Entropic characterization of optimal rates for learning Gaussian mixtures

Zeyu Jia ZYJIA@MIT.EDU
Massachusetts Institute of Technology

Yury Polyanskiy YP@MIT.COM
Massachusetts Institute of Technology

Yihong Wu YIHONG.WU@YALE.EDU
Yale University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We consider the question of estimating multi-dimensional Gaussian mixtures (GM) with compactly supported or subgaussian mixing distributions. Minimax estimation rate for this class (under Hellinger, TV and KL divergences) is a long-standing open question, even in one dimension. In this paper we characterize this rate (for all constant dimensions) in terms of the metric entropy of the class. Such characterizations originate from seminal works of Le Cam (1973); Birgé (1983); Haussler and Opper (1997); Yang and Barron (1999). However, for GMs a key ingredient missing from earlier work (and widely sought-after) is a comparison result showing that the KL and the squared Hellinger distance are within a constant multiple of each other uniformly over the class. Our main technical contribution is in showing this fact, from which we derive entropy characterization for estimation rate under Hellinger and KL. Interestingly, the sequential (online learning) estimation rate is characterized by the global entropy, while the single-step (batch) rate corresponds to local entropy, paralleling a similar result for the Gaussian sequence model recently discovered by Neykov (2022) and Mourtada (2023). Additionally, since Hellinger is a proper metric, our comparison shows that GMs under KL satisfy the triangle inequality within multiplicative constants, implying that proper and improper estimation rates coincide.

Keywords: KL divergence, Hellinger distances, Gaussian mixtures, estimation rates

1. Introduction

Gaussian mixtures are among the most popular and useful classes of distributions for modeling real data with heterogeneity. Specifically, each d-dimensional *mixing distribution* π induces a Gaussian *mixture* f_π, which is the convolution of π with the d-dimensional standard Gaussian distribution $\mathcal{N}(0, I_d)$, namely

$$f_\pi(x) = (\pi * \varphi)(x) = \int_{\mathbb{R}^d} \varphi(x - z)\pi(dz),$$

where $\varphi(z) = \frac{1}{\sqrt{2\pi}^d} \exp\left(-\|z\|_2^2/2\right)$ is the standard normal density.

There is a vast literature in statistics and machine learning on various aspects of mixture models such as parameter estimation and clustering. In this paper we focus on learning the mixture model in the sense of density estimation. To this end, it is necessary to impose tail conditions on the mixing distribution. Specifically, we consider two classes of Gaussian mixtures classes, wherein the mixing distribution is either compactly supported or subgaussian.

To measure the density estimation error, it is common to use f-divergences, notably, Kullback-Leibler (KL) divergence $\mathrm{KL}(f\|g) = \int f \log \frac{f}{g}$, the squared Hellinger distance $H^2(f,g) = \int(\sqrt{f} - \sqrt{g})^2$, and the total variation distance $\mathrm{TV}(f,g) = \frac{1}{2}\int |f-g|$. In this paper, we are chiefly concerned with mainly focus on the estimation rates under the KL-divergence and the squared Hellinger distance, as opposed to the L_2 distance due to its lack of operational meaning.[1]

Estimating Gaussian mixture densities is a classical topic in nonparametric statistics. Under the squared Hellinger loss, the minimax lower bound $\Omega((\log n)^d/n)$ was proved in Kim (2014); Kim and Guntuboyina (2022) the lower bound for subgaussian mixing distributions. On the constructive side, nonparametric maximum likelihood estimator (NPMLE) and sieve MLE have been analyzed in van de Geer (1993); Wong and Shen (1995); van de Geer (1996); Genovese and Wasserman (2000); Ghosal and van der Vaart (2001, 2007); Zhang (2009). In particular, the NPMLE, which offers a practical algorithm for properly learning the mixture model, is shown to achieve a near-parametric rate of $O((\log n)^2/n)$ for the subgaussian case Zhang (2009), which is subsequently generalized to $O(\log^{d+1} n/n)$ in d dimensions Saha and Guntuboyina (2020). Similar results for compactly supported mixing distribution are also obtained in (Polyanskiy and Wu, 2021, Theorem 20) Despite these advances, determining the optimal rate remains a long-standing open question even in one dimension.

Departing from maximum likelihood, there is a long line of work that aims at characterizing density estimation rates in terms of metric entropy of the class. These general entropic upper bounds originate from the seminal work of Le Cam (1973); Birgé (1983); Birgé (1986) for the Hellinger loss, Yatracos (1985) for the TV loss, and Yang and Barron (1999) for the KL loss. (We refer to in (Polyanskiy and Wu, 2022+, Chapter 33) for a detailed exposition on these results.) On the other hand, entropic lower bounds for KL and Hellinger losses were established for both the batch Haussler and Opper (1997) and sequential estimation Yang and Barron (1999). However, these entropy-based upper and lower bounds in general do not match unless extra conditions are imposed on the behavior on the model class (those conditions are satisfied, most notably, for the Hölder density class on $[0,1]^d$). Notably, a simple condition that ensures a sharp entropic determination of the minimax rate is the comparability of the Hellinger and KL divergence, namely, for any density f and g in the model class:

$$\mathrm{KL}(f\|g) \asymp H^2(f,g) \tag{1}$$

where $\asymp$ denotes equality within constant multiplicative factors. Note that the one-sided inequality $\mathrm{KL}(f\|g) \geq H^2(f,g)$ is always true cf. e.g. (Polyanskiy and Wu, 2022+, Eq. (7.30)). As such, whenever KL is dominated by H^2, the sharp minimax rate is determined by the local Hellinger entropy of the model class.

Indeed, this entropy-based approach has been successfully taken in Doss et al. (2020) to determine the sharp rate for the special case of *finite-component* Gaussian mixtures in general dimensions. Specifically, for the class of k-component GMs, (Doss et al., 2020, Theorem 4.2) shows that

$$\mathrm{KL}(f_\pi\|f_\eta) \asymp_k H^2(f_\pi, f_\eta), \tag{2}$$

1. Indeed, for densities supported on the entire real line, it is possible that two densities are arbitrarily close in L_2 distance but separated by a large TV distance and hence easily distinguishable. In fact, for the entire class of Gaussian mixtures, Kim (2014) showed ignoring the mixture structure and simply applying the kernel density estimator designed for analytic densities Ibragimov (2001) achieves the optimal rate in L_2. On the other hand, consistent estimation of Gaussian mixtures in more meaningful loss function such as TV is impossible unless tail conditions on the mixing distribution are imposed.

where π and η are k-atomic distributions supported on a Euclidean ball of bounded radius in $\mathbb{R}^d$ and $\asymp_k$ hides constants depending only on k. The proof of this result is based on the method of moments which shows both distances are proportional to the Euclidean distance between the moment tensors of mixing distributions up to degree $2k-1$. The crucial part of (2) is that it does not depend on the ambient dimension d. As such, this allows the optimal squared Hellinger rate to be determined by the local entropy, which, in turn, can be tightly estimated via the low rank of the moment tensor, leading to the sharp rate of $\Theta_k(\frac{d}{n})$ that holds even in high dimensions. On the other hand, (2) is not fully dimension-free in that the proportionality constant therein is in fact *exponential* in k, the number of components, a limitation of the moment-based approach. As such, it is unclear whether (2) continues to hold for continuous GMs even in one dimension.

We review related results on upper-bounding the KL divergence by Hellinger distance. (Birgé and Massart, 1998, Lemma 5) shows that $D_{KL}(f\|g) \lesssim H^2(f,g)$ if $\operatorname{ess\,sup} \frac{df}{dg} < \infty$. This results was further generalized to α-generalized Hellinger divergence in (Sason and Verdú, 2016, Theorem 9). However, ratios between two Gaussian mixture densities are not bounded. (Wong and Shen, 1995, Theorem 5) points out that if $\int_{f/g \geq \exp(1/\delta)} f^{\delta+1}/g^{\delta} < \infty$ for some $\delta > 0$, then we have $D_{KL} \lesssim H^2 \log(1/H^2)$. This method was extended by Haussler and Opper (1997) and we also use it in our Theorem 3. Note, however, that this method is unable to produce a linear upper bound: $D_{KL}(f\|g) \lesssim H^2(f,g)$. Yet another result follows by choosing $\eta = 1/2$ and $\bar{\eta} = 1$ in (Grünwald and Mehta, 2020, Lemma 13), which proves $D_{KL} \leq c_u H^2(f,g)$ with $c_u = \frac{u+2}{c}$ provided that $f, g \in \mathcal{F}$ and $\mathcal{F}$ satisfies the so-called (u,c)-witness condition, i.e. $\int f \log(f/g) \mathbf{1}_{f/g \leq \exp(u)} \geq c \cdot \int f \log(f/g)$. However, Gaussian mixtures again do not satisfy this condition. In particular, the left side of the above inequality can even be negative in some cases.[2]

In this paper we resolve the question of KL to Hellinger comparison and show that with a constant factor that depends (at most linearly) on the dimension, by proving that

$$\mathrm{KL}(f_\pi \| f_\eta) \asymp_d H^2(f_\pi, f_\eta), \tag{3}$$

where π and η are arbitrary distributions supported on a bounded ball in $\mathbb{R}^d$; furthermore, this result can be made dimension-free with an extra logarithmic factor. In addition, we show that (3) holds for $(1-\epsilon)$-subgaussian mixing distributions but fails for $(1+\epsilon)$-subgaussian distributions. Curiously, our method does not rely on comparing moments of mixing measures, the prevailing method for analyzing statistical distances between mixture distributions (cf. e.g. Wu and Yang (2020a,b); Bandeira et al. (2020); Fan et al. (2021); Doss et al. (2020); Chen and Niles-Weed (2021)).

The new comparison result has various statistical consequences, of which we report here one (see Corollary 11). To estimate the GM density with compactly supported or $(1-\epsilon)$-subgaussian mixing distributions based on an iid sample of size n, the minimax proper or improper density estimation risks under KL divergence or squared Hellinger distance are tightly characterized by the *local* Hellinger entropy of the density class, thereby reducing the question of optimal rates to that of computing the local entropy. Furthermore, the minimax risks in the sequential version (as opposed to the batch setting above) of this problem are tightly characterized by the *global* Hellinger entropy of the class. A similar phenomenon of local-vs-global entropy has been observed in a pair of recent works on Gaussian sequence model: Neykov (2022) showed that batch risk is controlled by the local entropy and Mourtada Mourtada (2023) showed that sequential risk is controlled by the global entropy.

2. To see this, consider $f = \mathcal{N}(0,1)$ and $g = \mathcal{N}(-\delta, 1)$ with $\delta > 0$. Then $\int f \log(f/g) \mathbf{1}_{f/g \leq \exp(u)} = \frac{\delta^2}{2} - \frac{1}{\sqrt{2\pi}} \delta \exp(-u^2/2)$. Hence for any choice of u, there always exists a δ close to zero such that this integral is negative.

Notation Let $B_2(r) = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le r\right\}$ denote the Euclidean ball of radius r centered at 0. Denote by $\mathtt{supp}(\pi)$ the support of a probability measure π. We say a distribution π on $\mathbb{R}^d$ is K-subgaussian if for $X \sim \pi$,

$$\mathbf{P}[\|X\|_2 > t] \le \exp\left(-\frac{t^2}{2K^2}\right), \quad \forall t \ge 0.$$

Organization The rest of the paper is organized as follows. Section 2 states our main results by providing upper bounds on KL divergence according to squared Hellinger for Gaussian mixtures. To illustrate the main ideas, the proof for one dimension is provide in Section 3 as a warm-up. The dimension-free bound for Gaussian mixtures where mixing distribution is compactly supported is provided in Section 4. Finally, the proof of Corollary 11, showing that the estimation rates are tightly characterized by Hellinger entropies, is given in Section 5. Proofs of other results are deferred to appendices.

2. Main Results

Before discussing their statistical consequences, we first state the main comparison results that control the KL divergence between Gaussian mixtures using their Hellinger distance.

For compactly supported mixing distributions, our main result is as follows:

Theorem 1 *Let π and η be supported on $B_2(M)$ in $\mathbb{R}^d$ where $M \ge 2$. Then*

$$\mathrm{KL}(f_\pi \| f_\eta) \le 5154(M^2 \vee d) H^2(f_\pi, f_\eta).$$

In Section 3 we provide a proof of Theorem 1 in one dimension. The proof of general cases is included in Appendix A.

Remark 2 *The bound in Theorem 1 is tight up to constant factors depending on the dimension d. To see this, consider $\pi = \delta_{\mathbf{u}}$ where $\mathbf{u} = (M, 0, 0, \cdots, 0)$ and $\eta = \delta_{-\mathbf{u}}$. Then we have $f_\pi = \mathcal{N}(\mathbf{u}, I)$ and $f_\eta = \mathcal{N}(-\mathbf{u}, I)$. A direct computation shows that $\mathrm{KL}(f_\pi \| f_\eta) = \frac{\|\mathbf{u}-(-\mathbf{u})\|^2}{2} = 2M^2$, while $H^2(f_\pi, f_\eta) = 2 - 2\exp\left(-\frac{M^2}{2}\right) \le 2$.*

Complementing Theorem 1, we also have the following dimension-free upper bound at the price of a mere logarithmic factor. This theorem is a direct corollary of (Wong and Shen, 1995, Theroem 5), if we notice that f_π, f_η satisfy their condition $\int_{f_\pi/f_\eta \ge e} f_\pi(f_\pi/f_\eta) \le \exp(4M^2) < \infty$ for any π, η supported on $B_2(M)$. For completeness, we include a proof in Section 4:

Theorem 3 *Let π and η be supported on $B_2(M)$ in $\mathbb{R}^d$ where $M \ge 1$. Then*

$$\mathrm{KL}(f_\pi \| f_\eta) \le 200M^2 H^2(f_\pi, f_\eta) + 16H^2(f_\pi, f_\eta) \log \frac{1}{H^2(f_\pi, f_\eta)}.$$

Next we consider the class of subgaussian mixing distributions. We discover a dichotomy depending on the subgaussian constant K: When $K < 1$, the KL divergence is indeed proportional to the squared Hellinger distance. When $K > 1$, such upper bound does not exist.

Theorem 4 *Let π, η be two d-dimensional K-subgaussian distributions where $K < 1$. Then*

$$\mathrm{KL}(f_\pi \| f_\eta) \le 1660056 \left(\frac{1}{(1-K)^3} \vee 8d^3 \right) H^2(f_\pi, f_\eta).$$

Theorem 5 *Fix $K > 1$. For any $C > 0$, there exists a 1-dimensional K-subgaussian distribution π such that*

$$\mathrm{KL}(f_\pi \| \mathcal{N}(0,1)) \ge C \cdot H^2(f_\pi, \mathcal{N}(0,1)).$$

Remark 6 *Notice that this phenomenon of dichotomy of $K > 1$ and $K < 1$ for Gaussian mixtures with K-subgaussian mixing distribution was also observed in Block et al. (2022). Therein, it is shown that the convergence rate of smoothed n-point empirical distribution to the smoothed population distribution under Wasserstein distance is $O(1/\sqrt{n})$ for $K < 1$, and $\omega(1/\sqrt{n})$ for $K > 1$.*

We further have the following dimension-free upper bound that holds for all $K > 0$.

Theorem 7 *Let π and η be K-subgaussian distributions on $\mathbb{R}^d$. Then*

$$\mathrm{KL}(f_\pi \| f_\eta) \le (10240K^4 + 652) H^2(f_\pi, f_\eta) \log \frac{4}{H^2(f_\pi, f_\eta)}.$$

The results presented so far are structural results on the information geometry of Gaussian mixture, whose proof are included in Appendix A-D. Next we discuss their statistical consequences. We start with the definition of covering/local covering number and minimax risks of density estimation.

Definition 8 (Covering Number and Local Covering Number) *Let $\mathcal{P}$ be a set of distributions over some measurable space $\mathcal{X}$. The Hellinger covering number of $\mathcal{P}$ is*

$$\mathcal{N}_H(\mathcal{P}, \epsilon) \triangleq \min \left\{ N : \exists Q_1, \cdots, Q_n \in \Delta(\mathcal{X}), \ \sup_{P \in \mathcal{P}} \inf_{1 \le i \le N} H(P, Q_i) \le \epsilon \right\},$$

where $\Delta(\mathcal{X})$ denotes the collection of all probability distributions on $\mathcal{X}$. The local Hellinger covering number of $\mathcal{P}$ is

$$\mathcal{N}_{loc,H}(\mathcal{P}, \epsilon) \triangleq \sup_{P \in \mathcal{P}, \eta \ge \epsilon} \mathcal{N}_H(B_H(P, \eta) \cap \mathcal{P}, \eta/2),$$

where $B_H(P, \eta)$ is the Hellinger ball of radius η centered at P.

We further define the minimax risks for proper and improper density estimation as well as the minimax risk in a sequential setting.

Definition 9 (Proper and Improper Density Estimation Minimax Risk) *For a given class $\mathcal{P}$ of distributions over $\mathcal{X}$, we define the improper minimax risk $R_{H^2,n}, R_{KL,n}$ and the proper minimax risk $\tilde{R}_{KL,n}$ with sample size n as follows: for $d \in \{H^2, \mathrm{KL}\}$, we define*

$$R_{d,n}(\mathcal{P}) \triangleq \inf_{\hat{f}_n} \sup_{f \in \mathcal{P}} \mathbb{E}_f \left[d(f, \hat{f}_n) \right],$$

and also[3]

$$\tilde{R}_{KL,n}(\mathcal{P}) \triangleq \inf_{\hat{f}_n \in \mathcal{P}} \sup_{f \in \mathcal{P}} \mathbb{E}_f \left[\mathrm{KL}(f \| \hat{f}_n) \right],$$

where $\hat{f}_n(\cdot) = \hat{f}_n(\cdot; X_1, \ldots, X_n)$ *is a density estimator based on* $X_1, \ldots, X_n$ *drawn iid from* P.

Definition 10 (Sequential Density Estimation Minimax Risk (Improper)) *For a given class* $\mathcal{P}$ *of distributions over* $\mathcal{X}$*, the sequential minimax risks* $C_{H^2,n}$ *and* $C_{KL,n}$ *are defined as: for* $d \in \{H^2, \mathrm{KL}\}$,

$$C_{d,n}(\mathcal{P}) = \inf_{\hat{f}_1, \cdots, \hat{f}_n} \sup_{f \in \mathcal{P}} \sum_{t=1}^{N} \mathbb{E}[d(f, \hat{f}_t(\cdot | X_1, \cdots, X_{t-1}))]$$

where $\hat{f}_t : \mathcal{X}^{t-1} \to \Delta(\mathcal{X})$ *denotes the density estimator at time* t *based on observations* $X_1, \cdots, X_{t-1}$.

We refer to Definitions 9 and 10 as the batch and online settings, respectively. The following corollary shows that the minimax density estimation risks in these settings can be characterized by the local and global Hellinger entropy up to constant factors. Furthermore, we show that proper and improper density estimation rates coincide. As explained earlier this is well-known for Hellinger loss but far from clear for KL loss which does not satisfy triangle inequality. In fact, the celebrated Yang-Baron construction Yang and Barron (1999) produces an improper density estimate. Nevertheless, we show that for Gaussian mixture class there is no gain in stepping outside the model class.

Corollary 11 *Let* $\mathcal{P}_{com}(M)$ *and* $\mathcal{P}_{sub}(K)$ *denote the collection of* d*-dimensional Gaussian mixtures where the mixing distribution is supported on* $B_2(M)$ *and* K*-subgaussian, respectively, i.e.,*

$$\mathcal{P}_{com}(M) = \{\pi * \mathcal{N}(0, I_d) | \mathrm{supp}(\pi) \subset B_2(M)\},$$
$$\mathcal{P}_{sub}(K) = \{\pi * \mathcal{N}(0, I_d) | \pi \textit{ is } K\textit{-subgaussian}\}.$$

Then for any compact (under Hellinger) subset $\mathcal{P}$ *where* $\mathcal{P} \subset \mathcal{P}_{com}(M)$ *or* $\mathcal{P} \subset \mathcal{P}_{sub}(K)$ *with* $K < 1$*, we have the following characterization on the proper or improper minimax risk:*

$$R_{H^2,n}(\mathcal{P}) \asymp R_{KL,n}(\mathcal{P}) \asymp \tilde{R}_{KL,n}(\mathcal{P}) \asymp \inf_{\epsilon > 0} \epsilon^2 + \frac{1}{n} \log \mathcal{N}_{loc,H}(\mathcal{P}, \epsilon),$$

and also for sequential minimax risk:

$$C_{H^2,n}(\mathcal{P}) \asymp C_{KL,n}(\mathcal{P}) \asymp \inf_{\epsilon > 0} n\epsilon^2 + \log \mathcal{N}_H(\mathcal{P}, \epsilon).$$

Here $\asymp$ *hides constants that may depend on* M, K*, or* d *but not on* n.

As we mentioned previously, a recent pair of works Neykov (2022); Mourtada (2023) established the same phenomenon: the sequential rate is given by global entropy, while the batch rate is given by the local entropy, though, their work is for a very different setting of a Gaussian sequence model.

Apart from the KL divergence and Hellinger distance, we also obtained comparison results for other distances between distributions, e.g. χ^2-divergence, TV and L_2 distances. See Appendix E.

We close this section with a list of related open problems.

3. Since Hellinger distance is a valid metric, for proper and improper density estimation, the minimax squared Hellinger risks coincide within a factor of four, as any estimator can be made proper by its Hellinger projection on the model class.

1. **Fully dimension-free comparison:** Currently our upper bound on KL/χ_2 according to Hellinger is depending on the dimension of the distributions, can we remove this dependence on the dimension? Suppose π and η are two d-dimensional distributions supported on $B_2(M)$, is there a constant $C_{KL}(M), C_{\chi_2}(M)$ such that

$$\mathrm{KL}(f_\pi \| f_\eta) \le C_{KL}(M) \cdot H^2(f_\pi, f_\eta),$$

which would have the best of both worlds of (2) and (3). Note that from Theorem 4.2 in Doss et al. (2020) we can obtain the following dimension-free bound

$$\mathrm{KL}(f_\pi \| f_\eta) \lesssim e^{Ck^2} H^2(f_\pi, f_\eta)$$

for some constant C, if we assume π and η are k-atomic distributions. But this bound depending exponentially on the number of components.

2. **Minimax rate for estimating Gaussian mixtures:** Find the sharp rate of

$$R_{H^2,n}(\mathcal{P}_{com}(M)) = \inf_{\hat{f}_n} \sup_{f \in \mathcal{P}_{com}(M)} \mathbb{E}_f[H^2(\hat{f}_n, f)].$$

Thanks to the comparison inequality in Theorem 1, Corollary 11 reduces this problem to computing the local Hellinger entropy of the mixture class $\mathcal{P}_{com}(M)$. The best known estimates for this in one dimension are

$$\log(1/\epsilon) \lesssim \log \mathcal{N}_{loc,H}(\mathcal{P}_{com}(M), \epsilon) \lesssim \left(\frac{\log(1/\epsilon)}{\log\log(1/\epsilon)}\right)^{3/2}.$$

Here the lower bound is from Theorem 1.3 in Kim (2014), which shows that $R_{H^2,n} = \Omega(\log n/n)$; the upper bound is from Nie and Wu (2021) by constructing a covering of the truncated moment space of the mixing distributions. The upper bound leads to an upper bound $O\left((\log n/\log\log n)^{1.5}/n\right)$ on the minimax risk, improving the $O((\log n)^2/n)$ result of Kim (2014).

3. **Linear comparison between TV and Hellinger:** It is well-known that $H^2 \lesssim \mathrm{TV} \lesssim H$ in general. Can we show that $\mathrm{TV} \asymp H$ for Gaussian mixtures? Specifically, for any two π and η supported on $[-M, M]$, can we show that there exists some constant $C = C(M)$ such that

$$\mathrm{TV}(f_\pi, f_\eta) \ge C \cdot H(f_\pi, f_\eta).$$

We notice that it is impossible to lower bound the L_2-distance $\|f_\pi - f_\eta\|_2$ linearly in $H(f_\pi, f_\eta)$, because Kim (2014) showed that for subgaussian mixing distributions, the minimax squared L_2 risk for estimating the mixture density is at most $O(\sqrt{\log n}/n)$ and the squared Hellinger risk is at least $\Omega(\log n/n)$. Thus the best comparison between L_2 and H will involve log factors. Similarly, the best known comparisons for L_2 and TV, which we derive in Section E, also involve log-factors. It is an open problem to find tight log-factors in these comparisons of L_2, H and TV.

3. Proof of Theorem 1 in one dimension

In this section, we provide a proof of Theorem $\mathrm{KL} \lesssim H^2$ for one-dimensional Gaussian mixtures where the mixing distribution is compactly supported. Similar proof techniques can also be applied in multiple dimensions; see Appendix A for the proof of Theorem 1 in general dimensions.

Theorem 12 (One-dimensional version of Theorem 1) *Let π and η be supported on $B_2(M)$ in $\mathbb{R}$ where $M \geq 2$. Then*

$$\mathrm{KL}(f_\pi \| f_\eta) \leq 1563 M^2 H^2(f_\pi, f_\eta).$$

For simplicity we abbreviate $f_\pi(\cdot)$ and $f_\eta(\cdot)$ as $p(\cdot), q(\cdot)$. Then we have

$$\mathrm{KL}(p\|q) = \int_{-\infty}^{\infty} q(x) \cdot \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} dx, \quad H^2(p,q) = \int_{-\infty}^{\infty} q(x) \cdot \left(\sqrt{\frac{p(x)}{q(x)}} - 1 \right)^2 dx.$$

We first state several lemmas. The first is a straightforward computation.

Lemma 13 *Let $p = \pi * \mathcal{N}(0,1)$ where $\mathtt{supp}(\pi) \subset [-M, M]$. Then we have $\forall y \geq r \geq x \geq M$,*

$$p(y) \exp\left(\frac{(y-M)^2 - (r-M)^2}{2} \right) \leq p(r) \leq p(x).$$

The following result bounds the grownth of the "score function" in the Gaussian mixture model.

Lemma 14 *Let $p = \pi * \mathcal{N}(0,1)$ where $\mathtt{supp}(\pi) \subset [-M, M]$. Then*

$$|\nabla \log p(r)| \leq 3|r| + 4M, \quad \forall r \in \mathbb{R}.$$

Proof By (Polyanskiy and Wu, 2016, Proposition 2), we have for all $r \in \mathbb{R}$, $|\nabla \log p(r)| \leq 3|r| + 4|\mathbb{E}[X]|$, where $X \sim \pi$. Since π is on $[-M, M]$, we have $|\mathbb{E}[X]| \leq M$. ∎

Lemma 15 *For every $0 \leq t \leq \exp\left(8M^2\right)$ with $M \geq 1$, we have*

$$t \log t - t + 1 \leq 9M^2 \left(\sqrt{t} - 1 \right)^2.$$

Proof We define

$$g(t) \triangleq \frac{t \log t - t + 1}{(\sqrt{t} - 1)^2}.$$

Then we have $g'(t) = \frac{t - 1 - \sqrt{t} \log t}{\sqrt{t}(\sqrt{t}-1)^3}$. The numerator $h(t) = t - 1 - \sqrt{t} \log t$ within satisfies that $h'(t) = 1 - \frac{\log t}{2\sqrt{t}} - \frac{1}{\sqrt{t}} = \frac{\sqrt{t} - 1 - \log \sqrt{t}}{\sqrt{t}} \geq 0$. Hence for $0 \leq t \leq 1$ we have $h(t) \leq h(1) = 0$ and for $t \geq 1$ we have $h(t) \geq h(1) = 0$. Therefore, we have $g'(t) = \frac{h(t)}{\sqrt{t}(\sqrt{t}-1)^3} \geq 0$ for all $t \geq 0$, which indicates that g is non-decreasing on $t \geq 0$. Hence for $0 \leq t \leq \exp\left(8M^2\right)$ and $M \geq 1$, we have

$$g(t) \leq g(\exp(8M^2)) \leq \frac{\exp(8M^2) \cdot 8M^2}{(\sqrt{\exp(8M^2)} - 1)^2} = \frac{8M^2}{(1 - \exp(-4M^2))^2} \leq 9M^2,$$

which indicates that

$$t\log t - t + 1 \le 9M^2\left(\sqrt{t}-1\right)^2$$

■

Proof [Proof of Theorem 12] It is easy to see that for every $x \in [-2M, 2M]$, we have

$$\frac{1}{\sqrt{2\pi}}\exp\left(-8M^2\right) \le p(x), q(x) \le \frac{1}{\sqrt{2\pi}}.$$

Let r be the smallest positive number (possibly infinite) such that $\log\frac{p(r)}{q(r)} \ge 8M^2$, then we have $r \ge 2M$. Without loss of generality we assume $r < \infty$. (Otherwise $\int_r^\infty p(x)\log\frac{p(x)}{q(x)} - p(x) + q(x)dx = 0$ and there is nothing to prove.) Since $\log\frac{p(\cdot)}{q(\cdot)}$ is a continuous function, we have $\log\frac{p(r)}{q(r)} = 8M^2$. According to Lemma 13, we have for every $x \ge r$,

$$p(x) \le p(r)\exp\left(-\frac{(x-M)^2-(r-M)^2}{2}\right), \quad q(x) \le q(r)\exp\left(-\frac{(x-M)^2-(r-M)^2}{2}\right)$$

and according to Lemma 14 we have

$$\left|\log\frac{p(x)}{q(x)}\right| \le \left|\log\frac{p(r)}{q(r)}\right| + \int_r^x (3|t|+4M)dt = 8M^2 + (x-r)(3x+3r+8M), \quad \forall x \ge r \ge 0.$$

Therefore, we obtain that

$$\begin{aligned}
&\int_r^\infty p(x)\log\frac{p(x)}{q(x)} - p(x) + q(x)dx \le \int_r^\infty p(x)\log\frac{p(x)}{q(x)} + q(x)dx \\
&\le p(r)\exp\left(\frac{(r-M)^2}{2}\right)\int_r^\infty (8M^2 + (x-r)(3x+3r+8M))\exp\left(-\frac{(x-M)^2}{2}\right)dx \\
&\quad + q(r)\exp\left(\frac{(r-M)^2}{2}\right)\int_r^\infty \exp\left(-\frac{(x-M)^2}{2}\right)dx \\
&\le p(r)\cdot\left(\frac{6}{(r-M)^3} + \frac{6r+8M}{(r-M)^2} + \frac{8M^2}{r-M}\right) + \frac{q(r)}{r-M} \le \frac{36M^2}{r-M}p(r) + \frac{q(r)}{r-M} \\
&\le \frac{37M^2}{r-M}p(r),
\end{aligned}$$

where the last inequality uses the fact $M \ge 1$ and $\log\frac{p(r)}{q(r)} = 8M^2 \ge 0$ hence $p(r) \ge q(r)$.

Moreover, according to Lemma 14 we also notice that for $0 \le x \le r$ we have $\left|\nabla\log\frac{p(x)}{q(x)}\right| \le 6x + 8M \le 6r + 8M$. Hence noticing that $\log\frac{p(r)}{q(r)} = 8M^2$ and also $r \ge 2M$, we have for every $r - \frac{M^2}{r+M} \le x \le r$,

$$\log\frac{p(x)}{q(x)} \ge 8M^2 - \frac{M^2}{r+M}\cdot(6r+8M) \ge M^2$$

and also $p(x) \ge p(r)$ according to Lemma 13. Therefore, noticing $M \ge 1$, we have

$$H^2(p,q) \ge \int_{r-\frac{M^2}{r+M}}^r p(x)\cdot\left(\sqrt{\frac{q(x)}{p(x)}}-1\right)^2 dx \ge \frac{p(r)M^2}{(r+M)}\cdot\left(1-\frac{1}{e^{M^2/2}}\right)^2 \ge \frac{p(r)M^2}{7(r+M)}.$$

Since $r \geq 2M$, we obtain that

$$\int_r^\infty p(x) \log \frac{p(x)}{q(x)} - p(x) + q(x) dx \leq 777 H^2(p, q).$$

Similarly, if we let s to be the largest negative number (possibly negative infinite) such that $\log \frac{p(s)}{q(s)} \geq 8M^2$, then we will also have

$$\int_{-\infty}^s p(x) \log \frac{p(x)}{q(x)} - p(x) + q(x) dx \leq 777 H^2(p, q).$$

Next we consider those $s \leq x \leq r$. For those x we have $\log \frac{p(x)}{q(x)} \leq 8M^2$. Hence according to Lemma 15, we have

$$\frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} - \frac{p(x)}{q(x)} + 1 \leq 9M^2 \left(\sqrt{\frac{p(x)}{q(x)}} - 1 \right)^2.$$

Therefore,

$$\int_s^r p(x) \log \frac{p(x)}{q(x)} - p(x) + q(x) dx = \int_s^r q(x) \cdot \left(\frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} - \frac{p(x)}{q(x)} + 1 \right) dx$$
$$\leq 9M^2 \int_s^r q(x) \cdot \left(\sqrt{\frac{p(x)}{q(x)}} - 1 \right)^2 dx \leq 9M^2 H^2(p, q).$$

Overall, we have shown that

$$\mathrm{KL}(p \| q) \leq (777 + 777 + 9M^2) H^2(p, q) \leq 1563 M^2 H^2(p, q),$$

which finishes the proof of Theorem 12. ■

4. Proof of Theorem 3

First of all, we notice that (Haussler and Opper, 1997, Lemma 5) shows that for any $\delta, \lambda > 1$ such that

$$0 < \delta < \exp(-1/2) \quad \text{and} \quad \log \log(1/\delta) / \log(1/\delta) \leq (\lambda - 1)/2, \tag{4}$$

and any probability measures $\mathbb{P}, \mathbb{Q}, \mathbb{S}$ and $\mathbb{Q}' = (1 - \delta)\mathbb{Q} + \delta \mathbb{S}$, we have

$$\mathrm{KL}(\mathbb{P} \| \mathbb{Q}') \leq \frac{2 \log(1/\delta)}{(1 - \delta)^2} H^2(\mathbb{P}, \mathbb{Q}) + \frac{4\delta \log(1/\delta)}{(1 - \delta)^2} + \delta^{\frac{\lambda - 1}{2}} \cdot \int_{\mathbb{R}^d} \frac{(d\mathbb{P})^\lambda}{(d\mathbb{S})^{\lambda - 1}}. \tag{5}$$

Let $\lambda = 3$, then as long as $0 < \delta < 1/2$, (4) holds. Choose, $\mathbb{P} = f_\pi$ and $\mathbb{S} = \mathbb{Q} = f_\eta$, we get

$$\mathrm{KL}(f_\pi \| f_\eta) \leq \frac{2 \log(1/\delta)}{(1 - \delta)^2} H^2(f_\pi, f_\eta) + \frac{4\delta \log(1/\delta)}{(1 - \delta)^2} + \delta \cdot \int_{\mathbb{R}^d} \frac{f_\pi(\mathbf{x})^3}{f_\eta(\mathbf{x})^2} d\mathbf{x}. \tag{6}$$

Notice that the last term is an f-divergence with $f = x^3$, which is a convex function, hence $\int_{\mathbb{R}^d} \frac{f_\pi(\mathbf{x})^3}{f_\eta(\mathbf{x})^2} d\mathbf{x}$ is convex in (f_π, f_η). Define the set $\mathscr{P}(M)$ of Gaussian mixtures with mixing distributions supported on the ball $B_2(M)$:

$$\mathscr{P}(M) = \{\pi * \mathcal{N}(0, I_d) : \mathtt{supp}(\pi) \subset B_2(M)\}$$

which is a convex set. Hence the maximum value of $\int_{\mathbb{R}^d} \frac{f_\pi(\mathbf{x})^3}{f_\eta(\mathbf{x})^2} d\mathbf{x}$ where $f_\pi, f_\eta \in \mathscr{P}(M)$ is attained when f_π, f_η are both at the boundary of $\mathscr{P}(M)$, i.e. $\exists \mathbf{u}, \mathbf{v}$ with $\|\mathbf{u}\|_2, \|\mathbf{v}\|_2 \le M$ and we have $f_\pi = \delta_\mathbf{u} * \mathcal{N}(0, I_d), f_\eta = \delta_\mathbf{v} * \mathcal{N}(0, I_d)$. (This is because any $f_\pi \in \mathcal{P}(M)$ can be written as $\int_{B_2(M)} \pi(\mathbf{u}) f_{\delta_\mathbf{u}} d\mathbf{u}$.) Therefore, we have

$$\begin{aligned}
\int_{\mathbb{R}^d} \frac{f_\pi(\mathbf{x})^3}{f_\eta(\mathbf{x})^2} d\mathbf{x} &\le \sup_{\mathbf{u},\mathbf{v}:\|\mathbf{u}\|,\|\mathbf{v}\|_2 \le M} \int_{\mathbb{R}^d} \frac{\left(\frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{\|\mathbf{x}+\mathbf{u}\|_2^2}{2}\right)\right)^3}{\left(\frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{\|\mathbf{x}+\mathbf{v}\|_2^2}{2}\right)\right)^2} d\mathbf{x} \\
&= \sup_{\mathbf{u},\mathbf{v}:\|\mathbf{u}\|,\|\mathbf{v}\|_2 \le M} \frac{1}{\sqrt{2\pi}^d} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\left(\mathbf{x}^T\mathbf{x} + 6\mathbf{x}^T\mathbf{u} - 4\mathbf{x}^T\mathbf{v} + 3\mathbf{u}^T\mathbf{u} - 2\mathbf{v}^T\mathbf{v}\right)\right) d\mathbf{x} \\
&= \sup_{\mathbf{u},\mathbf{v}:\|\mathbf{u}\|,\|\mathbf{v}\|_2 \le M} \frac{1}{\sqrt{2\pi}^d} \exp\left(3\|\mathbf{u}-\mathbf{v}\|_2^2\right) \cdot \int_{\mathbb{R}^d} \exp\left(-\frac{\|\mathbf{x}+3\mathbf{u}-2\mathbf{v}\|_2^2}{2}\right) d\mathbf{x} \\
&= \sup_{\mathbf{u},\mathbf{v}:\|\mathbf{u}\|,\|\mathbf{v}\|_2 \le M} \exp\left(3\|\mathbf{u}-\mathbf{v}\|_2^2\right) = \exp\left(12M^2\right).
\end{aligned}$$

Therefore, according to (6), we have for any $\delta \in [0, 1/2]$,

$$\mathrm{KL}(f_\pi \| f_\eta) \le \frac{2\log(1/\delta)}{(1-\delta)^2} H^2(f_\pi, f_\eta) + \frac{4\delta \log(1/\delta)}{(1-\delta)^2} + \exp(12M^2)\delta.$$

Choosing $\delta = \exp(-12M^2) H^2(f_\pi, f_\eta) \in [0, 1/2]$ and noticing that $(1-\delta)^2 \ge \frac{1}{2}$, we get

$$\begin{aligned}
\mathrm{KL}(f_\pi \| f_\eta) &\le H^2(f_\pi, f_\eta) + 96M^2 H^2(f_\pi, f_\eta) + 16H^2(f_\pi, f_\eta) \log \frac{1}{H^2(f_\pi, f_\eta)} \\
&\le 97M^2 H^2(f_\pi, f_\eta) + 16H^2(f_\pi, f_\eta) \log \frac{1}{H^2(f_\pi, f_\eta)}.
\end{aligned}$$

This finishes the proof of Theorem 3.

5. Proof of Corollary 11

For convenience, denote by $\tilde{R}_{H^2,n}$ the minimax squared Hellinger risk for improper density estimation, similar to $\tilde{R}_{KL,n}$. First, notice that for $\mathcal{P} \subset \mathcal{P}_{com}(M, d)$ or $\mathcal{P} \subset \mathcal{P}_{sub}(K, d)$

$$R_{H^2,n}(\mathcal{P}) \le \tilde{R}_{H^2,n}(\mathcal{P}), \quad R_{KL,n}(\mathcal{P}) \le \tilde{R}_{KL,n}(\mathcal{P}),$$

and also

$$R_{H^2,n}(\mathcal{P}) \lesssim R_{KL,n}(\mathcal{P})$$

since $H^2(\mathbb{P},\mathbb{Q}) \le \mathrm{KL}(\mathbb{P}\|\mathbb{Q})$ holds for all distributions $\mathbb{P}$ and $\mathbb{Q}$. Moreover, according to Theorems 1 and 4, for $\mathbb{P},\mathbb{Q}\in\mathcal{P}$, we have $\mathrm{KL}(\mathbb{P}\|\mathbb{Q}) \lesssim H^2(\mathbb{P},\mathbb{Q})$, which indicates that

$$\tilde{R}_{KL,n}(\mathcal{P}) \lesssim \tilde{R}_{H^2,n}(\mathcal{P}).$$

Therefore, we have

$$R_{H^2,n}(\mathcal{P}) \lesssim R_{KL,n}(\mathcal{P}) \le \tilde{R}_{KL,n}(\mathcal{P}) \lesssim \tilde{R}_{H^2,n}(\mathcal{P}).$$

Next, we notice that

$$R_{H^2,n}(\mathcal{P}) = \inf_{\hat{f}_n}\sup_{f\in\mathcal{P}} \mathbb{E}_f\left[H^2(\hat{f}_n,f)\right].$$

For one estimator $\hat{f}_n$, suppose $\tilde{f}_n$ is the projection of $\hat{f}_n$ into $\mathcal{P}$ under Hellinger distance (since $\mathcal{P}$ is convex, such projection always exists). Then for every $f\in\mathcal{P}$, we have

$$H(\tilde{f}_n,f) \le H(\tilde{f},\hat{f}) + H(\hat{f},f) \le 2H(\hat{f},f),$$

where the last inequality uses the fact that $H(\tilde{f},\hat{f}) \le H(\hat{f},f)$ due to projection. Here $\tilde{f}$ is a proper estimator. Therefore, we have

$$R_{H^2,n}(\mathcal{P}) = \inf_{\hat{f}_n}\sup_{f\in\mathcal{P}} \mathbb{E}_f\left[H^2(\hat{f}_n,f)\right] \ge \frac{1}{4}\inf_{\hat{f}_n\in\mathcal{P}}\sup_{f\in\mathcal{P}} \mathbb{E}_f\left[H^2(\hat{f}_n,f)\right] = \frac{1}{4}\tilde{R}_{H^2,n}(\mathcal{P}).$$

Hence we have proved that

$$R_{H^2,n}(\mathcal{P}) \lesssim R_{KL,n}(\mathcal{P}) \le \tilde{R}_{KL,n}(\mathcal{P}) \lesssim \tilde{R}_{H^2,n}(\mathcal{P}) \lesssim R_{H^2,n}(\mathcal{P}),$$

so we have

$$R_{H^2,n}(\mathcal{P}) \asymp R_{KL,n}(\mathcal{P}) \asymp \tilde{R}_{KL,n}(\mathcal{P}) \asymp \tilde{R}_{H^2,n}(\mathcal{P}).$$

Similarly, for sequential density estimation minimax risks, we can also show that

$$C_{H^2,n}(\mathcal{P}) \asymp C_{KL,n}(\mathcal{P}) \asymp \tilde{C}_{KL,n}(\mathcal{P}) \asymp \tilde{C}_{H^2,n}(\mathcal{P}),$$

where $\tilde{C}_{KL,n}(\mathcal{P}), \tilde{C}_{H^2,n}(\mathcal{P})$ are the proper sequential density estimation minimax risks (where we restrict $\hat{f}_1,\cdots,\hat{f}_n$ to be in the class $\mathcal{P}$ in Definition 10. Therefore, to prove Corollary 11, we only need to show:

$$R_{H^2,n} \asymp \inf_{\epsilon>0}\epsilon^2 + \frac{1}{n}\log\mathcal{N}_{loc,H}(\mathcal{P},\epsilon),$$
$$C_{KL,n} \asymp \inf_{\epsilon>0} n\epsilon^2 + \log\mathcal{N}_H(\mathcal{P},\epsilon).$$

For the first inequality above, the upper bound part follows directly from the celebrated Le Cam-Birgé construction Le Cam (1973); Birgé (1983); Birgé (1986). The lower bound follows from applying Fano's inequality to a local Hellinger ball and the fact that $\mathrm{KL}(\mathbb{P}\|\mathbb{Q}) \asymp H^2(\mathbb{P},\mathbb{Q})$ for $\mathbb{P},\mathbb{Q}\in\mathcal{P}$; see Corollary 33.2 in Polyanskiy and Wu (2022+).

The second inequality (on $C_{KL,n}$) follows directly from Lemma 6 and Lemma 7 in Haussler and Opper (1997) after noticing that the coefficient $b(\epsilon)$ in Lemma 7 of Haussler and Opper (1997) satisfies that

$$b(\epsilon) = \sup\left\{\frac{\mathrm{KL}(\mathbb{P}\|\mathbb{Q})}{H^2(\mathbb{P},\mathbb{Q})} : \mathbb{P},\mathbb{Q}\in\mathcal{P}, H^2(\mathbb{P},\mathbb{Q}) \le \epsilon\right\} \lesssim 1.$$

Acknowledgments

ZJ and YP were supported in part by the MIT-IBM Watson AI Lab and by the National Science Foundation under Grant No CCF-2131115. YW is supported in part by the NSF Grant CCF-1900507, an NSF CAREER award CCF-1651588, and an Alfred Sloan fellowship. The authors thank the anonymous referees for pointing out (Wong and Shen, 1995, Theroem 5) and other helpful comments.

References

Hassan Ashtiani, Shai Ben-David, Nicholas JA Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *Journal of the ACM (JACM)*, 67(6):1–42, 2020.

Afonso S Bandeira, Jonathan Niles-Weed, and Philippe Rigollet. Optimal rates of estimation for multi-reference alignment. *Mathematical Statistics and Learning*, 2(1):25–75, 2020.

Lucien Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237, 1983. ISSN 0044-3719. doi: 10.1007/BF00532480. URL http://dx.doi.org.offcampus.lib.washington.edu/10.1007/BF00532480.

Lucien Birgé. On estimating a density using hellinger distance and some other strange facts. *Probability theory and related fields*, 71(2):271–291, 1986.

Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998. ISSN 1350-7265. doi: 10.2307/3318720. URL http://dx.doi.org/10.2307/3318720.

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

Hong-Bin Chen and Jonathan Niles-Weed. Asymptotics of smoothed wasserstein distances. *Potential Analysis*, pages 1–25, 2021.

Natalie Doss, Yihong Wu, Pengkun Yang, and Harrison H Zhou. Optimal estimation of high-dimensional location gaussian mixtures. *arXiv preprint arXiv:2002.05818*, 2020.

Zhou Fan, Yi Sun, Tianhao Wang, and Yihong Wu. Likelihood landscape and maximum likelihood estimation for the discrete orbit recovery model. *Communications on Pure and Applied Mathematics*, 2021.

Christopher R Genovese and Larry Wasserman. Rates of convergence for the gaussian mixture sieve. *The Annals of Statistics*, 28(4):1105–1127, 2000.

Subhashis Ghosal and Aad van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723, 2007. ISSN 0090-5364. doi: 10.1214/009053606000001271. URL http://dx.doi.org/10.1214/009053606000001271.

Subhashis Ghosal and Aad W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5):1233–1263, 2001. ISSN 0090-5364. doi: 10.1214/aos/1013203453. URL https://doi.org/10.1214/aos/1013203453.

Peter D Grünwald and Nishant A Mehta. Fast rates for general unbounded loss functions: from erm to generalized bayes. *The Journal of Machine Learning Research*, 21(1):2040–2119, 2020.

David Haussler and Manfred Opper. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492, 1997.

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

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

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

Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53, 1973. ISSN 0090-5364. URL http://links.jstor.org.offcampus.lib.washington.edu/sici?sici=0090-5364(197301)1:1<38:COEUDR>2.0.CO;2-V&origin=MSN.

Jaouad Mourtada. Coding convex bodies under gaussian noise, and the wills functional. *Draft*, 2023.

Matey Neykov. On the minimax rate of the gaussian sequence model under bounded convex constraints. *IEEE Transactions on Information Theory*, 2022.

Yutong Nie and Yihong Wu. Improved rates for estimating gaussian mixtures. *Draft*, Dec 2021.

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*, 2022+. URL https://people.lids.mit.edu/yp/homepage/data/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.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Sara van de Geer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.*, 21(1):14–44, 1993. ISSN 0090-5364. doi: 10.1214/aos/1176349013. URL https://doi.org/10.1214/aos/1176349013.

Sara van de Geer. Rates of convergence for the maximum likelihood estimator in mixture models. *Journal of Nonparametric Statistics*, 6(4):293–310, 1996.

Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.*, 23(2):339–362, 1995. ISSN 0090-5364. doi: 10.1214/aos/1176324524. URL http://dx.doi.org/10.1214/aos/1176324524.

Yihong Wu and Pengkun Yang. Optimal estimation of gaussian mixtures via denoised method of moments. *The Annals of Statistics*, 48(4):1981–2007, 2020a.

Yihong Wu and Pengkun Yang. Polynomial methods in statistical inference: Theory and practice. *Foundations and Trends® in Communications and Information Theory*, 17(4):402–586, 2020b. ISSN 1567-2190. doi: 10.1561/0100000095. URL http://dx.doi.org/10.1561/0100000095.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999. ISSN 0090-5364. doi: 10.1214/aos/1017939142. URL http://dx.doi.org/10.1214/aos/1017939142.

Yannis G Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics*, 13(2):768–774, 1985.

Cun-Hui Zhang. Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, pages 1297–1318, 2009.

Appendix A. Proof of Theorem 1

Without loss of generality, we assume $d \le M^2$ (otherwise we use $\sqrt{d} \ge M$ to replace M, and since $\mathtt{supp}(\pi), \mathtt{supp}(\eta) \subset B_2(M) \subset B_2(\sqrt{d})$, the results still hold). For simplicity we abbreviate $f_\pi(\cdot)$ and $f_\eta(\cdot)$ as $p(\cdot), q(\cdot)$. Then we can write

$$\mathrm{KL}(p\|q) = \int_{\mathbb{R}^d} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}, \quad H^2(p,q) = \int_{\mathbb{R}^d} p(\mathbf{x}) \cdot \left(\sqrt{\frac{q(\mathbf{x})}{p(\mathbf{x})}} - 1 \right)^2 d\mathbf{x}. \tag{7}$$

Denoting by Ω the unit sphere in $\mathbb{R}^d$, each $\mathbf{x}$ in $\mathbb{R}^d$ can be written as $\mathbf{x}(r,\omega)$, with $r = \|\mathbf{x}\|_2$ and ω to be the vector parallel to $\mathbf{x}$ in Ω.

To prove Theorem 1, we need the following lemmas.

Lemma 16 *Suppose $p = \pi * \mathcal{N}(0, I_d)$, where $\mathtt{supp}(\pi) \subset B_2(M)$. Then for any $\omega \in \Omega$, we have:*

1. *$\forall r' \in [M, r]$, we have $p(\mathbf{x}(r',\omega)) \ge p(\mathbf{x}(r,\omega))$.*
2. *$\forall r' \ge r \ge M$, we have $p(\mathbf{x}(r',\omega)) \le p(\mathbf{x}(r,\omega)) \exp\left(-\frac{(r'-M)^2-(r-M)^2}{2}\right)$*

Proof We can write

$$p(\mathbf{x}(r',\omega)) = \int_{B_2(M)} \varphi(\mathbf{x}(r',\omega) - \mathbf{u})\pi(\mathbf{u}) d\mathbf{u}, \qquad p(\mathbf{x}(r,\omega)) = \int_{B_2(M)} \varphi(\mathbf{x}(r,\omega) - \mathbf{u})\pi(\mathbf{u}) d\mathbf{u},$$

where we use $\pi(\cdot)$ to denote the density distribution of π (which can be a generalized function), and $\varphi(\cdot)$ to denote the density distribution of $\mathcal{N}(0, I_d)$. To prove this lemma, we only need to verify the following two inequalities:

1. For any $\forall r' \in [M, r]$ and any $u \in B_2(M)$, we have $\varphi(\mathbf{x}(r',\omega) - \mathbf{u}) \ge \varphi(\mathbf{x}(r,\omega) - \mathbf{u})$;
2. For any $\forall r' \ge r \ge M$ and any $u \in B_2(M)$, we have $\varphi(\mathbf{x}(r',\omega) - \mathbf{u}) \le \varphi(\mathbf{x}(r,\omega) - \mathbf{u}) \exp\left(-\frac{(r'-M)^2-(r-M)^2}{2}\right)$.

Without loss of generality, we assume $\omega = (1, 0, \cdots, 0)$. Then for any $\mathbf{u} = (u_1, u_2, \cdots, u_d)$, we have

$$\varphi(\mathbf{x}(r,\omega) - \mathbf{u}) = \frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{(r-u_1)^2 + \sum_{i=2}^d u_i^2}{2} \right)$$
$$\varphi(\mathbf{x}(r',\omega) - \mathbf{u}) = \frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{(r'-u_1)^2 + \sum_{i=2}^d u_i^2}{2} \right).$$

When $M \le r' \le r$, it is easy to see that $|r - u_1| \ge |r' - u_1|$ for any $|u_1| \le M$. The first inequality is verified. As for the second inequality, since $|u_1| \le M \le r \le r'$, we have $(r'-u_1)^2 - (r-u_1)^2 \ge (r'-M)^2 - (r-M)^2$, which indicates that

$$-\frac{(r'-u_1)^2 + \sum_{i=2}^d u_i^2}{2} \le -\frac{(r-u_1)^2 + \sum_{i=2}^d u_i^2}{2} - \frac{(r'-M)^2 - (r-M)^2}{2}.$$

■

Lemma 17 *Suppose $p = \pi * \mathcal{N}(0, I_d)$ where $\mathtt{supp}(\pi) \subset B_2(M)$. Then we have*

$$\|\nabla \log p(\mathbf{x})\|_2 \leq 3\|\mathbf{x}\|_2 + 4M, \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

Proof According to Proposition 2 in Polyanskiy and Wu (2016), we have $\forall \mathbf{x} \in \mathbb{R}^d$,

$$\|\nabla \log p(\mathbf{x})\|_2 \leq 3\|\mathbf{x}\|_2 + 4\|\mathbb{E}[X]\|_2,$$

where $X \sim \pi$. Since the support of π is a subset of $B_2(M)$, we have $\|\mathbb{E}[X]\|_2 \leq M$. ■

Proof [Proof of Theorem 1] According to (7), we have

$$\begin{aligned}
\mathrm{KL}(p\|q) &= \int_\Omega \int_0^\infty r^{d-1} p(\mathbf{x}(r,\omega)) \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} dr d\omega \\
&= \int_\Omega \int_0^\infty r^{d-1} \left(p(\mathbf{x}(r,\omega)) \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - p(\mathbf{x}(r,\omega)) + q(\mathbf{x}(r,\omega)) \right) dr d\omega \qquad (8) \\
H^2(p,q) &= \int_\Omega \int_0^\infty r^{d-1} p(\mathbf{x}(r,\omega)) \left(\sqrt{\frac{q(\mathbf{x}(r,\omega))}{p(\mathbf{x}(r,\omega))}} - 1 \right)^2 dr d\omega
\end{aligned}$$

For every $\omega \in \Omega$, we define r_ω as

$$r_\omega \triangleq \inf \left\{ r : \log \frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))} \geq 8M^2 \right\}.$$

Notice that for any $r \leq 2M$ and $\omega \in \Omega$, we have

$$\begin{aligned}
p(\mathbf{x}(r,\omega)) &= \int_{B_2(M)} \pi(\mathbf{u}) \varphi(\mathbf{x}(r,\omega), \mathbf{u}) d\mathbf{u} \leq \frac{1}{\sqrt{2\pi}^d} \\
q(\mathbf{x}(r,\omega)) &= \int_{B_2(M)} \eta(\mathbf{u}) \varphi(\mathbf{x}(r,\omega), \mathbf{u}) d\mathbf{u} \geq \frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{(2M+M)^2}{2} \right),
\end{aligned}$$

which indicates that

$$\log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega)} \leq \frac{9M^2}{2} < 8M^2.$$

Hence for every $\omega \in \Omega$, we all have $r_\omega \geq 2M$. And if $r_\omega \neq \infty$, we have that $\log \frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))} = 8M^2$. According to Lemma 16 and Lemma 17, we know that for every $r \geq r_\omega$, we all have

$$\begin{aligned}
p(\mathbf{x}(r,\omega)) &\leq p(\mathbf{x}(r_\omega,\omega)) \exp\left(-\frac{(r-M)^2 - (r_\omega - M)^2}{2} \right), \\
\log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} &\leq \log \frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))} + (r - r_\omega)(3r + 3r_\omega + 8M) \\
&\leq 8M^2 + (r - r_\omega)(3r + 3r_\omega + 8M).
\end{aligned}$$

Therefore, we obtain that

$$\begin{aligned}
&\int_{r_\omega}^\infty r^{d-1} p(\mathbf{x}(r,\omega)) \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} dr \\
&\leq p(\mathbf{x}(r_\omega,\omega)) \int_{r_\omega}^\infty r^{d-1} (8M^2 + (r - r_\omega)(3r + 3r_\omega + 8M)) \exp\left(-\frac{(r-M)^2 - (r_\omega - M)^2}{2} \right) dr. \qquad (9)
\end{aligned}$$

We adopt the changes of variables from r to $t = r - r_\omega$, and obtain that

$$\begin{aligned} &8M^2 \int_{r_\omega}^{\infty} r^{d-1} \exp\left(-\frac{(r-M)^2 - (r_\omega - M)^2}{2}\right) dr \\ &= 8M^2 r_\omega^{d-1} \int_0^{\infty} \exp\left((d-1)\log\frac{t+r_\omega}{r_\omega} - \frac{t^2}{2} - (r_\omega - M)t\right) dt \end{aligned} \tag{10}$$

and

$$\begin{aligned} &\int_{r_\omega}^{\infty} r^{d-1}(r - r_\omega)(3r + 3r_\omega + 8M) \exp\left(-\frac{(r-M)^2 - (r_\omega - M)^2}{2}\right) dr \\ &= r_\omega^{d-1} \int_0^{\infty} \exp\left((d-1)\log\frac{t+r_\omega}{r_\omega} + \log t + \log(3t + 6r_\omega + 8M) - \frac{t^2}{2} - (r_\omega - M)t\right) dt. \end{aligned} \tag{11}$$

We first prove the upper bound (10). We define

$$f(t) \triangleq (d-1)\log\frac{t+r_\omega}{r_\omega} - \frac{1}{2}(r_\omega - M)t.$$

Since

$$d - 1 < d \le M^2 = \frac{1}{2} \cdot (2M) \cdot (2M - M) \le \frac{1}{2} r_\omega (r_\omega - M),$$

the first-order derivative of f satisfies that

$$f'(t) = \frac{d-1}{t + r_\omega} - \frac{1}{2}(r_\omega - M) \le \frac{d-1}{r_\omega} - \frac{1}{2}(r_\omega - M) \le 0, \quad \forall t \ge 0.$$

Therefore we have

$$(d-1)\log\frac{t+r_\omega}{r_\omega} - \frac{1}{2}(r_\omega - M)t = f(t) \le f(0) = 0, \quad \forall t \ge 0.$$

This directly indicates the following upper bound on (10).

$$\begin{aligned} &\frac{1}{r_\omega^{d-1}} \int_{r_\omega}^{\infty} r^{d-1} \exp\left(-\frac{(r-M)^2 - (r_\omega - M)^2}{2}\right) dr \le \int_0^{\infty} \exp\left(-\frac{t^2}{2} - \frac{1}{2}(r_\omega - M)t\right) dt \\ &\le \int_0^{\infty} \exp\left(-\frac{1}{2}(r_\omega - M)t\right) dt = \frac{2}{r_\omega - M} \end{aligned} \tag{12}$$

Next we prove the upper bound (11). We define

$$g(t) \triangleq \log t + \log(3t + 6r_\omega + 8M) - \frac{1}{3}(r_\omega - M)t.$$

Then we have

$$g'(t) = \frac{1}{t} + \frac{1}{3t + 6r_\omega + 8M} - \frac{r_\omega - M}{3},$$

which has a single root t_0 on $(0, \infty)$, which is also the maximum of $g(t)$ for $t \ge 0$. Further notice

$$0 = g'(t_0) = \frac{1}{t_0} + \frac{1}{3t_0 + 6r_\omega + 8M} - \frac{r_\omega - M}{3} \le \frac{4}{3t_0} - \frac{r_\omega - M}{3}.$$

Hence we get $t_0 \leq \frac{4}{r_\omega - M} \leq \frac{4}{M}$, and

$$3t_0 + 6r_\omega + 8M \leq \frac{12}{M} + 6r_\omega + 8M \leq 6r_\omega + 20M \leq 32(r_\omega - M).$$

This gives the following upper bound on $g(t)$ for $t \geq 0$:

$$g(t) \leq g(t_0) \leq \log t_0(3t_0 + 6r_\omega + 8M) - \frac{4}{3} \leq \log 128 - \frac{4}{3},$$

Combining this result with our upper bound on the function f, we get

$$\begin{aligned}
&(d-1)\log\frac{t + r_\omega}{r_\omega} + \log t + \log(3t + 6r_\omega + 8M) - \frac{t^2}{2} - (r_\omega - M)t \\
&= f(t) + g(t) - \frac{t^2}{2} - \frac{(r_\omega - M)t}{6} \leq \log 128 - \frac{4}{3} - \frac{(r_\omega - M)t}{6}.
\end{aligned}$$

This directly indicates the following upper bound on (11).

$$\begin{aligned}
&\int_{r_\omega}^{\infty} r^{d-1}(r - r_\omega)(3r + 3r_\omega + 8M)\exp\left(-\frac{(r-M)^2 - (r_\omega - M)^2}{2}\right)dr \\
&\leq r_\omega^{d-1}\int_0^{\infty}\exp\left(\log 128 - \frac{4}{3} - \frac{(r_\omega - M)t}{6}\right)dt = 128 r_\omega^{d-1}\cdot\frac{6e^{-4/3}}{r_\omega - M} \\
&\leq \frac{210 r_\omega^{d-1}}{r_\omega - M} \leq \frac{210 M^2 r_\omega^{d-1}}{r_\omega - M}.
\end{aligned}$$

We combine these two upper bounds together. According to (9) we obtain that

$$\int_{r_\omega}^{\infty} r^{d-1} p(\mathbf{x}(r,\omega))\log\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}dr \leq \frac{212 M^2 r_\omega^{d-1}}{r_\omega - M} p(\mathbf{x}(r_\omega,\omega)),$$

Similarly, we can also obtain bound on $\int_{r_\omega}^{\infty} r^{d-1} q(\mathbf{x}(r,\omega))dr$: According to Lemma 16 we obtain that

$$q(\mathbf{x}(r,\omega)) \leq q(\mathbf{x}(r_\omega,\omega))\exp\left(-\frac{(r-M)^2 - (r_\omega - M)^2}{2}\right).$$

According to (12) we have

$$\int_{r_\omega}^{\infty} r^{d-1} q(\mathbf{x}(r,\omega))dr \leq q(\mathbf{x}(r_\omega,\omega))\int_{r_\omega}^{\infty} r^{d-1}\exp\left(-\frac{(r-M)^2 - (r_\omega - M)^2}{2}\right)dr \leq \frac{2 r_\omega^{d-1}}{r_\omega - M} q(\mathbf{x}(r_\omega,\omega)).$$

We further notice that

$$\log\frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))} = 8M^2 \geq 0,$$

which indicates that $q(\mathbf{x}(r_\omega,\omega)) \leq p(\mathbf{x}(r_\omega,\omega))$. Therefore, since $M \geq 1$, we have

$$\int_{r_\omega}^{\infty} r^{d-1} q(\mathbf{x}(r,\omega))dr \leq \frac{2M^2 r_\omega^{d-1}}{r_\omega - M} p(\mathbf{x}(r_\omega,\omega)).$$

Therefore, we have the following upper bound

$$\begin{aligned}
&\int_\Omega \int_{r_\omega}^\infty r^{d-1} \left(p(\mathbf{x}(r,\omega)) \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - p(\mathbf{x}(r,\omega)) + q(\mathbf{x}(r,\omega)) \right) dr d\omega \\
\le &\int_\Omega \int_{r_\omega}^\infty r^{d-1} p(\mathbf{x}(r,\omega)) \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} dr d\omega + \int_\Omega \int_{r_\omega}^\infty r^{d-1} q(\mathbf{x}(r,\omega)) dr d\omega \\
\le &\int_\Omega \frac{(212+2) M^2 r_\omega^{d-1} p(\mathbf{x}(r_\omega,\omega))}{r_\omega - M} d\omega = \int_\Omega \frac{214 M^2 r_\omega^{d-1} p(\mathbf{x}(r_\omega,\omega))}{r_\omega - M} d\omega.
\end{aligned}$$

Next, according to Lemma 17, for $\forall \omega \in \Omega$ and $r \in [0, r_\omega]$ we have

$$\begin{aligned}
\left\| \nabla \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} \right\|_2 &= \|\nabla \log p(\mathbf{x}(r,\omega))\|_2 + \|\nabla \log q(\mathbf{x}(r,\omega))\|_2 \\
&\le 6r + 8M \le 6r_\omega + 8M \le 8r_\omega + 8M.
\end{aligned}$$

Notice that we also have $\log \frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))} = 8M^2$. Hence for $\forall r_\omega - \frac{1}{r_\omega + M} \le r \le r_\omega$,

$$\log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} \ge 8M^2 - (8r_\omega + 8M) \cdot \frac{1}{r_\omega + M} = 8M^2 - 8 \ge 2,$$

which indicates that

$$\left(\sqrt{\frac{q(\mathbf{x}(r,\omega))}{p(\mathbf{x}(r,\omega))}} - 1 \right)^2 \ge \left(\frac{1}{e^2} - 1 \right)^2 \ge \frac{1}{2}.$$

We further notice that $r_\omega \ge 2M \ge M$ and $d - 1 < d \le M^2$. Hence for $\forall r_\omega - \frac{1}{r_\omega + M} \le r \le r_\omega$,

$$r^{d-1} \ge r_\omega^{d-1} \cdot \left(1 - \frac{1}{r_\omega (r_\omega + M)} \right)^{d-1} \ge r_\omega^{d-1} \cdot \left(1 - \frac{d}{r_\omega (r_\omega + M)} \right) \ge r_\omega^{d-1} \left(1 - \frac{d-1}{2M^2} \right) \ge \frac{1}{2} r_\omega^{d-1}.$$

After noticing that $p(\mathbf{x}(r,\omega)) \ge p(\mathbf{x}(r_\omega,\omega))$ according to Lemma 16, we obtain

$$\begin{aligned}
&\int_{r_\omega - \frac{1}{r_\omega + M}}^{r_\omega} r^{d-1} p(\mathbf{x}(r,\omega)) \cdot \left(\sqrt{\frac{q(\mathbf{x}(r,\omega))}{p(\mathbf{x}(r,\omega))}} - 1 \right)^2 dr \\
\ge &\frac{1}{r_\omega + M} \cdot \min_{r_\omega - \frac{1}{r_\omega + M} \le r \le r_\omega} r^{d-1} p(\mathbf{x}(r,\omega)) \cdot \left(\sqrt{\frac{q(\mathbf{x}(r,\omega))}{p(\mathbf{x}(r,\omega))}} - 1 \right)^2 \\
\ge &\frac{1}{r_\omega + M} \cdot \frac{1}{2} r_\omega^{d-1} \cdot \frac{1}{2} p(\mathbf{x}(r_\omega,\omega)) \ge \frac{r_\omega^{d-1} p(\mathbf{x}(r_\omega,\omega))}{12(r_\omega - M)} \\
\ge &\frac{1}{2568 M^2} \int_{r_\omega}^\infty r^{d-1} \left(p(\mathbf{x}(r,\omega)) \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - p(\mathbf{x}(r,\omega)) + q(\mathbf{x}(r,\omega)) \right) dr.
\end{aligned}$$

Therefore, according to (8), we have

$$\begin{aligned}
H^2(p,q) &= \int_\Omega \int_0^\infty r^{d-1} p(\mathbf{x}(r,\omega)) \left(\sqrt{\frac{q(\mathbf{x}(r,\omega))}{p(\mathbf{x}(r,\omega))}} - 1 \right)^2 dr d\omega \\
&\geq \int_\Omega \int_{r_\omega - \frac{1}{r_\omega + M}}^{r_\omega} r^{d-1} p(\mathbf{x}(r,\omega)) \left(\sqrt{\frac{q(\mathbf{x}(r,\omega))}{p(\mathbf{x}(r,\omega))}} - 1 \right)^2 dr d\omega \\
&\geq \frac{1}{2568 M^2} \int_\Omega \int_{r_\omega}^\infty r^{d-1} \left(p(\mathbf{x}(r,\omega)) \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - p(\mathbf{x}(r,\omega)) + q(\mathbf{x}(r,\omega)) \right) dr d\omega.
\end{aligned}$$

Next we consider those $\mathbf{x}(r,\omega)$ with $0 \leq r \leq r_\omega$. According to the definition of r_ω, we know that for any such r, we have

$$\log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} \leq 8M^2.$$

Hence from Lemma 15, we have for any $\omega \in \Omega$ and $r \in [0, r_\omega]$,

$$\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} + 1 \leq 9M^2 \left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}} - 1 \right)^2,$$

which indicates that

$$\begin{aligned}
&\int_\Omega \int_0^{r_\omega} r^{d-1} \left(p(\mathbf{x}(r,\omega)) \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - p(\mathbf{x}(r,\omega)) + q(\mathbf{x}(r,\omega)) \right) dr d\omega \\
&= \int_{\omega \in \Omega} \int_0^{r_\omega} r^{d-1} q(\mathbf{x}(r,\omega)) \cdot \left(\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} + 1 \right) dr d\omega \\
&\leq 9M^2 \int_\Omega \int_0^{r_\omega} r^{d-1} q(\mathbf{x}(r,\omega)) \cdot \left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}} - 1 \right)^2 dr d\omega \\
&\leq 9M^2 \int_\Omega \int_0^{\infty} r^{d-1} q(\mathbf{x}(r,\omega)) \cdot \left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}} - 1 \right)^2 dr d\omega = 9M^2 H^2(p,q).
\end{aligned}$$

Combine the above two cases, we obtain that

$$\begin{aligned}
\mathrm{KL}(p\|q) &= \int_\Omega \int_0^\infty r^{d-1} \left(p(\mathbf{x}(r,\omega)) \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - p(\mathbf{x}(r,\omega)) + q(\mathbf{x}(r,\omega)) \right) dr d\omega \\
&= \int_\Omega \int_0^{r_\omega} r^{d-1} \left(p(\mathbf{x}(r,\omega)) \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - p(\mathbf{x}(r,\omega)) + q(\mathbf{x}(r,\omega)) \right) dr d\omega \\
&\quad + \int_\Omega \int_{r_\omega}^\infty r^{d-1} \left(p(\mathbf{x}(r,\omega)) \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - p(\mathbf{x}(r,\omega)) + q(\mathbf{x}(r,\omega)) \right) dr d\omega \\
&\leq 18M^2 H^2(p,q) + 5136 M^2 H^2(p,q) = 5154 M^2 H^2(p,q).
\end{aligned}$$

This completes the proof of Theorem 1. ■

Appendix B. Proof of Theorem 5

Given some constant $r > 1$, we consider the following distribution:

$$\pi_r = (1 - h_r)\delta_0 + h_r\delta_r,$$

where $h_r = \exp\left(-\frac{r^2}{2K^2}\right)$. Then for any $r > 1$, π_r is a K-subgaussian distribution.

We let $p_r = \pi_r * \mathcal{N}(0,1)$. Since $x \log x - x + 1 \geq 0$ holds for all $x > 0$, we have

$$\begin{aligned}
&\mathrm{KL}(p_r \| \mathcal{N}(0,1)) \\
&= \int_{-\infty}^{\infty} p_r(x) \log \frac{p_r(x)}{\varphi(x)} dx = \int_{-\infty}^{\infty} \varphi(x) \cdot \left(\frac{p_r(x)}{\varphi(x)} \log \frac{p_r(x)}{\varphi(x)} - \frac{p_r(x)}{\varphi(x)} + 1 \right) dx \\
&\geq \int_{r}^{r+1} \varphi(x) \cdot \left(\frac{p_r(x)}{\varphi(x)} \log \frac{p_r(x)}{\varphi(x)} - \frac{p_r(x)}{\varphi(x)} + 1 \right) dx \geq \int_{r}^{r+1} p_r(x) \log \frac{p_r(x)}{\varphi(x)} - p_r(x) dx.
\end{aligned}$$

According to our construction, for $r \leq x \leq r+1$ we have

$$p_r(x) = (1 - h_r)\varphi(x) + h_r\varphi(r - x) \geq h_r \cdot \varphi(1) \quad \text{and also} \quad \varphi(x) \leq \varphi(r).$$

Therefore, we obtain that for $r \leq x \leq r + 1$,

$$\log \frac{p_r(x)}{\varphi(x)} \leq \log \frac{p_r \varphi(1)}{\varphi(r)} = \log \exp\left(-\frac{r^2}{2K^2} - \frac{1}{2} + \frac{r^2}{2} \right) = \frac{r^2}{2} - \frac{r^2}{2K^2} - \frac{1}{2}.$$

Noticing that $\varphi(1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\right) \geq \frac{1}{5}$, we obtain that

$$\mathrm{KL}(p_r \| \mathcal{N}(0,1)) \geq \int_{r}^{r+1} \frac{h_r}{5} \cdot \left(\frac{r^2}{2} - \frac{r^2}{2K^2} - \frac{1}{2} - 1 \right) dx = \left(\frac{r^2}{10} - \frac{r^2}{10K^2} - \frac{3}{10} \right) h_r. \tag{13}$$

Next, we write

$$H^2(p_r, \mathcal{N}(0,1)) = \int_{-\infty}^{\infty} \left(\sqrt{p_r(x)} - \sqrt{\varphi(x)} \right)^2 dx.$$

We divide the integral domain into three regions: $(-\infty, -r], [-r, r]$ and $[r, \infty)$, and upper bound the contribution from each region separately.

Noticing that $p_r(x) = (1 - h_r)\varphi(x) + h_r\varphi(r - x)$, we have for any $x \leq -r$,

$$0 \leq p_r(x) \leq \varphi(x).$$

Therefore,

$$\begin{aligned}
&\int_{-\infty}^{-r} \left(\sqrt{p_r(x)} - \sqrt{\varphi(x)} \right)^2 dx \leq \int_{-\infty}^{-r} \varphi(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} \exp\left(-\frac{(x+r)^2}{2} \right) dx \\
&\leq \exp\left(-\frac{r^2}{2} \right) \int_{-\infty}^{0} \exp(-rx) dx = \frac{1}{r} \exp\left(-\frac{r^2}{2} \right) \leq \exp\left(-\frac{r^2}{2K^2} \right) = h_r.
\end{aligned}$$

Noticing that for those $x \geq r$,

$$(1 - h_r)\varphi(x) \leq \varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} \right) \leq \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{r^2 + (x-r)^2}{2} \right) \leq h_r \varphi(r - x),$$

we obtain

$$0 \le \varphi(x) \le (1-h_r)\varphi(x) + h_r\varphi(r-x) = p_r(x) \le 2h_r\varphi(r-x) \qquad \forall x \ge r,$$

which indicates that

$$\int_r^\infty \left(\sqrt{p_r(x)} - \sqrt{\varphi(x)}\right)^2 dx = \int_r^\infty p_r(x)dx \le 2h_r \int_r^\infty \varphi(r-x)dx = h_r.$$

Finally for those $-r \le x \le r$, we notice that

$$p_r(x) - \varphi(x) = (1-h_r)\varphi(x) + h_r\varphi(r-x) - \varphi(x) = h_r(\varphi(r-x) - \varphi(x)),$$

hence $|p_r(x) - \varphi(x)| \le h_r \cdot |\varphi(r-x) - \varphi(x)| \le \frac{h_r}{\sqrt{2\pi}} \le h_r$. Therefore, if either $p_r(x) \ge h_r$ or $\varphi(x) \ge h_r$, we have

$$\left(\sqrt{p_r(x)} - \sqrt{\varphi(x)}\right)^2 = \frac{(p_r(x)-\varphi(x))^2}{\left(\sqrt{p_r(x)} + \sqrt{\varphi(x)}\right)^2} \le \frac{h_r^2}{\sqrt{h_r}^2} = h_r.$$

And if neither $p_r(x) \ge h_r$ nor $\varphi(x) \ge h_r$ holds, then we have $0 \le p_r(x), \varphi(x) \le h_r$ and hence

$$\left(\sqrt{p_r(x)} - \sqrt{\varphi(x)}\right)^2 \le h_r.$$

Overall, we have $\left(\sqrt{p_r(x)} - \sqrt{\varphi(x)}\right)^2 \le h_r$ and hence

$$\int_{-r}^r \left(\sqrt{p_r(x)} - \sqrt{\varphi(x)}\right)^2 dx \le 2rh_r.$$

Combining the contributions from these regions, we obtain that

$$H^2(p_r, \mathcal{N}(0,1)) = \int_{-\infty}^\infty \left(\sqrt{p_r(x)} - \sqrt{\varphi(x)}\right)^2 dx \le h_r + 2rh_r + h_r = (2+2r)h_r. \tag{14}$$

Finally, applying (13) and (14), we choose the parameter r so that KL/H^2 exceeds an arbitrary constant C. Noticing that $K > 1$, we have $\frac{r^2}{10} - \frac{r^2}{10K^2} \ge 0$, hence there exists $r > 1$ such that

$$\frac{r^2}{10} - \frac{r^2}{10K^2} - \frac{3}{10} \ge c \cdot (2+2r).$$

And for this r, we have

$$\mathrm{KL}(p_r \| \mathcal{N}(0,1)) \ge c \cdot H^2(p_r, \mathcal{N}(0,1)).$$

This finishes the proof of Theorem 5.

Remark 18 *Notice that with similar proving techniques, we can also show that*

Appendix C. Proof of Theorem 4

Without loss of generality, we assume $d \le \frac{1}{2(1-K)}$ (otherwise we use $1-\frac{1}{2d} \ge K$ to replace K, and since π, η are K-subgaussian, hence they are also $\left(1-\frac{1}{2d}\right)$-subgaussian. The results still hold). We abbreviate $f_\pi(\cdot), f_\eta(\cdot)$ as $p(\cdot), q(\cdot)$.

Lemma 19 *Suppose $p = \pi * \mathcal{N}(0, I_d)$, where π is a K-subgaussian distribution with $K < 1$. Then for every $r \ge \frac{2\sqrt{K}}{1-\sqrt{K}}$ and any $\omega \in \Omega$, we have the following propositions:*

1. *$\forall r' \in \left[\frac{2\sqrt{K}}{1-\sqrt{K}}, r\right]$, we have*
$$p(\mathbf{x}(r',\omega)) \ge \frac{p(\mathbf{x}(r,\omega))}{7}. \tag{15}$$
2. *$\forall r' \ge r$, we have*
$$p(\mathbf{x}(r',\omega)) \le 7p(\mathbf{x}(r,\omega)) \cdot \exp\left(-\frac{1-K}{4} r(r'-r)\right). \tag{16}$$

Proof For every $r' \ge r$ and $\omega \in \Omega$, notice that we can write $p(\cdot)$ as the following integral:

$$p(\mathbf{x}(r',\omega)) = \int_{\mathbb{R}^d} \varphi(\mathbf{x}(r',\omega)-\mathbf{u})\pi(\mathbf{u})d\mathbf{u}, \qquad p(\mathbf{x}(r,\omega)) = \int_{\mathbb{R}^d} \varphi(\mathbf{x}(r,\omega)-\mathbf{u})\pi(\mathbf{u})d\mathbf{u},$$

where we use $\pi(\cdot)$ to denote the density distribution of π (which can be a generalized function), and $\varphi(\cdot)$ to denote the density distribution of $\mathcal{N}(0, I_d)$.

Since π is a K-subgaussian distribution, we have $\mathbf{P}[\|X\|_2 \ge \alpha] \le \exp\left(-\frac{\alpha^2}{2K^2}\right)$ for $\alpha \ge 0$ and $\mathbf{P}[\|X\|_2 \ge 1] \le \frac{1}{\sqrt{e}} \le \frac{2}{3}$, where $X \sim \pi$. Therefore, $\mathbf{P}[\|X\|_2 \le 1] \ge \frac{1}{3}$. Given $0 \le \alpha \le r \le \beta \le r'$ (α, β will be specified later), we have

$$\begin{aligned}
\int_{\alpha \le \|\mathbf{u}\| \le \beta} \varphi(\mathbf{x}(r',\omega)-\mathbf{u})\pi(\mathbf{u})d\mathbf{u} &\le \frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{(r'-\beta)^2}{2}\right) \cdot \mathbf{P}[\|X\|_2 \ge \alpha] \\
&\le \frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{\alpha^2}{2K^2} - \frac{(r'-\beta)^2}{2}\right); \\
p(\mathbf{x}(r,\omega)) &\ge \frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{(r+1)^2}{2}\right) \cdot \mathbf{P}[\|X\|_2 \le 1] \\
&\ge \frac{1}{3} \cdot \frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{(r+1)^2}{2}\right),
\end{aligned}$$

which indicates that

$$\int_{\alpha \le \|\mathbf{u}\| \le \beta} \varphi(\mathbf{x}(r',\omega)-\mathbf{u})\pi(\mathbf{u})d\mathbf{u} \le p(\mathbf{x}(r,\omega)) \cdot 3\exp\left(\frac{(r+1)^2}{2} - \frac{\alpha^2}{2K^2} - \frac{(r'-\beta)^2}{2}\right).$$

We further have

$$\int_{\|\mathbf{u}\|_2 \ge \beta} \varphi(\mathbf{x}(r',\omega)-\mathbf{u})\pi(\mathbf{u})d\mathbf{u} \le \frac{1}{\sqrt{2\pi}^d} \cdot \mathbf{P}[\|X\|_2 \ge \beta] \le \frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{\beta^2}{2K^2}\right),$$

which indicates that

$$\int_{\alpha\le\|\mathbf{u}\|\le\beta}\varphi(\mathbf{x}(r',\omega)-\mathbf{u})\pi(\mathbf{u})d\mathbf{u}\le p(\mathbf{x}(r,\omega))\cdot 3\exp\left(\frac{(r+1)^2}{2}-\frac{\beta^2}{2K^2}\right)$$

Finally, for every $\mathbf{x}$ such that $\|\mathbf{x}\|_2\le\alpha$, similar to previous proof we have

$$\|\mathbf{x}-\mathbf{x}(r,\omega)\|_2^2-(r-\alpha)^2\le\|\mathbf{x}-\mathbf{x}(r',\omega)\|_2^2-(r'-\alpha)^2,$$

which indicates that

$$\begin{aligned}\int_{\|u\|\le\alpha}\varphi(\mathbf{x}(r',\omega)-\mathbf{u})\pi(\mathbf{u})d\mathbf{u}&\le\int_{\|u\|\le\alpha}\varphi(\mathbf{x}(r,\omega)-\mathbf{u})\pi(\mathbf{u})d\mathbf{u}\cdot\exp\left(\frac{(r-\alpha)^2}{2}-\frac{(r'-\alpha)^2}{2}\right)\\&\le p(\mathbf{x}(r,\omega))\exp\left(\frac{(r-\alpha)^2}{2}-\frac{(r'-\alpha)^2}{2}\right).\end{aligned}$$

Above all, we get

$$\begin{aligned}p(\mathbf{x}(r',\omega))\le p(\mathbf{x}(r,\omega))\cdot\Bigg(&3\exp\left(\frac{(r+1)^2}{2}-\frac{\alpha^2}{2K^2}-\frac{(r'-\beta)^2}{2}\right)\\&+3\exp\left(\frac{(r+1)^2}{2}-\frac{\beta^2}{2K^2}\right)+\exp\left(\frac{(r-\alpha)^2}{2}-\frac{(r'-\alpha)^2}{2}\right)\Bigg)\end{aligned}$$

Further noticing that when $r\ge\frac{2\sqrt{K}}{1-\sqrt{K}}$, we have

$$r-\sqrt{K}(1+r)\ge\frac{1-\sqrt{K}}{2}r\ge 0.$$

Therefore, choosing $\alpha=\sqrt{K}(r+1)$ and $\beta=\frac{r'+r}{2}$, and noticing that $2(1-\sqrt{K})\ge 1-K$ holds for all $0<K<1$, we will get

$$\begin{aligned}&\frac{(r-\alpha)^2}{2}-\frac{(r'-\alpha)^2}{2}=\exp\left(-\frac{(r'-r)^2}{2}-(r'-r)(r-\sqrt{K}(r+1))\right)\\&\le\exp\left(-\frac{1-\sqrt{K}}{2}r(r'-r)\right)\le\exp\left(-\frac{1-K}{4}r(r'-r)\right),\\&\frac{(r+1)^2}{2}-\frac{\alpha^2}{2K^2}-\frac{(r'-\beta)^2}{2}=-\frac{(r+1)^2(1-K)}{2K}-\frac{(r'-r)^2}{8}\le-\frac{(1-K)r^2}{2}-\frac{(r'-r)^2}{8},\\&\frac{(r+1)^2}{2}-\frac{\beta^2}{2K^2}\le\frac{(r+1)^2}{2}-\frac{r^2}{2K^2}-\frac{(r'-r)^2}{8K^2}\le\frac{(r+1)^2}{2}-\frac{(r+1)^2}{2K}-\frac{(r'-r)^2}{8K^2}\\&\le-\frac{(1-K)r^2}{2}-\frac{(r'-r)^2}{8}.\end{aligned}$$

Hence we get

$$p(\mathbf{x}(r',\omega))\le p(\mathbf{x}(r,\omega))\cdot\left(6\exp\left(-\frac{(1-K)r^2}{2}-\frac{(r'-r)^2}{8}\right)+\exp\left(-\frac{1-K}{4}r(r'-r)\right)\right).$$

Next we notice that

$$\frac{(1-K)r^2}{2}+\frac{(r'-r)^2}{8} \geq (1-K)\cdot\left(\frac{r^2}{2}+\frac{(r'-r)^2}{8}\right) \geq 2(1-K)\frac{r(r'-r)}{4} \geq \frac{(1-K)r(r'-r)}{4},$$

which indicates that

$$p(\mathbf{x}(r',\omega)) \leq 7p(\mathbf{x}(r,\omega))\cdot\exp\left(-\frac{1-K}{4}r(r'-r)\right).$$

This proves (16).

Finally, swapping r and r' in (16), and noticing that

$$0 \leq \exp\left(-\frac{1-K}{4}r(r'-r)\right) \leq 1,$$

we get (15). ■

Lemma 20 *Suppose $p = \pi * \mathcal{N}(0, I_d)$, where π is a K-subgaussian distribution with $K < 1$. For $\mathbf{x} \in \mathbb{R}^d$, we have*

$$\|\nabla \log p(\mathbf{x})\|_2 \leq 3\|\mathbf{x}\|_2 + 12, \quad \|\nabla \log p(\mathbf{x})\|_2 \leq 3\|r\|_2 + 12.$$

Proof According to Proposition 2 in Polyanskiy and Wu (2016), we only need to verify $\mathbb{E}[\|X\|_2] \leq 3$, where $X \sim \pi$. Indeed, applying K-subgaussianity, we have

$$\mathbb{E}[\|X\|_2] = \int_0^\infty \mathbf{P}[\|X\|_2 \geq r]dr \leq \int_0^\infty \exp\left(-\frac{r^2}{2K^2}\right)dr = \sqrt{2\pi}K \leq 3.$$

■

Proof [Proof of Theorem 4] First we can write

$$\mathrm{KL}(p\|q) = \int_\Omega \int_0^\infty r^{d-1}p(\mathbf{x}(r,\omega))\log\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}drd\omega \tag{17}$$

$$= \int_\Omega \int_0^\infty r^{d-1}\left(p(\mathbf{x}(r,\omega))\log\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - p(\mathbf{x}(r,\omega)) + q(\mathbf{x}(r,\omega))\right)drd\omega \tag{18}$$

$$H^2(p,q) = \int_\Omega \int_0^\infty r^{d-1}p(\mathbf{x}(r,\omega))\left(\sqrt{\frac{q(\mathbf{x}(r,\omega))}{p(\mathbf{x}(r,\omega))}}-1\right)^2 drd\omega \tag{19}$$

For every $\omega \in \Omega$, we define r_ω as

$$r_\omega = \inf\left\{r:\ \log\frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))} \geq \log 3 + \frac{(3-\sqrt{K})^2}{2(1-\sqrt{K})^2}\right\}.$$

We notice that for every $\mathbf{x} \in \mathbb{R}^d$,

$$p(\mathbf{x}) = \int_{\mathbb{R}^d}\pi(\mathbf{y})\varphi(\mathbf{x}-\mathbf{y})d\mathbf{y} \leq \frac{1}{\sqrt{2\pi}^d}\int_{\mathbb{R}^d}\pi(\mathbf{y})d\mathbf{y} = \frac{1}{\sqrt{2\pi}^d},$$

and we further have

$$q(\mathbf{x}) = \int_{\mathbb{R}^d} \eta(\mathbf{y})\varphi(\mathbf{x}-\mathbf{y})d\mathbf{y} \geq \int_{\|\mathbf{y}\|_2\leq 1} \eta(\mathbf{y})\varphi(\mathbf{x}-\mathbf{y})d\mathbf{y}$$
$$\geq (1-e^{-1/2}) \cdot \min_{\|\mathbf{y}\|\leq 1} \varphi(\mathbf{x}-\mathbf{y}) \geq \frac{1}{3}\cdot\frac{1}{\sqrt{2\pi}^d}\exp\left(-\frac{(\|\mathbf{x}\|+1)^2}{2}\right).$$

Therefore, for all $\mathbf{x}$ such that $\|\mathbf{x}\|_2 < \frac{2}{1-\sqrt{K}}$, we have

$$\log\frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))} < \log 3 + \frac{(3-\sqrt{K})^2}{2(1-\sqrt{K})^2} \leq \log 3 + \frac{18}{(1-K)^2} \leq \frac{20}{(1-K)^2}. \tag{20}$$

Hence for every $\omega\in\Omega$, we have

$$r_\omega \geq \frac{2}{1-\sqrt{K}} = \frac{2(1+\sqrt{K})}{1-K} \geq \frac{2}{1-K}. \tag{21}$$

Then according to Lemma 19 and Lemma 20, we know that for every $r \geq r_\omega$, we have

$$p(\mathbf{x}(r,\omega)) \leq 7p(\mathbf{x}(r_\omega,\omega))\exp\left(-\frac{1-K}{4}r_\omega(r-r_\omega)\right),$$
$$\log\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} \leq T + (r-r_\omega)(3r+3r_\omega+24),$$

where T is defined in (20). Therefore, we obtain that

$$\int_{r_\omega}^\infty r^{d-1}p(\mathbf{x}(r,\omega))\log\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}dr$$
$$\leq 7p(\mathbf{x}(r_\omega,\omega))\int_{r_\omega}^\infty r^{d-1}\left(\frac{20}{(1-K)^2}+(r-r_\omega)(3r+3r_\omega+24)\right)\cdot\exp\left(-\frac{1-K}{4}r_\omega(r-r_\omega)\right)dr. \tag{22}$$

We adopt the changes of variables from r to $t = r - r_\omega$, and obtain that

$$\int_{r_\omega}^\infty r^{d-1}\exp\left(-\frac{1-K}{4}r_\omega(r-r_\omega)\right)dr = r_\omega^{d-1}\int_0^\infty \exp\left((d-1)\log\frac{t+r_\omega}{r_\omega} - \frac{(1-K)tr_\omega}{4}\right)dt$$

and

$$\int_{r_\omega}^\infty r^{d-1}(r-r_\omega)(3r+3r_\omega+24)\exp\left(-\frac{1-K}{4}r_\omega(r-r_\omega)\right)dr$$
$$= r_\omega^{d-1}\int_0^\infty \exp\left((d-1)\log\frac{t+r_\omega}{r_\omega} + \log t + \log(3t+6r_\omega+24) - \frac{(1-K)tr_\omega}{4}\right)dt$$

We first bound the first term in (22). We define

$$f(t) \triangleq (d-1)\log\frac{t+r_\omega}{r_\omega} - \frac{1-K}{8}r_\omega t,$$

then noticing that $d \le \frac{1}{2(1-K)}+1$ and hence $8(d-1) \le \frac{4}{1-K} \le (1-K)r_\omega^2$ holds for $r_\omega \ge \frac{2}{1-\sqrt{K}} \ge \frac{2}{1-K}$, its derivative satisfies that

$$f'(t) = \frac{d-1}{t+r_\omega} - \frac{(1-K)r_\omega}{8} \le \frac{d-1}{r_\omega} - \frac{(1-K)r_\omega}{8} \le 0, \quad \forall t \ge 0.$$

Therefore, for every $t \ge 0$, we have

$$(d-1)\log\frac{t+r_\omega}{r_\omega} - \frac{1-K}{8}r_\omega t = f(t) \le f(0) = 0,$$

which indicates that

$$\frac{1}{r_\omega^{d-1}}\int_{r_\omega}^{\infty} r^{d-1}\exp\left(-\frac{1-K}{4}r_\omega(r-r_\omega)\right)dr \le \int_0^\infty \exp\left(-\frac{1-K}{8}r_\omega t\right)dt = \frac{8}{(1-K)r_\omega} \tag{23}$$

We next bound the second term in (22). We define

$$g(t) \triangleq \log t + \log(3t+6r_\omega+24) - \frac{1-K}{16}r_\omega t,$$

then we have

$$g'(t) = \frac{1}{t} + \frac{1}{3t+6r_\omega+24} - \frac{1-K}{16}r_\omega,$$

which has a single root t_0 on $(0,\infty)$. And we have $g(t) \le g(t_0)$ holds for all $t \ge 0$. We further have

$$0 = g'(t_0) = \frac{1}{t_0} + \frac{1}{3t_0+6r_\omega+24} - \frac{1-K}{16}r_\omega \le \frac{4}{3t_0} - \frac{1-K}{16}r_\omega,$$

which indicates that $t_0 \le \frac{64}{3(1-K)r_\omega}$. Next noticing that $r_\omega \ge \frac{2}{1-\sqrt{K}}$, we have $(1-K)r_\omega \ge 2(1+\sqrt{K}) \ge 2$, hence we get

$$3t_0+6r_\omega+24 \le \frac{64}{(1-K)r_\omega} + 6r_\omega + 24 \le 6r_\omega + 56 \le 34 r_\omega.$$

Therefore for all $t \ge 0$,

$$g(t) \le g(t_0) \le \log t_0(3t_0+6r_\omega+24) - \frac{4}{3} \le \log\frac{2176}{3(1-K)} - \frac{4}{3},$$

Combine this result with our previous estimation on f, we get

$$\begin{aligned}&(d-1)\log\frac{t+r_\omega}{r_\omega} + \log t + \log(3t+6r_\omega+24) - \frac{1-K}{4}r_\omega t \\ &= f(t)+g(t) - \frac{1-K}{16}r_\omega t \le \log\frac{2176}{3(1-K)} - \frac{4}{3} - \frac{(1-K)r_\omega t}{16}.\end{aligned}$$

Hence we obtain

$$\begin{aligned}&\int_{r_\omega}^\infty r^{d-1}(r-r_\omega)(3r+3r_\omega+24)\exp\left(-\frac{1-K}{4}r_\omega(r-r_\omega)\right)dr \\ &\le r_\omega^{d-1}\int_0^\infty \exp\left(\log\frac{2176}{3(1-K)} - \frac{4}{3} - \frac{(1-K)r_\omega t}{16}\right)dt = \frac{2176e^{-4/3}}{3(1-K)}r_\omega^{d-1}\cdot\frac{16}{(1-K)r_\omega} \le \frac{3100 r_\omega^{d-2}}{(1-K)^2}.\end{aligned}$$

Therefore, according to (22) we have

$$\int_{r_\omega}^{\infty} r^{d-1} p(\mathbf{x}(r,\omega)) \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} dr \leq \left(\frac{160 r_\omega^{d-2}}{(1-K)^3} + \frac{3100 r_\omega^{d-2}}{(1-K)^2} \right) \cdot 7 p(\mathbf{x}(r_\omega,\omega)) \leq \frac{23000 r_\omega^{d-2} p(\mathbf{x}(r_\omega,\omega))}{(1-K)^3},$$

Similarly, we can also obtain bound on $\int_{r_\omega}^{\infty} r^{d-1} q(\mathbf{x}(r,\omega)) dr$: According to Lemma 19 we obtain that

$$q(\mathbf{x}(r,\omega)) \leq 7 q(\mathbf{x}(r_\omega,\omega)) \exp\left(-\frac{1-K}{4} r_\omega (r - r_\omega) \right),$$

which indicates that

$$\int_{r_\omega}^{\infty} r^{d-1} q(\mathbf{x}(r,\omega)) dr \leq q(\mathbf{x}(r_\omega,\omega)) \int_{r_\omega}^{\infty} r^{d-1} \exp\left(-\frac{1-K}{4} r_\omega (r - r_\omega) \right) dr.$$

According to (23), we get

$$\int_{r_\omega}^{\infty} r^{d-1} \exp\left(-\frac{1-K}{4} r_\omega (r - r_\omega) \right) dr \leq \frac{8 r_\omega^{d-2}}{1-K},$$

which indicates that

$$\int_{r_\omega}^{\infty} r^{d-1} q(\mathbf{x}(r,\omega)) dr \leq \frac{56 r_\omega^{d-2} q(\mathbf{x}(r_\omega,\omega))}{1-K}$$

Next noticing $\log \frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))} = \log 3 + \frac{(3-\sqrt{K})^2}{2(1-\sqrt{K})^2} \geq 0$. we have $q(\mathbf{x}(r_\omega,\omega)) \leq p(\mathbf{x}(r_\omega,\omega))$. Therefore, we have

$$\int_{r_\omega}^{\infty} r^{d-1} q(\mathbf{x}(r,\omega)) dr \leq \frac{56 r_\omega^{d-2} p(\mathbf{x}(r_\omega,\omega))}{1-K} \leq \frac{56 r_\omega^{d-2} p(\mathbf{x}(r_\omega,\omega))}{(1-K)^3}.$$

Therefore, we have the following upper bound

$$\begin{aligned}
&\int_\Omega \int_{r_\omega}^{\infty} r^{d-1} \left(p(\mathbf{x}(r,\omega)) \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - p(\mathbf{x}(r,\omega)) + q(\mathbf{x}(r,\omega)) \right) dr d\omega \\
&\leq \int_\Omega \int_{r_\omega}^{\infty} r^{d-1} p(\mathbf{x}(r,\omega)) \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} dr d\omega + \int_\Omega \int_{r_\omega}^{\infty} r^{d-1} q(\mathbf{x}(r,\omega)) dr d\omega \\
&\leq \int_\Omega \frac{23056 r_\omega^{d-2} p(\mathbf{x}(r_\omega,\omega))}{(1-K)^3} d\omega.
\end{aligned}$$

Next, according to (20), we notice that for any $\omega \in \Omega$ and $0 \leq r \leq r_\omega$ we have

$$\nabla \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} \leq 6r + 24 \leq 6 r_\omega + 24.$$

According to our choice of r_ω, we have $\log \frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))} = \log 3 + \frac{(3-\sqrt{K})^2}{2(1-\sqrt{K})^2}$. Hence noticing that $r_\omega \geq 2$ for every $\omega \in \Omega$ according to (21), we have for any $r_\omega - \frac{1}{18 r_\omega} \leq r \leq r_\omega$,

$$\log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} \geq \log 3 + \frac{(3-\sqrt{K})^2}{2(1-\sqrt{K})^2} - \frac{6 r_\omega + 24}{18 r_\omega} \geq 1 + 2 - 1 = 2,$$

which indicates that

$$\left(\sqrt{\frac{q(\mathbf{x}(r,\omega))}{p(\mathbf{x}(r,\omega))}}-1\right)^2 \geq \left(\frac{1}{e^1}-1\right)^2 \geq \frac{1}{3}.$$

Additionally, according to Lemma 16, for every $r_\omega - \frac{1}{18r_\omega} \leq r \leq r_\omega$, we have $p(\mathbf{x}(r,\omega)) \geq \frac{p(\mathbf{x}(r_\omega,\omega))}{7}$. We further adopt the assumption $d \leq \frac{1}{2(1-K)} + 1$ and also use (21) to get

$$r^{d-1} \geq r_\omega^{d-1}\cdot\left(1-\frac{1}{18r_\omega^2}\right)^{d-1} \geq r_\omega^{d-1}\cdot\left(1-\frac{d-1}{18r_\omega^2}\right) \geq r_\omega^{d-1}\left(1-\frac{1}{36}\right) \geq \frac{1}{2}r_\omega^{d-1}.$$

Therefore, we have

$$\begin{aligned}
&\int_\Omega \int_{r_\omega-\frac{1}{18r_\omega}}^{r_\omega} r^{d-1}p(\mathbf{x}(r,\omega))\cdot\left(\sqrt{\frac{q(\mathbf{x}(r,\omega))}{p(\mathbf{x}(r,\omega))}}-1\right)^2 dr d\omega \\
\geq &\int_\Omega \frac{1}{18r_\omega}\cdot\frac{1}{2}r_\omega^{d-1}\cdot p(\mathbf{x}(r_\omega,\omega))\cdot\frac{1}{2} \geq \frac{r_\omega^{d-2}p(\mathbf{x}(r_\omega,\omega))}{72}d\omega \\
\geq &\frac{(1-K)^3}{1660032}\int_\Omega\int_{r_\omega}^{\infty} r^{d-1}\left(p(\mathbf{x}(r,\omega))\log\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}-p(\mathbf{x}(r,\omega))+q(\mathbf{x}(r,\omega))\right)dr d\omega.
\end{aligned}$$

Therefore, according to (17), we have

$$\begin{aligned}
H^2(p,q) &= \int_\Omega\int_0^\infty r^{d-1}p(\mathbf{x}(r,\omega))\left(\sqrt{\frac{q(\mathbf{x}(r,\omega))}{p(\mathbf{x}(r,\omega))}}-1\right)^2 dr d\omega \\
&\geq \int_\Omega \int_{r_\omega-\frac{1}{r_\omega+M}}^{r_\omega} r^{d-1}p(\mathbf{x}(r,\omega))\left(\sqrt{\frac{q(\mathbf{x}(r,\omega))}{p(\mathbf{x}(r,\omega))}}-1\right)^2 dr d\omega \\
&\geq \frac{(1-K)^3}{1660032}\int_\Omega\int_{r_\omega}^{\infty} r^{d-1}\left(p(\mathbf{x}(r,\omega))\log\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}-p(\mathbf{x}(r,\omega))+q(\mathbf{x}(r,\omega))\right)dr d\omega.
\end{aligned}$$

Next we consider those $\mathbf{x}(r,\omega)$ with $0 \leq r \leq r_\omega$. According to the definition of r_ω, we know that for any such r, we have

$$\log\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} \leq \log 3 + \frac{(3-\sqrt{K})^2}{2(1-\sqrt{K})^2} \leq \frac{20}{(1-K)^2}.$$

According to Lemma 15, we have

$$\begin{aligned}
\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}\log\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} + 1 &\leq \frac{22.5}{(1-K)^2}\left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}}-1\right)^2 \\
&\leq \frac{24}{(1-K)^2}\left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}}-1\right)^2.
\end{aligned}$$

Hence we obtain that

$$\begin{aligned}
&\int_0^{r_\omega} r^{d-1}\left(p(\mathbf{x}(r,\omega))\log\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - p(\mathbf{x}(r,\omega)) + q(\mathbf{x}(r,\omega))\right)dr \\
&= \int_0^{r_\omega} r^{d-1} q(\mathbf{x}(r,\omega))\cdot\left(\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}\log\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} + 1\right)dr \\
&\le \int_0^{r_\omega} r^{d-1} q(\mathbf{x}(r,\omega))\cdot\frac{24}{(1-K)^2}\left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}} - 1\right)^2 dr.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\int_\Omega\int_0^{r_\omega} r^{d-1}\left(p(\mathbf{x}(r,\omega))\log\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - p(\mathbf{x}(r,\omega)) + q(\mathbf{x}(r,\omega))\right)drd\omega \\
&\le \frac{24}{(1-K)^2}\int_\Omega\int_0^{r_\omega} r^{d-1} q(\mathbf{x}(r,\omega))\cdot\left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}} - 1\right)^2 drd\omega \\
&\le \frac{24}{(1-K)^2}\int_\Omega\int_0^{\infty} r^{d-1} q(\mathbf{x}(r,\omega))\cdot\left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}} - 1\right)^2 drd\omega = \frac{24}{(1-K)^2}M^2H^2(p,q).
\end{aligned}$$

Combine these two analysis together, we obtain that

$$\begin{aligned}
\mathrm{KL}(p\|q) &= \int_\Omega\int_0^{\infty} r^{d-1}\left(p(\mathbf{x}(r,\omega))\log\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - p(\mathbf{x}(r,\omega)) + q(\mathbf{x}(r,\omega))\right)drd\omega \\
&= \int_\Omega\int_0^{r_\omega} r^{d-1}\left(p(\mathbf{x}(r,\omega))\log\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - p(\mathbf{x}(r,\omega)) + q(\mathbf{x}(r,\omega))\right)drd\omega \\
&\quad + \int_\Omega\int_{r_\omega}^{\infty} r^{d-1}\left(p(\mathbf{x}(r,\omega))\log\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - p(\mathbf{x}(r,\omega)) + q(\mathbf{x}(r,\omega))\right)drd\omega \\
&\le \frac{24}{(1-K)^2}H^2(p,q) + \frac{1660032}{(1-K)^3}H^2(p,q) = \frac{1660056}{(1-K)^3}H^2(p,q).
\end{aligned}$$

This completes the proof of Theorem 4. ■

Appendix D. Proof of Theorem 7

Applying (5) with $\mathbb{P} = f_\pi$ and $\mathbb{S} = \mathbb{Q} = f_\eta$, as long as (4) holds we get

$$\mathrm{KL}(f_\pi\|f_\eta) \le \frac{4\log(1/\delta)}{(1-\delta)^2}H^2(f_\pi,f_\eta) + \frac{4\delta\log(1/\delta)}{(1-\delta)^2} + \delta^{\frac{\lambda-1}{2}}\cdot\int_{\mathbb{R}^d}\frac{f_\pi(\mathbf{x})^\lambda}{f_\eta(\mathbf{x})^{\lambda-1}}d\mathbf{x}. \tag{24}$$

Notice that the last term is an f_λ-divergence D_{f_λ} with $f_\lambda(x) = x^\lambda$, which is a convex function for $\lambda \ge 1$, hence $\int_{\mathbb{R}^d}\frac{f_\pi(\mathbf{x})^\lambda}{f_\eta(\mathbf{x})^{\lambda-1}}d\mathbf{x}$ is convex in (f_π, f_η). Therefore, by Jensen's inequality we have

$$\begin{aligned}
\int_{\mathbb{R}^d}\frac{f_\pi(\mathbf{x})^\lambda}{f_\eta(\mathbf{x})^{\lambda-1}}d\mathbf{x} &= D_{f_\lambda}(\mathbb{E}[\delta_X * \mathcal{N}(0,I_d)]\|\mathbb{E}[\delta_{X'} * \mathcal{N}(0,I_d)]) \\
&\le \mathbb{E}[D_{f_\lambda}(\mathcal{N}(X,I_d)\|\mathcal{N}(X',I_d))] \\
&= \mathbb{E}\left[\exp\left(\frac{\lambda(\lambda-1)}{2}\|X-X'\|_2^2\right)\right]
\end{aligned}$$

for any possible coupling between (X, X') where $X \sim \pi, X' \sim \eta$.

Now according to the definition of subgaussian distributions, we have

$$\mathbf{P}[\|X\|_2 \geq t] \leq \exp\left(-\frac{t^2}{2K^2}\right), \quad \mathbf{P}[\|X'\|_2 \geq t] \leq \exp\left(-\frac{t^2}{2K^2}\right), \quad \forall t \geq 0,$$

which indicates that $\mathbf{P}[\|X\|_2 \geq 2K], \mathbf{P}[\|X'\|_2 \geq 2K] < \frac{1}{2}$. Therefore, we can construct the coupling between (X, X') so that if $\|X'\|_2 \geq 2K$ we always have $\|X\|_2 < 2K$, and if $\|X\|_2 \geq 2K$ we always have $\|X'\|_2 < 2K$. And we have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\lambda(\lambda-1)}{2}\|X-X'\|_2^2\right)\right] \\
&= \mathbb{E}\left[\exp\left(\frac{\lambda(\lambda-1)}{2}\|X-X'\|_2^2\right)\mathbf{1}_{\|X\|_2\geq 2K}\right] + \mathbb{E}\left[\exp\left(\frac{\lambda(\lambda-1)}{2}\|X-X'\|_2^2\right)\mathbf{1}_{\|X\|_2< 2K}\right] \\
&\leq \mathbb{E}\left[\exp\left(\frac{\lambda(\lambda-1)}{2}\|X-X'\|_2^2\right)\mathbf{1}_{\|X'\|_2< 2K}\right] + \mathbb{E}\left[\exp\left(\frac{\lambda(\lambda-1)}{2}\|X-X'\|_2^2\right)\mathbf{1}_{\|X\|_2< 2K}\right] \\
&\leq \mathbb{E}\left[\exp\left(\frac{\lambda(\lambda-1)}{2}(\|X\|_2+2K)^2\right)\right] + \mathbb{E}\left[\exp\left(\frac{\lambda(\lambda-1)}{2}(\|X'\|_2+2K)^2\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\lambda(\lambda-1)\left(\|X\|_2^2+4K^2\right)\right)\right] + \mathbb{E}\left[\exp\left(\lambda(\lambda-1)\left(\|X'\|_2^2+4K^2\right)\right)\right]
\end{aligned}$$

for any $\delta > 0$.

Since π and η are K-subgaussian distributions, we have,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\|X\|_2^2}{4K^2}\right)\right] = \int_0^\infty \exp\left(\frac{t^2}{4K^2}\right) d\pi[\|X\| \leq t] \leq \int_0^\infty \exp\left(\frac{t^2}{4K^2}\right)\cdot\frac{t}{K^2}\exp\left(-\frac{t^2}{2K^2}\right)dt \\
&= \int_0^\infty \exp\left(-\frac{t^2}{4K^2}\right) d\left(\frac{t^2}{2K^2}\right) = 2,
\end{aligned}$$

and similarly we also have

$$\mathbb{E}\left[\exp\left(\frac{\|X'\|_2^2}{4K^2}\right)\right] \leq 2.$$

We choose $\lambda = \frac{1+\sqrt{1+1/K^2}}{2}$, and we have $\lambda(\lambda-1) = \frac{1}{4K^2}$. Hence we get

$$\mathbb{E}\left[\exp\left(\frac{\lambda(\lambda-1)}{2}\|X-X'\|_2^2\right)\right] \leq \left(\mathbb{E}\left[\exp\left(\frac{\|X\|_2^2}{4K^2}\right)\right] + \mathbb{E}\left[\exp\left(\frac{\|X'\|_2^2}{4K^2}\right)\right]\right)\cdot\exp(1) \leq 4e \leq 12.$$

Therefore, according to (24), as long as (4) holds we get

$$\mathrm{KL}(f_\pi\|f_\eta) \leq \frac{4\log(1/\delta)}{(1-\delta)^2}H^2(f_\pi, f_\eta) + \frac{4\delta\log(1/\delta)}{(1-\delta)^2} + 12\delta^{\frac{\lambda-1}{2}}.$$

Choosing $\delta = \left(\frac{H^2(f_\pi,f_\eta)}{4}\right)^{\frac{8}{(\lambda-1)^2}\vee 1}$, and we will get

$$\delta \leq \frac{H^2(f_\pi, f_\eta)}{4} \leq \frac{1}{2} \quad \text{and} \quad \delta^{\frac{\lambda-1}{2}} \leq H^2(f_\pi, f_\eta).$$

Therefore, the first inequality in (4) holds. As for the second one, we notice that if $\lambda \geq 2$, then $\frac{\lambda-1}{2} \geq \frac{1}{2}$ and it always holds. For $1 < \lambda \leq 2$, we have $\frac{8}{(\lambda-1)^2} \vee 1 = \frac{8}{(\lambda-1)^2}$, and hence $\log \frac{1}{\delta} \geq \frac{8\log(2)}{(\lambda-1)^2} \geq \frac{4}{(\lambda-1)^2} > 4$. Since $\frac{\log t}{t}$ is decreasing for $t \geq 4$, we have $\frac{\log\log(1/\delta)}{\log(1/\delta)} \leq \frac{\log(4/(\lambda-1)^2)}{4/(\lambda-1)^2}$. Moreover, using the fact that $x > 2\log x$ holds for all $x \geq 0$, we have $\frac{\lambda-1}{2} \cdot \frac{4}{(\lambda-1)^2} - \log \frac{4}{(\lambda-1)^2} = \frac{2}{\lambda-1} - 2\log\frac{2}{\lambda-1} > 0$. Hence we get $\frac{\log\log(1/\delta)}{\log(1/\delta)} \leq \frac{\lambda-1}{2}$, which proves that the second inequality in (4) always holds.

Therefore, noticing that $(1-\delta)^2 \geq \frac{1}{4}$ and also

$$\begin{aligned}\log(1/\delta) &\leq \left(\frac{8}{(\lambda-1)^2} \vee 1\right) \log \frac{4}{H^2(f_\pi, f_\eta)} = (32K^2(K+\sqrt{K^2+1})^2 \vee 1) \log \frac{4}{H^2(f_\pi, f_\eta)} \\ &\leq (512K^4 + 32) \log \frac{4}{H^2(f_\pi, f_\eta)},\end{aligned}$$

we get

$$\mathrm{KL}(f_\pi \| f_\eta) \leq (10240K^4 + 652) H^2(f_\pi, f_\eta) \log \frac{4}{H^2(f_\pi, f_\eta)}.$$

Appendix E. Comparison inequalities for other distances

In this appendix, we discuss comparison inequalities for other popular distances between densities, namely, the χ^2-divergence, the TV distance, and the L_2 distance.

First we presents the results of $\chi^2 \lesssim H^2$, where $\chi^2(f\|g) = \int \frac{(f-g)^2}{g}$.

Theorem 21 *For d-dimensional distributions π, η supported on $B_2(M)$ with $M \geq 2$, we have*

$$\chi^2(f_\pi \| f_\eta) \leq 2\exp\left(50(M^2 \vee d)\right) H^2(f_\pi, f_\eta).$$

Next, we show that for one-dimensional Gaussian mixtures where the mixing distribution is compact supported, TV distance and L_2 distance are close to each other up to log factors.

Theorem 22 *Suppose π and η are one-dimensional distributions supported on $[-M, M]$ with $M \geq 1$. Then we have*

$$\mathrm{TV}(f_\pi, f_\eta) \leq \left(8\sqrt{M} + 2\log^{1/4} \frac{1}{\|f_\pi - f_\eta\|_2}\right) \|f_\pi - f_\eta\|_2$$

Theorem 23 *For any one-dimensional distributions π, η (that need not be compactly supported),*

$$\|f_\pi - f_\eta\|_2 \leq \left(\log^{1/4} \frac{1}{\mathrm{TV}(f_\pi, f_\eta)} \vee 3\right) \mathrm{TV}(f_\pi, f_\eta).$$

We discuss a statistical application of these results. The L_2 squared minimax estimation rates for all Gaussian mixtures are shown in Kim (2014); Kim and Guntuboyina (2022) to be $\Theta\left(\log^{d/2} n/n\right)$, which is sharp for all constant d. Therefore, equipped with the above comparison theorems, we can also get an upper bound on the minimax estimation rates under the TV distance. Previously, Ashtiani et al. (2020) showed a rate $\tilde{\mathcal{O}}(\sqrt{kd^2/n})$ for k-atomic Gaussian mixtures, where $\tilde{O}$ hides polylog

factors. For one-dimensional Gaussian mixtures with compactly supported mixing distributions, the best TV upper bound so far is $\mathcal{O}\left(\log^{3/8} n/\sqrt{n}\right)$, which in fact follows from combining the sharp L_2 rate and Theorem 22.

More details and proofs are provided in Section E.1 and E.2. Throughout this appendix, for simplicity we abbreviate $p \equiv f_\pi$ and $q \equiv f_\eta$.

E.1. Proof of Theorem 21

Lemma 24 *Suppose $p = \pi * \mathcal{N}(0, I_d), q = \eta * \mathcal{N}(0, I_d)$ where $\mathtt{supp}(\pi), \mathtt{supp}(\eta) \subset B_2(M)$, then for every $r \geq r_\omega \geq M$, we have*

$$\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} \leq \frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))} \exp\left(2(r-r_\omega)M\right).$$

Proof We first prove that for any $\omega \in \Omega$ and $r \geq r_\omega \geq M$,

$$q(\mathbf{x}(r,\omega)) \geq q(\mathbf{x}(r_\omega,\omega)) \exp\left(-\frac{(r+M)^2-(r_\omega+M)^2}{2}\right).$$

Without loss of generality, we assume $\omega = (1, 0, \cdots, 0)$. Then for any $\mathbf{u} = (u_1, u_2, \cdots, u_d) \in B_2(M)$, we have $|u_1| \leq M$ and

$$\varphi(\mathbf{x}(r,\omega)-\mathbf{u}) = \frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{(r-u_1)^2+\sum_{i=2}^d u_i^2}{2}\right)$$
$$\varphi(\mathbf{x}(r_\omega,\omega)-\mathbf{u}) = \frac{1}{\sqrt{2\pi}^d} \exp\left(-\frac{(r_\omega-u_1)^2+\sum_{i=2}^d u_i^2}{2}\right).$$

Noticing that $|u_1| \leq M \leq r_\omega \leq r$, we have $(r-u_1)^2-(r_\omega-u_1)^2 \leq (r+M)^2-(r_\omega+M)^2$, which indicates that

$$-\frac{(r-u_1)^2+\sum_{i=2}^d u_i^2}{2} \geq -\frac{(r_\omega-u_1)^2+\sum_{i=2}^d u_i^2}{2} - \frac{(r+M)^2-(r_\omega+M)^2}{2},$$

and hence

$$\varphi(\mathbf{x}(r,\omega)-\mathbf{u}) \geq \varphi(\mathbf{x}(r_\omega,\omega)-\mathbf{u}) \exp\left(-\frac{(r+M)^2-(r_\omega+M)^2}{2}\right).$$

Since we can write

$$q(\mathbf{x}(r,\omega)) = \int_{B_2(M)} \eta(\mathbf{u})\varphi(\mathbf{x}(r,\omega)-\mathbf{u})d\mathbf{u},$$

we can verify that

$$q(\mathbf{x}(r,\omega)) \geq q(\mathbf{x}(r_\omega,\omega)) \exp\left(-\frac{(r+M)^2-(r_\omega+M)^2}{2}\right).$$

Next, we notice that according to Lemma 16, we have

$$p(\mathbf{x}(r,\omega)) \leq p(\mathbf{x}(r_\omega,\omega)) \exp\left(-\frac{(r-M)^2-(r_\omega-M)^2}{2}\right).$$

This indicates that

$$\begin{aligned}\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} &\leq \frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))}\exp\left(-\frac{(r+M)^2-(r_\omega+M)^2}{2}+\frac{(r-M)^2-(r_\omega-M)^2}{2}\right)\\ &= \frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))}\exp\left(2(r-r_\omega)M\right).\end{aligned}$$

■

Proof [Proof of Theorem 21] Without loss of generality, we assume $d \leq M^2$. We write

$$\begin{aligned}\chi^2(p\|q) &= \int_\Omega \int_0^\infty r^{d-1} q(\mathbf{x}(r,\omega)\left(\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}-1\right)^2 dr d\omega\\ H^2(p,q) &= \frac{1}{2}\int_\Omega \int_0^\infty r^{d-1} q(\mathbf{x}(r,\omega))\left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}}-1\right)^2 dr d\omega\end{aligned} \tag{25}$$

For every $\omega \in \Omega$, we define r_ω as

$$r_\omega \triangleq \inf\left\{r \middle| \frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))} \geq \exp\left(25M^2\right)\right\}.$$

Notice that for any $r \leq 6M$ and $\omega \in \Omega$, we have

$$\begin{aligned}p(\mathbf{x}(r,\omega)) &= \int_{B_2(M)} \pi(\mathbf{u})\varphi(\mathbf{x}(r,\omega),\mathbf{u})d\mathbf{u} \leq \frac{1}{\sqrt{2\pi}^d}\\ q(\mathbf{x}(r,\omega)) &= \int_{B_2(M)} \eta(\mathbf{u})\varphi(\mathbf{x}(r,\omega),\mathbf{u})d\mathbf{u} \geq \frac{1}{\sqrt{2\pi}^d}\exp\left(-\frac{(6M+M)^2}{2}\right),\end{aligned}$$

which indicates that

$$\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega)} \leq \exp\left(\frac{49M^2}{2}\right) < \exp\left(25M^2\right).$$

Hence for every $\omega \in \Omega$, we all have $r_\omega \geq 6M$. And if $r_\omega \neq \infty$, we have that $\frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))} = \exp\left(25M^2\right)$. According to Lemma 16 and Lemma 24, we know that for every $r \geq r_\omega$, we all have

$$\begin{aligned}q(\mathbf{x}(r,\omega)) &\leq q(\mathbf{x}(r_\omega,\omega))\exp\left(-\frac{(r-M)^2-(r_\omega-M)^2}{2}\right),\\ \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} &\leq \frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))}\exp\left(2(r-r_\omega)M\right).\end{aligned}$$

Since $\frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))} = \exp(25M^2) \geq 1$, and $\exp\left(2(r-r_\omega)M\right) \geq 1$ for every $r \geq r_\omega$, we have

$$\left(\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}-1\right)^2 \leq \max\left\{1,\left(\frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))}\exp\left(2(r-r_\omega)M\right)\right)^2\right\} = \exp\left(50M^2+4(r-r_\omega)M\right).$$

Therefore, we obtain that

$$\int_{r_\omega}^{\infty} r^{d-1} q(\mathbf{x}(r,\omega)) \cdot \left(\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - 1\right)^2 dr$$
$$\le q(\mathbf{x}(r_\omega,\omega)) \int_{r_\omega}^{\infty} r^{d-1} \exp\left(50M^2 + 4(r-r_\omega)M - \frac{(r-M)^2 - (r_\omega - M)^2}{2}\right) dr.$$

We adopt the changes of variables from r to $t = r - r_\omega$, and obtain that

$$\int_{r_\omega}^{\infty} r^{d-1} \exp\left(50M^2 + 4(r-r_\omega)M - \frac{(r-M)^2 - (r_\omega - M)^2}{2}\right) dr$$
$$= r_\omega^{d-1} \exp(50M^2) \int_0^{\infty} \exp\left((d-1)\log\frac{t+r_\omega}{r_\omega} - \frac{t^2}{2} - (r_\omega - M)t + 4Mt\right) dt.$$

Next, we define

$$f(t) \triangleq (d-1)\log\frac{t+r_\omega}{r_\omega} - \frac{1}{2}(r_\omega - 5M)t.$$

Since

$$d - 1 < d \le M^2 < \frac{1}{2}\cdot(6M)\cdot(6M-5M) \le \frac{1}{2} r_\omega (r_\omega - 5M),$$

the first-order derivative of f satisfies that

$$f'(t) = \frac{d-1}{t+r_\omega} - \frac{1}{2}(r_\omega - 5M) \le \frac{d-1}{r_\omega} - \frac{1}{2}(r_\omega - 5M) \le 0, \quad \forall t \ge 0.$$

Therefore we have

$$(d-1)\log\frac{t+r_\omega}{r_\omega} - \frac{1}{2}(r_\omega - 5M)t = f(t) \le f(0) = 0, \quad \forall t \ge 0.$$

This directly indicates that

$$\int_0^{\infty} \exp\left((d-1)\log\frac{t+r_\omega}{r_\omega} - \frac{t^2}{2} - (r_\omega - M)t + 4Mt\right) dt$$
$$\le \int_0^{\infty} \exp\left(-\frac{t^2}{2} - \frac{1}{2}(r_\omega - 5M)t\right) dt \le \int_0^{\infty} \exp\left(-\frac{1}{2}(r_\omega - 5M)t\right) dt = \frac{2}{r_\omega - 5M}$$

Therefore, we obtain that

$$\int_{r_\omega}^{\infty} r^{d-1} q(\mathbf{x}(r,\omega)) \cdot \left(\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - 1\right)^2 dr \le \frac{2 r_\omega^{d-1} \exp(50M^2) q(\mathbf{x}(r_\omega,\omega))}{r_\omega - 5M}$$

Next, according to Lemma 17, for $\forall \omega \in \Omega$ and $r \in [0, r_\omega]$ we have

$$\nabla \log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} = \nabla \log p(\mathbf{x}(r,\omega)) - \nabla \log q(\mathbf{x}(r,\omega)) \le 6r + 8M \le 6r_\omega + 8M \le 8r_\omega + 8M.$$

Notice that we also have $\log \frac{p(\mathbf{x}(r_\omega,\omega))}{q(\mathbf{x}(r_\omega,\omega))} = 25M^2$. Hence for $\forall r_\omega - \frac{1}{r_\omega + M} \le r \le r_\omega$,

$$\log \frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} \ge 25M^2 - (8r_\omega + 8M)\cdot\frac{1}{r_\omega + M} = 25M^2 - 8 \ge 2,$$

which indicates that

$$\left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}}-1\right)^2 \geq \left(e^2-1\right)^2 \geq 40.$$

We further notice that $r_\omega \geq 5M \geq M$ and $d-1 < d \leq M^2$. Hence for $\forall r_\omega - \frac{1}{r_\omega+M} \leq r \leq r_\omega$,

$$r^{d-1} \geq r_\omega^{d-1}\cdot\left(1-\frac{1}{r_\omega(r_\omega+M)}\right)^{d-1} \geq r_\omega^{d-1}\cdot\left(1-\frac{d-1}{r_\omega(r_\omega+M)}\right) \geq r_\omega^{d-1}\left(1-\frac{d-1}{30M^2}\right) \geq \frac{1}{2}r_\omega^{d-1}.$$

After noticing that $p(\mathbf{x}(r,\omega)) \geq p(\mathbf{x}(r_\omega,\omega))$ according to Lemma 16, we obtain

$$\begin{aligned}
&\int_{r_\omega-\frac{1}{r_\omega+M}}^{r_\omega} r^{d-1}q(\mathbf{x}(r,\omega))\cdot\left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}}-1\right)^2 dr\\
&\geq \frac{1}{r_\omega+M}\cdot\min_{r_\omega-\frac{1}{r_\omega+M}\leq r\leq r_\omega} r^{d-1}q(\mathbf{x}(r,\omega))\cdot\left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}}-1\right)^2\\
&\geq \frac{1}{r_\omega+M}\cdot\frac{1}{2}r_\omega^{d-1}\cdot 40q(\mathbf{x}(r_\omega,\omega)) \geq \frac{2r_\omega^{d-1}q(\mathbf{x}(r_\omega,\omega))}{r_\omega-5M}\\
&\geq \exp\left(-50M^2\right)\int_{r_\omega}^{\infty} r^{d-1}q(\mathbf{x}(r,\omega))\cdot\left(\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}-1\right)^2 dr.
\end{aligned}$$

Therefore, according to (25), we have

$$\begin{aligned}
H^2(p,q) &= \int_\Omega\int_0^\infty r^{d-1}p(\mathbf{x}(r,\omega))\left(\sqrt{\frac{q(\mathbf{x}(r,\omega))}{p(\mathbf{x}(r,\omega))}}-1\right)^2 dr d\omega\\
&\geq \int_\Omega\int_{r_\omega-\frac{1}{r_\omega+M}}^{r_\omega} r^{d-1}p(\mathbf{x}(r,\omega))\left(\sqrt{\frac{q(\mathbf{x}(r,\omega))}{p(\mathbf{x}(r,\omega))}}-1\right)^2 dr d\omega\\
&\geq \exp\left(-50M^2\right)\int_\Omega\int_{r_\omega}^\infty r^{d-1}q(\mathbf{x}(r,\omega))\cdot\left(\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}-1\right)^2 dr d\omega.
\end{aligned}$$

Next we consider those $\mathbf{x}(r,\omega)$ with $0 \leq r \leq r_\omega$. According to the definition of r_ω, we know that for any such r, we have

$$\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} \leq \exp\left(25M^2\right).$$

Notice the inequality

$$(t-1)^2 = (\sqrt{t}+1)^2(\sqrt{t}-1)^2 \leq \exp\left(50M^2\right)\left(\sqrt{t}-1\right)^2, \quad \forall 0 \leq t \leq \exp\left(25M^2\right).$$

Hence we obtain that

$$\begin{aligned}
&\int_0^{r_\omega} r^{d-1}q(\mathbf{x}(r,\omega))\cdot\left(\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}-1\right)^2 dr\\
&\leq \int_0^{r_\omega} r^{d-1}q(\mathbf{x}(r,\omega))\cdot\exp\left(50M^2\right)\left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}}-1\right)^2 dr.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\int_\Omega \int_0^{r_\omega} r^{d-1} q(\mathbf{x}(r,\omega)) \cdot \left(\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))} - 1\right)^2 dr d\omega \\
&\le \exp\left(50M^2\right) \int_\Omega \int_0^{r_\omega} r^{d-1} q(\mathbf{x}(r,\omega)) \cdot \left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}} - 1\right)^2 dr d\omega \\
&\le \exp\left(50M^2\right) \int_\Omega \int_0^{\infty} r^{d-1} q(\mathbf{x}(r,\omega)) \cdot \left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}} - 1\right)^2 dr d\omega = \exp\left(50M^2\right) H^2(p,q).
\end{aligned}$$

Combine these two analysis together, we obtain that

$$\begin{aligned}
\mathrm{KL}(p\|q) &= \int_\Omega \int_0^{\infty} r^{d-1} q(\mathbf{x}(r,\omega)) \cdot \left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}} - 1\right)^2 dr d\omega \\
&= \int_\Omega \int_0^{r_\omega} r^{d-1} q(\mathbf{x}(r,\omega)) \cdot \left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}} - 1\right)^2 dr d\omega \\
&\quad + \int_\Omega \int_{r_\omega}^{\infty} r^{d-1} q(\mathbf{x}(r,\omega)) \cdot \left(\sqrt{\frac{p(\mathbf{x}(r,\omega))}{q(\mathbf{x}(r,\omega))}} - 1\right)^2 dr d\omega \\
&\le \exp\left(50M^2\right) H^2(p,q) + \exp\left(50M^2\right) H^2(p,q) = 2\exp\left(50M^2\right) H^2(p,q).
\end{aligned}$$

This completes the proof of Theorem 21. ■

E.2. Proof of Theorem 22 and 23

Proof [Proof of Theorem 22] First we notice that for any $|t| \ge M$, we have

$$0 \le p(t) = \int_{|x|\ge M} \varphi(t-x) d\pi(x) \le \max_{|x|\ge M} \varphi(t-x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(|t|-M)^2}{2}\right).$$

Similarly we have the same estimation for $q(t)$. Hence we get

$$|p(t) - q(t)| \le \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(|t|-M)^2}{2}\right) \quad \forall |t| \ge M,$$

which indicates that for any $m \ge M$

$$\begin{aligned}
\int_{|x|\ge t} |p(x) - q(x)| dx &\le \frac{1}{\sqrt{2\pi}} \int_{|x|_2 \ge t} \exp\left(-\frac{(|x|-M)^2}{2}\right) ds \\
&= \frac{2}{\sqrt{2\pi}} \int_t^{\infty} \exp\left(-\frac{(s-M)^2}{2}\right) ds \le \frac{2}{t-M} \exp\left(-\frac{(t-M)^2}{2}\right).
\end{aligned}$$

Hence for $t \ge M + \frac{1}{3}$ we have

$$\int_{|x|\ge t} |p(x) - q(x)| dx \le 6 \exp\left(-\frac{(t-M)^2}{2}\right).$$

Additionally, according to Cauchy-Schwarz inequality we have

$$\left(\int_{|x|\le t}|p(t)-q(t)|dt\right)^2 \le 2t\cdot\left(\int_{|x|\le t}|p(x)-q(x)|^2dx\right),$$

which indicates that

$$\int_{|x|\le t}|p(x)-q(x)|dx \le \sqrt{2t}\cdot\sqrt{\int_{|x|\le t}|p(x)-q(x)|^2dx} \le \sqrt{2t}\cdot\|p-q\|_2.$$

Therefore, we obtain that for every $t \ge M+\frac{1}{3}$,

$$\mathrm{TV}(p,q)=\int_{|x|\ge t}|p(x)-q(x)|dx+\int_{|x|\le t}|p(x)-q(x)|dx \le 6\exp\left(-\frac{(t-M)^2}{2}\right)+\sqrt{2t}\cdot\|p-q\|_2.$$

Finally, since for any $x\in\mathbb{R}$, we have $0\le p(x),q(x)\le \max_{x\in\mathbb{R}^d}\varphi(x)=\frac{1}{\sqrt{2\pi}}\le\frac{2}{5}$, we obtain that

$$\|p-q\|_2^2=\int_{-\infty}^{\infty}(p(x)-q(x))^2dx \le \int_{-\infty}^{\infty}\frac{2(p(x)+q(x))}{5}dx=\frac{4}{5},$$

which indicates that

$$\log\frac{1}{\|p-q\|_2}=\frac{1}{2}\log\frac{1}{\|p-q\|_2^2}\ge\frac{1}{2}\log\frac{5}{4}\ge\frac{1}{10}.$$

Therefore, choosing

$$t=M+\sqrt{2\log\frac{1}{\|p-q\|_2}}\ge M+\sqrt{\frac{1}{5}}\ge M+\frac{1}{3},$$

we get

$$\begin{aligned}\mathrm{TV}(p,q)&\le 6\|p-q\|_2+\sqrt{2M+2\sqrt{2\log\frac{1}{\|p-q\|_2}}}\|p-q\|_2\\&\le 6\|p-q\|_2+\left(\sqrt{2M}+\sqrt{2\sqrt{2\log\frac{1}{\|p-q\|_2}}}\right)\|p-q\|_2\\&\le\left(8\sqrt{M}+2\log^{1/4}\frac{1}{\|p-q\|_2}\right)\|p-q\|_2.\end{aligned}$$

■

Proof [Proof of Theorem 23] For any distribution $\mathbb{P}$, we define its characteristic function $\Psi_{\mathbb{P}}:\mathbb{R}\to\mathbb{C}$ as:

$$\Psi_{\mathbb{P}}(t)\triangleq\mathbb{E}\left[e^{itX}\right],\quad X\sim\mathbb{P}.$$

Suppose the characteristic function of π, η, p, q are $\Psi_\pi, \Psi_\eta, \Psi_p, \Psi_q$, respectively. Then by Gaussian convolution

$$\Psi_p(t) = \Psi_\pi(t) \exp\left(-\frac{t^2}{2}\right), \quad \Psi_q(t) = \Psi_\eta(t) \exp\left(-\frac{t^2}{2}\right).$$

We notice that for every $t \in \mathbb{R}$, we have from Plancherel's identity

$$|\Psi_p(t) - \Psi_q(t)| = \left| \int_{-\infty}^{\infty} (p(x) - q(x)) e^{itx} dx \right| \le \int_{-\infty}^{\infty} |(p(x) - q(x)) e^{itx}| dx = \mathrm{TV}(p, q).$$

Similarly, we also have $|\Psi_\pi(t) - \Psi_\eta(t)| \le \|\pi - \eta\|_1 \le 2$, which indicates that for every $t \in \mathbb{R}$,

$$|\Psi_p(t) - \Psi_q(t)| = e^{-\frac{t^2}{2}} |\Psi_\pi(t) - \Psi_\eta(t)| \le 2 \exp\left(-\frac{t^2}{2}\right).$$

Therefore, we obtain that for any $s > 0$,

$$\begin{aligned}
\|\Psi_p - \Psi_q\|_2^2 &= \int_{-\infty}^{\infty} |\Psi_p(t) - \Psi_q(t)|^2 \, dt \\
&= \int_{-\infty}^{-s} |\Psi_p(t) - \Psi_q(t)|^2 \, dt + \int_{-s}^{s} |\Psi_p(t) - \Psi_q(t)|^2 \, dt + \int_{s}^{\infty} |\Psi_p(t) - \Psi_q(t)|^2 \, dt \\
&\le \int_{-\infty}^{-s} \exp\left(-\frac{t^2}{2}\right) dt + \int_{s}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt + 2s \cdot \mathrm{TV}(p, q)^2 \\
&= 2 \int_{0}^{\infty} \exp\left(-\frac{(t+s)^2}{2}\right) dt + 2s \cdot \mathrm{TV}(p, q)^2 = \frac{2}{s} \exp\left(-\frac{s^2}{2}\right) + 2s \cdot \mathrm{TV}(p, q)^2
\end{aligned}$$

When $\mathrm{TV}(p, q) \ge \frac{1}{e}$. We further notice that $|\Psi_p(t) - \Psi_q(t)| \le 2 \exp\left(-\frac{t^2}{2}\right)$, which indicates that

$$\|\Psi_p - \Psi_q\|_2^2 = \int_{-\infty}^{\infty} |\Psi_p(t) - \Psi_q(t)|^2 dt \le \int_{-\infty}^{\infty} 4 \exp(-t^2) dt = 4\sqrt{\pi} \le \frac{36}{e} \le 36 \mathrm{TV}(p, q)$$

When $\mathrm{TV}(p, q) \le \frac{1}{e}$, we have $s = \sqrt{2 \log \frac{1}{\mathrm{TV}(p,q)^2}} = 2\sqrt{\log \frac{1}{\mathrm{TV}(p,q)}} \ge 2$, we get

$$\|\Psi_p - \Psi_q\|_2^2 \le \left(1 + 2\sqrt{\log \frac{1}{\mathrm{TV}(p, q)}}\right) \cdot \mathrm{TV}(p, q)^2 \le 4\sqrt{\log \frac{1}{\mathrm{TV}(p, q)}} \cdot \mathrm{TV}(p, q)^2.$$

Above all, we get

$$\|\Psi_p - \Psi_q\|_2^2 \le 4 \left(\log^{1/4} \frac{1}{\mathrm{TV}(p, q)} \vee 3\right)^2 \cdot \mathrm{TV}(p, q),$$

which indicates that

$$\|p - q\|_2^2 = \sqrt{\frac{1}{2\pi} \|\Psi_p - \Psi_q\|_2} \le \left(\log^{1/4} \frac{1}{\mathrm{TV}(p, q)} \vee 3\right) \mathrm{TV}(p, q).$$

■