On Learning Linear Separators with Large $L_\infty L_1$ Margins

Maria-Florina Balcan and Christopher Berlind

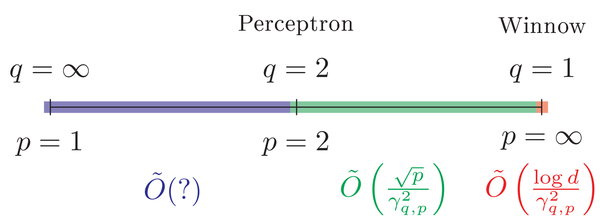
Georgia Tech College of Computing

Margins

The L_qL_p margin of $x \mapsto \text{sign}(w \cdot x)$ w.r.t. D is

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

where $1 \leq p, q \leq \infty$ and $1/p + 1/q = 1$.

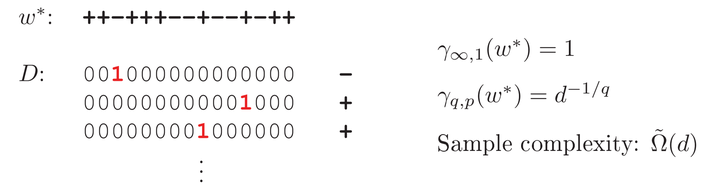


[Grove et al., 2001; Servedio, 2002]

Problem

For $p < 2$, large margin **not** sufficient for fast rates.

Example: Basis vectors with $w^* \in \{-1, 1\}^d$.



We address the following questions:
- Are $L_\infty L_1$ margins ever helpful?
- If so, can we quantify when they help?

Experiments

L_qL_p SVM: Given a set of n training examples, we can efficiently solve the convex program

$$\min_w \quad \|w\|_q + C \sum_{i=1}^n \xi_i$$
$$\text{s.t.} \quad \frac{y^i (w \cdot x^i)}{\|x^i\|_p} \geq 1 - \xi_i, \quad 1 \leq i \leq n.$$

Equivalent to minimizing the hinge loss of an L_p-normalized data set using L_q-norm regularization.

Empirical advantage for $p = 1$ on synthetic data and on several real data sets from the UCI repository.

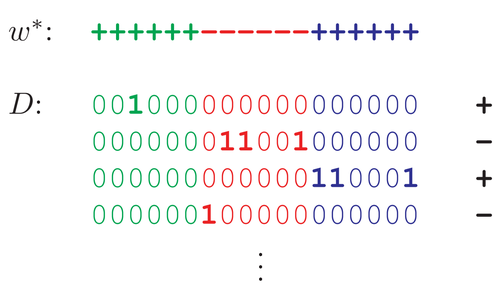
Making the Case for $L_\infty L_1$ Margins

A Family of Learning Problems

Divide d boolean coordinates evenly into k blocks.

Distribution D randomly picks a block and either
- sets to 1 a single variable in the block or
- sets to 1 exactly $d/(2k)$ variables in the block.

Target w^* is random and maximizes $L_\infty L_1$ margin.



Notation

- Set of large L_qL_p margin separators:
  $W_p = \{w \in \mathbb{R}^d : \|w\|_\infty = 1,\ \gamma_{q,p}(w) \geq \gamma_{q,p}(w^*)\}$
- Covering number of W: $\mathcal{N}(\epsilon, W, D)$
- Binary entropy:
  $H(\alpha) = -\alpha \log(\alpha) - (1 - \alpha) \log(1 - \alpha)$

Covering Number Bounds

Proposition 1. $\mathcal{N}(\epsilon, W_1, D) \leq 2^k$.
Proof.
- $\gamma_{\infty,1}(w) = 1$ if and only if each block is unanimous

- $\mathcal{N}(\epsilon, W_1, D) \leq |W_1| = 2^k$

Proposition 2. *If* $1 < p < \infty$ *then*
$\mathcal{N}(\epsilon, W_p, D) \geq 2^{(1/2 - H(2\epsilon))d - k^{1/q}(d/2)^{1/p} - k}$.
Proof.
- $w \in W_p$ iff each block has $\leq \frac{1}{2}(\frac{d}{2k} - (\frac{d}{2k})^{1/p})$ dissenters

- $|W_p| \geq 2^{d/2 - k^{1/q}(d/2)^{1/p} - k}$
- Volume argument: $\mathcal{N}(\epsilon, W_p, D) \geq |W_p| / 2^{H(2\epsilon)d}$

Proposition 3. $\mathcal{N}(\epsilon, W_\infty, D) \geq 2^{(1 - H(2\epsilon))d}$.
Proof.
- $W_\infty = \{-1, 1\}^d$

- Volume argument: $\mathcal{N}(\epsilon, W_\infty, D) \geq |W_\infty| / 2^{H(2\epsilon)d}$

Sample Complexity Bounds

Distribution-specific bounds [Benedek & Itai, 1991]:

$$\text{Upper:} \quad O\left(\frac{1}{\epsilon} \log \mathcal{N}(\epsilon, W, D) + \frac{1}{\epsilon} \log \frac{1}{\delta}\right)$$
$$\text{Lower:} \quad \Omega\Big(\log \mathcal{N}(2\epsilon, W, D) + \log(1 - \delta)\Big)$$

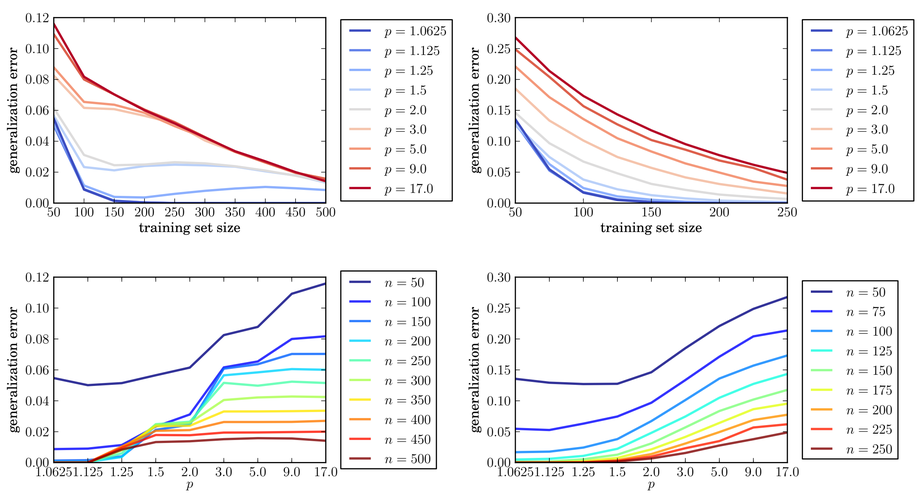
Using our covering number bounds, we obtain

$$p = 1: \quad \tilde{O}\left(\frac{k}{\epsilon}\right)$$
$$p = 2: \quad \tilde{\Omega}\Big((\tfrac{1}{2} - H(4\epsilon))d - \sqrt{kd} - k\Big)$$
$$p = \infty: \quad \tilde{\Omega}\Big((1 - H(4\epsilon))d\Big)$$

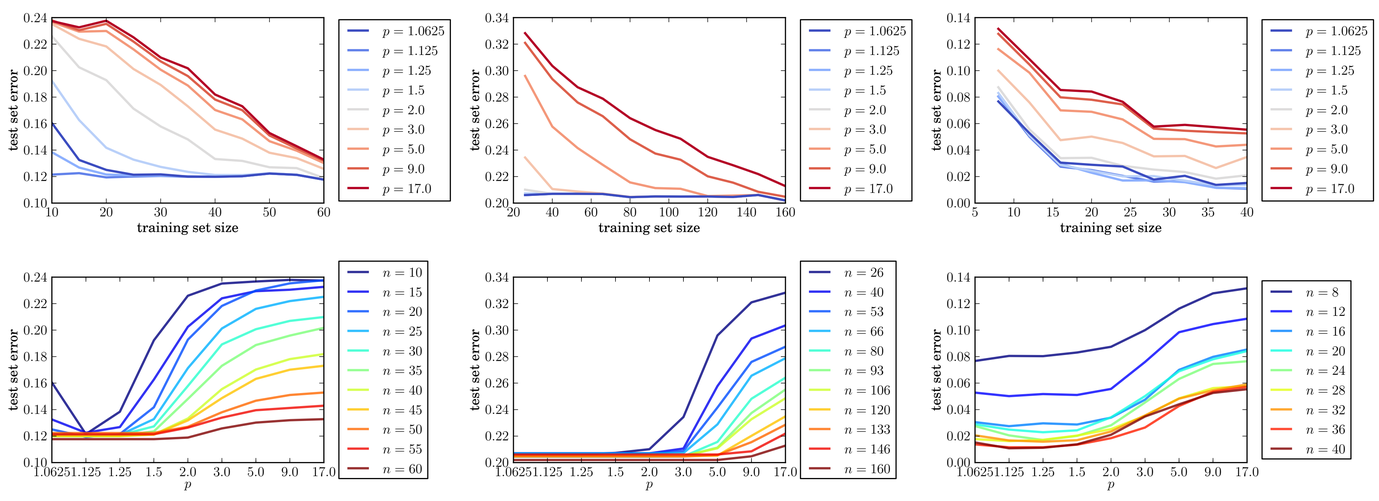
Example. If $k = O(d^{1/4})$ and $\Omega(d^{-1/4}) \leq \epsilon \leq 1/40$, then we have sample complexity bounds

$$p = 1: \quad \tilde{O}(\sqrt{d})$$
$$p > 1: \quad \tilde{\Omega}(d)$$

Synthetic Data Results



Real Data Results



Subsequent Work

Theorem 1. *Let* $\|X\|_p = \left(\sup_{x \sim D} \|x\|_p\right)$ *and*

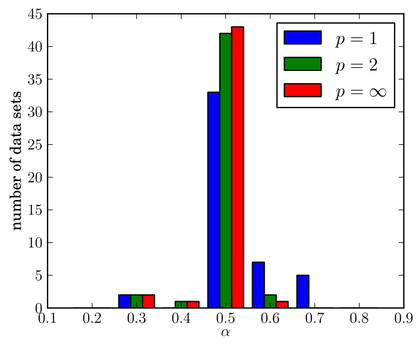
$$\|\mathbf{X}\|_{2,p} := \left(\sum_{i=1}^d \left(\sum_{j=1}^n |x_i^j|^2\right)^{p/2}\right)^{1/p}$$

If there are constants $C = C(d, p)$ *and* $0 \leq \alpha < 1$ *such that* $\|\mathbf{X}\|_{2,p} \leq C n^\alpha \|X\|_p$ *for any data set from* D*, then*

$$\tilde{O}\left(\frac{1}{\epsilon}\left(\frac{C\sqrt{p}}{\gamma_{q,p}}\right)^{\frac{1}{1-\alpha}}\right)$$

samples suffices to achieve error ϵ *for* ***any*** $1 \leq p < \infty$.

Estimation of α on real data sets



Discussion

- Different notions of margin yield different algorithm behavior
- Performance highly dependent on properties of data
- Either side of the spectrum can outperform the other
- $p < 2$ regime not as simple as $p \geq 2$

Future Work:
- Extend example to algorithms using finite samples
- Generalization bound for non-realizable case
- Explore relationship with structured sparsity