A New Perspective on Learning Linear Separators with Large L_qL_p Margins

Maria-Florina Balcan
Georgia Institute of Technology

Christopher Berlind
Georgia Institute of Technology

Abstract

We give theoretical and empirical results that provide new insights into large margin learning. We prove a bound on the generalization error of learning linear separators with large L_qL_p margins (where L_q and L_p are dual norms) for any finite $p \geq 1$. The bound leads to a simple data-dependent sufficient condition for fast learning in addition to extending and improving upon previous results. We also provide the first study that shows the benefits of taking advantage of margins with $p < 2$ over margins with $p \geq 2$. Our experiments confirm that our theoretical results are relevant in practice.

1 INTRODUCTION

The notion of "margin" arises naturally in many areas of machine learning. Margins have long been used to motivate the design of algorithms, to give sufficient conditions for fast learning, and to explain unexpected performance of algorithms in practice. Here we are concerned with learning the class of homogeneous linear separators in $\mathbb{R}^d$ over distributions with large margins. We use a general notion of margin, the L_qL_p margin, that captures, among others, the notions used in the analyses of Perceptron ($p = q = 2$) and Winnow ($p = \infty$, $q = 1$). For $p, q \in [1, \infty]$ with $1/p + 1/q = 1$, the L_qL_p margin of a linear classifier $x \mapsto \text{sign}(w \cdot x)$ with respect to a distribution D is defined as

$$\gamma_{q,p}(D, w) = \inf_{x \sim D} \frac{|w \cdot x|}{\|w\|_q \|x\|_p}.$$

While previous work has addressed the case of $p \geq 2$ both theoretically (Grove et al., 2001; Servedio, 2000; Gentile, 2003) and experimentally (Zhang, 2002), the $p < 2$ case has been mentioned but much less explored. This gap in the literature is possibly due to the fact that when $p < 2$ a large margin alone does not guarantee small sample complexity (see Example 1 below for such a case). This leads to the question of whether large L_qL_p margins with $p < 2$ can lead to small sample complexity, and if so, under what conditions will this happen? Furthermore, are there situations in which taking advantage of margins with $p < 2$ can lead to better performance than using margins with $p \geq 2$?

In this work, we answer these three questions using both theoretical and empirical evidence. We first give a bound on the generalization error of linear separators with large L_qL_p margins that holds for any finite $p \geq 1$. The result is proved through a new bound on the fat-shattering dimension of linear separators with bounded L_q norm. The bound improves upon previous results by removing a factor of $\log d$ when $2 \leq p < \infty$ and extends the previously known bounds to the case of $1 \leq p < 2$. A highlight of this theoretical result is that it gives a simple sufficient condition for fast learning even for the $p < 2$ case. The condition is related to the $L_{2,p}$ norm of the data matrix and can be estimated from the data.

We then give a concrete family of learning problems in which using the $L_\infty L_1$ margin gives significantly better sample complexity guarantees than for L_qL_p margins with $p > 1$. We define a family of distributions over labeled examples and consider the sample complexity of learning the class W_p of linear separators with large L_qL_p margins. By bounding covering numbers, we upper bound the sample complexity of learning W_1 and lower bound the complexity of learning W_p when $p > 1$, and we show that the upper bound can be significantly smaller than the lower bound.

In addition, we give experimental results supporting our claim that taking advantage of large $L_\infty L_1$ margins can lead to faster learning. We observe that in the realizable case, the problem of finding a consistent linear separator that maximizes the L_qL_p margin

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

is a convex program (similar to SVM). An extension of this method to the non-realizable case is equivalent to minimizing the L_q-norm regularized hinge loss. We apply these margin-maximization algorithms to both synthetic and real-world data sets and find that maximizing the $L_\infty L_1$ margin can result in better classifiers than maximizing other margins. We also show that the theoretical condition for fast learning that appears in our generalization bound is favorably satisfied on many real-world data sets.

Related Work

It has long been known that the classic algorithms Perceptron (Rosenblatt, 1958) and Winnow (Littlestone, 1988) have mistake bounds of $1/\gamma_{2,2}^2$ and $\tilde{O}(1/\gamma_{1,\infty}^2)$, respectively. A family of "quasi-additive" algorithms (Grove et al., 2001) interpolates between the behavior of Perceptron and Winnow by defining a Perceptron-like algorithm for any $p > 1$. While this gives an algorithm for any $1 < p \leq \infty$ the mistake bound of $\tilde{O}(1/\gamma_{q,p}^2)$ only applies for $p \geq 2$. For small values of p, these algorithms can be used to learn nonlinear separators by using factorizable kernels (Gentile, 2013). A related family (Servedio, 2000) was designed for learning in the PAC model rather than the mistake bound model, but again, guarantees were only given for $p \geq 2$.

Other works (Kakade et al., 2009; Cortes et al., 2010; Kloft and Blanchard, 2012; Maurer and Pontil, 2012) bound the Rademacher complexity of classes of linear separators under general forms of regularization. Special cases of each of these regularization methods correspond to L_q-norm regularization, which is closely related to maximizing L_qL_p margin. Kakade et al. (2009) directly consider the case of L_q-norm regularization but only give Rademacher complexity bounds for the case of $p \geq 2$. Both Cortes et al. (2010) and Kloft and Blanchard (2012) give Rademacher complexity bounds that cover the entire range $1 \leq p \leq \infty$ in the context of multiple kernel learning, but their discussion of excess risk bounds for different choices of p is limited to the $p \geq 2$ case while our work discusses the generalization error over the entire range. Maurer and Pontil (2012) consider the more general setting of block-norm regularized linear classes but only give bounds for the case of $p \geq 2$. In contrast to our work, none of the above works give lower bounds on the sample complexity or give concrete evidence of when some values of p will result in faster learning than others.

There are a few other cases in the literature where the $p \leq 2$ regime is discussed. Zhang (2002) deals with algorithms for L_q-norm regularized loss minimization and discusses cases in which L_∞-norm regularization may be appropriate. Balcan et al. (2013) study the problem of learning two-sided disjunctions in the semi-supervised setting and note that the regularity assumption inherent in the problem can be interpreted as a large $L_\infty L_1$ margin assumption. A related problem posed by Blum and Balcan (2007), the "two-sided majority with margins" problem, has a slightly modified form which constitutes another natural occurrence of a large $L_\infty L_1$ margin (see supplementary material).

2 PRELIMINARIES

Let D be a distribution over a bounded instance space $X \subseteq \mathbb{R}^d$. A linear separator over X is a classifier $h(x) = \text{sign}(w \cdot x)$ for some weight vector $w \in \mathbb{R}^d$. We use h^* and w^* to denote the target function and weight vector, respectively, so that $h^*(x) = \text{sign}(w^* \cdot x)$ gives the label for any instance x and $\mathsf{err}_D(h) = \Pr_{x\sim D}[h(x) \neq h^*(x)]$ is the generalization error of any hypothesis h. We will often abuse notation and refer to a classifier and its corresponding weight vector interchangeably. We will overload the notation $\mathbf{X}$ to represent either a set of n points in $\mathbb{R}^d$ or the $d \times n$ matrix of containing one point per column.

For any point $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and $p \geq 1$, the L_p-norm of x is

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p\right)^{\frac{1}{p}}$$

and the L_∞-norm is $\|x\|_\infty = \max_i |x_i|$. Let $\|X\|_p$ denote $\sup_{x\in X} \|x\|_p$, which is finite for any p by our assumption that X is bounded. The L_q-norm is the dual of the L_p-norm if $1/p + 1/q = 1$ (so the L_∞-norm and L_1-norm are duals). In this work, p and q will always denote dual norms.

For any weight vector w, the L_qL_p margin of w with respect to D is defined as

$$\gamma_{q,p}(D, w) = \inf_{x\sim D} \frac{|w \cdot x|}{\|w\|_q \|x\|_p}.$$

We can similarly define $\gamma_{q,p}(\mathbf{X}, w)$ for a set $\mathbf{X}$. We will drop the first argument when referring to the distribution-based definition and the distribution is clear from context. We assume that D has a positive margin; that is, there exists w such that $\gamma_{q,p}(D, w) > 0$. Note that by Hölder's inequality, $|w \cdot x| \leq \|w\|_q \|x\|_p$ for dual p and q, so $\gamma_{q,p}(D, w) \leq 1$.

We also define the $L_{a,b}$ matrix norm

$$\|\mathbf{M}\|_{a,b} = \left(\sum_{i=1}^{r} \left(\sum_{j=1}^{c} |m_{ij}|^a\right)^{b/a}\right)^{1/b}$$

for any $r \times c$ matrix $\mathbf{M} = (m_{ij})$. In other words, we take the L_a-norm of each row in the matrix and then take the L_b-norm of the resulting vector of L_a-norms.

2.1 L_qL_p Support Vector Machines

Given a linearly separable set $\mathbf{X}$ of n labeled examples, we can solve the convex program

$$\begin{aligned} \min_{w} \quad & \|w\|_q \\ \text{s.t.} \quad & \frac{y^i(w \cdot x^i)}{\|x^i\|_p} \geq 1, \quad 1 \leq i \leq n. \end{aligned} \tag{1}$$

to maximize the L_qL_p margin. Observe that a solution $\hat{w}$ to this problem has $\gamma_{q,p}(\mathbf{X}, \hat{w}) = 1/\|\hat{w}\|_q$. We call an algorithm that outputs a solution to (1) an L_qL_p SVM due to the close relationship between this problem and the standard support vector machine.

If $\mathbf{X}$ is not linearly separable, we can introduce non-negative slack variables in the usual way and solve

$$\begin{aligned} \min_{\substack{w, \\ \xi \geq 0}} \quad & \|w\|_q + C\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & \frac{y^i(w \cdot x^i)}{\|x^i\|_p} \geq 1 - \xi_i, \quad 1 \leq i \leq n. \end{aligned} \tag{2}$$

which is equivalent to minimizing the hinge loss with respect to an L_p-normalized data set using L_q-norm regularization on the weight vector space.

3 GENERALIZATION BOUND

In this section we give an upper bound on the generalization error of learning linear separators over distributions with large L_qL_p margins. The proof follows from combining a theorem of Bartlett and Shawe-Taylor (1999) with a new bound on the fat-shattering dimension of the class of linear separators with small L_q-norm. We begin with the following definitions.

Definition. For a set $\mathcal{F}$ of real-valued functions on X, a finite set $\{x^1, \ldots, x^n\} \subseteq X$ is said to be *γ-shattered by $\mathcal{F}$* if there are real numbers $r_1, \ldots, r_n$ such that for all $b = (b_1, \ldots, b_n) \in \{-1, 1\}^n$, there is a function $f_b \in \mathcal{F}$ such that

$$f_b(x^i) \begin{cases} \geq r_i + \gamma & \text{if } b_i = 1 \\ \leq r_i - \gamma & \text{if } b_i = -1. \end{cases}$$

The *fat-shattering dimension of $\mathcal{F}$ at scale γ*, denoted $\text{fat}_{\mathcal{F}}(\gamma)$, is the size of the largest subset of X which is γ-shattered by $\mathcal{F}$.

Our bound on the fat-shattering dimension will use two lemmas analogous to Lemmas 11 and 12 in (Servedio, 2000).

Lemma 1. *Let $\mathcal{F} = \{x \mapsto w \cdot x : \|w\|_q \leq \|W\|_q\}$ with $1 \leq p \leq \infty$. If the set $\{x^1, \ldots, x^n\} \subseteq X^n$ is γ-shattered by $\mathcal{F}$ then every $b = (b_1, \ldots, b_n) \in \{-1, 1\}^n$ satisfies $\left\|\sum_{i=1}^n b_i x^i\right\|_p \geq \frac{\gamma n}{\|W\|_q}$.*

Proof. The proof is identical to that of Lemma 11 in (Servedio, 2000), replacing the radius $1/\|X\|_p$ of $\mathcal{F}$ in their lemma with $\|W\|_q$. □

The next lemma will depend on the following classical result from probability theory known as the Khintchine inequality.

Theorem 1 (Khintchine)**.** *If the random variable $\sigma = (\sigma_1, \ldots, \sigma_n)$ is uniform over $\{-1, 1\}^n$ and $0 < p < \infty$, then any finite set $\{z_1, \ldots, z_n\} \in \mathbb{C}$ satisfies*

$$A_p \sqrt{\sum_{i=1}^{n} |z_i|^2} \leq \left(\mathbb{E}\left[\left|\sum_{i=1}^{n} \sigma_i z_i\right|^p\right]\right)^{\frac{1}{p}} \leq B_p \sqrt{\sum_{i=1}^{n} |z_i|^2}$$

where A_p and B_p are constants depending only on p.

The precise optimal constants for A_p and B_p were found by Haagerup (1982), but for our purposes, it suffices to note that when $p \geq 1$ we have $1/2 \leq A_p \leq 1$ and $1 \leq B_p \leq \sqrt{p}$.

Lemma 2. *For any set $\mathbf{X} = \{x^1, \ldots, x^n\} \subseteq X^n$ and any $1 \leq p < \infty$, there is some $b = (b_1, \ldots, b_n) \in \{-1, 1\}^n$ such that $\left\|\sum_{i=1}^n b_i x^i\right\|_p \leq B_p \|\mathbf{X}\|_{2,p}$.*

Proof. We will bound the expectation of $\left\|\sum_{i=1}^n b_i x^i\right\|_p$ when $b = (b_1, \ldots, b_n)$ is uniformly distributed over $\{-1, 1\}^n$. We have

$$\begin{aligned} \mathbb{E}\left[\left\|\sum_{i=1}^{n} b_i x^i\right\|_p\right] &= \mathbb{E}\left[\left(\sum_{j=1}^{d}\left|\sum_{i=1}^{n} \epsilon_i x_j^i\right|^p\right)^{1/p}\right] \\ &\leq \left(\sum_{j=1}^{d} \mathbb{E}\left[\left|\sum_{i=1}^{n} \epsilon_i x_j^i\right|^p\right]\right)^{1/p} \\ &\leq \left(\sum_{j=1}^{d} B_p^p \left(\sum_{i=1}^{n} |x_j^i|^2\right)^{p/2}\right)^{1/p} \\ &= B_p \|\mathbf{X}\|_{2,p} \end{aligned}$$

where the first inequality is an application of Jensen's inequality and the second uses the Khintchine inequality. The proof is completed by noting that there must be some choice of b for which the value of $\left\|\sum_{i=1}^n b_i x^i\right\|_p$ is smaller than its expectation. □

We can use these two lemmas to give an upper bound on the fat-shattering dimension for any finite p.

Theorem 2. *Let $\mathcal{F} = \{x \mapsto w \cdot x : \|w\|_q \leq \|W\|_q\}$ with $1 \leq p < \infty$. If there is a constant $C = C(d,p)$ independent of n such that $\|\mathbf{X}\|_{2,p} \leq Cn^\alpha \|X\|_p$ for any set $\mathbf{X}$ of n examples drawn from D, then*

$$\text{fat}_\mathcal{F}(\gamma) \leq \left(\frac{CB_p \|W\|_q \|X\|_p}{\gamma}\right)^{\frac{1}{1-\alpha}}.$$

Proof. Combining Lemmas 1 and 2, we have that any set $\mathbf{X} = \{x^1, \ldots, x^n\} \subseteq X^n$ that is γ-shattered by $\mathcal{F}$ satisfies $\frac{\gamma n}{\|W\|_q} \leq B_p \|\mathbf{X}\|_{2,p} \leq CB_p n^\alpha \|X\|_p$. Solving for n gives us $n \leq (\frac{CB_p\|W\|_q\|X\|_p}{\gamma})^{1/(1-\alpha)}$ as an upper bound on the maximum size of any γ-shattered set. □

This bound extends and improves upon Theorem 8 in (Servedio, 2000). In their specific setting, $\|W\|_q = 1/\|X\|_p$, so we can directly compare their bound

$$\text{fat}_\mathcal{F}(\gamma) \leq \frac{2\log 4d}{\gamma^2} \tag{3}$$

for $2 \leq p \leq \infty$ to our bound

$$\text{fat}_\mathcal{F}(\gamma) \leq \left(\frac{CB_p}{\gamma}\right)^{1/(1-\alpha)} \tag{4}$$

for $1 \leq p < \infty$. Observe that by Minkowski's inequality, any set $\mathbf{X}$ satisfies $\|\mathbf{X}\|_{2,p} \leq n^{1/2}\|X\|_p$ if $p \geq 2$. In this case, (4) simplifies to $(B_p/\gamma)^2$ which is dimension-independent and improves upon (3) by a factor of $\log d$ when p is constant. When $1 \leq p < 2$, (3) does not apply, but (4) still gives a bound that can be small in many cases depending on the relationship between $\|\mathbf{X}\|_{2,p}$ and γ. We will give specific examples in Section 3.1.

The fat-shattering dimension is relevant due to the following theorem of Bartlett and Shawe-Taylor (1999) that relates the generalization performance of a classifier with large margin to the fat-shattering dimension of the associated real-valued function class at a scale of roughly the margin of the classifier.

Theorem 3 (Bartlett & Shawe-Taylor)**.** *Let $\mathcal{F}$ be a collection of real-valued functions on a set X and let D be a distribution over X. Let $\mathbf{X} = \{x^1, \ldots, x^n\}$ be a set of examples drawn i.i.d. from D with labels $y_i = h^*(x^i)$ for each i. With probability at least $1-\delta$, if a classifier $h(x) = \text{sign}(f(x))$ with $f \in \mathcal{F}$ satisfies $y_i f(x^i) \geq \gamma > 0$ for each $x^i \in \mathbf{X}$, then*

$$\text{err}_D(h) \leq \frac{2}{n}\left(k\log\frac{8en}{k}\log(32n) + \log\frac{8n}{\delta}\right),$$

where $k = \text{fat}_\mathcal{F}(\gamma/16)$.

Now we can state and prove the following theorem which bounds the generalization performance of the L_qL_p SVM algorithm.

Theorem 4. *For any distribution D and target w^* with $\gamma_{q,p}(D, w^*) \geq \gamma_{q,p}$, if there is a constant $C = C(d,p)$ such that $\|\mathbf{X}\|_{2,p} \leq Cn^\alpha \|X\|_p$ for any set $\mathbf{X}$ of n examples from D then there is a polynomial time algorithm that outputs, with probability at least $1-\delta$, a classifier h such that*

$$\text{err}_D(h) = O\left(\frac{1}{n}\left(\left(\frac{CB_p}{\gamma_{q,p}}\right)^{\frac{1}{1-\alpha}}\log^2 n + \log\frac{n}{\delta}\right)\right).$$

Proof. By the definition of L_qL_p margin, there exists a w (namely, $w^*/\|w^*\|_q$) with $\|w\|_q = 1$ that achieves margin $\gamma_{q,p}$ with respect to D. This w has margin at least $\gamma_{q,p}$ with respect to any set $\mathbf{X}$ of n examples from D. A vector $\hat{w}$ satisfying these properties can be found in polynomial time by solving the convex program (1) and normalizing the solution. Notice that if the sample is normalized to have $\|x\|_p = 1$ for every $x \in \mathbf{X}$ then the L_qL_p margin of $\hat{w}$ does not change but becomes equal to the functional margin $y(\hat{w}\cdot x)$ appearing in the Theorem 3. Applying Theorem 2 with $\|W\|_q = 1$ and $\|X\|_p = 1$ yields $\text{fat}_\mathcal{F}(\gamma) \leq (\frac{CB_p}{\gamma})^{1/(1-\alpha)}$ and applying Theorem 3 to $\hat{w}$ gives us the desired result. □

Theorem 4 tells us that if the quantity

$$\left(\frac{CB_p}{\gamma_{q,p}}\right)^{\frac{1}{1-\alpha}} \tag{5}$$

is small for a certain choice of p and q then the L_qL_p SVM will have good generalization. This gives us a data-dependent bound, as (5) depends on data; specifically, C and α depend on the distribution D alone, while $\gamma_{q,p}$ depends on the relationship between D and the target w^*.

As mentioned, if $p \geq 2$ then we can use $C = 1$ and $\alpha = 1/2$ for any distribution, in which case the bound depends solely on the margin $\gamma_{q,p}$ (and to a lesser extent on B_p). If $p \leq 2$ then any set has $\|\mathbf{X}\|_{2,p} \leq n^{1/p}\|X\|_p$ (this follows by subadditivity of the function $z \mapsto z^{p/2}$ when $p \leq 2$) and we can obtain a similar dimension-independent bound with $C = 1$ and $\alpha = 1/p$. Achieving dimension independence for all distributions comes at the price of the bound becoming uninformative as $p \to 1$, as (5) simplifies to $(B_p/\gamma_{q,p})^q$ for these values. More interesting situations arise when we consider the quantity (5) for specific distributions, as we will show in the next section.

3.1 Examples

Here we will give some specific learning problems showing when large margins can be helpful and when they are not helpful. We focus on the $p \leq 2$ case, as large margins are always helpful when $p \geq 2$.

Example 1. Unhelpful margins. First, let D_1 be the uniform distribution over the standard basis vectors in $\mathbb{R}^d$ and let w^* be any weight vector in $\{-1,1\}^d$. In this case $\gamma_{q,p} = d^{-1/q}$, which is a large margin for small p. If $n \leq d$, then $\|\mathbf{X}\|_{2,p}$ is roughly $n^{1/p}\|X\|_p$ (ignoring log factors), and (5) simplifies to $B_p^q d$. We could also choose to simplify (5) using $C = d^{1/p}$ and $\alpha = 0$, which gives us $B_p d$. Either way, the bound in Theorem 4 becomes $\tilde{O}(d/n)$ which is uninformative since $n \leq d$. If we take $n \geq d$, we can still obtain a bound of $\tilde{O}(d/n)$, but this is the same as the worst-case bound based on VC dimension, so the large margin has no advantage. In fact, this example provides a lower bound: even if an algorithm knows the distribution D_1 and is allowed a $1/2$ probability of failure, an error of ϵ cannot be guaranteed with fewer than $(1-2\epsilon)d$ examples because any algorithm can hope for at best an error rate of $1/2$ on the examples it has not yet seen.

Example 2. Helpful margins. As another example, divide the d coordinates into $k = o(d)$ disjoint blocks of equal size and let D_2 be the uniform distribution over examples that have 1's for all coordinates within some block and 0's elsewhere. Taking w^* to be a vector in $\{-1,1\}^d$ that has the same sign within each block, we have $\gamma_{q,p} = k^{-1/q}$. If $k < n < d$ then $\|\mathbf{X}\|_{2,p}$ is roughly $k^{1/p-1/2}\sqrt{n}\|X\|_p$, and (5) simplifies to $B_p^2 k$. When $k = o(d)$ this is a significant improvement over worst case bounds for any constant choice of p.

Example 3. An advantage for $p < 2$. Consider a distribution that is a combination of the previous two examples: with probability $1/2$ it returns an example drawn from D_1 and otherwise returns an example from D_2. By including the basis vectors, we have made the margin $\gamma_{q,p} = d^{-1/q}$ but as long as $k = o(d)$ the bound on $\|\mathbf{X}\|_{2,p}$ does not change significantly from Example 2, and we can still use $C = k^{1/p-1/2}$ and $\alpha = 1/2$. Now (5) simplifies to $B_p^2 k$ for $p = 1$, but becomes $B_p^2 k^{2/p-1} d^{2/q}$ in general. When $k = \sqrt{d}$ this gives us an error bound of $\tilde{O}(\sqrt{d}/n)$ for $p = 1$ but $\tilde{O}(d/n)$ or worse for $p \geq 2$. While this upper bound does not imply that generalization error will be worse for $p \geq 2$ than it is for $p = 1$, we show in the next section that for a slightly modified version of this distribution we can obtain sample complexity lower bounds for large margin algorithms with $p \geq 2$ that are significantly greater than the upper bound for $p = 1$.

4 THE CASE FOR $L_\infty L_1$ MARGINS

Here we give a family of learning problems to show the benefits of using $L_\infty L_1$ margins over other margins. We do this by defining a distribution D over unlabeled examples in $\mathbb{R}^d$ that can be consistently labeled by a variety of potential target functions w^*. We then consider a family of large $L_q L_p$ margin concept classes W_p and bound the sample complexity of learning a concept in W_p using covering number bounds. We show that learning W_1 can be much easier than learning W_p for $p > 1$; for example, with certain parameters for D having $O(\sqrt{d})$ examples is sufficient for learning W_1, while learning any other W_p requires $\Omega(d)$ examples.

Specifically, let $W_p = \{w \in \mathbb{R}^d : \|w\|_\infty = 1,\ \gamma_{q,p}(w) \geq \gamma_{q,p}(w^*)\}$, where w^* maximizes the $L_\infty L_1$ margin with respect to D. We restrict our discussion to weight vectors with unit L_∞ norm because normalization does not change the margin (nor does it affect the output of the corresponding classifier). Let the covering number $\mathcal{N}(\epsilon, W, D)$ be the size of the smallest set $V \subseteq W$ such that for every $w \in W$ there exists a $v \in V$ with $d_D(w,v) := \Pr_{x\sim D}[\text{sign}(w \cdot x) \neq \text{sign}(v \cdot x)] \leq \epsilon$.

Define the distribution D over $\{0,1\}^d$ as follows. Divide the d coordinates into k disjoint blocks of size d/k (assume $d/(2k)$ is an odd integer). Flip a fair coin. If heads, pick a random block and return an example with exactly $d/(2k)$ randomly chosen coordinates set to 1 within the chosen block and all other coordinates set to 0. If tails, return a standard basis vector (exactly one coordinate set to 1) chosen uniformly at random. The target function will be determined by any weight vector w^* achieving the maximum $L_\infty L_1$ margin with respect to D. As we will see, w^* can be any vector in $\{-1,1\}^d$ with complete agreement within each block.

We first give an upper bound on the covering of W_1.

Proposition 1. *For any $\epsilon > 0$, $\mathcal{N}(\epsilon, W_1, D) \leq 2^k$.*

Proof. Let V_k be the following set. Divide the d coordinates into k disjoint blocks of size d/k. A vector $v \in \{-1,1\}^d$ is a member of V_k if and only if each block in v is entirely $+1$'s or entirely -1's. We will show that $W_1 = V_k$, and since $|V_k| = 2^k$ we will have $\mathcal{N}(\epsilon, W_1, D) \leq 2^k$ for any ϵ.

Note that by Hölder's inequality, $\gamma_{q,p}(w) \leq 1$ for any $w \in \mathbb{R}^d$. For any $w \in V_k$ and any example x drawn from D, we have $|w \cdot x| = \|x\|_1$, so $\gamma_{\infty,1}(w) = 1$, the maximum margin. If $w \notin V_k$ then either $w \notin \{-1,1\}^d$ or w has sign disagreement within at least one of the k blocks. If $w \notin \{-1,1\}^d$ then $\gamma_{\infty,1}(w) = \min_x |w \cdot x| / \|x\|_1 \leq \min_i |w \cdot e_i| = \min_i |w_i| < 1$. If w has sign disagreement within a block then $|w \cdot x| < \|x\|_1$ for any x with 1's in disagreeing coordinates of w, and this results in a margin strictly less than 1. □

Now we will give lower bounds on the covering numbers for W_p with $p > 1$. In the following let $H(\alpha) = -\alpha\log(\alpha) - (1-\alpha)\log(1-\alpha)$, the binary entropy function.

Proposition 2. *If $1 < p < \infty$ then for any $\epsilon > 0$, $\mathcal{N}(\epsilon, W_p, D) \geq 2^{(1/2-H(2\epsilon))d-k^{1/q}(d/2)^{1/p}-k}$.*

Proof. First we show that $|W_p| \geq 2^{d/2-k^{1/q}(d/2)^{1/p}-k}$. Any $w^* \in W_1$ has margin $\gamma_{q,p}(w^*) = d^{-1/q}$, so $W_p = \{w \in \mathbb{R}^d : \gamma_{q,p}(w) \geq d^{-1/q}\}$. Note that $W_p \subseteq \{-1,1\}^d$ because if $w \notin \{-1,1\}^d$ then $\gamma_{q,p}(w) \leq \min_i |w \cdot e_i| / \|w\|_q = \min_i |w_i| / \|w\|_q < d^{-1/q}$. Let $w \in \{-1,1\}^d$ be a weight vector such that in each block there are at least $d/k - r$ positive values and at most r negative values or vice versa (there are at most r values with whichever sign is in the minority). Clearly w has large margin with respect to any of the basis vectors drawn from D. For the rest of D we have $\inf_x |w \cdot x| = \max(1, n/(2k) - 2r)$, so $w \in W_p$ if and only if $\max(1, d/(2k) - 2r) \geq (d/(2k))^{1/p}$. For $p < \infty$ and $d > 2k$, this happens if and only if $r \leq \frac{1}{2}(\frac{d}{2k} - (\frac{d}{2k})^{1/p})$. Letting $r^* = \left\lfloor \frac{1}{2}(\frac{d}{2k} - (\frac{d}{2k})^{1/p}) \right\rfloor$, we have $|W_p| = (2\sum_{i=0}^{r^*} \binom{d/k}{i})^k \geq 2^{d/2-k^{1/q}(d/2)^{1/p}-k}$.

Now we can lower bound the covering number using a volume argument by noting that if m is the cardinality of the largest ϵ-ball around any $w \in W_p$ then $\mathcal{N}(\epsilon, W_p, D) \geq |W_p|/m$. For any pair $w, w' \in W_p$, $d_D(w, w') \geq h(w, w')/(2d)$ where $h(w, w')$ is the hamming distance (number of coordinates in which w and w' disagree). Therefore, in order for $d_D(w, w') \leq \epsilon$ we need $h(w, w') \leq 2\epsilon d$. For any $w \in W_p$ the number of w' such that $h(w, w') \leq 2\epsilon d$ is at most $\sum_{i=0}^{\lfloor 2\epsilon d \rfloor} \binom{d}{i} \leq 2^{H(2\epsilon)d}$. It follows that $\mathcal{N}(\epsilon, W_p, D) \geq |W_p|/2^{H(2\epsilon)d} \geq 2^{d/2-H(2\epsilon)d-k^{1/q}(d/2)^{1/p}-k}$. □

Proposition 3. *For any $\epsilon > 0$, $\mathcal{N}(\epsilon, W_\infty, D) \geq 2^{(1-H(2\epsilon))d}$.*

Proof. First we show that $W_\infty = \{-1,1\}^d$. Any $w^* \in W_1$ has margin $\gamma_{1,\infty}(w^*) = 1/d$, so $W_\infty = \{w \in \mathbb{R}^d : \gamma_{1,\infty}(w) \geq 1/d\}$. For any $w \in \{-1,1\}^d$ and example x drawn from D, we have $\|w\|_1 = d$, $\|x\|_\infty = 1$, and $|w \cdot x| \geq 1$ (since x has an odd number of coordinates set to 1) resulting in margin $\gamma_{1,\infty}(w) \geq 1/d$. If $w \notin \{-1,1\}^d$ then $\gamma_{1,\infty}(w) = \min_x |w \cdot x| / \|w\|_1 \leq \min_i |w \cdot e_i| / \|w\|_1 = \min_i |w_i| / \|w\|_1 < 1/d$.

To bound the covering number, we use the same volume argument as above. Again, the size of the largest ϵ-ball around any $w \in W_\infty$ is at most $2^{H(2\epsilon)d}$ (since this bound only requires that every pair $w, w' \in W_\infty$ has $d_D(w, w') \geq h(w, w')/(2d)$). It follows that $\mathcal{N}(\epsilon, W_\infty, D) \geq 2^d/2^{H(2\epsilon)d}$. □

Using standard distribution-specific sample complexity bounds based on covering numbers (Itai and Benedek, 1991), we have an upper bound of $O((1/\epsilon)\ln(\mathcal{N}(\epsilon, W, D)/\delta))$ and lower bound of $\ln((1-\delta)\mathcal{N}(2\epsilon, W, D))$ for learning, with probability at least $1-\delta$, a concept in W to within ϵ error. Thus, we have the following results for the sample complexity m of learning W_p with respect to D. If $p = 1$ then

$$m \leq O\left(\frac{1}{\epsilon}\left(k + \ln\frac{1}{\delta}\right)\right),$$

if $1 < p < \infty$ then

$$m \geq \left(\frac{1}{2} - H(4\epsilon)\right) d - k^{1/q}\left(\frac{d}{2}\right)^{1/p} - k + \ln(1-\delta),$$

and if $p = \infty$ then

$$m \geq (1 - H(4\epsilon))\, d + \ln(1-\delta).$$

For appropriate values of k and ϵ relative to d, the the sample complexity can be much smaller for the $p = 1$ case. For example, if $k = O(d^{1/4})$ and $\Omega(d^{-1/4}) \leq \epsilon \leq 1/40$, then (assuming δ is a small constant) having $O(\sqrt{d})$ examples is sufficient for learning W_1 while at least $\Omega(d)$ examples are required to learn W_p for any $p > 1$.

5 EXPERIMENTS

We performed two empirical studies to support our theoretical results. First, to give further evidence that using $L_\infty L_1$ margins can lead to faster learning than other margins, we ran experiments on both synthetic and real-world data sets. Using the L_qL_p SVM formulation defined in (1) for linearly separable data and the formulation defined in (2) for non-separable data, both implemented using standard convex optimization software, we ran our algorithms for a range of values of p and a range of training set sizes n on each data set. We report several cases in which maximizing the $L_\infty L_1$ margin results in faster learning (i.e., smaller sample complexity) than maximizing other margins.

Figure 1 shows results on two synthetic data sets. One is generated using the "Blocks" distribution family from Section 4 with $d = 90$ and $k = 9$. The other uses examples generated from a standard Gaussian distribution in $\mathbb{R}^{100}$ subject to having $L_\infty L_1$ margin at least 0.075 with respect to a fixed random target vector in $\{-1,1\}^d$ (in other words, Gaussian samples with margin smaller than 0.075 are rejected). In both cases, the error decreases much faster for $p < 2$ than for large p.

Figure 2 shows results on three data sets from the UCI Machine Learning Repository (Bache and Lichman, 2013). The Fertility data set consists of 100 training examples in $\mathbb{R}^{10}$, the SPECTF Heart data set has 267 examples in $\mathbb{R}^{44}$, and we used a subset of the CNAE-9 data set with 240 examples in $\mathbb{R}^{857}$. In all three cases, better performance was achieved by algorithms with $p < 2$ than by those with $p > 2$.

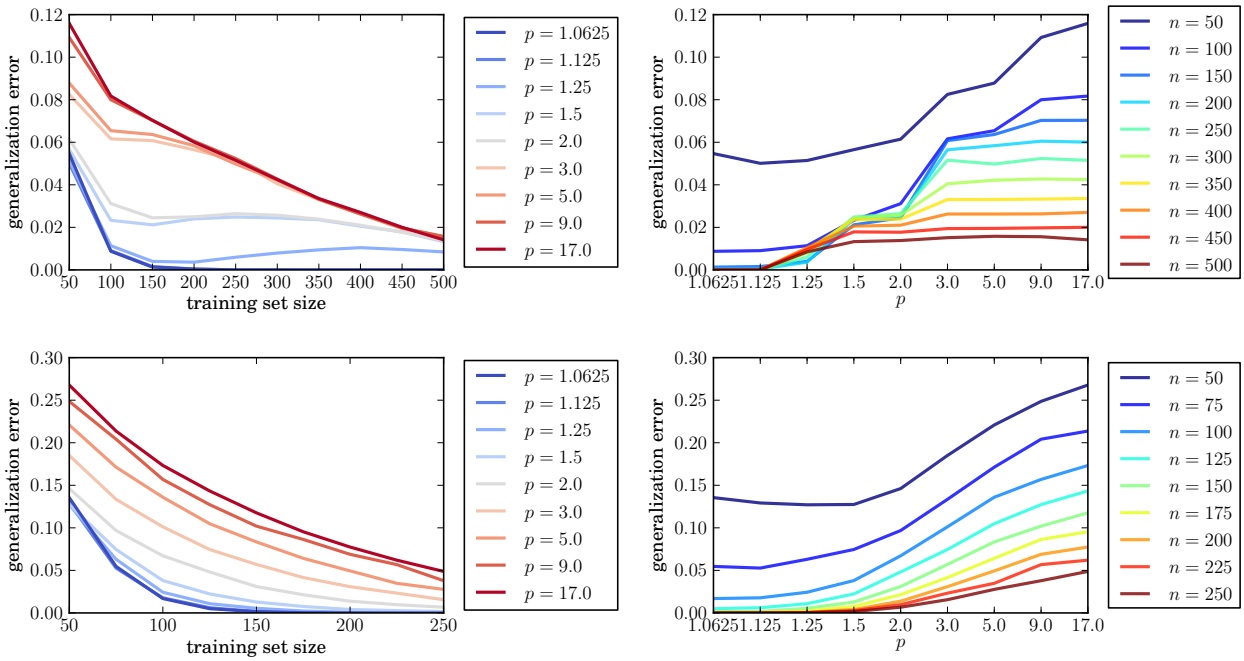

Figure 1: Synthetic data results for Blocks distribution (top) and Gaussian with margin (bottom). The left column plots generalization error (averaged over 500 trials with different training sets) versus number of training examples n while the right column plots error versus p.

The goal of our second experiment was to determine, for real-world data, what parameter α can be used in the bound on $\|\mathbf{X}\|_{2,p}$ in Theorem 4. Specifically, for each data set we want to find $\alpha_{\min} = \inf\{\alpha : \|\mathbf{X}\|_{2,p} \leq n^{\alpha} \|X\|_p\}$, the smallest value of α so the bound holds with $C = 1$. Recall that for $p = 1$, $\alpha_{\min}$ can theoretically be as great as 1, while for $p \geq 2$ it is at most $1/2$. We would like to see whether $\alpha_{\min}$ is often small in real data sets or whether it is close to the theoretical upper bound.

We can estimate $\alpha_{\min}$ for a given set of data by creating a sequence $\{\mathbf{X}_m\}_{m=1}^n$ of data matrices by adding to the matrix one point from the data set at a time. For each point in the sequence we can compute

$$\alpha_m = \frac{\log(\|\mathbf{X}_m\|_{2,p} / \|X\|_p)}{\log m},$$

a value of α that realizes the bound with equality for this particular data matrix. We repeat this process T times, each with a different random ordering of the data, to find T sequences α_m^i, where $1 \leq i \leq T$ and $1 \leq m \leq n$. We can then compute $\hat{\alpha}_{\min} = \max_{i,m} \alpha_m^i$, a value of α which realizes the bound for every data matrix considered and which causes the bound to hold with equality in at least one instance.

Figure 3 shows a histogram of the resulting estimates on a variety of data sets and for three values of p.

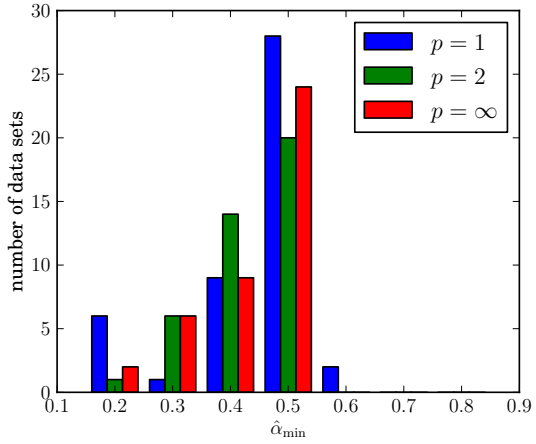

Figure 3: A histogram showing the values of $\hat{\alpha}_{\min}$ on 47 data sets from the UCI repository.

Notice that in the vast majority of cases, the estimate of $\alpha_{\min}$ is less than $1/2$. As expected there are more values above $1/2$ for $p = 1$ than for $p \geq 2$, but none of the estimates were above 0.7. This gives us evidence that many real data sets are much more favorable for learning with large $L_\infty L_1$ margins than the worst-case bounds may suggest.

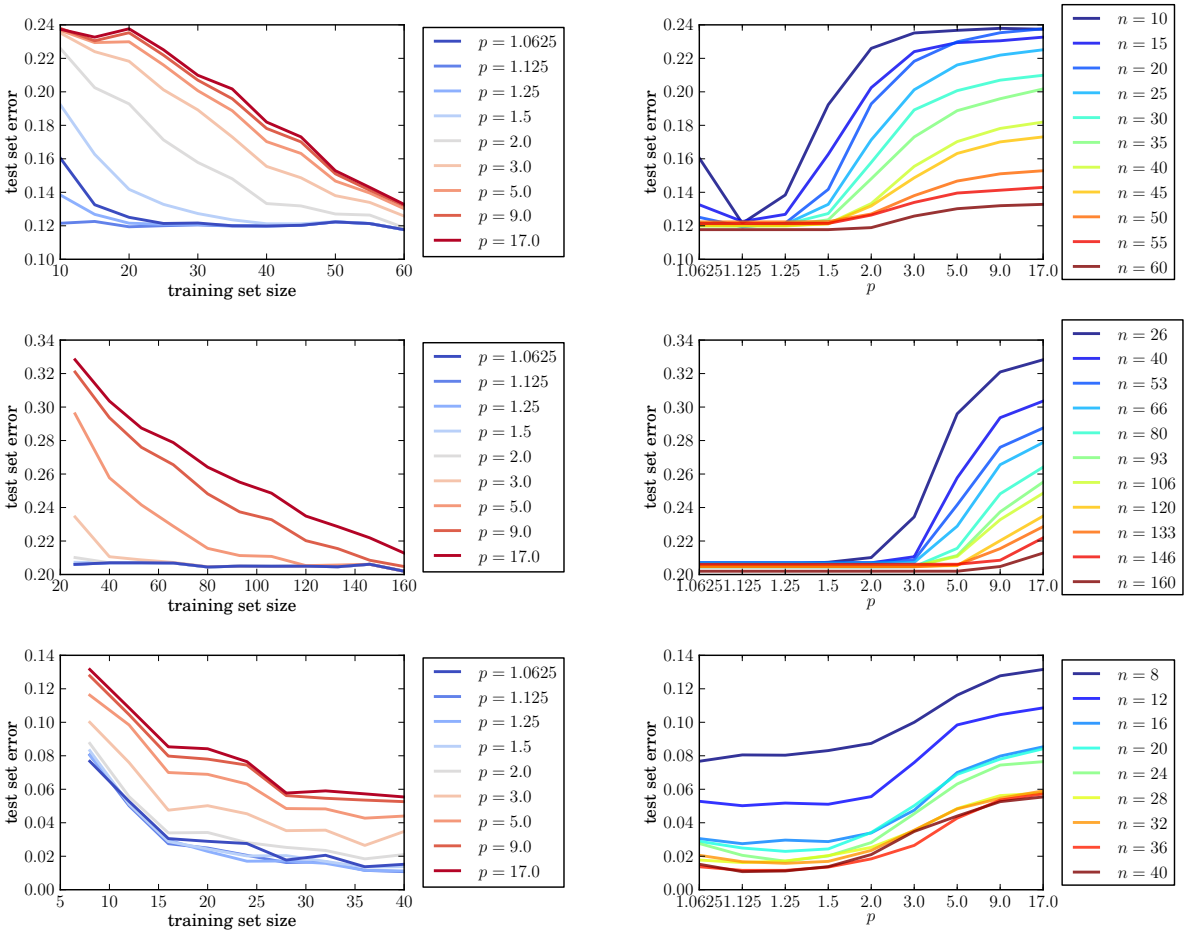

Figure 2: Results for Fertility data (top), SPECTF Heart data (middle), and CNAE-9 (bottom). The left column plots error (averaged over 100 trials with different training sets and tested on all non-training examples) versus number of training examples n while the right column plots error versus p.

6 DISCUSSION

Theorem 4 applies to the realizable case in which the two classes are linearly separable by a positive L_qL_p margin. The result can be extended to the non-realizable case by bounding the empirical Rademacher complexity in terms of the $L_{2,p}$-norm of the data using the Khintchine inequality. This bound can be seen as a special case of Proposition 2 of Kloft and Blanchard (2012) as our setting is a special case of multiple kernel learning (see supplementary material for details). Kloft and Blanchard (2012) also prove bounds on the population and local Rademacher complexities, although in doing so they introduce an explicit dependence on dimension (number of kernels).

This highlights a difference in goals between our work and much of the MKL literature. The MKL literature provides bounds that apply to all data distributions while focusing on the $p \geq 2$ regime, as this is the most relevant range for MKL. On the other hand, we are interested in understanding what kinds of data lead to fast (and dimension-independent) learning for different notions of margin, and furthermore when one type of margin can be provably better than another. We make steps toward understanding this through our condition on the $L_{2,p}$-norm and concrete lower bounds on the generalization error, neither of which has been explored in the context of MKL. Carrying over these perspectives into more general settings such as MKL is an exciting direction for further research.

Acknowledgments

We thank Ying Xiao for insightful discussions and the anonymous reviewers for their helpful comments. This work was supported in part by NSF grants CCF-0953192 and CCF-1101215, AFOSR grant FA9550-09-1-0538, ONR grant N00014-09-1-0751, and a Microsoft Research Faculty Fellowship.

References

K. Bache and M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

M.-F. Balcan, C. Berlind, S. Ehrlich, and Y. Liang. Efficient Semi-supervised and Active Learning of Disjunctions. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

P. L. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods*, pages 43–54. MIT Press, 1999.

A. Blum and M.-F. Balcan. Open Problems in Efficient Semi-Supervised PAC Learning. In *Proceedings of the Conference on Learning Theory (COLT)*, 2007.

C. Cortes, M. Mohri, and A. Rostamizadeh. Generalization Bounds for Learning Kernels. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

C. Gentile. The Robustness of the p-Norm Algorithms. *Machine Learning*, 53:265–299, 2003.

C. Gentile. Personal communication, 2013.

A. J. Grove, N. Littlestone, and D. Schuurmans. General Convergence Results for Linear Discriminant Updates. *Machine Learning*, 43:173–210, 2001.

U. Haagerup. The best constants in the Khintchine inequality. *Studia Mathematica*, 1982.

A. Itai and G. M. Benedek. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, 1991.

S. M. Kakade, K. Sridharan, and A. Tewari. On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2009.

M. Kloft and G. Blanchard. On the Convergence Rate of ℓ_p-Norm Multiple Kernel Learning. *Journal of Machine Learning Research*, 13:2465–2501, 2012.

N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988.

A. Maurer and M. Pontil. Structured Sparsity and Generalization. *Journal of Machine Learning Research*, 13:671–690, 2012.

F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386–408, Nov. 1958.

R. A. Servedio. PAC Analogues of Perceptron and Winnow via Boosting the Margin. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 130–151, 2000.

T. Zhang. On the dual formulation of regularized linear systems with convex risks. *Machine Learning*, 46:91–129, 2002.