Budgeted PAC Learning with Two Noisy Annotators

Dinesh Garg, S. Sundararajan, Sourangshu Bhattacharya DINESHG, SSRAJAN, SOURANGB@YAHOO-INC.COM
Yahoo! Labs, Bangalore, 560071 India

Shirish Shevade SHIRISH@CSA.IISC.ERNET.IN
Dept. of CSA, IISc, Bangalore, 560012 India

Abstract

We consider the problem of learning from noisy labeled examples where noisy labels are obtained from two annotators. Each annotator is characterized by its classification noise rate and annotation price per example. On the other hand, a learner has a fixed budget to meet the annotation cost. In this scenario, given a concept class and a learning algorithm, we derive theoretical bounds on the number of examples (labeled from each annotator), learning budget, and annotation price (given noise rate) from probably approximately correct (PAC) learning perspective. We also present a game theoretic view of the learner and annotators joint optimization problem and give a result on Nash equilibrium solution.

1. Background

We consider a typical binary classification problem in supervised setting, where we have a training set $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{M}$ containing M examples, with $x_i \in \mathscr{X}$ and $y_i \in \{0, 1\}$ and the goal is to build a classifier that generalizes well on unseen examples. In practice, the class labels y_i's are often acquired through a manual labeling process, where multiple annotators label the examples. It is natural to assume that the class labels y_i's thus obtained are noisy (due to human involvement). Further, the noise rate of the annotators can vary since the degrees of expertise and competence typically vary from person to person (Dekel & Shamir, 2009), (Raykar et al., 2009), (Sheng et al., 2008). The problem of learning from noisy labeled examples has been studied by several researchers from both theoretical (Angluin & Laird, 1988), (Bshouty et al., 2002), (Aslam & Decatur, 1996) and practical perspectives (Dekel & Shamir, 2009), (Raykar et al., 2009), (Sheng et al., 2008). Our motivation for this work comes from the following observations with regard to the characteristics of the problem and the related literature.

The scenario of multiple annotators is more prevalent in practice than a single annotator for several reasons. For example, in a web context, annotation tool (like Amazon's Mechanical Turk) can be simultaneously made available to several volunteer web users to acquire the labels cheaply, or, there are multiple vendors (competitors) available in the market who provide annotating resources at some specified prices. Manual annotation comes with varying degrees of noise and cost (charged by vendors). Also, the cost of getting high quality (accurate) labels will be typically higher compared to medium or low quality labels. Although a designer (learner) prefers high quality labels, he/she has a fixed learning budget to meet the annotation cost.

Although the problem of designing algorithms to build classifiers using noisy labeled examples from multiple annotators have been studied in the literature (Dekel & Shamir, 2009), (Raykar et al., 2009), (Sheng et al., 2008), these works do not study the problem from theoretical (Probably Approximately Correct (PAC) learning) perspective. From theoretical perspective, (Angluin & Laird, 1988) showed that one can still learn from noisy labeled examples under certain conditions and provided sample complexity bounds for PAC learning (Valiant, 1984), (Blumer et al., 1989). The results of (Angluin & Laird, 1988) revealed that as the noise rate of the annotator increases the number of labeled examples needed for PAC learning also increases. (Aslam & Decatur, 1996) presented a general lower bound on the number of examples required for PAC learning in the same noise modeling framework of (Angluin & Laird, 1988). However, all these results are applicable only in a *single annotator* scenario. From budgeted learning perspective, (Lizotte et al., 2003) and (Kapoor & Greiner, 2005) address the problem of building classifiers within budget constraint, where there is a cost associated with acquiring each feature value of the training examples. Unlike

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

these works, we are interested in the problem of budgeted PAC learning, where there is a fixed annotation cost for each example depending on the quality of annotators.

Based on these observations, this work addresses the problem of learning *with two noisy annotators* for the first time (to the best of our knowledge) from three perspectives, namely, PAC learning, Budgeted PAC learning, and a game theoretic view of learner's and annotators strategies. By introducing the notions of feasible and infeasible annotation plans, prices and learning budget, we derive theoretical bounds on the number of examples, annotation prices, and learning budget from PAC learning perspective. Finally, we present a Nash equilibrium (Myerson, 1997) result from a game theoretic viewpoint.

2. PAC Learning with Two Noisy Annotators

We begin by providing the basic definitions related to the PAC learning model with two noisy annotators followed by results on the feasibility of PAC learning. This model comprises of an *instance space* $\mathscr{X}$ and a *concept class* $\mathscr{C}$. The instance space $\mathscr{X}$ is a fixed set. It could be finite, countably infinite, $\{0,1\}^d$, or $\mathbb{R}^d$ (d-dimensional feature space), for some $d \geq 1$. The concept class $\mathscr{C}$ is a set of *concepts*, where a concept c is a subset of $\mathscr{X}$.

The task of the learner is to determine a close approximation to an unknown target concept c_t, from the labeled examples. We assume that $c_t \in \mathscr{C}$. The learner has access to two noisy annotators as the source of its training data. Each call to an annotator returns a labeled example $\langle x, y \rangle$, where example x is drawn randomly and independently according to some unknown (to the learner) sampling distribution D. The learner gets $m_1 \geq 0$ labeled examples from annotator 1 and $m_2 \geq 0$ labeled examples from annotator 2, which together constitute the training dataset. Finally, the learner employs a learning algorithm to output a hypothesis $h \in \mathscr{C}$, based on the training data.

The annotator $i, (i = 1, 2)$ reports the label y which is subject to an independent random mistake with a known probability η_i. So, the reported label is $y = \neg\, c_t(x)$ with probability η_i and $y = c_t(x)$ with probability $(1-\eta_i)$. This noise model known as *random classification noise* and was first studied by (Angluin & Laird, 1988) for the single noisy annotator case. The probabilities η_1 and η_2 are known as *noise rates* of the annotators 1 and 2, respectively. In this paper, we assume that $0 < \eta_1, \eta_2 < 1/2$.

2.1. PAC Bound, Annotation Plan, and MDA

For any hypothesis $h \in \mathscr{C}$, the error rate (or generalization error) is defined to be the probability that $h(x) \neq c_t(x)$ for an instance $x \in \mathscr{X}$ that is randomly drawn according to D. Then the error rate of a hypothesis h is given by $\mathbf{Pr}^D(c_t \Delta h)$, where $c_t \Delta h$ is the symmetric difference between c_t and h, and $\mathbf{Pr}^D(\cdot)$ is the probability of this event (calculated with respect to D). A hypothesis h is said to be ϵ-bad if its error rate is more than ϵ, i.e. $\mathbf{Pr}^D(c_t \Delta h) > \epsilon$. In the classical PAC model of (Valiant, 1984), the learner's goal is to come up with a learning algorithm which outputs an ϵ-bad hypothesis h with probability at most δ, where the probability is defined with respect to the distribution of training examples of a fixed size. Such a learning algorithm is known as PAC learning algorithm. In general, the error rate of the hypothesis chosen by a learning algorithm critically depends on the number of training examples supplied to the algorithm. Thus, a learning algorithm with single annotator is said to satisfy PAC bound with respect to the *sample size* $m(\epsilon, \delta)$ if the following condition holds true:

$$\mathbf{Pr}^{m(\epsilon,\delta)}(\mathbf{Pr}^D(c_t \Delta h) > \epsilon) < \delta \tag{1}$$

where, h is the hypothesis output by the learning algorithm when trained on the $m(\epsilon, \delta)$ training examples. The outer probability $\mathbf{Pr}^{m(\epsilon,\delta)}(\cdot)$ is taken over the distribution of $m(\epsilon, \delta)$ training examples (noisy or non-noisy). For a given PAC learning algorithm, the smallest sample size $m^*(\epsilon, \delta)$ for which this algorithm still satisfies PAC bound is known as its *sample complexity*.

In this paper, we extend the notion of PAC bound for the case of two noisy annotators. For this, we need the following definitions.

Definition 1 (Problem Instance) *An instance of the PAC learning problem is set of specifications of instance space $\mathscr{X}$, concept class $\mathscr{C}$, true concept c_t, and sampling distribution D.*

Definition 2 (Annotation Plan:) *A vector of non-negative integer valued sample sizes $(m_1(\epsilon, \delta), m_2(\epsilon, \delta))$ is called annotation plan, where each component corresponds to an annotator.*

Definition 3 (PAC Bound for Two Noisy Annotators case) *An annotation plan $(m_1(\epsilon, \delta), m_2(\epsilon, \delta))$ for two noisy annotators, along with a learning algorithm, satisfies PAC bound if the following condition holds true:*

$$\mathbf{Pr}^{(m_1(\epsilon,\delta), m_2(\epsilon,\delta))}(\mathbf{Pr}^D(c_t \Delta h) > \epsilon) < \delta \tag{2}$$

where h is the hypothesis chosen by the learning algorithm when trained using $m_1(\epsilon, \delta) + m_2(\epsilon, \delta)$ examples (obtained from annotator 1 and 2).[1] Note that the noise rates η_1, η_2 of the two annotators could be quite different. Hence the PAC bound depends not just on $m_1 + m_2$, but on the individual numbers m_1, m_2 also. Motivated by this, we define our notions of *feasible and infeasible annotation plans* below.

[1] In rest of the paper, the notations m_i and $m_i(\epsilon, \delta), i = 1, 2$ denote number of samples received from annotator i.

Algorithm 1 Minimum Disagreement Algorithm

Input: m_1 and m_2 examples from annotators 1 and 2.
Output: A hypothesis $h^* \in \mathscr{C}$
Algorithm: Let $\{\langle x_j^i, y_j^i\rangle \;\; i = 1,2; \;\; j = 1,\ldots,m_i\}$ be the training data supplied by annotator i in j^{th} call. For the given training dataset, choose a hypothesis h^* that minimizes the number of examples for which h^* disagrees with the given labels. That is,

$$h^* \in \arg\min_{h\in\mathscr{C}} |\{x_{ij}|h(x_{ij}) \neq y_{ij}; i = 1,2; j = 1,\ldots,m_i\}|$$

Definition 4 (Feasible Annotation Plans) *An annotation plan $(m_1(\epsilon,\delta), m_2(\epsilon,\delta))$ is said to be feasible for a given learning algorithm, if the learning algorithm satisfies PAC bound (2), for every instance of the problem when training data is supplied as per the plan.*

Definition 5 (Infeasible Annotation Plans) *An annotation plan $(m_1(\epsilon,\delta), m_2(\epsilon,\delta))$ is said to be infeasible for a given learning algorithm, if the algorithm fails to satisfy PAC bound (2) for at least one instance of the problem when training data is supplied as per the plan.*

Here we analyze *Minimum Disagreement Algorithm (MDA)* (Algorithm 1), which is simple, intuitive and also analyzed in (Laird, 1988), for feasibility and infeasibility of PAC learning in the two noisy annotators case[2].

2.2. Characterization of Feasible Annotation Plans

We now give a characterization of the feasible annotation plans for MDA. In order to derive such a characterization, we define a few events and their corresponding probabilities, assuming a finite concept class $\mathscr{C}$ having $|\mathscr{C}| = N$. The events are defined for an annotation plan (m_1, m_2) and a hypothesis h. We assume that samples of size m_1 and m_2 are drawn randomly and independently, according to the distribution D by the annotators 1 and 2, respectively, and then labels are flipped independently with noise rates η_1 and η_2. We also assume $h \in \mathscr{C}$. Let $L_e(h)$ be the empirical error (number of disagreements) for hypothesis h. The events E_1, E_2, E_3, and E_4, of our interest are defined as:

- $E_1(h, m_1, m_2)$: The empirical error of a given hypothesis $h \in \mathscr{C}$ is no more than the empirical error of the true hypothesis c_t, i.e. $L_e(h) \leq L_e(c_t)$.
- $E_2(h, m_1, m_2)$: The empirical error of a given hypothesis $h \in \mathscr{C}$ is the minimum across all hypotheses in the class $\mathscr{C}$, i.e. $L_e(h) \leq L_e(h')\; \forall h' \in \mathscr{C}$
- $E_3(h, m_1, m_2)$: MDA outputs a given hypothesis h.
- $E_4(\epsilon, m_1, m_2)$: MDA outputs an ϵ-bad hypothesis.

The probability of events $E_i, i = 1, 2, 3$ and E_4 are denoted by $\mathbf{Pr}^{(m_1,m_2)}[E_i(h)]$ and $\mathbf{Pr}^{(m_1,m_2)}[E_4(\epsilon)]$,respectively. We have the following useful lemmas in view of the above definitions.

Lemma 1 *Given a concept class $\mathscr{C}$ such that $c_t \in \mathscr{C}$ and $N = |\mathscr{C}|$, the following hold true for any given annotation plan (m_1, m_2).*

$$\mathbf{Pr}^{(m_1,m_2)}[E_4(\epsilon)] \leq (N-1)\left[\max_{h\in\mathscr{C}, h \text{ is } \epsilon\text{-bad}} \mathbf{Pr}^{(m_1,m_2)}[E_1(h)]\right]$$

Proof: By definition of the events, for any hypothesis $h \in \mathscr{C}$ that is ϵ-bad (for any $\epsilon > 0$), we have $E_3(h, m_1, m_2) \subseteq E_2(h, m_1, m_2) \subseteq E_1(h, m_1, m_2)$. Also, $E_4(\epsilon, m_1, m_2) = \bigcup_{h\in\mathscr{C}; h \text{ is } \epsilon\text{-bad}} E_3(h, m_1, m_2)$.

Taking probabilities of the events, we get $\mathbf{Pr}^{(m_1,m_2)}[E_3(h)] \leq \mathbf{Pr}^{(m_1,m_2)}[E_1(h)]$ and $\mathbf{Pr}^{(m_1,m_2)}[E_4(\epsilon)] \leq \sum_{h\in\mathscr{C}, h \text{ is } \epsilon\text{-bad}} \mathbf{Pr}^{(m_1,m_2)}[E_3(h)]$. The lemma follows from these two inequalities. *Q.E.D.*

Lemma 2 *Consider an instance of PAC learning problem where all $(N-1)$ concepts in $\mathscr{C}$ (except true concept c_t) are ϵ-bad. If MDA uses a hostile tie breaking rule (tie is broken by choosing a hypothesis whose error rate is highest) then the following holds true for any given annotation plan (m_1, m_2).*

$$\left[\max_{h\in\mathscr{C}, h \text{ is } \epsilon\text{-bad}} \mathbf{Pr}^{(m_1,m_2)}[E_1(h)]\right] \leq \mathbf{Pr}^{(m_1,m_2)}[E_4(\epsilon)]$$

Proof: Let $h \in \mathscr{C}$ be an ϵ-bad hypothesis. For the above concept class, MDA will output an ϵ-bad hypothesis for any training dataset where empirical error of this h is no more than the empirical error of true hypothesis c_t. Thus, for this h, we have $E_1(h, m_1, m_2) \subseteq E_4(\epsilon, m_1, m_2)$ for any annotation plan (m_1, m_2). Taking probabilities on both sides, followed by taking max over $h \in \mathscr{C}, h$ is ϵ-bad, proves the claim. *Q.E.D.*

Lemma 2 shows the existence of a problem instance (concept class) which does not satisfy PAC bound under one particular tie breaking rule. Finding other scenarios for which similar result holds is under progress. Next, we state our first result that characterizes the feasible annotation plans for the MDA.

Theorem 1 *Consider the PAC learning model with two noisy annotators and the MDA (Algorithm 1). Let $N = |\mathscr{C}| < \infty$. Then, for any given $0 < \epsilon, \delta < 1$ and $0 < \eta_1, \eta_2 < 1/3$, if m_1 and m_2 satisfy the following inequality then the MDA will satisfy PAC bound.*

$$\log(N/\delta) \leq m_1\psi_1 + m_2\psi_2 \tag{3}$$

[2]Note that there could be ties in terms of output choice for the MDA. However, MDA is free to choose any tie breaking rule.

where $\forall i = 1, 2$, we have

$$\psi_i = \log\left[1 - \epsilon\left(1 - \exp\left(-(1 - 3\eta_i)/8\right)\right)\right]^{-1} \tag{4}$$

Thus, the inequality (3) characterizes the set of feasible annotation plans.

Remark: This characterization is independent of the problem instance and tie breaking rule of the MDA.

Proof: MDA satisfies PAC bound iff there exists an annotation plan $(m_1(\epsilon,\delta), m_2(\epsilon,\delta))$ such that $\mathbf{Pr}^{(m_1(\epsilon,\delta),m_2(\epsilon,\delta))}[E_4(\epsilon)] < \delta$. From Lemma 1, it can be seen that for any $0 < \epsilon, \delta < 1$, if an annotation plan (m_1, m_2) satisfies the following condition, then MDA will satisfy PAC bound.

$$\left[\max_{h \text{ is } \epsilon\text{-bad}} \mathbf{Pr}^{(m_1,m_2)}[E_1(h)]\right] \leq \delta/N \tag{5}$$

LHS in above expression is an upper bound for the RHS of the expression in Lemma 1 (excluding $N-1$), because now h may not belong to $\mathscr{C}$. This is done to make the bound independent of the problem instance, although MDA will only output $h \in \mathscr{C}$.

Now, we upper bound the LHS of (5). To do this, we derive an upper bound for $\mathbf{Pr}^{(m_1,m_2)}[E_1(h)]$ where hypothesis h has an error rate of ϵ (for any $\epsilon \in (0,1)$). Later, in expression (9), we show that this bound is indeed an upper bound for LHS in (5). To derive this bound, note that for any random and independent sample (x, y) that is delivered by annotator 1 or 2, the probability of its agreeing (or disagreeing) with a true hypothesis c_t and a hypothesis h (having error rate ϵ) is given by a probability tree as shown in Figure 1. In view of this tree, it is easy to verify that

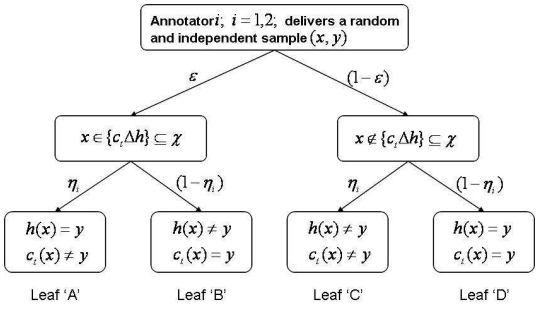

Figure 1. Probabilities of hypothesis c_t (and h) agreeing and disagreeing with a random sample.

the quantity $\mathbf{Pr}^{(m_1,m_2)}[E_1(h)]$ can be written as follows: *Probability that the number of samples (out of m_1, m_2) that fall under leaf B in the probability tree is at most the number of samples that fall under leaf A, i.e h disagrees on no more number of examples than c_t.*

In order to compute the above quantity, we need to first compute the following conditional probability: *If k_i examples from annotator i ($0 \leq k_i \leq m_i$), $i = 1, 2$ come from the set $(c_t \Delta h) \subseteq \mathscr{X}$, then what is the probability that out of these (k_1+k_2) examples, h disagrees on no more number of examples than c_t? That is, empirical error of h, denoted by $L_e(h)$, is at most the empirical error of c_t, denoted by $L_e(c_t)$.* This conditional probability can be given as follows (assuming without loss of generality that $k_1 \leq k_2$)

$$\mathbf{Pr}^{(k_1,k_2)}(L_e(h) \leq L_e(c_t)) = \sum_{i=0}^{k_1} \binom{k_1}{i}(1-\eta_1)^i(\eta_1)^{k_1-i} \sum_{j=0}^{\lfloor (k_1+k_2)/2 \rfloor - i} \binom{k_2}{j}(1-\eta_2)^j(\eta_2)^{k_2-j} \tag{6}$$

If we let Z^1 and Z^2 be two Bernoulli random variables with mean $1-\eta_1$ and $1-\eta_2$, respectively then the above expression gives the probability that sum of k_1 sample of Z^1 and k_2 samples of Z^2 is no more than $(k_1+k_2)/2$. If we write $Z = \sum_{i=1}^{k_1} Z_i^1 + \sum_{j=1}^{k_2} Z_j^2$ and $(1-\nu)\mu = \lfloor (k_1+k_2)/2 \rfloor$ where $\mu = k_1(1-\eta_1) + k_2(1-\eta_2)$ (mean of Z) then we can upper bound the above expression by applying the multiplicative form of Chernoff bound (see e.g. Theorem 4.2 in (Motwani & Raghavan, 1995)), which says $P[Z \geq (1-\nu)\mu] \leq \exp(-\mu\nu^2/2)$. This gives:

$$\mathbf{Pr}^{(k_1,k_2)}(L_e(h) \leq L_e(c_t)) \leq \exp\left(-\frac{1}{8}\frac{[k_1(1-2\eta_1) + k_2(1-2\eta_2)]^2}{[k_1(1-\eta_1) + k_2(1-\eta_2)]}\right) \tag{7}$$

We loosen this bound to get a separable form in k_1 and k_2. Assuming $0 < \eta_1, \eta_2 < 1/3$, the loosened bound becomes:

$$\exp\left(-[k_1(1-3\eta_1) + k_2(1-3\eta_2)]/8\right) \tag{8}$$

Summing up the above conditional probability bound over all possible values of k_1 and k_2, the total probability $\mathbf{Pr}^{(m_1,m_2)}[E_1(h)]$ becomes:

$$\sum_{k_1=0}^{m_1}\sum_{k_2=0}^{m_2}\binom{m_1}{k_1}\binom{m_2}{k_2}\epsilon^{k_1+k_2}(1-\epsilon)^{m_1+m_2-k_1-k_2} \mathbf{Pr}^{(k_1,k_2)}(L_e(h) \leq L_e(c_t))$$

Using the bound in (8), we get the following upper bound on $\mathbf{Pr}^{(m_1,m_2)}[E_1(h)]$.

$$\sum_{k_1=0}^{m_1}\sum_{k_2=0}^{m_2}\binom{m_1}{k_1}\binom{m_2}{k_2}\epsilon^{k_1+k_2}(1-\epsilon)^{m_1+m_2-k_1-k_2} \exp\left(-[k_1(1-3\eta_1) + k_2(1-3\eta_2)]/8\right)$$

Using the moment generating function of the Binomial distribution, above expression can be written as:

$$\mathbf{Pr}^{(m_1,m_2)}[E_1(h)] \leq \left[1 - \epsilon\left(1 - \exp\left(-(1-3\eta_1)/8\right)\right)\right]^{m_1} \left[1 - \epsilon\left(1 - \exp\left(-(1-3\eta_2)/8\right)\right)\right]^{m_2} \tag{9}$$

In above bound, h has an error rate exactly equal to ϵ. One can see that above bound is valid for an ϵ-bad hypothesis also because the expression decrease as ϵ increases. Substituting this upper bound on the LHS of (5), we get the desired claim. *Q.E.D.*

2.3. Characterization of Infeasible Annotation Plan

Next, we present a characterization of infeasible annotation plans for the MDA.

Theorem 2 *Consider the same setting as described in Theorem 1. Then, for any given $0 < \epsilon < 1, 0 < \delta < 1/4$, and $0 < \eta_1, \eta_2 < 1/2$, if (m_1, m_2) satisfy the following inequality then there exist problem instance for which the MDA will* **fail** *to satisfy PAC bound under the hostile tie breaking rule.*

$$m_1\theta_1 + m_2\theta_2 \leq \log(1/4\delta) \tag{10}$$

where $\forall i = 1, 2$ *we have*

$$\theta_i = \log\left[1 - \epsilon\left(1 - 2\eta_i\right)\right]^{-1} \tag{11}$$

Thus, the inequality (10) characterizes the set of infeasible annotation plans for the MDA.

Proof: Note that the MDA satisfies PAC bound iff there exists an annotation plan $(m_1(\epsilon,\delta), m_2(\epsilon,\delta))$ such that the following holds true: $\mathbf{Pr}^{(m_1(\epsilon,\delta),m_2(\epsilon,\delta))}[E_4(\epsilon)] < \delta$. By invoking Lemma 2 here, we can say that for any $0 < \epsilon < 1$, $0 < \delta < 1/4$, and any ϵ-bad hypothesis $h \in \mathscr{C}$, if the annotation plan (m_1, m_2) satisfies the following condition then for that plan, there exist an instance of the problem (defined in Lemma 2) for which the MDA will **fail** to satisfy the PAC bound under hostile tie breaking rule.

$$\delta \leq \mathbf{Pr}^{(m_1,m_2)}[E_1(h)] \tag{12}$$

The idea behind the rest of the proof is to lower bound the RHS of the above inequality. Recall the definition of the RHS of the above inequality from the previous section. The RHS is the probability that the empirical error of an ϵ-bad hypothesis is at most the empirical error of the true hypothesis for the annotation plan of (m_1, m_2). In order to lower bound this probability, we will first lower bound the conditional probability $\mathbf{Pr}^{(k_1,k_2)}(L_e(h) \leq L_e(c_t))$ which is given by Equation (6). If we assume that $0 \leq k_1 \leq k_2$ (there is no loss of generality in this) then the following lower bound would follow trivially from Equation (6) due to the fact that $(1/2 > \eta_1, \eta_2 > 0)$.

$$\begin{aligned}
&\mathbf{Pr}^{(k_1,k_2)}(L_e(h) \leq L_e(c_t)) \geq \\
&\eta_1^{k_1}\eta_2^{k_2}\sum_{i=0}^{k_1}\binom{k_1}{i}\sum_{j=0}^{\lfloor\frac{(k_1+k_2)}{2}\rfloor - i}\binom{k_2}{j} \\
&= \eta_1^{k_1}\eta_2^{k_2}\left[\sum_{i=0}^{k_1}\binom{k_1}{i}\sum_{j=0}^{\lfloor(k_2-k_1)/2\rfloor}\binom{k_2}{j} + \right. \\
&\left.\sum_{i=0}^{k_1}\binom{k_1}{i}\sum_{j=\lfloor(k_2-k_1)/2\rfloor+1}^{\lfloor(k_1+k_2)/2\rfloor - i}\binom{k_2}{j}\right]
\end{aligned} \tag{13}$$

For $i = k_1$ and $k_1 = k_2$ the second term in the above expression is 0. One can see that the second term inside the brackets in the above inequality is bounded below by:

$$\frac{1}{2}\sum_{i=0}^{k_1}\binom{k_1}{i}\sum_{j=\lfloor(k_2-k_1)/2\rfloor+1}^{\lfloor k_2/2\rfloor}\binom{k_2}{j} \tag{14}$$

Substituting this value of (14) in (13), we get

$$\begin{aligned}
\mathbf{Pr}^{(k_1,k_2)}(L_e(h) \leq L_e(c_t)) &\geq \eta_1^{k_1}\eta_2^{k_2}\tfrac{2^{k_1}}{2}\tfrac{2^{k_2}}{2} \\
&= \tfrac{1}{4}\exp(k_1\log(2\eta_1) + k_2\log(2\eta_2))
\end{aligned} \tag{15}$$

Summing the above conditional probability over all values of k_1 and k_2, we get the total probability which is the RHS of (12). Thus,

$$\begin{aligned}
RHS \geq \tfrac{1}{4}\textstyle\sum_{k_1=0}^{m_1}\sum_{k_2=0}^{m_2}\epsilon^{k_1+k_2}(1-\epsilon)^{m_1+m_2-k_1-k_2} \\
\exp(k_1\log(2\eta_1) + k_2\log(2\eta_2))
\end{aligned} \tag{16}$$

Using the moment generating function of the Binomial distribution, the above expression can be written as

$$RHS \geq \frac{1}{4}\left[1 - \epsilon\left(1 - 2\eta_1\right)\right]^{m_1}\left[1 - \epsilon\left(1 - 2\eta_2\right)\right]^{m_2} \tag{17}$$

By substituting this lower bound on RHS in the inequality (12), we get the desired claim. *Q.E.D.*

2.4. Single Noisy Annotator Case and Comparison with Existing Bounds

Note that if we let $m_1 = m_2 = m/2$ and $\eta_1 = \eta_2 = \eta$ then the proposed model would reduce to the single noisy annotator model and for such a case, the set of feasible (infeasible) annotation plans would correspond to the upper (lower) bounds on the sample complexity. Theorem 1 and Theorem 2 would give us the upper bound and the lower bound, respectively, as follows:

$$\log(N/\delta) \leq m\log\left[1 - \epsilon\left(1 - \exp\left(-\tfrac{1}{8}(1-3\eta)\right)\right)\right]^{-1}$$
$$\log(1/4\delta) \geq m\log\left[1 - \epsilon\left(1 - 2\eta\right)\right]^{-1}$$

Philip Laird (Laird, 1988) had proposed the following bounds on the sample complexity for the same scenario.

$$\log(N/\delta) \leq m\log\left[1 - \epsilon\left(1 - \exp\left(-\tfrac{1}{2}(1-2\eta)^2\right)\right)\right]^{-1}$$
$$\log(1/2\delta) \geq m\log\left[1 - \epsilon\left(1 - 2\eta\right)\right]^{-1}$$

Recall that our upper bound is valid for $\eta < 1/3$ and all other bounds are valid for $\eta < 1/2$. We compare our bounds with the bounds proposed in (Laird, 1988). The comparison is shown in Figure 2. It is clear from Figure 2 that both our upper bound and our lower bound are loose as compared to the bounds given in (Laird, 1988). The reason behind our upper bound being loose is inequality (8),

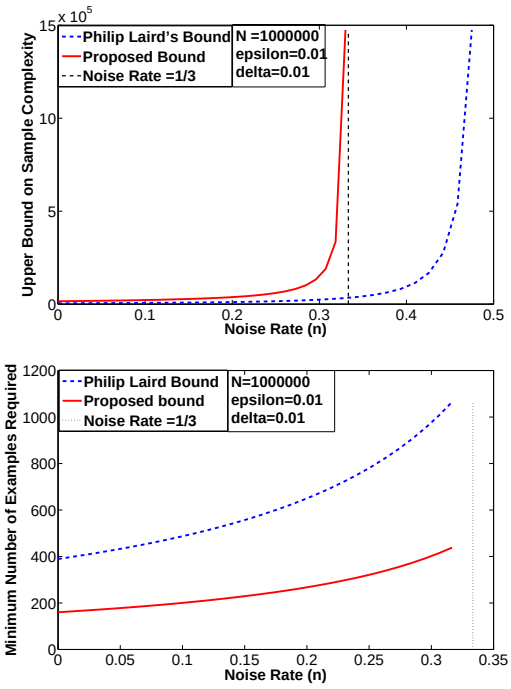

Figure 2. Comparison Sample Complexity Bounds

where we loosened the upper bound in order to convert it to a separable form. Also, because of the same step, our upper bound is valid only when $\eta < 1/3$ as opposed to $\eta < 1/2$ in (Laird, 1988). Nevertheless, it is important to note that for $\eta < 1/4$, our upper bound is comparable with the upper bound proposed in (Laird, 1988) and behavior of our bound closer to $\eta = 1/3$ is very much similar to the behavior of the other upper bound near $\eta = 1/2$. Similarly, the reason behind our lower bound being loose is inequality (14), where we loosen the lower bound in order to convert it to a separable form. Thus, we can say that if one can improve the bounds given in inequality (8) and/or (14) then it will improve our characterization of the feasible and/or infeasible annotation plans and hence the upper bound and/or lower bound on the sample complexity.

3. Budgeted PAC Learning

In this section, we turn our attention towards the learner's budget constraint and annotators' price constraints. We assume that the learner is constrained by a fixed *learning budget* B available with him and annotator $i, (i = 1, 2)$, charges a fixed *annotation price* π_i for providing every single labeled example. For such a setting, we define the notion of *feasible and infeasible learning budget* and *feasible and infeasible annotation prices* for any given algorithm as follows.

Definition 6 (Feasible and Infeasible Learning Budget) *For a given learning algorithm, parameters* $\epsilon, \delta, \eta_1, \eta_2$ $(0 < \epsilon, \delta < 1,\ 0 < \eta_1, \eta_2 < 1)$*, and annotation prices* (π_1, π_2)*, we say that the learning budget* B *is feasible if it is feasible for the learner to buy a feasible annotation plan at the given annotation prices. Similarly, we say that a learning budget* B *is infeasible if any annotation plan that can be bought by the learner in this budget is an infeasible annotation plan.*

Definition 7 (Feasible and Infeasible Annotation Prices) *For a given learning algorithm, parameters* $\epsilon, \delta, \eta_1, \eta_2$ $(0 < \epsilon, \delta < 1,\ 0 < \eta_1, \eta_2 < 1)$*, and learning budget* B*, we say that the given annotation prices* (π_1, π_2) *are feasible (infeasible) if the given learning budget* B *is feasible (infeasible) with respect to these prices.*

In what follows, we study the feasibility of MDA being a PAC learning algorithm in the presence of the finite learning budget B and the annotation prices (π_1, π_2). We present the characterization of the feasible and infeasible learning budget as well as the feasible and infeasible annotation prices for MDA.

3.1. Feasibility of Learning Budget

Theorem 3 *Let* (ψ_1, ψ_2) *and* (θ_1, θ_2) *are as defined in Theorem 1 and Theorem 2, respectively. Then, for a given values of parameters* $0 < \epsilon < 1, 0 < \delta < 1/4, 0 < \eta_1, \eta_2 < 1/3$*, and a given set of annotation prices* (π_1, π_2)*, the following holds true.*
(1) A learning budget B *is a feasible budget if* $\left[\log\left(\frac{N}{\delta}\right) \min\left(\frac{\pi_1}{\psi_1}, \frac{\pi_2}{\psi_2}\right)\right] \leq B$
(2) A learning budget B *is an infeasible budget if* $B \leq \left[\log\left(\frac{1}{4\delta}\right) \min\left(\frac{\pi_1}{\theta_1}, \frac{\pi_2}{\theta_2}\right)\right]$

Proof: In Figure 3, we have plotted the region of feasible and infeasible annotation plans for the MDA (as derived in Theorem 1 and Theorem 2, respectively) on a two dimensional plane (for a fixed values of parameters ϵ, δ, η_1,and η_2). Each point on this plane corresponds to one possible annotation plan (m_1, m_2). It is easy to verify that for any annotation plan that lies in the upper shaded region of this graph, the MDA satisfies PAC bound. On the other hand, for any annotation plan that lies in the lower shaded region of this graph, the MDA fails to satisfy PAC bound for some instance of the problem. Also, note that the boundary line of the feasible region would always lie above the boundary line of the infeasible region. This is due to the fact that $\frac{\log(1/4\delta)}{\theta_i} \leq \frac{\log(N/\delta)}{\psi_i}\ \forall i = 1, 2$ which can be verified by plugging in the expressions of ψ_i and θ_i. Now consider a budget constraint line $m_1\pi_1 + m_2\pi_2 = \alpha$ on the same plane for some fixed value of α. It is easy to see that as α increases from zero then depending upon whether $-\frac{\pi_1}{\pi_2} > (< or =) - \frac{\psi_1}{\psi_2}$, this line would become a supporting hyperplane of the feasible region at point A (or B or

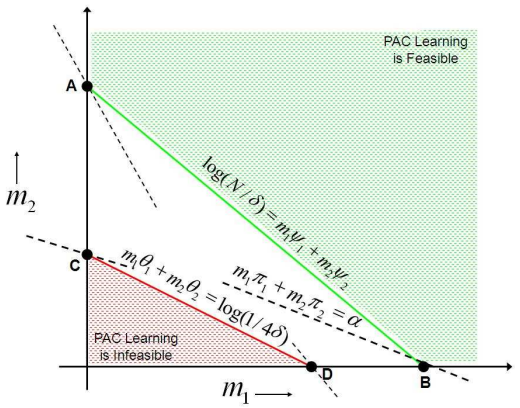

Figure 3. Idea Behind Deriving Bounds on the Learning Budget

both). The α corresponding to that line will serve as a lower bound on the feasible learning budget. It is easy to verify that such an α would be equal to $\left\lceil \log\left(\frac{N}{\delta}\right) \min\left(\frac{\pi_1}{\psi_1}, \frac{\pi_2}{\psi_2}\right)\right\rceil$. This proves the lower bound on the feasible learning budget. For proving the upper bound bound on the infeasible learning budget, we apply the same trick with the boundary line of the infeasible region. *Q.E.D.*

3.2. Feasibility of Annotation Prices

Theorem 4 *For a given set of parameters, $0 < \epsilon < 1$, $0 < \delta < 1/4$, $0 < \eta_1, \eta_2 < 1/3$, and a given learning budget $B > 0$, an annotation price vector (π_1, π_2) is*
(1) a feasible price vector if at least one of the price satisfies the following condition: $\pi_i \leq \frac{B\psi_i}{\log(N/\delta)},\ i = 1, 2$
(2) an infeasible price vector if it satisfies the following condition: $\pi_i \geq \frac{B\theta_i}{\log(1/4\delta)}\ \forall i = 1, 2$

Proof: This follows from Theorem 3 and the definitions of the feasible and infeasible annotation price vectors. *Q.E.D.*

It is interesting to plot the above result on a two dimensional plane of annotation prices π_1 and π_2 as shown in Figure 4. The intuition behind having an unbounded 'L' shaped feasible region is as follows: As long as the price of annotator 1 is within the range $\left[0, \frac{B\psi_1}{\log(N/\delta)}\right]$, then the price of the annotator 2 is immaterial because at this price the learner can always buy a feasible annotation plan $(m_1, 0)$. The same logic applies if we switch the roles of the annotators. Also, note that unbounded rectangular infeasible region always lies at the top-right side of the unbounded 'L' shaped feasible region because under the conditions of Theorem 4, we always have $\frac{B\psi_i}{\log(N/\delta)} \leq \frac{B\theta_i}{\log(1/4\delta)}\ \forall i = 1, 2$.

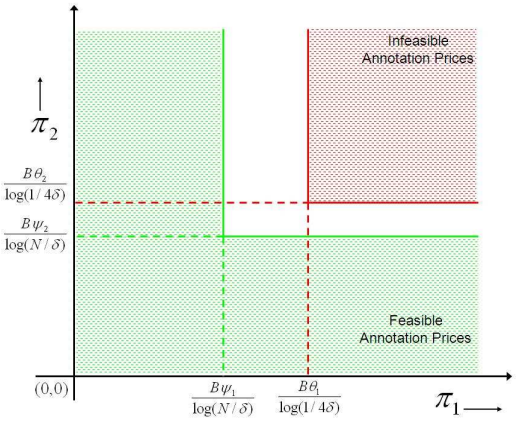

Figure 4. Idea Behind Deriving Bounds on the Annotation Prices

4. A Game Theoretic View of PAC Learning

In this section, we view the learning problem as a simultaneous decision making problem where the learner is trying to buy annotation services and the annotators are selling the services. This problem faced by the learner and the annotators can be modeled by means of a 3-player, non-cooperative, complete information game (Myerson, 1997). We assume that the learning budget B, the parameters ϵ and δ, and noise rates η_1 and η_2 are common knowledge among all the 3 players. In this 3-player game, all the players need to choose their strategy in a simultaneous and non-cooperative manner in order to maximize their individual utilities (defined below). The strategy space of the learner is all possible annotation plans and the strategy space of an annotator is all possible annotation prices.

Learner's Problem: Minimize $(m_1 + m_2)$ subject to $m_1\pi_1 + m_2\pi_2 \leq B$, $\log(N/\delta) \leq m_1\psi_1 + m_2\psi_2$, and $m_1, m_2 \geq 0$. Here, the learner wants to minimize the computational cost of MDA (typically, proportional to some polynomial in the number of examples) subject to the constraints with PAC learning and budget.

Annotator 1's Problem: Maximize $(m_1\pi_1)$ subject to the constraints that $0 \leq \pi_1$.

The problem of annotator 2 can be given in a similar manner. Here, each annotator aims at maximizing his revenue and not the profit, which is revenue minus cost. The formulation for profit maximization would be quite different and many other issues would also creep in. We are keeping off these considerations from the scope of this paper. Note that the three decision problems are inter-wined in the sense that the variables m_1 and m_2 that appear in the annotators' problems are controlled by the learner and the prices π_1 and π_2 that appear in the learner's problem are controlled by the annotators.

Definition 8 (A Nash Equilibrium of the 3-Player Game) *We say that the annotation plan (m_1^*, m_2^*) and the price vector (π_1^*, π_2^*) form a Nash equilibrium of the above discussed 3-player game if (m_1^*, m_2^*) solves the learner's problem (whenever $\pi_i = \pi_i^*,\ i = 1, 2$) and $\pi_i^*,\ i = 1, 2$ solves the annotator $i's$ problem (whenever*

$m_i = m_i^*,\ \forall i = 1, 2$*).*

The next theorem characterizes a Nash equilibrium of the above game.

Theorem 5 (Characterization of Nash Equilibrium) *For given values of parameters,* $0 < \epsilon, \delta < 1$, $0 < \eta_1, \eta_2 < 1/3$, $\eta_1 \neq \eta_2$, $0 < B$, *and* $1 < N$, *the 3-player game has (infinitely) many Nash equilibriums. Any Nash equilibrium* $(m_1^*, m_2^*, \pi_1^*, \pi_2^*)$ *of this game is of the following form:*

$$\pi_i^* = \begin{cases} B\psi_i/\log(N/\delta) & : \textit{If } \eta_i = \min(\eta_1, \eta_2) \\ \textit{any non -ve number} & : \textit{otherwise} \end{cases}$$

$$m_i^* = \begin{cases} 0 & : \textit{If } \eta_i = \max(\eta_1, \eta_2) \\ \log(N/\delta)/\psi_i & : \textit{otherwise} \end{cases}$$

Remark: If $\eta_1 = \eta_2 = \eta$ (i.e. $\psi_1 = \psi_2 = \psi$) then also the 3-player game has (infinitely) many Nash equilibriums. Any Nash equilibrium of this game is given by: $\pi_1^* = \pi_2^* = B\psi/\log(N/\delta)$, $m_1^* + m_2^* = \log(N/\delta)/\psi$, and $m_1^*, m_2^* \geq 0$. The proof given below is for the case when $\eta_1 \neq \eta_2$.

Proof: Note that the maximum possible values of the objective functions for either of the annotator is B due to learner's budget constraint. For the learner, in the absence of budget constraint, the minimum possible value of the objective function is $[\log(N/\delta)]/\max(\psi_1, \psi_2)$. Therefore, it is strongly dominant strategy (Myerson, 1997) for the annotator i, whose noise rate is lower (i.e. ψ_i is maximum), to announce a price $\pi_i = B\psi_i/\log(N/\delta)$. At this price, the learner can attain the minimum possible value of the objective function without violating the budget constraint and it will also result in this annotator's revenue being B. Setting a price higher or lower than this would result in lower revenue for this annotator. The price of the other annotator is immaterial because the learner can never attain the minimum possible value of his objective function if he uses examples of this annotator. *Q.E.D.*

5. Conclusion

We extended PAC learning results (under the random class noise model) for two noisy annotators case. Based on these results, we addressed the problem of budgeted PAC learning, and obtained a result that has an interesting economic interpretation from a game theoretic viewpoint. Future works include (1) relaxing the assumption of finite concept class by making use of Vapnik-Chervonenkis dimension (Blumer et al., 1989), (2) improving the quality of bounds, and (3) extending to multiple annotators (more than 2) case.

Acknowledgement

We are thankful to Chiranjib Bhattacharyya for helpful discussions and reviewers for useful comments.

References

Angluin, Dana and Laird, Philip. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Aslam, Javed A. and Decatur, Scott E. On the sample complexity of noise-tolerant learning. *Information Processing Letters*, 57(4):189–195, 1996.

Blumer, A., Ehrenfeucht, Andrzej, Haussler, David, and Warmuth, Manfred K. Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

Bshouty, Nader H., Eiron, Nadav, and Kushilevitz, Eyal. PAC learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002.

Dekel, Ofer and Shamir, Ohad. Good learners for evil teachers. In *ICML*, 2009.

Kapoor, Aloak and Greiner, Russell. Learning and classifying under hard budgets. In *ECML*, 2005.

Laird, Philip D. *Learning from good and bad data*. Kluwer Academic Publishers, Norwell, MA, USA, 1988.

Lizotte, Dan, Madani, Omid, and Greiner, Russell. Budgeted learning of naive-bayes classifiers. In *UAI*, 2003.

Motwani, Rajeev and Raghavan, Prabhakar. *Randomized Algorithms*. Cambridge University Press, NY, USA, 1995.

Myerson, R. B. *Game Theory: Analysis of Conflict*. Harvard University Press, Cambridge, Massachusetts, 1997.

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

Sheng, Victor S., Provost, Foster, and Ipeirotis, Panagiotis G. Get another label? improving data quality and data mining using multiple, noisy labelers. In *KDD*, 2008.

Valiant, L.G. A theory of learnable. *Communications of the ACM*, 27:1134–1142, 1984.