Efficient algorithms for learning functions with bounded variation

Philip M. Long

Genome Institute of Singapore
1 Science Park Road
The Capricorn, #05-01
Singapore 117528, Republic of Singapore
gislongp@nus.edu.sg

Abstract

We show that the class $\mathcal{F}_{\mathrm{BV}}$ of $[0,1]$-valued functions with total variation at most 1 can be agnostically learned with respect to the absolute loss in polynomial time from $O\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$ examples, matching a known lower bound to within a constant factor. We establish a bound of $O\left(\frac{1}{m}\right)$ on the expected error of a polynomial-time algorithm for learning $\mathcal{F}_{\mathrm{BV}}$ in the prediction model, also matching a known lower bound to within a constant factor. Applying a known algorithm transformation to our prediction algorithm, we obtain a polynomial-time PAC learning algorithm for $\mathcal{F}_{\mathrm{BV}}$ with a sample complexity bound of $O\left(\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$; this also matches a known lower bound to within a constant factor.

Key words: Statistical learning theory, computational learning theory, sample complexity, bounded variation, nonparametric regression.

1 Introduction

The total variation of a function can be viewed as the overall tendency for similar inputs to yield similar outputs. In this paper, we present polynomial-time algorithms for learning arbitrary members of the class $\mathcal{F}_{\mathrm{BV}}$ of $[0,1]$-valued functions with total variation at most 1 according to three theoretical models of the learning problem. The number of examples needed by each of the algorithms is within a constant factor of optimal. Throughout, we will measure the error of a prediction $\hat{y}$ of a real-valued quantity y with $|\hat{y}-y|$.

In the agnostic learning model [8,13], random examples $(x_1,y_1),...,(x_m,y_m)$ are drawn from an arbitrary joint distribution P, and the goal of the learning

algorithm is to output a function h such that the expected value of $|h(x) - y|$ for another pair (x, y) drawn according to P is nearly as small as that for the best function in $\mathcal{F}$.

We show that an algorithm, given $O\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$ examples, outputs a hypothesis with error at most ϵ worst than the best in $\mathcal{F}_{\mathrm{BV}}$ with probability at least $1-\delta$. This analysis uses a technique called Chaining (see [19,20]) from Empirical Process Theory. In [17], we applied this technique to obtain improved bounds for agnostic concept learning in a drifting environment. Please refer to that paper and [19] for high-level descriptions of Chaining.

A packing number for a class of functions measures the number of significantly different behaviors that functions in the class can have on a certain number of domain elements. While packing bounds for $\mathcal{F}_{\mathrm{BV}}$ were known [1,4,3], we needed new bounds for our application (the difference is described immediately after the proof of Lemma 3).

Our agnostic learning bound improves on the bound of $O\left(\frac{1}{\epsilon^2}\left(\frac{1}{\epsilon}+\log\frac{1}{\delta}\right)\right)$ that is obtained by combining packing bounds from [3] with the most commonly applied uniform convergence bounds in terms of packing numbers (see [19,8]). Straightforward application of Simon's [23] techniques yields a lower bound that matches our upper bound to within a constant factor (see Proposition 2).

Lee, Bartlett and Williamson [15] proved a bound of $\tilde{O}(d/\epsilon)$ on the sample complexity of agnostically learning any convex class $\mathcal{F}$ of functions with respect to the *quadratic* loss, where d is the pseudo-dimension [19] of $\mathcal{F}$. One can apply a bound implicit in this analysis (in terms of packing numbers for $\mathcal{F}$) together with known packing bounds [1,4,3] to get bounds on the sample complexity of agnostically learning $\mathcal{F}_{\mathrm{BV}}$ with respect to the quadratic loss similar to the bounds we present in this paper for the absolute loss.[1] However, the bounds of [15] for learning convex classes with respect to the quadratic loss do not appear to have a counterpart when the absolute loss is used. The class of all constant functions has pseudo-dimension 1 and is convex, but, again, straightforward application of Simon's [23] techniques yields a lower bound of $\Omega(\frac{1}{\epsilon^2})$ on the sample complexity of agnostically learning this class with respect to the absolute loss (see Proposition 2). Our analysis does not use the convexity of $\mathcal{F}_{\mathrm{BV}}$: the same bound holds for any function class with a packing bound like that we prove for $\mathcal{F}_{\mathrm{BV}}$.

Our sample complexity bound holds for any algorithm that outputs a hypothesis that minimizes the error on the examples. We show how to achieve this in polynomial time using linear programming.

[1] We thank Peter Bartlett for pointing this out.

In the prediction model [10], an algorithm is given examples

$$(x_1, f(x_1)), ..., (x_{m-1}, f(x_{m-1}))$$

of the behavior of an unknown function f chosen from a known class $\mathcal{F}$, and outputs a hypothesis h. A learning algorithm is evaluated by the expectation, over $x_1, ..., x_m$ drawn independently at random from a fixed, arbitrary probability distribution, of $|h(x_m) - f(x_m)|$. We prove a $\frac{1}{m} + \frac{1}{m(m-1)}$ upper bound on the expected error of a polynomial-time algorithm for learning $\mathcal{F}_{\mathrm{BV}}$ in this model, improving on the best previously known bound of $O\left(\frac{\log m}{m}\right)$ [21], and matching a known lower bound [21] of $\frac{1}{2m}$ to within a constant factor. Our algorithm is new, but one can modify our proof to establish an upper bound of $\frac{2}{m}$ for the nearest-neighbor algorithm.

Applying a known algorithm transformation [9] to our prediction algorithm, one gets a bound of $O(\frac{1}{\epsilon}\log\frac{1}{\delta})$ on the sample complexity of learning $\mathcal{F}_{\mathrm{BV}}$ in the PAC model; i.e., given $O(\frac{1}{\epsilon}\log\frac{1}{\delta})$ independent examples of the behavior of any $f \in \mathcal{F}_{\mathrm{BV}}$, the resulting algorithm, with probability at least $1-\delta$, outputs a hypothesis h such that the expectation of $|h(x) - f(x)|$ is at most ϵ. This improves on the best previously known bound of $O\left(\frac{1}{\epsilon}\left(\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$ [24], and matches a known lower bound [24] to within a constant factor.

2 Preliminaries

Denote the reals by $\mathbf{R}$, the rationals by $\mathbf{Q}$ and the positive integers by $\mathbf{N}$. Let $Y = \mathbf{Q} \cap [0, 1]$.

Define an *example* to be an element of $\mathbf{Q} \times Y$, and a *sample* to be a finite sequence of examples. A *learning algorithm* takes a sample as input, and outputs a *hypothesis*, which is a function from $\mathbf{Q}$ to Y. We will refer to a learning algorithm and the corresponding mapping from inputs to outputs interchangeably.

Choose a set X. For a metric ρ on X, $\epsilon > 0$ and $S \subseteq X$, define $\mathcal{M}(\rho, \epsilon, S)$ to be the size of the largest subset of S whose elements are pairwise at a distance greater than ϵ, as measured by ρ. Define $\mathcal{N}(\rho, \epsilon, S)$ to be the size of the smallest set $T \subseteq X$ such that each element of S is within distance ϵ (as measured by ρ) of some element of T. We will use the following general inequalities [14]:

$$\mathcal{M}(\rho, 2\epsilon, S) \leq \mathcal{N}(\rho, \epsilon, S) \leq \mathcal{M}(\rho, \epsilon, S). \tag{1}$$

For $d, p \in \mathbf{N}$, $\vec{v}, \vec{w} \in \mathbf{R}^d$, define

$$\ell_p(\vec{v}, \vec{w}) = \left(\frac{1}{d} \sum_{i=1}^{d} |v_i - w_i|^p \right)^{1/p}.$$

If P is a probability distribution, denote by P^m the distribution obtained by sampling m times independently from P.

Let $\mathcal{F}_{\mathrm{BV}}$ be the set of all functions f from $\mathbf{Q}$ to Y for which for all $x_1 < ... < x_n$, $\sum_{i=1}^{n-1} |f(x_i) - f(x_{i+1})| \leq 1$.

3 Agnostic learning

We begin by studying $\mathcal{F}_{\mathrm{BV}}$ in the agnostic learning model [8]. For a probability distribution P over $\mathbf{Q} \times Y$ and a function f from $\mathbf{Q}$ to Y, the error of f is defined by $\mathbf{er}_P(f) = \int |f(x) - y| \; dP(x, y)$. For $\epsilon, \delta > 0$, and $m \in \mathbf{N}$, we say a class $\mathcal{F}$ of functions from $\mathbf{Q}$ to Y is (ϵ, δ)-*agnostically learnable from* m *examples* if there is a learning algorithm A such that, for all probability distributions P on $\mathbf{Q} \times Y$, if a sample S is obtained by drawing m times independently at random according to P, and is passed to algorithm A, then, with probability at least $1-\delta$, the resulting output $A(S)$ satisfies $\mathbf{er}_P(A(S)) \leq \epsilon + \inf_{f \in \mathcal{F}} \mathbf{er}_P(f)$.

The algorithm that we will consider minimizes the total absolute loss on the examples from among hypotheses in $\mathcal{F}_{\mathrm{BV}}$. As usual [6,8], our analysis of this algorithm will proceed by showing that uniformly good estimates of the errors of the hypotheses in $\mathcal{F}_{\mathrm{BV}}$ can be obtained.

Choose a countable set Z.

Lemma 1 (see [19]) *Choose a set $\mathcal{G}$ of functions from Z to $[0,1]$, $\epsilon > 0$, $m \in \mathbf{N}$ for which $m \geq 3/\epsilon^2$, and a probability distribution D over Z. Then if U is the uniform distribution over $\{-1,1\}^m$, we have*

$$D^m \left\{ \vec{z} : \exists g \in \mathcal{G}, \left| \left(\frac{1}{m} \sum_{i=1}^{m} g(z_i) \right) - \int_Z g(z) \; dD(z) \right| > \epsilon \right\}$$
$$\leq 2 \sup_{\vec{z} \in Z^{2m}} U \left\{ \vec{\sigma} : \exists g \in \mathcal{G}, \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i (g(z_i) - g(z_{m+i})) \right| > \epsilon/2 \right\}.$$

Lemma 2 (see [19]) *Let $Y_1, ..., Y_m$ be independent random variables taking*

values in $[a_1, b_1], ..., [a_m, b_m]$ *respectively. Then*

$$\Pr\left(\left|\left(\sum_{i=1}^{m} Y_i\right) - \left(\sum_{i=1}^{m} \mathbf{E}(Y_i)\right)\right| > \eta\right) \leq 2\exp\left(\frac{-2\eta^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right).$$

The following lemma, which is proved using a chaining argument (see [20] and [16] for descriptions of Chaining), is the main part of our analysis.

Lemma 3 *Choose* $m \in \mathbf{N}$ *and* $G \subseteq [-1,1]^m$. *If there is a constant* $k \geq 1$ *such that for all* $0 < \alpha \leq 1/2$,

$$\mathcal{M}(\ell_2, G, \alpha) \leq \exp\left(\frac{k}{\alpha}\ln\frac{1}{\alpha}\right),$$

and if U *is the uniform distribution over* $\{-1,1\}^m$, *then for all* $m \geq 288k/\eta^2$,

$$U\left\{\vec{\sigma} : \exists g \in G, \left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i g_i\right| > \eta\right\} \leq 4e^{-\eta^2 m/288}.$$

Proof: Construct a sequence $G_0, G_1, ...$ of approximations to G as follows. Let $G_0 = \emptyset$, and for each $j \in \mathbf{N}$, construct G_j by initializing it to G_{j-1}, and as long as there is a $g \in G$ that has ℓ_2 distance greater than $\frac{1}{2^j}$ from each element of G_j, choosing such a g and adding it to G_j.

Note that $G_0 \subseteq G_1 \subseteq ....$ For each $g \in G$ and $j \in \mathbf{N}$, choose an element $\psi_j(g)$ of G_j from among those that minimize the ℓ_2 distance to g; note that $\ell_2(g, \psi_j(g)) \leq 1/2^j$. Let $H_1 = G_1$. For each $j > 1$, define H_j by

$$H_j = \{g - \psi_{j-1}(g) : g \in G_j\}.$$

Note that since for all $g \in G$, $\ell_2(g, \psi_{j-1}(g)) \leq 1/2^{j-1}$, for each $h \in H_j$, $\sum_{i=1}^{m} h_i^2 \leq m/4^{j-1}$.

By induction, for each $n \in \mathbf{N}$, for each $g \in G_n$, there exist $h_{g,1} \in H_1, ..., h_{g,n} \in H_n$ such that $g = \sum_{j=1}^{n} h_{g,j}$. Let $G_* = \cup_n G_n$. Since $G_1, G_2, ...$ form arbitrary fine covers of G, G_* is dense in G with respect to ℓ_2. Thus, for each $g \in G$ and each $n \in \mathbf{N}$, there exist $h_{g,1} \in H_1, ..., h_{g,n} \in H_n$ such that $\ell_2(g, \sum_{j=1}^{n} h_{g,j}) \leq 1/2^n$. Therefore, there exist $h_{g,1} \in H_1, h_{g,2} \in H_2, ...$ such that $g = \sum_{j=1}^{\infty} h_{g,j}$. Let

$$p = U\left\{\vec{\sigma} : \exists g \in G, \left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i g_i\right| > \eta\right\}$$

denote the quantity we wish to upper bound. Since G_* is dense in G,

$$p = U\left\{\vec{\sigma} : \exists g \in G_*,\ \left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i g_i\right| > \eta\right\}$$

Expressing g as $\sum_{j=1}^{\infty} h_{g,j}$, we get

$$p = U\left\{\vec{\sigma} : \exists g \in G_*,\ \left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i \left(\sum_{j=1}^{\infty}(h_{g,j})_i\right)\right| > \eta\right\}.$$

Pulling out the sum over j, we get

$$p = U\left\{\vec{\sigma} : \exists g \in G_*,\ \left|\sum_{j=1}^{\infty}\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h_{g,j})_i\right| > \eta\right\},$$

which, applying the triangle inequality, implies that

$$p \leq U\left\{\vec{\sigma} : \exists g \in G_*,\ \sum_{j=1}^{\infty}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h_{g,j})_i\right| > \eta\right\}.$$

For each $j \in \mathbf{N}$, let $\eta_j = (\eta/6)\sqrt{j/2^{j-1}}$. Then $\sum_{j=1}^{\infty}\eta_j \leq \eta$, and therefore

$$p \leq U\left\{\vec{\sigma} : \exists g \in G_*, j \in \mathbf{N}, \left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h_{g,j})_i\right| > \eta_j\right\}.$$

Replacing the disjunction over j with a sum, we get

$$p \leq \sum_{j=1}^{\infty} U\left\{\vec{\sigma} : \exists g \in G_*, \left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h_{g,j})_i\right| > \eta_j\right\}.$$

Since each $h_{g,j} \in H_j$, we have

$$p \leq \sum_{j=1}^{\infty} U\left\{\vec{\sigma} : \exists h \in H_j, \left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i h_i\right| > \eta_j\right\}.$$

Choose $j \in \mathbf{N}$. Since for each $h \in H_j$, $\sum_{i=1}^{m} h_i^2 \leq m/4^{j-1}$, applying Lemma 2,

$$U\left\{\vec{\sigma} : \left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i h_i\right| > \eta_j\right\} \leq 2e^{-\eta_j^2 4^{j-1} m/2}.$$

Thus

$$p \leq \sum_{j=1}^{\infty} 2|H_j|e^{-\eta_j^2 4^{j-1}m/2},$$

and substituting the definition of η_j yields

$$p \leq \sum_{j=1}^{\infty} 2|H_j|e^{-\eta^2 j 2^j m/144}.$$

By construction, each pair of elements of G_j have ℓ_2 distance more than $1/2^j$. By the assumed bounds on $\mathcal{M}(\ell_2, G, \cdot)$,

$$|H_j| \leq |G_j| \leq e^{kj2^j},$$

which implies, twice using the bound $m \geq 288k/\eta^2$, that

$$\begin{aligned}
p &\leq 2\sum_{j=1}^{\infty} e^{(k-\eta^2 m/144)j2^j} \\
&\leq 2\sum_{j=1}^{\infty} e^{-(\eta^2 m/288)j2^j} \\
&\leq 2\sum_{j=1}^{\infty} e^{-(\eta^2 m/288)j} \\
&= 2\frac{e^{-\eta^2 m/288}}{1-e^{-\eta^2 m/288}} \\
&\leq 4e^{-\eta^2 m/288},
\end{aligned}$$

completing the proof. □

Packing bounds for $\mathcal{F}_{\mathrm{BV}}$ are known [1,4,3], but to apply Lemma 3 we need bounds for ℓ_2 that are independent of m, and we are not aware of previously known bounds of this type.

For each $m \in \mathbf{N}$, define

$$A_m = \{\vec{a} \in [0,1]^m : a_1 \leq ... \leq a_m\}$$

and

$$C_m = \{\vec{c} \in [0,1]^m : c_1 = ... = c_m\}.$$

For each $x_1 < ... < x_m$, each

$$\vec{f} \in \{(f(x_1), ..., f(x_m)) : f \in \mathcal{F}_{\mathrm{BV}}\}$$

has $\vec{a}_1, \vec{a}_2 \in A_m$ and $\vec{c} \in C_m$ such that

$$\vec{f} = \vec{c} + \vec{a}_1 - \vec{a}_2$$

(see [22]), so we will work on A_m (C_m is easy).

As in [3], we will make use of an approximation to A_m by a class of piecewise constant functions.[2] For $\kappa > 0$, construct $A_{\kappa,m}$ by dividing the indices $\{1, ..., m\}$ into bins, putting roughly the first κm indices into the first bin, the next κm indices into the second bin, and so on, then letting $A_{\kappa,m}$ be the subset of A_m for which the components in each bin are equal. Specifically,

$$A_{\kappa,m} = \{\vec{a} \in A_m : \forall i, j, \lfloor i/(\kappa m) \rfloor = \lfloor j/(\kappa m) \rfloor \Rightarrow a_i = a_j\}.$$

First, we bound how well $A_{\kappa,m}$ approximates A_m.

Lemma 4 *For all $\kappa > 0$, for each $\vec{a} \in A_m$, there is an $\vec{a}' \in A_{\kappa,m}$ such that $\ell_2(\vec{a}, \vec{a}') \leq \sqrt{\kappa}$.*

Proof: For each j, let $B_j = \{i : \lfloor i/(\kappa m) \rfloor = j\}$ be the jth bin. Choose $\vec{a} \in A_m$. Define $\vec{a}'$ by, for each bin B_j, for each index $i \in B_j$, setting a'_i to be the average of the components of $\vec{a}$ whose indices are in B_j; i.e., $a'_i = \frac{1}{|B_j|} \sum_{i \in B_j} a_i$. (See Figure 1.) Then

$$\begin{aligned}
\ell_2(\vec{a}, \vec{a}') &= \sqrt{\frac{1}{m} \sum_j \sum_{i \in B_j} (a_i - a'_i)^2} \\
&\leq \sqrt{\frac{1}{m} \sum_j \sum_{i \in B_j} ((\max_{i \in B_j} a_i) - (\min_{i \in B_j} a_i))^2} \\
&= \sqrt{\frac{1}{m} \sum_j |B_j| ((\max_{i \in B_j} a_i) - (\min_{i \in B_j} a_i))^2} \\
&\leq \sqrt{\kappa \sum_j ((\max_{i \in B_j} a_i) - (\min_{i \in B_j} a_i))^2} \\
&= \sqrt{\kappa} \sqrt{\sum_j ((\max_{i \in B_j} a_i) - (\min_{i \in B_j} a_i))^2}
\end{aligned}$$

[2] Kearns and Schapire [12] described an algorithm for learning monotone p-concepts using piecewise constant hypotheses.

Fig. 1. A plot of an example of $\vec{a}$ (pictured using circles) and the corresponding $\vec{a}'$ (pictured using squares) from the proof of Lemma 4. The bin boundaries are shown using dotted lines.

$$\leq \sqrt{\kappa} \sum_j ((\max_{i \in B_j} a_i) - (\min_{i \in B_j} a_i))$$
$$\leq \sqrt{\kappa},$$

since $\vec{a} \in A_m$, completing the proof. $\Box$

Choose $m \in \mathbf{N}$. Say that $G \subseteq \mathbf{R}^m$ shatters a sequence $(i_1, r_1), ..., (i_d, r_d)$ of elements of $\{1, ..., m\} \times \mathbf{R}$ if for each $\vec{b} \in \{0,1\}^d$, there is a $g \in G$ such that for all $j \in \{1, ..., d\}$, $b_j = 1 \Leftrightarrow g_{i_j} \geq r_j$. The *pseudo-dimension* [19] of G is the length of the longest sequence shattered by G.

Lemma 5 ([11]) *If $V \subseteq \mathbf{R}$ is finite and $G \subseteq V^m$ has pseudo-dimension d, then $|G| \leq (em|V|/d)^d$.*

Let $A_{\kappa,\beta,m} = A_{\kappa,m} \cap \{\beta, ..., \beta\lfloor 1/\beta \rfloor\}^m$.

Lemma 6 *The pseudo-dimension of $A_{\kappa,\beta,m}$ is at most $\lfloor 1/\beta \rfloor$.*

Proof: Suppose $(i_1, r_1), ..., (i_d, r_d)$ with $i_1 < ... < i_d$ is shattered by $A_{\kappa,\beta,m}$. Since each component of each element of $A_{\kappa,\beta,m}$ is a positive multiple of β, we may assume without loss of generality that each of $r_1, ..., r_d$ is a positive multiple of β.

We claim that $r_1 < ... < r_d$. Assume for contradiction that there was a k such that $r_k \geq r_{k+1}$. The definition of shattering implies that there is a $\vec{a} \in A_{\kappa,\beta,m}$ such that such that $a_{i_j} \geq r_j$ and $a_{i_{j+1}} < r_{j+1}$, which then implies that $a_{i_j} > a_{i_{j+1}}$. But since $i_j < i_{j+1}$, this contradicts that fact that $\vec{a} \in A_{\kappa,\beta,m} \subseteq A_m$.

Since $r_1 < ... < r_d$, each of them are multiples of β, and they are all in $(0, 1]$, $d \leq \lfloor 1/\beta \rfloor$, completing the proof. $\Box$

Lemma 7 *For any $x_1 \leq ... \leq x_m$, and any $0 < \alpha \leq 1/8$, if*

$$F = \{(f(x_1), ..., f(x_m)) : f \in \mathcal{F}_{\mathrm{BV}}\},$$

then

$$\mathcal{N}(\ell_2, \alpha, F) \leq \exp\left(\frac{110}{\alpha} \ln \frac{1}{4\alpha}\right).$$

Proof: Let $\beta = \alpha/8$, $\kappa = \alpha^2/64$. Lemmas 5 and 6 imply that

$$|A_{1/m,\beta,m}| \leq (em)^{\lfloor 1/\beta \rfloor}. \tag{2}$$

Recall that in the definition of $A_{\kappa,\beta,m}$, the indices $1, ..., m$ are divided into $\lceil 1/\kappa \rceil$ bins, and all elements of $A_{\kappa,\beta,m}$ are constrained to have the same value in components whose indices are in the same bin (see Figure 1). Thus, by replacing each bin with a single component, elements of $A_{\kappa,\beta,m}$ can be put in 1-1 correspondence with elements of $A_{1/\lceil 1/\kappa \rceil,\beta,\lceil 1/\kappa \rceil}$. Therefore (2) implies that

$$|A_{\kappa,\beta,m}| = |A_{1/\lceil 1/\kappa \rceil,\beta,\lceil 1/\kappa \rceil}| \leq (e\lceil 1/\kappa \rceil)^{\lfloor 1/\beta \rfloor}.$$

Each $\vec{f} \in F$ has $\vec{c} \in C_m, \vec{a}_1, \vec{a}_2 \in A_m$ such that $\vec{f} = \vec{c} + \vec{a}_1 - \vec{a}_2$ [22]. Thus, if $\hat{C}_m$ is an $\alpha/2$-cover of C_m, and $\hat{A}_m$ is an $\alpha/4$-cover of A_m, then $\{\vec{c} + \vec{a}_1 - \vec{a}_2 : \vec{c} \in \hat{C}_m, \vec{a}_1, \vec{a}_2 \in \hat{A}_m\}$ is an α-cover of F. This implies that

$$\mathcal{N}(\ell_2, \alpha, F) \leq \mathcal{N}(\ell_2, \alpha/2, C_m)\mathcal{N}(\ell_2, \alpha/4, A_m)^2. \tag{3}$$

By Lemma 4,

$$\mathcal{N}(\ell_2, \sqrt{\kappa} + \beta, A_m) \leq |A_{\kappa,\beta,m}|.$$

Substituting the definitions of β and κ, we get

$$\mathcal{N}(\ell_2, \alpha/4, A_m) \leq |A_{\kappa,\beta,m}|,$$

and plugging into (3), we get

$$\mathcal{N}(\ell_2, \alpha, F) \leq \mathcal{N}(\ell_2, \alpha/2, C_m)|A_{\kappa,\beta,m}|^2.$$

Since $\alpha \leq 1/8$, substituting the values of β and κ and carrying out simple calculations shows that

$$\begin{aligned}\mathcal{N}(\ell_2, \alpha, F) &\leq \lfloor 2/\alpha \rfloor (e\lceil 1/\kappa \rceil)^{\lfloor 1/\beta \rfloor} \\ &\leq \exp\left(\frac{110}{\alpha} \ln \frac{1}{4\alpha}\right),\end{aligned}$$

completing the proof. $\square$

Lemma 8 ([18]) *For any $(x_1, y_1), ..., (x_m, y_m) \in \mathbf{R} \times [0,1]$, for any set $\mathcal{F}$ of functions from $\mathbf{R}$ to $[0,1]$,*

$$\begin{aligned}&\mathcal{N}(\ell_2, \alpha, \{(|f(x_1) - y_1|, ..., |f(x_m) - y_m|) : f \in \mathcal{F}\}) \\ &\leq \mathcal{N}(\ell_2, \alpha, \{(f(x_1), ..., f(x_m)) : f \in \mathcal{F}\}).\end{aligned}$$

Lemma 9 *For any $G \subseteq [0,1]^{2m}$,*

$$\begin{aligned}&\mathcal{N}(\ell_2, \alpha, \{(g_1 - g_{m+1}, ..., g_m - g_{2m}) : g \in G\}) \\ &\leq \mathcal{N}(\ell_2, \alpha/2, G).\end{aligned}$$

Proof: Choose $g, h \in [0,1]^{2m}$.

$$\begin{aligned}&\ell_2((g_1 - g_{m+1}, ..., g_m - g_{2m}), (h_1 - h_{m+1}, ..., h_m - h_{2m})) \\ &= \sqrt{\frac{1}{m}\sum_{i=1}^{m}((g_i - g_{m+1}) - (h_i - h_{m+i}))^2} \\ &= \sqrt{\frac{1}{m}\sum_{i=1}^{m}((g_i - h_i) + (h_{m+i} - g_{m+i}))^2} \\ &\leq \sqrt{\frac{1}{m}\sum_{i=1}^{m}2((g_i - h_i)^2 + (h_{m+i} - g_{m+i})^2)} \quad (\text{since } \forall u, v, (u+v)^2 \leq 2(u^2+v^2)) \\ &= 2\ell_2(g, h).\end{aligned}$$

Thus, an $\alpha/2$-cover for G can be used to construct an α-cover for $\{(g_1 - g_{m+1}, ..., g_m - g_{2m}) : g \in G\}$. $\square$

Theorem 10 *$\mathcal{F}_{\mathrm{BV}}$ is (ϵ, δ)-agnostically learnable from $O\left(\frac{1}{\epsilon^2}\ln\frac{1}{\delta}\right)$ examples.*

Proof: For any function $f : \mathbf{Q} \to Y$, define $L_f : \mathbf{Q} \times Y \to \mathbf{Q}$ by $L_f(x, y) = |f(x) - y|$.

Choose $0 < \alpha \leq 1/2$. Applying Lemmas 8, 9, and 7 together with (1), for all $(x_1, y_1), ..., (x_{2m}, y_{2m}) \in \mathbf{Q} \times Y$,

$$\begin{aligned}&\mathcal{M}(\ell_2, \alpha, \{(L_f(x_1, y_1) - L_f(x_{m+1}, y_{m+1}), ..., L_f(x_m, y_m) - L_f(x_{2m}, y_{2m})) \\ &\qquad\qquad : f \in \mathcal{F}_{\mathrm{BV}}\}) \qquad (4) \\ &\leq \exp\left(\tfrac{440}{\alpha} \ln \tfrac{1}{\alpha}\right).\end{aligned}$$

Assume without loss of generality that $\epsilon \leq 1/2$. Let $m = \left\lceil \frac{1}{\epsilon^2}\left(126720 + 288 \ln \frac{8}{\delta}\right)\right\rceil$. Applying Lemmas 1 and 3, and (4),

$$P^m\{\vec{z} : \exists f \in \mathcal{F}_{\mathrm{BV}}, \left|\left(\frac{1}{m}\sum_{t=1}^{m} L_f(z_t)\right) - \int L_f(u)\ dP(u)\right| > \epsilon/2\} \leq \delta. \qquad (5)$$

Consider some algorithm A that outputs an element of $\mathcal{F}_{\mathrm{BV}}$ which minimizes error on the examples. Then, for any $f_* \in \mathcal{F}_{\mathrm{BV}}$, the triangle inequality and (5) imply that

$$P^m\left\{\vec{z} : \int_{\mathbf{Q}\times Y} L_{A(\vec{z})}(u)\ dP(u) - \int_{\mathbf{Q}\times Y} L_{f_*}(u)\ dP(u) > \epsilon\right\} \leq \delta$$

completing the proof. □

Theorem 10 provides a sample complexity bound for any algorithm that outputs a hypothesis in $\mathcal{F}_{\mathrm{BV}}$ minimizing the error on the sample. Here, using standard techniques, we describe such an algorithm that uses linear programming. Applying efficient linear programming algorithms (e.g. [26]), this algorithm takes time polynomial in the size of its input, where rationals are represented by writing their numerators and denominators in binary.

Suppose the input sample is $(x_1, y_1), ..., (x_m, y_m)$ and that the x_i's have been sorted as a preprocessing step. Then $y_1, ..., y_m$ are treated as constants in the following linear program:

$$\text{minimize } \sum_{i=1}^{m} e_i^+ + e_i^-$$

subject to

$$y_i - \rho_i = e_i^+ - e_i^-, \forall 1 \leq i \leq m \qquad (6)$$

$$\rho_{i+1} - \rho_i = d_i^+ - d_i^-, \forall 1 \leq i \leq m-1 \qquad (7)$$

$$\sum_{i=1}^{m-1} d_i^+ + d_i^- \leq 1 \qquad (8)$$

$$\rho_i = \rho_j, \forall i, j \text{ such that } x_i = x_j \tag{9}$$
$$e_i^+, e_i^-, d_i^+, d_i^- \geq 0, \forall i. \tag{10}$$

Algorithm A^{LP} defines its output hypothesis h as follows: for some x, if x_i is the closest element of $\{x_1, ..., x_m\}$ (with ties broken in favor of the smaller neighbor), then $h(x) = \rho_i$. The constraints in (9) ensure that h is well-defined.

Proposition 1 *For any input* $(x_1, y_1), ..., (x_m, y_m)$*, Algorithm* A^{LP} *outputs* $h \in \mathcal{F}_{\mathrm{BV}}$ *that minimizes* $\sum_{i=1}^m |h(x_i) - y_i|$.

Proof: Fix $(x_1, y_1), ..., (x_m, y_m)$ with $x_1 \leq ... \leq x_m$, and fix the values of the variables in the linear program used by A^{LP} at their optimal values.

First, we claim that A^{LP}'s hypothesis h is in $\mathcal{F}_{\mathrm{BV}}$. Choose $u_1 < ... < u_n$. Then the triangle inequality implies that

$$\sum_{k=2}^{n} |h(u_k) - h(u_{k-1})| \leq \sum_{i=2}^{m} |\rho_i - \rho_{i-1}|.$$

Since the constraints in (10) ensure that the d_i^+'s and d_i^-'s are nonnegative, the constraints in (7), together with the previous inequality, imply that

$$\sum_{k=2}^{n} |h(u_k) - h(u_{k-1})| \leq \sum_{i=2}^{m} \max\{d_i^+, d_i^-\}$$

which in turn implies that

$$\sum_{k=2}^{n} |h(u_k) - h(u_{k-1})| \leq \sum_{i=2}^{m} d_i^+ + d_i^- \leq 1,$$

because of (8). Thus $h \in \mathcal{F}_{\mathrm{BV}}$.

Now, choose some $f \in \mathcal{F}_{\mathrm{BV}}$. Define an alternative setting of the variables in the linear program of A^{LP} as follows. For $i = 1, ..., m$, define $\underline{\rho_i} = f(x_i)$, and define $\underline{e_i^+} = \max\{0, y_i - \underline{\rho_i}\}$ and $\underline{e_i^-} = \max\{0, \underline{\rho_i} - y_i\}$. For $i = 1, ..., m-1$, define $\underline{d_i^+} = \max\{0, \underline{\rho_{i+1}} - \underline{\rho_i}\}$ and $\underline{d_i^-} = \max\{0, \underline{\rho_i} - \underline{\rho_{i+1}}\}$. It is straightforward to verify that this is a feasible solution to A^{LP}'s linear program, and therefore

$$\begin{aligned}\sum_{i=1}^{m} |f(x_i) - y_i| &= \sum_{i=1}^{m} \underline{e_i^+} + \underline{e_i^-} \\ &\geq \sum_{i=1}^{m} e_i^+ + e_i^-\end{aligned}$$

$$
\begin{aligned}
&\geq \sum_{i=1}^{m} \max\{e_i^+, e_i^-\} \\
&\geq \sum_{i=1}^{m} |\rho_i - y_i| \qquad \text{(see (10) and (6))} \\
&= \sum_{i=1}^{m} |h(x_i) - y_i|,
\end{aligned}
$$

completing the proof. □

4 The prediction model

In this section, we consider the prediction model of learning.

For a learning algorithm A, a function f from $\mathbf{Q}$ to Y, a distribution D over $\mathbf{Q}$, and a number m of domain elements, define

$$M(A, f, D, m) = \int_{\mathbf{Q}^m} |(h_{\vec{x},f,A}(x_m) - f(x_m)|\, dD^m(\vec{x}),$$

where

$$h_{\vec{x},f,A} = A((x_1, f(x_1)), ..., (x_{m-1}, f(x_{m-1}))).$$

That is, $M(A, f, D, m)$ is the expected absolute error of A's hypothesis, given $m-1$ random examples of f at domain elements independently drawn according to D. Then, for a set $\mathcal{F}$ of functions from $\mathbf{Q}$ to Y, define

$$M(A, \mathcal{F}, m) = \sup_{f \in \mathcal{F}, D} M(A, f, D, m).$$

Define A^* to be the algorithm that, given $(x_1, y_1), ..., (x_{m-1}, y_{m-1})$, constructs its hypothesis h by first sorting $(x_1, y_1), ..., (x_{m-1}, y_{m-1})$ by the x_i's, yielding $(u_1, v_1), ..., (u_{m-1}, v_{m-1})$. If $x \in \{u_1, ..., u_{m-1}\}$, $f(x)$ is known, so it sets $h(x) = f(x)$. For $x < u_1$, it sets $h(x) = v_1$, and for $x > u_m$ it sets $h(x) = v_m$. Finally, if $x \in (u_i, u_{i+1})$, it sets $h(x) = \frac{i}{m-1} v_i + \left(1 - \frac{i}{m-1}\right) v_{i+1}$. (See Figure 2.) Obviously, A^* is a polynomial-time algorithm.

We will make use of the following lemma. While it is well known, we have included a proof in an appendix for completeness.

Fig. 2. An example of a hypothesis h output by A^*. The value of $h(x)$ is a weighted average of values of a target function at previously seen points on either side of x. The point closer to the middle of the sample is weighted more. Loosely speaking, this is so that each example has equal "influence" on the final hypothesis. (Note that the examples on the ends completely determine the hypothesis' value for domain elements that fall to the left or the right of the sample respectively.)

Lemma 11 *Choose $m \in \mathbf{N}$, a distribution D on $\mathbf{Q}$, and a bounded $\phi : \mathbf{Q}^m \to \mathbf{R}$. Let D' be any distribution on the set Γ of permutations of $\{1, ..., m\}$. Then*

$$\int_{\mathbf{Q}^m} \phi(\vec{x}) dD^m(\vec{x}) \leq \sup_{(x_1,...,x_m) \in X^m} \int_{\Gamma} \phi(x_{\sigma(1)}, ..., x_{\sigma(m)}) D'(\sigma).$$

Proof: In Appendix A. □

For $\vec{x} \in \mathbf{Q}^m, j \in \{1, ..., m\}$, define switch$(\vec{x}, j)$ to be the result of exchanging x_j and x_m. The idea of analyzing a prediction algorithm by averaging over permutations of the domain elements is from [10].

Theorem 12 $M(A^*, \mathcal{F}_{\mathrm{BV}}, m) \leq \frac{1}{m} + \frac{1}{m(m-1)}$.

Proof: Fix an arbitrary $f \in \mathcal{F}_{\mathrm{BV}}$ and a distribution D over $\mathbf{Q}$. Define error : $\mathbf{Q}^m \to Y$ by

$$\mathrm{error}(\vec{x}) = |h_{\vec{x},f,A^*}(x_m) - f(x_m)|.$$

Applying Lemma 11 with the uniform distribution over permutations that

switch some element with the last, we get

$$\int_{\mathbf{Q}^m} \text{error}(\vec{x})\, dD(\vec{x}) \leq \sup_{x_1,\ldots,x_m} \frac{1}{m} \sum_{j=1}^{m} \text{error}(\text{switch}(\vec{x}, j)).$$

Choose $x_1, ..., x_m \in \mathbf{Q}$. Let $u_1, ..., u_m$ be $x_1, ..., x_m$ in sorted order.

$$\begin{aligned}
&\sum_{j=1}^{m} \text{error}(\text{switch}(\vec{x}, j)) \\
&= |f(u_1) - f(u_2)| + |f(u_{m-1}) - f(u_m)| \\
&\qquad + \sum_{i=2}^{m-1} \left| \left(\frac{i-1}{m-1} f(u_{i-1}) + \left(1 - \frac{i-1}{m-1}\right) f(u_{i+1}) \right) - f(u_i) \right| \\
&= |f(u_1) - f(u_2)| + |f(u_{m-1}) - f(u_m)| \\
&\qquad + \sum_{i=2}^{m-1} \left| \frac{i-1}{m-1} (f(u_{i-1}) - f(u_i)) + \left(1 - \frac{i-1}{m-1}\right) (f(u_{i+1}) - f(u_i)) \right| \\
&\leq |f(u_1) - f(u_2)| + |f(u_{m-1}) - f(u_m)| \\
&\qquad + \sum_{i=2}^{m-1} \left(\frac{i-1}{m-1} |f(u_{i-1}) - f(u_i)| + \left(1 - \frac{i-1}{m-1}\right) |f(u_i) - f(u_{i+1})| \right) \\
&= \left(1 + \frac{1}{m-1}\right) \sum_{i=1}^{m-1} |f(u_i) - f(u_{i+1})|
\end{aligned}$$

which is at most $1 + \frac{1}{m-1}$ since $f \in \mathcal{F}_{\text{BV}}$. $\square$

5 The PAC model

In this section, we show that Theorem 12 implies an improved bound for learning $\mathcal{F}_{\text{BV}}$ in the PAC model [27].

For some countable set X, and some class $\mathcal{F}$ of functions from X to $[0, 1]$, following [27], we say that a learning algorithm A *(ϵ, δ)-PAC learns $\mathcal{F}$ from m examples* for all probability distributions D on X and all $f \in \mathcal{F}$, if A is given $(x_1, f(x_1)), ..., (x_m, f(x_m))$ for $x_1, ..., x_m$ generated according to D^m, then with probability at least $1 - \delta$, A outputs a hypothesis h such the $\int_X |h(x) - f(x)|\, dD(x) \leq \epsilon$.

This model is like the agnostic model studied in Section 3, except with the added assumption that there is a function in $\mathcal{F}$ capable of perfect classification.

Lemma 13 ([9,10]) *For any set $\mathcal{F}$ of functions from $\mathbf{Q}$ to Y, if there is a*

polynomial-time algorithm A such that $M(A, \mathcal{F}, m) = O(1/m)$, then there is a polynomial-time algorithm that (ϵ, δ)-PAC learns $\mathcal{F}$ from $O\left(\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ examples.

Combining this with Theorem 12 implies the following.

Theorem 14 *There is a polynomial-time algorithm that (ϵ, δ)-PAC learns $\mathcal{F}_{\mathrm{BV}}$ from $O\left(\frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ examples.*

6 A lower bound

The following lemma follows from a lower bound of [25] (see [5,2]).

Lemma 15 *There are constants $c_1, c_2, c_3 > 0$ such that, for any $0 < \beta \leq c_1$, if a coin with probability $1/2 + \beta$ of coming up heads is flipped m independent times, the probability that it comes up heads fewer than $m/2$ times is at least $c_2 e^{-c_3\beta^2 m}$.*

The following proof makes heavy use of Simon's [23] ideas, and the result can easily be generalized in many ways. Since we don't know how to use a subset of Simon's proof to establish the result, we've included a proof here. It implies an $\Omega\left(\frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$ lower bound on the sample complexity of agnostically learning the set of all constant functions on $[0, 1]$, and therefore for $\mathcal{F}_{\mathrm{BV}}$.

Proposition 2 *Choose a set $\mathcal{F}$ of functions from $\mathbf{Q}$ to Y such that there is an $x \in \mathbf{Q}$ and $f_0, f_1 \in \mathcal{F}$ for which $f_0(x) = 0$ and $f_1(x) = 1$. If for all $\epsilon, \delta > 0$, $\mathcal{F}$ is (ϵ, δ)-agnostically learnable from $m(\epsilon, \delta)$ examples, then $m(\epsilon, \delta) = \Omega\left(\frac{1}{\epsilon^2}\ln\frac{1}{\delta}\right)$.*

Proof: Choose m and $\epsilon > 0$. Let P_0 and P_1 be the distributions over $X \times [0, 1]$ such that

$$
\begin{aligned}
P_0(\{(x,0)\}) &= 1/2 + 2\epsilon \\
P_0(\{(x,1)\}) &= 1/2 - 2\epsilon \\
P_1(\{(x,0)\}) &= 1/2 - 2\epsilon \\
P_1(\{(x,1)\}) &= 1/2 + 2\epsilon.
\end{aligned}
$$

Suppose b is chosen uniformly at random from $\{0, 1\}$, then m examples are generated according to P_b and passed to an algorithm, which outputs a hypothesis h. The overall probability that $\mathbf{er}_{P_b}(h) - \inf_{f\in\mathcal{F}} \mathbf{er}_{P_b}(f) > \epsilon$ is known to be minimized by any algorithm that, for each input, minimizes the a posteriori probability that this happens given the examples [7].

For any function h from $\mathbf{Q}$ to Y, and either $b \in \{0, 1\}$,

$$\begin{aligned}\mathbf{er}_{P_b}(h) &= (1/2+2\epsilon)|h(x)-b| + (1/2-2\epsilon)(1-|h(x)-b|) \\ &= (1/2-2\epsilon) + 4\epsilon|h(x)-b|.\end{aligned}$$

Thus, $\inf_{f\in\mathcal{F}} \mathbf{er}_{P_b}(f) = 1/2 - 2\epsilon$, and, to ensure $\mathbf{er}_{P_b}(h) - \inf_{f\in\mathcal{F}} \mathbf{er}_{P_b}(f) \leq \epsilon$, one needs $|h(x) - b| \leq 1/4$.

Since the a posteriori probability that $b = 1$ given a sample $(x, y_1), ..., (x, y_m)$ is at least $1/2$ if and only if more than half of the y_i's are 1, an optimal algorithm outputs some h with $h(x) \geq 3/4$ if this is the case, and otherwise outputs some h with $h(x) \leq 1/4$. The probability that such an algorithm has $\mathbf{er}_{P_b}(h) - \inf_{f\in\mathcal{F}} \mathbf{er}_{P_b}(f) > \epsilon$ is then the probability that a coin with bias $1/2 + 2\epsilon$ toward heads comes up heads fewer than $m/2$ times in m flips. Applying Lemma 15, requiring that this probability is at most δ and solving for m completes the proof. □

Acknowledgements

We thank Peter Bartlett, Sanjay Jain, and the members of the COLT'98 program committee for their comments on drafts of this paper. We gratefully acknowledge the support of National University of Singapore Academic Research Fund Grant RP960625.

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the Association for Computing Machinery*, 44(4):616–631, 1997.

[2] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[3] P. L. Bartlett, S. R. Kulkarni, and S. E. Posner. Covering numbers for real-valued function classes. *IEEE Transactions on Information Theory*, 43(5):1721–1724, 1997.

[4] P. L. Bartlett and P. M. Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *Journal of Computer and System Sciences*, 56(2):174–190, 1998.

[5] R. D. Barve and P. M. Long. On the complexity of learning from drifting distributions. *Information and Computation*, 138(2):101–123, 1997.

[6] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *JACM*, 36(4):929–965, 1989.

[7] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

[8] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[9] D. Haussler, M. Kearns, N. Littlestone, and M. K. Warmuth. Equivalence of models for polynomial learnability. *Information and Computation*, 95:129–161, 1991.

[10] D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):129–161, 1994.

[11] D. Haussler and P. M. Long. A generalization of Sauer's Lemma. *Journal of Combinatorial Theory, Series A*, 71(2):219–240, 1995.

[12] M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

[13] M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.

[14] A. N. Kolmogorov and V. M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Society Translations (Ser. 2)*, 17:277–364, 1961.

[15] W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1977–1980, 1998.

[16] P. M. Long. The complexity of learning according to two models of a drifting environment. *Proceedings of the 1998 Conference on Computational Learning Theory*, 1998.

[17] P. M. Long. The complexity of learning according to two models of a drifting environment. *Machine Learning*, 37(3):337–354, 1999.

[18] B. K. Natarajan. Occam's razor for functions. In *Proceedings of the 1993 ACM Conference on Computational Learning Theory*, pages 370–376. ACM Press, 1993.

[19] D. Pollard. *Convergence of Stochastic Processes*. Springer Verlag, 1984.

[20] D. Pollard. *Empirical Processes : Theory and Applications*, volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Math. Stat. and Am. Stat. Assoc., 1990.

[21] S. E. Posner and S. R. Kulkarni. On-line learning of functions of bounded variation under various sampling schemes. In *Proceedings of the 1993 Conference on Computational Learning Theory*, pages 439–445, 1993.

[22] H. L. Royden. *Real Analysis.* Macmillan, 1963.

[23] H. U. Simon. General lower bounds on the number of examples needed for learning probabilistic concepts. *Journal of Computer and System Sciences*, 52(2):239–254, 1996.

[24] H. U. Simon. Bounds on the number of examples needed for learning functions. *SIAM Journal on Computing*, 26(3):751–763, 1997.

[25] E. Slud. Distribution inequalities for the binomial law. *Annals of Probability*, 5:404–412, 1977.

[26] P. Vaidya. A new algorithm for minimizing convex functions over convex sets. *Proceedings of the 30th Annual Symposium on the Foundations of Computer Science*, pages 338–343, 1989.

[27] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

A Proof of Lemma 11

Fix a permutation $\sigma : \{1, ..., m\} \to \{1, ..., m\}$. We have

$$\int_{\mathbf{Q}^m} \phi(\vec{x})\, dD^m(\vec{x}) = \sum_{\vec{x} \in \mathbf{Q}^m} \phi(\vec{x}) \prod_{i=1}^{m} D(x_i) = \sum_{\vec{x} \in \mathbf{Q}^m} \phi(\vec{x}) \left(\prod_{i=1}^{m} D(x_{\sigma^{-1}(i)}) \right). \quad \text{(A.1)}$$

Note that $\psi : Q^m \to Q^m$ defined by $\psi(x_1, ..., x_m) = (x_{\sigma(1)}, ..., x_{\sigma(m)})$ maps onto Q^m, and $\psi(x_{\sigma^{-1}(1)}, ..., x_{\sigma^{-1}(m)}) = (x_1, ..., x_m)$. Thus (A.1) implies

$$\int_{\mathbf{Q}^m} \phi(\vec{x})\, dD^m(\vec{x}) = \sum_{\vec{x} \in \mathbf{Q}^m} \phi(x_{\sigma(1)}, ..., x_{\sigma(m)}) \left(\prod_{i=1}^{m} D(x_i) \right) \quad \text{(A.2)}$$

because the each term of the RHS of (A.1) can be paired with an equal term in the RHS of (A.2). By definition, (A.2) implies

$$\int_{\mathbf{Q}^m} \phi(\vec{x})\, dD^m(\vec{x}) = \int_{\mathbf{Q}^m} \phi(x_{\sigma(1)}, ..., x_{\sigma(m)}) D^m(\vec{x}).$$

So, for any set Γ of permutations on $\{1, ..., m\}$, and any probability distribution D' over Γ,

$$\begin{aligned}\int_{Q^m} \phi(\vec{x})\ dD^m(\vec{x}) &= \int_{\Gamma} \int_{Q^m} \phi(x_{\sigma(1)}, ..., x_{\sigma(m)})\ dD^m(\vec{x})\ dD'(\sigma) \\ &= \int_{Q^m} \int_{\Gamma} \phi(x_{\sigma(1)}, ..., x_{\sigma(m)})\ dD'(\sigma)\ dD^m(\vec{x}) \\ &\leq \sup_{x_1,...,x_m} \int_{\Gamma} \phi(x_{\sigma(1)}, ..., x_{\sigma(m)})\ dD'(\sigma),\end{aligned}$$

completing the proof. $\square$