A Uniform Lower Error Bound for Half-space Learning

Andreas Maurer

[1] Adalbertstrasse 55
D-80799 München, Germany
andreasmaurer@compuserve.com
[2] Dept. of Computer Science
University College London
Malet Pl., WC1E, London, UK
m.pontil@cs.ucl.ac.uk

Abstract. We give a lower bound for the error of any unitarily invariant algorithm learning half-spaces against the uniform or related distributions on the unit sphere. The bound is uniform in the choice of the target half-space and has an exponentially decaying deviation probability in the sample. The technique of proof is related to a proof of the Johnson Lindenstrauss Lemma. We argue that, unlike previous lower bounds, our result is well suited to evaluate the benefits of multi-task or transfer learning, or other cases where an expense in the acquisition of domain knowledge has to be justified.

1 Introduction

We will prove the following lower bound for half-space learning from the uniform distribution σ on the unit sphere $\mathcal{S}$ in $\mathbb{R}^N$.

Theorem 1. *Let $m < N$ and suppose that $f : \mathcal{S}^m \times \{-1,1\}^m \to \mathcal{S}$ is any learning algorithm such that*

$$f(V\mathbf{x}, \mathbf{y}) = Vf(\mathbf{x}, \mathbf{y}), \ \forall \textit{unitary } V \textit{ on } \mathbb{R}^N. \tag{1}$$

Then for every $u \in \mathcal{S}$

$$\Pr_{\mathbf{x}\sim\sigma^m}\left\{\mathrm{err}_{\sigma,u}\left(f\left(\mathbf{x}, u\left(\mathbf{x}\right)\right)\right) < \frac{1}{\pi}\sqrt{\frac{N-m}{N}} - t\right\} \leq e^{-N(t\pi)^2}.$$

Here $u \in \mathcal{S}$ defines the target function $u(x) = \mathrm{sign}\langle u, x\rangle$, $x \in \mathcal{S}$ and $f(\mathbf{x}, u(\mathbf{x}))$ is the hypothesis returned from the algorithm trained on the sample $\mathbf{x} = (x_1, ..., x_m)$ labeled by u, where $u(\mathbf{x}) = (u(x_1), ..., u(x_m))$. The classification error $\mathrm{err}_{\sigma,u}(f(\mathbf{x}, u(\mathbf{x})))$ is the σ-measure of the set of $x \in \mathcal{S}$ where the sign of $\langle u, x\rangle$ and that of $\langle f(\mathbf{x}, u(\mathbf{x})), x\rangle$ disagree.

The symmetry condition (1) plays an important role in the interpretation of our result. For the proof we only use the fact that symmetry of f implies that

$f(\mathbf{x}, \mathbf{y})$ lies in the subspace spanned by the sample $\mathbf{x}$. Clearly equation (1) is satisfied by all kernel-based algorithms which only depend on the Gramian of $\mathbf{x}$.

Our bound appears weaker than existing lower bounds in [4] and [6] in the sense that it only applies to symmetric algorithms. In another sense it is stronger, because it applies uniformly to *all* target functions, not just to a target function mischievously designed to make the algorithm f fail. It is precisely because of these differences that our bound is suitable for the evaluation of transfer learning or other methods to obtain domain knowledge, where the previous bounds in [4] and [6] are not applicable. A lower bound which holds for *all* algorithms cannot be used to justify the choice of any algorithm. A lower bound which holds for all *symmetric* algorithms can justify the use of an asymmetric algorithm if symmetry-breaking side information is available at a tolerable cost. This will be explained in detail in Section 5.

Theorem 1 is weaker and stronger than the results in [4] and [6] in two other respects. It is weaker, because it is restricted to small sample sizes $m < N$. When we expect small sample sizes and high-dimensional phenomena, this is not a problem. On the other hand Theorem 1 exhibits an exponential concentration property. If the ambient dimension N is large, the quantity $\sqrt{(N-m)/N}/\pi$ becomes an effective performance barrier, as smaller errors have neglegible probability.

A simple technique adapts our result to the case when the input marginal μ is not equal to, but absolutely continuous with respect to σ. If the corresponding density function η satisfies $0 < a \leq \eta(x) \leq b$ for almost all $x \sim \sigma$, then the above bound reads

$$\Pr_{\mathbf{x} \sim \mu^m} \left\{ \mathrm{err}_{\mu,u}\left(f\left(\mathbf{x}, u\left(\mathbf{x}\right)\right)\right) < \frac{a}{\pi}\sqrt{\frac{N-m}{N}} - t \right\} \leq \exp\left(-N\left(\frac{t\pi}{a}\right)^2 + m \ln b \right),$$

which reduces to the bound in Theorem 1 for $\eta = 1$, i.e. $\mu = \sigma$. We will prove this more general version below (Theorem 2).

We introduce some notation in the next section and give a proof of Theorem 1 in Section 3. Sections 4 and Section 5 briefly discuss previous work and the application to the evaluation of domain knowledge.

2 Notation

We will work in the space $\mathbb{R}^N$ with euclidean inner product $\langle \cdot, \cdot \rangle$, norm $\|x\| = \sqrt{\langle x, x \rangle}$ and euclidean metric $d(x, y) = \|x - y\|$. For $x \in \mathbb{R}^N$ and $F \subseteq \mathbb{R}^N$ we write $d(x, F) = \inf\{d(x, y) : y \in F\}$ and we denote with $F^\perp$ the subspace $F^\perp = \{x : \langle x, y \rangle = 0, \forall y \in F\}$. We write $G_{N,m}$ for the Grassman manifold of m-dimensional subspaces of $\mathbb{R}^N$. If M is a subspace of $\mathbb{R}^N$ then P_M is the orthogonal projection operator onto M.

With $\mathcal{S}$ we denote the unit sphere in $\mathbb{R}^N$, that is $\mathcal{S} = \left\{x \in \mathbb{R}^N : \|x\| = 1\right\}$. If $u, v \in \mathcal{S}$ we denote with $\rho(u, v)$ the (shortest) angle between u and v, so that $\rho(\cdot, \cdot)$ is the geodesic metric on $\mathcal{S}$. There is a unique unitarily invariant

probability measure σ (called the Haar measure) on $\mathcal{S}$. If γ is a standard $N(0,1)$-distributed random variable, then σ is defined by

$$\mathbb{E}_{x\sim\sigma}\left[\psi\left(x\right)\right]=\mathbb{E}_{x\sim\gamma^{N}}\psi\left(\frac{x}{\left\|x\right\|}\right)$$

for every Borel function ψ on $\mathcal{S}$.

An m-tuple $\mathbf{x}=(x_1,\ldots,x_m)\in\mathcal{S}^m$ is called a sample. A labeled sample is a member $(\mathbf{x},\mathbf{y})\in\mathcal{S}^m\times\{-1,1\}^m$. If $\mathbf{x}=(x_1,\ldots,x_m)$ is a sample then we use $[\mathbf{x}]$ to denote the linear span $\{x_1,\ldots,x_m\}$, and if V is a unitary transformation on $\mathbb{R}^N$, we denote with $V\mathbf{x}$ the sample $V\mathbf{x}=(Vx_1,\ldots,Vx_m)$.

For labeling we use the sign function

$$\operatorname{sgn}\left(t\right)=\begin{cases}1\text{ if }x>0,\\-1\text{ otherwise}\end{cases}$$

which differs from the usual definition, but this will simplify notation and otherwise be immaterial in the following. If $u,x\in\mathcal{S}$ we let $u\left(x\right)=\operatorname{sgn}(\langle u,x\rangle)$. The target function $u\left(\cdot\right)$ defines the open half-space

$$\left\{x:u\left(x\right)=1\right\}=\left\{x:\langle u,x\rangle>0\right\}.$$

If $\mathbf{x}\in\mathcal{S}^m$ is a sample then we denote

$$u\left(\mathbf{x}\right)=\left(u\left(x_1\right),\ldots,u\left(x_m\right)\right)\in\left\{-1,1\right\}^m.$$

Every $u\in\mathcal{S}$ induces a labeled sample $(\mathbf{x},u\left(\mathbf{x}\right))$.

A learning algorithm is a function $f:\mathcal{S}^m\times\{-1,1\}^m\to\mathcal{S}$, which assigns to every labeled sample $(\mathbf{x},\mathbf{y})$ the hypothesis $f\left(\mathbf{x},\mathbf{y}\right)\in\mathcal{S}$. A learning algorithm is called symmetric if

$$f\left(V\mathbf{x},\mathbf{y}\right)=Vf\left(\mathbf{x},\mathbf{y}\right)$$

for all unitary V. A symmetric algorithm has no preferred coordinate system. Note that all kernel-based algorithms are symmetric.

For $u,v\in\mathcal{S}$ we denote

$$\Delta\left(u,v\right)=\left\{x:u\left(x\right)\neq v\left(x\right)\right\}\subseteq\mathcal{S}.$$

If μ is a probability measure on $\mathcal{S}$ and $u,v\in\mathcal{S}$ then

$$\operatorname{err}_{\mu,u}\left(v\right)=\mu\left(\Delta\left(u,v\right)\right)$$

is the error probability for the hypothesis v when the true half-space is u and the underlying input probability is μ.

3 Proofs

In this section we prove the results announced in the introduction.

The idea is the following: The expected error of any hypothesis v is equal to the geodesic distance from v to the target u, divided by π. The hypothesis generated by a symmetric algorithm lies in $[\mathbf{x}]$, the span of the data, so the error of this hypothesis is lower bounded by the euclidean distance from u to $[\mathbf{x}]$, divided by π. This distance is sharply concentrated at $\sqrt{(N-m)/N}$, as follows from a result of Dasgupta and Gupta [3], given in their proof of the Johnson-Lindenstrauss Lemma.

Proposition 1. *Let μ be a probability measure on $\mathcal{S}$, such that $d\mu(x) = \eta(x)\, d\sigma(x)$ with $0 < a \leq \eta(x) \leq b$ for almost all $x \sim \sigma$. Then for every symmetric learning algorithm f, every $u \in \mathcal{S}$ and every $t > 0$ we have*

$$\Pr_{\mathbf{x}\sim\mu^m}\left\{\mathrm{err}_{\mu,u}\left(f\left(\mathbf{x}, u\left(\mathbf{x}\right)\right)\right) < t\right\} \leq b^m \sup_{M\in G_{N,m}} \Pr_{w\sim\sigma}\left\{d\left(w, M\right) < \frac{t\pi}{a}\right\}.$$

Proof. For any $v \in \mathcal{S}$ we have

$$\mathrm{err}_{\mu,u}\left(v\right) = \int_{\Delta(u,v)} \eta d\sigma \geq a\ \sigma\left(\Delta\left(u,v\right)\right).$$

Since σ is invariant under rotations in the u-v-plane, it is easily seen that $\sigma\left(\Delta\left(u,v\right)\right)$ is just the angle between u and v in radians, divided by π, that is $\sigma\left(\Delta\left(u,v\right)\right) = \rho\left(u,v\right)/\pi$. Since $\rho\left(u,v\right) \geq d\left(u,v\right)$, it follows that

$$\mathrm{err}_{\mu,u}\left(v\right) \geq d\left(u,v\right)\frac{a}{\pi}. \tag{2}$$

Let $(\mathbf{x},\mathbf{y})$ be an arbitrary labeled sample and let V be the unitary map $V = I$ on $[\mathbf{x}]$ and $V = -I$ on $[\mathbf{x}]^{\perp}$. By symmetry of f we must have $f\left(\mathbf{x},\mathbf{y}\right) = f\left(V\mathbf{x},\mathbf{y}\right) = V\ f\left(\mathbf{x},\mathbf{y}\right)$, which clearly implies that $f\left(\mathbf{x},\mathbf{y}\right) \in [\mathbf{x}]$. Combining this observation with (2), we have that

$$\mathrm{err}_{\mu,u}\left(f\left(\mathbf{x}, u\left(\mathbf{x}\right)\right)\right) \geq d\left(u, [\mathbf{x}]\right)\frac{a}{\pi}.$$

We then have

$$\begin{aligned}\Pr_{\mathbf{x}\sim\mu^m}\left\{\mathrm{err}_{\mu,u}\left(f\left(\mathbf{x}, u\left(\mathbf{x}\right)\right)\right) < t\right\} &\leq \Pr_{\mathbf{x}\sim\mu^m}\left\{d\left(u, [\mathbf{x}]\right) < \frac{t\pi}{a}\right\} \\ &\leq b^m \Pr_{\mathbf{x}\sim\sigma^m}\left\{d\left(u, [\mathbf{x}]\right) < \frac{t\pi}{a}\right\},\end{aligned} \tag{3}$$

where we used the upper bound on the density function η in the second inequality. We now use the unitary symmetry of the Haar measure σ. For any $w \in \mathcal{S}$ we denote with $V_{w\to u}$ the rotation which takes w to u and $V_{u\to w}$ its inverse. Then

$$\begin{aligned}\Pr_{\mathbf{x}\sim\sigma^m}\left\{d\left(u, [\mathbf{x}]\right) < \frac{t\pi}{a}\right\} &= \mathbb{E}_{w\sim\sigma} \Pr_{\mathbf{x}\sim\sigma^m}\left\{d\left(u, V_{w\to u}\left[\mathbf{x}\right]\right) < \frac{t\pi}{a}\right\} \\ &= \mathbb{E}_{w\sim\sigma} \Pr_{\mathbf{x}\sim\sigma^m}\left\{d\left(V_{u\to w}u, [\mathbf{x}]\right) < \frac{t\pi}{a}\right\} \\ &= \mathbb{E}_{w\sim\sigma}\mathbb{E}_{\mathbf{x}\sim\sigma^m}\mathbf{1}\left\{d\left(w, [\mathbf{x}]\right) < \frac{t\pi}{a}\right\}.\end{aligned}$$

Exchanging the two expectations and bounding the expectation in $\mathbf{x}$ by a supremum gives

$$\Pr_{\mathbf{x}\sim\sigma^m}\left\{d\left(u,[\mathbf{x}]\right)<\frac{t\pi}{a}\right\}\leq\sup_{M\in G_{N,m}}\Pr_{w\sim\sigma}\left\{d\left(w,M\right)<\frac{t\pi}{a}\right\},$$

which, together with (3), gives the conclusion. □

In their proof of the Johnson-Lindenstrauss Theorem Dasgupta and Gupta [3, Lemma 2.2] give the following lemma.

Lemma 1. *Let $k<N$ and M be a proper k-dimensional subspace of $\mathbb{R}^N$ and $\beta\in(0,1)$. Then*

$$\Pr_{x\sim\sigma}\left\{\left\|P_M x\right\|^2\leq\frac{\beta k}{N}\right\}\leq\exp\left(\frac{k}{2}\left(1-\beta+\ln\beta\right)\right).$$

We bring this result into a weaker but simpler form which is better suited for our purposes.

Lemma 2. *Let $k<N$ and M be a proper k-dimensional subspace of $\mathbb{R}^N$ and $t\in\left(0,\sqrt{(N-k)/N}\right)$. Then*

$$\Pr_{x\sim\sigma}\left\{d\left(x,M\right)\leq\sqrt{\frac{N-k}{N}}-t\right\}\leq e^{-Nt^2}.$$

Proof. For $s\in(0,1)$ let

$$g\left(s\right)=\left(1-s\right)^2-1-2\ln\left(1-s\right).$$

Now, if $h\left(s\right)=g\left(s\right)-2s^2$, then $h\left(0\right)=0$ and $h'\left(s\right)=2s^2/\left(1-s\right)\geq 0$. This shows that $g\left(s\right)\geq 2s^2$. Denote $k'=\dim M^{\perp}=N-k$ and let $s=t\sqrt{N/k'}$. Then $\left(1-s\right)^2\in(0,1)$, so by Lemma 1 applied to $M^{\perp}$ we get

$$\begin{aligned}
&\Pr_{x\sim\sigma}\left\{d\left(x,M\right)\leq\sqrt{\frac{k'}{N}}-t\right\}\\
&=\Pr_{x\sim\sigma}\left\{\left\|P_{M^{\perp}}x\right\|^2\leq\left(1-s\right)^2\frac{k'}{N}\right\}\\
&\leq\exp\left(\frac{k'}{2}\left(1-\left(1-s\right)^2+2\ln\left(1-s\right)\right)\right)\\
&=\exp\left(\frac{-k'g\left(s\right)}{2}\right)\leq e^{-k's^2}=e^{-Nt^2}.
\end{aligned}$$

□

Now we can prove our main result.

Theorem 2. *Let μ be a probability measure on $\mathcal{S}$, such that $d\mu(x) = \eta(x)\, d\sigma(x)$ with $0 < a \leq \eta(x) \leq b$ for almost all $x \sim \sigma$. Then for every symmetric learning algorithm f and for every target $u \in \mathcal{S}$*

$$\Pr_{\mathbf{x}\sim\mu^m}\left\{ err_{\mu,u}\left(f\left(\mathbf{x}, u\left(\mathbf{x}\right)\right)\right) < \frac{a}{\pi}\sqrt{\frac{N-m}{N}} - t \right\} \leq \exp\left(-N\left(\frac{t\pi}{a}\right)^2 + m\ln b \right).$$

Proof. Denoting $t = r\pi/a$ we have

$$\begin{aligned}
&\Pr_{\mathbf{x}\sim\mu^m}\left\{ \mathrm{err}_{\mu,u}\left(f\left(\mathbf{x}, u\left(\mathbf{x}\right)\right)\right) < \frac{a}{\pi}\sqrt{\frac{N-m}{N}} - r \right\} \\
&\leq b^m \sup_{M\in G_{N,m}} \Pr_{w\sim\sigma}\left\{ d\left(w, M\right) < \sqrt{\frac{N-m}{N}} - t \right\} \\
&\leq b^m e^{-Nt^2},
\end{aligned}$$

where we used Proposition 1 in the first and Lemma 2 in the second inequality. The result follows. □

4 Previous lower bounds

In learning theory a lot of attention has been devoted to upper error bounds for learning algorithms, and comparatively little work has been done on lower error bounds. For a deeper understanding of the foundations of the subject, however, lower bounds are interesting and they can serve to establish the tightness of upper bounds.

Ehrenfeucht, Haussler, Kearns and Valiant [4] gave a lower bound on the sample complexity of distribution free PAC learning of a function class $\mathcal{F}$ of VC-dimension d on a domain $\mathcal{X}$. They showed that there is a probability distribution μ on $\mathcal{X}$ such that any learning algorithm requires at least

$$\Omega\left(\frac{d}{\epsilon} + \frac{1}{\epsilon}\ln\frac{1}{\delta}\right)$$

examples in order to learn every function in $\mathcal{F}$, with an error at most ϵ (as measured by μ) and probability at least $1-\delta$ in the examples drawn i.i.d. from μ. While this result is a milestone in statistical learning theory, the method of proof, as in [4] or [1], constructs a special distribution μ concentrated in a particularly mischievous way on a set shattered by $\mathcal{F}$, and it can be expected that the distributions underlying realistic learning problems are less pathological.

This deficiency motivated Phil Long [6] to prove the following result pertaining to the learning of the class $\mathcal{F}$ of half-spaces in $\mathbb{R}^N$ from the uniform distribution σ on the unit sphere $\mathcal{S} \subseteq \mathbb{R}^N$: For any learning algorithm it takes at least

$$\Omega\left(\frac{N}{\epsilon} + \frac{1}{\epsilon}\ln\frac{1}{\delta}\right)$$

examples in order to learn every half space on $\mathbb{R}^N$ with an error at most ϵ (as measured by μ) and probability at least $1-\delta$ in the examples drawn i.i.d. from σ. This result replaces the pathological distribution above by a particularly well behaved one and is of considerable importance, because the notion of halfspace learning is central to many learning techniques (support vector machines, perceptron, etc.).

If the sample size m is smaller than the asserted complexity, then these results can be reformulated as follows: For every algorithm f there *exists* a target vector $u \in \mathcal{S}$ such that the probability, that the error of f with respect to u is less than ϵ, is upper bounded by δ. This is substantially different from our result, which restricts f to be symmetric but holds uniformly for *all* target vectors.

5 Evaluation of domain knowledge

We now return to the case, where the marginal distribution of the data is given by the Haar measure σ and describe circumstances under which our bound is preferable to the results above.

Complete ignorance of the nature of potential target functions can be expressed as a maximal entropy assumption, which in our case corresponds to the uniform prior σ and assigns the same a-priori probability to all halfspaces. Under this assumption it is reasonable to use an algorithm f^* which is optimal in the sense that it minimizes the expected error of the hypotheses it generates, on average over all training samples and target functions drawn from the uniform distribution. This algorithm, which would correspond to the Bayes-point algorithm as in [5], should therefore minimize the functional

$$\Omega\left(f\right) = \mathbb{E}_{u\sim\sigma}\mathbb{E}_{\mathbf{x}\sim\sigma^m}\mathrm{err}_{\sigma,u}\left(f\left(\mathbf{x},u\left(\mathbf{x}\right)\right)\right).$$

For a labeled sample $(\mathbf{x},\mathbf{y}) \in \left(\mathcal{S}^{n-1}\right)^m \times \{-1,1\}^m$ we denote

$$C\left(\mathbf{x},\mathbf{y}\right) = \left\{u \in \mathcal{S}^{n-1} : u\left(\mathbf{x}\right) = \mathbf{y}\right\}.$$

$C\left(\mathbf{x},\mathbf{y}\right)$ is thus the set of all hypotheses consistent with $(\mathbf{x},\mathbf{y})$, sometimes also called the version-space. Observe that, given $\mathbf{x}$ and u, there is exactly one $\mathbf{y}$ such that $\mathbf{y} = u\left(\mathbf{x}\right)$, that is $u \in C\left(\mathbf{x},\mathbf{y}\right)$. We therefore obtain

$$\begin{aligned}
\Omega\left(f\right) &= \pi^{-1}\mathbb{E}_{u\sim\sigma}\mathbb{E}_{\mathbf{x}\sim\sigma^m}\rho\left(f\left(\mathbf{x},u\left(\mathbf{x}\right)\right),u\right) \\
&= \pi^{-1}\mathbb{E}_{\mathbf{x}\sim\sigma^m}\sum_{\mathbf{y}\in\{-1,1\}^m}\mathbb{E}_{u\sim\sigma}\rho\left(f\left(\mathbf{x},u\left(\mathbf{x}\right)\right),u\right)1_{C(\mathbf{x},\mathbf{y})}\left(u\right) \\
&= \pi^{-1}\mathbb{E}_{\mathbf{x}\sim\sigma^m}\sum_{\mathbf{y}\in\{-1,1\}^m}\mathbb{E}_{u\sim\sigma}\rho\left(f\left(\mathbf{x},\mathbf{y}\right),u\right)1_{C(\mathbf{x},\mathbf{y})}\left(u\right),
\end{aligned}$$

and, so, the optimal algorithm is given by

$$f^*\left(\mathbf{x},\mathbf{y}\right) = \arg\min_{w\in\mathcal{S}^{n-1}}\mathbb{E}_{u\sim\sigma}\rho\left(w,u\right)1_{C(\mathbf{x},\mathbf{y})}\left(u\right).$$

The minimizer exists and is unique [7], so that this algorithm is indeed well defined. We also have, for any unitary matrix V, that

$$\begin{aligned}\mathbb{E}_{u\sim\sigma}\rho\left(w,u\right)1_{C(V\mathbf{x},\mathbf{y})}\left(u\right) &= \mathbb{E}_{u\sim\sigma}\rho\left(w,u\right)1_{C(\mathbf{x},\mathbf{y})}\left(V^{-1}u\right)\\ &= \mathbb{E}_{u\sim\sigma}\rho\left(V^{-1}w,u\right)1_{C(\mathbf{x},\mathbf{y})}\left(u\right),\end{aligned}$$

so that $f^{*}\left(V\mathbf{x},\mathbf{y}\right)=Vf^{*}\left(\mathbf{x},\mathbf{y}\right)$. The optimal algorithm f^{*} is therefore symmetric and the lower bound in Theorem 1 applies.

In summary these considerations show that in the absence of domain knowledge one is led to the use of a symmetric algorithm, with the limitations implied by Theorem 1. These limitations then also imply lower bounds on the functional Ω, valid for *every* algorithm f, for example

$$\Omega\left(f\right)\geq\Omega\left(f^{*}\right)\geq\frac{1}{2\pi}\left(1-e^{-\frac{N-m}{4}}\right)\sqrt{\frac{N-m}{N}},$$

as can be obtained by setting $t=\left(1/2\pi\right)\sqrt{\left(N-m\right)/N}$ in Theorem 1. Similar bounds cannot be derived from the results in [4] and [6], because they only hold for single target functions constructed in response to the algorithm f.

We were led to the use of the symmetric algorithm f^{*} by our ignorance of the true distribution of target functions. Suppose now that this distribution, rather than being uniform, is concentrated on a small subset M of the sphere. An example would be the intersection of a low-dimensional subspace with the sphere. As long as we do not know M the uniform prior still represents our knowledge on the potential target functions, and therefore the optimal algorithm is still given by f^{*} as above. On the other hand, if we have knowledge of M, we can adopt an algorithm f' which only searches M. Since M is small there will be a considerable improvement incurred by replacing f^{*} with f'. A quantitative guarantee on this improvement can be calculated by applying an upper error bound (using standard techniques) to f', and by an application of Theorem 1 to f^{*}.

As a simple example, suppose that M is finite. Theorem 1, a standard result and a union bound show that for every target function $u\in M$, with probability at least $1-\delta$ in $\mathbf{x}\sim\sigma^{m}$, the difference of the errors of the two algorithms satisfies

$$\mathrm{err}_{\sigma,u}\left(f^{*}\left(\mathbf{x},u\left(\mathbf{x}\right)\right)\right)-\mathrm{err}_{\sigma,u}\left(f'\left(\mathbf{x},u\left(\mathbf{x}\right)\right)\right)\geq\frac{1}{\pi}\sqrt{\frac{N-m-2\ln\left(\frac{2}{\delta}\right)}{2N}}-\frac{\ln\left|M\right|+\ln\frac{2}{\delta}}{m},$$

which can be rather large if $\ln\left|M\right|\ll m\ll N$. If M is infinite similar high probability bounds for the difference between the errors of the two algorithms could be derived using the VC-dimension of the hypothesis class corresponding to M. Such results cannot be obtained from the bounds in [4] and [6], because they do not hold for every u and may only be valid for a target function outside M. Observe also that the classical lower bounds cannot distinguish between f^{*} and f', while Theorem 1 holds only for f^{*} but not for f', which is not symmetric.

In practice the symmetry breaking knowledge of M will not come for free, but at a sometimes considerable cost. A case in point is multi-task or transfer-learning (as in [2]), where knowledge of M is obtained from a large number of tasks, with corresponding target functions drawn from M, and the cost of the acquired knowledge takes the form of the sampling burden for these tasks and the increased computational complexity of the transfer learning algorithm.

To justify such an expense it is necessary to compute the savings made in moving from f^* to f'. A rigorous computation of the *guaranteed* savings is possible only by comparing an upper error bound for f' to a lower error bound for f^*, as described above.

References

1. M. Anthony, P. Bartlett. *Learning in Neural Networks: Theoretical Foundations*. Cambridge University Press, 1999.
2. J. Baxter, A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12: 149–198, 2000.
3. S. Dasgupta, A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22: 60-65, 2003.
4. A. Ehrenfeucht, D. Haussler, M. Kearns, L. G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3): 247-251, 1989.
5. R. Herbrich, T. Graepel, C. Campbell. Bayes Point Machines. *Journal of Machine Learning Research*, 1: 245–279, 2001.
6. P. M. Long. On the sample complexity of PAC learning halfspaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556-1559, 1995.
7. A. Maurer. An optimization problem on the sphere. Technical Report arXiv:0805.2362.