The Rademacher Complexity of Co-Regularized Kernel Classes

David S. Rosenberg
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720
drosen@stat.berkeley.edu

Peter L. Bartlett
Department of Statistics and Computer Science Divison
University of California, Berkeley
Berkeley, CA 94720
bartlett@stat.berkeley.edu

Abstract

In the multi-view approach to semi-supervised learning, we choose one predictor from each of multiple hypothesis classes, and we "co-regularize" our choices by penalizing disagreement among the predictors on the unlabeled data. We examine the co-regularization method used in the co-regularized least squares (CoRLS) algorithm [12], in which the views are reproducing kernel Hilbert spaces (RKHS's), and the disagreement penalty is the average squared difference in predictions. The final predictor is the pointwise average of the predictors from each view. We call the set of predictors that can result from this procedure the co-regularized hypothesis class. Our main result is a tight bound on the Rademacher complexity of the co-regularized hypothesis class in terms of the kernel matrices of each RKHS. We find that the co-regularization reduces the Rademacher complexity by an amount that depends on the distance between the two views, as measured by a data dependent metric. We then use standard techniques to bound the gap between training error and test error for the CoRLS algorithm. Experimentally, we find that the amount of reduction in complexity introduced by co-regularization correlates with the amount of improvement that co-regularization gives in the CoRLS algorithm.

1 Introduction

In the multi-view approach to semi-supervised learning, we have several classes of predictors, or "views." The goal is to find a predictor in each view that performs well on the labeled data, such that all the chosen predictors give similar predictions on the unlabeled data. This approach is motivated by the assumption that each view contains a predictor that's approximately correct. Roughly speaking, predictors that are approximately correct are also approximately equal. Thus we can reduce the complexity of our learning problem by eliminating from the search space all predictors that don't have matching predictors in each of the other views. Because of this reduction in complexity, it's reasonable to expect better test performance for the same amount of labeled training data. This paper provides a more precise understanding of the ways in which the agreement constraint and the choice of views affect complexity and generalization.

Early theoretical results on the multi-view approach to semi-supervised learning, in particular [5] and the original co-training paper [3], assume that the predictors of each view are conditionally independent given the labels they try to predict. This assumption is difficult to justify in practice, yet there are scant other theoretical results to guide the choice of views. In [1], the authors present a theoretical framework for semi-supervised learning that nicely contains multi-view learning as a special case. Although their results do not assume independent views, their sample complexity results are in terms of the complexity of the space of "compatible predictors," which in the case of multi-view learning corresponds to those predictors that have matching predictors in the other views. To apply these results to a particular multi-view learning algorithm, one must compute the complexity of the class of compatible predictors. This problem is addressed to some extent in [6], in which they compute an upper bound on the Rademacher complexity of the space of compatible predictors. However, their bound is given as the solution to an optimization problem.

In this paper, we consider the co-regularized least squares (CoRLS) algorithm, a two-view, semi-supervised version of regularized least squares (RLS).

The algorithm was first discussed in [12], and a similar algorithm was given earlier in [10]. Although CoRLS has been shown to work well in practice for both classification [12] and regression [4], many would-be users of the algorithm are deterred by the requirement of choosing two views, as this often seems an arbitrary process. We attempt to improve this situation by showing how the choice of views affects generalization performance, even in settings where we can't make any probabilistic assumptions about the views.

In CoRLS, the two views are reproducing kernel Hilbert spaces (RKHS's), call them $\mathcal{F}$ and $\mathcal{G}$. We find predictors $f^* \in \mathcal{F}$ and $g^* \in \mathcal{G}$ that minimize an objective function of the form

$$\begin{aligned}&\text{Labeled Loss}(f,g) + \text{Complexity}(f,g)\\ &+\lambda \sum_{x \in \{\text{unlabeled points}\}} [f(x) - g(x)]^2.\end{aligned}$$

The last term is the "co-regularization," which encourages the selection of a pair of predictors (f^*, g^*) that agree on the unlabeled data. We follow [4, 6] and consider the average predictor $\varphi^* := (f^* + g^*)/2$, which comes from the class

$$\mathcal{J}_{\text{ave}} := \{x \mapsto [f(x) + g(x)]/2 : (f,g) \in \mathcal{F} \times \mathcal{G}\}.$$

For typical choices of $\mathcal{F}$ and $\mathcal{G}$, this class is too large to admit uniform generalization bounds. In Section 3.1, however, we see that under a boundedness condition on the labeled loss, the complexity and coregularization terms force φ^* to come from a much smaller class $\mathcal{J}_\lambda$, where $\lambda \geq 0$ is the coefficient of the co-regularization term in the objective function. As λ increases, it is clear that the size of $\mathcal{J}_\lambda$ decreases, and we'd expect this to improve generalization. We make this precise in Theorem 2, where we use standard arguments to bound the gap between training error and test error in terms of the Rademacher complexity of $\mathcal{J}_\lambda$.

The main contribution of this paper is Theorem 3, which gives an explicit expression for the Rademacher complexity of $\mathcal{J}_\lambda$, up to a small constant factor. For ordinary kernel RLS (i.e. single view and fully supervised), it is known that the squared Rademacher complexity is proportional to the trace of the kernel matrix (see e.g. [11, Thm 7.39, p. 231]). We find that for the two-view case without coregularization (i.e. $\lambda = 0$), $\mathcal{J}_\lambda$ has squared Rademacher complexity equal to the average of the traces of the two labeled-data kernel matrices. When $\lambda > 0$, the coregularization term reduces this quantity by an amount that depends on how different the two views are, and in particular on the average distance between the two views' representations of the labeled data, where the distance metric is determined by the unlabeled data.

In Section 2, we give a formal presentation of the CoRLS algorithm. Our results are presented in Section 3, discussed in Section 4, and proved in Section 5. In Section 6, we present an empirical investigation of whether the effect of co-regularization on test performance is correlated with its effect on Rademacher complexity.

2 Co-Regularized Least Squares

We consider the case of two views, though both the algorithm [4] and the analysis can be extended to multiple views. Our views are RKHS's $\mathcal{F}$ and $\mathcal{G}$ of functions mapping from an arbitrary space $\mathcal{X}$ to the reals. The CoRLS algorithm takes labeled points $(x_1, y_1), \ldots, (x_\ell, y_\ell) \in \mathcal{X} \times \mathcal{Y}$, and unlabeled points $x_{\ell+1}, \ldots, x_{\ell+u} \in \mathcal{X}$, and solves the following minimization problem:

$$\begin{aligned}(f^*, g^*) = \operatorname*{arg\,min}_{f \in \mathcal{F}, g \in \mathcal{G}} &\hat{L}(f,g) + \gamma_{\mathcal{F}} ||f||^2_{\mathcal{F}} + \gamma_{\mathcal{G}} ||g||^2_{\mathcal{G}}\\ &+ \lambda \sum_{i=\ell+1}^{\ell+u} [f(x_i) - g(x_i)]^2\end{aligned}$$

for some loss functional $\hat{L}$ that depends only on the labeled data, and regularization parameters $\gamma_{\mathcal{F}}$, $\gamma_{\mathcal{G}}$, and λ. The final output is $\varphi^* := (f^* + g^*)/2$.

In [12, 4], the loss functional considered was

$$\hat{L}(f,g) = \frac{1}{2\ell} \sum_{i=1}^{\ell} \left(\big[f(x_i) - y_i\big]^2 + \big[g(x_i) - y_i\big]^2 \right)$$

If we use this loss and set $\lambda = 0$, the objective function decouples into two single-view, fully-supervised kernel RLS regressions. We also propose the loss functional

$$\hat{L}(f,g) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\frac{f(x_i) + g(x_i)}{2} - y_i \right)^2$$

as one that seems natural when the final prediction function is $\frac{1}{2}(f+g)$, as in [4, 6]. Our analysis applies to both of these loss functionals, as well as many more that depend on the labeled data only and that satisfy a boundedness condition specified in Section 3.1. Thus we take the "S" in CoRLS to refer to the squares in the complexity and co-regularization terms, which our analysis requires, rather than to the squares in the loss functional, which we don't require.

2.1 Notation and Preliminaries

We'll denote the reproducing kernels corresponding to $\mathcal{F}$ and $\mathcal{G}$ by $k_{\mathcal{F}} : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ and $k_{\mathcal{G}} : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$, respectively. It's convenient to introduce notation for the "span of the data" in each

space: $\mathcal{L}_\mathcal{F} := \text{span}\{k_\mathcal{F}(x_i,\cdot)\}_{i=1}^{\ell+u} \subset \mathcal{F}$ and $\mathcal{L}_\mathcal{G} := \text{span}\{k_\mathcal{G}(x_i,\cdot)\}_{i=1}^{\ell+u} \subset \mathcal{G}$. By the Representer Theorem, it's clear that $(f^*, g^*) \in \mathcal{L}_\mathcal{F} \times \mathcal{L}_\mathcal{G}$. That is, we can write the CoRLS solution as

$$\begin{aligned} f^*(\cdot) &= \sum_{i=1}^{u+\ell} \alpha_i k_\mathcal{F}(x_i,\cdot) \in \mathcal{L}_\mathcal{F} \\ g^*(\cdot) &= \sum_{i=1}^{u+\ell} \beta_i k_\mathcal{G}(x_i,\cdot) \in \mathcal{L}_\mathcal{G} \\ \text{for} \qquad & \alpha = (\alpha_1,\ldots,\alpha_{u+\ell}) \in \mathbf{R}^{u+\ell} \\ \text{and} \qquad & \beta = (\beta_1,\ldots,\beta_{u+\ell}) \in \mathbf{R}^{u+\ell}. \end{aligned}$$

We'll denote an arbitrary element of $\mathcal{L}_\mathcal{F}$ by $f_\alpha = \sum_{i=1}^{u+\ell} \alpha_i k_\mathcal{F}(x_i,\cdot)$, and similarly for elements of $\mathcal{L}_\mathcal{G}$.

Define the kernel matrices $(K_\mathcal{F})_{ij} = k_\mathcal{F}(x_i,x_j)$ and $(K_\mathcal{G})_{ij} = k_\mathcal{G}(x_i,x_j)$, and partition them into blocks corresponding to labeled and unlabeled points:

$$K_\mathcal{F} = \begin{pmatrix} A_{u\times u} & C_{u\times \ell} \\ C'_{\ell\times u} & B_{\ell\times\ell} \end{pmatrix} \quad K_\mathcal{G} = \begin{pmatrix} D_{u\times u} & F_{u\times \ell} \\ F'_{\ell\times u} & E_{\ell\times\ell} \end{pmatrix}.$$

We can now write the agreement term as

$$\sum_{i=\ell+1}^{\ell+u} [f_\alpha(x_i) - g_\beta(x_i)]^2 = ||(A\ C)\,\alpha - (D\ F)\,\beta||^2,$$

and it follows from the reproducing property that $||f_\alpha||_\mathcal{F}^2 = \alpha' K_\mathcal{F} \alpha$ and $||g_\beta||_\mathcal{G}^2 = \beta' K_\mathcal{G} \beta$. For each of the loss functionals presented in the beginning of this section, the whole objective function is quadratic in α and β, and thus a solution (f^*, g^*) can be found by differentiating and solving a system of linear equations. See [12, 4] for more details.

3 Results

3.1 Bounding the CoRLS Function Class

We assume the loss functional $\hat{L} : \mathcal{F} \times \mathcal{G} \to [0,\infty)$ satisfies

$$\hat{L}(0,0) \le 1.$$

That is, $\hat{L}(f,g) \le 1$ for $f \equiv 0$ and $g \equiv 0$. This is true, for example, for the two loss functionals given in Section 2, provided that $\mathcal{Y} = [-1,1]$. Assuming $\hat{L}(0,0) \le 1$, we now derive the "co-regularized" function class[1] $\mathcal{J} \subset \mathcal{J}_{\text{ave}}$, from which the CoRLS predictors are drawn.

Recall that our original problem was to minimize

$$\begin{aligned} Q(f,g) :=& \hat{L}(f,g) + \gamma_\mathcal{F}||f||_\mathcal{F}^2 + \gamma_\mathcal{G}||g||_\mathcal{G}^2 \\ &+ \lambda \sum_{i=\ell+1}^{\ell+u} [f(x_i) - g(x_i)]^2 \end{aligned}$$

over $\mathcal{F} \times \mathcal{G}$. Plugging in the trivial predictors $f \equiv 0$ and $g \equiv 0$ gives the following upper bound:

$$\min_{f,g} Q(f,g) \le Q(0,0) = \hat{L}(0,0) \le 1$$

Since all terms of $Q(f,g)$ are nonnegative, we conclude that any (f^*, g^*) minimizing $Q(f,g)$ is contained in

$$\begin{aligned} \mathcal{H} :=& \Bigg\{ (f,g) \,:\, \gamma_\mathcal{F}||f||_\mathcal{F}^2 + \gamma_\mathcal{G}||g||_\mathcal{G}^2 \\ &+ \lambda \sum_{i=\ell+1}^{\ell+u} |f(x_i) - g(x_i)|^2 \le 1 \Bigg\}, \end{aligned}$$

and the final predictor for the CoRLS algorithm is chosen from the class

$$\mathcal{J} := \{x \mapsto [f(x) + g(x)]/2 \,:\, (f,g) \in \mathcal{H}\}.$$

Note that the function classes $\mathcal{H}$ and $\mathcal{J}$ do not depend on the labeled data, and thus are deterministic after conditioning on the unlabeled data.

3.2 Setup for the Theorems

We will use empirical Rademacher complexity as our measure of the size of a function class. The empirical Rademacher complexity of a function class $\mathcal{F} = \{\varphi : \mathcal{X} \to \mathcal{Y}\}$ for a sample $x_1,\ldots,x_\ell \in \mathcal{X}$ is defined as

$$\hat{R}_\ell(\mathcal{F}) = E^\sigma \left[\sup_{\varphi\in\mathcal{F}} \left| \frac{2}{\ell} \sum_{i=1}^{\ell} \sigma_i \varphi(x_i) \right| \right],$$

where the expectation is with respect to $\sigma = \{\sigma_1,\ldots,\sigma_\ell\}$, and the σ_i are i.i.d. Rademacher random variables[2].

In our semi-supervised context, we assume that the labeled points $(x_1,y_1),\ldots,(x_\ell,y_\ell)$ are drawn i.i.d. from a distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, but make no assumptions about the unlabeled points $x_{\ell+1},\ldots,x_{\ell+u} \in \mathcal{X}$. Indeed, our claims are conditional on the unlabeled data, and thus remain true no matter what distribution the unlabeled points are drawn from.

For a given loss function $L : \mathcal{Y}^2 \to [0,1]$, and for any choice of $\varphi \in \mathcal{J}$, we are interested in bounds on the expected loss $E_\mathcal{D} L(\varphi(X), Y)$. Typically, L would be the loss used to define the labeled empirical risk functional $\hat{L}$, but it need not be.

Our generalization bound in Theorem 2 is based on the following theorem (see e.g. [11, Thm 4.9, p. 96]):

[1] We now suppress λ in the $\mathcal{J}_\lambda$ from Section 1.

[2] We say σ is a Rademacher random variable if $P(\sigma = 1) = P(\sigma = -1) = \frac{1}{2}$.

Theorem 1. *Fix $\delta \in (0,1)$, and let $\mathcal{Q}$ be a class of functions mapping from $\mathcal{X} \times \mathcal{Y}$ to $[0,1]$. With probability at least $1-\delta$ over the sample $(X_1,Y_1),\ldots,(X_\ell,Y_\ell)$ drawn i.i.d. from $\mathcal{D}$, every $q \in \mathcal{Q}$ satisfies*

$$E_{\mathcal{D}}q(X,Y) \leq \frac{1}{\ell}\sum_{i=1}^{\ell} q(X_i,Y_i) + \hat{R}_\ell(\mathcal{Q}) + 3\sqrt{\frac{\ln(2/\delta)}{2\ell}}.$$

The expression $E_{\mathcal{D}}q(X,Y)$ is deterministic, but unknown to us because we do not know the data generating distribution $\mathcal{D}$. The terms $\frac{1}{\ell}\sum_{i=1}^{\ell} q(X_i,Y_i)$ and $\hat{R}_\ell(\mathcal{Q})$ are random, but with probability at least $1-\delta$, the inequality holds for the observed values of these random quantities, and for every $q \in \mathcal{Q}$.

3.3 Theorems

In Theorem 2, we give generalization bounds for the class $\mathcal{J}$ in terms of the empirical Rademacher complexity $\hat{R}_\ell(\mathcal{J})$. In Theorem 3, we give upper and lower bounds on $\hat{R}_\ell(\mathcal{J})$ that can be written explicitly in terms of blocks of the kernel matrices $K_{\mathcal{F}}$ and $K_{\mathcal{G}}$.

Theorem 2. *Suppose that $L : \mathcal{Y}^2 \to [0,1]$ satisfies the following uniform Lipschitz condition: for all $y \in \mathcal{Y}$ and all $\hat{y}_1, \hat{y}_2 \in \mathcal{Y}$ with $\hat{y}_1 \neq \hat{y}_2$,*

$$\frac{|L(\hat{y}_1,y) - L(\hat{y}_2,y)|}{|\hat{y}_1 - \hat{y}_2|} \leq B.$$

Then conditioned on the unlabeled data, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the sample of labeled points $(X_1,Y_1),\ldots,(X_\ell,Y_\ell)$ drawn i.i.d. from $\mathcal{D}$, we have for any predictor $\varphi \in \mathcal{J}$ that

$$E_{\mathcal{D}}L(\varphi(X),Y) \leq \frac{1}{\ell}\sum_{i=1}^{\ell} L(\varphi(X_i),Y_i) + 2B\hat{R}_\ell(\mathcal{J}) + \frac{1}{\sqrt{\ell}}\left(2 + 3\sqrt{\ln(2/\delta)/2}\right)$$

Note that for $\mathcal{Y} = [-1,1]$, the conditions of this theorem are satisfied by the loss function

$$L(\hat{y},y) = (\tau(\hat{y}) - y)^2/4,$$

where $\tau(y) = \min(1, \max(-1, y))$.

The following theorem is the main result of the paper. Recall that A and D are the unlabeled kernel submatrices, B and E are the labeled kernel submatrices, and C and F involve the cross-terms.

Theorem 3. *For the CoRLS function class $\mathcal{J}$,*

$$\frac{1}{\sqrt[4]{2}}\frac{U}{\ell} \leq \hat{R}_\ell(\mathcal{J}) \leq \frac{U}{\ell},$$

where

$$U^2 = \gamma_{\mathcal{F}}^{-1}tr(B) + \gamma_{\mathcal{G}}^{-1}tr(E) - \lambda tr\left(J'\left(I + \lambda M\right)^{-1}J\right),$$

with I the identity matrix, and

$$J = \gamma_{\mathcal{F}}^{-1}C - \gamma_{\mathcal{G}}^{-1}F \qquad M = \gamma_{\mathcal{F}}^{-1}A + \gamma_{\mathcal{G}}^{-1}D.$$

4 Discussion

4.1 Unlabeled Data Improves the Bound

The regularization parameter λ controls the amount that the unlabeled data constrains the hypothesis space. It's obvious from the definition of the hypothesis class $\mathcal{J}$ that if $\lambda_1 \geq \lambda_2 \geq 0$, then $\mathcal{J}_{\lambda_1} \subseteq \mathcal{J}_{\lambda_2}$, and thus $\hat{R}_\ell(\mathcal{J}_{\lambda_1}) \leq \hat{R}_\ell(\mathcal{J}_{\lambda_2})$. That is, increasing λ reduces the Rademacher complexity $\hat{R}_\ell(\mathcal{J}_\lambda)$. The amount of this reduction is characterized by the expression

$$\Delta(\lambda) := \lambda \mathrm{tr}\left(J'\left(I + \lambda M\right)^{-1}J\right)$$

from Theorem 3. When $\lambda = 0$, the algorithm ignores the unlabeled data, and the reduction is indeed $\Delta(0) = 0$. As we would expect, $\Delta(\lambda)$ is nondecreasing in λ and has a finite limit as $\lambda \to \infty$. We collect these properties in a proposition:

Proposition 1. *$\Delta(0) = 0$, $\Delta(\lambda)$ is nondecreasing on $\lambda \geq 0$, and $\lim_{\lambda\to\infty}\Delta(\lambda) = tr(J'M^{-1}J)$, provided the inverse exists.*

Proof. The limit claim is clear if we write the reduction as $\Delta(\lambda) = \mathrm{tr}(J'(\lambda^{-1}I + M)^{-1}J)$. Since A and D are Gram matrices, their positive combination M is positive semidefinite (psd). Thus we can write $M = Q'DQ$, with diagonal $D \geq 0$ and orthogonal Q. Then

$$\begin{aligned}\Delta(\lambda) &= \mathrm{tr}\left(J'Q'\left(\lambda^{-1}I + D\right)^{-1}QJ\right) \\ &= \sum_{i=1}^{\ell}\sum_{j=1}^{u}(QJ)_{ij}^2\left(\lambda^{-1} + D_{jj}\right)^{-1}\end{aligned}$$

From this expression, it's clear that $\Delta(\lambda)$ is nondecreasing in λ on $(0,\infty)$. Since $\Delta(\lambda)$ is continuous at $\lambda = 0$, it's nondecreasing on $[0,\infty)$. □

4.2 Interpretation of Improvement

Here we take some steps towards interpreting the reduction in complexity $\Delta(\lambda)$. For simplicity, take $\gamma_{\mathcal{F}} = \gamma_{\mathcal{G}} = 1$. Then the reduction is given by

$$\Delta(\lambda) = \lambda(C - F)'(I + \lambda M)^{-1}(C - F)$$

Note that the jth column of matrix C gives a representation of the jth labeled point by its $\mathcal{F}$-inner

product with each of the unlabeled points. That is, for $j = 1, \ldots, \ell$ and for $i = 1, \ldots, u$, we have $C_{ij} = \langle k_{\mathcal{F}}(x_{\ell+i}, \cdot), k_{\mathcal{F}}(x_j, \cdot) \rangle$. Similarly, F represents each labeled point by its $\mathcal{G}$-inner product with each of the unlabeled points. Since $(I + \lambda M)^{-1}$ is psd, it defines a semi-norm. Thus we can write the reduction in complexity as

$$\begin{aligned} \Delta(\lambda) &= \lambda \sum_{i=1}^{\ell} ||C_{\cdot i} - F_{\cdot i}||^2_{(I+\lambda M)^{-1}} \\ &= \sum_{i=1}^{\ell} ||C_{\cdot i} - F_{\cdot i}||^2_{(I/\lambda + M)^{-1}} \quad (\text{for } \lambda > 0) \end{aligned}$$

We see that $\Delta(\lambda)$ grows with the distance between the two different representations of the labeled points. For very small λ, this distance is essentially measured using the Euclidean norm. As λ grows, the distance approaches that determined by M^{-1}, where M is the sum of the two unlabeled data kernel matrices. Loosely summarized, the reduction is proportional to the difference between the representations of the labeled data in the two different views, where the measure of difference is determined by the unlabeled data.

5 Proofs

5.1 Proof of Theorem 2.

Define the loss class

$$\mathcal{Q} = \{(x, y) \mapsto L(\varphi(x), y) \ : \ \varphi \in \mathcal{J}\}.$$

By assumption, any function in $\mathcal{Q}$ maps into $[0, 1]$. Applying Theorem 1, we have for any $\varphi \in \mathcal{J}$, with probability at least $1-\delta$ over the labeled sample $(X_i, Y_i)_{i=1}^{\ell}$, that

$$\begin{aligned} E_{\mathcal{D}} L(\varphi(X), Y) \leq & \frac{1}{\ell} \sum_{i=1}^{\ell} L(\varphi(X_i), Y_i) \\ & + \hat{R}_\ell(\mathcal{Q}) + 3\sqrt{\frac{\ln(2/\delta)}{2\ell}}. \end{aligned}$$

The following lemma completes the proof:

Lemma 1. $\hat{R}_\ell(\mathcal{Q}) \leq 2B\hat{R}_\ell(\mathcal{J}) + \frac{2}{\sqrt{\ell}}$.

Proof. Define the functions $g_y = L(0, y)$ and $h_y(\hat{y}) = L(\hat{y}, y) - L(0, y)$. Then $L(\varphi(x), y) = g_y + h_y(\varphi(x))$, and

$$\mathcal{Q} = g_y + h_y \circ \mathcal{J}.$$

Since $|g_y| \leq 1$ for all y, we have

$$\hat{R}_\ell(\mathcal{Q}) \leq \hat{R}_\ell(h_y \circ \mathcal{J}) + \frac{2}{\sqrt{\ell}},$$

by a property of Rademacher complexity (see e.g. [11, Thm 4.15(v), p. 101]). For all y, $h_y(\cdot)$ is Lipschitz with constant B, and $h_y(0) = 0$. Thus by the Ledoux-Talagrand contraction inequality [9, Thm 4.12, p.112] we have

$$\hat{R}_\ell(h_y \circ \mathcal{J}) \leq 2B\hat{R}_\ell(\mathcal{J})$$

□

The bound in Lemma 1 is sometimes of the right order of magnitude and sometimes quite loose. Consider the space $\mathcal{X} \times \mathcal{Y} = \mathbf{R} \times \{-1, 0, 1\}$ and the loss $L(\hat{y}, y) = (\tau(\hat{y}) - y)^2/4$, where $\tau(y) = \min(1, \max(-1, y))$. Assume $P(y = 0) = 1$. Then for any class $\mathcal{J}$ of predictors that map into $\{-1, 1\}$, with probability 1 we have $L(\varphi(x), y) = 1$, and thus $\hat{R}(\mathcal{Q}) = \frac{2}{\ell} E^\sigma \left| \sum_{i=1}^{\ell} \sigma_i \right|$. By the Kahane-Khintchine inequality (c.f. Section 5.2.3), we conclude $\hat{R}(\mathcal{Q}) = \Theta(\ell^{-1/2})$. If we choose $\mathcal{J}$ small, say $\mathcal{J} = \{x \mapsto 1\}$, then $\hat{R}_\ell(\mathcal{J}) = \frac{2}{\ell} E^\sigma \left| \sum_{i=1}^{\ell} \sigma_i \right| = \Theta(\ell^{-1/2})$, and the bound is tight. If we choose $\mathcal{J}$ as large as possible, and we assume that x has a continuous distribution, then $\hat{R}_\ell(\mathcal{J}) = 2$ almost surely, and the bound is loose.

5.2 Proof of Theorem 3

We prove this theorem in several steps, starting from the definition

$$\hat{R}_\ell(\mathcal{J}) = E^\sigma \left[\sup_{(f,g) \in \mathcal{H}} \left| \frac{1}{\ell} \sum_{i=1}^{\ell} \sigma_i (f(x_i) + g(x_i)) \right| \right],$$

where as usual the expectation is w.r.t. σ. We first convert from a supremum over the function space $\mathcal{H}$ to a supremum over a finite-dimensional Euclidean space that we can solve directly. Next, we use the Kahane-Khintchine inequality to bound the expectation over σ above and below by expectations that we can compute explicitly. Finally, with some matrix algebra we can write $\hat{R}_\ell(\mathcal{J})$ in terms of blocks of the original kernel matrices.

5.2.1 Converting to Euclidean Space

Since $(f, g) \in \mathcal{H}$ implies $(-f, -g) \in \mathcal{H}$, we can drop the absolute value. Next, note that the expression inside the supremum depends only on the values of f and g at the sample points. By the reproducing kernel property, it's easy to show that $f(x_j) = (\text{Proj}_{\mathcal{L}_{\mathcal{F}}} f)(x_j)$ and $g(x_j) = (\text{Proj}_{\mathcal{L}_{\mathcal{G}}} g)(x_j)$ for any sample point x_j. Thus the supremum in the expression for $\hat{R}_\ell(\mathcal{J})$ is unchanged if we restrict the supremum to $(f, g) \in$

$(\mathcal{L}_\mathcal{F} \times \mathcal{L}_\mathcal{G}) \cap \mathcal{H}$. Applying these observations we get

$$\hat{R}_\ell(\mathcal{J}) = \frac{1}{\ell} E^\sigma \sup \left\{ \sum_{i=1}^{\ell} \sigma_i (f(x_i) + g(x_i)) : (f,g) \in (\mathcal{L}_\mathcal{F} \times \mathcal{L}_\mathcal{G}) \cap \mathcal{H} \right\}.$$

Finally, we can write the set $(\mathcal{L}_\mathcal{F} \times \mathcal{L}_\mathcal{G}) \cap \mathcal{H}$ as

$$\begin{aligned} &\left\{ (f_\alpha, g_\beta) : \gamma_\mathcal{F} \alpha' K_\mathcal{F} \alpha + \gamma_\mathcal{G} \beta' K_\mathcal{G} \beta \right. \\ &\qquad \left. + \lambda \sum_{i=\ell+1}^{\ell+u} |f_\alpha(x_i) - g_\beta(x_i)|^2 \le 1 \right\} \\ &= \left\{ (f_\alpha, g_\beta) : (\alpha' \ \beta') N \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \le 1 \right\}, \end{aligned}$$

where

$$N := \begin{pmatrix} \gamma_\mathcal{F} K_\mathcal{F} & 0 \\ 0 & \gamma_\mathcal{G} K_\mathcal{G} \end{pmatrix} + \lambda K_\mathcal{C},$$

$$K_\mathcal{C} := \begin{pmatrix} A \\ C' \\ -D \\ -F' \end{pmatrix} (A \ C \ -D \ -F)$$

Now we can write

$$\hat{R}_\ell(\mathcal{J}) = \frac{1}{\ell} E^\sigma \left[\sup_{\alpha,\beta} \left\{ \sigma'(C' \ B)\alpha + \sigma'(F' \ E)\beta \right. \right. \left. \left. \text{s.t. } (\alpha' \ \beta') N \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \le 1 \right\} \right].$$

5.2.2 Evaluating the Supremum

For a symmetric positive definite (spd) matrix M, it's easy to show that $\sup_{\alpha:\alpha' M \alpha \le 1} v'\alpha = \left|\left| M^{-1/2} v \right|\right|$. However, our matrix N may not have full rank. Note that each entry of the column vector $(C' \ B)\alpha$ is an inner product between α and a row (or column, by symmetry) of $K_\mathcal{F}$. Thus if $\alpha_{||} = \text{Proj}_{\text{ColSpace}(K_\mathcal{F})} \alpha$, then $(C' \ B)\alpha = (C' \ B)\alpha_{||}$. Similar reasoning shows that $(F' \ E)\beta = (F' \ E)\beta_{||}$, for $\beta_{||} = \text{Proj}_{\text{ColSpace}(K_\mathcal{G})} \beta$, and that the quadratic form $(\alpha' \ \beta') N (\alpha' \ \beta')'$ is unchanged when we replace (α, β) by $(\alpha_{||}, \beta_{||})$. Thus the supremum can be rewritten as

$$\sup_{\substack{\alpha_{||} \in \text{ColSpace}(K_\mathcal{F}) \\ \beta_{||} \in \text{ColSpace}(K_\mathcal{G})}} \left\{ \left| \sigma'(C' \ B)\alpha_{||} + \sigma'(F' \ E)\beta_{||} \right| \right. \left. \text{s.t. } (\alpha'_{||} \ \beta'_{||}) N (\alpha'_{||} \ \beta'_{||})' \le 1 \right\}$$

Changing to eigenbases cleans up this expression and clears the way for substantial simplifications in later sections. Diagonalize the psd kernel matrices to get orthonormal bases for the column spaces of $K_\mathcal{F}$ and $K_\mathcal{G}$:

$$V' K_\mathcal{F} V = \Sigma_\mathcal{F} \qquad W' K_\mathcal{G} W = \Sigma_\mathcal{G}$$

where $\Sigma_\mathcal{F}$ and $\Sigma_\mathcal{G}$ are diagonal matrices containing the nonzero eigenvalues, and the columns of V and W are bases for the column spaces of $K_\mathcal{F}$ and $K_\mathcal{G}$, respectively. Now introduce a and b such that

$$\alpha_{||} = Va \qquad \beta_{||} = Wb$$

Applying this change of variables to the quadratic form, we get

$$(\alpha'_{||} \ \beta'_{||}) N \begin{pmatrix} \alpha_{||} \\ \beta_{||} \end{pmatrix} = (a' \ b') T \begin{pmatrix} a \\ b \end{pmatrix}$$

where

$$T = \Sigma + \lambda R R'$$

with

$$\Sigma := \begin{pmatrix} \gamma_\mathcal{F} \Sigma_\mathcal{F} & 0 \\ 0 & \gamma_\mathcal{G} \Sigma_\mathcal{G} \end{pmatrix} \qquad R := \begin{pmatrix} V' & 0 \\ 0 & W' \end{pmatrix} \begin{pmatrix} A \\ C' \\ -D \\ -F' \end{pmatrix}$$

The matrix T is spd, since it's the sum of the spd diagonal matrix Σ and the psd matrix $\lambda R R'$. For compactness, define $\mathcal{W} = (C' \ B \ F' \ E) \left(\begin{smallmatrix} V & 0 \\ 0 & W \end{smallmatrix} \right)$. We can now write

$$\hat{R}_\ell(\mathcal{J}) = \frac{1}{\ell} E^\sigma \left[\sup_{a,b} \left\{ \left| \sigma' \mathcal{W} \left(\begin{smallmatrix} a \\ b \end{smallmatrix} \right) \right| \right. \left. \text{s.t. } (a' \ b') T (a' \ b')' \le 1 \right\} \right].$$

Since T is spd, we can evaluate the supremum as described above to get

$$\hat{R}_\ell(\mathcal{J}) = \frac{1}{\ell} E^\sigma \left|\left| T^{-1/2} \mathcal{W}' \sigma \right|\right|$$

5.2.3 Bounding $\hat{R}_\ell(\mathcal{J})$ above and below

We make use of the following lemma[3]:

Lemma 2 (Kahane-Khintchine inequality)**.** *For any vectors $a_1, \ldots, a_n$ in a Hilbert space and independent Rademacher random variables $\sigma_1, \ldots, \sigma_n$, we have*

$$\frac{1}{\sqrt{2}} E \left|\left| \sum_{i=1}^n \sigma_i a_i \right|\right|^2 \le \left(E \left|\left| \sum_{i=1}^n \sigma_i a_i \right|\right| \right)^2 \le E \left|\left| \sum_{i=1}^n \sigma_i a_i \right|\right|^2$$

Taking the columns of $T^{-1/2}\mathcal{W}'$ to be the a_i's, we can apply this lemma to our expression for $\hat{R}_\ell(\mathcal{J})$ to get

$$\frac{1}{\sqrt[4]{2}} \frac{U}{\ell} \le \hat{R}_\ell(\mathcal{J}) \le \frac{U}{\ell}$$

[3] See [8] for a proof of the lower bound. The upper bound is Jensen's inequality.

where

$$\begin{aligned} U^2 &:= E^\sigma \left|\left| T^{-1/2}\mathcal{W}'\sigma \right|\right|^2 \\ &= E^\sigma \mathrm{tr}\left[\mathcal{W}T^{-1}\mathcal{W}'\sigma\sigma'\right] \\ &= \mathrm{tr}\left[\mathcal{W}T^{-1}\mathcal{W}'\right] \end{aligned}$$

To get the second line we expanded the squared norm, took the trace of the scalar quantity inside the expectation, and rotated the factors inside the trace. To get the last equality we interchanged the trace and the expectation and noted that $E\sigma\sigma'$ is the identity matrix.

5.2.4 Writing our Expression in terms of the Original Kernel Matrices

It will be helpful to divide V and W into labeled and unlabeled parts. We note the dimensions of V and W are $(\ell+u)\times r_\mathcal{F}$ and $(\ell+u)\times r_\mathcal{G}$, where $r_\mathcal{F}$ and $r_\mathcal{G}$ are the ranks of $K_\mathcal{F}$ and $K_\mathcal{G}$, respectively. So we have

$$K_\mathcal{F} = \begin{pmatrix} A & C \\ C' & B \end{pmatrix} = \begin{pmatrix} V_u \\ V_\ell \end{pmatrix} \Sigma_\mathcal{F} (V_u'\ V_\ell') \tag{1}$$

$$K_\mathcal{G} = \begin{pmatrix} D & F \\ F' & E \end{pmatrix} = \begin{pmatrix} W_u \\ W_\ell \end{pmatrix} \Sigma_\mathcal{G} (W_u'\ W_\ell') \tag{2}$$

Rearranging the diagonalization, we also have

$$V' \begin{pmatrix} A & C \\ C' & B \end{pmatrix} = \Sigma_\mathcal{F} (V_u'\ V_\ell') \tag{3}$$

$$W' \begin{pmatrix} D & F \\ F' & E \end{pmatrix} = \Sigma_\mathcal{G} (W_u'\ W_\ell') \tag{4}$$

By equating blocks in these four matrix equations, we attain all the substitutions we need to write U^2 in terms of the original kernel submatrices A, B, C, D, E, and F. For example, by equating the top left submatrices in Equation 1, we get $A = V_u \Sigma_\mathcal{F} V_u'$. Using these substitutions, we can write:

$$\mathcal{W}' = \begin{pmatrix} V' & 0 \\ 0 & W' \end{pmatrix} \begin{pmatrix} C \\ B \\ F \\ E \end{pmatrix} = \begin{pmatrix} \Sigma_\mathcal{F} & 0 \\ 0 & \Sigma_\mathcal{G} \end{pmatrix} \begin{pmatrix} V_\ell' \\ W_\ell' \end{pmatrix}$$

$$R = \begin{pmatrix} V' & 0 \\ 0 & W' \end{pmatrix} \begin{pmatrix} A \\ C' \\ -D \\ -F' \end{pmatrix} = \begin{pmatrix} \Sigma_\mathcal{F} & 0 \\ 0 & -\Sigma_\mathcal{G} \end{pmatrix} \begin{pmatrix} V_u' \\ W_u' \end{pmatrix}$$

We now work on the T^{-1} factor in our expression $U^2 = \mathrm{tr}(\mathcal{W}T^{-1}\mathcal{W})$. Using the Sherman-Morrison-Woodbury formula [4], we expand $T^{-1} = (\Sigma + \lambda RR')^{-1}$ as

$$T^{-1} = \Sigma^{-1} - \lambda\Sigma^{-1}R\left(I + \lambda R'\Sigma^{-1}R\right)^{-1} R'\Sigma^{-1}$$

[4] $(A+UU')^{-1} = A^{-1} - A^{-1}U(I+U^TA^{-1}U)^{-1}U^TA^{-1}$, provided the inverses exist [7, p. 50].

Since Σ and $I + \lambda R'\Sigma^{-1}R$ are spd, our inverses exist and the expansion is justified. Substituting this expansion into our expression for U^2, we get

$$\begin{aligned} U^2 = \mathrm{tr}\left(\mathcal{W}\Sigma^{-1}\mathcal{W}'\right) - \lambda\mathrm{tr}\Big(&\mathcal{W}\Sigma^{-1}R \\ &\times \left(I + \lambda R'\Sigma^{-1}R\right)^{-1} R'\Sigma^{-1}\mathcal{W}\Big) \end{aligned}$$

We'll have our final form once we can express $\mathcal{W}\Sigma^{-1}\mathcal{W}'$, $R'\Sigma^{-1}R$, and $R'\Sigma^{-1}\mathcal{W}$ in terms of the original kernel matrix blocks. We have

$$\begin{aligned} &\mathcal{W}\Sigma^{-1}\mathcal{W}' \\ &= (V_\ell\ W_\ell) \begin{pmatrix} \Sigma_\mathcal{F} & 0 \\ 0 & \Sigma_\mathcal{G} \end{pmatrix} \begin{pmatrix} \gamma_\mathcal{F}^{-1}\Sigma_\mathcal{F}^{-1} & 0 \\ 0 & \gamma_\mathcal{G}^{-1}\Sigma_\mathcal{G}^{-1} \end{pmatrix} \\ &\times \begin{pmatrix} \Sigma_\mathcal{F} & 0 \\ 0 & \Sigma_\mathcal{G} \end{pmatrix} \begin{pmatrix} V_\ell' \\ W_\ell' \end{pmatrix} \\ &= (V_\ell\ W_\ell) \begin{pmatrix} \gamma_\mathcal{F}^{-1}\Sigma_\mathcal{F} & 0 \\ 0 & \gamma_\mathcal{G}^{-1}\Sigma_\mathcal{G} \end{pmatrix} \begin{pmatrix} V_\ell' \\ W_\ell' \end{pmatrix} \\ &= \gamma_\mathcal{F}^{-1} V_\ell \Sigma_\mathcal{F} V_\ell' + \gamma_\mathcal{G}^{-1} W_\ell \Sigma_\mathcal{G} W_\ell' \\ &= \gamma_\mathcal{F}^{-1} B + \gamma_\mathcal{G}^{-1} E \end{aligned}$$

The last equality follows by equating submatrices in Equations 1 and 2. Using very similar steps, but with different substitutions read from Equations 1 and 2, we also get

$$\begin{aligned} R'\Sigma^{-1}R &= \gamma_\mathcal{F}^{-1}A + \gamma_\mathcal{G}^{-1}D = M \\ R'\Sigma^{-1}\mathcal{W}' &= \gamma_\mathcal{F}^{-1}C - \gamma_\mathcal{G}^{-1}F = J \end{aligned}$$

Putting things together, we get

$$U^2 = \mathrm{tr}\left(\gamma_\mathcal{F}^{-1}B + \gamma_\mathcal{G}^{-1}E\right) - \lambda\mathrm{tr}\Big(J'\left(I + \lambda M\right)^{-1} J\Big)$$

□

6 Experiments

The objective of our experiments was to investigate whether the reduction in hypothesis space complexity due to co-regularization correlates with an improvement in test performance. We closely followed the experimental setup used in [4] on the UCI repository data sets [2]. We selected those data sets with continuous target values, between 5 and 500 examples, and at least 5 features. For each of these 29 data sets, we generated two views by randomly splitting the features into two sets of equal size. To get our performance numbers, we averaged over 10 randomly chosen feature splits. To evaluate the performance of each split, we performed 10-fold 'inverse' cross validation, in which one fold is used as labeled data, and the other nine folds are used as unlabeled data.

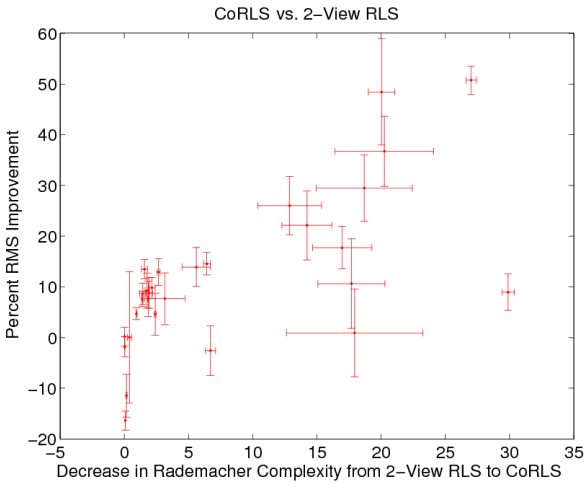

Figure 1: The percent improvement in RMS error of the CoRLS ($\lambda = 1/10$) algorithm over the 2-view RLS ($\lambda = 0$) algorithm vs. the decrease in Rademacher complexity.

For each data set, we used the CoRLS algorithm with loss functional

$$\hat{L}(f,g) = \frac{1}{2\ell}\sum_{i=1}^{\ell}\Big(\big[f(x_i)-y_i\big]^2 + \big[g(x_i)-y_i\big]^2\Big),$$

as in [12, 4]. In [4], CoRLS is compared to RLS. Here, we compare CoRLS with co-regularization parameter $\lambda = 1/10$ to the performance with $\lambda = 0$. In Figure 1, for each data set we plot the percent improvement in RMS error when going from $\lambda = 0$ to $\lambda = 1/10$ against the size of the decrease in the Rademacher complexity. The correlation between these two quantities is $r = .67$. The error bars extend two standard errors from the mean.

7 Conclusions

We have given tight bounds for the Rademacher complexity of the co-regularized hypothesis space arising from two RKHS views, as well as a generalization bound for the CoRLS algorithm. While our theorems bound the gap between training and test performance, it says nothing about the absolute performance: If neither view has good predictors, then we'll have poor performance, regardless of the generalization bound. Nevertheless, experimentally we found a correlation between improved generalization bounds and improved test performance. This may suggest that for typical parameter settings, or at least for those used in [4], reduction in Rademacher complexity is a good predictor of improved performance. We leave this question for further study, as well as the question of whether our expression for Rademacher complexity can help guide the choice of views.

Acknowledgements

We gratefully acknowledge the support of the NSF under awards DMS-0434383 and DMS-01030526.

References

[1] Maria-Florina Balcan and Avrim Blum. A PAC-style model for learning from labeled and unlabeled data. In *COLT*, pages 111–126, June 2005.

[2] Blake and C. J. Merz. UCI repository of machine learning databases, 1998.

[3] Avrim Blum and Tom M. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, pages 92–100, 1998.

[4] Ulf Brefeld, Thomas Gärtner, Tobias Scheffer, and Stefan Wrobel. Efficient co-regularised least squares regression. In *ICML*, pages 137–144, 2006.

[5] S. Dasgupta, M. L. Littman, and D. Mcallester. PAC generalization bounds for co-training. In *NIPS*, 2001.

[6] Jason D. R. Farquhar, David R. Hardoon, Hongying Meng, John S. Taylor, and Sándor Szedmák. Two view learning: SVM-2K, theory and practice. In *NIPS*, 2005.

[7] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)*. The Johns Hopkins University Press, October 1996.

[8] Rafal Latala and Krzysztof Oleszkiewicz. On the best constant in the Khintchine-Kahane inequality. *Studia Mathematica*, 109(1):101–104, 1994.

[9] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

[10] Tom Mitchell. Machine learning and extracting information from the web. NAACL Invited Talk, 2001.

[11] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, June 2004.

[12] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *ICML*, pages 824–831, 2005.