Proceedings of Machine Learning Research vol 201:1–36, 2023 34th International Conference on Algorithmic Learning Theory

Robust Estimation of Discrete Distributions under Local Differential Privacy

Julien Chhor* JULIEN.CHHOR@ENSAE.FR
CREST/ENSAE

Flore Sentenac* FLORE.SENTENAC@ENSAE.FR
CREST/ENSAE

Editors: Shipra Agrawal and Francesco Orabona

Abstract

Although robust learning and local differential privacy are both widely studied fields of research, combining the two settings is just starting to be explored. We consider the problem of estimating a discrete distribution in total variation from n contaminated data batches under a local differential privacy constraint. A fraction $1-\alpha$ of the batches contain k i.i.d. samples drawn from a discrete distribution p over d elements. To protect the users' privacy, each of the samples is privatized using an ϵ-locally differentially private mechanism. The remaining αn batches are an adversarial contamination. The minimax rate of estimation under contamination alone, with no privacy, is known to be $\alpha/\sqrt{k} + \sqrt{d/kn}$. Under the privacy constraint alone, the minimax rate of estimation is $\sqrt{d^2/\epsilon^2 kn}$. We show, up to a $\sqrt{\log(1/\alpha)}$ factor, that combining the two constraints leads to a minimax estimation rate of $\alpha\sqrt{d/\epsilon^2 k} + \sqrt{d^2/\epsilon^2 kn}$, larger than the sum of the two separate rates. We provide a polynomial-time algorithm achieving this bound, as well as a matching information theoretic lower bound.

Keywords: Privacy, Robustness, Adversarial Contamination, Multionmial Distributions, Statistical Optimality

1. Introduction

In recent machine learning developments, the growing need to analyze potentially corrupted, biased or sensitive data has given rise to unprecedented challenges. To extract relevant information from today's data, studying algorithms under new learning constraints has emerged as a major necessity. To name a few, let's mention learning from incomplete data, transfer learning, fairness, robust learning or privacy. Although each one of them has been subject to intense progress in recent works, combining learning constraints has just started to become an important topic of interest. In this work, we propose to study how to estimate discrete distributions under the constraint of both being robust to adversarial contamination and of ensuring local differential privacy.

On the one hand, robust learning has received considerable attention over the past decades. This recent research has been developing in two main directions. The first one deals with robustness to heavy tails, see Catoni (2012), see also Lugosi and Mendelson (2019) for an excellent review. The second one explores robustness to outliers. It mainly considers two contamination models, which are the Huber contamination model Huber (1968), Huber (1992), Huber (2004), Huber and Ronchetti (2009), where the outliers are iid with an unknown probability distribution, and the *adversarial* contamination, where the outliers are added by a malicious adversary who knows the estimation procedure, the underlying distribution and the data and seeks to deteriorate the procedure's estimation performance (Diakonikolas et al. (2019); Rousseeuw and Hubert (2011); Dalalyan and

* Equal contributions

Minasyan (2020); Lecué et al. (2020)).

On the other hand, preserving the privacy of individuals has emerged as a major concern, as more and more sensitive data are collected and processed. The most commonly used privatization framework is that of differential privacy (Dwork et al. (2006), Acharya et al. (2019), Butucea et al. (2020), Lam-Weil et al. (2020), Berrett and Butucea (2020), Cai et al. (2019), Barnes et al. (2020a), Barnes et al. (2020b), Barnes et al. (2019)). Both central and local models of privacy are considered in the field. In the centralized case, a global entity collects the data and analyzes it before releasing a privatized result, from which the original data should not be possible to infer. In local privacy, the data themselves are released and should remain private (Duchi et al. (2014)). The paper focuses on the latter notion. A vast line of work also studies private mechanism under communication constraints (Acharya et al. (2020a), Acharya et al. (2020b), Acharya et al. (2020c)), which we do not consider here, but adding a communication constraint would be interesting future work.

Connections between robustness and *global* differential privacy have been recently discussed in (Pinot et al. (2019); Lecuyer et al. (2019), Naseri et al. (2020)). These papers show that the two notions rely on the same theoretical concepts, and that results in the two fields are related. In other words, robustness and *global* differential privacy work well together. Several papers developed algorithms under robustness and *global* differential privacy constraints (Liu et al. (2021a), Hopkins et al. (2021), Liu et al. (2021b), Ashtiani and Liaw (2021)).

In this paper, we study how *local* differential privacy interacts with robustness. This interaction has been studied previously in Cheu et al. (2019), where the authors provide upper and lower bound for estimating discrete distributions under the two constraints. The lower bound was later tightened in Acharya et al. (2021). The papers also study testing. We detail in Section 1.1 how our setting is a generalization of theirs. The work of Li et al. (2022) also considers this interaction. We explain in more details the differences between their setting and ours in Section 1.1.

In this paper, we study how to combine robust statistics with local differential privacy for estimating discrete distributions over finite domains. Assume that we want to gather information from n data centers (think of n hospitals for instance). For each of them, we collect k iid observations with unknown discrete distribution p to be estimated. To protect the users' privacy (patients data in the hospital example), each single one of the nk observations is privatized using an ϵ-locally differentially private mechanism (see the formal definition of local differential privacy in Subsection 2.1). However, an α-fraction of the data centers are untrustworthy and can send adversarially chosen data. The goal is to estimate p in total variation distance (or ℓ_1 distance) from these n corrupted and privatized batches of size k. This setting is quite natural, as in many applications, the data are collected in batches, some of which may be untrustworthy or even adversarial.

1.1. Related work

With the local differential privacy constraint only (i.e. without contamination), the problem of estimating discrete distributions has been solved in (Evfimievski et al. (2003), Kasiviswanathan et al. (2008),Ye and Barg (2017)) where the authors propose a polynomial-time and minimax optimal algorithm for estimation under ℓ_1 and ℓ_2 losses. Note that *without privacy and outliers*, the minimax estimation rate in ℓ_1 is known to be $\sqrt{\frac{d}{N}}$, where N is the number of iid samples with a discrete distribution over d elements (see, e.g. Han et al. (2015), Devroye and Lugosi (2001)). The paper Duchi et al. (2014) shows that under privacy alone, the ℓ_1 min-

imax rate scales as $\frac{d}{\epsilon\sqrt{N}}$. We give an alternative proof of the lower bound of Duchi et al. (2014), in Appendix 9.

With the robustness constraint only (i.e. with n adversarially corrupted batches but without local differential privacy), the problem of estimating discrete distributions has been considered in Qiao and Valiant (2017). For $k = 1$, it is well known that $\Omega(\alpha)$ error is unavoidable. However, Qiao and Valiant (2017) surprisingly prove that the error can be reduced provided that k is large enough. More precisely, they show that with no privacy but under contamination, the minimax risk of estimation under ℓ_1 loss from n batches of size k and α adversarial corruption on the batches scales as $\sqrt{\frac{d}{N}} + \frac{\alpha}{\sqrt{k}}$, where $N = nk$. Qiao and Valiant (2017) both provide an information theoretic lower bound and a minimax optimal algorithm, unfortunately running in exponential time in either k or d. Polynomial-time algorithms were later proposed by Chen et al. (2020), Jain and Orlitsky (2020) and were shown to reach the information theoretic lower bound up to an extra $\sqrt{\log(\frac{1}{\alpha})}$ factor. In this specific setting, it is not known if this extra factor represents a computational gap between polynomial-time and exponential-time algorithms. However, for the problem of robust mean estimation of *normal distributions*, some lower bounds suggest that this exact quantity cannot be removed from the rate of computationally tractable estimators (see Diakonikolas et al. (2017)).

Closer to our setting, the papers by Cheu et al. (2019), Acharya et al. (2021) and Li et al. (2022) combine robustness with local differential privacy. The problem studied here is a generalization of the first two papers where the authors consider un-batched data, which corresponds to $k = 1$ in our setting. The setting considered by Li et al. (2022) is not the same as ours, as they do not consider discrete distributions and implicitly assume $k = 1$. More importantly, in their setting, contamination comes *before* privacy: some of the raw data $X_1, \dots, X_n$ are outliers themselves, and the privacy mechanism is applied on each X_i. Conversely, in our work and in the previous two papers, contamination occurs *after* privacy: none of the raw data are outliers and the adversary is allowed to choose the contamination directly on the set of privatized data. As we will highlight below, this difference yields fundamentally different phenomena compared to the results in Li et al. (2022).

1.2. Summary of the contributions

In this paper, we study the interplay between local differential privacy and *adversarial* contamination, when the contamination comes *after* the data have been privatized. In this case, we prove that the resulting estimation rate is not merely the sum of the two estimation rates stated in Duchi et al. (2014) and Qiao and Valiant (2017) but is always slower. More specifically, the term due to the contamination in the bound suffers a multiplicative inflation of $\sqrt{d}/\epsilon$. This generalizes a phenomenon first observed in Cheu et al. (2019). This phenomenon stands in contrast with Li et al. (2022), for which the resulting rate is exactly the sum of the rate with privacy but no contamination, plus the rate with contamination but no privacy. The reason is that in Li et al. (2022), contamination occurs *before* privacy. We provide an explicit algorithm that returns an estimator achieving the optimal bound up to a factor $\sqrt{\log(1/\alpha)}$, and runs polynomially in all parameters. This algorithm is an adaptation to our setting of methods that were previously used for robust estimation of discrete distributions (Jain and Orlitsky (2020, 2021)). On a side note, the algorithms introduced in Cheu et al. (2019) and Acharya et al. (2021), treating case $k = 1$, require the use of a public coin. The proposed algorithm also holds in their setting and relieves this assumption, also it has the downside that the bound then has this extra $\sqrt{\log(1/\alpha)}$ factor.

2. Setting

2.1. Definitions

For any integer $d \geq 2$, we write $[d] = \{1, \dots, d\}$, and we denote by $\mathcal{P}_d = \left\{p \in \mathbb{R}_+^d \;\middle|\; \sum_{j=1}^d p_j = 1\right\}$ the set of probability vectors over $[d]$. For any $x \in \mathbb{R}^d$, we write $\|x\|_1 = \sum_{j\in[d]} |x_j|$ and $\|x\|_2^2 = \sum_{j\in[d]} x_j^2$. For any two probability distributions p, q over some measurable space $(\mathcal{X}, \mathcal{A})$, we denote by $TV(p,q) = \sup_{A\in\mathcal{A}} |p(A) - q(A)|$ the total variation between p and q. Fix $\epsilon \in (0,1)$ and consider two measurable spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Z}, \mathcal{B})$. A Markov transition kernel $Q : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Z}, \mathcal{B})$ is said to be a (non-interactive) ϵ-*locally differentially private* mechanism if it satisfies

$$\sup_{B\in\mathcal{B}} \sup_{x,x'\in\mathcal{X}} \frac{Q(B|x)}{Q(B|x')} \leq e^\epsilon. \tag{1}$$

For any $x \in \mathcal{X}$, we say that the random variable Z is a privatized version of x if $Z \sim Q(\cdot|x)$. The measurable space $(\mathcal{Z}, \mathcal{B})$ is called the *image space* of Q. In what follows, we use the Landau notation O which hides an absolute constant, independent of $d, \alpha, n, k, \epsilon, Q, p$.

2.2. Model

We consider the problem of learning a discrete distribution p over a finite set $\{1, \dots, d\}$, $d \geq 3$ under two learning constraints: a) ensuring ϵ-local differential privacy and b) being robust to adversarial contamination. To this end, we assume that the data are generated as follows. For some small enough absolute constant $c \in (0, \frac{1}{100})$ and for some known corruption level $\alpha \in (0, c)$, we will use the notation $n' = n(1-\alpha)$ throughout and assume that $n' \in \mathbb{N}$.

1. First, n' iid *batches* of observations $X^1, \dots, X^{n'}$ are collected. More precisely, each batch X^b can be written as $X^b = (X_1^b, \dots, X_k^b)$ and consists of k iid random observations with an unknown discrete distribution $p \in \mathcal{P}_d$, i.e. $\forall (b,l,j) \in [n'] \times [k] \times [d] : \mathbb{P}(X_l^b = j) = p_j$.

2. Second, we privatize each of the $n'k$ observations using an ϵ-LDP mechanism Q, yielding n' iid batches $Y^1, \dots, Y^{n'}$ such that $Y^b = (Y_1^b, \dots, Y_k^b)$ where $Y_l^b | X_l^b \sim Q(\cdot|X_l^b)$. We denote by Qp the distribution of any random variable Y_l^b. We then have: $Qp(dz) = \sum_{j\in[d]} p_j Q(dz|j)$, where $Q(dz|j)$ is a shorthand for $Q(dz|X = j)$. The mechanism Q is chosen by the statistician in order to preserve statistical performance while ensuring privacy.

3. An adversary is allowed to build $n\alpha$ batches $Y^{n'+1}, \dots, Y^n$ on which no restriction is imposed. Then, he shuffles the set of n batches $(Y_1, \dots, Y_n)$. The resulting set of observations, denoted as $B = (Z^1, \dots, Z^n)$, is referred to as the α-corrupted family of batches.

The observed dataset therefore consists of $n = |B|$ batches of k samples each. Among these batches is an unknown collection of *good batches* $B_G \subset B$ of size $n(1-\alpha)$, corresponding to the non-contaminated batches. The remaining set $B_A = B \setminus B_G$ of size $n\alpha$, denotes the unknown set of adversarial batches.

The statistician never has access to the actual observations $X^1, \dots, X^{n'}$, but only to $Z^1, \dots, Z^n$ where $Z^b = (Z_1^b, \dots, Z_k^b)$. Each batch is assumed to be either entirely clean or adversarially corrupted. Note that

observing n batches of size k encompasses the classical case where $k = 1$, for which the data consist of n iid and α-corrupted *single observations* rather than batches. On top of being more general, the setting with general k allows us to derive faster rates for large k than for the classical case $k = 1$. Note also that in our setting, the contamination comes after the data have been privatized, which is one of the main differences with Li et al. (2022), where the authors assume that the Huber contamination comes before privacy. The examples considered by the authors are 1-dimensional mean estimation and density estimation without batches (i.e. for $k = 1$). In these settings, the authors surprisingly prove that the algorithm that would be used in absence of corruption is automatically robust to Huber contamination.

In our setting, we would like to answer the following questions:

1. In the presence of batches, does the term due to the contamination in the bound suffer a multiplicative inflation of $\sqrt{d}/\epsilon$, as it is the case when there are no batches (Cheu et al. (2021))?
2. If Q_α denotes the optimal privacy mechanism for α-contamination, how does Q_α depend on α?

We answer these questions as follows:

1. The term due to the contamination does suffer a multiplicative inflation of $\sqrt{d}/\epsilon$.
2. The optimal privacy mechanism Q_α does not depend on α, whereas the optimal estimator does.

We introduce the minimax framework. An *estimator* $\widehat{p}$ is a measurable function of the data taking values in $\mathcal{P}_d$: $\widehat{p} : \mathcal{Z}^{nk} \longrightarrow \mathcal{P}_d$. For any set of n' clean batches $Y^1, \dots, Y^{n'}$ where $Y^b = (Y_1^b, \dots, Y_k^b)$ and $n' = n(1-\alpha)$, we define the set of α-contaminated families of n batches as

$$\mathcal{C}(Y^1, \dots, Y^{n'}) = \left\{ (Z^b)_{b=1}^n \,\middle|\, \exists J \subset [n] \text{ s.t. } |J| = n\alpha \text{ and } \{Z^b\}_{b \notin J} = \{Y^1, \dots, Y^{n'}\} \right\}. \tag{2}$$

We are interested in estimating $p \in \mathcal{P}_d$ with guarantees in high probability. We therefore introduce the minimax estimation rate of p in high probability as follows.

Definition 1 *Given $\delta > 0$, the minimax rate of estimation rate of $p \in \mathcal{P}_d$ given the privatized and α-corrupted batches $(Z^b)_{b=1}^n$ where $\forall i \in \{1, \dots, n\} : Z^b = (Z_1^b, \dots Z_k^b)$ is defined as the quantity $\psi_\delta^*(n, k, \epsilon, d, \alpha)$ satisfying*

$$\psi_\delta^*(n, k, \epsilon, d, \alpha) = \inf \left\{ \psi > 0 \,\middle|\, \inf_{\hat{p}, Q} \sup_{p \in \mathcal{P}_d} \mathbb{P}\Big(\sup_{z \in \mathcal{C}(Y)} \big\| \widehat{p}(z) - p \big\|_1 > \psi \Big) \leq \delta \right\}. \tag{3}$$

where the infimum is taken over all estimators $\widehat{p}$ and all ϵ-LDP mechanisms Q, and the expectation is taken over all collections of n' clean batches $Y^1, \dots, Y^{n'}$ where $Y^b = (Y_1^b, \dots, Y_k^b)$ and $Y_l^b \overset{iid}{\sim} Qp$. Informally, ψ_δ^* represents the infimal distance such that there exists an estimator $\hat{p}$ able to estimate any $p \in \mathcal{P}_d$ within total variation ψ_δ^* with probability $\geq 1 - \delta$. The ℓ_1 norm is a natural metric for estimating discrete distributions since $TV(p, q) = \frac{1}{2}\|p - q\|_1$ for any $p, q \in \mathcal{P}_d$ (see Tsybakov (2008)).

3. Results

We now state our main Theorem.

Theorem 2 *Assume $d \geq 3$. There exist absolute constants $c, C, C', C'' > 0$ such that for $\delta = C'e^{-d}$, we have:*

$$\psi^*_\delta(n,k,\epsilon,\alpha,d) \geq c\left\{\left(\frac{d}{\epsilon\sqrt{kn}} + \frac{\alpha}{\epsilon}\sqrt{\frac{d}{k}}\right) \wedge 1\right\},$$

and if $n \geq C''d$ then

$$\psi^*_\delta(n,k,\epsilon,\alpha,d) \leq C\left\{\left(\frac{d}{\epsilon\sqrt{kn}} + \frac{\alpha\sqrt{\log(1/\alpha)}}{\epsilon}\sqrt{\frac{d}{k}}\right) \wedge 1\right\}.$$

In short, we prove that with probability at least $1 - O(e^{-d})$, it is possible to estimate any $p \in \mathcal{P}_d$ within total variation of the order of $\left(\frac{d}{\epsilon\sqrt{kn}} + \frac{\alpha}{\epsilon}\sqrt{\frac{d}{k}}\right) \wedge 1$ up to log factors and provided that $n \geq C''d$. We can compare this rate with existing results in the literature.

- As shown in Duchi et al. (2014), the term $\frac{d}{\epsilon\sqrt{kn}} \wedge 1$ corresponds to the estimation rate under privacy if there were no outliers, with a total number of observations of $N = nk$.
- The term $\frac{\alpha\sqrt{d}}{\epsilon\sqrt{k}} \wedge 1$ reveals an interesting interplay between contamination and privacy. In absence of privacy, Qiao and Valiant (2017) proved that the contribution of the contamination is of the order of $\frac{\alpha}{\sqrt{k}} \wedge 1$. The effect of the corruption therefore becomes more dramatic when it occurs after privatization.
- Letting $k' = \frac{\epsilon^2}{d}k$, our rate rewrites $\psi^*(n,k,\epsilon,\alpha,d) \asymp \left(\sqrt{\frac{d}{k'n}} + \frac{\alpha}{\sqrt{k'}}\right) \wedge 1$. Noticeably, this rate exactly corresponds to the rate from Qiao and Valiant (2017) if we had an α-corrupted family of n *non-privatized* batches $X_1, \ldots, X_n$, and if each batch contained k' observations. The quantity k' therefore acts as an effective sample size and the effect of privacy amounts to shrinking the number of observations by a factor ϵ^2/d.
- For the upper bound, the assumption $n \geq C''d$ is classical in the robust statistics literature, even in the gaussian setting (see e.g. Dalalyan and Minasyan (2020)).

3.1. Lower bound

The following Proposition yields an information theoretic lower bound on the best achievable estimation accuracy under local differential privacy and adversarial contamination.

Proposition 3 *Assume $d \geq 3$. There exist two absolute constants $C, c > 0$ such that for all $\alpha \in (0, \frac{1}{2})$, for all estimator $\hat{p}$ and all ϵ-LDP mechanism Q, there exists a probability vector $p \in \mathcal{P}_d$ satisfying*

$$\mathbb{P}_p\left[\sup_{z' \in \mathcal{C}(Y)} \left\|\widehat{p}(z') - p\right\|_1 \geq c\left\{\left(\frac{d}{\epsilon\sqrt{kn}} + \frac{\alpha\sqrt{d}}{\epsilon\sqrt{k}}\right) \wedge 1\right\}\right] \geq Ce^{-d},$$

where the probability $\mathbb{P}_p$ is taken over all collections of n' clean batches $Y = (Y^1, \ldots, Y^{n'})$ where $Y^b = (Y^b_1, \ldots, Y^b_k)$ and $Y^b_l \overset{iid}{\sim} Qp$.

The proof is given in Appendix 8. At a high level, the term $\frac{d}{\epsilon\sqrt{kn}} \wedge 1$ comes from the classical lower bound given in Duchi et al. (2014). The proof of the second term $\frac{\alpha\sqrt{d}}{\epsilon\sqrt{k}} \wedge 1$ is new. It is based on the fact that for any ϵ-LDP mechanism Q, it is possible to find two probability vectors $p, q \in \mathcal{P}_d$ such that $\|p-q\|_1 \gtrsim \frac{\alpha\sqrt{d}}{\epsilon\sqrt{k}} \wedge 1$ and $TV(Qp^{\otimes k}, Qq^{\otimes k}) \leq \alpha$. In other words, we prove:

$$\inf_{Q} \sup_{\substack{(p,q)\in\mathcal{P}_d: \\ TV(Qp^{\otimes k},Qq^{\otimes k})\leq\alpha}} \|p-q\|_1 \gtrsim \frac{\alpha\sqrt{d}}{\epsilon\sqrt{k}} \wedge 1.$$

In the proof, we argue that $TV(Qp^{\otimes k}, Qq^{\otimes k}) \leq \alpha$ represents an indistinguishability condition under α-contamination. Namely, it implies that, even if we had arbitrarily many clean batches drawn from p or q, the adversary could add $n\alpha$ corrupted batches such that the resulting family of batches has the same distribution under p or q. By observing this limiting distribution, it is therefore impossible to recover the underlying probability vector so that an error of $\|p-q\|_1/2$ is unavoidable.

To exhibit two vectors $p, q \in \mathcal{P}$ satisfying this, we restrict ourselves to vectors satisfying $\chi^2(Qp||Qq) \leq C\frac{\alpha^2}{k}$ for some small enough absolute constant $C > 0$, which implies that $TV(Qp^{\otimes k}, Qq^{\otimes k}) \leq \alpha$ for any ϵ-locally differenially private mechanism Q (see Tsybakov (2008) section 2.4). In particular, we prove the relation $\chi^2(Qp\,||\,Qq) = \Delta^T\Omega\Delta$, where $\Delta = p-q$ and $\Omega = \Omega(Q) = \left[\int_{\mathcal{Z}} \left(\frac{Q(z|i)}{Q(z|1)} - 1\right)\left(\frac{Q(z|j)}{Q(z|1)} - 1\right) Q(z|1)dz\right]_{i,j\in[d]}$ is a nonnegative symmetric matrix for any Q. The eigenvectors of Ω play an important role. Namely, we prove that we can choose a vector Δ in the span of the first $\lceil\frac{2d}{3}\rceil$ eigenvectors of Ω such that $\Delta^T\Omega\Delta \leq C\frac{\alpha^2}{k}$ and $\|\Delta\|_1 \gtrsim \frac{\alpha\sqrt{d}}{\epsilon\sqrt{k}} \wedge 1$. Defining the vectors $p = \left(\frac{|\Delta_j|}{\|\Delta\|_1}\right)_{j=1}^d \in \mathcal{P}_d$ and $q = p - \Delta$ ends the proof.

4. Upper bound

We now address the upper bound by proposing an ϵ-LDP mechanism Q for privatizing the clean data $X^1, \ldots, X^{n'}$ as well as an algorithm $\widehat{p}$ for robustly estimating vector p given an α-contaminated family of n batches $Z^1, \ldots, Z^n$. Each non-private data point $X_i^b \in [d]$ is privatized using the RANDOMIZED RESPONSE algorithm (Duchi et al. (2014); Kairouz et al. (2016)). In this procedure, the *privatization channel* Q randomly maps each point $X \in [d]$ to a point $Z \in \{0,1\}^d$ by flipping its coordinates independently at random with probability $\lambda = \frac{1}{e^{\epsilon/2}+1}$:

$$\forall j \in [d]: \quad Z(j) = \begin{cases} \mathbb{1}_{X=j} & \text{with probability } 1-\lambda, \\ 1 - \mathbb{1}_{X=j} & \text{otherwise.} \end{cases}$$

We now derive a polynomial-time algorithm taking as input the α-contaminated family of batches $(Z^b)_{b\in[n]}$ and returning an estimate $\hat{p}$ for p with the following properties.

Theorem 4 (Upper Bound) *For any $\alpha \in (0, 1/100]$, $\epsilon \in (0,1]$, if $n \geq \frac{4d}{\alpha^2\ln(e/\alpha)}$, Algorithm 1 runs in polynomial time in all parameters and its estimate $\widehat{p}$ satisfies $||\widehat{p}-p||_1 \lesssim \frac{\alpha}{\epsilon}\sqrt{\frac{d\ln(1/\alpha)}{k}}$ w.p. at least $1-O(e^{-d})$.*

If $n \geq O(d)$, then there exists $\alpha' \in (0, 1/100]$ s.t. $n = \frac{4d}{(\alpha')^2 \ln(1/\alpha')}$. Running the algorithm with that parameter α' rather than the true α gives the following result.

Corollary 5 *If $n \geq O(d)$, then the algorithm's estimate satisfies $||\widehat{p} - p||_1 \lesssim \frac{d}{\epsilon}\sqrt{\frac{e}{nk}}$ with probability at least $1 - O(e^{-d})$.*

Theorem 4 and Corollary 5 yield the upper bound. We have not seen the regime $n \leq d$ explored in the literature, even with robustness only. This would be an interesting research direction for future work. Note that for the estimate $\hat{p}$ given by Algorithm 1 we can have $\|\hat{p}\|_1 \neq 1$. The next corollary, proved in Appendix 7, states that normalizing $\hat{p}$ yields an estimator in $\mathcal{P}_d$ with the same estimation guarantees as in Theorem 4.

Corollary 6 *Let the assumptions of Theorem 4 be satisfied and let $\hat{p}$ denote the output of Algorithm 1. Define $\widehat{p}^* = \frac{\widehat{p}_+}{\|\widehat{p}_+\|_1}$ where $\widehat{p}_+(j) = 0 \vee \widehat{p}(j)$ for all $j \in [d]$, then $||\widehat{p}^* - p||_1 \leq 2||\widehat{p} - p||_1 \lesssim \frac{\alpha}{\epsilon}\sqrt{\frac{d\ln(1/\alpha)}{k}}$ holds with probability at least $1 - O(e^{-d})$.*

4.1. Description of the algorithm

We now give a high level description of our algorithm. It is based on algorithms for robust discrete distribution estimation, Jain and Orlitsky (2020, 2021). For each $S \subseteq [d]$, define $q(S) = \sum_{j \in S} q_j$ and $p(S) = \sum_{j \in S} p_j$. The quantities $\widehat{q}$, $\widehat{p}$ will respectively denote the estimators of p and q. Recalling that $TV(p, \widehat{p}) = \sup_{S \subseteq [d]} \left|p(S) - \widehat{p}(S)\right|$, we aim at finding $\widehat{p}$ satisfying $\left|p(S) - \widehat{p}(S)\right| \lesssim \frac{\alpha}{\epsilon}\sqrt{\frac{d\ln(1/\alpha)}{k}}$ for all $S \subseteq [d]$. To this end, it is natural to first estimate the auxiliary quantity

$$q(j) := \mathbb{E}_p\big[Z(j)\big|\ Z \text{ is a good sample}\big] \quad \text{for all } j \in [d],$$

as it is linked with $p(j)$ through the formula $p(j) = \frac{q(j)-1}{1-2\lambda}$. Our algorithm therefore first focuses on robustly estimating q and outputs $\widehat{p} = \frac{\widehat{q}-\mathbb{1}}{1-2\lambda}$. If there were no outliers, we would estimate $q(j)$ by $\frac{1}{nk}\sum_{b \in [n]}\sum_{l \in [k]} Z_l^b(j)$. In the presence of outliers, our algorithm iteratively deletes the batches that are likely to be contaminated, and returns the empirical mean of the remaining data. More precisely, at each iteration, the current collection of remaining batches B' is processed as follows:

1. Compute the *contamination rate* $\sqrt{\tau_{B'}}$ (defined in equation 9) of the collection B'. If $\sqrt{\tau_{B'}} \leq 200$, return the empirical mean of the elements in B'.
2. If $\sqrt{\tau_{B'}} \geq 200$, compute the *corruption score* ε_b (defined in equation 10) of each batch $b \in B'$. Select the subset B^o of the $n\alpha$ batches of B' with top corruption scores. Iteratively delete one batch in B^o: at each step, choose a batch b with probability proportional to ε_b, until the sum of all ε_b in B^o has been halved.

At a high level, the *contamination rate* $\sqrt{\tau_{B'}}$ quantifies how much contamination is left in the current collection B'. The *corruption score* ε_b quantifies how likely it is for batch b to be an outlier. Both the *contamination rate* and the *corruption scores* can be computed in polynomial time (see Remark 11). The algorithm therefore terminates in polynomial time, as it removes at least one batch per iteration. We give its pseudo-code below.
We now give a high level description of our algorithm's theoretical guarantees. Recall that B_G denotes the set of non-contaminated batches and B_A the set of adversarial batches. Throughout the paper, for any collection of batches $B' \subseteq [d]$, we will use the following shorthands:

$$B'_G = B' \cap B_G \text{ and } B'_A = B' \cap B_A.$$

Algorithm 1: ROBUST ESTIMATION PROCEDURE

input: Corruption level α, Batch collection B
$B' \leftarrow B$
while *contamination rate of* B', $\sqrt{\tau_{B'}} \geq 200$ **do**
  $\forall b \in B'$ compute corruption score ε_b
  $B^o \leftarrow \{\alpha|B|$ Batches with top corruption scores$\}$
  $\alpha_{\text{tot}} = \sum_{b\in B^o} \varepsilon_b$
  while $\sum_{b\in B^o} \varepsilon_b \geq \alpha_{tot}/2$ **do**
    Delete a batch from B^o, picking batch b with probability proportional to ε_b
  end
end
$\widehat{q}_{B'} = \frac{1}{|B'|}\sum_{b\in B'}\sum_{l=1}^{k} Z_l^b$ and $\hat{p} = \frac{\widehat{q}-\mathbb{1}}{1-2\lambda}$
output: Estimation $\hat{p}$

Assume that $n \geq O\left(\frac{d}{\alpha^2 \log(e/\alpha)}\right)$.

- In Lemma 13, we show that each deletion step has a probability at least $3/4$ of removing an adversarial batch. By a direct Chernoff bound, there is only a probability $\leq O(e^{-\alpha|B|}) \leq O(e^{-d})$ of removing more than $2\alpha|B_G|$ clean batches before having removed all the corrupted batches. In other words, our algorithm keeps at least $(1-2\alpha)n$ of the good batches with high probability.

- As proved in equations 11 and 12, as soon as a subset B' contains at least $(1-2\alpha)n$ good batches, it holds with probability $\geq 1 - O(e^{-d})$ that for all $S \subseteq [d]$

$$\begin{cases} |\widehat{q}_{B'}(S) - q(S)| \lesssim \left(1+\sqrt{\tau_{B'}}\right)\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}}, & \text{(i)} \\ \sqrt{\tau_{B'_G}} \leq 200. & \text{(ii)} \end{cases} \tag{4}$$

  There are two cases. If the algorithm has eliminated all the outliers, then it has kept at least $(1-2\alpha)n$ clean batches with probability $1 - O(e^{-d})$. Then condition (i) $\sqrt{\tau_{B'}} = \sqrt{\tau_{B'_G}} \leq 200$ ensures that the algorithm terminates. Otherwise, the algorithm stops before removing all of the outliers, but in this case, the termination condition guarantees that $\sqrt{\tau_{B'}} \leq 200$. In both cases, condition (ii) yields that the associated estimator $\widehat{q} := \widehat{q}_{B_{\text{out}}}$ has an estimation error satisfying $\sup_{S\subseteq[d]} |\widehat{q}(S) - q(S)| \lesssim \alpha\sqrt{\frac{d\ln(e/\alpha)}{k}}$ with probability $\geq 1 - O(e^{-d})$.

- Finally, we link the estimation error of $\widehat{q}$ to that of $\widehat{p}$

$$\begin{aligned} ||\widehat{p} - p||_1 &\leq 2\max_{S\in[d]} |\widehat{p}(S) - p(S)| \quad \text{(see Lemma 20)} \\ &\leq 2\max_{S\subseteq[d]} \left| \sum_{j\in S} \frac{1}{1-2\lambda}\left(\hat{q}_j - 1\right) - \frac{1}{1-2\lambda}\left(q_j - 1\right) \right| \\ &\leq \frac{1}{1-2\lambda}\max_{S\in[d]} |\hat{q}(S) - q(S)| \leq \frac{5}{\epsilon}\max_{S\in[d]} |\hat{q}(S) - q(S)| \end{aligned}$$

$$\lesssim \frac{\alpha}{\epsilon}\sqrt{\frac{d\ln(e/\alpha)}{k}} \quad \text{with probability } \geq 1 - O(e^{-d}),$$

which yields the estimation guarantee over $\widehat{p}$ and proves Theorem 4.

We now move to the formal definitions of the quantities involved in the algorithm and state all the technical results mentioned.

4.2. Technical results

Wlog, assume that $6\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}} \leq 1$. Otherwise the upper bound of the theorem is clear. For any set $S \subseteq [d]$ and any observation Z_i^b, we define the empirical weight of S in Z_i^b as $Z_i^b(S) := \sum_{j\in S} Z_i^b(j)$. This quantity is an estimator of $q(S)$. For each batch Z^b and each collection of batches $B' \subseteq B$, we aggregate these estimators by building

$$\widehat{q}_b(S) := \frac{1}{k}\sum_{i=1}^{k} Z_i^b(S) \quad \text{and} \quad \widehat{q}_{B'}(S) := \frac{1}{|B'|}\sum_{b\in B'} \widehat{q}_b(S).$$

Our goal is to remove batches Z^b that do not satisfy some concentration properties verified by clean batches. To this end, we introduce empirical estimators of the second order moment:

$$\widehat{\mathrm{Cov}}^{B'}_{S,S'}(b) := \left[\widehat{q}_b(S) - \widehat{q}_{B'}(S)\right]\left[\widehat{q}_b(S') - \widehat{q}_{B'}(S')\right] \tag{5}$$

$$\widehat{\mathrm{Cov}}_{S,S'}(B') := \frac{1}{|B'|}\sum_{b\in B'} \widehat{\mathrm{Cov}}^{B'}_{S,S'}(b)\,. \tag{6}$$

In Appendix 6.4, we give the expression of $\mathrm{Cov}_{S,S'}(q)$ s.t.

$$\mathrm{Cov}_{S,S'}(q) = \mathbb{E}\left[\widehat{\mathrm{Cov}}_{S,S'}\left(B'_G\right)\right].$$

We are now ready to define the essential concentration properties satisfied by the clean batches with high probability (see Lemma 8).

Definition 7 (Nice properties of good batches)
1. For all $S, S' \subseteq [d]$, all sub-collections $B'_G \subseteq B_G$ of good batches of size $|B'_G| \geq (1-2\alpha)|B_G|$,

$$\left|\widehat{q}_{B'_G}(S) - q(S)\right| \leq 6\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}}, \tag{7}$$

$$\left|\widehat{Cov}_{S,S'}\left(B'_G\right) - Cov_{S,S'}(\widehat{q}_{B'_G})\right| \leq \frac{250d\alpha\ln\left(\frac{e}{\alpha}\right)}{k}. \tag{8}$$

2. For all $S, S' \subseteq [d]$, for any sub collection of good batches B''_G s.t. $|B''_G| \leq \alpha|B_G|$,

$$\sum_{b\in B''_G}\left[\widehat{q}_b(S) - q(S)\right]\left[\widehat{q}_b(S') - q(S')\right] \leq \frac{33\alpha d|B_G|\ln(e/\alpha)}{k}.$$

Lemma 8 (Nice properties of good batches) *If* $|B_G| \geq \frac{3d}{\alpha^2 \ln(e/\alpha)}$, *the nice properties of the good batches hold with probability* $1 - 10e^{-d}$.

The proof is very similar to that of Lemma 3 in Jain and Orlitsky (2020), and can be found in Appendix 6.2 where we clarify which technical elements change.

In the case where $S' = S$, we use the shorthands $\widehat{\mathrm{Cov}}_{S,S}(B') = \widehat{\mathbf{V}}_S(B')$ and $\mathrm{Cov}_{S,S}(B') = \mathbf{V}_S(B')$. The following Lemma states that the quality of estimator $\widehat{q}_{B'}$ is controlled by the concentration of $\left|\widehat{\mathbf{V}}_S(B') - \mathbf{V}_S(\widehat{q})\right|$.

Lemma 9 (Variance gap to estimation error) *If conditions 1 and 2 hold and* $\max_{S\subseteq[d]} |\widehat{q}_{B'}(S) - q(S)| \leq 11$, *then for any subset* B' *s.t.* $|B'_G| \geq (1-2\alpha)|B_G|$ *and for any* $S \subset [d]$, *we have:*

$$|\widehat{q}_{B'}(S) - q(S)| \leq 28\alpha\sqrt{\frac{d\ln(6e/\alpha)}{k}} + 2\sqrt{\alpha\left|\widehat{\mathbf{V}}_S(B') - \mathbf{V}_S(\widehat{q})\right|}.$$

This Lemma is proved in Appendix 6.3. Together with equation (8), this Lemma ensures that removing enough outliers yields an estimator $\widehat{q}_{B'}$ with estimation guarantee $\sup_{S\subseteq[d]} |\widehat{q}_{B'}(S) - q(S)| \lesssim \alpha\sqrt{\frac{d\ln(1/\alpha)}{k}}$.

The adversarial batch deletion is achieved by identifying the batches Z^b for which $\widehat{\mathrm{Cov}}^{B'}_{S,S'}(b)$ (defined in equation (5)) is at odds with Definition 7 for some $S, S' \subset [d]$. Searching through all possible $S, S' \subseteq [d]$ would yield an exponential-time algorithm. A way around this is to introduce a semi-definite program that can be approximated in polynomial time. To this end, we prove the next Lemma, stating that the quantities $\widehat{\mathrm{Cov}}_{S,S'}(q)$ and $\mathrm{Cov}_{S,S'}(q)$ can be computed as scalar products of matrices.

Lemma 10 (Matrix expression) *Denote by* $\mathbb{1}_S$ *the indicator vector of the elements in* S. *For each vector* q, *there exists a matrix* $\mathbf{C}(\widehat{q})$ *s.t. for any* $S, S' \subseteq [d]$,

$$Cov_{S,S'}(\widehat{q}) = \left\langle \mathbb{1}_S \mathbb{1}_{S'}^T, \mathbf{C}(\widehat{q}) \right\rangle.$$

$$\widehat{Cov}^{B'}_{S,S'}(b) = \left\langle \mathbb{1}_S \mathbb{1}_{S'}^T, \widehat{\mathbf{C}}_{b,B'} \right\rangle \quad and \quad \widehat{Cov}_{S,S'}(B') = \left\langle \mathbb{1}_S \mathbb{1}_{S'}^T, \widehat{\mathbf{C}}(B') \right\rangle,$$

with $\widehat{\mathbf{C}}(B') = \frac{1}{|B'|}\sum_{b\in B'} \widehat{\mathbf{C}}_{b,B'}$ *and* $\widehat{\mathbf{C}}_{b,B'} = (\hat{q}_b - \hat{q}_{B'})(\hat{q}_b - \hat{q}_{B'})^\top$.

The proof of the Lemma and the precise expressions of the matrices can be found in Appendix 6.4. To define the semi-definite program, we introduce the following space of Gram matrices:

$$\mathcal{G} := \left\{ M \in \mathbb{R}^{d\times d}, M_{ij} = \langle u^{(i)}, v^{(j)} \rangle \;\middle|\; (u^{(i)})_{i=1}^d, (v^{(i)})_{j=1}^d \text{unit vectors in } (\mathbb{R}^d, \|\cdot\|_2) \right\}.$$

For a subset B', let us define $D_{B'} = \widehat{\mathbf{C}}(B') - \mathbf{C}(\widehat{q}_{B'})$, and define $M^*_{B'}$ as any matrix s.t.

$$\langle M^*_{B'}, D_{B'} \rangle \geq \max_{M\in\mathcal{G}} \langle M, D_{B'} \rangle - c\frac{\alpha d \ln(e/\alpha)}{k},$$

for some small enough absolute constant $c > 0$.

Remark 11 *Note that the quantity* $\max_{M\in\mathcal{G}} \langle M, D_{B'} \rangle$ *is an SDP. For all desired precision* $\delta > 0$, *it is possible to find the solution of this program up to an additive constant* δ *in polynomial time in all the parameters of the program and in* $\log(1/\delta)$. *Thus,* $M^*_{B'}$ *can be computed in polynomial time, as well as the contamination rate and the corruption score, defined below.*

Definition of the *contamination rate* and *corruption scores*. When $\widehat{q}(S) \gg \lambda|S|$ for some $S \subseteq [d]$, the *contamination rate* and *corruption scores* have special definitions. Formally, let $A = \left\{j \in [d] \,\middle|\, \widehat{q}_{B'}(j) \geq \lambda\right\}$ and $S^* = \arg\max_{S \subseteq [d]} \left|\widehat{q}_{B'}(S) - \lambda|S|\right|$. We have $S^* = A$ or $S^* = [d] \setminus A$, which can be computed in polynomial time. In the special case where $\left|\widehat{q}_{B'}(S^*) - \lambda|S^*|\right| \geq 11$, the *contamination rate* $\sqrt{\tau_{B'}}$ of the collection B' is defined as $\tau_{B'} = \infty$ and the corruption score of a batch is defined as $\varepsilon_b(B') = \left|\widehat{q}_b(S^*) - \lambda|S^*|\right|$.

Otherwise, the *contamination rate* $\sqrt{\tau_{B'}}$ of the collection B' is defined through the quantity satisfying

$$\langle M^*_{B'}, D_{B'}\rangle = \tau_{B'} \frac{\alpha d \ln(e/\alpha)}{k}. \tag{9}$$

Define the *corruption score* of a batch as

$$\varepsilon_b(B') = \langle M^*_{B'}, \widehat{\mathbf{C}}_{b,B'}\rangle. \tag{10}$$

The following Lemma guarantees that the quantity $\langle M^*_{B'}, D_{B'}\rangle$ is a good approximation of $\max_{S,S' \subseteq [d]} \left|\langle \mathbb{1}_S \mathbb{1}_{S'}^T, D_{B'}\rangle\right|$, with the advantage that it can be computed in polynomial time.

Lemma 12 (Grothendieck's inequality corollary) *Assume $d \geq 3$. For all symmetric matrix $A \in \mathbb{R}^{d\times d}$, it holds that*

$$\max_{S,S' \subseteq [d]} \left|\langle \mathbb{1}_S \mathbb{1}_{S'}^T, A\rangle\right| \leq \max_{M \in \mathcal{G}} \langle M, A\rangle \leq 8 \max_{S,S' \subseteq [d]} \left|\langle \mathbb{1}_S \mathbb{1}_{S'}^T, A\rangle\right|.$$

The proof of the Lemma can be found in Appendix 6.5. Together with Lemma 9, this Lemma implies that if conditions 1 and 2 hold, then for any subset B' s.t. $|B'_G| \geq (1-2\alpha)|B_G|$ and for any $S \subset [d]$, we have:

$$|\widehat{q}_{B'}(S) - q(S)| \leq \left(30 + 2\sqrt{\tau_{B'}}\right) \alpha \sqrt{\frac{d \ln(e/\alpha)}{k}}. \tag{11}$$

This Lemma implies that if equation (8) holds, then, for any B' s.t. $|B'_G| \geq (1-2\alpha)|B_G|$

$$\sqrt{\tau_{B'_G}} \leq 200. \tag{12}$$

Lemma 13 (Score good vs. adversarial batches) *If $\sqrt{\tau_{B'}} \geq 200$ and condition 1-2 hold, then for any collection of batches B' s.t. $|B' \cap B_G| \geq (1-2\alpha)|B_G|$, for any sub-collection of good batches $B''_G \subseteq B$, $|B''_G| \leq \alpha n$, we have:*

$$\sum_{b \in B''_G} \varepsilon_b(B') < \frac{1}{8} \sum_{b \in B'_A} \varepsilon_b(B').$$

This Lemma is proved in Appendix 6.6, where we justify that this Lemma ensures that each batch deletion has a probability at least $\frac{3}{4}$ of removing an adversarial batch.

5. Discussion and future work

We studied the problem of estimating discrete distributions in total variation, with both privacy and robustness constraints. We obtained an information theoretic lower bound of $\alpha\sqrt{d/\epsilon^2 k} + \sqrt{d^2/\epsilon^2 kn}$. We proposed an algorithm running in polynomial time and returning an estimated parameter such that the estimation error is within $\sqrt{\log(1/\alpha)}$ of the information theoretic lower bound. It would be interesting to explore if polynomial algorithms could achieve the optimal bound without this extra factor. We do not consider the adaptation to unknown contamination α and leave it for future work. It would also be interesting to explore what happens with contamination occurring before privacy rather than after, as studied in Li et al. (2022) in other settings. Indeed, they do not consider batched data, and it would be interesting to check if their result remains valid in that case. Also, the upper bound holds only if $n \geq O(d)$. Exploring the regime where $n \leq d$ would be an interesting research direction, which has not been done to our knowledge, even in the case of the sole robustness constraint. Finally, we could study the combination of the robustness and privacy constraints in other settings, such as density estimation.

References

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1120–1129. PMLR, 2019.

Jayadev Acharya, Clément L. Canonne, Yuhan Liu, Ziteng Sun, and Himanshu Tyagi. Interactive inference under information constraints, 2020a. URL https://arxiv.org/abs/2007.10976.

Jayadev Acharya, Clément L. Canonne, Ziteng Sun, and Himanshu Tyagi. Unified lower bounds for interactive high-dimensional estimation under information constraints, 2020b. URL https://arxiv.org/abs/2010.06562.

Jayadev Acharya, Peter Kairouz, Yuhan Liu, and Ziteng Sun. Estimating sparse discrete distributions under local privacy and communication constraints, 2020c. URL https://arxiv.org/abs/2011.00083.

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Robust testing and estimation under manipulation attacks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 43–53. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/acharya21a.html.

Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning gaussians and beyond, 2021. URL https://arxiv.org/abs/2111.11320.

Leighton P. Barnes, Yanjun Han, and Ayfer Özgür. Fisher information for distributed estimation under a blackboard communication protocol. In *ISIT*, pages 2704–2708. IEEE, 2019.

Leighton Pate Barnes, Wei-Ning Chen, and Ayfer Özgür. Fisher information under local differential privacy. *IEEE J. Sel. Areas Inf. Theory*, 1(3):645–659, 2020a.

Leighton Pate Barnes, Yanjun Han, and Ayfer Özgür. Lower bounds for learning distributions under communication constraints via fisher information. *J. Mach. Learn. Res.*, 21:Paper No. 236, 30, 2020b. ISSN 1532-4435.

Thomas Berrett and Cristina Butucea. Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms. *Advances in Neural Information Processing Systems*, 33:3164–3173, 2020.

Cristina Butucea, Amandine Dubois, Martin Kroll, and Adrien Saumard. Local differential privacy: Elbow effect in optimal density estimation and adaptation over besov ellipsoids. *Bernoulli*, 26(3):1727–1764, 2020.

T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*, 2019.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

Sitan Chen, Jerry Li, and Ankur Moitra. Efficiently learning structured distributions from untrusted batches. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 960–973, 2020.

Albert Cheu, Adam Smith, and Jonathan Ullman. Manipulation attacks in local differential privacy, 2019. URL https://arxiv.org/abs/1909.09630.

Albert Cheu, Adam Smith, and Jonathan Ullman. Manipulation attacks in local differential privacy. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 883–900. IEEE, 2021.

Arnak S Dalalyan and Arshak Minasyan. All-in-one robust estimator of the gaussian mean. *arXiv preprint arXiv:2002.01432*, 2020.

Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability, 2019.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy, data processing inequalities, and statistical minimax rates, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, page 211–222, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581136706. doi: 10.1145/773153.773174. URL https://doi.org/10.1145/773153.773174.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Transactions on Information Theory*, 61(11):6343–6354, 2015.

Samuel B. Hopkins, Gautam Kamath, and Mahbod Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism, 2021. URL https://arxiv.org/abs/2111.12981.

Peter J Huber. Robust confidence limits. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 10 (4):269–278, 1968.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

Peter J Huber and EM Ronchetti. Robust statistics. 2nd john wiley & sons. *Hoboken, NJ*, 2, 2009.

Ayush Jain and Alon Orlitsky. Optimal robust learning of discrete distributions from batches, 2020.

Ayush Jain and Alon Orlitsky. Robust density estimation from batches: The best things in life are (nearly) free. In *International Conference on Machine Learning*, pages 4698–4708. PMLR, 2021.

Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy, 2016.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? 2008. doi: 10.48550/ARXIV.0803.0924. URL https://arxiv.org/abs/0803.0924.

Joseph Lam-Weil, Béatrice Laurent, and Jean-Michel Loubes. Minimax optimal goodness-of-fit testing for densities under a local differential privacy constraint. 2020.

Guillaume Lecué, Matthieu Lerasle, and Timlothée Mathieu. Robust classification via mom minimization. *Machine Learning*, 109(8):1635–1665, 2020.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy, 2019.

Mengchu Li, Thomas B Berrett, and Yi Yu. On robustness and local differential privacy. *arXiv preprint arXiv:2201.00751*, 2022.

Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. *Advances in Neural Information Processing Systems*, 34, 2021a.

Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions, 2021b. URL https://arxiv.org/abs/2111.06578.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Mohammad Naseri, Jamie Hayes, and Emiliano De Cristofaro. Toward robustness and privacy in federated learning: Experimenting with local and central differential privacy. *arXiv e-prints*, pages arXiv–2009, 2020.

Rafael Pinot, Florian Yger, Cédric Gouy-Pailler, and Jamal Atif. A unified view on differential privacy and robustness to adversarial examples, 2019.

Mingda Qiao and Gregory Valiant. Learning discrete distributions from untrusted batches. *arXiv preprint arXiv:1711.08113*, 2017.

Peter J Rousseeuw and Mia Hubert. Robust statistics for outlier detection. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 1(1):73–79, 2011.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under locally differential privacy, 2017. URL https://arxiv.org/abs/1702.00610.

6. Appendix

6.1. Proof of Lemma 14, Law of the sum

Lemma 14 (Law of the sum) *For any subset $S \subseteq [d]$, we have:*

$$\sum_{j\in S} Z(j) \sim \sum_{j=1}^{|S|-1} b_j + b^S,$$

with the $(b_j)_{j=1}^{|S|-1}$ independent Bernoulli variables s.t. $\mathrm{P}(b_j = 1) = \lambda$ and b^S a Bernoulli independent of the others s.t.

$$\mathrm{P}(b^S = 1) = \lambda + (1-2\lambda)p_S.$$

For any $t \in [d]$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{j\in S} Z(j) = t\right) &= \binom{|S|-1}{t-1}(1-\lambda)^{|S|-t+1}\lambda^{t-1}p(S) + \big(1-p(S)\big)\binom{|S|}{t}(1-\lambda)^{|S|-t}\lambda^t \\
&\quad + \binom{|S|-1}{t}(1-\lambda)^{|S|-t-1}\lambda^{t+1}p(S) \\
&= \binom{|S|-1}{t-1}(1-\lambda)^{|S|-t}\lambda^{t-1}\Big[(1-\lambda)p(S) + \lambda\big(1-p(S)\big)\Big] \\
&\quad + \binom{|S|-1}{t}(1-\lambda)^{|S|-t-1}\lambda^{t}\Big[(1-\lambda)\big(1-p(S)\big) + \lambda p(S)\Big] \\
&= \binom{|S|-1}{t-1}(1-\lambda)^{|S|-t}\lambda^{t-1}\big[(1-2\lambda)p(S) + \lambda\big] \\
&\quad + \binom{|S|-1}{t}(1-\lambda)^{|S|-t-1}\lambda^{t}\big[1-\lambda-(1-2\lambda)p(S)\big].
\end{aligned}$$

Note that we have:

$$q(S) = (1-2\lambda)\,p(S) + \lambda|S|. \tag{13}$$

6.2. Proof of Lemma 8, Essential properties of good batches

We start with the following intermediary Lemma.

Lemma 15 *If* $\left|B_G\right| \geq \frac{2d}{\alpha^2 \ln(e/\alpha)}$, *then* $\forall S \subseteq [d]$ *and* $\forall B'_G \subseteq B_G$ *of size* $\left|B'_G\right| \geq (1-2\alpha)\left|B_G\right|$, *with probability at least* $1-4e^{-d}$,

$$\left|\widehat{q}_{B'_G}(S) - q(S)\right| \leq 6\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}}.$$

Proof : The proof of this lemma is exactly part of that of lemma 11 in Jain and Orlitsky (2020) with different constants, we repeat it for completeness. From Hoeffding's inequality, for any $S \subseteq [d]$,

$$\mathbb{P}\left[\left|B_G\right|\left|\widehat{q}_{B_G}(S) - q(S)\right| \geq \frac{\left|B_G\right|}{\sqrt{2}}\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}}\right] \leq 2e^{-\alpha^2\left|B_G\right|\ln(e/\alpha)} \leq 2e^{-2d}.$$

Similarly, for a fixed sub-collection $U_G \subseteq B_G$ of size $1 \leq \left|U_G\right| \leq 2\alpha\left|B_G\right|$,

$$\mathbb{P}\left[\left|U_G\right| \cdot \left|\widehat{q}_{U_G}(S) - q(S)\right| \geq 2\alpha\left|B_G\right|\sqrt{\frac{d\ln(e/\alpha)}{k}}\right] \leq 2e^{-8\frac{\alpha^2|B_G|^2}{|U_G|}\ln(e/\alpha)} \leq 2e^{-4\alpha|B_G|\ln(e/\alpha)}. \tag{14}$$

We now bound the number of subsets of cardinality smaller than $2\alpha\left|B_G\right|$:

$$\begin{aligned}\sum_{j=1}^{\lfloor 2\alpha|B_G|\rfloor}\binom{|B_G|}{j} &\leq 2\alpha\left|B_G\right|\binom{|B_G|}{\lfloor 2\alpha|B_G|\rfloor} \leq 2\alpha\left|B_G\right|\left(\frac{e\left|B_G\right|}{2\alpha\left|B_G\right|}\right)^{2\alpha|B_G|} \\ &\leq e^{2\alpha|B_G|\ln(e/\alpha)+\ln\left(2\alpha|B_G|\right)} < e^{3\alpha|B_G|\ln(e/\alpha)}. \end{aligned} \tag{15}$$

Thus, by union bound,

$$\mathbb{P}\left[\exists\left|U_G\right| \leq 2\alpha\left|B_G\right| : \left|U_G\right|\left|\widehat{q}_{U_G}(S) - q(S)\right| \geq 2\alpha\left|B_G\right|\sqrt{\frac{d\ln(e/2\alpha)}{k}}\right] \leq 2e^{-\alpha|B_G|\ln(e/\alpha)} \leq 2e^{-2d}.$$

For any sub-collection $B'_G \subseteq B_G$ with $\left|B'_G\right| \geq (1-2\alpha)\left|B_G\right|$,

$$\begin{aligned}\left|\sum_{b\in B'_G}\left[\widehat{q}_b(S) - q(S)\right]\right| &= \left|\sum_{b\in B_G}\left[\widehat{q}_b(S) - q(S)\right] - \sum_{b\in B_G\setminus B'_G}\left[\widehat{q}_b(S) - q(S)\right]\right| \\ &\leq \left|\sum_{b\in B_G}\left[\widehat{q}_b(S) - q(S)\right]\right| + \left|\sum_{b\in B_G\setminus B'_G}\left[\widehat{q}_b(S) - q(S)\right]\right| \\ &\leq \left|B_G\right| \times \left|\widehat{q}_{B_G}(S) - q(S)\right| + \max_{\substack{U_G \text{ s.t.} \\ \left|U_G\right|\leq 2\alpha\left|B_G\right|}} \left|U_G\right| \times \left|\widehat{q}_{U_G}(S) - q(S)\right| \\ &\leq \left(2 + \frac{1}{\sqrt{2}}\right)\alpha\left|B_G\right|\sqrt{\frac{d\ln(e/\alpha)}{k}}. \end{aligned}$$

where the last inequality holds with probability at least $1 - 4e^{-2d}$. We conclude by using a union bound over the 2^d possible subsets and by noting that $(2 + \frac{1}{\sqrt{2}})\frac{\left|B_G\right|}{\left|B'_G\right|} \leq 6$. ■

We now move to the following result.

Lemma 16 *If* $\left|B_G\right| \geq \frac{3d}{\alpha^2 \ln(e/\alpha)}$, *then* $\forall S, S' \subseteq [d]$ *and* $\forall B'_G \subseteq B_G$ *of size* $\left|B'_G\right| \geq (1-2\alpha)\left|B_G\right|$, *with probability at least* $1 - 2e^{-d}$,

$$\left| \frac{1}{|B'_G|} \sum_{b \in B'_G} \left[\widehat{q}_b(S) - q(S)\right]\left[\widehat{q}_b(S') - q(S')\right] - Cov_{S,S'}(q) \right| \leq \frac{140 d\alpha \ln\left(\frac{e}{\alpha}\right)}{k}.$$

Proof : Let $U_b(S,S') = \left(\frac{\widehat{q}_b(S)-q(S)}{d}\right)\left(\frac{\widehat{q}_b(S')-q(S')}{d}\right) - \frac{\mathrm{Cov}_{S,S'}(q)}{d^2}$. For $b \in B_G$, $\frac{\widehat{q}_b(S)-q(S)}{d} \sim \mathrm{subG}(1/4dk)$, therefore

$$\left(\frac{\widehat{q}_b(S) - q(S)}{d}\right)\left(\frac{\widehat{q}_b(S') - q(S')}{d}\right) - \mathbb{E}\left[\left(\frac{\widehat{q}_b(S) - q(S)}{d}\right)\left(\frac{\widehat{q}_b(S') - q(S')}{d}\right)\right] = Y_b \sim \mathrm{subE}\left(\frac{16}{4kd}\right).$$

Here subE is sub exponential distribution. For any $S, S' \subseteq [d]$, Bernstein's inequality gives:

$$\mathbb{P}\left[\left|\sum_{b \in B_G} U_b(S,S')\right| \geq 6\alpha |B_G| \frac{\ln(e/\alpha)}{kd}\right] \leq 2e^{-\alpha^2 \left|B_G\right| \ln^2(e/\alpha)} \leq 2e^{-3d}.$$

Next, for a fixed sub-collection $B''_G \subseteq B_G$ of size $1 \leq \left|B''_G\right| \leq \alpha\left|B_G\right|$,

$$\begin{aligned}\Pr\left[\left|\sum_{b \in B''_G} U_b(S,S')\right| \geq 64\alpha\left|B_G\right|\frac{\ln(e/\alpha)}{n}\right] &\leq 2e^{-\frac{64\alpha|B_G|\ln(e/\alpha)}{2\times 2\times 4/n}} \\ &\leq 2e^{-4\alpha|B_G|\ln(e/\alpha)}.\end{aligned}$$

The same steps as the previous lemma terminate the proof, except that there are now 2^{2d} sets $S, S' \subseteq [d]$. ■

By Lemma 15 and 16, if $\left|B_G\right| \geq \frac{2d}{\alpha^2 \ln(e/\alpha)}$, then $\forall S \subseteq [d]$ and $\forall B'_G \subseteq B_G$ of size $\left|B'_G\right| \geq (1-2\alpha)\left|B_G\right|$, with probability at least $1 - 8e^{-d}$:

$$\left|\widehat{q}_{B'_G}(S) - q(S)\right| \leq 6\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}}$$

and

$$\left| \frac{1}{|B'_G|} \sum_{b \in B'_G} \left[\widehat{q}_b(S) - q(S)\right]\left[\widehat{q}_b(S') - q(S')\right] - \mathrm{Cov}_{S,S'}(q) \right| \leq \frac{140 d\alpha \ln\left(\frac{6e}{\alpha}\right)}{k}.$$

Additionally Lemma 19, this implies:

$$\left|\mathrm{Cov}_{S,S'}(q) - \mathrm{Cov}_{S,S'}(\widehat{q}_{B'_G})\right| \leq 66\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}}.$$

Moreover:

$$\begin{aligned}
\frac{1}{|B'_G|}\sum_{b\in B'_G}\Big[\widehat{q}_b(S)-q(S)\Big]\Big[\widehat{q}_b(S')-q(S')\Big] = {} & \frac{1}{|B'_G|}\sum_{b\in B'_G}\Big[\widehat{q}_b(S)-\widehat{q}_{B'}(S)\Big]\Big[\widehat{q}_b(S')-\widehat{q}_{B'}(S')\Big] \\
& + \Big[q(S)-\widehat{q}_{B'}(S)\Big]\Big[q(S')-\widehat{q}_{B'}(S')\Big] \\
& + \frac{1}{|B'_G|}\sum_{b\in B'_G}\Big[\widehat{q}_b(S)-\widehat{q}_{B'}(S)\Big]\Big[q(S')-\widehat{q}_{B'}(S')\Big] \\
& + \frac{1}{|B'_G|}\sum_{b\in B'_G}\Big[q(S)-\widehat{q}_{B'}(S)\Big]\Big[\widehat{q}_b(S')-\widehat{q}_{B'}(S')\Big] \\
= {} & \frac{1}{|B'_G|}\sum_{b\in B'_G}\Big[\widehat{q}_b(S)-\widehat{q}_{B'}(S)\Big]\Big[\widehat{q}_b(S')-\widehat{q}_{B'}(S')\Big] \\
& + \Big[q(S)-\widehat{q}_{B'}(S)\Big]\Big[q(S')-\widehat{q}_{B'}(S')\Big].
\end{aligned}$$

Therefore:

$$\left|\widehat{\mathrm{Cov}}_{S,S'}\left(B'_G\right) - \mathrm{Cov}_{S,S'}(\widehat{q}_{B'_G})\right| \leq \frac{242d\alpha\ln\left(\frac{e}{\alpha}\right)}{k}.$$

Note that we also have:

$$\left|\widehat{\mathrm{Cov}}_{S,S'}\left(B'_G\right) - \mathrm{Cov}_{S,S'}(q)\right| \leq \frac{176d\alpha\ln\left(\frac{e}{\alpha}\right)}{k}. \tag{16}$$

The following Lemma gives condition 2.

Lemma 17 *If* $\left|B_G\right| \geq \frac{3d}{\alpha^2\ln(e/\alpha)}$, *then* $\forall S, S' \subseteq [d]$ *and* $\forall B''_G \subseteq B_G$ *of size* $\left|B''_G\right| \leq \alpha\left|B_G\right|$, *with probability at least* $1-2e^{-d}$,

$$\left|\sum_{b\in B''_G}\Big[\widehat{q}_b(S)-q(S)\Big]\Big[\widehat{q}_b(S')-q(S')\Big]\right| \leq \frac{33\alpha d\left|B_G\right|\ln(e/\alpha)}{k}.$$

Proof : For any $S, S' \subseteq [d]$ and any $B'_G \subseteq B_G$ Bernstein's inequality gives:

$$\mathbb{P}\left[\left|\sum_{b\in B''_G} U_b(S,S')\right| \geq 32\alpha\left|B_G\right|\frac{\ln(e/\alpha)}{kd}\right] \leq 2e^{-4\alpha\left|B_G\right|\ln(e/\alpha)}.$$

We have :

$$\left|\sum_{b\in B''_G}\Big[\widehat{q}_b(S)-q(S)\Big]\Big[\widehat{q}_b(S')-q(S')\Big]\right| = \left|\sum_{b\in B''_G} d^2U_b(S,S') + |B''_G|\mathrm{Cov}_{S,S'}(q)\right|$$

$$\leq \left| \sum_{b \in B''_G} d^2 U_b(S, S') \right| + \alpha \frac{d|B_G|}{k}.$$

A union bound over all the possible B''_G and the 2^{2d} sets S, S' terminates the proof. ■

Combining the three Lemmas of the section gives Lemma 8.

6.3. Proof of Lemma 9, Variance gap to estimation error

Proof: By condition 1 and Cauchy-Schwartz:

$$\begin{aligned}\left|\widehat{q}_{B'}(S) - q(S)\right| &\leq \frac{1}{|B'|}\left|\sum_{b \in B'_G} \widehat{q}_b(S) - q(S)\right| + \frac{1}{|B'|}\left|\sum_{b \in B'_A} \widehat{q}_b(S) - q(S)\right| \\ &\leq 6\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}} + \sqrt{\frac{|B'_A|}{|B'|}}\sqrt{\frac{1}{|B'|}\sum_{b \in B'_A}\left[\widehat{q}_b(S) - q(S)\right]^2}. \end{aligned} \tag{17}$$

We can decompose the second term:

$$\frac{1}{|B'|}\sum_{b \in B'_A}\left[\widehat{q}_b(S) - q(S)\right]^2 = \frac{1}{|B'|}\sum_{b \in B'}\left[\widehat{q}_b(S) - q(S)\right]^2 - \frac{1}{|B'|}\sum_{b \in B'_G}\left[\widehat{q}_b(S) - q(S)\right]^2.$$

By Lemma 16,

$$\left|\frac{1}{|B'_G|}\sum_{b \in B'_G}\left[\widehat{q}_b(S) - q(S)\right]^2 - \mathbf{V}_S(q)\right| \leq 140\frac{\alpha d\ln(e/\alpha)}{k}.$$

Thus,

$$\begin{aligned}\frac{1}{|B'|}\sum_{b \in B'_G}\left[\widehat{q}_b(S) - q(S)\right]^2 &= \frac{|B'_G|}{|B'|}\frac{1}{|B'_G|}\sum_{b \in B'_G}\left[\widehat{q}_b(S) - q(S)\right]^2 \\ &\geq (1-2\alpha)\left(\mathbf{V}_S(q) - 140\frac{\alpha d\ln(e/\alpha)}{k}\right) \\ &\geq \mathbf{V}_S(q) - 2\alpha\mathbf{V}_S(q) - 140\frac{\alpha d\ln(e/\alpha)}{k} \\ &\geq \mathbf{V}_S(\widehat{q}_{B'}) - 15\frac{\left|\widehat{q}_{B'}(S) - q(S)\right|}{k} - 142\frac{\alpha d\ln(e/\alpha)}{k},\end{aligned}$$

where the last inequality comes from Lemma 19 and $\mathbf{V}_S(q) \leq d/k$. Now, we have

$$\left|\widehat{q}_{B'}(S) - \widehat{q}_{B'_G}(S)\right| \leq \left|\left(\frac{1}{|B'_G|} - \frac{1}{|B'|}\right)\sum_{b \in B'_G}\widehat{q}_b(S)\right| + \left|\frac{1}{|B'|}\sum_{b \in B' \setminus B'_G}\widehat{q}_b(S)\right|$$

$$\leq \frac{2d\alpha}{1-\alpha} \leq 3d\alpha.$$

Thus,

$$\begin{aligned}\frac{\left|\widehat{q}_{B'}(S) - q(S)\right|}{k} &\leq \frac{\left|\widehat{q}_{B'_G}(S) - q(S)\right|}{k} + \frac{\left|\widehat{q}_{B'}(S) - \widehat{q}_{B'_G}(S)\right|}{k} \\ &\leq \frac{6\alpha}{k}\sqrt{\frac{d\ln(e/\alpha)}{k}} + \frac{3d\alpha}{k} \leq 3\frac{d\alpha}{k}\ln(e/\alpha).\end{aligned}$$

This implies

$$\frac{1}{|B'|}\sum_{b\in B'_G}\left[\widehat{q}_b(S) - q(S)\right]^2 \geq \mathbf{V}_S(\widehat{q}_{B'}) - 187\frac{\alpha d\ln(e/\alpha)}{k}. \tag{18}$$

On the other hand,

$$\begin{aligned}\frac{1}{|B'|}\sum_{b\in B'}\left[\widehat{q}_b(S) - q(S)\right]^2 &= \frac{1}{|B'|}\sum_{b\in B'}\left[\widehat{q}_b(S) - \widehat{q}_{B'}(S)\right]^2 \\ &\quad + \left[q(S) - \widehat{q}_{B'}(S)\right]^2 + 2\Big[q(S) - \widehat{q}_{B'}(S)\Big]\frac{1}{|B'|}\sum_{b\in B'}\Big[\widehat{q}_b(S) - \widehat{q}_{B'}(S)\Big] \\ &= \frac{1}{|B'|}\sum_{b\in B'}\left[\widehat{q}_b(S) - \widehat{q}_{B'}(S)\right]^2 + \left[q(S) - \widehat{q}_{B'}(S)\right]^2.\end{aligned}$$

Combining this equation with equations 18 and 17 gives

$$\begin{aligned}\left|\widehat{q}_{B'}(S) - q(S)\right| &\leq \frac{1}{|B'|}\left|\sum_{b\in B'_G}\widehat{q}_b(S) - q(S)\right| + \frac{1}{|B'|}\left|\sum_{b\in B'_A}\widehat{q}_b(S) - q(S)\right| \\ &\leq 6\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}} + \sqrt{2\alpha}\sqrt{\widehat{\mathbf{V}}_S(B') - \mathbf{V}_S(\widehat{q}_{B'}) + 187\frac{\alpha d\ln(e/\alpha)}{k} + \left[q(S) - \widehat{q}_{B'}(S)\right]^2} \\ &\leq 26\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}} + \sqrt{2\alpha\left|\widehat{\mathbf{V}}_S(B') - \mathbf{V}_S(\widehat{q}_{B'})\right|} + \sqrt{2\alpha}\left|\widehat{q}_{B'}(S) - q(S)\right|.\end{aligned}$$

Noting that $2\alpha \leq 1/8$ terminates the proof:

$$\left|\widehat{q}_{B'}(S) - q(S)\right| \leq 30\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}} + 2\sqrt{\alpha\left|\widehat{\mathbf{V}}_S(B') - \mathbf{V}_S(\widehat{q}_{B'})\right|}.$$

□

6.4. Proof of Lemma 10, Matrix expression

For each batch $b \in B$, define matrix $C^{EV}_{b,B'}$ as:

$$\widehat{\mathbf{C}}_{b,B'}(j,l) = \left[\widehat{q}_b(j) - \widehat{q}_{B'}(j)\right]\left[\widehat{q}_b(l) - \widehat{q}_{B'}(l)\right],\ \forall (j,l) \in [d]^2. \tag{19}$$

For each collection of batches B' define

$$\widehat{\mathbf{C}}(B') = \frac{1}{|B'|}\sum_{b\in B'}\widehat{\mathbf{C}}_{b,B'}.$$

For a set $S \subseteq [d]$, define $\mathbb{1}_S$ as the indicator vector of the elements in S. For any $S, S' \subseteq [d]$

$$\begin{aligned}\left\langle \widehat{\mathbf{C}}(B'), \mathbb{1}_S\mathbb{1}_{S'}^T\right\rangle &= \frac{1}{|B'|}\sum_{b\in B'}\sum_{j\in S}\sum_{l\in S'}\left[\widehat{q}_b(j) - \widehat{q}_{B'}(j)\right]\left[\widehat{q}_b(l) - \widehat{q}_{B'}(l)\right]\\ &= \frac{1}{|B'|}\sum_{b\in B'}\left(\sum_{j\in S}\widehat{q}_b(j) - \sum_{j\in S}\widehat{q}_{B'}(j)\right)\left(\sum_{l\in S'}\widehat{q}_b(l) - \sum_{l\in S'}\widehat{q}_{B'}(l)\right)\\ &= \widehat{\mathrm{Cov}}_{S,S'}\left(B'\right).\end{aligned}$$

We can compute

$$\begin{aligned}\mathbb{E}\left[\sum_{j\in S}Z(j)\middle|X\right] &= \lambda|S|\mathbb{1}_{X\notin S} + \left(\lambda(|S|-1)+1-\lambda\right)\mathbb{1}_{X\in S}\\ &= \lambda|S| + (1-2\lambda)\,\mathbb{1}_{X\in S}.\end{aligned}$$

For a set S, let us define $Y_S = \left(\sum_{j\in S}Z(j)\right) - q(S)$ and $\Delta_S = \lambda|S| - q(S)$. For any sets $S, S' \subseteq [d]$ s.t. $S\cap S' = \emptyset$, we have:

$$\begin{aligned}\mathbb{E}\left[Y_SY_{S'}\right] &= \mathbb{E}\left[\mathbb{E}\left[Y_S\middle|X\right]\mathbb{E}\left[Y_{S'}\middle|X\right]\right]\\ &= \mathbb{E}\left[\left(\Delta_S + (1-2\lambda)\,\mathbb{1}_{X\in S}\right)\left(\Delta_{S'} + (1-2\lambda)\,\mathbb{1}_{X\in S'}\right)\right]\\ &= \Delta_{S'}\Delta_S + \Delta_S(1-2\lambda)p(S') + \Delta_{S'}(1-2\lambda)p(S) \quad \text{since } S\cap S' = \emptyset\\ &= -\Delta_{S'}\Delta_S \quad \text{since by (13) we have } (1-2\lambda)p(S) = -\Delta_S.\end{aligned}$$

On the other hand, using the notation from Lemma 14, we have:

$$\begin{aligned}\mathbb{E}\left[Y_S^2\right] &= \mathbb{E}\left[\left(\sum_{j=1}^{|S|-1}(b_j - \mathbb{E}b_j) + b^S - \mathbb{E}b^S\right)^2\right] = \sum_{j=1}^{|S|-1}\mathbb{V}[b_j] + \mathbb{V}[b^S]\\ &= (|S|-1)\lambda(1-\lambda) + \left(\lambda + (1-2\lambda)p(S)\right)\left(1-\lambda-(1-2\lambda)p(S)\right)\\ &= (|S|-1)\lambda(1-\lambda) + (\lambda-\Delta_S)\,(1-\lambda+\Delta_S)\\ &= -\Delta_S^2 + |S|\lambda(1-\lambda) - (1-2\lambda)\Delta_S.\end{aligned}$$

For any $S, S' \subseteq [d]$, we thus have:

$$\begin{aligned}\mathbb{E}\left[Y_SY_{S'}\right] &= \mathbb{E}\left[\left(Y_{(S\cap S')} + Y_{(S\setminus S')}\right)\left(Y_{(S\cap S')} + Y_{(S'\setminus S)}\right)\right]\\ &= \mathbb{E}\left[Y_{(S\cap S')}^2\right] + \mathbb{E}\left[Y_{(S\cap S')}Y_{(S\setminus S')}\right] + \mathbb{E}\left[Y_{(S\cap S')}Y_{(S'\setminus S)}\right] + \mathbb{E}\left[Y_{(S\setminus S')}Y_{(S'\setminus S)}\right]\end{aligned}$$

$$= -\left(\Delta_{(S\cap S')} + \Delta_{(S\setminus S')}\right)\left(\Delta_{(S\cap S')} + \Delta_{(S'\setminus S)}\right) + |S\cap S'|\lambda(1-\lambda) - (1-2\lambda)\Delta_{(S\cap S')}$$
$$= -\Delta_S\Delta_{S'} + |S\cap S'|\lambda(1-\lambda) - (1-2\lambda)\Delta_{(S\cap S')}.$$

For a vector q, define

$$k\mathbf{C}(q) = -\left(\lambda\mathbb{1} - q\right)\left(\lambda\mathbb{1} - q\right)^T + \lambda(1-\lambda)I_d - (1-2\lambda)\mathrm{Diag}(\lambda\mathbb{1} - q). \tag{20}$$

For any two sets $S, S' \subseteq [d]$, we have: $\mathbb{E}\left[Y_S Y_{S'}\right] = \mathbb{1}_S^T k\mathbf{C}(q)\mathbb{1}_{S'}$, so that $\mathbb{E}\left[\widehat{\mathrm{Cov}}_{S,S'}(B'_G)\right] = \mathbb{1}_S^T\mathbf{C}(q)\mathbb{1}_{S'}$. We now define:

$$\mathrm{Cov}_{S,S'}(B') := \mathbb{1}_S^T\mathbf{C}(\widehat{q}_{B'})\mathbb{1}_{S'}. \tag{21}$$

6.5. Proof of Lemma 12, Grothendieck's inequality corollary

Proof [Proof of Lemma 12]

- For the first inequality, fix any $x, y \in \{0,1\}^d$ and three orthonormal vectors $e_0, e_1, e_2 \in \mathbb{R}^d$. Define the following vectors:

$$\forall j \in \{1,\dots,d\} : u^{(j)} = \begin{cases} e_0 & \text{if } x_j = 1, \\ e_1 & \text{otherwise,} \end{cases} \qquad \text{and} \quad v^{(j)} = \begin{cases} e_0 & \text{if } y_j = 1, \\ e_2 & \text{otherwise.} \end{cases}$$

  Then the matrix $M = \left[\langle u^{(i)}, v^{(j)}\rangle\right]_{ij}$ belongs to $\mathcal{G}$ and we have by construction $M = xy^T$ which proves the first inequality.

- For the second inequality, we have by Grothendieck's inequality

$$\max_{M\in\mathcal{G}}\langle M, A\rangle \le 2\max_{x,y\in\{\pm1\}^d}\langle xy^T, A\rangle.$$

  For all $a \in \mathbb{R}$, define $a^+ = a \vee 0$ and $a^- = (-a)\vee 0$ and for all vector $x \in \mathbb{R}^d$, define $x^+ = (x_j^+)_j$ and $x^- = (x_j^-)_j$. Note that if $x \in \{\pm1\}^d$, then $x^+, x^- \in \{0,1\}^d$. We therefore have:

$$\max_{x,y\in\{\pm1\}^d}\langle xy^T, A\rangle = \max_{x,y\in\{\pm1\}^d}\left|\langle x^+y^{+T}, A\rangle - \langle x^-y^{+T}, A\rangle - \langle x^+y^{-T}, A\rangle + \langle x^-y^{-T}, A\rangle\right|$$
$$\le 4\max_{a,b\in\{0,1\}^d}\left|\langle ab^T, A\rangle\right|,$$

  which proves the second inequality.

■

6.6. Proof of Lemma 13, Score good vs. adversarial batches

We first note that the Lemma implies the desired property for the batches in B^o, namely that each batch deletion has a probability at least $\frac{3}{4}$ of removing an adversarial batch. Indeed, we have:

$$\sum_{b\in B^o}\varepsilon_b = \sum_{b\in B^o_G}\varepsilon_b + \sum_{b\in B^o_A}\varepsilon_b.$$

If we had $\sum_{b\in B^o_A}\varepsilon_b < 7\sum_{b\in B^o_G}\varepsilon_b$, this would imply:

$$\sum_{b\in B^o}\varepsilon_b < 8\sum_{b\in B^o_G}\varepsilon_b < \sum_{b\in B'_A}\varepsilon_b,$$

where the last inequality comes from the Lemma. However this is in contradiction with the definition of B^o, which is the sub-collection of $\alpha|B|$ batches with top ε_b scores, since $|B'_A| \le \alpha|B|$. We therefore have that $\sum_{b\in B^o_A}\varepsilon_b \ge 7\sum_{b\in B^o_G}\varepsilon_b$ hence $\sum_{b\in B^o_G}\varepsilon_b \le \frac{1}{8}\sum_{b\in B^o}\varepsilon_b$. Denote by $B^o(t)$ the current set obtained from B^o after having removed t batches (and before $\sum_{b\in B^o}\varepsilon_b$ has been halved). We keep deleting batches from B^o until $\sum_{b\in B^o(t)}\varepsilon_b \le \frac{1}{2}\sum_{b\in B^o}\varepsilon_b$. At each step, we therefore have that $\sum_{b\in B^o_G(t)}\varepsilon_b \le \frac{1}{4}\sum_{b\in B^o(t)}\varepsilon_b$ hence the probability of removing a good batch from $B^o(t)$ is always less than $\frac{1}{4}$.

Subcase 1 We first prove the Lemma in the case where $\max_{S\subseteq[d]}\left|\widehat{q}_{B'}(S) - \lambda|S|\right| \ge 11$.
We have:

$$\begin{aligned}
|\widehat{q}_{B'}(S) - \lambda|S|| &\le \frac{1}{|B'|}\left|\sum_{b\in B'_G}\widehat{q}_b(S) - \lambda|S|\right| + \frac{1}{|B'|}\left|\sum_{b\in B'_A}\widehat{q}_b(S) - \lambda|S|\right| \\
&\le \frac{|B'_G|}{|B'|}\frac{1}{|B'_G|}\left|\sum_{b\in B'_G}\widehat{q}_b(S) - q(S)\right| + \frac{|B'_G|}{|B_G|}\left|\lambda|S| - q(S)\right| + \frac{1}{|B'|}\left|\sum_{b\in B'_A}\widehat{q}_b(S) - \lambda|S|\right| \\
&\le 6\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}} + 1 + \frac{1}{|B'|}\left|\sum_{b\in B'_A}\widehat{q}_b(S) - \lambda|S|\right| \quad \text{by equation (7).}
\end{aligned}$$

Let $S^* = \arg\max_{S\subseteq[d]}|\widehat{q}_{B'}(S) - \lambda|S||$. We have:

$$\left|\sum_{b\in B'_A}\widehat{q}_b(S^*) - \lambda|S^*|\right| \ge 9(1-2\alpha)\big|B_G\big|.$$

On the other hand, by equation 14, we have for any B''_G s.t. $|B''_G| \le \alpha\big|B_G\big|$:

$$\left|\sum_{b\in B''_G}\widehat{q}_b(S^*) - \lambda|S^*|\right| \le \left|\sum_{b\in B''_G}\widehat{q}_b(S^*) - q(S)\right| + |B''_G|$$

$$\leq \Big(\alpha + 2\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}}\Big)\big|B_G\big|$$
$$\leq \big(1+\alpha\big)\big|B_G\big|.$$

Thus we have:

$$\frac{\Big|\sum_{b\in B'_A}\widehat{q}_b(S^*) - \lambda|S^*|\Big|}{\Big|\sum_{b\in B''_G}\widehat{q}_b(S^*) - \lambda|S^*|\Big|} \geq \frac{9(1-2\alpha)}{1+\alpha} > 8.$$

Subcase 2 In the case where $\max_{S\subseteq[d]}\big|\widehat{q}_{B'}(S) - \lambda|S|\big| \leq 11$, the proof relies on the following intermediary Lemma 18.

Lemma 18 *If conditions 1 and 2 hold, then, for any $B' \subset [B]$, for any two sets S, S':*

$$\big(\tau_{B'} - 11\sqrt{\tau_{B'}} - 1313\big)\frac{\alpha d\ln(e/\alpha)}{k} \leq \frac{1}{|B'|}\sum_{b\in B'_A}\langle M^*, \widehat{\mathbf{C}}_{b,B'}\rangle.$$

Proof : In this proof only, we use the shorthand:

$$\gamma := \frac{\alpha d\ln(e/\alpha)}{k}.$$

We have

$$\big\langle M^*, D_{B'}\big\rangle = \Big\langle M^*, \widehat{\mathbf{C}}(B') - \mathbf{C}(q)\Big\rangle + \Big\langle M^*, \mathbf{C}(q) - \mathbf{C}(\widehat{q}_{B'})\Big\rangle.$$

We analyse separately each term. For any S', S, according to Lemmas 19 and equation 11, we have:

$$\Big|\mathrm{Cov}_{S,S'}(\widehat{q}_{B'}) - \mathrm{Cov}_{S,S'}(q)\Big| \leq \frac{11}{k}\max_{S''}\Big|\widehat{q}_{B'}(S'') - q(S'')\Big|$$
$$\leq \frac{330 + 22\sqrt{\tau_{B'}}}{k}\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}}$$
$$\leq (330 + 22\sqrt{\tau_{B'}})\gamma\sqrt{\frac{1}{d\ln(e/\alpha)k}}$$
$$\leq (96 + 7\sqrt{\tau_{B'}})\gamma.$$

Where the last line come from $d \geq 3$, $\alpha \leq \frac{1}{20}$. Thus, by Lemma 12, we have:

$$\arg\max_{M\in\mathcal{G}}\langle M, \mathbf{C}(q) - \mathbf{C}(\widehat{q}_{B'})\rangle \leq 8\max_{S,S'}\Big|\mathrm{Cov}_{S,S'}(\widehat{q}_{B'}) - \mathrm{Cov}_{S,S'}(q)\Big|$$
$$\leq \frac{88}{k}\Big(30 + 2\sqrt{\tau'_B}\Big)\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}}$$
$$\leq (763 + 51\sqrt{\tau_{B'}})\gamma. \tag{22}$$

On the other hand,

$$\widehat{\mathbf{C}}(B') - \mathbf{C}(q) = \frac{1}{|B'|}\sum_{b\in B'}\widehat{\mathbf{C}}(b, B') - \mathbf{C}(q)$$

$$= \frac{1}{|B'|}\sum_{b\in B'_G}\widehat{\mathbf{C}}(b,B') - \mathbf{C}(q) + \frac{1}{|B'|}\sum_{b\in B'_A}\widehat{\mathbf{C}}(b,B') - \mathbf{C}(q).$$

From Lemma 12 and 16, we have:

$$\left|\left\langle M^*, \frac{1}{|B'_G|}\sum_{b\in B'_G}\widehat{\mathbf{C}}(b,B') - \mathbf{C}(q)\right\rangle\right| \le \left|\left\langle M^*, \frac{1}{|B'_G|}\sum_{b\in B'_G}\widehat{\mathbf{C}}(b,B'_G) - \widehat{\mathbf{C}}(b,B')\right\rangle\right| + \left|\left\langle M^*, \frac{1}{|B'_G|}\sum_{b\in B'_G}\widehat{\mathbf{C}}(b,B'_G) - \mathbf{C}(q)\right\rangle\right|.$$

We start by bounding the first term $A = \left|\left\langle M^*, \frac{1}{|B'_G|}\sum_{b\in B'_G}\widehat{\mathbf{C}}(b,B'_G) - \widehat{\mathbf{C}}(b,B')\right\rangle\right|$. By Lemma 12:

$$A \le \frac{8}{|B'_G|}\max_{S,S'\in[d]}\left|\sum_{b\in B'_G}\left[\widehat{q}_b(S) - \widehat{q}_{B'}(S)\right]\left[\widehat{q}_b(S') - \widehat{q}_{B'}(S')\right] - \left[\widehat{q}_b(S) - \widehat{q}_{B'_G}(S)\right]\left[\widehat{q}_b(S') - \widehat{q}_{B'_G}(S')\right]\right|$$
$$= \frac{8}{|B'_G|}\max_{S,S'\in[d]}\left|\left[\widehat{q}_{B'_G}(S) - \widehat{q}_{B'}(S)\right]\left[\widehat{q}_{B'_G}(S') - \widehat{q}_{B'}(S')\right]\right|.$$

By equation 4 and condition 1, for any $S \subseteq [d]$, we have:

$$\left|\widehat{q}_{B'_G}(S) - \widehat{q}_{B'}(S)\right| \le \left|\widehat{q}_{B'_G}(S) - q(S)\right| + \left|q(S) - \widehat{q}_{B'}(S)\right| \le \left(36 + 2\sqrt{\tau_{B'}}\right)\alpha\sqrt{\frac{d\ln(e/\alpha)}{k}}.$$

Thus,

$$A \le 8\left(36 + 2\sqrt{\tau_{B'}}\right)^2\alpha\gamma.$$

By equation 16 and Lemma 12, we have:

$$\left|\left\langle M^*, \frac{1}{|B'_G|}\sum_{b\in B'_G}\widehat{\mathbf{C}}(b,B'_G) - \mathbf{C}(q)\right\rangle\right| \le 1408\gamma.$$

Thus:

$$\left|\left\langle M^*, \frac{1}{|B'|}\sum_{b\in B'_G}\widehat{\mathbf{C}}(b,B') - \mathbf{C}(q)\right\rangle\right| = \frac{|B'_G|}{|B'|}\left|\left\langle M^*, \frac{1}{|B'_G|}\sum_{b\in B'_G}\widehat{\mathbf{C}}(b,B') - \mathbf{C}(q)\right\rangle\right| \le 1408\gamma + 8\left(36 + 2\sqrt{\tau_{B'}}\right)^2\alpha\gamma.$$

Finally, for any q, we have:

$$\langle M^*, C(q)\rangle \le 8\max_{S,S'}\mathrm{Cov}_{S,S'}(q) \le \frac{8d}{k}.$$

This gives:

$$\begin{aligned}\left|\left\langle M^*, \widehat{\mathbf{C}}(B') - \mathbf{C}(q)\right\rangle\right| &\leq \left|\left\langle M^*, \frac{1}{|B'|}\sum_{b\in B'_G}\widehat{\mathbf{C}}(b,B') - \mathbf{C}(q)\right\rangle\right| + \left|\left\langle M^*, \frac{1}{|B'|}\sum_{b\in B'_A}\widehat{\mathbf{C}}(b,B')\right\rangle\right| \\ &\quad + \frac{|B'_A|}{|B'|}\left|\left\langle M^*, \mathbf{C}(q)\right\rangle\right| \\ &\leq 1408\gamma + 8\left(36 + 2\sqrt{\tau_{B'}}\right)^2\alpha\gamma + \frac{\alpha}{1-2\alpha}\frac{8d}{k} + \left|\left\langle M^*, \frac{1}{|B'|}\sum_{b\in B'_A}\widehat{\mathbf{C}}(b,B')\right\rangle\right| \\ &\leq 1409\gamma + 8\left(36 + 2\sqrt{\tau_{B'}}\right)^2\alpha\gamma + \left|\left\langle M^*, \frac{1}{|B'|}\sum_{b\in B'_A}\widehat{\mathbf{C}}(b,B')\right\rangle\right|.\end{aligned}$$

We can now combine this with equations 22:

$$\begin{aligned}\tau_{B'}\gamma &= \langle M^*, D_{B'}\rangle \\ &\leq \left|\left\langle M^*, \widehat{\mathbf{C}}(B') - \mathbf{C}(q)\right\rangle\right| + \left|\left\langle M^*, \mathbf{C}(q) - \mathbf{C}(\widehat{q}_{B'})\right\rangle\right| \\ &\leq 2200\gamma + 8\left(36 + 2\sqrt{\tau_{B'}}\right)^2\alpha\gamma + 51\sqrt{\tau_{B'}}\gamma + \frac{1}{|B'|}\left|\left\langle M^*, \sum_{b\in B'_A}\widehat{\mathbf{C}}(b,B')\right\rangle\right|.\end{aligned}$$

Thus:

$$\left|\langle M^*, \sum_{b\in B'_A}\widehat{\mathbf{C}}(b,B')\rangle\right| \geq (1-2\alpha)\left[(1-32\alpha)\tau_{B'} - (\alpha 1152 + 51)\sqrt{\tau_{B'}} - 2200 - 8*36^2\alpha\right]|B_G|\gamma$$

With $\alpha \leq 1/100$, we get:

$$\left|\langle M^*, \sum_{b\in B'_A}\widehat{\mathbf{C}}(b,B')\rangle\right| \geq (0.66\tau_{B'} - 62\sqrt{\tau_{B'}} - 2260)|B_G|\gamma$$

■

On the other hand, for any collection of good batches $B''_G \subseteq B'$ s.t. $|B''_G| \leq \alpha|B_G|$, we have by Lemma 12:

$$\begin{aligned}\sum_{b\in B''_G}\langle M^*, \widehat{\mathbf{C}}_{b,B'}\rangle &\leq 8\max_{S,S'\in[d]}\sum_{b\in B''_G}\langle \mathbb{1}_S\mathbb{1}_{S'}^T, \widehat{\mathbf{C}}_{b,B'}\rangle \\ &= 8\max_{S,S'\in[d]}\sum_{b\in B''_G}\left[\widehat{q}_b(S) - \widehat{q}_{B'}(S)\right]\left[\widehat{q}_b(S') - \widehat{q}_{B'}(S')\right].\end{aligned}$$

We can decompose the terms in the sum:

$$\left[\widehat{q}_b(S) - \widehat{q}_{B'}(S)\right]\left[\widehat{q}_b(S') - \widehat{q}_{B'}(S')\right] = \left[\widehat{q}_b(S) - q(S)\right]\left[\widehat{q}_b(S') - q(S')\right] + \left[q(S) - \widehat{q}_{B'}(S)\right]\left[q(S') - \widehat{q}_{B'}(S')\right]$$

$$+\Big[\widehat{q}_b(S)-q(S)\Big]\Big[q(S')-\widehat{q}_{B'}(S')\Big]+\Big[q(S)-\widehat{q}_{B'}(S)\Big]\Big[\widehat{q}_b(S')-q(S')\Big].$$

By condition 1:

$$\max_{S,S'\in[d]}\sum_{b\in B''_G}\Big[\widehat{q}_b(S)-q(S)\Big]\Big[\widehat{q}_b(S')-q(S')\Big]\le 33\big|B_G\big|\gamma.$$

By equation 11,

$$\max_{S,S'\in[d]}\sum_{b\in B''_G}\Big[q(S)-\widehat{q}_{B'}(S)\Big]\Big[q(S')-\widehat{q}_{B'}(S')\Big]\le |B''_G|(33+2\sqrt{\tau_{B'}})^2\alpha\gamma.$$

By equations 14 and 11,

$$\begin{aligned}\max_{S,S'\in[d]}\sum_{b\in B''_G}\Big[q(S)-\widehat{q}_{B'}(S)\Big]\Big[\widehat{q}_b(S')-q(S')\Big]&=\max_{S,S'\in[d]}|B''_G|\Big[\widehat{q}_{B''_G}(S')-q(S')\Big]\Big[q(S)-\widehat{q}_{B'}(S)\Big].\\&\le 2(33+2\sqrt{\tau_{B'}})\big|B_G\big|\alpha\gamma.\end{aligned}$$

Combining the three bounds we have:

$$\begin{aligned}\sum_{b\in B''_G}\langle M^*,\widehat{\mathbf{C}}_{b,B'}\rangle&\le 8\big|B''_G\big|(33+2\sqrt{\tau_{B'}})^2\alpha\gamma+32(33+2\sqrt{\tau_{B'}})\big|B_G\big|\alpha\gamma+264\big|B_G\big|\gamma\\&\le 8\big|B_G\big|(33+2\sqrt{\tau_{B'}})^2\alpha^2\gamma+32(33+2\sqrt{\tau_{B'}})\big|B_G\big|\alpha\gamma+264\big|B_G\big|\gamma\\&\le\Big[32\tau_{B'}\alpha^2+(1056\alpha^2+64)\sqrt{\tau_{B'}}+(1056\alpha+264)\Big]\big|B_G\big|\gamma.\end{aligned}$$

Which gives with $\alpha\le 1/100$:

$$\sum_{b\in B''_G}\langle M^*,\widehat{\mathbf{C}}_{b,B'}\rangle\le(0.0032\tau_{B'}+65\sqrt{\tau_{B'}}+275)\big|B_G\big|\gamma.$$

Thus, we have:

$$\frac{\sum_{b\in B'_A}\langle M^*,\widehat{\mathbf{C}}_{b,B'}\rangle}{\sum_{b\in B''_G}\langle M^*,\widehat{\mathbf{C}}_{b,B'}\rangle}\ge\frac{0.66\tau_{B'}-62\sqrt{\tau_{B'}}-2260}{0.02\tau_{B'}+65\sqrt{\tau_{B'}}+275}.$$

With $\sqrt{\tau_{B'}}\ge 200$,

$$\frac{\sum_{b\in B'_A}\langle M^*,\widehat{\mathbf{C}}_{b,B'}\rangle}{\sum_{b\in B''_G}\langle M^*,\widehat{\mathbf{C}}_{b,B'}\rangle}\ge 8.$$

6.7. Auxiliary Lemmas

Lemma 19 (Covariance is Lipschitz) *Let $q,q'\in\mathbb{R}^d$ and define $\alpha=q'-q$. For any $S,S'\subset[d]$, if $\big|\alpha(S)\big|\vee\big|\alpha(S')\big|\le 12$, then*

$$\Big|Cov_{S,S'}(q)-Cov_{S,S'}\left(q'\right)\Big|\le\frac{15}{k}\max\left(\big|\alpha(S)\big|,\big|\alpha(S')\big|\right).$$

Proof [Proof of Lemma 19] By equation (13), we have $|\Delta_S| \leq 1$ for all $S \subset [d]$. Therefore, by Lemma 10 and equation (20):

$$\begin{aligned}
\left|\mathrm{Cov}_{S,S'}(q) - \mathrm{Cov}_{S,S'}(q')\right| &= \left|\left\langle \mathbb{1}_S \mathbb{1}_{S'}^T, \mathbf{C}(q) - \mathbf{C}(q')\right\rangle\right| \\
&= \frac{1}{k}\left|\left\langle \mathbb{1}_S \mathbb{1}_{S'}^T, qq^T - (q+\alpha)(q+\alpha)^T + \lambda \mathbb{1}\alpha^T + \lambda\alpha\mathbb{1}^T + (1-2\lambda)\mathrm{Diag}(\alpha)\right\rangle\right| \\
&= \frac{1}{k}\left|\alpha(S)\Delta_{S'} + \alpha(S')\Delta_S + (1-2\lambda)\alpha(S\cap S') - \alpha(S)\alpha(S')\right| \\
&\leq \frac{1}{k}\Big(\big|\alpha(S)\big| + \big|\alpha(S')\big| + \big|\alpha(S)\big| + 12\big|\alpha(S)\big|\Big) \\
&\leq \frac{15}{k}\max\Big(|\alpha(S)|, |\alpha(S')|\Big).
\end{aligned}$$

■

If $22\alpha\sqrt{\frac{d\ln(e/\alpha}{k}} \geq 1$, the proven bound for the algorithm is trivially true. Else, whenever condition 1 holds, we have for any $|B'_G| \geq (1-2\alpha)\big|B_G\big|$:

$$\max\left|\widehat{q}_{B'_G}(S) - q(S)\right| \leq 1.$$

Thus, Lemma 19 may be applied to $\mathrm{Cov}_{S,S'}(\widehat{q}_{B'_G}) - \mathrm{Cov}_{S,S'}(q)$.

7. Proof of Corollary 6

Lemma 20 *Let $p \in \mathcal{P}_d$ and $p' \in \mathbb{R}^d$. Then* $\sup_{S\subseteq[d} |p(S) - p'(S)| \leq \|p - p'\|_1 \leq 2\sup_{S\subseteq[d} |p(S) - p'(S)|$.

Proof [Proof of Lemma 7] The first inequality follows from the triangle inequality. For the second one, letting $A = \{j \in [d] : p_j \geq p'_j\}$, we have: $\|p - p'\|_1 = p(A) - p'(A) + p'(A^c) - p(A^c) \leq 2\sup_{S\subseteq[d} |p(S) - p'(S)|$. ■

Proof [Proof of Corollary 6] Let $\widehat{p}$ be the output of Algorithm 1 and $\widehat{p}^* = \frac{\widehat{p}}{\|\widehat{p}\|_1}$. Then

$$\|p - \widehat{p}^*\|_1 \leq \|\widehat{p} - p\|_1 + \|\widehat{p} - \widehat{p}^*|_1 = \|\widehat{p} - p\|_1 + \big|\|\widehat{p}\|_1 - 1\big| \leq 2\|p - \widehat{p}\|_1.$$

■

8. Lower bound: Proof of Proposition 3

For any two probability distributions p, q over some measurable space $(\mathcal{X}, \mathcal{A})$, we denote by

$$\chi^2(p||q) = \begin{cases} \int_{\mathcal{X}} \frac{p}{q}\, dp - 1 \text{ if } p \ll q \\ +\infty \text{ otherwise} \end{cases}$$

the χ^2 divergence between p and q. We start with the following Lemma.

Lemma 21 *Assume that $d \geq 3$. There exists an absolute constant $c > 0$ such that for all estimator $\hat{p}$ and all ϵ-LDP mechanism Q, there exists a probability vector $p \in \mathcal{P}_d$ satisfying*

$$\mathbb{E}\left[\sup_{z' \in \mathcal{C}(Z)} \left\|\hat{p}(z') - p\right\|_1\right] \geq c\left\{\left(\frac{d}{\epsilon\sqrt{kn}} + \frac{\alpha\sqrt{d}}{\epsilon\sqrt{k}}\right) \wedge 1\right\},$$

where the expectation is taken over all collections of n' clean batches $Z^1, \dots, Z^{n'}$ where $Z^b = (Z_1^b, \dots, Z_k^b)$ and $Z_l^b \overset{iid}{\sim} Qp$.

This Lemma is the analog of Proposition 3 but with the guarantee in expectation rather than with high probability. We first prove this Lemma before moving to the proof of Proposition 3.

Proof [Proof of Lemma 21] We first show that $R^*_{n,k}(\epsilon, \alpha, d) \geq c\left(\frac{d}{\epsilon\sqrt{kn}} \wedge 1\right)$ for some small enough absolute constant $c > 0$. Informally, this amounts to saying that the estimation problem under both contamination and privacy is more difficult than just under privacy. Formally:

$$R^*_{n,k}(\epsilon, \alpha, d) = \inf_{\hat{p}, Q} \sup_{p \in \mathcal{P}_d} \mathbb{E}\left[\sup_{z' \in \mathcal{C}(Z))} \left\|\hat{p}(z') - p\right\|_1\right] \geq \inf_{\hat{p}, Q} \sup_{p \in \mathcal{P}_d} \mathbb{E}\left[\left\|\hat{p} - p\right\|_1\right] \geq c\left(\frac{d}{\epsilon\sqrt{kn}} \wedge 1\right),$$

where the last inequality follows from Duchi et al. (2014) Proposition 6. We also give a simpler proof of this fact in Appendix 9, using Assouad's lemma.

We now prove $R^*_{n,k}(\epsilon, \alpha, d) \geq c\left(\frac{\alpha\sqrt{d}}{\epsilon\sqrt{k}} \wedge 1\right)$. For any ϵ-LDP mechanism Q and probability vector $p \in \mathcal{P}_d$, denote by Qp the density of the privatized random variable Z defined by $Z|X \sim Q(\cdot|X)$ and by $Qp^{\otimes k}$ the density of the joint distribution of k iid observations with distribution Qp. Define the set of pairs of probability vectors that are indistinguishable after privatization by Q and adversarial contamination

$$\mathcal{A}(Q) = \left\{(p, q) \in \mathcal{P}_d \;\middle|\; TV(Qp^{\otimes k}, Qq^{\otimes k}) \leq \alpha\right\}. \tag{23}$$

To derive the adversarial rate, it suffices to prove

$$\inf_{Q} \sup_{p,q \in \mathcal{A}(Q)} \|p - q\|_1 \geq c\left\{\frac{\alpha\sqrt{d}}{\epsilon\sqrt{k}} \wedge 1\right\}. \tag{24}$$

To understand why (24) is a natural program to consider, fix an ϵ-LDP mechanism Q and denote by $(\mathcal{Z}, \mathcal{U}, \nu)$ its image space. If $(p, q) \in \mathcal{A}(Q)$, then letting

$$A = \frac{Qp^{\otimes k} \vee Qq^{\otimes k}}{1 + TV(Qp^{\otimes k}, Qq^{\otimes k})}, \quad N^{(p)} = \frac{A - (1-\alpha)Qp^{\otimes k}}{\alpha}, \quad \text{and } N^{(q)} = \frac{A - (1-\alpha)Qq^{\otimes k}}{\alpha},$$

we can directly check that $A, N^{(p)}$ and $N^{(q)}$ are probability measures over $(\mathcal{Z}, \mathcal{U})$ (for $N^{(p)}$ and $N^{(q)}$, we use the fact that $(p, q) \in \mathcal{A}(Q)$ to prove that $N^{(p)}(dz) \geq 0$ and $N^{(q)}(dz) \geq 0$). Moreover, it holds that $A = (1-\alpha)Qp^{\otimes k} + \alpha N^{(p)} = (1-\alpha)Qq^{\otimes k} + \alpha N^{(q)}$. This is exactly equivalent to saying that any clean family of n batches with distribution $Qp^{\otimes k}$ or $Qp^{\otimes k}$ can be transformed into an α-contaminated family of

n batches with distribution A through α adversarial contamination. By observing such a contaminated family, it is therefore impossible to determine whether the underlying distribution is p or q, so that the quantity $\|p-q\|_1/2$ is a lower bound on the minimax estimation risk.

We now prove (24). For all $j \in \{1,\dots,d\}$ and $z \in \mathcal{Z}$, set

$$q_j(z) = \frac{Q(z|j)}{Q(z|1)} - 1, \quad d\mu(z) = Q(z|1)d\nu(z), \tag{25}$$

and

$$\Omega_Q = \big(\Omega_Q(j,j')\big)_{jj'} = \left(\int_{\mathcal{Z}} q_j(z)q_{j'}(z)d\mu(z)\right)_{ij} \tag{26}$$

Given Q, we first prove that a sufficient condition for (p,q) to belong to $\mathcal{A}(Q)$ is that $(p-q)^T\Omega(p-q) \le C\alpha^2/k$ for some small enough absolute constant $C>0$. Fix $p,q \in \mathcal{P}_d$ and define $\Delta = p-q$. By Tsybakov (2008), Section 2.4, we have

$$TV(Qp^{\otimes k}, Qq^{\otimes k}) \le \sqrt{-1 + \big(1+\chi^2(Qp||Qq)\big)^k}. \tag{27}$$

Now,

$$\begin{aligned}
\chi^2(Qp||Qq) &= \int_{\mathcal{Z}} \frac{\big(Qp(z)-Qq(z)\big)^2}{Qq(z)}dz = \int_{\mathcal{Z}} \frac{\left(\sum_{j=1}^d Q(z|j)\ \Delta_j\right)^2}{\sum_{j=1}^d Q(z|j)\ q_j}dz \\
&= \int_{\mathcal{Z}} \frac{\left(\sum_{j=1}^d \left(\frac{Q(z|j)}{Q(z|1)} - 1\right)\Delta_j\right)^2}{\sum_{j=1}^d \frac{Q(z|j)}{Q(z|1)}\ q_j} Q(z|1)d\nu(z) \quad \text{since } \sum_{j=1}^d \Delta_j = 0 \\
&\le e^\epsilon \int_{\mathcal{Z}} \sum_{j,j'=1}^d \Delta_j\Delta_{j'} q_j(z)q_{j'}(z)d\mu(z) \\
&= e^\epsilon \Delta^T\Omega_Q\Delta.
\end{aligned}$$

Write $\Omega = \Omega_Q$ and assume that $\Delta^T\Omega_Q\Delta \le C\alpha^2/k$ for $C \le e^{-2}$. Then equation (27) yields:

$$TV(Qp^{\otimes k}, Qq^{\otimes k}) \le \sqrt{-1+\big(1+e^\epsilon\Delta^T\Omega\Delta\big)^k} \le \sqrt{-1+\exp\big(e^\epsilon k\Delta^T\Omega\Delta\big)} \le \sqrt{-1+\exp\big(Ce^\epsilon\alpha^2\big)} \le \alpha.$$

Defining

$$\mathcal{A}_{\chi^2}(Q) = \left\{(p,q) \in \mathcal{P} \;\middle|\; (p-q)^T\Omega(p-q) \le \frac{C\alpha^2}{k}\right\}, \tag{28}$$

it follows that $\mathcal{A}_{\chi^2}(Q) \subset \mathcal{A}(Q)$ for all Q, so that

$$\inf_Q \sup_{(p,q)\in\mathcal{A}(Q)} \|\Delta\|_1 \ge \inf_Q \sup_{(p,q)\in\mathcal{A}_{\chi^2}(Q)} \|\Delta\|_1.$$

Fix Q and note that Ω_Q is symmetric and nonnegative. We sort its eigenvalues as $\{\lambda_1 \le \cdots \le \lambda_d\}$ and denote by $v_1,\dots,v_d$ the associated eigenvectors. We also define $j_0 = \max\big\{j \in \{1,\dots,d\} : \lambda_j \le 3e^2\epsilon^2\big\}$. Noting

that $\forall j : |q_j| \leq e\epsilon$ and that μ is a probability measure, we get that $Tr(\Omega) = \sum_{j=1}^{d} \int_{\mathcal{Z}} q_j^2 d\mu \leq de^2\epsilon^2$, so that $(d-j_0)3e^2\epsilon^2 \leq de^2\epsilon^2$ hence $j_0 \geq 2d/3$.

Let $H = \left\{x \in \mathbb{R}^d : x^T\mathbb{1} = 0\right\}$, and note that $V := \text{span}(v_j)_{j\leq j_0} \cap H$ is of dimension at least $m = \frac{2d}{3} - 1 \geq \frac{d}{3}$. Therefore by Lemma 22, there exists $\Delta \in V$ such that $\|\Delta\|_2^2 = \frac{C\alpha^2}{2e^2\epsilon^2 k} \wedge \frac{1}{d}$ and $\|\Delta\|_1 \geq C_{22}\sqrt{m}\|\Delta\|_2 \gtrsim \frac{\alpha\sqrt{d}}{\epsilon\sqrt{k}}$. Noting that over $\mathbb{R}^d$, $\|\cdot\|_1 \leq \sqrt{d}\|\cdot\|_2$, we also have $\|\Delta\|_1 \leq \frac{\alpha}{\epsilon\sqrt{k}} \wedge 1 \leq 1$.

This allows us to define the following vectors: $p = \left(\frac{|\Delta_j|}{\|\Delta\|_1}\right)_{j=1}^{d} \in \mathcal{P}_d$ and $q = p - \Delta$. To check that $q \in \mathcal{P}_d$, note that the condition $\Delta^T\mathbb{1} = 0$ ensures that $q^T\mathbb{1} = 1$. Moreover, for all $j \in \{1, \dots, d\}$ we have $q_j = \frac{|\Delta_j|}{\|\Delta\|_1} - \Delta_j \geq 0$ since $\|\Delta\|_1 \leq 1$.
Since by construction, we have $\Delta\Omega_Q\Delta \leq 2e^2\epsilon^2\|\Delta\|_2^2 \leq \frac{C\alpha^2}{k}$ and $p, q \in \mathcal{P}_d$, we have $(p, q) \in \mathcal{A}_{\chi^2}(Q)$. For all ϵ-LDP mechanism Q, it therefore holds that $\sup_{(p,q)\in\mathcal{A}_{\chi^2}(Q)} \|\Delta\|_1 \gtrsim \frac{\alpha\sqrt{d}}{\epsilon\sqrt{k}} \wedge 1$. Taking the infimum over all Q, the result is proven. ■

Lemma 22 *There exists an absolute constant C_{22} such that for all $m \in \{\lceil \frac{d}{3} \rceil, \dots, d\}$ and all linear subspace $V \subset \mathbb{R}^d$ of dimension m, it holds:*

$$\sup_{v\in V} \frac{\|v\|_1}{\|v\|_2} \geq C_{22}\sqrt{m}.$$

Proof [Proof of Lemma 22] Let V be a linear subspace of $\mathbb{R}^d$ of dimension m and denote by $\Pi_V := \left(\Pi_V(i,j)\right)_{ij}$ the orthogonal projector onto V. Let $X \sim \mathcal{N}(0, \Pi_V)$. For some large enough absolute constant $C > 0$ we have:

$$\begin{aligned}\sup_{v\in V} \frac{\|v\|_1}{\|v\|_2} &\geq \mathbb{E}\left[\frac{\|X\|_1}{\|X\|_2}\right] \geq \mathbb{E}\left[\frac{\|X\|_1}{\|X\|_2}\mathbb{1}\left\{\|X\|_2 \leq C\sqrt{m}\right\}\right] \\ &\geq \underbrace{\frac{1}{C\sqrt{m}}\mathbb{E}\left[\|X\|_1\right]}_{\text{Principal term}} - \underbrace{\frac{1}{C\sqrt{m}}\mathbb{E}\left[\|X\|_1\mathbb{1}\left\{\|X\|_2 > C\sqrt{m}\right\}\right]}_{\text{Residual term}} \end{aligned} \tag{29}$$

We first analyze the principal term.

$$\mathbb{E}\|X\|_1 = \sum_{i,j=1}^{d} \mathbb{E}|X_{ij}| = \sqrt{\frac{2}{\pi}} \sum_{i,j=1}^{d} \left|\Pi_V(i,j)\right|^{1/2}$$

Note that $\forall i, j \in \{1, \dots, d\} : |\Pi_V(i,j)| \leq 1$ and that $\sum_{i,j=1}^{d} \Pi_V^2(i,j) = m$. Therefore:

$$\inf_{dim(V)=m} \sum_{i,j=1}^{d} \left|\Pi_V(i,j)\right|^{1/2} \geq \inf_{A\in\mathbb{R}^{d\times d}} \sum_{i,j=1}^{d} |a_{ij}|^{1/2} \text{ s.t. } \begin{cases} \|A\|_2^2 = m \\ \forall i,j : |a_{ij}| \leq 1. \end{cases}$$

$$= \inf_{a\in\mathbb{R}^{d\times d}} \sum_{i,j=1}^{d} a_{ij} \;\text{ s.t. }\; \begin{cases} \sum_{i,j=1}^{d} a_{ij}^4 = m \\ \forall i,j : 0 \le a_{ij} \le 1. \end{cases} \tag{30}$$

The last optimization problem amounts to minimizing an affine function over a convex set, hence the solution, denoted by $(a^*_{ij})_{ij}$, is attained on the boundaries of the domain. Therefore, $\forall i,j \in \{1,\dots,d\} : a^*_{ij} \in \{0,1\}$. It follows from $\sum_{ij} a_{ij}^4 = m$ that the family a^*_{ij} contains exactly m nonzero coefficients, which are all equal to 1. Therefore, the value of the last optimization problem is m, which yields that the principal term is lower bounded by $\frac{\sqrt{m}}{C}$.

We now move to the residual term. Writing $X = \sum_{j=1}^m x_j e_j$ where $(e_j)_{j=1}^m$ is an orthonormal basis of V, we have:

$$\begin{aligned}\mathbb{E}\left[\|X\|_1 \mathbb{1}\left\{\|X\|_2 > C\sqrt{m}\right\}\right] &\le \sqrt{d}\,\mathbb{E}\left[\|X\|_2 \mathbb{1}\left\{\|X\|_2 > C\sqrt{m}\right\}\right] \le \sqrt{d}\left\{\mathbb{E}\left[\|X\|_2^2 \mathbb{1}\left\{\|X\|_2^2 > C^2 m\right\}\right]\right\}^{1/2} \\ &\le \sqrt{d}\left\{m\,\mathbb{E}\left[x_1^2 \mathbb{1}\left\{\sum_{j=1}^m x_j^2 \ge C^2 m\right\}\right]\right\}^{1/2}. \end{aligned} \tag{31}$$

Moreover

$$\begin{aligned}\mathbb{E}\left[x_1^2 \mathbb{1}\left\{\sum_{j=1}^m x_j^2 \ge C^2 m\right\}\right] &\le \mathbb{E}\Big[x_1^2\,\mathbb{1}\{x_1 \ge C\}\Big] + \mathbb{E}\left[x_1^2\,\mathbb{1}\left\{\sum_{j=2}^m x_j^2 \ge C^2(m-1)\right\}\right] \\ &\le \mathbb{E}\Big[x_1^2\,\mathbb{1}\{x_1 \ge C\}\Big] + \mathbb{E}\Big[x_1^2\Big]\,\mathbb{P}\left(\Big|\sum_{j=2}^m x_j^2 - \mathbb{E}x_1^2\Big| \ge \big(C^2 - \mathbb{E}x_1^2\big)(m-1)\right)\end{aligned} \tag{32}$$

By the dominated convergence Theorem, $\lim_{C\to+\infty} \mathbb{E}\big[x_1^2\,\mathbb{1}\{x_1 \ge C\}\big] = 0$. Moreover, by Chebyshev's inequality:

$$\mathbb{P}\left(\Big|\sum_{j=2}^m x_j^2 - \mathbb{E}x_1^2\Big| \ge \big(C^2 - \mathbb{E}x_1^2\big)(m-1)\right) \le \frac{\mathbb{V}(x_1^2)}{\big(C^2 - \mathbb{E}x_1^2\big)^2 (m-1)} \underset{C\to+\infty}{\to} 0. \tag{33}$$

By (31), (32) and (33), we conclude that for all absolute constant $c > 0$, there exists a large enough absolute constant $C > 0$ such that the residual term is at most $\frac{c\sqrt{d}}{C}$. Take $c = \frac{1}{2}$ and $m \ge \frac{d}{3}$, then by equation (29) we get:

$$\sup_{v\in V} \frac{\|v\|_1}{\|v\|_2} \ge \frac{\sqrt{m}}{C} - \frac{c\sqrt{d}}{C} \ge \Big(1 - \frac{\sqrt{3}}{2}\Big)\frac{\sqrt{m}}{C} =: C_{22}\sqrt{m}.$$

■

Proof [Proof of Proposition 3]
We distinguish between two cases.

1. **First case** If $\frac{d}{\epsilon\sqrt{nk}} \leq \frac{\alpha}{\epsilon}\sqrt{\frac{d}{k}}$ i.e. if the dominating term comes from the contamination, taking $p, q \in \mathcal{P}_d$ like in the proof of Proposition 21 and $t \in \{p, q\}$ uniformly at random yields that

$$\inf_{\hat{p}} \sup_{p\in\mathcal{P}_d} \mathbb{P}\left(\sup_{Z\in\mathcal{C}(Y)} \|\hat{p}(Z) - p\|_1 \geq \|p-q\|_1/2\right)$$
$$\geq \inf_{\hat{p}} \mathbb{E}_{t\in\{p,q\}}\mathbb{P}_t\left(\sup_{Z\in\mathcal{C}(Y)} \|\hat{p}(Z) - p\|_1 \geq \|p-q\|_1/2\right) \geq \frac{1}{2} \geq O(e^{-d}),$$

where $\|p-q\|_1 \gtrsim \frac{\alpha}{\epsilon}\sqrt{\frac{d}{k}} \wedge 1$.

2. **Second case** If $\frac{d}{\epsilon\sqrt{nk}} \geq \frac{\alpha}{\epsilon}\sqrt{\frac{d}{k}}$ i.e. if the dominating term comes from the privacy constraint, then we set $N = nk$ and assume that we observe $Z_1, \ldots, Z_N$ iid with probability distribution $Z|X \sim Q(\cdot|X)$ such that X has a discrete distribution over $\{1, \ldots, d\}$. In other words, the random variables Z_i are no longer batches, but rather we have nk iid clean samples that are privatized versions of iid samples with distribution p. By section 9, it holds that

$$\inf_{\hat{p}} \sup_{p\in\mathcal{P}_d} \mathbb{E}\left[\sup_{\text{contamination}} \|\hat{p} - p\|_1\right] \geq \inf_{\hat{p}} \sup_{p\in\mathcal{P}_d} \mathbb{E}\|\hat{p} - p\|_1 \geq c\frac{d}{\epsilon\sqrt{N}},$$

for some small enough absolute constant $c > 0$. We use the definition of γ and of the cubic set of hypotheses $\mathcal{P}$ from (37). Let $\hat{p}$ be any estimator of the probability parameter and, for some small enough absolute constant $c > 0$, define

$$r = c\,\frac{d}{\sqrt{kn}}. \tag{34}$$

We first justify that for this particular set of hypotheses, it is possible to assume *wlog* that

$$\|p - \hat{p}\|_1 \leq 6\gamma d \;\leq\; 6\,c_\gamma r. \tag{35}$$

Indeed, define $u = \left(\frac{1}{d}\right)_{j=1}^{d}$. If for some observation $Z = (Z_1, \ldots Z_N)$ the estimate $\hat{p}(Z)$ satisfies $\|\hat{p}(Z) - u\|_1 > 4\gamma d$, then it is possible to improve $\hat{p}$ by replacing it with the estimator $\bar{p}$ satisfying $\|\bar{p}(Z) - p\|_1 \leq 6\gamma d$ and defined as:

$$\bar{p} := \hat{p}\mathbb{1}\left\{\left\|\hat{p} - u\right\|_1 \leq 4\gamma d\right\} + u\mathbb{1}\left\{\left\|\hat{p} - u\right\|_1 > 4\gamma d\right\}.$$

Indeed, recalling that $\forall p \in \mathcal{P} : \left\|u - p\right\|_1 = 2\gamma d$, there are two cases.

- If $\left\|\hat{p}(Z) - u\right\|_1 \leq 4\gamma d$, then $\hat{p} = \bar{p}$ so that $\left\|p - \bar{p}\right\|_1 \leq \left\|p - u\right\|_1 + \left\|u - \bar{p}\right\|_1 \leq 2\gamma d + 4\gamma d = 6\gamma d$.
- Otherwise, $\bar{p} = u$ and we get

$$\left\|\bar{p}(Z) - p\right\|_1 = 2\gamma d = 4\gamma d - 2\gamma d < \left\|\hat{p}(Z) - u\right\|_1 - \left\|u - p\right\|_1 \leq \left\|\hat{p}(Z) - p\right\|_1,$$

which proves that (35) can be assumed *wlog*. Now, from the proof of Lemma 23, we also have

$$\sup_{p\in\mathcal{P}} \mathbb{E}_p\left\|\hat{p}-p\right\|_1 \;\geq\; \frac{c_\gamma}{4} r =: Cr.$$

Fix any $p \in \mathcal{P}$ and write $\pi := \sup_{p\in\mathcal{P}} \mathbb{P}_p\left(\|\hat{p}-p\|_1 \geq cr\right)$ for $c = c_\gamma\left(\frac{1}{4} - 6\delta\right) > 0$ for $\delta < \frac{1}{24} =: c'$.

It follows that:

$$\begin{aligned}
Cr &\leq \sup_{p\in\mathcal{P}} \mathbb{E}_p\left\|\hat{p}-p\right\|_1 \\
&= \sup_{p\in\mathcal{P}} \left\{ \mathbb{E}_p\left[\left\|\hat{p}-p\right\|_1 \mathbb{1}\left\{\left\|\hat{p}-p\right\|_1 \geq cr\right\}\right] \;+\; \mathbb{E}_p\left[\left\|\hat{p}-p\right\|_1 \mathbb{1}\left\{\left\|\hat{p}-p\right\|_1 < cr\right\}\right] \right\} \\
&\leq 6c_\gamma r\cdot\pi \;+\; cr \quad \text{by equation (35), so that } \pi \geq \frac{C-c}{6c_\gamma} \geq \delta \geq O(e^{-d}).
\end{aligned}$$

■

9. Simpler proof of the lower bound with privacy and no outliers

Here, we assume that $k = 1$ and that we observe $Z_1, \ldots, Z_n$ that are n iid with probability distribution $Z|X \sim Q(\cdot|X)$ and X has a discrete distribution over $\{1, \ldots, d\}$. We prove the following Lemma

Lemma 23 *In this setting, it holds*

$$\inf_{\hat{p}} \sup_{p\in\mathcal{P}_d} \mathbb{E}\|\hat{p}-p\|_1 \geq c\frac{d}{\epsilon\sqrt{n}}, \tag{36}$$

for some small enough absolute constant $c > 0$.

For all $\alpha \in \{\pm 1\}^{\lfloor d/2\rfloor}$, define the probability vector $p_\alpha \in \mathcal{P}_d$ such that

$$\forall j \in \{1,\ldots,d\} : p_\alpha(j) = \begin{cases} \frac{1}{d} + \alpha_j\,\gamma & \text{if } j \leq \frac{d}{2}, \\ \frac{1}{d} & \text{if } d \text{ is odd and } j = \frac{d+1}{2}, \\ \frac{1}{d} - \alpha_{d-j+1}\gamma & \text{otherwise,} \end{cases} \tag{37}$$

where $\gamma = \frac{c_\gamma}{\epsilon\sqrt{n}} \wedge \frac{c_\gamma}{d}$ and c_γ is a small enough absolute constant. Consider the cubic set of hypotheses

$$\mathcal{P} = \left\{ p_\alpha \;\middle|\; \alpha \in \{\pm 1\}^{\lfloor d/2\rfloor} \right\}. \tag{38}$$

This set $\mathcal{P}$ consists of $M = 2^{\lfloor d/2\rfloor}$ hypotheses. Over $\mathcal{P}$, the ℓ_1 distance simplifies as follows:

$$\forall \alpha, \alpha' \in \{\pm 1\}^{\lfloor d/2\rfloor} : \|p_\alpha - p_{\alpha'}\|_1 = 4\gamma\,\rho(\alpha,\alpha'), \tag{39}$$

where $\rho(\alpha, \alpha') = \sum_{j=1}^{\lfloor d/2 \rfloor} \mathbb{1}_{\alpha_j \neq \alpha'_j}$ denotes the Hamming distance between α and α'.

To apply Assouad's Lemma (see e.g. Tsybakov (2008) Theorem 2.12.(ii)), let $\alpha, \alpha' \in \{\pm 1\}^{\lfloor d/2 \rfloor}$ such that $\rho(\alpha, \alpha') = 1$. Recall that the observations $Z_1, \dots, Z_n$ are iid and follow the distribution $Z|X \sim Q(\cdot|X)$ where Q is an ϵ-locally differentially private mechanism. Fix any such mechanism Q, and denote by q_α and $q_{\alpha'}$ the respective densities of Z when $X \sim p_\alpha$ and $X \sim p_{\alpha'}$. We therefore have $\forall z \in \mathcal{Z} : q_\alpha(z) = \int Q(z|x)\ p_\alpha(x) d\nu(x)$ where ν denotes the counting measure over $\{1, \dots, d\}$. For some probability distribution P, we also denote by $P^{\otimes n}$ the law of the probability vector $(X_1, \dots, X_n)$ when $X_i \overset{iid}{\sim} P$. Now, we have:

$$TV\left(q_\alpha^{\otimes n}, q_{\alpha'}^{\otimes n}\right) \leq \sqrt{\chi^2\left(q_\alpha^{\otimes n} \,||\, q_{\alpha'}^{\otimes n}\right)} = \sqrt{\left(1 + \chi^2\left(q_\alpha \,||\, q_{\alpha'}\right)\right)^n - 1}, \tag{40}$$

and defining $\Delta p(x) = p_\alpha(x) - p_{\alpha'}(x)$ for all $x \in \{1, \dots, d\}$, we can write:

$$\begin{aligned}
\chi^2\left(p_\alpha \,||\, p_{\alpha'}\right) &= \int_{\mathcal{Z}} \frac{\left(q_\alpha(z) - q_{\alpha'}(z)\right)^2}{q_{\alpha'}(z)} dz = \int_{\mathcal{Z}} \frac{\left(\int Q(z|x)\ \Delta p(x) d\nu(x)\right)^2}{\int Q(z|x)\ p_{\alpha'}(x) d\nu(x)} dz \\
&= \int_{\mathcal{Z}} Q(z|1)\ \frac{\left(\int_{\mathcal{X}} \left(\frac{Q(z|x)}{Q(z|1)} - 1\right) \Delta p(x) d\nu(x)\right)^2}{\int_{\mathcal{X}} \frac{Q(z|x)}{Q(z|1)}\ p_{\alpha'}(x) d\nu(x)} dz \quad \text{since } \int \Delta p(x) d\nu(x) = 0 \\
&\leq \int_{\mathcal{Z}} Q(z|1)\ \frac{\left(\int_{\mathcal{X}} \left|\frac{Q(z|x)}{Q(z|1)} - 1\right| \left|\Delta p(x)\right| d\nu(x)\right)^2}{\int_{\mathcal{X}} e^{-\epsilon}\ p_{\alpha'}(x) d\nu(x)} dz \\
&\leq \int_{\mathcal{Z}} Q(z|1)\ \frac{\left(C\epsilon\ TV(p_\alpha, p_{\alpha'})\right)^2}{e^{-\epsilon}} dz = e^\epsilon C^2 \epsilon^2 \left(2\gamma\ \rho(p_\alpha, p_{\alpha'})\right)^2 \\
&\leq 12 C^2 \epsilon^2 \gamma^2 \leq \frac{12 C^2 c_\gamma^2}{n},
\end{aligned}$$

where $C > 0$ is an absolute constant such that for all $\epsilon \in (0, 1)$ we have $e^\epsilon - 1 \leq C\epsilon$ and $1 - e^{-\epsilon} \leq C\epsilon$. Now by (40), we have:

$$TV\left(q_\alpha^{\otimes n}, q_{\alpha'}^{\otimes n}\right) \leq \sqrt{\left(1 + \chi^2\left(q_\alpha \,||\, q_{\alpha'}\right)\right)^n - 1} \leq \sqrt{\exp\left(12 C^2 c_\gamma^2\right) - 1}.$$

Choosing c_γ small enough therefore ensures that $TV\left(q_\alpha^{\otimes n}, q_{\alpha'}^{\otimes n}\right) \leq \frac{1}{2}$, so that by Assouad's lemma, the minimax risk is lower bounded as:

$$\inf_{\widehat{p}} \sup_{p \in \mathcal{P}_d} \mathbb{E}\|\widehat{p} - p\|_1 \geq \left\lfloor \frac{d}{2} \right\rfloor \frac{1}{2} 4\gamma \left(1 - \frac{1}{2}\right) \geq \frac{c_\gamma}{10} \left(\frac{d}{\epsilon\sqrt{n}} \wedge 1\right).$$