Learning from Weak Teachers

Ruth Urner
School of Computer Science
University of Waterloo
Waterloo, Ontario
Canada N2L 3G1
rurner@cs.uwaterloo.ca

Shai Ben-David
School of Computer Science
University of Waterloo
Waterloo, Ontario
Canada N2L 3G1
shai@cs.uwaterloo.ca

Ohad Shamir
Microsoft Research
New England
Cambridge MA
USA
ohadsh@microsoft.com

Abstract

This paper addresses the problem of learning when high-quality labeled examples are an expensive resource, while samples with error-prone labeling (for example generated by crowdsourcing) are readily available. We introduce a formal framework for such learning scenarios with label sources of varying quality, and we propose a parametric model for such label sources (“weak teachers”), reflecting the intuition that their labeling is likely to be correct in label-homogeneous regions but may deteriorate near classification boundaries. We consider learning when the learner has access to weakly labeled random samples and, on top of that, can actively query the correct labels of sample points of its choice. We propose a learning algorithm for this scenario, analyze its sample complexity and prove that, under certain conditions on the underlying data distribution, our learner can utilize the weak labels to reduce the number of expert labels it requires. We view this paper as a first step towards the development of a theory of learning from labels generated by teachers of varying accuracy, a scenario that is relevant in various practical applications.

1 Introduction

This work was motivated by the following problem: There is considerable interest in the development of automated programs for diagnosis of brain tumors from CT scans of the skull. Machine learning is a natural approach for the development of such programs, but machine learning tools require the input of large labeled samples. Thus, we would need large amounts of images classified according to whether, say, tumors appearing in the images are benign or malignant. However, it is very difficult and expensive to obtain such classifications from top human experts. Alternatively, one could have medical students label the images, which are available both in larger numbers and at a much lower cost. The labeling provided by students may be erroneous though. Especially so for images that are difficult to classify. A possible compromise may be to consult students for the vast majority of the training images, but refer a few of tho images to an expert–those that are most challenging to classify (or most crucial for the design of the classification program).

Similar scenarios of utilizing such “weak” but readily available teachers in the process of learning arise in many other practical domains. For instance, weak labels might be generated using Amazon’s Mechanical Turk. While such crowdsourcing mechanisms are becoming more and more popular, no formal model for learning with label sources of varying quality has been developed yet. This work aims at providing a first step in extending classical statistical learning theory to this important area of applications. Many questions arise in this context: How should such weak teachers be modeled? (Clearly, their errors are not just random noise - the likelihood of mislabeling an instance varies between instances, and averaging multiple queries to the same instance will not wash off an error.) How should an output classifier be computed from a mixture of expert-labeled and novice-labeled training data? How should one decide which instances to refer to an expert? Can such a paradigm save calls to an expert without compromising prediction accuracy by too much?

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. Copyright 2012 by the authors.

1.1 Outline of our model and results

We propose a formal model of what we call *learning from weak teachers*. In this model, the learner gets a randomly generated set of "weakly labeled" examples (according to some mechanism for generating weak labels), and may then pick some of the training examples and query their correct label. The learner's goal is to find a good label predictor while making as few as possible such queries.

We address situations in which the quality of labels provided by the weak teacher varies with the degree of label homogeneity in a certain neighborhood of labeled points, presuming that a non-expert labeler is more likely to err when the correct label depends on subtle differences in the features than in clear-cut cases that are far from a decision boundary. We formalize this intuition as two requirements on the labeling rule applied by the weak teacher. Our requirements are formulated in terms of the expected label of 'neighboring instances' (weighted by their similarity to the instance we wish to label). The first requirement is that for instances having that expected label close to either 0 or 1, the weak teacher should provide that value (with high probability). The second requirement states that the weak teacher should "hesitate" when labeling instances for which that expected label on the 'neighbors' is close to 1/2 (that is, the similarity neighborhood is label-heterogeneous). We then show that, under some mild conditions concerning the data generating distribution, any labeling rule satisfying these requirements can be utilized as a weak teacher for learning that will result in saving of exact-label queries.

Our algorithmic paradigm is based on estimating the confidence that a learner can have in the quality of a weak label for a point, and querying exact labels only for points of low confidence. We introduce a precise notion of that confidence and show that it can be reliably estimated from finite samples. We hope that the modeling of the problem, the formalization of required data assumptions and the demonstration of algorithmic success will stimulate more research on this type of learning tasks.

After introducing our notion of weak teachers (more precisely, the requirements such teachers should meet) in Section 3, we formulate some natural "mildness" conditions on the underlying data-generating distribution in Section 4. We then propose a learning algorithm that utilizes weakly labeled examples in Section 5, and in Section 6 we analyze its the sample complexity and show that our proposed learning paradigm succeeds in utilizing a weakly labeled training sample to generate label predictors of high accuracy while reducing the required number of correctly labeled examples.

1.2 Previous work

While we are not aware of any theoretical work that proposes and analyzes a general formal model for learning with a mixture of weak and strong supervision, there have been experimental studies investigating similar scenarios. Vijayanarasimhan and Grauman [11] consider learning from two types of teachers in the context of image annotation tasks. They develop a learning algorithm that allows the learner to choose between labels provided by strong and weak types of supervision, based on the expected information gain from each such label. There are various recent experimental studies on the use of crowdsourcing and the problem of detecting levels of expertise of label-sources, e.g., [7], [8], [12], [13]. In motivation maybe most similar to our work is a study by Sheng et al. [10]. They, experimentally, investigate multiple labeling as method for replacing correct supervision with less reliable but vastly available weak labels.

Most of the work on supervised classification learning theory assumes that the supervision, the labels provided for the training examples, are correct labels. There are two established directions in which this assumption is relaxed that are related to our work. The first is learning from noisy data. The *learning with noise* model assumes (in its most common version) that there is some small parameter η (the *noise level*) such that each example in the training set gets its correct label with probability $(1-\eta)$, and with probability η its label is flipped [2], [4], [6]. However, while this is an appropriate model for, say, transmission noise, where the noise level is independent of the specific example, we are interested in modeling labeling flaws that stem from non-expert human labelers, where we will encounter varying label quality depending on the difficulty to label the example in question.

The second relevant framework is *active learning*, where the learner gets an unlabeled sample generated by the environment, but can then choose a subset from the sample for which the labels should be disclosed [5]. The difference between our proposed weak teachers model and active learning is that in our model the learner does get some label-related information (in the form of weak teacher labels) for each of the sample points, not just for those actively queried.

2 Preliminaries

Let $(\mathcal{X}, \Delta)$ be some domain set endowed with a metric $\Delta : \mathcal{X}^2 \to \mathbb{R}^+$ ($\mathbb{R}^+$ denoting the set of non-negative real numbers) and $\{0, 1\}$ a label set (often, the domain set is a subset of a Euclidean space and $\Delta(x, y)$ is the Euclidean distance $||x - y||$). Given some probability distribution, P over $\mathcal{X} \times \{0,1\}$, we denote its marginal distribution over $\mathcal{X}$ by D and let p denote the conditional probability distribution over the labels, defined by $p(x) = \Pr_{(X,Y)\sim P}(Y = 1 \mid X = x)$. For this, we use the notation $P = (D, p)$. We define the *β-ball* around a point $x \in \mathcal{X}$ by $B_\beta(x) = \{y \in \mathcal{X} \mid \Delta(x,y) \le \beta\}$.

Let $h : \mathcal{X} \to \{0,1\}$ be a hypothesis. We define the *error* of h with respect to P as $\mathrm{Err}_P(h) = \Pr_{(x,y)\sim P}(y \neq h(x))$. For a class H of hypotheses on $\mathcal{X}$, we denote the smallest error of a hypothesis $h \in H$ with respect to P by $\mathrm{opt}_H(P) := \min_{h\in H} \mathrm{Err}_P(h)$.

Definition 1. *An algorithm A is an* agnostic learner *for some hypothesis class H over $\mathcal{X}$ if for all $\epsilon > 0$ and $\delta > 0$ there exists a sample size $m = m_A(\epsilon, \delta)$ such that, for all distributions P over $\mathcal{X} \times \{0,1\}$, when given an i.i.d. sample of size m from P, then with probability at least $1 - \delta$ over the sample, A outputs a classifier $h \in H$ with error at most $\mathrm{opt}_H(P) + \epsilon$.*

By basic VC-dimension theory (see, for example, [3]) there is an agnostic learner for a class H if and only if H has finite VC-dimension. In case of finite VC-dimension, d, the basic ERM (empirical risk minimization) paradigm is an agnostic learner for H with $m_{\mathrm{ERM}}(\epsilon,\delta) = \tilde{O}(\frac{d+\log(1/\delta)}{\epsilon^2})$.

2.1 Similarity notions

Many machine learning paradigms are based on the postulate that "similar points tend to have similar labels". Our modeling of weak teachers is based on having such a measure of similarity between instances (which might be related, but not necessarily equal, to the metric Δ over the domain of instances). We start our discussion of weak teachers with the notion of similarity between domain points. This notion of similarity is part of the modeling of the weak teacher, and we do not require that it is known to the learner.

Definition 2. *Let $\mathcal{X}$ be a domain set. A* similarity function *is a symmetric function $s : \mathcal{X}^2 \to [0,1]$ with $\max_{y\in\mathcal{X}} s(x,y) = s(x,x) = 1$ for all $x \in \mathcal{X}$.*

In this work, we consider similarity functions of the form $s(x,y) = f(\Delta(x,y))$ where $f : \mathbb{R}^+ \to [0,1]$ is a continuous non-increasing function. Note that most of the kernels over Euclidean spaces can be cast as such.

Example 1. *Let our domain $\mathcal{X}$ be some Euclidean space $\mathbb{R}^n$*

1. *We can model the notion of similarity using an n-variate Gaussian distribution with variance α around a domain point x, i.e. for all $x, y \in \mathcal{X}$*

$$s(x,y) = \frac{1}{\sqrt{(2\pi\alpha)^n}} \mathrm{e}^{\frac{||x-y||^2}{2\alpha}}$$

2. *One can model the similarity as a "landfill" function. For some radius α, threshold $0 < t \le 1$ and slope $c > 0$, we have for all $x, y \in \mathcal{X}$*

$$s(x,y) = \begin{cases} 1 & \text{if } ||y-x|| \le \alpha \\ 1 - c(||y-x|| - \alpha) & \text{if } \alpha \le ||y-x|| \\ & \le \alpha + \frac{1-t}{c} \\ t & \text{otherwise.} \end{cases}$$

3 Our modeling of weak teachers

Learning tasks utilizing label supervision from a non-perfectly-reliable source (or, "weak teachers") may arise in a variety of practical learning scenarios. The nature of labeling errors by such supervision may vary between such tasks. Consequently, it is unlikely that there exists a generic precise way in which such "weak labels" are modeled. Just the same, the following properties seem to be common to many practical examples of weak labelings and they distinguish such labelings from the common model of random labeling noise (as discussed in [2]).

1. The probability of mislabeling an instance varies between instances (some instances are easy to label, and even novice supervisors are not likely to err on them).

2. The labeling is error is rather persistent - it cannot be canceled out by averaging repeated labelings of a given instance.

These properties are, of course, not sufficient to determine the nature of weak-teacher's errors. In this work we consider one possible way of modeling weak teachers. As an intuitive motivation for our modeling, consider the example of diagnostic labeling of medical brain images described in the introduction.

Our modeling of weak teachers is based on a notion of instance similarity. We postulate that the chance of mislabeling an instance depends on how likely it is that instances similar to it have a different label. A novice doctor is likely to label images, which are "surrounded" by similarly labeled images, correctly. Likewise, the labels provided by weak teachers will be more often wrong for images whose similarity neighborhood is heterogeneous in terms of the (true) labels. Our formal definitions below aim to capture this intuition.

3.1 The requirements from weak teachers

We define our notion of weak labeling by imposing two requirements. Any labeling rule that meets these requirements can be utilized by our learning paradigm to save label queries to a strong teacher. We consider probabilistic labeling rules, $w : \mathcal{X} \to [0,1]$, where for any instance x, $w(x)$ is the probability that the rule assigns to x the label 1. Our first requirement is to assign the majority label of a neighborhood (with high probability) if this majority is sufficiently large. Our second requirement is to assign both labels (that is, have $w(x)$ in the central region of the $[0,1]$ interval) if the neighborhood is not homogeneous.

The formalization of these requirements relies on the following definition of *the similarity neighborhood* of an instance and the *s-smoothed labeling function*:

Definition 3. *Given a probability distribution D over $\mathcal{X}$ and a similarity function s over $\mathcal{X}$,*

1. *The* neighborhood distribution *of an instance $x \in \mathcal{X}$ is a probability distribution $D_{s(x)}$ over $\mathcal{X}$. Assuming D has density function d the density $d_{s(x)}$ of $D_{s(x)}$ is defined as*

$$d_{s(x)}(y) = \frac{d(y)s(x,y)}{\mathbb{E}_{y\sim D} d(y)s(x,y)}.$$

   Namely, the neighborhood distribution of a point x is a version of D biased towards points similar to x.

2. *Given a distribution, $P = (D,p)$ over $\mathcal{X} \times \{0,1\}$, the s-*smoothed labeling function *$\pi_s : \mathcal{X} \to \{0,1\}$ is the expectation of the label value over the neighborhood distribution:*

$$\pi_s(x) = \mathbb{E}_{z\sim D_{s(x)}} p(z).$$

The value $\min\{\pi_s(x), 1-\pi_s(x)\}$ indicates the *heterogeneity of the neighborhood* around x[1].

In this paper we assume that the similarity notion is such that $\pi_s(x)$ satisfies a Lipschitz condition. We prove in the supplementary material [1] that this holds for a large family of similarity notions (including the ones in Example 1) with a Lipschitz constant that is independent of the underlying data distribution.

The labeling function $w : \mathcal{X} \to [0,1]$ (where $w(x)$ is the probability of assigning the label 1 to the instance x) of the weak teacher is required to comply with the s-smoothed labeling function π_s of the neighborhoods in the following sense:

Definition 4. *Let $\nu \le \mu \in [0,1/2]$. A probabilistic labeling function, $w : \mathcal{X} \to [0,1]$, qualifies as a weak teacher with threshold μ and slack ν whenever, for all $x \in \mathcal{X}$,*

$$w(x) \begin{cases} \le \mu+\nu & \text{if } \pi_s(x) \le \mu \\ \in (\mu-\nu, 1-\mu+\nu) & \text{if } \pi_s(x) \in (\mu, 1-\mu) \\ \ge 1-\mu-\nu & \text{if } \pi_s(x) \ge 1-\mu \end{cases}$$

In other words, when the neighborhood of an instance x is rather label homogeneous (as reflected by $\pi_s(x) \le \mu$ for the label 0 and $\pi_s(x) \ge 1-\mu$ for the label 1), the weak teacher should give high probability ($\ge 1-\mu-\nu$) to assigning that value. On the other hand, when the neighborhood of x is heterogeneous in terms of its labels ($\pi_s(x)$ in the central region of the $[0,1]$ interval), the weak teacher is expected to assign non-negligible probability ($\ge \mu-\nu$) to each of the two labels.

In contrast to this, a teacher is *strong* if it always assigns the label according to the distribution P.

Example 2.

1. *We can model a weak teacher that chooses a random point according to $D_s(x)$ and returns its label. In this case, we have a weak teacher without threshold (i.e. it satisfies the definition for any threshold $\mu \in [0,1/2]$) and slack* 0 *with $w(x) = \pi_s(x)$ for all $x \in \mathcal{X}$.*

2. *We can allow a weak teacher small deviations from $\pi_s(x)$, thus a weak teacher without threshold and with slack ν could be modeled by any function which satisfies $w(x) \in (\pi_s(x)-\nu, \pi_s(x)+\nu)$ for all $x \in \mathcal{X}$.*

3. *A weak teacher could also use μ as a threshold of confidence to decide whether to label a point deterministically according to the majority of similar points or to flip an unbiased coin for the label. In this case we have for all $x \in \mathcal{X}$*

$$w(x) = \begin{cases} 0 & \text{if } \pi_s(x) \le \mu \\ 1/2 & \text{if } \pi_s(x) \in (\mu, 1-\mu) \\ 1 & \text{if } \pi_s(x) \ge 1-\mu \end{cases}$$

For the sake of simplicity of our arguments, we will from now on assume that the weak teacher labels exactly according to the heterogeneity of the neighborhoods, *i.e.* $w(x) = \pi_s(x)$ as in the first of these examples. We would like to stress that our analysis holds for the more general notion of a weak teacher as formalized by definition 4.

4 Mildness properties of distributions

Many learning paradigms can be viewed as utilizing a notion of similarity between instances for which similar points are likely to assume similar labels (nearest neighbor algorithms make the most explicit use of

[1] Slightly abusing the notions, we will often refer to the *neighborhood* of a point x instead of points distributed according to the neighborhood distribution. Note the distinction between the *neighborhood* of a point, which refers to the neighborhood distribution induced by a similarity and the (metric) *ball* around a point.

such an assumption). We now introduce properties of distributions that capture a certain compliance of the labeling function with the similarity.

"Nice" distributions

The following definition formalizes the property of having not too many label-heterogeneous neighborhoods. This can be viewed as stating that the underlying marginal distribution over the instances is sparse around the label decision boundaries.

Definition 5. *For values* $0 \leq \lambda \leq 1/2$ *and* $0 \leq \kappa \leq 1$, *we say that a distribution* $P = (D, p)$ *is* (λ, κ)-nice *with respect to a similarity function* s *if*

$$\Pr_{x \sim D}(\min\{\pi_s(x), 1 - \pi_s(x)\} \geq \lambda) \leq \kappa$$

That is, the average label over a local neighborhoods is, for most of the instances, either close to 0 or close to 1.

For a function $\psi : [0, 1/2] \to [0, 1]$, *we say that a distribution* P *is* ψ-nice *if it is* $(\lambda, \psi(\lambda))$*-nice for all* $0 \leq \lambda \leq 1/2$.

Note that ψ-niceness always holds for ψ being the constant 1 function. We require ψ to be a monotonically decreasing function with $\psi(0) = 1$. A distribution is intuitively very nice, if there exists a small λ and a small κ such that the distribution is (λ, κ)-nice (in other words, if the average label in local neighborhoods is λ-close to either 0 or 1 for all but a κ mass of the instances).

"Local conservative" distributions

Our second mildness requirement bounds the probability of points that have a label different from the vast majority of their neighbors (w.r.t. our measure of similarity). We call this property "local conservativeness".

Definition 6. *For values* $0 \leq \lambda \leq 1$ *and* $0 \leq \kappa \leq 1$, *we say that a distribution* $P = (D, p)$ *is* (λ, κ)-locally conservative *with respect to a similarity function* s, *if*

$$\Pr_{x \sim D}(|\pi_s(x) - p(x)| > 1 - \lambda) \leq \kappa$$

That is, for most of the instances, the average label over their neighborhood is a good approximation of the probability of having label 1 for that instance.

For a function $\varphi : [0, 1] \to [0, 1]$, *we say that a distribution* P *is* φ-locally conservative *if it is* $(\lambda, \varphi(\lambda))$*-locally conservative for all* $0 \leq \lambda \leq 1$.

Note that every distribution is φ-locally conservative if φ is the constant 1 function. In order to let φ-local conservativeness be a meaningful property we therefore need to require that φ is a monotonically increasing function with $\varphi(0) = 0$. Intuitively, a distribution is very locally conservative, if there exists a not too small λ and a very small κ such that the distribution is (λ, κ)-locally conservative. See the supplementary material [1] for further discussion of these properties.

5 Our learning algorithm

Our algorithmic paradigm is based on evaluating the confidence that we should assign labels provided by the weak teacher. Once we have a method for estimating that confidence, the learning algorithm runs as follows:

1. Obtain two random samples of domain points, S and T (sampled i.i.d. according to the marginal distribution) and query the weak teacher for the labels of the points in S.
2. Use these labels to estimate our confidence in the correctness of the labels that would be assigned by the weak teacher to each point of T.
3. Query the strong teacher about the labels of points in T for which that confidence is low.
4. Label the high confidence points of T using the weak teacher's labels on the sample S.
5. Pass T with the labels obtained (by this procedure) to an agnostic learner.

In our setting the confidence (or uncertainty) of the weak teacher corresponds to the homogeneity of the labels of neighborhoods. However, the learner does not know the weak teacher's notion of similarity for the neighborhoods. It thus needs to estimate the uncertainty about the weak teacher's labels using the metric of the input feature space. For each point x in T, we let the algorithm estimate the homogeneity by averaging the weak labels of points in S in a ball of radius β around it. Given a labeled set of weakly labeled points, $S \subseteq \mathcal{X} \times \{0, 1\}$, and some radius $\beta > 0$ we define the (S, β)-estimate of $\pi_s(x)$ by

$$w_{S,\beta}(x) = \frac{1}{|S|} \sum_{(z,y) \in B_\beta(x) \cap S} y.$$

The algorithm uses a threshold parameter η for the uncertainty of the weak teacher's labels in order to decide whether or not to call the strong teacher. We fix some radius, $\beta > 0$, and, given a weakly labeled sample S and a point $x \in T$, use $w_{S,\beta}(x)$ to estimate $\pi_s(x)$. The algorithm calls the strong teacher for x if $\min\{w_{S,\beta}(x), 1 - w_{S,\beta}(x)\} \geq \eta$ and otherwise labels x with 1 or 0 depending on whether $\min\{w_{S,\beta}(x), 1 - w_{S,\beta}(x)\} \geq 1/2$ or not.

6 Analysis

In this section we analyze the algorithm outlined above.

We start by showing in 6.1 that for a sufficiently large sample S we can guarantee that with high probability $w_{S,\beta}$ is a close approximation of the smoothed labeling function π_s, thus we can estimate the uncertainty from samples of distribution independent size. In 6.2 we then prove that the algorithm is guaranteed to output a good label predictor while saving queries to the strong teacher.

More concretely, we show, in Theorem 1 that a large enough sample of weakly-labeled points suffices to provide a reliable estimate of the s-smooth labeling function, π_s. Then, in Lemma 3, we show that having such an estimate, Step 4 of our algorithm provides a labeling that is very close to that provided by the true labeling function (where the quality of this approximation depends on the local conservativeness of the data-generating distribution, and the algorithm's choice of confidence threshold, η). Theorem 2 then combines the previous results to show that, given any learnable hypothesis class and a batch learning algorithm for it (that uses usual correctly labeled examples), using that algorithm as a subroutine, our algorithm is guaranteed to learn the class with a number of strong-teacher queries that is smaller than the number of labeled examples the subroutine algorithm requires (where the sample size gain depends on the niceness of the underlying data distribution).

Throughout this section, we fix some ψ and φ and consider only distributions P over $\mathcal{X} \times \{0,1\}$ that are ψ-nice, φ-locally conservative and have deterministic labels, *i.e.* $p(x) \in \{0,1\}$ for all $x \in \mathcal{X}$.

6.1 Estimation of the uncertainty

Here we prove that by choosing a sufficiently large size for the weakly labeled set S, we can obtain a close approximation of the s-smoothed labeling function (and thus of the confidence or the uncertainty in the weak teachers labels). We start by showing that we can obtain such an approximation if we assume π_s to be Lipschitz *i.e.* if we have $\frac{|\pi_s(x)-\pi_s(y)|}{\Delta(x,y)} \leq L$ for all $x, y \in \mathcal{X}$ for some constant $L \in \mathbb{R}$. We then claim that for a large family of similarity notions, the s-smoothed labeling function π_s is guaranteed to be Lipschitz with a constant that is independent of the underlying distribution. This claim is formalized and proved in the supplementary material [1].

Given a set of weakly labeled points, $S \subseteq \mathcal{X} \times \{0,1\}$, and some parameter $\beta > 0$ recall that we define the (S,β)-estimate of $\pi_s(x)$ by $w_{S,\beta}(x) = \frac{1}{|S|}\sum_{(z,y)\in B_\beta(x)\cap S} y$. We start by showing that for a sufficiently large set S, we can guarantee that with high probability the β-ball around a domain point x contains many points from S.

Lemma 1. *For every δ, every $k \in \mathbb{N}$, every $\beta > 0$ and every $\gamma > 0$, there exists an $M \in \mathbb{N}$ such that for an i.i.d. D-sample S of size at least M we have $\Pr_{x\sim D}[|S \cap B_\beta(x)| \geq k] \geq 1-\gamma$ with probability at least $1-\delta$ (over the choice of S).*

Proof. We use the following result from [9] (Lemma 21, p. 68): If D is a distribution over some domain $\mathcal{X}$ and $C_1, \ldots, C_r$ is a sequence of subsets of $\mathcal{X}$, then $\mathbb{E}_{S\sim D^n}[\sum_{C_i:S\cap C_i=\emptyset} D(C_i)] \leq \frac{r}{n\mathrm{e}}$. Let R be the number of balls of radius $\beta/2$ that are needed to cover the space $\mathcal{X}$ and let $C_1, \ldots, C_R$ be such a cover. Now, if we divide the sample S into k subsets $S_1, \ldots, S_k$, each of size n (we can consider this as taking k samples of size n), we obtain that

$$\begin{aligned}
&\mathbb{E}_{S_1,\ldots,S_k\sim D^n}[\sum_{C_i:S\cap C_i\leq k} D(C_i)] \\
&\leq \mathbb{E}_{S_1,\ldots,S_k\sim D^n}[\sum_{C_i:(\exists S_j:C_i\cap S_j=\emptyset)} D(C_i)] \\
&\leq \mathbb{E}_{S_1,\ldots,S_k\sim D^n}[\sum_{j=1}^{k}(\sum_{C_i:(C_i\cap S_j=\emptyset)} D(C_i))] \leq \frac{kR}{n\mathrm{e}}.
\end{aligned}$$

Thus, by Markov's inequality, it follows that for every $\delta > 0$ and every $\gamma > 0$ there is a size $M(= kn)$ such that we have $\Pr_{x\sim D}[|S \cap B_\beta(x)| \geq k] \geq 1-\gamma$ with probability at least $(1-\delta)$ over i.i.d. samples S of size M. □

Next, we show that if the β-ball around a domain point x contains sufficiently many points from S then $w_{S,\beta}(x)$ is close to $\pi_s(x)$ with high probability.

Lemma 2. *Assume that the s-smoothed labeling rule, $\pi_s : \mathcal{X} \to [0,1]$ satisfies the Lipschitz condition with a constant L, let x be any domain point and $\beta > 0$ and let S be an i.i.d. D-sample, that is labeled by the weak teacher, then, for every $\epsilon > 0$ the probability (over the choice of S) that $|w_{S,\beta}(x) - \pi_s(x)| \geq \beta L + \epsilon$ is less than $\mathrm{e}^{-2k\epsilon^2}$, where $k = |S \cap B_\beta(x)|$.*

Proof. By the Lipschitz condition, for every $u \in |S \cap B_\beta(x)|$ we have $|\pi_s(x) - \pi_s(u)| \leq \beta L$. We can view S as generated by first sampling the instances according to D, and then sampling the labels using the weak teacher. Once the instances have been fixed, we can view the labels assigned to the various points u by the weak teacher as an independent Bernoulli random variables with expectation $\pi_s(u)$. Therefore, the expectation of $w_{S,\beta}(x) = \frac{1}{k}\Sigma_{u \in S \cap B_\beta(x)} \ell_w(u)$ equals $\frac{1}{k}\Sigma_{u \in S \cap B_\beta(x)} \pi_s(u)$. Since for each such u we have $|\pi_s(x) - \pi_s(u)| \leq \beta L$, we get that $|\mathbb{E}(w_{S,\beta}(x) - \pi_s(x))| \leq \beta L$. The result now follows from Hoeffding's inequality. □

These lemmas imply that we can approximate $\pi_s(x)$ with arbitrarily high precision by choosing a sufficiently large size for the set S and a sufficiently small radius β:

Theorem 1. *Assume that for some underlying data distribution D the s-smoothed labeling rule, $\pi_s : \mathcal{X} \to [0,1]$ satisfies the Lipschitz condition with a constant L. Then for every $\epsilon > 0$, $\delta > 0$, every $\gamma > 0$ there are values M and β determined by ϵ, δ, L and γ such that for a weakly labeled i.i.d. D-sample S of size at least M we have $\Pr_{x \sim D}[|w_{S,\beta}(x) - \pi_s(x)| \leq \epsilon] \geq 1 - \gamma$ with probability at least $1 - \delta$ (over the choice of S).*

Proof. For a sample size M and parameters β and k, say that a pair $(x, S) \in \mathcal{X} \times \mathcal{X}^M$ is (β, k)-good, if $|S \cap B_\beta(x)| \geq k$ and that such a pair is (ϵ, β)-approx, if $|w_{S,\beta}(x) - \pi_s(x)| \leq \epsilon$. Lemma 1 implies that for any γ and δ, we can choose M large enough so that $\Pr_{x \sim D,\ S \sim D^M}[(x, S) \text{ is } (\beta, k)\text{-good}] \geq 1 - \gamma - \delta$. From this, we get that for every η

$$\Pr_{x \sim D}\left[\Pr_{S \sim D^M}[(x,S) \text{ is not } (\beta,k)\text{-good }] \geq \eta\right] \leq (\gamma + \delta)/\eta.$$

Thus, we changed the order of quantifiers (in Lemma 1) and have that for any δ and η, we can choose a sample size M such that

$$\Pr_{x \sim D}\left[\Pr_{S \sim D^M}[(x,S) \text{ is } (\beta,k)\text{-good }] \geq 1 - \eta\right] \geq 1 - \delta.$$

Choosing $\beta = \epsilon/L$, Lemma 2 implies that, for every x and every k, conditioned upon S being such that the pair (x, S) is (β, k)-good, the probability over $S \sim D^M$ that (x, S) is $(2\epsilon, \beta)$-approx is at least $1 - e^{-2k\epsilon^2}$. For a fixed x with $\Pr_{S \sim D^M}[(x, S) \text{ is } (\beta, k)\text{-good }] \geq 1 - \eta$ this yields

$$\Pr_{S \sim D^M}[(x,S) \ \text{ is } (2\epsilon,\beta)\text{-approx}] \geq (1-\eta)(1 - e^{-2k\epsilon^2}).$$

For a given ϵ and γ, we can choose k and η such $(1 - \eta)(1 - e^{-2k\epsilon^2}) \geq 1 - \gamma$. Thus we have shown that for any ϵ, γ and δ, we can choose β and M, so that

$$\Pr_{x \sim D}\left[\Pr_{S \sim D^M}[(x,S) \text{ is } (2\epsilon,\beta)\text{-approx}] \geq 1 - \gamma\right] \geq 1 - \delta.$$

As above, we can change the order of quantification over x and S, which yields the claim of the theorem. □

6.2 Sample complexity analysis

Given a weakly labeled set S and a point x in $\mathcal{X}$, we let p_S denote the label that x receives by the procedure that we use for members of the set T (before being passed to the agnostic learner). We denote the induced probability distribution over $\mathcal{X} \times \{0,1\}$ by $P_S = (D, p_S)$. We start by showing that if $w_{S,\beta}$ is a good estimate for π_s then the distribution P_S is close to the true distribution P in the following sense:

Lemma 3. *Let η be the threshold for the uncertainty the algorithm uses and assume that S is such that we have $\Pr_{x \sim D}[|w_{S,\beta}(x) - \pi_s(x)| \leq \epsilon_0] \geq 1 - \gamma$, then for every $l : \mathcal{X} \to \{0,1\}$,*

$$|\mathrm{Err}_{P_S}(l) - \mathrm{Err}_P(l)| \leq \varphi(\eta + \epsilon_0) + \gamma.$$

Proof. We have $|w_{S,\beta}(x) - \pi_s(x)| \leq \epsilon_0$ with probability at least $1 - \gamma$. We now assume that this inequality holds, which yields the additive factor γ.

Consider a random x drawn according to D. It suffices to bound the probability that $p_S(x) \neq p(x)$. If we have $\min\{\pi_s(x), 1 - \pi_s(x)\} \leq \eta + \epsilon_0$ and the true label $p(x)$ does correspond to the majority of labels in the neighborhood, then the algorithm assigns the correct label $p_S(x) = p(x)$ (assuming that η and ϵ_0 were chosen small enough for $w_{S,\beta}(x) \leq \eta + 2\epsilon_0 \leq 1/2$ to hold). Further, if we have $\min\{\pi_s(x), 1 - \pi_s(x)\} \geq \eta + \epsilon_0$, the algorithm calls the strong teacher as this implies $\min\{w_{S,\beta}(x), 1 - w_{S,\beta}(x)\} \geq \eta$, thus in this case we have $p_S(x) = p(x)$ as well.

The probability that we have $\min\{\pi_s(x), 1 - \pi_s(x)\} \leq \eta + \epsilon_0$ and the true label $p(x)$ not corresponding to the majority of labels in the neighborhood (*i.e.* $p(x) = 1$ and $\min\{\pi_s(x), 1 - \pi_s(x)\} = \pi_s(x)$ or $p(x) = 0$ and $\min\{\pi_s(x), 1 - \pi_s(x)\} = 1 - \pi_s(x)$), is bounded by $\varphi(\eta + \epsilon_0)$ because of the local conservativeness.

□

Now we bound the error of our algorithm if it uses an agnostic learner as a subroutine (see Definition 1).

Lemma 4. *Let H be a hypothesis class of finite VC-dimension and let B be an agnostic learner for H. Then, for all $\epsilon > 0$ and $\delta > 0$, there is a finite sample size M and a threshold η for our algorithm such that, for a weakly labeled sample S of size at least M and an unlabeled sample T of size $m_B(\epsilon/2, \delta/2)$, we have*

$$\mathrm{Err}_P(h) \leq \mathrm{opt}_H(P) + \epsilon$$

with probability at least $1 - \delta$, where $h \in H$ is the hypothesis that our algorithm outputs.

Proof. We can choose a threshold η for the algorithm, a $\gamma > 0$ and an ϵ_0 such that $2(\varphi(\eta + \epsilon_0) + \gamma) < \epsilon/2$, where φ is the local conservativeness of the distribution. Further, we can choose an M such that with probability $1-\delta/2$ we have $\Pr_{x\sim D}[|w_{S,\beta}(x)-\pi_s(x)| \leq \epsilon_0] \geq 1-\gamma$ by Theorem 1. Let h^* be the hypothesis with minimal error with respect to P, *i.e.* $\mathrm{Err}_P(h^*) = \mathrm{opt}_H(P)$. Then Lemma 3 implies

$$\mathrm{Err}_{P_S}(h^*) \leq \mathrm{Err}_P(h^*) + \varphi(\eta+\epsilon_0) + \gamma$$

Thus, the guarantee on the agnostic learner B yields that with probability at least $1-\delta/2$ we have

$$\mathrm{Err}_{P_S}(h) \leq \mathrm{Err}_P(h^*) + \varphi(\eta+\epsilon_0) + \gamma + \epsilon/2$$

In order to get a bound on the error of h with respect to the original distribution P we use Lemma 3 again. This implies that with probability at least $1-\delta$ we have $\mathrm{Err}_P(h) \leq \mathrm{Err}_P(h^*) + 2(\varphi(\eta+\epsilon_0)+\gamma) + \epsilon/2 \leq \mathrm{opt}_H(P) + \epsilon$.

□

We now bound the expected number of calls to the strong teacher that the algorithm makes.

Lemma 5. *Let P be a ψ-nice distribution. If η is the threshold for the uncertainty the algorithm uses and the weakly labeled sample S is such that we have $\Pr_{x\sim D}[|w_{S,\beta}(x)-\pi_s(x)| \leq \epsilon_0] \geq 1-\gamma$ for some $\gamma > 0$, then the expected number of calls to the strong teacher the algorithm makes is bounded by $|T|(\psi(\eta-\epsilon_0)+\gamma)$.*

Proof. The algorithm makes a call to the strong teacher if $\min\{w_{S,\beta}(x), 1-w_{S,\beta}(x)\} \geq \eta$. We have $\Pr_{x\sim D}(\min\{\pi_s(x), 1-\pi_s(x)\} \geq \eta-\epsilon_0) \leq \psi(\eta-\epsilon_0)$ by the niceness. As we have $|w_{S,\beta}(x)-\pi_s(x)| \leq \epsilon_0$ with probability $1-\gamma$ we get $\Pr_{x\sim D}(\min\{w_{S,\beta}(x), 1-w_{S,\beta}(x)\} \geq \eta) \leq \psi(\eta-\epsilon_0)+\gamma$. This implies the claim. □

The following theorem summarizes our algorithms sample complexity, as established in this section:

Theorem 2. *Let H be a hypothesis class of finite VC-dimension and let B be an agnostic learner for H. Then, for every $\epsilon > 0$, $\delta > 0$, functions ψ and φ, and every γ, ϵ_0 and η satisfying $2(\varphi(\eta+\epsilon_0)+\gamma) \leq \epsilon/2$, there exists a sample size M, such that our algorithm, given a weakly labeled random sample S of size at least M and a random unlabeled sample T of size $m_B(\epsilon/2, \delta/2)$ generated by a distribution P over $\mathcal{X}\times\{0,1\}$ that is φ-locally conservative and ψ-nice, with probability exceeding $(1-\delta)$ outputs a hypothesis with error at most $\mathrm{opt}_H(P)+\epsilon$ using at most $(\psi(\eta-\epsilon_0)+\gamma)m_B(\epsilon/2,\delta/2)$ calls to the strong teacher.*

Note that the sample complexity of ERM for classes of finite VC dimension depends quadratically on $1/\epsilon$ and only logarithmically on $1/\delta$. Thus we can approximate $m_B(\epsilon/2,\delta/2)$ with $1/4m_B(\epsilon,\delta)$ and get:

Corollary 1. *For every $\epsilon_0 > 0$ and $\gamma > 0$, if the parameter η, chosen by the learning algorithm, is such that $4(\psi(\eta-\epsilon_0)+\gamma) < 1$, then the expected number of calls to the strong teacher that our algorithm makes is less than the sample size of strongly-labeled examples that would be needed by the agnostic learner B without access to weakly labeled points.*

Discussion While the condition $2(\varphi(\eta+\epsilon_0)+\gamma) \leq \epsilon/2$ in Theorem 2 requires the parameters η, ϵ_0 and γ to be set sufficiently small, the corollary shows that we save queries to the strong teacher in comparison to the agnostic learner B only if $4(\psi(\eta-\epsilon_0)+\gamma) < 1$ (and, by the conclusion of the theorem, the saving is proportional to that quantity) which requires these parameters to be not too small.

In other words, we need to set the threshold for the algorithm so small, that the local conservativeness guarantees that there are not too many points with a label different from the (roughly) $1-\eta$ majority of their neighborhood. On the other hand, we need to leave the threshold large enough for our algorithm not to call the strong teacher too often.

Can both these requirement be met simultaneously? We believe that in "natural" situations, the first requirement is not an issue. In other words, we believe that "natural" distributions are highly locally conservative. For example, if our domain is the unit ball in $\mathbb{R}^n$, we assume that the labels are defined by a homogeneous halfspace and similarity corresponds to the Euclidean metric, then for any $\lambda \leq 1/2$, this distribution is $(\lambda, 0)$-locally conservative. This enables the learner to choose a relatively large value for the threshold η, so that the second requirement is met (allowing significant savings in the number of strong labels).

7 Concluding remarks

We view this paper as a first step in a practically relevant research direction that, as far as we can tell, has not received much attention terms of theoretical foundations. There are many directions in which this work can be extended. This paper focuses on a specific model of weak teachers, while it is not hard to come up with practical scenarios that call for a different modelling. Another natural direction for extending this research is the consideration of a hierarchy of teachers, and connecting this with budgeted learning.

Acknowledgements

We thank Russ Greiner for introducing us to the problem and Phil Long for insightful discussions on this work.

References

[1] Supplementary material. www.cs.uwaterloo.ca/~rurner/WTSuppl.pdf.

[2] Dana Angluin and Philip D. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1987.

[3] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[4] Nicolò Cesa-Bianchi, Eli Dichterman, Paul Fischer, Eli Shamir, and Hans-Ulrich Simon. Sample-efficient strategies for learning in the presence of noise. *J. ACM*, 46(5):684–719, 1999.

[5] Sanjoy Dasgupta. Two faces of active learning. *Theor. Comput. Sci.*, 412(19):1767–1781, 2011.

[6] Scott E. Decatur and Rosario Gennaro. On learning from noisy and incomplete examples. In *COLT*, pages 353–360, 1995.

[7] Vikas C. Raykar and Shipeng Yu. Ranking annotators for crowdsourced labeling tasks. In *NIPS*, pages 1809–1817. 2011.

[8] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Anna K. Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *ICML*, page 112, 2009.

[9] Shai Shalev-Shwartz. Introduction to machine learning, lecture notes, 2010. www.cs.huji.ac.il/~shais/Handouts.pdf.

[10] Victor S. Sheng, Foster J. Provost, and Panagiotis G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *KDD*, pages 614–622, 2008.

[11] Sudheendra Vijayanarasimhan and Kristen Grauman. Cost-sensitive active visual category learning. *International Journal of Computer Vision*, 91(1):24–44, 2011.

[12] Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. The multidimensional wisdom of crowds. In *NIPS*, pages 2424–2432, 2010.

[13] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*, pages 2035–2043, 2009.