Approximately Correct Learning under Adversarial Design

Daniel Berend
Dept. of Math. and Dept. of Comp. Sci.
Ben-Gurion University
Beer Sheva 84105, Israel
berend@cs.bgu.ac.il

Aryeh Kontorovich
Dept. of Comp. Sci.
Ben-Gurion University
Beer Sheva 84105, Israel
karyeh@cs.bgu.ac.il

Lev Reyzin
Dept. of Math., Stats., & Comp. Sci.
University of Illinois at Chicago
Chicago, IL 60607, USA
lreyzin@math.uic.edu

Abstract

We propose a new learning model, which attempts to capture situations where the training data is sufficiently informative so as to pick out only very few potential candidate hypotheses. In such cases, the training points need not be independent and identically distributed — or indeed, be drawn from any distribution. Instead, we derive accuracy guarantees from the geometric configuration of the training data that happened to be provided to the learner. Our model allows for white label noise and in general has a connection to noisy PAC. In at least one instance, we obtain an improvement of known noisy PAC results. We discuss the case of learning thresholds in detail, and show how the analysis extends to intervals and rectangles in $\mathbb{R}^d$. An unusual feature of our analysis in the learning-theoretic setting is the appearance of random walks. Challenging open problems are posed.

1 Introduction

The celebrated Probably Approximately Correct (PAC) model (Valiant, 1984) has been enormously influential in setting the learning-theoretic agenda over the past thirty years. Indeed, this model has laid the foundation for a clean and elegant theory while retaining some measure of empirical plausibility. Regarding the latter criterion, numerous results have aimed at whittling away at the model's initially somewhat restrictive formulation. The original requirement of clean labels was relaxed to encompass a benign type of label noise (Angluin and Laird, 1987; Kearns, 1998), as well as considerably more adversarial noise models (Kearns and Schapire, 1994; Kearns et al., 1994). Similarly, the i.i.d. sampling assumption — which early learning theory papers often took pains to apologize for — has by now been subsumed by far less restrictive mixing conditions (Gamarnik, 2003; Karandikar and Vidyasagar, 2002; London et al., 2012, 2013; Mohri and Rostamizadeh, 2008, 2010; Rostamizadeh and Mohri, 2007; Shalizi and Kontorovich, 2013; Steinwart and Christmann, 2009; Steinwart et al., 2009; Zou et al., 2014). In the *online* learning model (Cesa-Bianchi and Lugosi, 2006), one dispenses with a sampling distribution entirely, and instead assumes an adversarially chosen sequence of labeled examples. Due to this model's worst-case nature, one can only prove *regret* bounds as opposed to absolute error estimates.

In this paper, we propose a distribution-less variant of learning. Unlike the online setting, training data is provided in batch and we use its structure to draw conclusions about the range of possible target hypotheses. In this sense, our learning model conceptually resembles the so-called "algorithmic luckiness" framework (Shawe-Taylor et al., 1998; Herbrich and Williamson, 2002), where the generalization bound depends on the empirical configuration of the training sample (such as it having a large margin). The salient difference is that the former requires i.i.d. samples, while we allow an arbitrary set of points. Our model attempts to capture situations in which the training data is sufficiently informative so as to pick out only very few potential candidate hypotheses. If, furthermore, all of these candidates are "close" in some metric, it stands to reason that all of them are in fact close to the target concept in that metric. Linear regression provides a canonical example of this situation (worked out in detail in Section 2). Indeed, when the response variable $y = \mathbf{w}^\intercal \mathbf{x}$ is a noiseless linear function of the d-dimensional predictor variables $\mathbf{x}$, it suffices to observe d labeled points in general position in order to recover $\mathbf{w}$ exactly. When the observations are perturbed by additive noise ($y = \mathbf{w}^\intercal \mathbf{x} + \xi$), it will be possible to recover $\mathbf{w}$ up to an error that depends on the configuration of the training points as well as the magnitude of the noise (essentially, a signal to noise ratio).

Main results. We define a learning model which we call Adversarial Design Approximately Correct (ADAC). This model tries to capture the phenomenon that, when learning a restricted class of hypotheses, it is often the case that a few arbitrarily chosen noiseless examples pin down the target function uniquely. We begin, as in the PAC model, with an instance space $\mathcal{X}$, a label space $\mathcal{Y}$, and a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. Unlike PAC, however, there is no distribution over $\mathcal{X}$ from which a training sample would be drawn; instead, some arbitrary data set $S \subseteq \mathcal{X}$ is provided. A teacher chooses a target concept $c \in \mathcal{H}$ and labels every $x \in S$ with its true label $c(x)$. These labels are then corrupted by a noise process η, and the learner ultimately receives the set of pairs (x, y), with $x \in S$ and $y = \eta(c(x))$.[1] To finalize our specification of an ADAC problem, we need a *loss* function $\ell : \mathcal{H} \times \mathcal{H} \to [0, \infty)$ over the hypotheses. The learner observes the labeled data $\{(x_i, y_i) : x_i \in S\}$ and produces a hypothesis $h \in \mathcal{H}$. This induces the random variable $L = \ell(h, c)$ — where the only source of randomness is the label noise process. We say that the quadruple $(\mathcal{H}, S, \eta, \ell)$ is (ε, δ)-*ADAC learnable* if $\mathbb{P}(L > \varepsilon) < \delta$.

We initiate the study of ADAC learnability by giving a positive result for the concept class of thresholds $h_a : x \mapsto \mathbb{1}_{\{x \geq a\}}$ indexed by $a \in \mathbb{R}$, where the noise process flips a label with probability $\eta < 1/2$ and $\ell(a, a') = |a - a'|$. While this target class is rather simple, its analysis already turns out to be nontrivial. In Theorem 2, we show that, as long as the training data contains

$$\Omega\left(\log(1/\delta)/(1-2\eta)^2\right)$$

points within a distance of ε from the target threshold, the (ε, δ)-ADAC learnability condition is satisfied. We further show in Corollary 3 that this recovers and sharpens a known noisy PAC learnability result for thresholds.

2 Warm-up: regression

In this section, we use linear regression as a vehicle for building intuition regarding the ADAC learning model. The simplest case is one-dimensional: an affine function determines the y value from the x coordinate of a point. Notice that if there is no noise in the y values, any two distinct points will exactly determine the line. If the y coordinate is corrupted by additive Gaussian noise, two "far" points are more informative than two close ones, and more points are more informative than fewer. With this in mind, let us consider the general d-dimensional case. Let us arrange our data $S = \{\mathbf{x}^1, \ldots, \mathbf{x}^m\} \subset \mathbb{R}^d$ as an $m \times d$ matrix $\mathbf{X}$, and assume that $\mathbf{X}^\intercal\mathbf{X}$ is non-singular. The learner's hypothesis $\hat{\mathbf{w}}$ will be the ordinary least squares (OLS) estimate:

$$\hat{\mathbf{w}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y},$$

where $\mathbf{y} \in \mathbb{R}^m$ is the "labels" vector. We assume Gaussian white label noise,

$$\mathbf{y} = \mathbf{X}\mathbf{w}^\star + \boldsymbol{\xi},$$

where $\boldsymbol{\xi} \sim N(0, \eta^2 I_m)$ for a certain noise parameter $\eta > 0$. Thus,

$$\begin{aligned}
\hat{\mathbf{w}} &= (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} \\
&= (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal(\mathbf{X}\mathbf{w}^\star + \boldsymbol{\xi}) \\
&= (\mathbf{X}^\intercal\mathbf{X})^{-1}(\mathbf{X}^\intercal\mathbf{X})\mathbf{w}^\star + (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\xi} \\
&= \mathbf{w}^\star + (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\xi}.
\end{aligned}$$

Hence, the error vector $\mathbf{z} := \hat{\mathbf{w}} - \mathbf{w}^\star$ is distributed as $N(0, \eta^2 BB^\intercal)$, where $B = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$. Simplifying $BB^\intercal = (\mathbf{X}^\intercal\mathbf{X})^{-1}$, this yields

$$\mathbf{z} \sim N(0, \eta^2(\mathbf{X}^\intercal\mathbf{X})^{-1}),$$

and hence

$$\begin{aligned}
\mathbb{E}\,\|\mathbf{z}\|_2^2 &= \sum_{i=1}^{d} \mathbb{E} z_i^2 \\
&= \eta^2 \sum_{i=1}^{d} [(\mathbf{X}^\intercal\mathbf{X})^{-1}]_{ii} \\
&= \eta^2 \operatorname{tr}((\mathbf{X}^\intercal\mathbf{X})^{-1}) \\
&= \eta^2 \sum_{i=1}^{d} \sigma_i^{-2}(\mathbf{X}),
\end{aligned}$$

where $\sigma_i(\mathbf{X})$ is the ith singular value.

Let us first make the connection to classical statistics, which makes additional assumptions on $\mathbf{X}$. In the *random design* setting, the data points $\mathbf{x}^i$ are assumed to be sampled from some distribution. In the simple case where $\mathbf{x}^i \sim N(0, I_d)$ and $m \gg d$, all of the singular values $\sigma_i(\mathbf{X})$ are of order of magnitude $\sqrt{m}$ (Rudelson and Vershynin, 2009), which yields the estimate

$$\mathbb{E}\,\|\hat{\mathbf{w}} - \mathbf{w}^\star\|_2^2 = O\left(\frac{d\eta^2}{m}\right). \tag{1}$$

Analogous estimates hold in typical *fixed design* settings (Tsybakov, 2004). An ADAC result is also readily obtained from (1). If the $m \times d$ data matrix $\mathbf{X}$ satisfies $m \geq d$ and $\sigma_d = \Omega(\sqrt{m})$, then Markov's inequality applied to (1) yields an (ε, δ)-ADAC learnability result for linear regression, with $\varepsilon = O(d\eta^2/\sqrt{m})$, $\delta = O(\frac{1}{\sqrt{m}})$, and loss function $\ell(\mathbf{w}, \mathbf{w}') = \|\mathbf{w} - \mathbf{w}'\|^2$.

[1] The noise processes in this paper will be specified by a single parameter, which we will also denote by $\eta \in \mathbb{R}$. No confusion should arise.

3 Learning thresholds

In this section we give an analysis for learning thresholds in the ADAC model and describe connections to PAC learning. Our main analysis comes from drawing a connection to the theory of biased random walks.

3.1 Biased random walks

We begin by taking a detour to the theory of random walks and proving the following lemma, which contains the crux of the argument for learning thresholds. This lemma characterizes the expectation (and deviation bounds) of the last time a drunkard reaches the bottommost point of a random walk on the vertical line with an upward bias.

Lemma 1. *Given a random walk on the integer line with upward and downward step probabilities of $p > 1/2$ and $q = 1-p$, respectively, let T be the last time the bottommost point is visited. Then the moment generating function of T is given by*

$$\mathbb{E}[e^{sT}] = \frac{2(p-q)}{1+\sqrt{1-4pqe^{2s}}-2qe^{s}},$$

where

$$0 \le s \le \frac{1}{2}\log\frac{1}{4pq}.$$

In particular,

$$\mathbb{E}[T] = \frac{q(3-4q)}{(p-q)^2}$$

and

$$\mathbb{P}(T \ge t) = O\left((4pq)^{t/2}\right).$$

Proof. It will be instructive to actually start by calculating directly the first two moments, $\mu_1 := \mathbb{E}[T]$ and $\mu_2 := \mathbb{E}[T^2]$ of T. If on the first step, the drunkard moves upward and never visits 0 again (which happens with probability $p-q$), then $T = 0$. If he moves upward and returns to 0 for the first time in $2k$ steps (which happens with probability $p\binom{2k-2}{k-1}(pq)^{k-1}q/k$), then the conditional expectation of T is $2k+\mu_1$. If on the first step he moves down, the conditional expectation is μ_1+1. Hence:

$$\begin{aligned}\mu_1 &= (p-q)\cdot 0 + \sum_{k=1}^{\infty} p\frac{\binom{2k-2}{k-1}}{k}(pq)^{k-1}q(\mu_1+2k) \\ &\quad + q\cdot(\mu_1+1),\end{aligned}$$

so that

$$\begin{aligned}p\mu_1 &= p\sum_{k=1}^{\infty}\frac{\binom{2k-2}{k-1}}{k}(pq)^{k-1}q\mu_1 \\ &\quad + 2p\sum_{k=1}^{\infty}\binom{2k-2}{k-1}(pq)^k + q \\ &= \frac{1}{2}\mu_1\left(1-\sqrt{1-4pq}\right) + 2\frac{pq}{1-\sqrt{1-4pq}} + q,\end{aligned}$$

whence

$$\mu_1 = \frac{q(3-4q)}{(p-q)^2}.$$

Let us now calculate $\mu_2 := \mathbb{E}[T^2]$. By reasoning similar to above,

$$\begin{aligned}\mu_2 &= (p-q)\cdot 0^2 + \sum_{k=1}^{\infty}\frac{\binom{2k-2}{k-1}}{k}(pq)^k\cdot\mathbb{E}[(T+2k)^2] \\ &\quad + q\cdot\mathbb{E}[(T+1)^2] \\ &= q\mu_2 + 4\sum_{k=1}^{\infty}\binom{2k-2}{k-1}(pq)^k x \\ &\quad + 4\sum_{k=1}^{\infty}k\binom{2k-2}{k-1}(pq)^k + q\mu_2 + 2qx + q,\end{aligned}$$

whence

$$\mu_2 = \frac{q(1-8p+28p^2-16p^3)}{(p-q)^4}.$$

In this manner we can compute the moment generating function of T:

$$\begin{aligned}\mathbb{E}[e^{sT}] &= (p-q)\cdot 1 + \sum_{k=1}^{\infty}\frac{\binom{2k-2}{k-1}}{k}(pq)^k\cdot\mathbb{E}[e^{s(T+2k)}] \\ &\quad + q\cdot\mathbb{E}[e^{s(T+1)}].\end{aligned}$$

Thus,

$$\mathbb{E}[e^{sT}] = \frac{2(p-q)}{1+\sqrt{1-4pqe^{2s}}-2qe^{s}},$$

where

$$0 \le s \le \frac{1}{2}\log\frac{1}{4pq}.$$

By Markov's inequality, we have

$$\mathbb{P}(T \ge t) \le \frac{2(p-q)e^{-st}}{1+\sqrt{1-4pqe^{2s}}-2qe^{s}}$$

for all $0 \le s \le \frac{1}{2}\log\frac{1}{4pq}$. The choice

$$s = \frac{1}{2}\log\frac{1}{4pq}$$

yields

$$\mathbb{P}(T \ge t) \le \frac{2(p-q)(4pq)^{t/2}}{1+\sqrt{q/p}},$$

which completes the proof. □

3.2 ADAC learnability of thresholds

Consider the class $\mathcal{H}$ of thresholds over $\mathbb{R}$. Each function $h_a \in \mathcal{H}$ can be represented by a scalar value $a \in \mathbb{R}$ and assigns the positive label to all points in $[a, \infty)$. Perhaps the most natural learner for this problem is one that chooses a hypothesis $\hat{a}$ so as to minimize the number of mistakes on the sample. Formally, a hypothesis h is *dominated* by a hypothesis h' if the set of examples that h gets correct is strictly contained in the set of examples h' gets correct (note that this definition makes sense for infinite samples also). At this point, for notational simplicity, we will write a instead of h_a for the hypothesis indexed by $a \in \mathbb{R}$. An Undominated Empirical Risk (UER) hypothesis $\hat{a}$ is one that is not dominated by any other hypothesis; we note in passing that this notion is closely related but not identical to Empirical Risk Minimization (ERM) .

In pathological cases, a UER hypothesis might not exist. More commonly, it will exist but not be unique. In this section, we will prove the following theorem.

Theorem 2. *Let h_{a^*} be the target threshold. There exists an $m_0 = m_0(\delta, \eta)$, of magnitude*

$$m_0 = O\left(\frac{\log(1/\delta)}{\log(1/(\eta - \eta^2))}\right) \quad (2)$$

$$\subset O\left(\frac{\log 1/\delta}{(1-2\eta)^2}\right), \quad (3)$$

such that if the sample contains $m \geq m_0$ data points both in the interval $[a^* - \varepsilon, a^*)$ and also in $(a^*, a^* + \varepsilon]$, then any UER learner will output a hypothesis $h_{\hat{a}}$ such that $|a^* - \hat{a}| \leq \varepsilon$ with probability at least $1 - \delta$.

Proof. First, we introduce a simplifying assumption, which we will later remove: assume our (infinite) sample consists of all the integers $\mathbb{Z}$. We will now analyze the given UER classifier. Consider the question: when would the UER classifier choose an integer value $\hat{a} > a^*$? This can happen only if the number of negative examples exceeds the number of positive examples in $[a^*, \hat{a}]$. This event can be analyzed from the viewpoint of a random walk in 1 dimension, starting at 0 and moving "up" by 1 upon seeing a positively labeled point and "down" by 1 upon seeing a negatively labeled point, which will happen for each point with probability $1 - \eta$ and η, respectively. It is easy to see that the event whereby the UER value at $\geq a^*$ will happen at time $\hat{a}$ where the drunkard has reached his bottommost point, as illustrated in Figure 1. Now we will analyze the "deviation" random variable $D := |\hat{a} - a^*|$. (In case of non-unique $\hat{a}$, define D to be the "worst" deviation.) Let D_+ be the distance from a^* to the farthest empirical optimum to its right and define D_- analogously on the left. Clearly,

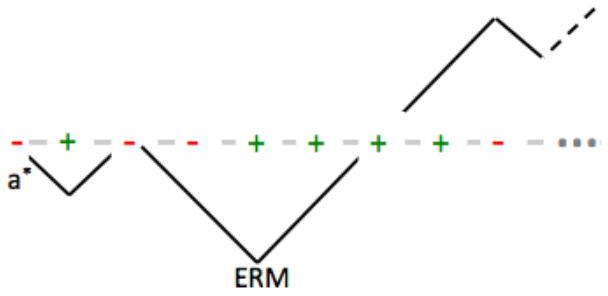

Figure 1: An illustration of the UER learner as an extreme point of a random walk. Only points to the r.h.s. of a^* are depicted. Positive points are correctly labeled; the labels of negative points are due to noise.

D_+, D_- are independent and identically distributed, and $D = \max(D_+, D_-)$. Hence

$$\mathbb{E}[D] \leq 2\mathbb{E}[D_+].$$

We will define the random variable T as the **last** time that the drunkard visits the bottommost point (i.e., minimum) of his entire walk. Note that T and D_+ have the same distribution. Now we use Lemma 1 and substitute $4pq = 4\eta - 4\eta^2$, and we get that with probability $1 - \delta$, the UER algorithm will produce a hypothesis $\hat{a}$ such that

$$|a^* - \hat{a}| \in O\left(\frac{\log(1/\delta)}{\log(1/(\eta - \eta^2))}\right)$$

$$\subset O\left(\frac{\log 1/\delta}{(1-2\eta)^2}\right)$$

Finally, we can get rid of our assumption that the data lies on all the integer points of the line by making the following two observations:

1. Our argument does not require the sample to be on integer points. Rather, the "drunkard" takes a step upon encountering a new point, so as long as he sees

   $$m = O\left(\frac{\log(1/\delta)}{\log(1/(\eta - \eta^2))}\right)$$

   data points within ε from a^*, the UER hypothesis will also be within distance ε.

2. Given that the algorithm has seen sufficiently many points (denoted m above) within ε of a^*, seeing additional data is not necessary for the algorithm to succeed (in fact, it only increases the probability of failure). Hence, an infinite sample is not needed.

This completes the proof. □

Remark. The inequality

$$\frac{1}{\log\left(1/(4\eta - 4\eta^2)\right)} \leq \frac{1}{(1-2\eta)^2}$$

follows from the elementary estimate

$$1 - x \leq \log \frac{1}{x}, \qquad x > 0$$

applied to $x = 4\eta - 4\eta^2$. It was mainly invoked to obtain a bound in a form familiar from noisy PAC (see below). Observe, however, that as $\eta \to 0$, the two bounds (2) and (3) become qualitatively different. In this regime, the estimate in (2) becomes $O(1)$ — which makes sense, since without noise, a single pair of sample points trapping the target threshold within ε from left and right suffices to achieve the desired accuracy. In contradistinction, even for $\eta = 0$, the estimate in (3) remains of order $\log(1/\delta)$.

Open problem. A natural high-dimensional analogue of thresholds is the concept class of half-spaces:

$$\mathcal{H} = \left\{\mathbf{x} \mapsto \operatorname{sign}(\mathbf{w} \cdot \mathbf{x}) : \mathbf{w} \in \mathbb{R}^d\right\}.$$

For some $p \geq 1$, define the loss

$$\begin{aligned} \ell_p(\mathbf{w}, \mathbf{w}') &= \|\mathbf{w} - \mathbf{w}'\|_p \\ &= \left(\sum_{i=1}^{d} |w_i - w_i'|^p\right)^{1/p}. \end{aligned}$$

The noise process is the same as for the thresholds: each label is flipped with probability $0 \leq \eta < 1/2$. What non-trivial property must the training data satisfy in order to assure ADAC learnability? One possibly helpful fact is that a half-space also imposes an "ordering" (similar to the ordering implicitly used by our threshold analysis) on the points along the normal to its hyperplane, and that n points in d dimensions only admit $O(n^d)$ different orderings when projected onto lines (Cover, 1967).

3.3 Relationship to PAC learning

Let us briefly recall the (proper[2]) noisy PAC learning model (Angluin and Laird, 1987; Kearns and Vazirani, 1997). A teacher and a learner agree on the *instance space* $\mathcal{X}$ and concept class $\mathcal{C} \subset \{0,1\}^{\mathcal{X}}$. The teacher privately chooses any $c \in \mathcal{C}$ and any distribution D over $\mathcal{X}$. He proceeds to draw m examples i.i.d. $\sim D$ and label each example x_i with $c(x_i)$. Each label is then flipped independently with probability $\eta < 1/2$, and the learner gets to see the m examples along with their (potentially corrupted) labels. Based on this noisy sample, the learner produces a hypothesis $h \in \mathcal{C}$, with its associated *generalization error*

$$\operatorname{err}(h) := \mathbb{P}_{X \sim D}[h(X) \neq c(X)].$$

The learner is (ε, δ)-successful if

$$\mathbb{P}(\operatorname{err}(h) > \varepsilon) < \delta.$$

A PAC learner is one who is (ε, δ)-successful whenever $m > m_0$, where m_0 depends on $\mathcal{C}, \varepsilon, \delta$ but, crucially, not c or D.

We will now show that the ADAC result we proved for thresholds in Theorem 2 has implications for noisy PAC learnability of this concept class under the uniform distribution. Formally, we take $\mathcal{X} = [0,1]$ with uniform distribution and $\mathcal{C}$ to be the collection of thresholds c_a, as defined above.

Corollary 3. *Thresholds are (properly) PAC learnable, with label noise at level $0 \leq \eta < 1/2$, under the uniform distribution over $[0,1]$, by any UER algorithm that has access $m \geq m_0$ i.i.d. examples, for*

$$m_0 = O\left(\frac{\log^2(1/\delta)}{\varepsilon(1-2\eta)^2}\right).$$

Proof. Theorem 2 says that we require

$$m = \Omega\left(\frac{\log(1/\delta)}{(1-2\eta)^2}\right)$$

data points within ε to the left and right of the target, a^*. This will happen after $O(\log(1/\delta)m/\varepsilon)$ points are drawn from the uniform distribution, giving a sample complexity of

$$O\left(\frac{\log^2(1/\delta)}{\varepsilon(1-2\eta)^2}\right)$$

and proving the PAC learnability of thresholds under the uniform distribution with an UER algorithm. Note that in the event that the target threshold a^* lies within ε of a boundary (0 or 1), we only need points to one side of the threshold, and the same analysis goes through. □

Note that this also improves the dependence on ε of the general argument of Angluin and Laird (1987) and Talagrand (1994) bound of

$$O\left(\frac{d\log(1/\delta)}{\varepsilon^2(1-2\eta)^2}\right)$$

examples being sufficient — the reason being that the errors of the different hypotheses are highly correlated,

[2]The qualifier "proper" means that the teacher and learner both work with the same concept class $\mathcal{C}$. More generally, the learner might choose to produce hypotheses $h \in \mathcal{H} \supsetneq \mathcal{C}$ from a strictly richer collection, but we will not consider this "improper" setting here.

as the random walk analysis reveals. Moreover, our bound, specialized to the PAC setting matches Aslam and Decatur's state of the art bound for learning noisy thresholds (Aslam and Decatur, 1998), The random walk analysis perhaps surprisingly re-derives the factor of $(1-2\eta)$ via a rather different technique. Our slightly worse dependence on δ can be improved via a finer analysis simultaneously considering the two failure events (not having enough points in the interval, and failure from the noise), but we are not especially concerned with polylog factors.

4 Other classes

Our random walk analysis also extends to other hypothesis classes. For instance (to learn all parameters within ε), it is not difficult to see that intervals on the line require sufficiently many points within ε to the right and left of both boundaries, and more generally, unions of k intervals would require

$$m = \Omega\left(\log(k/\delta)/(1-2\eta)^2\right)$$

points within each of the $2k$ boundaries, which requires $\tilde{O}(km/\varepsilon)$ samples. This yields a PAC bound of

$$\tilde{O}\left(\frac{k\log^2(1/\delta)}{\varepsilon(1-2\eta)^2}\right)$$

on the number of examples sufficient for learning under the uniform distribution.

For the class of axis-aligned rectangles in $\mathbb{R}^d$, the requirement would be that sufficiently many points should lie within ε of each boundary, when the points are projected to each of the d dimensions. This reduces the problem into d separate problems of ADAC learning intervals, in each of the dimensions. This again would yield a PAC bound of

$$\tilde{O}\left(\frac{d\log^2(1/\delta)}{\varepsilon(1-2\eta)^2}\right)$$

for learning under the uniform.

As in the case of thresholds, both the above bounds improve, by a factor of $1/\varepsilon$, on the generic noisy PAC bounds of Angluin and Laird (1987).

5 Discussion

We have defined a learning model that may be roughly characterized as "distribution-less proper noisy PAC." Indeed, as in proper PAC, we assume a fixed, known concept class; we also allow label corruption as in noisy PAC. Our substantial departure from the PAC framework consists of admitting arbitrarily chosen training points, and giving an accuracy guarantee in terms of their configuration. Although ADAC does not strictly imply proper noisy PAC (the two are not directly comparable), we have given an example where the ADAC analysis recovers and actually sharpens the noisy PAC version. One cannot help but wonder about a more fundamental connection between learning in the ADAC and noisy PAC models.

References

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

Javed A. Aslam and Scott E. Decatur. Specification and simulation of statistical query algorithms for efficiency and noise tolerance. *J. Comput. Syst. Sci.*, 56(2):191–208, 1998. doi: 10.1006/jcss.1997.1558. URL `http://dx.doi.org/10.1006/jcss.1997.1558`.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, Cambridge, 2006.

Thomas M Cover. The number of linearly inducible orderings of points in d-space. *SIAM Journal on Applied Mathematics*, 15(2):434–439, 1967.

David Gamarnik. Extension of the PAC framework to finite and countable Markov chains. *IEEE Trans. Inform. Theory*, 49(1):338–345, 2003.

Ralf Herbrich and Robert C. Williamson. Algorithmic luckiness. *Journal of Machine Learning Research*, 3: 175–212, 2002.

Rajeeva L. Karandikar and Mathukumalli Vidyasagar. Rates of uniform convergence of empirical means with mixing processes. *Statist. Probab. Lett.*, 58(3): 297–307, 2002. ISSN 0167-7152.

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

Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *J. Comput. Syst. Sci.*, 48(3):464–497, 1994. ISSN 0022-0000. doi: http://dx.doi.org/10.1016/S0022-0000(05)80062-5.

Michael J. Kearns, Robert E. Schapire, and Linda Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

Micheal Kearns and Umesh Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, 1997.

Ben London, Bert Huang, and Lise Getoor. Improved generalization bounds for large-scale structured prediction. In *NIPS Workshop on Algorithmic and Statistical Approaches for Large Social Networks*, 2012.

Ben London, Bert Huang, Benjamin Taskar, and Lise Getoor. Collective stability in structured prediction: Generalization from one example. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary phi-mixing and beta-mixing processes. *Journal of Machine Learning Research*, 11:789–814, 2010.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In *Neural Information Processing Systems (NIPS)*, 2008.

Afshin Rostamizadeh and Mehryar Mohri. Stability bounds for non-i.i.d. processes. In *Neural Information Processing Systems (NIPS)*, 2007.

Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Commun. Pure Appl. Math.*, 62(12):1707–1739, 2009.

Cosma Rohilla Shalizi and Aryeh Kontorovich. Predictive PAC learning and process decompositions. In *Neural Information Processing Systems (NIPS)*, 2013.

John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5): 1926–1940, 1998.

Ingo Steinwart and Andreas Christmann. Fast learning from non-i.i.d. observations. In *NIPS*, pages 1768–1776, 2009.

Ingo Steinwart, Don Hush, and Clint Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175 – 194, 2009. ISSN 0047-259X. doi: http://dx.doi.org/10.1016/j.jmva.2008.04.001. URL `http://www.sciencedirect.com/science/article/pii/S0047259X08001097`.

Michel Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994.

Alexandre B. Tsybakov. *Introduction à l'estimation non-paramétrique*, volume 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2004. ISBN 3-540-40592-5.

Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

Bin Zou, Zong-ben Xu, and Jie Xu. Generalization bounds of ERM algorithm with Markov chain samples. *Acta Mathematicae Applicatae Sinica (English Series)*, pages 1–16, 2014. ISSN 0168-9673. URL `http://dx.doi.org/10.1007/s10255-011-0096-4`.