Journal of Machine Learning Research 16 (2015) 1275-1304 Submitted 12/12; Revised 3/15; Published 7/15

Learning Sparse Low-Threshold Linear Classifiers

Sivan Sabato SABATOS@CS.BGU.AC.IL
Ben-Gurion University of the Negev
Beer Sheva, 8410501, Israel

Shai Shalev-Shwartz SHAIS@CS.HUJI.AC.IL
Benin School of Computer Science and Engineering
The Hebrew University
Givat Ram, Jerusalem 91904, Israel

Nathan Srebro NATI@TTIC.EDU
Toyota Technological Institute at Chicago
6045 S. Kenwood Ave.
Chicago, IL 60637

Daniel Hsu DJHSU@CS.COLUMBIA.EDU
Department of Computer Science
Columbia University
1214 Amsterdam Avenue, #0401
New York, NY 10027

Tong Zhang TZHANG@STAT.RUTGERS.EDU
Department of Statistics
Rutgers University
Piscataway, NJ 08854

Editor: Koby Crammer

Abstract

We consider the problem of learning a non-negative linear classifier with a ℓ_1-norm of at most k, and a fixed threshold, under the hinge-loss. This problem generalizes the problem of learning a k-monotone disjunction. We prove that we can learn efficiently in this setting, at a rate which is linear in both k and the size of the threshold, and that this is the best possible rate. We provide an efficient online learning algorithm that achieves the optimal rate, and show that in the batch case, empirical risk minimization achieves this rate as well. The rates we show are tighter than the uniform convergence rate, which grows with k^2.

Keywords: linear classifiers, monotone disjunctions, online learning, empirical risk minimization, uniform convergence

1. Introduction

We consider the problem of learning non-negative, low-ℓ_1-norm linear classifiers *with a fixed (or bounded) threshold*. That is, we consider hypothesis classes over instances $x \in [0,1]^d$ of the following form:

$$\mathcal{H}_{k,\theta} = \left\{ x \mapsto \langle w, x \rangle - \theta \;\middle|\; w \in \mathbb{R}^d_+, \|w\|_1 \leq k \right\}, \tag{1}$$

where we associate each (real valued) linear predictor in $\mathcal{H}_{k,\theta}$ with a binary classifier:[1]

$$x \mapsto \text{sign}(\langle w, x\rangle - \theta) = \begin{cases} 1 & \text{if } \langle w, x\rangle > \theta \\ -1 & \text{if } \langle w, x\rangle < \theta \end{cases}. \tag{2}$$

Note that the hypothesis class is specified by both the ℓ_1-norm constraint k *and* the fixed threshold θ. In fact, the main challenge here is to understand how the complexity of learning $\mathcal{H}_{k,\theta}$ changes with θ.

The classes $\mathcal{H}_{k,\theta}$ can be seen as a generalization and extension of the class of k-monotone-disjunctions and r-of-k-formulas. Considering binary instances $x \in \{0,1\}^d$, the class of k-monotone-disjunctions corresponds to linear classifiers with binary weights, $w \in \{0,1\}^d$, with $\|w\|_1 \leq k$ and a fixed threshold of $\theta = \frac{1}{2}$. That is, a restriction of $\mathcal{H}_{k,\frac{1}{2}}$ to integer weights and integer instances. More generally, the class of r-of-k formulas (i.e., formulas which are true if at least r of a specified k variables are true) corresponds to a similar restriction, but with a threshold of $\theta = r - \frac{1}{2}$.

Studying k-disjunctions and r-of-k formulas, Littlestone (1988) presented the efficient Winnow online learning rule, which admits an online mistake bound (in the separable case) of $O(k \log d)$ for k-disjunctions and $O(rk \log d)$ for r-of-k-formulas. In fact, in this analysis, Littlestone considered also the more general case of real-valued weights, corresponding to the class $\mathcal{H}_{k,\theta}$ over binary instances $x \in \{0,1\}^d$ and for separable data, and showed that Winnow enjoys a mistake bound of $O(\theta k \log d)$ in this case as well. By applying a standard online-to-batch conversion (see, e.g., Shalev-Shwartz, 2012), one can also achieve a sample complexity upper bound of $O(\theta k \log(d)/\epsilon)$ for batch supervised learning of this class in the separable case.

In this paper, we consider the more general case, where the instances x can also be fractional, i.e., where $x \in [0,1]^d$ and in the agnostic, non-separable, case. It should be noted that Littlestone (1989) also studied a limited version of the non-separable setting.

In order to move on to the fractional and agnostic analysis, we must clarify the loss function we will use, and the related issue of separation with a margin. When the instances x and weight vectors w are integer-valued, we have that $\langle w, x\rangle$ is always integer. Therefore, if positive and negative instances are at all separated by some predictor w (i.e., $\text{sign}(\langle w, x\rangle - \theta) = y$ where $y \in \{\pm 1\}$ denotes the target label), they are necessarily separated by a margin of half. That is, setting $\theta = r - \frac{1}{2}$ for an integer r, we have $y(\langle w, x\rangle - \theta) \geq \frac{1}{2}$. Moving to fractional instances and weight vectors, we need to require such a margin explicitly. And if considering the agnostic case, we must account not only for misclassified points, but also for margin violations. As is standard both in online learning (e.g., the agnostic Perceptron guarantee of Gentile 2003) and in statistical learning using convex optimization (e.g., support vector machines), we will rely on the hinge loss at margin half,[2] which is equal to: $2 \cdot \left[\frac{1}{2} - yh(x)\right]_+$. The hinge loss is a convex upper bound to the zero-one loss (that is, the misclassification rate) and so obtaining learning guarantees for it translates to guarantees on the misclassification error rate.

1. The value of the mapping when $\langle w, x\rangle = \theta$ can be arbitrary, as our results and our analysis do not depend on it.
2. Measuring the hinge loss at a margin of half rather than a margin of one is an arbitrary choice, which corresponds to a scaling by a factor of two, which fits better with the integer case discussed above.

Phrasing the problem as hinge-loss minimization over the hypothesis class $\mathcal{H}_{k,\theta}$, we can use Online Exponentiated Gradient (EG) (Kivinen and Warmuth, 1994) or Online Mirror Descent (MD) (e.g., Shalev-Shwartz, 2007; Srebro et al., 2011), which rely only on the ℓ_1-bound and hold for any threshold. In the statistical setting, we can use Empirical Risk Minimization (ERM), in this case minimizing the empirical hinge loss, and rely on uniform concentration for bounded ℓ_1 predictors (Schapire et al., 1997; Zhang, 2002; Kakade et al., 2009), again regardless of the threshold.

However, these approaches yield mistake bounds or sample complexities that scale quadratically with the ℓ_1 norm, that is with k^2 rather than with θk. Since the relevant range of thresholds is $0 \leq \theta \leq k$, a scaling of θk is always better than k^2. When θ is large, that is, roughly $k/2$, the Winnow bound agrees with the EG and MD bounds. But when we consider classification with a small threshold (for instance, $\theta = \frac{1}{2}$) in the case of disjunctions, the Winnow analysis clarifies that this is a much simpler class, with a resulting smaller mistake bound and sample complexity, scaling with k rather than with k^2. This distinction is lost in the EG and MD analyses, and in the ERM guarantee based on uniform convergence arguments. For small thresholds, where $\theta = O(1)$, the difference between these analyses and the Winnow guarantee is a factor of k.

Our starting point and our main motivation for this paper is to understand this gap between the EG, MD and uniform concentration analyses and the Winnow analysis. Is this gap an artifact of the integer domain or the separability assumption? Or can we obtain guarantees that scale as θk rather then k^2 also in the non-integer non-separable case? In the statistical setting, must we use an online algorithm (such as Winnow) and an online-to-batch conversion in order to ensure a sample complexity that scales with θk, or can we obtain the same sample complexity also with ERM? This is an important question, since the ERM algorithm is considered the canonical batch learning algorithm, and understanding its scope and limitations is of theoretical and practical interest. A related question is whether it is possible to establish uniform convergence guarantees with a dependence on θk rather then k^2, or do the learning guarantees here arise from a more delicate argument.

If an ERM algorithm obtains similar bounds to the ones of the online algorithm with online-to-batch convergence, then any algorithm that can minimize the risk on the sample can be used for learning in this setting. Moreover, this advances our theoretical understanding of the limitations and scope of the canonical ERM algorithm.

The gap between the Winnow analysis and the more general ℓ_1-norm-based analyses is particularly interesting since we know that, in a sense, online mirror descent always provides the best possible rates in the online setting (Srebro et al., 2011). It is thus desirable to understand whether mirror descent is required here to achieve the best rates, or can it be replaced by a simple regularized loss minimization.

Answering the above questions, our main contributions are:

- We provide a variant of online Exponentiated Gradient, for which we establish a regret bound of $O(\sqrt{\theta k \log(d) T})$ for $\mathcal{H}_{k,\theta}$, improving on the $O(\sqrt{k^2 \log(d) T})$ regret guarantee ensured by the standard EG analysis. We do so using a more refined analysis based on local norms. Using a standard online-to-batch conversion, this yields a sample complexity of $O(\theta k \log(d)/\epsilon^2)$ in the statistical setting. This result is given in Corollary 5, Section 3.

- In the statistical agnostic PAC setting, we show that the rate of uniform convergence of the empirical hinge loss of predictors in $\mathcal{H}_{k,\theta}$ is indeed $\Omega(\sqrt{k^2/m})$ where m is the sample size, corresponding to a sample complexity of $\Omega(k^2/\epsilon^2)$, even when θ is small. We show this in Theorem 21 in Section 5. Nevertheless, we establish a learning guarantee for empirical risk minimization which matches the online-to-batch guarantee above (up to logarithmic factors), and ensures a sample complexity of $\tilde{O}(\theta k \log(d)/\epsilon^2)$ also when using ERM. This is obtained by a more delicate local analysis, focusing on predictors which might be chosen as empirical risk minimizers, rather than a uniform analysis over the entire class $\mathcal{H}_{k,\theta}$. The result is given in Theorem 6, Section 4.

- We also establish a matching lower bound (up to logarithmic factors) of $\Omega(\theta k/\epsilon^2)$ on the required sample complexity for learning $\mathcal{H}_{k,\theta}$ in the statistical setting. This shows that our ERM analysis is tight (up to logarithmic factors), and that, furthermore, the regret guarantee we obtain in the online setting is likewise tight up to logarithmic factors. This lower bound is provided in Theorem 17, Section 5.

1.1 Related Prior Work

We discussed Littlestone's work on Winnow at length above. In our notation, Littlestone (1988) established a mistake bound (that is, a regret guarantee in the separable case, where there exists a predictor with zero hinge loss) of $O(k\theta \log(d))$ for $\mathcal{H}_{k,\theta}$, when the instances are integer $x \in \{0,1\}^d$. Littlestone also established a lower bound of $k \log(d/k)$ on the VC-dimension of k-monotone-disjunctions, corresponding to the case $\theta = \frac{1}{2}$, thus implying a $\Omega(k \log(d/k)/\epsilon^2)$ lower bound on learning $\mathcal{H}_{k,\frac{1}{2}}$. However, the question of obtaining a lower bound for other values of the threshold θ was left open by Littlestone.

In the agnostic case, Auer and Warmuth (1998) studied the discrete problem of k-monotone disjunctions, corresponding to $\mathcal{H}_{k,\frac{1}{2}}$ with integer instances $x \in \{0,1\}^d$ and integer weights $w \in \{0,1\}^d$, under the *attribute loss*, defined as the number of variables in the assignment that need to be flipped in order to make the predicted label correct. They provide an online algorithm with an expected mistake bound of $A^* + 2\sqrt{A^* k \ln(d/k)} + O(k \ln(d/k))$, where A^* is the best possible attribute loss for the given online sequence. An online-to-batch conversion thus achieves here a zero-one loss which converges to the optimal attribute loss on this problem at the rate of $O(k \ln(d/k)/\epsilon^2)$. Since the attribute loss is upper bounded by the hinge loss, a similar result, in which A^* is replaced with the optimal hinge-loss for the given sequence, also holds for the same algorithm. This establishes an agnostic guarantee of the desired form, for a threshold of $\theta = \frac{1}{2}$, and when both the instances and weight vectors are integers.

2. Notations and Definitions

For a real number q, we denote its positive part by $[q]_+ := \max\{0, q\}$. We denote universal positive constants by C. The value of C may be different between statements or even between lines of the same expression. We denote by $\mathbb{R}^d_+$ the non-negative orthant in $\mathbb{R}^d$. The all-zero vector in $\mathbb{R}^d$ is denoted by $\mathbf{0}$. For an integer n, we denote $[n] = \{1, \ldots, n\}$. For a vector $x \in \mathbb{R}^d$, and $i \in [d]$, $x[i]$ denotes the i'th coordinate of x.

We will slightly overload notation and use $\mathcal{H}_{k,\theta}$ to denote both the set of linear predictors $x \mapsto \langle w, x\rangle - \theta$ and the set of vectors $w \in \mathbb{R}^d_+$ such that $\|w\|_1 \leq k$. We will use w to denote both the vector and the linear predictor associated with it.

For convenience we will work with *half* the hinge loss at margin half, and denote this loss, for a predictor $w \in \mathcal{H}_{k,\theta}$, for $\theta \in [0, k]$, by

$$\ell_\theta(x, y, w) := \left[\frac{1}{2} - y(\langle w, x\rangle - \theta)\right]_+.$$

The subscript θ will sometimes be omitted when it is clear from context. We term ℓ_θ the *Winnow loss*.

Echoing the half-integer thresholds for k-monotone-disjunctions, r-of-k formulas, and the discrete case more generally, we will denote $r = \theta + \frac{1}{2}$, so that $\theta = r - \frac{1}{2}$. In the discrete case r is integer, but in this paper $\frac{1}{2} \leq r \leq k - \frac{1}{2}$ can also be fractional. We will also sometimes refer to $r' = \frac{1}{2} - \theta$. Note that r' can be negative.

In the statistical setting, we refer to some fixed and unknown distribution D over instance-label pairs (X, Y), where we assume access to a sample (training set) drawn i.i.d. from D, and the objective is to minimize the expected loss:

$$\ell_\theta(w, D) = \mathbb{E}_{X,Y\sim D}[\ell_\theta(X, Y, w)]. \tag{3}$$

When the distribution D is clear from context, we simply write $\ell_\theta(w)$, and we might also omit the subscript θ. For fixed D and θ we let $w^* \in \operatorname{argmin}_{w\in\mathcal{H}_{k,\theta}} \mathbb{E}[\ell(X, Y, w)]$. This is the true minimizer of the loss on the distribution.

For a set of predictors (hypothesis class) H, we denote $\ell^*_\theta(H, D) := \min_{w\in H} \ell_\theta(w, D)$. For a sample $S \in ([0, 1]^d \times \{\pm 1\})^*$, we use the notation

$$\hat{\mathbb{E}}_S[f(X, Y)] = \frac{1}{|S|}\sum_{i=1}^{|S|} f(x_i, y_i) \tag{4}$$

and again sometimes drop the subscript S when it is clear from context. For a fixed sample S, and fixed θ and D, the empirical loss of a predictor w on the sample is denoted $\hat{\ell}(w) = \hat{\mathbb{E}}_S[\ell_\theta(X, Y, w)]$.

2.1 Rademacher Complexity

The empirical Rademacher complexity of the Winnow loss for a class $W \subseteq \mathbb{R}^d$ with respect to a sample $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in ([0, 1]^d \times \{\pm 1\})^m$ is

$$\mathcal{R}(W, S) := \frac{2}{m}\mathbb{E}\left[\sup_{w\in W}\left|\sum_{i=1}^{m} \epsilon_i \ell(x_i, y_i, w)\right|\right] \tag{5}$$

where the expectation is over the *Rademacher random variables* $\epsilon_1, \ldots, \epsilon_m$. These are defined as independent random variables drawn uniformly from $\{\pm 1\}$. The average Rademacher complexity of the Winnow loss for a class $W \subseteq \mathbb{R}^d$ with respect to a distribution D over $[0, 1]^d \times \{\pm 1\}$ is denoted by

$$\mathcal{R}_m(W, D) := \mathbb{E}_{S\sim D^m}[\mathcal{R}(W, S)]. \tag{6}$$

We also define the average Rademacher complexity of W with respect to the *linear loss* by

$$\mathcal{R}_m^L(W, D) := \frac{2}{m}\mathbb{E}\left[\sup_{w\in W}\left|\sum_{i=1}^{m}\epsilon_i Y_i\langle w, X_i\rangle\right|\right] \tag{7}$$

where the expectation is over $\epsilon_1, \ldots, \epsilon_m$ as above and $((X_1, Y_1), \ldots, (X_m, Y_m)) \sim D^m$.

2.2 Probability Tools

We use the following variation on Bernstein's inequality.

Proposition 1 *Let $B > 0$. For a random variable $X \in [0, B]$, $\delta \in (0, 1)$ and n an integer, with probability at least $1 - \delta$ over n i.i.d. draws of X,*

$$\left|\hat{\mathbb{E}}[X] - \mathbb{E}[X]\right| \leq 2B\sqrt{\frac{\ln(1/\delta)}{n} \cdot \max\left\{\frac{\mathbb{E}[X]}{B}, \frac{\ln(1/\delta)}{n}\right\}}.$$

Proof By Bernstein's inequality (Bernstein, 1946), if $Z_1, \ldots, Z_n$ are i.i.d. draws from a random variable $Z \in [-1, 1]$ such that $\mathbb{E}[Z] = 0$, and $\mathrm{Var}[Z^2] = \sigma^2$, then

$$\mathbb{P}[\hat{\mathbb{E}}[Z] \geq \epsilon] \leq \exp\left(-\frac{n\epsilon^2}{2(\sigma^2 + \epsilon/3)}\right). \tag{8}$$

Fix $\delta \in (0, 1)$ and an integer n. If $\ln(1/\delta)/n \leq \sigma^2$ then let $\epsilon = 2\sqrt{\frac{\ln(1/\delta)}{n} \cdot \sigma^2} \leq 2\sigma^2$. In this case

$$\frac{n\epsilon^2}{2\sigma^2 + 2\epsilon/3} \geq \frac{n\epsilon^2}{10\sigma^2/3} \geq \ln(1/\delta).$$

If $\ln(1/\delta)/n > \sigma^2$ then let $\epsilon = 2\ln(1/\delta)/n$. Then $\sigma^2 \leq \ln(1/\delta)/n = \epsilon/2$. In this case

$$\frac{n\epsilon^2}{2\sigma^2 + 2\epsilon/3} \geq \frac{n\epsilon^2}{5\epsilon/3} \geq n\epsilon/4 = \ln(1/\delta).$$

In both cases, the RHS of Eq. (8) is at most δ. Therefore, with probability at least $1 - \delta$,

$$\hat{\mathbb{E}}[Z] \leq 2\sqrt{\frac{\ln(1/\delta)}{n}\max\left\{\sigma^2, \frac{\ln(1/\delta)}{n}\right\}}.$$

where the last inequality follows from the range of Z. Now, for a random variable X with range in $[0, B]$, let $Z = (X - \mathbb{E}[X])/B$. We have $\sigma^2 = \mathrm{Var}[Z] = \mathrm{Var}[X]/B^2 \leq \mathbb{E}[X^2/B^2] \leq \mathbb{E}[X/B]$, where the last inequality follows from the range of X. Therefore

$$\frac{\hat{\mathbb{E}}[X] - \mathbb{E}[X]}{B} \leq 2\sqrt{\frac{\ln(1/\delta)}{n}\max\left\{\frac{\mathbb{E}[X]}{B}, \frac{\ln(1/\delta)}{n}\right\}}.$$

The same bound on $\mathbb{E}[X] - \hat{\mathbb{E}}[X]$ can be derived similarly by considering $Z = (\mathbb{E}[X] - X)/B$. ■

We further use the following fact, which bounds the ratio between the empirical fraction of positive or negative labels and their true probabilities. We will apply this fact to make sure that enough negative and positive labels can be found in a random sample.

Proposition 2 *Let B be a binomial random variable, $B \sim Binomial(m,p)$. If $p \geq 8\ln(1/\delta)/m$ then with probability of at least $1-\delta$, $B \geq mp/2$.*

Proof This follows from a multiplicative Chernoff bound (Angluin and Valiant, 1979). ■

3. Online Algorithm

Consider the following algorithm:

Unnormalized Exponentiated Gradient (unnormalized-EG)

parameters: $\eta, \lambda > 0$
input: $z_1, \ldots, z_T \in \mathbb{R}^d$
initialize: $w_1 = (\lambda, \ldots, \lambda) \in \mathbb{R}^d$
update rule: $\forall i, w_{t+1}[i] = w_t[i] e^{-\eta z_t[i]}$

The following theorem provides a regret bound with local-norms for the unnormalized EG algorithm (for a proof, see Theorem 2.23 of Shalev-Shwartz, 2012).

Theorem 3 *Assume that the unnormalized EG algorithm is run on a sequence of vectors such that for all t, i we have $\eta z_t[i] \geq -1$. Then, for all $u \in \mathbb{R}^d_+$,*

$$\sum_{t=1}^{T} \langle w_t - u, z_t \rangle \leq \frac{d\lambda + \sum_{i=1}^{d} u[i] \ln(u[i]/(e\,\lambda))}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{d} w_t[i] z_t[i]^2 \ .$$

Now, let us apply it to a case in which we have a sequence of convex functions $f_1, \ldots, f_T$, and z_t is the sub-gradient of f_t at w_t. Additionally, set $\lambda = k/d$ and consider u s.t. $\|u\|_1 \leq k$. We obtain the following.

Theorem 4 *Assume that the unnormalized EG algorithm is run with $\lambda = k/d$. Assume that for all t, we have $z_t \in \partial f_t(w_t)$, for some convex function f_t. Further assume that for all t, i we have $\eta z_t[i] \geq -1$, and that for some positive constants α, β, it holds that $\eta = \sqrt{k \ln(d)/(\beta T)}$, $T \geq 4\alpha^2 k \ln(d)/\beta$, and*

$$\sum_{i=1}^{d} w_t[i] z_t[i]^2 \leq \alpha f_t(w_t) + \beta \ . \tag{9}$$

Then, for all $u \in \mathbb{R}^d_+$, with $\|u\|_1 \leq k$ we have

$$\sum_{t=1}^{T} f_t(w_t) \leq \sum_{t=1}^{T} f_t(u) + \sqrt{\frac{4\alpha^2 k \ln(d)}{\beta T}} \cdot \sum_{t=1}^{T} f_t(u) + \sqrt{4\beta k \ln(d) T} + 4\alpha k \ln(d).$$

Proof Using the convexity of f_t and the assumption that $z_t \in \partial f_t(w_t)$ we have that

$$\sum_{t=1}^{T} (f_t(w_t) - f_t(u)) \leq \sum_{t=1}^{T} \langle w_t - u, z_t \rangle \ .$$

Combining with Theorem 3 we obtain

$$\sum_{t=1}^{T}(f_t(w_t) - f_t(u)) \leq \frac{d\lambda + \sum_{i=1}^{d} u[i]\ln(u[i]/(e\,\lambda))}{\eta} + \eta \sum_{t=1}^{T}\sum_{i=1}^{d} w_t[i] z_t[i]^2 \ .$$

Using the assumption in Eq. (9), the definition of $\lambda = k/d$, and the assumptions on u, we obtain

$$\sum_{t=1}^{T}(f_t(w_t) - f_t(u)) \leq \frac{k\ln(d)}{\eta} + \eta\beta T + \eta\alpha \sum_{t=1}^{T} f_t(w_t) \ .$$

Rearranging the above we conclude that

$$\sum_{t=1}^{T} f_t(w_t) \leq \frac{1}{1-\alpha\eta}\left(\sum_{t=1}^{T} f_t(u) + \frac{k\ln(d)}{\eta} + \eta\beta T\right).$$

Now, since $1/(1-x) \leq 1+2x$ for $x \in [0, 1/2]$ and $\alpha\eta \leq \frac{1}{2}$, we conclude, by substituting for the definition of η, that

$$\sum_{t=1}^{T} f_t(w_t) \leq \sum_{t=1}^{T} f_t(u) + 2\sqrt{k\ln(d)\beta T} + 2\alpha\sqrt{\frac{k\ln(d)}{\beta T}} \cdot \sum_{t=1}^{T} f_t(u) + 4\alpha k \ln(d).$$

■

We can now derive the desired regret bound for our algorithm. We also provide a bound for the statistical setting, using online-to-batch conversion.

Corollary 5 *Let $\ell \equiv \ell_\theta$ for some $\theta \in [0,k]$. Fix any sequence $(x_1,y_1),(x_2,y_2),\ldots,(x_T,y_T) \in [0,1]^d \times \{\pm 1\}$ and assume $T \geq 4k\ln(d)/r$. Suppose the unnormalized EG algorithm listed in Section 3 is run using $\eta := \sqrt{\frac{k\ln(d)}{rT}}$, $\lambda := k/d$, and any $z_t \in \partial_w \ell(x_t, y_t, w_t)$ for all t. Define $L_{\mathrm{UEG}} := \sum_{t=1}^{T} \ell(x_t, y_t, w_t)$, let $L(u) := \sum_{t=1}^{T} \ell(x_t, y_t, u)$, and let $u^* \in \operatorname{argmin} L(u)$. Then the following regret bound holds.*

$$L_{\mathrm{UEG}} - L(u^*) \leq \sqrt{16rk\ln(d)T} + 4k\ln(d). \tag{10}$$

Moreover, for $m \geq 1$, assume that a random sample $S = ((x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m))$ is drawn i.i.d. from an unknown distribution D over $[0,1]^d \times \{\pm 1\}$. Then there exists an online-to-batch conversion of the UEG algorithm that takes S as input and outputs $\bar{w}$, such that

$$\mathbb{E}[\ell(\bar{w}, D)] \leq \ell(w^*, D) + \sqrt{\frac{16rk\ln(d)}{m}} + \frac{4k\ln(d)}{m}, \tag{11}$$

where the expectation is over the random draw of S.

Proof Every sub-gradient $z_t \in \partial_w \ell(x_t, y_t, w_t)$ is of the form $z_t = a_t x_t$ for some $a_t \in \{-1, 0, +1\}$. Since $0 \leq x_t[i] \leq 1$ and $w_t[i] \geq 0$ for all i, it follows that $\sum_{i=1}^{d} w_t[i] z_t[i]^2 = |a_t| \sum_{i=1}^{d} w[i] x_t[i]^2 \leq |a_t| \langle w_t, x_t \rangle$. Now consider three disjoint cases.

- Case 1: $\langle w_t, x_t \rangle \leq r$. Then $\sum_{i=1}^{d} w_t[i] z_t[i]^2 \leq \langle w_t, x_t \rangle \leq r$.

- Case 2: $\langle w_t, x_t\rangle > r$ and $y_t = 1$. Then $a_t = 0$ and $\sum_{i=1}^d w_t[i]z_t[i]^2 = 0$.
- Case 3: $\langle w_t, x_t\rangle > r$ and $y_t = -1$. Then $\sum_{i=1}^d w_t[i]z_t[i]^2 \le \langle w_t, x_t\rangle \le [r' + \langle w_t, x_t\rangle]_+ - r' \le [r' + \langle w_t, x_t\rangle]_+ + r$.

In all three cases, the final upper bound on $\sum_{i=1}^d w_t[i]z_t[i]^2$ is at most $\ell(x_t, y_t, w_t) + r$. Therefore, Eq. (9) from Theorem 4 is satisfied with $f_t(w) := \ell(x_t, y_t, w)$, $\alpha := 1$, and $\beta := r$. From Theorem 4 with this choice of f_t and the given settings of η, λ, and z_t, we get that for any u such that $\|u\|_1 \le k$,

$$L_{\mathrm{UEG}} \le L(u) + L(u)\sqrt{\frac{4k\ln(d)}{rT}} + \sqrt{4rk\ln(d)T} + 4k\ln(d). \tag{12}$$

Observing that $L(u^*) \le L(\mathbf{0}) \le rT$, we conclude the regret bound in Eq. (10).

For the statistical setting, a simple approach for online-to-batch conversion is to run the UEG algorithm as detailed in Corollary 5, with $T = m$, and to return the average predictor $\bar{w} = \frac{1}{m}\sum_{i\in[m]} w_i$. By standard analysis (e.g., Shalev-Shwartz, 2012, Theorem 5.1), $\mathbb{E}[\ell_\theta(\bar{w}, D)] \le \frac{1}{m}\mathbb{E}[L_{UEG}]$, where the expectation is over the random draw of S. Setting $u = w_*$, Eq. (12) gives

$$\mathbb{E}[\ell_\theta(\bar{w}, D)] \le \mathbb{E}\left[\hat{\ell}(w^*) + \sqrt{\hat{\ell}(w^*)^2 \cdot \frac{4k\ln(d)}{rm}} + \sqrt{\frac{4rk\ln(d)}{m}} + \frac{4k\ln(d)}{m}\right].$$

Since $\mathbb{E}[\hat{\ell}(w^*)] = \ell(w^*)$ and $\ell(w^*) \le r$, Eq. (11) follows. ■

In the online setting a simple version of the canonical mirror descent algorithm thus achieves the postulated regret bound of $O(\sqrt{rk\log(d)T}) \equiv O(\sqrt{\theta k\log(d)T})$. For the statistical setting, an online-to-batch conversion provides the desired rate of $O(rk\log(d)/\epsilon^2) \equiv O(\theta k\log(d)/\epsilon^2)$. Is this online-to-batch approach necessary, or is a similar rate for the statistical setting achievable also using standard ERM? Moreover, this online-to-batch approach leads to an improper algorithm, that is, the output w might not be in $\mathcal{H}_{k,\theta}$, since it might not satisfy the norm bound. In the next section we show that standard, proper, ERM, leads to the same learning rate.

4. ERM Upper Bound

We now proceed to analyze the performance of empirical risk minimization in the statistical batch setting. As above, assume a random sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ of pairs drawn i.i.d. according to a distribution D over $[0,1]^d \times \{\pm 1\}$. An empirical risk minimizer on the sample is denoted $\hat{w} \in \mathrm{argmin}_{w\in\mathcal{H}_{k,\theta}} \frac{1}{m}\sum_{i\in[m]} \ell(x_i, y_i, w)$. We wish to show an upper bound on $\ell(\hat{w}) - \ell(w^*)$. We will prove the following theorem:

Theorem 6 *For $k \ge r \ge 0$, and $m \ge k$, with probability $1-\delta$ over the random draw of S,*

$$\ell(\hat{w}) \le \ell(w^*) + \sqrt{\frac{O(rk(\ln(d)\ln^3(3m) + \ln(1/\delta)))}{m}} + \frac{O(r\log(1/\delta))}{m}. \tag{13}$$

The proof strategy is based on considering the loss on negative examples and the loss on positive examples separately. Denote

$$\ell_-(w, D) = \mathbb{E}_{(X,Y)\sim D}[\ell(X, Y, w) \mid Y = -1], \text{ and}$$
$$\ell_+(w, D) = \mathbb{E}_{(X,Y)\sim D}[\ell(X, Y, w) \mid Y = +1].$$

For a given sample, denote $\hat{\ell}_-(w) = \hat{\mathbb{E}}[\ell(X, Y, w) \mid Y = -1]$ and similarly for $\hat{\ell}_+(w)$. Denote $p_+ = \mathbb{E}_{(X,Y)\sim D}[Y = +1]$ and $\hat{p}_+ = \hat{\mathbb{E}}[Y = +1]$, and similarly for p_- and $\hat{p}_-$.

As Theorem 21 in Section 5 below shows, the rate of uniform convergence of $\hat{\ell}_-(w)$ to $\ell_-(w)$ for all $w \in \mathcal{H}_{k,\theta}$ is $\tilde{\Omega}(\sqrt{k^2/m})$, which is slower than the desired $\tilde{O}(\sqrt{\theta k/m})$. Therefore, uniform convergence analysis for $\mathcal{H}_{k,\theta}$ cannot provide a tight result. Instead, we define a subset $U_b \subseteq \mathcal{H}_{k,\theta}$, such that with probability at least $1 - \delta$, the empirical risk minimizer of a random sample is in U_b. We show that a uniform convergence rate of $\tilde{O}(\sqrt{\theta k/m})$ does in fact hold for all $w \in U_b$. The analysis of uniform convergence of the negative loss is carried out in Section 4.1.

For positive labels, uniform convergence rates over $\mathcal{H}_{k,\theta}$ in fact suffice to provide the desired guarantee. This analysis is provided in Section 4.2. The analysis uses the results in Section 3 for the online algorithm to construct a small cover of the relevant function class. This then bounds the Rademacher complexity of the class and leads to a uniform convergence guarantee. In Section 4.3, the two convergence results are combined, while taking into account the mixture of positive and negative labels in D.

4.1 Convergence on Negative Labels

We now commence the analysis for negative labels. Denote by D_- the distribution of $(X, Y) \sim D$ conditioned on $Y = -1$, so that $\mathbb{P}_{(X,Y)\sim D_-}[Y = -1] = 1$, and $\mathbb{P}_{(X,Y)\sim D_-}[X = x] = \mathbb{P}_{(X,Y)\sim D}[X = x \mid Y = -1]$. For $b \geq 0$ define

$$U_b(D) = \{w \in \mathbb{R}^d_+ \mid \|w\|_1 \leq k, \mathbb{E}_D[\langle w, X\rangle \mid Y = -1] \leq b\}.$$

Note that $U_b(D) \subseteq \mathcal{H}_{k,\theta}$.

We now bound the rate of convergence of $\hat{\ell}_-$ to ℓ_- for all $w \in U_b(D)$. We will then show that b can be set so that with high probability $\hat{w} \in U_b(D)$. Our technique is related to local Rademacher analysis (Bartlett et al., 2005), in that the latter also proposes to bound the Rademacher complexity of subsets of a function class, and uses these bounds to provide tighter convergence rates. Our analysis is better tailored to the Winnow loss, by taking into account the different effects of the negative and positive labels.

The convergence rate for $U_b(D)$ is bounded by first bounding $\mathcal{R}^L_m(U_b(D), D_-)$, the Rademacher complexity of the linear loss for the distribution over the examples with negative labels, and then concluding a similar bound on $\mathcal{R}_m(U_b(D), D)$. We start with a more general bound on $\mathcal{R}^L_m$.

Lemma 7 *For a fixed distribution over D over $[0, 1]^d \times \{\pm 1\}$, let $\alpha_j = \mathbb{E}_{(X,Y)\sim D}[X[j]]$, and let $\mu \in \mathbb{R}^d_+$. Define $U^\mu = \{w \in \mathbb{R}^d_+ \mid \langle w, \mu\rangle \leq 1\}$. Then if $dm \geq 3$,*

$$\mathcal{R}^L_m(U^\mu, D) \leq \max_{j:\alpha_j>0} \frac{1}{\mu_j} \sqrt{\frac{32 \ln(d)}{m} \cdot \max\left\{\alpha_j, \frac{\ln(dm)}{m}\right\}}$$

Proof Assume w.l.o.g that $\alpha_j > 0$ for all j (if this is not the case, dimensions with $\alpha_j = 0$ can be removed because this implies that $X[j] = 0$ with probability 1).

$$\begin{aligned}\frac{m}{2} R_m^L(U^\mu, S) &= \mathbb{E}_\sigma \left[\sup_{w:\langle w,\mu\rangle \le 1} \sum_{i=1}^m \sigma_i \langle w, x_i\rangle \right] \\ &= \mathbb{E}_\sigma \left[\sup_{w:\langle w,\mu\rangle \le 1} \langle w, \sum_{i=1}^m \sigma_i x_i\rangle \right] \\ &= \mathbb{E}_\sigma \left[\max_{j\in[d]} \sum_{i=1}^m \sigma_i \frac{x_i[j]}{\mu[j]} \right] .\end{aligned}$$

Therefore, using Massart's lemma (Massart, 2000, Lemma 5.2) and denoting $\hat{\alpha}_j = \frac{1}{m}\sum_{i\in[m]}^m x_i[j]$, we have:

$$\begin{aligned}R_m^L(U^\mu, S) &\le \frac{\sqrt{8\ln(d)}}{m} \cdot \max_j \frac{\sqrt{\sum_i x_i[j]^2}}{\mu[j]} \\ &\le \frac{\sqrt{8\ln(d)}}{m} \cdot \max_j \frac{\sqrt{\sum_i x_i[j]}}{\mu[j]} \\ &= \sqrt{\frac{8\ln(d)}{m}} \cdot \max_j \frac{\sqrt{\hat{\alpha}_j}}{\mu[j]} \\ &= \sqrt{\frac{8\ln(d)}{m} \cdot \max_j \frac{\hat{\alpha}_j}{\mu[j]^2}} .\end{aligned}$$

Taking expectation over S and using Jensen's inequality we obtain

$$R_m^L(U^\mu, D) = \mathbb{E}_S[R_m^L(U^\mu, S)] \le \sqrt{\frac{8\ln(d)}{m} \cdot \mathbb{E}_S[\max_j \frac{\hat{\alpha}_j}{\mu[j]^2}]}$$

By Bernstein's inequality (Proposition 1), with probability $1-\delta$ over the choice of $\{x_i\}$, for all $j \in [d]$

$$\hat{\alpha}_j \le \alpha_j + 2\sqrt{\frac{\ln(d/\delta)}{m} \cdot \max\left\{\alpha_j, \frac{\ln(d/\delta)}{m}\right\}}.$$

And, in any case, $\hat{\alpha}_j \le 1$. Therefore,

$$\mathbb{E}_S\left[\max_j \frac{\hat{\alpha}_j}{\mu[j]^2}\right] \le \max_j \frac{1}{\mu[j]^2}\left(\delta + \alpha_j + 2\sqrt{\frac{\ln(d/\delta)}{m} \cdot \max\left\{\alpha_j, \frac{\ln(d/\delta)}{m}\right\}}\right)$$

Choose $\delta = 1/m$ and let j be a maximizer of the above. Consider two cases. If $\alpha_j < \ln(dm)/m$ then

$$\mathbb{E}_S\left[\max_j \frac{\hat{\alpha}_j}{\mu[j]^2}\right] \le \max_j \frac{1}{\mu[j]^2} \cdot \frac{4\ln(dm)}{m}.$$

Otherwise,

$$\mathbb{E}_S\left[\max_j \frac{\hat{\alpha}_j}{\mu[j]^2}\right] \le \max_j \frac{1}{\mu[j]^2}(\delta + 3\alpha_j) \le \max_j \frac{4\alpha_j}{\mu[j]^2}.$$

All in all, we have shown

$$R_m^L(U^\mu, D) \leq \max_j \frac{1}{\mu[j]} \sqrt{\frac{32\ln(d)}{m} \cdot \max\left\{\alpha_j, \tfrac{\ln(dm)}{m}\right\}}.$$

■

The lemma above can now be used to bound the Rademacher complexity of the linear loss for D_-.

Lemma 8 *For any distribution D over $(X,Y) \in [0,1]^d \times \{\pm 1\}$, if $dm \geq 3$,*

$$\mathcal{R}_m^L(U_b(D), D_-) \leq \sqrt{\frac{128k\ln(d)}{m} \max\left\{b, \frac{k\ln(dm)}{m}\right\}}.$$

Proof Let $\alpha_j = \mathbb{E}_{(X,Y)\sim D_-}[X[j]]$. Let $J = \{j \in [d] \mid \alpha_j \geq \frac{b}{k}\}$, and $\bar{J} = \{j \in [d] \mid \alpha_j < \frac{b}{k}\}$. For a vector $v \in \mathbb{R}^d$ and a set $I \subseteq [d]$, denote by $v[I]$ the vector which is obtained from v by setting the coordinates not in I to zero. Let $((X_1,Y_1),\ldots,(X_m,Y_m)) \sim D_-^m$. By the definition of $\mathcal{R}_m^L$, with Rademacher random variables $\epsilon_1,\ldots,\epsilon_m$ (see Eq. 7), we have

$$\begin{aligned}
&\mathcal{R}_m^L(U_b(D), D_-) \\
&= \frac{2}{m}\mathbb{E}\left[\sup_{w\in U_b(D)}\left|\sum_{i=1}^m \epsilon_i Y_i \langle w, X_i\rangle\right|\right] \\
&= \frac{2}{m}\mathbb{E}\left[\sup_{w\in U_b(D)}\left|\sum_{i=1}^m \epsilon_i Y_i \langle w[J], X_i[J]\rangle + \sum_{i=1}^m \epsilon_i Y_i \langle w[\bar{J}], X_i[\bar{J}]\rangle\right|\right] \\
&\leq \frac{2}{m}\mathbb{E}\left[\sup_{w\in U_b(D)}\left|\sum_{i=1}^m \epsilon_i Y_i \langle w[J], X_i[J]\rangle\right|\right] + \frac{2}{m}\mathbb{E}\left[\sup_{w\in U_b(D)}\left|\sum_{i=1}^m \epsilon_i Y_i \langle w[\bar{J}], X_i[\bar{J}]\rangle\right|\right] \\
&= \mathcal{R}_m^L(U_b(D), D_1) + \mathcal{R}_m^L(U_b(D), D_2), \qquad (14)
\end{aligned}$$

where D_1 is the distribution of $(X[J], Y)$, where $(X,Y) \sim D_-$, and D_2 is the distribution of $(X[\bar{J}], Y)$. We now bound the two Rademacher complexities of the right-hand side using Lemma 7.

To bound $\mathcal{R}_m^L(U_b(D), D_1)$, define U^μ as in Lemma 7 for $\mu \in \mathbb{R}_+^d$, and define $\mu_1 \in \mathbb{R}_+^d$ by $\mu_1[j] = \alpha_j/b$. It is easy to see that $U_b(D) \subseteq U^{\mu_1}$. Therefore $\mathcal{R}_m^L(U_b(D), D_1) \leq \mathcal{R}_m^L(U^{\mu_1}, D_1)$. By Lemma 7 and the definition of μ_1

$$\begin{aligned}
\mathcal{R}_m^L(U^{\mu_1}) &\leq \max_{j\in J} \frac{1}{\mu_1[j]} \sqrt{\frac{32\ln(d)}{m} \max\left\{\alpha_j, \frac{\ln(dm)}{m}\right\}} \\
&= \max_{j\in J} \frac{b}{\alpha_j} \sqrt{\frac{32\ln(d)}{m} \max\left\{\alpha_j, \frac{\ln(dm)}{m}\right\}} \\
&= \max_{j\in J} \sqrt{\frac{b}{\alpha_j}\frac{32\ln(d)}{m} \max\left\{b, \frac{b}{\alpha_j}\frac{\ln(dm)}{m}\right\}}.
\end{aligned}$$

By the definition of J, for all $j \in J$ we have $\frac{b}{\alpha_j} \leq k$. It follows that

$$\mathcal{R}_m^L(U^{\mu_1}, D_1) \leq \sqrt{\frac{32k\ln(d)}{m} \max\left\{b, \frac{k\ln(dm)}{m}\right\}}. \tag{15}$$

To bound $\mathcal{R}_m^L(U_b(D), D_2)$, define $\mu_2 \in \mathbb{R}_+^d$ by $\mu_2[j] = \frac{1}{k}$. Note that $U^{\mu_2} = \mathcal{H}_{k,\theta}$ and $U_b(D) \subseteq \mathcal{H}_{k,\theta}$, hence $\mathcal{R}_m^L(U_b(D), D_2) \leq \mathcal{R}_m^L(U^{\mu_2}, D_2)$. By Lemma 7 and the definition of μ_2

$$\begin{aligned}\mathcal{R}_m^L(U^{\mu_2}, D_2) &\leq \max_{j\in\bar{J}} \frac{1}{\mu_2[j]} \sqrt{\frac{32\ln(d)}{m} \max\left\{\alpha_j, \frac{\ln(dm)}{m}\right\}} \\ &= \max_{j\in\bar{J}} \sqrt{\frac{32k\ln(d)}{m} \max\left\{k\alpha_j, \frac{k\ln(dm)}{m}\right\}}.\end{aligned}$$

By the definition of $\bar{J}$, for all $j \in J$ we have $k\alpha_j \leq b$. Therefore

$$\mathcal{R}_m^L(U^{\mu_2}, D_2) \leq \sqrt{\frac{32k\ln(d)}{m} \max\left\{b, \frac{k\ln(dm)}{m}\right\}}. \tag{16}$$

Combining Eq. (14), Eq. (15) and Eq. (16) we get the statement of the theorem. ■

Finally, the bound on $\mathcal{R}_m^L(U_b(D), D)$ is used in the following theorem to obtain a uniform convergence result of the negative loss for predictors in $U_b(D)$.

Theorem 9 *Let $b \geq 0$. There exists a universal constant C such that for any distribution D over $[0,1]^d \times \{\pm 1\}$, with probability $1-\delta$ over samples of size m, for any $w \in U_b(D)$,*

$$\ell_-(w) \leq \hat{\ell}_-(w) + C\left(\sqrt{\frac{kb\ln(d/\delta) + |r'|}{m\hat{p}_-}} + \frac{k\ln(dm\hat{p}_-/\delta)}{m\hat{p}_-}\right). \tag{17}$$

Proof Define $\phi : \mathbb{R} \to \mathbb{R}$ by $\phi(z) = [r' - z]_+$. Since $\mathbb{P}_{(X,Y)\sim D}[Y = -1] = 1$, the Winnow loss on pairs (X, Y) drawn from D is exactly $\phi(Y\langle w, X\rangle)$. Note that ϕ is an application of a 1-Lipschitz function to a translation of the linear loss. Thus, by the properties of the Rademacher complexity (Bartlett and Mendelson, 2002) and by Lemma 8 we have, for $dm \geq 3$,

$$\begin{aligned}\mathcal{R}_m(U_b(D), D_-) &\leq \mathcal{R}_m^L(U_b(D), D_-) \\ &\leq \sqrt{\frac{128k\ln(d)}{m} \max\left\{b, \frac{k\ln(dm)}{m}\right\}}.\end{aligned} \tag{18}$$

Assume that $r' \leq 0$. By Talagrand's inequality (see, e.g., Boucheron et al., 2005, Theorem 5.4), with probability $1-\delta$ over samples of size m drawn from D_-, for all $w \in U_b(D)$

$$\ell(w) \leq \hat{\ell}(w) + 2\mathcal{R}_m(U_b(D), D_-) + \sqrt{\frac{2\sup_{w\in U_b(D)} \mathrm{Var}_{D_-}[\ell(X,Y,w)]\ln(1/\delta)}{m}} + \frac{4k\ln(1/\delta)}{3m}. \tag{19}$$

To bound $\mathrm{Var}_{D_-}[\ell(X,Y,w)]$, note that $\ell(X,Y,w) \in [0,k]$. In addition, $\mathbb{P}_{D_-}[Y=-1]=1$, thus with probability 1, $\ell(X,Y,w) = [r' + \langle w, X\rangle]_+ \leq \langle w, x\rangle$, where the last inequality follows from the assumption $r' \leq 0$. Therefore, for any $w \in U_b(D)$

$$\mathrm{Var}_{D_-}[\ell(X,Y,w)] \leq \mathbb{E}[\ell^2(X,Y,w)] \leq \mathbb{E}_{D_-}[k\ell(X,Y,w)] \leq k \cdot \mathbb{E}_{D_-}[\langle w, X\rangle] \leq kb. \tag{20}$$

Combining Eq. (18), Eq. (19) and Eq. (20) we conclude that there exists a universal constant C such that for any $w \in U_b(D)$, if a sample of size m is drawn i.i.d. from D_-, then

$$\ell(w) \leq \hat{\ell}(w) + C\left(\sqrt{\frac{kb\ln(d/\delta)}{m}} + \frac{k\ln(dm/\delta)}{m}\right).$$

If $r' > 0$, $\hat{\ell}_-(w) - \ell_-(w)$ is identical to the case $r' = 0$, thus the same result holds.

To get Eq. (17), consider a sample of size m drawn from D instead of D_-. In this case, $\ell(w, D_-) = \ell_-(w, D)$, $\hat{\ell}(w, D_-) = \hat{\ell}_-(w, D)$, and the effective sample size for D_- is $m\hat{p}_-$. ■

We now show that with an appropriate setting of b, $\hat{w} \in U_b(D)$ with high probability over the draw of a sample from D. First, the following lemma provides a sample-dependent guarantee for $\hat{w}$.

Lemma 10 *Let $\hat{w}$ and $\hat{p}_-$ be defined as above and let $\hat{E} := \hat{E}_S$ for the fixed sample S defined above. Then*

$$\hat{\mathbb{E}}[\langle \hat{w}, X\rangle \mid Y=-1] \leq \frac{r}{\hat{p}_-}.$$

Proof Let $m_+ = |\{i \mid y_i = +1\}|$, and $m_- = |\{i \mid y_i = -1\}|$. By the definition of the hinge function and the fact that $\langle x_i, \hat{w}\rangle \geq 0$ for all i we have that

$$\begin{aligned} m_- r' + \sum_{y_i=-1} \langle x_i, \hat{w}\rangle &\leq \sum_{y_i=-1} (r' + \langle x_i, \hat{w}\rangle) \\ &\leq \sum_{y_i=+1} [r - \langle x_i, \hat{w}\rangle]_+ + \sum_{y_i=-1} [r' + \langle x_i, \hat{w}\rangle]_+ \\ &= \sum_{i\in[m]} \ell(x_i, y_i, \hat{w}). \end{aligned}$$

By the optimality of $\hat{w}$, $\sum_{i\in[m]} \ell(x_i, y_i, \hat{w}) \leq \sum_{i\in[m]} \ell(x_i, y_i, \mathbf{0}) = m_+ r + m_- [r']_+$. Therefore

$$\sum_{y_i=-1} \langle x_i, \hat{w}\rangle \leq m_+ r + m_-([r']_+ - r') = m_+ r + m_-[-r']_+ \leq (m_+ + m_-)r = mr,$$

where we have used the definitions of r' and r to conclude that $[-r']_+ \leq r$. Dividing both sides by m_- we conclude our proof. ■

The following lemma allows converting the sample-dependent restriction on $\hat{w}$ given in Lemma 10 to one that holds with high probability over samples.

Lemma 11 *For any distribution over $[0,1]^d$, with probability $1-\delta$ over samples of size n, for any $w \in \mathcal{H}_{k,\theta}$*

$$\mathbb{E}[\langle w, X\rangle] \leq 2\hat{\mathbb{E}}[\langle w, X\rangle] + \frac{16k\ln(\frac{d}{\delta})}{n}.$$

Proof For every $j \in [d]$, denote $\alpha_j = \mathbb{E}[X[j]]$. Denote $\hat{\alpha}_j = \hat{\mathbb{E}}[X[j]]$. By Bernstein's inequality (Proposition 1), with probability $1-\delta$,

$$\alpha_j \leq \hat{\alpha}_j + 2\sqrt{\frac{\ln(1/\delta)}{n} \cdot \max\left\{\alpha_j, \frac{\ln(1/\delta)}{n}\right\}} \leq \hat{\alpha}_j + \max\left\{\frac{\alpha_j}{2}, \frac{8\ln(1/\delta)}{n}\right\},$$

where the last inequality can be verified by considering the cases $\alpha_j \leq \frac{16\ln(1/\delta)}{n}$ and $\alpha_j \geq \frac{16\ln(1/\delta)}{n}$. Applying the union bound over $j \in [d]$ we obtain that with probability of $1-\delta$ over samples of size n, for any $w \in \mathcal{H}_{k,\theta}$

$$\begin{aligned}\mathbb{E}[\langle w, X\rangle] = \langle w, \alpha\rangle &\leq \sum_{j\in[d]} w_j \left(\hat{\alpha}_j + \frac{\alpha_j}{2} + \frac{8\ln(d/\delta)}{n}\right) \\ &\leq \hat{\mathbb{E}}[\langle w, X\rangle] + \frac{1}{2}\mathbb{E}[\langle w, X\rangle] + \frac{8\ln(d/\delta)}{n} \cdot k.\end{aligned}$$

Thus $\mathbb{E}[\langle w, X\rangle] \leq 2\hat{\mathbb{E}}\langle w, X\rangle + \frac{16k\ln(d/\delta)}{n}$. ■

Combining the two lemmas above, we conclude that with high probability, $\hat{w} \in U_b$ for an appropriate setting of b.

Lemma 12 *If $p_- \geq \frac{8\ln(1/\delta)}{m}$, then with probability $1-\delta$ over samples of size m, $\hat{w} \in U_b(D)$, where*

$$b = \frac{4r}{p_-} + \frac{32k\ln(2d/\delta)}{mp_-}. \tag{21}$$

Proof Apply Lemma 11 to D_-. With probability of $1-\delta$ over samples of size n drawn from D_-,

$$\mathbb{E}_{D_-}[\langle w, X\rangle] \leq 2\hat{\mathbb{E}}_{D_-}[\langle w, X\rangle] + \frac{16k\ln(d/\delta)}{n}.$$

Now, consider a sample of size m drawn according to D. Then $\mathbb{E}_{D_-}[\cdot] = \mathbb{E}_D[\cdot \mid Y=-1]$, and $n = m\hat{p}_-$. Therefore, with probability $1-2\delta$,

$$\begin{aligned}\mathbb{E}[\langle w, X\rangle \mid Y=-1] &\leq 2\hat{\mathbb{E}}[\langle w, X\rangle \mid Y=-1] + \frac{16k\ln(d/\delta)}{m\hat{p}_-} \\ &\leq \frac{2r}{\hat{p}_-} + \frac{16k\ln(d/\delta)}{m\hat{p}_-} \\ &\leq \frac{4r}{p_-} + \frac{32k\ln(d/\delta)}{mp_-},\end{aligned} \tag{22}$$

where the second inequality follows from Lemma 10, and the last inequality follows from the assumption on p_- and Proposition 2. ■

This lemma shows that to bound the sample complexity of an ERM algorithm for the Winnow loss, it suffices to bound the convergence rates of the empirical loss for $w \in U_b(D)$, with b defined as in Eq. (21). Thus, we will be able to use Theorem 9 to bound the convergence of the loss on negative examples.

4.2 Convergence on Positive Labels

For positive labels, we show a uniform convergence result that holds for the entire class $\mathcal{H}_{k,\theta}$. The idea of the proof technique below is as follows. First, following a technique in the spirit of the one given by Zhang (2002), we show that the regret bound for the online learning algorithm presented in Section 3 can be used to construct a small cover of the set of loss functions parameterized by $\mathcal{H}_{k,\theta}$. Second, we convert the bound on the size of the cover to a bound on the Rademacher complexity, thus showing a uniform convergence result. This argument is a refinement of Dudley's entropy bound (Dudley, 1967), which is stated in explicit terms by Srebro et al. (2010, Lemma A.3).

We first observe that by Theorem 4, if the conditions of the theorem hold and there is u such that $f_t(u) = 0$ for all t, then

$$\frac{1}{T}\sum_{t=1}^{T} f_t(w_t) \leq 4\sqrt{\frac{\beta k \ln(d)}{T}}. \tag{23}$$

Let $k \geq r \geq 0$ be two real numbers and let $W \subseteq \mathbb{R}^d_+$. Let ϕ_w denote the function defined by $\phi_w(x, y) = \ell(x, y, w)$, and consider the class of functions $\Phi_W = \{\phi_w \mid w \in W\}$. Given $S = ((x_1, y_1), \ldots, (x_m, y_m))$, where $x_i \in [0, 1]^d$ and $y_i \in \{\pm 1\}$, we say that (Φ_W, S) is (∞, ϵ)-properly-covered by a set $V \subseteq \Phi_W$ if for any $f \in \Phi_W$ there is a $g \in V$ such that

$$\|(f(x_1, y_1), \ldots, f(x_m, y_m)) - (g(x_1, y_1), \ldots, g(x_m, y_m))\|_\infty \leq \epsilon.$$

We denote by $\mathbb{N}_\infty(W, S, \epsilon)$ the minimum value of an integer N such that exists a $V \subseteq \Phi_W$ of size N that (∞, ϵ)-properly-covers (Φ_W, S).

The following lemma bounds the covering number for F_W, for sets S with all-positive labels y_i.

Lemma 13 *Let $S = ((x_1, 1), \ldots, (x_m, 1))$, where $x_i \in [0, 1]^d$. Then,*

$$\ln \mathbb{N}_\infty(\mathcal{H}_{k,\theta}, S, \epsilon) \leq 16 \cdot rk \ln(d) \ln(3m)/\epsilon^2.$$

Proof We use a technique in the spirit of the one given by Zhang (2002). Fix some u, with $u \geq 0$ and $\|u\|_1 \leq k$. For each i let

$$g_i^u(w) = \begin{cases} |\langle w, x_i\rangle - \langle u, x_i\rangle| & \text{if } \langle u, x_i\rangle \leq r \\ [r - \langle w, x_i\rangle]_+ & \text{o.w.} \end{cases}$$

and define the function

$$G_u(w) = \max_i g_i^u(w) \ .$$

It is easy to verify that for any w,

$$\|(\phi_w(x_1, 1), \ldots, \phi_w(x_m, 1)) - (\phi_u(x_1, 1), \ldots, \phi_u(x_m, 1))\|_\infty \leq G_u(w).$$

Now, clearly, $G_u(u) = 0$. In addition, for any $w \geq 0$, a sub-gradient of G_u at w is obtained by choosing i that maximizes $g_i^u(w)$ and then taking a sub-gradient of g_i^u, which is of the form $z = \alpha x_i$ where $\alpha \in \{-1, 0, 1\}$. If $\alpha \in \{-1, 1\}$, it is easy to verify that

$$\sum_j w[j] z[j]^2 \leq \langle w, x_i\rangle \leq g_i^u(w) + r = G_u(w) + r \ .$$

If $\alpha = 0$ then clearly $\sum_j w[j]z[j]^2 \leq G_u(w) + r$ as well.

We can now use Eq. (23) by setting $f_t = G_u$ for all t, setting $\alpha = 1$ and $\beta = r$ in Eq. (9), and noting that since $x_i \in [0,1]^d$, we have $z_t \in [-1,1]^d$ for all t. If $\eta \leq 1$ we have $\eta z_t[i] \geq -1$ for all t, i as needed. Since $\eta = \sqrt{\frac{k\ln(d)}{rT}}$, this holds for all $T \geq k\ln(d)/r$.

We conclude that if we run the unnormalized EG algorithm with $T \geq k\ln(d)/r$ and η and λ as required, we get

$$\sum_{t=1}^{T} G_u(w_t) \leq 4\sqrt{rk\ln(d)T}.$$

Dividing by T and using Jensen's inequality we conclude

$$G_u\left(\tfrac{1}{T}\sum_t w_t\right) \leq 4\sqrt{\frac{rk\ln(d)}{T}}.$$

Denote $w_u = \frac{1}{T}\sum_t w_t$. Setting $\epsilon = 4\sqrt{\frac{rk\ln(d)}{T}}$, it follows that the following set is a (∞,ϵ)-proper-cover for $(F_{\mathcal{H}_{k,\theta}}, S)$:

$$V = \{w_u \mid u \in \mathcal{H}_{k,\theta}\}.$$

Now, we only have left to bound the size of V. Consider again the unnormalized EG algorithm. Since $z_t = \alpha x_i$ for some $\alpha \in \{-1,0,+1\}$ and $i \in \{1,\ldots,m\}$, at each round of the algorithm there are only two choices to be made: the value of i and the value of α. Therefore, the number of different vectors produced by running unnormalized EG for T iterations on G_u for different values of u is at most $(3m)^T$. Thus $|V| \leq (3m)^T$. By our definition of ϵ,

$$\ln|V| \leq T\ln(3m) \leq 16rk\ln(d)\ln(3m)/\epsilon^2.$$

This concludes our proof. ■

Using this result we can bound from above the covering number defined using the Euclidean norm: We say that (Φ_W, S) is $(2,\epsilon)$-properly-covered by a set $V \subseteq \Phi_W$ if for any $f \in \Phi_W$ there is a $g \in V$ such that

$$\frac{1}{\sqrt{m}}\|(f(x_1,y_1),\ldots,f(x_m,y_m)) - (g(x_1,y_1),\ldots,g(x_m,y_m))\|_2 \leq \epsilon.$$

We denote by $\mathbb{N}_2(W,S,\epsilon)$ the minimum value of an integer N such that exists a $V \subseteq \Phi_W$ of size N that $(2,\epsilon)$-properly-covers (Φ_W, S). It is easy to see that for any two vectors $u, v \in \mathbb{R}^m$, $\frac{1}{\sqrt{m}}\|u-v\|_2 \leq \|u-v\|_\infty$. It follows that for any W and S, we have $\mathbb{N}_2(W,S,\epsilon) \leq \mathbb{N}_\infty(W,S,\epsilon)$.

The $\mathbb{N}_2$ covering number can be used to bound the Rademacher complexity of (Φ_W, S) using a refinement of Dudley's entropy bound (Dudley, 1967), which is stated explicitly by Srebro et al. (2010, Lemma A.3). The lemma states that for any $\epsilon \geq 0$,

$$\mathcal{R}(W,S) \leq 4\epsilon + \frac{10}{\sqrt{m}}\int_\epsilon^B \sqrt{\ln \mathbb{N}_2(W,S,\gamma)}\,d\gamma,$$

where B is an upper bound on the possible values of $f \in \Phi_W$ on members of S. For S with all-positive labels we clearly have $B \leq r$.

Combining this with Lemma 13, we get

$$\mathcal{R}(\mathcal{H}_{k,\theta}, S) \leq C \cdot \left(\epsilon + \frac{1}{\sqrt{m}} \int_{\epsilon}^{r} \sqrt{rk \ln(d) \ln(3m)}/\gamma \, d\gamma \right) = C \cdot \left(\epsilon + \sqrt{\frac{rk \ln(d) \ln(3m)}{m}} \ln(r/\epsilon) \right).$$

Setting $\epsilon = rk/m$ we get

$$\mathcal{R}(\mathcal{H}_{k,\theta}, S) \leq C \cdot \sqrt{\frac{rk \ln(d) \ln^3(3m)}{m}}.$$

Thus, for any distribution D over $[0,1]^d \times \{\pm 1\}$ that draws only positive labels, we have

$$\mathcal{R}_m(\mathcal{H}_{k,\theta}, D) \leq C \left(\sqrt{\frac{rk \ln(d) \ln^3(3m)}{m}} \right).$$

By Rademacher sample complexity bounds (Bartlett and Mendelson, 2002), and since ℓ for positive labels is bounded by r, we can immediately conclude the following:

Theorem 14 *Let $k \geq r \geq 0$. For any distribution D over $[0,1]^d \times \{\pm 1\}$ that draws only positive labels, with probability $1 - \delta$ over samples of size m, for any $w \in \mathcal{H}_{k,\theta}$,*

$$\begin{aligned} \ell_+(w) &\leq \hat{\ell}_+(w) + C \cdot \left(\sqrt{\frac{rk \ln(d) \ln^3(3m)}{m}} + \sqrt{\frac{r^2 \ln(1/\delta)}{m}} \right) \\ &\leq \hat{\ell}_+(w) + C \cdot \left(\sqrt{\frac{rk(\ln(d) \ln^3(3m) + \ln(1/\delta))}{m}} \right). \end{aligned}$$

4.3 Combining Negative and Positive Losses

We have shown separate convergence rate results for the loss on positive labels and for the loss on negative labels. We now combine these results to achieve a convergence rate upper bound for the full Winnow loss. To do this, the convergence results given above must be adapted to take into account the fraction of positive and negative labels in the true distribution as well as in the sample. The following theorems accomplish this for the negative and the positive cases. First, a bound is provided for the positive part of the loss.

Theorem 15 *There exists a universal constant C such that for any distribution D over $[0,1]^d \times \{\pm 1\}$, with probability $1 - \delta$ over samples of size m*

$$p_+ \ell_+(\hat{w}) \leq \hat{p}_+ \hat{\ell}_+(\hat{w}) + C \cdot \sqrt{\frac{rk(\ln(kd) \ln^3(m) + \ln(3/\delta))}{m}}.$$

Proof First, if $p_+ \leq \frac{8\ln(1/\delta)}{m}$ then the theorem trivially holds. Therefore we assume that $p_+ \geq \frac{8\ln(1/\delta)}{m}$. We have

$$p_+\ell_+(\hat{w}) = \hat{p}_+\hat{\ell}_+(\hat{w}) + (p_+ - \hat{p}_+)\hat{\ell}_+(\hat{w}) + p_+(\ell_+(\hat{w}) - \hat{\ell}_+(\hat{w})). \tag{24}$$

To prove the theorem, we will bound the two rightmost terms. First, to bound $(p_+ - \hat{p}_+)\hat{\ell}_+(\hat{w})$, note that by definition of the loss function for positive labels we have that $\hat{\ell}_+(\hat{w}) \in [0, r]$. Therefore, Bernstein's inequality (Proposition 1) implies that with probability $1-\delta/3$

$$(p_+ - \hat{p}_+)\hat{\ell}_+(\hat{w}) \leq 2r\sqrt{\frac{\ln(3/\delta)}{m}\max\left\{p_+, \frac{\ln(3/\delta)}{m}\right\}} \leq \sqrt{\frac{4r\ln(3/\delta)}{m}}. \tag{25}$$

Second, to bound $p_+(\ell_+(\hat{w}) - \hat{\ell}_+(\hat{w}))$, we apply Theorem 14 to the conditional distribution induced by D on X given $Y = 1$, to get that with probability $1 - \delta/3$

$$p_+(\ell_+(\hat{w}) - \hat{\ell}_+(\hat{w})) \leq p_+ \cdot C \cdot \sqrt{\frac{rk(\ln(d)\ln^3(3m) + \ln(3/\delta))}{m\hat{p}_+}}.$$

Using our assumption on p_+ we obtain from Proposition 2 that with probability $1 - \delta/3$, $p_+/\hat{p}_+ \leq 2$. Therefore, $p_+/\sqrt{\hat{p}_+} \leq \sqrt{2p_+} \leq \sqrt{2}$. Thus, with probability $1 - 2\delta/3$,

$$p_+(\ell_+(\hat{w}) - \hat{\ell}_+(\hat{w})) \leq C \cdot \sqrt{\frac{rk(\ln(d)\ln^3(3m) + \ln(3/\delta))}{m}}. \tag{26}$$

Combining Eq. (24), Eq. (25) and Eq. (26) and applying the union bound, we get the theorem. ■

Second, a bound is provided for the negative part of the loss.

Theorem 16 *There exists a universal constant C such that for any distribution D over $[0,1]^d \times \{\pm 1\}$, with probability $1 - \delta$ over samples of size m*

$$p_-\ell_-(\hat{w}) \leq \hat{p}_-\hat{\ell}_-(\hat{w}) + C\left(\sqrt{\frac{rk\ln(d/\delta)}{m}} + \frac{k\ln(dm/\delta)}{m}\right). \tag{27}$$

Proof First, if $p_- \leq \frac{8\ln(1/\delta)}{m}$ then the theorem trivially holds (since $\ell_-(\hat{w}) \in [0, r+k]$). Therefore we assume that $p_- \geq \frac{8\ln(1/\delta)}{m}$. Thus, by Proposition 2, $\hat{p}_- \geq p_-/2$. We have

$$p_-\ell_-(\hat{w}) = \hat{p}_-\hat{\ell}_-(\hat{w}) + (p_- - \hat{p}_-)\hat{\ell}_-(\hat{w}) + p_-(\ell_-(\hat{w}) - \hat{\ell}_-(\hat{w})). \tag{28}$$

To prove the theorem, we will bound the two rightmost terms. First, to bound $(p_- - \hat{p}_-)\hat{\ell}_-(\hat{w})$, note that by Bernstein's inequality (Proposition 1) and our assumption on p_-, with probability $1 - \delta$

$$p_- - \hat{p}_- \leq 2\sqrt{\frac{\ln(1/\delta)}{m}\max\left\{p_-, \frac{\ln(1/\delta)}{m}\right\}} = 2\sqrt{\frac{p_-\ln(1/\delta)}{m}}.$$

By Lemma 10 and Proposition 2, $\hat{\ell}_-(\hat{w}) \leq \frac{2r}{\hat{p}_-} \leq \frac{4r}{p_-}$. In addition, by definition $\hat{\ell}_-(\hat{w}) \leq r + k \leq 2k$. Therefore

$$(p_- - \hat{p}_-)\hat{\ell}_-(\hat{w}) \leq 4 \min\left\{\frac{2r}{p_-}, k\right\} \sqrt{\frac{p_- \ln(1/\delta)}{m}}. \tag{29}$$

Now, if $k > 2r/p_-$, then the right-hand of the above becomes

$$8\frac{r}{p_-}\sqrt{\frac{p_- \ln(1/\delta)}{m}} = 8\sqrt{\frac{(r/p_-) \cdot r \ln(1/\delta)}{m}} \leq 8\sqrt{\frac{k \cdot r \ln(1/\delta)}{m}} .$$

Otherwise, $k \leq 2r/p_-$ and the right-hand of Eq. (29) becomes

$$4k\sqrt{\frac{p_- \ln(1/\delta)}{m}} \leq 4k\sqrt{\frac{(2r/k) \ln(1/\delta)}{m}} \leq 8\sqrt{\frac{k \cdot r \ln(1/\delta)}{m}} .$$

All in all, we have shown that

$$(p_- - \hat{p}_-)\hat{\ell}_-(\hat{w}) \leq 8\sqrt{\frac{rk \ln(1/\delta)}{m}}. \tag{30}$$

Second, to bound $p_-(\ell_-(\hat{w}) - \hat{\ell}_-(\hat{w}))$, recall that by Lemma 12, we have $\hat{w} \in U_b(D)$, where

$$b = \frac{4r}{p_-} + \frac{32k \ln(d/\delta)}{mp_-} \leq \frac{C}{p_-}\left(2r + \frac{k \ln(d/\delta)}{m}\right).$$

Thus, by Theorem 9, with probability $1 - \delta$

$$\ell_-(w) \leq \hat{\ell}_-(w) + C\left(\sqrt{\frac{kb \ln(d/\delta)}{m\hat{p}_-}} + \frac{k \ln(dm/\delta)}{m\hat{p}_-}\right).$$

Since $\hat{p}_- \geq p_-/2$,

$$\ell_-(w) \leq \hat{\ell}_-(w) + C\left(\sqrt{\frac{kb \ln(d/\delta)}{mp_-}} + \frac{k \ln(dm/\delta)}{mp_-}\right).$$

for some other constant C. Therefore, substituting b for its upper bound we get

$$p_-(\ell_-(w) - \hat{\ell}_-(w)) \leq C\left(\sqrt{\frac{kr \ln(d/\delta)}{m}} + \frac{k \ln(dm/\delta)}{m}\right). \tag{31}$$

Combining Eq. (28), Eq. (30) and Eq. (31) we get the statement of the theorem. ■

Finally, we prove our main result for the sample complexity of ERM algorithms for Winnow.

Proof (Proof of Theorem 6) From Theorem 15 and Theorem 16 we conclude that with probability $1-\delta$,

$$\begin{aligned}\ell(\hat{w}) &= p_-\ell_-(\hat{w}) + p_+\ell_+(\hat{w}) \\ &\leq \hat{p}_-\hat{\ell}_-(\hat{w}) + \hat{p}_+\hat{\ell}_+(\hat{w}) + \sqrt{\frac{O(rk(\ln(d)\ln^3(3m)+\ln(1/\delta)))}{m}}. \end{aligned} \tag{32}$$

Now,

$$\hat{p}_-\hat{\ell}_-(\hat{w}) + \hat{p}_+\hat{\ell}_+(\hat{w}) = \hat{\ell}(\hat{w}) \leq \hat{\ell}(w^*). \tag{33}$$

We have $\mathbb{E}[\ell(X,Y,w^*)] = \ell(w^*) \leq \ell(\mathbf{0}) \leq r$. By Bernstein's inequality (Proposition 1), with probability $1-\delta$

$$\begin{aligned}\hat{\ell}(w^*) = \hat{\mathbb{E}}[\ell(X,Y,w^*)] &\leq \mathbb{E}[\ell(X,Y,w^*)] + 2r\sqrt{\frac{\ln(1/\delta)}{m}\max\left\{\frac{\mathbb{E}[\ell(X,Y,w^*)]}{r}, \frac{\ln(1/\delta)}{m}\right\}} \\ &\leq \ell(w^*) + 2\sqrt{\frac{r^2\ln(1/\delta)}{m}} + 2\frac{r\ln(1/\delta)}{m}.\end{aligned}$$

Combining this with Eq. (33), we get that with probability $1-\delta$

$$\hat{p}_-\hat{\ell}_-(\hat{w}) + \hat{p}_+\hat{\ell}_+(\hat{w}) \leq \ell(w^*) + 2\sqrt{\frac{r^2\ln(1/\delta)}{m}} + 2\frac{r\ln(1/\delta)}{m}.$$

In light of Eq. (32), we conclude Eq. (13) ■

Theorem 6 shows that using empirical risk minimization, the loss of the obtained predictor converges to the loss of the optimal predictor at a rate of the order

$$\tilde{O}\left(\sqrt{\frac{rk\log(d)}{m}}\right) \equiv \tilde{O}\left(\sqrt{\frac{\theta k\log(d)}{m}}\right).$$

Up to logarithmic factors, this is the best possible rate for learning in the generalized Winnow setting. This is shown in the next section, in Theorem 17. We also show, in Theorem 21, that this rate cannot be obtain via standard uniform convergence analysis.

5. Lower Bounds

In this section we provide lower bounds for the learning rate and for the uniform convergence rate of the Winnow loss ℓ_θ.

5.1 Learning Rate Lower Bound

Fix a threshold θ. The best Winnow loss for a distribution D over $[0,1]^d \times \{\pm 1\}$ using a hyperplane from a set $W \subseteq \mathbb{R}^d_+$ is denoted by $\ell^*_\theta(W) = \min_{w\in W}\ell_\theta(w)$. The following result shows that even if the data domain is restricted to the discrete domain $\{0,1\}^d$, the number of samples required for learning with the Winnow loss grows at least linearly in θk. This resolves an open question posed by Littlestone (1988).

Theorem 17 *Let $k \geq 1$ and let $\theta \in [1, k/2]$. The sample complexity of learning $\mathcal{H}_{k,\theta}$ with respect to the loss ℓ_θ is $\Omega(\theta k/\epsilon^2)$. That is, for all $\epsilon \in (0, 1/2)$ if the training set size is $m = o(\theta k/\epsilon^2)$, then for any learning algorithm, there exists a distribution such that the classifier, $h : \{0,1\}^d \to \mathbb{R}_+$, that the algorithm outputs upon receiving m i.i.d. examples satisfies $\ell_\theta(h) - \ell^*_\theta(\mathcal{H}_{k,\theta}) > \epsilon$ with a probability of at least $1/4$.*

The construction which shows the lower bound proceeds in several stages: First, we prove that there exists a set of size k^2 in $\{\pm 1\}^{k^2}$ which is shattered on the linear loss with respect to predictors with a norm bounded by k. Then, apply a transformation on this construction to show a set in $\{0,1\}^{2k^2+1}$ which is shattered on the linear loss with a threshold of $k/2$. In the next step, we adapt the construction to hold for any value of the threshold. Finally, we use the resulting construction to prove Theorem 17.

The construction uses the notion of a *Hadamard matrix*. A Hadamard matrix of order n is an $n \times n$ matrix H_n with entries in $\{\pm 1\}$ such that $H_n H_n^T = n I_n$. In other words, all rows in the matrix are orthogonal to each other. Hadamard matrices exist at least for each n which is a power of 2 (Sylvester, 1867). The first lemma constructs a shattered set for the linear loss on $\{\pm 1\}^{k^2}$.

Lemma 18 *Assume k is a power of 2, and let $d = k^2$. Let $x_1, \ldots, x_d \subseteq \{\pm 1\}^d$ be the rows of the Hadamard matrix of order d. For every $y \in \{\pm 1\}^d$, there exists a $w \in W' = \{w \in [-1,1]^d \mid \|w\| \leq k\}$ such that for all $i \in [d]$, $y[i]\langle w, x_i\rangle = 1$.*

Proof By the definition of a Hadamard matrix, for all $i \neq j$, $\langle x_i, x_j\rangle = 0$. Given $y \in \{\pm 1\}^d$, set $w = \frac{1}{d}\sum_{j\in[d]} y_j x_j$. Then for each i,

$$y_i\langle w, x_i\rangle = y_i \frac{1}{d}\sum_{j\in[d]} y_j\langle x_i, x_j\rangle = \frac{1}{d} y_i^2 \langle x_i, x_i\rangle = \frac{1}{d}\|x_i\|_2^2 = 1.$$

It is left to show that $w \in W'$. First, for all $i \in [d]$, we have

$$|w[i]| = |\frac{1}{d}\sum_{j\in[d]} y_j x_j[i]| \leq \frac{1}{d}\sum_{j\in[d]} |x_j[i]| = 1,$$

which yields $w \in [-1,1]^d$. Second, using $\|w\|_1 \leq \sqrt{d}\|w\|_2$ and

$$\|w\|_2^2 = \langle w, w\rangle = \frac{1}{d^2}\sum_{i,j\in[d]} \langle y_i x_i, y_j x_j\rangle = \frac{1}{d^2}\sum_{i\in[d]} y_i^2\langle x_i, x_i\rangle = \frac{1}{d^2}\sum_{i\in[d]} d = 1,$$

we obtain that $\|w\|_1 \leq \sqrt{d} = k$. ■

The next lemma transforms the construction from Lemma 18 to a linear loss with a threshold of $k/2$.

Lemma 19 *Let k be a power of 2 and let $d = 2k^2 + 1$. There is a set $\{x_1, \ldots, x_{k^2}\} \subseteq \{0,1\}^d$ such that for every $y \in \{\pm 1\}^{k^2}$, there exists $w \in \mathcal{H}_{k,\theta}$ such that for all $i \in [k^2]$, $y[i](\langle w, x_i\rangle - k/2) = \frac{1}{2}$.*

Proof From Lemma 18 we have that there is a set $X = \{x_1, \ldots, x_{k^2}\} \subseteq \{\pm 1\}^{k^2}$ such that for each labeling $y \in \{\pm 1\}^{k^2}$, there exists a $w_y \in [-1, 1]^d$ with $\|w_y\|_1 \leq k$ such that for all $i \in [k^2]$, $y[i]\langle w_y, x_i\rangle = 1$. We now define a new set $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_{k^2}\} \subseteq \{0,1\}^d$ based on X that satisfies the requirements of the lemma.

For each $i \in [k^2]$ let $\tilde{x}_i = [\frac{\vec{1}+x_i}{2}, \frac{\vec{1}-x_i}{2}, 1]$, where $[\cdot,\cdot,\cdot]$ denotes a concatenation of vectors and $\vec{1}$ is the all-ones vector. In words, each of the first k^2 coordinates in $\tilde{x}_i$ is 1 if the corresponding coordinate in x_i is 1, and zero otherwise. Each of the next k^2 coordinates in $\tilde{x}_i$ is 1 if the corresponding coordinate in x_i is -1, and zero otherwise. The last coordinate in $\tilde{x}_i$ is always 1.

Now, let $y \in \{\pm 1\}^{k^2}$ be a desired labeling. We defined $\tilde{w}_y$ based on w_y as follows: $\tilde{w}_y = [[w_y]_+, [-w_y]_+, \frac{k-\|w_y\|_1}{2}]$, where by $z = [v]_+$ we mean that $z[j] = \max\{v[j], 0\}$. In words, the first k^2 coordinates of $\tilde{w}_y$ are copies of the positive coordinates of w_y, with zero in the negative coordinates, and the next k^2 coordinates of $\tilde{w}_y$ are the absolute values of the negative coordinates of w_y, with zero in the positive coordinates. The last coordinate is a scaling term.

We now show that $\tilde{w}_y$ has the desired property on $\tilde{X}$. For each $i \in [k^2]$,

$$\begin{aligned}\langle \tilde{w}_y, \tilde{x}_i\rangle &= \left\langle \frac{\vec{1}+x_i}{2}, [w_y]_+ \right\rangle + \left\langle \frac{\vec{1}-x_i}{2}, [-w_y]_+ \right\rangle + \frac{k-|w_y|_1}{2} \\ &= \frac{|w_y|_1}{2} + \frac{\langle x_i, w_y\rangle}{2} + \frac{k-|w_y|_1}{2} = \frac{\langle x_i, w_y\rangle}{2} + \frac{k}{2} = \frac{y_i}{2} + \frac{k}{2}.\end{aligned}$$

It follows that $y_i(\langle \tilde{w}_y, \tilde{x}_i\rangle - k/2) = y_i^2/2 = 1/2$.

Now, clearly $\tilde{w}_y \in \mathbb{R}_+^d$. In addition,

$$\|\tilde{w}_y\|_1 = \|w_y\|_1 + \frac{k-\|w_y\|_1}{2} = \frac{\|w_y\|_1}{2} + \frac{k}{2} \leq k.$$

Hence $\tilde{w}_y \in \mathcal{H}_{k,\theta}$ as desired. ■

The last lemma adapts the previous construction to hold for any threshold.

Lemma 20 *Let z be a power of 2 and let k such that z divides k. Let $d = 2kz + k/z$. There is a set $\{x_1, \ldots, x_{zk}\} \subseteq \{0,1\}^d$ such that for every $y \in \{\pm 1\}^{zk}$, there exists a $w \in \mathcal{H}_{k,\theta}$ such that for all $i \in [zk]$, $y[i](\langle w, x_i\rangle - z/2) = \frac{1}{2}$.*

Proof By Lemma 19 there is a set $X = \{x_1, \ldots, x_{z^2}\} \subseteq \{0,1\}^{2z^2+1}$ such that for all $y \in \{\pm 1\}^{z^2}$, there exists a $w_y \in \mathbb{R}_+^{2z^2+1}$ such that $\|w_y\|_1 \leq z$ and for all $i \in [z^2]$, $y[i](\langle w_y, x_i\rangle - z/2) = \frac{1}{2}$.

We now construct a new set $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_{zk}\} \subseteq \{0,1\}^{2kz+k/z}$ as follows: For $i \in [zk]$, let $n = \lfloor i/z^2 \rfloor$ and $m = i \mod z^2$, so that $i = nz^2 + m$.The vector $\tilde{x}_i$ is the concatenation of $\frac{kz}{z^2} = \frac{k}{z}$ vectors, each of which is of dimension $2z^2+1$, where all the vectors are the all-zeros vector, except the $(n+1)$'th vector which equals to x_{m+1}. That is:

$$\tilde{x}_i = [\overbrace{0}^{\in \mathbb{R}^{2z^2+1}}, \ldots, \overbrace{0}^{\in \mathbb{R}^{2z^2+1}}, \overbrace{x_{m+1}}^{\text{block } n+1}, \overbrace{0}^{\in \mathbb{R}^{2z^2+1}}, \ldots, \overbrace{0}^{\in \mathbb{R}^{2z^2+1}}] \in \mathbb{R}^{\frac{k}{z}(2z^2+1)}.$$

Given $\tilde{y} \in \{\pm 1\}^{kz}$, let us rewrite it as a concatenation of k/z vectors, each of which in $\{\pm 1\}^{z^2}$, namely,

$$\tilde{y} = [\overbrace{\tilde{y}(1)}^{\in\{\pm 1\}^{z^2}}, \ldots, \overbrace{\tilde{y}(k/z)}^{\in\{\pm 1\}^{z^2}}] \in \{\pm 1\}^{kz} .$$

Define $\tilde{w}_{\tilde{y}}$ as the concatenation of k/z vectors in $\{\pm 1\}^{z^2}$, using w_y defined above for each $y \in \{\pm 1\}^{z^2}$, as follows:

$$\tilde{w}_{\tilde{y}} = [\overbrace{w_{\tilde{y}(1)}}^{\in\mathbb{R}_+^{2z^2+1}}, \ldots, \overbrace{w_{\tilde{y}(k/z)}}^{\in\mathbb{R}_+^{2z^2+1}}] \in \mathbb{R}_+^{\frac{k}{z}(2z^2+1)} .$$

For each i such that $n = \lfloor i/z^2 \rfloor$ and $m = i \mod z^2$, we have

$$\langle \tilde{w}_{\tilde{y}}, \tilde{x}_i \rangle - z/2 = \langle w_{\tilde{y}(n+1)}, x_{m+1} \rangle - z/2 = \frac{1}{2}\tilde{y}(n+1)[m+1].$$

Now $\tilde{y}(n+1)[m+1] = \tilde{y}[i]$, thus we get $\tilde{y}[i](\langle \tilde{w}_{\tilde{y}}, \tilde{x}_i \rangle - z/2) = \frac{1}{2}$ as desired. Finally, we observe that $\|\tilde{w}_{\tilde{y}}\|_1 = \sum_{n\in[k/z]} \|w_{\tilde{y}(n)}\|_1 \leq k/z \cdot z = k$, hence $\tilde{w}_{\tilde{y}} \in \mathcal{H}_{k,\theta}$. ■

Finally, the construction above is used to prove the convergence rate lower bound.

Proof (Proof of Theorem 17) Let $k \geq 1$, $\theta \in [\frac{1}{2}, \frac{k}{2}]$. Define $z = 2\theta$. Let $n = \max\{n \mid 2^n \leq z\}$, and let $m = \max\{m \mid m2^n \leq k\}$. Define $\tilde{z} = 2^n$ and $\tilde{k} = m2^n$. We have that $\tilde{z}$ is a power of 2 and $\tilde{z}$ divides $\tilde{k}$. Let $\tilde{d} = 2\tilde{k}\tilde{z} + \tilde{k}/\tilde{z}$. By Lemma 20, there is a set $X = \{x_1, \ldots, x_{\tilde{z}\tilde{k}}\} \subseteq \{0,1\}^{\tilde{d}}$ such that for every $y \in \{\pm 1\}^{|X|}$, there exists a $w_y \in \mathcal{H}_{k,\theta}$ such that for all $i \in [\tilde{z}\tilde{k}]$, $y[i](\langle w_y, x_i \rangle - \tilde{z}/2) = \frac{1}{2}$.

Now, let $d = \tilde{d} + 1$, and define $\tilde{w}_y = [w_y, \frac{z-\tilde{z}}{2}]$ and $\tilde{x}_i = [x_i, 1]$. It follows that

$$\begin{aligned} y[i](\langle \tilde{w}_y, \tilde{x}_i \rangle - \theta) &= y[i](\langle \tilde{w}_y, \tilde{x}_i \rangle - z/2) \\ &= y[i](\langle w_y, x_i \rangle + z/2 - \tilde{z}/2 - z/2) \\ &= y[i](\langle w_y, x_i \rangle - \tilde{z}/2) = \frac{1}{2}. \end{aligned}$$

We conclude that for all $i \in [\tilde{z}\tilde{k}]$, $\ell_\theta(\tilde{x}_i, y[i], \tilde{w}_y) = 0$ and $\ell_\theta(\tilde{x}_i, 1 - y[i], \tilde{w}_y) = 1$. Moreover, $\text{sign}(\langle \tilde{w}_y, \tilde{x}_i \rangle - \theta) = y[i]$.

Now, for a given w define $h_w(x) = \text{sign}(\langle w, x_i \rangle - \theta)$, and consider the binary hypothesis class $H = \{h_w \mid w \in \mathcal{H}_{k,\theta}\}$ over the domain X. Our construction of $\tilde{w}_y$ shows that the set X is shattered by this hypothesis class, thus its VC dimension is at least $|X|$. By VC-dimension lower bounds (e.g., Anthony and Bartlett, 1999, Theorem 5.2), it follows that for any learning algorithm for H, if the training set size is $o(|X|/\epsilon^2)$, then there exists a distribution over X so that with probability greater than $1/64$, the output $\hat{h}$ of the algorithm satisfies

$$\mathbb{E}[\hat{h}(x) \neq y] > \min_{w\in\mathcal{H}_{k,\theta}} \mathbb{E}[h_w(x) \neq y] + \epsilon . \tag{34}$$

Next, we show that the existence of a learning algorithm for $\mathcal{H}_{k,\theta}$ with respect to ℓ_θ whose sample complexity is $o(|X|/\epsilon^2)$ would contradict the above statement. Indeed, let w^* be a minimizer of the right-hand side of Eq. (34), and let y^* be the vector of predictions

of w^* on X. As our construction of $\tilde{w}_{y^*}$ shows, we have $\ell_\theta(\tilde{w}_{y^*}) = \mathbb{E}[h_{w^*}(x) \neq y]$. Now, suppose that some algorithm learns $\hat{w} \in \mathcal{H}_{k,\theta}$ so that $\ell_\theta(\hat{w}) \leq \ell^*_\theta(\mathcal{H}_{k,\theta}) + \epsilon$. This implies that

$$\ell_\theta(\hat{w}) \leq \ell_\theta(\tilde{w}_{y^*}) + \epsilon = \mathbb{E}[h_{w^*}(x) \neq y] + \epsilon \ .$$

In addition, define a (probabilistic) classifier, $\hat{h}$, that outputs the label $+1$ with probability $p(\hat{w}, x)$ where $p(\hat{w}, x) = \min\{1, \max\{0, 1/2 + (\langle \hat{w}, x\rangle - \theta)\}\}$. Then, it is easy to verify that

$$\mathbb{P}[\hat{h}(x) \neq y] \leq \ell_\theta(x, y, \hat{w}) \ .$$

Therefore, $\mathbb{E}[\hat{h}(x) \neq y] \leq \ell_\theta(\hat{w})$, and we obtain that

$$\mathbb{E}[\hat{h}(x) \neq y] \leq \mathbb{E}[h_{w^*}(x) \neq y] + \epsilon \ ,$$

which leads to the desired contradiction. ■

We next show that the uniform convergence rate for our problem is in fact slower than the achievable learning rate.

5.2 Uniform Convergence Lower Bound

The next theorem shows that the rate of uniform convergence for our problem is asymptotically slower than the rate of convergence of the empirical loss minimizer given in Theorem 6, even if the drawn label in a random pair is negative with probability 1. This indicates that indeed, a more subtle argument than uniform convergence is needed to show that ERM learns at a rate of $\tilde{O}(\sqrt{\theta k/m})$, as done in Section 4.

Theorem 21 *Let $k \geq 1$, and assume $\theta \leq k/2$. There exists a distribution D over $\{0,1\}^{k^2+1} \times Y$ such that $\forall x \in \{0,1\}^d, \mathbb{P}[Y = -1 \mid X = x] = 1$, and $\ell^*(\mathcal{H}_{k,\theta}, D) = [r']_+$, and such that with probability at least $1/2$ over samples $S \sim D^m$,*

$$\exists w \in \mathcal{H}_{k,\theta}, \quad |\ell(w, S) - \ell(w, D)| \geq \Omega(\sqrt{k^2/m}). \tag{35}$$

This claim may seem similar to well-known uniform convergence lower bounds for classes with a bounded VC dimension (see, e.g., Anthony and Bartlett, 1999, Chapter 5). However, these standard results rely on constructions with non-realizable distributions, while Theorem 21 asserts the existence of a realizable distribution which exhibits this lower bound.

To prove this theorem we first show two useful lemmas. The first lemma shows that a lower bound on the uniform convergence of a function class can be derived from a lower bound on the Rademacher complexity of a related function class.

Lemma 22 *Let Z be a set, and consider a function class $F \subseteq [0,1]^Z$. Let D be a distribution over Z. Let $\bar{F} = \{(x_1, x_2) \to f(x_1) - f(x_2) \mid f \in F\}$. With probability at least $1 - \delta$ over samples $S \sim D^m$,*

$$\exists f \in F, \quad |\mathbb{E}_{X \sim S}[f(X)] - \mathbb{E}_{X \sim D}[f(X)]| \geq \frac{1}{4}\mathcal{R}_m(\bar{F}, D \times D) - \sqrt{\frac{\ln(1/\delta)}{8m}}. \tag{36}$$

Proof Denote $E[f,S] = \mathbb{E}_{X\sim S}[f(X)]$, and $E[f,D] = \mathbb{E}_{X\sim D}[f(X)]$. Consider two independent samples $S = (X_1,\ldots,X_m), S' = (X'_1,\ldots,X'_m) \sim D^m$. Let $\sigma = (\sigma_1,\ldots,\sigma_m)$ be Rademacher random variables, and let $S \sim (D\times D)^m$. We have

$$\begin{aligned}
2\cdot\mathbb{E}_S\left[\sup_{f\in F}|E[f,S]-E[f,D]|\right] &= \mathbb{E}_{S,S'}\left[\sup_{f\in F}|E[f,S]-E[f,D]| + \sup_{f\in F}|E[f,S']-E[f,D]|\right] \\
&\geq \mathbb{E}_{S,S'}\left[\sup_{f\in F}|E[f,S]-E[f,D]| + |E[f,S']-E[f,D]|\right] \\
&\geq \mathbb{E}_{S,S'}\left[\sup_{f\in F}|E[f,S]-E[f,S']|\right] \\
&= \frac{1}{m}\mathbb{E}_{S,S'}\left[\sup_{f\in F}\left|\sum_{i\in[m]} f(X_i)-f(X'_i)\right|\right] \\
&= \frac{1}{m}\mathbb{E}_{\sigma,\bar{S}}\left[\sup_{\bar{f}\in\bar{F}}\left|\sum_{i\in[m]}\sigma_i\bar{f}(X_i)\right|\right] = \mathcal{R}_m(\bar{F}, D\times D)/2.
\end{aligned}$$

We have left to show a lower bound with high probability. Define $g(S) = \sup_{f\in F}|E[f,S] - E[f,D]|$. Any change of one element in S can cause $g(S)$ to change by at most $1/m$, Therefore, by McDiarmid's inequality, $\mathbb{P}[g(S) \leq \mathbb{E}[g(S)] - t] \leq \exp(-2mt^2)$. Eq. (36) thus holds with probability $1-\delta$. ■

The next lemma provides a uniform convergence lower bound for a universal class of binary functions.

Lemma 23 *Let $H = \{0,1\}^{[n]}$ be the set of all binary functions on $[n]$. Let D be the uniform distribution over $[n]$. For any $n \geq 45$ and $m \geq 32n$, with probability of at least $\frac{1}{2}$ over i.i.d. samples of size m drawn from D,*

$$\exists h\in H, \quad |\mathbb{E}_{X\sim S}[h(X)] - \mathbb{E}_{X\sim D}[h(X)]| \geq \sqrt{\frac{n}{512m}}.$$

Proof Let $n \geq 45$ and $m \geq 32n$. By Lemma 22, it suffices to provide a lower bound for $\mathcal{R}_m(\bar{H}, D\times D)$. Fix a sample $S = ((x_1,x'_1),\ldots,(x_m,x'_m)) \sim (D\times D)^m$. We have

$$\frac{m}{2}\mathcal{R}(\bar{H},S) = \mathbb{E}_\sigma\left[\left|\sup_{h\in H}\sum_{i=1}^m \sigma_i(h(x_i)-h(x'_i))\right|\right],$$

where $\sigma = (\sigma_1, \ldots, \sigma_m)$ are Rademacher random variables. For a given $\sigma \in \{\pm 1\}^m$, define $h_\sigma \in H$ such that $h_\sigma(j) = \text{sign}(\sum_{i:x_i=j} \sigma_i - \sum_{i:x'_i=j} \sigma_i)$. Then

$$\begin{aligned}\frac{m}{2}\mathcal{R}(\bar{H}, S) &\geq \mathbb{E}_\sigma \left[\left| \sum_{i\in[m]} \sigma_i (h_\sigma(x_i) - h_\sigma(x'_i)) \right| \right] \\ &= \mathbb{E}_\sigma \left[\left| \sum_{j\in[n]} \left(\sum_{i:x_i=j} \sigma_i - \sum_{i:x'_i=j} \sigma_i \right) h_\sigma(j) \right| \right] \\ &= \sum_{j\in[n]} \mathbb{E}_\sigma \left[\left| \sum_{i:x_i=j} \sigma_i - \sum_{i:x'_i=j} \sigma_i \right| \right].\end{aligned}$$

Let $c_j(S)$ be the number of indices i such that exactly one of $x_i = j$ and $x'_i = j$ holds. Then $\mathbb{E}_\sigma[|\sum_{i:x_i=j} \sigma_i - \sum_{i:x'_i=j} \sigma_i|]$ is the expected distance of a random walk of length $c_j(S)$, which can be bounded from below by $\sqrt{c_j(S)/2}$ (Szarek, 1976). Therefore,

$$\mathcal{R}(\bar{H}, S) \geq \frac{\sqrt{2}}{m} \sum_{j\in[n]} \sqrt{c_j(S)}.$$

Taking expectation over samples, we get

$$\mathcal{R}(\bar{H}, D \times D) = \mathbb{E}_{S\sim(D\times D)^m}[\mathcal{R}(\bar{H}, S)] \geq \frac{\sqrt{2}}{m} \sum_{j\in[n]} \mathbb{E}_S \left[\sqrt{c_j(S)} \right]. \tag{37}$$

Our final step is to bound $\mathbb{E}_S \left[\sqrt{c_j(S)} \right]$. We have

$$\mathbb{E}_S[c_j(S)] = m \left(\frac{1}{n} - \frac{1}{n^2} \right) \geq \frac{m}{2n},$$

and

$$\text{Var}_S[c_j(S)] = m \left(\frac{1}{n} - \frac{1}{n^2} \right) \left(1 - \frac{1}{n} + \frac{1}{n^2} \right) \leq \frac{m}{n}.$$

Thus, by Chebyshev's inequality,

$$\mathbb{P} \left[c_j(S) \leq \frac{m}{2n} - t \right] \leq \frac{m}{nt^2}.$$

Therefore

$$\mathbb{E}_S \left[\sqrt{c_j(S)} \right] \geq \left(1 - \frac{m}{nt^2} \right) \sqrt{\frac{m}{2n} - t}.$$

Setting $t = \frac{m}{4n}$, and since $m/n \geq 32$, $\mathbb{E}_S \left[\sqrt{c_j(S)} \right] \geq \sqrt{\frac{m}{16n}}$. Plugging this into Eq. (37), we get that $\mathcal{R}(\bar{H}, D \times D) \geq \sqrt{\frac{n}{8m}}$. By Lemma 22, it follows that with probability at least $1 - \delta$ over samples,

$$\exists f \in F, \quad |\mathbb{E}_{X\sim S}[f(X)] - \mathbb{E}_{X\sim D}[f(X)]| \geq \sqrt{\frac{n}{128m}} - \sqrt{\frac{\ln(1/\delta)}{8m}}.$$

Fixing $\delta = 1/2$, we get that since $n \geq 64\ln(2)$, the RHS is at least $\sqrt{\frac{n}{512m}}$. ■

Using the two lemmas above, we are now ready to prove our uniform convergence lower bound. This is done by mapping a subset of $\mathcal{H}_{k,\theta}$ to a universal class of binary functions over $\Theta(k^2)$ elements from our domain. Note that for this lower bound it suffices to consider the more restricted domain of binary vectors.

Proof (Proof of Theorem 21) Let q be the largest power of 2 such that $q \leq k$. By Lemma 19, there exists a set of vectors $Z = \{z_1, \ldots, z_{q^2}\} \subseteq \{0,1\}^{q^2+1}$ such that for every $t \in \{\pm 1\}^{q^2}$ there exists a $w_t \in \mathcal{H}_{k,\theta}$ such that for all i, $t[i](\langle w, z_i\rangle - q/2) = \frac{1}{2}$. Denote $U = \{w_t \mid t \in \{\pm 1\}^{q^2}\}$. It suffices to prove a lower bound on the uniform convergence of U, since this implies the same lower bound for $\mathcal{H}_{k,\theta}$. Define the distribution D over $Z \times \{\pm 1\}$ such that for $(X, Y) \sim D$, X is drawn uniformly from $z_1, \ldots, z_{q^2}$ and $Y = -1$ with probability 1.

Consider the set of functions $H = \{0,1\}^Z$, and for $h \in H$ define $t_h \in \{\pm 1\}^{q^2}$ such that for all $i \in [q^2]$, $t_h[i] = 2h(z_i) - 1$. For any $i \in q^2$, we have

$$\ell(z_i, -1, w_{t_h}) = [r' + \langle w, z_i\rangle]_+ = [r' + (t[i]+k)/2]_+ = [r' + (k-1)/2 + h(i)]_+ = r' + (k-1)/2 + h(z_i).$$

The last equality follows since $r' \geq \frac{1-k}{2}$. It follows that for any $h \in H$ and any sample S drawn from D,

$$|\ell(w_{t_h}, S) - \ell(w_{t_h}, D)| = |\mathbb{E}_{X\sim S}[h(X)] - \mathbb{E}_{X\sim D}[h(X)]|.$$

By Lemma 23, with probability of at least $\frac{1}{2}$ over the sample $S \sim D^m$,

$$\exists h \in H, \quad |\mathbb{E}_{X\sim S}[h(X)] - \mathbb{E}_{X\sim D}[h(X)]| \geq \Omega(\sqrt{q^2/m}) = \Omega(\sqrt{k^2/m}).$$

Thus, with probability at least $1/2$,

$$\exists w \in \mathcal{H}_{k,\theta}, \quad |\ell(w_{t_h}, S) - \ell(w_{t_h}, D)| \geq \Omega(\sqrt{k^2/m}).$$

■

Acknowledgements

Tong Zhang is supported by the following grants: NSF IIS1407939, NSF IIS1250985, and NIH R01AI116744.

References

D. Angluin and L. G. Valiant. Fast probabilistic algorithms for Hamiltonian circuits and matchings. *Journal of Computer and System Sciences*, 18(2):155–193, April 1979.

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

P. Auer and M.K. Warmuth. Tracking the best disjunction. *Machine Learning*, 32(2): 127–150, 1998.

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

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

S. Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow, 1946.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

R.M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290 – 330, 1967.

C. Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53:265–299, 2003.

S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Proceedings of NIPS*, 2009.

J. Kivinen and M. Warmuth. Additive versus exponentiated gradient updates for learning linear functions. Technical Report UCSC-CRL-94-16, University of California Santa Cruz, Computer Research Laboratory, 1994.

N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

N. Littlestone. *Mistake Bounds and Logarithmic Linear-Threshold Learning Algorithms*. PhD thesis, U. C. Santa Cruz, March 1989.

P. Massart. Some applications of concentration inequalities to statistics. In *Annales de la Faculté des Sciences de Toulouse*, volume 9:2, pages 245–303. Université Paul Sabatier, 2000.

R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Machine Learning: Proceedings of the Fourteenth International Conference*, pages 322–330, 1997. To appear, *The Annals of Statistics*.

S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University, 2007.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low-noise and fast rates. *CoRR*, abs/1009.3896, 2010.

N. Srebro, K. Sridharan, and A. Tewari. On the universality of online mirror descent. *Advances in Neural Information Processing Systems (NIPS)*, 2011.

J.J. Sylvester. Thoughts on inverse orthogonal matrices, simultaneous signsuccessions, and tessellated pavements in two or more colours, with applications to newton's rule, ornamental tile-work, and the theory of numbers. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 34(232):461–475, 1867.

S.J. Szarek. On the best constants in the Khinchin inequality. *Studia Math*, 58(2), 1976.

T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.