Vox Populi: Collecting High-Quality Labels from a Crowd

Ofer Dekel
Microsoft Research
One Microsoft Way, Redmond, WA 98052, USA
oferd@microsoft.com

Ohad Shamir
The Hebrew University
Jerusalem 91904, Israel
ohadsh@cs.huji.ac.il

Abstract

With the emergence of search engines and crowdsourcing websites, machine learning practitioners are faced with datasets that are labeled by a large heterogeneous set of teachers. These datasets test the limits of our existing learning theory, which largely assumes that data is sampled i.i.d. from a fixed distribution. In many cases, the number of teachers actually scales with the number of examples, with each teacher providing just a handful of labels, precluding any statistically reliable assessment of an individual teacher's quality. In this paper, we study the problem of pruning low-quality teachers in a crowd, in order to improve the label quality of our training set. Despite the hurdles mentioned above, we show that this is in fact achievable with a simple and efficient algorithm, which does not require that each example be repeatedly labeled by multiple teachers. We provide a theoretical analysis of our algorithm and back our findings with empirical evidence.

1 Introduction

In recent years, new distributed data-collection techniques have emerged, which harness the power of the Internet and its vast audience. While providing large amounts of cheap labeled data for our machine learning algorithms, these techniques violate some of the fundamental assumptions of our existing theoretical models of learning. Traditionally, we assume that training examples are sampled i.i.d. from a fixed probability distribution, while in these scenarios examples are labeled by a heterogeneous set of teachers. Different teachers have different levels of dedication, expertise, and competence, and provide labels of different quality.

For example, *crowdsourcing* websites allow members of the public to perform various simple labeling tasks, either voluntarily or for pay. *Galaxy Zoo* is a website where visitors help classify galaxies, and *Stardust@home* asks its visitors to detect interstellar dust particles in astronomical images. Amazon.com's *Mechanical Turk* is an online system on which individuals can publish crowdsourcing tasks and offer payments for their completion. By design, the crowdsourcing approach distributes the label collection task across numerous different individuals, while providing little means of quality control.

An even more extreme example of a multi-teacher setting occurs when data is harvested from search engine logs, with the intent of improving search engine results. A search engine log records the links that were clicked-on by each user; we think of each user as a teacher and of each click as a label. The number of distinct search engine users can easily run into the millions, and clearly the clicks of some users convey more useful information than the clicks of others. In this example, the number of teachers actually scales with the number of examples, and each individual teacher labels a tiny fraction of the data.

While most of the theoretical literature on machine learning focuses on the single-teacher setting, we are faced with the problem of *learning from a crowd*. A common first step is to identify and remove low-quality teachers that provide worthless labels. This data cleaning step is the focus of our paper. In dealing with this task, we are faced with a number of theoretical hurdles. First, we have no prior knowledge on the identity or the quality of any teacher. Additionally, it is unlikely that we have access to a large gold-set of perfectly labeled examples, which could be used to evaluate the quality of each teacher. Even if such a set were available, we assume that a typical teacher labels only a handful of examples, far from sufficient to perform a statistically reliable estimation of that teacher's quality. More often than not, we lack the ability to control which examples are assigned to which teachers. This prevents us from actively probing the quality of certain teachers or using convenient *repeated labeling* techniques, in which each example is labeled multiple times and inter-judge agreement is measured. Ideally, even when repeated labeling is possible, it should be avoided, since it increases the cost per example significantly. Ultimately, all we have to work with is the raw labeled data itself, with a single noisy label per example. Despite these difficulties, our goal is to detect and eliminate low-quality teachers in a principled and effective manner, resulting in improved label quality in our training set.

For simplicity, we focus here on binary classification and margin-based classifiers. We address our problem with a very simple algorithm, indirectly inspired by the most straightforward repeated labeling technique. Imagine for a moment that we could collect multiple labels for each example. If most of the teachers are reasonably good, we could eliminate much of the noise in the data by taking the average or the majority over the repeated labels of each example. Treating this aggregate label as an approximate ground-truth, we could count the number of incorrect labels provided by each teacher and identify those teachers that tend to make errors. Once the low-quality teachers are identified, we can make sure to ignore any labels they provide in the future. However, in our setting we do not have repeated labels and we cannot generate aggregate labels, so instead we simulate aggregate labels. Specifically, we train a hypothesis on the entire unfiltered dataset and regard the predictions of this hypothesis as the approximate ground-truth. Intuitively, fitting a hypothesis to the entire dataset is similar to aggregating multiple labels per example. With this hypothesis playing the role of our ground-truth, and pretending that we can reliably estimate a teacher's quality from the handful of labels it provides, we evaluate each teacher and prune away the low-quality ones.

We show that this simple technique effectively reduces the noise level in our data. Specifically, we show that this technique is effective even when the number of instances per teacher is nowhere near enough to reliably estimate an individual teacher's quality. Intuitively, our technique may mistake some good teachers for bad teachers and remove them, but more bad teachers are removed, and the overall noise level decreases. In this setting, a reasonable number of false negatives (good teachers that are removed) is easily tolerated, since the crowd is large and teachers are abundant.

Once the data is cleaned, the natural next step is to retrain a classifier on the cleaned dataset. The obvious difficulty here is that the examples in the cleaned dataset are no longer independent. Thus, it is not clear whether or not the standard generalization analysis of our learning algorithm still applies. Fortunately, a small modification to our basic cleaning procedure resolves this problem. Instead of applying our cleaning algorithm directly to the original dataset, we first split our data in two, and then use each half of the data to determine which teachers should be removed from the other half. Using this modification, any standard generalization analysis still holds, with a small penalty in the bounds.

Our paper is organized as follows. We conclude the introduction by mentioning important related work. In Sec. 2 we set the stage for our theoretical analysis, in Sec. 3 we state our main theoretical results, and formally prove them in Sec. 4. In Sec. 5 we empirically demonstrate the effectiveness of our technique on a real dataset obtained using Mechanical Turk. Finally, we conclude the paper in Sec. 6.

1.1 Related Work

To our knowledge, the literature on multi-teacher learning discusses two main approaches: using prior information and repeated labeling. For example, the work in [BCK$^+$07] and [CKW08] assumes that labeled examples are obtained from multiple heterogeneous sources, and that we have explicit prior knowledge on the relationships between these sources. Repeated labeling (see [SFB$^+$94, SPI08] and the references therein) is the practice of having each example labeled by multiple teachers, and then aggregating these labels in a way that cleans noise and identifies bad teachers. As discussed earlier, both approaches make problematic assumptions for the setting we have in mind. A somewhat related problem is designing machine learning algorithms that withstand specific types of label-noise, either on the training set [KL93, Kea98], or on the test set [TGRS07, DS08]. However, these approaches do not assume multiple teachers, and do not attempt to identify bad teachers. We also note the related work in [DFP08], which addresses the multi-teacher learning problem from a mechanism design perspective, and incentivises teachers to be good.

2 Setting and Notation

We focus on a binary classification setting. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be an instance space and let $\mathcal{D}$ be a probability distribution on $\mathcal{X} \times \{-1, +1\}$. Let $p_{\mathcal{D}}$ be the marginal probability of $\mathcal{D}$ on $\mathcal{X}$. We assume that $\mathcal{D}$ is the distribution on which we are tested after a classifier is learned.

We assume that the learning method we use belongs to the large family of algorithms that optimize a regularized empirical risk functional [SS02]. More specifically, given a dataset $S = \{\mathbf{x}_i, y_i\}_{i=1}^m$, we assume that the algorithm minimizes the convex function

$$\hat{F}_\lambda(\mathbf{w}, S) = \lambda \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^m \ell(f(\mathbf{w}, \mathbf{x}_i), y_i) \ ,$$

where $\mathbf{w}$ is a vector in a reproducing kernel Hilbert space H and $\lambda > 0$ is a regularization parameter. Additionally, $f(\mathbf{w}, \mathbf{x}_i) = \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle$ represents the application of the classifier $\mathbf{w}$ to the instance $\mathbf{x}_i$ after applying a feature mapping $\phi(\cdot) : \mathcal{X} \to H$, and ℓ is a loss function assumed to be convex and L-Lipschitz in its first argument. Also, we assume that the binary label predicted by the classifier $\mathbf{w}$ is the sign of $f(\mathbf{w}, \mathbf{x}_i)$, and that $\sup_{\mathbf{x} \in \mathcal{X}} \sqrt{\langle \phi(\mathbf{x}), \phi(\mathbf{x}) \rangle} \le R$ for some constant R. We restrict our discussion to classifiers that do not include a bias term, for the purpose of simplicity.

Unlike the typical supervised learning setting, where we assume that a training set S is sampled i.i.d. from $\mathcal{D}$, here we include a stage where the data is labeled by a set of k teachers. As discussed in the introduction, we are interested in a setting where both m and k are large and scale together. Namely, as $m, k \to \infty$, we assume that $m/k = \Theta(1)$.

Formally, the labeling process proceeds as follows: we assume that there exist k (possibly randomized) classifiers $\{h_1(\mathbf{x}), \ldots, h_k(\mathbf{x})\} : \mathcal{X} \to \{-1, +1\}$, which represent the way each teacher labels data. We also assume that we have a distribution $(q_1, \ldots, q_k)$ over $\{1, \ldots, k\}$. First, an *unlabeled* dataset is sampled i.i.d. according to $p_{\mathcal{D}}$. For each unlabeled instance $\mathbf{x}$, we choose a teacher $t \in \{1, \ldots, k\}$ at random, according to the distribution $(q_1, \ldots, q_k)$. This results in splitting the sample into k subsets, $S_1, \ldots, S_k$. Each instance in S_t is then labeled using h_t. As previously mentioned, $\mathcal{D}$ is the distribution on which we are ultimately *tested*, and may represent either the ground-truth or the subjective opinions of the tester. Notice that in the process, as described above, we assume that each teacher's labeling strategy is fixed before observing any data. For simplicity, we focus in this paper on the setting where $(q_1, \ldots, q_k)$ is the uniform distribution over $\{1, \ldots, k\}$, and note that our approach can be generalized.

Equivalently, we can view the process described above as sampling an unlabeled dataset and labeling it using $\bar{h}(\mathbf{x})$, where $\bar{h}(\mathbf{x})$ is the random classifier defined by randomly choosing a hypothesis from $h_1, \ldots, h_k$. Taking this view, the optimization problem corresponding to $\hat{F}_\lambda(\mathbf{w}, S)$ can be seen as the empirical counterpart of minimizing the convex function

$$F_\lambda(\mathbf{w}) = \lambda\|\mathbf{w}\|^2 + \mathbb{E}\left[\ell\left(f(\mathbf{w}, \mathbf{x}), \bar{h}(\mathbf{x})\right)\right] \ .$$

Let $\mathbf{w}^\star$ denote the minimizer of $F_\lambda(\mathbf{w})$. In general, optimizing $F_\lambda(\mathbf{w})$ is not the same as finding the optimal classifier with respect to $\mathcal{D}$, since $\bar{h}(\mathbf{x})$ and $\mathcal{D}$ conditioned on $\mathbf{x}$ are not necessarily the same.

Our goal is to identify and prune away low-quality teachers, based on a random sample. Namely, we seek to remove the teachers that are deemed harmful to the learning process. After pruning, we are left with a subset of hopefully high-quality teachers. This subset induces a new labeling distribution, namely, given an instance $\mathbf{x}$, we randomly pick a teacher from the remaining teachers, and label the instance according to the selected teacher. We denote this random classifier (whose form depends on which teachers were pruned) as $\bar{h}_T(\cdot)$. Now, let $e_t(\mathbf{x}) = \Pr_{y\sim\mathcal{D}}(yh_t(\mathbf{x}) < 0|\mathbf{x})$, and define

$$\begin{aligned} e_t &= \Pr\nolimits_{(\mathbf{x},y)\sim\mathcal{D}}(yh_t(\mathbf{x}) < 0) \ , \\ \bar{e} &= \Pr\nolimits_{(\mathbf{x},y)\sim\mathcal{D}}(y\bar{h}(\mathbf{x}) < 0) \ , \\ \bar{e}_T &= \Pr\nolimits_{(\mathbf{x},y)\sim\mathcal{D}}(y\bar{h}_T(\mathbf{x}) < 0 \,\big|\, S) \ . \end{aligned}$$

In words, $e_t(\mathbf{x})$ is the probability that teacher t incorrectly labels instance $\mathbf{x}$, where the correct label is drawn according to $\mathcal{D}$. $e_t = \mathbb{E}_\mathbf{x}[e_t(\mathbf{x})]$ is the average error-rate of teacher t over the entire instance space. Finally, $\bar{e}$ and $\bar{e}_T$ are the average error-rate of the entire crowd, before and after pruning. Note that $\bar{e}_T$ is a random variable, as it depends on the random sampling of S.

How can the quality of each teacher be measured? Ideally, since we are tested on the distribution $\mathcal{D}$, we want to keep teachers with a low error-rate e_t. Unfortunately, we do not know $\mathcal{D}$ nor h_t, and we cannot calculate e_t directly. Even worse, there is no overlap in the instances labeled by the different teachers, so we cannot use, say, a voting mechanism in order to estimate the correct label of any given instance. Therefore, a different approach is needed.

The key idea is to evaluate a somewhat different quantity, which is the error-rate of each teacher *with respect to* $\mathbf{w}^\star$, the optimal classifier for the labeling distribution induced by all the teachers together. Formally, we define

$$\begin{aligned} \epsilon_t &= \Pr(h_t(\mathbf{x})f(\mathbf{w}^\star, \mathbf{x}) < 0) \ , \\ \bar{\epsilon} &= \Pr(\bar{h}(\mathbf{x})f(\mathbf{w}^\star, \mathbf{x}) < 0) \ , \\ \bar{\epsilon}_T &= \Pr(\bar{h}_T(\mathbf{x})f(\mathbf{w}^\star, \mathbf{x}) < 0 \,\big|\, S) \ . \end{aligned}$$

The probabilities are taken with respect to $p_\mathcal{D}(\mathbf{x})$ and the randomization of $\bar{h}$ and $h_t(\mathbf{x})$. From the definitions, it is straightforward to verify that

$$\bar{\epsilon} = \frac{\sum_{t=1}^k \epsilon_t}{k}, \quad \bar{\epsilon}_T = \frac{\sum_{t=1}^k \mathbf{1}(t \text{ not pruned})\epsilon_t}{|\{t : \text{t not pruned}\}|}.$$

To gain some intuition, suppose that $\mathbf{w}^\star$ has an error-rate of zero with respect to $\mathcal{D}$. Then ϵ_t directly measures the test error-rate of teacher t, and $\bar{\epsilon}, \bar{\epsilon}_T$ represent the average test error-rate of the crowd of teachers before and after pruning. In this case, our goal should obviously be to make $\bar{\epsilon}_T$ small compared to $\bar{\epsilon}$, by pruning away teachers with relatively high ϵ_t. Although $\mathbf{w}^\star$ will practically never have a zero error-rate, we will see that, under mild conditions, reducing $\bar{\epsilon}_T$ still serves the purpose of reducing test error-rate.

3 Results

3.1 Relating $\bar{\epsilon}_T$ and Classification Error

As discussed earlier, the connection between $\bar{\epsilon}_T$ and the average error-rate of the teachers after pruning is trivial if $\mathbf{w}^\star$ is a Bayes optimal classifier. However, this is not the case in general, especially because in such a setting the given set of teachers is already optimal, so no pruning is really needed.

On the other hand, we clearly have no hope to learn if there is no correlation between the teachers and the test distribution, so it is realistic to assume at least some correlation. We begin our analysis by assessing how mild our assumptions can be while still allowing for $\bar{\epsilon}_T$ to serve as a reasonable proxy for $\bar{e}$. This justifies the rest of our analysis, which focuses on discarding teachers with high ϵ_t in order to reduce $\bar{\epsilon}_T$.

To simplify the presentation, assume that the conditional label distribution $p_\mathcal{D}(y|\mathbf{x})$ with respect to $\mathcal{D}$ is deterministic (namely, that the label y is a deterministic function of the instance $\mathbf{x}$). This assumption is not necessary, and the analysis can be easily extended to the general case.

Theorem 1 *Assuming $p_\mathcal{D}(y|\mathbf{x}) \in \{0, 1\}$, it holds for any teacher t that ϵ_t equals*

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}}(yf(\mathbf{w}^\star, \mathbf{x}) < 0) + \mathbb{E}_{(\mathbf{x},y)}\left[e_t(\mathbf{x})\mathrm{sign}(yf(\mathbf{w}^\star, \mathbf{x}))\right].$$

Proof: For any given $(\mathbf{x}, y)$ sampled from $\mathcal{D}$, and teacher hypothesis $h_t(\cdot)$, it holds that $h_t(\mathbf{x}) = y$ with probability $1 - e_t(\mathbf{x})$, and $h_t(\mathbf{x}) \neq y$ with probability $e_t(\mathbf{x})$. Therefore,

$$\begin{aligned} \epsilon_t &= \mathbb{E}\left[\mathbf{1}(h_t(\mathbf{x})f(\mathbf{w}^\star, \mathbf{x}) < 0)\right] \\ &= \mathbb{E}\left[(1 - e_t(\mathbf{x}))\mathbf{1}(yf(\mathbf{w}^\star, \mathbf{x}) < 0)\right] \\ &\quad + \mathbb{E}\left[e_t(\mathbf{x})\mathbf{1}(yf(\mathbf{w}^\star, \mathbf{x}) \geq 0)\right] \\ &= \Pr(yf(\mathbf{w}^\star, \mathbf{x}) < 0) + \mathbb{E}\left[e_t(\mathbf{x})\mathbf{1}(yf(\mathbf{w}^\star, \mathbf{x}) \geq 0)\right] \\ &\quad - \mathbb{E}\left[e_t(\mathbf{x})\mathbf{1}(yf(\mathbf{w}^\star, \mathbf{x}) < 0)\right]. \end{aligned}$$

The theorem follows after a slight simplification. ■

To get some feeling for what this theorem tells us, let us examine the case where for each teacher t, $e_t(\mathbf{x}) \equiv e_t$ is a constant independent of $\mathbf{x}$. This holds in the important case of a uniform noise model for each teacher: given an instance to label, teacher t flips a coin with bias $e_t \in [0, 1]$, and gives either the correct label or the incorrect label depending on the outcome of the flip. In such a scenario, we have the following straightforward corollary:

Corollary 2 *Assume that for any teacher t, $e_t(\mathbf{x}) \equiv e_t$ is a constant independent of* $\mathbf{x}$. *If* $\Pr(\mathit{sign}(f(\mathbf{w}^\star, \mathbf{x})) \neq y) < 1/2$, *then $\{\epsilon_t\}, \bar{\epsilon}, \bar{\epsilon}_T$ are equivalent to $\{e_t\}, \bar{e}, \bar{e}_T$ respectively, up to a uniform, monotonically increasing linear transformation.*

Proof: Applying Thm. 1, we have that ϵ_t equals

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}}(yf(\mathbf{w}^\star,\mathbf{x})<0)+e_t\big(1-2\Pr(yf(\mathbf{w}^\star,\mathbf{x})<0)\big).$$

If $\Pr(yf(\mathbf{w}^\star,\mathbf{x})<0)<1/2$, this gives us a uniform, monotonically increasing linear relation between ϵ_t and e_t which does not depend on t. By averaging over all teachers or over all remaining teachers after pruning, we get a similar relation between $\bar{\epsilon},\bar{e}$ and $\bar{\epsilon}_T,\bar{e}_T$. ■

Thus, we see that $\mathbf{w}^\star$, induced by learning with the aggregate of all teachers, does not have to be a particularly good classifier with respect to the test distribution $\mathcal{D}$, if we want $\bar{\epsilon},\bar{\epsilon}_T$ to be closely related to $\bar{e},\bar{e}_T$. In the setting above, an error-rate smaller than $1/2$ suffices. In more general cases of teacher labeling models, the relationship is less direct, but Thm. 1 still implies that unless $\mathbf{w}^\star$ is highly misleading, $\bar{\epsilon},\bar{\epsilon}_T$ and $\bar{e},\bar{e}_T$ are closely related. These results justify our approach, where we attempt to reduce $\bar{\epsilon}_T$ as a proxy to the uncomputable $\bar{e}_T$.

3.2 Pruning Can Help

Motivated by Thm. 1, we consider the following simple algorithm to prune teachers: Train a classifier $\mathbf{w}'$ on the entire dataset and prune away any teacher for which

$$\frac{\sum_{i\in S_t}\mathbf{1}(h_t(\mathbf{x}_i)f(\mathbf{w}',\mathbf{x}_i)<0)}{|S_t|}\;>\;T \qquad (1)$$

for some threshold $T\in(0,1)$, and assuming the denominator is positive. For a discussion of how T should be set, see Subsection 3.4

This procedure essentially calculates a rough empirical estimate of ϵ_t, and removes all teachers where this estimate exceeds the threshold T. Note however that this estimate of ϵ_t is not reliable or even unbiased: the fixed classifier $\mathbf{w}^\star$ is replaced by the empirically-learned classifier $\mathbf{w}'$, and the probability in the definition of ϵ_t is replaced by empirical averaging over a small sample.

We now focus on whether this pruning procedure can actually help at reducing $\bar{\epsilon}_T$ compared to $\bar{\epsilon}_t$. Combined with the insights of Subsection 3.1, this should lead to a reduction in the average error-rate of the remaining teachers, compared to the average error-rate of all teachers.

An interesting insight that can be gleaned from the following results is that the sample size m *need not* be much larger than the number of teachers k. In fact, we get a meaningful improvement even when $m/k\leq 1$ - namely, even when each teacher labels on average at most one instance! However, the amount of improvement is closely related to how $\epsilon_1,\ldots,\epsilon_k$ are distributed, and is hard to express in closed form. We also note that in the next section, a more general "safety guarantee" is provided, which implies that for *any* set of $\{\epsilon_1,\ldots,\epsilon_k\}$, our pruning procedure will not lead to $\bar{\epsilon}_T$ being significantly larger than $\bar{\epsilon}$.

Our main technical result is the following.

Theorem 3 *Assume we use the pruning procedure described above. Also, let $F:[0,1]\to[0,1]$ be a cumulative distribution function, such that $F(a)=\frac{1}{k}\sum_{t=1}^{k}\mathbf{1}(\epsilon_t\leq a)$. Let $P\sim F(\cdot)$, and let $N\sim$ Poi(m/k) be a Poisson random variable with parameter m/k. If we assume $m/k=\Theta(1)$ as m,k increase, it holds that*

$$\bar{\epsilon}=\mathbb{E}_P[P],$$

and with probability at least $1-\delta$ over the training sample,

$$\bar{\epsilon}_T\leq\frac{\mathbb{E}_{P,N}[\Pr(X_N^P\leq NT)P]+r(m,\delta)}{\mathbb{E}_{P,N}[\Pr(X_N^P\leq NT)]-r(m,\delta)}, \qquad (2)$$

where X_N^P is a binomial random variable, representing a sum of N independent Bernoulli random variables with parameter P, and

$$r(m,\delta)=O\left(\sqrt{\frac{\log(6/\delta)}{m}}\right).$$

The O-notation hides dependencies on L,R,λ and a certain smoothness assumption[1] *on the underlying distribution close enough to the decision surface of the classifier $\mathbf{w}^\star$, which can be either assumed a-priori or replaced by a quantity that can be reliably estimated from the sample.*

Note that for any finite k, $F(\cdot)$ is a step function. However, as $k\to\infty$ (scaling with $m\to\infty$), we would like to approximate the histogram of $\{\epsilon_t\}$ by a continuous distribution, allowing us to model various intuitive settings, such as a setting where the histogram of the values of $\{\epsilon_1,\ldots,\epsilon_k\}$ has an approximately Gaussian mixture distribution. It is not hard to show that if we consider a sequence of distribution functions $(F_k(\cdot))_{k=1}^{\infty}$, converging to the distribution of a continuous density function $p(\cdot)$, in the sense that

$$\sup_a\left|F_k(a)-\int_0^a p(x)dx\right|\;=\;O\left(\frac{1}{k}\right),$$

then we can replace expectation with respect to $F(\cdot)$ in Eq. (3) with expectation with respect to $p(\cdot)$, up to $O(1/k)$ terms.

The main drawback of Thm. 3 is that it cannot really be expressed in closed form, since it intimately depends on $F(\cdot)$. However, it is easily calculated numerically. By trying out different distributions, we can gain confidence that indeed our pruning procedure makes $\bar{\epsilon}_T$ considerably smaller than $\bar{\epsilon}$ in various scenarios.

In Fig. 1, numerical results are displayed for various distributions $p(\cdot)$. In all cases, we get that the pruning procedure effectively makes $\bar{\epsilon}_f$ significantly smaller than $\bar{\epsilon}$, even when m/k is very small. These examples may indicate that the improvement, even for small m/k, is quite robust.

One important special case, corresponding to the first row in Fig. 1, when we are able to get somewhat more explicit analytical results from Thm. 3, is when $F(\cdot)$ is a uniform distribution. Intuitively, this corresponds to a setting where we have a uniform spectrum of teachers, ranging from very bad to very good. In this case, we have the following result:

[1] Specifically, an upper bound on the Lipschitz parameter of the distribution function of $f(\mathbf{w}^\star,\mathbf{x})$ close enough to 0 - see the proof for more details. If we use a linear kernel, this condition follows from boundedness of $p_{\mathcal{D}}(\mathbf{x})$.

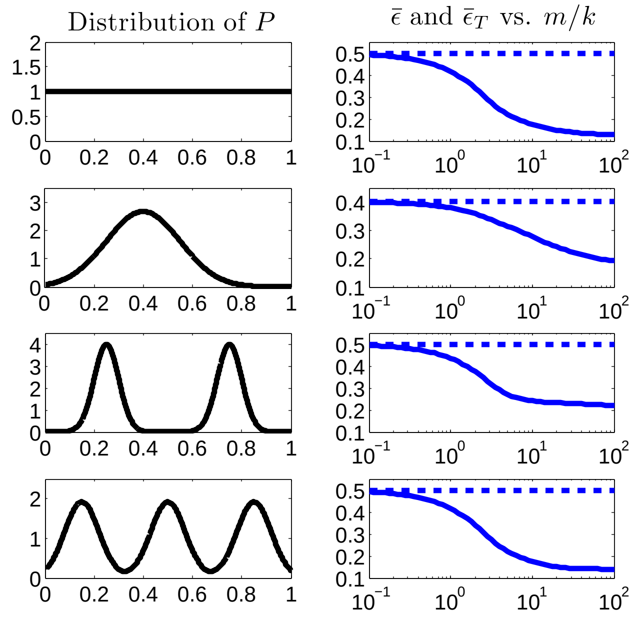

Figure 1: In each row, the left plot represents the distribution density function of P as described in Thm. 3. In the right plot, the solid line represents $\bar{\epsilon}_T$ as a function of the ratio m/k, for that particular distribution of P, and the dashed line represents $\bar{\epsilon}$ for reference. For any value of m/k, the difference between the dashed and the solid line represents the improvement obtained by pruning the teachers. In all cases, the threshold is set to $T = 0.25$.

Theorem 4 *Assume that in the setting of Thm. 3, $F_k(\cdot)$ converges (in the sense described above) to the uniform distribution on $[0,1]$, i.e. $\sup_{a\in[0,1]} |F_k(a) - a| \leq O(1/k)$. Ignoring $r(m,\delta)$ and other $O(1/m), O(1/k)$ terms, it holds that $\bar{\epsilon} = 1/2$, whereas $\bar{\epsilon}_T$ is upper bounded (with arbitrarily high probability as $m, k \to \infty$) by*

$$\mathbb{E}_V\left[\frac{1 + \lfloor VT \rfloor/2}{2+V}\right],$$

where $V \in \{0, 1, \ldots\}$ is distributed according to

$$\Pr(V = v) = \frac{1}{Z}\left(\frac{m}{k}\right)^v \frac{\lfloor vT \rfloor + 1}{(v+1)!},$$

and Z is a normalization term.

The upper bound on $\bar{\epsilon}_T$ is the expected value over V of $(1 + \lfloor VT \rfloor/2)/(2+V)$. For $V = 0$, this is equal to $1/2$. As V increases, it quickly becomes smaller, converging to $T/2$ as $V \to \infty$. The distribution of V depends on m/k. As it increases, the distribution puts more weight on larger values of V. Overall, we get that up to negligible factors, $\bar{\epsilon}_T$ is less than $1/2$, and gets arbitrarily close to $T/2$ as m/k increases. Comparing this to $\bar{\epsilon} = 1/2$, we see that we indeed get a considerable reduction in the average ϵ_t by pruning. Again, notice that m/k does not need to be large at all in order to obtain good results - even for $m/k \approx 1$ we get a noticeable improvement. For a graphical illustration in the case $T = 0.25$, see the first row of Fig. 1.

3.3 Pruning Can't Hurt

In the previous subsection, we discussed several cases where pruning indeed makes $\bar{\epsilon}_T$ considerably smaller than $\bar{\epsilon}$. The obvious question is whether something can be guaranteed in general, without assuming a specific distribution on $\{\epsilon_t\}$? It is not hard to see that there are cases where $\bar{\epsilon}_T$ cannot be considerably smaller than $\bar{\epsilon}$. A simple example is when ϵ_t is the same for all t. However, can we guarantee that $\bar{\epsilon}_T$ is never considerably *larger* than $\bar{\epsilon}$?

The following theorem answers this question in the affirmative.

Theorem 5 *In the setting of Thm. 3, it holds for any $\{\epsilon_t\}$ that*

$$\bar{\epsilon}_T \leq \bar{\epsilon} + \frac{2r(m,\delta)}{\mathbb{E}_{P,N}[\Pr(X_N^P \leq NT)] - r(m,\delta)}.$$

Recalling the motivating examples discussed in the introduction, we can safely assume that the sample size m is large enough for the excess term in the bound to be rather insignificant. Thus, regardless of how $\{\epsilon_t\}$ are distributed, $\bar{\epsilon}_T$ will not be significantly larger than $\bar{\epsilon}$.

3.4 Setting the Threshold T

The corollary of Thm. 1 below answers the following question: for arbitrary teacher noise models, how can we reliably detect whether $e_t > \bar{e}$ (namely, a teacher's error-rate is greater than average) using ϵ_t?

Corollary 6 *In the setting of Thm. 1, a sufficient condition for $e_t > \bar{e}$ is*

$$\epsilon_t > \Pr_{(\mathbf{x},y)\sim\mathcal{D}}(yf(\mathbf{w}^\star, \mathbf{x}) < 0) + \bar{e}. \qquad (3)$$

Proof: Assume that $e_t \leq \bar{e}$. Applying Thm. 1, we have that

$$\epsilon_t \leq \Pr_{(\mathbf{x},y)\sim\mathcal{D}}(yf(\mathbf{w}^\star, \mathbf{x}) < 0) + \mathbb{E}_{\mathbf{x}}[e_t(\mathbf{x})].$$

Our assumption that $\mathbb{E}_{\mathbf{x}}[e_t(\mathbf{x})] = e_t \leq \bar{e}$ implies that

$$\epsilon_t \leq \Pr_{(\mathbf{x},y)\sim\mathcal{D}}(yf(\mathbf{w}^\star, \mathbf{x}) < 0) + \bar{e}\ \ ,$$

and we have reached a contradiction to Eq. (3) ■

This simple corollary implies that if for some teacher t, ϵ_t is larger than a certain quantity, then it is definitely worse-than-average in terms of its error-rate.

This suggests a reasonable conservative criterion for setting the threshold T in our pruning procedure: we should set it to

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}}(yf(\mathbf{w}^\star, \mathbf{x}) < 0) + \bar{e}.$$

Although $\mathbf{w}^\star$ is not known, the probability above is well approximated by $\Pr_{(\mathbf{x},y)\sim\mathcal{D}}(yf(\mathbf{w}', \mathbf{x}) < 0)$, where $\mathbf{w}'$ is the actual learned classifier, if m is large enough. The proof is similar to the proof of Lemma 10 below, and we skip the details due to lack of space. As to $\bar{e}$, we cannot calculate it based on the sample, but it can be easily estimated with a small held-out set of examples drawn from the true underlying distribution $\mathcal{D}$. This is not too harsh a requirement, since we use this set just for a single validation, and it need not scale with the complexity of the problem.

3.5 Reusing the Cleaned Dataset

After removing low-quality teachers, a natural next step is to retrain a classifier on the cleaned data, using any learning algorithm we like. If the pruning stage was successful, we expect the remaining data to be cleaner than the original data, and the resulting classifier should be more accurate. However, the pruning was done in a data-dependent manner, and it is not clear that we can simply reuse the training set and still get a generalization guarantee, as if we used a fresh sample. Fortunately, we can show that by adding a small twist to our basic algorithm, the dataset may be safely reused for learning, and any standard generalization analysis still holds.

The twist in our algorithm is as follows: first, we randomly split the sample S into two subsets S_1, S_2. Next, we use our technique to detect low-quality teachers in each subset independently, but stop short of actually removing any examples from either set. The result is two sets of low-quality teachers: $B_1 \subseteq \{1, \ldots, k\}$ is the set of low-quality teachers according to S_1 and $B_2 \subseteq \{1, \ldots, k\}$ is the set of low-quality teachers according to S_2. Next, we define S'_1 as the subset of S_1 obtained by removing examples that are labeled by the teachers in B_2, and we define S'_2 as the subset of S_2 obtained by removing examples from teachers in B_1. Finally, we set $S' = S'_1 \cup S'_2$ and we train a classifier (using any learning algorithm we like) on S'.

To state our guarantee formally, we need some additional notation. Let $\bar{h}_{T_1}(\cdot)$ to be the classifier that chooses a random teacher t where $t \notin B_1$ and then returns the label according to h_t. In other words, $\bar{h}_{T_1}(\cdot)$ is the classifier induced by the high-quality teachers according to S_1. Similarly, let $\bar{h}_{T_2}(\cdot)$ be the classifier induced by the high-quality teachers according to S_2. Finally, let $\bar{h}_{T_1 \vee T_2}(\cdot)$ be the classifier that, given any instance $\mathbf{x}$, randomly chooses and returns either $\bar{h}_{T_1}(\mathbf{x})$ or $\bar{h}_{T_2}(\mathbf{x})$ with probability $1/2$. Note that if our cleaning procedure is successful, we expect both $\bar{h}_{T_1}(\cdot)$ and $\bar{h}_{T_2}(\cdot)$, hence $\bar{h}_{T_1 \vee T_2}(\cdot)$ as well, to be relatively good classifiers, in the sense that their labeling policy is closer to the unknown test distribution, compared to the set of all teachers before pruning.

Theorem 7 *Assume that our learning algorithm returns a classifier from the space $\mathcal{H}$, and has a uniform generalization analysis, in the following sense: For any distribution $\mathcal{D}$ and any m, with probability at least $1-\delta$ over the sampling of S from $\mathcal{D}^m$, it holds that*

$$\forall h \in \mathcal{H} \quad \Pr_{(\mathbf{x},y)\sim\mathcal{D}}(h(\mathbf{x}) \neq y) \leq \hat{R}(h, S) + b(|S|, \delta),$$

where $\hat{R}(h, S)$ is the the empirical risk on the dataset S and $b(\cdot,\cdot)$ is the generalization bound, a function of $|S|$ and δ. Then, with probability at least $1-\delta$ over the sampling of S from $\mathcal{D}^m$, it holds for all $h \in \mathcal{H}$ that

$$\Pr_{\mathbf{x}}\left(h(\mathbf{x}) \neq \bar{h}_{T_1 \vee T_2}(\mathbf{x})\right) \leq \frac{\hat{R}(h, S'_1) + \hat{R}(h, S'_2)}{2} + \frac{b(|S'_1|, 2(|S_1|+1)\delta) + b(|S'_2|, 2(|S_2|+1)\delta)}{2}.$$

This theorem implies that we can use S', the cleaned dataset, to learn a classifier h with a generalization guarantee that bounds its expected disagreement rate with $\bar{h}_{T_1 \vee T_2}(\cdot)$.

Proof: By the definition of $\bar{h}_{T_1 \vee T_2}$, we have that $\Pr_{\mathbf{x}}(h(\mathbf{x}) \neq \bar{h}_{T_1 \vee T_2}(\mathbf{x}))$ equals

$$\frac{\Pr_{\mathbf{x}}(h(\mathbf{x}) \neq \bar{h}_{T_1}(\mathbf{x})) + \Pr_{\mathbf{x}}(h(\mathbf{x}) \neq \bar{h}_{T_2}(\mathbf{x}))}{2}.$$

By construction, the set B_2 is independent of S_1. Therefore, S'_1 can be seen as an i.i.d. sample from the distribution induced by $\bar{h}_{T_2}$. The only technical subtlety is that $|S'_1|$ is a random quantity (based on the randomness of S_2 and the resulting pruning procedure). However, we can take a union bound over the $|S_1|+1$ possible values of $|S'_1|$, and use the uniform generalization assumption to get that with probability at least $1-\delta/2$,

$$\Pr_{\mathbf{x}}(h(\mathbf{x}) \neq \bar{h}_{T_2}(\mathbf{x})) \leq \hat{R}(h, S'_1) + b(|S'_1|, 2(|S_1|+1)\delta)$$

Applying a similar analysis on S'_2 and combining the result with a union bound, the theorem follows. ■

4 Proofs of Results from Sec. 3

We begin by proving Thm. 3, on which the other results are based. First, we need a sequence of auxiliary lemmas.

Lemma 8 *Let S be a sample of size m drawn i.i.d. from some distribution on $\mathcal{X} \times \{-1, +1\}$. Let $\mathbf{w}' = \arg\min_{\mathbf{w}} \hat{F}_\lambda(\mathbf{w}, S)$, and $\mathbf{w}^\star = \arg\min_{\mathbf{w}} F_\lambda(\mathbf{w})$. Then with probability of at least $1-\delta$, it holds for any $\mathbf{x} \in \mathcal{X}$ that*

$$|f(\mathbf{w}', \mathbf{x}) - f(\mathbf{w}^\star, \mathbf{x})| \leq g(m, \delta),$$

where

$$g(m, \delta) = \frac{2R^2 L \log(2/\delta)\left(1 + \sqrt{2m/\log(2/\delta)}\right)}{m\lambda}.$$

Proof: The proof is based on showing that $\|\mathbf{w}' - \mathbf{w}^\star\|$ converges to zero with the sample size. This is provided by a deep result in [Ste03, Prop. 33], which implies that

$$\Pr(\|\mathbf{w}' - \mathbf{w}^\star\| \geq \epsilon) \leq 2\exp\left(-\frac{m\lambda^2\epsilon^2}{8R^2L^2 + 2\lambda RL\epsilon}\right).$$

Fixing a confidence parameter δ, requiring it to equal the right-hand side, and solving for ϵ, we get

$$\epsilon = \frac{RL\log(2/\delta)}{m\lambda}\left(1 + \sqrt{1 + \frac{8m}{\log(2/\delta)}}\right) \leq \frac{RL\log(2/\delta)}{m\lambda}\left(2 + \sqrt{\frac{8m}{\log(2/\delta)}}\right),$$

from which it follows that with probability at least $1-\delta$,

$$\|\mathbf{w}' - \mathbf{w}^\star\| \leq \frac{2RL\log(2/\delta)\left(1 + \sqrt{2m/\log(2/\delta)}\right)}{m\lambda}. \quad (4)$$

If this indeed occurs, then for any $\mathbf{x} \in \mathcal{X}$ it holds that

$$|\langle\phi(\mathbf{x}), \mathbf{w}'\rangle - \langle\phi(\mathbf{x}), \mathbf{w}^\star\rangle| \leq \|\phi(\mathbf{x})\|\|\mathbf{w}' - \mathbf{w}^\star\|$$

by the Cauchy-Schwartz inequality. From Eq. (4), the fact that $\|\phi(\mathbf{x})\| \leq R$, and definition of $f(\cdot,\cdot)$, the lemma follows. ■

Lemma 9 *Let X_n^p be a binomial random variable with parameters (n,p). Then for any $T > 0$ and n, the function $g(p) = \Pr(X_n^p \leq T)$ is monotonically decreasing in p, and is Lipschitz continuous with a Lipschitz parameter upper bounded by n.*

Proof: If $T < 1$, the probability reduces to $(1-p)^n$, which decreases monotonically with p and with Lipschitz parameter upper bounded by n. The same thing happens trivially if $T \geq n$. If $n > T \geq 1$, we have by a technical but not difficult calculation that the derivative of $\Pr(X_n^p \leq T)$ is

$$\frac{\partial}{\partial p}\left(\sum_{i=0}^{\lfloor T\rfloor}\binom{n}{i}p^i(1-p)^{n-i}\right)$$
$$= -n\binom{n-1}{\lfloor T\rfloor}p^{\lfloor T\rfloor}(1-p)^{n-\lfloor T\rfloor-1}.$$

For any $p \in [0,1]$, the expression above is non-positive, and therefore $\Pr(X_n^p \leq T)$ decreases monotonically with p as required. Moreover, the absolute value of this expression constitutes an upper bound on the Lipschitz parameter of the function:

$$n\binom{n-1}{\lfloor T\rfloor}p^{\lfloor T\rfloor}(1-p)^{n-\lfloor T\rfloor-1} \leq n(p+(1-p))^{n-1} = n.$$

■

To continue, we require some additional notation, which allows us to relate the pruning procedure based on the learned classifier, and a virtual pruning procedure based on $\mathbf{w}^\star$. Given an i.i.d. sample $\{\mathbf{x}_i\}_{i=1}^m$, we turn it into a labeled dataset S using the teachers as described earlier in the paper, and learn a classifier $\mathbf{w}'$ by minimizing $\hat{F}_\lambda(\mathbf{w}, S)$. Then we define:

$$\hat{\epsilon}'_t = \frac{1}{|S_t|}\sum_{i\in S_t}\mathbf{1}\left(h_t(\mathbf{x})f(\mathbf{w}',\mathbf{x}_i) < 0\right)\ ,$$

$$\hat{\epsilon}_t = \frac{1}{|S_t|}\sum_{i\in S_t}\mathbf{1}\left(h_t(\mathbf{x})f(\mathbf{w}^\star,\mathbf{x}_i) < 0\right)\ ,$$

$$\hat{\epsilon}_{t,c} = \frac{1}{|S_t|}\sum_{i\in S_t}\mathbf{1}\left(h_t(\mathbf{x}_i)f(\mathbf{w}^\star,\mathbf{x}_i) < c\right)\ .$$

where $c \in \mathbb{R}$ is some parameter. We also define these quantities to be 0 in case $S_t = \emptyset$.

Lemma 10 *For any $\delta \in (0,1]$, it holds with probability at least $1-\delta$ that uniformly for all t, $\mathbf{1}(t \text{ not pruned})$ is upper bounded by $\mathbf{1}(\hat{\epsilon}_{t,-g(m,\delta)} \leq T)$, and lower bounded by $\mathbf{1}(\hat{\epsilon}_{t,g(m,\delta)} \leq T)$, where $g(m,\delta)$ is defined in Lemma 8.*

Moreover, if $m/k = \Theta(1)$, and $f(\mathbf{w}^\star, \mathbf{x})$ has a Lipschitz continuous distribution (based on the randomness of $\mathbf{x}$ according to the underlying distribution $p_{\mathcal{D}}(\mathbf{x})$) in a sufficiently small neighborhood around 0, then both $|\Pr(\hat{\epsilon}_{t,g(m,\delta)}) - \Pr(\hat{\epsilon}_t)|$ and $|\Pr(\hat{\epsilon}_{t,-g(m,\delta)}) - \Pr(\hat{\epsilon}_t)|$ are at most a constant times $g(m,\delta)$, for any m.

Proof: We have that $\mathbf{1}(t \text{ not pruned}) = \mathbf{1}(\hat{\epsilon}'_t \leq T)$. By Lemma 8, $|f(\mathbf{w}',\mathbf{x}_i) - f(\mathbf{w}^\star,\mathbf{x}_i)| \leq g(m,\delta)$ for all $\mathbf{x}_i$ holds with probability at least $1-\delta$. If this indeed happens, then whenever $h_t(\mathbf{x}_i)f(\mathbf{w}^\star,\mathbf{x}_i) < -g(m,\delta)$ occurs, $h_t(\mathbf{x}_i)f(\mathbf{w}',\mathbf{x}_i) < 0$ occurs as well. Recalling the definition of $\hat{\epsilon}'_t$ and $\hat{\epsilon}_{t,c}$ above, this implies that $\mathbf{1}(\hat{\epsilon}'_t \leq T)$ is at most $\mathbf{1}(\hat{\epsilon}_{t,-g(m,\delta)} \leq T)$. Repeating the same argument to get a lower bound on $\mathbf{1}(\hat{\epsilon}'_t \leq T)$, the first half of the lemma follows.

To prove the second part of the lemma, it suffices to show that $\Pr(\hat{\epsilon}_{t,c} \leq T)$ is Lipschitz continuous as a function of c. We begin by noticing that if $f(\mathbf{w}^\star,\mathbf{x})$ has a Lipschitz continuous distribution, then $h_t(\mathbf{x})f(\mathbf{w}^\star,\mathbf{x})$ also has a Lipschitz continuous distribution (with some trivial measurability requirements on $h_t(\cdot)$). As a result, $\mathbf{1}(h_t(\mathbf{x})f(\mathbf{w}^\star,\mathbf{x}) < c)$ is a Bernoulli random variable (based on the randomness of $\mathbf{x}$), whose parameter is Lipschitz continuous in c. Now, we examine

$$\Pr(\hat{\epsilon}_{t,c} \leq T) = \sum_{n=0}^{\infty}\Pr(|S_t| = n)\Pr\left(\hat{\epsilon}_{t,c} \leq T\Big||S_t| = n\right),$$

and consider how fast can the right-hand side change as we change c. By definition of $\hat{\epsilon}_{t,c}$, for any fixed n in the sum, the conditional probability is simply the probability that a Binomially distributed random variable is smaller than some threshold, with a parameter that has a Lipschitz relation to c (say with Lipschitz parameter l). Invoking Lemma 9, we have that the Lipschitz parameter of this conditional probability with respect to c is at most nl. As a result, we have that the Lipschitz parameter of the entire sum with respect to c is

$$\sum_{n=0}^{\infty}\Pr(|S_t| = n)ln \ = \ l\mathbb{E}[|S_t|] \ = \ l\frac{m}{k}.$$

Since we assume that $m/k = \Theta(1)$, we get that $\Pr(\hat{\epsilon}_{t,c} \leq T)$ is indeed Lipschitz continuous as a function of c. ■

Proof of Thm. 3: The assertion that $\bar{\epsilon} = \mathbb{E}_P[P]$ is immediate from the definitions, so all the work lies in proving Eq. (2). By definition, we have that

$$\bar{\epsilon}_T = \frac{\frac{1}{k}\sum_{t=1}^k \mathbf{1}(t \text{ not pruned})\bar{\epsilon}_t}{\frac{1}{k}\sum_{t=1}^k \mathbf{1}(t \text{ not pruned})}. \tag{5}$$

where we take this fraction to be 1 if all teachers are pruned.

Applying the first half of Lemma 10, we can upper bound this expression with probability at least $1-\delta$ by

$$\frac{\frac{1}{k}\sum_{t=1}^k \mathbf{1}(\hat{\epsilon}_{t,-g(m,\delta)} \leq T)\epsilon_t}{\frac{1}{k}\sum_{t=1}^k \mathbf{1}(\hat{\epsilon}_{t,g(m,\delta)} \leq T)}. \tag{6}$$

We now apply McDiarmid's inequality [McD89], to show that both the numerator and the denominator are close to their expectation with high probability. Let us focus for instance on the numerator. Notice that the numerator is a function of $2m$ independent random variables: the m unlabeled instances in the sample (and the label given to them by each teacher), and the assignment of each instance to each teacher. For any assignment of these variables, changing any one of them can change only two of the indicator functions in the numerator, which leads to a change of at most $2/k$ in the value of the numerator (recalling that $\epsilon_t \in [0,1]$). Thus, the numerator has a bounded differences property with parameter $2/k$. A similar reasoning shows that the denominator also has a bounded differences property with the same parameter.

Applying McDiarmid's inequality separately on each, and using a union bound, we get that with probability at least $1 - 2\delta$, Eq. (6) is upper bounded by

$$\frac{\frac{1}{k}\sum_{t=1}^{k}\Pr(\hat{\epsilon}_{t,-g(m,\delta)} \leq T)\epsilon_t + 2\sqrt{\log(1/\delta)m/k^2}}{\frac{1}{k}\sum_{t=1}^{k}\Pr(\hat{\epsilon}_{t,g(m,\delta)} \leq T) - 2\sqrt{\log(1/\delta)m/k^2}}. \quad (7)$$

Now, applying the second half of Lemma 10, we can upper bound this by

$$\frac{\frac{1}{k}\sum_{t=1}^{k}\Pr(\hat{\epsilon}_t \leq T)\epsilon_t + r(m,\delta)}{\frac{1}{k}\sum_{t=1}^{k}\Pr(\hat{\epsilon}_t \leq T) - r(m,\delta)}, \quad (8)$$

where $r(m,\delta)$ is the same as in the theorem statement. Now, we reduce $\Pr(\hat{\epsilon}_t \leq T)$ to an expression involving standard probability distributions. We start by noticing that

$$\Pr(\hat{\epsilon}_t \leq T) = \\ \sum_{n=0}^{\infty}\Pr(|S_t| = n)\Pr\left(\hat{\epsilon}_t \leq T \middle| |S_t| = n\right).$$

By definition of $\hat{\epsilon}_t$, the conditional probability in each sum is simply $\Pr(X_n^{\epsilon_t})$, the probability that a binomial random variable with parameters (ϵ_t, n) is at most T. Therefore, we can reduce the expression above to

$$\sum_{n=0}^{\infty}\Pr(|S_t| = n)\Pr(X_n^{\epsilon_t} \leq T).$$

The distribution of $\Pr(|S_t| = n)$ is the distribution of the number of instances assigned to teacher i. Its expected value is m/k. As m increases and with k scaling accordingly, we have that this distribution converges to a Poisson distribution, $N \sim Poi(m/k)$. An explicit bound on the rate of convergence is given by Le Cam's inequality (e.g. proposition 2.8 in [RP07]), which implies that $\sum_n |\Pr(|S_t| = n|) - \Pr(N = n)| \leq m/k^2$. Therefore, we have that

$$\left|\sum_{n=0}^{\infty}|\Pr(|S_t| = n)\Pr(X_n^{\epsilon_t} \leq T) \right. \\ \left. - \sum_{n=0}^{\infty}|\Pr(N = n)\Pr(X_n^{\epsilon_t} \leq T)\right| \leq \frac{m}{k^2}.$$

Applying this on Eq. (8), absorbing the $m/k^2 = O(1/m)$ term into $r(m,\delta)$, and substituting $\delta/3$ for δ, to have the result hold with overall probability $1-\delta$, we get Eq. (2).

In the proof of the theorem (to be exact, as part of proving the second half of Lemma 10), we made a smoothness assumption which essentially allowed us to upper bound

$$\frac{1}{k}\sum_{t=1}^{k}\left(\Pr(\hat{\epsilon}_{t,-g(m,\delta)} \leq T) - \Pr(\hat{\epsilon}_t \leq T)\right) \quad (9)$$

and

$$\frac{1}{k}\sum_{t=1}^{k}\left(\Pr(\hat{\epsilon}_t \leq T) - \Pr(\hat{\epsilon}_{t,g(m,\delta)} \leq T)\right).$$

We now show that these quantities, if we wish, can actually be estimated empirically, without making any a-priori assumptions. We focus on Eq. (9), as the reasoning for the other expression is similar.

The semi-empirical counterpart of Eq. (9), based on the sample but using the non-empirical $\mathbf{w}^\star$, is

$$\frac{1}{k}\sum_{t=1}^{k}\left(\mathbf{1}\left(\frac{1}{|S_t|}\sum_{i\in S_t}\mathbf{1}(h_t(\mathbf{x})f(\mathbf{w}^\star,\mathbf{x}_i) < -g(m,\delta)) \leq T\right)\right. \\ \left. -\mathbf{1}\left(\frac{1}{|S_t|}\sum_{i\in S_t}\mathbf{1}(h_t(\mathbf{x})f(\mathbf{w}^\star,\mathbf{x}_i) < 0) \leq T\right)\right). \quad (10)$$

By McDiarmid's inequality, this is larger than its expectation (Eq. (9)) by at most $2\sqrt{\log(1/\delta)m/k^2}$ with probability at least $1-\delta$ over the training set (the reasoning is similar to our use of McDiarmid's inequality earlier). Since we already assume that the inequality in Lemma 8 holds, we have that Eq. (10) is at most

$$\frac{1}{k}\sum_{t=1}^{k}\left(\mathbf{1}\left(\frac{1}{|S_t|}\sum_{i\in S_t}\mathbf{1}(h_t(\mathbf{x})f(\mathbf{w}',\mathbf{x}_i) < -2g(m,\delta)) \leq T\right)\right. \\ \left. -\mathbf{1}\left(\frac{1}{|S_t|}\sum_{i\in S_t}\mathbf{1}(h_t(\mathbf{x})f(\mathbf{w}',\mathbf{x}_i) < g(m,\delta)) \leq T\right)\right),$$

where we have replaced $\mathbf{w}^\star$ by the actual learned classifier $\mathbf{w}'$. This expression can be evaluated based on the training data, and we therefore get a high-probability empirical estimate of Eq. (9). ■

Proof of Thm. 4: The assertion that $\bar{\epsilon} = 1/2$ is immediate from Thm. 3 and the way we define $F(\cdot)$. Also from Thm. 3, we have that

$$\bar{\epsilon}_T \leq \frac{\mathbb{E}_N\left[\int_{p=0}^{1}\Pr(X_N^p \leq NT)p\,dp\right] + r(m,\delta)}{\mathbb{E}_N\left[\int_{p=0}^{1}\Pr(X_N^p \leq NT)\,dp\right] - r(m,\delta)}. \quad (11)$$

Let us start by analyzing the numerator in Eq. (11). By definition of a Binomial distribution, we have that

$$\int_{p=0}^{1}\Pr(X_N^p \leq NT)p\,dp \\ = \sum_{i=0}^{\lfloor NT\rfloor}\binom{N}{i}\int_{p=0}^{1}p^{i+1}(1-p)^{N-i}\,dp.$$

It turns out that the integral in the expression above is the so-called Euler Beta function [AS72], with parameters $i+2, N-i+1$. Using well known results on this function, we have that the expression above is equal to

$$\sum_{i=0}^{\lfloor NT\rfloor}\binom{N}{i}\frac{\Gamma(i+2)\Gamma(N-i+1)}{\Gamma(N+3)},$$

where $\Gamma(\cdot)$ is the Gamma function. Since $\Gamma(n) = (n-1)!$ for any positive integer n, the expression above is equal in turn to

$$\sum_{i=0}^{\lfloor NT\rfloor}\frac{N!}{i!(N-i)!}\frac{(i+1)!(N-i)!}{(N+2)!} \\ = \sum_{i=0}^{\lfloor NT\rfloor}\frac{i+1}{(N+1)(N+2)} = \frac{(\lfloor NT\rfloor+1)(\lfloor NT\rfloor+2)}{2(N+1)(N+2)}.$$

Therefore, the numerator in Eq. (11) is equal (up to the $r(m, \delta)$ term) to

$$\sum_{n=0}^{\infty} \Pr(N = n) \frac{(\lfloor nT \rfloor + 1)(\lfloor nT \rfloor + 2)}{2(n+1)(n+2)} \qquad (12)$$

Applying the same kind of analysis on the denominator in Eq. (11), we get that it is equal (up to the $r(m, \delta)$ term) to

$$\sum_{n=0}^{\infty} \Pr(N = n) \frac{\lfloor nT \rfloor + 1}{n+1}. \qquad (13)$$

Substituting Eq. (12) and Eq. (13) into Eq. (11), using the fact that $\Pr(N = n) = \exp(-m/k)(m/k)^n/n!$, and simplifying, proves the theorem. ■

Proof of Thm. 5: The main step consist of showing that

$$\frac{\mathbb{E}_{P,N}[\Pr(X_N^P \le NT)P]}{\mathbb{E}_{P,N}[\Pr(X_N^P \le NT)]} \le \mathbb{E}_P[P] = \bar{\epsilon}. \qquad (14)$$

For that, we use a result from the correlation inequalities literature, which can be traced back to Chebyshev (see thm. 43 in [HLP88]): Given two real functions $f(\cdot), g(\cdot)$ on $[0, 1]$, one monotonically increasing and the other monotonically decreasing, it holds that $\mathbb{E}[f(P)g(P)] \le \mathbb{E}[f(P)]\mathbb{E}[g(P)]$ for any random variable P on $[0, 1]$. In our case, by Lemma 9, we have that for any specific $P' > P$ and any N,

$$\Pr(X_N^{P'} \le NT) \;\le\; \Pr(X_N^P \le NT),$$

so

$$\mathbb{E}_N[\Pr(X_N^{P'} \le NT)] \;\le\; \mathbb{E}_N[\Pr(X_N^P \le NT)].$$

Applying Chebyshev's result, we get

$$\begin{aligned} &\mathbb{E}_P\left[\mathbb{E}_N\left[\Pr(X_N^P \le NT)\right] P\right] \qquad (15) \\ &\le\; \mathbb{E}_{P,N}\left[\Pr(X_N^P \le NT)\right] \mathbb{E}_P\left[P\right], \end{aligned}$$

from which Eq. (14) follows. Applying Eq. (14) on Eq. (2) in Thm. 3 and simplifying, the theorem follows. ■

5 Experiments

We tested our data-pruning approach on a large dataset, labeled using Amazon.com's *Mechanical Turk*. We first created an unlabeled set of over 8K examples, where each example included a search engine query (e.g. "doggy treats") and an Internet URL (e.g. "www.doggydelightz.com"), and the task was to determine if the query-URL pair is a relevant match or not. Each example was labeled by 15 different teachers, for a total of over 120K labels in the entire dataset. Other than requiring 15 labels from different teachers for each example, we had no control over the assignment of examples to teachers. A total of 375 individual teachers contributed labels to our dataset. A few of these teachers labeled thousands of examples, while others labeled only a handful.

We derived two additional datasets from the original dataset, with the goal of making the data resemble our theoretical framework as much as possible (since the data was already labeled, we could not follow the theoretical framework verbatim). We set a parameter μ and simulated a situation where each teacher labels at most μ examples. If teacher t labeled m_t examples and $m_t > \mu$, we partitioned those examples into $\lceil m_t/\mu \rceil$ sets and considered each set as an independent teacher. We created such datasets for $\mu = 200$ and $\mu = 50$. For uniformity, we refer to the original dataset as $\mu = \infty$. As mentioned above, the original dataset has 375 unique teachers. With $\mu = 200$ we have 881 teachers, and with $\mu = 50$ we have 2509 teachers.

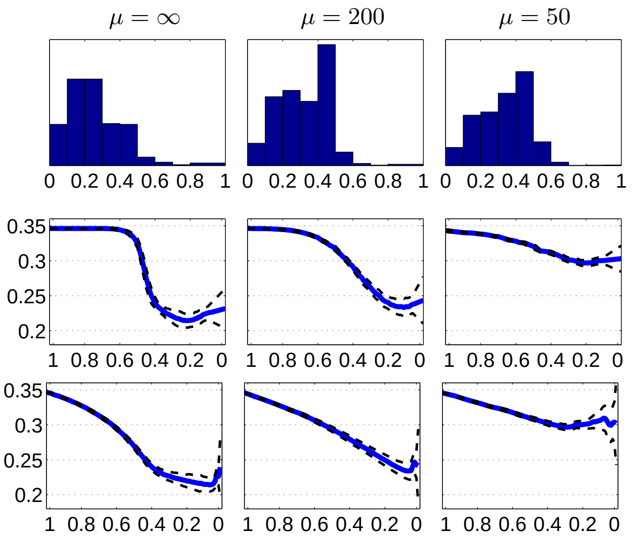

Figure 2: (Top) Distribution of teacher quality in each dataset. (Middle) Noise level in the data after pruning low-quality teachers, as a function of T, averaged over 1000 randomized repetitions of the experiment, with standard deviation. (Bottom) Noise level as a function of the fraction of unpruned examples. When $\mu = 200$, a typical teacher labels 14 examples, when $\mu = 50$ a typical teacher labels 4 examples.

The 15 labels collected for each example were used to define the ground-truth labeling of our dataset, which we used to evaluate the performance of our algorithm. Specifically, we took the majority over the 15 labels of each example, and defined that as the correct label. These correct labels were calculated once and were fixed across all of our experiments. Clearly, these majority labels were hidden from our algorithm, which was given *only one* label per example, chosen at random from the available 15. In other words, our algorithm was applied to a random 15th of the data, and as a consequence, most teachers contribute roughly $\mu/15$ labels.

Having established a ground-truth labeling, we can calculate the fraction of incorrect labels provided by each teacher (irrespective of the number of labels provided by the teacher). A histogram of these values is given for each of our three datasets in the top row of Fig. 2. These histograms demonstrate the very high variability in teacher quality. We believe that teachers that correctly labeled close to 0.5 of the data are actually automated scripts that spam Mechanical Turk and assign random labels. The noise across the entire dataset is 0.35.

As mentioned above, we randomly chose one label per example and trained a well-tuned linear SVM using this data. We then compared the labels provided by each teacher with the predictions of the SVM classifier to obtain an estimate of teacher quality. We examined what happens when we remove all teachers whose empirical error-rate exceeds T, for all values of T between 0 and 1. With $T = 1$ no data is removed, and with $T = 0$ all of the data is removed. For each value of T, we used the majority labels to determine the noise level in the data that remains after pruning. We repeated this randomized experiment 1000 times for each of our datasets. The results are presented in the bottom two rows of Fig. 2. These two rows convey the same data in two different ways: The middle row shows noise level as a function of T, while the bottom row shows noise level as a function of the fraction of examples remaining after pruning.

The figures show a steady decline in noise level as T decreases from 1 to 0, and as the dataset size decreases. Clearly, we now face a tradeoff between data quantity and quality. The best results were obtained on the $\mu = \infty$ dataset, where removing half of the dataset reduces the noise level from 0.35 to 0.27. The quality of the results deteriorates slightly when we move to the $\mu = 200$ dataset. Keep in mind that with $\mu = 200$, each teacher contributes roughly $\mu/15 = 14$ labels. The quality of our results deteriorates more significantly on the $\mu = 50$ dataset, where a typical teacher contributes around 4 labels. However, even with so few labels per teacher, the positive effects of our technique are still apparent.

6 Conclusions and Future Research

In this paper, we considered the problem of identifying low-quality teachers in a large heterogeneous crowd. Our theoretical approach was inspired by the real-world experiences of machine learning practitioners that collect data using crowdsourcing websites and search engine logs. With the right noise-filtering tools, the practical value of these seemingly chaotic sources of data increases significantly.

Despite the fact that the crowd is large and that each teacher labels only a handful of examples, we presented a simple data cleaning algorithm that does not require any prior knowledge and does not resort to the expensive practice of repeated labeling. Moreover, we showed that the cleaned data can be safely reused for the actual learning process.

There are many opportunities for future research. One possible direction is to design and analyze additional techniques such as ours. For example, we could try to directly estimate the effect a teacher has on a classifier by removing the teacher's examples, retraining, and measuring the change in the classifier. If the change is significant, this may indicate that the teacher should be removed. Also, it would be highly beneficial to design learning algorithms that are specifically tailored to datasets that are labeled by crowds, rather than adding a teacher-pruning preprocessing step to an existing algorithm. Overall, we believe that learning from crowds poses many novel, relevant, and interesting theoretical challenges.

Acknowledgment

We thank Raman Chandrasekar for giving us the idea of splitting the data in two and using each half to clean the other.

References

[AS72] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions*. Dover, 1972.

[BCK+07] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *Proc. of NIPS 21*, pages 129–136, 2007.

[CKW08] K. Crammer, M. Kearns, and J. Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9:1757–1774, 2008.

[DFP08] O. Dekel, F. Fischer, and A. Procaccia. Incentive compatible regression learning. In *Proc. of SODA 19*, pages 884–893, 2008.

[DS08] O. Dekel and O. Shamir. Learning to classify with missing and corrupted features. In *Proc. of ICML 25*, pages 216–223, 2008.

[HLP88] G. Hardy, J. Littlewood, and G. Pólya. *Inequalities*. cambridge University Press, 2nd edition, 1988.

[Kea98] M. Kearns. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006, 1998.

[KL93] M. Kearns and M. Li. Learning in the presence of malicious errors. *SIAM J. Comput.*, 22(4):807–837, 1993.

[McD89] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, volume 141, pages 148–188. Cambridge University Press, 1989.

[RP07] S. Ross and E. Pekoz. *A Second Course in Probability*. Pekozbooks, 2007.

[SFB+94] P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. Inferring ground truth from subjective labeling of Venus images. In *Proc. of NIPS 8*, pages 1085–1092, 1994.

[SPI08] V. Sheng, F. Provost, and P. Ipeirotis. Get another label? Improving data quality using multiple, noisy labelers. In *Proc. of KDD 14*, pages 614–622, 2008.

[SS02] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

[Ste03] I. Steinwart. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4:1071–1105, 2003.

[TGRS07] C. Teo, A. Globerson, S. Roweis, and A. Smola. Convex learning with invariances. In *Proc. of NIPS 21*, pages 1489–1496, 2007.