Information and Inference: A Journal of the IMA (2021) **00**, 1–78
doi:10.1093/imaiai/iaab013

Estimating location parameters in sample-heterogeneous distributions

Ankit Pensia[†]
Department of Computer Sciences, University of Wisconsin-Madison, Madison, Wisconsin, USA

Varun Jog
Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, UK

and

Po-Ling Loh
Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, UK
[†]Corresponding author: ankitp@cs.wisc.edu

[Received on 15 December 2020; accepted on 11 March 2021]

Estimating the mean of a probability distribution using i.i.d. samples is a classical problem in statistics, wherein finite-sample optimal estimators are sought under various distributional assumptions. In this paper, we consider the problem of mean estimation when independent samples are drawn from d-dimensional non-identical distributions possessing a common mean. When the distributions are radially symmetric and unimodal, we propose a novel estimator, which is a hybrid of the modal interval, shorth and median estimators and whose performance adapts to the level of heterogeneity in the data. We show that our estimator is near optimal when data are i.i.d. and when the fraction of 'low-noise' distributions is as small as $\Omega\left(\frac{d\log n}{n}\right)$, where n is the number of samples. We also derive minimax lower bounds on the expected error of any estimator that is agnostic to the scales of individual data points. Finally, we extend our theory to linear regression. In both the mean estimation and regression settings, we present computationally feasible versions of our estimators that run in time polynomial in the number of data points.

Keywords: location estimation; robust statistics; sample heterogeneity.

1. Introduction

Heterogeneity is prevalent in many modern data sets, leading to new challenges in estimation and prediction. The i.i.d. assumption imposed in much of classical statistics is unlikely to hold in practice, creating a need to develop new theory under relaxed assumptions allowing for heterogeneous data [12, 15, 32, 40, 49]. In this paper, we consider the problem of estimating a common mean when independent data are drawn from non-identical distributions.

A version of this problem for Gaussian distributions was recently studied in Chierichetti *et al.* [8], who motivated their work using the following crowdsourcing application: suppose the quality of an item is obtained by soliciting ratings from several agents, who are assumed to provide unbiased ratings. However, the rating distributions may vary across agents depending, e.g., on their expertise. In the Gaussian setting, this translates into data drawn from independent distributions with a common mean but possibly different variances. Chierichetti *et al.* [8] proposed a mean estimator based on calculating the 'shortest gap' between samples and derived upper bounds on the estimation error of their algorithm. Naturally, one might ask whether the estimators proposed by Chierichetti *et al.* [8] also perform provably

well for non-Gaussian settings; furthermore, although Chierichetti *et al.* [8] derived some lower bounds for the behavior of the best possible estimator in the unknown variance setting, the optimality of their proposed estimator was only partially addressed.

The work of this paper revisits the problem of common mean estimation and generalizes the case of Gaussian mixtures considered in Chierichetti *et al.* [8] to settings where the component distributions are only assumed to be symmetric and unimodal about a common mean. Although the estimators studied in our paper resemble the estimators proposed by Chierichetti *et al.* [8], our method of analysis is substantially different and allows us to obtain bounds without assuming Gaussianity, sub-Gaussianity or even finite variances of individual distributions. In the multivariate mean estimation setting, this leads to sharper estimation error rates than those obtained in Chierichetti *et al.* [8] for isotropic Gaussian data. The upper bounds we derive are stated in terms of percentiles of the overall mixture distribution and may be finite even in the case of heavy-tailed distributions.

The aforementioned model of non-i.i.d. data has even older roots in the statistics literature, under the name of *sample heterogeneity*. Initial research in sample heterogeneity focused on understanding the asymptotic distribution of order statistics and linear functions thereof [18, 37–39, 41, 47]. More recent work has established necessary and sufficient conditions for consistency of the sample median [16, 17, 34]. In particular, Hallin and Mizera [17] established the optimality of the median over a certain class of M-estimators (having a bounded, non-decreasing, skew-symmetric score function). However, as explained in more detail later (cf. Section 3.2), certain cases exist where the median itself is not optimal in comparison to more complicated estimators. For example, redescending M-estimators do not lie in the class studied by Hallin and Mizera [17]. We show that, under certain conditions, the modal interval estimator (Estimator 1)—which may be viewed as an extreme case of a redescending M-estimator—has smaller error than the median (cf. Table 1). Sample heterogeneity was also studied in the linear regression setting, where previous work focused on the least absolute deviation estimator [13, 24]. Inspired by the modal estimator, we propose and analyze a related estimator for linear regression (cf. Section 8). Note that we are chiefly interested in estimators which have minimal assumptions and allow the fraction of low-variance points to be as small as $\frac{\log n}{n}$, whereas the estimators studied in previous work required the fraction to be $\Omega\left(\frac{\sqrt{n}}{n}\right)$ [13, 17, 17, 24, 34].

We also briefly mention classical work on the modal interval estimator [7] and shorth estimator [4], which are used as building blocks for our hybrid estimator. Notably, a previous analysis has focused on asymptotic results for i.i.d. data, where both the modal interval and shorth estimators were proven to have an $n^{-\frac{1}{3}}$ convergence rate [23], in contrast to the faster $n^{-\frac{1}{2}}$ convergence rate of the sample mean. The results of this paper show the benefit of these estimators when a substantial fraction of the component distributions have large (or even infinite) variances, underscoring the general fact that robustness may need to be traded off for efficiency in clean-data settings.

The main contributions of our paper may be summarized as follows:

- Provide a rigorous analysis of the modal interval (Theorems 3.1, 4.1 and 4.2), shorth (Theorems 3.2 and 4.3) and hybrid (Theorems 3.3 and 4.4) estimators for multivariate, radially symmetric distributions. We also show how to relax the symmetry conditions further (Theorem 7.1). These estimation error guarantees hold with high probability.
- Derive upper bounds on the expected error of the estimators (Theorem 5.3). Along the way, we demonstrate the need for additional conditions on the tails of the mixture components in order to derive expected error bounds of the same order as the high-probability results.
- Derive minimax lower bounds on the error rate of any estimator (Theorem 5.4), and prove that the hybrid estimator is nearly optimal in various regimes of interest (Theorem 5.5).

- Extend the methodology for multivariate mean estimation to linear regression (Theorem 8.2).
- Provide computationally efficient versions of the multivariate mean estimator (Theorem 6.1) and linear regression estimator (Theorem 8.3) in high dimensions.

We also note that while our work vastly generalizes the results of Chierichetti *et al.* [8] for mean estimation in Gaussian mixtures, our derivations bypass some critical technical gaps in their proofs using a very different approach via empirical process theory. Finally, we comment that preliminary work on this topic appeared in our earlier conference paper [35] but was limited to the univariate case (Theorems 3.1, 3.2, 3.3 and 4.2) and did not discuss optimality, regression or any computational aspects.[1] Furthermore, all examples and counterexamples illustrating various phenomena, including the detailed theoretical derivations (Propositions 3.7–5.2), are new to this paper.

We end with a few remarks regarding parameter estimation in mixture models. The setting studied in our paper is markedly different from the canonical setting [2, 5, 9, 22, 31], since the number of components in the mixture distribution is allowed to be as large as the number of observations. Furthermore, the parameters of the component mixtures are 'entangled' in the sense that they share a common mean, which we wish to estimate. Notably, this allows us to obtain meaningful error guarantees without imposing strong distributional assumptions such as Gaussianity or log-concavity, which are prevalent in the literature on parameter estimation for mixture models.

The roadmap of the paper is as follows: in Section 2, we define notation and the basic estimators we will consider in the univariate case, which are subsequently analyzed in Section 3. In Section 4, we present results for the multivariate analog of these estimators. In Section 5, we derive expected error bounds on the performance of our estimators and also present minimax lower bounds on the estimation error of any estimator, thus providing settings in which our proposed estimators are provably optimal. In Section 6, we present computationally feasible variants of our estimators in higher dimensions and prove that the error rates of these estimators are of the same order as those derived earlier. In Section 7, we discuss various relaxations of the symmetry assumptions on the mixture components. In Section 8, we describe our results for linear regression. Simulation results reporting the relative performance of different estimators are contained in Section 9. All proofs are contained in the supplementary appendix.

Notation: we regularly use the standard big-O notation: for two real-valued non-negative functions $f(n)$ and $g(n)$, we write $f = O(g)$, when there exists constants n_0 and $C > 0$ such that for all $n \geqslant n_0$, $f(n) \leqslant Cg(n)$. We say $f = \Omega(g)$ if $g = O(f)$, and say $f = \Theta(g)$ when $f = O(g)$ and $g = O(f)$. We write $f = \omega(g)$ if for every real constant $c > 0$, there exists $n_0 \geqslant 1$ such that $f(n) > c \cdot g(n)$ for every integer $n \geqslant n_0$. We write $f = o(g)$, when $g = \omega(f)$. We use $\tilde{O}(\cdot)$, $\tilde{\Omega}(\cdot)$ and $\tilde{\omega}(\cdot)$ to hide polylogarithmic factors. We write w.h.p., or 'with high probability,' to mean with probability tending to 1 as the sample size increases. We use C, c, C' and c' to represent absolute positive constants which may vary from place to place, and their exact values can be found in the proofs. Similarly, we use C_t to represent positive numbers that depend only on t. For a real-valued random variable X, we use $\mathbb{V}X$ to denote its variance.

We will use $\|\cdot\|_2$ to denote the Euclidean norm. We use $B(x, r)$ to denote the Euclidean ball of radius r centered around x, and we also write B_r in place of $B(0, r)$. We denote the $d \times d$ identity matrix by I_d. We use $P(X, \epsilon)$ to denote the ϵ-packing number of a set X with respect to Euclidean distance, and we use $N(X, \epsilon)$ to denote the ϵ-covering number. We write $\mathrm{Diam}(X)$ to denote the diameter of the set with respect to Euclidean distance, i.e., $\mathrm{Diam}(X) := \sup_{x,y \in X} \|x - y\|_2$.

[1] We also mention follow-up papers by Liang and Yuan [30] and Devroye *et al.* [11], which appeared after the initial posting of our conference paper.

2. Problem setup

We begin by introducing the entangled mean estimation problem. Suppose we have n independent samples $X_i \sim P_i$, where each P_i is a distribution in $\mathbb{R}^d$ with a density. Furthermore, we assume that each density p_i is radially symmetric and unimodal with a common mean (and median) μ^*. Our goal is to estimate the location parameter μ^* from the n samples, where the P_is are unknown *a priori* and may even come from different classes of (non-)parametric distributions. Since the estimators we consider are translation invariant, we can assume without loss of generality that $\mu^* = 0$, so the error of an estimator $\widehat{\mu}$ is measured by $\|\widehat{\mu}\|_2$.

A natural estimator to use is the empirical mean, which is certainly an unbiased estimator of μ^*. However, it is a well-known fact that the mean is not 'robust', in the sense that one outlying observation can have a massive impact on the estimation error of the mean. In our setting, one P_i with a very large variance can dramatically inflate the error of the mean, even if the remaining $n-1$ distributions are well behaved. Due to the symmetry assumption on the P_is, we could consider a (multivariate) median as a more robust alternative. Our theory in Section 3 below shows that using a median estimator can somewhat improve the estimation error so that it depends only on the spread of the $\sqrt{n}\log n$ distributions with the smallest quantiles; however, other more cleverly constructed estimators can reduce this dependence to $O(d\log n)$ distributions, meaning that the remaining mixture components may have arbitrarily large (or even infinite) variances, yet have a bounded effect on the behavior of the estimator.

Another potential estimator when the mixing components come from a sufficiently nice parametric family (e.g., Gaussians) is the maximum likelihood estimator. However, since we do not assume knowledge of which observations are drawn from which mixture components, the MLE calculation becomes considerably more complicated. Nonetheless, it is sometimes informative to compare the error rate of the MLE—assuming side information of which observations correspond to which mixture components—to the error rates obtained using various agnostic estimators. In particular, if the former error rate diverges with n, we know that a diverging error rate for a proposed estimator is reasonable.

We will focus on the simple setting where the overall mixture distribution is radially symmetric, e.g., we have multivariate Gaussian observations $X_i \sim \mathcal{N}(0_d, \sigma_i^2 I_d)$. Throughout this paper, we focus on the setting where $d = O(\log n)$; as shown in Chierichetti *et al.* [8], when $d = \Omega(\log n)$, the problem reduces to the case of known variances, since these can be estimated accurately. We shall discuss how to replace the spherical symmetry assumption by log-concavity in Section 7. As the covariance matrix of a radially symmetric distribution is of the form $\sigma^2 I_d$, we denote the covariance matrix of X_i by $\sigma_i^2 I_d$.

We now define the central objects in our analysis:

DEFINITION 2.1 (Order statistics). Let the covariance of X_i be $\sigma_i^2 I_d$. Define $\sigma_{(i)}$ to be the corresponding order statistic. Let s_i denote the interquartile range of X_i, so that $\mathbb{P}(\|X_i - \mu^*\|_2 \leqslant s_i) = \frac{1}{2}$. Define $s_{(i)}$ to be the corresponding order statistic.

DEFINITION 2.2 (Indicator functions on balls). For $x \in \mathbb{R}^d$ and $r \in \mathbb{R}$, let $f_{x,r}(z) := \mathbb{1}_{\|x-z\|_2 \leqslant r}$ denote the indicator function of the ℓ_2-ball $B(x,r)$. For $s \in \mathbb{R}$, we will also use $f_{s,r}(z)$ to denote the indicator of the ball of radius r centered at the vector with first coordinate s and all other coordinates equal to 0.

Note that when $d = 1$, the function $f_{x,r}$ is simply the indicator function of the interval $[x-r, x+r]$.

DEFINITION 2.3 (Function class). Let

$$\mathcal{H}_r := \{f_{x,r'} : x \in \mathbb{R}^d, r' \in \mathbb{R}, 0 \leqslant r' \leqslant r\}.$$

Note that $\mathscr{H}_r$ has VC dimension $d+1$ [48].

As in prior analysis of sample heterogeneous models [37, 39], most of our arguments will be in terms of the mixture distribution $\overline{P} := \frac{1}{n}\sum_{i=1}^n P_i$, which is again unimodal and symmetric. We will write $\overline{P}_n$ to denote the empirical distribution of $X_1, \ldots, X_n$.

DEFINITION 2.4 (Risk). For a function f, we use $R_n(f) := \frac{1}{n}\sum_{i=1}^n f(X_i)$ to denote the expectation of f with respect to the empirical distribution of $X_1, \ldots, X_n$. Let

$$R(f) := \frac{1}{n}\sum_{i=1}^n \mathbb{E} f(X_i).$$

Thus, $R(f)$ is the expectation of f with respect to $\overline{P}$. Define

$$R_r^* := \sup_{f \in \mathscr{H}_r} R(f) = R(f_{0,r}),$$

where the second equality follows by symmetry and unimodality.

Note that $R(f_{0,r})$ also equals the probability of the ball $B(0,r)$ under $\overline{P}$. The spherical symmetry assumption readily gives $R(f_{x,r}) = R(f_{s,r})$ for all x such that $\|x\|_2 = s$.

We first state several useful properties of radially symmetric distributions. The proof of the following result is contained in Appendix A.1.

LEMMA 2.1 Recall Definitions 2.1, 2.2 and 2.4 of $\sigma_{(i)}, s_{(i)}, f_{x,r}$ and R_r^*. Suppose the density of $\overline{P}$ is radially symmetric and unimodal. We have the following properties:

(i) For any $r > 0$ and $x, x' \in \mathbb{R}^d$, if $\|x\|_2 < \|x'\|_2$, then $R(f_{x,r}) \geqslant R(f_{x',r})$.

(ii) For any $x \in \mathbb{R}^d$, if $r < r'$, then $R(f_{x,r}) \leqslant R(f_{x,r'})$.

(iii) If $0 < r_1 < r_2$, then $\frac{R_{r_1}^*}{r_1^d} > \frac{R_{r_2}^*}{r_2^d}$.

(iv) If $0 < r_1 < r_2$, then

$$R(f_{r_2,r_1}) < \frac{1}{P(B_{r_2-r_1}, r_1)} R_{r_2}^* \leqslant \left(\frac{2r_1}{r_2 - r_1}\right)^d R_{r_2}^*,$$

where $P(B_{r_2-r_1}, r_1)$ denotes the packing number of $B_{r_2-r_1}$ with respect to B_{r_1}. In particular, if $r_1 \leqslant \frac{r_2}{2}$, then $R(f_{r_2,r_1}) \leqslant \left(\frac{4r_1}{r_2}\right)^d R_{r_2}^*$.

(v) If $1 \leqslant k \leqslant n/2$, then $\frac{k}{n} < R_{s_{(2k)}}^*$ and $\frac{k}{n} < R_{2\sqrt{d}\sigma_{(2k)}}^*$.

2.1 *Estimators*

We now proceed to define the estimators that will be studied in our paper.

ESTIMATOR 1 (r-modal interval). The r-modal interval estimator, introduced for the (univariate) i.i.d. setting by Chernoff [7], outputs the center of the most populated ball of radius r, with ties broken

arbitrarily:

$$\widehat{\mu}_{M,r} \in \arg\max_x R_n(f_{x,r}). \tag{2.1}$$

ESTIMATOR 2 (k-shortest gap/shorth estimator). For $k \geqslant 2$, the k-shortest gap (k-shorth) estimator, $\widehat{\mu}_{S,k}$, outputs the center of the smallest ball containing at least k points. More precisely, we define

$$\widehat{r}_k := \inf\left\{r : \sup_x R_n(f_{x,r}) \geqslant \frac{k}{n}\right\}, \qquad \widehat{\mu}_{S,k} := \widehat{\mu}_{M,\widehat{r}_k}. \tag{2.2}$$

The traditional (univariate) shorth estimator [4, 23] corresponds to $k = \frac{n}{2}$, whereas choosing $k = 2$ outputs the midpoint of the shortest interval between any two points. As we will see, the choice of $k = C \log n$ will be convenient for our setting and is more suitable than $k = \frac{n}{2}$ if data are not i.i.d.

Note that a type of 'shortest interval' estimator has also been employed in the work on mean estimation for contaminated i.i.d. data [25], but was used as an outlier screening step in that context, rather than a mean estimator. Incidentally, our hybrid estimator to be introduced later will employ a different screening approach based on the median and then use the shorth estimator to return a more accurate mean estimate.

DEFINITION 2.5 Recall Definitions 2.2 and 2.4 for the quantity $R(f_{x,r})$. Define

$$r_k := \inf\left\{r : \sup_x R(f_{x,r}) \geqslant \frac{k}{n}\right\} = \inf\left\{r : R(f_{0,r}) \geqslant \frac{k}{n}\right\},$$

where the second equality follows from unimodality and radial symmetry.

The quantity r_k measures the spread of $\overline{P}$, and $r_{n/2}$ is the interquartile range of $\overline{P}$. Furthermore, since $\overline{P}$ has a density, we have $R^*_{r_k} = \frac{k}{n}$. Note that r_k is problem dependent, since its magnitude depends on the relative dispersion of the mixing components; in particular, we will be interested in $r_{\Theta(d \log n)}$. As the fraction of 'nice' points increases, r_k becomes smaller. However, r_k does not depend too strongly on the high-variance distributions (cf. Lemma 2.1(i) and Proposition 3.7).

The univariate k-median outputs an element from the centermost k points of the data. According to our definition, the k-median outputs a set rather than a point estimator, which will be used as a preprocessing step before applying the modal interval or shorth estimators to obtain a hybrid estimator with superior rates.

ESTIMATOR 3 (k-median). In the univariate setting, the k-median estimator outputs an arbitrary element $\widehat{\mu}_{\text{med},k}$ from the subset S_k, defined as $X_i \in S_k$ if and only if $\widehat{\theta}_{\text{med},-k} \leqslant X_i \leqslant \widehat{\theta}_{\text{med},k}$, where

$$\widehat{\theta}_{\text{med},k} := \inf\left\{\theta : \psi_n(\theta) \geqslant \frac{k}{n}\right\},$$
$$\widehat{\theta}_{\text{med},-k} := \sup\left\{\theta : \psi_n(\theta) \leqslant \frac{-k}{n}\right\},$$

and $\psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n \text{sign}(\theta - X_i)$. The sample median corresponds to taking $k = 0$.

Various multivariate extensions of the median exist, with different robustness properties and computational complexity; for our purposes, it will suffice to consider the simplest version of the multivariate median, which simply operates componentwise on the data points.

DEFINITION 2.6 (Multivariate median). Define the set $S_{k,i}$ as follows: For each dimension i, consider the k median points in that dimension; i.e.,

$$S_{k,i} := \{X_j(i) : X_j(i) \text{ belongs to the } k\text{-median of } (X_j(i))_{j=1}^n\},$$

where $X_j(i)$ denotes the ith coordinate of the vector X_j. Define S_k^∞ to be the cuboid based on $S_{k,i}$, for each dimension i:

$$S_k^\infty := \prod_{i=1}^d [\min(S_{k,i}), \max(S_{k,i})].$$

ESTIMATOR 4 (Hybrid estimator). The hybrid algorithm consists of the following steps, summarized in Algorithm 1:

(i) Compute the cuboid $S_{k_1}^\infty$ with $k_1 = \sqrt{n}\log n$.

(ii) Compute the k_2-shorth estimator $\widehat{\mu}_{S,k_2}$ with $k_2 = Cd\log n$.

(iii) If $\widehat{\mu}_{S,k_2} \notin S_{k_1}^\infty$, return the projection of $\widehat{\mu}_{S,k_2}$ on $S_{k_1}^\infty$. Otherwise, return $\widehat{\mu}_{S,k_2}$.

ALGORITHM 1 Hybrid mean estimator (d-dimensional)

function HYBRIDMULTIDIMENSIONAL($X_{1:n}, k_1, k_2, d$)
  $S_{k_1}^\infty \leftarrow \text{kCuboid}(X_{1:n}, k_1)$.
  $\widehat{\mu}_{S,k_2} \leftarrow \text{Shorth}(X_{1:n}, k_2)$.
  if $\widehat{\mu}_{S,k_2} \in S_{k_1}^\infty$ **then**
    $\widehat{\mu}_{k_1,k_2} \leftarrow \widehat{\mu}_{S,k_2}$
  else
    $\widehat{\mu}_{k_1,k_2} \leftarrow \arg\min_{x \in S_{k_1}^\infty} \|x - \widehat{\mu}_{S,k_2}\|_2$
  end if
  return $\widehat{\mu}_{k_1,k_2}$
end function

Note that the projection in step (iii) is easy to accomplish, since ℓ_2-projection onto the cuboid may be done componentwise, hence computed in $O(d)$ time. Our theoretical results show that replacing the shorth estimator by the modal interval estimator produces similar statistical error rates.

2.2 *Concentration inequality*

The following concentration inequality will be a key technical ingredient for deriving results concerning our estimators. The proof is contained in Appendix A.2.

LEMMA 2.2 Recall the Definitions 2.3 and 2.4 of the terms $R_n(f), R(f), R_r^*$ and $\mathscr{H}_r$. For any fixed $t \in (0,1]$ and $n > 1$, we have

$$\mathbb{P}\left\{\sup_{f \in \mathscr{H}_r} |R_n(f) - R(f)| \geqslant tR_r^*\right\} \leqslant 2\exp\left(-cnR_r^* t^2\right),$$

provided r is large enough so that $nR_r^* \geqslant C_t \frac{d+1}{2} \log n$, where $C_t = \left(\frac{144}{t}\right)^2$ and $c = \frac{1}{200}$.

This theorem is useful because the bounds rely on R_r^*; i.e., they are adaptive to the problem, compared to the traditional $O\left(\frac{1}{\sqrt{n}}\right)$ distribution-independent bound. We also note that Lemma 2.2 requires the mass R_r^* lying around the true mode to be sufficiently large, and while the theorem requires R_r^* to increase with d, we will work in settings where $d = O(\log n)$.

3. Univariate mean estimation

We now state several theoretical guarantees for the aforementioned estimators in the univariate setting. Some of these results appeared in our preliminary work [35], but we provide complete proofs of all statements in Appendix B.

In the univariate setting, we assume that we have n independent samples $X_i \sim P_i$, where each P_i is a univariate distribution with a density p_i which is symmetric and decreasing around μ^*. Let q_i and σ_i denote the interquartile range and standard deviation of P_i, respectively, and recall that the interquartile range satisfies $\mathbb{P}(|X_i - \mu^*| \leqslant q_i) = \frac{1}{2}$. We use $q_{(i)}$ and $\sigma_{(i)}$ to denote the ith smallest interquartile range and standard deviation, respectively (cf. Definition 2.1). By Lemma B.1(v) below, we have $r_k \leqslant q_{(2k)}$ and $r_k \leqslant 2\sigma_{(2k)}$, although these bounds may be loose (for instance, r_k could be finite even if $\sigma_{(1)}$ is infinite). However, we are guaranteed that r_k will be small if $2k$ points come from 'nice' (low-variance) distributions.

THEOREM 3.1 (Pensia *et al.* [35, Theorem 2]). Recall Definitions 2.2 and 2.4 of the terms R_r^*, $R(\cdot)$ and $f_{r',r}$. Let r be a fixed number such that $R_r^* = \Omega\left(\frac{\log n}{n}\right)$. Then with probability at least $1 - 2\exp(-c'nR_r^*)$, the modal interval estimator (Estimator 1) satisfies

$$|\widehat{\mu}_{M,r}| \leqslant r', \tag{3.1}$$

for any r' that satisfying $R(f_{r',r}) < \frac{R_r^*}{2}$. In particular, we can always choose $r' = \frac{2r}{R_r^*}$ to obtain the bound

$$|\widehat{\mu}_{M,r}| \leqslant \frac{2r}{R_r^*}. \tag{3.2}$$

The proof of Theorem 3.1 is contained in Appendix B.1 and proceeds by using Lemma 2.2 to bound the ratio between $R(f_{\widehat{\mu}_{M,r},r})$ and R_r^*, and then using Lemma B.1 to turn this into a deviation bound on

$|\widehat{\mu}_{M,r}|$. Although the bound (3.2) in Theorem 3.1 is simple to state, it may be looser than the bound (3.1).

REMARK 1 Importantly, by Lemma B.1(v), we know that the choice $r = \sigma_{(C \log n)}$ always guarantees the condition $R_r^* = \Omega\left(\frac{\log n}{n}\right)$. Hence, inequality (3.2) implies that

$$|\widehat{\mu}_{M,r}| \leqslant \frac{2\sigma_{(C \log n)}}{R_r^*} \leqslant \frac{2n\sigma_{(C \log n)}}{\log n}, \tag{3.3}$$

with a similar inequality involving $q_{(C' \log n)}$. Note that this bound holds regardless of the magnitude of the standard deviations of the latter $n - C \log n$ mixture components.

At the same time, one might be wary of the fact that the bound in inequality (3.3) could *increase* with n if we fix $\sigma_{(C \log n)}$; for i.i.d. data, $R_r^* = \Theta(1)$, so even the first expression in the bound is of constant order. This is rather alarming, compared to the $O(n^{-1/2})$ error rate of the median. However, it should be noted that if the variances of the mixture components increase sufficiently rapidly with n, even the error rate of the MLE in the Gaussian case (which knows the distribution of each sample) will have a diverging error rate. Thus, although the error bounds of the modal interval estimator in Theorem 3.1 may be rather unsatisfactory in the case of i.i.d. data, they can lead to more meaningful error bounds when the mixture distribution involves a sizable portion of high-variance points. We will explore the question of optimality in more detail in Section 5.2 below.

Guarantees for the shorth estimator are similar to the modal interval estimator. Further, note that as the proofs of the results in this section reveal, the technical machinery we have developed to derive guarantees for the error of the modal interval estimator may also be used to derive estimation error bounds for the shorth estimator.

The proof of the following result is provided in Appendix B.2.

THEOREM 3.2 (Pensia *et al.* [35, Theorem 4]). Recall Definitions 2.1 and 2.5 of the terms $\sigma_{(i)}, q_{(i)}$ and r_k. Suppose $2k \geqslant C_{0.25} \log n$. With probability at least $1 - 2\exp(-c'k)$, the shorth estimator (Estimator 2) satisfies

$$|\widehat{\mu}_{S,k}| \leqslant \frac{2nr_{2k}}{k} < \frac{2n \min\left(q_{(4k)}, 2\sigma_{(4k)}\right)}{k}.$$

REMARK 2 Lemma A.1(iii) shows that the upper bound is actually tighter for small k: for $k' > k$, we have $kr_{2k'} > k'r_{2k}$. The smallest value permissible from our theory would be $k = \Theta(\log n)$. Also note that the upper bound in Theorem 3.2 for the shorth estimator resembles the bound in Theorem 4.2, except for the fact that the bound for the modal interval estimator involves the quantity $r_{C_{0.25} \log n}$ rather than $r_{2C_{0.25} \log n}$, and the latter could be larger depending on the spread of $\overline{P}$. Furthermore, both upper bounds in Theorem 3.2 may sometimes be loose: In particular, if the X_is were i.i.d., r_{2k} would be of order $\Theta\left(\frac{k}{n}\right)$ for small k, so the bound $\frac{nr_{2k}}{k}$ would be of constant order, whereas it is known [23] that the shorth estimator is consistent for $k = 0.5n$.

We now turn to theoretical guarantees from the hybrid estimator, which combines the shorth and k-median in order to obtain superior performance for both fast and slow decay of $\overline{P}$. Recall from Table 1 that the median has superior performance when there is less heterogeneity in the data and $\overline{P}$ decays fast enough. However, the superior performance of the modal interval estimator is apparent in the presence of

large number of high-variance points. It is then desirable to have an estimator that adapts to the problem and enjoys the best of both worlds without any prior information. Indeed, as outlined in Proposition 3.10, the hybrid estimator achieves this rate. The key point is that if the true mean lies inside a convex set (defined with respect to the k-median), then projecting any other point (e.g., the shorth) onto the set will only move the point closer to the mean, so the hybrid estimator can leverage the better of the two rates enjoyed by the median and shorth.

Algorithm 2 specializes the hybrid estimator of Algorithm 1 to the univariate setting. The algorithm proceeds by separately computing the k_1-shorth estimator and k_2-median. If the shorth estimator lies within the median interval, the algorithm outputs the shorth; otherwise, it outputs the closest endpoint of the median interval. Note that this estimator resembles the estimator proposed by Chierichetti *et al.* [8] since it employs the median as a screening step for points with very large variance. However, the shorth estimator is computed separately and then projected onto an interval around the median. In contrast, the estimator proposed by Chierichetti *et al.* [8] first computes the k_2-median and then computes the shorth on the remaining points, leading to a delicate conditioning argument in the analysis and creating some technical gaps in the proofs.

ALGORITHM 2 Hybrid mean estimator

function HYBRIDMEANESTIMATOR($X_{1:n}, k_1, k_2$)
  $S_{k_1} \leftarrow \text{kMedian}(X_{1:n}, k_1)$.
  $\widehat{\mu}_{S,k_2} \leftarrow \text{Shorth}(X_{1:n}, k_2)$.
  if $\widehat{\mu}_{S,k_2} \in [\min(S_{k_1}), \max(S_{k_1})]$ **then**
    $\widehat{\mu}_{k_1,k_2} \leftarrow \widehat{\mu}_{S,k_2}$
  else
    $\widehat{\mu}_{k_1,k_2} \leftarrow \text{closestPoint}(S_{k_1}, \widehat{\mu}_{S,k_2})$
  end if
  return $\widehat{\mu}_{k_1,k_2}$
end function

THEOREM 3.3 (Pensia *et al.* [35,Theorem 5]). Recall the Definition 2.5 and Estimator 1 for the terms r_k, S_k and $\widehat{\mu}_{S,k}$. If $k_1 = \sqrt{n}\log n$ and $k_2 \geqslant C\log n$, the error of the hybrid estimator (Estimator 4) in Algorithm 2 is bounded by

$$|\widehat{\mu}_{k_1,k_2}| \leqslant \min\left(\text{Diam}(S_{k_1}), |\widehat{\mu}_{S,k_2}|\right) \leqslant \frac{4\sqrt{n}\log n}{k_2} r_{2k_2},$$

with probability at least $1 - 2\exp(-c'k_2) - 4\exp(-c\log^2 n)$.

The proof of Theorem 3.3 is provided in Appendix B.3. Importantly, the bound in Theorem 3.3 is finite even for heavy-tailed distributions with infinite variance. Finally, note that in Algorithm 2, we could replace the shorth estimator by the modal interval estimator with adaptively chosen interval width to obtain similar error guarantees.

3.1 *Examples*

We illustrate this below in several cases for $r_{\log n}$, assuming Gaussian distributions for simplicity. The following examples will reappear throughout the paper to illustrate the error of our proposed estimators in various regimes of interest:

EXAMPLE 3.4 (i.i.d. observations). $P_i = \mathcal{N}(0, \sigma^2)$, so $\overline{P}$ is again $\mathcal{N}(0, \sigma^2)$.

EXAMPLE 3.5 (quadratic variance). $P_i = \mathcal{N}(0, c^2 i^2)$, for some small $c > 0$.

EXAMPLE 3.6 (α-mixture distributions).

$$P_i = \begin{cases} \mathcal{N}(0,1), & \text{if } i \leqslant c\lceil \log n \rceil, \\ \mathcal{N}(0, n^{2\alpha}), & \text{otherwise,} \end{cases}$$

for some $\alpha > 0$ and some large $c > 0$.

Example 3.6 is similar to the 'contamination model' in prior work [39, 41], but with a specific scaling of variances to highlight the difference between multiple estimators by varying α. The following proposition, proved in Appendix C.1, will be useful in our development:

PROPOSITION 3.7 We have the following bounds for $r_{\log n}$ when $n = \Omega(1)$:

1. For Example 3.4 (i.i.d. observations), we have $r_{\log n} = \Theta\left(\frac{\sigma \log n}{n}\right)$.
2. For Example 3.5 (quadratic variance) and sufficiently small $c > 0$, we have $r_{\log n} = \Theta(1)$.
3. For Example 3.6 (α-mixture distributions) and sufficiently large $c > 0$, we have

$$r_{\log n} = \begin{cases} \Theta\left(\frac{\log n}{n^{1-\alpha}}\right), & \text{if } \alpha < 1, \\ \Theta(1), & \text{if } \alpha \geqslant 1. \end{cases}$$

Note that these bounds are tighter than the ones provided by Lemma 2.1(v); the latter states that $r_k \leqslant \sigma_{(2k)}$. This is because Lemma 2.1(v) is a worst-case bound which does not account for the contributions of high-variance points.

3.1.1 *Guarantees for individual estimators*

We now revisit the examples above and calculate the bounds that follow from Theorem 3.1 by choosing $r = r_{C \log n}$ for a large constant $C > 0$. We also mention the cases where the bound (3.2) is weaker than the bound (3.1). The proof of the following proposition is contained in Appendix C.2.

PROPOSITION 3.8 Recall Definition 2.5 of r_k. Suppose $r = r_{C \log n}$. We have the following bounds for the modal interval estimator $|\widehat{\mu}_{M,r}|$ (Estimator 1):

1. For Example 3.4 (i.i.d. observations), we have $|\widehat{\mu}_{M,r}| \leqslant \Theta(\sigma)$, w.h.p.
2. For Example 3.5 (quadratic variance), we have $|\widehat{\mu}_{M,r}| \leqslant O(n^{\epsilon})$, w.h.p., for any $\epsilon > 0$. Inequality (3.2) results in a weaker bound of the form $O(n)$, w.h.p.

3. For Example 3.6 (α-mixture distributions), we have

$$|\widehat{\mu}_{M,r}| = \begin{cases} O(n^{\alpha}), & \text{if } \alpha < 1 \\ O(1), & \text{if } \alpha \geqslant 1, \end{cases}$$

w.h.p. For $\alpha \geqslant 1$, inequality (3.2) results in a weaker bound of the form $O(n^{\alpha})$.

REMARK 3 As discussed in Remark 1 above, the guarantees for the modal interval estimator are somewhat unsatisfactory for i.i.d. data, since Proposition 3.8(i) gives an error rate of $\Theta(\sigma)$, rather than the optimal rate $\Theta\left(\frac{\sigma}{\sqrt{n}}\right)$ achievable by the sample mean. On the other hand, Proposition 3.8 shows that for other problem settings with more widely varying variances—such as the α-mixture with $\alpha \geqslant 1$—the modal interval estimator results in constant error, whereas the sample mean would have $\Theta(n^{\alpha-0.5})$ error. These differences are summarized in more detail in Table 1 below.

The modal interval estimator is a 'local' estimator that only considers the value of $\overline{P}_n$ in small windows. As we increase the variance of noisy points, the distribution $\overline{P}$ approaches 0 around μ^*. The modal interval estimator makes mistakes when $\overline{P}$ is flat after normalization, meaning that the density at $x + \mu^*$ is within a $(1 - \epsilon)$-factor of its density at μ^*, for $\epsilon = o(1)$. If this is the case, $\overline{P}_n$ might assign higher mass at $x + \mu^*$ than μ^* due to stochasticity introduced by sampling, so a local method would mistakenly choose $x + \mu^*$ over μ^*.

More concretely, consider the setting of Example 3.6. If an adversary tried to alter the estimator by making the variance of the points very high ($\alpha \gg 1$), then although $\overline{P}$ would approach 0, the normalized density would not be flat. An extreme example of this can be seen when variance of noisy points is '∞': Near μ^*, the distribution $\overline{P}$ would behave like $\mathcal{N}(\mu^*, 1)$ scaled by $O\left(\frac{\log n}{n}\right)$, which is *not* flat after normalization although $\overline{P}$ approaches 0 very rapidly, so that the mean or median would behave poorly. As Proposition 3.8 shows, the modal interval estimator would only suffer $O(1)$ error in this case.

REMARK 4 Examining the bound in Proposition 3.8 for Example 3.6, we see the possible emergence of a 'phase transition' phenomenon: for $\alpha < 1$, the modal interval estimator has error growing with n,

TABLE 1 *The table below summarizes the performance of various estimators on our three running examples. We have ignored poly-logarithmic factors for simplicity, and we use n^{ϵ} to denote $O(n^{\epsilon})$ error for any $\epsilon > 0$ and c to denote an error bounded by a constant. The radius for the modal estimator and the k for the shorth estimator are adjusted to be optimal for each particular example; i.e., the estimators are assumed to know which example data are coming from. Observe that mean and median estimators outperform the modal and shorth estimators when the outliers have relatively small variances. On the other hand, the modal and shorth estimators are better when the outliers have large variances. Simulations in Section 9 show that the rates provided above are indeed observed in practice. Our hybrid estimator achieves the best performance in all cases without knowing which example is under consideration*

	Mean	Median	Modal/shorth	Hybrid
Example 1 (i.i.d. samples)	$\boldsymbol{n^{-0.5}}$	$\boldsymbol{n^{-0.5}}$	$n^{-1/3}$	$\boldsymbol{n^{-0.5}}$
Example 2 (quadratic variances)	$\sqrt{n}$	$\sqrt{n}$	$\boldsymbol{n^{\epsilon}}$	$\boldsymbol{n^{\epsilon}}$
Example 3 ($\alpha < 1$-mixture distributions)	$\boldsymbol{n^{\alpha-0.5}}$	$\boldsymbol{n^{\alpha-0.5}}$	n^{α}	$\boldsymbol{n^{\alpha-0.5}}$
Example 3 ($\alpha \geqslant 1$-mixture distributions)	$n^{\alpha-0.5}$	$n^{\alpha-0.5}$	$\boldsymbol{c}$	$\boldsymbol{c}$

whereas for $\alpha \geqslant 1$, the modal interval estimator only incurs constant error. This suggests that for $\alpha < 1$, high-variance points are more effectively hidden within the mixture distribution, so the accuracy of the modal interval estimator is more severely compromised than in the case when $\alpha \geqslant 1$, where the modal interval estimator can distinguish between low-variance and high-variance points. This phase transition phenomenon is established rigorously in Section 3.1.2 below, where we prove a lower bound of $\Omega(n^{\alpha})$ in the case when $\alpha < 1$.

The performance of the $\Theta(\log n)$-shorth estimator is similar to the modal interval estimator with $r = r_{\Theta(\log n)}$ (cf. inequality (3.3) in Remark 1). Consequently, the error guarantees derived for the running examples in Proposition 3.8 also hold for the $\Theta(\log n)$-shorth.

For completeness, we calculate the bounds of the $(\sqrt{n}\log n)$-median estimator on the recurring examples, proved in Appendix C.3:

Proposition 3.9 We have the following bounds on the $(\sqrt{n}\log n)$-median estimator (Estimator 1):

1. For Example 3.4 (i.i.d. observations), $|\widehat{\mu}_{\text{med},\sqrt{n}\log n}| = O\left(\frac{\sigma \log n}{\sqrt{n}}\right)$, w.h.p.
2. For Example 3.5 (quadratic variance), $|\widehat{\mu}_{\text{med},\sqrt{n}\log n}| = O(n^{0.5}\log n)$, w.h.p.
3. For Example 3.6 (α-mixture distributions), $|\widehat{\mu}_{\text{med},\sqrt{n}\log n}| = O(n^{\alpha-0.5}\log n)$, w.h.p.

The following proposition translates the error guarantees of Theorem 3.3 into our running examples. These bounds are a direct result of Theorem 3.3 and Propositions 3.8 and 3.9.

Proposition 3.10 When k_1 and k_2 are chosen as in Theorem 3.3, we have the following bounds on the hybrid estimator (Estimator 4):

1. For Example 3.4 (i.i.d. observations), $|\widehat{\mu}_{k_1,k_2}| = O\left(\frac{\sigma \log n}{\sqrt{n}}\right)$, w.h.p.
2. For Example 3.5 (quadratic variance), $|\widehat{\mu}_{k_1,k_2}| = O(n^{\epsilon})$, w.h.p., for any $\epsilon > 0$.
3. For Example 3.6 (α-mixture distributions), with high probability,

$$|\widehat{\mu}_{k_1,k_2}| = \begin{cases} O(n^{\alpha-0.5}), & \text{if } \alpha < 1, \\ O(1), & \text{if } \alpha \geqslant 1. \end{cases}$$

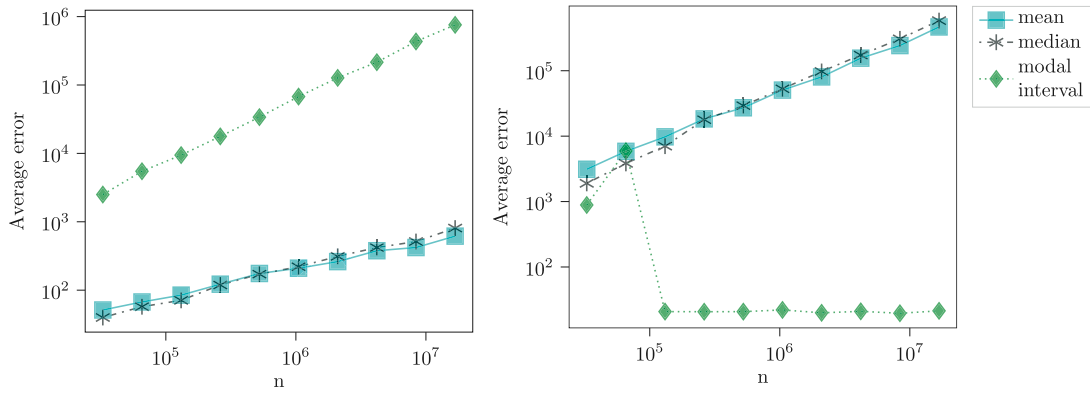
3.1.2 *Phase transition behavior* In this subsection, we focus on verifying the statement in Remark 4 above, namely the existence of a phase transition for the modal interval estimator depending on whether $\alpha < 1$ or $\alpha \geqslant 1$. This phenomenon is illustrated via simulations in the plots of Fig. 1.

For ease of analysis, we tweak the setting of Example 3.6 slightly: instead of having different distributions for high variance and low variance points, we assume that the points are sampled i.i.d. from a mixture distribution, with weights resembling their original fraction in Example 3.6. Moreover, we assume that individual distributions are uniform rather than Gaussian.

Example 3.11 (Modified α-mixture distributions). Let $c > 0$ be a large enough constant. For each i, $P_i = Q_n$, where

$$Q_n = \frac{c\log n}{n} U[-1,1] + \frac{n - c\log n}{n} U[-n^{\alpha}, n^{\alpha}],$$

and $U[-a,a]$ is the uniform distribution on $[-a,a]$.



(a) $\alpha = 0.9$ (b) $\alpha = 1.3$

FIG. 1. Plots comparing average error of the mean, median and modal interval estimators on Example 3.6 (α-mixture distributions) for different values of α. As shown in Proposition 3.8, the modal interval estimator undergoes a phase transition at $\alpha = 1$, where the error of modal interval estimator drops from the increasing function $\Omega(n^{\alpha})$ to the constant function $\Theta(1)$. Moreover, as shown in Proposition 3.9, the median has better performance than the modal interval estimator for $\alpha < 1$, motivating our hybrid estimator. The average error, $\frac{1}{T}\sum_{i=1}^{T} |\widehat{\mu} - \mu^*|$, is calculated using $T = 200$ runs for each n. Both of the axes are on the log scale. More details can be found in Section 9.

Note that if we sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Q_n$, the number of points with variance $\Theta(1)$ is $\Theta(\log n)$, w.h.p. It is easy to see that the upper bounds for Example 3.11 are the same as that of Example 3.6 in Proposition 3.8, i.e.,

$$|\widehat{\mu}_{M,r_{C\log n}}| = \begin{cases} O(n^{\alpha}), & \text{if } \alpha < 1 \\ O(1), & \text{if } \alpha \geqslant 1, \end{cases}$$

w.h.p. The following proposition, proved in Appendix C.4, establishes a lower bound of $\Omega(n^{\alpha})$ on the error:

PROPOSITION 3.12 For $\frac{1}{3} \leqslant \alpha < 1$ in Example 3.11, the 1-modal interval estimator (Estimator 1) incurs $\Omega(n^{\alpha})$ error, with a constant non-zero probability.

Proposition 3.12 proves rigorously that the apparent phase transition of the modal interval estimator is not simply an artifact of the argument used to prove Proposition 3.8. Indeed, the modal interval estimator experiences a sharp phase transition depending on the relative variance of the mixture component with the higher variance, which is governed by the parameter α. Moreover, this phase transition is not specific to just modal interval estimator. As stated in Theorem 5.4, all agnostic estimators must have error $\Omega(n^{\alpha-0.5})$ for $\alpha < 1$. Thus, Example 3.6 is indeed a difficult problem for $\alpha < 1$, but a surprisingly easy one for $\alpha > 1$.

As a final remark, note that in Examples 3.6 and 3.11, the sample median and even the mean would have an error of $\tilde{O}(n^{\alpha-0.5})$. When $\alpha < 1$, this rate is much better than the $O(n^{\alpha})$ guarantee of the modal interval estimator. This motivates the hybrid estimator proposed above, which is able to combine the 'best of both worlds' for the modal interval and median estimators.

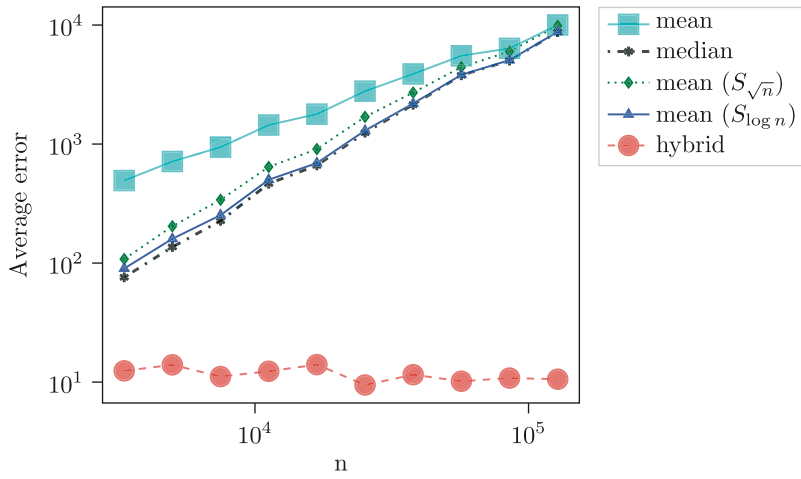


FIG. 2. Plot comparing the average error of various estimators on Example 3.6 with $\alpha = 1.3$. Both of the axes are on the log scale to show the rate. As mentioned in Table 1, both the mean and median have an $n^{\alpha-0.5}$ error rate. The error rates of the α-trimmed mean, with $\alpha = \frac{1}{2} - \frac{\sqrt{n}}{n}$ and $\alpha = \frac{1}{2} - \frac{\log n}{n}$, are similar to the median. Note that the hybrid estimator has far superior performance. More simulations and details are available in Section 9.

3.2 *Comparison to common estimators*

We briefly mention some common univariate estimators and contrast their performance with the performance guarantees of our proposed estimators. For simplicity, we focus on mixtures of univariate Gaussian distributions in which $\Theta(\log n)$ of the samples are drawn from distributions with bounded variance. The primary reason why the estimators mentioned below have suboptimal guarantees is because they are designed to guard against a constant fraction of arbitrarily corrupted or heavy-tailed points. In such cases, the sample median is the optimal estimator; in contrast, the sample median can be shown to be suboptimal in our setting (see Table 1 or Figure 2).

1. *Sample median:* Hallin and Mizera [17] established necessary and sufficient conditions for the consistency of the median for sample heterogeneous distributions. Although the sample median is more robust than the sample mean, this result shows that sample median is consistent if and only if $R^*_\epsilon = \omega\left(\frac{1}{\sqrt{n}}\right)$ for every $\epsilon > 0$. In particular, it implies that if the median is consistent, then $r_{\sqrt{n}} \to 0$. Focusing on particular Example 3.6, the error rate of the median is $O(n^{\alpha-0.5})$ (cf. Table 1 and Fig. 2).
   Moreover, Hallin and Mizera [17] established the optimality of the median among all M-estimators with score functions $\psi(\cdot)$ satisfying the following conditions:
   a. $\psi(\cdot)$ is non-decreasing and skew-symmetric.
   b. $\psi(\infty) = 1$.
   c. The set of discontinuity points of $\psi(\cdot)$ is finite.
   Therefore, one must consider broader classes of estimators beyond this family of M-estimators in order to obtain better error guarantees than the median.
2. *Huber's M-estimator:* for any finite truncation parameter, Huber's M-estimator [19, 20] falls in the class of M-estimators considered by Hallin and Mizera [17], since normalizing the score

function of a bounded score function does not change the final estimate. Thus, the error rate of Huber's M-estimator cannot be any better than the error rate of the median.

3. *k-median of means:* this estimator [33] divides the n data points into k disjoint blocks $(B_1, \ldots, B_k)$ of equal size (assuming n/k is an integer). For each $i \in \{1, \ldots, k\}$, we define Z_i to be the mean of the samples in B_i, and then define the estimator

$$\widehat{\mu} := \text{Median}_{1 \leqslant i \leqslant k}(Z_i) = \text{Median}_{1 \leqslant i \leqslant k}\left(\frac{X_{(i-1)\frac{n}{k}+1} + \cdots + X_{i\frac{n}{k}}}{n/k}\right).$$

The median of means is robust to a constant fraction of outliers and sub-Gaussian tails even for heavy-tailed i.i.d. distributions [26, 33]; however, we argue that the median of means estimator is also not robust to substantial sample heterogeneity. If each $X_i \sim \mathcal{N}(\mu, \sigma_i^2)$, then $Z_i \sim \mathcal{N}\left(\mu, \frac{\sum_{l=(i-1)k+1}^{ik} \sigma_l^2}{n^2/k^2}\right)$. Therefore, the final estimator behaves like the median of independent Gaussian samples. Furthermore, a single 'high-variance' point in the set B_i can increase the variance of Z_i arbitrarily, hiding the signal from 'low-variance' samples. The best case would thus be when each block contains either all 'low-variance' samples or all 'high-variance' samples. However, in that case, $\widehat{\mu}$ behaves essentially like the median of a smaller set with rescaled standard deviations. As argued above, regimes exist where the median estimator is suboptimal.

4. *α-trimmed mean:* let $\alpha \in [0, 0.5)$ be such that αn is an integer. Given samples $\{X_1, \ldots, X_n\}$, the α-trimmed mean [19, 33] discards the largest and smallest αn samples and returns the mean of the remaining $(1 - 2\alpha)n$ samples:

$$\widehat{\mu}_\alpha = \frac{1}{(1-2\alpha)n} \sum_{i=\alpha n+1}^{n-\alpha n} X_{(i)}.$$

The trimmed means estimator is robust to a constant fraction of outliers and has sub-Gaussian tails even for heavy-tailed distributions [33]. As the fraction of 'low-variance' points can be as small as $\frac{\log n}{n}$ in our sample-heterogeneous setting, the estimator $\widehat{\mu}_\alpha$ would have a large variance for any constant $\alpha > 0$. Thus, our choice of α should depend on n, going to 0.5 as $n \to \infty$.
Recall that in the definition of the k-median (cf. Estimator 3), S_k was defined as the k centermost points of the data. Thus, $\widehat{\mu}_\alpha$ is the mean of the set S_k with $k = n(1 - 2\alpha)$. In the extreme case of $\alpha = 0.5 - \frac{1}{2n}$, the trimmed means estimator $\widehat{\mu}_\alpha$ is the same as the median, which is not optimal. As Fig. 2 shows, the trimmed mean behaves like the median for large α, and decreasing α (i.e., increasing k) degrades the performance. Note that we bound the error of the k-median by bounding the range of S_k (cf. Lemma B.4). Therefore, the bounds for the k-median also imply bounds for $\widehat{\mu}_{\frac{n-k}{2n}}$. However, the k-median primarily allows us to define a hybrid estimator by projecting onto the set S_k, which performs better than the k-median alone.

4. Multivariate case

In the following sections, we derive the main results of our paper, which generalize the theorems in Section 3 to d dimensions.

4.1 *Modal interval estimator*

The following result provides an error bound for the modal interval estimator. The proof is in Appendix D.1.

THEOREM 4.1 Recall Definitions 2.1, 2.4 and Estimator 1. Suppose $R_r^* \geqslant C_{0.5}\left(\frac{(d+1)\log n}{n}\right)$. The multidimensional modal interval estimator satisfies the error bounds

$$\|\widehat{\mu}_{M,r}\|_2 \leqslant 4r\left(\frac{2}{R_r^*}\right)^{\frac{1}{d}}, \tag{4.1}$$

$$\|\widehat{\mu}_{M,r}\|_2 \leqslant 8\sqrt{d}\sigma_{(2Cd\log n)}\left(\frac{2}{R_r^*}\right)^{\frac{1}{d}} \leqslant 8\sqrt{d}\left(\frac{n}{C'd\log n}\right)^{\frac{1}{d}}\sigma_{(2Cd\log n)}, \tag{4.2}$$

with probability at least $1-2\exp(-c'd\log n)$.

As the proof of Theorem 4.1 reveals, inequality (4.2) could also be stated using $s_{(2k)}$ in place of $2\sqrt{d}\sigma_{(2k)}$, since it is obtained from inequality (4.1) simply by substituting the bounds of Lemma 2.1(v). Note that when $d=1$, the bound (4.1) in Theorem 4.1 reduces to the bound (3.2) in Theorem 3.1, up to constant factors.

REMARK 5 Our bound (4.2) may be compared with Theorem 5.1 in Chierichetti *et al.* [8]: note that we have removed a factor of polylog(n), although their bound depends on $\sigma_{(\log n)}$ rather than $\sigma_{(d\log n)}$. Nonetheless, we emphasize the fact that our results hold for general radially symmetric distributions, whereas the proofs in Chierichetti *et al.* [8] are Gaussian specific.

Note that by Lemma 2.1(iii), the bound in Theorem 4.1 is tighter for smaller values of r. Thus, the choice of r which optimizes the bound satisfies $R_r^* = C\left(\frac{d\log n}{n}\right)$. As discussed in Pensia *et al.* [35] for the univariate setting, an estimator with near-optimal performance may be obtained via Lepski's method [28] even without knowledge of $\overline{P}$: define r^* to be the interval radius satisfying $R_{r^*}^* = C_{0.5}\left(\frac{(d+1)\log n}{n}\right)$, and suppose we have rough initial estimates $r_{\min}$ and $r_{\max}$ such that $r_{\min} \leqslant r^* \leqslant r_{\max}$. Define $r_j := r_{\min}2^j$, and define

$$\mathscr{J} := \left\{j \geqslant 1 : r_{\min} \leqslant r_j < 2r_{\max}\right\}.$$

We then define the index j_* to be

$$\min\left\{j \in \mathscr{J} : \forall i > j \text{ s.t.} i \in \mathscr{J}, \|\widehat{\mu}_{M,r_i} - \widehat{\mu}_{M,r_j}\|_2 \leqslant 8r_i\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}\right\},$$

which may be calculated using pairwise comparisons of the modal interval estimator computed over the gridding of $[r_{\min}, r_{\max}]$. We then have the following result, proved in Appendix D.2:

THEOREM 4.2 Recall the definition of Estimator 1. With probability at least $1 - 2\left(1 + \log_2\left(\frac{2r_{\max}}{r_{\min}}\right)\right)\exp(-c' \log n)$, we have $j_* < \infty$ and

$$\|\widehat{\mu}_{M,r_{j_*}}\|_2 \leqslant 24 r^* \left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}. \tag{4.3}$$

Note that the cost of using Lepski's method is a factor of 6 in the estimation error. Finally, the following lemma shows that the shorth estimator can be used to obtain rough initial bounds on r^*:

LEMMA 4.1 Recall Definition 2.5 of r_k. For $k \geqslant C_{0.5} d \log n$, with probability at least $1 - 2\exp(-ck)$, we have $r_{k/2} \leqslant \widehat{r}_k \leqslant r_{2k}$.

The proof of Lemma 4.1 is in Appendix D.3 and uses Lemma 2.2 to control the fluctuations of $\widehat{r}_k$ from its empirical counterpart. In particular, the lemma shows that we may use $r_{\min} = \widehat{r}_{C_{0.5}(d+1)\log n/2}$ and $r_{\max} = \widehat{r}_{C_{0.5}(d+1)\log n}$.

4.2 *Shorth estimator*

We now derive error bounds for the multidimensional shorth estimator. The proof is contained in Appendix D.4.

THEOREM 4.3 Recall Definition 2.5 of r_k. Suppose $k \geqslant C(d+1)\log n$. The multidimensional shorth estimator (Estimator 2) satisfies the error bound

$$\|\widehat{\mu}_{S,k}\|_2 \leqslant 4 r_{2k} \left(\frac{2n}{k}\right)^{1/d},$$

with probability at least $1 - 2\exp(-c' d \log n)$.

As in the univariate case, the estimation error guarantees for the multidimensional modal interval and shorth estimators are similar. In particular, for the 'optimal' choice of r such that $R_r^* = \frac{cd\log n}{n}$, inequality (4.1) in Theorem 4.1 gives the bound $\|\widehat{\mu}_{M,r}\|_2 = O\left(r_{c'd\log n}\left(\frac{n}{Cd\log n}\right)^{1/d}\right)$, which is of the same form as the guarantee from Theorem 4.3 when $k = C(d+1)\log n$.

4.3 *Hybrid estimator*

We now prove that the hybrid estimator produces an estimator with rates of $O(\sqrt{n}^{1/d})$, rather than the rate $O(n^{1/d})$ obtained in Theorems 4.1 and 4.3. Since the overall mixture distribution is radially symmetric, all the marginal distributions are identical and symmetric about 0. Accordingly, we denote the common marginal distribution by $\overline{P}_1$, and define $r_{k,1}$ to be the smallest interval (centered at 0) that contains $\frac{k}{n}$ mass under $\overline{P}_1$.

We then have the following result, proved in Appendix D.5:

THEOREM 4.4 Recall Definition 2.6, Estimator 1 and Definition 2.5 of the terms S_k^∞, $\widehat{\mu}_{S,k}$ and r_k. Suppose $k_1 = \sqrt{n}\log n$ and $k_2 \geqslant Cd\log n$. Then the error of the hybrid algorithm (Estimator 4) is bounded by

$$\|\widehat{\mu}_{k_1,k_2}\|_2 \leqslant \min\left\{\mathrm{Diam}(S_{k_1}^\infty), \|\widehat{\mu}_{S,k_2}\|_2\right\} \leqslant C' \min\left\{\sqrt{d} r_{2k_1,1}, \sqrt{n}^{1/d} r_{k_2}\right\},$$

with probability at least $1 - 2\exp(-c'k_2) - 4d\exp(-c\log^2 n)$.

REMARK 6 Similar to the univariate case, the multivariate hybrid estimator achieves good error guarantees for both slow and fast decay of $\overline{P}$. In particular, when data are i.i.d. Gaussian with distribution $\mathcal{N}(0, \sigma^2 I_d)$, as in Example 3.4, the error of the hybrid estimator is of the order $O\left(\frac{\sigma\sqrt{d}\log n}{\sqrt{n}}\right)$. This is within log factors of the optimal $\frac{\sqrt{d}\sigma}{\sqrt{n}}$ error rate. At the same time, the worst-case error guarantee is of the form $O\left(\sqrt{d}\sqrt{n}^{1/d}\sigma_{(Cd\log n)}\right)$.

We also briefly comment on the error guarantees of the hybrid estimator on the multivariate analog of Example 3.6. We can show that $r_{2k_1,1} = \tilde{O}\left(n^{\alpha-0.5}\right)$ and $r_{k_2} = \tilde{O}\left(\sqrt{d}n^{\alpha-\frac{1}{d}}\right)$, so Lemma D.1 implies a bound of $\tilde{O}(\sqrt{d}n^{\alpha-0.5})$ for the median estimator. On the other hand, Theorem 4.3 leads to a bound of $\tilde{O}(\sqrt{d}n^{\alpha})$ for the shorth estimator. This bound can be improved for $\alpha \geqslant \frac{1}{d}$: if $\alpha \geqslant \frac{1}{d}$, we have $\|\widehat{\mu}_{S,k_2}\|_2 = O(\sqrt{d})$ (cf. Theorem 5.5). The second expression in Theorem 4.4 then implies that the error of the hybrid estimator is $\tilde{O}\left(\sqrt{d}\min(n^{\alpha-0.5}, 1)\right)$ for $\alpha \geqslant \frac{1}{d}$ and $\tilde{O}(\sqrt{d}n^{\alpha-0.5})$ for $\alpha \leqslant \frac{1}{d}$. This improves upon the error rates of both the median and shorth estimators.

5. Bounds in expectation

Thus far, we have focused on high-probability bounds. We now briefly discuss how to convert the upper bounds into bounds on the expected error of the estimator. We then derive lower bounds on the estimation error of any estimator, thus addressing the question of optimality in certain regimes.

5.1 *Imposing additional assumptions*

We first show that unlike high-probability bounds, expected error bounds of a similar order *cannot* be derived for modal interval estimator without any assumptions on the high-variance mixture components. To illustrate this point, we provide a univariate example in which it is possible to derive high-probability bounds of $O(1)$ for the modal interval estimator without further assumptions, whereas bounds in expectation of a similar order provably require additional tail assumptions, since $\mathbb{E}\,|\widehat{\mu}_{M,1}| \to \infty$ as $q_n \to \infty$.

EXAMPLE 5.1 For any n, let the densities of the P_is be defined as follows: for $i \leqslant C\log n$, let

$$p_i(x) = \begin{cases} \frac{1}{6i}, & |x| \leqslant 3i, \\ 0, & \text{otherwise.} \end{cases}$$

For $i > C\log n$ and $\alpha \in (0,1)$, let

$$p_i(x) = \begin{cases} n^{-\alpha}, & |x| \leqslant 1, \\ h_n, & 1 < |x| \leqslant q_n, \\ 0, & \text{otherwise,} \end{cases}$$

where the $\{h_n\}$ and $\{q_n\}$ are constrained such that the total area is 1, i.e., $2n^{-\alpha} + 2(q_n - 1)h_n = 1$ and $h_n \leqslant \frac{n^{-\alpha}}{2}$. In particular, for an $\alpha > 0$, we can still choose q_n arbitrarily large; we will take $q_n = \Omega(n)$.

The proof of the following statement is contained in Appendix E.1:

PROPOSITION 5.2 For Example 5.1, we have $\mathbb{E}\,|\widehat{\mu}_{M,1}| \to \infty$ as $q_n \to \infty$. Moreover, $|\widehat{\mu}_{M,1}| = O(1)$, w.h.p.

As seen by the example above, additional assumptions need to be imposed to prove the bounds in expectation. Suppose the variances $\{\sigma_i\}$ are all finite. We will consider two types of assumptions: either (i) 'high-noise' points do not have very large variances, or (ii) 'low-noise' points have small support.

We state a result for the modal interval estimator in d dimensions; similar proofs hold for the shorth, median and hybrid estimators. The following result is proved in Appendix E.2.

THEOREM 5.3 Recall Definitions 2.1, 2.4 and Estimator 1 for the terms $R_r^*, \sigma_{(i)}$ and $\widehat{\mu}_{M,r}$. Let $nR_r^* = \Omega\,(d \log n)$. The following upper bounds hold for the expected error of the modal interval estimator:

(i) Suppose

$$\log\left(\frac{\sigma_{(n)}}{r}\right) = O\left(nR_r^*\right). \tag{5.1}$$

Then the modal interval estimator satisfies the expected error bound

$$\mathbb{E}\,\|\widehat{\mu}_{M,r}\|_2 = O\left(r\left(\frac{c}{R_r^*}\right)^{1/d}\right).$$

(ii) In the case $d = 1$, suppose the support of $\Omega\,(nR_r^*)$ points lies in $[-r, r]$. Then

$$\mathbb{E}\,|\widehat{\mu}_{M,r}| = O\left(\frac{r}{R_r^*}\right).$$

REMARK 7 The condition (5.1) in Theorem 5.3(i) can be translated into the inequality $\sigma_{(n)} \leqslant r\exp(CnR_r^*)$ and provides an upper bound on the variance of the worst mixture components. If we choose $r = \sigma_{(d\log n)}$, we obtain the requirement that $\sigma_{(n)}$ is at most a factor of $O(n^{Cd})$ larger than the variance $\sigma_{(d\log n)}$ of the 'good' points. This can be compared to the assumption $\sigma_{(n)} = \sigma_{(1)}\text{poly}(n)$ imposed by Chierichetti *et al.* [8] when proving upper bounds on expected error in the univariate case. As the proof of Theorem 5.3 reveals, we could also convert the tighter version of the estimation error guarantee (cf. Theorem 3.1 in the univariate setting) into an expected error bound in a similar manner: if condition (5.1) holds in Theorem 5.3 and we additionally assume that $r' = \Omega\,(r)$, then $\mathbb{E}\,|\widehat{\mu}_{M,r}| = O(r')$.

Note that the condition in Theorem 5.3(ii) imposes no constraints on the behavior of the large-variance mixture components. The proof proceeds by integrating the tail probability of the modal interval estimator and showing that it must decay sufficiently quickly by considering the mass of intervals lying far from the true mean. An extension to the multivariate case is possible but would require somewhat more refined technical analysis.

5.2 *Minimax bounds*

We are now ready to discuss the optimality of our hybrid estimator, which we will consider in the context of expected error bounds. We state our results in the case of a general dimension $d \geqslant 1$. The goal of this section is to describe a general setting in which it is possible to show that the hybrid estimator is (nearly) minimax optimal.

We will consider the class of distributions $\mathscr{P}(\sigma_1, \sigma_2, p)$, containing symmetric, unimodal distributions $\{P_i\}_{i=1}^n$ with common mean μ, such that at least np distributions have marginal variance bounded by σ_2^2 and the remaining distributions have marginal variance bounded by σ_1^2. Note that σ_1, σ_2 and p may all be functions of n, e.g., $p = \frac{\log n}{n}$.

We call an algorithm agnostic if applying the algorithm does not require knowledge of the variance of individual points (e.g., the sample mean or median). We have the following minimax lower bound, proved in Appendix E.3:

THEOREM 5.4 Suppose $p \leqslant \frac{1}{3}$, $\sigma_2 \leqslant \sigma_1$ and $p = \Omega\left(\frac{\log n}{n}\right)$.

(i) The minimax error of any agnostic algorithm is

$$\min_{\widehat{\mu}} \max_{\{P_i\} \subseteq \mathscr{P}(\sigma_1, \sigma_2, p)} \mathbb{E}\left[\|\widehat{\mu} - \mu\|_2\right] \geqslant C_\ell \sqrt{d} \min\left\{\frac{\sigma_2}{\sqrt{np}}, \frac{\sigma_1}{\sqrt{n}}\right\}. \tag{5.2}$$

(ii) In the case $d = 1$, suppose in addition we have

$$\frac{\sigma_1}{\sigma_2} = O\left(\frac{1}{np^2}\right). \tag{5.3}$$

Then the algorithm of any agnostic algorithm satisfies that

$$\min_{\widehat{\mu}} \max_{\{P_i\} \subseteq \mathscr{P}(\sigma_1, \sigma_2, p)} \mathbb{E}\left[\|\widehat{\mu} - \mu\|_2\right] \geqslant \frac{C'_\ell \sigma_1}{\sqrt{n}}. \tag{5.4}$$

REMARK 8 In the $d = 1$ case, the lower bound in Theorem 5.4 when condition (5.3) is satisfied matches the lower bound derived by Chierichetti *et al.* [8]. On the other hand, our proof technique is somewhat more direct and proceeds via a straightforward (albeit lengthy) calculation.

We now state our general upper bound, achieved by the hybrid estimator. Under the specific regimes, we impose mild regularity conditions on the distributions to obtain cleaner expressions:

(i) Let $q_i(x)$ denote the marginal distribution of P_i, where $q_i : \mathbb{R} \to \mathbb{R}$ (since P_i is radially symmetric, all marginals are equal). Let v_i^2 denote the marginal variance of P_i. Then

$$q_i(v_i) \geqslant \frac{c}{v_i}. \tag{5.5}$$

(ii) Let each density be written as $p_i(x) = f_i(\|x\|_2)$, where $f_i : \mathbb{R} \to \mathbb{R}$ is a decreasing function on the positive reals. Then

$$f_i(0) \leqslant \left(\frac{c'}{v_i}\right)^d, \quad \text{and} \quad \int_{B(K\sqrt{d}v_i, 2\sqrt{d}v_i)} p_i(y)\,\mathrm{d}y \leqslant C_1 \exp\left(-C_2K^2\right), \quad \forall K \geqslant C_3 > 1. \tag{5.6}$$

Condition (5.5) assumes that the marginal densities do not decrease too rapidly around the mean and implies the accuracy of the median filtering step. Condition (5.6) assumes that the joint densities do not have too much mass concentrated around any single point (e.g., the mean), from which we may derive tighter error bounds on the shorth estimator when we have sufficiently separated variances, i.e., $\frac{\sigma_1}{\sigma_2} = \Omega\left(n^{1/d}\right)$. Note that conditions (i) and (ii) hold for Gaussian distributions; furthermore, condition (ii) holds more broadly when the norm of $p_i(\cdot)$ has right $c'v_i\sqrt{d}$-sub-Gaussian tails around $\sqrt{d}v_i$. Then this expression can be upper bounded by $\mathbb{P}\{\|X\| - \sqrt{d}v_i \geqslant cK\sqrt{d}v_i\} \leqslant \exp(-c'K^2)$ using the sub-Gaussian assumption.

We also define $\mathscr{Q}(\sigma_1, \sigma_2, p)$ to be the class of symmetric, unimodal distributions with $\{P_i\}_{i=1}^n$ with common mean μ, such that at least np distributions have marginal variances bounded by σ_2^2 and remaining distributions have marginal variance at least $\Omega(\sigma_1^2)$ and at most σ_1^2. Thus, $\mathscr{Q}(\sigma_1, \sigma_2, p)$ is the class of distributions with sufficient division between high-variance and low-variance points, and we clearly have $\mathscr{Q}(\sigma_1, \sigma_2, p) \subseteq \mathscr{P}(\sigma_1, \sigma_2, p)$. Finally, in order to derive bounds in expectation, we impose the additional growth condition (5.1) on the variance of the mixture components.

The following result is proved in Appendix E.4.

THEOREM 5.5 If $p = \Omega\left(\frac{d\log n}{n}\right)$ and condition (5.5) holds, then the hybrid estimator satisfies the upper bound

$$\max_{\{P_i\} \subseteq \mathscr{P}(\sigma_1, \sigma_2, p)} \mathbb{E}\left[\|\widehat{\mu} - \mu\|_2\right] \leqslant C_u\sqrt{d}\min\left\{\sqrt{n}^{1/d}\sigma_2, \frac{\log n}{\sqrt{n}}\sigma_1\right\}. \tag{5.7}$$

We also have the following special cases if we impose additional assumptions:

(a) If $p = \Omega\left(\frac{\sqrt{n}\log n}{n}\right)$, we have the tighter bound

$$\max_{\{P_i\} \subseteq \mathscr{P}(\sigma_1, \sigma_2, p)} \mathbb{E}\left[\|\widehat{\mu} - \mu\|_2\right] \leqslant C'_u\sqrt{d}\min\left\{\frac{\log n}{p\sqrt{n}}\sigma_2, \frac{\log n}{\sqrt{n}}\sigma_1\right\}. \tag{5.8}$$

(b) If $\frac{\sigma_1}{\sigma_2} = \Omega\left(n^{\frac{1}{d}}\right)$ and condition (5.6) holds, then

$$\max_{\{P_i\} \subseteq \mathscr{Q}(\sigma_1, \sigma_2, p)} \mathbb{E}\left[\|\widehat{\mu} - \mu\|_2\right] \leqslant C''_u\sqrt{d}\min\left\{\sigma_2\sqrt{\log n}, \frac{\log n}{\sqrt{n}}\sigma_1\right\}. \tag{5.9}$$

It is instructive to compare the upper bounds for the hybrid estimator in Theorem 5.5 with the lower bounds derived in Theorem 5.4. (Note that the same class of distributions used to obtain the minimax lower bounds over $\mathscr{P}$ falls into the class $\mathscr{Q}$, so the upper bounds in Theorem 5.5 may be directly

TABLE 2 *Comparison of upper and lower bounds for estimation error, given by Theorems 5.4 and 5.5, in three regimes of interest. In all of these cases, we assume* $p = \Omega\left(\frac{d\log n}{n}\right)$. *For simplicity, we set* $\sigma_2 = 1$ *and ignore multiplicative factors which are logarithmic in n. We provide more details regarding these calculations in Appendix E.5*

	Large heterogeneity $\mathcal{Q}\left(\Omega\left(\frac{1}{\sqrt{p}} + n^{1/d}\right), 1, o(n^{-0.5})\right)$	Mild heterogeneity $\mathcal{P}\left(O\left(\frac{1}{\sqrt{p}}\right)\right), 1, p$	Large p $\mathcal{P}\left(\sigma_1, 1, \Omega\left(n^{-0.5}\log n\right)\right)$
Hybrid estimator	$\sqrt{d}$	$\frac{\sigma_1\sqrt{d}}{\sqrt{n}}$	$\sqrt{d}\min\left\{\frac{1}{p\sqrt{n}}, \frac{\sigma_1}{\sqrt{n}}\right\}$
Lower bound	$\frac{\sqrt{d}}{\sqrt{np}}$	$\frac{\sigma_1\sqrt{d}}{\sqrt{n}}$	$\sqrt{d}\min\left\{\frac{1}{\sqrt{pn}}, \frac{\sigma_1}{\sqrt{n}}\right\}$

compared with the lower bounds in inequality (5.9), as well.) In particular, we can see that the hybrid estimator is nearly minimax optimal in three somewhat different regimes of interest, which can be derived directly from the bounds in the theorems. The results are summarized in Table 2:

1. Large heterogeneity: when σ_1 is very large compared to σ_2 and p is very small (still satisfying $p = \Omega(\frac{d\log n}{n})$), a direct application of the median would lead to large error. However, the shorth estimator is able to focus on the low-variance points due to the sufficiently large separation in variances. As p becomes smaller, the gap between the upper and lower bounds reduces, reaching within $\log n$ factors when $p = \Theta\left(\frac{d\log n}{n}\right)$.
2. Mild heterogeneity: since σ_1 is relatively small, the median and even mean are minimax optimal. The hybrid estimator is able to achieve these rates (including the i.i.d. case).
3. Large p: as p increases, the number of good points increase and we expect to obtain vanishing error for reasonable values of σ_1 (e.g., under condition (5.1)). Indeed, the hybrid estimator achieves vanishing error for large $p = \Omega\left(\frac{\sqrt{n}\log n}{n}\right)$ irrespective of the magnitude of σ_1. Also, the gap between the upper bound and the lower bound decreases as either $p \to 1$ or $\sigma_1 \to \sigma_2$.

REMARK 9 Although we have shown that the hybrid estimator is indeed optimal in several diverse regimes, the preceding discussion leaves open the question of optimality in other settings. In particular, although our general upper bounds (e.g., inequality (5.7)) suggests the presence of a $\sqrt{n}^{1/d}$ factor when using the hybrid estimator, our lower bound techniques do not show that such a factor is unavoidable for $d \geqslant 2$. As argued by Chierichetti *et al.* [8], a factor of $\sqrt{n}$ is unavoidable in $d = 1$ (cf. Theorem 5.4).

6. Computation in high dimensions

We now discuss how to make our estimators computationally feasible when d is large. The main idea is that both the modal interval and shorth estimators involve finding optimal balls in $\mathbb{R}^d$. To save on computation, we will show that restricting the search to balls centered at one of the n data points leads to estimators with similar performance guarantees. This is an idea previously introduced in the literature on mode estimation in i.i.d. scenarios [1, 10, 21].

Concretely, the modal interval and shorth estimators are replaced by the following:

ESTIMATOR 5 The computationally efficient modal interval estimator is defined by

$$\widetilde{\mu}_{M,r} := \arg \max_{x \in \{x_1,\ldots,x_n\}} R_n(f_{x,r}). \tag{6.1}$$

ESTIMATOR 6 The computationally efficient shorth estimator is defined by

$$\widetilde{r}_k := \inf_r \sup_{x \in \{x_1,\ldots,x_n\}} \left\{ R_n(f_{x,r}) \geqslant \frac{k}{n} \right\}, \qquad \widetilde{\mu}_{S,k} := \widetilde{\mu}_{M,\widetilde{r}_k}. \tag{6.2}$$

In other words, we select the data point such that the smallest ball centered around that point containing at least k points has the minimum radius.

Note that both estimators (6.1) and (6.2) may be computed in $O(n^2 d)$ time. In contrast, computing the modal interval or shorth estimators directly would correspond to solving the circle placement problem or smallest enclosing ball problem, for which the best-known exact algorithms are $\Omega(n^d)$ [3, 14, 27].

Using a peeling argument [43], we can obtain a more refined concentration result than Theorem 2.2. The proof of the following result is contained in Appendix A.3. Note that the proof critically leverages radial symmetry of R, whereas the concentration inequality in Lemma 2.2 does not require R to be radially symmetric.

LEMMA 6.1 Recall Definitions 2.2 and 2.4 for the terms $f_{x,r}$, $R_n(\cdot)$ and $R(\cdot)$. For any $t \in (0, 1]$, radii $\bar{r}, r > 0$, and $n > 1$, we have the following inequalities:

$$\mathbb{P}\Big(|R_n(f_{x,r}) - R(f_{x,r})| \leqslant 2tR(f_{x,r}), \quad \forall x \text{ s.t.} \|x\|_2 \leqslant \bar{r} \Big) \geqslant 1 - \frac{2\exp(-cnt^2 R(f_{\bar{r},r}))}{1 - \exp(-cnt^2 R(f_{\bar{r},r}))}, \tag{6.3}$$

$$\mathbb{P}\left(\sup_{\|x\|_2 \geqslant \bar{r}} |R_n(f_{x,r}) - R(f_{x,r})| \geqslant tR(f_{\bar{r},r}) \right) \leqslant 2\exp(-cnt^2 R(f_{\bar{r},r})), \tag{6.4}$$

provided $\bar{r}$ and r are such that $R(f_{\bar{r},r}) \geqslant \frac{C_t d \log n}{n}$.

Using Lemma 6.1, we can derive the following results for the computationally efficient modal interval and shorth estimators. The proof is contained in Appendix D.6.

THEOREM 6.1 Recall Definition 2.5 of the term r_k. For the computationally efficient estimators, we have the following error guarantees:

(i) Suppose $r \geqslant 2r_{6Cd\log n}$. Then the modal interval estimator satisfies the bound $\|\widetilde{\mu}_{M,r}\|_2 \leqslant 4r\left(\frac{n}{Cd\log n}\right)^{1/d}$, with probability at least $1 - 6\exp(-c_3 d \log n)$.

(ii) Suppose $k \geqslant 2C_{0.5}(d+1)\log n$. Then the shorth estimator satisfies the bound $\|\widetilde{\mu}_{S,k}\|_2 \leqslant 4r_{2k}\left(\frac{2n}{k}\right)^{1/d}$, with probability at least $1 - 2\exp(-c'k)$.

REMARK 10 Comparing Theorem 6.1(i) with Theorem 4.1, we see that the computationally efficient modal interval essentially incurs an additional factor of 2 in the error bound, since we require $r \geqslant 2r_{C'd\log n}$. If we take $k = Cd\log n$, the error bound in Theorem 6.1(ii) is very similar to the error guarantee for the modal interval estimator (4.2) derived in Theorem 4.3, except for an extra factor of 2.

Of course, the quality of the guarantee in Theorem 6.1(i) worsens as r increases. As discussed in Section 4.1, we can use Lepski's method to calibrate the modal interval radius. Note that we can again use the shorth estimator to obtain rough upper and lower bounds. Using a similar argument as in the proof of Lemma 4.1, we are guaranteed that $\frac{1}{2}\widetilde{r}_{3Cd\log n} \leqslant r_{6Cd\log n} \leqslant \widetilde{r}_{6Cd\log n}$, w.h.p. Essentially, the same argument as in Theorem 4.2 then shows that the error of the modal interval estimator with Lepski calibration is guaranteed to be upper-bounded by $12r_{6Cd\log n}\left(\frac{n}{Cd\log n}\right)^{1/d}$.

As discussed in Section 4.3, the projection step for the hybrid screening procedure can be computed in $O(d)$ time. The construction of the cuboid S_k^{∞} itself can clearly be computed in $O(nd)$ time. Thus, one can also easily obtain the $O(\sqrt{n}^{1/d})$ rates using a computationally efficient hybrid estimator, as well.

7. Relaxing radial symmetry

We now consider the case when the population-level distribution $\overline{P} = \frac{1}{n}\sum_{i=1}^{n} P_i$ is not symmetric. In the case $d = 1$, we can obtain the same estimation error rates only assuming that density p_i is log-concave with a unique mode at 0. In the case $d > 1$, we can obtain weaker estimation error guarantees of the order $O(\sqrt{n})$ rather than $O(\sqrt{n}^{1/d})$ if we only assume that the mixture components are centrally symmetric. Furthermore, it is possible to obtain $O(n^{1/d})$ rates if we assume that a certain fraction of the components are radially symmetric.

Although radial symmetry is a strict assumption, it provides us an $O(\sqrt{n}^{\frac{1}{d}})$ error. Whereas if we just assume central asymmetry, a union bound argument gives $O(\sqrt{dn})$ error. This factor of $O(\sqrt{dn})$ cannot be improved in general. To see this, note that there exists a problem instance in single dimension where the lower bound is a factor of $\tilde{\Omega}(\sqrt{n})$. Central symmetry allows for having the same 'hard' problem on each dimension separately, forcing an $\tilde{\Omega}(\sqrt{n})$ error in each dimension.

We can relax the radial symmetry assumptions slightly. In particular, Theorem 2.2 only relies on the fact that R_r^*, the mass of the interval centered around the true mode 0, is $\Omega\left(\frac{\log n}{n}\right)$ (with no additional symmetry assumptions). We do need $R(f_{x,r})$ to satisfy some additional monotonicity assumptions along rays as x moves away from 0.

7.1 *General theory*

In place of radial symmetry, we impose the following condition (stated with respect to a fixed radius r):

(C1) The population-level quantity $R(f_{x,r})$ is maximized at $x = 0$, and otherwise monotonically decreasing along rays from the origin.

Note that condition (C1) is satisfied if the same property holds for all components p_i in the mixture. We now define the function

$$g(a, r) := \sup_{\|x\|_2 = a} R(f_{x,r}), \tag{7.1}$$

for $a, r > 0$. By Lemma 2.1, we can argue that under radial symmetry of R, we have $g(a, r) \leqslant \frac{1}{N(B_a, r)} \leqslant \left(\frac{r}{a}\right)^d$, which can then be plugged into the argument of Theorem 4.1. The proof of the following statement is contained in Appendix F.1.

THEOREM 7.1 Suppose condition (C1) holds.

(i) Recall Definition 2.4 of R_r^*. Suppose r is such that $R_r^* = \Omega\left(\frac{d \log n}{n}\right)$, and r' is chosen sufficiently large such that $g(r', r) < \frac{R_r^*}{2}$. Then the modal interval estimator satisfies $\|\widehat{\mu}_{M,r}\|_2 \leqslant r'$, w.h.p.

(ii) Recall Definition 2.5 of r_k. Suppose r' is chosen such that $g(r', r_{8d \log n}) \leqslant \frac{8d \log n}{4n}$. With high probability, the error of the shorth estimator satisfies $\|\widehat{\mu}_{S,k}\|_2 \leqslant r'$, and the error of the hybrid algorithm with $k_2 = r_{8d \log n}$ is bounded by $\min(r', \sqrt{d} r_{4\sqrt{n \log n}, 1})$.

REMARK 11 For radially symmetric distributions, note that $g(r', r) \leqslant \left(\frac{r}{r'}\right)^d$, so we can take $r' = r\left(\frac{2}{R_r^*}\right)^{1/d}$ and $r' = r_{2k}\left(\frac{4}{R_{2k}^*}\right)^{1/d}$ to obtain the results of Theorems 4.1 and 4.3 for the modal interval and shorth estimators, respectively. Furthermore, by Lemma 2.1(iii), we have $r_{\sqrt{n} \log n} \leqslant \left(\frac{\sqrt{n}}{8d}\right)^{1/d} r_{8d \log n}$. Thus, we also recover the analog of Theorem 4.4 for the hybrid estimator.

Finally, note that an analog of Theorem 7.1 holds when we use the computationally efficient modal interval and shorth estimators described in Section 6, with minor proof modifications.

7.2 *Sufficient conditions*

Condition (C1) may be a bit difficult to interpret. We define two related conditions:

(C2) Each component density p_i is log-concave with a unique mode at 0. Recall that a distribution with density p is log-concave if $p(x) \propto e^{-\phi(x)}$ for a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$.

(C3) For all $x \in \mathbb{R}^d$ and all $1 \leqslant i \leqslant n$, we have $p_i(x) = p_i(-x)$.

Note that condition (C3) only requires symmetry of the density around 0, rather than radial symmetry; in particular, it holds for Gaussian distributions that are not necessarily isotropic.

We have the following result, proved in Appendix F.2:

PROPOSITION 7.2 Suppose conditions (C2) and (C3) hold. Then condition (C1) also holds. Furthermore, $g(a, r) \leqslant \frac{1}{\lfloor a/2r \rfloor}$.

In fact, we can even derive a result only assuming condition (C2) in the case $d = 1$. As argued in the proof of Theorem 7.1, we may establish that $R(f_{\widehat{\mu}_{M,r}}, r) \geqslant \frac{R_r^*}{2}$, w.h.p. Thus, there exists some i such that $R_i(f_{\widehat{\mu}_{M,r}}, r) \geqslant \frac{R_r^*}{2}$. By properties of log-concave convolutions (cf. proof of Proposition 7.2), we know that $R_i(f_{x,r})$ is decreasing along rays originating from some point x_i^*, and also $\|\widehat{\mu}_{M,r} - x_i^*\|_2 \leqslant \frac{4r}{R_r^*}$, since we could otherwise pack too many intervals into the ray between x_i^* and $\widehat{\mu}_{M,r}$, thus contradicting the inequality $R_i(f_{\widehat{\mu}_{M,r}}, r) \geqslant \frac{R_r^*}{2}$. Finally, note that due to the unimodality of p_i at 0, we clearly have

$\|x_i^*\|_2 \leqslant r$. Altogether, we obtain the error bound

$$\|\widehat{\mu}_{M,r}\|_2 \leqslant \frac{4r}{R_r^*} + r,$$

which is of the same order as the guarantees in Theorem 3.1. A similar conclusion could be reached if we replaced condition (C2) by the condition that each p_i has a unique median and mode at 0, since $R_i(f_{x,r})$ is decreasing along rays originating from r $(-r)$ in the positive (negative) direction.

7.3 *Examples*

We now describe two examples to illustrate concrete use cases of our more general theory.

EXAMPLE 7.3 (Elliptically symmetric distributions). We now consider the case where the components of the mixture are not spherical but have the same axes of symmetry. Concretely, suppose that for a fixed matrix $\Sigma \succ 0$, the density of each X_i is of the form $f_i\left((x-\mu)^T\Sigma^{-1}(x-\mu)\right)$, where $f_i : \mathbb{R} \to \mathbb{R}$ is a decreasing function defined on the positive reals. The goal is to estimate the common parameter $\mu \in \mathbb{R}^d$. As a specific example, we might have a mixture of non-isotropic Gaussian distributions where the covariance matrices are all scalar multiples of Σ. This strictly generalizes the case of radially symmetric distributions, which corresponds to the case $\Sigma = I$.

Suppose we employ the modal interval, shorth or hybrid estimators described above. Note that these estimators do not require knowledge of the matrix Σ. We wish to analyze the behavior of the quantity $g(a, r)$ defined in equation (7.1), which is relevant for Theorem 7.1. Indeed, we can derive an analog of Lemma 2.1 that applies in this setting. The main step is to understand bound the quantity $g(r_2, r_1)$ when $r_1 < r_2$. We have the following result, proved in Appendix F.3:

PROPOSITION 7.4 Let $r_1 < r_2$. For an elliptically symmetric distribution, we have

$$g(r_2, r_1) \leqslant C\left(\frac{r_1\lambda_{\max}(\Sigma)}{r_2\lambda_{\min}(\Sigma)}\right)^d.$$

Clearly, taking $C = 1$ and $\Sigma = I$ in Proposition 7.4 recovers the result for radially symmetric distributions.

REMARK 12 Similar arguments as in Example 7.3 could be applied in the case when the probability density functions of the distributions are proportional to $\exp(-\|x-\mu\|/\sigma)$, for a different norm $\|\cdot\|$ besides the squared ℓ_2-norm or the Mahalanobis norm. Also note that if the matrix Σ (accordingly, the norm $\|\cdot\|$) were known *a priori*, it might be possible to obtain better rates by using a modal interval/shorth estimator based on the level sets of the norm rather than spheres of varying radii.

EXAMPLE 7.5 (Mixture of radially and centrally symmetric distributions). For another interesting special case, suppose we have s points drawn from radially symmetric distributions, and $n-s$ points drawn from centrally symmetric distributions. Suppose we have $f(n)$ points which are well behaved in the sense that the interquartile range of the corresponding distributions is small. (These distributions need not coincide with the radially symmetric distributions.) We have the following result, proved in Appendix F.4:

PROPOSITION 7.6 For $r = q_{(f(n))}$ and $r' = 2rn^{1/d}$, we have

$$g(r', r) \leqslant \frac{R_r^*}{2},$$

provided $s \geqslant n - 2n^{1/d}(f(n) - 4)$.

Thus, as the proportion of well-behaved points increases, the required proportion of radially symmetric distributions required to obtain a specific error guarantee becomes smaller. In particular, if $f(n) = \Omega(n^{1-1/d})$, we do not need any radially symmetric distributions; recall, however, that the coordinatewise median already performs well on a mixture of centrally symmetric distributions if $f(n) = \Omega(\sqrt{n}\log n)$.

8. Linear regression

We now shift our focus to the problem of linear regression and demonstrate how the methodology developed thus far may be adapted to parameter estimation in multivariate regression. Suppose we have observations $\{(x_i, y_i)\}_{i=1}^n$ from the linear model

$$y_i = x_i^T \beta^* + \epsilon_i, \qquad \forall 1 \leqslant i \leqslant n, \tag{8.1}$$

where the pairs $\{(x_i, \epsilon_i)\}_{i=1}^n$ are independent but not necessarily identically distributed and x_i and ϵ_i are independent for each i.

Following the theme of our paper, we assume that the probability density function of ϵ_is is symmetric and unimodal. We want to study the behavior of the modal interval regression estimator

$$\widehat{\beta} \in \operatorname*{argmax}_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n 1\left\{|y_i - x_i^T \beta| \leqslant r\right\}, \tag{8.2}$$

for an appropriate choice of $r > 0$.

A natural question is whether the true parameter β^* is the unique population-level maximizer in the regression setting. As the following proposition shows, this is indeed the case when the densities of the x_is are absolutely continuous with respect to Lebesgue measure. The proof is contained in Appendix G.1.

PROPOSITION 8.1 Consider the linear model in equation (8.1), where the distributions of x_is and ϵ_is have Lebesgue density. Then the population-level maximizer is given by

$$\beta^* = \arg\max_{\beta} \sum_{i=1}^n \mathbb{E}\left[1\left\{|y_i - x_i^T \beta| \leqslant r\right\}\right], \qquad \forall r > 0. \tag{8.3}$$

Importantly, Proposition 8.1, and the ensuing theory, does not require specific assumptions on the form of the distribution of the x_is. However, in order to derive easily interpretable error bounds on the modal interval regression estimator, we will assume further distributional assumptions (cf. the statement of Theorem 8.2 below).

8.1 *Estimation error*

In order to obtain error bounds on $\|\widehat{\beta} - \beta^*\|_2$, we need to analyze the behavior of the quantities

$$R_\beta := \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(|y_i - x_i^T\beta| \leqslant r\right),$$

for a fixed value of r, chosen sufficiently large that $R_{\beta^*} \geqslant \frac{Cd\log n}{n}$. In particular, we want to show that for $\|\beta - \beta^*\|_2$ larger than a certain value, we will have $R_\beta < \frac{R_{\beta^*}}{2} = \frac{1}{2n}\sum_{i=1}^{n}\mathbb{P}(|\epsilon_i| \leqslant r)$.

As before, the key ingredient for deriving error bounds is a uniform concentration result. This is proved in the following lemma:

LEMMA 8.1 Let $t \in (0, 1]$, and suppose r is large enough so that $R_{\beta^*} \geqslant \frac{Cd\log n}{n}$. Then

$$\mathbb{P}\left(\sup_{\beta\in\mathbb{R}^d, r'\leqslant r}\left|\frac{1}{n}\sum_{i=1}^{n} 1\left\{|y_i - x_i^T\beta| \leqslant r'\right\} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[1\left\{|y_i - x_i^T\beta| \leqslant r'\right\}\right]\right| \geqslant tR_{\beta^*}\right) \leqslant 2\exp(-cnR_{\beta^*}t^2). \quad (8.4)$$

Since the proof is directly analogous to the proof of Theorem 2.2, we only provide a sketch: the key point is to consider the VC dimension of the class of functions $f(x, y) = 1\{|y - x^T\beta| \leqslant r\}$, indexed by the pair (β, r). Note that the subset of points in $\mathbb{R}^{d+1}$ associated with the indicator function $f(x, y)$ is an intersection of two halfspaces. Using results on the VC dimension of an intersection of concept classes [44], we see that the VC dimension of the desired hypothesis class is bounded by $C'd$. The concentration result then follows by the same arguments used to derive Theorem 2.2.

It is generally difficult to state general bounds on estimation error that depend only on order statistics of quantiles, since as in the case of mean regression, the error bounds one can derive will be largely problem-dependent. In order to simplify our presentation, we will only discuss the case where the ϵ_is and x_is are Gaussian: $\epsilon_i \sim N(0, \sigma_i^2)$ and $x_i \sim N(\mu_i', \Sigma_i')$. We have the following result, proved in Appendix G.2:

THEOREM 8.2 Let $\lambda_{\min} := \min_i \lambda_{\min}(\Sigma_i')$, and suppose $\lambda_{\min} > 0$. Suppose $r > 0$ is chosen such that $R_{\beta^*} \geqslant \frac{Cd\log n}{n}$. Then the regression estimator (8.2) satisfies

$$\|\widehat{\beta} - \beta^*\|_2 \leqslant \frac{c' n\sigma_{(cd\log n)}}{\sqrt{\lambda_{\min}}},$$

w.h.p.

We conjecture that it is possible to decrease this upper bound to $O(\sqrt{n}\sigma_{(c\log n)})$ by an appropriate hybrid screening procedure, but we leave this to future work. Also note that in order for the bound in Theorem 8.2 to be useful, the quantity $\lambda_{\min}$ must either be a constant, or else not decrease too rapidly with n.

8.2 *Computation*

A natural question is whether the modal interval regression estimator (8.2) is actually computationally feasible. We claim that an estimator may be obtained in $O(n^d)$ time, using Algorithm ??. The proof is in Appendix G.3.

ALGORITHM 3 Modal interval regression estimator

function MODALINTERVALREGRESSION($X_{1:n}, k_1, k_2, d$)

Construct the set of hyperplanes

$$\mathcal{S}_r = \{y_i = x_i^T\beta + r\} \bigcup \{y_i = x_i^T\beta - r\}.$$

Let $\{S_1, \ldots, S_N\}$ denote the set of subsets of $\mathcal{S}_r$ cardinality d.

for $j = 1, \ldots, N$

Solve the system of linear equations given by S_j. Let B_j be a solution (if one exists).

end for

$j^* \leftarrow \arg\max_{1 \le j \le N} \frac{1}{n} \sum_{i=1}^n 1\left\{|y_i - x_i^T\beta_j| \le r\right\}$

return β_{j^*}

end function

THEOREM 8.3 The output of Algorithm 3 is a maximizer of equation (8.2).

REMARK 13 Correct application of Algorithm 3 would assume that r is chosen appropriately. It is less clear how this parameter might be calibrated based on the data, perhaps using an appropriate variant of Lepski's method. We leave this important open question to future work.

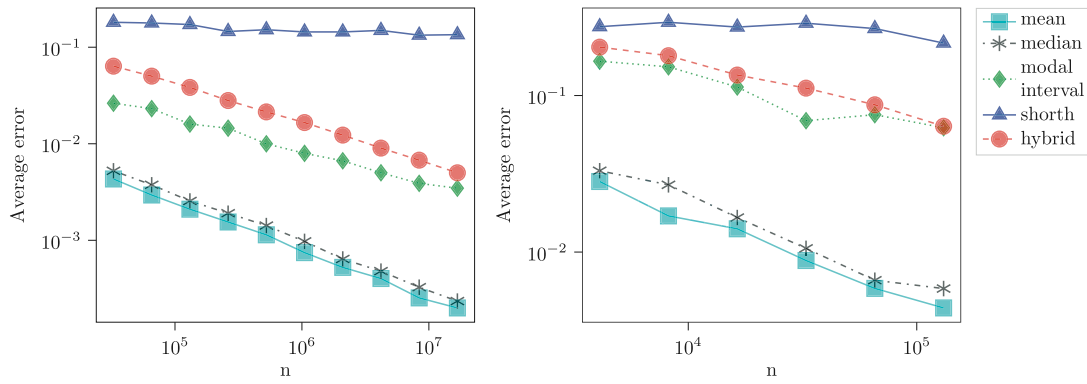
9. Simulations

We now present the results of simulations on the recurring examples to validate our theoretical predictions (cf. Table 1). Although our theorem statements involve large constants, we empirically observe that smaller constants suffice to elicit the same behavior predicted by our theory. We run the k-shorth estimator with $k = 5d \log n$ and k-median with $k = \sqrt{n} \log n$. We use these estimators for the hybrid estimator, i.e., the $(\sqrt{n} \log n, 5d \log n)$-hybrid estimator. The mean estimator corresponds to the simple average, whereas the median estimator refers to the (coordinatewise) sample median.

For each n, we run $T = 200$ simulations for univariate data and $T = 20$ simulation for multivariate data and report the average error $\frac{1}{T} \sum_{i=1}^T |\widehat{\mu} - \mu^*|$ of various estimators. Both axes in all of the plots are in a log-scale. In particular, the slope of the curves indicates the power of n in the estimation error, and vertical shifts correspond to constant prefactors.

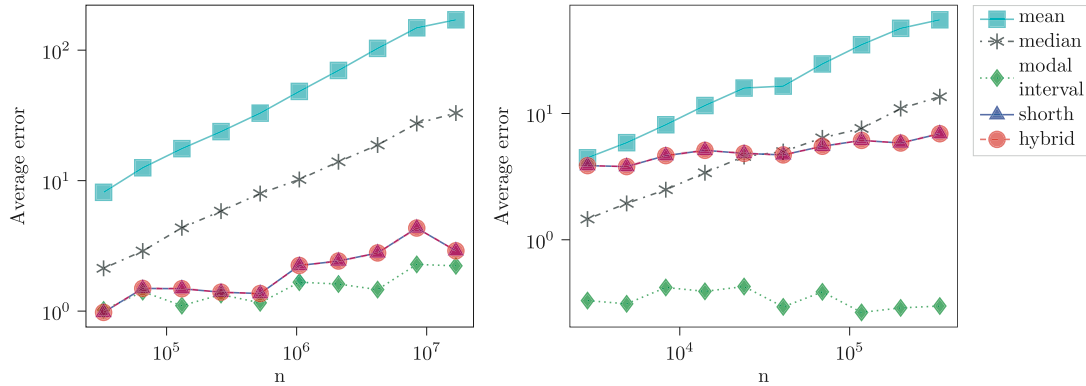
9.1 *Univariate data*

We first present simulation results when $d = 1$. We use $r = 1$ for the simulations involving r-modal interval estimators, since $R_1^* = \Omega\left(\frac{\log n}{n}\right)$ in each of the recurring examples, although the constant prefactors do not exactly align with our theory.



(a) $d = 1$ (b) $d = 3$

FIG. 3. Plot comparing average error of various estimators on Example 3.4. Both the mean and median exhibit the familiar $O(n^{-0.5})$ error rate. The modal interval has errors of order $n^{-1/3}$. As suggested by our theoretical bounds, the $(\log n)$-shorth has constant error. The hybrid estimator improves the rate of the shorth estimator, with a similar error decay as the median estimator as n increases.



(a) $d = 1$ (b) $d = 3$

FIG. 4. Plot comparing average error of various estimators on Example 3.5. As mentioned in Table 1, both the mean and median have $\sqrt{n}$ error rate. The error rates of the modal interval, shorth (with $k = 5d \log n$) and hybrid estimators are superior to the median in the univariate case, and the hybrid estimator is clearly superior when $d = 3$.

In the case of Example 3.4 (i.i.d. observations), we generate $x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. As seen in Fig. 3(a), the mean and median estimators perform optimally in this setting, giving an error rate of $O(n^{-0.5})$. In contrast, the shorth estimator (with $k = 5 \log n$) has a flat trend line indicative of constant error, as suggested by Remark 2 and the phase transition arguments in Section 3.1.2. On the other hand, the error of the hybrid estimator decays at a rate more comparable to the mean and median. As discussed in Remark 6, the hybrid estimator is indeed optimal up to log factors. We see that the performance of the modal interval estimator is better than the shorth but worse than the hybrid estimator and exhibits the cube-root asymptotic decay encountered in classical statistics [23]. Furthermore, the estimation error of the hybrid estimator behaves more like the error of the median estimator as n increases. Note that although our bounds for the shorth and modal interval estimators are tighter for smaller values of k and

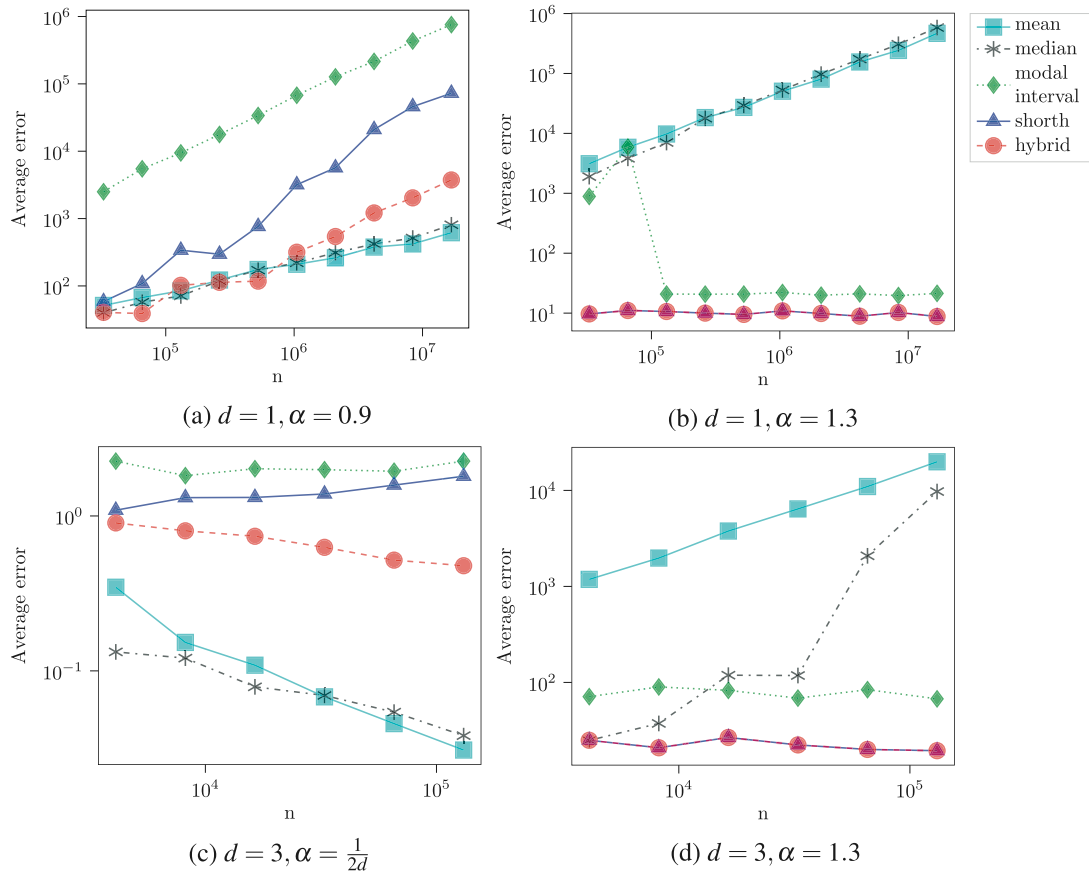


FIG. 5. Plots comparing average error of various estimators on Example 3.6 for different values of α. As suggested by Proposition 3.9, the median and mean have superior performance to the modal interval and shorth estimators for $\alpha < 1$. Moreover, the hybrid estimator exhibits similar behavior to the median when $\alpha < 1$ and to the shorth when $\alpha > 1$.

r, choosing larger values results in better performance when the data are homogeneous, which is not a valid assumption in our general use case.

For Example 3.5 (quadratic variance), we generate $x_i \sim \mathcal{N}(0, i^2)$. In Fig. 4(a), we see that the both the median and mean have similar slopes: Proposition 3.9 predicts that the median would have $\tilde{O}(\sqrt{n})$ error, compared to the $\Theta\left(\sqrt{\frac{1}{n^2}\sum_{i=1}^{n} i^2}\right) = \Theta\left(\sqrt{n}\right)$ error of the mean; indeed, the curves are roughly parallel. However, the error rate of the modal interval, shorth and hybrid estimators is significantly smaller. As stated in Propositions 3.8 and 3.10, the error of these estimators is upper bounded by $O(n^{\epsilon})$, for $\epsilon > 0$.

For Example 3.6 (α-mixture distributions), we generate $\lceil 10 \log n \rceil$ samples from a $\mathcal{N}(0, 4 \times 10^{-4})$ distribution and the remaining samples from a $\mathcal{N}(0, n^{\alpha})$ distribution, with $\alpha = 0.9$ and 1.3. The plots in Fig. 5 add additional curves to the phase transition plots in Fig. 1. As suggested by Propositions 3.8 and 3.10, the modal, shorth and hybrid estimators have constant error for $\alpha > 1$, whereas the error increases with n when $\alpha < 1$. Furthermore, the hybrid estimator performs better than the shorth estimator when $\alpha < 1$, with an error rate of $O(n^{\alpha - 0.5})$ rather than $O(n^{\alpha})$, while the modal interval estimator seems to perform comparably to the hybrid. Finally, note that the behavior of the hybrid

estimator is similar to the behavior of the median estimator when $\alpha < 1$ and to the modal interval/shorth estimator when $\alpha > 1$, showing that it indeed enjoys the better of the two rates in different regimes.

9.2 *Multivariate*

We now present simulation results for multivariate data, using $d = 3$. The data for all three recurring examples are generated with the same parameters as in the univariate case, except with isotropic distributions. We run the computationally efficient versions of the shorth and modal interval estimators described in Section 6, with $k = 5d \log n$ and $r = \sqrt{d}$.

The trends for i.i.d. data, shown in Fig. 3(b), are analogous to the univariate case. Similarly, the plots in Fig. 4(b) for the quadratic variance example resemble the plots in Fig. 4(a), with the hybrid, shorth and modal interval estimators performing noticeably better than the mean or median. Note that for these experiments, the modal interval estimator appears to behave better than either the shorth or hybrid estimators by a constant factor. For the multivariate version of the α-mixture distribution, we run simulations with $\alpha = \frac{1}{2d} < 1$ and $\alpha = 1.3$, where we have chosen the first value of α so that the upper bound in Theorem 5.5 gives $O\left(n^{\alpha - \frac{1}{2}}\right) = O(n^{\frac{1}{2d} - \frac{1}{2}})$ error for the hybrid estimator, whereas the derived bounds for the modal interval and shorth are $O(n^{\alpha}) = O(n^{\frac{1}{2d}})$ (cf. Remark 6). Indeed, we see in Fig. 5(c) that the estimation error of the hybrid estimator decreases with n, like the mean and median estimators, whereas the shorth estimator has an increasing trend line. The curve for the modal interval estimator appears to be roughly constant (or possibly slightly increasing). The curves in Fig. 5(d) are very similar to the curves in Fig. 5(b), suggesting the existence of a phase transition for $\alpha \in \left(\frac{1}{2d}, 1\right]$ in the multivariate case, as well.

10. Conclusion

We have studied the problem of mean estimation of a heterogeneous mixture when the fraction of clean points tends to 0. We have shown that the modal interval and shorth estimators, which perform suboptimally in i.i.d. settings, are superior to the sample mean in such settings. We have also shown that these estimators and the k-median have complementary strengths that may be combined into a single hybrid estimator, which adapts to the given problem and is nearly optimal in certain settings. An important question for further study is whether the proposed hybrid estimator is always near-optimal, or optimal, for more general collections of variances.

Our discussion of linear regression estimators has been fairly brief. Some issues that we have not addressed include derivations for non-Gaussian error distributions and regression estimators in the case of a fixed design matrix. We leave these questions, and a derivation of optimal error rates in the linear regression setting, for future work.

Funding

National Science Foundation (DMS-1749857 to A.P. and P.L., CCF-1841190 to V.J., CCF-1740707 to A.P.).

Acknowledgements

The authors thank the reviewers for their detailed feedback, which helped improve the manuscript. P.L. thanks Gabor Lugosi for introducing her to the entangled mean estimation problem at the 2017 probability and combinatorics workshop in Barbados.

References

1. Abraham, C., Biau, G. & Cadre, B. (2004) On the asymptotic properties of a simple estimate of the mode. *ESAIM Probab. Stat.*, **8**, 1–11.
2. Achlioptas, D. & McSherry, F. (2005) On spectral learning of mixtures of distributions. *International Conference on Computational Learning Theory*. Berlin, Heidelberg: Springer, pp. 458–469.
3. Agarwal, P. K. & Sharir, M. (1998) Efficient algorithms for geometric optimization. *ACM Comput. Surv.*, New York, USA **30**, 412–458.
4. Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. & Tukey, J. W. (1972) *Robust Estimates of Location: Survey and Advances*. Princeton University Press.
5. Arora, S. & Kannan, R. (2001) Learning mixtures of arbitrary Gaussians. *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing*, New York, USA: Association for Computing Machinery, pp. 247–257.
6. Boucheron, S., Lugosi, G. & Massart, P. (2016) *Concentration Inequalities: A Nonasymptotic Theory of Independence*, 1st edn. Oxford, UK: Oxford University Press.
7. Chernoff, H. (1964) Estimation of the mode. *Ann. Inst. Stat. Math.*, **16**, 31–41.
8. Chierichetti, F., Dasgupta, A., Kumar, R. & Lattanzi, S. (2014) Learning entangled single-sample Gaussians. *Proceedings of the 25th Annual Symposium on Discrete Algorithms, SODA*, Philadelphia, USA: SIAM pp. 511–522.
9. Dasgupta, S. (1999) Learning mixtures of Gaussians. *40th Annual Symposium on Foundations of Computer Science*. Los Alamitos, USA: IEEE, pp. 634–644.
10. Dasgupta, S. & Kpotufe, S. (2014) Optimal rates for k-NN density and mode estimation. *Advances in Neural Information Processing Systems*, New York, USA: Curran Associates, Inc. pp. 2555–2563.
11. Devroye, L., Lattanzi, S., Lugosi, G. & Zhivotovskiy, N. (2020) On mean estimation for heteroscedastic random variables. arXiv:2010.11537.
12. Dundar, M., Krishnapuram, B., Bi, J. & Rao, R. B. (2007) Learning classifiers when the training data is not IID. *IJCAI*, San Francisco, USA: Morgan Kaufmann Publishers Inc. pp. 756–761.
13. El Bantli, F. & Hallin, M. (1999) L_1-estimation in linear models with heterogeneous white noise. *Statist. Probab. Lett.*, **45**, 305–315.
14. Eppstein, D. & Erickson, J. (1994) Iterated nearest neighbors and finding minimal polytopes. *Discrete Comput. Geom.*, **11**, 321–350.
15. Flaxman, S. R., Neill, D. B. & Smola, A. J. (2016) Gaussian processes for independence tests with non-iid data in causal inference. *ACM Trans. Intell. Syst. Technol.*, **7**, 22.
16. Hallin, M. & Mizera, I. (1997) Unimodality and the asymptotics of M-estimators. *Lecture Notes-Monograph Series*. Institute of Mathematical Statistics **31**, 47–56. https://www.jstor.org/stable/4355966
17. Hallin, M. & Mizera, I. (2001) Sample heterogeneity and M-estimation. *J. Statist. Plann. Inference*, **93**, 139–160.
18. Hoeffding, W. (1956) On the distribution of the number of successes in independent trials. *Ann. Math. Stat.*, **27**, 713–721.
19. Huber, P. J. (1964) Robust estimation of a location parameter. *Ann. Math. Stat.*, **35**, 73–101.
20. Huber, P. J. & Ronchetti, E. M. (2009) *Robust Statistics*. New York, USA: John Wiley & Sons, Inc.
21. Jiang, H. (2017) Uniform convergence rates for kernel density estimation. *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research (PMLR) pp. 1694–1703.
22. Kannan, R., Salmasian, H. & Vempala, S. (2005) The spectral method for general mixture models. *International Conference on Computational Learning Theory*. Berlin, Heidelberg: Springer, pp. 444–457.
23. Kim, J. & Pollard, D. (1990) Cube root asymptotics. *Ann. Statist.*, **18**, 191–219.
24. Knight, K. (1999) Asymptotics for L_1-estimators of regression parameters under heteroscedasticity. *Can. J. Stat.*, **27**, 497–507.

25. Lai, K. A., Rao, A. B. & Vempala, S. (2016) Agnostic estimation of mean and covariance. *57th Annual Symposium on Foundations of Computer Science (FOCS)*, Los Alamitos, USA: IEEE Computer Society pp. 665–674.

26. Lecué, G. & Depersin, J. (2019) Robust subgaussian estimation of a mean vector in nearly linear time. arXiv:1906.03058.

27. Lee, D.-T. & Preparata, F. P. (1984) Computational geometry—a survey. *IEEE Trans. Comput.*, **12**, 1072–1101.

28. Lepskii, O. V. (1991) On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.*, **35**, 454–466.

29. Li, J., Marsiglietti, A. & Melbourne, J. (2020) Further investigations of Rényi entropy power inequalities and an entropic characterization of s-concave densities In: Klartag B., Milman E. (eds) Geometric Aspects of Functional Analysis. Lecture Notes in Mathematics, vol 2266. Springer, Cham.

30. Liang, Y. & Yuan, H. (2020) Learning entangled single-sample Gaussians in the subset-of-signals model. *Proceedings of Machine Learning Research*, vol. 125. PMLR, pp. 2712–2737.

31. Lindsay, B. G. (1995) Mixture models: theory, geometry and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*. JSTOR, pp. i–163.

32. Liu, R. Y. (1988) Bootstrap procedures under some non-iid models. *Ann. Statist.*, **16**, 1696–1708.

33. Lugosi, G. & Mendelson, S. (2019) Robust multivariate mean estimation: the optimality of trimmed mean *Ann. Statist.* **49**, 393–410.

34. Mizera, I. & Wellner, J. A. (1998) Necessary and sufficient conditions for weak consistency of the median of independent but not identically distributed random variables. *Ann. Statist.*, **26**, 672–691.

35. Pensia, A., Jog, V. & Loh, P. (2019) Mean estimation for entangled single-sample distributions. *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE.

36. Raskutti, G., Wainwright, M. J. & Yu, B. (2010) Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.*, **11**, 2241–2259.

37. Sen, P. K. (1968) Asymptotic normality of sample quantiles for m-dependent processes. *Ann. Math. Stat.*, **39**, 1724–1730.

38. Sen, P. K. (1970) A note on order statistics for heterogeneous distributions. *Ann. Math. Stat.*, **41**, 2137–2139.

39. Shorack, G. R. & Wellner, J. A. (2009) *Empirical Processes with Applications to Statistics*. Classics in Applied Mathematics. Philadelphia, USA: Society for Industrial and Applied Mathematics.

40. Steinwart, I. & Christmann, A. (2009) Fast learning from non-iid observations. *Advances in NIPS*, New York, USA: Curran Associates, Inc. pp. 1768–1776.

41. Stigler, S. M. (1976) The effect of sample heterogeneity on linear functions of order statistics, with applications to robust estimation. *J. Am. Stat. Assoc.*, **71**, 956–960.

42. Tsybakov, A. B. (2008) *Introduction to Nonparametric Estimation*. Springer Series in Statistics. New York, USA: Springer-Verlag, New York.

43. Van de Geer, S. A. (2000) *Empirical Processes in M-Estimation*, vol. 6. Cambridge, UK: Cambridge University Press.

44. Van Der Vaart, A. & Wellner, J. A. (2009) A note on bounds for VC dimensions. *Inst. Math. Stat. Collect.*, **5**, 103.

45. Vershynin, R. (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge, UK: Cambridge University Press.

46. Wainwright, M. J. (2019) *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

47. Weiss, L. (1969) The asymptotic distribution of quantiles from mixed samples. *Sankhya A*, **31**, 313–318.

48. Wenocur, R. S. & Dudley, R. M. (1981) Some special Vapnik–Chervonenkis classes. *Discrete Math.*, **33**, 313–318.

49. Zhu, T., Xiong, P., Li, G. & Zhou, W. (2015) Correlated differential privacy: Hiding information in non-iid data set. *IEEE Trans. Inf. Forensics Secur.*, **10**, 229–242.

A. Proofs of preliminaries

In this appendix, we provide proofs for the preliminary lemma concerning properties of radially symmetric distributions in Section 2, as well as the concentration results used in the paper.

A.1 *Proof of Lemma 2.1*

1. Note that $R(f_{x,r})$ can be written as convolution of $\overline{P}$ with indicator function of B_r, both of which are unimodal and radially symmetric. The desired result then follows by Li *et al.* [29, Proposition 8], which implies that $R(f_{x,r})$ is also unimodal and radially symmetric.
2. This follows from the non-negativity of the density.
3. As $\overline{P}$ is radially symmetric, let the density of $\overline{P}$ at x be given by $p(\|x\|)$. R_r^* can be written as $R_r^* = C\int_0^r p(s)s^{d-1}ds$ where C is a constant for a fixed dimension. Define $g(r) := \frac{R_r^*}{Cr^d} = \frac{\int_0^r p(s)s^{d-1}ds}{r^d}$ for $r > 0$. Property (iii) is equivalent to showing that $\frac{d}{dr}g(r) < 0$. By unimodality of $p(\cdot)$, it follows that $g(r) > \frac{p(r)}{d}$. Differentiating $g(\cdot)$, we get
$$\frac{d}{dr}g(r) = \frac{p(r)r^{d-1}r^d - dr^{d-1}\int_0^r p(s)s^{d-1}ds}{r^{2d}} = \frac{p(r) - dg(r)}{r} < 0.$$
4. Note that any r_1-packing of $B(0, r_2 - r_1)$ has the property that all balls in the packing must be entirely contained within the larger ball B_{r_2}. Furthermore, by Lemma 2.1(i) above, we know that $R(f_{x,r_1}) \geqslant R(f_{r_2,r_1})$ when $\|x\|_2 \leqslant r_2$. Hence, by summing up the densities of all balls in the packing, we obtain
$$R(f_{0,r_2}) \geqslant P(B_{r_2-r_1}, r_1)R(f_{r_2,r_1}),$$
from which the first inequality follows.
To obtain the second inequality, we use the sphere-packing lower bound
$$P(B_{r_2-r_1}, r_1) \geqslant N(B_{r_2-r_1}, 2r_1) \geqslant \left(\frac{r_2 - r_1}{2r_1}\right)^d,$$
where $N(\cdot,\cdot)$ denotes the covering number (cf. Vershynin [45, Proposition 4.2.12]).
5. The proof of the first inequality is the same as the proof of the corresponding statement in Lemma B.1. The second inequality follows by noting that $\mathbb{E}\|X_i - \mu\|_2^2 = \mathrm{Tr}(\Sigma_i) = d\sigma_i^2$. By Chebyshev's inequality, we have
$$\tilde{R}_i(f_{0,2\sigma_i\sqrt{d}}) = \mathbb{P}(\|X_i - \mu\|_2 \leqslant 2\sqrt{d}\sigma_i) \geqslant \frac{3}{4},$$
for each i. Thus, $B_{2\sigma_{(2k)}\sqrt{d}}$ covers at least $\frac{3}{4}$ of the mass of at least $2k$ distributions, implying the desired result.

A.2 *Proof of Lemma 2.2*

Recall that $\tilde{R}_i(f) = \mathbb{E}f(X_i)$. We define the random variables
$$Y_{f,i} := f(X_i) - \tilde{R}_i(f).$$

Note that $\mathbb{E}_i[Y_{f,i}] = 0$ and $|Y_{f,i}| \leqslant 1$. Furthermore, the variables $(Y_{f,i})_{i=1}^n$ are independent for each fixed f. Let

$$Z := \sup_{f \in \mathscr{H}_r} \big(R_n(f) - R(f)\big) = \sup_{f \in \mathscr{H}_r} \frac{1}{n} \sum_{i=1}^n Y_{f,i}.$$

We will apply Lemma H.2 to obtain a high-probability upper bound on Z. Here $V = d + 1$, the VC dimension of balls.

Since its application requires a bound on the expectation, we first derive the following lemma:

LEMMA A.1 If $nR_r^* \geqslant 1300V \log n$ with both $n > 1$ and $d \geqslant 1$, then

$$\mathbb{E}\, Z \leqslant 72\sqrt{V \frac{R_r^* \log n}{2n}}.$$

Proof. We will use Theorem H.1 from Appendix H, with $\sigma^2 = \sup_{x,r' \leqslant r} R(f_{x,r'}) = R_r^*$. In particular, note that since $n\sigma^2 \geqslant 1300V \log n$, we have

$$\log\left(\frac{4e^2}{\sigma}\right) = \frac{1}{2}\log\left(\frac{16e^4}{\sigma^2}\right) \leqslant \frac{1}{2}\log\left(\frac{16e^4 n}{1300V \log n}\right) \leqslant \frac{\log n}{2},$$

so

$$\left(24\sqrt{\frac{V}{5n}\log\left(\frac{4e^2}{\sigma}\right)}\right)^2 = \frac{576V}{5n}\log\left(\frac{4e^2}{\sigma}\right) \leqslant \frac{576V}{5n} \cdot \frac{\log n}{2} = 57.6V\frac{\log n}{n} \leqslant \sigma^2.$$

Thus, Theorem H.1 is applicable and leads to the following bound:[2]

$$\mathbb{E}\, Z \leqslant 72\frac{\sqrt{R_r^*}}{\sqrt{n}}\sqrt{V\log\left(\frac{4e^2}{\sigma}\right)} \leqslant 72\sqrt{\frac{VR_r^* \log n}{2n}}.$$

□

We now apply Theorem 12.9 from Boucheron *et al.* [6] (stated in Lemma H.2 in Appendix H) with $W_{i,s} = Y_{i,f}$ and

$$\rho^2 = \sup_{f \in \mathscr{H}_r} \sum_{i=1}^n \mathbb{E}\, Y_{i,f}^2 = \sup_{f \in \mathscr{H}_r} \sum_{i=1}^n \mathbb{V}[f(X_i)]$$

$$\leqslant \sup_{f \in \mathscr{H}_r} \sum_{i=1}^n \mathbb{E}[f(X_i)] = \sup_{f \in \mathscr{H}_r} nR(f) = nR_r^*,$$

[2] Note that the definition of Z in Theorem H.1 has a factor of $1/\sqrt{n}$ as opposed to the factor of $1/n$ here.

where the inequality holds because the variance of a Bernoulli random variable is bounded by its expectation. Hence, using Lemma A.1 and the assumption $nR_r^* \geqslant 1300V \log n$, we have

$$\begin{aligned} v &= 2n\,\mathbb{E}\,Z + \rho^2 \leqslant 2n\,\mathbb{E}\,Z + nR_r^* \\ &\leqslant 144\sqrt{0.5VnR_r^* \log n} + nR_r^* \leqslant nR_r^*\left(144\sqrt{\frac{0.5V\log n}{nR_r^*}} + 1\right) \\ &\leqslant nR_r^*\left(144\sqrt{\frac{0.5V\log n}{1300V\log n}} + 1\right) < 6nR_r^*. \end{aligned}$$

Thus, $\frac{ntR_r^*}{2v} > \frac{t}{12}$, so

$$\log\left(1 + 2\log\left(1 + \frac{ntR_r^*}{2v}\right)\right) \geqslant \log\left(1 + 2\log\left(1 + \frac{t}{12}\right)\right) \geqslant \frac{t}{50}, \tag{A.1}$$

using the fact that $t \leqslant 1$.

Now suppose $nR_r^* \geqslant C_t \frac{V}{2}\log n$ for the constant $C_t = \left(\frac{144}{t}\right)^2$. Note that for $t \leqslant 1$, we have $nR_r^* \geqslant 1300V\log n$, so all the previous results are also valid. Moreover, we have

$$\begin{aligned} \frac{\mathbb{E}\,Z}{0.5tR_r^*} = \frac{n\,\mathbb{E}\,Z}{0.5tnR_r^*} &\leqslant \frac{72\sqrt{0.5VnR_r^*\log n}}{0.5tnR_r^*} = \frac{144\sqrt{0.5V\log n}}{t\sqrt{nR_r^*}} \\ &\leqslant \frac{144\sqrt{0.5V\log n}}{t\sqrt{0.5C_tV\log n}} = \frac{144}{t\sqrt{C_t}} < 1. \end{aligned}$$

Now we have all the ingredients required for the application of Theorem 12.9:

$$\begin{aligned} \mathbb{P}\{Z \geqslant tR_r^*\} &\leqslant \mathbb{P}\{Z \geqslant \mathbb{E}\,Z + 0.5tR_r^*\} \\ &\leqslant \exp\left(-\frac{ntR_r^*}{4}\log\left(1 + 2\log\left(1 + \frac{ntR_r^*}{2v}\right)\right)\right) \\ &\leqslant \exp\left(-\frac{1}{200}nt^2R_r^*\right), \end{aligned}$$

where the last inequality follows by inequality (A1).

An identical argument can be used to upper bound the quantity

$$\sup_{f\in\mathscr{H}_r}\left(R(f) - R_n(f)\right),$$

concluding the proof.

A.3 *Proof of Lemma 6.1*

We begin by proving inequality (6.3). First consider the following peeling lemma, an adaptation of Raskutti *et al.* [36, Lemma 3]:

LEMMA A.2 Let $A \subseteq \mathbb{R}^p$, and suppose $\{Y_x\}_{x\in A}$ is a collection of random variables indexed by x. Also suppose $g : \mathbb{R} \to \mathbb{R}_+$ is a strictly increasing function such that $\inf_{x\in A} g(h(\|x\|_2)) \geqslant \mu$, for some $\mu > 0$,

and $h : \mathbb{R}_+ \to \mathbb{R}_+$ is a constraint function, and the tail bound

$$\mathbb{P}\left(\sup_{x \in A : h(\|x\|_2) \leqslant s} Y_x \geqslant g(s) \right) \leqslant 2 \exp\big(- cg(s) \big)$$

holds for all $s \in \text{range}(h)$. Then

$$\mathbb{P}\Big(Y_x \leqslant 2g(h(\|x\|_2)), \quad \forall x \in A \Big) \geqslant 1 - \frac{2\exp(-c\mu)}{1 - \exp(-c\mu)}. \tag{A.2}$$

Proof. We define the sets

$$A_m := \left\{ x \in A : 2^{m-1}\mu \leqslant g(h(\|x\|_2)) \leqslant 2^m \mu \right\},$$

for $m \geqslant 1$. By a union bound, we have

$$\mathbb{P}\Big(\exists x \in A \text{ s.t.} Y_x > 2g(h(\|x\|_2)) \Big) \leqslant \sum_{m=1}^{M} \mathbb{P}\Big(\exists x \in A_m \text{ s.t.} Y_x > 2g(h(\|x\|_2)) \Big),$$

where $M = \sup_{m \geqslant 1} g^{-1}(2^{m-1}\mu) \in \text{range}(h)$.

Further note that if $x \in A_m$ satisfies $Y_x > 2g(h(\|x\|_2))$, then $g(h(\|x\|_2)) \geqslant 2^{m-1}\mu$, so

$$\begin{aligned}
\mathbb{P}\left(\sup_{x \in A_m} Y_x > 2g(h(\|x\|_2)) \right) &\leqslant \mathbb{P}\left(\sup_{x \in A_m} Y_x > 2 \cdot 2^{m-1}\mu \right) \\
&\leqslant \mathbb{P}\left(\sup_{x \in A : g(h(\|x\|_2)) \leqslant 2^m \mu} Y_x > 2^m \mu \right) \\
&= \mathbb{P}\left(\sup_{x \in A : h(\|x\|_2) \leqslant g^{-1}(2^m \mu)} Y_x > 2^m \mu \right) \\
&\leqslant 2\exp\left(-c \cdot 2^m \mu \right),
\end{aligned}$$

if $m < M$. If $m = M$, the same logic shows that

$$\mathbb{P}\left(\sup_{x \in A_m} Y_x > 2g(h(\|x\|_2)) \right) \leqslant \mathbb{P}\left(\sup_{x \in A : h(\|x\|_2) \leqslant \nu} Y_x > 2^m \mu \right),$$

where $\nu = \sup_{x \in A} h(\|x\|_2)$. Furthermore, $2^{m-1}\mu \leqslant g(\nu) \leqslant 2^m \mu$, so the last probability is upper-bounded by

$$\mathbb{P}\left(\sup_{x \in A : h(\|x\|_2) \leqslant \nu} Y_x \geqslant g(\nu) \right) \leqslant 2\exp(-cg(\nu)) \leqslant 2\exp(-c \cdot 2^{m-1}\mu).$$

It follows that

$$\mathbb{P}\left(\sup_{x \in A_m} Y_x > 2g(h(\|x\|_2)) \right) \leqslant 2\exp(-c \cdot 2^{m-1}\mu),$$

for all $m \geqslant 1$, so summing up over m then gives

$$\mathbb{P}\Big(\exists x \in A \text{ s.t.} Y_x > 2g(h(\|x\|_2))\Big) \leqslant \sum_{m=1}^{\infty} 2\exp(-c \cdot 2^{m-1}\mu) \leqslant \frac{2\exp(-c\mu)}{1-\exp(-c\mu)},$$

implying inequality (A2). □

We apply Lemma A.2 with $A = \{x : \|x\|_2 \leqslant \bar{r}\}$, and

$$Y_x = |R_n(f_{x,r}) - R(f_{x,r})|, \qquad h(\|x\|_2) = R(f_{x,r}), \qquad g(s) = ts,$$

for fixed values of $\bar{r}, r > 0$ and $t \in (0, 1]$. Clearly, g is monotonically increasing and satisfies $\inf_{x\in A} g(h(\|x\|_2)) \geqslant tR(f_{\bar{r},r})$. Note that for any $s \in \text{range}(h)$, we have $s = R(f_{x_s,r})$ for some x_s, and

$$\begin{aligned}\mathbb{P}\left(\sup_{x\in A: h(\|x\|_2)\leqslant s} |R_n(f_{x,r}) - R(f_{x,r})| \geqslant g(s)\right) &= \mathbb{P}\left(\sup_{\|x_s\|_2 \leqslant \|x\|_2 \leqslant \bar{r}} |R_n(f_{x,r}) - R(f_{x,r})| \geqslant tR(f_{x_s,r})\right) \\ &\leqslant 2\exp(-cnR(f_{x_s,r})t^2) \\ &= 2\exp(-cntg(s)),\end{aligned}$$

assuming $nR(f_{\bar{r},r}) \geqslant C_t d \log n$, where we use a slight modification of Lemma 2.2 where $\mathscr{H}_r$ is the set of balls centered around points in $\{\|x\|_2 \geqslant \|x_s\|_2\}$. Lemma A.2 then implies the desired concentration inequality.

To establish inequality (6.4), note that we can simply use a modification of Theorem 2.2, where $\mathscr{H}_r$ is now the set of balls centered around points in $\{\|x\|_2 > \bar{r}\}$.

B. Proofs for univariate estimators

We begin with the following lemma, also appearing as Pensia *et al.* [35, Lemma 1].

LEMMA B.1 We have the following properties:

(i) For any $r > 0$ and $x, x' \in \mathbb{R}$, if $|x| < |x'|$, then $R(f_{x,r}) \geqslant R(f_{x',r})$.

(ii) For any $x \in \mathbb{R}$, if $r < r'$, then $R(f_{x,r}) \leqslant R(f_{x,r'})$.

(iii) If $0 < r < r'$, then $\frac{R^*_r}{r} > \frac{R^*_{r'}}{r'}$.

(iv) If $0 < r, r'$, then $R(f_{r',r}) < \frac{r}{r'} R^*_{r'}$.

(v) If $1 \leqslant k \leqslant n$, then $\frac{k}{n} < R^*_{q_{(2k)}}$ and $\frac{k}{n} < R^*_{2\sigma_{(2k)}}$.

Proof. The proofs proceed using simple calculus and algebraic manipulations, relying only on the properties of symmetry and unimodality.

(i) Property (i) follows directly by unimodality and symmetry of $\overline{P}$.

(ii) Property (ii) is true by the non-negativity of density.

(iii) Let $p(x)$ be the density of $\overline{P}$. Then $R^*_x = 2\int_0^x p(y)\,dy$. Define $g(x) := \frac{R^*_x}{x}$ for $x > 0$. Property (iii) is equivalent to showing that $\frac{d}{dx} g(x) < 0$. By unimodality of $p(\cdot)$, we have $g(x) > 2p(x)$

for $x > 0$. By differentiation, we have

$$\frac{d}{dx}g(x) = \frac{2xp(x) - 2\int_0^x p(y)dy}{x^2} = \frac{2p(x) - g(x)}{x} < 0,$$

as wanted.

(iv) Note that r' can be written as $r' = (K + \alpha)r$, where $K \in \mathbb{N}$ and $\alpha \in [0, 1)$. We need to show that $R^*_{r'} > (K + \alpha)R(f_{r',r})$. We may write

$$\begin{aligned} R^*_{r'} &= 2\int_0^{r'} p(x)\,\mathrm{d}x \\ &= 2\int_0^{\alpha r} p(x)\,\mathrm{d}x + \sum_{k=1}^{K} 2\int_{r'-kr}^{r'-(k-1)r} p(x)\,\mathrm{d}x. \end{aligned}$$

The second term is 0 if $K = 0$. By (iii) above, we have $R^*_{\alpha r} > \alpha R^*_r$. Therefore,

$$\begin{aligned} R^*_{r'} &> 2\alpha\int_0^{r} p(x)\,\mathrm{d}x + \sum_{k=1}^{K} 2\int_{r'-kr}^{r'-(k-1)r} p(x)\,\mathrm{d}x \\ &> \alpha\int_{r'-r}^{r'+r} p(x)\,\mathrm{d}x + \sum_{k=1}^{K}\int_{r'-r}^{r'+r} p(x)\,\mathrm{d}x \\ &= (\alpha + K)R(f_{r',r}), \end{aligned}$$

where the last inequality again uses unimodality of $\overline{P}$ and the second term is 0 if $K = 0$.

(v) Note that

$$R^*_{q_{(2k)}} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(|X_i| \leqslant q_{(2k)}) > \frac{1}{2}\cdot\frac{2k}{n} = \frac{k}{n}.$$

Let $\tilde{R}_i(f)$ be the expectation of f under P_i, i.e., $\tilde{R}_i(f) = \mathbb{E}f(X_i)$. For the second inequality, note that by Chebyshev's inequality,

$$\tilde{R}_i(f_{0,2\sigma_i}) = \mathbb{P}(|X_i - \mu| \leqslant 2\sigma_i) \geqslant \frac{3}{4},$$

for all i. Therefore, an interval of length $4\sigma_{(2k)}$ covers at least $\frac{3}{4}$ mass of at least $2k$ distributions, implying that

$$R^*_{2\sigma_{(2k)}} = R(f_{0,2\sigma_{(2k)}}) = \frac{1}{n}\sum_{i=1}^{n}\tilde{R}_i(f) \geqslant \frac{1}{n}\cdot\frac{3\times 2k}{4} > \frac{k}{n}.$$

□

Lemma B.1 shows that we can use $\overline{P}$ as a measure of distance between two intervals. In particular, if two intervals with the same center/radius are close under R, the respective radii/centers must also be close.

B.1 *Proof of Theorem 3.1*

We begin with the following result, which follows from Lemma 2.2:

LEMMA B.2 Let $t \in (0,1]$, and let r be such that $R_r^* \geqslant C_{0.5t}\left(\frac{\log n}{n}\right)$. Then with probability at least $1 - 2\exp(-c'nR_r^*t^2)$, we have $R(f_{\widehat{\mu}_{M,r},r}) \geqslant (1-t)R_r^*$.

Proof. This will follow from Lemma 2.2 by choosing $0.5t$ instead of t. If $R_r^* \geqslant C_{0.5t}\frac{\log n}{n}$, then with probability $1 - 2\exp(-cnR_r^*t^2/4)$, we have

$$|R_n(f) - R(f)| \leqslant \frac{tR_r^*}{2},$$

uniformly over $f \in \mathscr{H}_r$. Assume that this event happens. Note that $R(f_{0,r}) = R_r^*$ and $R_n(f_{\widehat{\mu}_{M,r},r}) \geqslant R_n(f_{0,r})$ by maximality of the modal interval estimator. Since $f_{\widehat{\mu}_{M,r},r}, f_{0,r} \in \mathscr{H}_r$, we have

$$\begin{aligned} R(f_{\widehat{\mu}_{M,r},r}) &\geqslant R_n(f_{\widehat{\mu}_{M,r},r}) - \frac{tR_r^*}{2} \geqslant R_n(f_{0,r}) - \frac{tR_r^*}{2} \\ &\geqslant R(f_{0,r}) - tR_r^* = R_r^* - tR_r^*, \end{aligned}$$

as wanted. □

Lemma B.2 states that if r is small, then $R(f_{x,r})$ behaves like a (scaled) density of the mixture distribution $\overline{P}$. Indeed, the density of $\overline{P}$ at the empirical mode, $\widehat{\mu}_{M,r}$, is within a constant factor of the density at μ^*.

Turning to the proof of the theorem, note that by Lemma B.1(i), we know that if $R(f_{r',r}) < R(f_{\widehat{\mu}_{M,r},r})$, then $|\widehat{\mu}_{M,r}| \leqslant r'$. Furthermore, taking $t = \frac{1}{2}$ in Lemma B.2, we have $R(f_{\widehat{\mu}_{M,r},r}) \geqslant \frac{R_r^*}{2}$, with probability at least $1 - 2\exp(-c'nR_r^*/4)$. Thus, inequality (3.1) holds provided $R(f_{r',r}) < \frac{R_r^*}{2}$.

Now suppose Let $r' = \frac{2r}{R_r^*}$. By Lemma B.1(iv) and noting that $R_{r'}^* \leqslant 1$, we have

$$R(f_{r',r}) < \frac{r}{r'}R_{r'}^* \leqslant \frac{r}{r'} = \frac{r}{\frac{2r}{R_r^*}} = \frac{R_r^*}{2}.$$

This establishes inequality (3.2).

B.2 *Proof of Theorem 3.2*

The proof of Theorem 3.2 is similar in spirit to the proof of Theorem 3.1. We begin by proving a lemma, which replaces Lemma B.2:

LEMMA B.3 For $2k \geqslant C_{0.5t}\log n$ and $t \in (0,1]$, with probability at least $1 - 2\exp(-c'kt^2)$, we have

$$R(f_{\widehat{\mu}_{S,k},r_{2k}}) \geqslant (1-t)R_{r_k}^* = (1-t)\frac{k}{n}.$$

Proof. By assumption, we have $nR_{r_{2k}}^* = 2k \geqslant C_{0.5t}\log n$. Applying Lemma 2.2 with $t = 0.5t$ and $r = r_{2k}$, we know that with probability at least $1 - \exp(-c2kt^2/4)$, we have

$$\sup_{x, r \leqslant r_{2k}} R_n(f_{x,r}) - R(f_{x,r}) < \frac{t}{2}R_{r_{2k}}^*.$$

Combined with the guarantee of Lemma 4.1, we conclude that

$$R_n(f_{\widehat{\mu}_{S,k},\widehat{r}_k}) - R(f_{\widehat{\mu}_{S,k},\widehat{r}_k}) < \frac{t}{2} R^*_{r_{2k}},$$

with probability at least $1 - \exp(-ckt^2/2) - \exp(-k/8)$.

Furthermore, since all the distributions have densities, all the X_is are distinct with probability 1, so $R_n(f_{\widehat{\mu}_{S,k},\widehat{r}_k}) = \frac{k}{n}$. We thus conclude that

$$\frac{k}{n} - R(f_{\widehat{\mu}_{S,k},\widehat{r}_k}) < \frac{t}{2} \cdot \frac{2k}{n},$$

so $R(f_{\widehat{\mu}_{S,k},\widehat{r}_k}) > (1-t)\frac{k}{n} = (1-t)R^*_{r_k}$. Again using the fact that $\widehat{r}_k \leqslant r_{2k}$, we can use Lemma B.1(ii) to conclude that $R(f_{\widehat{\mu}_{S,k},\widehat{r}_k}) \leqslant R(f_{\widehat{\mu}_{S,k},r_{2k}})$, so the required statement holds. □

Let $r' = \frac{2nr_{2k}}{k}$. Taking $t = \frac{1}{2}$ in Lemma B.3 and using Lemma B.1(i), it suffices to show that $R(f_{r',r_{2k}}) < \frac{k}{2n}$, which follows by Lemma B.1(iv) and the fact that $R^*_{r'} \leqslant 1$.

B.3 *Proof of Theorem 3.3*

We first prove the following result:

LEMMA B.4 With probability at least $1 - 4\exp(-ck^2/n)$, both of the following statements hold:

1. S_k contains the origin in the sense that $0 \in [\min(S_k), \max(S_k)]$.
2. $\mathrm{Diam}(S_k) \leqslant 2r_{2k}$

Proof. The k-median was defined using ψ_n. It is therefore instructive to study the properties of the population-level quantity $\psi(\theta) := \mathbb{E}\,\psi_n(\theta)$. For $\theta > 0$, we have

$$\psi(\theta) := \mathbb{E}\,\psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathrm{sign}(\theta - X_i)]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(-\theta \leqslant X_i < \theta) = R(f_{0,\theta}) = R^*_\theta.$$

In particular, $\psi(r_k) = R^*_{r_k} = \frac{k}{n}$. Similarly, for $\theta < 0$, we have $\psi(\theta) = -R^*_{|\theta|}$.

1. It suffices to show the events $\widehat{\theta}_{\mathrm{med},k} \leqslant r_{2k}$ and $\widehat{\theta}_{\mathrm{med},-k} \geqslant -r_{2k}$ hold with the required probability. We will focus only on the error on the positive side, i.e., $\widehat{\theta}_{\mathrm{med},k} > r_{2k}$. The analysis for $\widehat{\theta}_{\mathrm{med},-k} < -r_{2k}$ is similar by symmetry. Recall that $\psi_n(\widehat{\theta}_{\mathrm{med},k}) = \frac{k}{n}$ a.s., so by monotonicity of ψ_n, it follows that

$$\mathbb{P}\left(\widehat{\theta}_{\mathrm{med},k} > r_{2k}\right) \leqslant \mathbb{P}\left(\psi_n(r_{2k}) \leqslant \frac{k}{n}\right)$$

$$= \mathbb{P}\left(\psi_n(r_{2k}) - \psi(r_{2k}) \leqslant -\frac{k}{n}\right).$$

Since $\psi_n(\cdot) - \psi(\cdot)$ is a centered sum of independent bounded random variables, we may apply Hoeffding's inequality on its negative tail. Therefore,

$$\mathbb{P}\left(\widehat{\theta}_{\text{med},k} > r_{2k}\right) \leqslant \exp\left(-cn\left(\frac{k}{n}\right)^2\right) \leqslant \exp(-ck^2/n).$$

2. We bound the probability that $\max(S_k) < 0$; the bound for $\min(S_k) > 0$ is analogous. If $\max(S_k) < 0$, then $\psi_n(0) \geqslant \frac{k}{n}$ by monotonicity of ψ_n and the fact that $\max(S_k) = \widehat{\theta}_{\text{med},k}$ and $\psi_n(\widehat{\theta}_{\text{med},k}) = \frac{k}{n}$. By Hoeffding's inequality, we then have

$$\begin{aligned}\mathbb{P}\left(\max(S_k) < 0\right) &\leqslant \mathbb{P}\left(\psi_n(0) \geqslant \frac{k}{n}\right) = \mathbb{P}\left(\psi_n(0) - \psi(0) \geqslant \frac{k}{n}\right) \\ &\leqslant \exp\left(-cn \cdot \frac{k^2}{n^2}\right) = \exp\left(-c\frac{k^2}{n}\right).\end{aligned}$$

□

By Lemma B.4 and Theorem 3.2, with probability with probability at least $1 - 4\exp(-c\log^2 n)$, both of the following events happen simultaneously:

1. $0 \in [\min(S_k), \max(S_k)]$.
2. $\text{Diam}(S_k) \leqslant 2r_{k_1}$.

As the set $[\min(S_k), \max(S_k)]$ is convex and 0 belongs to the set, $|\widehat{\mu}_{k_1,k_2}| \leqslant |\widehat{\mu}_{S,k_2}|$. As $\widehat{\mu}_{k_1,k_2} \in [\min(S_k), \max(S_k)]$, $|\widehat{\mu}_{k_1,k_2}|$ is less than the diameter of S_k. This proves the first inequality of the statement.

Let $r' := \frac{4\sqrt{n}\log n}{k_2} r_{2k_2}$. To prove the second inequality, we break down the analysis in two cases:

Case 1: Suppose $R^*_{r'} \geqslant \frac{2\log n}{\sqrt{n}}$. This implies that $r_{2k_1} \leqslant r'$ and thus desired holds. Since the final prediction is always within the set spanned by S_{k_1}, we must have $|\widehat{\mu}_{k_1,k_2}| \leqslant r'$ with probability at least $1 - 4\exp(-c\log^2 n)$.

Case 2: Suppose $R^*_{r'} < \frac{2\log n}{\sqrt{n}}$. We will first show that $|\widehat{\mu}_{S,k_2}| \leqslant r'$. Similar to the proof of Theorem 3.2, it suffices to show that $R(f_{r',r_{2k_2}}) < \frac{k_2}{2n}$. Indeed, we have by Lemma B.1(iv)

$$R(f_{r',r_{2k_2}}) < \frac{r_{2k_2}}{r'} R^*_{r'} < \frac{1}{\frac{4\sqrt{n}\log n}{k_2}} \frac{2\log n}{\sqrt{n}} = \frac{k_2}{2n},$$

with probability at least $1 - 2\exp(-c'k_2)$.

Altogether, we conclude that $|\widehat{\mu}_{k_1,k_2}| \leqslant r'$, with probability at least $1 - 2\exp(-c'k_2) - 4\exp(-c\log^2 n)$.

C. Proofs for examples

In this appendix, we provide the proofs for the propositions regarding the examples discussed in Section 3.1.

C.1 *Proof of Proposition 3.7*

Using the symmetry and unimodality of $\overline{p}$, we have the following relation:

$$2\overline{p}(0)r \geqslant R(f_{[0,r]}) \geqslant 2r\overline{p}(r).$$

Using the first inequality above and choosing $r = r_k$, we obtain $r_k \geqslant \frac{k}{2\overline{p}(0)}$. The second inequality implies that if $2\overline{p}(y)y \geqslant \frac{k}{n}$, then $r_k \leqslant y$. In the remainder of the proof, we will show the bounds for each example using this approach:

1. The lower bound follows by noting that the density at 0 is $\frac{1}{\sqrt{2\pi}\sigma}$. As a result, $r_{\log n} \geqslant \frac{\sqrt{2\pi}\sigma \log n}{2n}$. The upper bound follows by noting that density at $x = |\sigma|$ is within constant factor of the density at 0. Let $r = (\sigma\sqrt{2\pi e}\log n)/n$. For large enough n, we have that $r \leqslant \sigma$. Thus,

$$2\overline{p}(r)r \geqslant 2\overline{p}(\sigma)r = 2\frac{e^{-1/2}}{\sqrt{2\pi}\sigma}\frac{\sigma\sqrt{2\pi e}\log n}{n} = \frac{2\log n}{n}.$$

   Therefore, $r_k \leqslant (\sigma\sqrt{2\pi e}\log n)/n$.

2. The lower bound follows by noting that the density at $x = 0$ is

$$\overline{p}(0) = \left(\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}cin}\right) = \Theta\left(\frac{\log n}{cn}\right),$$

   where we use that $\log n \leqslant \sum_{i=1}^{n} i^{-1} \leqslant (\log n + 1)$. Thus, $r_{\log n} \geqslant \frac{\log n}{2n\overline{p}(0)} = \Theta(1)$. The upper bound follows by noting that the density at $x = 1$ is

$$\overline{p}(1) = \left(\sum_{i=1}^{n}\frac{e^{-\frac{1}{i^2c^2}}}{\sqrt{2\pi}cin}\right) \geqslant \left(\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}cin}\left(1-\frac{1}{i^2c^2}\right)\right) = \overline{p}(0) - \frac{1}{\sqrt{2\pi}c^3n}\sum_{i=1}^{n}\frac{1}{i^3},$$

   where the inequality uses that for all $x \in \mathbb{R}$, $e^x \geqslant 1 + x$. As $\sum_{i=1}^{n} i^3$ converges, we let $C = \lim_n \sum_{i=1}^{n}\frac{1}{i^3}$. We thus have that

$$2\overline{p(1)}1 \geqslant 2\left(\overline{p}(0) - \frac{C}{c^3n}\right) \geqslant \frac{2\log n}{\sqrt{2\pi}n}\left(\frac{1}{c} - \frac{C}{c^3\log n}\right).$$

   This last expression is greater than $(\log n)/n$, when c is less than (say) $\sqrt{1/2\pi}$ and n is large enough such that $\log n > 2C/c^2$.

3. We first consider the case $\alpha \geqslant 1$. The lower bound follows by noting that the density at 0 is

$$\frac{c\log n}{n}\frac{1}{\sqrt{2\pi}} + \frac{n - c\log n}{n}\frac{1}{n^\alpha} = \Theta\left(\frac{\log n}{n}\right).$$

   The upper bound follows from the fact that at least $c\log n$ distributions have variance 1. Thus, the interval $[-1, 1]$ contains more than 0.6 probability of at least $c\log n$ distributions. As $R(f_{0,1}) \geqslant 0.6c(\log n)/n$, which is larger than $(\log n)/n$ for $c \geqslant 5/3$, implying that $r_{\log n} \leqslant 1$.

4. We now consider the case when $\alpha < 1$. The density at 0 is

$$\frac{c\log n}{n}\frac{1}{\sqrt{2\pi}} + \frac{n - c\log n}{n}\frac{1}{n^\alpha} = \Theta\left(\frac{1}{n^\alpha}\right),$$

which implies the desired upper bound. For the desired lower bound, we note that the density at $x = 1$ is also $\Theta\left(\frac{1}{n^\alpha}\right)$. Using a similar calculation to that of Example 1 above, we get the desired upper bound on r_k.

C.2 *Proof of Proposition 3.8*

Since $r = r_{C\log n}$, we have $R^*_r = \frac{C\log n}{n}$. By inequality (3.2) of Theorem 3.1, we have

$$|\widehat{\mu}_{M,r}| \leqslant \frac{2nr_{C\log n}}{C\log n}, \tag{C.1}$$

w.h.p.

1. Analogously to Proposition 3.7, we have $r_{C\log n} = \Theta\left(\frac{C\sigma\log n}{n}\right)$. Inequality (C.1) then gives the result.
2. The bound of $\tilde{O}(n)$ follows by inequality (C.1) and noting that $r_{C\log n} = O(1)$ for a fixed C and sufficiently small $c > 0$. We now focus on how to obtain the tighter bound of $O(n^\epsilon)$ for an $\epsilon > 0$, using inequality (3.1).
3. Let $\tilde{R}_i(f)$ be the expectation of f under P_i, i.e., $\tilde{R}_i(f) = \mathbb{E}f(X_i)$. Fix an $\epsilon > 0$. Let $r' = n^\epsilon$ and $r = 1$. Then it suffices to show that $R^*_r - R(f_{r',r}) \geqslant C'R^*_r$ where $C' > 0$ might depend on ϵ but not on n.
4. We will show that
   a. $R^*_r - R(f_{r',r}) \geqslant c_1 \sum_{i\leqslant \frac{r'}{10c}} \tilde{R}_i(f_{0,r})$,
   b. $\sum_{i\leqslant \frac{r'}{5c}} \tilde{R}_i(f_{0,r}) \geqslant c_2 nR^*_r$.

5. To derive the first inequality, note that

$$\begin{aligned} nR_r^* - nR(f_{r',r}) &\geqslant \sum_{i \leqslant \frac{r'}{10c}} R(f_{0,1}) - R(f_{r',1}) \\ &\geqslant \sum_{i \leqslant \frac{r'}{10c}} 2 \int_0^1 \frac{1}{\sqrt{2\pi}ci} \left(e^{-\frac{x^2}{2c^2i^2}} - e^{-\frac{(0.5r'+x)^2}{2c^2i^2}} \right) \mathrm{d}x \\ &\geqslant \sum_{i \leqslant \frac{r'}{10c}} 2 \int_0^1 \frac{(1 - e^{-\frac{0.25r'^2}{2c^2i^2}})}{\sqrt{2\pi}ci} e^{-\frac{x^2}{2c^2i^2}} \, \mathrm{d}x \\ &\geqslant (1 - e^{-10}) \sum_{i \leqslant \frac{r'}{10c}} \tilde{R}_i(f_{0,r}). \end{aligned}$$

Now it remains to show that $\sum_{i \leqslant \frac{r'}{10c}} \tilde{R}_i(f_{0,r}) \geqslant c_2 R_r^*$. First note that $nR_r^* \leqslant \frac{\log n}{c}$. Hence,

$$\sum_{i \leqslant \frac{r'}{10c}} \tilde{R}_i(f_{0,1}) \geqslant \sum_{i: \frac{1}{c} < i \leqslant \frac{r'}{10c}} \tilde{R}_i(f_{0,1}) \geqslant \sum_{i: \frac{1}{c} < i \leqslant \frac{r'}{10c}} \frac{2e^{-0.5}}{\sqrt{2\pi}ci} \geqslant c_3 \log\left(\frac{r'}{10e}\right) \geqslant c_4 \log n^{\epsilon} \geqslant c_5 \epsilon n R_r^*.$$

6. For $\alpha < 1$, let $r' = \Theta(n^{\alpha})$. Then it is easy that $R(f_{r',r}) \leqslant \frac{R_r^*}{2}$. This follows by observing that the density of a Gaussian distribution decreases by more than half at a distance of σ from the mean.

7. For $\alpha \geqslant 1$, let $r' = 10$. Then $R_r^* \geqslant 0.5 \frac{C \log n}{n}$, as a Gaussian distribution contains about 0.68 mass within 1 standard deviation of the mean. Moreover,

$$R(f_{r',r}) \leqslant 0.1 \frac{C \log n}{n} + \frac{n}{\sqrt{2\pi} n^{\alpha}} \leqslant 0.2 \frac{C \log n}{n} \leqslant \frac{R_r^*}{2}.$$

Inequality (3.1) then implies the result.

C.3 *Proof of Proposition 3.9*

In the following, we will show the bounds on $r_{2\sqrt{n}\log n}$, which gives us the result:

1. As in the proof of Proposition 3.7, we have $r_k = \Theta\left(\frac{\sigma k}{n}\right)$ for small k.
2. By Lemma B.1(i), we have $r_{2\sqrt{n}\log n} \leqslant 2\sigma_{(4\sqrt{n}\log n)} = O\left(\sqrt{n}\log n\right)$.
3. Note that for any fixed k, the value of r_k for Example 3.6 is smaller than the value of r_k for Example 3.4 with $\sigma = n^{\alpha}$. Thus, we have $r_{2\sqrt{n}\log n} = O\left(\frac{n^{\alpha}\sqrt{n}\log n}{n}\right) = O\left(n^{\alpha - 0.5}\log n\right)$.

C.4 *Proof of Proposition 3.12*

We first provide the main steps of the proof. Proofs of supporting lemmas are contained in further sub-sections.

C.4.1 *Main argument*

Proof. (Proof of Proposition 3.12) Let W be a generic random variable with distribution Q_n as defined in Example 3.11. Let $A = [-2, 2]$. Consider two disjoint set of hypothesis classes $\mathscr{K}$ and $\mathscr{J}$, with $\mathscr{K} = \{f_{x,1} : x \in A\}$ and $\mathscr{J} = \{f_{x,1} : x \notin A\}$. The hypothesis class $\mathscr{J}$ contains the intervals that are far from 0. Define the following random variables:

$$Z_1 = \sup_{f \in \mathscr{K}} R_n(f), \qquad Z_2 = \sup_{f \in \mathscr{J}} R_n(f).$$

We would show that with constant non-zero probability: (i) $Z_1 < Z_2$ and (ii) the maximum is achieved in Z_2 at intervals that are far from 0.

Note that $R_1^* = \sup_{f \in \mathscr{K}} R(f) = \Theta\left(n^{-\alpha}\right)$. Define $R^*_{\mathscr{J}} \stackrel{\text{def}}{=} \sup_{f \in \mathscr{J}} R(f)$. Note that supremum is achieved in both the cases and $R^*_{\mathscr{J}} < R_1^*$. Moreover, we have the following straightforward relations:

1. $2R_1^* \geqslant \mathbb{P}(W \in A) \geqslant R_1^*$.
2. $nR^*_{\mathscr{J}} = \Theta\left(n^{1-\alpha}\right)$.
3. $\mathbb{P}(W \in A)\sqrt{nR^*_{\mathscr{J}}} = O(1)$.
4. For every constant C', there exists another constant $C > 0$ such that

$$R^*_{\mathscr{J}} + C\left(\sqrt{\frac{R^*_{\mathscr{J}}}{n}}\right) \geqslant R_1^* + C'\left(\sqrt{\frac{R_1^*}{n}}\right).$$

These relations suffice for showing that $Z_1 < Z_2$ with constant probability. To this end, we would show that with constant probability both (1) $Z_1 = R_1^* + O\left(\sqrt{\frac{R_1^*}{n}}\right)$ and (2) $Z_2 \geqslant R^*_{\mathscr{J}} + C\left(\sqrt{\frac{R^*_{\mathscr{J}}}{n}}\right)$, for any $C > 0$. Note that these events are dependent and thus we had use the following lemma, which shows that conditioned on the inclusion of points in each of two disjoint intervals, the distributions of the histograms on each of the intervals behave independently:

Lemma C.1 Let $\{x_1, \ldots, x_n\}$ be i.i.d. draws from a distribution with density p_i. Consider two disjoint intervals A and B. For any two disjoint subsets $S, T \subseteq \{1, \ldots, n\}$, we use x_S to denote the vector $(x_i : i \in S)$, and we define x_T similarly. Let E denote the event that $x_i \in A$ for all $i \in S$, and $x_i \in B$ for all $\in T$. Then for $x_S \subseteq A$ and $x_T \subseteq B$, we have

$$p_{S,T}(x_S, x_T \mid E) = p_S(x_S \mid E) p_T(x_T \mid E).$$

Furthermore,

$$p_S(x_S \mid E) = \prod_{i \in S} \frac{p_i(x_i)}{\mathbb{P}(X_i \in A)}, \quad \text{and}$$

$$p_T(x_T \mid E) = \prod_{i \in T} \frac{p_i(x_i)}{\mathbb{P}(X_i \in B)}$$

are the joint densities of independent draws from the renormalized distributions of the points lying in each interval.

Let $S \subset \{1, \ldots, n\}$ be an index set. For a fixed index set S, let the event E_S be $E_S = \{X_S \subset A, X_{S^c} \subset A^c\}$, where X_S is the vector $(X_i : i \in S)$ and A is defined above.

Conditioned on E_S, Lemma C.1 states that X_is are independent. Thus, conditioned on E_S, the random variables Z_1 and Z_2 are independent.

LEMMA C.2 Consider the setting in Proposition 3.12. Let $S \subset [n]$ be such that $|S| \leqslant n\mathbb{P}(A)$. Conditioned on the event E_S, we have that for some $C' > 0$,

$$Z_1 \leqslant R_1^* + C'\sqrt{\frac{R_1^*}{n}}$$

with a constant non-zero probability.

LEMMA C.3 Consider the setting in Proposition 3.12. Let $S \subset [n]$ be such that $|S^c| \geqslant n\mathbb{P}(A^c)$. Conditioned on the event E_S, we have that for all $C > 0$,

$$Z_2 \geqslant R^*_{\mathscr{J}_n} + C\left(\sqrt{\frac{R^*_{\mathscr{J}_n}}{n}}\right)$$

with a constant, non-zero probability depending on the constant C.

LEMMA C.4 Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$, where P is a uniform distribution over $[-b, -a] \bigcup [a, b]$ for some $0 \leqslant a < b$. For a $r > 0$, let $Z = \sup_{f \in \mathscr{H}_r} R_n(f)$ and $k \in \mathbb{N}$ such that $E = \{Z = k\}$ is an event of non-zero probability. If $\frac{b-a}{r} > C$, then

1. $\mathbb{P}(|\widehat{\mu}_{M,r}| \geqslant \frac{b-a}{2}) \geqslant c > 0$.
2. $\mathbb{P}(|\widehat{\mu}_{M,r}| \geqslant \frac{b-a}{2} | Z \geqslant k) \geqslant c > 0$.

Lemmas C.2, C.3 and C.4 give us the required lower bound on the probability of error. Let $\widehat{\mu}_{M,1,\mathscr{J}} := \arg\max_{f \in \mathscr{J}} R_n(f)$. Clearly, we can write

$$\begin{aligned}
\mathbb{P}\left\{|\widehat{\mu}_{M,1}| \geqslant \frac{n^\alpha}{2}\right\} &= \mathbb{P}\left\{Z_1 < Z_2, |\widehat{\mu}_{M,1,\mathscr{J}}| \geqslant \frac{n^\alpha}{2}\right\} \\
&= \sum_{S \subset [n]} \mathbb{P}(E_S)\mathbb{P}\left(Z_1 \leqslant Z_2, |\widehat{\mu}_{M,1,\mathscr{J}}| \geqslant \frac{n^\alpha}{2} \middle| E_S\right) \\
&\geqslant \sum_{S \subset [n] : |S| \leqslant n\mathbb{P}(A)} \mathbb{P}(E_S)\mathbb{P}\left(Z_1 \leqslant nR_1^* + C\sqrt{nR^*}, Z_2 \geqslant nR_1^* + C\sqrt{nR_1^*}, |\widehat{\mu}_{M,1,\mathscr{J}}| \geqslant \frac{n^\alpha}{2} \middle| E_S\right).
\end{aligned}$$

Furthermore, note that since Z_1 is computed over the points lying in A and Z_2 and $\widehat{\mu}_{M,1,\mathscr{J}}$ is computed over the points lying in A^c, Lemma C.1 implies that

$$\begin{aligned}
&\mathbb{P}\left(Z_1 \leqslant nR_1^* + C\sqrt{nR_1^*}, Z_2 \geqslant nR_1^* + C\sqrt{nR_1^*}, |\widehat{\mu}_{M,1,\mathscr{J}}| \geqslant \frac{n^\alpha}{2}\middle| E_S\right) \\
&\quad = \mathbb{P}\left(Z_1 \leqslant nR_1^* + C\sqrt{nR_1^*}\middle| E_S\right)\mathbb{P}\left(Z_2 \geqslant nR_1^* + C\sqrt{nR_1^*}, |\widehat{\mu}_{M,1,\mathscr{J}}| \geqslant \frac{n^\alpha}{2}\middle| E_S\right) \\
&= \mathbb{P}\left(Z_1 \leqslant nR_1^* + C\sqrt{nR_1^*}\middle| E_S\right)\mathbb{P}\left(Z_2 \geqslant nR_1^* + C\sqrt{nR_1^*}\middle| E_S\right) \\
&\qquad \cdot \mathbb{P}\left(|\widehat{\mu}_{M,1,\mathscr{J}}| \geqslant \frac{n^\alpha}{2}\middle| Z_2 \geqslant nR_1^* + C\sqrt{nR_1^*}, E_S\right).
\end{aligned}$$

Finally, note that conditioned on E_S, the points in A^c are certainly still uniformly distributed by the construction. Hence, we can apply Lemmas C.2, C.3 and C.4 to lower bound each of the three factors by a constant. We conclude that

$$\mathbb{P}\left\{|\widehat{\mu}_{M,r}| \geqslant \frac{n^\alpha}{2}\right\} \geqslant \sum_{S\subset[n]:|S|\leqslant n\mathbb{P}(A)} \mathbb{P}(E_S)\Theta(1) = \Theta(1),$$

where the final equality uses the fact that for $X \sim \mathrm{Bin}(n,p)$, we have $\mathbb{P}(X \leqslant \mathbb{E}X) = \Theta(1)$. This concludes the proof of Proposition 3.12. □

C.4.2 *Proof of Lemma C.1*

Proof. (Proof of Lemma C.1) Clearly, we have

$$p_{S,T}(x_S, x_T \mid E) = \frac{p_{S,T}(x_S, x_T)}{\mathbb{P}(E)} = \frac{\prod_{i\in S} p_i(x_i)\prod_{j\in T} p_i(x_j)}{\mathbb{P}(E)}.$$

Similarly, we may write

$$\begin{aligned}
p_S(x_S \mid E) &= \frac{p_i(x_S)\prod_{j\in T}\mathbb{P}(X_j \in B)}{\mathbb{P}(E)}, \\
p_T(x_T \mid E) &= \frac{p_i(x_T)\prod_{i\in S}\mathbb{P}(X_i \in A)}{\mathbb{P}(E)}.
\end{aligned}$$

Using the fact that

$$\mathbb{P}(E) = \prod_{i\in S}\mathbb{P}(X_i \in A)\prod_{j\in T}\mathbb{P}(X_j \in B)$$

implies the desired statements. □

C.4.3 *Proof of Lemma C.2*

Proof. (Proof of Lemma C.2) Throughout the whole proof, we will condition on the set E_S. Conditioned on E_S, Lemma C.1 states that X_S is a vector of $|S|$ i.i.d. points with distribution, say, $Q_{n|A}$. Under $Q_{n|A}$, $\sup_{f\in\mathscr{K}} R(f) = \frac{R_1^*}{\mathbb{P}(W\in A)} \geqslant \frac{1}{2}$.

Using Theorem H.2 (Vershynin [45, Theorem 8.3.23]), we get that

$$\mathbb{E}\left[\left|\sup_{f\in\mathcal{K}}\sum_{i\in S}f(X_i)-\mathbb{E}[f(X_i)|E_S]\right|\right]\leqslant C\sqrt{|S|}\leqslant C\sqrt{2|S|\frac{R_1^*}{\mathbb{P}(W\in A)}},$$

where the last step uses that $2R_1^*\geqslant\mathbb{P}(W\in A)$. Thus, with constant positive probability,

$$\begin{aligned}Z_1=\sup_{f\in\mathcal{K}}\sum_i f(X_i)&\leqslant|S|\frac{R_1^*}{\mathbb{P}(A)}+C'\sqrt{|S|\frac{R_1^*}{\mathbb{P}(A)}}\\&\leqslant nR_1^*+C'\sqrt{nR_1^*},\end{aligned}$$

where we use Markov's inequality and the assumption that $|S|\leqslant n\mathbb{P}(A)$. □

C.4.4 *Proof of Lemma C.3*

Proof. (Proof of Lemma C.3) We will condition on the event E_S throughout the proof. Once we have conditioned on E_S, there are $|S^c|$ points distributed over A^c according to Lemma C.1, i.e., i.i.d. with a uniform distribution, say, $Q_{n|A^c}$. Consider a fixed function $f\in\mathcal{J}$. As the distribution is uniform, $R(f)=R^*_{\mathcal{J}}$.

For each $i\in S^c$, let $Y_i=f(X_i)-\frac{R^*_{\mathcal{J}}}{\mathbb{P}(A^c)}$. Y_is are centered i.i.d. Bernoulli random variables. We calculate the following quantities required for the Berry–Esseen theorem:

$$\begin{aligned}\mathbb{E}[Y_i]&=0\\\mathbb{V}[Y_i]&=\frac{R^*_{\mathcal{J}}}{\mathbb{P}(A^c)}\left(1-\frac{R^*_{\mathcal{J}}}{\mathbb{P}(A^c)}\right)\geqslant\frac{R^*_{\mathcal{J}}}{2\mathbb{P}(A^c)}\\\mathbb{E}\,|Y_i|^3&=\frac{R^*_{\mathcal{J}}}{\mathbb{P}(A^c)}\left|1-\frac{R^*_{\mathcal{J}}}{\mathbb{P}(A^c)}\right|^3+\left(1-\frac{R^*_{\mathcal{J}}}{\mathbb{P}(A^c)}\right)\left|\frac{R^*_{\mathcal{J}}}{\mathbb{P}(A^c)}\right|^3\\&\leqslant\frac{R^*_{\mathcal{J}}}{\mathbb{P}(A^c)}+\left(\frac{R^*_{\mathcal{J}}}{\mathbb{P}(A^c)}\right)^3\leqslant 2\frac{R^*_{\mathcal{J}}}{\mathbb{P}(A^c)}\end{aligned}$$

By the Berry–Esseen theorem [45], we have

$$\begin{aligned}\mathbb{P}\left\{\frac{\sum_{i\in S^c}Y_i}{\sqrt{|S^c|\mathbb{V}[Y_i]}}\geqslant t\right\}&\geqslant\phi(t)-\frac{\mathbb{E}\,|Y_i|^3}{\sqrt{\mathbb{V}[Y_i]^3|S^c|}}\geqslant\phi(t)-\frac{\frac{2R^*_{\mathcal{J}}}{\mathbb{P}(A^c)}}{\sqrt{\frac{\left(R^*_{\mathcal{J}}\right)^3}{8\mathbb{P}(A^c)^3}n\mathbb{P}(A^c)}}\\&\geqslant\phi(t)-\frac{c'}{\sqrt{nR^*_{\mathcal{J}}}}=\phi(t)-o_n(1),\end{aligned}$$

where $\phi(t) \stackrel{\text{def}}{=} \mathbb{P}(g \leqslant t)$ and $g \sim \mathcal{N}(0,1)$. Therefore,

$$\begin{aligned}
\mathbb{P}\left\{Z_2 \geqslant R^*_{\mathscr{J}} + C\left(\sqrt{\frac{R^*_{\mathscr{J}}}{n}}\right)\right\} &\geqslant \mathbb{P}\left\{\sum_{i\in S^c} f(X_i) \geqslant nR^*_{\mathscr{J}} + C\sqrt{nR^*_{\mathscr{J}}}\right\} \\
&= \mathbb{P}\left\{\sum_{i\in S^c} Y_i \geqslant |S|R^*_{\mathscr{J}} + C\sqrt{nR^*_{\mathscr{J}}}\right\} \\
&= \mathbb{P}\left\{\frac{1}{\sqrt{|S^c|\mathbb{V}[Y_i]}}\sum_{i\in S^c} Y_i \geqslant \frac{|S|R^*_{\mathscr{J}} + C\sqrt{nR^*_{\mathscr{J}}}}{\sqrt{|S^c|\mathbb{V}[Y_i]}}\right\} \\
&\geqslant \phi\left(\frac{|S|R^*_{\mathscr{J}} + C\sqrt{nR^*_{\mathscr{J}}}}{\sqrt{|S^c|\mathbb{V}[Y_i]}}\right) - o_n(1) \\
&\geqslant \phi\left(\frac{n\mathbb{P}(A)R^*_{\mathscr{J}} + C\sqrt{nR^*_{\mathscr{J}}}}{\sqrt{n\mathbb{P}(A^c)\frac{R^*_{\mathscr{J}}}{2\mathbb{P}(A^c)}}}\right) - o_n(1) \\
&\geqslant \phi\left(\mathbb{P}(A)\sqrt{nR^*_{\mathscr{J}}} + \sqrt{2}C\right) - o_n(1) \geqslant c\phi\left(C' + \sqrt{2}C\right)
\end{aligned}$$

where we use that for $\alpha \geqslant \frac{1}{3}$, $\mathbb{P}(A)\sqrt{nR^*_{\mathscr{J}}} = \Theta\left(n^{-\alpha+\frac{1-\alpha}{2}}\right) = O(1)$. □

C.4.5 *Proof of Lemma C.4*

Proof. (Proof of Lemma C.4) Let $\mathscr{H}$ be the set of intervals of width equal to $2r$. Currently, the intervals near the endpoints have less probability mass. We will replace such intervals with bigger intervals to make the process symmetric. First consider the intervals near $\pm a$ which have less probability mass: we can instead focus on bigger intervals to include the middle interval $[-a, a]$. Let $\mathscr{J} := \{\mathbb{1}_{[x,y]} : |x-y| = 2r + 2(b-a), |x+a| \leqslant 2r\}$. Next we can consider warping the number line and 'joining' the two endpoints, i.e., let $\mathscr{K} := \{\mathbb{1}_{[-\infty,x]\cup[y,\infty]} : 0 \leqslant b - y \leqslant 2r, 0 \leqslant x + b \leqslant 2r, y - x = 2b - 2r\}$.

Let $\mathscr{H}' := \mathscr{J} \cup \mathscr{K} \cup \mathscr{H} \setminus \{f \in \mathscr{H} : R(f) < \frac{2r}{2(b-a)}\}$ and $\widehat{\mu}'_{M,r} = \arg\max_{f\in\mathscr{H}'} R_n(f)$. Note that every function in $\mathscr{H}'$ contains equal mass and the distribution is uniform. Moreover, for $|x| \in [\frac{b-a}{2}, \frac{3(b-a)}{4}]$, $f_{x,r} \in \mathscr{H}' \cap \mathscr{H}$ because $b - a \geqslant Cr$. Thus, we have not removed a lot of functions from $\mathscr{H}$.

The problem of the location of $\widehat{\mu}'_{M,r}$ is equivalent to a uniform distribution on a circle of circumference $2(b-a)$, where we form the circle by joining $-a$ and a at a single point, and join $-b$ to b. By symmetry, we obtain that $|\widehat{\mu}'_{M,r}|$ is uniform on $[a, b]$. Thus, $\mathbb{P}\left(|\widehat{\mu}'_{M,r}| \in [\frac{b-a}{2}, \frac{3(b-a)}{4}]\right) = \frac{1}{4}$.

$$\begin{aligned}
\mathbb{P}\left(|\widehat{\mu}_{M,r}| \geqslant \frac{b-a}{2}\right) &\geqslant \mathbb{P}\left(|\widehat{\mu}_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right]\right) \\
&\geqslant \mathbb{P}\left(|\widehat{\mu}'_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right]\right) = \frac{1}{4}.
\end{aligned}$$

This proves the first statement. Now, we consider the case when we condition on the value of Z. Note that if $|\widehat{\mu}'_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right]$, then $Z = Z'$.

$$\begin{aligned}
\mathbb{P}\left(|\widehat{\mu}_{M,r}| \geqslant \frac{b-a}{2} \middle| Z \geqslant k\right) &\geqslant \mathbb{P}\left(|\widehat{\mu}_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right] \middle| Z \geqslant k\right) \\
&\geqslant \mathbb{P}\left(|\widehat{\mu}'_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right] \middle| Z \geqslant k\right) \\
&= \frac{\mathbb{P}\left(|\widehat{\mu}'_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right], Z \geqslant k\right)}{\mathbb{P}(Z \geqslant k)} \\
&\geqslant \frac{\mathbb{P}\left(|\widehat{\mu}'_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right], Z \geqslant k\right)}{\mathbb{P}(Z' \geqslant k)} \\
&= \frac{\mathbb{P}\left(|\widehat{\mu}'_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right], Z' \geqslant k\right)}{\mathbb{P}(Z' \geqslant k)} \\
&= \mathbb{P}\left(|\widehat{\mu}'_{M,r}| \in \left[\frac{b-a}{2}, \frac{3(b-a)}{4}\right] \middle| Z' \geqslant k\right) = \frac{1}{4}
\end{aligned}$$

where we use the following Lemma C.5 for independence of $\widehat{\mu}'_{M,r}$ and Z'. □

LEMMA C.5 Suppose $X_1, \ldots, X_n$ are i.i.d. uniform points on a circle. Let E be the event that the maximum number of points contained in an arc of a certain length is equal to k. Then the joint distribution $p(x_1, \ldots, x_n)$ is rotationally invariant.

Proof. Suppose without loss of generality that the circle has circumference 1. Note that the law of $(X_1, \ldots, X_n)$ can be equivalently generated as follows: first generate $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \mathit{Unif}[0,1]$. Next, generate $R \sim \mathit{Unif}[0,1]$, and define $X_i = Y_i + R$ for all $1 \leqslant i \leqslant n$, where the addition is taken modulo 1. We want to show that

$$p(x_1, \ldots, x_n \mid E) = p(x_1 + r, \ldots, x_n + r \mid E) \tag{C2}$$

for any $r \in [0,1]$, where addition is again taken modulo 1. Clearly, it suffices to consider configurations $(x_1, \ldots, x_n)$ that are consistent with E.

We can calculate

$$p(x_1, \ldots, x_n \mid E) = \frac{\int_{E'} p(x_1, \ldots, x_n, y_1, \ldots, y_n)\, \mathrm{d}y}{P(E)},$$

where the integral is taken over the region of $[0,1]^n$ containing points $(y_1, \ldots, y_n)$ that can be obtained from $(x_1, \ldots, x_n)$ via some rotation. Importantly, note that

$$p(x_1, \ldots, x_n, y_1, \ldots, y_n) = p(x_1, \ldots, x_n \mid y_1, \ldots, y_n) p(y_1, \ldots, y_n) = p(y_1, \ldots, y_n),$$

since R is uniform, so we have

$$p(x_1, \ldots, x_n \mid E) = \frac{\int_{E'} p(y_1, \ldots, y_n)\, \mathrm{d}y}{P(E)}.$$

Similarly, we can write

$$p(x_1 + r, \ldots, x_n + r \mid E) = \frac{\int_{E'} p(x_1 + r, \ldots, x_n + r, y_1, \ldots, y_n)\, \mathrm{d}y}{P(E)} = \frac{\int_{E'} p(y_1, \ldots, y_n)\, \mathrm{d}y}{P(E)}.$$

This establishes the desired equality (C.2) and completes the proof. □

D. Proofs for multivariate estimators

In this appendix, we provide proofs of the various theorems and lemmas for multivariate mean estimation.

D.1 *Proof of Theorem 4.1*

The initial steps in the proof parallel the proof of Theorem 3.1, where Lemma B.2 is proved using the concentration inequality in Lemma 2.2. It then follows that if we choose r such that $R^*_r \geqslant C_{0.5}\left(\frac{(d+1)\log n}{n}\right)$, we have $R(f_{\widehat{\mu}_{M,r},r}) \geqslant \frac{R^*_r}{2}$, w.h.p.

Now let $r_2 = 4r\left(\frac{2}{R^*_r}\right)^{\frac{1}{d}}$. By Lemma 2.1(i), the desired result will follow if we can show that $R(f_{r_2,r}) \leqslant \frac{R^*_r}{2}$. By Lemma 2.1(iv), we have

$$R(f_{r_2,r}) \leqslant \frac{R^*_r}{2} \cdot R^*_{r_2} \leqslant \frac{R^*_r}{2}.$$

To obtain inequality (4.2), note that using Lemma 2.1(v), we know that $r = 2\sqrt{d}\sigma_{(2Cd\log n)}$ satisfies the assumption on R^*_r. Plugging into inequality (4.1) then produces the desired bound.

D.2 *Proof of Theorem 4.2*

Let $j' := \min\{j \in \mathscr{J} : r_j \geqslant r^*\}$. Then

$$\begin{aligned}
\mathbb{P}(j_* > j') &= \mathbb{P}\left(\bigcup_{i \in \mathscr{J}: i > j'} \left\{\|\widehat{\mu}_{M,r_i} - \widehat{\mu}_{M,r_{j'}}\|_2 > 8r_i\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}\right\}\right) \\
&\leqslant \mathbb{P}\left(\|\widehat{\mu}_{M,r_{j'}}\|_2 > 4r_{j'}\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}\right) \\
&\quad + \sum_{i \in \mathscr{J}: i > j'} \mathbb{P}\left(\|\widehat{\mu}_{M,r_i}\|_2 > 4r_i\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}\right),
\end{aligned}$$

using a union bound and the triangle inequality. We may use Theorem 4.1 to bound each individual term, so that the probability of the bad event

$$E := \bigcup_{i \in \mathscr{J}: i > j'} \left\{\|\widehat{\mu}_{M,r_i}\|_2 > 4r_i\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}\right\} \cup \left\{\|\widehat{\mu}_{M,r_{j'}}\|_2 > 4r_{j'}\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}\right\}$$

is bounded by

$$\begin{aligned}\mathbb{P}(E) &\leqslant (1+|\mathscr{J}|)\cdot 2\exp(-c'd\log n)\\ &\leqslant 2\left(1+\log_2\left(\frac{2r_{\max}}{r_{\min}}\right)\right)\exp(-c'd\log n).\end{aligned}$$

Finally, note that on the event E^c, we have $j_* \leqslant j'$ (establishing that j_* is finite), so

$$\|\widehat{\mu}_{M,r_{j_*}} - \widehat{\mu}_{M,r_{j'}}\|_2 \leqslant 8r_{j'}\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}.$$

Combined with the inequality $\|\widehat{\mu}_{M,r_{j'}}\|_2 < 4r_{j'}\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}$, we conclude that

$$\begin{aligned}\|\widehat{\mu}_{M,r_{j_*}}\|_2 &\leqslant 8r_{j'}\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d} + 4r_{j'}\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}\\ &\leqslant 12r_{j'}\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d}\\ &\leqslant 24r^*\left(\frac{2n}{C_{0.5}(d+1)\log n}\right)^{1/d},\end{aligned}$$

using the fact that $r_{j'} < 2r^*$.

D.3 *Proof of Lemma 4.1*

We first prove the upper bound. Note that $R(f_{0,r_{2k}}) = R^*_{r_{2k}} = \frac{2k}{n}$. It suffices to show that this ball contains at least k points, with high probability. By the multiplicative form of the Chernoff bound (Lemma H.1 in Appendix H),

$$\begin{aligned}\mathbb{P}\left(R_n(f_{0,r_{2k}}) \leqslant \frac{k}{n}\right) &= \mathbb{P}\left(R_n(f_{0,r_{2k}}) \leqslant \frac{1}{2}R(f_{0,r_{2k}})\right)\\ &\leqslant \exp\left(-n\cdot\frac{k}{n}\cdot\frac{1}{8}\right) = \exp(-k/8).\end{aligned}$$

Therefore, with probability at least $1-\exp(-k/8)$, a ball of radius r_{2k} contains at least k points, implying that the shortest gap, $\widehat{r}_k \leqslant r_{2k}$.

We now turn to verifying the lower bound. We will prove that with high probability, no ball of radius $r_{k/2}$ contains at least k points, so that $\widehat{r}_k > r_{k/2}$. By definition, $nR^*_{r_{k/2}} = \frac{k}{2}$. Thus, assuming $k \geqslant C_{0.5}d\log n$, we may apply Lemma 2.2 to conclude that

$$\sup_{f\in\mathscr{H}_{r_{k/2}}} R_n(f) - R(f) \leqslant \frac{R^*_{r_{k/2}}}{2},$$

with probability at least

$$1-\exp\left(-\frac{cn}{4}R^*_{r_{k/2}}\right) = 1-\exp(-ck/8).$$

This implies that

$$\sup_{f \in \mathscr{H}_{r_{k/2}}} R_n(f) \leqslant \frac{3}{2} \cdot R^*_{r_{k/2}} = \frac{3}{2} \cdot \frac{k}{2n} < \frac{k}{n},$$

which is exactly what we want.

D.4 *Proof of Theorem 4.3*

We parallel the proof of Theorem 3.2. Note that the guarantees of Lemma 4.1 and Lemma B.3 continue to hold in d dimensions, except that we have the lower bound $k \geqslant 2C_{0.5}(d+1)\log n$ instead. We then conclude that $R(f_{\widehat{\mu}_{S,k}, r_{2k}}) \geqslant \frac{k}{2n}$, with probability at least $1 - 2\exp(-c'd\log n)$.

Setting $r' = 4r_{2k}\left(\frac{2n}{k}\right)^{1/d}$, it thus suffices to show that $R(f_{r', r_{2k}}) \leqslant \frac{k}{2n}$. By Lemma 2.1(iv), we have

$$R(f_{r', r_{2k}}) \leqslant \frac{k}{2n} \cdot R^*_{r'} \leqslant \frac{k}{2n},$$

as wanted.

D.5 *Proof of Theorem 4.4*

We begin with the following result, which can be proved directly via a union bound on Lemma B.4:

LEMMA D.1 With probability at least $1 - 4d\exp(-ck^2/n)$:

(i) The cuboid S_k^∞ contains the origin.

(ii) We have the bound $\mathrm{Diam}(S_k^\infty) \leqslant 2\sqrt{d} r_{2k,1}$.

Lemma D.1 will be critical in our analysis of the hybrid estimator proposed below. In particular, the estimator will consist of projecting the modal interval/shorth estimator onto the cuboid S_k^∞, and Lemma D.1(i) guarantees that the estimation error of the projected estimator will be no larger than the estimation error of the initial estimator without projection. On the other hand, Lemma D.1(ii) bounds the error of an estimator based on the k-median alone.

We first derive an upper bound of $\sqrt{d} r_{2\sqrt{n}\log n, 1}$. We begin by deriving the following lemma, relating the statistics of marginal distributions to the statistics of the overall distribution:

LEMMA D.2 We have that $r_{\frac{k}{2},1} \leqslant \frac{C}{\sqrt{d}} r_k$, for some absolute constant $C > 0$ and any $k \leqslant n$.

Proof. Consider a uniform distribution on a sphere (or shell) of radius r in $\mathbb{R}^d$. Vershynin [45, Theorem 3.4.6] provides a concentration result which states that most of the probability of such a distribution lies close to the equator; i.e., the set $\left[-\frac{Cr}{\sqrt{d}}, \frac{Cr}{\sqrt{d}}\right] \times \mathbb{R}^{d-1}$ contains at least half the probability for some absolute constant $C > 0$. Notice that a radially symmetric distribution is simply a weighted sum of uniform distributions on spheres. Thus, given a radially symmetric distribution restricted to the ball of radius r, the set $\left[-\frac{Cr}{\sqrt{d}}, \frac{Cr}{\sqrt{d}}\right] \times \mathbb{R}^{d-1}$ will contain at least half the total probability assigned to the ball.

By our definition of r_k, the ball of radius r_k centered at origin, $\mathbb{B}_{r_k}$, contains $\frac{k}{n}$ probability mass. The above argument implies that the set $\left[-\frac{Cr_k}{\sqrt{d}}, \frac{Cr_k}{\sqrt{d}}\right] \times \mathbb{R}^{d-1}$ will contain at least half the probability of the total probability contained in $\mathbb{B}_{r_k}$. Equivalently, $r_{\frac{k}{2},1} \leqslant \frac{C}{\sqrt{d}} r_k$. □

Since the output of the hybrid algorithm must lie within the cuboid $S^{\infty}_{\sqrt{n}\log n}$, it is clear that we have the error bound

$$\|\widehat{\mu}_{k_1,k_2}\|_2 \leqslant \sqrt{n}^{1/d} \cdot \sqrt{d} r_{2\sqrt{n}\log n,1}.$$

To obtain the second upper bound expression, we parallel the proof of Theorem 3.3, by splitting into two cases:

Case 1: $r_{4\sqrt{n}\log n} \leqslant \sqrt{n}^{1/d} r_{8d\log n}$. By Lemma D.2, we therefore have

$$r_{2\sqrt{n}\log n,1} \leqslant \frac{C}{\sqrt{d}} r_{4\sqrt{n}\log n} \leqslant \frac{C}{\sqrt{d}} \cdot \sqrt{n}^{1/d} r_{8d\log n}.$$

By Lemma D.1, w.h.p., the cuboid $S^{\infty}_{\sqrt{n}\log n}$ is entirely contained in the ℓ_2-ball of radius $\sqrt{d} r_{2\sqrt{n}\log n,1}$ around the origin. This ball in turn lies inside the ℓ_2-ball of radius $C\sqrt{n}^{1/d} r_{8d\log n}$ around the origin. Since the output of the hybrid algorithm must also lie within this ball, the desired result follows.

Case 2: $r_{4\sqrt{n}\log n} > \sqrt{n}^{1/d} r_{8d\log n}$. Denoting $r' = \sqrt{n}^{1/d} r_{8d\log n}$, we therefore have the relation $R^*_{r'} < \frac{4\sqrt{n}\log n}{2n}$. In particular, since

$$R(f_{\widehat{\mu}_{S,8d\log n}, r_{8d\log n}}) \geqslant R_n(f_{\widehat{\mu}_{S,8d\log n}, r_{8d\log n}}) - \frac{1}{2} R^*_{r_{8d\log n}} = \frac{8d\log n}{4n},$$

w.h.p., by Lemma 2.2, we have

$$R(f_{r', r_{8d\log n}}) \leqslant \left(\frac{1}{\sqrt{n}^{1/d}}\right)^d R^*_{r'} < \frac{1}{\sqrt{n}} \cdot \frac{2\log n}{\sqrt{n}} = \frac{8d\log n}{4n} \leqslant R(f_{\widehat{\mu}_{S,8d\log n}, r_{8d\log n}}).$$

This implies that $\widehat{\mu}_{S,8d\log n}$ is within r' of the origin.

Finally, we need to show that projecting the shorth estimator on the cuboid does not increase its distance from the origin. Note that ℓ_2-projection onto a cuboid is simply a componentwise operation of projection on each interval defining an edge of the cuboid. Furthermore, Lemma D.1 guarantees that the origin lies within the cuboid, w.h.p., in which case each interval contains 0. As argued in the proof of Theorem 3.3, the distance from the shorth estimator to the origin computed along any dimension will not increase after the projection. Therefore, the ℓ_2-norm of the projected estimator is also upper bounded by r'.

Hence, if we take $C' = \max\{C, 1\}$, we have the desired bound in both cases. This concludes the proof.

D.6 *Proof of Theorem 6.1*

We begin by deriving the proof for the modal interval estimator. Let $s_1 = \frac{r}{2}$, and define s_2 such that $R(f_{s_2,r}) = \frac{1}{3} R(f_{s_1,r})$. Note that

$$R(f_{s_1,r}) \geqslant R(f_{0,r/2}) \geqslant \frac{3C_{1/6} d\log n}{n},$$

so $R(f_{s_2,r}) \geqslant \frac{C_{1/6} d\log n}{n}$. Applying Lemma 6.1 with $\bar{r} = s_1$ and $t = \frac{1}{6}$, we conclude that

$$R_n(f_{x,r}) \geqslant \frac{2}{3} R(f_{x,r}) \geqslant \frac{2}{3} R(f_{s_1,r}), \tag{D.1}$$

uniformly over $\|x\|_2 \leqslant s_1$, with probability at least $1 - \frac{2\exp(-cnR(f_{s_1,r})/36)}{1-\exp(-cnR(f_{s_1,r})/36)}$, which is in turn lower bounded by $1 - 4\exp(-c_1 d\log n)$.

Furthermore, inequality (6.4) implies that

$$R_n(f_{x,r}) \leqslant R(f_{x,r}) + \frac{1}{3}R(f_{s_2,r}) \leqslant \frac{4}{3}R(f_{s_2,r}) = \frac{4}{9}R(f_{s_1,r}), \tag{D.2}$$

uniformly over $\|x\|_2 > s_2$, with probability at least $1 - 2\exp(-cnR(f_{s_2,r})/9) \geqslant 1 - 2\exp(-c_2 d\log n)$. Thus, combining inequalities (D.1) and (D.2), we conclude that

$$\sup_{\|x\|_2 > s_2} R_n(f_{x,r}) < \inf_{\|x\|_2 \leqslant s_1} R_n(f_{x,r}), \tag{D.3}$$

with probability at least $1 - 6\exp(-c_3 d\log n)$.

Now note that by inequality (D.1), we also have $R_n(f_{0,s_1}) \geqslant \frac{2}{3}R(f_{0,s_1}) > 0$, implying that $\{x_1,\dots,x_n\} \cap B(0,s_1) \neq \emptyset$. In particular,

$$\sup_{x\in\{x_1,\dots,x_n\}} R_n(f_{x,r}) \geqslant \inf_{\|x\|_2 \leqslant s_1} R_n(f_{x,r}).$$

Together with inequality (D.3), we conclude that $\|\widetilde{\mu}_{M,r}\|_2 < s_2$.

Finally, we claim that $s_2 \leqslant 4r\left(\frac{n}{C_{1/6}d\log n}\right)^{1/d}$. To see this, let $\tilde{s}_2 := 4r\left(\frac{n}{C_{1/6}d\log n}\right)^{1/d}$, and note that by Lemma 2.1(iv), we have

$$R(f_{\tilde{s}_2,r}) \leqslant \frac{C_{1/6}d\log n}{n}\cdot R^*_{\tilde{s}_2} \leqslant \frac{C_{1/6}d\log n}{n}.$$

Since the last quantity is upper bounded by $R(f_{s_2,r})$, we conclude that $s_2 \leqslant \tilde{s}_2$, as claimed.

Turning to the analysis of the computationally efficient shorth estimator, we adapt the argument in the proof of Theorem 3.2. By Lemma 2.2, if $R^*_{2r_{2k}} \geqslant \frac{C_{0.5}(d+1)\log n}{n}$, we have

$$\sup_x \sup_{r\leqslant 2r_{2k}} \big(R_n(f_{x,r}) - R(f_{x,r})\big) < \frac{t}{2}R^*_{2r_{2k}},$$

with probability at least $1 - 2\exp(-cnR^*_{2r_{2k}}t^2) \geqslant 1 - 2\exp\left(-cnt^2\cdot\frac{2k}{n}\right)$.

We know that $\frac{k}{n} = R_n(f_{\widetilde{\mu}_{S,k},\widetilde{r}_k}) \leqslant R_n(f_{\widetilde{\mu}_{S,k},2r_{2k}})$. Let s be defined such that $R(f_{s,2r_{2k}}) = \frac{k}{2n}$. By inequality (6.4), we know that

$$\sup_{\|x\|_2 \geqslant s} \big|R_n(f_{x,2r_{2k}}) - R(f_{x,2r_{2k}})\big| \leqslant \frac{1}{2}R(f_{s,2r_{2k}}),$$

with probability at least $1 - 2\exp(-ck)$, implying that for $\|x\|_2 \geqslant s$, we have

$$R_n(f_{x,2r_{2k}}) \leqslant R(f_{x,2r_{2k}}) + \frac{1}{2}R(f_{s,2r_{2k}}) \leqslant \frac{3}{2}R(f_{s,2r_{2k}}) = \frac{3k}{4n}.$$

Since this is strictly smaller than $R_n(f_{\widetilde{\mu}_{S,k},2r_{2k}})$, we conclude that $\|\widetilde{\mu}_{S,k}\|_2 \leqslant s$, w.h.p. , which also implies that $R(f_{\widetilde{\mu}_{S,k},2r_{2k}}) \geqslant \frac{k}{2n}$.

Finally, let $r' = 4r_{2k}\left(\frac{2n}{k}\right)^{1/d}$. By Lemma 2.1(iv), we have

$$R(f_{r',2r_{2k}}) < \frac{k}{2n} \cdot R^*_{r'} \leqslant \frac{k}{2n} < R(f_{\widetilde{\mu}_{S,k},2r_{2k}}).$$

Applying Lemma 2.1(i), we conclude that $\|\widetilde{\mu}_{S,k}\|_2 \leqslant r'$.

E. Proofs for expected error bounds

In this appendix, we prove the results stated in Section 5.

E.1 *Proof of Proposition 5.2*

The proof sketch is that we will show that with finite probability, no interval contains more than one low-variance point, and all the high-variance points lie far from origin. Conditioned on this event, the modal interval estimator incurs a high error.

Let $E = A \cap B$, where we define the events

$$A = \{R_n(f_{x,1}) \leqslant 1, \quad \forall x : |x| \leqslant 3C\log n\},$$
$$B = \{X_i \notin [-4C\log n, 4C\log n], \quad \forall i > C\log n\}.$$

Hence, on the event E, no interval overlapping with $[-3C\log n, 3C\log n]$ contains two low-variance points or a single high-variance point. Then $\mathbb{P}(E)$ is lower bounded by

$$\begin{aligned}
\mathbb{P}(E) &\geqslant \left(\prod_{i=1}^{C\log n} \mathbb{P}\{X_i \in [3i-3, 3i-2]\}\right)\left(\prod_{i>C\log n} \mathbb{P}\{X_i \notin [-4C\log n, 4C\log n]\}\right) \\
&= \left(\prod_{i=1}^{C\log n} \frac{1}{6i}\right)\left(\prod_{i>C\log n} (1 - n^{-\alpha} - h_n(8C\log n - 2))\right) \\
&\geqslant \frac{1}{6^{C\log n}\Gamma(3C\log n)} e^{-cn^{1-\alpha}} \\
&\geqslant \exp\left(-cn^{1-\alpha} - O\left(\log^2 n\right)\right),
\end{aligned}$$

assuming $h_n \log n \ll n^{-\alpha}$, which happens for $q_n = \Omega(n)$.

However, conditioned on E, the points $\{X_i\}_{i>C\log n}$ are i.i.d. with the following distribution:

$$p_{i,E}(x) = \begin{cases} 0, & |x| \leqslant 4C\log n, \\ \frac{h_n}{(1-n^{-\alpha}-h_n(8C\log n-2))}, & 4C\log n < |x| \leqslant q_n, \\ 0, & \text{otherwise.} \end{cases}$$

We can now apply the symmetry arguments of Lemma C.4. Note that no interval lying inside $[-3C\log n, 3C\log n]$ can contain more than one point. Thus, unless a tie occurs, the mode will be located outside the interval $[-3C\log n, 3C\log n]$, and hence a distance of $\Theta(q_n)$ away from the mean in expectation. Even if we were to break ties randomly, a large error would occur with probability at least

$\frac{1}{n}$, since at most n ties can occur. Thus,

$$\mathbb{E}[|\widehat{\mu}_{M,1}||E] \geqslant \mathbb{P}(E)\,\mathbb{E}[|\widehat{\mu}_{M,1}||E] \geqslant \exp(-cn^{1-\alpha})\Theta(q_n).$$

The bounds in high probability follow from Lemma B.2, by noting that $nR_r^* = \Omega\left(n^{-\alpha}\right) = \Omega(\log n)$. Moreover, the density drops by at least half at $x > 1$.

E.2 *Proof of Theorem 5.3*

We begin by proving (i). By Theorem 4.1, we have

$$\|\widehat{\mu}_{M,r}\|_2 = O\left(r\left(\frac{c}{R_r^*}\right)^{1/d}\right),$$

with probability at least most $1 - O(\exp\left(-c'nR_r^*\right))$. In the worst case, the modal interval estimator returns the point which is furthest from the origin, which has expected value bounded as

$$\mathbb{E}\left[\max_i \|X_i\|_2\right] \leqslant \mathbb{E}\left[\sqrt{\sum_{i=1}^n \|X_i\|_2^2}\right] \leqslant \sqrt{\sum_{i=1}^n \mathbb{E}[\|X_i\|_2^2]} \leqslant \sqrt{n \cdot d\sigma_{(n)}^2}.$$

Using the assumption that $\sigma_n \leqslant r\exp(CnR_r^*)$, for some constant $C > 0$, we then have

$$\begin{aligned}
\mathbb{E}\,\|\widehat{\mu}_{M,r}\|_2 &\leqslant O\left(r\left(\frac{c}{R_r^*}\right)^{1/d}\right) + O(\exp\left(-c'nR_r^*\right))\sqrt{nd}\sigma_{(n)} \\
&\leqslant O\left(r\left(\frac{c}{R_r^*}\right)^{1/d}\right) + O\left(\exp\left(-c'nR_r^*\right) r\sqrt{nd}\exp(CnR_r^*)\right) \\
&= O\left(r\left(\frac{c}{R_r^*}\right)^{1/d}\right),
\end{aligned}$$

where in the last inequality, we use the facts that

$$\exp(-c'nR_r^*)\sqrt{nd} = O(\exp(-c''nR_r^*))$$

and $nR_r^* = \Omega\left(d\log n\right)$, and choose $C < c''$.

Turning to (ii), we first prove the following concentration result, which may be viewed as a refinement of Lemma 2.2 that is suitable for our settings. For example, note that if $R^*_{\mathscr{J}} = O\left(\frac{1}{n}\right)$, the derivations from Lemma 2.2 would not be meaningful since $R^*_{\mathscr{J}} = o\left(\frac{\log n}{n}\right)$. On the other hand, if $KR^*_{\mathscr{J}} = \Theta\left(\frac{\log n}{n}\right)$, Lemma E.1 gives a vanishing upper bound.

LEMMA E.1 Let $\mathcal{J}$ be a set of intervals and define $R^*_{\mathcal{J}} := \sup_{f\in\mathcal{J}} R(f)$. If $R^*_{\mathcal{J}} \leqslant \frac{1}{3}$, then for any $K \geqslant 8$, we have

$$\mathbb{P}\left\{\sup_{f\in\mathcal{J}} R_n(f) \geqslant KR^*_{\mathcal{J}}\right\} \leqslant \frac{2}{R^*_{\mathcal{J}}} \exp\left(-cnR^*_{\mathcal{J}} K \log K\right).$$

Proof. For a given $f \in \mathcal{J}$, the desired bound follows from Chernoff's inequality. We want to upper bound the probability that any one interval in $\mathcal{J}$ has too many points. In general, the set $\mathcal{J}$ may be infinite, so a direct union bound is not feasible. We thus create a new finite set of intervals $\mathcal{F}$, not necessarily a subset of $\mathcal{J}$, satisfying the following properties:

1. For each $f \in \mathcal{F}$, we have $\frac{R^*_{\mathcal{J}}}{2} \leqslant R(f) \leqslant R^*_{\mathcal{J}}$.
2. $|\mathcal{F}| \leqslant \frac{2}{R^*_{\mathcal{J}}}$.
3. $\mathcal{F}$ covers $\mathcal{J}$ in the sense that $\forall f \in \mathcal{J}, \exists f_1, f_2 \in \mathcal{F} : f(x) \leqslant f_1(x) + f_2(x)$.

It follows that if any interval in $\mathcal{J}$ contains at least k points, then at least one interval in $\mathcal{F}$ contains at least $\frac{k}{2}$ points. We construct $\mathcal{F}$ of cardinality $|\mathcal{F}| = \lceil \frac{1}{R^*_{\mathcal{J}}} \rceil \leqslant \frac{2}{R^*_{\mathcal{J}}}$, as follows: to create the first interval ($i = 1$), define $x_1 \in \mathbb{R}$ such that $R(\mathbb{1}_{(-\infty,x_1]}) = \frac{1}{|\mathcal{F}|}$. (Such an x_1 exists because $\overline{P}$ is assumed to have a density.) Then iteratively, for each $i \geqslant 1$, define x_i such that $R(\mathbb{1}_{(x_{i-1},x_i]}) = \frac{1}{|\mathcal{F}|}$. For the final interval, add $\mathbb{1}_{[x_{i-1},\infty)}$ to $\mathcal{F}$ and terminate the construction. Note that for each $f \in \mathcal{F}$, we have $R(f) = \frac{1}{\lceil 1/R^*_{\mathcal{J}} \rceil}$, which clearly lies in $\left[\frac{R^*_{\mathcal{J}}}{2}, R^*_{\mathcal{J}}\right]$ under the assumptions.

We are now ready to apply the union bound on $\mathcal{F}$ using Lemma H.1(ii):

$$\begin{aligned}
\mathbb{P}\left\{\sup_{f\in\mathcal{J}} R_n(f) \geqslant KR^*_{\mathcal{J}}\right\} &\leqslant \mathbb{P}\left\{\sup_{f\in\mathcal{F}} R_n(f) \geqslant \frac{KR^*_{\mathcal{J}}}{2}\right\} \\
&\leqslant |\mathcal{F}|\mathbb{P}\left\{R_n(f) \geqslant \frac{KR^*_{\mathcal{J}}}{2} \text{ for a fixed} f \text{ with} R(f) \leqslant R^*_{\mathcal{J}}\right\} \\
&\leqslant \frac{2}{R^*_{\mathcal{J}}} \exp\left(-cnR^*_{\mathcal{J}} K \log K\right).
\end{aligned}$$

□

For an $s \geqslant 0$, let $\mathcal{J}_s = \{f_{x,r} : \|x\|_2 \geqslant s\}$, i.e., the set of intervals which incur large error. By assumption, the support of at least CnR^*_r points is contained in $[-r, r]$, implying that $R_n(f_{0,r}) \geqslant CR^*_r$, a.s. If $\|\widehat{\mu}_{M,r}\|_2 \geqslant s$, then $\sup_{f\in\mathcal{J}_s} R_n(f) \geqslant CR^*_r$. However as s increases, the quantity $R^*_{\mathcal{J}_s} := \sup_{f\in\mathcal{J}_s} R(f) = R(f_{s,r})$ decreases. We can then use Lemma E.1 to control this probability of error.

For $s \geqslant \frac{Kr}{CR_r^*}$, it follows from Lemma 2.1(iv) that $R^*_{\mathscr{J}_s} = R(f_{s,r}) \leqslant \frac{CR_r^*}{K}$. Taking $K \geqslant C'$, we then have

$$\begin{aligned}
\mathbb{P}\{|\widehat{\mu}_{M,r}| \geqslant s\} &\leqslant \mathbb{P}\left(\sup_{f \in \mathscr{J}_s} R_n(f) \geqslant CR_r^*\right) \\
&= \mathbb{P}\left(\sup_{f \in \mathscr{J}_s} R_n(f) \geqslant \frac{CR_r^*}{R^*_{\mathscr{J}_s}} R^*_{\mathscr{J}_s}\right) \\
&\leqslant \frac{2}{R^*_{\mathscr{J}_s}} \exp\left(-cnR^*_{\mathscr{J}_s} \frac{CR_r^*}{R^*_{\mathscr{J}_s}} \log\left(\frac{CR_r^*}{R^*_{\mathscr{J}_s}}\right)\right) \\
&= \frac{2}{R_r^*} \exp\left(-cCnR_r^* \log\left(\frac{CR_r^*}{R^*_{\mathscr{J}_s}}\right) + \log\left(\frac{R_r^*}{R^*_{\mathscr{J}_s}}\right)\right) \\
&\leqslant \frac{2}{R_r^*} \exp\left(-c'nR_r^* \log\left(\frac{R_r^*}{R^*_{\mathscr{J}_s}}\right)\right),
\end{aligned}$$

where we have applied Lemma E.1 in the second inequality. Thus,

$$\begin{aligned}
\mathbb{E}\,|\widehat{\mu}_{M,r}| &\leqslant \frac{4r}{CR_r^*} + \int_{\frac{4r}{CR_r^*}}^{\infty} \mathbb{P}\{|\widehat{\mu}_{M,r}| \geqslant s\}\,\mathrm{d}s \\
&\leqslant O\left(\frac{r}{R_r^*}\right) + \frac{2}{R_r^*} \int_{\frac{4r}{CR_r^*}}^{\infty} \exp\left(-c'nR_r^* \log\left(\frac{R_r^*}{R^*_{\mathscr{J}_s}}\right)\right) \mathrm{d}s \\
&\leqslant O\left(\frac{r}{R_r^*}\right) + \frac{2}{R_r^*} \int_{\frac{4r}{CR_r^*}}^{\infty} \exp\left(-c'nR_r^* \log\left(\frac{sR_r^*}{r}\right)\right) \mathrm{d}s \\
&\leqslant O\left(\frac{r}{R_r^*}\right) + \frac{r}{R_r^*}\frac{2}{R_r^*} \int_{4/C}^{\infty} \exp\left(-c'nR_r^* \log s_1\right)\,\mathrm{d}s_1 \\
&= O\left(\frac{r}{R_r^*}\right) + \frac{r}{R_r^*}\frac{2}{R_r^*} \int_{4/C}^{\infty} s_1^{-c'nR_r^*}\,\mathrm{d}s_1 \\
&\leqslant O\left(\frac{r}{R_r^*}\right) + \frac{r}{R_r^*}\frac{2}{R_r^*} \cdot \frac{1}{c'nR_r^* - 1} (4/C)^{1-c'nR_r^*} \\
&= O\left(\frac{r}{R_r^*}\right),
\end{aligned}$$

where the third inequality uses the fact that $R^*_{\mathscr{J}_s} = R(f_{s,r}) \leqslant \frac{r}{s}$ and the last equality follows from an appropriately small choice of C.

E.3 *Proof of Theorem 5.4*

Note that for any $s > 0$, Markov's inequality gives

$$\min_{\widehat{\mu}} \max_{\{P_i\} \subseteq \mathscr{P}(\sigma_1, \sigma_2, p)} \mathbb{E}[\|\widehat{\mu} - \mu\|_2] \geqslant \min_{\widehat{\mu}} \max_{\{P_i\} \subseteq \mathscr{P}(\sigma_1, \sigma_2, p)} s \cdot \mathbb{P}(\|\widehat{\mu} - \mu\|_2 \geqslant s).$$

Clearly, the right-hand expression is lower bounded by the maximum over any specific collection of distributions in the class $\mathscr{P}(\sigma_1, \sigma_2, p)$. In particular, let $\mathscr{P}_m^{\mu}$ be the collection of multivariate distributions where each distribution is either $N(\mu, \sigma_1^2 I)$ or $N(\mu, \sigma_2^2 I)$, with m distributions of the latter type. We then have

$$\begin{aligned}\min_{\widehat{\mu}} \max_{\{P_i\} \subseteq \mathscr{P}(\sigma_1, \sigma_2, p)} \mathbb{P}(\|\widehat{\mu} - \mu\|_2 \geqslant s) &\geqslant \min_{\widehat{\mu}} \max_{\mu} \max_{np \leqslant m \leqslant 2np} \mathbb{P}(\|\widehat{\mu} - \mu\|_2 \geqslant s \mid \{P_i\} = \mathscr{P}_m^{\mu}) \\ &\geqslant \min_{\widehat{\mu}} \max_{\mu} \sum_{np \leqslant m \leqslant 2np} \mathbb{P}\left(\|\widehat{\mu} - \mu\|_2 \geqslant s \mid \{P_i\} = \mathscr{P}_m^{\mu}\right) p_m,\end{aligned}$$

where $\{p_m\}$ is any allocation of probabilities defined over $\{\mathscr{P}_{np}^{\mu}, \ldots, \mathscr{P}_{2np}^{\mu}\}$, such that $0 \leqslant p_m \leqslant 1$ for all m and $\sum_m p_m \leqslant 1$. In particular, consider the probability mass function $\{q_m\}_{m=1}^n$ over $\{\mathscr{P}_1^{\mu}, \ldots, \mathscr{P}_n^{\mu}\}$ corresponding to the Binomial(n, p) distribution, and define $p_m = q_m$ for all $np \leqslant m \leqslant 2np$.

Now let $\mathbb{P}_{\text{Bin}}^{\mu}$ denote the probability distribution when the P_is are chosen i.i.d. in the following manner: with probability $p' := 1.5p$, the distribution is $N(\mu, \sigma_2^2 I)$, and with probability $1 - 1.5p$, the distribution is $N(\mu, \sigma_1^2 I)$. Then

$$\mathbb{P}_{\text{Bin}}^{\mu}(\|\widehat{\mu} - \mu\|_2 \geqslant s) = \sum_{m=1}^{n} \mathbb{P}\left(\|\widehat{\mu} - \mu\|_2 \geqslant s \mid \{P_i\} = \mathscr{P}_m^{\mu}\right) q_m.$$

Hence,

$$\begin{aligned}\left| \sum_{np \leqslant m \leqslant 2np} \mathbb{P}\left(\|\widehat{\mu} - \mu\|_2 \geqslant s \mid \{P_i\} = \mathscr{P}_m^{\mu}\right) p_m - \mathbb{P}_{\text{Bin}}^{\mu}(\|\widehat{\mu} - \mu\|_2 \geqslant s) \right| &\leqslant \sum_{m < np} q_m + \sum_{m > 2np} q_m \\ &\leqslant 2\exp(-cnp) \\ &\leqslant 2\exp(-c' \log n),\end{aligned}$$

where second inequality follows from the multiplicative Chernoff bound (Lemma H.1) and the last inequality follows by the assumption $p = \Omega\left(\frac{\log n}{n}\right)$. Combining the inequalities, we conclude that

$$\min_{\widehat{\mu}} \max_{\{P_i\} \subseteq \mathscr{P}(s_1, s_2, p)} \mathbb{E}[\|\widehat{\mu} - \mu\|_2] \geqslant s\left(\min_{\widehat{\mu}} \max_{\mu} \mathbb{P}_{\text{Bin}}^{\mu}(\|\widehat{\mu} - \mu\|_2 \geqslant s) - 2\exp(-c' \log n)\right).$$

Thus, it suffices to find s such that the expression $\min_{\widehat{\mu}} \max_{\mu} \mathbb{P}_{\text{Bin}}^{\mu}(\|\widehat{\mu} - \mu\|_2 \geqslant s)$ can be lower bounded by a constant.

For part (i), using standard techniques [42, 46], we may obtain such a lower bound via Fano's inequality. In particular, if we can construct a set $\{\mu_1, \ldots, \mu_M\} \subseteq \mathbb{R}^d$ such that $\|\mu_j - \mu_k\|_2 \geqslant 2s$ and $KL(\mathbb{P}_{\text{Bin}}^{\mu_j}, \mathbb{P}_{\text{Bin}}^{\mu_k}) \leqslant \alpha$ for all $j \neq k$, then

$$\min_{\widehat{\mu}} \max_{\mu} \mathbb{P}_{\text{Bin}}^{\mu}(\|\widehat{\mu} - \mu\|_2 \geqslant s) \geqslant \left(1 - \frac{\alpha + \log 2}{\log M}\right).$$

Note that by tensorization and convexity of the KL divergence, we have the upper bound

$$KL(\mathbb{P}_{\text{Bin}}^{\mu_j}, \mathbb{P}_{\text{Bin}}^{\mu_k}) \leqslant n(1-p')KL\left(N(\mu_j, \sigma_1^2 I), N(\mu_k, \sigma_1^2 I)\right) + np'KL\left(N(\mu_j, \sigma_2^2 I), N(\mu_k, \sigma_2^2 I)\right), \quad \text{(E.1)}$$

where the KL divergences in the right-hand expression are computed with respect to single samples from the respective multivariate normal distributions. Furthermore, the right-hand side of inequality (E.1) is easily calculated to be

$$n(1-p') \cdot \frac{\|\mu_j - \mu_k\|_2^2}{2\sigma_1^2} + np' \cdot \frac{\|\mu_j - \mu_k\|_2^2}{2\sigma_2^2} = n\|\mu_j - \mu_k\|_2^2 \left(\frac{1-p'}{2\sigma_1^2} + \frac{p'}{2\sigma_2^2}\right).$$

In particular, suppose $\{\mu_1, \ldots, \mu_M\}$ is a $2s$-packing of the ball of radius $4s$ in ℓ_2-norm, with $s = C\sqrt{d}\min\left\{\frac{\sigma_1}{\sqrt{n}}, \frac{\sigma_2}{\sqrt{np'}}\right\}$. Then $\log M \geqslant cd$ and

$$KL(\mathbb{P}_{\text{Bin}}^{\mu_j}, \mathbb{P}_{\text{Bin}}^{\mu_k}) \leqslant 4ns^2\left(\frac{1-p'}{2\sigma_1^2} + \frac{p'}{2\sigma_2^2}\right) \leqslant 4C^2 d := \alpha.$$

For a sufficiently small choice of C, we conclude that $\min_{\widehat{\mu}} \max_{\mu} \mathbb{P}_{\text{Bin}}^{\mu}(\|\widehat{\mu} - \mu\|_2 \geqslant s) \geqslant \frac{1}{2}$. Hence, we arrive at the desired bound (5.2).

We now turn to part (ii). We derive the tighter lower bound (5.4) for the case $d = 1$ by evaluating $KL(\mathbb{P}_{\text{Bin}}^{\mu_1}, \mathbb{P}_{\text{Bin}}^{\mu_2})$ more directly. By Tsybakov [42, Theorem 2.2], we know that if we have a pair $\mu_1, \mu_2 \in \mathbb{R}^d$ such that $\|\mu_1 - \mu_2\|_2 \geqslant 2s$ and

$$KL(\mathbb{P}_{\text{Bin}}^{\mu_1}, \mathbb{P}_{\text{Bin}}^{\mu_2}) \leqslant \alpha < \infty, \quad \text{(E.2)}$$

then

$$\min_{\widehat{\mu}} \max_{\mu} \mathbb{P}_{\text{Bin}}^{\mu}(\|\widehat{\mu} - \mu\|_2 \geqslant s) \geqslant \max\left\{\frac{\exp(-\alpha)}{4}, \frac{1 - \sqrt{\alpha/2}}{2}\right\}.$$

Again, since the KL divergence tensorizes, it suffices to compute the KL divergence between a single sample from the distributions $\mathbb{P}_{\text{Bin}}^{\mu_1}$ and $\mathbb{P}_{\text{Bin}}^{\mu_2}$, which we denote by $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively.

We provide the details of the calculation for general d, with the assumption (5.3) replaced by the condition

$$\left(\frac{\sigma_1}{\sigma_2}\right)^d = O\left(\frac{1}{np^2}\right). \quad \text{(E.3)}$$

By a straightforward calculation, we have

$$\log\left(\frac{d\mathbb{P}_1(x)}{d\mathbb{P}_2(x)}\right) = \log\left(\frac{(1-p')\frac{1}{(\sqrt{2\pi}\sigma_1)^d}\exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_1^2}\right) + p'\frac{1}{(\sqrt{2\pi}\sigma_2)^d}\exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2}\right)}{(1-p')\frac{1}{(\sqrt{2\pi}\sigma_1)^d}\exp\left(\frac{-\|x-\mu_2\|_2^2}{2\sigma_1^2}\right) + p'\frac{1}{(\sqrt{2\pi}\sigma_2)^d}\exp\left(\frac{-\|x-\mu_2\|_2^2}{2\sigma_2^2}\right)}\right)$$

$$= \left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_1^2} + \frac{\|x-\mu_2\|_2^2}{2\sigma_1^2}\right) + \log\left(\frac{1+y}{1+z}\right),$$

where

$$y := \frac{p'}{1-p'}\left(\frac{\sigma_1}{\sigma_2}\right)^d \exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2} + \frac{\|x-\mu_1\|_2^2}{2\sigma_1^2}\right),$$
$$z := \frac{p'}{1-p'}\left(\frac{\sigma_1}{\sigma_2}\right)^d \exp\left(\frac{-\|x-\mu_2\|_2^2}{2\sigma_2^2} + \frac{\|x-\mu_2\|_2^2}{2\sigma_1^2}\right).$$

Hence,

$$KL(\mathbb{P}_1, \mathbb{P}_2) = \mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}\left[\frac{-\|x-\mu_1\|_2^2}{2\sigma_1^2} + \frac{\|x-\mu_2\|_2^2}{2\sigma_1^2}\right] + \mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}\left[\log\left(\frac{1+y}{1+z}\right)\right]$$
$$\leqslant \frac{\|\mu_1-\mu_2\|^2}{2\sigma_1^2} + \mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[y] - \mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[z] + \mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[z^2],$$

using the fact that

$$\log\left(\frac{1+y}{1+z}\right) = \log\left(1 + \frac{y-z}{1+z}\right) \leqslant \frac{y-z}{1+z} \leqslant y - z + z^2,$$

since $y, z > 0$. We now write

$$\mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[y] = \frac{p'}{1-p'}\left(\frac{\sigma_1}{\sigma_2}\right)^d \left((1-p')\int \exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2} + \frac{\|x-\mu_1\|_2^2}{2\sigma_1^2}\right)\frac{1}{(\sqrt{2\pi}\sigma_1)^d}\exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_1^2}\right)\mathrm{d}x\right.$$
$$\left. + p'\int \exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2} + \frac{\|x-\mu_1\|_2^2}{2\sigma_1^2}\right)\frac{1}{(\sqrt{2\pi}\sigma_2)^d}\exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2}\right)\mathrm{d}x\right)$$
$$:= A_y + B_y,$$

and

$$\mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[z] = \frac{p'}{1-p'}\left(\frac{\sigma_1}{\sigma_2}\right)^d \left((1-p')\int \exp\left(\frac{-\|x-\mu_2\|_2^2}{2\sigma_2^2} + \frac{\|x-\mu_2\|_2^2}{2\sigma_1^2}\right)\frac{1}{(\sqrt{2\pi}\sigma_1)^d}\exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_1^2}\right)\mathrm{d}x\right.$$
$$\left. + p'\int \exp\left(\frac{-\|x-\mu_2\|_2^2}{2\sigma_2^2} + \frac{\|x-\mu_2\|_2^2}{2\sigma_1^2}\right)\frac{1}{(\sqrt{2\pi}\sigma_2)^d}\exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2}\right)\mathrm{d}x\right)$$
$$:= A_z + B_z.$$

Now, we may calculate

$$A_y = p'\left(\frac{1}{\sqrt{2\pi}\sigma_2}\right)^d \int \exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2}\right)\mathrm{d}x = p',$$

and

$$B_y = \frac{(p')^2}{1-p'}\left(\frac{\sigma_1}{\sqrt{2\pi}\sigma_2^2}\right)^d \int \exp\left(\frac{-\|x-\mu_1\|_2^2}{\sigma_2^2} + \frac{\|x-\mu_1\|_2^2}{2\sigma_1^2}\right) \mathrm{d}x$$
$$= \frac{(p')^2}{1-p'}\left(\frac{\sigma_1}{\sqrt{2\pi}\sigma_2^2}\right)^d \left(\frac{\pi}{\frac{1}{\sigma_2^2}-\frac{1}{2\sigma_1^2}}\right)^{d/2} \leqslant \frac{(p')^2}{1-p'}\left(\frac{\sigma_1}{\sigma_2}\right)^d,$$

using the fact that $\frac{1}{2\sigma_1^2} \leqslant \frac{1}{2\sigma_2^2}$. Under the assumption (E3), we get that $B_y = O(1/n)$.

For ease of calculation, we now set

$$\mu_1^T = (\mu, 0, \ldots, 0),$$
$$\mu_2^T = (-\mu, 0, \ldots, 0). \tag{E.4}$$

Using the formula

$$\int \exp\left(-x^T A x + b^T x + c\right) \mathrm{d}x = \sqrt{\frac{\pi^d}{\det(A)}} \exp\left(\frac{1}{4} b^T A^{-1} b + c\right), \tag{E.5}$$

we have

$$A_z = p'\left(\frac{1}{\sqrt{2\pi}\sigma_2}\right)^d \int \exp\left(\frac{-\|x-\mu_2\|_2^2}{2\sigma_2^2} + \frac{\|x-\mu_2\|_2^2}{2\sigma_1^2} - \frac{\|x-\mu_1\|_2^2}{2\sigma_1^2}\right) \mathrm{d}x$$
$$= p' \exp\left(\frac{\sigma_2^2}{2}\left(\left\|\frac{\mu_2}{\sigma_2^2} - \frac{\mu_2}{\sigma_1^2} + \frac{\mu_1}{\sigma_1^2}\right\|_2^2 - \frac{\mu_2^T\mu_2}{2\sigma_2^2} + \frac{\mu_2^T\mu_2}{2\sigma_1^2} - \frac{\mu_1^T\mu_1}{2\sigma_1^2}\right)\right)$$
$$= p' \exp\left(-2\mu^2\left(\frac{1}{\sigma_1^2} - \frac{\sigma_2^2}{\sigma_1^4}\right)\right).$$

In particular, using the fact that $\exp(-x) \geqslant 1 - x$ for $x \geqslant 0$, we have

$$A_y - A_z = p' - A_z \leqslant p' \cdot 2\mu^2\left(\frac{1}{\sigma_1^2} - \frac{\sigma_2^2}{\sigma_1^4}\right) \leqslant \frac{2\mu^2}{\sigma_1^2}.$$

We can use the simple fact that $B_z \geqslant 0$ to ensure that $B_y - B_z \leqslant B_y = O(1/n)$.

Combining the inequalities, we conclude that

$$\mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[y] - \mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[z] = O\left(\frac{\mu^2}{\sigma_1^2}\right) + O\left(\frac{1}{n}\right).$$

Finally, we compute

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[z^2] &= \left(\frac{p'}{1-p'}\right)^2\left(\frac{\sigma_1}{\sigma_2}\right)^{2d}\left((1-p')\int \exp\left(\frac{-\|x-\mu_2\|_2^2}{\sigma_2^2}+\frac{\|x-\mu_2\|_2^2}{\sigma_1^2}\right)\right.\\ &\qquad\cdot\frac{1}{(\sqrt{2\pi}\sigma_1)^d}\exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_1^2}\right)\mathrm{d}x\\ &\qquad\left.+p'\int\exp\left(\frac{-\|x-\mu_2\|_2^2}{\sigma_2^2}+\frac{\|x-\mu_2\|_2^2}{\sigma_1^2}\right)\frac{1}{(\sqrt{2\pi}\sigma_2)^d}\exp\left(\frac{-\|x-\mu_1\|_2^2}{2\sigma_2^2}\right)\mathrm{d}x\right)\\ &:= A_z'+B_z'.\end{aligned}$$

Again using the designation (E.4) and the formula (E.5), we have

$$\begin{aligned}A_z' &= \frac{(p')^2}{1-p'}\left(\frac{\sigma_1}{\sqrt{2\pi}\sigma_2^2}\right)^d\int\exp\left(\frac{-\|x-\mu_2\|_2^2}{\sigma_2^2}+\frac{\|x-\mu_2\|_2^2}{\sigma_1^2}-\frac{\|x-\mu_1\|_2^2}{2\sigma_1^2}\right)\mathrm{d}x\\ &= \frac{(p')^2}{1-p'}\left(\frac{\sigma_1}{\sqrt{2}\sigma_2^2\sqrt{\frac{1}{\sigma_2^2}-\frac{1}{2\sigma_1^2}}}\right)^d\\ &\quad\exp\left(\frac{1}{4\left(\frac{1}{\sigma_2^2}-\frac{1}{2\sigma_1^2}\right)}\left\|\frac{2\mu_2}{\sigma_2^2}-\frac{2\mu_2}{\sigma_1^2}+\frac{\mu_1}{\sigma_1^2}\right\|_2^2-\frac{\mu_2^T\mu_2}{\sigma_2^2}+\frac{\mu_2^T\mu_2}{\sigma_1^2}-\frac{\mu_1^T\mu_1}{2\sigma_1^2}\right)\\ &= \frac{(p')^2}{1-p'}\left(\frac{\sigma_1}{\sqrt{2}\sigma_2^2\sqrt{\frac{1}{\sigma_2^2}-\frac{1}{2\sigma_1^2}}}\right)^d\exp\left(\frac{(-2\mu/\sigma_2^2+3\mu/\sigma_1^2)^2}{4\left(\frac{1}{\sigma_2^2}-\frac{1}{2\sigma_1^2}\right)}-\frac{\mu^2}{\sigma_2^2}+\frac{\mu^2}{2\sigma_1^2}\right)\\ &\leqslant \frac{(p')^2}{1-p'}\left(\frac{\sigma_1}{\sigma_2}\right)^d\exp\left(\frac{(-2\mu/\sigma_2^2+3\mu/\sigma_1^2)^2}{4\left(\frac{1}{\sigma_2^2}-\frac{1}{2\sigma_1^2}\right)}-\frac{\mu^2}{\sigma_2^2}+\frac{\mu^2}{2\sigma_1^2}\right),\end{aligned}$$

and

$$\begin{aligned}
B'_z &= \frac{(p')^3}{1-p'}\left(\frac{\sigma_1^2}{\sqrt{2\pi}\sigma_2^3}\right)^d \int \exp\left(\frac{-\|x-\mu_2\|_2^2}{\sigma_2^2} + \frac{\|x-\mu_2\|_2^2}{\sigma_1^2} - \frac{\|x-\mu_1\|_2^2}{2\sigma_2^2}\right) \mathrm{d}x \\
&= \frac{(p')^3}{1-p'}\left(\frac{\sigma_1^2}{\sqrt{2}\sigma_2^3\sqrt{\frac{3}{2\sigma_2^2}-\frac{1}{\sigma_1^2}}}\right)^d \\
&\quad \exp\left(\frac{1}{4\left(\frac{3}{2\sigma_2^2}-\frac{1}{\sigma_1^2}\right)}\left\|\frac{2\mu_2}{\sigma_2^2}-\frac{2\mu_2}{\sigma_1^2}+\frac{\mu_1}{\sigma_2^2}\right\|_2^2 - \frac{\mu_2^T\mu_2}{\sigma_2^2}+\frac{\mu_2^T\mu_2}{\sigma_1^2}-\frac{\mu_1^T\mu_1}{2\sigma_2^2}\right) \\
&= \frac{(p')^3}{1-p'}\left(\frac{\sigma_1^2}{\sqrt{2}\sigma_2^3\sqrt{\frac{3}{2\sigma_2^2}-\frac{1}{\sigma_1^2}}}\right)^d \exp\left(\frac{(-\mu/\sigma_2^2+2\mu/\sigma_1^2)^2}{4\left(\frac{3}{2\sigma_2^2}-\frac{1}{\sigma_1^2}\right)} - \frac{3\mu^2}{2\sigma_2^2}+\frac{\mu^2}{\sigma_1^2}\right) \\
&\leqslant \frac{(p')^3}{1-p'}\left(\frac{\sigma_1}{\sigma_2}\right)^{2d} \exp\left(\frac{(-\mu/\sigma_2^2+2\mu/\sigma_1^2)^2}{4\left(\frac{3}{2\sigma_2^2}-\frac{1}{\sigma_1^2}\right)} - \frac{3\mu^2}{2\sigma_2^2}+\frac{\mu^2}{\sigma_1^2}\right).
\end{aligned}$$

Considering the exponential terms in the expressions for A'_z and B'_z, note that for A'_z, we have

$$\frac{(-2\mu/\sigma_2^2+3\mu/\sigma_1^2)^2}{4\left(\frac{1}{\sigma_2^2}-\frac{1}{2\sigma_1^2}\right)} - \frac{\mu^2}{\sigma_2^2} = \frac{\mu^2}{\sigma_2^2}\left(\frac{\left(2-\frac{3\sigma_2^2}{\sigma_1^2}\right)^2}{4\left(1-\frac{\sigma_2^2}{2\sigma_1^2}\right)} - 1\right) < 0,$$

assuming $\sigma_2 \leqslant \sigma_1$, whereas for B'_z, we have

$$\frac{(-\mu/\sigma_2^2+2\mu/\sigma_1^2)^2}{4\left(\frac{3}{2\sigma_2^2}-\frac{1}{\sigma_1^2}\right)} - \frac{3\mu^2}{2\sigma_2^2} = \frac{\mu^2}{\sigma_2^2}\left(\frac{\left(1-\frac{2\sigma_2^2}{\sigma_1^2}\right)^2}{4\left(\frac{3}{2}-\frac{\sigma_2^2}{\sigma_1^2}\right)} - \frac{3}{2}\right) < 0,$$

using the fact that $\sigma_2 \leqslant \sigma_1$. Thus, using the assumption (E.3), we obtain

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim\mathbb{P}_1}[z^2] = A'_z + B'_z &= O\left(\frac{1}{n}\right)\exp\left(\frac{\mu^2}{2\sigma_1^2}\right) + O\left(\frac{1}{n^2 p}\right)\exp\left(\frac{\mu^2}{\sigma_1^2}\right) \\
&= O\left(\frac{1}{n}\right)\exp\left(\frac{\mu^2}{\sigma_1^2}\right).
\end{aligned}$$

Finally, we take $\mu = \frac{\sigma_1}{\sqrt{n}}$ to obtain the desired bound (E.2). This completes the proof.

E.4 *Proof of Theorem 5.5*

By a similar argument used to derive the bound in Theorem 5.3, the following expected error bound may be derived from the high-probability bound in Theorem 4.4 for the hybrid estimator:

$$\mathbb{E}\,\|\widehat{\mu}_{k_1,k_2}\|_2 \leqslant \min\left\{\sqrt{d}r_{2k_1,1}, \sqrt{n}^{1/d} r_{k_2}\right\}. \tag{E.6}$$

In what follows, we will bound these expressions to obtain the desired results.

As shown in the proof of Lemma 2.1(v), a ball of radius $C\sigma_2\sqrt{d}$ around the origin will contain at least $\frac{1}{2}$ of the mass of np distributions. Thus, if $np \geqslant 2k_2$, we will have $r_{k_2} \leqslant C\sigma_2\sqrt{d}$.

We now claim that $r_{2k_1,1} \leqslant \frac{C\sigma_1 \log n}{\sqrt{n}} := r'$, which we will show by integrating the marginal densities on the interval $[-r', r']$. Note that $v_i \leqslant \sigma_1$ for all i. We consider two cases: if $v_i \geqslant r'$, then $q_i(r') \geqslant \frac{c}{v_i} \geqslant \frac{c}{\sigma_1}$, using inequality (5.5), so $\int_{[-r',r']} q_i(x)\,\mathrm{d}x \geqslant \frac{2cr'}{\sigma_1} \geqslant \frac{2\log n}{\sqrt{n}}$ for large enough C. If $v_i < r'$, then $\int_{[-v_i,v_i]} q(x)\,\mathrm{d}x \geqslant c' \geqslant \frac{2\log n}{\sqrt{n}}$, as well. Thus,

$$\sum_{i=1}^{n} \int_{[-r',r']} q_i(x)\,\mathrm{d}x \geqslant \sum_{i=1}^{n} \frac{2\log n}{\sqrt{n}} \geqslant 2\sqrt{n}\log n = 2k_1. \tag{E.7}$$

Combining the results with inequality (E.6) proves inequality (5.7).

We now consider the special cases:

(a) In the case when $p = \Omega\left(\frac{\sqrt{n}\log n}{n}\right)$, we can use fact that at least $np = \Omega(\sqrt{n}\log n)$ points have marginal variance at most σ_2. Let $r' := \frac{C\sigma_2 \log n}{p\sqrt{n}}$. By similar reasoning as above, for at least np distributions, we have $\int_{[-r',r']} q_i(x)\,\mathrm{d}x \geqslant \frac{\log n}{p\sqrt{n}}$. Thus, we can replace inequality (E7) by

$$\sum_{i=1}^{n} \int_{[-r',r']} q_i(x)\,\mathrm{d}x \geqslant np \cdot \frac{2\log n}{p\sqrt{n}} \geqslant 2\sqrt{n}\log n,$$

to conclude that $r_{2k_1,1} = O\left(\frac{\sigma_2 \log n}{p\sqrt{n}}\right)$. This leads to the stated bound.

(b) In this case, we will obtain a better bound by showing that $\|\widehat{\mu}_{S,k_2}\|_2 \leqslant r_{2k_2}$, w.h.p., rather than the looser bound $\|\widehat{\mu}_{S,k_2}\|_2 \leqslant C'\sqrt{n}^{1/d} r_{k_2}$ used to derive inequality (E.6) (cf. Theorem 4.4). Since $r_{2k_2} \leqslant C\sigma_2\sqrt{d}$, the tighter bound will then follow.

(b) Let $r' := C'\sqrt{d\log n}\sigma_2$. As argued in the proof of Theorem 4.3, it suffices to show that $R(f_{r',r_{2k}}) \leqslant \frac{k}{2n}$, where $k = k_2$. We will deal with low-variance and high-variance points separately.

(b) First, consider i such that $v_i = \Omega(\sigma_1) = \Omega(\sigma_2 n^{\frac{1}{d}}) \geqslant C''\sigma_2 n^{\frac{1}{d}}$ for large C'', and let v_d denote the volume of the ball of radius 1. Then

$$\mathbb{P}\left(X_i \in B(r', r_{2k})\right) \leqslant \mathbb{P}\left(X_i \in B(0, r_{2k})\right) \leqslant f_i(0) v_d r_{2k}^d \leqslant \left(\frac{c'}{C''\sigma_2 n^{1/d}}\right)^d v_d \sigma_2^d C^d \sqrt{d}^d \leqslant \frac{1}{n},$$

where we use condition (5.6) and the fact that $\frac{v_d\sqrt{d}^d}{\tilde{C}^d} \leqslant 1$ for a sufficiently large constant $\tilde{C}$.

(b) Now consider i such that $v_i \leqslant \sigma_2$. By condition (5.6), we have

$$\mathbb{P}\left(X_i \in B(r', r_{2k})\right) \leqslant \exp(-c_1 \log n) \leqslant \frac{1}{n^{c_1}}.$$

For large enough C', we can ensure that $c_1 \geqslant 1$. Altogether, we conclude that

$$R(f_{r',r_{2k}}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(X_i \in B(r', r_{2k})\right) \leqslant \frac{1}{n} < \frac{k_2}{2n},$$

which concludes the proof.

E.5 *Details for Table 2*

1. Large heterogeneity
   - Upper bound: we have $\frac{\sigma_1}{\sigma_2} = \Omega(n^{1/d})$. Since $\sigma_2 = 1$, Theorem 5.5(b) states that the error of the hybrid estimator is bounded as follows:

$$\mathbb{E}\,\|\widehat{\mu} - \mu\|_2 \leqslant C''_u \sqrt{d}\sqrt{\log n}.$$

   - Lower bound: as remarked after Theorem 5.5, the lower bounds for the class $\mathscr{P}(\sigma_1, \sigma_2, p)$ also hold for the class $\mathscr{Q}(\sigma_1, \sigma_2, p)$ because these families share the class of distributions used in the proof of Theorem 5.4. Using Theorem 5.4(a), the error of any estimator $\widehat{\mu}$ is bounded from below as follows:

$$\mathbb{E}\,\|\widehat{\mu} - \mu\|_2 \geqslant C_\ell \sqrt{d} \min\left(\frac{1}{\sqrt{np}}, \frac{\sigma_1}{\sqrt{n}}\right) = C_\ell \frac{\sqrt{d}}{\sqrt{np}} \min\left(1, \sigma_1\sqrt{p}\right) = \Omega\left(\frac{\sqrt{d}}{\sqrt{np}}\right),$$

   where we use the fact that $\sigma_1\sqrt{p} = \Omega(1)$ by assumption.

2. Mild heterogeneity
   - Upper bound: since $\sigma_2 = 1$, inequality (5.7) in Theorem 5.5 states that the error of the hybrid estimator is bounded as follows:

$$\mathbb{E}\,\|\widehat{\mu} - \mu\|_2 \leqslant C''_u \sqrt{d}\sigma_1 \frac{\log n}{\sqrt{n}}.$$

   - Lower bound: using Theorem 5.4(a), the error of any estimator $\widehat{\mu}$ is bounded from below as follows:

$$\mathbb{E}\,\|\widehat{\mu} - \mu\|_2 \geqslant C_\ell \sqrt{d} \min\left(\frac{1}{\sqrt{np}}, \frac{\sigma_1}{\sqrt{n}}\right) = C_\ell \frac{\sqrt{d}}{\sqrt{n}} \min\left(\frac{1}{\sqrt{p}}, \sigma_1\right) = \Omega\left(\frac{\sqrt{d}\sigma_1}{\sqrt{n}}\right),$$

   where we use the fact that $\sigma_1 = O(1/\sqrt{p})$ by assumption.

3. Large p

   - Upper bound: as $p = \Omega\left(\frac{\sqrt{n}\log n}{n}\right)$, Theorem 5.5(a) states that the error of the hybrid estimator is bounded as follows:

$$\mathbb{E}\,\|\widehat{\mu} - \mu\|_2 \leqslant C'_u\sqrt{d}\log n \min\left(\frac{1}{p\sqrt{n}}, \frac{\sigma_1}{\sqrt{n}}\right).$$

   - Lower bound: the lower bound follows directly from Theorem 5.4(a).

F. Proofs for alternative conditions

In this appendix, we prove the statements of the results in Section 7.

F.1 *Proof of Theorem 7.1*

We first prove claim (i). Note that the result of Lemma B.2 will still hold, since it only depends on the uniform concentration bound and optimality of the modal interval estimator. Thus, $R(f_{\widehat{\mu}_{M,r},r}) \geqslant \frac{R^*_r}{2}$, w.h.p.

For a fixed value of r', define $\widehat{\mu}' = \frac{\widehat{\mu}_{M,r}}{\|\widehat{\mu}_{M,r}\|_2} \cdot r'$ to be the rescaled version of $\widehat{\mu}_{M,r}$. By condition (C1), we will have $\|\widehat{\mu}_{M,r}\|_2 \leqslant r'$ if we can show that $R(f_{\widehat{\mu}',r}) \leqslant R(f_{\widehat{\mu}_{M,r},r})$. Note that

$$R(f_{\widehat{\mu}',r}) \leqslant g(r', r),$$

so if we choose r' sufficiently large so that $g(r', r) < \frac{R^*_r}{2}$, the desired inequality will hold.

Turning to claim (ii), note that Lemma 4.1 continues to hold, since it only relies on the uniform concentration bound and a Chernoff bound. We thus conclude that $R(f_{\widehat{\mu}_{S,k},r_{2k}}) \geqslant \frac{k}{4n} = \frac{R^*_{2k}}{4}$, w.h.p. For a fixed value of r', we define $\widehat{\mu}' = \frac{\widehat{\mu}_{M,r_{2k}}}{\|\widehat{\mu}_{M,r_{2k}}\|_2} \cdot r'$. By condition (C1) (which we only need to assume holds for $r = r_{2k}$), if $R(f_{\widehat{\mu}',r_{2k}}) \leqslant R(f_{\widehat{\mu}_{M,r_{2k}},r_{2k}})$, then $\|\widehat{\mu}_{M,r_{2k}}\|_2 \leqslant r'$. Furthermore, $R(f_{\widehat{\mu}',r_{2k}}) \leqslant g(r', r_{2k})$, so we simply need to choose r' such that $g(r', r_{2k}) < \frac{1}{4}$.

For the hybrid estimator, note that Lemma D.1 shows that the output is always within $\sqrt{d}r_{4\sqrt{n\log n},1}$ of the output. Furthermore, the output of shorth estimator is always with r' of the origin by part (ii). If the shorth estimator lies outside the $S^\infty_{\sqrt{n\log n}}$, then its ℓ_2 projection on $S^\infty_{\sqrt{n\log n}}$ will only decrease its distance from the origin because (1) the origin belongs to $S^\infty_{\sqrt{n\log n}}$; and (2) $S^\infty_{\sqrt{n\log n}}$ is convex.

F.2 *Proof of Proposition 7.2*

We first show that for each $r > 0$, the functions $R_i(f_{x,r})$ are unimodal as functions of $x \in \mathbb{R}^d$. Let q be the uniform distribution on the Euclidean ball of radius r. Then $p_i \star q$, being a convolution of two log-concave densities, is also log-concave. Log-concave densities by definition are proportional to $e^{-\phi(x)}$ for some convex function ϕ, and therefore they are unimodal and monotonically decreasing along rays from the mode. Now note that if condition (C3) holds, then $R_i(f_{x,r})$ must also be symmetric around 0. Hence, if $R_i(f_{x,r})$ is unimodal, its unique mode must clearly occur at 0. This proves that conditions (C2) and (C3) together imply condition (C1).

For the second statement, it suffices to verify the inequality

$$\sup_{\|x\|_2=a} R_i(f_{x,r}) \leqslant \frac{1}{\lfloor a/2r \rfloor}, \qquad \forall i. \tag{F.1}$$

Indeed, we would then have

$$g(a,r) = \sup_{\|x\|_2=a} \frac{1}{n}\sum_{i=1}^{n} R_i(f_{x,r}) \leqslant \frac{1}{n}\sum_{i=1}^{n} \sup_{\|x\|_2=a} R_i(f_{x,r}) \leqslant \frac{1}{\lfloor a/2r \rfloor}.$$

Thus, it remains to verify inequality (F.1). Focusing on a particular i, consider $x \in \mathbb{R}^d$ such that $\|x\|_2 = a$. We know that $R_i(f_{x,r})$ is decreasing on the ray from 0 to x. Furthermore, we can pack $\lfloor \frac{a}{2r} \rfloor$ balls of radius r on the ray, including the balls $B(x_i^*, r)$ and $B(x, r)$ at the endpoints. The total mass of these balls is clearly upper bounded by 1. Hence,

$$\left\lfloor \frac{a}{2r} \right\rfloor \cdot R_i(f_{x,r}) \leqslant 1,$$

implying the desired result.

F.3 *Proof of Proposition 7.4*

Let X have an elliptically symmetric density defined as $p_X(x) = f(x^T \Sigma^{-1} x)$ for a decreasing function $f : \mathbb{R} \to \mathbb{R}$. Consider a point $x_0 \in \mathbb{R}^d$ such that $\|x_0\|_2 = r_2$, and consider the ball $B(x_0, r_1) = \{x \in \mathbb{R}^d \ : \ \|x - x_0\| \leqslant r_1\}$. For analysis purposes, we first transform the elliptically symmetric density to a spherically symmetric, decreasing density. This may be achieved by applying the linear transformation $\Sigma^{-1/2} : \mathbb{R}^d \to \mathbb{R}^d$. Define $Y := \Sigma^{-1/2} X$, let $\Sigma^{-1/2} x_0 = y_0$, and let $\hat{B}$ be the image of $B(x_0, r_1)$ under the transformation $\Sigma^{-1/2}$. Note that

$$\hat{B} = \left\{ y \in \mathbb{R}^d \ : \ (y - y_0)^T \Sigma (y - y_0) \leqslant r_1 \right\},$$

and further note that $R(f_{x_0,r_1})$ is equal to the integral of $p_Y(\cdot)$ over $\hat{B}$; i.e., $\mathbb{P}(Y \in \hat{B})$. It is easy to see that $\hat{B} \subseteq B\left(y_0, \frac{r_1}{\lambda_{\min}(\Sigma)}\right)$. Hence,

$$R(f_{x_0,r_1}) = \mathbb{P}(Y \in \hat{B}) \leqslant \mathbb{P}(Y \in B\left(y_0, \frac{r_1}{\lambda_{\min}(\Sigma)}\right).$$

We may now use the strategy from Lemma 2.1, to obtain

$$\begin{aligned}
1 &\geqslant \mathbb{P}(Y \in B(0, \|y_0\|_2)) \\
&\geqslant P\left(B(0, \|y_0\|_2), \frac{r_1}{\lambda_{\min}(\Sigma)}\right) \cdot \mathbb{P}\left(Y \in B\left(y_0, \frac{r_1}{\lambda_{\min}(\Sigma)}\right)\right) \\
&\geqslant P\left(B\left(0, \frac{r_2}{\lambda_{\max}(\Sigma)}\right), \frac{r_1}{\lambda_{\min}(\Sigma)}\right) \cdot R(f_{x_0,r_1}).
\end{aligned}$$

Since this inequality holds for any x_2 with $\|x_2\|_2 = r_2$, we conclude that

$$g(r_2, r_1) \leqslant \frac{1}{P\left(B\left(0, \frac{r_2}{\lambda_{\max}(\Sigma)}\right), \frac{r_1}{\lambda_{\min}(\Sigma)}\right)} \leqslant C\left(\frac{r_1\lambda_{\max}(\Sigma)}{r_2\lambda_{\min}(\Sigma)}\right)^d.$$

F.4 *Proof of Proposition 7.6*

We index the distributions so that $\{R_i\}_{i=1}^s$ are radially symmetric. Note that

$$g(a, r) = \sup_{\|x\|_2=a} R(f_{x,r}) \leqslant \frac{1}{n}\sum_{i=1}^n \sup_{\|x\|_2=a} R_i(f_{x,r}).$$

Furthermore, for each $1 \leqslant i \leqslant s$, we have

$$\sup_{\|x\|_2=a} R_i(f_{x,r}) \leqslant \left(\frac{r}{a}\right)^d R_i(f_{0,a}) \leqslant \left(\frac{r}{a}\right)^d.$$

On the other hand, for $i > s$, we have

$$\sup_{\|x\|_2=a} R_i(f_{x,r}) \leqslant \frac{r}{a}.$$

Hence,

$$g(a, r) \leqslant \frac{s}{n}\left(\frac{r}{a}\right)^d + \frac{n-s}{n}\left(\frac{r}{a}\right).$$

Now note that $R^*_{q_{(f(n))}} \geqslant \frac{f(n)}{2n}$. Thus,

$$g(r', r) \leqslant \frac{s}{n}\cdot\frac{1}{2^d n} + \frac{n-s}{n}\cdot\frac{1}{2n^{1/d}} \leqslant \frac{1}{n} + \frac{n-s}{n}\cdot\frac{1}{2n^{1/d}} < \frac{f(n)}{4n} \leqslant \frac{R^*_r}{2},$$

using the assumed lower bound on s.

G. Proofs for regression

In this appendix, we provide the proofs of the statements in Section 8.

G.1 *Proof of Proposition 8.1*

We write

$$\sum_{i=1}^n \mathbb{E}\left[1\left\{|y_i - x_i^T\beta| \leqslant r\right\}\right] = \sum_{i=1}^n \mathbb{P}\left(|y_i - x_i^T\beta| \leqslant r\right) = \sum_{i=1}^n \mathbb{P}\left(|x_i^T(\beta^* - \beta) + \epsilon_i| \leqslant r\right).$$

Note that conditioned on x_i, each summand is maximized uniquely when $x_i^T(\beta^* - \beta) = 0$, since the distribution of ϵ_i is symmetric and unimodal. Since

$$\sum_{i=1}^{n} \mathbb{E}\left[1\left\{|y_i - x_i^T\beta| \leqslant r\right\}\right] = \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}\left[1\left\{|y_i - x_i^T\beta| \leqslant r\right\} \mid \{x_i\}_{i=1}^n\right]\right], \tag{G.1}$$

we see that the right-hand expression in equation (G.1) is therefore maximized when $\beta = \beta^*$. On the other hand, we can also argue that the maximizer is unique. Indeed, suppose $\beta \in \mathbb{R}^d$ were such that $\beta \neq \beta^*$. The set $\mathscr{S} := \left\{\{x_i\}_{i=1}^n \subseteq (\mathbb{R}^d)^n : x_i^T(\beta - \widehat{\beta}) = 0 \quad \forall i\right\}$ has Lebesgue measure 0. We can write

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}\left[1\left\{|y_i - x_i^T\beta| \leqslant r\right\} \mid \{x_i\}_{i=1}^n\right]\right] = \int_{\{x_i\}\in\mathscr{S}} \mathbb{E}\left[1\left\{|y_i - x_i^T\beta| \leqslant r\right\} \mid \{x_i\}_{i=1}^n\right] \mathrm{d}\mathbb{P}(\{x_i\})$$
$$+ \int_{\{x_i\}\notin\mathscr{S}} \mathbb{E}\left[1\left\{|y_i - x_i^T\beta| \leqslant r\right\} \mid \{x_i\}_{i=1}^n\right] \mathrm{d}\mathbb{P}(\{x_i\}).$$

Noting that

$$\mathbb{E}\left[1\left\{|y_i - x_i^T\beta| \leqslant r\right\} \mid \{x_i\}_{i=1}^n\right] = \mathbb{E}\left[1\left\{|y_i - x_i^T\beta^*| \leqslant r\right\} \mid \{x_i\}_{i=1}^n\right], \quad \forall\{x_i\} \in \mathscr{S},$$
$$\mathbb{E}\left[1\left\{|y_i - x_i^T\beta| \leqslant r\right\} \mid \{x_i\}_{i=1}^n\right] < \mathbb{E}\left[1\left\{|y_i - x_i^T\beta^*| \leqslant r\right\} \mid \{x_i\}_{i=1}^n\right], \quad \forall\{x_i\} \notin \mathscr{S},$$

completes the proof.

G.2 *Proof of Theorem 8.2*

The proof follows the same approach used to prove estimation error bounds for the modal interval estimator throughout the paper (e.g., Theorem 3.1). By Lemma 8.1, we know that $R_{\widehat{\beta}} \geqslant \frac{R_{\beta^*}}{2}$, w.h.p. We will be done if we can show that $R_\beta < \frac{R_{\beta^*}}{2}$ for all β satisfying

$$\|\beta - \beta\|_2 > \frac{c' n \sigma_{(cd\log n)}}{\lambda_{\min}}. \tag{G.2}$$

First note that

$$R_{\beta^*} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(|\epsilon_i| \leqslant r).$$

Hence, as argued for mean estimation, we certainly have $r \leqslant C'\sigma_{(Cd\log n)}$.

Also note that for any $\beta \in \mathbb{R}^d$, we have

$$y_i - x_i^T\beta = \epsilon_i + x_i^T(\beta^* - \beta) \sim N\left((\beta^* - \beta)^T\mu_i', (\beta^* - \beta)^T\Sigma_i'(\beta^* - \beta)\right).$$

Let $\mathscr{J}$ denote the set of indices of the smallest $d\log n$ of the σ_is. Note that

$$R_{\beta^*} \geqslant \frac{1}{n}\sum_{i\in\mathscr{J}} \mathbb{P}(|\epsilon_i| \leqslant r) \geqslant 2r \cdot \frac{c}{n}\sum_{i=1}^{d\log n} \frac{1}{\sqrt{2\pi}\sigma_{(i)}},$$

since the Gaussian pdf decreases by a factor of $\approx 68\%$ within one standard deviation of 0.

Now suppose $\beta \in \mathbb{R}^d$ satisfies inequality (G.2). We have

$$R_\beta \leqslant \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(|z_i| \leqslant r\right),$$

where $z_i \sim N\left(0, \sigma_i^2 + (\beta^* - \beta)^T \Sigma_i'(\beta^* - \beta)\right)$. For $i \notin \mathscr{J}$, we write

$$\mathbb{P}(|z_i| \leqslant r) \leqslant 2r \cdot \frac{1}{\sqrt{2\pi}\sqrt{\sigma_i^2 + (\beta^* - \beta)^T \Sigma_i'(\beta^* - \beta)}} \leqslant \frac{2r}{n\sigma_{(d\log n)}\sqrt{2\pi}},$$

since by the choice of β, we have

$$(\beta^* - \beta)^T \Sigma_i'(\beta^* - \beta) \geqslant \lambda_{\min}\|\beta - \beta^*\|_2^2 \geqslant n^2\sigma_{(d\log n)}^2.$$

For $i \in \mathscr{J}$, we write

$$\mathbb{P}(|z_i| \leqslant r) \leqslant 2r \cdot \frac{1}{\sqrt{2\pi}\sqrt{\sigma_i^2 + (\beta^* - \beta)^T \Sigma_i'(\beta^* - \beta)}} \leqslant \frac{2r}{3\sigma_i^2\sqrt{2\pi}},$$

since by the choice of β, we have

$$(\beta^* - \beta)^T \Sigma_i'(\beta^* - \beta) \geqslant 2\sigma_{(d\log n)}^2 \geqslant 2\sigma_i^2.$$

Thus, we conclude that

$$R_\beta \leqslant \frac{2r}{\sqrt{2\pi}} \cdot \frac{1}{n}\left(\sum_{i\in\mathscr{J}} \frac{1}{3\sigma_i^2} + \sum_{i\notin\mathscr{J}} \frac{1}{n\sigma_{(d\log n)}}\right) \leqslant \frac{R_{\beta^*}}{3} + \frac{c'}{n} < \frac{R_{\beta^*}}{2},$$

as wanted. This concludes the proof.

G.3 *Proof of Theorem 8.3*

For $i \in [n]$, consider the sets

$$U_i := \{\beta \subseteq \mathbb{R}^d : -r \leqslant x_i^T\beta \leqslant +r\}.$$

The set U_i is sandwiched between the two hyperplanes $x_i^T\beta = y_i - r$ and $x_i^T\beta = y_i + r$. Denote these hyperplanes by $H_-(U_i)$ and $H_+(U_i)$, respectively. These $2n$ hyperplanes partition $\mathbb{R}^d$ into a finite number of (possibly unbounded) convex regions, which we denote by $\{R_1, \ldots, R_M\}$. Define the function $f(\beta) := \sum_{i=1}^n \mathbb{1}_{U_i}(\beta)$. Our goal is to find $\hat{\beta} = \operatorname{argmax}_{\beta\in\mathbb{R}^d} f(\beta)$, where $\mathbb{1}_{U_i}$ is the indicator function of U_i. It is easy to see that $f(\cdot)$ is constant when restricted to the interior of any fixed region R_j for $j \in [M]$. Also, since $\mathbb{1}_{U_i}$ is an upper-semicontinuous function for each $i \in [n]$, so is f. Thus, the value of $f(\cdot)$ at the vertices R_j is at least as large as the value of f in its interior. Thus, to find the maximum of $f(\cdot)$, we may only consider $\beta \in \mathbb{R}^d$ that correspond to vertices of R_j for $j \in [M]$. All such vertices may be obtained by choosing any d (mutually non-parallel) hyperplanes from among $\{H_-(U_1), \ldots, H_-(U_n), H_+(U_1), \ldots, H_+(U_M)\}$ and considering their point of intersection. The total number of such points is bounded above by $\binom{2n}{d}$, and our algorithm may simply list such points and evaluate f at each point in the list.

H. Auxiliary results

This appendix contains several technical results invoked throughout the paper.

We will employ the following multiplicative Chernoff bound, which is standard (cf. Vershynin [45] or Boucheron *et al.* [6]):

LEMMA H.1 Let $X_1, \dots, X_n$ be independent Bernoulli random variables with parameters $\{p_i\}$. Let $S_n = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[S_n]$.

(i) For any $\delta \in (0, 1]$, we have

$$\mathbb{P}\left(S_n \geqslant (1+\delta)\mu\right) \leqslant \exp\left(-\frac{\mu\delta^2}{3}\right).$$

and

$$\mathbb{P}\left(S_n \leqslant (1-\delta)\mu\right) \leqslant \exp\left(-\frac{\mu\delta^2}{2}\right).$$

(ii) For $\delta \geqslant 4$, we have

$$\mathbb{P}\left(S_n \geqslant \delta\mu\right) \leqslant \exp\left(-c\mu\delta\log\delta\right).$$

We will also use the following result from Boucheron *et al.* [6]:

LEMMA H.2 (Boucheron *et al.* [6, Theorem 12.9]) Let $W_1, \dots, W_n$ be independent vector-valued random variables and let $Z = \sup_{s\in\mathcal{T}} \sum_{i=1}^n W_{i,s}$. Assume that for all $i \leqslant n$ and $s \in \mathcal{T}$, we have $\mathbb{E}\, W_{i,s} = 0$, and $|W_{i,s}| \leqslant 1$. Let

$$v := 2\,\mathbb{E}\, Z + \rho^2,$$

$$\rho^2 := \sup_{t\in T} \sum_{i=1}^n \mathbb{E}\, W_{i,s}^2.$$

Then $\mathbb{V}(Z) \leqslant v$ and

$$\mathbb{P}\{Z \geqslant \mathbb{E}\, Z + t\} \leqslant \exp\left(-\frac{t}{4}\log\left(1 + 2\log\left(1 + \frac{t}{v}\right)\right)\right).$$

We now state and prove a generalization of Boucheron *et al.* [6, Theorem 13.7]:

THEOREM H.1 Let $\mathcal{A} = \{A_t : t \in \mathcal{T}\}$ be a countable class of measurable subsets of $\mathcal{X}$ with VC dimension V, such that $A_0 = \emptyset \in \mathcal{A}$. Let $X_1, \dots, X_n$ be independent random variables taking values in $\mathcal{X}$, with distributions $P_1, \dots, P_n$, respectively. Assume that for some $\sigma > 0$, we have

$$\frac{1}{n}\sum_{i=1}^n P_i(A_t) \leqslant \sigma^2, \text{ for every} t \in \mathcal{T}.$$

Let Z and Z^- be defined as follows:

$$Z = \frac{1}{\sqrt{n}} \sup_{t\in\mathcal{T}} \sum_{i=1}^n \left(\mathbb{1}_{X_i\in A_t} - P_i(A_t)\right), \quad \text{and}$$

$$Z^- = \frac{1}{\sqrt{n}} \sup_{t\in\mathcal{T}} \sum_{i=1}^n \left(P_i(A_t) - \mathbb{1}_{X_i\in A_t}\right).$$

If $\sigma \geqslant 24\sqrt{\frac{V}{5n}\log\left(\frac{4e^2}{\sigma}\right)}$, then

$$\max\left(\mathbb{E}\,Z, \mathbb{E}\,Z^{-}\right) \leqslant 72\sigma\sqrt{V\log\frac{4e^2}{\sigma}}.$$

Proof. The following proof is an adaptation of the proof of Theorem 13.7 in Boucheron *et al.* [6]. The generalization from identical to non-identical distributions is possible because (1) independence suffices for symmetrization inequality and (2) after conditioning on $X_1, \ldots, X_n$, it is no longer relevant whether the distributions of the random variables are identical. We include the initial steps of the proof for completeness and direct the reader to Boucheron *et al.* [6] for more details.

By the symmetrization inequalities of Boucheron *et al.* [6, Lemma 11.4], we have

$$\begin{aligned}&\mathbb{E}\,\frac{1}{\sqrt{n}}\sup_{t\in\mathscr{T}}\sum_{i=1}^{n}\left(\mathbb{1}_{X_i\in A_t} - P(A_t)\right)\\ &\qquad\leqslant 2\,\mathbb{E}\left[\mathbb{E}\left[\frac{1}{\sqrt{n}}\sup_{t\in\mathscr{T}}\sum_{i=1}^{n}\epsilon_i\mathbb{1}_{X_i\in A_t}\middle| X_1,\ldots,X_n\right]\right],\end{aligned} \tag{H.1}$$

where the ϵ_is are independent Rademacher variables. Define the random variable

$$\delta_n^2 = \max\left(\sup_{t\in\mathscr{T}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i\in A_t},\ \sigma^2\right).$$

Clearly, $\delta_n^2 \leqslant \frac{Z}{\sqrt{n}} + \sigma^2$, so by Jensen's inequality,[3]

$$\mathbb{E}\,\delta_n \leqslant \sqrt{\mathbb{E}\left(\frac{Z}{\sqrt{n}}\right) + \sigma^2}.$$

Now let $Z_t = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i\mathbb{1}_{X_i\in A_t}$. Noting that the Rademacher averages are sub-Gaussian, conditioned on the X_is, we have

$$\begin{aligned}&\log\mathbb{E}\left[e^{\lambda(Z_t - Z_{t'})}\middle| X_1,\ldots,X_n\right]\\ &\qquad\leqslant \frac{\lambda^2\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}_{X_i\in A_t} - \mathbb{1}_{X_i\in A_{t'}})^2\right)}{2}\\ &\qquad= \frac{\lambda^2\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}_{X_i\in A_t} \neq \mathbb{1}_{X_i\in A_{t'}})\right)}{2}.\end{aligned}$$

Let $d(t,t') = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}_{X_i\in A_t} \neq \mathbb{1}_{X_i\in A_{t'}})}$, and let $H(\delta, \mathscr{T})$ denote the universal δ-metric entropy (with respect to $d(\cdot,\cdot)$). Since the zero function (corresponding to $\emptyset$) belongs to the function class, we

[3] Note that both Z and Z^{-} are non-negative since $\phi \in \mathscr{A}$.

have

$$\sup_{t\in\mathscr{T}} d(t,0) = \sup_{t\in\mathscr{T}} \sqrt{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i\in A_t}} \leqslant \delta_n.$$

Therefore, we can apply the discrete version of Dudley's inequality (Boucheron *et al.* [6, Lemma 13.1]) with δ_n as the maximum radius. Since $\delta_n \geqslant \sigma$, we can upper bound the random quantity $H(a\delta_n)$ by the fixed quantity $H(a\sigma)$, for any $a > 0$. This implies that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\sqrt{n}}\sup_{t\in\mathscr{T}}\sum_{i=1}^{n}\epsilon_i \mathbb{1}_{X_i\in A_t}\middle| X_1,\ldots,X_n\right] \\
&\leqslant 3\sum_{j=0}^{\infty}\delta_n 2^{-j}\sqrt{H(\delta_n 2^{-j-1},\mathscr{T})} \\
&\leqslant 3\sum_{j=0}^{\infty}\delta_n 2^{-j}\sqrt{H(\sigma 2^{-j-1},\mathscr{T})}.
\end{aligned}$$

Taking the expectation with respect to $X_1,\ldots,X_n$ and combining with inequality (H.1) we then obtain

$$\begin{aligned}
\mathbb{E}\,Z &\leqslant 6\,\mathbb{E}\,\delta_n \cdot \sum_{j=1}^{\infty} 2^{-j}\sqrt{H(\sigma 2^{-j-1},\mathscr{T})} \\
&\leqslant 6\sqrt{\mathbb{E}\left(\frac{Z}{\sqrt{n}}\right)+\sigma^2}\left(\sum_{j=1}^{\infty} 2^{-j}\sqrt{H(\sigma 2^{-j-1},\mathscr{T})}\right).
\end{aligned}$$

From this step onward, the proof is identical to the proof of Boucheron *et al.* [6, Theorem 13.7]. □

THEOREM H.2 (Vershynin [45, Theorem 8.3.23]) Let $\mathscr{F}$ be a class of Boolean functions on a probability space (Ω,Σ,μ) with finite VC dimension $V \geqslant 1$. Let $X, X_1, X_2, \ldots, X_n$ be independent random points in Ω distributed according to the law μ. Then

$$\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X)\right|\right] \leqslant C\sqrt{\frac{V}{n}}.$$