Robust Density Estimation from Batches: The Best Things in Life are (Nearly) Free

Ayush Jain [1] **Alon Orlitsky** [1]

Abstract

In many applications data are collected in batches, some potentially biased, corrupt, or even adversarial. Learning algorithms for this setting have therefore garnered considerable recent attention. In particular, a sequence of works has shown that all approximately piecewise polynomial distributions—and in particular all Gaussian, Gaussian-mixture, log-concave, low-modal, and monotone-hazard distributions—can be learned robustly in polynomial time. However, these results left open the question, stated explicitly in (Chen *et al.*, 2020), about the best possible sample complexity of such algorithms. We answer this question, showing that, perhaps surprisingly, up to logarithmic factors, the optimal sample complexity is the same as for genuine, non-adversarial, data! To establish the result, we reduce robust learning of approximately piecewise polynomial distributions to robust learning of the probability of all subsets of size at most k of a larger discrete domain, and learn these probabilities in optimal sample complexity linear in k regardless of the domain size. In simulations, the algorithm runs very quickly and estimates distributions to essentially the accuracy achieved when all adversarial batches are removed. The results also imply the first polynomial-time sample-optimal algorithm for robust interval-based classification based on batched data.

1. Overview

1.1. Robust learning

In many learning applications, some samples are inadvertently or maliciously corrupted. A natural and intuitive example shows that regardless of the number of samples available, such corruption severely curtails the learning accuracy even for the simplest of tasks, a binary hypothesis test.

Consider independent binary samples distributed either all $\text{Ber}(1/2+\beta/2)$ or all $\text{Ber}(1/2-\beta/2)$. With genuine samples, the underlying distribution can be identified with error that plummets to 0 exponentially fast in the number of samples.

However, if an adversary can observe a fraction $1-\beta$ of the samples and select the rest, our best error is destined to remain a half, regardless of the number of samples available. The poltergeist could simply use the observed samples to determine the underlying distribution, and set the rest so the whole sequence appears to be generated by a $\text{Ber}(1/2)$ distribution, leaving us with no better than a random guess.

This elemental example propagates to essentially all learning tasks, hard-limiting the performance of all learning algorithms. For example, the total variation (TV) distance between the two indistinguishable distributions above is β. Hence the triangle inequality implies that for any number of samples, if a β fraction are adversarial, then even binary, let alone general discrete and continuous, distributions cannot be learned to TV distance less than $\beta/2$. Similar hard limits follow for classification and other learning tasks.

The foregoing seems to suggest the discouraging conclusion that with a β fraction of adversarial data, an $\Omega(\beta)$ loss is inevitable, which as real-life β may be quite large, could be rather foreboding. Fortunately, that is not necessarily so.

In the following and many other applications, data are collected from multiple sources, most typically genuine, but some possibly corrupted or adversarial. Data may be gathered by sensors, each providing a large amount of data, and some sensors may be faulty. The word frequency of an author may be estimated from several large texts, some of which are mis-attributed. User preferences may be learned by querying several individuals, some intentionally biasing their feedback. Multiple agents may contribute to a crowdsourcing platform, but some may be unreliable or malicious.

The collection of data generated by each source, or during a time period, is called a *batch*. Interestingly, for data generated in batches, a fraction β of which are corrupted or

[1]University of California, San Diego. Correspondence to: Ayush Jain <ayjain@eng.ucsd.edu>, Alon Orlitsky <alon@eng.ucsd.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

adversarial, significantly higher accuracy can be achieved.

1.2. Robust learning from batches

To formalize this setting, (Qiao & Valiant, 2017) considered estimating an unknown distribution p over the finite domain $[\ell] = \{1, \ldots, \ell\}$ in TV-distance. Estimation is based on m batches with $\geq n$ samples each. In most batches, the samples are drawn independently according to p, but a fraction $\beta < 0.5$ of the batches are *adversarial* and may be arbitrarily corrupted, possibly even with knowledge of the good batches.

Unlike the strict $\Theta(\beta)$ accuracy limit for individual samples, they derived a batch-setting algorithm that approximates p to a much lower TV-distance $\mathcal{O}(\beta/\sqrt{n})$, where the implied constant factor is independent of ℓ. They also showed a matching *adversarial lower bound* (for batches), that even for binary distributions, and hence for general finite ones, the lowest achievable TV distance with any number of batches is $\geq \Delta_{\min} := \Delta_{\min}(\beta, n) := \beta/(2\sqrt{2n})$.

However, their estimator had some limitations as well. When all samples are genuine, and none is adversarial, estimating p to TV distance ϵ requires $\Theta(\ell/\epsilon^2)$ samples, *e.g.,* (Kamath *et al.*, 2015). Since robust learning is at least as hard, this also forms a *statistical lower bound* on the number of samples required to achieve error ϵ with adversarial batches.

To achieve TV distance $\mathcal{O}(\beta/\sqrt{n}) = \mathcal{O}(\Delta_{\min})$, the estimator in (Qiao & Valiant, 2017) required $\Omega(\frac{n+\ell}{n\cdot\Delta_{\min}^2})$ batches, hence $\Omega(\frac{n+\ell}{\Delta_{\min}^2})$ samples that for $n \gg \ell$ exceeds the statistical lower bound. Crucially, and much more significantly, its run-time was exponential in the domain size ℓ, rendering its application, or even simulation, infeasible for even moderate size domains.

The first polynomial-time, and practical, algorithm for the problem was derived in (Jain & Orlitsky, 2019). The algorithm efficiently finds and removes, or *filters*, "outlier" adversarial batches that significantly perturb the empirical distribution away from the underlying p, and then estimates p as the empirical distribution of the remaining batches. It achieves TV distance $\mathcal{O}(\Delta)$, where $\Delta := \Delta(\beta, n) := \Delta_{\min} \cdot \sqrt{\ln(1/\beta)}$ is essentially the adversarial lower bound. To achieve this error they require $\mathcal{O}(\ell/\Delta^2)$ samples, matching the statistical lower bound even when all samples are genuine.

1.3. Robust learning large and continuous distributions

Since many modern applications utilize very large, often continuous, domains, even linear ℓ/Δ^2 dependence of the sample complexity on the domain size may be prohibitive.

Fortunately, common distributions often possess some structure that facilitates more efficient learning. One of the most popular, and important structures is piecewise polynomials.

A distribution q over $[a, b]$ is *t-piecewise degree-d* if for some partition of $[a, b]$ into t intervals $I_1, \ldots, I_t$, and degree-d polynomials $r_1, \ldots, r_t$, $\forall j$, $x \in I_j$, $q(x) = r_j(x)$. Let $\mathcal{P}_{t,d}$ denote the set of all t-piecewise degree d distributions.

Piecewise-polynomials include important distribution families, *e.g.,* $\mathcal{P}_{t,0}$ for histograms and $\mathcal{P}_{t,1}$ for piecewise-linear distributions. They can also approximate any piecewise continuous distribution. Importantly, with very low t and d, they arbitrarily closely approximate many staple one-dimensional distribution families, including Gaussians and their mixtures, log-concave, low-modal, and monotone hazard *e.g.,* (Acharya *et al.*, 2017).

For genuine, non-adversarial, samples, several works, *e.g.,* (Acharya *et al.*, 2017; Hao *et al.*, 2020), derived efficient algorithms that learn t-piecewise degree-d polynomials to TV distance ϵ, with optimal $\mathcal{O}(td/\epsilon^2)$ sample complexity.

$\mathcal{P}_{t,d}$ can be similarly defined as discrete distributions over the interval domain $[\ell]$. (Chen *et al.*, 2019) showed that these distributions can be robustly learned from batches to TV-distance $\mathcal{O}(\Delta)$ with sample complexity only quasi-polylogarithmic in ℓ. However their sample complexity was quasi-polynomial in the other parameters t, d, batch size n, and $1/\beta$. And the algorithm's computational complexity was quasi-polynomial in these parameters and the domain size ℓ.

If computation time is no object, (Jain & Orlitsky, 2020) presented an exponential-time estimator that achieves TV distance $\mathcal{O}(\Delta)$ with $\tilde{\mathcal{O}}(td/\Delta^2)$ samples, the same, up to log logarithmic factors, as the minimum genuine samples required.

To obtain polynomial-time algorithms with low sample complexity, subsequent works adapted the filtering approach of (Jain & Orlitsky, 2019). For example, (Chen *et al.*, 2020) achieved TV-distance $\mathcal{O}(\Delta)$ using $\tilde{\mathcal{O}}((td\log\ell)^2/\Delta^2)$ samples, and concurrently (Jain & Orlitsky, 2020) achieved the same TV distance using $\tilde{\mathcal{O}}(td/\Delta^3)$ samples, in particular, removing the dependence of sample complexity on the domain size, and for the first time, enabling robust learning over infinite and continuous domains.

1.4. Overview of results and applications

Still, both $\tilde{\mathcal{O}}((td\log\ell)^2/\Delta^2)$ and $\tilde{\mathcal{O}}(td/\Delta^3)$ exceed the $\mathcal{O}(td/\Delta^2)$ optimal sample complexity of genuine samples, leading (Chen *et al.*, 2020) to raise the open question of the optimal sample complexity of robust polynomial $\mathcal{P}_{t,d}$ estimators.

This paper essentially answers this question. We derive a filter-based polynomial-time algorithm that achieves TV

distance $\mathcal{O}(\Delta)$ using only $\tilde{\mathcal{O}}(td/\Delta^2)$ samples, that up to poly-logarithmic factors matches the statistical lower bound of even genuine samples. It therefore essentially determines the sample complexity of robust and efficient learning of piecewise polynomials to optimal accuracy. It also shows that for this large and general class, robustness can be achieved at the small cost of at most a poly-logarithmic increase in the number of samples.

These results apply to both continuous and discreet distributions, and as described in Subsection 2.2 also learn distributions that can be approximated by $\mathcal{P}_{t,d}$, hence apply to monotone, log-concave, Gaussian, Gaussian mixtures, and other fundamental distribution classes.

While we present the results in terms of robust density estimation, their distance to other fundamental learning staples is minute. We demonstrate two such applications.

The first is to robust classification. We show that a simple extension of the results yields the first sample-optimal, polynomial-time, robust, classifier based on batched training data. We demonstrate the method's efficacy on the fundamental and practical problem of interval-based classification over the real line.

The second application is to the common *top* k or *heavy hitters* problem that calls for finding the k highest-probability elements in a distribution over a large domain. The problem arises in many applications ranging from caching, to recommendation systems, and vaccine design. We show that in the batch setting, the top k elements can be approximated robustly with sample complexity linear in k regardless of the domain size.

1.5. Other related works

This paper builds on several long and impressive lines of work, briefly summarized herein. Structured-distribution estimation was studied in (Chan *et al.*, 2014; O'Brien, 2016; Diakonikolas, 2016; Ashtiani & Mehrabian, 2018; Acharya *et al.*, 2017; Hao *et al.*, 2020). Robust-statistics was introduced in the classical works of (Tukey, 1960; Huber, 1992). Efficient algorithms for learning the mean and covariance matrices of high-dimensional sub-gaussian and other distributions with bounded fourth moments in the presence of the adversarial samples was studied in (Lai *et al.*, 2016; Diakonikolas *et al.*, 2016). When more than half of the samples are adversarial, the underlying distribution cannot be estimated well, and instead, (Charikar *et al.*, 2017) returned a small set of candidate distributions one of which is a good approximate of the underlying distribution. For extensive surveys on robust learning algorithms see (Steinhardt *et al.*, 2017; Diakonikolas *et al.*, 2019).

The filtering approach to robust estimation was introduced in (Diakonikolas *et al.*, 2016), and used in several subsequent applications including high dimensional estimation (Diakonikolas *et al.*, 2017; 2018; Steinhardt *et al.*, 2017; Diakonikolas *et al.*, 2019). These estimators applied to single samples and learned in L_2 distance. By contrast, the results in this paper and those in (Jain & Orlitsky, 2019; 2020) address batch learning under TV-distance.

Several recent works considered related "multi-source" or "collaborative" PAC learning scenarios. As in our setting, they assume multiple sources, some genuine and others possibly adversarial, where each source provides multiple labeled samples, but some specific assumptions differ. (Awasthi *et al.*, 2017) considers only the realizable case and allow actively acquisition of more data from the source of choice. (Qiao, 2018) also focuses on realizable case where sources share a common labelling function, but may have different input distributions. (Konstantinov *et al.*, 2020) considers the setting that most closely resembles ours and the more general prior work (Jain & Orlitsky, 2020), but they do not present efficient algorithms and incur sub-optimal $\mathcal{O}(\sqrt{k}\Delta)$ excess loss, higher than the $\mathcal{O}(\Delta)$ we achieve.

1.6. Organization of the paper

In the next section we describe the main results we obtain, the techniques used to derive them, and some of their applications. In Section 3, we simplify the learning problem to that of learning all k-element subsets of a large discrete set. In Section 4 we describe an efficient filtering algorithm for this problem. In Section 5 we describe the experiments. The appendix contains most of the proofs.

2. Main techniques, results, and applications

While we would like to learn continuous distribution in TV distance, as in (Chen *et al.*, 2019; Jain & Orlitsky, 2020; Chen *et al.*, 2020), it will prove advantageous to first learn them in a weaker (smaller) distance.

2.1. Density estimation in $\mathcal{A}_k$ distance

Recall that the TV distance between two real distributions q and q' is the maximum of $|q(S) - q'(S)|$ over all Borel sets $S \subseteq \mathbb{R}$. This notion generalizes to arbitrary collections $\mathcal{S}$ of real sets. The $\mathcal{S}$-*distance* between q and q' is

$$||q - q'||_{\mathcal{S}} := \max_{S \in \mathcal{S}} |q(S) - q'(S)|.$$

For $k \geq 1$, let $\mathcal{A}_k$ be the collection of all unions of at most k real intervals. Clearly $||q-q'||_{\mathcal{A}_k} \leq ||q-q'||_{\text{TV}}$ for any q, q', with equality when the domain size $\ell \leq k$. Hence from now on we assume $\ell > k$, and can also be infinite.

One nice property of $\mathcal{A}_k$ distance is that with only genuine samples, the empirical distribution itself, already estimates any discrete or continuous distribution to $\mathcal{A}_k$ distance ϵ with

$\mathcal{O}(k/\epsilon^2)$ samples, which is also optimal.

To learn distributions in $\mathcal{A}_k$ distance, (Jain & Orlitsky, 2020) and (Chen *et al.*, 2020) adapted the filtering algorithm in (Jain & Orlitsky, 2019). First removing outlier batches, and retaining batches whose empirical distribution approximates p in $\mathcal{A}_k$, rather than TV, distance.

However, both algorithms had suboptimal sample complexity. For $\Delta = \Delta_{\min}\sqrt{\ln(1/\beta)} = \Theta(\beta\sqrt{\ln(1/\beta)/n})$, essentially the best $\mathcal{A}_k$ distance achievable with n-sample batches, they required $\tilde{\mathcal{O}}(k/\Delta^3)$, and $\tilde{\mathcal{O}}((k\log\ell)^2/\Delta^2)$ samples, respectively.

Our fundamental contribution is an algorithm that learns any discrete or continuous distribution to $\mathcal{A}_k$-distance Δ with sample complexity $\tilde{\mathcal{O}}(k/\Delta^2)$, optimal up to logarithmic factors.

Theorem 1. *For some constants $c < 1/2$ and $C > 1$, for any k, $\beta < c$, $\delta < 1$, $n > \Omega(\log^C(1/\beta))$, and discrete or continuous p, the algorithm uses $m \cdot n = \tilde{\mathcal{O}}(\frac{k+\log(1/\delta)}{\Delta^2})$ total samples, and in time $\mathrm{poly}(k, n, m, \beta, \delta)$ outputs an estimate $\hat{p}$ that with probability $\geq 1-\delta$ satisfies*

$$||\hat{p} - p||_{\mathcal{A}_k} \leq \mathcal{O}(\Delta).$$

Remark. Theorem 1 achieves $\mathcal{A}_K$ distance $\mathcal{O}(\Delta)$, within a small $O(\sqrt{\log(1/\beta)})$ factor from the adversarial lower bound for unlimited samples. The algorithm uses a poly-logarithmic factor more samples than the min-max number required for this distance even with strictly genuine data. When the number of samples does not suffice to achieve the minimal $\mathcal{A}_K$ distance of $\mathcal{O}(\Delta)$, the algorithm can be modified to achieve $\mathcal{A}_K$ distance $\tilde{\mathcal{O}}(\sqrt{k/(mn)})$, again within a poly-logarithmic factor from the statistical lower bound as achieving $\mathcal{A}_K$ distance ϵ requires at least k/ϵ^2 genuine samples. This result can be derived by augmenting Theorem 1 with the steps taken in the derivation of Theorem 2 in (Jain & Orlitsky, 2019). A similar observation also holds for all the applications stated next.

To derive the algorithm, we first reduce robust $\mathcal{A}_k$ learning over any domain, even continuous, to robust learning the probability of all $2k$-element subsets of discrete distributions over large domains. We propose a filtering algorithm that learns these probabilities with optimal sample complexity linear in k and independent of the domain's size. The new, simpler, formulation allows for a tight SDP relaxation, that with more refined analysis yields near optimal sample complexity.

The algorithm has several important implications. We apply it to three robust-learning tasks using batched data: (i) learning distributions in or near $\mathcal{P}_{t,d}$, (ii) interval-based binary classification, (iii) learning the top-k heavy hitters. For all three problems we achieve the nearly best possible TV distance $\mathcal{O}(\Delta)$ with the same sample complexity as with genuine samples up to logarithmic factors.

2.2. Density estimation in TV distance

Theorem 1 described an optimal, robust, batch-based, algorithm for learning any distribution over the reals in $\mathcal{A}_k$ distance. Yet $||q - q'||_{\mathcal{A}_k} \leq ||q - q'||_{\mathrm{TV}}$ for any q, q'. This section extends the results to robustly learn in the more standard, and stringent, TV-distance.

In Theorem 3 we present a batch-based algorithm that robustly learns $\mathcal{P}_{t,d}$ and related distributions $\mathcal{P}_{t,d}$, including monotone, log-concave, Gaussian, Gaussian mixtures, and other fundamental distributions.

For real distribution p, let $\mathrm{opt}_{t,d}(p) := \inf_{q\in\mathcal{P}_{t,d}} ||p-q||_{TV}$ be p's TV-distance to its nearest distribution in $\mathcal{P}_{t,d}$. We wish to find a distribution $\hat{p}$ such that for a small ϵ and universal constant α, with probability $\geq 1-\delta$,

$$||\hat{p} - p||_{TV} \leq \alpha \cdot \mathrm{opt}_{t,d}(p) + \epsilon.$$

This ensures that we learn distributions not just in $\mathcal{P}_{t,d}$, but also nearby. While not emphasized here, the α we derive is roughly 3, and same as the best known factor for learning $\mathcal{P}_{t,d}$ with only genuine samples.

To convert learning $\mathcal{A}_k$- to TV-distance, we use a transformation that maps $\mathcal{A}_k$ neighborhoods of distributions in or near $\mathcal{P}_{t,d}$ to TV-neighborhoods.

Theorem 2. *(Acharya et al., 2017) For a constant α (roughly 3) and any t, d, and ϵ, an algorithm they describe runs in time $\mathrm{poly}(t, d, \epsilon)$ and converts any real distribution p' to a distribution p'' such that for every distribution p,*

$$||p - p''||_{TV} \leq \alpha \cdot \mathrm{opt}_{t,d}(p) + \mathcal{O}(||p - p'||_{\mathcal{A}_{t(d+1)}}) + \epsilon.$$

The theorem shows that if p is near $\mathcal{P}_{t,d}$, then an $\mathcal{A}_{t(d+1)}$ distance approximation of p can be converted to a TV-distance approximation of p, hence it suffices to approximate p in the weaker $\mathcal{A}_{t(d+1)}$ distance.

Combining Theorems 1 and 2 for $k = t(d+1)$, we derive a polynomial-time algorithm that robustly estimates any real distribution nearly as well as its best $\mathcal{P}_{t,d}$ approximation, using the optimal number of samples.

Theorem 3. *For some constants α (roughly 3), $c < 1/2$, and $C > 1$, for any t, d, $\beta < c$, $\delta < 1$, $n > \Omega(\log^C(\frac{1}{\beta}))$, and real or discrete distribution p, a simple combination of the above algorithms uses $m \cdot n = \tilde{\mathcal{O}}(\frac{t(d+1)+\log(1/\delta)}{\Delta^2})$ total samples, and in time $\mathrm{poly}(t, d, n, m, \beta, \delta)$ outputs an estimate $\hat{p}$ that with probability $\geq 1-\delta$ satisfies*

$$||\hat{p} - p||_{TV} \leq \alpha \cdot \mathrm{opt}_{t,d}(p) + \mathcal{O}(\Delta).$$

Note that the adversarial-batch lower bound on the approximation's TV distance is $\Delta_{\min} = \beta/(2\sqrt{2n})$, while the theorem, like all other robust-learning results so far, applies to a slightly higher TV distance $\Delta = \mathcal{O}(\Delta_{\min}\sqrt{\log(1/\beta)})$. Based on evidence from Gaussian robust mean esti-

mation, (Chen *et al.*, 2020) suggested that the extra $\mathcal{O}(\sqrt{\log(1/\beta)})$ factor may be necessary for any polynomial time algorithm.

2.3. Application to interval-based classification

We now show that though presented for density estimation, a simple extension of our results yields the first polynomial-time, sample-optimal, robust batch classifier, and demonstrate it on the fundamental and practical problem of interval-based binary classification over the reals.

Without loss of generality let the observations be distributed over $[0,1]$. Each good batch therefore contain n labeled samples from a distribution p over $[0,1]\times\{-1,1\}$, while the adversarial batches contain n arbitrary pairs.

Consider a hypothesis family of Boolean functions $\mathcal{H}_k : [0,1]\to\{-1,1\}$ whose decision regions, the inverse images of -1 and 1, consist of at most k-intervals. The loss of classifier $h\in\mathcal{H}_k$ for any distribution q over $[0,1]\times\{-1,1\}$ is $r_q(h) := \Pr_{(X,Y)\sim q}[h(X)\neq Y]$. The *optimal* $\mathcal{H}_k$ *classifier* for a distribution q is $h^{\text{opt}}(q) := \arg\min_{h\in\mathcal{H}_k} r_q(h)$, and the *optimal loss* is $r_q^{\text{opt}}(\mathcal{H}_k) := r_q(h^{\text{opt}}(q))$.

Given samples from an underlying distribution p, the goal is to return a classifier $h\in\mathcal{H}_k$ whose *excess loss* $r_p(h) - r_p^{\text{opt}}(\mathcal{H}_k)$ relative to the optimal loss is small.

Map any distribution q over $[0,1]\times\{-1,1\}$, to a new distribution $q^{[-1,1]}$ over $[-1,1]$, where $q^{[-1,1]}(z) := \Pr(X\cdot Y = z)$ for $(X,Y)\sim q$. Note that there is a 1-1 correspondence between q and $q^{[-1,1]}$, and that we can define $\mathcal{A}_k$ distance over the new domain $[-1,1]$.

Lemma 6 in (Jain & Orlitsky, 2020) upper bounds the excess loss when the optimal classifier for distribution q is applied to distribution p in terms of $\mathcal{A}_k$ distance between $p^{[-1,1]}$ and $q^{[-1,1]}$. For completeness we present a short proof in Appendix F.

Lemma 4. *For any distributions p, q over $[0,1]\times\{-1,1\}$,*
$$r_p(h^{opt}(q)) - r_p^{opt}(\mathcal{H}_k) \le 2||p^{[-1,1]} - q^{[-1,1]}||_{\mathcal{A}_{2k}}.$$

Furthermore, (Maass, 1994) derived an algorithm that for any empirical distribution q over $[0,1]\times\{-1,1\}$ finds the optimal classifier $h^{\text{opt}}(q)$ in polynomial time in the number of samples and k. Then from the above Lemma to obtain an excess loss $\mathcal{O}(\Delta)$ it suffices to estimate $p^{[-1,1]}$ to $\mathcal{A}_k$ distance $\mathcal{O}(\Delta)$.

Theorem 1 provides an algorithm to learn any real distribution to $\mathcal{A}_k$ distance $\mathcal{O}(\Delta)$ using $\tilde{\mathcal{O}}(k/\Delta^2)$ samples, implying the following.

Theorem 5. *For some constants $c < 1/2$ and $C > 1$, for any k, $\beta < c$, $\delta < 1$, $n > \Omega(\log^C(1/\beta))$, and p over $[0,1]\times\{-1,1\}$, the above algorithm uses $m\cdot n = \tilde{\mathcal{O}}(\frac{k+\log(1/\delta)}{\Delta^2})$ pairs, and in* $\text{poly}(k,n,m,\beta,\delta)$ *time outputs a classification h^* with excess loss $r_p(h^*) - r_p^{\text{opt}}(\mathcal{H}_k) \le \mathcal{O}(\Delta)$.*

Since the VC-dimension of the collection $\mathcal{H}_k$ is $\mathcal{O}(k)$, any algorithm achieving excess loss ϵ requires $\Omega(k/\epsilon^2)$ samples, even with genuine data. Therefore, achieving excess loss $\mathcal{O}(\Delta)$ requires $\Omega(k/\Delta^2)$ samples, even with genuine data, showing that our algorithm is sample optimal up to logarithmic factors.

(Jain & Orlitsky, 2020) showed that the best possible excess loss for this problem is $\Omega(\Delta_{\text{min}})$. They used a similar reduction from $\mathcal{A}_k$ distance, to derive a polynomial-time algorithm with $\mathcal{O}(\Delta)$ excess loss, but required a suboptimal $\tilde{\mathcal{O}}(k/\Delta^3)$ number of samples.

2.4. Application to the top k heavy hitters problem

Our last application is to the prevalent *top k*, or *heavy hitters*, problem. Given samples from a distribution over a large domain, we would like to find the k elements with highest probability. This problem arises in numerous applications including deciding which pages to store in a cache, results to show on the front page of a web search, viruses to inoculate for in an influenza vaccine (Wikipedia, 2020; Center for Disease Control, 2020), and products to recommend to online shoppers.

As in the rest of the paper, we consider samples that arrive in batches, some possibly corrupt or adversarial. For example, some shoppers biasing consumer ratings towards select products.

The top k elements clearly have the highest total probability among all k-element subsets. However, this set cannot always be found as some elements with nearly identical probabilities cannot be identified. Instead, we therefore aim to robustly find a k-element subset whose total probability is maximal up to a $\mathcal{O}(\Delta)$ difference.

The results in this section apply to all discrete distributions, that without loss of generality we assume range over the integers. They can be trivially extended to mixed distributions over the reals as well.

A natural approach may be to learn p robustly to TV distance Δ as in (Jain & Orlitsky, 2019), and return the k element subset with highest estimated probability. However, this approach would require number of samples proportional to the domain size, while in a typical k-hitter problem, k is significantly smaller.

Instead, we first estimate p to an $\mathcal{A}_k$ distance $\mathcal{O}(\Delta)$, which from Theorem 1 can be done efficiently using $\tilde{\mathcal{O}}(k/\Delta^2)$ samples. We then return the k-element subset with highest estimated probability. Since the collection $\mathcal{A}_k$ is a superset of the collection of all subsets of size $\le k$, learning to an $\mathcal{A}_k$ distance $\mathcal{O}(\Delta)$ implies learning the probability of all

such subsets to accuracy $\mathcal{O}(\Delta)$. By the triangle inequality, the k-element subset with highest estimated probability is maximal up to a $2\mathcal{O}(\Delta)$ probability difference.

3. Two simplifications of A_k-distance learning

3.1. Discretization using partitioning

Let B denote a collection of all m batches. Recall that each batch has n samples. For $s = n \cdot m$, let $x^s = x_1, x_2, \ldots, x_s \in \mathbb{R}$ be the samples of B sorted in non-decreasing order, and define $\bar{p}_B$ to be the empirical distribution of x^s.

Given samples x^s, for $j \geq 1$, let $\mathcal{P}^j = P_1^j, P_2^j, \ldots, P_{k\cdot j}^j$, partition $\mathbb{R}$ into $k \cdot j$ disjoint intervals, or *parts*, each containing $\approx \frac{s}{k\cdot j}$ samples, and given by

$$P_i^j := \begin{cases} (-\infty, x_{\lfloor \frac{s}{k\cdot j} \rfloor}] & i = 1, \\ (x_{\lfloor \frac{(i-1)s}{k\cdot j} \rfloor}, x_{\lfloor \frac{i\cdot s}{k\cdot j} \rfloor}] & 2 \leq i < k \cdot j, \\ (x_{\lfloor \frac{(i-1)s}{k\cdot j} \rfloor}, \infty) & i = k \cdot j. \end{cases}$$

Let $\mathcal{C}(\mathcal{P}^j)$ be the collection of real subsets formed by unions of parts of $\mathcal{P}^j$. Unions of consecutive parts of $\mathcal{P}^j$ are themselves intervals in $\mathbb{R}$ that we call *intervals over* $\mathcal{P}^j$.

Let $\mathcal{A}_k(\mathcal{P}^j)$ be the collection of all unions of at most k intervals over $\mathcal{P}^j$. Clearly, $\mathcal{A}_k(\mathcal{P}^j) \subseteq \mathcal{A}_k$, hence $||q - q'||_{\mathcal{A}_k(\mathcal{P}^j)} \leq ||q - q'||_{\mathcal{A}_k}$ for any distributions q and q'. Interestingly a reverse relation holds for the underlying distribution p.

Lemma 6. *For $m \cdot n = \tilde{\Omega}(k/\Delta^2)$, w.h.p., for all $j \geq \frac{1}{\Delta}$ and all distributions q over $\mathbb{R}$,*

$$||q - p||_{\mathcal{A}_k} \leq ||q - p||_{\mathcal{A}_k(\mathcal{P}^j)} + \mathcal{O}(\Delta).$$

To prove the lemma we need the following results, proved in Appendix F.

Lemma 7. *For any subset $S \in \mathcal{A}_k$, there are sets $S', S'' \in \mathcal{A}_k(\mathcal{P}^j)$ such that $S' \subseteq S \subseteq S''$ and $\bar{p}_B(S'' \setminus S') \leq 2/j$.*

Lemma 8. *For $\beta < 1/2$, $m \cdot n = \tilde{\Omega}(\frac{k + \log 1/\delta}{\Delta^2})$, with probability $> 1 - \delta$, for all $S \in \mathcal{C}(\mathcal{P}^j)$, $p(S) \leq 2 \cdot \bar{p}_B(S) + \mathcal{O}(\Delta)$.*

Proof of Lemma 6. From Lemma 7 for any subset $S \in \mathcal{A}_k$, let $S', S'' \in \mathcal{A}_k(\mathcal{P}^j)$ be the sets such that $S' \subseteq S \subseteq S''$ and $\bar{p}_B(S'' \setminus S') \leq \mathcal{O}(1/j)$. Clearly $S'' \setminus S' \subseteq \mathcal{C}(\mathcal{P}^j)$, then from Lemma 8, *w.h.p.*, $p(S'' \setminus S') \leq 2 \cdot \bar{p}_B(S'' \setminus S') + \mathcal{O}(\Delta) \leq \mathcal{O}(1/j + \Delta)$. Then

$$\begin{aligned} p(S) - q(S) &\leq p(S) - q(S') \\ &= p(S') - q(S') + p(S \setminus S') \\ &\leq p(S') - q(S') + p(S'' \setminus S') \\ &\leq ||q - p||_{\mathcal{A}_k(\mathcal{P}^j)} + \mathcal{O}(1/j + \Delta). \end{aligned}$$

A similar bound for $q(S) - p(S)$ completes the proof. ■

The lemma shows that to approximate p in $\mathcal{A}_k$-distance it suffices to estimate it in $\mathcal{A}_k(\mathcal{P}^j)$-distance for any $j = \Omega(\frac{1}{\Delta})$. The advantage of this reduction is that the set $\mathcal{A}_k(\mathcal{P}^j)$ is finite in contrast to $\mathcal{A}_k$.

Given a distribution q on $\mathbb{R}$, for any $j \geq 1$ let $q^j \in \mathbb{R}^{k\cdot j}$ be the discrete distribution over the indices of partition $\mathcal{P}^j$, defined by $q^j(i) = q(P_i^j)$ for $i \in [k \cdot j]$.

Map every subset $S \in \mathcal{C}(\mathcal{P}^j)$ to the binary vector $v_S \in \{0,1\}^{k\cdot j}$ whose ith coordinate indicates whether $P_i^j \subseteq S$. Observe that for any distribution q over $\mathbb{R}$, we can express $q(S)$ as the inner product $q^j \cdot v_S$. Let $\mathcal{V}_k^\ell$ denotes the collection of binary vectors $\{0,1\}^\ell$ with at most k runs of ones. Since each interval over $\mathcal{P}^j$ corresponds to a single run of ones, if $S \in \mathcal{A}_k(\mathcal{P}^j)$, then $v_S \in \mathcal{V}_k^{k\cdot j} \subseteq \{0,1\}^{kj}$.

This discussion and Lemma 6 show that if for the discretized versions of an estimator $\hat{p}$ and underlying distribution p, $\max_{v \in \mathcal{V}_k^{k\cdot j}} |\hat{p}^j \cdot v - p^j \cdot v| \leq \mathcal{O}(\Delta)$ then $||\hat{p} - p||_{\mathcal{A}_k} \leq \mathcal{O}(\Delta)$. However, the collection $\mathcal{V}_k^{k\cdot j}$ is rather complicated and does not have a tight convex relaxation. Previous relaxations of $\mathcal{V}_k^{k\cdot j}$ (Chen *et al.*, 2020) lead to sub-optimal sample complexities. Instead, we show in the next section that this problem can be further reduced to robust learning of the probabilities of all subsets of a fixed size $2k$ over a large discrete domain. In Section 4, we show that these probabilities can be robustly estimated with optimal sample-complexity $\tilde{\mathcal{O}}(k)$.

3.2. Reduction to learning k element subset

Let $\mathcal{I}(\mathcal{P}^j) \subseteq \mathcal{C}(\mathcal{P}^j)$ consist of all unions of at most $2k$ parts of $\mathcal{P}^j$. Let $\{0,1\}_k^\ell$ denote the set of binary vectors of length ℓ with at most k ones. Observe that every subset in $S \in \mathcal{I}(\mathcal{P}^j)$ corresponds to a binary vector $v_S \in \{0,1\}_{2k}^{k\cdot j}$. Note that $\mathcal{I}(\mathcal{P}^2) = \mathcal{C}(\mathcal{P}^2)$, as $\{0,1\}_{2k}^{2k} = \{0,1\}^{2k}$.

We now show that to estimate p in $\mathcal{A}_k$ distance it suffices find a q such that $\forall j \in 2^{[\log(1/\Delta)]}$, the powers of two between 2 and $1/\Delta$, the distances $||p - q||_{\mathcal{I}(\mathcal{P}^j)}$ are small.

Theorem 9. *For every $m \cdot n = \tilde{\Omega}(\frac{k + \log 1/\delta}{\Delta^2})$ and distribution q over $\mathbb{R}$, with probability $> 1 - \delta$,*

$$||q - p||_{\mathcal{A}_k} \leq \sum_{j \in 2^{[\log(1/\Delta)]}} \max_{v \in \{0,1\}_{2k}^{k\cdot j}} |q^j \cdot v - p^j \cdot v| + \mathcal{O}(\Delta).$$

Note that for any j, the set $\mathcal{I}(\mathcal{P}^j) \subset \mathcal{A}_{2k}$, therefore the sample complexity of estimating p in $\mathcal{I}(\mathcal{P}^j)$ distance is at most that of learning in $\mathcal{A}_{2k}$-distance.

Importantly, this reduces the more complicated set $\mathcal{V}_k^{k\cdot j}$ to more manageable sets $\{0,1\}_{2k}^{k\cdot j}$, which, as we see in the next section, have nice convex relaxations.

To prove Theorem 9, note a simple geometric observation, proved in Appendix F.

Lemma 10. *For any $i \geq 1$, any interval over partition $\mathcal{P}^{2^i}$ is the union of at-most 2 parts from each partition*

$\mathcal{P}^{2^i}, \mathcal{P}^{2^{i-1}}, ..., \mathcal{P}^{2^2}$ *and one interval over* $\mathcal{P}^2$.

The following result is a simple consequence.

Lemma 11. *For any* $i \geq 1$, *any subset in* $\mathcal{A}_k(\mathcal{P}^{2^i})$ *is the union of one subset from each of* $\mathcal{I}(\mathcal{P}^{2^i})$, $\mathcal{I}(\mathcal{P}^{2^{i-1}})$,...,$\mathcal{I}(\mathcal{P}^{2^1})$.

Proof of Lemma 11. Any subset in $\mathcal{A}_k(\mathcal{P}^{2^i})$ is a union of at most k intervals over partition $\mathcal{P}^{2^i}$, and Lemma 10 implies that it can be expressed as a union of at-most $2k$ parts from each partition $\mathcal{P}^{2^i}, \ldots, \mathcal{P}^{2^2}$ and at most k intervals over $\mathcal{P}^2$. The lemma follows as any union of intervals over $\mathcal{P}^2$ is in $\mathcal{C}(\mathcal{P}^2)$, and $\mathcal{C}(\mathcal{P}^2) = \mathcal{I}(\mathcal{P}^2)$. ■

Proof of Theorem 9. For any distribution q over $\mathbb{R}$, Lemma 11 and the triangle inequality imply

$$||q-p||_{\mathcal{A}_k(\mathcal{P}^{2^i})} = \max_{S \in \mathcal{A}_k(\mathcal{P}^{2^i})} |q(S)-p(S)| \leq \sum_{\ell=1}^{i} \max_{S \in \mathcal{I}(\mathcal{P}^{2^\ell})} |q(S)-p(S)|.$$

Letting $i = \lfloor \log_2(\frac{2}{\Delta}) \rfloor$ and Lemma 6 complete the proof.■

4. Filtering algorithm for $\mathcal{A}_k$ distance

4.1. Notation

We begin with notation that helps describe the filtering algorithm. Recall that B is the collection of m batches, each consisting of $\geq n$ samples. Let B_G denote the collection of all *good* batches in B whose samples are drawn independently from common unknown real distribution p. We refer to the batches in remaining set $B_A := B \setminus B_G$ as *adversarial*. Note that $|B_A| \leq \beta m$.

Let $\bar{\mu}_b$ denote the empirical distribution of samples in batch $b \in B$. Note that $\bar{\mu}_b$ is a collection of n Dirac delta functions. Let B' denote any sub-collection of B. For a batch sub-collection $B' \subseteq B$, consider the average of the empirical distributions of batches in B'.

$$\bar{p}_{B'} \triangleq \frac{1}{|B'|} \sum_{b \in B'} \bar{\mu}_b.$$

Note that $\bar{p}_{B'}$ is also the empirical distribution of all samples in batches of B'.

Recall that for any distribution q over $\mathbb{R}$, $q^j \in \mathbb{R}^{k \cdot j}$ is the discrete distribution induced over the the parts of partition $\mathcal{P}^j$, and let $\bar{\mu}_b^j$ and $\bar{p}_{B'}^j$ be the corresponding empirical distributions of batch b and batch collection B', respectively.

For any discrete distribution, or normalized frequency vector, q, let $\text{Mul}^N(q,n)$ denote the distribution of a normalized multinomial frequency vector μ, where $n \cdot \mu \sim \text{Mul}(q,n)$. Also, let $C(q) := \frac{1}{n}(\text{Diag}(q) - qq^\intercal)$ be the covariance of $\text{Mul}^N(q,n)$.

Let $\mu_1, \ldots, \mu_m \sim \text{Mul}^N(q,n)$ be m i.i.d. normalized frequency vectors, and let $\bar{\mu}$ and V be the mean and covariance of the μ_i's. Intuitively speaking, both V and $C(\bar{\mu})$ converge to the covariance of $\text{Mul}^N(q,n)$, hence their difference tends to zero.

If the partition $\mathcal{P}^j$ was fixed beforehand, not after obtaining the samples, then for $b \in B_G$, the frequency vector $\bar{\mu}_b^j$ would follow a normalized multinomial distribution $\text{Mul}^N(p^j, n)$. Even though the partition depends on the samples, the above multinomial-distribution intuition is still useful as the distribution of $\bar{\mu}_b^j$ is still essentially $\text{Mul}^N(p^j, n)$.

For any batch b, and sub-collection B', let $C_{b,B'}^j := (\bar{\mu}_b^j - \bar{p}_{B'}^j)(\bar{\mu}_b^j - \bar{p}_{B'}^j)^\intercal$ be the *deviation* of batch b relative to batch collection B'.

The *filtering statistics* of a batch b w.r.t. a sub collection B', $F_{b,B'}^j = C_{b,B'}^j - C(\bar{p}_{B'}^j)$ is the difference between the deviation of batch b relative to batch collection B' and covariance matrix of a frequency vector μ generated using the distribution $\mu \sim \text{Mul}(\bar{p}_{B'}^j, n)$. Finally, the *filtering statistics* of a batch sub collection $B' \subseteq B$ is the average $F_{B'}^j := \frac{1}{|B'|} \sum_{b \in B'} F_{b,B'}^j$ of the filtering scores of all batches $b \in B'$ w.r.t. this sub collection B'.

Note that $F_{B'}^j = \frac{1}{|B'|} \sum_{b \in B'} C_{b,B'}^j - C(\bar{p}_{B'}^j)$ is the difference between the empirical covariance matrix of $\{\bar{\mu}_b^j\}_{b \in B'}$, and the covariance matrix of the normalized multinomial distribution with parameter $q = \bar{p}_{B'}^j$, the mean of frequency vectors $\bar{\mu}_b^j$ in B'.

We note that this filtering statistics was first used in (Jain & Orlitsky, 2019) to robustly learn discrete distributions in TV distance, and later used in (Jain & Orlitsky, 2020; Chen *et al.*, 2020) for learning in $\mathcal{A}_k$ distance.

4.2. The filtering algorithm

If there were no adversarial batches, the empirical distribution $\bar{p}_B$ of all batches would estimate p in $\mathcal{A}_k$ distance. However, the presence of adversarial outlier batches can move the empirical distribution $\bar{p}_B$ away from p.

We derive a filtering algorithm that finds a sub-collection B' of batches such that $\forall\, j \in 2^{[\log(1/\Delta)]}$

$$\max_{v \in \{0,1\}_{2k}^{k \cdot j}} |\bar{p}_{B'}^j \cdot v - p^j \cdot v| \leq \mathcal{O}\left(\frac{\beta}{\log^2 j}\sqrt{\frac{\log(\frac{1}{\beta})}{n}}\right) = \mathcal{O}\left(\frac{\Delta}{\log^2 j}\right). \tag{1}$$

Note that $\sum_{j \in 2^{[\log(1/\Delta)]}} \frac{1}{\log^2 j} \leq \sum_i \frac{1}{i^2} = \mathcal{O}(1)$ and $\Delta = \beta\sqrt{(1/n) \cdot \log(1/\beta)}$. Hence Theorem 9 implies that $\bar{p}_{B'}$ estimates p to $\mathcal{A}_k$ distance $\mathcal{O}(\Delta)$.

Inequality (1) characterizes B' whose empirical distribution approximates the underlying distribution p in $\mathcal{A}_k$ distance. However, its definition involves the unknown p itself. It

is naturally more convenient to work with inequalities that does not include p.

One attempt at such an inequality is

$$\max_{v\in\{0,1\}^{k\cdot j}_{2k}} \langle vv^\intercal, F^j_{B'}\rangle \le \mathcal{O}(\tfrac{\beta\log\frac{1}{\beta}}{n\cdot\log^4 j}) = \mathcal{O}(\tfrac{\Delta^2}{\beta\log^4 j}).$$

While under mild conditions this inequality can be shown to imply (1), it is still not easy to use as the set $\{vv^\intercal : v\in\{0,1\}^{k\cdot j}_{2k}\}$ is not convex, hence it is unclear how to efficiently optimize the left hand side.

To circumvent this difficulty, we define a semi-definite programing (SDP) relaxation of $\{vv^\intercal : v\in\{0,1\}^{k\cdot j}_{2k}\}$ as

$$\boldsymbol{R}^j := \{M\in\mathbb{R}^{k\cdot j\times k\cdot j} : M\succcurlyeq 0, M_{ii}\le 1, \textstyle\sum_i M_{ii}\le 2k\}.$$

This leads to the following B' inequality, $\forall\, j\in 2^{[\log(1/\Delta)]}$,

$$\max_{M\in\boldsymbol{R}^j}\langle M, F^j_{B'}\rangle \le \mathcal{O}(\tfrac{\beta\log\frac{1}{\beta}}{n\cdot\log^4 j}) = \mathcal{O}(\tfrac{\Delta^2}{\beta\log^4 j}). \quad (2)$$

Lemma 16 in the appendix shows that any B' with $|B_G\cap B'|\le(1-2\beta)|B_G|$ that satisfies this inequality also satisfies Inequality (1).

Next, we describe a filtering algorithm that finds $B'\subseteq B$ satisfying the new inequality.

To find such a batch sub-collection, we show that for all $B'\subseteq B$ such that $|B_G\cap B'|\le(1-2\beta)|B_G|$ good batches, the following conditions hold:

1. There is a computationally efficient algorithm for finding $\operatorname{argmax}\{\langle M, F^j_{B'}\rangle : M\in\boldsymbol{R}^j\}$.
2. Given an M for which $\langle M, F^j_{B'}\rangle$ is large, we can delete batches from B' such that in expectation we delete 3 times more adversarial batches than good.
3. If B' has no adversarial batches, it satisfies (2).

The algorithm consists of a main part (Algorithm 1) that sequentially over $j\in 2^{[\log(1/\Delta)]}$ checks if Equation (2) is satisfied for partition $\mathcal{P}^j$. If not, it iteratively calls subroutine Batch-Deletion (Algorithm 2), to delete the appropriate batches. Due to space limitations we present the pseudo code for Algorithm 1 and Algorithm 2 in the appendix. Next, we argue that the algorithm identifies B' for which (2) holds.

It starts with $B'=B$, and sequentially over $j\in 2^{[\log(1/\Delta)]}$, perform the following recursive algorithm. Efficiently find M maximizing $\langle M, F^j_{B'}\rangle$ (condition 1). Use M to delete batches $b\in B'$ for which $\langle M, C^j_{b,B'}\rangle$ is high. Continue until Equation (2) holds for j. As the algorithm proceeds, so long as Equation (2) fails to hold, Condition 2 ensures that the algorithm removes more adversarial batches than good batches (in expectation). Observe that without adversarial batches, Equation (2) holds. Hence, at the latest, when all adversarial batches are removed, the condition 3 ensures Equation (2) will hold and algorithm will stop. The second condition ensures that w.h.p. the algorithm does not remove more than more than $|B_A|/2=\beta(1-\beta)B_G/2$, which for $\beta\le 1/6$ is $\le 2\beta B_G$ good batches, before removing all adversarial batches.

Hence in the end B' will satisfy Equation (2), and therefore Equation (1). The empirical distribution $\bar{p}_{B'}$ achieves the guarantee in Theorem 1.

In the appendix, we derive the above filtering conditions by using the following concentration properties of good batches.

Essential Properties of good batches: For all sub-collections $B'_G\subseteq B_G$ of good batches, $j\in 2^{[\log(1/\Delta)]}$, and $M\in\boldsymbol{R}^j$:

1. If $|B'_G|\ge(1-2\beta)|B_G|$, then
   (a) $\langle M, (\bar{p}^j_{B'_G}-p^j)^{\otimes 2}\rangle \le \mathcal{O}\Big(\frac{\Delta^2}{\log^4 j}\Big)$,
   (b) $\langle M, F^j_{B'_G}\rangle \le \mathcal{O}\Big(\frac{\Delta^2}{\beta\log^4 j}\Big)$.
2. If $|B'_G|\le 2\beta|B_G|$, then
   $\sum_{b\in B'_G}\langle M, (\bar{\mu}^j_b-p^j)^{\otimes 2}\rangle \le \mathcal{O}\Big(|B_G|\cdot\frac{\Delta^2}{\beta\log^4 j}\Big)$.

The next theorem shows that w.h.p. the good batch collection B_G satisfies the above properties.

Theorem 12. *For some constants $c<1/2$ and $C>1$, for any k, $\beta<c$, $\delta<1$, $n>\Omega(\log^C(1/\beta))$, and discrete or continuous p. If $|B_G|\cdot n=\tilde{\Omega}(\frac{k+\log(1/\delta)}{\Delta^2})$, then the essential properties hold with probability $\ge 1-\delta$.*

Crucially, the Theorem shows that for carefully chosen SDP relaxation $\boldsymbol{R}^j$ of the set of $2k$ sparse binary vectors, the filtering properties hold with only $\tilde{\Omega}(k)$ samples. By comparison, (Chen *et al.*, 2020) used a convex relaxation of binary vectors that are sparse in Haar basis, and for that relaxation they showed $\tilde{\mathcal{O}}(k^2)$ sample complexity.

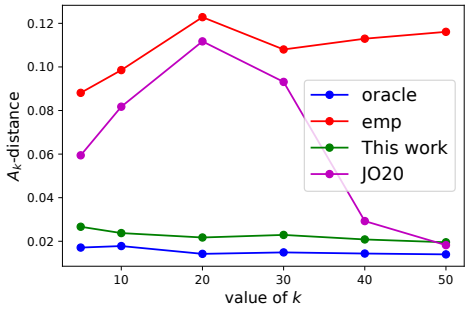
Let $\mathcal{L}^j_i : \{v\in\mathbb{R}^j : ||v||_\infty=1, ||v||_2^2\le i\}$. The next theorem shows that to prove that the above properties hold for all elements in $\boldsymbol{R}^j$ it suffices to show that the property holds for the following strictly smaller set $\{vv^\intercal : v\in\mathcal{L}^{k\cdot j}_{2k}\}$.

Theorem 13. *Consider an $n\times n$ symmetric matrix A of real numbers. Then there is a universal constant $K_G\le 1.7822$ such that*

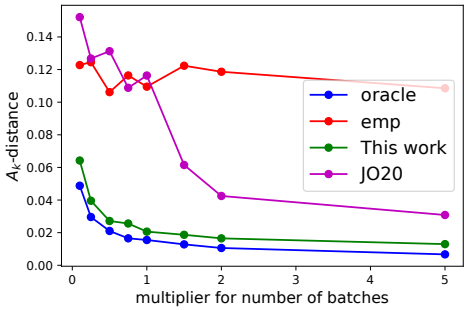
$$\max_{M\in\boldsymbol{R}^j}|\langle M, A\rangle| \le 2\cdot K_G \max_{v\in\mathcal{L}^{k\cdot j}_{2k}}|\langle vv^\intercal, A\rangle|.$$

We derive the above theorem in Appendix G using Grothendieck's inequality.

The set $\{vv^\intercal : v\in\mathcal{L}^{k\cdot j}_{2k}\}$ is still infinite. Even its $o(1)$ cover can be shown to have size exponential in $\tilde{\Omega}(k\cdot j)$. Taking the union bound on the cover elements, as in the previous



(a) $\mathcal{A}_k$ distance vs. k with constant no. of samples to k ratio



(b) $\mathcal{A}_k$ distance vs. no. of samples (batches)

Figure 1. Learning distributions in $\mathcal{A}_k$ distance

works, would yield only a sub-optimal $\max_j \tilde{\mathcal{O}}(k \cdot j/\Delta^2) = \tilde{\mathcal{O}}(k/\Delta^3)$ sample complexity. But applying a much more nuanced and complex technique, we obtain the optimal sample complexity $\tilde{\mathcal{O}}(k/\Delta^2)$. Due to space constraint we leave the details to Appendix G and I.

5. Experiments

We corroborate our results by performing simulations.

We present here experiments for our main technical contribution, robustly learning arbitrary distributions to $\mathcal{A}_k$ distance using just $\mathcal{O}(k)$ samples, even when the domain size is much larger than k. The simulations for learning continuous distribution in TV distance are relegated to the appendix.

For discrete distributions we set the domain size ℓ to 500. We select this rather large value to show that the algorithm is practical for large domains, where exploiting the structure becomes more important.

We show two plots, for both we set the fraction of adversarial batches to a relatively high value $\beta = 0.4$ and the batch size to a moderate value of 500. This shows that the algorithms perform well even when corruption is high and batch size is only moderate. Note that the algorithm's performance will improve if we increase the batch size or decrease β.

We compare the performance of our algorithm with three other estimators. The first is a powerful oracle, who knows which batches are good batches and uses their empirical distribution as its estimate. The performance of Oracle shows the information theoretic limit in absence of adversarial batches. The second estimator is the standard empirical estimator that simply returns the empirical distribution of all samples in B. The third estimator is the (Jain & Orlitsky, 2020) filtering-based estimator. We also considered the estimator of (Chen *et al.*, 2020), however for the large domain size we test our algorithm on, the implementation of their algorithm provided with their paper took several hours even for a single run, while our estimator took on average less than three minutes.

The simulations were performed on a laptop with a configuration of 2.3 GHz Intel Core i7 CPU and 16 GB of RAM. We took the average of 10 runs to plot the results. For both plots we select p by generating a random vector in $[0, 1]^\ell$ and normalizing it. We tried various adversarial distribution: a randomly chosen distribution similar to p; a randomly generated k piecewise histogram; and their linear combination with p. For each estimator we plot the results for worst adversarial distribution.

In our first simulation we verify that our algorithm can learn large discrete distributions in $\mathcal{A}_k$ distance, with a number of samples only linear in k. We choose the a rather large alphabet size $\ell = 500$ and test for various values of k from $10, 20, 30, 40, 50$. For each k we choose the number of good batches to be k/β^2. Our plots show that the error achieved by our algorithm essentially remains the same as k increases, demonstrating the linear dependence of the sample complexity on k. Our algorithm nearly achieves the performance of the oracle that enjoys the best statistical guarantee, even for the non-adversarial setting. Note that results in $\mathcal{A}_k$ learning imply the other results.

In the second plot we keep k constant and increase number of good batches as fk/β^2, for factor $f = [0.01, 0.25, 0.5, 0.75, 1, 1.5, 2, 5]$.

Acknowledgements

We thank Vaishakh Ravindrakumar for running the SURF (Hao *et al.*, 2020) algorithm used to conduct the continuous density estimation experiments in the appendix, the authors of (Chen *et al.*, 2020) for making their code for computing $\mathcal{A}_k$ distance available on Github, the anonymous reviewers for helpful comments, and the National Science Foundation for supporting this work through grants CIF-1564355 and CIF-1619448.

References

Acharya, Jayadev, Diakonikolas, Ilias, Li, Jerry, & Schmidt, Ludwig. 2017. Sample-optimal density estimation in nearly-linear time. *Pages 1278–1289 of: Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM.

Anthony, Martin, & Shawe-Taylor, John. 1993. A result of Vapnik with applications. *Discrete Applied Mathematics*, **47**(3), 207–217.

Ashtiani, Hassan, & Mehrabian, Abbas. 2018. Some techniques in density estimation. *arXiv preprint arXiv:1801.04003*.

Awasthi, Pranjal, Blum, Avrim, Haghtalab, Nika, & Mansour, Yishay. 2017. Efficient PAC learning from the crowd. *Pages 127–150 of: Conference on Learning Theory*. PMLR.

Center for Disease Control. 2020. *CDC Influenza Vaccine 2020/2021*. https://www.cdc.gov/flu/season/faq-flu-season-2020-2021.htm.

Chan, Siu-On, Diakonikolas, Ilias, Servedio, Rocco A, & Sun, Xiaorui. 2014. Efficient density estimation via piecewise polynomial approximation. *Pages 604–613 of: Proceedings of the forty-sixth annual ACM symposium on Theory of computing*.

Charikar, Moses, Steinhardt, Jacob, & Valiant, Gregory. 2017. Learning from untrusted data. *Pages 47–60 of: Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM.

Chen, Sitan, Li, Jerry, & Moitra, Ankur. 2019. Efficiently Learning Structured Distributions from Untrusted Batches. *arXiv preprint arXiv:1911.02035*.

Chen, Sitan, Li, Jerry, & Moitra, Ankur. 2020. Learning Structured Distributions From Untrusted Batches: Faster and Simpler. *arXiv preprint arXiv:2002.10435*.

Devroye, Luc, & Lugosi, Gabor. 2001. *Combinatorial Methods in Density Estimation*. Springer Science & Business Media.

Diakonikolas, Ilias. 2016. Learning Structured Distributions. *Handbook of Big Data*, **267**.

Diakonikolas, Ilias, Kamath, Gautam, Kane, Daniel M, Li, Jerry, Moitra, Ankur, & Stewart, Alistair. 2016. Robust Estimators in High Dimensions without the Computational Intractability. *Pages 655–664 of: 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE.

Diakonikolas, Ilias, Kamath, Gautam, Kane, Daniel M, Li, Jerry, Moitra, Ankur, & Stewart, Alistair. 2017. Being robust (in high dimensions) can be practical. *Pages 999–1008 of: Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

Diakonikolas, Ilias, Kamath, Gautam, Kane, Daniel M, Li, Jerry, Steinhardt, Jacob, & Stewart, Alistair. 2018. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*.

Diakonikolas, Ilias, Kamath, Gautam, Kane, Daniel, Li, Jerry, Moitra, Ankur, & Stewart, Alistair. 2019. Robust Estimators in High-Dimensions Without the Computational Intractability. *SIAM Journal on Computing*, **48**(2), 742–864.

Hao, Yi, Jain, Ayush, Orlitsky, Alon, & Ravindrakumar, Vaishakh. 2020. SURF: A Simple, Universal, Robust, Fast Distribution Learning Algorithm. *arXiv preprint arXiv:2002.09589*.

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

Jain, Ayush, & Orlitsky, Alon. 2019. Optimal Robust Learning of Discrete Distributions from Batches. *arXiv preprint arXiv:1911.08532*.

Jain, Ayush, & Orlitsky, Alon. 2020. A General Method for Robust Learning from Batches. *arXiv preprint arXiv:2002.11099*.

Kamath, Sudeep, Orlitsky, Alon, Pichapati, Dheeraj, & Suresh, Ananda Theertha. 2015. On learning distributions from their samples. *Pages 1066–1100 of: Conference on Learning Theory*.

Konstantinov, Nikola, Frantar, Elias, Alistarh, Dan, & Lampert, Christoph. 2020. On the sample complexity of adversarial multi-source pac learning. *Pages 5416–5425 of: International Conference on Machine Learning*. PMLR.

Lai, Kevin A, Rao, Anup B, & Vempala, Santosh. 2016. Agnostic estimation of mean and covariance. *Pages 665–674 of: 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE.

Maass, Wolfgang. 1994. Efficient agnostic pac-learning with simple hypothesis. *Pages 67–75 of: Proceedings of the seventh annual conference on Computational learning theory*.

O'Brien, Carl M. 2016. Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics. *International Statistical Review*, **84**(2), 318–319.

Qiao, Mingda. 2018. Do Outliers Ruin Collaboration? *Pages 4180–4187 of: International Conference on Machine Learning*. PMLR.

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

Rigollet, Philippe. 2015. 18.S997 High-Dimensional Statistics. *Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA*.

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

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

Vaart, Aad W, & Wellner, Jon A. 1996. *Weak convergence and empirical processes: with applications to statistics*. Springer.

Vapnik, Vladimir, & Chervonenkis, Alexey. 1974. *Theory of pattern recognition*.

Wikipedia. 2020. *Historical Annual Reformulations of the Influenza Vaccine*. https://en.wikipedia.org/wiki/Historical_annual_reformulations_of_the_influenza_vaccine.