Submitted to the Annales de l'Institut Henri Poincaré - Probabilités et Statistiques

On Mean Estimation for Heteroscedastic Random Variables

Luc Devroye[a], Silvio Lattanzi[b] Gábor Lugosi[c] and Nikita Zhivotovskiy[d]

[a]*School of Computer Science, McGill University, Montreal, Canada. E-mail:* lucdevroye@gmail.com
[b]*Google Research, Zürich, Switzerland. E-mail:* silviol@google.com
[c]*Department of Economics and Business, Pompeu Fabra University and Barcelona Graduate School of Economics, Barcelona, Spain. E-mail:* gabor.lugosi@upf.edu
[d]*Department of Mathematics, ETH Zürich, Switzerland. E-mail:* nikita.zhivotovskii@math.ethz.ch

Abstract. We study the problem of estimating the common mean μ of n independent symmetric random variables with different and unknown standard deviations $\sigma_1 \le \sigma_2 \le \cdots \le \sigma_n$. We show that, under some mild regularity assumptions on the distribution, there is an adaptive estimator $\widehat{\mu}$ such that it is invariant to permutations of the elements of the sample and satisfies that, up to logarithmic factors, with high probability,

$$|\widehat{\mu} - \mu| \lesssim \min\left\{ \sigma_{m^*}, \frac{\sqrt{n}}{\sum_{i=\sqrt{n}}^{n} \sigma_i^{-1}} \right\},$$

where the index $m^* \lesssim \sqrt{n}$ satisfies $m^* \approx \sqrt{\sigma_{m^*} \sum_{i=m^*}^{n} \sigma_i^{-1}}$.

Résumé. Nous étudions le problème de l'estimation de la moyenne commune μ de n variables aléatoires symétriques indépendantes avec des écarts types différents et inconnus $\sigma_1 \le \sigma_2 \le \cdots \le \sigma_n$. Nous montrons que, sous certaines hypothèses de régularité modérée sur la distribution, il existe un estimateur adaptatif $\widehat{\mu}$ tel qu'il est invariant aux permutations des éléments de l'échantillon et satisfait qu' à facteurs logarithmiques près, avec une probabilité élevée,

$$|\widehat{\mu} - \mu| \lesssim \min\left\{ \sigma_{m^*}, \frac{\sqrt{n}}{\sum_{i=\sqrt{n}}^{n} \sigma_i^{-1}} \right\},$$

où l'indice $m^* \lesssim \sqrt{n}$ satisfait $m^* \approx \sqrt{\sigma_{m^*} \sum_{i=m^*}^{n} \sigma_i^{-1}}$.

MSC2020 subject classifications: Primary 62G30, 62F25; secondary 62F35
Keywords: mean estimation, heteroscedastic observations, order statistic, robustness, adaptivity

In this note we study the problem of estimating the common mean $\mu \in \mathbb{R}$ of n independent real random variables $X_1, \ldots, X_n$. These random variables do not need to be identically distributed. Moreover, the variances of the X_i may greatly vary and therefore the information each observation carries about the mean may be different. For the sake of this introductory discussion, assume that the X_i all have normal distribution so that $X_i \sim \mathcal{N}(\mu, \sigma_i^2)$ for some $0 < \sigma_1 \le \cdots \le \sigma_n$.

If the values of the standard deviations σ_i were known, then one could choose the maximum likelihood estimator

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} \frac{X_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}},$$

leading to an expected error $\mathbb{E}|\widehat{\mu} - \mu| \le \left(\sum_{i=1}^{n} \sigma_i^{-2}\right)^{-1/2}$. The general study of such estimators goes back to Ibragimov and Has'minskii [14, Chapter 3, Section 4] where the estimation of a single parameter based on independent but non-identically distributed observations is studied. However, this idealistic estimator assumes that the standard deviation of each sample point is known to the statistician.

In this note we consider the situation where nothing is known about the values of the σ_i (or their assignments to the data points). In particular, we focus on the estimators invariant to permutations of the elements of the sample. Naturally, one may always compute the sample mean $(1/n)\sum_{i=1}^n X_i$. However, the sample mean has an error of the order of $(1/n)\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2}$ whose performance deteriorates even if a single data point has a large variance.

For symmetric distributions like the normal distribution, another – and more robust – natural estimator of the mean is the sample median. One of the contributions of this note is to provide new non-asymptotic performance guarantees for the sample median. In particular, we show that under some mild assumptions the error of the sample median is bounded, with high probability, by

$$\frac{c\sqrt{n\log n}}{\sum_{i=\sqrt{cn\log n}}^n \sigma_i^{-1}} \tag{1}$$

for some constant c (see Proposition 1 for the rigorous statement).

As simple as the sample median is, it has the disadvantage that it does not take advantage of the presence of data points with very small variance. Indeed, the performance of the sample median is essentially insensitive to the approximately $\sqrt{n}$ smallest variances. This can be demonstrated by the following argument: since we consider the symmetric distribution, each observation has an equal probability of being larger or smaller than the true mean. By an anti-concentration argument for Bernoulli random variables, we have that with constant probability the number of the observations larger (smaller) than μ minus the number of the observations smaller (larger) than μ is of order $\sqrt{n}$. Therefore, on this event, the median will not choose the true mean even if approximately $\sqrt{n}$ first variances are all equal to zero.

At the same time, the presence of data with very small variances makes the problem much easier. A simple way to exploit such situations is in using the so-called *modal interval estimator* introduced by Chernoff [8] for estimating the mode of a density function. The modal interval estimator looks for the most populated interval of a certain length $s > 0$ and outputs its mid-point. The main challenge in applying this method in our setting is that without any knowledge of the variances $\sigma_1, \sigma_2, \dots$ it is a hard to establish a good value of s a priori. In Proposition 2 below we establish a simple sufficient condition for the length s that guarantees that the modal interval contains the mean μ. Roughly speaking, this condition guarantees that random fluctuations of the data far from the mean cannot produce an interval of length s that has more points than the expected number of points in the interval of same length centered at the mean μ. We call such "good" values of s *admissible*. Admissibility of an interval length depends, in a complex way, on the entire sequence $\sigma_1, \dots, \sigma_n$. Ideally, one would like to use the modal interval estimator with the smallest possible admissible interval length. The main contribution of this note is an adaptive estimator that essentially achieves this goal. More precisely, without any previous knowledge of the σ_i, we show that one can construct a completely data-driven estimator that has a performance at least as good (up to constant factors in the error) as the best of the sample median and the modal interval estimator with the smallest admissible interval length.

In the remainder of this introduction we discuss previous related work. In Section 1 the analysis of the sample median is presented. We also show that an appropriately chosen *median interval* is a valid empirical confidence interval. This is important in the construction of the adaptive estimator. The modal interval estimator is analyzed in Section 2. The adaptive estimator is described in Section 3 and its performance guarantees are established in Theorem 3.1. In Section 4 we take a closer look at some concrete examples and compare our performance bounds with those of previous work.

Related work

For some classical references on the maximum likelihood estimator in our setup we refer to the work of Ibragimov and Has'minskii [15] and Beran [2]. The sample median has been analyzed in the literature in our setup. For example, Mizera and Wellner [19] provide necessary and sufficient conditions for the consistency of the sample median for triangular arrays of independent, not identically distributed random variables (in a more general setting than ours). The role of the sum of the reciprocals of the standard deviations as in (1) appears in early work. In particular, the result of Nevzorov [20, Theorem 2] can be used to provide rates of convergence of the sample median to the normal law for non-identically distributed Gaussian data that involves $\sum_{i=1}^n \sigma_i^{-1}$. The work of Gordon, Litvak, Schütt and Werner [12, Theorem 7] uses this quantity in the context of the moments of order statistics for non-identically distributed random variables. Moreover, the work of Xia [23, Corollary 6] makes direct connections between the sum of reciprocals and the performance of the sample median, see Section 4 for a detailed comparison. The same quantity appears in the analysis of the iterative trimming algorithm of Liang and Yuan [18, Remark following Theorem 1]. Importantly, in the context of the mean estimation problem, some of the above-mentioned results provide performance guarantees when the value of $\sum_{i=1}^n \sigma_i^{-1}$ is large, whereas our bounds provide sharp guarantees for the entire range of values of the sum of reciprocals of the standard deviations.

The most related papers are [9], [21], and [18]. For example, Chierichetti, Dasgupta, Kumar and Lattanzi [9] construct an estimator whose error is bounded, with high probability, by $\widetilde{O}\left(\sqrt{n}\sigma_{\log n}\right)$, where $\widetilde{O}(\cdot)$ suppresses multiplicative poly-logarithmic factors. The *hybrid* estimator of Pensia, Jog and Loh [21, Algorithm 2] uses a combination of the *shortest gap* with the median estimators, quite similar to our estimator. However, in contrast to these previous results, the estimator proposed here is adaptive to unknown parameters. Our estimator also compares favourably with the iterative trimming algorithm of Liang and Yuan [18], which does not cover the entire range of the values of $\sigma_1, \ldots, \sigma_n$ and depends on some tuning parameters and the initialization. Section 4 includes extensive comparisons with these papers. In particular, we show that, up to logarithmic factors, our bounds are never worse than the previous (non-adaptive) bounds.

Notation
In what follows, we denote $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. Given $X_1, \ldots, X_n$ let $X_{(1)}, \ldots, X_{(n)}$ denote the non-decreasing rearrangement of its elements. The value $X_{(i)}$ is usually referred to as the *i-th order statistic*. In what follows, $a \lesssim b$ and $b \gtrsim a$ denote the existence of a numerical constant c such that $a \leq cb$. The numerical constants are denoted by $c, c_1, c_2, \ldots > 0$. Their values may change from line to line. We also use the standard $O(\cdot)$ notation as well as its version $\widetilde{O}(\cdot)$ that suppresses multiplicative poly-logarithmic factors. Finally, let $[n]$ denote the set $\{1, \ldots, n\}$.

1. Analysis of the α-median interval

When the distribution of each random variable X_i is symmetric about the mean μ, the empirical median is a natural estimator of the mean. In this section we present an analysis of the empirical median. We assume the following regularity conditions.

Assumption A. *Let $X_1, \ldots, X_n$ be independent random variables and let $0 < \sigma_1 \leq \cdots \leq \sigma_n$. We assume that*

(i) $\mathbb{E}X_i = \mu$ for all $i \in [n]$;
(ii) Symmetry: for each $i \in [n]$, $X_i - \mu$ and $\mu - X_i$ have the same distribution ;
(iii) Tail assumption: for some constant $\beta > 0$, we have that for any $t > 0$,

$$\mathbb{P}\left\{|(X_i - \mu)/\sigma_i| \geq t\right\} \leq \exp(-\beta t) \,. \tag{2}$$

A canonical example satisfying Assumption A is when $X_i \sim \mathcal{N}(\mu, \sigma_i^2)$. In this case one may choose $\beta = \sqrt{\frac{2}{\pi}}$. Note that we do not need to assume that the $(X_i - \mu)/\sigma_i$ are identically distributed. It suffices that they are independent, symmetric, and satisfy the tail assumption (2). Note also that condition (iii) implies that $\mathbb{P}\{|(X_i - \mu)/\sigma_i| < t\}$ is lower bounded by $\beta t/2$ for $t \leq 2/\beta$. In particular, if X_i has an absolute continuous distribution, this assumption implies that the density of $(X_i - \mu)/\sigma_i$ is bounded away from zero near zero. An unfavorable example excluded by condition (iii) is the case of independent Rademacher random variables. Indeed, in this case if n is odd, the median is equal to either 1 or -1 and does not converge to the expected value 0. However, in this case there is no $\beta > 0$ such that $\mathbb{P}\{|X_i| \geq t\} = \mathbb{1}_{1 \geq t} \leq \exp(-\beta t)$ for all $t > 0$.

For reasons that will become apparent later, we consider not only the empirical median as a point estimator but also the so-called *median interval*, defined as the interval whose endpoints are $X_{(n/2-k)}$ and $X_{(n/2+k)}$ for an appropriately chosen value of k. This will allow us to obtain an empirical confidence interval that is essential for our adaptive procedure.

To define the median interval, assume, for simplicity, that n is even and recall that $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ denote the order statistics of $X_1, \ldots, X_n$. In order to avoid complications arising from ties, we assume that the X_i have a nonatomic distribution.

We fix $\alpha \in (0, \sqrt{n}/2)$ such that $\alpha\sqrt{n}$ is an integer. Consider the random interval

$$I_\alpha = \left[X_{(n/2-\alpha\sqrt{n})}, X_{(n/2+\alpha\sqrt{n})}\right] . \tag{3}$$

We refer to I_α as the *α-median interval*. Our first result provides two key properties of the median interval: if α is proportional to $\sqrt{\log(1/\delta)}$, the interval I_α contains the mean μ with probability at least $1 - \delta$. Moreover, we provide an upper bound for the length of I_α in terms of the sum of the reciprocals of the standard deviations.

Proposition 1. *Let Assumption A be satisfied. Fix $\delta \in (0, 1)$ such that $128 \log \frac{6}{\delta} \leq n$ and set $\alpha = \sqrt{2 \log \frac{6}{\delta}}$. The median interval I_α satisfies, with probability at least $1 - \delta$, that $\mu \in I_\alpha$ and*

$$|I_\alpha| \leq 8e\sqrt{2}\left(\log\frac{3}{\delta} \vee \log(n+1)\right)\beta^{-1} \max_{1 \leq j \leq 8\alpha\sqrt{n}} \frac{8\alpha\sqrt{n} + 1 - j}{\sum_{i=j}^{n} \sigma_i^{-1}} \,.$$

Note that by ignoring constant factors, Proposition 1 implies

$$|I_\alpha| \lesssim \beta^{-1} \log\left(\frac{n}{\delta}\right) \frac{\alpha\sqrt{n}}{\sum_{i=8\alpha\sqrt{n}}^{n} \sigma_i^{-1}} \,. \tag{4}$$

The key to the proof of Proposition 1 is following rearrangement inequality due to Gordon et al. [12, Theorem 7]. Let $|X|_{(1)}, \ldots, |X|_{(n)}$ denote the non-decreasing rearrangement of $|X_1|, \ldots, |X_n|$.

Lemma 1.1. *Let $X_1, \ldots, X_n$ be independent random variables such that for $0 < \sigma_1 \le \cdots \le \sigma_n$ and $\beta > 0$, for all $t > 0$, $\mathbb{P}(|X_i/\sigma_i| \ge t) \le \exp(-\beta t)$. Then for all $p \ge 1$ and $1 \le k \le n$,*

$$\left(\mathbb{E}(|X|_{(k)})^p\right)^{\frac{1}{p}} \le 4\sqrt{2}\max\{p, \log(k+1)\}\beta^{-1} \max_{1\le j\le k} \frac{k+1-j}{\sum_{i=j}^{n} \sigma_i^{-1}} \,.$$

Proof of Proposition 1. First, we show that $\mu \in I_\alpha$. Without loss of generality, we may assume that $\mu = 0$ for the rest of the proof. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables. Since the distribution of each X_i is assumed to be symmetric, $(\varepsilon_1|X_1|, \ldots, \varepsilon_n|X_n|)$ has the same distribution as $(X_1, \ldots, X_n)$. Conditioning on the $X_1, \ldots, X_n$, we have, by Hoeffding's inequality,

$$\mathbb{P}(\mu \notin I_\alpha) = \mathbb{P}\left(\left|\sum_{i=1}^{n} \varepsilon_i\right| > \alpha\sqrt{n}\right) \le 2\exp\left(-\frac{\alpha^2}{2}\right).$$

We denote the event that $\mu \in I_\alpha$ by E_1 and proceed with the bound on the length of the interval $|I_\alpha|$. Fix $k \le n$ and consider $|X|_{(1)}, \ldots, |X|_{(k)}$ — these are the absolute values of the k observations closest to $\mu = 0$. Note that, depending on the realizations of the random signs ε_i, the corresponding values $\varepsilon_1|X_i|$ may be on either side of $\mu = 0$. Let E_2 be the event that there are more than $k/4$ of these k observations on both sides of μ. By a simple binomial estimate,

$$\mathbb{P}(E_2) \ge 1 - 2\exp\left(-\frac{k}{8}\right) \,.$$

Consider the event $E_1 \cap E_2$ and choose $k = 8\alpha\sqrt{n}$ so that at least $2\alpha\sqrt{n} + 1$ of these closest observations are on both sides of μ. On this event since I_α contains $\mu = 0$ and exactly $2\alpha\sqrt{n} + 1$ observations, both $|X_{(n/2-\alpha\sqrt{n})}| \le |X|_{(8\alpha\sqrt{n})}$ and $|X_{(n/2+\alpha\sqrt{n})}| \le |X|_{(8\alpha\sqrt{n})}$ hold. Therefore, on the event $E_1 \cap E_2$,

$$|I_\alpha| \le 2|X|_{(8\alpha\sqrt{n})} \,. \tag{5}$$

Finally, we use Lemma 1.1 to control $|X|_{(8\alpha\sqrt{n})}$. By Markov's inequality and Lemma 1.1, we have

$$\mathbb{P}\left(|X|_{(8\alpha\sqrt{n})} \ge t\right) \le \frac{\mathbb{E}|X|^p_{(8\alpha\sqrt{n})}}{t^p} \le t^{-p}\left(4\sqrt{2}\max\{p, \log(8\alpha\sqrt{n}+1)\}\beta^{-1} \max_{1\le j\le 8\alpha\sqrt{n}} \frac{k+1-j}{\sum_{i=j}^{n} \sigma_i^{-1}}\right)^p \,.$$

Denote $\gamma = 4\sqrt{2}\beta^{-1} \max_{1\le j\le 8\alpha\sqrt{n}} \frac{k+1-j}{\sum_{i=j}^{n} \sigma_i^{-1}}$. Provided that $\frac{t}{\gamma}e^{-1} \ge \log(8\alpha\sqrt{n}+1)$, we may fix $p = \frac{t}{\gamma}e^{-1}$ and get

$$\mathbb{P}\left(|X|_{(8\alpha\sqrt{n})} \ge t\right) \le \exp\left(-\frac{t}{e\gamma}\right) \,.$$

Fixing $t = \left(\log\frac{3}{\delta} \vee \log(8\alpha\sqrt{n}+1)\right) e\gamma$ we have that, with probability at least $1 - \delta/3$,

$$|X|_{(8\alpha\sqrt{n})} \le 4e\sqrt{2}\left(\log\frac{3}{\delta} \vee \log(8\alpha\sqrt{n}+1)\right)\beta^{-1} \max_{1\le j\le 8\alpha\sqrt{n}} \frac{k+1-j}{\sum_{i=j}^{n} \sigma_i^{-1}}.$$

Denote this event by E_3.

Choosing $\alpha = \sqrt{2\log\frac{6}{\delta}}$ we have $\mathbb{P}(E_1) \ge 1 - \delta/3$. Since $\alpha\sqrt{n} \ge 2\alpha^2$, we have $\mathbb{P}(E_2) \ge 1 - 2\exp(-\alpha\sqrt{n}) \ge 1 - 2\exp\left(-4\log\frac{6}{\delta}\right) \ge 1 - \delta/3$. Therefore, we have by the union bound, that $E_1 \cap E_2 \cap E_3$ is of probability at least

$1-\delta$. On this event due to (5) we have

$$|I_\alpha| \le 8e\sqrt{2}\left(\log\frac{3}{\delta} \vee \log\left(8\alpha\sqrt{n}+1\right)\right)\beta^{-1} \max_{1\le j\le 8\sqrt{2n\log\frac{6}{\delta}}} \frac{k+1-j}{\sum_{i=j}^n \sigma_i^{-1}} .$$

The claim follows by observing that $k = 8\alpha\sqrt{n} \le n$ is equivalent to $128\log\frac{6}{\delta} \le n$.

Corollary 1. *Under the assumptions of Proposition 1 the median $X_{(n/2)}$ satisfies, with probability at least $1-\delta$,*

$$|X_{(n/2)} - \mu| \le 8e\sqrt{2}\left(\log\frac{3}{\delta} \vee \log(n+1)\right)\beta^{-1} \max_{1\le j\le 8\sqrt{2n\log\frac{6}{\delta}}} \frac{8\sqrt{2n\log\frac{6}{\delta}}+1-j}{\sum_{i=j}^n \sigma_i^{-1}}.$$

Proof. Indeed, with probability at least $1-\delta$, both μ and $X_{(n/2)}$ belong to I_α for α as in Proposition 1, and therefore, $|X_{(n/2)} - \mu| \le |I_\alpha|$. ■

2. Modal interval estimator

The second component of our adaptive estimator is the simple and natural estimator that looks for an interval of a given length containing the maximum number of data points. This is the so-called *modal interval estimator* introduced by Chernoff [8] for estimating the mode of a density function. Pensia et al. [21] also analyze this estimator though their bounds have some limitations for our purposes. We make a detailed comparison in Section 4 below.

In this section we work under the following assumptions.

Assumption B. *Let $X_1, \dots, X_n$ be independent random variables such that X_i has density $(1/\sigma_i)\phi((x-\mu)/\sigma_i)$ where ϕ is some fixed density function, μ is a location parameter and $\sigma_1 \le \dots \le \sigma_n$ are positive scale parameters. Assume that*

- *(i)* $\int x\phi(x)dx = 0$. *This implies that* $\mathbb{E}X_i = \mu$ *for all* $i \in [n]$.
- *(ii)* $\int x^2\phi(x)dx = 1$. *This implies that* $\mathrm{Var}(X_i) = \sigma_i^2$ *for all* $i \in [n]$.
- *(iii) Symmetry:* $\phi(-x) = \phi(x)$ *for all* $x \in \mathbb{R}$.
- *(iv) Unimodality:* $\phi(x)$ *is non-increasing for* $x > 0$ *and non-decreasing for* $x < 0$.

An important example satisfying Assumption B is the Gaussian case, that is, when $\phi(x) = (1/\sqrt{2\pi})e^{-x^2/2}$. However, in general, ϕ may have a heavy tail as long as the second moment exists. We also do not need to assume that ϕ is bounded. Introduce the notation

$$\Phi(t) = \int_{-t}^{t} \phi(x)dx .$$

For $s > 0$, denote the interval $A_s(x) = [x-s, x+s]$. Let

$$D_s(x) = \sum_{i=1}^n \mathbb{1}_{X_i \in A_s(x)}$$

be the number of points in the interval $A_s(x)$. Denoting $q_i(s) = \mathbb{P}\{X_i \in A_s(\mu)\} = \Phi(s/\sigma_i)$, we have

$$\mathbb{E}D_s(\mu) = \sum_{i=1}^n q_i(s) .$$

Define the *modal interval estimator* which returns the center of the densest interval of length $2s$. That is,

$$\widehat{\mu}_{n,s} \in \operatorname*{argmax}_{x\in\mathbb{R}} D_s(x) . \tag{6}$$

For the modal interval estimator to work (in the sense that it contains the common mean μ), the length s has to satisfy certain conditions. Such a sufficient condition is formulated in the following definition that intuitively captures the fact that the densest interval should contain μ, even after accounting for random fluctuations. In Proposition 2 below we prove the condition of admissibility specified here is indeed sufficient.

Definition 1. *Fix the confidence $\delta > 0$ and the interval length $s > 0$. Define*

$$m_s = \max\{m \in [n] : \sigma_m \le s\} .$$

We say that the length s is admissible if

$$m_s \ge \kappa \left(\sqrt{\mathbb{E} D_s(\mu) \log \frac{2n}{\delta}} + \log \frac{2n}{\delta} \right),$$

where $\kappa > 0$ is a numerical constant. Finally, we set

$$\overline{s}(\delta) = \inf \{ s > 0 : s \text{ is admissible} \} . \tag{7}$$

Remark 1. *The value of the constant $\kappa > 0$ depends on a universal constant appearing in Lemma A.1 below. While it is possible to extract a specific value, it is somewhat tedious and not crucial for our arguments, so we prefer to keep it unspecified. All results below hold for all values of $\kappa \ge \kappa_0$ for some constant κ_0. Changing the value only effects the constants in the results below.*

Remark 2. *Observe that if the density is bounded, that is, if $\phi(0)$ is finite, we have $q_i(s) \le \min\{1, 2\phi(0)s/\sigma_i\}$. Therefore, adjusting the constant κ, we may replace the admissibility criterion by the condition*

$$m_s \ge \kappa \left(\sqrt{\left(\sum_{i=1}^{n} \min\left\{ 1, 2\phi(0) \frac{s}{\sigma_i} \right\} \right) \log \frac{2n}{\delta}} + \log \frac{2n}{\delta} \right) ,$$

Roughly speaking, whenever $\phi(0)$ is finite one may think that $\overline{s}(\delta)$ is approximately equal to σ_{m^*}, where m^* is the smallest integer satisfying

$$m^* \gtrsim \sqrt{\sigma_{m^*} \left(\sum_{i=m^*}^{n} \frac{1}{\sigma_i} \right) \log \frac{n}{\delta}} .$$

Remark 3. *Our arguments can be immediately generalized to the case where each observation X_i has its own normalized density function denoted by ϕ_i. In particular, our analysis only requires that $\int_0^1 \phi_i(x)dx$ is the same for all $i = 1, \dots, n$ and minor modifications are needed if these quantities differ from each other by a multiplicative constant factor. However, to simplify the form of our bounds we assumed that observations come from a single family of distributions.*

The main result of this section is the following bound.

Proposition 2. *Let Assumption B be satisfied. Fix $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$, simultaneously for all admissible $s > 0$, it holds that*

$$|\widehat{\mu}_{n,s} - \mu| \le 4s .$$

Proof. We start by showing a simple lower bound for $\Phi(1) = 2\int_0^1 \phi(x)dx$. Fix any $t \ge 1$ and observe that by property (iv) in Assumption B, we have $t\Phi(1) \ge \Phi(t)$. At the same time, by Chebyshev's inequality and property (ii) we have $\Phi(t) > (1 - 1/t^2)$. Therefore,

$$\Phi(1) \ge \sup_{t \ge 1} \frac{1}{t} \left(1 - \frac{1}{t^2} \right) = \frac{2}{3\sqrt{3}} . \tag{8}$$

As the estimator is translation invariant, we may assume, without loss of generality, that $\mu = 0$. We show that, on the one hand, with probability at least $1 - \delta/2$, simultaneously for all admissible s,

$$\max_{x \in \mathbb{R}} D_s(x) \ge m_s \frac{3\Phi(1)}{4} + \sum_{i > m_s} q_i(s) , \tag{9}$$

and, on the other hand, with probability at least $1 - \delta/2$,

$$\max_{x \in \mathbb{R} : |x| \ge 4s} D_s(x) < m_s \frac{3\Phi(1)}{4} + \sum_{i > m_s} q_i(s) . \tag{10}$$

These two properties together imply the proposition. First, we show (9). Note that for $i \le m_s$ we have $\sigma_i \le s$ and $q_i(s) = \mathbb{P}\{X_i \in A_s(0)\} \ge \Phi(1)$, and therefore,

$$\mathbb{E}D_s(0) = \sum_{i=1}^{n} q_i(s) \ge m_s\Phi(1) + \sum_{i>m_s} q_i(s)\,. \tag{11}$$

Observe that, since s is admissible, we have

$$\begin{aligned} m_s\Phi(1) &\ge \Phi(1)\left(\kappa\sqrt{\mathbb{E}D_s(0)\log\frac{2n}{\delta}} + \kappa\log\frac{2n}{\delta}\right) \\ &\ge \frac{2}{3\sqrt{3}}\left(\kappa\sqrt{\mathbb{E}D_s(0)\log\frac{2n}{\delta}} + \kappa\log\frac{2n}{\delta}\right). \end{aligned} \tag{12}$$

Denote $\kappa' = \frac{2}{3\sqrt{3}}\kappa$. Using (12), we have

$$\begin{aligned} &\mathbb{P}\left\{\exists s>0 : s \text{ is admissible}, \max_{x\in\mathbb{R}} D_s(x) \le m_s\frac{3\Phi(1)}{4} + \sum_{i>m_s} q_i\right\} \\ &\le \mathbb{P}\left\{\exists s>0 : s \text{ is admissible}, D_s(0) \le m_s\frac{3\Phi(1)}{4} + \sum_{i>m_s} q_i\right\} \\ &\le \mathbb{P}\left\{\exists s>0 : s \text{ is admissible}, D_s(0) \le \mathbb{E}D_s(0) - m_s\frac{\Phi(1)}{4}\right\} \\ &\le \mathbb{P}\left\{\exists s>0 : s \text{ is admissible}, \mathbb{E}D_s(0) - D_s(0) \ge \frac{\kappa'}{4}\sqrt{\mathbb{E}D_s(0)\log\frac{2n}{\delta}} + \frac{\kappa'}{4}\log\frac{2n}{\delta}\right\}. \end{aligned}$$

The last quantity can be controlled by the uniform Bernstein-type inequality for non-identically distributed random variables. By Lemma A.1 in Appendix A, since the VC dimension of the family intervals in $\mathbb{R}$ equals 2, one may tune the value of κ' such that the last probability is bounded by $\frac{\delta}{2}$. For the convenience of the reader, we deffer the exact formulation of this Lemma as well as the definition of the VC dimension to Appendix.

We are now ready to analyze (10) for which it is enough to show that

$$\mathbb{P}\Big\{\exists s \ge 0 \text{ and } |x| \ge (1+\sqrt{2/\Phi(1)})s : s \text{ is admissible and } D_s(x) > m_s\frac{3\Phi(1)}{4} + \sum_{i>m_s} q_i\Big\} \le \frac{\delta}{2}\,. \tag{13}$$

Observe that by (8) we have $1+\sqrt{2/\Phi(1)} \le 4$. Given x such that $|x| > 4s > (1+\sqrt{2/\Phi(1)})s$, using the properties of the density ϕ together with $s \ge \sigma_{m_s}$ and Chebyshev's inequality, we have

$$\begin{aligned} \mathbb{E}D_s(x) &= \sum_{i\le m_s} \mathbb{P}\{X_i \in A_s(x)\} + \sum_{i>m_s} \mathbb{P}\{X_i \in A_s(x)\} \\ &\le \sum_{i\le m_s} \mathbb{P}\{|X_i| \ge \sqrt{2/\Phi(1)}s\} + \sum_{i>m_s} \mathbb{P}\{X_i \in A_s(x)\} \\ &\le m_s\frac{\Phi(1)}{2} + \sum_{i>m_s} q_i\,. \end{aligned}$$

Using this inequality together with (12) and recalling that $\kappa' = \frac{2}{3\sqrt{3}}\kappa$, we have

$$\begin{aligned} &\mathbb{P}\left\{\exists s \ge 0 \text{ and } |x| \ge 4s : s \text{ is admissible}, D_s(x) > m_s\frac{3\Phi(1)}{4} + \sum_{i>m_s} q_i\right\} \\ &\le \mathbb{P}\left\{\exists s \ge 0 \text{ and } |x| \ge 4s : s \text{ is admissible}, D_s(x) > \mathbb{E}D_s(x) + m_s\frac{\Phi(1)}{4}\right\} \end{aligned}$$

$$\leq \mathbb{P}\Big\{\exists s \geq 0 \text{ and } |x| \geq 4s : s \text{ is admissible}, D_s(x) > \mathbb{E}D_s(x) + \frac{\kappa'}{4}\sqrt{\mathbb{E}D_s(0)\log\frac{2n}{\delta}} + \frac{\kappa'}{4}\log\frac{2n}{\delta}\Big\}\,.$$

Using $\mathbb{E}D_s(x) \leq \mathbb{E}D_s(0)$, the last line is bounded by

$$\mathbb{P}\left\{\exists s \geq 0 \text{ and } |x| \geq 4s : s \text{ is admissible}, D_s(x) > \mathbb{E}D_s(x) + \frac{\kappa'}{4}\sqrt{\mathbb{E}D_s(x)\log\frac{2n}{\delta}} + \frac{\kappa'}{4}\log\frac{2n}{\delta}\right\}.$$

Finally, the last expression and Lemma A.1, which holds simultaneously for all x and s, implies (13) by adjusting the constant κ (and thus κ'). The proof is complete. ■

3. An adaptive estimator: combining the median and the modal Interval

Proposition 2 shows that, as long as s is an *admissible* value, the modal interval estimator has an error bounded by $4s$. Hence, to optimize the bound, one should choose s to be the smallest possible admissible value, that is, $\overline{s}(\delta)$ introduced in Definition 1. However, the value of $\overline{s}(\delta)$ depends on the values $\sigma_1, \ldots, \sigma_n$ and therefore one doesn't have access to $\overline{s}(\delta)$ unless the standard deviations are known (up to a permutation), a typically unrealistic requirement. In this section we introduce an adaptive estimator that is able to find an approximate value of $\overline{s}(\delta)$ based only on the available data $X_1, \ldots, X_n$. Furthermore, the adaptive estimator combines the α-median interval estimator with the modal interval estimator and achieves an error that is at least as good as the best of the median and the optimal modal interval estimator, up to a constant factor.

The key to making the estimator adaptive is an empirical criterion, based on which one can reject values of s that are not admissible. Once one has such a criterion, standard techniques of adaptive estimation may be applied (such as Lepski's method [17]).

Fix $\delta > 0$ and $s > 0$. Let $\eta, \xi > 0$ be numerical constants specified in the proof. Based on $X_1, \ldots, X_n$ let $\widehat{\mu}_{n,s}$ be any maximizer of $D_s(x)$ defined by (6).

- We ACCEPT the interval $A_s(\widehat{\mu}_{n,s})$ if $D_s(\widehat{\mu}_{n,s}) \geq \xi \log\frac{2n}{\delta}$ and

$$\max_{x\in\mathbb{R}, |x-\widehat{\mu}_{n,s}|\geq 8s} D_s(x) \leq D_s(\widehat{\mu}_{n,s}) - \eta\left(\sqrt{D_s(\widehat{\mu}_{n,s})\log\frac{2n}{\delta}} + \log\frac{2n}{\delta}\right).$$

- Otherwise, we REJECT this interval.

Remark 4. *Since we only consider $D_s(\widehat{\mu}_{n,s}) \geq \xi \log\frac{2n}{\delta}$ we may instead consider a criterion of the form*

$$\max_{x\in\mathbb{R}, |x-\widehat{\mu}_{n,s}|\geq 8s} D_s(x) \leq D_s(\widehat{\mu}_{n,s}) - \eta'\sqrt{D_s(\widehat{\mu}_{n,s})\log\frac{2n}{\delta}}\,,$$

for some $\eta' > 0$. However, the choice above makes the proof more transparent.

This criterion satisfies the following relation.

Proposition 3. *With probability at least $1-\delta$, simultaneously for all $s > 0$, no interval with $|\widehat{\mu}_{n,s} - \mu| > 8s$ is accepted and every admissible interval is accepted.*

Proof. Recall that, without the loss of generality, we set $\mu = 0$. From now on we work on the event E_1 where the inequalities of Lemma A.1 hold. We begin by proving that any admissible $s > 0$ is accepted with high probability. We have shown in the proof of Proposition 2 that on the event E_1, for all admissible values of s,

$$|\widehat{\mu}_{n,s}| \leq 4s\,.$$

Therefore, on this event any $x \in \mathbb{R}$ such that $|x - \widehat{\mu}_{n,s}| \geq 8s$ satisfies $|x| \geq 4s$. Also, by the argument in the proof of Proposition 2 and Lemma A.1 we have for all $|x| \geq 4s$,

$$\begin{aligned} D_s(x) &\leq m_s \frac{\Phi(1)}{2} + \sum_{i>m_s} q_i + c_1 \left(\sqrt{\mathbb{E}D_s(x) \log \frac{2n}{\delta}} + \log \frac{2n}{\delta} \right) \\ &\leq m_s \frac{\Phi(1)}{2} + \sum_{i>m_s} q_i + c_1 \left(\sqrt{\mathbb{E}D_s(0) \log \frac{2n}{\delta}} + \log \frac{2n}{\delta} \right), \end{aligned} \tag{14}$$

where $c_1 > 0$ is a numerical constant.

Observe that the function $y \mapsto y - \eta\sqrt{y \log \frac{2n}{\delta}}$ is increasing whenever $y > \eta^2 \log \frac{2n}{\delta}/4$. Thus, $D_s(0) \geq \eta^2 \log \frac{2n}{\delta}/4$ implies

$$D_s(\widehat{\mu}_{n,s}) - \eta \left(\sqrt{D_s(\widehat{\mu}_{n,s}) \log \frac{2n}{\delta}} + \log \frac{2n}{\delta} \right) \geq D_s(0) - \eta \left(\sqrt{D_s(0) \log \frac{2n}{\delta}} + \log \frac{2n}{\delta} \right) . \tag{15}$$

Observe also that the line (27) in the proof of Lemma A.1 implies that on the event E_1 it holds simultaneously for all x that

$$D_s(x) \leq 2\mathbb{E}D_s(x) + c_2 \log \frac{2n}{\delta} , \tag{16}$$

where $c_2 > 0$ is a numerical constant. By (9), on the same event, $D_s(0) \geq m_s \frac{3\Phi(1)}{4} + \sum_{i>m_s} q_i(s)$, and therefore, using the admissibility of s, the inequality (15) implies

$$\begin{aligned} &D_s(\widehat{\mu}_{n,s}) - \eta \left(\sqrt{D_s(\widehat{\mu}_{n,s}) \log \frac{2n}{\delta}} + \log \frac{2n}{\delta} \right) \\ &\geq m_s \frac{\Phi(1)}{2} + m_s \frac{\Phi(1)}{4} + \sum_{i>m_s} q_i(s) - \eta \left(\sqrt{D_s(0) \log \frac{2n}{\delta}} + \log \frac{2n}{\delta} \right) \\ &\geq m_s \frac{\Phi(1)}{2} + \frac{\kappa'}{4} \sqrt{\mathbb{E}D_s(0) \log \frac{2n}{\delta}} + \frac{\kappa'}{4} \log \frac{2n}{\delta} + \sum_{i>m_s} q_i(s) - \eta \left(\sqrt{D_s(0) \log \frac{2n}{\delta}} + \log \frac{2n}{\delta} \right) \\ &\geq m_s \frac{\Phi(1)}{2} + \frac{\kappa'}{4} \sqrt{\mathbb{E}D_s(0) \log \frac{2n}{\delta}} + \frac{\kappa'}{4} \log \frac{2n}{\delta} + \sum_{i>m_s} q_i(s) - \eta \left(\sqrt{2\mathbb{E}D_s(0) \log \frac{2n}{\delta}} + (1 + c_2) \log \frac{2n}{\delta} \right) , \end{aligned}$$

where κ' is defined in the proof of Proposition 2 and in the last line we used (16) together with $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$. Comparing this with (14) and choosing a sufficiently large value of κ in Definition 1, we prove that admissible intervals are accepted with high probability. It is only left to check that $D_s(\widehat{\mu}_{n,s}) \geq \xi \log \frac{2n}{\delta}/4$ and that, given that the constant ξ is properly adjusted, our additional acceptance assumption $D_s(\widehat{\mu}_{n,s}) \geq \xi \log \frac{2n}{\delta}/4$ implies, with high probability, that $D_s(0) \geq \eta^2 \log \frac{2n}{\delta}/4$ which was used in (15). This computation follows immediately from Lemma A.1 and the fact that $\mathbb{E}D_s(x)$ is maximized at $x = 0$.

It remains to prove that our empirical criterion can never accept the interval with its center $\widehat{\mu}_{n,s}$ satisfying $|\widehat{\mu}_{n,s}| > 8s$. To do so we observe that if $|\widehat{\mu}_{n,s}| > 8s$ then the interval $A_{8s}(\widehat{\mu}_{n,s})$ does not contain $\mu = 0$ and in the acceptance criterion we should compare with $D_s(0)$. Assuming that $\eta > 2\kappa_2$, where κ_2 is defined in Lemma A.1 and using that $\mathbb{E}D_s(x)$ is maximized at 0, we have on the event where the inequalities of Lemma A.1 hold

$$\begin{aligned} &D_s(\widehat{\mu}_{n,s}) - \eta \sqrt{D_s(\widehat{\mu}_{n,s}) \log \frac{2n}{\delta}} - \eta \log \frac{2n}{\delta} \\ &< \mathbb{E}\left[D_s(\widehat{\mu}_{n,s}) | X_1, \ldots, X_n \right] - \kappa_2 \sqrt{D_s(\widehat{\mu}_{n,s}) \log \frac{2n}{\delta}} - \kappa_2 \log \frac{2n}{\delta} \\ &\leq \mathbb{E}\left[D_s(\widehat{\mu}_{n,s}) | X_1, \ldots, X_n \right] - \kappa_2 \sqrt{D_s(0) \log \frac{2n}{\delta}} - \kappa_2 \log \frac{2n}{\delta} \end{aligned}$$

$$\leq \mathbb{E}D_s(0) - \kappa_2\sqrt{D_s(0)\log\frac{2n}{\delta}} - \kappa_2\log\frac{2n}{\delta}$$
$$\leq D_s(0)\ .$$

Therefore, $D_s(\widehat{\mu}_{n,s}) - \eta\sqrt{D_s(\widehat{\mu}_{n,s})\log\frac{2n}{\delta}} - \eta\log\frac{2n}{\delta} < \max_{x\in\mathbb{R}, |x-\widehat{\mu}_{n,s}|\geq 8s} D_s(x)$, which implies that the interval $A_s(\widehat{\mu}_{n,s})$ is rejected. ■

The adaptive estimator
We are now ready to define an adaptive estimator that achieves a performance that is at least as good – up to a constant factor – as the best of our bounds for the median (Proposition 1) and the modal interval with optimally chosen length (Proposition 2).

We observe a sample of independent random variables $X_1, \ldots, X_n$. Fix the desired confidence level $\delta \in (0,1)$. We output the estimator $\widehat{\mu}$ defined as follows:

- Fix $\alpha = \sqrt{2\log\frac{6}{\delta}}$ and compute the α-median interval I_α.
- Let $\widehat{\mu}$ be the midpoint of the interval

$$\left(\bigcap_{\substack{0\leq s\leq |I_\alpha| \\ A_s(\widehat{\mu}_{n,s}) \text{ is ACCEPTed}}} A_{8s}(\widehat{\mu}_{n,s})\right) \cap I_\alpha, \tag{17}$$

  where $\widehat{\mu}_{n,s}$ is defined by (6) and let $\widehat{\mu}$ be the midpoint of the interval I_α if the set (17) is empty.
- Return $\widehat{\mu}$.

Remark 5. *In practice there is no need to search through all $s > 0$ in (17). One may discretize and consider only $s_i = 2^{-i}|I_\alpha|$ for integers $i \geq 0$. Also, due to Lemma A.1 we may essentially replace $\mathbb{E}D_s(x)$ by $D_s(x)$ in the steps of the proof where admissibility is used. That is, one may instead consider the* random *admissibility condition of the form*

$$m_s \gtrsim \sqrt{D_s(\mu)\log\frac{2n}{\delta}} + \log\frac{2n}{\delta}\ .$$

Due to the discrete nature of the sample, only a finite number of values $D_s(\mu)$ is possible and in the set (17) one may consider only at most $\binom{n}{2}$ values of s that correspond to the distances between pairs of points. For the sake of brevity we omit the straightforward details of the analysis of the discretized estimator and focus on the estimator defined above.

Theorem 3.1. *Let Assumptions* **A** *and* **B** *be satisfied. Fix $\delta \in (0, 1/2)$ such that $128\log\frac{6}{\delta} \leq n$. There is a numerical constant $c_1 > 0$ such that, with probability at least $1 - 2\delta$, the estimator $\widehat{\mu}$ defined above satisfies*

$$|\widehat{\mu} - \mu| \leq c_1 \min\left\{\overline{s}(\delta), \beta^{-1}\log\left(\frac{n}{\delta}\right) \max_{1\leq j\leq 8\alpha\sqrt{n}} \frac{8\alpha\sqrt{n} + 1 - j}{\sum_{i=j}^n \sigma_i^{-1}}\right\}\ ,$$

where $\overline{s}(\delta)$ is given by Definition 1.

Proof. Recalling that $|I_\alpha|$ is a random variable, consider the event E_1,

$$\overline{s}(\delta) \leq |I_\alpha|\ .$$

On the complementary event $\overline{E}_1$ we have that the solution based on the α-median interval is better than what one can get with the modal interval estimator. In particular, since $\widehat{\mu}$ always returns a point in I_α, the proof is complete by Proposition 1.

Otherwise, we focus on the event E_1. Let E_2 be the event that every accepted interval $A_s(\widehat{\mu}_{n,s})$ satisfies $\mu \in A_{8s}(\widehat{\mu}_{n,s})$. By Proposition 3, it holds that $\mathbb{P}\{E_2\} \geq 1-\delta$. Therefore, on E_2, we have either

$$\mu \in \bigcap_{\substack{0 \leq s \leq |I_\alpha| \\ A_s(\widehat{\mu}_{n,s}) \text{ is ACCEPTed}}} A_{8s}(\widehat{\mu}_{n,s}) \,, \tag{18}$$

or there are no accepted intervals in this range. The latter cannot be true on the event $E_1 \cap E_2$ since $\overline{s}(\delta) \leq |I_\alpha|$ and $A_{\overline{s}(\delta)}(\widehat{\mu}_{n,\overline{s}(\delta)})$ is accepted. Therefore, the intersection of intervals in (18) is non-empty and its length is bounded by $16\overline{s}(\delta)$. Thus, on the event $E_1 \cap E_2$ we have $|\widehat{\mu} - \mu| \leq 16\overline{s}(\delta)$. The claim follows by the union bound. ■

4. Examples and a comparison with existing results

To demonstrate the meaning of the derived performance bounds, in this section we discuss several natural examples and compare our results with existing general bounds. As already mentioned, our adaptive estimator is closely related to the estimator of Chierichetti et al. [9] and to the hybrid estimator of Pensia et al. [21]. Let us emphasize some technical differences with the latter work which generalizes the results in [9]:

- The hybrid estimator of Pensia et al. [21, Algorithm 2] depends on the choice of the parameter k_2. This parameter counts the number of points in the k-shortest gap estimator. We believe that one can make an adaptive choice of k_2, though it is not immediately clear to what extent it affects the overall performance of their estimator.
- Even though the results in [21] work under milder assumptions, their bounds depend on the distribution through the quantity r_k which should be "manually" computed in each particular case. In contrast, our results require that Assumptions **A** and **B** hold, but because of this the resulting bound depends explicitly on the standard deviations $\sigma_1, \ldots, \sigma_n$. Moreover, Proposition 5 shows that our Theorem 3.1 is never worse than the best known bound written in terms of $\sigma_1, \ldots, \sigma_n$ [9, Theorem 4.1]
- Our analysis of the modal interval estimator is sharper. In particular, while by Pensia et al. [21, Theorem 3.1] the modal interval estimator can never choose a center that has on average less than $\frac{1}{2}\mathbb{E}D_s(\mu)$ observations, our analysis uses the sharper property that the modal interval estimator never chooses a center that has, on the average, less than $\mathbb{E}D_s(\mu) - c\sqrt{\mathbb{E}D_s(\mu)}$ observations, for some $c > 0$ up to logarithmic factors.

Our results can also be compared with the estimator of Liang and Yuan [18, Algorithm 1]. Their iterative truncation algorithm uses the initial approximation $\mu^{(0)}$ and the parameter B satisfying $|\mu - \mu^{(0)}| \leq B$. They also assume that the index m such that $\sigma_m \leq 1$ is known and their bound depends on m. In Section 4.1 we show that their bounds are implied by our median estimator alone.

4.1. Examples

Most of our examples appear in [21] and [18]. We show that our bounds written in terms of $\sigma_1, \ldots, \sigma_n$ are not worse than any of the previous bounds depending on some more involved distribution dependent quantities, often achieved by non-adaptive estimators. In all examples we only consider the Gaussian case, that is, we assume $X_i \sim \mathcal{N}(\mu, \sigma_i^2)$. Also, for the sake of presentation we fix the allowed probability of error to be $\delta = \frac{1}{n}$.

Example 1. *(Equal variances.) In the simplest case we have $\sigma_i = \sigma$ for $i \in [n]$. In this example the median interval alone recovers the optimal error rate $\widetilde{O}\left(\frac{\sigma}{\sqrt{n}}\right)$. Therefore, our adaptive algorithm mimics the optimal behavior of the sample mean in the i.i.d. scenario.*

Example 2. *(Two variances.) Consider the case where $\sigma_i = \sigma$ for $i \in [m]$ and $\sigma_i = \sigma' > \sigma$ for $i \in [n] \setminus [m]$.*

There are different cases and we consider the most interesting regimes. First, if $m \gtrsim \sqrt{n \log n}$ the median gives the rate $\widetilde{O}\left(\frac{\sqrt{n}}{m\sigma^{-1} + (n-m)(\sigma')^{-1}}\right)$ and the interval algorithm can always guarantee the error $O(\sigma)$ since the interval of length σ is admissible. Next we consider $m \lesssim \sqrt{n \log n}$. The median gives the rate $\widetilde{O}\left(\frac{\sigma'}{\sqrt{n}}\right)$ in this regime and the interval of length $O(\sigma)$ is admissible if $m \gtrsim \sqrt{\frac{\sigma}{\sigma'} n \log n} + \log n$. In particular, an application of Theorem 3.1 and shows that, with probability at least $1 - \frac{1}{n}$,

$$|\widehat{\mu} - \mu| = \begin{cases} \widetilde{O}\left(\frac{\sqrt{n}}{m\sigma^{-1} + (n-m)(\sigma')^{-1}}\right), & \text{if } m \gtrsim \sqrt{n \log n}; \\ \widetilde{O}\left(\sigma \wedge \frac{\sigma'}{\sqrt{n}}\right), & \text{if } \sqrt{\frac{\sigma}{\sigma'} n \log n} + \log n \lesssim m \lesssim \sqrt{n \log n}. \end{cases}$$

Example 3. *(α-mixture distributions.) This is a particular case of the example of two variances above, with $m = c\lfloor \log n \rfloor$, for some $c > 0$; $\sigma = 1$ and $\sigma' = n^\alpha$ for some $\alpha > 0$. This example was thoroughly studied in [21]. When $\alpha < 1$ the analysis of the sample median in the example above gives, with probability at least $1 - \frac{1}{n}$,*

$$|\widehat{\mu} - \mu| = \widetilde{O}(n^{\alpha - 1/2}) \, ,$$

otherwise, for $\alpha \geq 1$ provided that c is a large enough numerical constant we have

$$|\widehat{\mu} - \mu| = O(1) \, .$$

Therefore, our algorithm recovers the best known rates in [21, Table 1], in an adaptive manner.

Example 4. *(Quadratic variances.) In this setup we assume that for some constant $c > 0$, $\sigma_i^2 = c^2 i^2$. In this case, an interval of length $s = cj$ is admissible if*

$$j \gtrsim \sqrt{\sum_{i=j}^{n} \frac{j}{i} \log n + \log n} \, .$$

Using $\sum_{i=j}^{n} \frac{j}{i} \lesssim j \log \frac{n}{j}$, we see that an interval of length proportional to $\log n$ is admissible. A simple computation shows that the median interval can produce an error $\widetilde{O}(\sqrt{n})$. Finally, an application of Theorem 3.1 gives, with probability at least $1 - \frac{1}{n}$,

$$|\widehat{\mu} - \mu| = O(\log n) \, .$$

This improves upon the bound of Pensia et al. [21, Table 1] where for the same model an arbitrarily small polynomial error is established. We remark that this result coincides with the performance of the maximum likelihood estimator $\left(\sum_{i=1}^{n} \sigma_i^{-2}\right)^{-1/2}$ up to a logarithmic factor.

Example 5. *(The subset-of-signals model.) In this setup the only assumption is that, for some $m < n$, at least m out of n variances are less or equal to one. In other words, $\sigma_m \leq 1$. The subset-of-signals model was studied by Liang and Yuan [18]. The authors prove that if $m \gtrsim \sqrt{n \log n}$, then there is an estimator $\widetilde{\mu}$ based on iterative truncations (first studied in [24]) such that, with probability at least $1 - 1/n$,*

$$|\widetilde{\mu} - \mu| \lesssim \frac{\sqrt{n \log n}}{m} \, .$$

Assuming that $m \gtrsim \sqrt{n \log n}$, we have by Proposition 1 and Theorem 3.1, that, with probability at least $1 - \frac{1}{n}$,

$$|I_\alpha| \lesssim \frac{\sqrt{n} (\log n)^{3/2}}{m} \quad \text{and thus,} \quad |\widehat{\mu} - \mu| = \widetilde{O}\left(\frac{\sqrt{n}}{m}\right) \, .$$

This shows that the sample median (and hence our general adaptive estimator) performs as well as the algorithm of Liang and Yuan [18], up to a logarithmic factor. The advantage of the median is that its complexity is linear in the number of observations [3] whereas the iterative truncation algorithm is more complex. As we mentioned, the iterative truncation algorithm of Liang and Yuan [18] depends on some parameters of the problem as well as on an initialization. We additionally remark that according to [18] the hybrid estimator of Pensia et al. [21] also recovers the rate $\widetilde{O}\left(\frac{\sqrt{n}}{m}\right)$ in the subset-of-signals model.

4.2. A comparison with some general bounds

Finally, we compare our results with several recent general bounds. These bounds can also be explicitly written in terms of $\sigma_1, \ldots, \sigma_n$. Our main conclusion is that, apart from the logarithmic factors and at least in the case of Gaussian data, our adaptive estimator performs at least as well as the best known guarantees in the literature. We emphasize again that our estimator does not depend on any parameters of the problem whereas the best known algorithms require some kind of parameter tuning.

The result of Xia on the median of Gaussians

Xia [23] analyzed the sample median of independent, not necessarily identically distributed random variables with the same median. For the sake of an easier comparison, we only consider here the case of normal random variables. The following result appears in [23, Corollary 6].

Proposition 4. *Consider independent $X_1, \dots, X_i$ such that $X_i \sim \mathcal{N}(\mu, \sigma_i^2)$. Assume that $\delta \in (0,1)$ satisfies*

$$\frac{\sqrt{n \log \frac{1}{\delta}}}{\sum_{i=1}^n \sigma_i^{-1}} \leq \frac{7\sqrt{2}\sigma_1}{10} \,. \tag{19}$$

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

$$|X_{(n/2)} - \mu| \leq \frac{\frac{10}{7}\sqrt{2n \log \frac{1}{\delta}}}{\sum_{i=1}^n \sigma_i^{-1}} \,.$$

At first glance the result of Proposition 4 looks stronger than what is given by Corollary 1 in the special case of Gaussians. Indeed, it does not have the $\log n$ factor and has a better dependence on $\log \frac{1}{\delta}$. The main difference comes from the assumption (19) which is more restrictive than the only assumption $128 \log \frac{6}{\delta} \leq n$ of Corollary 1. Indeed, in the most favourable case when $\sigma_i = \sigma$ for $i \in [n]$, the condition (19) implies $\log \frac{1}{\delta} \leq \left(\frac{7\sqrt{2}}{10}\right)^2 n$ which coincides with our assumption up to absolute constants. However, for small σ_1 the assumption (19) requires $\delta \to 1$ whereas our bound is not sensitive to the approximately $\sqrt{n \log \frac{1}{\delta}}$ smallest variances. The following result shows that the condition (19) simplifies the bound of Proposition 1 making it almost the same as the result of Proposition 4, up to logarithmic factors.

Corollary 2. *Fix $\delta \in (0,1)$ such that $128 \log \frac{6}{\delta} \leq n$ and set $\alpha = \sqrt{2 \log \frac{6}{\delta}}$. Assume that there is $0 < c < 1$ such that*

$$\frac{8\sqrt{2n \log \frac{6}{\delta}}}{\sum_{i=1}^n \sigma_i^{-1}} \leq c\sigma_1 \,.$$

Under the assumptions of Proposition 1 we have, with probability at least $1-\delta$,

$$|I_\alpha| \leq 64e\sqrt{2} \left(\log \frac{3}{\delta} \vee \log(n+1) \right) \beta^{-1} \frac{\sqrt{2n \log \frac{6}{\delta}}}{(1-c)\sum_{i=1}^n \sigma_i^{-1}} \,.$$

Proof. The proof is based on elementary comparisons. Fix $j \leq 8\alpha\sqrt{n}$. Then

$$\begin{aligned}
\sum_{i=j}^n \sigma_i^{-1} &\geq \sum_{i=8\alpha\sqrt{n}+1}^n \sigma_i^{-1} = \sum_{i=1}^n \sigma_i^{-1} - \sum_{i=1}^{8\alpha\sqrt{n}} \sigma_i^{-1} \\
&\geq 8\alpha\sqrt{n}\, \sigma_1^{-1} c^{-1} - \sum_{i=1}^{8\alpha\sqrt{n}} \sigma_i^{-1} \geq (c^{-1} - 1) \sum_{i=1}^{8\alpha\sqrt{n}} \sigma_i^{-1}.
\end{aligned}$$

This implies

$$c^{-1} \sum_{i=j}^n \sigma_i^{-1} \geq (c^{-1}-1) \sum_{i=j}^n \sigma_i^{-1} + (c^{-1}-1) \sum_{i=1}^{8\alpha\sqrt{n}} \sigma_i^{-1} \geq (c^{-1}-1) \sum_{i=1}^n \sigma_i^{-1}.$$

Combining these inequalities, we obtain

$$\max_{1 \leq j \leq 8\alpha\sqrt{n}} \frac{8\alpha\sqrt{n} - j + 1}{\sum_{i=j}^n \sigma_i^{-1}} \leq \frac{1}{1-c} \max_{1 \leq j \leq 8\alpha\sqrt{n}} \frac{8\alpha\sqrt{n} - j + 1}{\sum_{i=1}^n \sigma_i^{-1}} = \frac{1}{1-c} \frac{8\alpha\sqrt{n}}{\sum_{i=1}^n \sigma_i^{-1}} \,.$$

The result follows. ∎

The bound in [9].

Chierichetti et al. [9, Theorem 4.1] introduce an estimator $\widetilde{\mu}$ such that for $X_i \sim \mathcal{N}(\mu, \sigma_i)$, with probability at least $1 - \frac{1}{n}$,

$$|\widetilde{\mu} - \mu| = \widetilde{O}(\sigma_{\log n}\sqrt{n}) \,. \tag{20}$$

The hybrid estimator of Pensia et al. [21] satisfies a similar performance bound if the parameters are chosen in a specific way. The next result shows that the adaptive estimator introduced in this note achieves this bound without any additional parameter tuning. Moreover, the result follows from our general bounds written in terms of $\sigma_1, \dots, \sigma_n$.

Proposition 5. *Let Assumptions* **A** *and* **B** *hold and assume that* $\log n$ *is integer. There is a constant* $c = c(\beta, \phi(0)) > 1$ *such that the adaptive estimator of Theorem 3.1 satisfies for large enough* n *that, with probability at least* $1 - \frac{1}{n}$,

$$|\widehat{\mu} - \mu| \le c\left(\sigma_{c\log n}\sqrt{n}\log^{3/2} n\right) \,.$$

The proof is based on some elementary but tedious computations, see Appendix B.

5. Concluding remarks

In this note we construct an adaptive estimator for the common mean of independent, not necessarily identically distributed random variables and provide performance guarantees that hold under certain assumptions for the underlying distribution. Among the key assumptions are that the distributions are symmetric around the mean and the underlying densities are unimodal. However, even in the simplest case of normal random variables, the problem is not fully understood. In particular, as far as we know, no general nontrivial lower bounds are available. It is not difficult to prove that no estimator can have an expected error smaller than that of the maximum likelihood estimator that "knows" the variance of each sample point, that is, $\left(\sum_{i=1}^n \sigma_i^{-2}\right)^{-1/2}$. In the absence of knowledge of the σ_i, the problem becomes significantly harder. It remains an interesting challenge to prove general lower bounds that are much larger than the trivial bound $\left(\sum_{i=1}^n \sigma_i^{-2}\right)^{-1/2}$. In fact, we think that, up to logarithmic factors, the upper bound of Theorem 3.1 is essentially tight for most interesting values of the parameters. However, the full picture is surely more complex. For example, in some particular ranges of the parameters it is easy to improve on Theorem 3.1. To illustrate such an example, consider the case of two variances discussed in Section 4, that is, when $\sigma_i = \sigma$ for $i \in [m]$ and $\sigma_i = \sigma' > \sigma$ for $i \in [n] \setminus [m]$. Suppose that $\sigma\sqrt{\log m} \ll \sigma'/n$. In this case, with high probability, the modal interval of length $s = 3\sigma\sqrt{\log m}$ contains all of $X_1, \dots, X_m$ but none of $X_{m+1}, \dots, X_n$. In this case, instead of outputting the center of the modal interval, by averaging the points falling in it, one obtains an error of the order $O(\sigma/\sqrt{m})$, as opposed to $O(\sigma)$ guaranteed by Theorem 3.1 in this case.

Even our analysis of the sample median leaves room for improvement. In particular, we think that part (iii) of Assumption **A** may be weakened. While it is obviously necessary to assume that the density of the X_i are bounded away from zero near the mean (consider the case of independent Rademacher random signs in the i.i.d. case), the exponential tail condition implied by this assumption seems unnecessary. Indeed, Xia [23, Corollary 12] deals with the heavy-tailed Cauchy distribution.

Another interesting challenge is to gain an understanding of more general cases when $X_1, \dots, X_n$ are independent, they have the same mean, but their distribution may not be symmetric or unimodal.

Finally, we mention that the model studied in this note is closely related to the model of heteroscedastic linear regression with fixed design. In this model it is assumed that one observes, for $i \in [n]$,

$$Y_i = \langle x_i, \beta\rangle + \xi_i \,,$$

where $\beta \in \mathbb{R}^d$ is the target parameter, $x_i \in \mathbb{R}^d$ are deterministic design vectors, and $\xi_i \sim \mathcal{N}(0, \sigma_i)$ are independent noise variables. In order to provide some reasonable guarantees for this model, one usually makes some additional assumptions. In a classical model (see, for instance, [10]) it is assumed that the values of σ_i are arbitrary, but there are enough repetitions of each observation available so that one can estimate the values of σ_i. Once the values of σ_i and their assignments to the observations are (almost) known, one may use the weighted mean described in the introduction which achieves (almost) optimal performance. Another line or research which can be attributed, among other papers, to the early work of Carroll and Ruppert [6], is where some additional assumptions on σ_i are made. For example, they are increasing according to some law. Our model can be seen as a particular case of heteroscedastic linear regression in dimension one, where we additionally assume that the design x_i is the same for all i. However, these simplifications are compensated by the fact that

we make neither the assumption on the repeated observations nor the assumption that the σ_i follow a particular functional form. Finally, our estimators are invariant to the permutation of the elements of the sample and thus cannot exploit the monotonicity of the standard deviations.

Appendix A: Ratio-type VC bounds for non-identically distributed entries

In this section we provide high probability ratio-type VC bounds. These results are originally due to Vapnik and Chervonenkis [22, Theorem 12.2]) and hold for identically distributed random variables. A bound of a similar type for non-identically distributed random variables was proved in [21, Lemma 2.2] though their result is not sufficient for our purposes[1]. We also note that similar bounds for non-identically distributed independent random variables were shown in [7]. Consider a set $\mathcal{F}$ of $\{0,1\}$-valued functions defined on a domain $\mathcal{X}$ such with VC-dimension equal to d. Recall that the VC dimension that is the largest integer d such that there are $x_1,\dots,x_d \in \mathcal{X}$ satisfying $\left|\{(f(x_1),\dots,f(x_d)) : f \in \mathcal{F}\}\right| = 2^d$. The proof of the next technical lemma is a quite straightforward generalization of similar bounds for the i.i.d. case. The analysis is based on localization techniques for empirical processes. We refer, for instance, to [1, Corollary 3.7] and to [5] for some similar results in the context of VC classes.

Lemma A.1. *Let $X_1,\dots,X_n$ be independent but not necessary identically distributed random variables taking their values in $\mathcal{X}$. Assume that the class $\mathcal{F}$ of $\{0,1\}$-valued functions has the VC dimension d. Then there are numerical constants $\kappa_1,\kappa_2 > 0$ such that for any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $f \in \mathcal{F}$,*

$$\left|\sum_{i=1}^n (f(X_i) - \mathbb{E}f(X_i))\right| \le \kappa_1 \left(\sqrt{\left(\sum_{i=1}^n \mathbb{E}f(X_i)\right)\left(d\log\frac{n}{d} + \log\frac{1}{\delta}\right)} + d\log\frac{n}{d} + \log\frac{1}{\delta}\right) \tag{21}$$

and

$$\left|\sum_{i=1}^n (f(X_i) - \mathbb{E}f(X_i))\right| \le \kappa_2 \left(\sqrt{\left(\sum_{i=1}^n f(X_i)\right)\left(d\log\frac{n}{d} + \log\frac{1}{\delta}\right)} + d\log\frac{n}{d} + \log\frac{1}{\delta}\right) . \tag{22}$$

Proof. Without loss of generality we may assume that $0 \in \mathcal{F}$ since by adding $f \equiv 0$ to the class the VC dimension increases by at most one which can be absorbed by choosing slightly larger values of $\kappa_1,\kappa_2 > 0$. Consider the star-shaped hull of $\mathcal{F}$ around zero, that is, the class $\mathcal{H}$ of $[0,1]$-valued functions defined as

$$\mathcal{H} = \{\alpha f : f \in \mathcal{F}, \alpha \in [0,1]\} .$$

For $h \in \mathcal{H}$, we denote $Ph^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}h(X_i)^2$. Fix any $\delta \in (0,1)$ and consider the fixed point

$$\gamma(\lambda,\delta) = \inf\left\{ s > 0 : \mathbb{P}\left(\sup_{h\in\mathcal{H}, Ph^2 \le s^2} \left|\sum_{i=1}^n (h(X_i) - \mathbb{E}h(X_i))\right| \le \lambda n s^2\right) \ge 1-\delta \right\} ,$$

where $\lambda > 0$ is a numerical constant specified below. By the definition of $\gamma(\lambda,\delta)$ we have, with probability at least $1-\delta$,

$$\sup_{h\in\mathcal{H}, Ph^2 \le \gamma(\lambda,\delta)^2} \left|\sum_{i=1}^n (h(X_i) - \mathbb{E}h(X_i))\right| \le \lambda n \gamma(\lambda,\delta)^2 . \tag{23}$$

Fix any $h \in \mathcal{H}$ such that $Ph^2 \ge \gamma(\lambda,\delta)^2$. Since $\mathcal{H}$ is star-shaped, we have that $h' = h\gamma(\lambda,\delta)/\sqrt{Ph^2} \in \mathcal{H}$ and $P(h')^2 = \gamma(\lambda,\delta)^2$, which, applying (23) for h', implies on the same event (and the same holds simultaneously for any such h)

$$\left|\sum_{i=1}^n (h(X_i) - \mathbb{E}h(X_i))\right| \le \lambda n \gamma(\lambda,\delta)\sqrt{Ph^2} .$$

The last inequality, combined with (23), implies simulteniously for all $h \in \mathcal{H}$,

$$\left|\sum_{i=1}^n (h(X_i) - \mathbb{E}h(X_i))\right| \le \lambda n \gamma(\lambda,\delta)\sqrt{Ph^2} + \lambda n \gamma(\lambda,\delta)^2 . \tag{24}$$

[1] In particular, our result covers some values of their parameter t that are not allowed in [21, Lemma 2.2].

Finally, we need to prove an upper bound for $\gamma(\lambda,\delta)$. Denoting $\mathcal{H}'=\mathcal{H}\cup(-\mathcal{H})$, we have

$$\sup_{h\in\mathcal{H},Ph^2\le s^2}\left|\sum_{i=1}^n(h(X_i)-\mathbb{E}h(X_i))\right| = \sup_{h\in\mathcal{H}',Ph^2\le s^2}\left(\sum_{i=1}^n(h(X_i)-\mathbb{E}h(X_i))\right) .$$

By [11, Theorem 3.3.16] (see inequality (3.128) there which is relaxed in what follows by using $\sqrt{2(2\mathbb{E}Z+\mathcal{V}_n)x}\le\sqrt{2\mathcal{V}_n x}+x+\mathbb{E}Z$), since almost surely $|h(X_i)-\mathbb{E}h(X_i)|\le 1$ and by fixing $x=\log\frac{1}{\delta}$, we have, with probability at least $1-\delta$,

$$\begin{aligned}&\sup_{h\in\mathcal{H}',Ph^2\le s^2}\left(\sum_{i=1}^n(h(X_i)-\mathbb{E}h(X_i))\right)\\ &\le 2\mathbb{E}\sup_{h\in\mathcal{H}',Ph^2\le s^2}\left(\sum_{i=1}^n(h(X_i)-\mathbb{E}h(X_i))\right)+s\sqrt{n\log\frac{1}{\delta}}+(5/2)\log\frac{1}{\delta} .\end{aligned} \tag{25}$$

Finally, using the symmetrization inequality [16] we have

$$\mathbb{E}\sup_{h\in\mathcal{H}',Ph^2\le s^2}\left(\sum_{i=1}^n(h(X_i)-\mathbb{E}h(X_i))\right)\le 2\mathbb{E}\sup_{h\in\mathcal{H}',Ph^2\le s^2}\left(\sum_{i=1}^n\varepsilon_i h(X_i)\right) ,$$

where $\varepsilon_1,\dots,\varepsilon_n$ are i.i.d. Rademacher random variables with $\mathbb{P}\{\varepsilon_i=1\}=\mathbb{P}\{\varepsilon_i=-1\}=1/2$. Conditioning on $X_1,\dots,X_n$, we may use Dudley's entropy integral bound (see, for instance, [4]). First, we estimate the covering numbers of the set $\mathcal{H}$ with respect to the (random) distance $\rho(f,g)=\sqrt{\sum_{i=1}^n(f(X_i)-g(X_i))^2/n}$. Denote

$$\mathrm{diam}(n,s)=\sup_{f,h\in\mathcal{H},Pf^2\le s^2,Ph^2\le s^2}\rho(f,h) .$$

By the bound of Haussler [13], the covering number of $\mathcal{F}$ at scale r is upper bounded by $e(d+1)\left(\frac{2e}{r^2}\right)^d$ and by a standard argument we have that the covering number of $\mathcal{H}$ is upper bounded by $e(d+1)\left(\frac{8e}{r^2}\right)^d\left(1+\lceil\frac{2}{r}\rceil\right)$ (see [1, Proof of Corollary 3.7]). Therefore, by the Dudley's bound we have, for some constants $c_1,c_2>0$,

$$\begin{aligned}\mathbb{E}\sup_{h\in\mathcal{H}',Ph^2\le s^2}\left(\sum_{i=1}^n\varepsilon_i h(X_i)\right) &= \mathbb{E}\sup_{h\in\mathcal{H},Ph^2\le s^2}\left|\sum_{i=1}^n\varepsilon_i h(X_i)\right|\\ &\le c_1\sqrt{n}\mathbb{E}\int_0^{\mathrm{diam}(n,s)}\sqrt{d\log\frac{e}{r}}dr\\ &\le c_2\sqrt{n}\mathbb{E}\mathrm{diam}(n,s)\sqrt{d\log\frac{e}{\mathrm{diam}(n,s)}}\left(\mathbb{1}_{\mathrm{diam}(n,s)\ge\sqrt{d/n}}+\mathbb{1}_{\mathrm{diam}(n,s)<\sqrt{d/n}}\right)\\ &\le c_2\left(\sqrt{n}\mathbb{E}\mathrm{diam}(n,s)\sqrt{d\log\frac{n}{d}}+d\sqrt{\log\frac{n}{d}}\right) .\end{aligned}$$

By Jensen's inequality combined with the standard symmetrization and contraction inequalities [16] we have, for some $c_3>0$,

$$\begin{aligned}\sqrt{n}\mathbb{E}\mathrm{diam}(n,s) &\le \sqrt{2\mathbb{E}\sup_{h\in\mathcal{H}',Ph^2\le s^2}\sum_{i=1}^n h^2(X_i)}\\ &\le \sqrt{2\mathbb{E}\sup_{h\in\mathcal{H}',Ph^2\le s^2}\sum_{i=1}^n(h^2(X_i)-\mathbb{E}h^2(X_i))+2ns^2}\\ &\le c_3\left(\sqrt{\mathbb{E}\sup_{h\in\mathcal{H}',Ph^2\le s^2}\left(\sum_{i=1}^n\varepsilon_i h(X_i)\right)}+\sqrt{n}s\right) .\end{aligned}$$

Combining the last two arguments, we have, for some $c_4 > 0$,

$$\mathbb{E} \sup_{h \in \mathcal{H}', Ph^2 \le s^2} \left(\sum_{i=1}^n \varepsilon_i h(X_i) \right) \le c_4 \left(s\sqrt{dn \log \frac{n}{d}} + d\sqrt{\log \frac{n}{d}} \right) . \tag{26}$$

Finally, combining (25), (26) and adjusting the constant λ we have, for some $c_5 > 0$, that

$$\gamma(\lambda, \delta) \le c_5 \sqrt{\frac{d \log \frac{n}{d} + \log \frac{1}{\delta}}{n}},$$

which implies our first bound (21) by (24).

To prove (22) we use that for $a, b, x > 0$, $\sqrt{ab} \le \frac{a}{2x} + \frac{bx}{2}$. This implies

$$\kappa_1 \sqrt{\left(\sum_{i=1}^n \mathbb{E} f(X_i) \right) \left(d \log \frac{n}{d} + \log \frac{1}{\delta} \right)} \le \frac{1}{2} \sum_{i=1}^n \mathbb{E} f(X_i) + \frac{\kappa_1^2}{2} \left(d \log \frac{n}{d} + \log \frac{1}{\delta} \right) , \tag{27}$$

which, by (21), implies that on the same event where (21) holds,

$$\frac{1}{2} \sum_{i=1}^n \mathbb{E} f(X_i) \le \sum_{i=1}^n f(X_i) + \left(\kappa_1^2/2 + \kappa_1 \right) \left(d \log \frac{n}{d} + \log \frac{1}{\delta} \right) .$$

Plugging this into (21) and adjusting the constant κ_2 proves (22). ■

Appendix B: Proof of Proposition 5

To simply the presentation we assume that the values $\log n, n^{1/3}, n^{1/6}, \ldots$ corresponding to the indexes are always integers. It follows from Theorem 3.1 that there exists a constant $C > 0$ (which only depends on β and $\phi(0)$) such that, with the same probability of error, the adaptive estimator has an error at most

$$C \min \left(\frac{\sqrt{n} \log^{3/2} n}{\sum_{i > C\sqrt{n \log n}} \frac{1}{\sigma_i}}, \sigma_m \right) ,$$

where m is any integer that satisfies

$$m \ge C \max \left(\sqrt{\sigma_m \sum_{i \ge m} \frac{1}{\sigma_i} \log n}, \log n \right) . \tag{28}$$

Therefore, it is sufficient to prove that for all sequences σ_i,

$$\min \left(\frac{\sqrt{n} \log^{3/2} n}{\sum_{i > C\sqrt{n \log n}} \frac{1}{\sigma_i}}, \sigma_m \right) \lesssim \sqrt{n} (\log^{3/2} n) \sigma_{C \log n} .$$

If

$$\frac{\sqrt{n} \log^{3/2} n}{\sum_{i > C\sqrt{n \log n}} \frac{1}{\sigma_i}} \le \sqrt{n} (\log^{3/2} n) \sigma_{C \log n} ,$$

then we are done, so we may assume

$$\frac{\sqrt{n} \log^{3/2} n}{\sum_{i > C\sqrt{n \log n}} \frac{1}{\sigma_i}} > \sqrt{n} (\log^{3/2} n) \sigma_{C \log n} ,$$

or, equivalently,

$$\sum_{i > C\sqrt{n \log n}} \frac{1}{\sigma_i} < \frac{1}{\sigma_{C \log n}} . \tag{29}$$

It suffices to show that, when (29) holds, then there exists a value of m satisfying (28) for which $\sigma_m \leq \sqrt{n}(\log^{3/2} n)\sigma_{C\log n}$. For any $m \leq C\sqrt{n\log n}$, we may write

$$\sum_{i>m} \frac{1}{\sigma_i} = \sum_{i>C\sqrt{n\log n}} \frac{1}{\sigma_i} + \sum_{i\in[m,C\sqrt{n\log n}]} \frac{1}{\sigma_i} \,.$$

Using (29), we see that m satisfies (28) whenever

$$\frac{m^2}{C^2\sigma_m} \geq \max\left(\frac{\log^2 n}{\sigma_m}, \frac{\log n}{\sigma_{C\log n}} + \log n \sum_{i\in[m,C\sqrt{n\log n}]} \frac{1}{\sigma_i}\right) \,. \tag{30}$$

First, note that if the first term dominates on the right-hand side of the above inequality, then $m = C\log n$ satisfies the inequality above, and therefore the new bound is at most $C\sigma_{C\log n}$ and our claim follows.

Hence, we may assume that the second term dominates and therefore we look for the values of m such that

$$\frac{m^2}{C^2\sigma_m} \geq \frac{\log n}{\sigma_{C\log n}} + \log n \sum_{i\in[m,C\sqrt{n\log n})} \frac{1}{\sigma_i} \,. \tag{31}$$

We distinguish two cases depending on which term dominates on the right-hand side: in case (i),

$$\frac{1}{\sigma_{C\log n}} > \sum_{i\in[m,C\sqrt{n\log n})} \frac{1}{\sigma_i} \,,$$

while in case (ii) the opposite holds. In case (i), the right-hand side of (31) is at most $2\log n/\sigma_{C\log n}$. Hence, we may take $m = C\log n$ to satisfy the inequality (28) for n large enough, leading to the bound $C\sigma_{C\log n}$ which proves our claim.

In case (ii), the right-hand side of (31) is bounded by

$$2\log n \sum_{i\in[m,C\sqrt{n\log n})} \frac{1}{\sigma_i} \leq 2C\sqrt{n}(\log^{3/2} n)\frac{1}{\sigma_m} \,. \tag{32}$$

This implies by (31) that the inequality (28) is satisfied when

$$m \geq \sqrt{2}C^{3/2}n^{1/4}(\log^{3/4} n) \,.$$

Since for n large enough

$$n^{1/3} \geq \sqrt{2}C^{3/2}n^{1/4}(\log^{3/4} n) \,,$$

this yields the upper bound σ_{m_1} with $m_1 = n^{1/3}$.

If $\sigma_{m_1} \leq \sqrt{n}(\log^{3/2} n)\sigma_{C\log n}$, then the proof is finished. Otherwise,

$$\sum_{i\in[m,C\sqrt{n\log n})} \frac{1}{\sigma_i} \leq \sum_{i\in[m,m_1)} \frac{1}{\sigma_i} + C\sqrt{n\log n}\frac{1}{\sigma_{m_1}} \leq \sum_{i\in[m,m_1)} \frac{1}{\sigma_i} + \frac{C}{(\log n)\sigma_{C\log n}} \,. \tag{33}$$

Plugging this back to (31), we see that in case (ii), the upper bound becomes σ_m for any m that satisfies

$$\frac{m^2}{C^2\sigma_m} \geq \frac{C+\log n}{\sigma_{C\log n}} + \log n \sum_{i\in[m,m_1)} \frac{1}{\sigma_i} \,. \tag{34}$$

This has the same form as (31) but with a reduced range in the summation on the right-hand side.

We proceed the same way as above. Once again, we consider two cases. In case (iii),

$$\frac{C+\log n}{\sigma_{C\log n}} > \log n \sum_{i\in[m,m_1)} \frac{1}{\sigma_i} \,,$$

while in case (iv),

$$\frac{C+\log n}{\sigma_{C\log n}} \le \log n \sum_{i\in[m,m_1)} \frac{1}{\sigma_i} ,$$

In case (iii), the right-hand side of (34) is at most $2(C+\log n)/\sigma_{C\log n}$, so, just like before, we may take $m = C\log n$ to satisfy the inequality (31), leading to the bound $C\sigma_{C\log n}$ whenever $\log n \gtrsim C$.

In case (iv), the right-hand side of (34) is bounded by

$$2\log n \sum_{i\in[m,m_1)} \frac{1}{\sigma_i} \le 2\log n \frac{m_1}{\sigma_m} = \frac{2n^{1/3}\log n}{\sigma_m} . \tag{35}$$

Thus, in this case (30) is satisfied for any $m \ge 2Cn^{1/6}\log^{3/2} n$, and in particular, for $m_2 = n^{2/9}$. If $\sigma_{m_2} \le \sqrt{n}(\log^{3/2} n)\sigma_{C\log n}$, then the proof is finished. Otherwise,

$$\sum_{i\in[m,m_1)} \frac{1}{\sigma_i} \le \sum_{i\in[m,m_2)} \frac{1}{\sigma_i} + \frac{m_1}{\sigma_{m_2}} \le \sum_{i\in[m,m_2)} \frac{1}{\sigma_i} + \frac{1}{n^{1/6}(\log^{3/2} n)\sigma_{C\log n}} .$$

Resubstituting into (31), we see that in case (iv), the upper bound becomes σ_m for any m that satisfies

$$\frac{m^2}{C^2\sigma_m} \ge \frac{C+\log n}{\sigma_{C\log n}} + \frac{1}{n^{1/6}(\log^{1/2} n)\sigma_{C\log n}} + \log n \sum_{i\in[m,m_2)} \frac{1}{\sigma_i} .$$

We may now continue the same fashion, at each step reducing the range of the sum on the right-hand side unless at the j-th iteration $\sigma_{m_j} \le \sqrt{n}(\log^{3/2} n)\sigma_{C\log n}$ and we are done. In general, at the j-th iteration, the summation is between m and $m_j = n^{(2/3)^j/2}$. If we reach the j-th iteration such that $m_j = C\log^{1/2} n$, we have

$$\log n \sum_{i\in[m,m_j)} \frac{1}{\sigma_i} \le \frac{C\log^{3/2} n}{\sigma_m} ,$$

so that one may choose $m = C\log n$ for large enough n. The claim follows.

Acknowledgments

The authors would like to thank the anonymous referees and an Associate Editor for their comments that improved the presentation of the paper.

Funding

Luc Devroye was supported by NSERC Discovery Grants and by an FRQNT Team Research Grant. Gábor Lugosi was supported by the Spanish Ministry of Economy and Competitiveness, Grant PGC2018-101643-B-I00 and FEDER, EU, and by "Google Focused Award Algorithms and Learning for AI". Nikita Zhivotovskiy is funded in part by ETH Foundations of Data Science (ETH-FDS).

References

[1] Bartlett, P. L., Bousquet, O. and Mendelson, S. [2005]. Local Rademacher complexities, *The Annals of Statistics* **33**(4): 1497–1537.
[2] Beran, R. [1982]. Robust estimation in models for independent non-identically distributed data, *The Annals of Statistics* pp. 415–428.
[3] Blum, M., Floyd, R. W., Pratt, V. R., Rivest, R. L. and Tarjan, R. E. [1973]. Time bounds for selection, *J. Comput. Syst. Sci.* **7**(4): 448–461.
[4] Boucheron, S., Lugosi, G. and Massart, P. [2013]. *Concentration Inequalities: A Nonasymptotic Theory Of Independence*, Oxford university press.
[5] Bousquet, O. and Zhivotovskiy, N. [2021]. Fast classification rates without standard margin assumptions, *Information and Inference: A Journal of the IMA* **10**(4): 1389–1421.
[6] Carroll, R. J. and Ruppert, D. [1982]. Robust estimation in heteroscedastic linear models, *The Annals of Statistics* pp. 429–441.
[7] Catoni, O. [2004]. Improved Vapnik Cervonenkis bounds, *arXiv preprint math/0410280* .
[8] Chernoff, H. [1964]. Estimation of the mode, *Annals of the Institute of Statistical Mathematics* **16**(1): 31–41.

[9] Chierichetti, F., Dasgupta, A., Kumar, R. and Lattanzi, S. [2014]. Learning entangled single-sample gaussians, *Proceedings of the Twenty-fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, SIAM, pp. 511–522.

[10] Fuller, W. A. and Rao, J. [1978]. Estimation for a linear regression model with unknown diagonal covariance matrix, *The Annals of Statistics* pp. 1149–1158.

[11] Giné, E. and Nickl, R. [2016]. *Mathematical Foundations of Infinite-dimensional Statistical Models*, Vol. 40, Cambridge University Press.

[12] Gordon, Y., Litvak, A., Schütt, C. and Werner, E. [2006]. On the minimum of several random variables, *Proceedings of the American Mathematical Society* **134**(12): 3665–3675.

[13] Haussler, D. [1995]. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension, *J. Comb. Theory, Ser. A* **69**(2): 217–232.

[14] Ibragimov, I. and Has'minskii, R. [1981]. *Statistical Estimation: Asymptotic Theory*, Springer.

[15] Ibragimov, I. and Has'minskii, R. [1976]. Local asymptotic normality for non-identically distributed observations, *Theory of Probability & Its Applications* **20**(2): 246–260.

[16] Ledoux, M. and Talagrand, M. [2013]. *Probability in Banach Spaces: Isoperimetry and Processes*, Springer Science & Business Media.

[17] Lepskii, O. [1992]. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates, *Theory of Probability & Its Applications* **36**(4): 682–697.

[18] Liang, Y. and Yuan, H. [2020]. Learning entangled single-sample gaussians in the subset-of-signals model, *Conference on Learning Theory* pp. 2712–2737.

[19] Mizera, I. and Wellner, J. A. [1998]. Necessary and sufficient conditions for weak consistency of the median of independent but not identically distributed random variables, *Annals of Statistics* pp. 672–691.

[20] Nevzorov, V. [1984]. Rate of convergence to the normal law of order statistics for nonidentically distributed random variables, *Journal of Soviet Mathematics* **27**(6): 3263–3270.

[21] Pensia, A., Jog, V. and Loh, P.-L. [2021]. Estimating location parameters in sample-heterogeneous distributions, *Information and Inference: A Journal of the IMA* .

[22] Vapnik, V. and Chervonenkis, A. [1974]. *Theory of Pattern Recognition*, Nauka. Moscow.

[23] Xia, D. [2019]. Non-asymptotic bounds for percentiles of independent non-identical random variables, *Statistics & Probability Letters* **152**: 111–120.

[24] Yuan, H. and Liang, Y. [2020]. Learning entangled single-sample distributions via iterative trimming, *In proceedings of AISTATS, 2020* pp. 2666–2676.