A New Perspective on Learning Linear Separators with Large L_qL_p Margins: Supplementary Material

Maria-Florina Balcan
Georgia Institute of Technology

Christopher Berlind
Georgia Institute of Technology

A Examples of $L_\infty L_1$ Margins

Here we give two natural learning settings where $L_\infty L_1$ margins play a key role.

A.1 Two-sided Disjunctions

A two-sided disjunction $h = (h_+, h_-)$ is a pair of disjunctions over a boolean instance space $X = \{0,1\}^n$ that labels points according to the positive disjunction h_+ and is also guaranteed to satisfy $h_+(x) = -h_-(x)$ for all examples $x \sim D$ (Balcan et al., 2013; Blum and Balcan, 2007). The variables included in the disjunction h_+ are positive indicators, those included in h_- are negative indicators, and the remaining are the k non-indicators. If there is a target two-sided disjunction labeling the data from D then we are guaranteed that every example from D has at least one indicator set to 1 and does not have indicators of both types set to 1.

We can represent the target by a linear separator $w^* \in \{-1, 0, 1\}^n$, where the nonzero values in w^* correspond to indicators (positive or negative) and remaining variables are the non-indicators. According to the two-sided disjunction assumption, $|w^* \cdot x| \geq 1$ for any $x \sim D$, so when $\|x\|_1 \leq k$ we immediately have

$$\frac{|w^* \cdot x|}{\|w^*\|_\infty \|x\|_1} \geq \frac{1}{k}$$

and when $\|x\|_1 > k$, $|w^* \cdot x|$ is minimized when x has all k non-indicators set to 1, so we have

$$\frac{|w^* \cdot x|}{\|w^*\|_\infty \|x\|_1} \geq \frac{\|x\|_1 - k}{\|x\|_1} \geq \frac{1}{k+1}.$$

Combining these two cases gives us $\gamma_{\infty,1}(w^*) \geq \frac{1}{k+1}$, so the $L_\infty L_1$ margin is roughly inversely proportional to the number of non-indicators.

A.2 Majority with Margins

As above, we have n boolean variables divided into positive and negative indicators (this time with no non-indicators), and the target is a majority function over the variables set to 1 in an example. The assumption of majority with margins ensures that for some constant $1/2 < \alpha \leq 1$, at least an α fraction of indicators in positive examples are positive (so at most a $1-\alpha$ fraction are negative) and at least an α fraction of indicators in negative examples are negative (at most a $1-\alpha$ fraction are positive). Representing the target as $w^* \in \{-1, 1\}^n$ and $X = \{0,1\}^n$, we have for every $x \sim D$,

$$|w^* \cdot x| \geq \alpha \|x\|_1 - (1-\alpha)\|x\|_1 = (2\alpha - 1)\|x\|_1 .$$

Thus, $\gamma_{\infty,1}(w^*) \geq 2\alpha - 1$ because

$$\frac{|w^* \cdot x|}{\|w^*\|_\infty \|x\|_1} \geq \frac{(2\alpha-1)\|x\|_1}{\|x\|_1} = 2\alpha - 1,$$

and we have a constant $L_\infty L_1$ margin.

B Generalization Bounds in the Non-realizable Case

The results in Section 3 apply to the realizable case—that is, when the two classes are linearly separable by a positive "hard margin." When the data is not linearly separable, convex program (1) has no solution, but convex program (2) remains solvable and we may still achieve good generalization performance in the presence of a "soft margin" (some small margin violations exist in the data, but the majority of points will be far from the optimal separator). In this non-realizable case, we can still obtain generalization bounds analogous to Theorem 4, but they will include an additional dependence on how far the data is from being separable by a large margin (the hinge loss).

B.1 Using Rademacher Complexity

The empirical Rademacher complexity of a class $\mathcal{F}$ of real-valued functions is

$$\mathcal{R}_n(\mathcal{F}) \;=\; \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i f(x^i)\right]$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$ is uniform over $\{-1, 1\}^n$. In the case of linear functions $x \mapsto w \cdot x$ with $\|w\|_q \leq \|W\|_q$, this is

$$\begin{aligned}\mathcal{R}_n(\mathcal{F}) &= \frac{1}{n}\mathbb{E}\left[\sup_w \sum_{i=1}^n \sigma_i (w \cdot x^i)\right] \\ &= \frac{1}{n}\mathbb{E}\left[\sup_w w \cdot \left(\sum_{i=1}^n \sigma_i x^i\right)\right] \\ &\leq \frac{\|W\|_q}{n}\mathbb{E}\left[\left\|\sum_{i=1}^n \sigma_i x^i\right\|_p\right] \\ &\leq \frac{B_p \|W\|_q \|\mathbf{X}\|_{2,p}}{n},\end{aligned}$$

where we have applied Jensen's inequality and the Khintchine inequality as in Section 3. This result is a special case of Proposition 2 of Kloft and Blanchard (2012). If $\|\mathbf{X}\|_{2,p} \leq Cn^\alpha \|X\|_p$, then this simplifies to

$$\mathcal{R}_n(\mathcal{F}) \leq \frac{CB_p \|W\|_q \|X\|_p}{n^{1-\alpha}}$$

which can be used to bound the Rademacher complexity term in several standard generalization bounds such as those in terms of convex loss functions.

B.2 Using Fat-shattering Dimension

Theorem VII.14 of Shawe-Taylor and Cristianini (2000) gives a generalization error bound in terms of the fat-shattering dimension of the concept class $\mathcal{F}$ and the sum of the slack variables ξ in convex program (2). The bound is of the form

$$\mathsf{err}(h) \leq \tilde{O}\left(\frac{1}{n}\left(\mathrm{fat}_{\mathcal{F}}(\gamma/16) + \frac{1}{\gamma}\sum_{i=1}^n \hat{\xi}_i\right)\right)$$

where h is the classifier corresponding to a solution $\hat{w}$ of (2) and where $\hat{\xi}_i = \max(0, \gamma - y^i(\hat{w} \cdot x^i))$. We can then use our bound from Theorem 2 to obtain a bound analogous to Theorem 4.

C L_q-norm Regularization and Multiple Kernel Learning

Here we give some details of the relationship between L_q-norm regularized loss minimization and L_r-norm multiple kernel learning (MKL). Throughout the following, we will use the notational conventions from our work on L_qL_p margins: p will index a norm on the instance space X and q will index a norm on the weight vector space. We will also use r to index a norm on the weights of the kernel combination, which corresponds to the q used by Kloft and Blanchard (2012). The p used by Kloft and Blanchard (2012) corresponds to the q used here.

C.1 Multiple Kernel Learning

L_r-norm MKL attempts to learn a nonnegative combination of M base kernels $k_1, \ldots, k_M$ subject to an L_r-norm penalty (for $1 \leq r \leq \infty$) on the combination weights $\theta_1, \ldots, \theta_M$. Specifically, L_r-norm MKL is an empirical risk minimization problem over the class of linear functions

$$\{x \mapsto w \cdot \phi_k(x) : \|w\|_k \leq D\}$$

(where the transformation ϕ_k and norm $\|\cdot\|_k$ are those of the RKHS defined by the combined kernel k) and over the class of kernels

$$\{k = \sum_{i=1}^M \theta_i k_i : \theta \geq 0,\ \|\theta\|_r \leq 1\}.$$

This problem is known to be equivalent to empirical risk minimization over the class of linear functions

$$\mathcal{H}_{q,D,M} = \{x \mapsto w \cdot \phi(x) : w = (w^{(1)}, \ldots, w^{(M)}), \|w\|_{2,q} \leq D\}$$

where $\phi(x) = (\phi_1(x), \ldots, \phi_M(x))$ (a representation of x in the product RKHS of the base kernels) and $\|w\|_{2,q} = \left\|\left(\|w^{(1)}\|_{k_1}, \ldots, \|w^{(M)}\|_{k_M}\right)\right\|_q$ with $1 \leq q = \frac{2r}{r+1} \leq 2$. Because of this equivalence, many risk bounds for MKL make use of the latter class which is easier to deal with theoretically.

C.2 L_q-norm Regularization as L_r-norm MKL

If we want to perform ERM over the L_q-norm regularized linear class

$$\mathcal{H} = \{x \mapsto w \cdot x : \|w\|_q \leq D\}$$

in $\mathbb{R}^d$, we can phrase this as an L_r MKL problem as follows. Use $M = d$ base kernels $k_1, \ldots, k_d$ where $k_i(x, x') = x_i x_i'$, the product of the i-th coordinates of the argument vectors. Then $\phi_i(x) = x_i$ is a valid kernel mapping for each i because each kernel is then the inner product in the corresponding space. Then the mapping $\phi(x)$ in the definition of $\mathcal{H}_{q,D,M}$ is the identity mapping, so the weight vector w in its definition is the same as that of $\mathcal{H}$. It only remains to verify that the norms are the same. The RKHS norm $\|\cdot\|_{k_i}$ can be expressed as $\|x\|_{k_i} = \sqrt{\phi_i(x) \cdot \phi_i(x)} = \sqrt{k_i(x, x)} = |x_i|$, so $\|w\|_{2,q} = \|(|w_i|, \ldots, |w_d|)\|_q = \|w\|_q$ and the two function classes are the same.

Proposition 2 of Kloft and Blanchard (2012) upper bounds the global empirical Rademacher complexity of $\mathcal{H}_{q,D,M}$ as

$$R(\mathcal{H}_{q,D,M}) \leq \frac{D\sqrt{p}}{n}\sqrt{\|(\mathrm{tr}(K_1), \ldots, \mathrm{tr}(K_M))\|_{p/2}}$$

where $\text{tr}(K_i)$ is the trace of the kernel matrix formed by $k_i(x^j, x^k)$ for each pair x^j, x^k in the data set[1]. In our specific setting, $\text{tr}(K_i) = \sum_{j=1}^{n} (x_i^j)^2$ so the bound becomes

$$\begin{aligned} R(\mathcal{H}_{q,D,M}) &\leq \frac{D\sqrt{p}}{n} \left(\sum_{i=1}^{d} |\text{tr}(K_i)|^{p/2} \right)^{1/p} \\ &= \frac{D\sqrt{p}}{n} \left(\sum_{i=1}^{d} \left(\sum_{j=1}^{n} (x_i^j)^2 \right)^{p/2} \right)^{1/p} \\ &= \frac{D\sqrt{p}}{n} \|\mathbf{X}\|_{2,p} \end{aligned}$$

where $\mathbf{X}$ is the $d \times n$ data matrix with one example in each column.

Note that in this setting, the rank of each base kernel is 1 (because the dimension of the ϕ-space of each kernel is 1). This means each kernel k_i will have one eigenvalue equal to $\mathbb{E}x_i^2$, and this setting satisfies the eigenvalue decay rate assumption $\lambda_j^{(i)} \leq d_i j^{-\alpha_i}$ with $d_i = \mathbb{E}x_i^2$ and arbitrarily large α_i.

References

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.

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

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.

J. Shawe-Taylor and N. Cristianini. On the Generalisation of Soft Margin Algorithms. *IEEE Transactions on Information Theory*, 48, 2000.

[1] Kloft and Blanchard (2012) use K_i to mean the *normalized* kernel matrix which is a factor $1/n$ times our unnormalized one. This accounts for the $1/\sqrt{n}$ in their Rademacher bound while ours here has a $1/n$.