On the Sample Complexity of Adversarial Multi-Source PAC Learning

Nikola Konstantinov [1] **Elias Frantar** [1 2] **Dan Alistarh** [1] **Christoph H. Lampert** [1]

Abstract

We study the problem of learning from multiple untrusted data sources, a scenario of increasing practical relevance given the recent emergence of crowdsourcing and collaborative learning paradigms. Specifically, we analyze the situation in which a learning system obtains datasets from multiple sources, some of which might be biased or even adversarially perturbed. It is known that in the single-source case, an adversary with the power to corrupt a fixed fraction of the training data can prevent PAC-learnability, that is, even in the limit of infinitely much training data, no learning system can approach the optimal test error. In this work we show that, surprisingly, the same is not true in the multi-source setting, where the adversary can arbitrarily corrupt a fixed fraction of the data sources. Our main results are a generalization bound that provides finite-sample guarantees for this learning setting, as well as corresponding lower bounds. Besides establishing PAC-learnability our results also show that in a cooperative learning setting sharing data with other parties has provable benefits, even if some participants are malicious.

1. Introduction

An important problem of current machine learning research is to make learned systems more *trustworthy*. One particular aspect of this is *robustness* against data of unexpected or even adversarial nature. Robustness at prediction time has recently received a lot of attention, in particular with work on the detection of *out-of-distribution* conditions (Hendrycks & Gimpel, 2017; Liang et al., 2018; Lee et al., 2018) and protection against *adversarial examples* (Raghunathan et al., 2018; Singh et al., 2018; Cohen et al., 2019). Robustness at training time, however, is represented less prominently, despite also being of great importance. One reason might be that learning from a potentially adversarial data source is very hard: a classic result states that when a fixed fraction of the training dataset is adversarially corrupted, successful learning in the PAC sense is not possible anymore (Kearns & Li, 1993). In other words, there exists no *robust learning algorithm* that could overcome the effects of adversarial corruptions in a constant fraction of the training dataset and approach the optimal model, even in the limit of infinite data.

In this work, we study the question of robust learning in the multi-source case, i.e. when more than one dataset is available for training. This is a situation of increasing relevance in the era of *big data*, where machine learning models tend to be trained on very large datasets. To create these, one commonly relies on distributing the task of collecting and annotating data, e.g. to crowdsourcing (Sheng & Zhang, 2019) services, or by adopting a collective or federated learning scenario (McMahan & Ramage, 2017).

Unfortunately, relying on data from other parties comes with the danger that some of the sources might produce data of lower quality than desired, be it due to negligence, bias or malicious behaviour. Consequently, the analogous question to the classic problem described above is the following, which we refer to as *adversarial multi-source learning*. *Given a number of i.i.d. datasets, a constant fraction of which might have been adversarially manipulated, is there a learning algorithm that overcomes the effect of the corruptions and approaches an optimal model?*

In this work, we study this problem formally and provide a positive answer. Specifically, *our main result is an upper bound on the sample complexity of adversarial multi-source learning, that holds as long as less than half of sources are manipulated (Theorem 1).*

A number of interesting results follow as immediate corollaries. First, we show that any hypothesis class that is uniformly convergent and hence PAC-learnable in the classical i.i.d. sense is also PAC-learnable in the adversarial multi-source scenario. This is in stark contrast to the single-source situation where, as mentioned above, no non-trivial hypothesis class is robustly PAC-learnable. As a second consequence, we obtain the insight that in a cooperative learning

[1]Institute of Science and Technology Austria, Klosterneuburg, Austria [2]Vienna University of Technology, Vienna, Austria. Correspondence to: Nikola Konstantinov <nkonstan@ist.ac.at>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

scenario, every honest party can benefit from sharing their data with others, as compared to using their own data only, even if some of the participants are malicious.

Besides our main result we prove two additional theorems that shed light on the difficulty of adversarial multi-source learning. First, we prove that the naïve but common strategy of simply merging all data sources and training with some robust procedure on the joint dataset cannot result in a robust learning algorithm (Theorem 2). Second, we prove a lower bound on the sample complexity under very weak conditions (Theorem 3). This result shows that under adversarial conditions a slowdown of convergence is unavoidable, and that in order to approach optimal performance, the number of samples per source must necessarily grow, while increasing the number of sources need not help.

2. Related work

To our knowledge, our results are the first that formally characterize the statistical hardness of learning from multiple i.i.d. sources, when a constant fraction of them might be adversarially corrupted. There are a number of conceptually related works, though, which we will discuss for the rest of this section.

Qiao & Valiant (2018), as well as the follow-up works of Chen et al. (2019); Jain & Orlitsky (2019), aim at estimating discrete distributions from multiple batches of data, some of which have been adversarially corrupted. The main difference to our results is the focus on finite data domains and estimating the underlying probability distribution rather than learning a hypothesis.

Qiao (2018) studies collaborative binary classification: a learning system has access to multiple training datasets and a subset of them can be adversarially corrupted. In this setup, the uncorrupted sources are allowed to have different input distributions, but share a common labelling function. The author proves that it is possible to robustly learn individual hypotheses for each source, but a single shared hypothesis cannot be learned robustly. For the specific case that all data distributions are identical, the setup matches ours, though only for binary classification in the realizable case, and with a different adversarial model.

In a similar setting, Mahloujifar et al. (2019) show, in particular, that an adversary can increase the probability of any "bad property" of the learned hypothesis by a term at least proportional to the fraction of manipulated sources. These results differ from ours, by their assumption that different sources have different distributions, which renders the learning problem much harder.

In Konstantinov & Lampert (2019), a learning system has access to multiple datasets, some of which are manipulated, and the authors prove a generalization bound and propose an algorithm based on learning with a weighted combination of all datasets. The main difference to our work is that their proposed method crucially relies on a trusted subset of the data being known to the learner. Their adversary is also weaker, as it cannot influence the data points directly, but only change the distribution from which they are sampled, and the work does not provide finite sample guarantees.

There are a number of classic results on the fundamental limits of PAC learning from a *single labelled set of samples*, a fraction of which can be arbitrarily corrupted, e.g. (Kearns & Li, 1993; Bshouty et al., 2002). We compare our results against this classic scenario in Section 4.1.

Another related general direction is the research on Byzantine-resilient distributed learning, which has seen significant interest recently, e.g. (Blanchard et al., 2017; Chen et al., 2017; Yin et al., 2018; 2019; Alistarh et al., 2018). There the focus is on learning by exchanging gradient updates between nodes in a distributed system, an unknown fraction of which might be corrupted by an omniscient adversary and may behave arbitrarily. These works tend to design defences for specific gradient-based optimization algorithms, such as SGD, and their theoretical analysis usually assumes strict conditions on the objective function, such as convexity or smoothness. Nevertheless, the (nearly) tight sample complexity upper and lower bounds developed for Byzantine-resilient gradient descent (Yin et al., 2018) and its stochastic variant (Alistarh et al., 2018) are relevant to our results and are therefore discussed in detail in Sections 4.2 and 5.2.

The work of Awasthi et al. (2017) considers learning from crowdsourced data, where some of the workers might behave arbitrarily. However, they only focus on label corruptions. Feng (2017) consider the fundamental limits of learning from adversarial distributed data, but in the case when *each of the nodes* can iteratively send corrupted updates with certain probability. Feng et al. (2014) provide a method for distributing the computation of any robust learning algorithm that operates on a single large dataset. There is also a large body of literature on attacks and defences for federated learning, e.g. (Bhagoji et al., 2019; Fung et al., 2018). Apart from focusing on iterative gradient-based optimization procedures, these works also allow for natural variability in the distributions of the uncorrupted data sources.

After the submission deadline, two further relevant works have appeared on ArXiv. In particular, (Jain & Orlitsky, 2020) extend the framework of (Qiao & Valiant, 2018) to learning of distributions over infinite domains, from untrusted batches of data. They also develop a robust algorithm for binary classification, achieving similar statistical rates to ours. Regarding hardness results, (Hanneke & Kpotufe,

2020) recently gave a number of sample complexity lower bounds for multi-task learning with N tasks that share an optimal hypothesis. Their work shows in particular that no learning procedure can achieve optimal risk as $N \to \infty$, even in this non-adversarial setting, as long as the number of data points per task is constant and the learner has no additional information about the relationship between the tasks, apart from the given data.

3. Preliminaries

In this section we introduce the technical definitions that are necessary to formulate and prove our main results. We start by reminding the reader of the classical notion of PAC-learnability and uniform convergence, as they can be found in most machine learning textbooks. We then introduce the setting of learning from multiple sources and notions of adversaries of different strengths.

3.1. Notation and Background

Let $\mathcal{X}$ and $\mathcal{Y}$ be given input and output sets, respectively, and $\mathcal{D} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be a fixed but unknown probability distribution. By $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ we denote a loss function, and by $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$ a set of hypotheses. All of these quantities are assumed arbitrary but fixed for the purpose of this work.

A *(statistical) learner* is a function $\mathcal{L} : \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$. In the classic supervised learning scenario, the learner has access to a *training set* of m labelled examples, $\{(x_1, y_1), \ldots, (x_m, y_m)\}$, sampled i.i.d. from D, and aims at learning a hypothesis $h \in \mathcal{H}$ with small *risk*, i.e. expected loss, under the unknown data distribution,

$$\mathcal{R}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}(\ell(h(x), y)). \tag{1}$$

PAC-learnability is a key property of the hypothesis set, which ensures the existence of an algorithm that performs successful learning:

Definition 1 (PAC-Learnability). We call $\mathcal{H}$ (agnostic) *probably approximately correct (PAC) learnable* with respect to ℓ, if there exists a learner $\mathcal{L}$ and a function $m_{\mathcal{H},\ell} : (0,1)\times(0,1) \to \mathbb{N}$, such that for any $\epsilon, \delta \in (0,1)$, whenever S is a set of $m \geq m_{\mathcal{H},\ell}(\epsilon, \delta)$ i.i.d. labelled samples from $\mathcal{D}$, then with probability at least $1 - \delta$ over the sampling of S:

$$\mathcal{R}(\mathcal{L}(S)) \leq \min_{h\in\mathcal{H}} \mathcal{R}(h) + \epsilon. \tag{2}$$

Another important concept related to PAC-learnability is that of *uniform convergence*.

Definition 2 (Uniform convergence). We say that $\mathcal{H}$ has the *uniform convergence* property with respect to ℓ with rate $s_{\mathcal{H},\ell}$, if there exists a function $s_{\mathcal{H},\ell} : \mathbb{N} \times (0,1) \times \bigcup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \to \mathbb{R}$, such that for any distribution $\mathcal{D} \in \mathcal{P}(X \times Y)$ and any $\delta \in (0,1)$:

- given m samples $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \overset{i.i.d.}{\sim} \mathcal{D}$, with probability at least $1 - \delta$ over the data :

$$\sup_{h\in\mathcal{H}} |\mathcal{R}(h) - \widehat{\mathcal{R}}(h)| \leq s_{\mathcal{H},\ell}(m, \delta, S), \tag{3}$$

  where $\hat{\mathcal{R}}(h)$ is the empirical risk of the hypothesis h.
- $s_{\mathcal{H},\ell}(m, \delta, S_m) \to 0$ as $m \to \infty$, for any sequence $(S_m)_{m\in\mathbb{N}}$ with $S_m \in (\mathcal{X} \times \mathcal{Y})^m$.

Throughout the paper we drop the dependence on $\mathcal{H}$ and ℓ and simply write s for $s_{\mathcal{H},\ell}$. Note that above definition is equivalent to the classic definition of uniform convergence (e.g. Chapter 4 in (Shalev-Shwartz & Ben-David, 2014)). We only introduce an explicit notation, s, for the sample complexity rate of uniform convergence, as this simplifies the layout of our analysis later. It is well-known that uniform convergence implies PAC-learnability and that the opposite is also true for agnostic binary classification (Shalev-Shwartz & Ben-David, 2014).

3.2. Multi-source learning

Our focus in this paper is on learning from multiple data sources. For simplicity of exposition, we assume that they all provide the same number of data points, i.e. the training data consists of N groups of m samples each, where $m, N \in \mathbb{N}$ are fixed integers.

Formally, we denote by $(\mathcal{X} \times \mathcal{Y})^{N\times m}$ the set of all possible collections (i.e. unordered sequences) of N groups of m datapoints each. A (statistical) *multi-source learner* is a function $\mathcal{L} : \cup_{N=1}^{\infty} \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^{N\times m} \to \mathcal{H}$ that takes such a collection of datasets and returns a predictor from $\mathcal{H}$.

3.3. Robust Multi-Source Learning

Informally, one considers a learning system *robust* if it is able to learn a good hypothesis, even when the training data is not perfectly i.i.d., but contains some artifacts, e.g. annotation errors, a selection bias or even malicious manipulations. Formally, one models this by assuming the presence of an *adversary*, that observes the original datasets and outputs potentially manipulated versions. The learner then has to operate on the manipulated data without knowledge of what the original one had been or what manipulations have been made.

Definition 3 (Adversary). An adversary is any function $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^{N\times m} \to (\mathcal{X} \times \mathcal{Y})^{N\times m}$.

Throughout the paper, we denote by $S' = \{S'_1, S'_2, \ldots, S'_N\}$ the original, uncorrupted datasets, drawn i.i.d. from $\mathcal{D}$, and by $S = \{S_1, S_2, \ldots, S_N\} = \mathcal{A}(S')$ the datasets returned by the adversary.

Different scenarios are obtained by giving the adversary different amounts of power. For example, a weak adversary

might only be able to randomly flip labels, i.e. simulate the presence of label noise. A much stronger adversary would be one that can potentially manipulate all data and do so with knowledge not only of all of the datasets but also of the underlying data distribution and the learning algorithm to be used later.

In this work, we adopt the latter view, as it leads to much stronger robustness guarantees. We define two adversary types that can make arbitrary manipulations to data sources, but only influence a certain subset of them.

Definition 4 (Fixed-Set Adversary). Let $G \subset [N]$. An adversary is called *fixed-set (with preserved set G)*, if it only influences the datasets outside of G. That is, $S_i = S_i^{'}$ for all $i \in G$.

Definition 5 (Flexible-Set Adversary). Let $k \in \{0, 1, \ldots, N\}$. An adversary is called *flexible-set* (with preserved size k), if it can influence any $N - k$ of the N given datasets. That is, there exists a set $G \subset [N]$, such that $|G| = k$ and $S_i = S_i'$ for all $i \in G$.

In both cases, we call the fraction α of corrupted datasets the *power* of the adversary, i.e. $\alpha = \frac{N-|G|}{N}$ for the fixed-set and $\alpha = \frac{N-k}{N}$ for the flexible-set adversaries.

While similarly defined, the fixed-set adversary is strictly weaker than the flexible-set one, as the latter one can first inspect all data and then choose which subset to modify, while the former one is restricted to a fixed, data-independent subset of sources. In particular, the flexible-set adversary can already bias the distribution of the data by throwing out a carefully chosen set of sources, before replacing them with new data.

Both adversary models are inspired by real-world considerations and analogs have appeared in a number of other research areas. The fixed-set adversaries can model a situation in which N parties collaborate on a single learning task, but an unknown and fixed set of them are compromised, e.g. by hackers, that can act maliciously and collude with each other. This is a similar reasoning as in *Byzantine-robust optimization*, where an unknown subset of computing nodes are assumed to behave arbitrarily, thereby disrupting the optimization progress.

The second adversary corresponds to a situation where a malicious party can observe all of the available datasets and choose which ones to corrupt, up to a certain budget. This is similar to classic models in the fields of robust PAC learning, e.g. (Bshouty et al., 2002), and robust mean estimation, e.g. (Diakonikolas et al., 2019), where the adversary itself can influence which subset of the data to modify once the whole dataset is observed.

Whether robust learning in the presence of an adversary is possible for a certain hypothesis set or not is captured by the following definition:

Definition 6. A hypothesis set, $\mathcal{H}$, is called *multi-source PAC-learnable* against the class of fixed-set adversaries (or flexible-set adversaries) and with respect to ℓ, if there exists a multi-source learner $\mathcal{L}$ and a function $m : (0,1)^2 \to \mathbb{N}$, such that for any $\epsilon, \delta \in (0,1)$ and any set $G \subset [N]$ of size $|G| > \frac{1}{2}N$ (or any $\alpha < \frac{1}{2}$), whenever $S^{'} \in (\mathcal{X} \times \mathcal{Y})^{N \times m}$ is a collection of N datasets of $m \geq m(\epsilon, \delta)$ i.i.d. labelled samples from $\mathcal{D}$ each, then with probability at least $1 - \delta$ over the sampling of $S^{'}$:

$$\mathcal{R}(\mathcal{L}(\mathcal{A}(S^{'})) \leq \min_{h \in \mathcal{H}} \mathcal{R}(h) + \epsilon, \tag{4}$$

uniformly against all fixed-set adversaries with preserved set G (or all flexible-set adversaries of power α). A learner, $\mathcal{L}$, with this property is called *robust multi-source learner* for $\mathcal{H}$.

In particular, the same learner $\mathcal{L}$ should work against any adversary and for any α or set G. In the same time, the adversary is arbitrary once $\mathcal{L}$ is fixed, so in particular it can depend on the learning algorithm.

Note that the robust learner should achieve optimal error as $m \to \infty$, while N can stay constant. This reflects that we want to study adversarial multi-source learning in the context of a constant and potentially not very large number of sources. In fact, our lower bound results in Section 5 show that the adversary can always prevent the learner from approaching optimal risk in the opposite regime of constant m and $N \to \infty$.

4. Sample Complexity of Robust Multi-Source Learning

In this section, we present our main result, a theorem that states that whenever $\mathcal{H}$ has the uniform convergence property, there exists an algorithm that guarantees a bounded excess risk against both the fixed-set and the flexible-set adversary. We then derive and discuss some instantiations of the general result that shed light on the sample complexity of PAC learning in the adversarial multi-source learning setting. Finally, we provide a high-level sketch of the theorem's proof.

4.1. Main result

Theorem 1. *Let $N, m, k \in \mathbb{N}$ be integers, such that $k \in (N/2, N]$. Let $\alpha = \frac{N-k}{N} < \frac{1}{2}$ be the proportion of corrupted sources. Assume that $\mathcal{H}$ has the uniform convergence property with rate function s. Then there exists a learner $\mathcal{L} : (\mathcal{X} \times \mathcal{Y})^{N \times m} \to \mathcal{H}$ with the following two properties.*

(a) Let G be a fixed subset of $[N]$ of size $|G| = k$. For

$S' = \{S'_1, \dots, S'_N\} \overset{i.i.d.}{\sim} \mathcal{D}$, *with probability at least* $1-\delta$ *over the sampling of* S'*:*

$$\mathcal{R}(\mathcal{L}(\mathcal{A}(S'))) - \min_{h\in\mathcal{H}} \mathcal{R}(h) \leq 2s\big(km, \frac{\delta}{2}, S_G\big) + 6\alpha \max_{i\in[N]} s\big(m, \frac{\delta}{2N}, S_i\big), \tag{5}$$

uniformly against all fixed-set adversaries with preserved set G*, where* $S = \{S_1, \dots, S_N\} = \mathcal{A}(S')$ *is the dataset modified the adversary and* $S_G = \cup_{i\in G} S_i$ *is the set of all uncorrupted data.*

(b) For $S' = \{S'_1, \dots, S'_N\} \overset{i.i.d.}{\sim} \mathcal{D}$*, with probability at least* $1-\delta$ *over the sampling of* S'*:*

$$\mathcal{R}(\mathcal{L}(\mathcal{A}(S'))) - \min_{h\in\mathcal{H}} \mathcal{R}(h) \leq 2s\big(km, \frac{\delta}{2\binom{N}{k}}, S_G\big) + 6\alpha \max_{i\in[N]} s\big(m, \frac{\delta}{2N}, S_i\big), \tag{6}$$

uniformly against all flexible-set adversaries with preserved size k*, where* $S = \{S_1, \dots, S_N\} = \mathcal{A}(S')$ *is the dataset returned by the adversary,* G *is the set of sources not modified by the adversary and* $S_G = \cup_{i\in G} S_i$ *is the set of all uncorrupted data.*

The learner $\mathcal{L}$ is in fact explicit, we define and discuss it in the proof sketch that we provide in Section 4.3. The complete proof is provided in the supplementary material.

As an immediate consequence we obtain:

Corollary 1. *Assume that* $\mathcal{H}$ *has the uniform convergence property. Then* $\mathcal{H}$ *is multi-source PAC-learnable against the class of fixed-set and the class of flexible-set adversaries.*

Proof. It suffices to show that for any $\delta \in (0,1)$, the right hand sides of (5) and (6) converge to 0 for $m \to \infty$. This it true, since $s(\bar{m}, \bar{\delta}, \bar{S}) \to 0$ as $\bar{m} \to \infty$ for any $\bar{\delta}$ and $\bar{S}$, by the definition of uniform convergence. Since the same learner works regardless of the choice of G and/or α, the result follows. □

Discussion. Corollary 1 is in sharp contrast with the situation of single dataset PAC robustness. In particular, Bshouty et al. (2002) study a setup where an adversary can manipulate a fraction α datapoints out of a dataset with m i.i.d.-sampled elements[1]. The authors show that in the binary realizable case, for any hypothesis space with at least two functions, no learning algorithm can learn a hypothesis with risk less than 2α with probability greater than $1/2$. Similarly, Kearns & Li (1993) showed that for an adversary that modifies each data point with constant probability α, no algorithm can learn a hypothesis with accuracy better than $\alpha/(1-\alpha)$. Both results hold regardless of the value of m, thus showing that PAC-learnability is not fulfilled.

[1] To be precise, the number of influenced points has to be binomially distributed with mean αm, but the difference between this and the deterministic setting becomes irrelevant for $m \to \infty$.

4.2. Rates of convergence

While Theorem 1 is most general, it does not yet provide much insight into the actual sample complexity of the adversarial multi-source PAC learning problem, because the rate function s might behave in different ways. In this section we give more explicit upper bounds in terms a standard complexity measure of hypothesis spaces – the Rademacher complexity. Let

$$\mathfrak{R}_S(\ell \circ \mathcal{H}) = \mathbb{E}_\sigma\Big(\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(h(x_i), y_i)\Big), \tag{7}$$

be the (empirical) Rademacher complexity of $\mathcal{H}$ with respect to the loss function ℓ on a sample $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$. Here $\{\sigma_i\}_{i=1}^n$ are i.i.d. Rademacher random variables. Let $S_G = \bigcup_{i\in G} S_i$, $\mathfrak{R}_i = \mathfrak{R}_{S_i}(\ell \circ \mathcal{H})$ and $\mathfrak{R}_G = \mathfrak{R}_{S_G}(\ell \circ \mathcal{H})$.

4.2.1. Rates for the fixed-set adversary.

An application of Theorem 1 with a standard uniform concentration result gives:

Corollary 2. *In the setup of Theorem 1, against any fixed-set adversary, it holds that*

$$\mathcal{R}(\mathcal{L}(\mathcal{A}(S'))) - \min_{h\in\mathcal{H}} \mathcal{R}(h) \leq 4\mathfrak{R}_G + 6\sqrt{\frac{\log(\frac{4}{\delta})}{2km}} + \alpha\Big(18\sqrt{\frac{\log\left(\frac{4N}{\delta}\right)}{2m}} + 12\max_{i\in[N]} \mathfrak{R}_i\Big). \tag{8}$$

The full proof is included in the supplementary material.

In many common learning settings, the Rademacher complexity scales as $\mathcal{O}(1/\sqrt{n})$ with the sample size n (see e.g. (Bousquet et al., 2004)). Thereby, we obtain the following rates against the fixed-set adversary:

$$\widetilde{\mathcal{O}}\Big(\frac{1}{\sqrt{km}} + \alpha\frac{1}{\sqrt{m}}\Big), \tag{9}$$

where the $\widetilde{\mathcal{O}}$-notation hides constant and logarithmic factors.

The results in Corollary 2 and Equation (9) allow us to reason about the type of guarantees that can be achieved given a certain amount of data. However, they also imply an explicit upper bound on the sample complexity of adversarial multi-source learning (i.e. an upper bound on the smallest

possible $m(\epsilon, \delta)$ in Definition 6) of the form:

$$m(\epsilon, \delta) \leq \mathcal{O}\left(\frac{\log(\frac{N}{\delta})}{\epsilon^2}\left(\frac{1}{\sqrt{(1-\alpha)N}} + \alpha\right)^2\right). \quad (10)$$

Discussion. We can make a number of observations from Equation (9). The $\sqrt{1/km}$-term is the rate one expects when learning from k (uncorrupted) sources of m samples each, that is from all the available uncorrupted data. The $\sqrt{1/m}$-term reflects the rate when learning from any single source of m samples, i.e. without the benefit of sharing information between sources. The latter enters weighted by α, i.e. it is directly proportional to the power of the adversary. In the limit of $\alpha \to 0$ (i.e. all N sources are uncorrupted, $k \to N$), the bound becomes $\widetilde{\mathcal{O}}(\sqrt{1/Nm})$. Thus, we recover the classic convergence rate for learning from Nm samples in the non-realizable case. This fact is interesting, as the robust learner of Theorem 1 actually does not need to know the value of α for its operation. Consequently, the same algorithm will work robustly if the data contains manipulations but without an unnecessary overhead (i.e. with optimal rate), if all data sources are in fact uncorrupted.

Another insight follows from the fact that for reasonably small α, we have:

$$\widetilde{\mathcal{O}}\Big(\frac{1}{\sqrt{km}} + \alpha\frac{1}{\sqrt{m}}\Big) \ll \widetilde{\mathcal{O}}\Big(\frac{1}{\sqrt{m}}\Big), \quad (11)$$

so learning from multiple, even potentially manipulated, datasets converges to a good hypothesis faster than learning from a single uncorrupted dataset. This fact can be interpreted as encouraging cooperation: any of the *honest* parties in the multi-source setting with fixed-set adversary will benefit from making their data available for multi-source learning, even if some of the other parties are malicious.

Comparison to Byzantine-robust optimization. Our obtained rates for the fixed-set adversary can also be compared to the state-of-art convergence results for Byzantine-robust distributed optimization, where the compromised nodes are also fixed, but unknown. Yin et al. (2018) and Alistarh et al. (2018) develop robust algorithms for gradient descent and stochastic gradient descent respectively, achieving convergence rates of order

$$\widetilde{\mathcal{O}}\Big(\frac{1}{\sqrt{km}} + \alpha\frac{1}{\sqrt{m}} + \frac{1}{m}\Big) \quad (12)$$

for $\alpha < 1/2$ unknown. Clearly, these rates resemble ours, except for the additional $1/m$-term, which matters when α is 0 or very small. As shown in Yin et al. (2018), this term can also be made to disappear if an upper bound $\beta \geq \alpha$ is assumed to be known a priori.

Overall, these similarities should not be over-interpreted, as the results for Byzantine-robust optimization describe practical gradient-based algorithms for distributed optimization under various technical assumptions, such as convexity, smoothness of the loss function and bounded variance of the gradients. In contrast, our work is purely statistical, not taking computational cost into account, but holds in a much broader context, for any hypothesis space that has the uniform convergence property of suitable rate and without constraints on the optimization method to be used. Additionally, our rates improve automatically in situations where uniform convergence is faster.

4.2.2. Rates for the flexible-set adversary

An analogous result to Corollary 2 holds also for flexible-set adversaries:

Corollary 3. *In the setup of Theorem 1, against any flexible-set adversary, it holds that*

$$\mathcal{R}(\mathcal{L}(\mathcal{A}(S'))) - \min_{h\in\mathcal{H}}\mathcal{R}(h) \quad (13)$$
$$\leq 4\mathfrak{R}_G + 12\alpha\max_{i\in[N]}\mathfrak{R}_i + \widetilde{\mathcal{O}}\left(\frac{\sqrt[4]{\alpha}}{\sqrt{m}}\right).$$

The proof is provided in the supplemental material.

Making the same assumptions as above, we obtain a sample complexity rate

$$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{km}} + \frac{\sqrt[4]{\alpha}}{\sqrt{m}}\right). \quad (14)$$

which differs from (9) only in the rate of dependence on α[2], which, if at all, matters only for very small (but non-zero) α. Despite the difference, most of our discussion above still applies. In particular, even for the flexible-set adversary the same learning algorithm exhibits robustness for $\alpha > 0$ and achieves optimal rates for $\alpha = 0$.

Moreover, an explicit upper bound on the sample complexity against a flexible-set adversary is given by:

$$m(\epsilon, \delta) \leq \widetilde{\mathcal{O}}\left(\frac{1}{\epsilon^2}\left(\frac{1}{\sqrt{(1-\alpha)N}} + \sqrt[4]{\alpha}\right)^2\right). \quad (15)$$

4.3. Proof Sketch for Theorem 1

The proof of Theorem 1 consists of two parts. First, we introduce a filtering algorithm, that attempts to determine which of the data sources can be trusted, meaning that it should be safe to use them for training a hypothesis. Note that this

[2] In fact, we believe the $\sqrt[4]{\alpha}$-term to be an artifact of our proof technique, but currently do not have a bound with improved dependence on α.

Algorithm 1

input Datasets $S_1, \ldots, S_N$
  Initialize $\mathrm{T} = \{\}$ // trusted sources
  for $i = 1, \ldots, N$ **do**
    if $d_{\mathcal{H}}(S_i, S_j) \leq s\left(m, \frac{\delta}{2N}, S_i\right) + s\left(m, \frac{\delta}{2N}, S_j\right)$,
      for at least $\lfloor \frac{N}{2} \rfloor$ values of $j \neq i$, **then**
      $\mathrm{T} = \mathrm{T} \cup \{i\}$
    end if
  end for
output $\bigcup_{i \in \mathrm{T}} S_i$ // all data of trusted sources

can be because they were not manipulated, or because the manipulations are too small to have negative consequences. The output of the algorithm is a new *filtered* training set, consisting of all data from the trusted sources only. Second, we show that training a standard single-source learner on the filtered training set yields the desired results.

Step 1. Pseudo-code for the filtering algorithm is provided in Algorithm 1. The crucial component is a carefully chosen notion of distance between the datasets, called *discrepancy*, that we define and discuss below. It guarantees that if two sources are close to each other then the difference of training on one of them compared to the other is small.

To identify the trusted sources, the algorithm checks for each source how close it is to all other sources with respect to the discrepancy distance. If it finds the source to be closer than a threshold to at least half of the other sources, it is marked as trusted, otherwise it is not. To show that this procedure does what it is intended to do it suffices to show that two properties hold with high probability: 1) all trusted sources are safe to be used for training, 2) at least all uncorrupted sources will be trusted.

Property 1) follows from the fact that if a source has small distance to at least half of the other datasets, it must be close to at least one of the uncorrupted sources. By the property of the discrepancy distance, including it in the training set will therefore not affect the learning of the hypothesis very negatively. Property 2) follows from a concentration of mass argument, which guarantees that for any uncorrupted source its distance to all other uncorrupted sources will approach zero at a well-understood rate. Therefore, with a suitably selected threshold, at least all uncorrupted sources will be close to each other and end up in the trusted subset with high probability.

Discrepancy Distance. For any dataset $S_i \in (\mathcal{X} \times \mathcal{Y})^m$, let

$$\widehat{\mathcal{R}}_i(h) = \frac{1}{m} \sum_{(x,y) \in S_i} \ell(h(x), y) \tag{16}$$

be the empirical risk of a hypothesis h with respect to the loss ℓ. The (empirical) *discrepancy distance* between two datasets, S_i and S_j, is defined as

$$d_{\mathcal{H}}(S_i, S_j) = \sup_{h \in \mathcal{H}} \left(|\widehat{\mathcal{R}}_i(h) - \widehat{\mathcal{R}}_j(h)| \right). \tag{17}$$

This is the empirical counterpart of the so-called discrepancy distance, which, together with its unsupervised form, is widely adopted within the field of domain adaptation (Kifer et al., 2004; Ben-David et al., 2010; Mohri & Medina, 2012). Typically, the discrepancy is used to bound the maximum possible effect of distribution drift on a learning system. The metric was also used in (Konstantinov & Lampert, 2019) to measure the effect of training on sources that have been sampled randomly, but from adversarially chosen distributions. As shown in Kifer et al. (2004); Ben-David et al. (2010), for randomly sampled datasets, the empirical discrepancy concentrates with known rates to its distributional value, i.e. to zero, if two sources have the same underlying data distributions. The empirical discrepancy is well-defined even for data not sampled from a distribution, though, and together with the uniform convergence property it allows us to bound the effect of training on one dataset rather than another.

Step 2. Let $S_T = \bigcup_{i \in T} S_i$ be the output of the filtering algorithm, i.e. the union of all trusted datasets. Then, for any $h \in \mathcal{H}$, the empirical risk over S_T can be written as

$$\widehat{\mathcal{R}}_{\mathrm{T}}(h) = \frac{1}{|\mathrm{T}|} \sum_{i \in \mathrm{T}} \widehat{\mathcal{R}}_i(h) \tag{18}$$

We need to show that training on S_{T}, e.g. by minimizing $\widehat{\mathcal{R}}_{\mathrm{T}}(h)$, with high probability leads to a hypothesis with small risk under the true data distribution $\mathcal{D}$.

By construction, we know that for any trusted source S_i, there exists an uncorrupted source S_j, such that the difference between $\widehat{\mathcal{R}}_i(h)$ and $\widehat{\mathcal{R}}_j(h)$ is bounded by a suitably chosen constant (that depends on the growth function s). By the uniform convergence property of $\mathcal{H}$, we know that for any uncorrupted source, the difference between $\widehat{\mathcal{R}}_i(h)$ and the true risk $\mathcal{R}(h)$ can also be bounded in terms of the growth function s. In combination, we obtain that $\widehat{\mathcal{R}}_{\mathrm{T}}(h)$ is a suitably good estimator of the true risk, uniformly over all $h \in \mathcal{H}$. Consequently, S_{T} can be used for successful learning.

For the formal derivations and, in particular, the choice of thresholds, please see the supplemental material.

5. Hardness of Robust Multi-Source Learning

We now take an orthogonal view compared to Section 4, and study where the hardness of the multi-source PAC learning stems from and what allows us to nevertheless overcome it. For this, we prove two additional results that describe fundamental limits of how well a learner can perform in the multi-source adversarial setting.

For simplicity of exposition we focus on binary classification. Let $Y = \{-1, 1\}$ and ℓ be the zero-one loss, i.e. $\ell(y, \bar{y}) = [\![y \neq \bar{y}]\!]$. Following Bshouty et al. (2002), we define:

Definition 7. A hypothesis space $\mathcal{H}$ over an input set $\mathcal{X}$ is said to be *non-trivial*, if there exist two points $x_1, x_2 \in \mathcal{X}$ and two hypotheses $h_1, h_2 \in \mathcal{H}$, such that $h_1(x_1) = h_2(x_1)$, but $h_1(x_2) \neq h_2(x_2)$.

5.1. What makes robust learning possible?

We show that if the learner does not make use of the multi-source structure of the data, i.e. it behaves as a single-source learner on the union of all data samples, then a (multi-source) fixed-set adversary can always *prevent* PAC-learnability.

Theorem 2. *Let $\mathcal{H}$ be a non-trivial hypothesis space. Let m and N be any positive integers and let G be a fixed subset of $[N]$ of size $k \in \{1, \dots, N-1\}$. Let $\mathcal{L} : (\mathcal{X} \times \mathcal{Y})^{N \times m} \to \mathcal{H}$ be a multi-source learner that acts by merging the data from all sources and then calling a single-source learner. Let $S' \in (\mathcal{X} \times \mathcal{Y})^{N \times m}$ be drawn i.i.d. from $\mathcal{D}$. Then there exists a distribution $\mathcal{D}$ with $\min_{h \in \mathcal{H}} \mathcal{R}(h) = 0$ and a fixed-set adversary $\mathcal{A}$ with index set G, such that:*

$$\mathbb{P}_{S' \sim \mathcal{D}}\Big(\mathcal{R}\big(\mathcal{L}(\mathcal{A}(S'))\big) > \frac{\alpha}{8(1-\alpha)}\Big) > \frac{1}{20}, \tag{19}$$

where $\alpha = \frac{N-k}{N}$ is the power of the adversary.

The proof is provided in the supplemental material. Note that, since the theorem holds for the fixed-set adversary, it automatically also holds for the stronger flexible-set adversary.

The theorem sheds light on why PAC-learnability is possible in the multi-source setting, while in the single source setting it is not. The reason is not simply that the adversary is weaker, because it is restricted to manipulating samples in a subset of datasets instead of being able to choose freely. Inequality (19) implies that even against such a weaker adversary, a single-source learner cannot be adversarially robust. Consequently, it is the additional information that the data comes in multiple datasets, some of which remain uncorrupted even after the adversary was active, that gives the multi-source learner the power to learn robustly.

An immediate consequence of Theorem 2 is also that the common practice of merging the data from all sources and performing a form of empirical risk minimization on the resulting dataset is not a robust learner and therefore suboptimal in the studied context.

5.2. How hard is robust learning?

As a tool for understanding the limiting factors of learning in the adversarial multi-source setting, we now establish a lower bound on the achievable excess risk in terms of the number of samples per source and the power of the adversary.

Theorem 3. *Let $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$ be a hypothesis space, let m and N be any integers and let G be a fixed subset of $[N]$ of size $k \in \{1, \dots, N-1\}$. Let $S' \in (\mathcal{X} \times \mathcal{Y})^{N \times m}$ be drawn i.i.d. from $\mathcal{D}$. Then the following statements hold for any multi-source learner $\mathcal{L}$:*

(a) Suppose that $\mathcal{H}$ is non-trivial. Then there exists a distribution $\mathcal{D}$ on $\mathcal{X}$ with $\min_{h \in \mathcal{H}} \mathcal{R}(h) = 0$, and a fixed-set adversary $\mathcal{A}$ with index set G, such that:

$$\mathbb{P}_{S'}\Big(\mathcal{R}\big(\mathcal{L}(\mathcal{A}(S'))\big) > \frac{\alpha}{8m}\Big) > \frac{1}{20}. \tag{20}$$

(b) Suppose that $\mathcal{H}$ has VC dimension $d \geq 2$. Then there exists a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ and a fixed-set adversary $\mathcal{A}$ with index set G, such that:

$$\mathbb{P}_{S'}\Bigg(\mathcal{R}\big(\mathcal{L}(\mathcal{A}(S'))\big) - \min_{h \in \mathcal{H}} \mathcal{R}(h) > \sqrt{\frac{d}{1280Nm}} + \frac{\alpha}{16m}\Bigg) > \frac{1}{64}. \tag{21}$$

In both cases, $\alpha = \frac{N-k}{N}$ is the power of the adversary.

The proof is provided in the supplemental material. As for Theorem 2, it is clear that the same result holds also for flexible-set adversaries with preserved size k.

Analysis. Inequality (20) shows that even in the realizable scenario, the risk might not shrink faster than with rate $\Omega(\alpha/m)$, regardless of how many data sources, and therefore data samples, are available. This is contrast to the i.i.d. situation, where the corresponding rate is $\Omega(1/Nm)$. The difference shows that robust learning with a constant fraction of corrupted sources is only possible if the number of samples per dataset grows. Conversely, if the number of corrupted datasets is constant, regardless of the total number of sources, i.e., $\alpha = \mathcal{O}(1/N)$, we recover the rates for learning without an adversary up to constants.

In inequality (21), the term $\Omega(\sqrt{d/Nm})$ is due to the classic lower bound on the sample complexity of binary classification (e.g. Theorem 3.23 in (Mohri et al., 2018)) and corresponds to the fundamental limits of learning, now in the non-realizable case. The $\Omega(\alpha/m)$-term appears as the price of robustness, and as before, it implies that for constant α, $m \to \infty$ is necessary in order to achieve arbitrarily small excess risk, while just $N \to \infty$ does not suffice.

Relation to prior work. Lower bounds of similar structure as in Theorem 3 have also been derived for Byzantine optimization and collaborative learning. In particular, Yin et al.

(2018) prove that in the case of distributed mean estimation of a d-dimensional Gaussian on N machines, an α fraction of which can be Byzantine, any algorithm would incur loss of $\Omega(\frac{\alpha}{\sqrt{m}} + \sqrt{\frac{d}{Nm}})$. Alistarh et al. (2018) construct specific examples of a Lipschitz continuous and a strongly convex function, such that no distributed stochastic optimization algorithm, working with an α-fraction of Byzantine machines, can optimize the function to error less than $\Omega(\frac{\alpha}{\sqrt{m}} + \sqrt{\frac{d}{Nm}})$, where d is the number of parameters. For realizable binary classification in the context of collaborative learning, Qiao (2018) prove that there exists a hypothesis space of VC dimension d, such that no learner can achieve excess risk less than $\Omega(\alpha d/m)$.

Besides the different application scenario, the main difference between these results and Theorem 3 is that our bounds hold for *any* hypothesis space $\mathcal{H}$ that is non-trivial (Ineq. (20)), or has VC-dimension $d \geq 2$ (Ineq. (21)), while the mentioned references construct explicit examples of hypothesis spaces or stochastic optimization problems where the bounds hold. In particular, our results show that the limitations on the learner due the finite total number of samples, the finite number of samples per source and the fraction of unreliable sources α are inherent and not specific to a subset of hard-to-learn hypotheses.

6. Conclusion

We studied the problem of robust learning from multiple unreliable datasets. Rephrasing this task as learning from datasets that might be adversarially corrupted, we introduced the formal problem of adversarial learning from multiple sources, which we studied in the classic PAC setting.

Our main results provide a characterization of the hardness of this learning task from above and below. First, we showed that adversarial multi-source PAC learning is possible for any hypothesis class with the uniform convergence property, and we provided explicit rates for the excess risk (Theorem 1 and Corollaries). The proof is constructive and shows also that integrating robustness comes at a minor statistical cost, as our robust learner achieves optimal rates when run on data without manipulations. Second, we proved that adversarial PAC learning from multiple sources is far from trivial. In particular, it is impossible to achieve for learners that ignore the multi-source structure of the data (Theorem 2). Third, we proved lower bounds on the excess risk under very general conditions (Theorem 3), which highlight an unavoidable slowdown of the convergence rate proportional to the adversary's strength compared to the i.i.d. (adversarial-free) case. Furthermore, in order to facilitate successful learning with a constant fraction of corrupted sources, the number of samples per source has to grow.

A second emphasis of our work was to highlight connections of the adversarial multi-source learning task to related methods in robust optimization, cryptography and statistics. We believe that a better understanding of these connections will allow us to come up with tighter bounds and to design algorithms that are not only statistically efficient (as was the focus of this work), but also obtain insight into the trade-offs with computational complexity.

Acknowledgements

Dan Alistarh is supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 805223 ScaleML). This research was supported by the Scientific Service Units (SSU) of IST Austria through resources provided by Scientific Computing (SciComp).

References

Alistarh, D., Allen-Zhu, Z., and Li, J. Byzantine stochastic gradient descent. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Awasthi, P., Blum, A., Mansour, Y., et al. Efficient pac learning from the crowd. In *Conference on Computational Learning Theory (COLT)*, 2017.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

Bhagoji, A. N., Chakraborty, S., Mittal, P., and Calo, S. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learing (ICML)*, 2019.

Blanchard, P., Guerraoui, R., Stainer, J., et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, 2017.

Bousquet, O., Boucheron, S., and Lugosi, G. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pp. 169–207. Springer, 2004.

Bshouty, N. H., Eiron, N., and Kushilevitz, E. Pac learning with nasty noise. *Theoretical Computer Science*, 288(2): 255–275, 2002.

Chen, S., Li, J., and Moitra, A. Efficiently learning structured distributions from untrusted batches. In *ACM Symposium on Theory of Computing (STOC)*, 2019.

Chen, Y., Su, L., and Xu, J. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and*

Analysis of Computing Systems (POMACS), 1(2):1–25, 2017.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learing (ICML)*, 2019.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

Feng, J. On fundamental limits of robust learning. *arXiv preprint arXiv:1703.10444*, 2017.

Feng, J., Xu, H., and Mannor, S. Distributed robust learning. *arXiv preprint arXiv:1409.5937*, 2014.

Fung, C., Yoon, C. J., and Beschastnikh, I. Mitigating sybils in federated learning poisoning. *arXiv preprint arXiv:1808.04866*, 2018.

Hanneke, S. and Kpotufe, S. A no-free-lunch theorem for multi-task learning. *arXiv preprint arXiv:2006.15785*, 2020.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassifier and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

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

Jain, A. and Orlitsky, A. A general method for robust learning from batches. *arXiv preprint arXiv:2002.11099*, 2020.

Kearns, M. and Li, M. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 1993.

Kifer, D., Ben-David, S., and Gehrke, J. Detecting change in data streams. In *VLDB*, 2004.

Konstantinov, N. and Lampert, C. H. Robust learning from untrusted sources. In *International Conference on Machine Learing (ICML)*, 2019.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Mahloujifar, S., Mahmoody, M., and Mohammed, A. Universal multi-party poisoning attacks. In *International Conference on Machine Learing (ICML)*, 2019.

McMahan, H. B. and Ramage, D. Federated learning: Collaborative machine learning without centralized training data. https://research.googleblog.com/2017/04/federated-learning-collaborative.html, 2017.

Mohri, M. and Medina, A. M. New analysis and algorithm for learning with drifting distributions. In *International Conference on Algorithmic Learning Theory (ALT)*, 2012.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Qiao, M. Do outliers ruin collaboration? In *International Conference on Machine Learing (ICML)*, 2018.

Qiao, M. and Valiant, G. Learning discrete distributions from untrusted batches. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 94. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.

Sheng, V. S. and Zhang, J. Machine learning with crowdsourcing: A brief summary of the past research and future directions. In *AAAI Conference on Artificial Intelligence*, 2019.

Singh, G., Gehr, T., Mirman, M., Püschel, M., and Vechev, M. Fast and effective robustness certification. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Yin, D., Chen, Y., Ramchandran, K., and Bartlett, P. Byzantine-robust distributed learning: Towards optimal statistical rates. *International Conference on Machine Learing (ICML)*, 2018.

Yin, D., Chen, Y., Kannan, R., and Bartlett, P. Defending against saddle point attack in Byzantine-robust distributed learning. In *International Conference on Machine Learing (ICML)*, 2019.