Electronic Journal of Statistics
Vol. 14 (2020) 2653–2677
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1731

Confidence regions and minimax rates in outlier-robust estimation on the probability simplex

Amir-Hossein Bateni and Arnak S. Dalalyan

CREST-ENSAE-IP Paris, 5 Avenue Le Chatelier, 91120 Palaiseau, France

Abstract: We consider the problem of estimating the mean of a distribution supported by the k-dimensional probability simplex in the setting where an ε fraction of observations are subject to adversarial corruption. A simple particular example is the problem of estimating the distribution of a discrete random variable. Assuming that the discrete variable takes k values, the unknown parameter $\boldsymbol{\theta}$ is a k-dimensional vector belonging to the probability simplex. We first describe various settings of contamination and discuss the relation between these settings. We then establish minimax rates when the quality of estimation is measured by the total-variation distance, the Hellinger distance, or the $\mathbb{L}^2$-distance between two probability measures. We also provide confidence regions for the unknown mean that shrink at the minimax rate. Our analysis reveals that the minimax rates associated to these three distances are all different, but they are all attained by the sample average. Furthermore, we show that the latter is adaptive to the possible sparsity of the unknown vector. Some numerical experiments illustrating our theoretical findings are reported.

MSC 2010 subject classifications: Primary 62F35; secondary 62H12.
Keywords and phrases: Robust estimation, discrete models, confidence regions.

Received January 2020.

Contents

1. Introduction

Assume $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are n independent and identically distributed random variables taking their values in the k-dimensional probability simplex $\Delta^{k-1} = \{\boldsymbol{v} \in \mathbb{R}^k_+ : v_1 + \ldots + v_k = 1\}$. Our goal is to estimate the unknown vector $\boldsymbol{\theta} = \mathbf{E}[\boldsymbol{X}_i]$ in the case where the observations are contaminated by outliers. An important particular case is the estimation of the distribution of a discrete random variable $\boldsymbol{X}$ taking k distinct values. In this particular case, $\boldsymbol{X}_i$'s take values in $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_k\}$, the set of the vectors of the canonical basis, which are also the extreme points of the simplex Δ^{k-1}.

In this introduction, to convey the main messages, we limit ourselves to the Huber contamination model, although our results apply to the more general adversarial contamination. Huber's contamination model assumes that there are two probability measures $\boldsymbol{P}$, $\boldsymbol{Q}$ on Δ^{k-1} and a real $\varepsilon \in [0, 1/2)$ such that $\boldsymbol{X}_i$ is drawn from

$$\boldsymbol{P}_i = (1-\varepsilon)\boldsymbol{P} + \varepsilon \boldsymbol{Q}, \qquad \forall i \in \{1, \dots, n\}.$$

This amounts to assuming that $(1-\varepsilon)$-fraction of observations, called inliers, are drawn from a reference distribution $\boldsymbol{P}$, whereas ε-fraction of observations are outliers and are drawn from another distribution $\boldsymbol{Q}$. In general, all the three parameters $\boldsymbol{P}$, $\boldsymbol{Q}$ and ε are unknown. The parameter of interest is some functional (such as the mean, the standard deviation, etc.) of the reference distribution $\boldsymbol{P}$, whereas $\boldsymbol{Q}$ and ε play the role of nuisance parameters.

When the unknown parameter lives on the probability simplex, there are many appealing ways of defining the risk. We focus on the following three metrics: total-variation, Hellinger and $\mathbb{L}^2$ distances[1]

$$\begin{aligned} d_{\mathrm{TV}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &:= {}^1\!/_2 \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1, \\ d_{\mathrm{H}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &:= {}^1\!/_{\sqrt{2}} \|\widehat{\boldsymbol{\theta}}^{1/2} - \boldsymbol{\theta}^{1/2}\|_2, \\ d_{\mathbb{L}^2}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &:= \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2. \end{aligned}$$

The Hellinger distance above is well defined when the estimator $\widehat{\boldsymbol{\theta}}$ is non-negative, which will be the case throughout this work. We will further assume that the dimension k may be large, but the vector $\boldsymbol{\theta}$ is s-sparse, for some $s \le k$, *i.e.* $\#\{j : \theta_j \neq 0\} \le s$. Our main interest is in constructing confidence regions and evaluating the minimax risk

$$\mathfrak{R}_\square(n, k, s, \varepsilon) := \inf_{\bar{\boldsymbol{\theta}}_n} \sup_{\boldsymbol{P}, \boldsymbol{Q}} \mathbf{E}[d_\square(\bar{\boldsymbol{\theta}}_n, \boldsymbol{\theta})], \tag{1}$$

where the *inf* is over all estimators $\bar{\boldsymbol{\theta}}_n$ built upon the observations $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \overset{\text{iid}}{\sim} (1-\varepsilon)\boldsymbol{P} + \varepsilon\boldsymbol{Q}$ and the *sup* is over all distributions $\boldsymbol{P}$, $\boldsymbol{Q}$ on the probability simplex such that the mean $\boldsymbol{\theta}$ of $\boldsymbol{P}$ is s-sparse. The subscript $\square$ of $\mathfrak{R}$ above refers to the distance used in the risk, so that $\square$ is TV, H, or $\mathbb{L}^2$.

[1] We write $\|\boldsymbol{u}\|_q = (\sum_{j=1}^k |u_j|^q)^{1/q}$ and $\boldsymbol{u}^q = (u_1^q, \dots, u_k^q)$ for any $\boldsymbol{u} \in \mathbb{R}^k_+$ and $q > 0$.

The problem described above arises in many practical situations. One example is an election poll: each participant expresses his intention to vote for one of k candidates. Thus, each θ_j is the true proportion of electors of candidate j. The results of the poll contain outliers, since some participants of the poll prefer to hide their true opinion. Another example related to elections, is the problem of counting votes across all constituencies. Each constituency communicates a vector of proportions to a central office, which is in charge of computing the overall proportions. However, in some constituencies (hopefully a small fraction only) the results are rigged. Hence, the set of observed vectors contains outliers.

We intend to provide non-asymptotic upper and lower bounds on the minimax risk that match up to numerical constants. In addition, we will provide confidence regions of the form $B_\square(\widehat{\boldsymbol{\theta}}_n, r_{n,\varepsilon,\delta}) = \{\boldsymbol{\theta} : d_\square(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}) \le r_{n,\varepsilon,\delta}\}$ containing the true parameter with probability at least $1-\delta$ and such that the radius $r_{n,\varepsilon,\delta}$ goes to zero at the same rate as the corresponding minimax risk.

When there is no outlier, *i.e.*, $\varepsilon = 0$, it is well known that the sample mean

$$\bar{\boldsymbol{X}}_n := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$$

is minimax-rate-optimal and the rates corresponding to various distances are

$$\mathfrak{R}_{\mathbb{L}^2}(n,k,s,0) \asymp (1/n)^{1/2} \quad \text{and} \quad \mathfrak{R}_\square(n,k,s,0) \asymp (s/n)^{1/2} \quad \text{for } \square \in \{\mathrm{TV}, \mathrm{H}\}.$$

This raises several questions in the setting where data contains outliers. In particular, the following three questions will be answered in this work:

Q1. How do the risks $\mathfrak{R}_\square$ depend on ε? What is the largest proportion of outliers for which the minimax rate is the same as in the outlier-free case?
Q2. Does the sample mean remain optimal in the contaminated setting?
Q3. What does happen if the unknown parameter $\boldsymbol{\theta}$ is s-sparse?

The most important step for answering these questions is to show that

$$\begin{aligned}
\mathfrak{R}_{\mathrm{TV}}(n,k,s,\varepsilon) &\asymp (s/n)^{1/2} + \varepsilon,\\
\mathfrak{R}_{\mathrm{H}}(n,k,s,\varepsilon) &\asymp (s/n)^{1/2} + \varepsilon^{1/2},\\
\mathfrak{R}_{\mathbb{L}^2}(n,k,s,\varepsilon) &\asymp (1/n)^{1/2} + \varepsilon.
\end{aligned}$$

It is surprising to see that all the three rates are different leading to important discrepancies in the answers to the second part of question Q1 for different distances. Indeed, it turns out that the minimax rate is not deteriorated if the proportion of the outliers is smaller than $(s/n)^{1/2}$ for the TV-distance, s/n for the Hellinger distance and $(1/n)^{1/2}$ for the $\mathbb{L}^2$ distance. Furthermore, we prove that the sample mean is minimax rate optimal. Thus, even when the proportion of outliers ε and the sparsity s are known, it is not possible to improve upon the sample mean. In addition, we show that all these claims hold true for the adversarial contamination and we provide corresponding confidence regions.

The rest of the paper is organized as follows. Section 2 introduces different possible ways of modeling data sets contaminated by outliers. Pointers to

relevant prior work are given in Section 3. Main theoretical results and their numerical illustration are reported in Section 4 and Section 5, respectively. Section 6 contains a brief summary of the obtained results and their consequences, whereas the proofs are postponed to the appendix.

2. Various models of contamination

Different mathematical frameworks have been used in the literature to model the outliers. We present here five of them, from the most restrictive one to the most general, and describe their relationship. We present these frameworks in the general setting when the goal is to estimate the parameter $\boldsymbol{\theta}^*$ of a reference distribution $\boldsymbol{P}_{\boldsymbol{\theta}^*}$ when ε proportion of the observations are outliers.

2.1. Huber's contamination

The most popular framework for studying robust estimation methods is perhaps the one of Huber's contamination. In this framework, there is a distribution $\boldsymbol{Q}$ defined on the same space as the reference distribution $\boldsymbol{P}_{\boldsymbol{\theta}^*}$ such that all the observations $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are independent and drawn from the mixture distribution $\boldsymbol{P}_{\varepsilon,\boldsymbol{\theta}^*,\boldsymbol{Q}} := (1-\varepsilon)\boldsymbol{P}_{\boldsymbol{\theta}^*} + \varepsilon\boldsymbol{Q}$.

This corresponds to the following mechanism: one decides with probabilities $(1-\varepsilon, \varepsilon)$ whether a given observation is an inlier or an outlier. If the decision is made in favor of being inlier, the observation is drawn from $\boldsymbol{P}_{\boldsymbol{\theta}^*}$, otherwise it is drawn from $\boldsymbol{Q}$. More formally, if we denote by $\widehat{O}$ the random set of outliers, then conditionally to $\widehat{O} = O$,

$$\{\boldsymbol{X}_i : i \notin O\} \overset{\text{iid}}{\sim} \boldsymbol{P}_{\boldsymbol{\theta}^*},\ \{\boldsymbol{X}_i : i \in O\} \overset{\text{iid}}{\sim} \boldsymbol{Q},\ \{\boldsymbol{X}_i : i \in O\} \perp\!\!\!\perp \{\boldsymbol{X}_i : i \notin O\}, \quad (2)$$

for every $O \subset \{1, \dots, n\}$. Furthermore, for every subset O of the observations, we have $\boldsymbol{P}(\widehat{O} = O) = (1-\varepsilon)^{n-|O|}\varepsilon^{|O|}$. We denote by[2] $\mathcal{M}_n^{\text{HC}}(\varepsilon, \boldsymbol{\theta}^*)$ the set of joint probability distributions $\boldsymbol{P}_n$ of the random variables $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ satisfying the foregoing condition.

2.2. Huber's deterministic contamination

The set of outliers as well as the number of outliers in Huber's model of contamination are random. This makes it difficult to compare this model to the others that will be described later in this section. To cope with this, we define here another model, termed Huber's deterministic contamination. As its name indicates, this new model has the advantage of containing a deterministic number of outliers, in the same time being equivalent to Huber's contamination in a sense that will be made precise below.

[2]The superscript HC refers to the Huber's contamination.

We say that the distribution $\boldsymbol{P}_n$ of $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ belongs to the Huber's deterministic contamination model denoted by $\mathcal{M}_n^{\mathrm{HDC}}(\varepsilon, \boldsymbol{\theta}^*)$, if there are a set $O \subset \{1, \dots, n\}$ of cardinality at most $n\varepsilon$ and a distribution $\boldsymbol{Q}$ such that (2) is true. The apparent similarity of models $\mathcal{M}_n^{\mathrm{HC}}(\varepsilon, \boldsymbol{\theta}^*)$ and $\mathcal{M}_n^{\mathrm{HDC}}(\varepsilon, \boldsymbol{\theta}^*)$ can also be formalized mathematically in terms of the orders of magnitude of minimax risks. To ease notation, we let $R_d^{\square}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}})$ to be the worst-case risk of an estimator $\widehat{\boldsymbol{\theta}}$, where $\square$ is either HC or HDC. More precisely, for $\mathcal{M}_n^{\square}(\varepsilon, \Theta) := \cup_{\boldsymbol{\theta} \in \Theta} \mathcal{M}_n^{\square}(\varepsilon, \boldsymbol{\theta})$, we set[3]

$$R_d^{\square}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}) := \sup_{P_n \in \mathcal{M}_n^{\square}(\varepsilon, \Theta)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)].$$

This definition assumes that the parameter space Θ is endowed with a pseudo-metric $d : \Theta \times \Theta \to \mathbb{R}_+$. When $\Theta = \{\boldsymbol{\theta}^*\}$ is a singleton, we write $R_{d,n}^{\square}(\varepsilon, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}})$ instead of $R_d^{\square}(n, \varepsilon, \{\boldsymbol{\theta}^*\}, \widehat{\boldsymbol{\theta}})$.

Proposition 1 Let $\widehat{\boldsymbol{\theta}}_n$ be an arbitrary estimator of $\boldsymbol{\theta}^*$. For any $\varepsilon \in (0, 1/2)$,

$$R_d^{\mathrm{HC}}(n, \varepsilon, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}_n) \le R_{d,n}^{\mathrm{HDC}}(2\varepsilon, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}_n) + e^{-n\varepsilon/3} R_{d,n}^{\mathrm{HDC}}(1, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}_n), \quad (3)$$

$$\sup_{\boldsymbol{P}_n \in \mathcal{M}_n^{\mathrm{HC}}(\varepsilon, \boldsymbol{\theta}^*)} r\mathbf{P}\big(d(\widehat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^*) > r\big) \le R_{d,n}^{\mathrm{HDC}}(2\varepsilon, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}_n) + re^{-n\varepsilon/3}. \quad (4)$$

Proof in the appendix, page 2667

Denote by $\mathcal{D}_\Theta$ the diameter of Θ, $\mathcal{D}_\Theta := \max_{\boldsymbol{\theta}, \boldsymbol{\theta}'} d(\boldsymbol{\theta}, \boldsymbol{\theta}')$. Proposition 1 implies that

$$\inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathrm{HC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) \le \inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathrm{HDC}}(n, 2\varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) + e^{-n\varepsilon/3} \mathcal{D}_\Theta. \quad (5)$$

When Θ is bounded, the last term is typically of smaller order than the minimax risk over $\mathcal{M}_n^{\mathrm{HDC}}(2\varepsilon, \Theta)$. Therefore, the minimax rate of estimation in Huber's model is not slower than the minimax rate of estimation in Huber's deterministic contamination model. This entails that a lower bound on the minimax risk established in HC-model furnishes a lower bound in HDC-model.

2.3. Oblivious contamination

A third model of contamination that can be of interest is the oblivious contamination. In this model, it is assumed that the set O of cardinality o and the joint distribution $\boldsymbol{Q}_O$ of outliers are determined in advance, possibly based on the knowledge of the reference distribution $\boldsymbol{P}_{\boldsymbol{\theta}^*}$. Then, the outliers $\{\boldsymbol{X}_i : i \in O\}$ are drawn randomly from $\boldsymbol{Q}_O$ independently of the inliers $\{\boldsymbol{X}_i : i \in O^c\}$. The set of

[3]The subscript d refers to the distance d used in the definition of the risk.

all the joint distributions $\boldsymbol{P}_n$ of random variables $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ generated by such a mechanism will be denoted by $\mathcal{M}_n^{\mathrm{OC}}(\varepsilon, \boldsymbol{\theta}^*)$. The model of oblivious contamination is strictly more general than that of Huber's deterministic contamination, since it does not assume that the outliers are iid. Therefore, the minimax risk over $\mathcal{M}_n^{\mathrm{OC}}(\varepsilon, \Theta)$ is larger than the minimax risk over $\mathcal{M}_n^{\mathrm{HDC}}(\varepsilon, \Theta)$:

$$\inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathrm{HDC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) \le \inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathrm{OC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n).$$

The last inequality holds true for any set Θ, any contamination level $\varepsilon \in (0, 1)$ and any sample size.

2.4. Parameter contamination

In the three models considered above, the contamination acts on the observations. One can also consider the case where the parameters of the distributions of some observations are contaminated. More precisely, for some set $O \subset \{1, \dots, n\}$ selected in advance (but unobserved), the outliers $\{\boldsymbol{X}_i : i \in O\}$ are independent and independent of the inliers $\{\boldsymbol{X}_i : i \in O^c\}$. Furthermore, each outlier $\boldsymbol{X}_i$ is drawn from a distribution $\boldsymbol{Q}_i = \boldsymbol{P}_{\boldsymbol{\theta}_i}$ belonging to the same family as the reference distribution, but corresponding to a contaminated parameter $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}^*$. Thus, the joint distribution of the observations can be written as $(\bigotimes_{i \in O^c} \boldsymbol{P}_{\boldsymbol{\theta}^*}) \otimes (\bigotimes_{i \in O} \boldsymbol{P}_{\boldsymbol{\theta}_i})$. The set of all such distributions $\boldsymbol{P}_n$ will be denoted by $\mathcal{M}_n^{\mathrm{PC}}(\varepsilon, \boldsymbol{\theta}^*)$, where PC refers to "parameter contamination".

2.5. Adversarial contamination

The last model of contamination we describe in this work, the adversarial contamination, is the most general one. It corresponds to the following two-stage data generation mechanism. In a first stage, iid random variables $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ are generated from a reference distribution $\boldsymbol{P}_{\boldsymbol{\theta}^*}$. In a second stage, an adversary having access to $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ chooses a (random) set $\widehat{O}$ of (deterministic) cardinality s and arbitrarily modifies data points $\{\boldsymbol{Y}_i : i \in \widehat{O}\}$. The resulting sample, $\{\boldsymbol{X}_i : i = 1, \dots, n\}$, is revealed to the Statistician. In this model, we have $\boldsymbol{X}_i = \boldsymbol{Y}_i$ for $i \notin \widehat{O}$. However, since $\widehat{O}$ is random and potentially dependent of $\boldsymbol{Y}_{1:n}$, it is not true that conditionally to $\widehat{O} = O$, $\{\boldsymbol{X}_i : i \in O^c\}$ are iid drawn from $\boldsymbol{P}_{\boldsymbol{\theta}^*}$ (for any deterministic set O of cardinality o).

We denote by $\mathcal{M}_n^{\mathrm{AC}}(\varepsilon, \boldsymbol{\theta}^*)$ the set of all the joint distributions $\boldsymbol{P}_n$ of all the sequences $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ generated by the aforementioned two-stage mechanism. This set $\mathcal{M}_n^{\mathrm{AC}}(\varepsilon, \boldsymbol{\theta}^*)$ is larger than all the four sets of contamination introduced in this section. Therefore, the following inequalities hold:

$$\inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathrm{PC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) \le \inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathrm{OC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) \le \inf_{\widehat{\boldsymbol{\theta}}_n} R_d^{\mathrm{AC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n),$$

for any n, ε, Θ and any distance d.

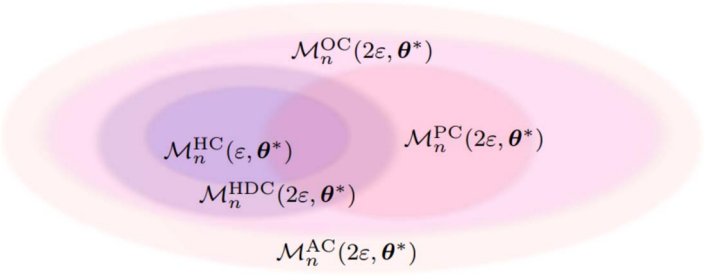

FIG 1. *Visual representation of the hierarchy between various contamination models. Note that the inclusion of* $\mathcal{M}_n^{HC}(\varepsilon, \boldsymbol{\theta}^*)$ *in* $\mathcal{M}_n^{HDC}(2\varepsilon, \boldsymbol{\theta}^*)$ *is somewhat heuristic, based on the relation on the worst-case risks reported in Proposition 1.*

2.6. Minimax risk "in expectation" versus "in deviation"

Most prior work on robust estimation focused on establishing upper bounds on the minimax risk in deviation,[4] as opposed to the minimax risk in expectation defined by (1). One of the reasons for dealing with the deviation is that it makes the minimax risk meaningful for models[5] having random number of outliers and unbounded parameter space Θ. The formal justification of this claim is provided by the following result.

Proposition 2 Let Θ be a parameter space such that $\mathcal{D}_\Theta = \sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \Theta} d(\boldsymbol{\theta}, \boldsymbol{\theta}') = +\infty$. Then, for every estimator $\widehat{\theta}_n$, every $\varepsilon > 0$ and $n \in \mathbb{N}$, we have $R_d^{\mathrm{HC}}(n, \varepsilon, \Theta, \widehat{\boldsymbol{\theta}}_n) = +\infty$.

Proof in the appendix, page 2668

This result shows, in particular, that the last term in (5), involving the diameter of Θ is unavoidable. Such an explosion of the minimax risk occurs because Huber's model allows the number of outliers to be as large as $n/2$ with a strictly positive probability. One approach to overcome this shortcoming is to use the minimax risk in deviation. Another approach is to limit theoretical developments to the models HDC, PC, OC or AC, in which the number of outliers is deterministic.

3. Prior work

Robust estimation is an area of active research in Statistics since at least five decades (Huber, 1964; Tukey, 1975; Donoho and Huber, 1983; Donoho and Gasko, 1992; Rousseeuw and Hubert, 1999). Until very recently, theoretical guarantees were almost exclusively formulated in terms of the notions of breakdown

[4]We call a risk bound in deviation any bound on the distance $d(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$ that holds true with a probability close to one, for any parameter value $\boldsymbol{\theta}^* \in \Theta$.

[5]This is the case, for instance, of the Gaussian model with Huber's contamination.

point, sensitivity curve, influence function, etc. These notions are well suited for accounting for gross outliers, observations that deviate significantly from the data points representative of an important fraction of data set.

More recently, various authors investigated (Nguyen and Tran, 2013; Dalalyan and Chen, 2012; Chen et al., 2013) the behavior of the risk of robust estimators as a function of the rate of contamination ε. A general methodology for parametric models subject to Huber's contamination was developed in Chen et al. (2018, 2016). This methodology allowed for determining the rate of convergence of the minimax risk as a function of the sample size n, dimension k and the rate of contamination ε. An interesting phenomenon was discovered: in the problem of robust estimation of the Gaussian mean, classic robust estimators such as the coordinatewise median or the geometric median do not attain the optimal rate $(k/n)^{1/2} + \varepsilon$. This rate is attained by Tukey's median, the maximaizer of Tukey's halfspace depth, the computation of which is costly in a high dimensional setting. Detailed analysis of Tukey's halfspace depth was conducted in Brunel (2019).

In the model analyzed in this paper, we find the same minimax rate, $(k/n)^{1/2} + \varepsilon$, only when the total-variation distance is considered. A striking difference is that this rate is attained by the sample mean which is efficiently computable in any dimension. This property is to some extent similar to the problem of robust density estimation (Liu and Gao, 2019), in which the standard kernel estimators are minimax optimal in contaminated setting. Note that the fact that in the sparse setting the improvement from $(k/n)^{1/2}$ to $(s/n)^{1/2}$ can be achieved without any penalization, and that the constraint of belonging to the probability simplex acts as a sparsity favoring penalty, was already known in the literature, see Xia and Koltchinskii (2016); Dalalyan and Sebbar (2018). Interestingly, similar phenomenon is observed in problems of estimation under shape constraints (Bellec, 2018; Guntuboyina and Sen, 2018). It is an interesting avenue of future research to analyze the robustness of the maximum likelihood estimator in this context.

Computational intractability of Tukey's median motivated a large number of studies that aimed at designing computationally tractable methods with nearly optimal statistical guarantees. Many of these works went beyond Huber's contamination by considering parameter contamination models (Bhatia et al., 2017; Collier and Dalalyan, 2019; Carpentier et al., 2018), oblivious contamination (Feng et al., 2014; Lai et al., 2016) or adversarial contamination (Diakonikolas et al., 2016; Balakrishnan et al., 2017; Diakonikolas et al., 2017, 2018; Dalalyan and Minasyan, 2020). Interestingly, in the problem of estimating the Gaussian mean, it was proven that the minimax rates under adversarial contamination are within a factor at most logarithmic in n and k of the minimax rates under Huber's contamination.[6] While each of the aforementioned papers introduced clearly the conditions on the contamination, to our knowledge, none of them described different possible models and the relationship between them.

[6] All these papers consider the risk in deviation, so that the minimax risk under Huber's contamination is finite.

Another line of growing literature on robust estimation aims at robustifying estimators and prediction methods to heavy tailed distributions, see Audibert and Catoni (2011); Minsker (2015); Donoho and Montanari (2016); Devroye et al. (2016); Joly et al. (2017); Minsker (2018); Lugosi and Mendelson (2019); Lecué and Lerasle (2017); Chinot et al. (2018). The results of those papers are of a different nature, as compared to the present work, not only in terms of the goals, but also in terms of mathematical and algorithmic tools.

In the case of the discrete distributions, Braess and Sauer (2004) established the minimax rates under the Kullback-Leibler divergence. More recently, Kamath et al. (2015) determined the minimax rates under other distances such as $\mathbb{L}^2$, TV and and the general family of f-divergence (including the $\boldsymbol{\chi}^2$-divergence and the Hellinger distance). The estimator proposed in Kamath et al. (2015), achieving the minimax rate for $\mathbb{L}^2$ and TV distances, is the sample mean while different estimators are proposed for the other distances. Concerning the robust estimation of discrete distributions, Qiao and Valiant (2018); Chen et al. (2020); Jain and Orlitsky (2019) studied the case of group-contamination. The distinctive feature of this setting is a better dependence of the minimax rate on the contamination rate ε. More precisely, if each group contains m samples, and ε fraction of groups are contaminated, the rates are obtained by replacing ε by $\varepsilon/\sqrt{m}$. The estimators achieving these rates are much more sophisticated than the sample mean.

4. Minimax rates on the "sparse" simplex and confidence regions

We now specialize the general setting of Section 2 to a reference distribution $\boldsymbol{P}$, with expectation $\boldsymbol{\theta}^*$, defined on the simplex Δ^{k-1}. Along with this reference model describing the distribution of inliers, we will use different models of contamination. More precisely, we will establish upper bounds on worst-case risks of the sample mean in the most general, adversarial, contamination setting. Then, matching lower bounds will be provided for minimax risks under Huber's contamination.

4.1. Upper bounds: worst-case risk of the sample mean

We denote by Δ_s^{k-1} the set of all $\boldsymbol{v} \in \Delta^{k-1}$ having at most s non-zero entries.

Theorem 1 For every triple of positive integers (k, s, n) and for every $\varepsilon \in [0, 1]$, the sample mean $\bar{\boldsymbol{X}}_n := \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$ satisfies

$$R_{\mathrm{TV}}^{\mathrm{AC}}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{X}}_n) \le (s/n)^{1/2} + 2\varepsilon,$$

$$R_{\mathrm{H}}^{\mathrm{AC}}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{X}}_n) \le (s/n)^{1/2} + \sqrt{2}\,\varepsilon^{1/2},$$

$$R_{\mathbb{L}^2}^{\mathrm{AC}}(n, \varepsilon, \Delta_s^{k-1}, \bar{\boldsymbol{X}}_n) \le (1/n)^{1/2} + \sqrt{2}\,\varepsilon.$$

Proof in the appendix, page 2669

An unexpected and curious phenomenon unveiled by this theorem is that all the three rates are different. As a consequence, the answer to the question "what is the largest possible number of outliers, $o_d^*(n,s)$, that does not impact the minimax rate of estimation of $\boldsymbol{\theta}^*$?" crucially depends on the considered distance d. Taking into account the relation $\varepsilon = o/n$, we get

$$o^*_{\mathrm{TV}}(n,s) \asymp (ns)^{1/2}, \quad o^*_{\mathrm{H}}(n,s) \asymp s, \quad o^*_{\mathbb{L}^2}(n,s) \asymp n^{1/2}.$$

Furthermore, all the claims concerning the total variation distance, in the considered model, yield corresponding claims for the Wasserstein distances W_q, for every $q \geq 1$. Indeed, one can see an element $\boldsymbol{\theta} \in \Delta^{k-1}$ as the probability distribution of a random vector $\boldsymbol{X}$ taking values in the finite set $\mathcal{A} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k\}$ of vectors of the canonical basis of $\mathbb{R}^k$. Since these vectors satisfy $\|\boldsymbol{e}_j - \boldsymbol{e}_{j'}\|_2^2 = 2\mathbb{1}(j \neq j')$, we have

$$\begin{aligned} W_q^q(\boldsymbol{\theta}, \boldsymbol{\theta}') &= \inf_{\Gamma} \mathbf{E}_{(\boldsymbol{X},\boldsymbol{X}')\sim\Gamma}[\|\boldsymbol{X} - \boldsymbol{X}'\|_2^q] \\ &= \inf_{\Gamma} 2^{q/2} \boldsymbol{P}(\boldsymbol{X} \neq \boldsymbol{X}') = 2^{q/2} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_{\mathrm{TV}}, \end{aligned} \tag{6}$$

where the *inf* is over all joint distributions Γ on $\mathcal{A} \times \mathcal{A}$ having marginal distributions $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. This implies that

$$R^{\mathrm{AC}}_{W_q}(n, \varepsilon, \Delta^{k-1}_s) \leq \sqrt{2}\big\{(s/n)^{1/2} + 2\varepsilon\big\}^{1/q}, \qquad \forall q \geq 1. \tag{7}$$

In addition, since the $\mathbb{L}_2$ norm is an upper bound on the $\mathbb{L}_\infty$-norm, we have $R^{\mathrm{AC}}_{\mathbb{L}_\infty}(n, \varepsilon, \Delta^{k-1}) \leq (1/n)^{1/2} + \sqrt{2}\,\varepsilon$. Thus, we have obtained upper bounds on the risk of the sample mean for all commonly used distances on the space of probability measures.

4.2. Lower bounds on the minimax risk

A natural question, answered in the next theorem, is how tight are the upper bounds obtained in the last theorem. More importantly, one can wonder whether there is an estimator that has a worst-case risk of smaller order than that of the sample mean.

Theorem 2 There are universal constants $c > 0$ and n_0, such that for any integers $k \geq 3$, $s \leq k \wedge n$, $n \geq n_0$ and for any $\varepsilon \in [0,1]$, we have

$$\begin{aligned} \inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathrm{HC}}_{\mathrm{TV}}(n, \varepsilon, \Delta^{k-1}_s, \bar{\boldsymbol{\theta}}_n) &\geq c\{(s/n)^{1/2} + \varepsilon\}, \\ \inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathrm{HC}}_{\mathrm{H}}(n, \varepsilon, \Delta^{k-1}_s, \bar{\boldsymbol{\theta}}_n) &\geq c\{(s/n)^{1/2} + \varepsilon^{1/2}\}, \\ \inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathrm{HC}}_{\mathbb{L}^2}(n, \varepsilon, \Delta^{k-1}_s, \bar{\boldsymbol{\theta}}_n) &\geq c\{(1/n)^{1/2} + \varepsilon\}, \end{aligned}$$

where $\inf_{\bar{\boldsymbol{\theta}}_n}$ stands for the infimum over all measurable functions $\bar{\boldsymbol{\theta}}_n$ from $(\Delta^{k-1})^n$ to Δ^{k-1}.

Proof in the appendix, page 2671

The main consequence of this theorem is that whatever the contamination model is (among those described in Section 2), the rates obtained for the MLE in Section 4.1 are minimax optimal. Indeed, Theorem 2 yields this claim for Huber's contamination. For Huber's deterministic contamination and and the TV-distance, on the one hand, we have

$$\begin{aligned} R_{\mathrm{TV}}^{\mathrm{HDC}}(n,\varepsilon,\Delta_s^{k-1},\bar{\boldsymbol{\theta}}_n) &\geq R_{\mathrm{TV}}^{\mathrm{HDC}}(n,0,\Delta_s^{k-1},\bar{\boldsymbol{\theta}}_n) \\ &\overset{(1)}{=} R_{\mathrm{TV}}^{\mathrm{HC}}(n,0,\Delta_s^{k-1},\bar{\boldsymbol{\theta}}_n) \overset{(2)}{\geq} c(s/n)^{1/2}, \end{aligned}$$

where (1) uses the fact that for $\varepsilon = 0$ all the sets $\mathcal{M}_n^{\square}(\varepsilon,\boldsymbol{\theta}^*)$ are equal, while (2) follows from the last theorem. On the other hand, in view of Proposition 1, for $\varepsilon \geq (6/n)\log(8n/c)$ (implying that $2e^{-n\varepsilon/6} \leq (c/4)\varepsilon$),

$$\begin{aligned} R_{\mathrm{TV}}^{\mathrm{HDC}}(n,\varepsilon,\Delta_s^{k-1},\bar{\boldsymbol{\theta}}_n) &\geq R_{\mathrm{TV}}^{\mathrm{HC}}(n,\varepsilon/2,\Delta_s^{k-1},\bar{\boldsymbol{\theta}}_n) - 2e^{-n\varepsilon/6} \\ &\geq (c/4)\big\{(s/n)^{1/2}+\varepsilon\big\}. \end{aligned}$$

Combining these two inequalities, for $n \geq (10+2\log(1/c))^2$, we get

$$R_{\mathrm{TV}}^{\mathrm{HDC}}(n,\varepsilon,\Delta_s^{k-1},\bar{\boldsymbol{\theta}}_n) \geq (c/4)\big\{(s/n)^{1/2}+\varepsilon\big\}$$

for every $k \geq 1$ and every $\varepsilon \in [0,1]$. The same argument can be used to show that all the inequalities in Theorem 2 are valid for Huber's deterministic contamination model as well. Since the inclusions $\mathcal{M}_n^{\mathrm{HDC}}(\varepsilon,\boldsymbol{\theta}^*) \subset \mathcal{M}_n^{\mathrm{OC}}(\varepsilon,\boldsymbol{\theta}^*) \cap \mathcal{M}_n^{\mathrm{PC}}(\varepsilon,\boldsymbol{\theta}^*) \subset \mathcal{M}_n^{\mathrm{AC}}(\varepsilon,\boldsymbol{\theta}^*)$ hold true, we conclude that the lower bounds obtained for HC remain valid for all the other contamination models and are minimax optimal.

The main tool in the proof of Theorem 2 is the following result (Chen et al., 2018, Theorem 5.1). There is a universal constant $c_1 > 0$ such that

$$\inf_{\bar{\boldsymbol{\theta}}_n} \sup_{\mathcal{M}_n^{\mathrm{HC}}(\varepsilon,\Delta)} \boldsymbol{P}\big(d(\bar{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*) \geq w_d(\varepsilon,\Delta)\big) \geq c_1, \quad \forall \varepsilon \in [0,1),$$

where $w_d(\varepsilon,\Delta)$ is the modulus of continuity defined by $w_d(\varepsilon,\Delta) = \sup\{d(\boldsymbol{\theta},\boldsymbol{\theta}') : d_{\mathrm{TV}}(\boldsymbol{\theta},\boldsymbol{\theta}') \leq \varepsilon/(1-\varepsilon)\}$. Choosing $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ to differ on only to coordinates, one can check that, for any $\varepsilon \leq 1/2$, $w_{\mathrm{TV}}(\varepsilon,\Delta_s^{k-1}) \geq \varepsilon$, $w_{\mathrm{H}}(\varepsilon,\Delta_s^{k-1}) \geq \varepsilon^{1/2}$ and $w_{\mathbb{L}^2}(\varepsilon,\Delta_s^{k-1}) \geq \sqrt{2}\varepsilon$. Combining with the lower bounds in the non-contaminated setting, this result yields the claims of Theorem 2. In addition, (6) combined with the results of this section implies that the rate in (7) is minimax optimal.

4.3. Confidence regions

In previous sections, we established bounds for the expected value of estimation error. The aim of this section is to present bounds on estimation error of the sample mean holding with high probability. This also leads to constructing confidence regions for the parameter vector $\boldsymbol{\theta}^*$. To this end, the contamination rate ε and the sparsity s are assumed to be known. It is an interesting open

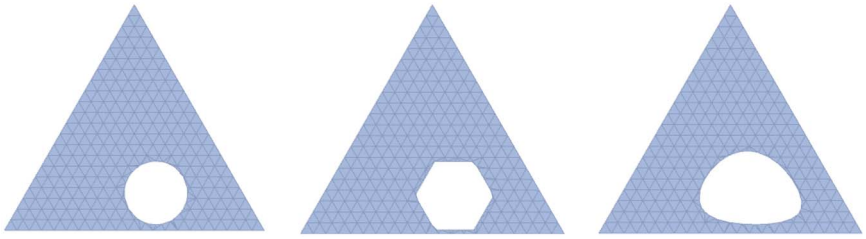

FIG 2. *The shape of confidence sets (white regions) for the distances* $\mathbb{L}^2$ *(left), TV (center), and Hellinger (right) when the sample mean is* $(1/3, 1/2, 1/6)$.

question whether one can construct optimally shrinking confidence regions for unknown ε and s.

Theorem 3 Let $\delta \in (0,1)$ be the tolerance level. If $\boldsymbol{\theta}^* \in \Delta_s^{k-1}$, then under any contamination model, the regions of Δ^{k-1} defined by each of the following inequalities

$$d_{\mathbb{L}^2}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}) \le (1/n)^{1/2} + \sqrt{2}\,\varepsilon + \big(\log(1/\delta)/n\big)^{1/2},$$
$$d_{\mathrm{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}) \le (s/n)^{1/2} + 2\,\varepsilon + \big(2\log(1/\delta)/n\big)^{1/2},$$
$$d_{\mathrm{H}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}) \le \sqrt{5}\big((s/n)\log(2s/\delta)\big)^{1/2} + \varepsilon^{1/2} + \big((1/2n)\log(2/\delta)\big)^{1/2},$$

contain $\boldsymbol{\theta}^*$ with probability at least $1-\delta$.

Proof in the appendix, page 2673

To illustrate the shapes of these confidence regions, we depicted them in Figure 2 for a three dimensional example, projected onto the plane containing the probability simplex. The sample mean in this example is equal to $(1/3, 1/2, 1/6)$.

While the estimator, $\bar{\boldsymbol{X}}_n$, used in the construction of confidence regions is adaptive both to the sparsity s and to the contamination rate ε, the confidence region itself is not adaptive. Indeed, the radius of the confidence region depends on s and/or on ε. Constructing adaptive confidence regions and adaptive tests is an important question in statistics, we refer to Cai and Low (2006); Hoffmann and Nickl (2011) for some precise results.

In the setting of shape constrained regression, Bellec (2016) proposed a method of constructing adaptive confidence regions that could be of interest for our setting as well. In particular, if there is no contamination, *i.e.*, $\varepsilon = 0$, we can define $\widehat{s} = \mathrm{Card}(j : (\bar{\boldsymbol{X}}_n)_j \neq 0)$, the observed sparsity. It is clear that $\widehat{s} \le s$. The analog of Bellec's method in our setting would consist in replacing s by $\widehat{s}$ in the radius of the ball given by Section 4.3. Of course, this does not inflate the ball. However, we did not manage to adapt the argument in Bellec (2016, Theorem 4.1) for proving that the modified region (with $\widehat{s}$ instead of s) has still the required coverage property and, therefore, is adaptive to s. The

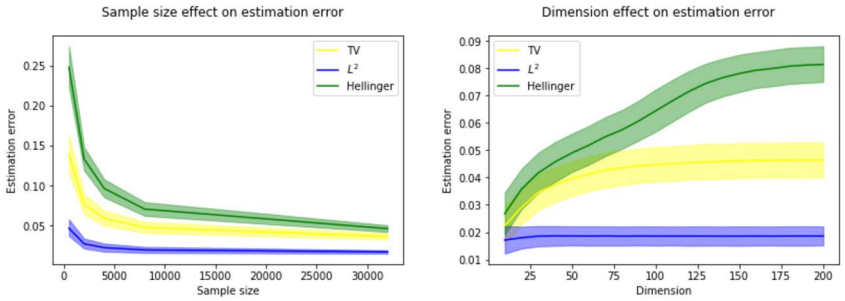

FIG 3. *Estimation error of $\bar{\boldsymbol{X}}_n$ measured by total variation, Hellinger, and $\mathbb{L}^2$ distances as a function of (left panel) number of observations with contamination rate 0.2 and dimension 10^2 and (right panel) dimension with contamination rate 0.2 and 10^4 samples. The interval between 5th and 95th quantiles of the error, obtained from 10^4 repetitions, is also depicted for every graph.*

main difficulty, at a very high-level, comes from the fact that the distance we consider, d_{TV}, is not induced by an inner product. Thus, constructing adaptive confidence regions even when there is no contamination is an open question.

5. Illustration on a numerical example

We provide some numerical experiments which illustrate theoretical results of Section 4. The data set is the collection of 38 books written by Alexandre Dumas (1802–1870) and 38 books written by Emile Zola (1840–1902).[7] To each author, we assign a parameter vector corresponding to the distribution of the number of words contained in the sentences used in the author's books. To be more clear, a sentence containing l words is represented by vector $\boldsymbol{e}_l$, and if the parameter vector of an author is $(\theta_1, \ldots, \theta_k)$, it means that a sentence used by the author is of size $l \in \{1, \ldots, k\}$ with probability θ_l. We carried out synthetic experiments in which the reference parameter to estimate is the probability vector of Dumas, while the distribution of outliers is determined by the probability vector of Zola. Ground truths for these parameters are computed from the aforementioned large corpus of their works. Only the dense case $s = k$ were considered. For various values of ε and n, a contaminated sample was generated by randomly choosing n sentences either from Dumas' works (with probability $1 - \varepsilon$) or from Zola's works (with probability ε). The sample mean was computed for this corrupted sample, and the error with respect to Dumas' parameter vector was measured by the three distances TV, $\mathbb{L}^2$ and Hellinger. This experiment was repeated 10^4 times for each special setting to obtain information on error's distribution. Furthermore, by grouping nearby outcomes we created samples of different dimensions for illustrating the behavior of the error as a function of k.

[7]The works of both authors are available from https://www.gutenberg.org/

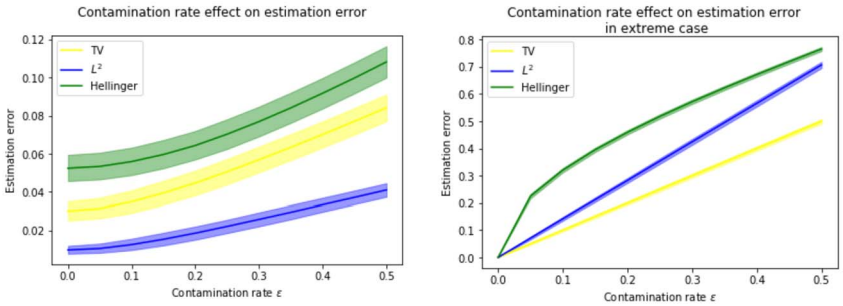

FIG 4. *The estimation error of $\bar{\boldsymbol{X}}_n$ measured by total variation, Hellinger, and $\mathbb{L}^2$ distances in terms of the contamination rate (with dimension 10^2 and 10^4 samples). At right, we plotted the error with respect to the contamination rate for an extreme case, where the reference and contamination distributions have the largest distance. The interval between 5th and 95th quantiles of the error, obtained from 10^4 trials, is also depicted.*

The error of $\bar{X}_n$ as a function of the sample size n, dimension k, and contamination rate ε is plotted in Figures 3 and 4. These plots are conform to the theoretical results. Indeed, the first plot in Figure 3 shows that the errors for the three distances is decreasing w.r.t. n. Furthermore, we see that up to some level of n this decay is of order $n^{-1/2}$. The second plot in Figure 3 confirms that the risk grows linearly in k for the TV and Hellinger distances, while it is constant for the $\mathbb{L}^2$ error.

Left panel of Figure 4 suggests that the error grows linearly in terms of contamination rate. This is conform to our results for the TV and $\mathbb{L}^2$ errors. But it might seem that there is a disagreement with the result for the Hellinger distance, for which the risk is shown to increase at the rate $\varepsilon^{1/2}$ and not ε. This is explained by the fact that the rate $\varepsilon^{1/2}$ corresponds to the worst-case risk, whereas here, the setting under experiment does not necessarily represent the worst case. When the parameter vectors of the reference and contamination distributions, respectively, are $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ with $i \neq j$ (*i.e.*, when these two distributions are at the largest possible distance, which we call an extreme case), the graph of the error as a function of ε (right panel of Figure 4) is similar to that of square-root function.

6. Summary and conclusion

We have analyzed the problem of robust estimation of the mean of a random vector belonging to the probability simplex. Four measures of accuracy have been considered: total variation, Hellinger, Euclidean and Wasserstein distances. In each case, we have established the minimax rates of the expected error of estimation under the sparsity assumption. In addition, confidence regions shrinking at the minimax rate have been proposed.

An intriguing observation is that the choice of the distance has much stronger impact on the rate than the nature of contamination. Indeed, while the rates

for the aforementioned distances are all different, the rate corresponding to one particular distance is not sensitive to the nature of outliers (ranging from Huber's contamination to the adversarial one). While the rate obtained for the TV-distance coincides with the previously known rate of robustly estimating a Gaussian mean, the rates we have established for the Hellinger and for the Euclidean distances appear to be new. Interestingly, when the error is measured by the Euclidean distance, the quality of estimation does not get deteriorated with increasing dimension.

Appendix A: Proofs of propositions

Proof of Proposition 1 on page 2657. Recall that $\widehat{O}$ is the set of outliers in the Huber model. Let O be any subset of $\{1,\dots,n\}$. It follows from the definition of Huber's model that if $\boldsymbol{P}_n^O$ stands for the conditional distribution of $(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)$ given $\widehat{O}=O$, when $(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)$ is drawn from $\boldsymbol{P}^n\in\mathcal{M}_n^{\mathrm{HC}}(\varepsilon,\boldsymbol{\theta}^*)$, then $\boldsymbol{P}_n^O\in\mathcal{M}_n^{\mathrm{HDC}}(|O|/n,\boldsymbol{\theta}^*)$. Therefore, for every O of cardinality $o\geq 2\varepsilon n$, we have

$$\begin{aligned}\sup_{\mathcal{M}_n^{\mathrm{HC}}(\varepsilon,\boldsymbol{\theta}^*)}\mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)\mathbb{1}(\widehat{O}=O)] &= \sup_{\mathcal{M}_n^{\mathrm{HC}}(\varepsilon,\boldsymbol{\theta}^*)}\mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)|\widehat{O}=O]\mathbf{P}(\widehat{O}=O)\\ &\leq \sup_{\mathcal{M}_n^{\mathrm{HDC}}(o/n,\boldsymbol{\theta}^*)}\mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)]\mathbf{P}(\widehat{O}=O)\\ &\overset{(1)}{\leq} \sup_{\mathcal{M}_n^{\mathrm{HDC}}(1,\boldsymbol{\theta}^*)}\mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)]\mathbf{P}(\widehat{O}=O). \qquad (8)\end{aligned}$$

Inequality (1) above is a direct consequence of the inclusion $\mathcal{M}_n^{\mathrm{HDC}}(o/n,\boldsymbol{\theta}^*)\subset\mathcal{M}_n^{\mathrm{HDC}}(1,\boldsymbol{\theta}^*)$. Summing the obtained inequality over all sets O of cardinality $\geq 2\varepsilon n$, we get

$$\sup_{\mathcal{M}_n^{\mathrm{HC}}(\varepsilon,\boldsymbol{\theta}^*)}\mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)\mathbb{1}(|\widehat{O}|\geq 2\varepsilon n)]\leq \sup_{\mathcal{M}_n^{\mathrm{HDC}}(1,\boldsymbol{\theta}^*)}\mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)]\mathbf{P}(|\widehat{O}|\geq 2\varepsilon n).$$

It follows from the multiplicative form of Chernoff's inequality that $\mathbf{P}(|\widehat{O}|\geq 2\varepsilon n)\leq e^{-n\varepsilon/3}$. This leads to the last term in inequality (3).

Using the same argument as for (8), for any O of cardinality $o<2n\varepsilon$, we get

$$\begin{aligned}\sup_{\mathcal{M}_n^{\mathrm{HC}}(\varepsilon,\boldsymbol{\theta}^*)}\mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)\mathbb{1}(|\widehat{O}|<2n\varepsilon)] &\leq \sup_{\mathcal{M}_n^{\mathrm{HDC}}(2\varepsilon,\boldsymbol{\theta}^*)}\mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)]\sum_{|O|\leq 2n\varepsilon}\mathbf{P}(\widehat{O}=O)\\ &= \sup_{\mathcal{M}_n^{\mathrm{HDC}}(2\varepsilon,\boldsymbol{\theta}^*)}\mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)].\end{aligned}$$

This completes the proof of (3).

One can use the same arguments along with the Tchebychev inequality to establish (4). Indeed, for every S of cardinality $o\leq 2\varepsilon n$, we have

$$\sup_{\mathcal{M}_n^{\mathrm{HC}}(\varepsilon,\boldsymbol{\theta}^*)} r\mathbf{P}\big(d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)>r \text{ and } \widehat{O}=O\big)$$

$$
\begin{aligned}
&= \sup_{\mathcal{M}_n^{\mathrm{HC}}(\varepsilon,\boldsymbol{\theta}^*)} r\mathbf{P}\big(d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*) > r \,|\, \widehat{O} = O\big)\mathbf{P}(\widehat{O} = O)\\
&\le \sup_{\mathcal{M}_n^{\mathrm{HDC}}(o/n,\boldsymbol{\theta}^*)} r\mathbf{P}\big(d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*) > r\big)\mathbf{P}(\widehat{O} = O)\\
&\le \sup_{\mathcal{M}_n^{\mathrm{HDC}}(2\varepsilon,\boldsymbol{\theta}^*)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)]\mathbf{P}(\widehat{O} = O).
\end{aligned}
$$

Summing the obtained inequality over all sets O of cardinality $o \le 2\varepsilon n$, we get

$$
\begin{aligned}
\sup_{\mathcal{M}_n^{\mathrm{HC}}(\varepsilon,\boldsymbol{\theta}^*)} r\mathbf{P}\big(d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*) > r \text{ and } |\widehat{O}| \le 2\varepsilon n\big) &\le \sup_{\mathcal{M}_n^{\mathrm{HDC}}(2\varepsilon,\boldsymbol{\theta}^*)} \mathbf{E}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*)]\\
&= R_d^{\mathrm{HDC}}(n, 2\varepsilon, \boldsymbol{\theta}^*, \widehat{\boldsymbol{\theta}}).
\end{aligned}
$$

On the other hand, it holds that

$$
\sup_{\mathcal{M}_n^{\mathrm{HC}}(\varepsilon,\boldsymbol{\theta}^*)} r\mathbf{P}\big(d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}^*) > r \text{ and } |\widehat{O}| > 2\varepsilon n\big) \le r\mathbf{P}\big(|\widehat{O}| > 2\varepsilon n\big) \le re^{-n\varepsilon/3},
$$

and the claim of the proposition follows. □

Proof of Proposition 2 on page 2659. Let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be two points in Θ. We have

$$
\begin{aligned}
2R_d^{\mathrm{HC}}(n,\varepsilon,\Theta,\widehat{\boldsymbol{\theta}}_n) &\ge R_d^{\mathrm{HC}}(n,\varepsilon,\boldsymbol{\theta}_1,\widehat{\boldsymbol{\theta}}_n) + R_d^{\mathrm{HC}}(n,\varepsilon,\boldsymbol{\theta}_2,\widehat{\boldsymbol{\theta}}_n)\\
&\ge \mathbf{E}_{(1-\varepsilon)\boldsymbol{P}_{\boldsymbol{\theta}_1}+\varepsilon\boldsymbol{P}_{\boldsymbol{\theta}_2}}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}_1)] + \mathbf{E}_{(1-\varepsilon)\boldsymbol{P}_{\boldsymbol{\theta}_1}+\varepsilon\boldsymbol{P}_{\boldsymbol{\theta}_2}}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}_2)].
\end{aligned}
$$

To ease writing, assume that n is an even number. Let O be any fixed set of cardinality $n/2$. It is clear that the set of outliers $\widehat{O}$ satisfies

$$
p_O = \boldsymbol{P}(\widehat{O} = O) = \boldsymbol{P}(\widehat{O} = O^c) > 0.
$$

Furthermore, if $\boldsymbol{X}_{1:n}$ is drawn from $((1-\varepsilon)\boldsymbol{P}_{\boldsymbol{\theta}_1} + \varepsilon\boldsymbol{P}_{\boldsymbol{\theta}_2})^{\otimes n} := \boldsymbol{P}_\varepsilon^{\otimes n}$, then its conditional distribution given $\widehat{O} = O$ is exactly the same as the conditional distribution of $\boldsymbol{X}_{1:n} \sim \boldsymbol{P}_\varepsilon^{\otimes n}$ given $\widehat{O} = O^c$. This implies that

$$
\begin{aligned}
&2R_d^{\mathrm{HC}}(n,\varepsilon,\Theta,\widehat{\boldsymbol{\theta}}_n)\\
&\quad\ge p_O\big(\mathbf{E}_{\boldsymbol{P}_\varepsilon}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}_1)|\widehat{O} = O] + \mathbf{E}_{\boldsymbol{P}_\varepsilon}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}_2)|\widehat{O} = O^c]\big)\\
&\quad= p_O\mathbf{E}_{\boldsymbol{P}_\varepsilon}[d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}_1) + d(\widehat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}_2)|\widehat{O} = S] \ge p_O d(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2),
\end{aligned}
$$

where in the last step we have used the triangle inequality. The obtained inequality being true for every $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, we can take the supremum to get

$$
R_d^{\mathrm{HC}}(n,\varepsilon,\Theta,\widehat{\boldsymbol{\theta}}_n) \ge (p_O/2) \sup_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\Theta} d(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = +\infty.
$$

This completes the proof. □

Appendix B: Minimax upper bounds over the sparse simplex

This section is devoted to the proof of the upper bounds on minimax risks in the discrete model with respect to various distances.

Proof of Theorem 1 on page 2661. To ease notation, we set

$$\bar{\boldsymbol{X}}_n = \frac{1}{n}\sum_i \boldsymbol{X}_i, \quad \bar{\boldsymbol{Y}}_n = \frac{1}{n}\sum_i \boldsymbol{Y}_i, \quad \bar{\boldsymbol{X}}_O = \frac{1}{o}\sum_{i\in O} \boldsymbol{X}_i, \quad \bar{\boldsymbol{Y}}_O = \frac{1}{o}\sum_{i\in O} \boldsymbol{Y}_i$$

In the adversarial model, we have $\boldsymbol{X}_i = \boldsymbol{Y}_i$ if $i \notin O$ where $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ are generated from the reference distribution $\boldsymbol{\theta}^*$.

$$\begin{aligned} d_{\mathbb{L}^2}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*) &= \big\|\bar{\boldsymbol{X}}_n - \boldsymbol{\theta}^*\big\|_2 = \Big\|\bar{\boldsymbol{Y}}_n - \boldsymbol{\theta}^* + \frac{1}{n}\sum_{i\in O}(\boldsymbol{X}_i - \boldsymbol{Y}_i)\Big\|_2 \\ &\le \big\|\bar{\boldsymbol{Y}}_n - \boldsymbol{\theta}^*\big\|_2 + \frac{|O|}{n}\sup_{\boldsymbol{x},\boldsymbol{y}\in\Delta^{k-1}} \|\boldsymbol{x}-\boldsymbol{y}\|_2 \\ &= \big\|\bar{\boldsymbol{Y}}_n - \boldsymbol{\theta}^*\big\|_2 + \sqrt{2}\varepsilon, \end{aligned}$$

which gives us

$$\sup_{\mathcal{M}_n^{\mathrm{AC}}(\varepsilon,\boldsymbol{\theta}^*)} \mathbf{E}[d_{\mathbb{L}^2}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*)] \le \sup_{\boldsymbol{\theta}^*} \mathbf{E}[d_{\mathbb{L}^2}(\bar{\boldsymbol{Y}}_n, \boldsymbol{\theta}^*)] + \sqrt{2}\varepsilon.$$

And for a fixed $\boldsymbol{\theta}^*$ it is well known that

$$\begin{aligned} \mathbf{E}[d^2_{\mathbb{L}^2}(\bar{\boldsymbol{Y}}_n, \boldsymbol{\theta}^*)] &= \sum_{j=1}^k \mathbf{Var}[\bar{\boldsymbol{Y}}_{n,j}] = \sum_{j=1}^k \mathbf{Var}\left[\frac{1}{n}\sum_{i=1}^n \mathbb{1}(Y_i = \boldsymbol{e}_j)\right] \\ &= \frac{1}{n}\sum_{j=1}^k \mathbf{Var}[\mathbb{1}(Y_1 = \boldsymbol{e}_j)] \le \frac{1}{n}\sum_{j=1}^k \mathbf{E}[\mathbb{1}(Y_1 = \boldsymbol{e}_j)] = \frac{1}{n}. \end{aligned}$$

Hence, we obtain $R^{\mathrm{AC}}_{\mathbb{L}^2}(n, \varepsilon, \Delta^{k-1}) \le (1/n)^{1/2} + \varepsilon$. Similarly,

$$\begin{aligned} d_{\mathrm{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*) &\le \big\|\bar{\boldsymbol{Y}}_n - \boldsymbol{\theta}^*\big\|_1 + \frac{|O|}{n}\sup_{\boldsymbol{x},\boldsymbol{y}\in\Delta^{k-1}} \|\boldsymbol{x}-\boldsymbol{y}\|_1 \\ &= \big\|\bar{\boldsymbol{Y}}_n - \boldsymbol{\theta}^*\big\|_1 + 2\varepsilon. \end{aligned}$$

This gives

$$\sup_{\mathcal{M}_n^{\mathrm{AC}}(\varepsilon,\boldsymbol{\theta}^*)} \mathbf{E}[d_{\mathrm{TV}}(\bar{\boldsymbol{X}}_n, \boldsymbol{\theta}^*)] \le \sup_{\boldsymbol{\theta}^*} \mathbf{E}[d_{\mathrm{TV}}(\bar{\boldsymbol{Y}}_n, \boldsymbol{\theta}^*)] + 2\varepsilon.$$

In addition, for every $\boldsymbol{\theta}^*$,

$$\mathbf{E}[d_{\mathrm{TV}}(\bar{\boldsymbol{Y}}_n, \boldsymbol{\theta}^*)] = \frac{1}{2}\sum_{j=1}^k \mathbf{E}\big[\big|\bar{\boldsymbol{Y}}_{n,j} - \boldsymbol{\theta}^*_j\big|\big]$$

$$
\begin{aligned}
&\le \frac{1}{2}\sum_{j=1}^{k}\Big(\mathbf{E}\big[\big|\bar{\boldsymbol{Y}}_{n,j}-\boldsymbol{\theta}^*_j\big|^2\big]\Big)^{1/2}\\
&= \frac{1}{2}\sum_{j=1}^{k}\Big(\mathbf{Var}\big[\bar{\boldsymbol{Y}}_{n,j}\big]\Big)^{1/2}\\
&\overset{(1)}{=} \frac{1}{2}\sum_{j\in J}\big(\frac{1}{n}\boldsymbol{\theta}^*_j(1-\boldsymbol{\theta}^*_j)\big)^{1/2}\\
&\overset{(2)}{\le} \frac{1}{2}\,s^{1/2}\Big(\sum_{j=1}^{k}\frac{1}{n}\boldsymbol{\theta}^*_j(1-\boldsymbol{\theta}^*_j)\Big)^{1/2} \le \frac{1}{2}\big(s/n\big)^{1/2},
\end{aligned}
$$

where in (1) we have used the notation $J=\{j:\theta^*_j\neq 0\}$ and in (2) we have used the Cauchy-Schwarz inequality. This leads to

$$R^{\mathrm{AC}}_{\mathrm{TV}}(n,\varepsilon,\Delta^{k-1}) \le (k/n)^{1/2}+\varepsilon.$$

Finally, for the Hellinger distance

$$d_{\mathrm{H}}(\bar{\boldsymbol{X}}_n,\boldsymbol{\theta}^*) \le d_{\mathrm{H}}(\bar{\boldsymbol{X}}_n,\bar{\boldsymbol{Y}}_n)+d_{\mathrm{H}}(\bar{\boldsymbol{Y}}_n,\boldsymbol{\theta}^*) \le d_{\mathrm{TV}}(\bar{\boldsymbol{X}}_n,\bar{\boldsymbol{Y}}_n)^{1/2}+d_{\mathrm{H}}(\bar{\boldsymbol{Y}}_n,\boldsymbol{\theta}^*),$$

where we have already seen that

$$d_{\mathrm{TV}}(\bar{\boldsymbol{X}}_n,\bar{\boldsymbol{Y}}_n)=\frac{o}{n}\big\|\bar{\boldsymbol{X}}_O-\bar{\boldsymbol{Y}}_O\big\|_1 \le 2\varepsilon.$$

This yields

$$\sup_{\mathcal{M}_n^{\mathrm{AC}}(\varepsilon,\boldsymbol{\theta}^*)}\mathbf{E}[d_{\mathrm{H}}(\bar{\boldsymbol{X}}_n,\boldsymbol{\theta}^*)] \le \sup_{\boldsymbol{\theta}^*}\mathbf{E}[d_{\mathrm{H}}(\bar{\boldsymbol{Y}}_n,\boldsymbol{\theta}^*)]+\sqrt{2\varepsilon}.$$

Note that $\mathbf{E}[\bar{\boldsymbol{Y}}_{n,j}]=\theta^*_j$ implies that $\bar{\boldsymbol{Y}}_{n,j}=\theta^*_j=0$ for every j that does not belong to the sparsity pattern J. Furthermore, for every $j\in J$, we have

$$\big|\sqrt{\bar{\boldsymbol{Y}}_{n,j}}-\sqrt{\boldsymbol{\theta}^*_j}\big| = \frac{|\bar{\boldsymbol{Y}}_{n,j}-\boldsymbol{\theta}^*_j|}{\sqrt{\bar{\boldsymbol{Y}}_{n,j}}+\sqrt{\boldsymbol{\theta}^*_j}} \le \frac{|\bar{\boldsymbol{Y}}_{n,j}-\boldsymbol{\theta}^*_j|}{\sqrt{\boldsymbol{\theta}^*_j}}.$$

This implies that for every $\boldsymbol{\theta}^*\in\Delta^{k-1}_s$,

$$
\begin{aligned}
\mathbf{E}[d^2_{\mathrm{H}}(\bar{\boldsymbol{Y}}_n,\boldsymbol{\theta}^*)] &= \mathbf{E}\Big[\frac{1}{2}\sum_{j=1}^{k}\Big(\sqrt{\bar{\boldsymbol{Y}}_{n,j}}-\sqrt{\boldsymbol{\theta}^*_j}\Big)^2\Big] \le \frac{1}{2}\mathbf{E}\Big[\sum_{j\in J}\frac{(\bar{\boldsymbol{Y}}_{n,j}-\boldsymbol{\theta}^*_j)^2}{\boldsymbol{\theta}^*_j}\Big]\\
&= \frac{1}{2}\sum_{j\in J}\frac{1}{\boldsymbol{\theta}^*_j}\mathbf{Var}\big[\bar{\boldsymbol{Y}}_{n,j}\big] = \frac{1}{2}\sum_{j\in J}\frac{1}{\boldsymbol{\theta}^*_j}\times\frac{\boldsymbol{\theta}^*_j(1-\boldsymbol{\theta}^*_j)}{n} = \frac{s-1}{2n}.
\end{aligned}
$$

Hence, by Jensen's inequality $\mathbf{E}[d_{\mathrm{H}}(\bar{\boldsymbol{Y}}_n,\boldsymbol{\theta}^*)]<\sqrt{s/n}$. Therefore, we infer that

$$R^{\mathrm{AC}}_{\mathrm{H}}(n,\varepsilon,\Delta^{k-1}_s) \le (s/n)^{1/2}+\sqrt{2}\varepsilon^{1/2}$$

and the last claim of the theorem follows. □

Appendix C: Minimax lower bounds over the sparse simplex

This section is devoted to the proof of the lower bounds on minimax risks in the discrete model with respect to various distances. Note that the rates over the high-dimensional "sparse" simplex Δ_s^{k-1} coincide with those for the dense simplex Δ^{s-1}. For this reason, all the lower bounds will be proved for Δ^{s-1} only (for $s \geq 2$ an even integer). In addition, we will restrict our attention to the distributions $\boldsymbol{P}$, $\boldsymbol{Q}$ over Δ^{s-1} that are supported by the set $\mathcal{A}$ of the elements of the canonical basis that is $\boldsymbol{P}(\mathcal{A}) = \boldsymbol{Q}(\mathcal{A}) = 1$.

Proof of Theorem 2 on page 2662. We denote by $\boldsymbol{e}_j$ the vector in $\mathbb{R}^s$ having all the coordinates equal to zero except the jth coordinate which is equal to one. Setting

$$\boldsymbol{\theta} = \boldsymbol{e}_1, \quad \text{and} \quad \boldsymbol{\theta}' = \Big(1 - \frac{\varepsilon}{1-\varepsilon}\Big)\boldsymbol{e}_1 + \frac{\varepsilon}{1-\varepsilon}\boldsymbol{e}_2$$

we have

$$d_{\mathrm{TV}}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{\varepsilon}{1-\varepsilon}, \quad d_{\mathbb{L}^2}(\boldsymbol{\theta}, \boldsymbol{\theta}') \geq \sqrt{2}\varepsilon, \quad \text{and} \quad d_{\mathrm{H}}(\boldsymbol{\theta}, \boldsymbol{\theta}') \geq \varepsilon^{1/2}.$$

Therefore, modulus of continuity defined by

$$w_d(\varepsilon, \Delta) = \sup\big\{d(\boldsymbol{\theta}, \boldsymbol{\theta}') : \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Delta,\ d_{\mathrm{TV}}(\boldsymbol{\theta}, \boldsymbol{\theta}') \leq \varepsilon/(1-\varepsilon)\big\}$$

for a distance d and a set Δ, satisfies for any $\varepsilon \leq 1/2$,

$$w_{\mathrm{TV}}(\varepsilon, \Delta^{k-1}) \geq \varepsilon, \quad w_{\mathbb{L}^2}(\varepsilon, \Delta^{k-1}) \geq \sqrt{2}\varepsilon, \quad \text{and} \quad w_{\mathrm{H}}(\varepsilon, \Delta^{k-1}) \geq \varepsilon^{1/2}.$$

These bounds on the modulus of continuity are the first ingredient we need for lower bounding the minimax risk using (Chen et al., 2018, Theorem 5.1).

The second ingredient is the minimax rate in the non-contaminated case. are well known. For each of the considered distances, this rate is well-known. However, for the sake of completeness, we provide below a proof those lower bounds using Fano's method. For this, we use the Varshamov-Gilbert lemma (see e.g. Lemma 2.9 in Tsybakov, 2009) and Theorem 2.5 in Tsybakov (2009). The Varshamov-Gilbert lemma guarantees the existence of a set $\boldsymbol{\omega}^{(1)}, \ldots, \boldsymbol{\omega}^{(M)} \in \{0,1\}^{\lfloor s/2 \rfloor}$ of cardinality $M \geq 2^{s/16}$ such that

$$\rho(\boldsymbol{\omega}^{(i)}, \boldsymbol{\omega}^{(j)}) \geq \frac{s}{16}, \quad \text{for all} \quad i \neq j,$$

where $\rho(.,.)$ stands for the Hamming distance. Using these binary vectors $\{\boldsymbol{\omega}_j\}$, a parameter $\beta \in [0, 1/s]$ to be specified later and the "baseline" vector $\boldsymbol{\theta}^{(0)} = (1/s, \ldots, 1/s)$, we define hypotheses $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}$ by the relations

$$\boldsymbol{\theta}^{(i)}_{2j-1} = \boldsymbol{\theta}^{(0)}_{2j-1} + \boldsymbol{\omega}^{(i)}_j \beta \quad \text{and} \quad \boldsymbol{\theta}^{(i)}_{2j} = \boldsymbol{\theta}^{(0)}_{2j} - \boldsymbol{\omega}^{(i)}_j \beta \quad \forall j \in \{0, \ldots, \lfloor s/2 \rfloor\}.$$

Remark that $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(M)}$ are all probability vectors of dimension s. Denoting the Kullback-Leibler divergence by $d_{\mathrm{KL}}(.,.)$, one can check the conditions of Theorem 2.5 in Tsybakov (2009):

$$d_{\mathbb{L}^2}(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(j)}) \geq \beta \frac{\sqrt{2s}}{4} \qquad \forall j \neq i,$$
$$d_{\mathrm{TV}}(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(j)}) \geq \frac{\beta s}{16} \qquad \forall j \neq i,$$

as well as

$$\begin{aligned} d_{\mathrm{KL}}(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^{(0)}) &\leq \sum_{j=1}^{\lfloor s/2 \rfloor} (\boldsymbol{\theta}^{(0)}_{2j-1} + \beta) \log \frac{\boldsymbol{\theta}^{(0)}_{2j-1} + \beta}{\boldsymbol{\theta}^{(0)}_{2j-1}} + (\boldsymbol{\theta}^{(0)}_{2j} - \beta) \log \frac{\boldsymbol{\theta}^{(0)}_{2j} - \beta}{\boldsymbol{\theta}^{(0)}_{2j}} \\ &\leq \sum_{j=1}^{\lfloor k/2 \rfloor} \frac{\beta}{\boldsymbol{\theta}^{(0)}_{2j-1}} (\boldsymbol{\theta}^{(0)}_{2j-1} + \beta) - \frac{\beta}{\boldsymbol{\theta}^{(0)}_{2j}} (\boldsymbol{\theta}^{(0)}_{2j} - \beta) \\ &= \beta^2 \sum_{j=1}^{k} \frac{1}{\boldsymbol{\theta}^{(0)}_j} = \beta^2 \sum_{j=1}^{s} s = \beta^2 s^2 \leq \frac{\alpha \log M}{n} \qquad \forall i \in \{1, \ldots, M\}, \end{aligned}$$

for $\beta = \sqrt{\alpha/(ns)}/4$, taking into account the fact that $2^{s/16} \leq M$. Now by applying the aforementioned theorem, we obtain for the non-contaminated setting ($\varepsilon = 0$)

$$\inf_{\bar{\boldsymbol{\theta}}_n} \sup_{\mathcal{M}_n^{\mathrm{HC}}(0, \Delta^{s-1})} \boldsymbol{P}\left(d_{\mathbb{L}^2}(\bar{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^*) \geq \frac{\beta\sqrt{2s}}{2} \right) \geq \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}} \right),$$
$$\inf_{\bar{\boldsymbol{\theta}}_n} \sup_{\mathcal{M}_n^{\mathrm{HC}}(0, \Delta^{s-1})} \boldsymbol{P}\left(d_{\mathrm{TV}}(\bar{\boldsymbol{\theta}}_n, \boldsymbol{\theta}^*) \geq \frac{\beta s}{8} \right) \geq \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}} \right).$$

Setting $M = 2^{s/16}$ and $\alpha = 1/32$, by Markov's inequality, one concludes

$$\inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathrm{HC}}_{\mathbb{L}^2}(n, 0, \Delta^{s-1}, \bar{\boldsymbol{\theta}}_n) \geq c(1/n)^{1/2},$$
$$\inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathrm{HC}}_{\mathrm{TV}}(n, 0, \Delta^{s-1}, \bar{\boldsymbol{\theta}}_n) \geq c(s/n)^{1/2},$$

where $c = 1/25600$. Since, $\sqrt{2}\, d_{\mathrm{H}}(\boldsymbol{\theta}, \boldsymbol{\theta}') \geq d_{\mathrm{TV}}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ for any pair of points $(\boldsymbol{\theta}, \boldsymbol{\theta}')$ on the simplex, we have

$$\inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathrm{HC}}_{\mathrm{H}}(n, 0, \Delta^{s-1}, \bar{\boldsymbol{\theta}}_n) \geq c(s/n)^{1/2}.$$

Finally, we apply (Chen et al., 2018, Theorem 5.1) stating in our case for any distance d

$$\inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathrm{HC}}_d(n, \varepsilon, \Delta^{s-1}, \bar{\boldsymbol{\theta}}_n) \geq c\big\{ \inf_{\bar{\boldsymbol{\theta}}_n} R^{\mathrm{HC}}_d(n, 0, \Delta^{s-1}, \bar{\boldsymbol{\theta}}_n) + w_d(\varepsilon, \Delta^{s-1}) \big\},$$

for an universal constant c, which completes the proof of Theorem 2. □

Appendix D: Proofs of bounds with high probability

Proof of Theorem 3 on page 2664. Suppose $\boldsymbol{X}_i = \boldsymbol{Y}_i$ if $i \notin O$ where $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ are independently generated from the reference distribution $\boldsymbol{P}$ so that $\mathbf{E}[\boldsymbol{Y}_i] = \boldsymbol{\theta}^*$. For any $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n$, let $\boldsymbol{\Phi}_\square(\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n) := d_\square(\sum_{i=1}^n \boldsymbol{Z}_i/n, \boldsymbol{\theta}^*)$, where $\square$ refers here to the distances $\mathbb{L}^2$ or TV. Given $\boldsymbol{Y}'_1, \dots, \boldsymbol{Y}'_n \in \Delta^{k-1}$ we have for every i

$$\boldsymbol{\Phi}_\square(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_i, \dots, \boldsymbol{Y}_n) - \boldsymbol{\Phi}_\square(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_{i-1}, \boldsymbol{Y}'_i, \boldsymbol{Y}_{i+1}, \dots, \boldsymbol{Y}_n) \le \frac{1}{n} d_\square(\boldsymbol{Y}_i, \boldsymbol{Y}'_i).$$

Furthermore, it can easily be shown that the last term is bounded by $\sqrt{2}/n$ and $2/n$ for the distances $\mathbb{L}^2$ and TV, respectively. By bounded difference inequality (see for example Theorem 6.2 of Boucheron et al., 2013) with probability at least $1 - \delta$

$$\begin{aligned}
\boldsymbol{\Phi}_{\mathbb{L}^2}(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n) &\le \mathbf{E}\boldsymbol{\Phi}_{\mathbb{L}^2}(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n) + \big(\log(1/\delta)/n\big)^{1/2} \\
&\le (1/n)^{1/2} + \big(\log(1/\delta)/n\big)^{1/2}, \\
\boldsymbol{\Phi}_{TV}(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n) &\le \mathbf{E}\boldsymbol{\Phi}_{TV}(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n) + \big(\log(2/\delta)/n\big)^{1/2} \\
&\le (s/n)^{1/2} + \big(\log(2/\delta)/n\big)^{1/2}.
\end{aligned}$$

Using $\boldsymbol{\Phi}_\square(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) \le \boldsymbol{\Phi}_\square(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n) + d_\square(\sum_{i \in O} \boldsymbol{X}_i/n, \sum_{i \in O} \boldsymbol{Y}_i/n)$, one can conclude the proof of the first two claims of the theorem.

For the Hellinger distance, the computations are more tedious. We have to separate the case of small θ^*_j. To this end, let $J = \{j : 0 < \theta^*_j < (1/n)\log(2s/\delta)\}$ and $J' = \{j : \theta^*_j \ge (1/n)\log(2s/\delta)\}$. We have

$$\begin{aligned}
\sum_{j \in J} \Big(\sqrt{\bar{\boldsymbol{Y}}_{n,j}} - \sqrt{\theta^*_j}\Big)^2 &\le \sum_{j \in J} (\bar{\boldsymbol{Y}}_{n,j} + \theta^*_j) \\
&\le \frac{1}{n} \sum_{i=1}^n \Big(\sum_{j \in J} \boldsymbol{Y}_{i,j} - \theta^*_j \Big) + \frac{2s \log(2s/\delta)}{n}.
\end{aligned}$$

Since the random variables $U_i := \Big(\sum_{j \in J} \boldsymbol{Y}_{i,j} \Big)$ are iid, positive, bounded by 1, the Bernstein inequality implies that

$$\frac{1}{n} \sum_{i=1}^n \big(U_i - \mathbf{E}[U_1]\big) \le \sqrt{\frac{2\mathbf{Var}(U_1)\log(2/\delta)}{n}} + \frac{\log(2/\delta)}{3n},$$

holds with probability at least $1 - \delta/2$ for $0 < \delta < 1$. One easily checks that $\sqrt{\mathbf{Var}(U_1)} \le \sum_{j \in J} \sqrt{\mathbf{Var}(\boldsymbol{Y}_{1,j})} \le s\sqrt{\log(2s/\delta)/n}$. Therefore, with probability at least $1 - \delta/2$, we have

$$\frac{1}{n} \sum_{i=1}^n \Big(\sum_{j \in J} \boldsymbol{Y}_{i,j} - \theta^*_j \Big) \le \frac{\sqrt{2}\, s \log(2s/\delta)}{n} + \frac{\log(2/\delta)}{3n}.$$

This yields

$$\sum_{j\in J}\left(\sqrt{\bar{\boldsymbol{Y}}_{n,j}}-\sqrt{\theta_j^*}\right)^2\le\frac{3.5s\log(2s/\delta)+\log(2/\delta)}{n},\tag{9}$$

with probability at least $1-\delta/2$.

On the other hand, we have

$$\sum_{j\in J'}\left(\sqrt{\bar{\boldsymbol{Y}}_{n,j}}-\sqrt{\theta_j^*}\right)^2\le\sum_{j\in J'}\frac{(\bar{\boldsymbol{Y}}_{n,j}-\theta_j^*)^2}{\theta_j^*}\le s\max_{j\in J'}\frac{(\bar{\boldsymbol{Y}}_{n,j}-\theta_j^*)^2}{\theta_j^*}.\tag{10}$$

The Bernstein inequality and the union bound imply that, with probability at least $1-\delta/2$, for all $j\in J'$,

$$\begin{aligned}|\bar{\boldsymbol{Y}}_{n,j}-\theta_j^*|&\le\sqrt{\frac{2\mathbf{Var}(\boldsymbol{Y}_{1,j})\log(2s/\delta)}{n}}+\frac{\log(2s/\delta)}{n}\\&\le\sqrt{\frac{2\theta_j^*\log(2s/\delta)}{n}}+\frac{\log(2s/\delta)}{n}\le 2.5\sqrt{\frac{\theta_j^*\log(2s/\delta)}{n}}.\end{aligned}\tag{11}$$

Combining (10) and (11), we obtain

$$\sum_{j\in J'}\left(\sqrt{\bar{\boldsymbol{Y}}_{n,j}}-\sqrt{\theta_j^*}\right)^2\le 2.5^2\frac{s\log(2s/\delta)}{n},\tag{12}$$

with probability at least $1-\delta/2$. Finally, inequalities (9) and (12) together lead to

$$d_{\mathrm{H}}^2(\bar{\boldsymbol{Y}}_n,\boldsymbol{\theta}^*)=\frac{1}{2}\sum_{j=1}^n\left(\sqrt{\bar{\boldsymbol{Y}}_{n,j}}-\sqrt{\theta_j^*}\right)^2\le\frac{5s\,\log(2s/\delta)}{n}+\frac{\log(2/\delta)}{2n},$$

which is true with probability at least $1-\delta$. Using the triangle inequality and the fact that the Hellinger distance is smaller than the square root of the TV-distance, we get

$$\begin{aligned}d_{\mathrm{H}}(\bar{\boldsymbol{X}}_n,\boldsymbol{\theta}^*)&\le d_{\mathrm{H}}(\bar{\boldsymbol{X}}_n,\bar{\boldsymbol{Y}}_n)+d_{\mathrm{H}}(\bar{\boldsymbol{Y}}_n,\boldsymbol{\theta}^*)\\&\le\sqrt{d_{\mathrm{TV}}(\bar{\boldsymbol{X}}_n,\bar{\boldsymbol{Y}}_n)}+d_{\mathrm{H}}(\bar{\boldsymbol{Y}}_n,\boldsymbol{\theta}^*)\\&\le\varepsilon^{1/2}+\sqrt{\frac{5s\,\log(2s/\delta)}{n}}+\sqrt{\frac{\log(2/\delta)}{2n}},\end{aligned}$$

with probability at least $1-\delta$. This completes the proof of the theorem. □

Acknowledgments

This work was partially supported by the grant Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

References

J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794, 10 2011. MR2906886

S. Balakrishnan, S. S. Du, J. Li, and A. Singh. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212, 2017.

P. C. Bellec. Adaptive confidence sets in shape restricted regression. *arXiv preprint* arXiv:1601.05766, 2016.

P. C. Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.*, 46(2):745–780, 04 2018. MR3782383

K. Bhatia, P. Jain, P. Kamalaruban, and P. Kar. Consistent robust regression. In *Advances in Neural Information Processing Systems*, pages 2110–2119, 2017.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP, Oxford, 2013. MR3185193

D. Braess and T. Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004. MR2068697

V.-E. Brunel. Concentration of the empirical level sets of Tukey's halfspace depth. *Probability Theory and Related Fields*, 173(3-4):1165–1196, 2019. MR3936153

T. T. Cai and M. G. Low. Adaptive confidence balls. *Ann. Statist.*, 34(1):202–228, 2006. MR2275240

A. Carpentier, S. Delattre, E. Roquain, and N. Verzelen. Estimating minimum effect with outlier selection. *arXiv e-prints* arXiv:1809.08330, Sept. 2018.

M. Chen, C. Gao, and Z. Ren. A general decision theory for Huber's ε-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774, 2016. MR3579675

M. Chen, C. Gao, and Z. Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.*, 46(5):1932–1960, 10 2018. MR3845006

S. Chen, J. Li, and A. Moitra. Efficiently learning structured distributions from untrusted batches. In *Proccedings of STOC 2020, June 22-26*, pages 960–973. ACM, 2020.

Y. Chen, C. Caramanis, and S. Mannor. Robust sparse regression under adversarial corruption. In *International Conference on Machine Learning*, pages 774–782, 2013.

G. Chinot, G. Lecué, and M. Lerasle. Statistical learning with Lipschitz and convex loss functions. *arXiv e-prints* arXiv:1810.01090, Oct. 2018. MR4087486

O. Collier and A. S. Dalalyan. Multidimensional linear functional estimation in sparse gaussian models and robust estimation of the mean. *Electron. J. Statist.*, 13(2):2830–2864, 2019. MR3998929

A. Dalalyan and Y. Chen. Fused sparsity and robust estimation for linear models with unknown variance. In *Advances in Neural Information Processing Systems 25*, pages 1259–1267. Curran Associates, Inc., 2012.

A. S. Dalalyan and A. Minasyan. All-in-one robust estimator of the gaussian mean. *math.ST*, arXiv:2002.01432, 2020.

A. S. Dalalyan and M. Sebbar. Optimal Kullback-Leibler aggregation in mixture density estimation by maximum likelihood. *Math. Stat. Learn.*, 1(1):1–35, 2018. MR4049449

L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-gaussian mean estimators. *Ann. Statist.*, 44(6):2695–2725, 12 2016. MR3576558

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, USA*, pages 655–664, 2016. MR3631028

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of ICML 2017*, pages 999–1008, 2017.

I. Diakonikolas, J. Li, and L. Schmidt. Fast and sample near-optimal algorithms for learning multidimensional histograms. In *COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pages 819–842, 2018.

D. L. Donoho and M. Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.*, 20(4):1803–1827, 1992. MR1193313

D. L. Donoho and P. J. Huber. The notion of breakdown point. Festschr. for Erich L. Lehmann, 157–184, 1983. MR0689745

D. L. Donoho and A. Montanari. High dimensional robust m-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969, Dec 2016. MR3568043

J. Feng, H. Xu, S. Mannor, and S. Yan. Robust logistic regression and classification. In *Advances in Neural Information Processing Systems 27*, pages 253–261. Curran Associates, Inc., 2014.

A. Guntuboyina and B. Sen. Nonparametric shape-restricted regression. *Statist. Sci.*, 33(4):568–594, 11 2018. MR3881209

M. Hoffmann and R. Nickl. On adaptive inference and confidence bands. *Ann. Statist.*, 39(5):2383–2409, 2011. MR2906872

P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 1964. MR0161415

A. Jain and A. Orlitsky. Robust learning of discrete distributions from batches. *CoRR*, arXiv:1911.08532, 2019.

E. Joly, G. Lugosi, and R. Imbuzeiro Oliveira. On the estimation of the mean of a random vector. *Electron. J. Statist.*, 11(1):440–451, 2017. MR3619312

S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh. On learning distributions from their samples. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3–6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 1066–1100. JMLR.org, 2015.

K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674, Oct 2016. MR3631029

G. Lecué and M. Lerasle. Learning from MOM's principles: Le Cam's approach. *Stoch. Proc. App.*, 129(11):4385–4410, 2019. MR4013866

H. Liu and C. Gao. Density estimation with contamination: minimax rates and

theory of adaptation. *Electron. J. Stat.*, 13(2):3613–3653, 2019. MR4013747

G. Lugosi and S. Mendelson. Sub-gaussian estimators of the mean of a random vector. *Ann. Statist.*, 47(2):783–794, 04 2019. MR3909950

S. Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 11 2015. MR3378468

S. Minsker. Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.*, 46(6A):2871–2903, 12 2018. MR3851758

N. H. Nguyen and T. D. Tran. Robust lasso with missing and grossly corrupted observations. *IEEE Trans. Inform. Theory*, 59(4):2036–2058, 2013. MR3043781

M. Qiao and G. Valiant. Learning discrete distributions from untrusted batches. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11–14, 2018, Cambridge, MA, USA*, volume 94 of *LIPIcs*, pages 47:1–47:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. MR3761783

P. J. Rousseeuw and M. Hubert. Regression depth. *Journal of the American Statistical Association*, 94(446):388–402, 1999. MR1702314

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009. MR2724359

J. W. Tukey. Mathematics and the picturing of data. In *Proc. int. Congr. Math., Vancouver 1974*, volume 2, pages 523–531, 1975. MR0426989

D. Xia and V. Koltchinskii. Estimation of low rank density matrices: Bounds in schatten norms and other distances. *Electron. J. Statist.*, 10(2):2717–2745, 2016. MR3546973