The Complexity of Improperly Learning Large Margin Halfspaces

Shai Shalev-Shwartz
TTI-Chicago
shai@tti-c.org

Ohad Shamir
The Hebrew University
ohadsh@cs.huji.ac.il

Karthik Sridharan
TTI-Chicago
karthik@tti-c.org

The main question A probabilistic classifier is a mapping $h : \mathcal{X} \to [0,1]$, where we interpret $h(\mathbf{x})$ as the probability to predict the label $+1$ and $1 - h(\mathbf{x})$ is the probability to predict the label 0. The error of h on a classification example $(\mathbf{x}, y) \in \mathcal{X} \times \{0,1\}$ is $|h(\mathbf{x}) - y|$, which is the expected 0-1 error. Given a distribution $\mathcal{D}$ on $\mathcal{X} \times \{0,1\}$, the (generalization) error of h is:

$$\mathrm{err}(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[|h(\mathbf{x}) - y|] \ .$$

Let $\mathcal{X}$ be the unit ℓ_2 ball of a Hilbert space, let $\phi : \mathbb{R} \to [0,1]$ be a transfer function, and consider the class of probabilistic classifiers:

$$\mathcal{H} \ = \ \{h(\mathbf{w}) = \phi(\langle \mathbf{w}, \mathbf{x}\rangle) : \|\mathbf{w}\|_2 \le 1\} \ ,$$

where $\langle \mathbf{w}, \mathbf{x}\rangle$ is the inner product between the vectors $\mathbf{x}$ and $\mathbf{w}$. For the 0-1 transfer function, $\phi_{0-1}(a) = \frac{\mathrm{sgn}(a)+1}{2}$, $\mathcal{H}$ becomes the class of halfspaces. We allow any transfer functions that satisfy the following (μ, ϵ) margin condition: $\max\{|\phi(a) - \phi_{0-1}(a)| : |a| > \mu\} \le \epsilon$. For example, the sigmoid function $\phi_{\mathrm{sig}}(a) = \frac{1}{1+e^{-a/\sigma}}$ satisfies the (μ, ϵ) condition if $\sigma \le \mu/(\log(1/\epsilon - 1)$. For an illustration see Figure 1.

An improper agnostic learning algorithm, A, receives as input a training set of m i.i.d. samples from $\mathcal{D}$, and returns a classifier (not necessarily from $\mathcal{H}$). The output classifier is a random variable and we denote it by $A(m)$. We use $\mathrm{err}(A(m))$ to denote the expected generalization error of the predictor returned by A, where expectation is with respect to the random choice of the training set. We denote by $\mathrm{time}(A, m)$ the expected runtime of the algorithm A when running on a training set of m examples.

Open Question 1 *Given $\epsilon, \mu > 0$, let ϕ be a transfer function that satisfies the (μ, ϵ) margin condition and let $\mathcal{H}$ be the corresponding hypothesis class. For which pairs (m, T), there exists an algorithm A such that $\mathrm{time}(A, m) \le T$ and*

$$\mathrm{err}(A(m)) \le \min_{h\in\mathcal{H}} \mathrm{err}(h) + \epsilon \ .$$

How do m and T depend on ϵ and on the margin μ ? In particular, does there exist a pair (m, T) such that T is sub-exponential in the margin parameter μ ?

Motivation and Importance Some of the most important machine learning tools are based on learning large-margin halfspaces. Examples include the Perceptron [10], Support Vector Machines [12], and AdaBoost [7]. The class of halfspace classifiers correspond to the class $\mathcal{H}$ with the 0-1 transfer function. While the expressive power of halfspaces seems to be rather restricted – for example, it is impossible to express XOR functions using linear classifiers – one can use the so-called kernel trick to implicitly map the instances into a higher dimension space and then learn a halfspace in that space. The kernel trick has had tremendous impact on machine learning theory and algorithms over the past decade (e.g. [4, 11]).

It is well known that the VC dimension of (homogeneous) halfspaces in an n-dimensional space equals n. This implies that the number of training examples required to learn the class of halfspaces with the 0-1 loss function scales linearly with the dimension n. Without imposing more assumptions, this bound on the number of examples is tight, namely, there exist distributions for which less examples will cause overfitting. When learning with kernels we in fact learn a halfspace in a possibly infinite dimensional inner product space. Since the VC dimension in this case is inifinite, we cannot learn with the 0-1 transfer function. A common solution is to require that the transfer function will be Lipschitz. This requirement is also closely related to the principle of *large margin* analysis because it is easy to construct Lipschitz transfer functions that satisfy the (μ, ϵ) margin condition with $L = O(1/\mu)$. Using the technique of Rademacher complexities [1], it was shown that from the statistical perspective it is possible to learn using $1/(\mu\epsilon)^2$ examples. Since this bound does not depend on the dimension we can learn even in infinite dimensional spaces, which is the heart of kernel-based learning.

From a computational perspective, practical algorithms such as support vector machines often use a convex surrogate objective function, and then apply convex optimization tools. However, there are no guarantees on how well the surrogate function approximates the 0-1 error function (there do exist some recent results on the *asymptotic* relationship between these error functions in some cases (cf. [2]), but these do not apply to the finite-sample finite-time setting we are studying).

Understanding which pairs of sample-time values allow us to learn an accurate classifier w.r.t. the 0-1 error is therefore an important question.

Known Partial Results There exist strong hardness of approximation results for *proper* learning *without* margin, i.e. with the 0-1 transfer function (see for example [8, 5] and the references therein). There are also hardness results for proper learning with sufficiently small margins [3]. We emphasize that we allow improper learning, which is just as useful for the purpose of learning good classifiers, and thus these hardness results do not apply.

As to positive results, it is well known that if the data is separable with margin μ then it is possible (e.g. using the kernel Perceptron of [6]) to learn a halfspace with excess error of ϵ in time $\text{poly}(1/(\mu\epsilon))$. For agnostically learning halfspaces when $\mathcal{X} = \mathbb{R}^n$, the best current result is the algorithm of [9], with complexity $\text{poly}(n)$ for any constant $\epsilon > 0$. However, this algorithm crucially assumes a restricted set of marginal distribution on $\mathcal{X}$ and also has an explicit dependence on the dimension of $\mathcal{X}$. A computational complexity analysis under margin assumptions for the agnostic case was first carried out in [3]. The technique used in [3] is the observation that in the noise-free case, an optimal halfpsace can be expressed as a linear sum of at most $1/\mu^2$ examples. Therefore, one can perform an exhaustive search over all sub-sequences of $1/\mu^2$ examples, and choose the optimal halfspace. Combining this with standard margin-based sample complexity bounds, we obtain learnability with $m = 1/(\mu\epsilon)^2$ examples and time $\text{poly}(\exp((\frac{1}{\mu})^2 \log(\frac{1}{\mu\epsilon})))$. We recently proposed a different approach for learning with respect to the sigmoid transfer function where both m and the runtime are $\text{poly}(\exp(\frac{1}{\mu}\log(\frac{1}{\mu\epsilon}))$ (see [**?**]).

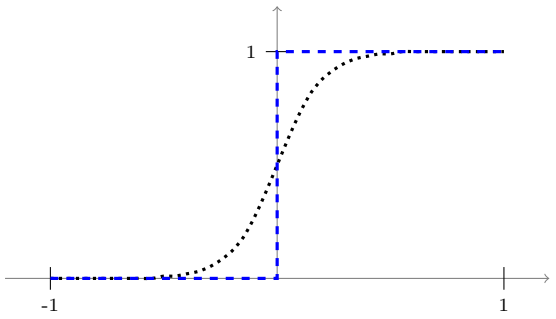

Figure 1: Illustrations of transfer functions: the 0-1 transfer function (dashed blue line) and the sigmoid transfer function (dotted black line, for $\sigma = 0.1$).

References

[1] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3: 463–482, 2002.

[2] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

[3] S. Ben-David and H. Simon. Efficient learning of linear perceptrons. In *Advances in Neural Information Processing Systems 14*, 2000.

[4] N. Cristianini and J. Shawe-Taylor. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[5] V. Feldman, P. Gopalan, S. Khot, and A.K. Ponnuswami. New results for learning noisy parities and halfspaces. In *In Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, 2006.

[6] Y. Freund and R. E. Schapire. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296, 1999.

[7] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

[8] V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. In *Proceedings of the 47th Foundations of Computer Science (FOCS)*, 2006.

[9] A. Kalai, A.R. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th Foundations of Computer Science (FOCS)*, 2005.

[10] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407, 1958. (Reprinted in *Neurocomputing* (MIT Press, 1988).).

[11] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

[12] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.