Boosting Conditional Probability Estimators

Dan Gutfreund and **Aryeh Kontorovich** and **Ran Levy** and **Michal Rosen-Zvi**

Abstract

In the standard agnostic multiclass model, <instance, label> pairs are sampled independently from some underlying distribution. This distribution induces a conditional probability over the labels given an instance, and our goal in this paper is to learn this conditional distribution. Since even *unconditional* densities are quite challenging to learn, we give our learner access to <instance, conditional distribution> pairs. Assuming a base learner oracle in this model, we might seek a boosting algorithm for constructing a strong learner. Unfortunately, without further assumptions, this is provably impossible. However, we give a new boosting algorithm that succeeds in the following sense: given a base learner guaranteed to achieve some average accuracy (i.e., risk), we efficiently construct a learner that achieves the same level of accuracy with arbitrarily high probability. We give generalization guarantees of several different kinds, including distribution-free accuracy and risk bounds. None of our estimates depend on the number of boosting rounds and some of them admit dimension-free formulations.

Introduction

The goal of multi-class supervised learning is to produce a classification rule given labeled examples drawn independently from some unknown distribution. The performance of a classifier is measured by its generalization error: the probability of misclassifying an example drawn from the underlying distribution. A much more daunting task is to learn the full conditional class probability function (rather than just the most likely class). Formally, a distribution D over $\mathcal{X} \times \mathcal{Y}$ (where $\mathcal{X}$ and $\mathcal{Y}$ are the instance and label spaces, respectively) induces the conditional distribution $p^x = \Pr[\cdot \,|\, X = x]$, and our goal is to learn the map $p^* : \mathcal{X} \ni x \mapsto p^x \in \mathbb{R}^{|\mathcal{Y}|}$. Given the difficulty of the task — even learning *unconditional* high-dimensional densities is notoriously hard — we consider the more benign setting where the learner is allowed to observe a training sample consisting of pairs (X, p^X). This scenario is not entirely idealized: some part-of-speech induction techniques work implicitly in this setting (Das and Petrov 2011; Toutanova and Cherry 2009), and p^x may also be estimated empirically from the more standard data arriving as (X, Y) pairs.

Given a hypothesis $f : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$, we consider loss functions of the general form $\ell(f, p^*) = \|f - p^*\|$, where $\|\cdot\|$ is some norm on distributions bounded above by 1. A *η-base learner* is an algorithm that, given a sample $\left(X, p^X\right)_{X \in S}$ and any distribution $\mathbf{w} \in \mathbb{R}^S$ as inputs, produces a hypothesis $h : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ whose loss is bounded by η in $\mathbf{w}$-weighted expectation over S; formally, $\mathbf{E}_{\mathbf{w}} \left\| h - p^X \right\| = \sum_{x \in S} w_x \left\| h(x) - p^x \right\| \le \eta$. One might perhaps attempt to convert a η-base learner (for fixed η) to one that achieves arbitrarily small expected sample loss. Unfortunately, as we show in Preliminaries, this is generally impossible to achieve by boosting in the standard sense. However, we are able to boost the learner in the following sense: given oracle access to a η-base learner, we construct an efficient learner that achieves $\Pr_S[\left\| h(X) - p^X \right\| > \eta] < \exp(-O(T))$, where T is the number of boosting rounds. For comparison, note that Markov's inequality yields the trivial estimate $\Pr_S[\left\| h(X) - p^X \right\| > \eta] \le \eta^{-1} \mathbf{E}_S \left\| h(X) - p^X \right\|$, which can be arbitrarily close to 1 and is insensitive to the number of boosting rounds.

Main results We propose a new boosting algorithm, `Adacond` (Figure 1), for learning conditional distributions. `Adacond` efficiently converts a learner that is accurate on average to one that is accurate with high probability: the sample probability of the boosted learner exceeding the base loss η on any training example decays exponentially with the number of boosting rounds (Theorem 1).

We give distribution-free generalization guarantees of several different kinds, including accuracy (Theorems 4 and 6) and risk bounds (Theorem 7). None of our estimates degrade with the number of boosting rounds and some of them admit dimension-free formulations (i.e., independent of $|\mathcal{Y}|$, Theorem 8).

The accuracy bounds are in terms of the fat-shattering dimension of the base learner (Alon et al. 1997). For the risk bounds, we take $\mathcal{X}$ to be a metric space with finite doubling dimension and the hypothesis class satisfies a natural Lipschitz condition.

Related work Boosting is by now a standard tool in supervised learning. The general paradigm is to combine many simple, weakly accurate classification rules into a single, highly accurate, classifier. The Adaboost algorithm was introduced in the seminal work of Freund and Schapire (Freund and Schapire 1997). Loosely speaking, given a training set, Adaboost iteratively asks a weak learner to generate classifiers with respect to different distributions over the training set. The final classifier is a convex combination of the weak classifiers generated during the iterations. The analysis in (Freund and Schapire 1997) shows that if the weak learner is able to produce at each iteration a classifier that does only slightly better than a random guess, then the training error of the final classifier decays exponentially with the number of iterations. Following (Freund and Schapire 1997), many variants of Adaboost has been suggested.

Boosting in the context of estimating conditional class probabilities has also been studied although to a much lesser degree. Friedman et al. (Friedman, Hastie, and Tibshirani 2000) showed that Adaboost can be interpreted as an algorithm for stagewise fitting an additive logistic regression model. They argue that substituting the result of the final classifier into the logistic regression formula gives a good estimate for the conditional class probability. Mease et al. (Mease, Wyner, and Buja 2007) argue, on the other hand, that since Adaboost classifies at the 50th percentile of the conditional class distribution (for two-class problems), it tends to put all the weight on the most likely class and thus performs badly as an estimator for this distribution, whether this is done directly or via a logistic regression link function as in (Friedman, Hastie, and Tibshirani 2000). Mease et al. support their arguments with simulations on synthetic data. As an alternative approach, they suggest estimating conditional class probability by running Adaboost multiple times to produce classifiers on a grid of percentiles using under/over sampling to achieve different ratios between class labels. They show via simulations that their approach works well but do not provide a theoretical analysis. A similar approach uses cost sensitive boosting such as Adacost (Fan et al. 1999) as an alternative to under/over sampling. In (Kanamori 2010), Kanamori proposes a new class of loss functions for multi-class classification. A boosting algorithm, minimizing this loss function is presented. If the sample size tends to infinity, it is shown that the conditional probability can be calculated using the obtained decision function. The rate of convergence is not discussed and generalization bounds are not obtained.

Our algorithm may also be viewed as a tool for boosted regression (Duffy and Helmbold 2002), although our choice of penalty makes the conditional distribution interpretation more natural.

Preliminaries

The learning problem we have described in the Introduction is formalized as follows. Our instance space $\mathcal{X}$ is a measurable space and the label set $\mathcal{Y}$ is finite. A training sample $S = \left(X_i, p^{X_i}\right)_{i=1}^{n}$ is generated by drawing n points in $\mathcal{X}$ independently from some unknown distribution D and labeling each example X_i with $p^{X_i} = p^*(X_i)$, where the target function $p^* : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ has the interpretation of being the conditional probability over the labels: $p^*(x) = \Pr[\cdot \mid X = x]$.

```
Input:
  • training sample $S = \left(X_i, p^{X_i}\right)_{i=1}^{n}$
  • $\eta$-base learning oracle BaseLearn
  • number of boosting rounds $T \in \mathbb{N}$
Initialize: weight vector $w_i^1 = 1/n$ for $1 \le i \le n$.
The algorithm: for $t = 1$ to $T$:
  1. call BaseLearn on the sample $S$ with distribution $w_i^t / \sum_{j=1}^{n} w_j^t$;
     get back a classifier $c_t : \mathcal{X} \to \mathbb{R}^k$
  2. calculate the average loss of $c_t$:
     $\epsilon_t := \sum_{i=1}^{n} w_i^t \left\| c_t(X_i) - p^{X_i} \right\| / \sum w_j^t$
  3. $\beta_t := ((1-\eta)\epsilon_t)/(\eta(1-\epsilon_t))$
  4. update $w_i^{t+1} := w_i^t \beta^{1-\left\| c_t(X_i) - p^{X_i} \right\|}$
Output: set $Z := \sum_{t=1}^{T} \log(1/\beta_t)$, output the hypothesis
  $h(x) := \frac{1}{Z} \sum_{t=1}^{T} \log(1/\beta_t) c_t(x)$
```

Figure 1: The algorithm Adacond for boosting conditional class probabilities estimators.

The sample S induces the following (so-called *empirical*) distribution: $\Pr_S = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$, where δ_{X_i} is a point mass of weight 1 at X_i. Unless a different distribution over S is explicitly specified, $\Pr_S$ and $\mathbf{E}_S$ denote the probability and expectation, respectively, under the empirical distribution. Random variables are capitalized and concrete values they take are written lowercase.

In this paper, all norms on probability distributions will be assumed to take values in $[0,1]$. One example of such a norm is the *total variation*: for two stochastic vectors ξ, ψ, this quantity is half of their ℓ_1 distance: $\|\xi - \psi\|_{\mathrm{TV}} = \frac{1}{2} \|\xi - \psi\|_1 = \frac{1}{2} \sum_i |\xi_i - \psi_i|$.

A η-base learner (with respect to some fixed norm $\|\cdot\|$) is an oracle that, given a sample S and a distribution $\mathbf{w} \in \mathbb{R}^S$ as inputs, produces a hypothesis $h : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ whose loss is bounded by η in $\mathbf{w}$-weighted expectation over S: $\mathbf{E}_{\mathbf{w}} \left\| h - p^X \right\| \le \eta$. We mentioned in the Introduction that it is in general impossible to boost a η-base learner to one that achieves arbitrarily small expected sample loss. Indeed, suppose a base learner always returns a hypothesis h that on every training element always outputs the same distribution and this distribution is *exactly* η away from the target function. Then for every distribution $\mathbf{w}$ over S, $\mathbf{E}_{\mathbf{w}} \|h - p^*\| = \eta$, which means that it is a η-base learner. Clearly, any convex combination of hypotheses generated by this η-base learner will also have a sample loss of exactly η, and there is no hope of proving a general theorem which guarantees an arbitrarily small expected sample loss. This motivates our alternate goal of boosting the learner from being η-close on average to being η-close with arbitrarily high probability (Theorem 1).

The boosting algorithm

Our boosting algorithm, Adacond, is given in Figure 1. It works for any norm $\|\cdot\|$ on the space of probability vectors taking values in $[0,1]$; one example of such a norm is $\|\cdot\|_{\mathrm{TV}}$.

Bounding `Adacond`'s training error

Given a sample $\left(X, p^X\right)_{X\in S}$, our boosting algorithm efficiently converts a learner achieving a sample loss of at most η on average into one that achieves this with high probability:

Theorem 1. *If* `Adacond` *is given oracle access to a η-base learner and run for T boosting rounds, it will output a hypothesis $h : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ that satisfies*

$$\Pr_{X\sim S}\left[\left\|h(X) - p^X\right\| > \eta\right] \le \prod_{t=1}^{T} \frac{\epsilon_t(1-\epsilon_t)}{((1-\eta)\epsilon_t)^{1-\eta}(\eta(1-\epsilon_t))^{\eta}}.$$

The bound on the training error given in Theorem 1 vanishes at an exponential rate if each term in the product on the right-hand side is bounded away from 1. To see this, put $\gamma_t = \eta - \epsilon_t \ge 0$. It is straightforward to verify that for $0 < \eta \le \frac{1}{2}$, $0 \le \gamma_t \le \eta$, we have

$$\frac{\epsilon_t(1-\epsilon_t)}{((1-\eta)\epsilon_t)^{1-\eta}(\eta(1-\epsilon_t))^{\eta}} \quad \le \quad \exp\left(-\frac{\gamma_t^2}{2\eta(1-\eta)}\right) \le \exp\left(-2\gamma_t^2\right)$$

and hence,

$$\Pr_S\left[\left\|h(X) - p^X\right\| > \eta\right] \quad \le \quad \exp\left(-\frac{\sum_{t=1}^T \gamma_t^2}{2\eta(1-\eta)}\right) \le \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right),$$

which decays exponentially with T if $\gamma_t \ge \gamma > 0$ for all t. In this case, `Adacond` is guaranteed to output a hypothesis that achieves a loss of at most η on every training example provided that $T > 2\eta(1-\eta)\log(n)/\gamma^2$.

The proof of Theorem 1 follows the outline of the proof from (Freund and Schapire 1997) with the appropriate adjustments to our setting. The following simple consequence of the convexity of a norm will be useful:

$$\left\|\sum_{t=1}^{T}\alpha_t\mathbf{w}^t - \mathbf{v}\right\| \le \sum_{t=1}^{T}\alpha_t\left\|\mathbf{w}^t - \mathbf{v}\right\| \qquad (1)$$

holds for all vectors $\mathbf{v}, \mathbf{w}^1, \ldots, \mathbf{w}^T$ and all $\alpha \in [0,1]^T$ with $\sum_t \alpha_t = 1$.

Proof of Theorem 1. By the update rule of the weight vector we have

$$\begin{aligned}\sum_{i=1}^{n} w_i^{t+1} &= \sum_{i=1}^{n} w_i^t \beta_t^{1-\left\|c_t(X_i)-p^{X_i}\right\|}\\ &\le \sum_{i=1}^{n} w_i^t\left(1-(1-\beta_t)(1-\left\|c_t(X_i)-p^{X_i}\right\|\right)\\ &= \left(\sum_{i=1}^{n} w_i^t\right)(1-(1-\epsilon_t)(1-\beta_t)).\end{aligned}$$

Here we used the fact that $0 \le \left\|c_t(X_i) - p^{X_i}\right\| \le 1$ and Bernoulli inequality: $\alpha^r \le 1-(1-\alpha)r$ for any $\alpha \ge 0$ and $0 \le r \le 1$.

Unraveling the recurrence on t, we get the following bound on the final weight function:

$$\sum_{i=1}^{n} w_i^{T+1} \le \prod_{i=1}^{T}(1-(1-\epsilon_t)(1-\beta_t)). \qquad (2)$$

The final hypothesis h is more than η-far from the correct probability vector on the i^{th} example if

$$\left\|\frac{1}{Z}\sum_{t=1}^{T}\log(1/\beta_t)c_t(X_i) - p^{X_i}\right\| > \eta. \qquad (3)$$

Since $\log(1/\beta_t) \ge 0$ for every $1 \le t \le T$, Inequality (1) implies

$$\left\|\frac{1}{Z}\sum_{t=1}^{T}\log(1/\beta_t)c_t(X_i) - p^{X_i}\right\| \le \frac{1}{Z}\sum_{t=1}^{T}\log(1/\beta_t)\left\|c_t(X_i) - p^{X_i}\right\|.$$

Combining with Inequality (3), we have

$$\sum_{t=1}^{T}\log(1/\beta_t)\left\|c_t(X_i) - p^{X_i}\right\| \ge \eta Z = \eta\sum_{t=1}^{T}\log(1/\beta_t),$$

which implies

$$\prod_{t=1}^{T}\beta_t^{-\left\|c_t(X_i)-p^{X_i}\right\|} \ge \left(\prod_{t=1}^{T}\beta_t\right)^{-\eta}. \qquad (4)$$

By the update rule for the weights, we get

$$w_i^{T+1} = \frac{1}{n}\prod_{t=1}^{T}\beta_t^{1-\left\|c_t(X_i)-p^{X_i}\right\|}. \qquad (5)$$

Let $\epsilon = \Pr_S[\left\|h(X) - p^X\right\| > \eta]$. Combining Equations (4) and (5) we get

$$\begin{aligned}\sum_{i=1}^{n} w_i^{T+1} &\ge \sum_{i:\left\|h(X_i)-p^{X_i}\right\|>\eta} w_i^{T+1}\\ &\ge \left(\sum_{i:\left\|h(X_i)-p^{X_i}\right\|>\eta}\frac{1}{n}\right)\left(\prod_{t=1}^{T}\beta_t\right)^{1-\eta}\\ &= \epsilon\cdot\left(\prod_{t=1}^{T}\beta_t\right)^{1-\eta}.\end{aligned}$$

Combining this with Inequality (2) we get the following bound on ϵ:

$$\epsilon \le \prod_{t=1}^{T}\frac{1-(1-\epsilon_t)(1-\beta)}{\beta_t^{1-\eta}}. \qquad (6)$$

Choosing each β_t to minimize the t^{th} factor in the product yields $\beta_t = ((1-\eta)\epsilon_t)/(\eta(1-\epsilon_t))$. Plugging this into Inequality (6) gives us the desired bound

$$\epsilon \le \prod_{t=1}^{T}\frac{\epsilon_t(1-\epsilon_t)}{((1-\eta)\epsilon_t)^{1-\eta}(\eta(1-\epsilon_t))^{\eta}}.$$

□

Generalization bounds

For the analysis, we specialize to a particular choice of norm over probability vectors — namely, the total variation norm

$\|\cdot\|_{\mathrm{TV}}$, which in some sense, is the most natural norm in this setting (Devroye and Lugosi 2001; Gibbs and Su 2002).

We henceforth take $|\mathcal{Y}| = k$ and consider a hypothesis class $\mathcal{H}$ of conditional probability estimators for the k-class problem. That is, $h \in \mathcal{H}$ is of the form $h : \mathcal{X} \to [0,1]^k$, where $\sum_{i=1}^{k} [h(x)]_i = 1$. We denote by $\mathcal{H}_i$ the projection of $\mathcal{H}$ onto the i^{th} coordinate. For positive integers N, T we define the T-convex hull of $\mathcal{H}$ to be

$$\mathrm{conv}_T(\mathcal{H}) = \left\{ \sum_{t=1}^{T} \alpha_t h_t : \alpha_1, \ldots, \alpha_T \in [0,1], \sum_{t=1}^{T} \alpha_t = 1, h_1, \ldots, h_T \in \mathcal{H} \right\}$$

and the set of N-averages over $\mathcal{H}$ to be

$$\mathrm{Avg}_N(\mathcal{H}) = \left\{ \frac{1}{N} \sum_{i=1}^{N} h_i : h_1, \ldots, h_N \in \mathcal{H} \right\};$$

in both definitions, multiple appearances of the same $h \in \mathcal{H}$ are allowed in the sum. Note that the final hypothesis of `Adacond` is in the class $\mathrm{conv}_T(\mathcal{H})$ where $\mathcal{H}$ is the class of hypotheses generated by the base learner.

We will need the following concentration bound in Hilbert spaces, which is a simple consequence of McDiarmid's inequality (McDiarmid 1989):

Lemma 2. *Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in a separable Hilbert space with norm $\|\cdot\|_H$. Suppose further that $\mathbf{E}[X_i] = \mathbf{0}$ and $\|X_i\|_H \le 1$ for every $1 \le i \le n$. Then for any $t \ge 4\sqrt{n}$,*

$$\Pr[\|X_1 + \cdots + X_n\|_H \ge t] \le e^{-\frac{t^2}{8n}}.$$

Corollary 3. *Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in $\mathbb{R}^k$. Suppose further that $\mathbf{E}[X_i] = \mathbf{0}$ and $\|X_i\|_1 \le 1$ for every $1 \le i \le n$. Then for any $t \ge 4\sqrt{nk}$,*

$$\Pr[\|X_1 + \cdots + X_n\|_1 \ge t] \le e^{-\frac{t^2}{8nk}}.$$

Proof. By Lemma 2 and the fact that $\|X_i\|_2 \le \|X_i\|_1 \le 1$, we have that for any $t \ge 4\sqrt{n}$,

$$\Pr[\|X_1 + \cdots + X_n\|_2 \ge t] \le e^{-\frac{t^2}{8n}}.$$

Since $\|X\|_1 \le \sqrt{k}\,\|X\|_2$, we get that for any $t \ge 4\sqrt{nk}$,

$$\begin{aligned} \Pr[\|X_1 + \cdots + X_n\|_1 \ge t] &\le \Pr\left[\|X_1 + \cdots + X_n\|_2 \ge \frac{t}{\sqrt{k}}\right] \\ &\le e^{-\frac{t^2}{8nk}}. \end{aligned}$$

□

Finite hypothesis classes

The proof of the following theorem as well as the infinite case (Theorem 6) is based on the proof technique of (Schapire et al. 1998) with the appropriate adjustments to our setting.

Theorem 4. *Let $\mathcal{H}$ be a finite hypothesis class of probability estimators. For every $0 < \epsilon, \epsilon', \delta < 1$, with probability at least $1 - \delta$ over the choice of a training sample S of size n, for every $f \in \mathrm{conv}_T(\mathcal{H})$*

$$\begin{aligned} \Pr_D[\|f(x) - p(x)\|_{\mathrm{TV}} > \epsilon + \epsilon'] &< \Pr_S[\|f(x) - p(x)\|_{\mathrm{TV}} > \epsilon] + \\ & O\left(\sqrt{\frac{1}{n} \left(\frac{k \log n \log |\mathcal{H}|}{\epsilon'^2} + \log \frac{1}{\delta} \right)} \right). \end{aligned}$$

Remark: The bound depends on k and thus is not dimension-free. In the section titled Dimension-free bounds, we discuss additional conditions that yield dimension-free bounds.

Proof. Fix $f \in \mathrm{conv}_T(\mathcal{H})$. We can write it as $f = \sum_{i=1}^{T} \alpha_i h_i$ for some probability vector $\alpha = (\alpha_1, \ldots, \alpha_T)$ and $h_1, \ldots, h_T \in \mathcal{H}$. We define a distribution over $\mathrm{Avg}_N(\mathcal{H})$ as follows: choose N elements $h_1, \ldots, h_N$ from $\mathcal{H}$ independently according to the distribution α and output $g = \frac{1}{N} \sum_{i=1}^{N} h_i$. Denote this distribution by A.

For every g chosen according to A it holds that

$$\begin{aligned} \Pr_D\left[\|f(x) - p(x)\|_{\mathrm{TV}} > \epsilon + \epsilon'\right] &\le \Pr_D\left[\|g(x) - p(x)\|_{\mathrm{TV}} > \epsilon + \epsilon'/2\right] + \\ & \Pr_D\left[\|g(x) - p(x)\|_{\mathrm{TV}} \le \epsilon + \epsilon'/2, \|f(x) - p(x)\|_{\mathrm{TV}} > \epsilon + \epsilon'\right]. \end{aligned}$$

In particular, this holds when we take expectation over the choice of g:

$$\begin{aligned} \Pr_D[\|f(x) - p(x)\|_{\mathrm{TV}} > \epsilon + \epsilon'] &\le \mathbf{E}_A\left[\Pr_D\left[\|g(x) - p(x)\|_{\mathrm{TV}} > \epsilon + \epsilon'/2\right]\right] + \\ & \mathbf{E}_A\left[\Pr_D\left[\|g(x) - p(x)\|_{\mathrm{TV}} \le \epsilon + \epsilon'/2, \|f(x) - p(x)\|_{\mathrm{TV}} > \epsilon + \epsilon'\right]\right] \\ &= \mathbf{E}_A\left[\Pr_D\left[\|g(x) - p(x)\|_{\mathrm{TV}} > \epsilon + \epsilon'/2\right]\right] + \\ & \mathbf{E}_D\left[\Pr_A\left[\|g(x) - p(x)\|_{\mathrm{TV}} \le \epsilon + \epsilon'/2, \|f(x) - p(x)\|_{\mathrm{TV}} > \epsilon + \epsilon'\right]\right], \end{aligned}$$

where the last equality follows from independence of A and D.

We bound the two terms above separately, starting with the probability inside the expectation of the second summand. Fix an $x \in \mathcal{X}$ and let $h_1, \ldots, h_N$ be the functions chosen independently in the construction of g. Consider the random variables $Z_i = \frac{1}{2}(h_i(x) - f(x))$. These $\mathbb{R}^k$-valued random variables are i.i.d. and have expectation $\mathbf{0}$ (since $f(x) = \mathbf{E}[h_i]$), and $\|Z_i\|_1 = \|h_i - f\|_{\mathrm{TV}} \le 1$. Thus, Corollary 3 applies: for $\epsilon' > 16\sqrt{\frac{k}{N}}$ we have

$$\Pr_A[\|g(x) - f(x)\|_{\mathrm{TV}} > \epsilon'/2] = \Pr[\|Z_1 + \cdots + Z_N\|_1 > N\epsilon'/4] \le \exp\left(-\frac{N\epsilon'^2}{128k}\right).$$

Using the triangle inequality, we get

$$\begin{aligned} \Pr_A[\|g(x) - p(x)\|_{\mathrm{TV}} < \epsilon + \epsilon'/2, \|f(x) - p(x)\|_{\mathrm{TV}} > \epsilon + \epsilon'] &\le \Pr_A[\|g(x) - f(x)\|_{\mathrm{TV}} > \epsilon'/2] \\ &< \exp\left(-\frac{N\epsilon'^2}{128k}\right). \end{aligned}$$

In order to bound the first summand, $\mathbf{E}_A[\Pr_D[\|g(x) - p(x)\|_{\mathrm{TV}} > \epsilon + \epsilon'/2]]$, we will relate the error of g with respect to D to its error on the training set S. For any fixed $g \in \mathrm{Avg}_N(\mathcal{H})$ and any $0 < \epsilon_2 \le 1$, it follows from Hoeffding's inequality that with probability at most $2\exp(-2\epsilon_2^2 n)$ over the choice of the set S, we have

$$\left| \Pr_D[\|g(x) - p(x)\|_{\mathrm{TV}} > \epsilon + \epsilon'/2] - \Pr_S[\|g(x) - p(x)\|_{\mathrm{TV}} > \epsilon + \epsilon'/2] \right| > \epsilon_2.$$

Here we used the fact that elements in S are chosen independently from D. By the union bound, the probability that this event happens for some g is at most

$$2\,|\mathrm{Avg}_N(\mathcal{H})| \exp(-2\epsilon_2^2 n) = 2\,|\mathcal{H}|^N \exp(-2\epsilon_2^2 n).$$

Next, we relate the error of g on S to the error of f on any fixed S (and therefore the bound below will also hold for a randomly chosen S). Similar to the argument above,

$$\begin{aligned}\Pr_{A,S}[\|g(x)-p(x)\|_{\text{TV}} > \epsilon+\epsilon'/2] &\leq \Pr_{S}[\|f(x)-p(x)\|_{\text{TV}} > \epsilon] + \\ &\Pr_{A,S}[\|g(x)-p(x)\|_{\text{TV}} > \epsilon+\epsilon'/2, \|f(x)-p(x)\|_{\text{TV}} < \epsilon].\end{aligned}$$

The second term can be bounded as before by $\exp\left(-\frac{N\epsilon'^2}{128k}\right)$.

Putting it all together, we have that with probability at least $1-2\left|\mathcal{H}\right|^N \exp(-2\epsilon_2^2 n)$ over the choice of S, it holds that for every $f \in \text{conv}_T(\mathcal{H})$

$$\Pr_D[\|f(x)-p(x)\|_{\text{TV}} > \epsilon+\epsilon'] \leq \Pr_S[\|f(x)-p(x)\|_{\text{TV}} > \epsilon] + 2\exp\left(-\frac{N\epsilon'^2}{128k}\right) + \epsilon_2.$$

Setting $\delta = 2\left|\mathcal{H}\right|^N \exp(-2\epsilon_2^2 n)$ gives $\epsilon_2 = \sqrt{\frac{1}{2n}(1+N\log|\mathcal{H}|+\log\frac{1}{\delta})}$. We then set $N = \frac{128k\log n}{\epsilon'^2}$ (this enforces the requirement $\epsilon' > 16\sqrt{\frac{k}{N}}$). For these values of δ and N, the last display implies that

$$\begin{aligned}\Pr_D[\|f(x)-p(x)\|_{\text{TV}} > \epsilon+\epsilon'] &\leq \Pr_S[\|f(x)-p(x)\|_{\text{TV}} > \epsilon] + \frac{2}{n} + \\ &\sqrt{\frac{1}{2n}\left(1+\frac{128k\log n\log|\mathcal{H}|}{\epsilon'^2}+\log\frac{1}{\delta}\right)} \\ &= \Pr_S[\|f(x)-p(x)\|_{\text{TV}} > \epsilon] + \\ &O\left(\sqrt{\frac{1}{n}\left(\frac{k\log n\log|\mathcal{H}|}{\epsilon'^2}+\log\frac{1}{\delta}\right)}\right)\end{aligned}$$

holds with probability at least $1-\delta$. □

Infinite hypothesis classes

Extending our results to infinite function classes will require some machinery from covering numbers and fat-shattering dimensions. We refer the reader to (Alon et al. 1997) — especially since one of their key results will be used in our proof. We define the γ-fat shattering dimension of a function class $\mathcal{F}$, denoted $\text{fat}_\gamma(\mathcal{F})$, to coincide with P_γ in (Alon et al. 1997).

Let $M = (\mathcal{S}, \rho)$ be a metric space. For $E \subseteq \mathcal{S}$, we say that $C \subseteq E$ is an ϵ-cover of E if for every $s \in E$ there exist an $s' \in C$ such that $\rho(s,s') \leq \epsilon$.

We consider real-valued function classes $\mathcal{F} = \{f : \mathcal{X} \to [0,1]\}$. For $x^n = (x_1,\ldots,x_n) \in \mathcal{X}^n$, we denote by $\mathcal{F}_{|x^n}$ the functions in $\mathcal{F}$ restricted to x^n. We define the ϵ-covering number of $\mathcal{F}$ at sample size n with respect to the ℓ_∞ norm:

$$\mathcal{N}_\infty(\epsilon,\mathcal{F},n) = \max_{x^n}\min_{C}\{|C| : C \text{ is an } \epsilon\text{-cover of } \mathcal{F}_{|x^n} \text{ w.r.t. } \ell_\infty\}.$$

If for some x^n there is no finite ϵ-cover, then $\mathcal{N}_\infty(\epsilon,\mathcal{F},n) = \infty$. The following uniform convergence lemma is implicit in (Bartlett and Shawe-Taylor 1999).

Lemma 5. *Let $\mathcal{F}$ be a function class from a domain $\mathcal{X}$ to $\mathbb{R}$. Let $\eta > \alpha > 0$ and $\epsilon > 0$ and let D be a distribution over X. Let $S = (x_1,\ldots,x_n)$ be a sample drawn independently according to D. Then with probability at least $1-2\mathcal{N}_\infty(\alpha/2,\pi_\alpha(\mathcal{F}),2n)\exp(-\epsilon^2 m/2)$ over the choice of S, for every $f \in \mathcal{F}$*

$$\Pr_D[f(x) > \eta] \leq \Pr_S[f(x) > \eta-\alpha] + \epsilon,$$

where the function $\pi_\alpha : \mathbb{R} \to \mathbb{R}$ is identity on $(-\alpha,\alpha)$ and constant $-\alpha, +\alpha$ to the left and to the right, respectively.

Theorem 6. *Let $\mathcal{H}$ be a hypothesis class of probability estimators. For every $0 < \epsilon, \epsilon', \delta < 1$, with probability at least $1-\delta$ over the choice of a training set S of size n, for every $f \in \text{conv}_T(\mathcal{H})$*

$$\begin{aligned}\Pr_D[\|f(x)-p(x)\|_{\text{TV}} > \epsilon+\epsilon'] &< \Pr_S[\|f(x)-p(x)\|_{\text{TV}} > \epsilon] + \\ &O\left(\sqrt{\frac{1}{n}\left(\frac{dk^2\log n}{\epsilon'^2}\log^2\left(\frac{nk}{\epsilon'}\right)+\log\frac{1}{\delta}\right)}\right)\end{aligned}$$

where $d = \max_{1\leq i\leq k}\{\text{fat}_{\epsilon'/32}(\mathcal{H}_i)\}$.

Proof. The proof follows the arguments of the proof of Theorem 4 until the point where we need to bound the term $\mathbf{E}_A[\Pr_D[\|g(x)-p(x)\|_{\text{TV}} > \epsilon+\epsilon'/2]]$. Consider the function class $\mathcal{F} = \{f_g(x) = \|g(x)-p(x)\|_{\text{TV}} : g \in \text{Avg}_N(\mathcal{H})\}$, and let $\mathcal{F}_i = \{f_g^i(x) = |[g(x)]_i - [p(x)]_i| : g \in \text{Avg}_N(\mathcal{H})\}$ be the function class induced by the i^{th} component in the loss $g-p$. By Lemma 5, for any $\epsilon_2 > 0$, with probability at least $1-2\mathcal{N}_\infty(\epsilon'/8,\pi_{\epsilon'/4}(\mathcal{F}),2n)\exp(-\epsilon_2^2 n/2)$ over the choice of S, we have

$$\Pr_D[f_g(X) > \epsilon+\epsilon'/2] \leq \Pr_S[f_g(X) > \epsilon+\epsilon'/4] + \epsilon_2 \tag{7}$$

for every $f_g \in \mathcal{F}$. We bound the covering number of $\mathcal{F}$ in terms of (among other parameters) the fat-shattering dimension of $\mathcal{H}$. Let $d = \max_i\{\text{fat}_{\epsilon'/32}(\mathcal{H}_i)\}$.

$$\begin{aligned}\mathcal{N}_\infty(\epsilon'/8,\pi_{\epsilon'/4}(\mathcal{F}),2n) &\leq \mathcal{N}_\infty(\epsilon'/8,\mathcal{F},2n) \leq \max_i\{\mathcal{N}_\infty(\epsilon'/8k,\mathcal{F}_i,2n)\}^k \\ &\leq \max_i\{\mathcal{N}_\infty(\epsilon'/8k,\text{Avg}_N(\mathcal{H}_i),2n)\}^k \leq \max_i\{\mathcal{N}_\infty(\epsilon'/8k,\mathcal{H}_i,2n)^N\}^k \\ &\leq 2^{kN}\left(\frac{512nk^2}{\epsilon'^2}\right)^{dkN\log(32enk/(d\epsilon'))}.\end{aligned}$$

The first four inequalities follow from standard facts regarding covering numbers, and the last from (Alon et al. 1997).

We conclude that with probability at least $1-2^{kN}(\frac{512nk^2}{\epsilon'^2})^{dkN\log(32enk/(d\epsilon'))}\exp(-\epsilon_2^2 n/2)$ over the choice of S Equation (7) holds.

Similar to the proof of Theorem 4, we have

$$\begin{aligned}\Pr_{A,S}[\|g(X)-p(X)\|_{\text{TV}} > \epsilon+\epsilon'/4] &\leq \Pr_S[\|f(X)-p(X)\|_{\text{TV}} > \epsilon] + \\ &\Pr_{A,S}[\|g(X)-p(X)\|_{\text{TV}} > \epsilon+\epsilon'/4, \|f(X)-p(X)\|_{\text{TV}} < \epsilon].\end{aligned}$$

The second term can be bounded as before by $\exp\left(-\frac{N\epsilon'^2}{512k}\right)$.

Putting it all together we have that with probability at least

$$1-2^{kN}\left(\frac{512nk^2}{\epsilon'^2}\right)^{dkN\log(32enk/(d\epsilon'))}\exp(-\epsilon_2^2 n/2)$$

over the choice of S, for every $f \in \text{conv}_T(\mathcal{H})$ it holds that

$$\Pr_D[\|f(X)-p(X)\|_{\text{TV}} > \epsilon+\epsilon'] \leq \Pr_S[\|f(X)-p(X)\|_{\text{TV}} > \epsilon] + 2\exp\left(-\frac{N\epsilon'^2}{512k}\right) + \epsilon_2. \tag{8}$$

Setting

$$\delta = 2^{kN}(\frac{512nk^2}{\epsilon'^2})^{dkN\log(32enk/(d\epsilon'))}\exp(-\epsilon_2^2 n/2)$$

gives

$$\epsilon_2 = \sqrt{\frac{2}{n}\left(kN + dkN\log\left(\frac{512nk^2}{\epsilon'^2}\right)\log\left(\frac{32nk}{\mathrm{d}\epsilon'}\right)+\log\frac{1}{\delta}\right)}.$$

Letting $N = \frac{512k \log n}{\epsilon'^2}$ and substituting into (8), we get that with probability at least $1 - \delta$ over the choice of S, for every $f \in \mathrm{conv}_T(\mathcal{H})$

$$\begin{aligned}\Pr_D[\|f(X) - p(X)\|_{\mathrm{TV}} > \epsilon + \epsilon'] \quad &\leq \quad \Pr_S[\|f(X) - p(X)\|_{\mathrm{TV}} > \epsilon] + \\ &O\left(\sqrt{\frac{1}{n}\left(\frac{dk^2 \log n}{\epsilon'^2} \log^2\left(\frac{nk}{\epsilon'}\right) + \log\frac{1}{\delta}\right)}\right).\end{aligned}$$

□

Risk bounds for Lipschitz functions in doubling spaces

We will make the following structural assumption about the instance space $\mathcal{X}$ and the hypothesis class $\mathcal{H}$. We take $(\mathcal{X}, \rho)$ to be a metric space and endow the space of distributions on $\mathcal{Y}$ with the $\|\cdot\|_{\mathrm{TV}}$ metric. Then our assumptions are

(i) the target hypothesis p^* and all members of $\mathcal{H}$ are Lipschitz-continuous with respect to ρ and $\|\cdot\|_{\mathrm{TV}}$: there is an $L > 0$ such that

$$\|h(x) - h(x')\|_{\mathrm{TV}} \leq L\rho(x, x'), \qquad x, x' \in \mathcal{X} \tag{9}$$

holds for each $h \in \mathcal{H}$

(ii) $(\mathcal{X}, \rho)$ has a finite doubling dimension: $\mathrm{ddim}(\mathcal{X}) < \infty$ (Gupta, Krauthgamer, and Lee 2003; Krauthgamer and Lee 2004).

Given our collection $\mathcal{H}$ of L-Lipschitz conditional distributions, consider an $h \in \mathcal{H}$ and its $\|\cdot\|_{\mathrm{TV}}$ loss with respect to the target conditional distribution function $p^* : \mathcal{X} \to [0,1]^{\mathcal{Y}}$ on an instance $x \in \mathcal{X}$:

$$f_{h,p^*}(x) = \|h(x) - p^*(x)\|_{\mathrm{TV}} \,. \tag{10}$$

For a fixed conditional distribution function p^*, define $\mathcal{F}_{p^*} = \{f_{h,p^*} : h \in \mathcal{H}\}$ to be the collection of all such functions; each has a Lipschitz constant of at most $2L$. It follows from Corollary 3 in (Gottlieb, Kontorovich, and Krauthgamer 2010) that

$$\mathrm{fat}_\gamma(\mathcal{F}_{p^*}) \leq \left\lceil \frac{L\,\mathrm{diam}(\mathcal{X})}{\gamma} \right\rceil^{\mathrm{ddim}(\mathcal{X})+1},$$

where $\mathrm{ddim}(\mathcal{X})$ is the doubling dimension of $\mathcal{X}$.

Define the *sample risk* of a hypothesis h by

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n} \left\|h(X_i) - p^{X_i}\right\|_{\mathrm{TV}}$$

and the expected risk by $R(h) = \mathbf{E}R_n(h)$.

Then we have, via (Alon et al. 1997), that the empirical risk cannot exceed the true risk by much:

Theorem 7. *For all $\epsilon \geq \sqrt{2/n}$,*

$$\Pr\left[\sup_{h \in \mathcal{H}}(R(h) - R_n(h)) > \epsilon\right] \quad \leq \quad 24n\left(\frac{288n}{\epsilon^2}\right)^{d\log(24en/\epsilon)} \exp(-\epsilon^2 n/36),$$

where $d = \mathrm{fat}_{\epsilon/24}(\mathcal{F}_{p^*}) \leq \left\lceil \frac{24L\,\mathrm{diam}(\mathcal{X})}{\epsilon} \right\rceil^{\mathrm{ddim}(\mathcal{X})+1}$.

Note that this bound is both distribution- and dimension-free.

Dimension-free bounds

We remarked that the accuracy bound in Theorem 4 is not dimension-free since it depends on the number of labels $k = |\mathcal{Y}|$. This quantity enters the bound via Corollary 3, which states that for i.i.d. vectors $Z_i \in \mathbb{R}^k$ with $\mathbf{E}[Z_i] = \mathbf{0}$ and $\|Z_i\|_1 \leq 1$, we have

$$\Pr\left[\left\|\sum_{i=1}^{n} Z_i\right\|_1 > t\right] \leq \exp\left(-\frac{t^2}{8nk}\right)$$

for $t \geq 4\sqrt{nk}$. Thus, the key to obtaining a dimension-free accuracy bound is a Banach-space version of Hoeffding's inequality. Such a result was obtained by (Talagrand 1996):

$$\Pr\left[\left|\left\|\sum_{i=1}^{n} Z_i\right\|_1 - \mathbf{E}\left\|\sum_{i=1}^{n} Z_i\right\|_1\right| > t\right] \leq C\exp\left(-\frac{ct^2}{n}\right) \tag{11}$$

for some universal constants $c, C > 0$.

In order to make use of (11), we need some estimate on $V_n = \mathbf{E}\left\|\sum_{i=1}^{n} Z_i\right\|_1$. Recall that in our applications, $Z_i = h_i(x) - f(x)$ is the deviation between a random element of $h \in \mathcal{H}$ and $f \in \mathrm{conv}_T(\mathcal{H})$. Thus, in some sense, V_n is a proxy for the variance of the random process. We would like for V_n to be $o(n)$, but this cannot be guaranteed in general. Indeed, suppose that $k \gg n$ and $Z_i = \pm\mathbf{e}_j$ chosen uniformly at random, where $\mathbf{e}_j$ is the j^{th} basis element of $\mathbb{R}^k$. Since k is large, cancellations in the sum can be made arbitrarily improbable, and in this case $V_n = \Omega(n)$.

This example, however, is degenerate, since the Z_i are extremely "peaked" or "concentrated". The situation becomes better if we make a smoothness assumption:

Theorem 8. *Let Z_i be i.i.d. random variables in $\mathbb{R}^k$ with $\mathbf{E}[Z_i] = \mathbf{0}$ and $\|Z_i\|_\infty = O(1/k)$. Then $\mathbf{E}\left\|\sum_{i=1}^{n} Z_i\right\|_1 = O(\sqrt{n})$.*

Remark: We thank Gideon Schechtman for this result and the example above.

Proof. Our probability space is actually the Hilbert space $L_2(\ell_2^k)$ with the inner product $\langle Z_i, Z_j\rangle = \mathbf{E}[Z_i \cdot Z_j]$, where $Z_i \cdot Z_j$, is the standard dot product in $\mathbb{R}^k$. By independence and $\mathbf{E}[Z_i] = \mathbf{0}$, we have $\langle Z_i, Z_j\rangle = \mathbf{E}[Z_i] \cdot \mathbf{E}[Z_j] = 0$ for $i \neq j$, and so the Z_i are orthogonal as elements of $L_2(\ell_2)$. Thus,

$$\left(\mathbf{E}\left\|\sum_{i=1}^{n} Z_i\right\|_2\right)^2 \leq \mathbf{E}\left\|\sum_{i=1}^{n} Z_i\right\|_2^2 = \mathbf{E}\sum_{i=1}^{n} \|Z_i\|_2^2 = O\left(\frac{n}{k}\right)$$

where the first inequality is Jensen's, the first identity follows by orthogonality, and the estimate $O(n/k)$ holds since $\|Z_i\|_\infty = O(1/k)$. Now $\|Z\|_1 \leq \sqrt{k}\,\|Z\|_2$, and so

$$\mathbf{E}\left\|\sum_{i=1}^{n} Z_i\right\|_1 \leq \sqrt{k}\mathbf{E}\left\|\sum_{i=1}^{n} Z_i\right\|_2 \leq \sqrt{k\mathbf{E}\left\|\sum_{i=1}^{n} Z_i\right\|_2^2} = O(\sqrt{n}).$$

□

Thus, if we can guarantee $\mathbf{E}\left\|\sum_{i=1}^{n} Z_i\right\|_1 = O(\sqrt{n})$ (either via the assumption $\|Z_i\|_\infty = O(1/k)$ or by exploiting some special structure) we can remove the dependence on k from the accuracy bound in Theorem 4.

References

Alon, N.; Ben-David, S.; Cesa-Bianchi, N.; and Haussler, D. 1997. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM* 44(4):615–631.

Bartlett, P., and Shawe-Taylor, J. 1999. Generalization performance of support vector machines and other pattern classifiers. 43–54.

Das, D., and Petrov, S. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, 600–609. Stroudsburg, PA, USA: Association for Computational Linguistics.

Devroye, L., and Lugosi, G. 2001. *Combinatorial methods in density estimation*. Springer Series in Statistics. New York: Springer-Verlag.

Duffy, N., and Helmbold, D. 2002. Boosting methods for regression. *Machine Learning* 47:153–200.

Fan, W.; Stolfo, S. J.; Zhang, J.; and Chan, P. K. 1999. Adacost: Misclassification cost-sensitive boosting. In *ICML*, 97–105.

Freund, Y., and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55(1):119–139.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2000. Additive logistic regression: A statistical view of boosting. *Annals of Statistics* 28:337–374.

Gibbs, A. L., and Su, F. E. 2002. On choosing and bounding probability metrics. *International Statistical Review* 70(3):419–435.

Gottlieb, L.-A.; Kontorovich, L.; and Krauthgamer, R. 2010. Efficient classification for metric data. In *COLT*, 433–440.

Gupta, A.; Krauthgamer, R.; and Lee, J. R. 2003. Bounded geometries, fractals, and low-distortion embeddings. In *FOCS*, 534–543.

Kanamori, T. 2010. Deformation of log-likelihood loss function for multiclass boosting. *Neural Networks* 23(7):843 – 864.

Krauthgamer, R., and Lee, J. R. 2004. Navigating nets: Simple algorithms for proximity search. In *15th Annual ACM-SIAM Symposium on Discrete Algorithms*, 791–801.

McDiarmid, C. 1989. On the method of bounded differences. In Siemons, J., ed., *Surveys in Combinatorics, volume 141 of LMS Lecture Notes Series*. Morgan Kaufmann Publishers, San Mateo, CA. 148–188.

Mease, D.; Wyner, A. J.; and Buja, A. 2007. Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research* 8:409–439.

Schapire, R. E.; Freund, Y.; Bartlett, P.; and Lee, W. S. 1998. Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Statist.* 26(5):1651–1686.

Talagrand, M. 1996. New concentration inequalities in product spaces. *Invent. Math.* 126(3):505–563.

Toutanova, K., and Cherry, C. 2009. A global model for joint lemmatization and part-of-speech prediction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ACL '09, 486–494.