Robust Learning of Discrete Distributions from Batches

Ayush Jain and Alon Orlitsky
University of California, San Diego
{ayjain,alon}@eng.ucsd.edu

Abstract

Let $d$ be the lowest $L_1$ distance to which a $k$-symbol distribution $p$ can be estimated from $m$ batches of $n$ samples each, when up to $\beta m$ batches may be adversarial. For $\beta < \frac{1}{2}$, Qiao and Valiant [12] showed that $d = \Omega(\beta / \sqrt{n})$ and requires $m = \Omega(k/\beta^2)$ batches. For $\beta < 1/900$, they provided a $d$ and $m$ order-optimal algorithm that runs in time exponential in $k$.

For $\beta < 0.5$, we propose an algorithm with comparably optimal $d$ and $m$, but run-time polynomial in $k$ and all other parameters.

1 Introduction

Estimating discrete distributions from their samples is a fundamental tenet of modern science. In many applications, part of the data inadvertently or maliciously corrupted, but the distribution still needs to be estimated as best possible.

If an adversary can corrupt a fraction $\beta$ of the samples, it could replace all samples it corrupts by the lowest-probability symbol $s$, resulting in the sample distribution $(1 - \beta)p + \beta \delta_s$, where $\delta_s$ is the singleton distribution $p(s) = 1$. The $L_1$ distance between this distribution and $p$ is $2(\beta + p_s(1 - \beta) - p_s) = 2\beta(1 - p_s)$, that for the worst distribution is $2\beta$. Conversely, it can be easily shown that any estimator for $p$ incurs an $L_1$ loss $\geq 2\beta$ for some distribution.

Fortunately, in many applications, the data is provided in batches. For example, when the data is collected by $m$ sensors of which a $\beta$ fraction are faulty, when trying to estimate the word frequencies of an author from texts of which some fraction is mis-attributed, or when learning user preferences and some fraction of users provides intentionally biased feedback. In these cases, the underlying distribution can be estimated to a much higher accuracy.

To formalize the problem, [12] considered learning a $k$-symbol distribution $p$ whose samples are provided in batches of size $n$. A total of $m$ batches are provided, of which a fraction $\leq \beta$ may be arbitrarily and adversarially corrupted, while every other batch $b$ consists of samples drawn according a distribution $p_b$ such that $||p_b - p||_1 \leq \eta$. Let $d^* = d(k, n, m, \beta, \eta)$ be the lowest $L_1$ distance between $p$ and its approximation achievable with high probability for all distributions.

For $\beta < 1/2$, they showed that for any alphabet size $k \geq 2$, and any number $m$ of batches, the lowest achievable $L_1$ distance is $d^* = \Omega(\eta + \beta/\sqrt{n})$. For $\beta < 1/900$, they also derived an estimation algorithm that achieves this distance lower-bound up to a constant factor, and uses $m = O(k/\beta^2)$ batches. When $s$ genuine samples are available and no corruptions occur, [7] showed that the expected $d^* \sim \sqrt{2(k - 1)/(\pi s)}$, since here, at most $mn$ genuine samples are available, the algorithm is also orderwise sample optimal.
One drawback of the proposed algorithm is that it runs in time exponential in the alphabet size $k$. In this paper we derive an algorithm that achieves the same optimal distortion and same optimal sample efficiency, but runs in time polynomial in all parameters.

In a paper concurrent and independent of this work, [4] propose an algorithm that estimates $p$ to the same distance as ours, but with $\tilde{O}(nk^{O(1/\beta)})$ batches and run-time $\tilde{O}(nk^{O(\log(1/\beta))/\beta^4})$. Both the sample complexity and run time are higher than ours, and become non-polynomial if $\beta$ decreases. They also consider certain structured distributions, not addressed in this paper, for which they provide an algorithm with similar run time, but lower sample complexity.

1.1 Related Work

The current results extend several long lines of work on learning distributions and their properties.

The best approximation of a distribution with a given number of samples was determined up to the exact first-order constant for KL loss loss [2], and $L_1$ and $\chi^2$ loss [7]. These settings do not allow adversarial examples, and some modification of the empirical estimates of the samples is often shown to be near optimal. This is not the case in the presence of adversarial samples, where the challenge is to devise algorithms that are efficient from both computational and sample viewpoints.

Our results also relate to classical robust-statistics work [6,14]. There has also been significant recent work leading to practical distribution learning algorithms that are robust to adversarial contamination of the data. For example, [5,8] presented algorithms for learning the mean and covariance matrix of high-dimensional sub-Gaussian and other distributions with bounded fourth moments in presence of the adversarial samples. Their estimation guarantees are typically in terms of $L_2$, and do not yield the $L_1$- distance results required for discrete distributions.

The work was extended in [3] to the case when more than half of the samples are adversarial. Their algorithm returns a small set of candidate distributions one of which is a good approximate of the underlying distribution. For more extensive survey on robust learning algorithms in the continuous setting, see [5,13].

Another motivation for this work derives from the practical federated-learning problem, where information arrives in batches [9,10].

1.2 Problem Formulation

Let $\Delta_k$ be the collection of all distributions over $[k] = \{1, \ldots, k\}$. The $L_1$ distance between two distributions $p, q \in \Delta_k$ is

$$||p - q||_1 \triangleq \sum_{i \in [k]} |p(i) - q(i)| = 2 \cdot \max_{S \subseteq [k]} |p(S) - q(S)|.$$ 

We would like to estimate an unknown target distribution $p \in \Delta_k$ to a small $L_1$ distance from samples, some of which may be corrupted or even adversarial. Specifically, let $B$ be a collection of $m$ batches, each comprising $n$ samples. For an unknown set $G \subseteq B$ of good batches, each batch consists of i.i.d. samples from a distribution $p_b$ that is within $L_1$ distance $\leq \eta$ from $p$, and for the unknown complement set $A \triangleq B - G$ of adversarial batches, each batch consists of arbitrary samples that may even be chosen by an adversary, possibly based on the samples in the good batches.

Let $\alpha = |G|/m$, and $\beta = |A|/m = 1 - \alpha$ be the fractions of good and adversarial batches, respectively. Our goal is to use the $m$ batches to return a distribution $p^*$ such that $||p^* - p||_1$ is small or equivalently $|p(S) - p^*(S)|$ is small for all $S \subseteq [k]$.

For the special case where the distance bound $\eta = 0$, all samples in the good batches are generated by the target distribution $p$. Since the proofs and techniques are essentially the same for $\eta = 0$ and $\eta > 0$, we first assume for simplicity that $\eta = 0$, and in section 1.3 we discuss $\eta > 0$.

1.3 Lower Bounds

Qiao and Valiant [12] showed a tight lower bound for the above problem when $\alpha \geq 1/2$. They also show an algorithm which achieves the lower bound within a constant factor although the run time of
the algorithm is exponential in $k$. For completeness, we present the lower bound and the outline of the proof given in [12]. Note that the lower bound is independent of the alphabet size $k$.

**Theorem 1.** ([12]) For every $\beta \in [0, 1]$, $k \geq 2$, $n \geq 1$, and $m$, for any estimate $\hat{p}$ there is a distribution $p$ such that with probability $\geq 1/2$, $||p - \hat{p}||_1 \geq \frac{\beta}{\sqrt{2n}}$.

**Proof.** Consider $k = 2$. Let $\gamma = \frac{\beta}{2\sqrt{2n}}$, and let $p$ be either $\text{Bern}(\frac{1}{2} + \gamma)$ or $\text{Bern}(\frac{1}{2} - \gamma)$.

For every batch, the number of 1’s is a sufficient statistic for estimating $p$, and it is distributed either $B(n, \frac{1}{2} + \gamma)$ or $B(n, \frac{1}{2} - \gamma)$. The $L_1$ distance between these distributions is small enough such that the adversary can choose distributions $q_1$ and $q_2$, over number of ones in the adversarial batches, such that

$$(1 - \beta)B(n, \frac{1}{2} + \gamma) + \beta q_1 = (1 - \beta)B(n, \frac{1}{2} - \gamma) + \beta q_2.$$  

We skip the simple proof of this statement. Hence, if the good batches are distributed as $B(n, \frac{1}{2} + \gamma)$ then adversary chooses $q_1$ as distribution of the adversarial batches and if good batches are distributed as $B(n, \frac{1}{2} - \gamma)$ then adversary chooses $q_2$ and in both the cases the resultant joint distribution of all the batches is same. Hence the two cases are indistinguishable. Then $||\text{Bern}(\frac{1}{2} + \gamma) - \text{Bern}(\frac{1}{2} - \gamma)||_1 = 2\frac{1}{2} + \gamma - (\frac{1}{2} - \gamma) = 4\gamma$. From the triangle inequality, any estimate $\hat{p}$ will be at a distance at least $\min\{||\text{Bern}(\frac{1}{2} + \gamma) - \hat{p}||, ||\text{Bern}(\frac{1}{2} - \gamma) - \hat{p}||\} \geq 2\gamma$ from one of the possible distribution of the batches.

Theorem implies that even with access to infinitely many batches, even for an alphabet of size as small as 2, no algorithm can estimate $p$ to $L_1$ distance below $\Omega(\beta/\sqrt{n})$ with probability $1/2$. In the next section, we show that this bound can essentially be met by a polynomial-time algorithm.

### 1.4 Results Summary

In section 4 we derive a polynomial-time algorithm that for $\beta < 0.5$ returns an estimate $p^*$ of $p$. To state concrete constant factors, the following theorem characterizes the algorithm’s performance for $\beta \leq 0.4$.

**Theorem 2.** For any given $\beta \leq 0.4$, $n \geq 1$ and $k \geq 2$, and $m \geq \max\left\{\frac{72k}{\beta^2 \ln(6e/\beta)}, \frac{192 \ln^2 n}{\beta \ln(e/\beta)}(k + \log \log n)\right\}$, Algorithm [1] runs in time polynomial in all parameters $k$, $n$, and $1/\beta$, and its estimate $p^*$ satisfies $||p^* - p||_1 \leq 30\beta \sqrt{\frac{\ln(6e/\beta)}{n}}$ with probability $\geq 1 - 8e^{-k}$.

Observe that $p^*$ approximates $p$ to $L_1$ distance that is within a small factor of $O(\sqrt{\log(1/\beta)})$ from the lower bound in Theorem 1.

Even in the absence of an adversary, estimating a $k$-element distribution to $L_1$ distance $\epsilon$ requires $\Omega(k/\epsilon^2)$ samples. Hence estimating $p$ to the distance $O(\frac{\beta}{\sqrt{n}})$ achieved in the above theorem requires $\tilde{O}(kn/\beta^2)$ samples. Note that Algorithm [1] achieves this sample complexity up to log factors even in the presence of adversarial samples.

Moreover, the algorithm uses $nm = \tilde{O}(kn/\beta^2)$ samples, which is within log factors from the best possible sample complexity to estimate a $k$-element distribution, even without the presence of adversarial data, up to a distance achieved by our algorithm in Theorem 2.

### 1.5 Preliminaries

We introduce notation that will help outline our approach and will be used throughout the paper.

Let $p(S) = \sum_{i \in S} p(i)$ denote the probability of a set $S \subseteq [k]$. To estimate $p(S)$, let $n_b(S)$ denote the number of samples in batch $b$ that belong to $S$. Let $\tilde{p}_b(S) = n_b(S)/n$ be the empirical estimate of $p(S)$ from batch $b$. Note that for any $S \subseteq [k]$ and $b \in G$, $n_b(S)$ is distributed according to binomial distribution $n_b(S) \sim B(n, p(S))$. Therefore, for $b \in G$,

$$E[\tilde{p}_b(S)] = p(S) \quad \text{and} \quad E[(\tilde{p}_b(S) - p(S))^2] = \frac{p(S)(1 - p(S))}{n}.$$
For any subset $U \subseteq B$ of the batches, let $\tilde{p}_U = (\tilde{p}_U(1), \tilde{p}_U(2), \ldots, \tilde{p}_U(k)) \in \mathbb{R}^k$ denote the empirical estimate of $p$ from the batches in $U$, where, here and below, we abbreviate singleton set such as $\{j\}$ by $j$. For any $S \subseteq [k]$, let
\[ \tilde{p}_U(S) = \frac{1}{|U|} \sum_{b \in U} \tilde{p}_b(S), \]
and define the sample variance of $\tilde{p}_b(S)$ in $U$ to be
\[ V_U(S) = \frac{1}{|U|} \sum_{b \in U} (\tilde{p}_b(S) - \tilde{p}_U(S))^2. \]
For $r \in [0, 1]$, let $V(r) \triangleq \frac{r(1-r)}{n}$ be the variance of a random variable $\frac{X}{n}$, where $X \sim B(n, r)$. Observe that $V(r) \leq \frac{1}{4n}$. Also $|V'(r)| \leq 1/n, \forall r \in [0, 1]$, hence
\[ \forall r, s \in [0, 1], |V(r) - V(s)| \leq \frac{|r - s|}{n}. \tag{1} \]
As above, if the samples in batches $b \in U$ were distributed according to $B(n, p)$ then the variance of the probability estimates $\tilde{p}_b(S)$ would be $V(p(S))$. And since $\tilde{p}_U(S)$ is an estimate of $p(S)$ based on batches in $U$, then $V(\tilde{p}_U(S))$ can be used as an estimate of the variance of $\tilde{p}_b(S)$ for the batches in $U$. We refer to it as the mean-induced variance estimate.

### 1.6 Roadmap of the Paper

If the set $G$ of the good batches is known, then using the standard empirical approach, $p$ can be estimated by $\hat{p}_G(i) = \frac{1}{|G|} \sum_{b \in G} \tilde{p}_b(S), \forall i \in [k]$. This approach guarantees that w.h.p., $||\hat{p}_G - p||_1 \leq O(\sqrt{k/|G|})$ that diminishes as the number of good batches $|G|$ increases.

Since $G$ is unknown, if we ignore the presence of the adversarial batches and estimate $p$ as $\tilde{p}_B$, the adversary can choose samples in its batches $A$ such that for some $S \subseteq [k]$, $|\tilde{p}_A(S) - p(S)| = \Omega(1)$. This will result in $||\tilde{p}_B - p||_1 \geq \Omega(1/\beta) \geq \Omega(\beta)$.

Next, we briefly describe the algorithm [12] proposed for $\beta \leq 1/900$, and show why it has exponential time complexity. For every $S \subseteq [k]$, they use a linear program to estimate $p(S)$ to within $\pm O(\beta/\sqrt{n})$. They further use a linear program to estimate a distribution over $[k]$ that is consistent with all the $2^k$ subset probability estimates. Since the algorithm involves computing $k^k$ probabilities, its running time is exponential in $k$.

We note that the first linear program, estimating the probability $p(S)$ of each subset $S \subseteq [k]$, can be replaced by simply the median of the estimates $\tilde{p}_b(S)$ of $p(S)$ over all the batches. The median approximates $p(S)$ to within $\pm O(\beta/\sqrt{n} + \frac{1}{n})$, which for $n = \Omega(1/\beta^2)$ achieves the desired $\Theta(\beta/\sqrt{n})$ accuracy.

We next review our algorithm that estimates $p$ w.h.p. to a distance $O(\beta \sqrt{\frac{\log(1/\beta)}{n}})$ with running time polynomial in all parameters. Like the standard estimate, the algorithm estimates the probability of each symbol by its empirical count, except that instead of considering all batches, the count is taken over a subset $U \subseteq B$ of the batches.

In Section 3, we prove a key robustness to deletion property of good batches, that holds with high probability. Their empirical statistics remain essentially the same when any small fraction, even adversarial, of these batches is removed. Specifically, for every subset $S \subseteq [k]$ and every collection of good batches $U_G \subseteq G$ that excludes at most $|G \setminus U_G| \leq O(\beta|G|)$ of the good batches, w.h.p. $|\tilde{p}_{U_G}(S) - p(S)| \leq O(\beta/\sqrt{n})$ and $|\hat{V}_{U_G}(S) - V(p(S))| \leq O(\beta/n)$. This property will play a key role in recovering $p$. Similar properties have been used in past robust-recovery algorithms but in different settings [3][5][13]. These bounds obtained are essential to establish our algorithm’s success.

For the algorithm to successfully learn $p$ to $L_1$ distance $\hat{O}(\beta/\sqrt{n})$ it suffices to find a collection $U$ of batches such that $|\tilde{p}_U(S) - p(S)| \leq \hat{O}(\beta/\sqrt{n})$ for all $S \subseteq [k]$. To achieve this goal, the algorithm finds a collection $U$ of batches that satisfies the following two objectives, that we argue are stronger:
1. \( U \) contains all but a small \( O(\beta) \) fraction of the good batches.

2. For every \( S \subseteq [k] \), the sum of the squared deviations of estimates of the adversarial batches \( \bar{p}_b(S) \) from \( p(S) \) over all adversarial batches in \( U \) is small. Specifically, \( \sum_{b \in U \cap A}(\bar{p}_b(S) - p(S))^2 \leq O(|A|/n) \).

The first objective and the robustness to deletion property ensure that \( \bar{p}_{U \cap G}(S) \) is close to \( p(S) \). The bound on the adversarial squared deviations in the second objective ensures that the batches in \( U \cap A \) have little influence on the overall empirical estimate \( \bar{p}_U(S) \).

To find a collection \( U \) that satisfies the two objectives, the algorithm starts with the set \( U = B \) and iteratively updates it by deleting batches from it while ensuring that Objective 1 is always met. To achieve this objective, in each update, the algorithm removes from \( U \) a collection of batches, most of which are adversarial. The removal of adversarial batches ensures that each update moves the collection \( U \) closer to Objective 2.

To update \( U \), the algorithm finds a set \( S \subseteq [k] \) not meeting Objective 2. It can be shown that for such \( S \), the estimates \( \bar{p}_b(S) \) are far from \( p \) for many adversarial batches \( b \in U \cap A \). Since for most good batches, the estimate \( \bar{p}_b(S) \) is near \( p(S) \), the algorithm can delete some batches in \( U \) for which the estimates \( \bar{p}_b(S) \) are far from \( p(S) \) in a way that more adversarial than good batches will be removed, ensuring that the smaller collection \( U \) continues to satisfy Objective 1. In deciding which batches to delete, the algorithm replaces the unknown \( p(S) \) by the median of \( \bar{p}_b(S) \).

The algorithm stops updating \( U \) when Objective 2 is met. As argued earlier when both the objectives are met, the empirical distribution of the samples of the remaining batches in \( U \) approximate \( p \) to the desired distance.

The main remaining challenge is to efficiently identify a subset \( S \) that does not meet Objective 2. There are two challenges. First, we don’t know \( U \cap A \), and second, there are \( 2^k \) different subsets \( S \subseteq [k] \). We propose a novel way to efficiently identify such subsets \( S \). When Objective 2 is not met due to some subset \( S \subseteq [k] \), the estimate \( \bar{p}_b(S) \) of adversarial batches has a high squared deviation around \( p(S) \), pushing the overall sample variance \( \bar{V}_U(S) \) higher than its expected value for good batches. A typical approach would find such a subset for which \( \bar{V}_U(S) \) is the highest, but this method is sub-optimal and falls short of achieving the desired guarantees, as for the good batches, different subsets \( \bar{p}_b(S) \) have different variance. Hence would not be able to detect the subsets corrupted by adversary for which the original variance was smaller.

To tackle this issue we use mean-induced estimate \( \bar{V}(\bar{p}_U(S)) \), as another estimator of the variance of \( \bar{p}_b(S) \). If all the batches were good then the difference between the sample variance \( \bar{V}_U(S) \) and \( \bar{V}(\bar{p}_U(S)) \) would be small. Since that the sample variance \( \bar{V}_U(S) \) is computed using the second moment, whereas \( \bar{V}(\bar{p}_U(S)) \) is computed using the first moments \( \bar{p}_b(S) \); due to this second moment dependence adversarial batches affect the sample variance \( \bar{V}_U(S) \) more severely. Using this observation, we show that difference between the two estimates of variance for a subset \( S \subseteq [k] \), \( \bar{V}_U(S) - \bar{V}(\bar{p}_U(S)) \) is high if and only if for \( S \) Objective 2 is not met. This reduces the problem to find a subset \( S \) such that the difference between two estimates \( \bar{V}_U(S) - \bar{V}(\bar{p}_U(S)) \) corresponding to \( S \) is large. This new objective function now depends on all the batches in \( U \) rather than on an unknown collection \( U \cap A \). In Section 2 we pose this problem of finding a subset maximizing this difference as an optimization problem and suggest an existing approximation algorithm to solve this optimization problem.

1.7 Organization of the Paper

In Section 2, we discuss a known approximation algorithm that, given a a collection \( U \) of batches, can identify a subset \( S \subseteq [k] \) such that the difference \( \bar{V}_U(S) - \bar{V}(\bar{p}_U(S)) \) between the sample variance and its estimate is high. Section 3 shows that with high probability, the collection of good batches satisfies the robustness to deletion and a few other useful properties. Section 4 assumes that the good batches display these properties, and proposes an algorithm that can always estimate \( p \) to a small \( L_1 \) distance for any choice of adversarial batches.
2 Efficient Detection

Recall that \( \bar{V}_U(S) \) is the empirical variance of the estimate \( \bar{p}_b(S) \) for a subset \( S \). And \( \bar{V}(\bar{p}_U(S)) \) is the mean-induced estimate of variance. We describe a polynomial-time algorithm that given a collection \( U \) of batches finds a subset \( \mathcal{S}'_U \) such that

\[
|\bar{V}_U(S'_U) - \bar{V}(\bar{p}_U(S'_U))| \geq 0.5 \max\{|\bar{V}_U(S) - \bar{V}(\bar{p}_U(S))| : S \subseteq [k]\}.
\]

Next for a subset of batches \( U \) we construct two covariance matrices \( C^\mathsf{EV}_U \) and \( C^\mathsf{EM}_U \) of size \( k \times k \). The first covariance matrix, \( C^\mathsf{EV}_U \), is sample covariance in \( U \), with entries

\[
C^\mathsf{EV}_U(j, l) = \frac{1}{|U|} \sum_{b \in U} (\bar{p}_b(j) - \bar{p}_U(j))(\bar{p}_b(l) - \bar{p}_U(l)) \quad \text{for} \quad j, l \in [k].
\]

The second covariance matrix \( C^\mathsf{EM}_U \), is an expected covariance matrix of a multinomial random variable (normalized by \( n \)) distributed according to the empirical distribution \( \bar{p}_U \) and the other parameter being \( n \). Therefore, its entries are

\[
C^\mathsf{EM}_U(j, l) = \frac{\bar{p}_U(j)\bar{p}_U(l)}{n} \quad \text{for} \quad j, l \in [k], \ j \neq l, \ \text{and} \quad C^\mathsf{EM}_U(j, j) = \frac{\bar{p}_U(j)(1 - \bar{p}_U(j))}{n}.
\]

The last matrix \( D_U \) is the difference between the two matrices:

\[
D_U = C^\mathsf{EV}_U - C^\mathsf{EM}_U.
\]

For a vector \( x \in \{0, 1\}^k \), let

\[
S(x) \triangleq \{ j \in [k] : x(j) = 1 \},
\]

be the subset of the alphabet \([k]\) corresponding to the vector \( x \).

Observations

1. The sum of elements in any row and or column for both covariance matrices, and hence also for the difference matrix, is zero, hence

\[
C^\mathsf{EV}_U \mathbf{1} = C^\mathsf{EM}_U \mathbf{1} = D_U \mathbf{1} = 0.
\]

\textbf{Proof.} We prove for \( C^\mathsf{EV}_U \), the proof for \( C^\mathsf{EM}_U \) is similar.

\[
\sum_{l \in [k]} C^\mathsf{EV}_U(j, l) = \frac{1}{|U|} \sum_{l \in [k]} \sum_{b \in U} (\bar{p}_b(j) - \bar{p}_U(j))(\bar{p}_b(l) - \bar{p}_U(l)) = \sum_{b \in U} (\bar{p}_b(j) - \bar{p}_U(j)) \sum_{l \in [k]} (\bar{p}_b(l) - \bar{p}_U(l)) = \sum_{b \in U} (\bar{p}_b(j) - \bar{p}_U(j))(1 - 1) = 0.
\]

2. It is easy to verify that for any vector \( x \in \{0, 1\}^k \),

\[
\langle C^\mathsf{EV}_U, xx^T \rangle = \frac{1}{|U|} \sum_{b \in U} (\bar{p}_b(S(x)) - \bar{p}_U(S(x)))^2 = \bar{V}_U(S(x)),
\]

the sample variance of \( \bar{p}_b(S(x)) \) in \( U \). Similarly,

\[
\langle C^\mathsf{EM}_U, xx^T \rangle = \frac{\bar{p}_U(S(x))(1 - \bar{p}_U(S(x)))}{n} = \bar{V}(\bar{p}_U(S(x))),
\]

which is the mean-induced estimate of variance of \( \bar{p}_b(S(x)) \) for batches in \( U \), defined earlier. Therefore,

\[
\langle D_U, xx^T \rangle = \langle C^\mathsf{EV}_U - C^\mathsf{EM}_U, xx^T \rangle = \bar{V}_U(S(x)) - \bar{V}(\bar{p}_U(S(x))).
\]

3. Note that \( y \rightarrow \frac{1}{2}(y + 1) \) is a 1-1 mapping from \( \{-1, 1\}^k \rightarrow \{0, 1\}^k \), and that

\[
\langle C^\mathsf{EV}_U, \frac{1}{2}(y + 1)\frac{1}{2}(y + 1)^T \rangle = \langle C^\mathsf{EV}_U, \frac{1}{4}(yy^T + 1y^T + y1^T + 11^T) \rangle = \frac{1}{4} \langle C^\mathsf{EV}_U, yy^T \rangle.
\]


Let 
\[ y = \arg\max_{y \in \{-1, 1\}^k} |\langle D_U, y y^T \rangle|. \]
Then from \( y \) one can recover a set \( S(x) \), with \( x = \frac{1}{2}(y + 1) \), maximizing
\[ |\bar{V}_U(S(x)) - \bar{V}(\bar{p}_U(S(x)))|. \]

Similar optimization problem arises in a number of different settings, and is NP-hard in general. In [1] Alon et al. derives a polynomial-time approximation algorithm for the above optimization problem. The algorithm first uses a semi-definite relaxation of the problem and then uses randomized integer rounding techniques based on Grothendieck’s Inequality. Their algorithm recovers \( y_U \) such that
\[ |\langle D_U, y_U y_U^T \rangle| \leq 0.56 \max_{y \in \{-1, 1\}^k} |\langle D_U, y y^T \rangle|. \]

Let \( x_U = \frac{1}{2}(y + 1) \). Then from observation 3 it follows that
\[ |\langle D_U, x_U x_U^T \rangle| \leq 0.56 \max_{x \in \{0, 1\}^k} |\langle D_U, x x^T \rangle|. \]

Therefore for \( S_U^* = S(x_U) \) we get
\[ |\bar{V}_U(S_U^*) - \bar{V}(\bar{p}_U(S_U^*)))| \geq 0.56 \max_{S \subseteq [k]} |\bar{V}_U(S) - \bar{V}(\bar{p}_U(S))|. \]

### 3 Properties of the Collection of Good Batches

The next lemma show that the empirical mean and variance are robust to removal of a small fraction of the batches, i.e. even after deleting any small fraction of good batches the empirical mean and the variance approximate the distribution mean and the variance well enough for all subsets \( S \subseteq [k] \).

**Lemma 1.** For any \( 0 < \epsilon < 1/4 \), and \( |G| \geq \frac{k}{e^{2e/\epsilon}} \). Then \( \forall S \subseteq [k] \) and \( \forall U_G \subseteq G \) of size \( |U_G| \geq (1 - \epsilon)|G| \), with probability \( \geq 1 - 6e^{-k} \),
\[ \left| \bar{p}_{U_G}(S) - p(S) \right| \leq 3e \sqrt{\frac{\ln(e/\epsilon)}{n}} \] (2)
and
\[ \left| \frac{1}{|U_G|} \sum_{b \in U_G} (\bar{p}_b(S) - p(S))^2 - \bar{V}(p(S)) \right| \leq 32e \frac{\ln(e/\epsilon)}{n}. \] (3)

**Proof.** For a good batch \( b \in G \) and \( S \subseteq [k] \), \( n_b(S) \) is distributed according to Binomial distribution and can be thought of as the sum of \( n \) i.i.d. Bernoulli random variables. Since Bernoulli random variables are \( \sim \text{subG}(1/4) \), then the (centered) average of \( n \) of them will satisfy \( \bar{p}_b(S) - p(S) \sim \text{subG}(n/4n^2) \). subG(\( \cdot \)) is used to denote a sub-Gaussian distribution. From Hoeffding’s inequality,
\[
\Pr \left[ |G| |\bar{p}_G(S) - p(S)| \geq |G| e \sqrt{\frac{\ln(e/\epsilon)}{n}} \right] = \Pr \left[ \left| \sum_{b \in G} (\bar{p}_b(S) - p(S)) \right| \geq |G| e \sqrt{\frac{\ln(e/\epsilon)}{n}} \right] \\
\leq 2e^{-2 \frac{|G|^2 e^2 \ln(e/\epsilon)}{2n}} = 2e^{-2|G|^2 \frac{e^2 \ln(e/\epsilon)}{n}} \leq e^{-2k}. \] (4)

Similarly, for a fix subset \( V_G \subseteq G \) of size \( 1 \leq |V_G| \leq |G| \),
\[
\Pr \left[ |V_G| |\bar{p}_{V_G}(S) - p(S)| \geq |V_G| e \sqrt{\frac{\ln(e/\epsilon)}{n}} \right] = \Pr \left[ \left| \sum_{b \in V_G} (\bar{p}_b(S) - p(S)) \right| \geq |V_G| e \sqrt{\frac{\ln(e/\epsilon)}{n}} \right] \\
\leq 2e^{-2 \frac{e^2 \ln(e/\epsilon)^2}{|V_G|^2}} \leq 2e^{-2|G| \ln(e/\epsilon)},
\]
where the last inequality used \( |V_G| \leq |G| \). Next, the number of subsets (non-empty) of \( G \) with size \( \leq |G| \) is bounded by
\[
\sum_{j=1}^{|G|} \left( \begin{array}{c} |G| \\ j \end{array} \right) \leq |G| \left( \begin{array}{c} |G| \\ |G| \end{array} \right) \leq e^{3|G|} \left( \frac{3|G|}{|G|} \right)^{|G|} < e^{3|G| \ln(e/\epsilon)}. \]
where last of the above inequality used \( \ln(e|G|) < e|G|/2 \) and \( \ln(e/e) \geq 1 \). Then, using the union bound, \( \forall V_G \subseteq G \) such that \( |V_G| \leq e|G| \), we get

\[
\Pr \left[ |V_G| \cdot |\bar{p}_{V_G}(S) - p(S)| \geq e|G| \sqrt{\frac{\ln(e/e)}{n}} \right] \leq 2e^{-\frac{e}{2}e|G|\ln(e/e)} < 2e^{-k} < 2e^{-2k}. \tag{5}
\]

For any subset \( U_G \subseteq G \) with \( |U_G| \geq (1 - \epsilon)|G| \),

\[
\left| \sum_{b \in U_G} (\bar{p}_b(S) - p(S)) \right| = \left| \sum_{b \in G} (\bar{p}_b(S) - p(S)) - \sum_{b \in G/U_G} (\bar{p}_b(S) - p(S)) \right|
\leq \left| \sum_{b \in G} (\bar{p}_b(S) - p(S)) \right| + \left| \sum_{b \in G/U_G} (\bar{p}_b(S) - p(S)) \right|
\leq |G| \times |\bar{p}_V(G) - p(S)| + \max_{V_G:|V_G| \leq e|G|} |V_G| \times |\bar{p}_{V_G}(S) - p(S)|
\leq 2e|G| \sqrt{\frac{\ln(e/e)}{n}},
\]
with probability \( \geq 1 - 2e^{-2k} - 2e^{-2k} \geq 1 - 4e^{-2k} \). Then

\[
|\bar{p}_{U_G}(S) - p(S)| = \frac{1}{|U_G|} \left| \sum_{b \in U_G} (\bar{p}_b(S) - p(S)) \right| \leq \frac{2e|G|}{|U_G|} \sqrt{\frac{\ln(e/e)}{n}}
\leq \frac{2e}{(1-\epsilon)\sqrt{\frac{\ln(e/e)}{n}}} < 3e \sqrt{\frac{\ln(e/e)}{n}},
\]
with probability \( \geq 1 - 4e^{-2k} \). The last step used \( \epsilon \leq 1/4 \). Since there are \( 2^k \) different choices for \( S \subseteq [k] \), from the union bound we get,

\[
\Pr \left[ \bigcup_{S \subseteq [k]} \left\{ |\bar{p}_{V_G}(S) - p(S)| > 4e \sqrt{\frac{\ln(e/e)}{n}} \right\} \right] \leq 4e^{-2k} \times 2^k = 4e^{-k}.
\]

This completes the proof of (2).

Let \( Y_b = (\bar{p}_b(S) - p(S))^2 - V(p(S)) \). For \( b \in G \), \( \bar{p}_b(S) - p(S) \sim \text{subE}(1/4n) \), therefore

\[
(\bar{p}_b(S) - p(S))^2 - E(\bar{p}_b(S) - p(S))^2 = Y_b \sim \text{subE}(\frac{16}{4n}) = \text{subE}(\frac{4}{n}).
\]

Here subE is sub exponential distribution \( \mathbb{I} \). Then Bernstein’s inequality gives:

\[
\Pr \left[ \left| \sum_{b \in G} Y_b \right| \geq 8|G| \frac{\epsilon}{n} \ln(e/e) \right] \leq 2e^{-\frac{1}{2} \left( \frac{8|G| \epsilon \ln(e/e)}{4n} \right)^2} = 2e^{-2|G|\epsilon^2 \ln^2(e/e)} \leq 2e^{-2k}.
\]

Next, for a fix subset \( V_G \subseteq G \) of size \( 1 \leq |V_G| \leq e|G| \),

\[
\Pr \left[ \left| \sum_{b \in V_G} Y_b \right| \geq 16\epsilon|G| \frac{\ln(e/e)}{n} \right] \leq 2e^{-\frac{16\epsilon|G| \ln(e/e)}{2^k \ln(e/e)}}
\leq 2e^{-2\epsilon|G| \ln(e/e)}.
\]

Then following the same steps as in the proof of (2) one can complete the proof of (3). \( \blacksquare \)

Next we divide the batches into the subsets based on the distance of their estimate of \( p(S) \) from the actual value.

\[
I_j(S) = \{ b \in G : |\bar{p}_b(S) - p(S)| \geq (2^j - 1)\sqrt{\ln(e/e)/n} \}.
\]

First note that \( I_j = \phi \) for \( j \geq \log(\sqrt{n} + 1) \), since \( |\bar{p}_b(S) - p(S)| \leq 1 \). Next lemma upper bounds the size of the set \( I_j(S) \cap G \), and show that set size \( I_j \) decreases sharply with \( j \), hence the tails become smaller for larger \( j \).
Lemma 2. For any \( \epsilon \in (0, 1/2) \) and \( |G| \geq \frac{96 \log^2(n)}{\epsilon^4 \ln(1/\epsilon)} (k + \log \log n) \), \( \forall S \subseteq [k] \) and \( \forall j \in \{2, 3, 4, \ldots, \log(\sqrt{n} + 1)\} \), with probability \( \geq 1 - e^{-k} \),
\[
|I_j(S)| \leq |G| \frac{e^{2 - 2j}}{24j^2}.
\]

Proof. From Hoeffding’s inequality, for a fix \( j \geq 2 \) and \( S \subseteq [k] \),
\[
\Pr \left[ |\bar{b}_h(S) - p(S)| \geq (2j - 1)\sqrt{\ln(1/\epsilon)/n} \right] \leq 2e^{-2(2j - 1)^2 \ln(1/\epsilon)} = 2\left(\frac{\epsilon}{e}\right)^{2(2j - 1)^2}.
\]
Let \( \mathbb{1}_j(i) \) be the indicator random variable for the event \( b \in I_j \). Therefore, for \( b \in G \), \( E[\mathbb{1}_j(i)] \leq 2\left(\frac{\epsilon}{e}\right)^{2(2j - 1)^2} \). Let \( \mu = |G|E[\mathbb{1}_j(i)] \) and \( \mu(1 + \delta) = |G|\left(\frac{e}{e}\right)^{2 - 2j} \). Then
\[
(1 + \delta) \geq \frac{|G|}{2}\left(\frac{e}{e}\right)^{2 - 2j} \cdot \frac{1}{24j^2}.
\]
where inequality (a) used \( j \leq 2 \) and \( \epsilon / \epsilon \geq 4 \), and inequality (b) follows since the expression is an increasing function of \( j \) for \( j \geq 2 \), and attains minimum at \( j = 2 \). Then using the Chernoff bound,
\[
\Pr \left( \sum_{b \in G} \mathbb{1}_j(i) \geq (1 + \delta)\mu \right) \leq \left(\frac{e\delta}{1 + \delta}\right)^{(1 + \delta)\mu} \leq \left(\frac{e(1 + \delta)}{1 + \delta}\right)^{(1 + \delta)\mu} = e^{-\mu(1 + \delta) \ln \left(\frac{1 + \delta}{\epsilon}\right)}
\]
\[
\leq e^{-\frac{|G|}{24j^2} \ln \left(\frac{1 + \delta}{\epsilon} (2j - 1)^2\right)}
\]
\[
= e^{-\frac{|G|}{24j^2} \ln \left(\frac{1 + \delta}{\epsilon} \right)} \leq e^{-\frac{|G|}{24j^2} \ln \left(\frac{1 + \delta}{\epsilon} \right)} \leq e^{-2(k + \log \log n)} \leq e^{-2k / \log n}
\]
Taking the union over \( j \), we get \( \forall j \geq 2 \)
\[
|I_j(S)| \leq |G| \frac{e^{2 - 2j}}{24j^2}.
\]
with probability \( \geq 1 - e^{-2k} \). Finally taking the union bound of the complement of the above event, over all \( 2^k \) subsets gives the statement of the Lemma.

The last lemma in the section shows that for most of the good batches their estimate \( \bar{b}_h(S) \) cluster within a small interval and \( p(S) \) in within interval.

Lemma 3. For any \( \epsilon \in (0, 1/4] \) and \( |G| \geq 12k / \epsilon \), \( \forall S \subseteq [k] \), with probability \( \geq 1 - e^{-k} \),
\[
|\{ b \in G : |\bar{b}_h(S) - p(S)| \leq \sqrt{\ln(1/\epsilon)/n}\}| \geq |G| (1 - \epsilon).
\]

Proof. From Hoeffding’s inequality, for \( b \in G \) and \( S \subseteq [k] \),
\[
\Pr \left[ |\bar{b}_h(S) - p(S)| \geq \ln(1/\epsilon)/n \right] \leq 2e^{-2\ln(1/\epsilon)} \leq 2\epsilon^2 \leq \epsilon / 2.
\]
Let \( \mathbb{1}_j(i) \) be the indicator random variable for the event \( |\bar{b}_h(S) - p(S)| \geq \sqrt{\ln(1/\epsilon)/n} \). Therefore, for \( b \in G \), \( E[I_h] \leq 2\epsilon / 2 \). Using the Chernoff bound,
\[
\Pr \left( \sum_{b \in G} \mathbb{1}_b \geq \epsilon |G| \right) \leq e^{-\epsilon^2 |G| / 2} \leq e^{-2k}.
\]
Taking the union bound over all \( 2^k \) subsets \( S \) give the statement of the Lemma.
4 Estimating Distribution in the Presence of Adversarial Batches

In this section, we focus on the case when good batches form the majority and the fraction of good batches $\alpha \geq 0.6$. This implies that $\beta \leq 0.4$. We also assume that the batch-size $n \geq 9$, as for a constant $n$ the results are trivial and achieved by an empirical estimate.

Let $\kappa \triangleq \frac{\beta \ln(6e/\beta)}{n}$. For the remainder of the section, the algorithm assumes that the samples in good batches $G$ satisfies the following properties. For all subsets of good batches, $U_G \subseteq G$, such that $|U_G| \geq (1 - \beta/6)|G|$ and $\forall S \subseteq [k]$, the following conditions hold

$$\left| \hat{p}_{U_G}(S) - p(S) \right| \leq \frac{\beta}{2} \sqrt{\frac{\ln(6e/\beta)}{n}} = \frac{1}{2} \sqrt{\beta \kappa}$$

(6)

$$\left| \sum_{b \in U_G} (\hat{p}_b(S) - p(S))^2 - |U_G|V(p(S)) \right| \leq 6|U_G|\frac{\beta \ln(6e/\beta)}{n} = 6|U_G|\kappa \leq 6|U|\kappa$$

(7)

$$\left\{ b \in G : |\hat{p}_b(S) - p(S)| \leq \sqrt{\ln(6)/n} \right\} \geq \frac{5}{6}|G|$$

(8)

$$\left\{ b \in G : |\hat{p}_b(S) - p(S)| \geq (2^j - 1)\sqrt{\ln(e/\beta)/n} \right\} \leq |G|\frac{2^j - 2j}{24j^2}, \text{ for } j \geq 2.$$

(9)

For $|G| \geq \max\{36k/\beta^2 \ln(6e/\beta), \frac{96k^2 \log^2(n)}{\beta \ln(e/\beta)}\}$, the above conditions hold with probability $\geq 1 - O(e^{-k})$. The first two conditions hold from Lemma 1 by choosing $\epsilon = \beta/6$. The condition in (8) holds from Lemma 2. The last condition holds from Lemma 2.

Next we present the algorithm that estimates $p$ to a small $L_1$ distance provided the above four conditions above are met.

4.1 Algorithm

Let $\text{med}(S)$ be the median of the set of estimates $\{\hat{p}_b(S) : b \in B\}$. The median of these estimates $\hat{p}_b(S)$ is used as an estimate of $p(S)$. Next, we divide the batches into the subsets based on how far their estimate $\hat{p}_b(S)$ is from the median.

$$I_1^m(S) = \{ b \in B : |\hat{p}_b(S) - \text{med}(S)| \leq 4\sqrt{\ln(6e/\beta)/n} \}$$

and for $j > 1$,

$$I_j^m(S) = \{ b \in B : |\hat{p}_b(S) - \text{med}(S)| \in (2^j \sqrt{\ln(6e/\beta)/n}, 2^{j+1} \sqrt{\ln(6e/\beta)/n}) \}.$$

Algorithm 1 Distribution Estimation Algorithm

**Input** : All batches $b \in B$, batch size $n$, alphabet size $k$, and $\beta$.

**Output** : Estimate $p^*$ of the distribution $p$.

1: $U \leftarrow B$.
2: while True do
3: Using the detection algorithm obtain the set $S_U^U$.
4: if $(\text{Var}_U(S_U^U) - V(\hat{p}_U(S_U^U))) \leq 40\kappa$ then Break;
5: end if
6: Update $\leftarrow$ False.
7: for $j \leftarrow 2$ to $\log(\sqrt{n} + 1)$ do
8: if $|I_j^m(S_U^U) \cap U| \geq |G|\frac{2^j - 2j}{24j^2}$ then
9: $U \leftarrow U \setminus I_j^m(S_U^U) \cap U$.
10: Update $\leftarrow$ True.
11: end if
12: end for
13: if (Update = False) then Break;
14: end if
15: end while
16: return ($p^* \leftarrow \hat{p}_U$).
4.2 Analysis of the Algorithm

In this sub-section, we show that the algorithm indeed learns the distribution. We will assume that the good batches satisfy the properties in the last subsection, and under those conditions, we show that the above algorithm always achieves the bounds in Theorem\[2\] Since the properties assumed of good batches has been shown to hold with high probability in the previous section, it implies that the algorithm succeeds with high probability. To prove this we establish each of the following claims:

1. In step (9), when the algorithm updates $U$, in each update, among the batches that are being removed, the ratio of good batches to bad batches is at most 0.1.
2. If $(\text{Var}_{\tilde{U}}(S) - V(\bar{p}_U(S)))$ is large for some $S$, then it implies that $\sum_{b \in U_A} (\bar{p}_b(S) - p(S))^2$ is large. Specifically $\text{Var}_{\tilde{U}}(S) - V(\bar{p}_U(S))) \geq 40\kappa$, then $\sum_{b \in U_A} (\bar{p}_b(S) - p(S))^2 \geq 30|U|\kappa$.
3. If for some subset $S$, $\sum_{b \in U_A} (\bar{p}_b(S) - p(S))^2 \geq 30|U|\kappa$, then
   
   $$|\mathcal{I}^m_j(S) \cap U| \geq |\mathcal{I}^m_j(S) \cap U_A| \geq |G|\frac{\beta^2 - 2^j}{2j^2}$$
   
   for at-least one $j \in \{2, ..., \log(\sqrt{n} + 1)\}$. Hence the algorithm exits from the while loop only via Step 4.
4. From Claim 3 above it follows that in every iteration of the while loop, the algorithm ends up deleting at-least $\min\{|G|\frac{\beta^2 - 2^j}{2j^2} : j \in \{2, ..., \log(\sqrt{n} + 1)\} \} \geq |G|\frac{\beta}{2n \log^2 n}$ batches. Then claim 1 implies that the algorithm deletes at-least $\frac{10G\beta}{22n \log^2 n}$ bad batches in each iteration. And since the number of bad batches are upper bounded by $\beta m$, the algorithm iterates over the while loop for at-most $\frac{\beta m}{\log^2(n)} \leq 4n \log^2 n$ times.
5. If the Algorithm exits from the while loop only in Step 5, from the properties of the detection algorithm we have that when the algorithm terminates $(\text{Var}_{\tilde{U}}(S) - V(\bar{p}_U(S))) \leq 80\kappa$ for all $S$.
6. If $\sum_{b \in U_A} (\bar{p}_b(S) - p(S))^2$ is small then $\sum_{b \in U_A} (\bar{p}_b(S) - p(S))^2$ is small as well. Specifically, if $(\text{Var}_{\tilde{U}}(S) - V(\bar{p}_U(S))) \leq 80\kappa$, then $\sum_{b \in U_A} (\bar{p}_b(S) - p(S))^2 \leq 169|U|\kappa$.
7. If $\sum_{b \in U_A} (\bar{p}_b(S) - p(S))^2$ is small then $|\bar{p}_U(S) - p(S)|$ will be small. In particular, if $\sum_{b \in U_A} (\bar{p}_b(S) - p(S))^2 \leq 169|U|\kappa$, then $|\bar{p}_U(S) - p(S)| \leq 15\beta \sqrt{\frac{\ln(6e/\beta)}{n}}$. Hence, the Algorithm achieves the guarantees in Theorem\[2\]

Note that Claims 4 and Claim 5 above are true given the others. We start by proving Claim 1 above, which implies Lemma\[5\]. Next we prove Lemma\[6\] which implies the last of the above claims. And Lemma\[7\] and Corollary\[8\] which relates $(\text{Var}_{\tilde{U}}(S) - V(\bar{p}_U(S)))$ and $\sum_{b \in U_A} (\bar{p}_b(S) - p(S))^2$, that proves Claims 2 and 6, are the main results of the section. Finally Lemma\[9\] shows Claim 3 to conclude the proof.

For the subset of batches $U$ let $U_G \triangleq U \cap G$ and $U_A \triangleq U \cap A$. The algorithm updates the set $U$ in step 9. To prove the first claim we show that whenever the algorithm performs the update $U \leftarrow U \setminus I^m_j(S^U_G) \cap U$, then $|I^m_j(S^U_G) \cap U_A| \geq 10|I^m_j(S^U_G) \cap U_G|$.

For prove this, in the next lemma, we first show that the median $\text{med}(S)$ is close to $p(S)$, for all $S \subseteq [k]$.

**Lemma 4.**

$$|p(S) - \text{med}(S)| \leq \sqrt{\ln 6/n}.$$  

**Proof.** Its easy to verify that $|p(S) - \text{med}(S)| \geq \sqrt{\ln 6/n}$, only if the size of the set $T = \{i : |p(S) - \bar{p}_b(S)| \geq \sqrt{\ln 6/n}\}$ is at-least 0.5m. But 

$$|T| = |T \cap G| + |T \cap A| \leq |G|/6 + |A| = \frac{m}{6} + \frac{5}{6} |A| \leq \frac{m}{6} + \frac{2m}{6} = 0.5m,$$

where inequality (a) follows from the condition \[9\] and (b) follows since $|A| \leq \beta m \leq 0.4m$. \[\blacksquare\]
Then the condition (9) implies,
\[ |I_j^m(S) \cap G| \leq |I_j^m(S) \cap G| \leq |G| \frac{\beta^2 - 2j}{24j^2}. \]

Therefore if for some subset \([k], |I_j^m(S^*_U) \cap U| \geq |G| \frac{\beta^2 - 2j}{24j^2}, \]
then \(|I_j^m(S^*_U) \cap U_A| \geq 10|G| \frac{\beta^2 - 2j}{24j^2} \geq 10|I_j^m(S^*_U) \cap U_G|. \)
Hence the number of bad batches that gets removed from \(U\) in each update are at-least 10-times the number of good batches that are removed. Therefore, at any stage of the algorithm we have

**Lemma 5.**
\[ |G \setminus U_G| \leq 0.1|A \setminus U_A|. \]

Next, we show that \(U\) will always all but at-most \(\beta/6\) fraction of the good batches.
\[ |U_G| \geq |G| - 0.1|A \setminus U_A| \geq |G| - 0.1|A| \geq |G| - 0.1|U| \geq |G|(1 - \frac{\beta}{6}), \]
where (a) follows since \(|A| \leq \beta|U|\) and (b) follows since \(|U| \geq \frac{|G|}{\beta}\). Next, we show that in such a collection \(U\) fraction of the bad batches is bounded by their fraction in the \(|m|\).
\[ \frac{|U_A|}{|U_G|} = \frac{|A| - |A \setminus U_A|}{|G| - |G \setminus U_G|} \leq \frac{|A|}{|G|}, \]

since \(|A| < |G|\) and \(|A \setminus U_A| \geq |G \setminus U_G|\). Therefore,
\[ \frac{|U_A|}{|U|} \leq \frac{|A|}{m} \leq \beta. \]

Let
\[ \sum_{b \in U_A} (\bar{p}_b(S) - p(S))^2 = \Delta|U|\beta \frac{\ln(e/\beta)}{n} = \Delta|U| \kappa, \]
for some \(\Delta \geq 0\). The next lemma upper bounds \(|\bar{p}_U(S) - p(S)|\) in terms of \(\sum_{b \in U_A} (\bar{p}_b(S) - p(S))^2\), which is sum of squared distance of \(\bar{p}_b(S)\) from \(p(S)\) for the adversarial batches, to show the claim (7).

**Lemma 6.** \(|\bar{p}_U(S) - p(S)| \leq (\frac{1}{2} + \sqrt{\Delta})\sqrt{\kappa} = (\frac{1}{2} + \sqrt{\Delta})\sqrt{\beta \kappa}. \)

**Proof.** The next inequality upper bounds the deviation of empirical average of \(\bar{p}_b(S)\) from \(p(S)\) for the collection of adversarial batches \(U_A\).
\[ |\bar{p}_{U_A}(S) - p(S)| = \left| \frac{1}{|U_A|} \sum_{b \in U_A} (\bar{p}_b(S) - p(S)) \right| \leq \frac{1}{|U_A|} \sum_{b \in U_A} |\bar{p}_b(S) - p(S)| \]
\[ \leq \sqrt{\frac{1}{|U_A|} \sum_{b \in U_A} (\bar{p}_b(S) - p(S))^2} = \sqrt{\frac{\Delta|U|}{|A|} \kappa}, \]
where the second inequality follows from the Cauchy-Schwarz inequality. The next equation expresses the deviation of \(p(S)\) from the overall empirical average \(\bar{p}_b(S)\) of batches in \(U\), from its deviation from the average of good and the adversarial batches.
\[ \bar{p}_U(S) - p(S) = \frac{|U_G|}{|U|} \bar{p}_{U_G}(S) + \frac{|U_A|}{|U|} \bar{p}_{U_A}(S) - p(S) \]
\[ = \frac{|U_G|}{|U|} (\bar{p}_{U_G}(S) - p(S)) + \frac{|U_A|}{|U|} (\bar{p}_{U_A}(S) - p(S)) \]
Combining the above two equations and the condition (6) gives
\[
|\tilde{p}_V(S) - p(S)| \leq \frac{|U_G|}{|U|} \frac{1}{2} \sqrt{\beta \kappa} + \frac{|U_A|}{|U|} \sqrt{\Delta |U_A| \kappa}
\]
\[
\leq \frac{1}{2} \sqrt{\beta \kappa} + \sqrt{\Delta |U_A| \kappa} = \frac{1}{2} \sqrt{\beta \kappa} + \sqrt{\Delta \beta \kappa} = \left( \frac{1}{2} + \sqrt{\Delta} \right) \sqrt{\beta \kappa}.
\]

The next lemma upper and lower bounds the difference between the sample variance and the mean induced variance estimate $V_U(S) - V(\tilde{p}_V(S))$ in terms of $\sum_{b \in U_A} (\tilde{p}_b(S) - p(S))^2$.

**Lemma 7.**

\[(0.6 \Delta - \sqrt{\Delta} - 7)\kappa \leq V_U(S) - V(\tilde{p}_V(S)) \leq (\Delta + 7 + \sqrt{\Delta}/3)\kappa.
\]

**Proof.** The next equation relates the sample variance of $\tilde{p}_b(S)$ to sum of their squared deviation from $p(S)$ over the batches in $U$.

\[
|U|V_U(S) = \sum_{b \in U} (\tilde{p}_b(S) - p(S))^2 = \sum_{b \in U} (\tilde{p}_b(S) - p(S) - (\tilde{p}_V(S) - p(S)))^2
\]
\[
= \sum_{b \in U} (\tilde{p}_b(S) - p(S))^2 + |U|((\tilde{p}_V(S) - p(S))^2 - 2(\tilde{p}_V(S) - p(S))(\tilde{p}_b(S) - p(S)))
\]
\[
= \sum_{b \in U} (\tilde{p}_b(S) - p(S))^2 + |U|((\tilde{p}_V(S) - p(S))^2 - 2(\tilde{p}_V(S) - p(S))\sum_{b \in U} (\tilde{p}_b(S) - p(S))
\]
\[
= \sum_{b \in U} (\tilde{p}_b(S) - p(S))^2 - |U|((\tilde{p}_V(S) - p(S))^2.
\]

Therefore,

\[
|U|V_U(S) = \sum_{b \in U} (\tilde{p}_b(S) - p(S))^2 - |U|((p(S) - \tilde{p}_V(S))^2
\]
\[
= \sum_{b \in U} (\tilde{p}_b(S) - p(S))^2 + \sum_{b \in U} (\tilde{p}_b(S) - p(S))^2 - |U|((p(S) - \tilde{p}_V(S))^2
\]
\[
= \Delta |U|\kappa + \sum_{b \in U} (\tilde{p}_b(S) - p(S))^2 - |U|((p(S) - \tilde{p}_V(S))^2
\]

Next set of inequalities lead to the upper bound in the Lemma.

\[
|U|((V_U(S) - V(\tilde{p}_V(S))) \leq (a) \Delta |U|\kappa + |U_G|V(p(S)) + 6|U|\kappa - |U|V(\tilde{p}_V(S))
\]
\[
\leq (b) (\Delta + 6)|U|\kappa + |U|V(p(S)) - |U|V(\tilde{p}_V(S))
\]
\[
\leq (c) (\Delta + 6)|U|\kappa + |U|\frac{|p(S) - \tilde{p}_V(S)|}{n},
\]

where the inequality (a) follows from (7) and (13), (b) follows since $|U| \geq |U_G|$ and $V(\cdot) \geq 0$, and inequality (c) uses (1). Next, from the lemma we have,

\[
|p(S) - \tilde{p}_V(S)| \leq \left( \frac{1}{2} + \sqrt{\Delta} \right) \sqrt{\beta \kappa} = \left( \frac{1}{2} + \sqrt{\Delta} \right) \sqrt{\beta \kappa} \sqrt{\frac{\ln(6e/\beta)}{n}} = \left( \frac{1}{2} + \sqrt{\Delta} \beta \kappa \sqrt{\frac{\ln(6e/\beta)}{n}}
\]
\[
= \left( \frac{1}{2} + \sqrt{\Delta} \beta \ln(6e/\beta) \sqrt{\frac{1}{n \ln(6e/\beta)}} = \left( \frac{1}{2} + \sqrt{\Delta} \right) \frac{1}{3} n \kappa,
\]

(14)
where the last inequality used the fact that the batch-size \( n \geq 9 \). Combining the above two equations gives the upper bound.

Next showing the lower bound,

\[
|U|(\bar{V}_U(S) - V(\bar{p}_U(S)))
\]

\[\geq \Delta |U|\kappa + |U_G|V(p(S)) - 6|U|\kappa - |U|V(\bar{p}_U(S)) - |U|(p(S) - \bar{p}_U(S))^2 \]

\[\geq (\Delta - 6)|U|\kappa + |U_G|V(p(S)) - |U_G|V(\bar{p}_U(S)) - |U_A|V(\bar{p}_U(S)) - |U|(p(S) - \bar{p}_U(S))^2 \]

\[\geq (\Delta - 6)|U|\kappa - |U_G|\frac{|p(S) - \bar{p}_U(S)|}{n} - \frac{|U_A|}{4n} - |U|(p(S) - \bar{p}_U(S))^2 \]

\[\geq (\Delta - 6)|U|\kappa - |U_G|\frac{|p(S) - \bar{p}_U(S)|}{n} - \frac{|U|}{4n}(\frac{1}{2} + \sqrt{\Delta})\sqrt{\kappa} \]

\[\geq (\Delta - 6)|U|\kappa - \frac{|U|}{4}(\frac{1}{2} + \sqrt{\Delta})\sqrt{\kappa} \]

\[\geq (\Delta - 6)|U|\kappa - |U|\beta(\frac{1}{4} + \sqrt{\Delta} + \epsilon) \]

\[\geq (0.6\Delta - \sqrt{\Delta} - \tau)|U|\kappa, \]

where the inequality (a) follows from (1) and (13), (b) follows from (1) and \( V(\cdot) \leq \frac{1}{\kappa} \), and inequality (c) uses that \( |U_A| \leq \beta|U| \), inequality (d) follows from (14) and lemma 8, inequality (e) uses that \( \kappa \geq \frac{2}{\kappa} \) and finally (f) follows from \( \beta \leq 0.4 \).

From the above lemma, we get the following:

**Corollary 8.** If \( |\bar{V}_U(S) - V(\bar{p}_U(S))| \geq 40\kappa \) only if \( \Delta \geq 30 \). And if \( |\bar{V}_U(S) - V(\bar{p}_U(S))| \leq 80\kappa \) then \( \Delta \leq 169 \).

Finally, we show that if the number of adversarial batches in the tail are small enough, then \( \sum_{b \in U_A}(\bar{p}_b(S) - p(S))^2 \) can’t be too large.

**Lemma 9.** If for the subset \( U_A \subseteq A \),

\[|I_j^m(S) \cap U_A| \leq |G|\frac{\beta^{2-j}}{2j^2}, \quad \forall \ j \in \{2, \ldots, \log(\sqrt{n} + 1)\}, \]

then

\[\sum_{b \in U_A}(\bar{p}_b(S) - p(S))^2 < 30|U|\kappa.\]

**Proof.**

\[
\sum_{b \in U_A}(\bar{p}_b(S) - \text{med}(S))^2 = \sum_{j \geq 1} \sum_{b \in U_A \cap I_j^m(S)} (\bar{p}_b(S) - \text{med}(S))^2
\]

\[\leq \sum_{j \geq 1} |U_A \cap I_j^m(S)|(2^{j+1}\sqrt{\ln(6e/\beta)/n})^2
\]

\[= |U_A \cap I_j^m(S)|(2\sqrt{\ln(6e/\beta)/n})^2 + \sum_{j \geq 2} |U_A \cap I_j^m(S)|(2^{j+1}\sqrt{\ln(6e/\beta)/n})^2
\]

\[\leq |U_A|(4\sqrt{\ln(6e/\beta)/n})^2 + \sum_{j \geq 2} |G|\frac{\beta^{2-j}}{2j^2}(2^{j+1}\sqrt{\ln(6e/\beta)/n})^2
\]

\[\leq 16|U|\beta\frac{\ln(6e/\beta)}{n} + \sum_{j \geq 2} |G|\frac{\beta^{2}}{2j^2} = 16|U|\kappa + \sum_{j \geq 2} |G|\frac{\kappa}{j^2}
\]
We are grateful to the National Science Foundation (NSF) for supporting this work through grants [10]. We thank Vaishakh Ravindrakumar and Yi Hao for their helpful comments in the preparation of this manuscript.

For any given \( S \subseteq [k] \), let \( p_b \) be the distribution of \( b \) in the subset \( S \). Then \( p_b \) is close to a common target distribution \( p \), such that \( \|p_b - p\| \leq \eta \). For simplicity, we have given the proof for only \( \eta = 0 \). The algorithm and the proof naturally extend to this more general case; here we get an extra dependence on \( \eta \) for the bounds in the lemmas and the theorems, and for the parameters of the algorithm. For \( \eta > 0 \), the techniques in the paper show that if for some subset \( S \subseteq [k] \), \( \|p_b(S) - p(S)\| \geq \Omega(\eta + \beta \sqrt{\log 1/\beta}/n) \), then such a subset \( S \) can be identified efficiently looking at the difference between the sample variance and its estimate, similar to the case when \( \eta = 0 \). And we can use this \( S \) to delete the adversarial batches as before. Hence with a slight modification to the parameters the algorithm can estimate the \( p \) for general \( \eta > 0 \). The modified algorithm for general \( \eta > 0 \) has the same sample and time complexity. The next theorem characterizes the performance of the modified algorithm. For general \( \eta \), [12] show that even with arbitrary many batches \( m \), any algorithm fails to approximate \( p \) to an \( L_1 \) distance better than \( O(\eta + \beta / \sqrt{n}) \). The modified algorithm will essentially achieves the lower bound for general \( \eta \) as well.

**Theorem 3.** For any given \( \eta \geq 0, \beta \leq 0.4, n \geq 1 \) and \( k \geq 2 \), and \( m \geq \Omega(\frac{\kappa}{\beta^2 \ln(1/\beta)} + \frac{\log^2(n)}{\beta \ln(1/\beta)} (k + \log \log n)) \), an algorithm runs in time polynomial in all parameters \( k, n, 1/\eta, \) and \( 1/\beta \), and its estimate \( p^* \) satisfies \( \|p^* - p\|_1 = O(\eta + \beta \sqrt{\ln(1/\beta) \log n}/n) \) with probability \( \geq 1 - O(e^{-k}) \).

**Acknowledgements**

We thank Vaishakh Ravindrakumar and Yi Hao for their helpful comments in the preparation of this manuscript.

We are grateful to the National Science Foundation (NSF) for supporting this work through grants CIF-1564355 and CIF-1619448.

**References**

[1] Noga Alon and Assaf Naor. Approximating the cut-norm via grothendieck’s inequality. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 72–80. ACM, 2004.

[2] Dietrich Braess and Thomas Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004.

[3] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60. ACM, 2017.
[4] Siton Chen, Jerry Li, and Ankur Moitra. Efficiently learning structured distributions from untrusted batches. arXiv preprint arXiv:1911.02035, 2019.

[5] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. SIAM Journal on Computing, 48(2):742–864, 2019.

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

[7] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In Conference on Learning Theory, pages 1066–1100, 2015.

[8] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), pages 665–674. IEEE, 2016.

[9] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. arXiv preprint arXiv:1602.05629, 2016.

[10] H Brendan McMahan and Daniel Ramage. https://research.google.com/pubs/pub44822.html. 2017.

[11] Rigollet Philippe. 18.s997 high-dimensional statistics. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA, 2015.

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

[13] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. arXiv preprint arXiv:1703.04940, 2017.

[14] John W Tukey. A survey of sampling from contaminated distributions. Contributions to probability and statistics, pages 448–485, 1960.