Journal of Machine Learning Research ? (2020) ?? Submitted 9/19; Published ???

On the Complexity of Proper Distribution-Free Learning of Linear Classifiers

Philip M. Long PLONG@GOOGLE.COM
Google

Raphael J. Long RAFILONG42@GMAIL.COM
University of Illinois at Urbana-Champaign

Editor: Aryeh Kontorovich and Gergely Neu

Abstract

For proper distribution-free learning of linear classifiers in d dimensions from m examples, we prove a lower bound on the optimal expected error of $\frac{d-o(1)}{m}$, improving on the best previous lower bound of $\frac{d/\sqrt{e}-o(1)}{m}$, and nearly matching a $\frac{d+1}{m+1}$ upper bound achieved by the linear support vector machine.

Keywords: Statistical learning theory, lower bounds, linear classifiers.

1. Introduction

This paper is about the following learning problem. A learner seeks to approximate an unknown linear classifier f in $F_d = \{f_{\mathbf{w},b} : \mathbf{w} \in \mathbf{R}^d, b \in \mathbf{R}\}$, where $f_{\mathbf{w},b}(\mathbf{x}) = +$ if $\mathbf{w} \cdot \mathbf{x} \geq b$, and otherwise $f_{\mathbf{w},b}(\mathbf{x}) = -$. For $\mathbf{x}_1, ..., \mathbf{x}_m$, drawn independently at random from a probability distribution D over $\mathbf{R}^d$, the learner receives examples $(\mathbf{x}_1, f(\mathbf{x}_1)), ..., (\mathbf{x}_m, f(\mathbf{x}_m))$, and outputs $h \in F_d$. The accuracy of the learner is measured using another independent draw $\mathbf{x}_{m+1}$ from D; its goal is to minimize the probability, with respect to all $m+1$ random draws and any internal randomization, that $h(\mathbf{x}_{m+1}) \neq f(\mathbf{x}_{m+1})$. Let $\mathrm{opt}_d(m)$ be best possible upper bound on this probability of error that a learner can achieve for every f and D. (A formal definition of $\mathrm{opt}_d(m)$ can be found in Section 2.)

It is known that, for all d,

$$\frac{d/\sqrt{e} - o(1)}{m} \leq \mathrm{opt}_d(m) \leq \frac{d+1}{m+1}. \tag{1}$$

The upper bound is achieved by the linear SVM algorithm (Boser et al., 1992). (Because we could not find a proof, we have included one in Appendix A.) The lower bound, which also holds for learners that may output arbitrary classifiers, is implicit in the analysis of (Helmbold and Long, 2012). When $d \leq 2$, a better upper bound is known: $\mathrm{opt}_d(m) \leq \frac{d+o(1)}{m}$ (Blumer and Littlestone, 1989; Haussler et al., 1994).

In this paper, we show that, for all d,

$$\mathrm{opt}_d(m) \geq \frac{d - o(1)}{m},$$

determining the leading constant for general d to within one, and matching the upper bound in the case $d \leq 2$ up to the leading constant.

We use ideas from (Haussler et al., 1994) to reduce the problem of proving lower bounds on $\text{opt}_d(m)$ to the case where $d = 1$. The core of our analysis is a new lower bound on $\text{opt}_1(m)$.

Since (Ehrenfeucht et al., 1989), a common lower bound technique is to (a) choose f and D randomly, (b) characterize the optimal algorithm for minimizing the probability of error with respect to the random choice of f and D along with the random data, and (c) analyze the probability of error of this "Bayes optimal algorithm". If we view the learning problem as a game between the learner and Nature, then, informally, adopting this strategy gives away the advantage that f and D can depend on the learner A, or, in game-theoretic terms, that Nature can "move last". It may be tempting to believe that no leverage is lost in this way, since the minimax theorem (von Neumann, 1928) may be loosely interpreted as saying that nothing is lost by moving first.

The minimax theorem holds for all finite games, but it has long been known that it can fail for some infinite games (Sion and Wolfe, 1957). Here is one example. Each player chooses a member of the open interval $(0, 1)$, and the winner is the player with the bigger number. The player who moves second can win with an arbitrarily high probability. For example, if Player B knows Player A's (mixed) strategy, and always outputs a value a tiny bit greater than the 99th percentile of A's distribution, then Player B will win at least 99% of the time.

The halfline learning problem at the core of this paper is somewhat like this: it can be helpful for Nature to put probability beyond the point where the learner is likely to put its decision boundary. Our lower bound proof constructs D and f as a function of the learner roughly in this way.

As we mentioned above, following (Haussler et al., 1994), we prove a lower bound for $\text{opt}_d(m)$ by embedding d copies of the problem of learning one-dimensional linear classifiers into the d-dimensional problem. Using another embedding from (Haussler et al., 1994), our new lower bound on $\text{opt}_1(m)$ implies a $\frac{2d-o(1)}{m}$ lower bound for proper learning of axis-aligned hyper-rectangles in $\mathbf{R}^d$, matching a known $\frac{2d}{m+1}$ upper bound (Haussler et al., 1994) up to the leading constant, and improving on the $\frac{2d/\sqrt{e}-o(1)}{m}$ lower bound implicit in (Helmbold and Long, 2012).

Related work. The most closely related previous work was mentioned earlier. For learning a class F of VC-dimension d without the constraint that the classifier comes from F, Li et al. (2001) proved lower bounds of $\frac{d-o(1)}{m}$ for classes that they constructed, matching the general upper bound of $\frac{d}{m+1}$ from (Haussler et al., 1994) up to the leading constant. Srebro et al. (2010) and Shamir (2015) proved lower bounds on the complexity of distribution-free linear regression, establishing the complexity of a formulation of this problem to within a constant factor. Some less closely related lower bound work includes (Opper and Haussler, 1991; Devroye and Lugosi, 1995; Long, 1995; Antos and Lugosi, 1998).

2. Preliminaries and main result

For $\mathbf{w} \in \mathbf{R}^d$ and $b \in \mathbf{R}$, the linear classifier $f_{\mathbf{w},b} : \mathbf{R}^d \to \{-,+\}$ parameterized by $\mathbf{w}$ and b outputs $f_{\mathbf{w},b}(\mathbf{x}) = +$ if and only if $\mathbf{w} \cdot \mathbf{x} \geq b$. Let $F_d = \{f_{\mathbf{w},b} : \mathbf{w} \in \mathbf{R}^d, b \in \mathbf{R}\}$.

For any d, an *example* is a member of $\mathbf{R}^d \times \{-,+\}$, a *training set* is a finite multiset of examples.

Informally, a *learner* is a randomized mapping from training sets to F_d. A more detailed definition, which includes a measurability constraint, is given in Appendix B.

Let D be any probability distribution over $\mathbf{R}^d$ with respect to the Lebesgue σ-algebra. For a learner A and $f \in F_d$, let $\mathrm{er}_{D,f}(A,m)$ be the probability that, for $\mathbf{x}_1, ..., \mathbf{x}_{m+1}$ drawn independently at random from D, if A is given $(\mathbf{x}_1, f(\mathbf{x}_1)), ..., (\mathbf{x}_m, f(\mathbf{x}_m))$, it produces h such that $h(\mathbf{x}_{m+1}) \neq f(\mathbf{x}_{m+1})$.

Define $\mathrm{er}_d(A,m) = \sup_{D,f} \mathrm{er}_{D,f}(A,m)$ and $\mathrm{opt}_d(m) = \inf_A \mathrm{er}_d(A,m)$.

The following is our main result.

Theorem 1 *For all d, $\mathrm{opt}_d(m) \geq \frac{d-o(1)}{m}$.*

As mentioned in the introduction, this nearly matches known upper bounds of $\mathrm{opt}_1(m) \leq \frac{d+o(1)}{m}$ for $d \leq 2$ and $\mathrm{opt}_d(m) \leq \frac{d+1}{m+1}$ for $d > 2$.

3. The $d = 1$ case

In this section, we prove Theorem 1 in the case that $d = 1$.

First of all, define $\widetilde{\mathrm{opt}}_1$ analogously to opt_1, with the additional constraint that the support of D is finite. Then $\widetilde{\mathrm{opt}}_1(m) \leq \mathrm{opt}_1(m)$, so it suffices to prove a lower bound for $\widetilde{\mathrm{opt}}_1(m)$. We will do this. This obviates any measurability issues.

Given a learner A and a number m of examples, we will describe a probability distribution D over $[0,1]$ with finite support and $f \in F_1$ such that $\mathrm{er}_{D,f}(A,m) \geq (1-o(1))/m$. Our construction only uses a subset of F_1 with a single parameter θ: classifiers that evaluate to $+$ on x iff $x \leq \theta$.

We define D as follows. First, $\mathbf{Pr}(x = 1) = 1 - \frac{1}{2\sqrt{m}}$. The remaining probability is distributed evenly among $\ell = m^3$ points in $[0,1)$. Let us call the set of these ℓ points $T = \{t_1, ..., t_\ell\}$ where $t_1 < ... < t_\ell$; each member of T thus has probability $\frac{1}{2\ell\sqrt{m}}$. The iterative construction of T will be described later. For all $t \in T$, $f(t) = +$, and $f(1) = -$. (The behavior of f outside $T \cup \{1\}$ does not matter.)

Let h be the output of A. Since $h \in F_1$, there are $v, a \in \mathbf{R}$ such that $h(x) = +$ iff $vx \geq a$. If $v = 0$, then h is either the all-$+$ classifier or the all-$-$ classifier. If $v > 0$, there is a threshold $\hat{\theta}$ such that $h(x) = +$ exactly when $x \geq \hat{\theta}$, and, otherwise, there is a $\hat{\theta}$ such that $h(x) = +$ exactly when $x \leq \hat{\theta}$.

The following lemma enables us to assume without loss of generality that there is an $\hat{\theta}$ such that $h(x) = +$ if and only if $x \leq \hat{\theta}$.

Lemma 2 *For any learner B, there is a learner A such that*

- *A always outputs h for which there is an $\hat{\theta}$ such that $h(x) = +$ if and only if $x \leq \hat{\theta}$, and*

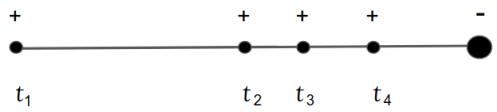

Figure 1: The distributions used in the proof of Theorem 1 concentrate probability on a negative example at 1, and spread probability evenly among positive examples in $[0, 1)$ that are chosen depending on the learner A.

- *for any choice of* T, $\mathrm{er}_{D,f}(A, m) \leq \mathrm{er}_{D,f}(B, m)$.

Proof Whenever B outputs a classifier that assigns all elements of $[0, 1]$ the same class, A can also do this, either by choosing $\hat{\theta} = 2$ or $\hat{\theta} = -1$. If B outputs a classifier h that predicts $+$ on $[\tilde{\theta}, \infty)$ for $\tilde{\theta} \leq 1$, then A can improve it using the all-$+$ classifier, since all of h's predictions on $T \cap [0, \theta)$ are incorrect. ■

For the rest of the proof, let $\hat{\theta}$ refer to the threshold associated with the output of A that is guaranteed by Lemma 2.

Let E_0 be the event that all examples (x_j, y_j) have $x_j = 1$, and let P_0 be the probability distribution on $\hat{\theta}$ obtained by conditioning D^m on E_0. Since E_0 is the event that none of the examples are members of T, if we change T, this does not effect P_0 – conditioning on E_0 removes any effect of the choice of T on $\hat{\theta}$.

The choice of t_1 depends on A as follows. The first case is where $\mathbf{Pr}(\hat{\theta} < 1 | E_0) \geq 1 - \frac{1}{\sqrt{m}}$. Since

$$\begin{aligned}
\lim_{n\to\infty} \mathbf{Pr}(\hat{\theta} < 1 - 1/n \mid E_0) &= \lim_{n\to\infty} \mathbf{Pr}(\hat{\theta} < 1 \mid E_0) - \mathbf{Pr}(\hat{\theta} \in [1 - 1/n, 1) \mid E_0) \\
&= \mathbf{Pr}(\hat{\theta} < 1 \mid E_0) - \lim_{n\to\infty} \mathbf{Pr}(\hat{\theta} \in [1 - 1/n, 1) \mid E_0) \\
&= \mathbf{Pr}(\hat{\theta} < 1 \mid E_0)
\end{aligned}$$

there is an n such that

$$\mathbf{Pr}(\hat{\theta} < 1 - 1/n \mid E_0) \geq 1 - \frac{2}{\sqrt{m}};$$

we choose $t_1 = 1 - 1/n$ for an arbitrary such n.

In the remaining case, where $\mathbf{Pr}(\hat{\theta} < 1 | E_0) < 1 - \frac{1}{\sqrt{m}}$, we set $t_1 = 0$.

For $j \in \{2, ..., \ell\}$, the choice of t_j is similar. The distribution over $\hat{\theta}$ obtained by conditioning the m independent draws from D on the event E_{j-1} that the greatest positive example is t_{j-1} is unaffected by the choices of $t_j, ..., t_\ell$, because conditioning on E_{j-1} removes any effect of $t_j, ..., t_\ell$ on the distribution over $\hat{\theta}$. We choose t_j as follows. First, if $\mathbf{Pr}(\hat{\theta} < 1 \mid E_{j-1}) \geq 1 - \frac{1}{\sqrt{m}}$, then, similarly to the case $j = 1$, we have

$$\lim_{n\to\infty} \mathbf{Pr}(\hat{\theta} < 1 - 1/n \mid E_{j-1}) = \mathbf{Pr}(\hat{\theta} < 1 \mid E_{j-1})$$

so there is an n such that $\mathbf{Pr}(\hat{\theta} < 1 - 1/n \mid E_{j-1}) \geq 1 - \frac{2}{\sqrt{m}}$ we set $t_j = 1 - 1/n$ for an arbitrary such n that also satisfies $1 - 1/n > t_{j-1}$. If $\mathbf{Pr}(\hat{\theta} < 1 \mid E_{j-1}) < 1 - \frac{1}{\sqrt{m}}$, then t_j is an arbitrary member of $(t_{j-1}, 1)$.

Now that we have defined f and D, let us bound $\mathrm{er}_{D,f}(A, m)$. If $\mathbf{Pr}(\hat{\theta} < 1 \mid E_j) \geq 1 - \frac{1}{\sqrt{m}}$, let us say that A is *reasonable at* j. (Note that this is a property of A and T, and not the random training and/or test data.) We have

$$\mathbf{Pr}(h(x_{m+1}) \neq y_{m+1}) = \sum_{j=0}^{\ell} \mathbf{Pr}(h(x_{m+1}) \neq y_{m+1} \wedge E_j).$$

Let us focus on a particular value of j. As a first case, suppose A is reasonable at j. Then, given E_j, with probability at least $1 - \frac{2}{\sqrt{m}}$, t_{j+1}, and therefore $t_{j'}$ for all $j' \geq j+1$, are all greater than $\hat{\theta}$. Thus, for j for which A is reasonable at j, we have

$$\mathbf{Pr}(h(x_{m+1}) \neq y_{m+1} \mid E_j) \geq (\ell - j)\left(1 - \frac{2}{\sqrt{m}}\right)\frac{1}{2\ell\sqrt{m}}. \tag{2}$$

Now, suppose A is unreasonable at j. Then

$$\begin{aligned}\mathbf{Pr}(h(x_{m+1}) \neq y_{m+1} \mid E_j) &= \mathbf{Pr}(\hat{\theta} \geq 1 \mid E_j)\left(1 - \frac{1}{2\sqrt{m}}\right) \\ &> \left(1 - \frac{1}{2\sqrt{m}}\right)\frac{1}{\sqrt{m}} \\ &> (\ell - j)\left(1 - \frac{2}{\sqrt{m}}\right)\frac{1}{2\ell\sqrt{m}}.\end{aligned}$$

Consider the event U that x_{m+1} is less than 1 but greater than all positive training examples. Note that

$$\mathbf{Pr}(U|E_j) = \frac{\ell - j}{2\ell\sqrt{m}}.$$

Thus, for every j, the probability of a mistake given E_j is at least $1 - 2/\sqrt{m}$ times $\mathbf{Pr}(U|E_j)$. Thus, overall, the probability of a mistake is at least $1 - 2/\sqrt{m}$ times $\mathbf{Pr}(U)$. Thus, it suffices to bound $\mathbf{Pr}(U)$ from below.

We bound $\mathbf{Pr}(U)$ by conditioning on the very likely event $\mathcal{E}$ that there is at least one positive example and that no positive example is seen twice. Conditioned on $\mathcal{E}$, any ordering of the $m+1$ examples is equally likely (because we have conditioned a permutation-invariant distribution on a permutation-invariant event). Given $\mathcal{E}$, U holds if x_{m+1} is the greatest positive example, which happens for a fraction $\frac{1}{m+1}$ of the random permutations of the data.

Finally, we claim that $\mathbf{Pr}(\mathcal{E}) = 1 - o(1)$. The probability that at least one $x_1, ..., x_{m+1}$ is labeled $+$ is at least

$$1 - (1 - 1/(2\sqrt{m}))^{m+1} = 1 - o(1).$$

The probability that any positive example is seen twice is at most ℓ times the probability that any particular $t \in T$ is seen twice. The latter probability is at most m^2 times the

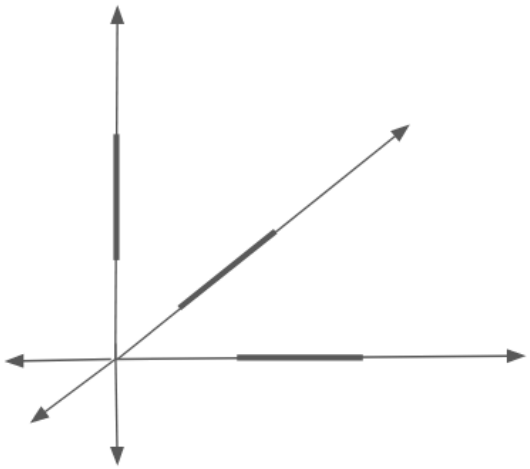

Figure 2: Haussler et al. (1994) proved lower bounds for linear classifiers in d dimensions by embedding d interval-learning problems.

probability that any particular pair of examples share x values of t, which is at most $1/\ell^2$. Therefore the probability that any positive example is seen twice is at most $\ell m^2/\ell^2 = 1/m = o(1)$, so the probability that no examples are seen twice is $1 - o(1)$.

Thus, overall, $\mathbf{Pr}(U) \geq (1 - o(1))/m$, which, as argued earlier, yields $\mathbf{Pr}(h(x_{m+1}) \neq y_{m+1}) \geq (1 - o(1))/m$.

4. The $d > 1$ case

As mentioned before, the extension to the $d > 1$ case uses ideas from (Haussler et al., 1994; Li et al., 2001).

First, we note that $\text{opt}(F_1, m)$ is a non-increasing function of m — if additional examples hurt the learner, it could be improved by ignoring the harmful examples.

Suppose D is some distribution supported on $\{c\mathbf{e}_i : i \in \{1, ..., d\}, c \in [1, 2]\}$, where $\mathbf{e}_i$ is the ith natural basis vector. Any linear classifier $f_{\mathbf{w},b}$ restricted to the support of D can be decomposed into d pieces that are applied to $X_i = \{c\mathbf{e}_i : c \in [1, 2]\}$ for different choices of i (see Figure 2). For each piece, the restriction of $f_{\mathbf{w},b}$ to X_i is isomorphic to a one-dimensional classifier on the interval $[0, 1]$. Thus, any learner A from F_d produces d learners $A_1, ..., A_d$ for the class of restrictions of the members of F_1 to $[0, 1]$. (In particular, the hypothesis of $A_1, ..., A_d$ are linear classifiers.)

Suppose we put negative examples on each of $\{2\mathbf{e}_i : i \in \{1, ..., d\}\}$, and, if T_i is the support set associated with A_i, put positive examples on $\{1 + t\mathbf{e}_i : i \in \{1, ..., d\}, t \in T_i\}$. This data is collectively linearly separable. We may therefore apply our construction from Section 3 independently to each piece, viewing examples from the other pieces as

randomization. If $h_1, ..., h_d$ are the classifiers produced by $A_1, ..., A_d$, we have

$$\mathbf{Pr}(h(x_{m+1}) \neq y_{m+1}) = \frac{1}{d}\sum_{i=1}^{d} \mathbf{Pr}(h_i(x_{m+1}) \neq y_{m+1} \mid y_{m+1} \in X_i).$$

Thus, it suffices to prove a lower bound for $\mathbf{Pr}(h_i(x_{m+1}) \neq y_{m+1} \mid y_{m+1} \in X_i)$. Applying a standard Hoeffding bound, with probability $1 - o(1)$, the number of examples falling in X_i is at most $m/d + \sqrt{m \ln m}$. Applying our lower bound construction from the case $d = 1$, we get

$$\mathbf{Pr}(h(x_{m+1}) \neq y_{m+1} \mid y_{m+1} \in X_i) \geq \frac{1 - o(1)}{m/d + \sqrt{m \ln m}} = \frac{d - o(1)}{m}.$$

Acknowledgments

We thank Peter Bartlett for a valuable conversation, and the reviewers for helpful comments.

Appendix A. Upper bound proof

This is a proof of an upper bound of $\frac{d+1}{m+1}$ for the linear SVM algorithm in $\mathbf{R}^d$. The linear SVM algorithm behaves as follows. When all training examples are the same class, the algorithm outputs that class. Otherwise, it predicts using the linear classifier that separates the positive examples from the negative examples while maximizing the distance from the closest example to its separating hyperplane.

Suppose the training and test examples are

$$(\mathbf{x}_1, y_1), ..., (\mathbf{x}_{m+1}, y_{m+1}),$$

where $y_1, ..., y_{m+1} \in \{-1, 1\}$. Since any permutation of these $m + 1$ examples is equally likely, it suffices to bound from above the probability of a mistake when a uniform random choice of these $m + 1$ examples is the test example.

It is known (see (Cristianini et al., 2000)), that the parameters $(\mathbf{w}^*, b^*)$ of the linear SVM applied to *all* of the data (both training and test) are the solution to the problem of choosing $\mathbf{w}$ and b to minimize $||\mathbf{w}||^2$ subject to

$$\forall t,\ y_t(\mathbf{w} \cdot \mathbf{x}_t - b) \geq 1.$$

It also is known that there are non-negative $\alpha_1, ..., \alpha_{m+1}$ such that

- $\mathbf{w}^* = \sum_{t=1}^{m+1} \alpha_t y_t \mathbf{x}_t$, and
- $\sum_{t=1}^{m+1} \alpha_t y_t = 0$.

Finally, the maximum-margin hyperplane is unique. If $\mathbf{w}' = (w_1^*, ..., w_d^*, 0)$, and, for each $\mathbf{x}_t = (x_{t,1}, ..., x_{t,d})$, we define $\mathbf{x}'_t = (x_{t,1}, ..., x_{t,d}, 1)$, the above two conditions can be consolidated into

$$\mathbf{w}' = \sum_{t=1}^{m+1} \alpha_t y_t \mathbf{x}'_t.$$

By Carathéodory's Theorem for cones, there is $U \subseteq \{1, ..., m+1\}$ of size $d+1$, and non-negative $\beta_1, ..., \beta_{m+1}$ such that $\mathbf{w}' = \sum_{t=1}^{m+1} \beta_t y_t \mathbf{x}'_t$ and $\beta_t = 0$ for all $t \notin U$, so that $\mathbf{w}' = \sum_{t \in U} \beta_t y_t \mathbf{x}'_t$. Unwrapping this, $\mathbf{w}^* = \sum_{t=1}^{m+1} \beta_t y_t \mathbf{x}_t$ and $\sum_{t=1}^{m+1} \beta_t y_t = 0$.

Now, consider the case, for some $s \notin U$, that $(\mathbf{x}_s, y_s)$ is the test example. Using the β_t's for $s \neq t$, in part since $\beta_s = 0$, $\mathbf{w}^*$ and b^* still satisfy the Karush-Kuhn-Tucker conditions for a global optimum for the optimization problem obtained by excluding example number s. Thus, the hyperplane parameterized by $\mathbf{w}^*$ and b^* is output when $(\mathbf{x}_s, y_s)$ is the test example, and $\mathbf{x}_s$ was classified correctly by $\mathbf{w}^*$ and b^*. Since this holds for all $s \notin U$, the linear SVM only makes mistakes on elements of U, and, since $|U| \leq d+1$, this completes the proof.

Appendix B. Detailed definition of a learner

A learner can be built using any probability space (Ω, Σ, P) as a source of randomness. It is a function from $\Omega \times (\mathbf{R}^d \times \{-,+\})^*$ to F_d. Associated with each learner A is a predictor ϕ_A that maps $\Omega \times (\mathbf{R}^d \times \{-,+\})^* \times \mathbf{R}^d$ to $\{+,-\}$ defined by $\phi_A(\omega, S, \mathbf{x}) = (A(\omega, S))(\mathbf{x})$. For any finite number m of examples, the restriction of ϕ_A to the case of training sets of size m must be measurable with respect to the product distribution of (Ω, Σ, P) and any $m+1$-fold product distribution used to generate the training examples and the test point; the distributions used to generate examples use the Lebesgue σ-algebra.

References

András Antos and Gábor Lugosi. Strong minimax lower bounds for learning. *Machine Learning*, 30(1):31–56, 1998.

Anselm Blumer and Nick Littlestone. Learning faster than promised by the vapnik-chervonenkis dimension. *Discrete Applied Mathematics*, 24(1-3):47–53, 1989.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

Luc Devroye and Gábor Lugosi. Lower bounds in pattern recognition and learning. *Pattern recognition*, 28(7):1011–1018, 1995.

Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

David Haussler, Nicholas Littlestone, and Manfred K Warmuth. Predicting {0, 1}-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

David P. Helmbold and Philip M. Long. New bounds for learning intervals with implications for semi-supervised learning. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 30.1–30.15, 2012.

Yi Li, Philip M Long, and Aravind Srinivasan. The one-inclusion graph algorithm is near-optimal for the prediction model of learning. *IEEE Transactions on Information Theory*, 47(3):1257–1261, 2001.

Philip M Long. On the sample complexity of PAC learning half-spaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

Manfred Opper and David Haussler. Calculation of the learning curve of bayes optimal classification algorithm for learning a perceptron with noise. In *Annual Workshop on Computational Learning Theory: Proceedings of the fourth annual workshop on Computational learning theory*, volume 5, pages 75–87, 1991.

Ohad Shamir. The sample complexity of learning linear predictors with the squared loss. *The Journal of Machine Learning Research*, 16(1):3475–3486, 2015.

Maurice Sion and Philip Wolfe. On a game without a value. *Contributions to the theory of games*, 3:299–306, 1957.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, pages 2199–2207, 2010.

J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.