu) - (X - \mu)\| \,|\, Y - \mu = y - \mu])^\beta \\
&\overset{(b)}{\leq} L(3\|y - \mu\| + 4\mathbb{E}[\|X - \mu\|])^\beta
\end{aligned}$$

where (a) is by Jensen's inequality; (b) applies Proposition 2 (in particular, Eq. (16)) in Polyanskiy and Wu (2016) to the random variable $Y - \mu = (X - \mu) + \sqrt{h}Z$. ■

The following lemma is a counterpart for Lemma 10:

Lemma 19 *Let s denote the score of p. Suppose s is (L, β)-Hölder continuous for some $0 < \beta \leq 1$. Then for all $t > 0$,*

$$\log \frac{p}{p_t}(y) \leq L\,(td)^{(1+\beta)/2}. \tag{45}$$

Furthermore, for all $a > 0$, $t > 0$, we have

$$\log \frac{p_a}{p_{a+t}}(y) \leq 5L(td)^{(1+\beta)/2} + 4L\sqrt{td}(\|y - \mu\|^\beta + A^\beta) \tag{46}$$

where $A = \mathbb{E}_{X\sim p}[\|X - \mu\|]$ and $\mu = \mathbb{E}_{X\sim p}[X]$.

Later in the proof of Theorem 2, we will apply (45) with $t = h$ (for change of measure, so that the likelihood ratio is bounded) or (46) with $a = t = h/2$ (for bounding the smoothed empirical distribution, in which case $\|y - \mu\| \lesssim \sqrt{\log n}$ and $h = 1/\mathrm{poly}(n)$ so it is also bounded).

Proof Following the proof of Lemma 10, we have

$$\begin{aligned}
\log \frac{p_a}{p_{a+t}}(y) = &\ \log p_a(y) - \log \mathbb{E}[p_a(y - \sqrt{t}Z)] \\
&\leq \mathbb{E} \int_{-\sqrt{t}}^{0} \langle -Z, s_a(y - uZ) - s(y) \rangle du \\
&= \mathbb{E} \int_{-\sqrt{t}}^{0} \langle -Z, s_a(y - uZ) - s(y - uZ) \rangle du + \mathbb{E} \int_{-\sqrt{t}}^{0} \langle -Z, s(y - uZ) - s(y) \rangle du.
\end{aligned}$$

Using the (L, β)-Hölder continuity of s and Jensen's inequality, the second term is upper bounded by

$$\int_{-\sqrt{t}}^{0} L\mathbb{E}[\|Z\|^{1+\beta}]\,|u|^\beta du = \frac{L}{1+\beta}\mathbb{E}[\|Z\|^{1+\beta}]\,t^{(1+\beta)/2} \leq \frac{L}{1+\beta} d^{(1+\beta)/2} t^{(1+\beta)/2}.$$

(This proves (45) for $a = 0$.) Applying Lemma 18, the first term is bounded by:

$$\begin{aligned}
4L\int_{-\sqrt{t}}^{0}\mathbb{E}[\|Z\|(\|y-\mu-uZ\|+A)^{\beta}]du &\le 4L\sqrt{t}\,\mathbb{E}[\|Z\|(\|y-\mu\|+\sqrt{t}\|Z\|+A)^{\beta}]\\
&\le 4L\sqrt{t}\left((\|y-\mu\|^{\beta}+A^{\beta})\,\mathbb{E}[\|Z\|]+t^{\beta/2}\mathbb{E}[\|Z\|^{1+\beta}]\right)\\
&\le 4L\sqrt{td}(\|y-\mu\|^{\beta}+A^{\beta})+4Lt^{(1+\beta)/2}d^{(1+\beta)/2}.
\end{aligned}$$

Combining the two terms above yields the bound in (46). ■

With the above lemma, we extend Lemma 9 on the smoothed empirical distribution to Hölder-continuous scores.

Lemma 20 *If ρ^* is α-subgaussian and s^* is (L,β)-Hölder continuous for some $0<\beta\le 1$, then for $h\le\frac{1}{4L}$ and $h\le\alpha^2$,*

$$\mathbb{E}H^2(\hat{\rho}_h,\rho_h^*)\le\frac{1}{n}\left(\frac{2\alpha^2\log n}{h}\right)^{d/2}\exp\left(C_1\sqrt{h}+C_2\sqrt{h}(\log n)^{\beta/2}\right)+\frac{4d}{n} \tag{47}$$

where $C_1=9L\alpha^{\beta}d^{\frac{1+\beta}{2}}$ and $C_2=8L\alpha^{\beta}d^{\frac{1+\beta}{2}}$.

Proof The proof of Lemma 20 follows that of Lemma 9, except that at the step (33) we apply the bound (46) from Lemma 19 with $a=t=h/2$. We bound $\|y-\mu\|\le 2\alpha\sqrt{d\log n}$ for all y in the box $B=\mu+[-2\alpha\sqrt{\log n},\,2\alpha\sqrt{\log n}]^d$ where $\mu=\mathbb{E}_{X\sim\rho^*}[X]$, we bound $A=\mathbb{E}_{X\sim\rho^*}[\|X-\mu\|]\le\sqrt{\mathbb{E}_{X\sim\rho^*}[\|X-\mu\|^2]}\le\sqrt{d}\alpha$ since ρ^* is α-subgaussian, and $h^{\frac{1+\beta}{2}}\le\sqrt{h}\alpha^{\beta}$ since $h\le\alpha^2$. ■

We are now ready to complete the proof of Theorem 2:

Proof Following the proof of Theorem 1, we perform a change of measure as in (14), by applying the bound (45) from Lemma 19 to get

$$\mathbb{E}\ell(\hat{s}_h^{\varepsilon},\rho^*)=\mathbb{E}\|\hat{s}_h^{\varepsilon}-s^*\|_{\rho^*}^2\le\exp\left(L(hd)^{(1+\beta)/2}\right)\mathbb{E}\|\hat{s}_h^{\varepsilon}-s^*\|_{\rho_h^*}^2. \tag{48}$$

Since we will choose $h=1/\mathrm{poly}(n)$, the exponential factor in the right-hand side above is bounded by a constant. Following (15), $\mathbb{E}\|\hat{s}_h^{\varepsilon}-s^*\|_{\rho_h^*}^2$ can be decomposed into the same three terms. For the second term, the bound (17) from Lemma 13 still applies as its proof only uses subgaussianity of ρ^*. We can bound the third term by Lemma 16 as follows:

$$\|s_h^*-s^*\|_{\rho_h^*}^2\le L^2(hd)^{\beta}. \tag{49}$$

To bound the first term, we mimic the proof of Lemma 12, which involves two crucial steps: The first step applies Lemma 8, which still holds since it does not rely on any assumption on the score function. The second step involves deriving a Hellinger rate between the Gaussian-smoothed empirical distribution to the population (analogous to Lemma 9), which can be extended as Lemma 20. Since we will choose $h=1/\mathrm{poly}(n)$, the bound in Lemma 20 can be further bounded by

$$\mathbb{E}H^2(\hat{\rho}_h,\rho_h^*)\le\frac{C_2}{n}\left(\frac{\alpha^2\log n}{h}\right)^{d/2}+\frac{4d}{n}, \tag{50}$$

where C_2 is a universal constant. Then the rest of the proof for bounding the first term follows that of Lemma 12 and the same bound holds with a different constant:

$$\mathbb{E}\|\hat{s}_h^{\varepsilon} - s_h^{*\varepsilon}\|_{\rho_h^*}^2 \leq \frac{C_3 d\left(C_{h,d,\alpha}+d\right)}{nh}\left[\left(\log\frac{(2\pi h)^{-d/2}}{\varepsilon}\right)^3 + \log\frac{n}{C_{h,d,\alpha}+d}\right]$$

where C_3 is a universal constant and $C_{h,d,\alpha} = \left(\frac{\alpha^2 \log n}{h}\right)^{d/2}$.

Similar to the proof of Theorem 1, by choosing $\varepsilon = n^{-2}$,

$$\mathbb{E}\|\hat{s}_h^{\varepsilon} - s^*\|_{\rho_h^*}^2 \leq \frac{C_3 d^4 (\alpha^2 \log n)^{d/2}}{n h^{d/2+1}}\left(\log\frac{n}{h}\right)^3 + L^2 (hd)^{\beta}.$$

Optimizing the bound by choosing $h = \left(\frac{d^{4-\beta}(\alpha^2\log n)^{d/2}}{L^2 n}\right)^{\frac{2}{d+2\beta+2}}$ and combining with the change of measure (48), we obtain the desired rate:

$$\mathbb{E}\ell(\hat{s}_h^{\varepsilon}, \rho^*) \leq C d^{\beta} L^2 \alpha^{2\beta} (\log n)^{\frac{d\beta}{d+2\beta+2}} n^{-\frac{2\beta}{d+2\beta+2}}$$

for some universal constant $C > 0$. ■

Appendix C. Proof of Lower Bound (Theorem 3)

Proof [Proof of Theorem 3.] The proof of Theorem 3 follows that of standard minimax lower bounds for nonparametric density estimation, with a few adjustments made for scores. By scaling, we can assume without loss of generality that α is some constant. Fix a reference density $f_0 = \varphi$, the standard normal density in d dimensions. We create a collection of perturbations to f_0 by modifying its values on the unit cube $D = [0,1]^d$. Fix some $\epsilon > 0$ to be specified later. Let $m = \frac{1}{\epsilon}$ and assume m is an integer. Fix some kernel $w : \mathbb{R} \to \mathbb{R}$ satisfying the following conditions:

- w is supported on $[0,1]$, with $\int_0^1 w(x)dx = 0$.
- w is twice differentiable, with $\max\{\|w\|_\infty, \|w'\|_\infty, \|w''\|_\infty\} \leq C$ for some absolute constant C. As a result, w satisfies the periodic boundary conditions $w(0) = w(1) = w'(0) = w'(1) = 0$.

(For example, a concrete choice is a sinusoid kernel, such as $w(x) = (1-\cos(4\pi x))\mathbf{1}\left\{0 \leq x < 1/2\right\} + (\cos(4\pi x) - 1)\mathbf{1}\left\{\frac{1}{2} \leq x \leq 1\right\}$.) We then extend w to $\mathbb{R}^d$ as follows: for every $x = (x_1, \ldots, x_d)$, $w(x) \triangleq \prod_{t=1}^d w(x_t)$. Then w is twice differentiable and supported on D, with bounded gradient and Hessian, satisfying the periodic boundary conditions $w(x) = 0$ and $\nabla w(x) = 0$ for all $x \in \partial D$.

For every $i = 0, \ldots, m-1$, let $x_i = \frac{i}{m}$. For a multi-index $\mathbf{i} = (i_1, \ldots, i_d) \in \mathcal{I} \triangleq \{0, \ldots, m-1\}^d$, let $x_{\mathbf{i}} = (x_{i_1}, \ldots, x_{i_d})$ and $D_{\mathbf{i}} = x_{\mathbf{i}} + \epsilon D$. Then $D = \cup_{\mathbf{i}\in\mathcal{I}} D_{\mathbf{i}}$. For each $b = (b_{\mathbf{i}} : \mathbf{i} \in \mathcal{I}) \in \mathcal{B} \triangleq \{0,1\}^{\mathcal{I}}$, let

$$f_b(x) = f_0(x) + \epsilon^2 \sum_{\mathbf{i}\in\mathcal{I}} b_{\mathbf{i}} w\left(\frac{x - x_{\mathbf{i}}}{\epsilon}\right).$$

Note that provided that ϵ is at most a constant,[5] each f_b satisfies the following:

[5] Here and below all constants depend on d.

- f_b is a valid density on $\mathbb{R}^d$.
- f_b is bounded from above and below by a universal constant on $D = [0,1]^d$ and $f_b = f_0$ outside of D. Thus f_b is α-subgaussian for some constant α.
- f_b is twice differentiable on $\mathbb{R}^d$. Furthermore, $\|\nabla f_b\| \triangleq \sup_{x\in\mathbb{R}^d} \|\nabla f_b(x)\|_\infty$ and $\|\nabla^2 f_b\| \triangleq \sup_{x\in\mathbb{R}^d} \|\nabla^2 f_b(x)\|_\infty$ are at most a constant.
- The score $s_b \triangleq \nabla \log f_b = \frac{\nabla f_b}{f_b}$ satisfies

$$s_b(x) = \begin{cases} -x & x \in D^c \\ \frac{\nabla f_0(x) + \epsilon b_{\mathbf{i}} \nabla w\left(\frac{x - x_{\mathbf{i}}}{\epsilon}\right)}{f_0(x) + \epsilon^2 b_{\mathbf{i}} w\left(\frac{x-x_{\mathbf{i}}}{\epsilon}\right)} & x \in D_{\mathbf{i}} \end{cases} \tag{51}$$

  Furthermore, s_b is L-Lipschitz on $\mathbb{R}^d$ for some constant L. To see this, note that its Jacobian is $Ds_b = \frac{\nabla^2 f_b}{f_b} - \frac{\nabla f_b \nabla f_b^\top}{f_b^2}$. Since f_b is lower bounded by a constant on D and both ∇f_b and $\nabla^2 f_b$ are entrywise upper bounded everywhere, Ds_b is bounded on D and hence s_b is L-Lipschitz on either D or D^c. This implies that s_b is L-Lipschitz on $\mathbb{R}^d$. (Indeed, for every $x \in D$ and $y \in D^c$, there exists $z \in \partial D$ such that $\|x - y\| = \|x - z\| + \|z - y\|$. So $\|s_b(x) - s_b(y)\| \le \|s_b(x) - s_b(z)\| + \|s_b(z) - s_b(y)\| \le L(\|x - z\| + \|z - y\|) = L\|x - y\|$.)

In view of the above properties, we have $\mathcal{F} = \{f_b : b \in \mathcal{B}\} \subset \mathcal{P}_{\alpha,L}$. So

$$\begin{aligned} \inf_{\hat s} \sup_{f\in\mathcal{P}_{\alpha,L}} \mathbb{E}_f \|\hat s - s_f\|^2_{L^2(\mathbb{R}^d, f)} &\ge \inf_{\hat s} \sup_{b\in\mathcal{B}} \mathbb{E}_{f_b} \|\hat s - s_b\|^2_{L^2(\mathbb{R}^d, f_b)} \\ &= \inf_{\hat s} \sup_{b\in\mathcal{B}} \mathbb{E}_{f_b} \|\hat s - s_b\|^2_{L^2(D, f_b)} + \mathbb{E}_{f_b} \|\hat s - s_b\|^2_{L^2(D^c, f_b)} \\ &\overset{(i)}{=} \inf_{\hat s} \sup_{b\in\mathcal{B}} \mathbb{E}_{f_b} \|\hat s - s_b\|^2_{L^2(D, f_b)} \\ &\gtrsim \inf_{\hat s} \sup_{b\in\mathcal{B}} \mathbb{E}_{f_b} \|\hat s - s_b\|^2_{L^2(D)}. \end{aligned} \tag{52}$$

Here the infimum in (i) is over estimators $\hat s(\cdot) = \hat s(\cdot; X_1, \ldots, X_n)$ such that $\hat s(x) = -x$ for $x \in D^c$, $\|\hat s - s_b\|^2_{L^2(D)} \triangleq \int_D dx \|\hat s(x) - s_b(x)\|_2^2$ stands for the unweighted squared L^2-norm on D, and the last inequality holds because $f_b \in \mathcal{F}$ is uniformly lower bounded on D.

After these reductions, the proof proceeds by a standard application of Fano's inequality as follows.

Separation of scores. For any $b, b' \in \mathcal{B}$,

$$\|s_b - s_{b'}\|^2_{L^2(D)} = \sum_{\mathbf{i}\in\mathcal{I}} \int_{D_{\mathbf{i}}} dx \|s_b(x) - s_{b'}(x)\|_2^2 \mathbf{1}\left\{b_{\mathbf{i}} \ne b'_{\mathbf{i}}\right\}.$$

For each $\mathbf{i} \in \mathcal{I}$ such that $b_{\mathbf{i}} \neq b'_{\mathbf{i}}$, say $b_{\mathbf{i}} = 1$ and $b'_{\mathbf{i}} = 0$, applying (51) yields

$$\begin{aligned}
\int_{D_{\mathbf{i}}} dx \|s_b(x) - s_{b'}(x)\|_2^2 &= \int_{D_{\mathbf{i}}} dx \left\| \frac{\nabla f_0(x) + \epsilon \nabla w\left(\frac{x-x_{\mathbf{i}}}{\epsilon}\right)}{f_0(x) + \epsilon^2 w\left(\frac{x-x_{\mathbf{i}}}{\epsilon}\right)} - \frac{\nabla f_0(x)}{f_0(x)} \right\|_2^2 \\
&= \epsilon^d \int_D dy \left\| \frac{\nabla f_0(x_{\mathbf{i}} + \epsilon y) + \epsilon \nabla w(y)}{f_0(x_{\mathbf{i}} + \epsilon y) + \epsilon^2 w(y)} - \frac{\nabla f_0(x_{\mathbf{i}} + \epsilon y)}{f_0(x_{\mathbf{i}} + \epsilon y)} \right\|_2^2 \\
&= \epsilon^d \int_D dy \left\| \frac{\epsilon f_0(x_{\mathbf{i}} + \epsilon y)\nabla w(y) - \epsilon^2 w(y) \nabla f_0(x_{\mathbf{i}} + \epsilon y)}{f_0(x_{\mathbf{i}} + \epsilon y)(f_0(x_{\mathbf{i}} + \epsilon y) + \epsilon^2 w(y))} \right\|_2^2 \\
&\asymp \epsilon^{d+2}
\end{aligned}$$

where the last step applies again the facts that f_0 and w are bounded from above and below on D and $\|\nabla w\|$ is bounded from above on D. So for sufficiently small ϵ, the numerator of the integrand scales as $\Theta(\epsilon^2)$, whereas the denominator scales as $\Theta(1)$. Thus

$$\|s_b - s_{b'}\|_{L^2(D)}^2 \asymp \epsilon^{d+2} d_{\mathrm{H}}(b, b')$$

where $d_{\mathrm{H}}(b, b') = \sum_{\mathbf{i} \in \mathcal{I}} \mathbf{1}\left\{b_{\mathbf{i}} \neq b'_{\mathbf{i}}\right\}$ is the Hamming distance.

Next, by the Gilbert-Varshamov bound (see e.g. (Tsybakov, 2009, Lemma 2.9)), there exists an exponentially large packing $\mathcal{B}' \subset \mathcal{B}$, whose minimum Hamming distance is linear in the dimension, namely, $|\mathcal{B}'| \geq \exp(c_0 m^d)$ and $\min_{b \neq b \in \mathcal{B}'} d_{\mathrm{H}}(b, b') \geq c_0 m^d$ for some universal constant c_0. Recalling that $m = 1/\epsilon$, we have

$$\min_{b \neq b \in \mathcal{B}'} \|s_b - s_{b'}\|_{L^2(D)}^2 \asymp \epsilon^2. \tag{53}$$

Bounding the KL radius. For every $b \in \mathcal{B}$, the Kullback-Leibler divergence satisfies

$$\begin{aligned}
\mathrm{KL}(f_b \| f_0) &\leq \chi^2(f_b \| f_0) \\
&= \int_{\mathbb{R}^d} dx \frac{(f_b(x) - f_0(x))^2}{f_0(x)} \\
&= \int_D dx \frac{(f_b(x) - f_0(x))^2}{f_0(x)} \\
&\asymp \int_D dx (f_b(x) - f_0(x))^2 \\
&\leq \sum_{\mathbf{i} \in \mathcal{I}} \int_{D_{\mathbf{i}}} dx \left(\epsilon^2 w\left(\frac{x - x_{\mathbf{i}}}{\epsilon}\right)\right)^2 \\
&\overset{(i)}{\asymp} \epsilon^4
\end{aligned}$$

where (i) is because

$$\sum_{\mathbf{i} \in \mathcal{I}} \int_{D_i} w^2\left(\frac{x - x_{\mathbf{i}}}{\epsilon}\right) dx = \sum_{\mathbf{i} \in \mathcal{I}} \epsilon^d \int_D w^2(y)\, dy = \sum_{\mathbf{i} \in \mathcal{I}} \epsilon^d \left(\int_0^1 w^2(y_i)\, dy_i\right)^d \asymp m^d \epsilon^d = 1.$$

Since the observations are i.i.d., we get $\max_{b\in\mathcal{B}} \mathrm{KL}(f_b^{\otimes n}\|f_0^{\otimes n}) \lesssim n\epsilon^4$. By choosing $\epsilon = cn^{-\frac{1}{d+4}}$ for sufficiently small constant c, we get $n\epsilon^4 \leq c|\mathcal{B}'|$. In view of (53), applying Fano's inequality (see e.g. (Tsybakov, 2009, Corollary 2.6)) yields

$$\inf_{\hat{s}} \sup_{b\in\mathcal{B}'} \mathbb{E}_f \|\hat{s} - s_b\|^2_{L^2(D)} \gtrsim n^{-\frac{2}{d+4}},$$

which, together with (52), implies the desired lower bound (20).

■

Appendix D. Proof of Upper Bound in DDPM (Theorem 5)

Proof [Proof of Theorem 5] By triangle inequality, we can decompose $L^2(\nu_t)$ error into two components

$$\mathbb{E}\|\hat{s}_t^\varepsilon - s_t\|^2_{\nu_t} \leq 2\mathbb{E}\|\hat{s}_t^\varepsilon - s_t^\varepsilon\|^2_{\nu_t} + 2\|s_t^\varepsilon - s_t\|^2_{\nu_t}. \tag{54}$$

The first term can be bounded in the same manner as the proof of Lemma 12, as follows. First, we bound $\|\hat{s}_t^\varepsilon - s_t^\varepsilon\|^2_{\nu_t}$ by Lemma 8: If $0 < \varepsilon \leq (2\pi\tau(t))^{-d/2}e^{-1/2}$, then

$$\|\hat{s}_t^\varepsilon - s_t^\varepsilon\|^2_{\nu_t} \leq \frac{Cd}{\tau(t)} \max\left\{\left(\log \frac{(2\pi\tau(t))^{-d/2}}{\varepsilon}\right)^3, |\log H(\hat{\nu}_t, \nu_t)|\right\} H^2(\hat{\nu}_t, \nu_t). \tag{55}$$

Let $X \sim \nu_0$ and $X' \sim \hat{\nu}_0$. Note that

$$H^2(\hat{\nu}_t, \nu_t) = H^2(\mathrm{Law}(X + \sqrt{e^{2t}-1}\,Z), \mathrm{Law}(X' + \sqrt{e^{2t}-1}\,Z))$$

where $Z \sim \mathcal{N}(0, I_d)$ and it is independent of X, X'. Hence by data processing inequality, $H^2(\hat{\nu}_t, \nu_t)$ is decreasing in t. It follows that

$$H^2(\hat{\nu}_t, \nu_t) \leq H^2(\hat{\nu}_\eta, \nu_\eta)$$

where η is the step size in DDPM (24).

Note that since X is α-subgaussian, Law$(e^{-\eta}X)$ is $(e^{-\eta}\alpha)$-subgaussian. Then by Lemma 9 we have: If $\eta \leq \frac{1}{2}\log\left(1 + \frac{1}{4L-1}\right)$,

$$\mathbb{E}H^2(\hat{\nu}_t, \nu_t) \leq \frac{1}{n}\left(\frac{\alpha^2 \log n}{e^{2\eta}-1}\right)^{d/2} + \frac{4d}{n}. \tag{56}$$

Therefore, similar to (34)-(36), we obtain

$$\mathbb{E}\|\hat{s}_t^\varepsilon - s_t^\varepsilon\|^2_{\nu_t} \leq \frac{Cd}{\tau(t)} \frac{C_{\eta,d,\alpha} + d}{n}\left[\left(\log \frac{1}{\varepsilon(2\pi\tau(t))^{d/2}}\right)^3 + \log \frac{n}{C_{\eta,d,\alpha} + d}\right] \tag{57}$$

where $C_{\eta,d,\alpha} = \left(\frac{\alpha^2 \log n}{e^{2\eta}-1}\right)^{d/2}$, if $\eta \gtrsim \frac{1}{2}\log\left(\frac{\alpha^2 \log n}{n^{2/d}} + 1\right)$.

The second term in (54) can be bounded by Lemma 13: If $0 \leq \varepsilon \leq (2\pi\tau(t))^{-d/2}/e$, then

$$\|s_t^\varepsilon - s_t\|_{\nu_t}^2 \leq \frac{2\varepsilon}{\tau(t)}(64e^{-2t}\alpha^2 \log n)^{d/2} \log \frac{1}{\varepsilon(2\pi\tau(t))^{d/2}} + \frac{2d^{3/2}}{\tau(t)n^2}. \tag{58}$$

Combining (57) and (58), and taking $\varepsilon = n^{-2}$, for $n = \Omega(e^d)$ we have

$$\mathbb{E}\|\hat{s}_t^\varepsilon - s_t\|_{\nu_t}^2 \leq \frac{Cd}{\tau(t)} \frac{C_{\eta,d,\alpha} + d}{n} \left(\log \frac{n}{(2\pi\tau(t))^{d/4}}\right)^3.$$

■