Review of "Learning from Untrusted Data,"
Charikar, Steinhardt, and Valiant
CS 229T: Statistical Learning Theory

Bo Liu , Weiyun Ma
{bliuxix, wyma}@cs.stanford.edu
December 9, 2018

1 Problem Statement

The chosen paper [1] studies the following problem of learning from partially trusted data:

> *Given n data points, of which αn are from the distribution of interest, p^*, and the remaining $(1-\alpha)n$ can be chosen arbitrarily, then to what extent are learning problems solvable?*

This problem can be formally set up as follows:

> *Observe convex functions $f_1, \ldots, f_n : \mathcal{H} \to \mathbb{R}$, where $\mathcal{H} \subseteq \mathbb{R}^d$ is a convex parameter space. For a subset $I_{good} \subseteq [n]$ of size αn, $f_i \overset{i.i.d.}{\sim} p^*$ for $i \in I_{good}$, while the remaining f_i are chosen arbitrarily. The goal is to minimize the population mean $\bar{f} = \mathbb{E}_{f \sim p^*}[f]$. Typically, f_i is the loss incurred on the observed data point x_i, so the goal becomes minimizing the expected loss.*

Despite many previous work that put emphasis on the case when $\alpha \approx 1$, the authors focus on the case when $\alpha \ll 1$. In particular, the authors contend two frameworks accommodated to the presence of significant fraction of arbitrary (or adversarial) data: the *list-decodable* and the *semi-verified* learning models. Roughly speaking, *list-decodable* algorithm returns a list of possible answers while *semi-verified* algorithm returns a single answer but is allowed to peek at a few "verified" data points. The formal definition of both settings follows.

Definition 1 (List Decodable Learning)**.** We say that a learning, estimation, or optimzation problem is (m, ϵ) *list-decodably solvable* if an efficient algorithm can output a set of at most m hypotheses/estimates/answers, with the guarantee that at least one is accurate to within error ϵ.

Definition 2 (Semi-Verified Model)**.** In the *semi-verified* model, we observe n data points, of which an unknown αn are "real" data reflecting an underlying distribution p^*, and the remaining $(1-\alpha)n$ points are arbitrary. Furthermore, we observe k "verified" data points that are guaranteed to be drawn from p^*.

Intuitions. Here we provide some intuitions on why this problem is hard but important, why the two frameworks are natural options, and why the problem is still solvable. Due to the vast growth of the Internet, data poisoning attacks have become prevalent and it is both

interesting and necessary to look for algorithms that can learn from arbitrarily corrupted data. However, consider the above setting when $\alpha < \frac{1}{2}$. The learning problem seems *infeasible* since the adversary can simply make $1/\alpha$ copies of the "real" data to prevent any algorithm from distinguishing among those clusters of data. Therefore, the best we can hope for is that the algorithm can render a list of $O(1/\alpha)$ possible answers, with one of which close to the target. On the other hand, if the algorithm is allowed to peek at a few (often a single) data point from the true distribution, it can hopefully quickly differentiate the "real" data from the adversarial ones. The above two intuitions lead to the two frameworks we have discussed above. To see why this problem is still solvable, notice that, *if the adversarial data are statistically different from the "real one," we could easily tell they are outliers. By contrast, if they look similar enough to the honest data, the resulting estimation might not be biased too much.*

2 Main Results

Spectral norm of gradients. The following quantity is a key player throughout the main results of the paper [1]:

$$S \stackrel{\text{def}}{=} \max_{w \in \mathcal{H}} \frac{1}{\sqrt{|I_{\text{good}}|}} \left\| \left[\nabla f_i(w) - \nabla \bar{f}(w) \right]_{i \in I_{\text{good}}} \right\|_{\text{op}},$$

where $\bar{f}(w) \stackrel{\text{def}}{=} \mathbb{E}_{f \sim p^*}[f(w)]$ is the mean of $f(w)$ under p^*, and $||\cdot||_{\text{op}}$ denotes the spectral or operator norm. Intuitively, S is a uniform-convergence-type property of the parameter space $\mathcal{H}$ on the good data. In particular, it measures the extent to which the empirical gradient deviates from the expected gradient in terms of the matrix operator norm for every parameter $w \in \mathcal{H}$.

Throughout this review, we use $r \stackrel{\text{def}}{=} \max_{w \in \mathcal{H}} \|w\|_2$ to denote the ℓ_2-radius of $\mathcal{H}$, and let w^* be the minimizer of $\bar{f}$. We now state the first main result of the paper. We make the assumptions in bold text for clarity.

Theorem 1. *Given n data points containing a set I_{good} of αn data points* ***with spectral norm bound S****, we can obtain an ellipse $\mathcal{E}_Y = \{w \,|\, ww^T \preceq Y\}$ such that $\text{tr}(Y) \leq \mathcal{O}(\frac{r^2}{\alpha})$ and*

$$\min_{w \in \mathcal{E}_Y} \bar{f}(w) - \bar{f}(w^*) \leq \mathcal{O}\big(\frac{Sr}{\sqrt{\alpha}}\big). \tag{1}$$

Theorem 1 says that even when a large portion of data is untrusted, it is still possible to narrow the entire parameter space down to an (ideally much smaller) ellipse containing a parameter w such that $\bar{f}(w)$ is close enough to the optimum, provided the spectral norm bound S is small. Observe that when α is too small, the upper bound on the trace of the ellipse may be too large to be useful. To be more precise, consider the case $\alpha \leq \frac{1}{d}$. By the definition of r, the parameter space $\mathcal{H}$ is contained in the ℓ_2-ball $B(0; r)$ of radius r centered

at the origin. Note that $B(0;r)$ is precisely the ellipse defined by the matrix $r^2 I$, whose trace is $dr^2 \leq \frac{r^2}{\alpha}$. Thus, $B(0;r)$ is a trivial candidate of the ellipse, and the bound (1) becomes vacuous in this case (i.e. the LHS is 0). This also suggests that the bound (1) will be more useful when d is large.

Algorithm behind Theorem 1. Theorem 1 is based on an algorithm which repeatedly adjusts the restricted parameter space $\mathcal{E}_Y$ through identification and removal of outliers. Specifically, the algorithm assigns a separate parameter w_i to each function f_i, with the regularization condition that all the w_i should lie in a small ellipse $\mathcal{E}_Y$. The goal then is to minimize

$$\sum_{i=1}^{n} c_i f_i(w_i) + \lambda \operatorname{tr}(Y) \qquad \text{subject to } w_i w_i^T \preceq Y, \tag{2}$$

where c_i is the weight corresponding to f_i and λ is the regularization weight depending on the spectral norm bound S. The intuition behind (2) is that, if we simply want to minimize the empirical loss $\sum_{i=1}^{n} f_i(w)$, then even a single adversary data point may significantly bias the minimizer $\hat{w}$. The setup of (2) avoids this issue by having individual parameters w_i for f_i.

Once a minimizer $(\{\hat{w}_i\}, \hat{Y})$ of (2) is obtained, if the ellipse $\mathcal{E}_{\hat{Y}}$ satisfies the trace bound in Theorem 1, then we are done. Otherwise, the algorithm will further shrink the ellipse by identifying outliers, adjusting the weights c_i accordingly, and solving (2) again with the updated weights. In particular, outliers are identified based on the following intuition: Around a good function f_i usually clusters a large number of other good functions f_j, $j \neq i$, since they are all drawn from the same distribution p^*. Therefore, one can expect to use a combination of the corresponding parameters $\hat{w}_j$ to effectively minimize f_i. On the contrary, an outlier f_i is not likely to be well-minimized by too many parameters $\hat{w}_j$ for other functions f_j, $j \neq i$.

Specializing to strongly convex losses. The authors use Theorem 1 to derive bounds in the more specific list-decodable and the semi-verified learning models. Here we focus on the following result for the list-decodable model when the functions f_i are strongly convex in w:

Theorem 2. *Suppose the functions* $\boldsymbol{f_i}$ ***are*** $\boldsymbol{\kappa}$***-strongly convex****, and suppose there is a set* I_{good} *of size* αn ***with spectral norm bound*** $\boldsymbol{S}$*. Then, for any* $\epsilon \leq \frac{1}{2}$*, it is possible to obtain a set of* $m \leq \lfloor \frac{1}{(1-\epsilon)\alpha} \rfloor$ *candidate parameters* $\hat{w}_1, \ldots \hat{w}_m$*, such that*

$$\min_{j=1,\ldots,m} \|\hat{w}_j - w^*\|_2 \leq \mathcal{O}\left(\frac{S}{\kappa}\sqrt{\frac{\log\left(\frac{2}{\alpha}\right)}{\alpha\epsilon}}\right). \tag{3}$$

A key distinction between Theorem 2 and Theorem 1 is that the bound (3) is on the ℓ_2-distance between the candidate parameters and w^*, rather than on the expected loss as in (1). Such a distinction is due to our assumption that the functions f_i are κ-strongly convex, i.e.

$$f_i(w') - f_i(w) \geq \langle \nabla f_i(w), w' - w \rangle + \frac{\kappa}{2}\|w' - w\|_2^2$$

for all $w, w' \in \mathcal{H}$. This allows us to relate the ℓ_2-distance between parameters to the difference in their losses.

With the strong convexity assumption, the authors are able to make the bound (3) independent of r - the ℓ_2-radius of the parameter space $\mathcal{H}$ - by repeatedly applying the algorithm above and eventually shrinking the parameter space to a small ball around w^*. But now, (3) depends on the strong convexity parameter κ: the bound is stronger if κ is larger. Note that (3) is vacuous if f_i is only convex but not strongly convex.

3 Examples

Linear classification with hinge loss. In the paper [1], the authors show that the bound (1) of Theorem 1 is vacuous in an example of linear classification with logistic loss. Here, we illustrate that (1) is also vacuous in an example of linear classification with *hinge loss*. Assume $r = \mathcal{O}(1)$. Suppose x_i are drawn from a distribution supported on the unit ball $B(0;1)$. Let $y_i = \text{sign}(u^T x_i)$ for some unknown vector u and $f_i(w) = \max\{0, 1 - y_i w^T x_i\}$. Notice that

$$\nabla f_i(w) = \begin{cases} -y_i x_i & \text{if } y_i w^T x_i \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Observe that $\nabla \bar{f}(w)$ and all $\nabla f_i(w)$ are contained in $B(0;1)$, so $\left\|\nabla f_i(w) - \nabla \bar{f}(w)\right\|_2 \le 2$. As a result, we have

$$\left\| \left[\nabla f_i(w) - \nabla \bar{f}(w)\right]_{i \in I_{\text{good}}} \right\|_{\text{op}} \le \left\| \left[\nabla f_i(w) - \nabla \bar{f}(w)\right]_{i \in I_{\text{good}}} \right\|_{\text{F}} \le 2\sqrt{|I_{\text{good}}|},$$

where $\|\cdot\|_{\text{F}}$ is the Frobenius norm. This shows that $S \le 2$. The bound (1) then becomes

$$\min_{w \in \mathcal{E}_Y} \bar{f}(w) - \bar{f}(w^*) \le \mathcal{O}\big(\frac{1}{\sqrt{\alpha}}\big) \tag{4}$$

for some ellipse $\mathcal{E}_Y$. But this is vacuous since $\bar{f}(0) = 1$ and $\bar{f}(w^*) \ge 0$, so the LHS of (4) is already bounded by $1 \le \frac{1}{\sqrt{\alpha}}$.

Mean estimation. It is interesting to see how the bound (3) might improve (1) in the example of mean estimation. Consider $x_i \sim \mathcal{N}(\mu, \sigma^2 I)$ and let $f_i(w) = \frac{1}{2}\|x_i - w\|_2^2$. Note that f_i is 1-strongly convex, so Theorem 2 applies. In this case, we have $w^* = \mu$, and

$$\bar{f}(w) - \bar{f}(\mu) = \frac{1}{2}\|w - \mu\|_2^2.$$

Thus, (1) and (3) essentially bound the same object, and the bounds are $\mathcal{O}(\frac{Sr}{\sqrt{\alpha}})$ and $\mathcal{O}(\frac{S\sqrt{\log(2/\alpha)}}{\sqrt{\alpha}})$ respectively. (For the latter we set $\epsilon = \frac{1}{2}$.) We see that when $r \gg \sqrt{\log(2/\alpha)}$, the latter bound significantly improves the former. In addition, Theorem 2 offers a finite list of candidate parameters for us to check, while Theorem 1 only ensures that we can find a

good parameter in some ellipse.

Density estimation for product distributions. Finally, we look at an example of how Theorem 1 can be used to estimate product distributions in the semi-verified setting. Suppose x_i is drawn from a product distribution on $\{0,1\}^d$ with $P(x_{ij}=1)=p_j^*$. Define the negative log-likelihood loss

$$f_i(w) = -\bigg(\sum_{j=1}^{d} x_{ij}\log(w_j) + (1-x_{ij})\log(1-w_j)\bigg).$$

In this case, $w_j^* = p_j^*$, and the authors show that $S=\mathcal{O}(1)$ if p^* is balanced, i.e. $p_j^* \in [1/4, 3/4]$. Here, we extend their idea and further bound the KL divergence between the estimated and the true distributions. We show that if we are allowed to look at a single verified sample $x\sim p^*$, then we can output a density estimate $\hat{w}$ such that $\mathbb{E}[\mathrm{KL}(w^*||\hat{w})] \le \mathcal{O}(\frac{r}{\sqrt{\alpha}})$. The proof is similar to that of Corollary 9.4 of [1].

Proof. We first show that this product distribution can be written in the form of an exponential family. Consider

$$p_w(x_i) = \exp(-f_i(w)) = \exp\bigg(\sum_{j=1}^{d} x_{ij}\log\frac{w_j}{1-w_j} + \log(1-w_j)\bigg). \tag{5}$$

With the change of variable from w to θ defined by $\theta_j = \log\frac{w_j}{1-w_j}$, equation (5) becomes

$$p_\theta(x_i) = \exp\big(\theta^T x_i - A(\theta)\big),$$

where $A(\theta) = \sum_j \log(1+\exp(\theta_j))$. We can compute that $\nabla f_i(\theta) - \nabla \bar{f}(\theta) = x_i - p^*$. Therefore, Proposition B.1 of [1] implies that w.p. $1-\exp(-\Omega(\alpha n))$, there is a set of at least $\frac{\alpha}{2}n$ samples x_i with spectral norm bound $S=\mathcal{O}(1)$. Theorem 1 then implies that we can run the algorithm described in Section 2 to obtain an ellipse $\mathcal{E}_Y$ such that $\mathrm{tr}(Y)\le\mathcal{O}(\frac{r^2}{\alpha})$.

We now take the verified sample $x\sim p^*$ and define $\hat{\theta} = \arg\min_{\theta\in\mathcal{E}_Y} A(\theta) - \theta^T\phi(x)$. Then, following similar steps as in Corollary 9.4 (105) - (112) in [1], we can show that

$$\mathbb{E}[\bar{f}(\hat{\theta}) - \bar{f}(\theta^*)] \le \mathbb{E}[(\theta^*-\hat{\theta})^T(p^*-x))] \le \mathcal{O}(\frac{r}{\sqrt{\alpha}}).$$

Now let $\hat{w} = \frac{1}{1+\exp(-\hat{\theta})}$. We then have

$$\begin{aligned}
\bar{f}(\hat{\theta}) - \bar{f}(\theta^*) &= (\theta^*-\hat{\theta})^T p^* + A(\hat{\theta}) - A(\theta^*) \\
&= \sum_{j=1}^{d} w_j^* \log\frac{w_j^*(1-\hat{w}_j)}{\hat{w}_j(1-w_j^*)} + \log\frac{1-w_j^*}{1-\hat{w}_j} \\
&= \sum_{j=1}^{d} w_j^* \log\frac{w_j^*}{\hat{w}_j} + (1-w_j^*)\log\frac{1-w_j^*}{1-\hat{w}_j} \\
&= \mathrm{KL}(w^*||\hat{w}).
\end{aligned}$$

Therefore, we have $\mathbb{E}[\mathrm{KL}(w^*||\hat{w})] \leq \mathcal{O}(\frac{r}{\sqrt{\alpha}})$ as desired.

□

4 Limitations and Future directions

One limitation of the paper [1] is that the list-decodable result depends on the strongly convex assumption and becomes less useful when $\kappa \to 0$. This results in a gap between Theorem 2 and other results (Theorem 1, and Lemma 2.4 in [1] in the single-verified setting) where only convexity is assumed. Therefore we wonder if there exist weaker assumptions or better bounds in the list-decodable setting when κ is small.

All the main results are based on the particular quantity of the spectral norm bound S, which seems to come out of the usage of Hölder's Inequality $\mathrm{tr}\big(A^T B\big) \leq \|A\|_{\mathrm{op}} \|B\|_*$ and other technicalities in the proof of Theorem 1. In many examples, S is bounded by the covariance of the data, and in the end we need to require bounded covariance. This makes the results less applicable to certain (supervised) learning problems that do not typically require the covariance to be small. Can we derive similar results using other quantities, perhaps those less dependent on the covariance, so that they can apply to more general settings?

Finally, for the semi-verified model, the paper only focuses on the single-verified case yet the bound (e.g. Lemma 2.4) and applications (e.g. Corollary 9.4 and our example of product distributions above) are already interesting. We wonder how much improvement we can have (on the bound) if we have k verified points.

References

[1] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data, `arXiv:1611.02315`, 2016.