Multi-View Regression via Canonical Correlation Analysis

Sham M. Kakade[1] and Dean P. Foster[2]

[1] Toyota Technological Institute at Chicago
Chicago, IL 60637
[2] University of Pennsylvania
Philadelphia, PA 19104

Abstract. In the multi-view regression problem, we have a regression problem where the input variable (which is a real vector) can be partitioned into two different views, where it is assumed that either view of the input is sufficient to make accurate predictions — this is essentially (a significantly weaker version of) the co-training assumption for the regression problem.

We provide a semi-supervised algorithm which first uses unlabeled data to learn a norm (or, equivalently, a kernel) and then uses labeled data in a ridge regression algorithm (with this induced norm) to provide the predictor. The unlabeled data is used via canonical correlation analysis (CCA, which is a closely related to PCA for two random variables) to derive an appropriate norm over functions. We are able to characterize the intrinsic dimensionality of the subsequent ridge regression problem (which uses this norm) by the correlation coefficients provided by CCA in a rather simple expression. Interestingly, the norm used by the ridge regression algorithm is derived from CCA, unlike in standard kernel methods where a special apriori norm is assumed (i.e. a Banach space is assumed). We discuss how this result shows that unlabeled data can decrease the sample complexity.

1 Introduction

Extracting information relevant to a task in an unsupervised (or semi-supervised) manner is one of the fundamental challenges in machine learning — the underlying question is how unlabeled data can be used to improve performance. In the "multi-view" approach to semi-supervised learning [Yarowsky, 1995, Blum and Mitchell, 1998], one assumes that the input variable x can be split into two different "views" $(x^{(1)}, x^{(2)})$, such that good predictors based on each view tend to agree. Roughly speaking, the common underlying multi-view assumption is that the best predictor from either view has a low error — thus the best predictors tend to agree with each other.

There are many applications where this underlying assumption is applicable. For example, object recognition with pictures form different camera angles — we expect a predictor based on either angle to have good performance. One can even consider multi-modal views, e.g. identity recognition where the task might be to identify a person with one view being a video stream and the other an audio stream — each of these views would be sufficient to determine the identity. In NLP, an example would be a paired document corpus, consisting of a document and its translation into another language. The motivating example in Blum and Mitchell [1998] is a web-page classification task, where one view was the text in the page and the other was the hyperlink structure.

A characteristic of many of the multi-view learning algorithms [Yarowsky, 1995, Blum and Mitchell, 1998, Farquhar et al., 2005, Sindhwani et al., 2005, Brefeld et al., 2006] is to force agreement between the predictors, based on either view. The idea is to force a predictor, $h^{(1)}(\cdot)$, based on view one to agree with a predictor, $h^{(2)}(\cdot)$, based on view two, i.e. by constraining $h^{(1)}(x^{(1)})$ to usually equal $h^{(2)}(x^{(2)})$. The intuition is that the complexity of the learning problem should be reduced by eliminating hypothesis from each view that do not agree with each other (which can be done using unlabeled data).

This paper studies the multi-view, linear regression case: the inputs $x^{(1)}$ and $x^{(2)}$ are real vectors; the outputs y are real valued; the samples $((x^{(1)}, x^{(2)}), y)$ are jointly distributed; and the prediction of y is *linear* in the input x. Our first contribution is to explicitly formalize a multi-view assumption for regression. The multi-view assumption we use is a *regret* based one, where we assume that the best linear predictor from each view is roughly as good as the best linear predictor based on both views. Denote the (expected) squared loss of a prediction function $g(x)$ to be $\mathrm{loss}(g)$. More precisely, the multi-view assumption is that

$$\begin{aligned}\mathrm{loss}(f^{(1)}) - \mathrm{loss}(f) \leq \epsilon \\ \mathrm{loss}(f^{(2)}) - \mathrm{loss}(f) \leq \epsilon\end{aligned}$$

where $f^{(\nu)}$ is the *best linear predictor* based on view $\nu \in \{1, 2\}$ and f is the *best linear predictor* based on both views (so $f^{(\nu)}$ is a linear function of $x^{(\nu)}$ and f is a linear function of $x = (x^{(1)}, x^{(2)})$). This assumption implies that (only on average) the predictors must agree (shown in Lemma 1). Clearly, if the both optimal predictors $f^{(1)}$ and $f^{(2)}$ have small error, then this assumption is satisfied, though this precondition is not necessary. This (average) agreement is explicitly used in the "co-regularized" least squares algorithms of Sindhwani et al. [2005], Brefeld et al. [2006], which directly constrain such an agreement in a least squares optimization problem.

This assumption is rather weak in comparison to previous assumptions [Blum and Mitchell, 1998, Dasgupta et al., 2001, Abney, 2004]. Our assumption can be viewed as weakening the original co-training assumption (for the classification case). First, our assumption is stated in terms of expected errors only and implies only expected approximate agreement (see Lemma 1). Second, our assumption is only in terms of regret — we do *not* require that the loss of any predictor be small. Lastly, we make no further distributional assumptions (aside from a bounded second moment on the output variable), such as the commonly used, overly-stringent assumption that the distribution of the views be conditionally independent given the label [Blum and Mitchell, 1998, Dasgupta et al., 2001, Abney, 2004]. In Balcan and Blum [2006], they provide a compatibility notion which also relaxes this latter assumption, though it is unclear if this compatibility notion (defined for the classification setting) easily extends to the assumption above.

Our main result provides an algorithm and an analysis under the above multi-view regression assumption. The algorithm used can be thought of as a ridge regression algorithm with regularization based on a norm that is determined by *canonical correlation analysis* (CCA). Intuitively, CCA [Hotelling, 1935] is an unsupervised method for analyzing jointly distributed random vectors. In our setting, CCA can be performed with the unlabeled data.

We characterize the expected regret of our multi-view algorithm, in comparison to the best linear predictor, as a sum of a bias and a variance term: the bias is 4ϵ so it is small if the multi-view assumption is good; and the variance is $\frac{d}{n}$, where n is the sample size and d is the *intrinsic dimensionality* which we show to be the sum of the squares of the correlation coefficients provided by CCA. The notion of intrinsic dimensionality we use is the related to that of Zhang [2005], which provides a notion of intrinsic dimensionality for kernel methods.

An interesting aspect to our setting is that no apriori assumptions are made about any special norm over the space of linear predictions, unlike in kernel methods which *apriori* impose a Banach space over predictors. In fact, our multi-view assumption is co-ordinate free — the assumption is stated in terms of the best linear predictor for the given linear *subspaces*, which has no reference to any co-ordinate system. Furthermore, no apriori assumptions about the dimensionality of our spaces are made — thus being applicable to infinite dimensional methods, including kernel methods. In fact, kernel CCA methods have been developed in Hardoon et al. [2004].

The remainder of the paper is organized as follows. Section 2 formalizes our multi-view assumption and reviews CCA. Section 3 presents the main results, where the bias-variance tradeoff and the intrinsic dimensionality are characterized. The Discussion expands on a number of points. The foremost issue addressed is how the multi-view assumption, with unlabeled data, could potentially allow a significant reduction in the sample size. Essentially, in the high (or infinite) dimensional case, the multi-view assumption imposes a norm which could coincide with a much lower intrinsic dimensionality. In the Discussion, we also examine two related multi-view learning algorithms: the SVM-2K algorithm of Farquhar et al. [2005] and the co-regularized least squares regression algorithm of Sindhwani et al. [2005].

2 Preliminaries

This first part of this section presents the multi-view regression setting and formalizes the multi-view assumption. As is standard, we work with a distribution $D(x, y)$ over input-output pairs. To abstract away the difficulties of analyzing the use of a *random* unlabeled set sampled from $D(x)$, we instead assume that the second order statistics of x are known. The transductive setting and the fixed design setting (which we discuss later in Section 3) are cases where this assumption is satisfied. The second part of this section reviews CCA.

2.1 Regression with Multiple Views

Assume that the input space X is a subset of a real linear space, which is of either finite dimension (i.e. $X \subset \mathbb{R}^d$) or countably infinite dimension. Also assume that each $x \in X$ is in ℓ_2 (i.e. x is a squared summable sequence). In the multi-view framework, assume each x has the form $x = (x^{(1)}, x^{(2)})$, where $x^{(1)}$ and $x^{(2)}$ are interpreted as the two views of x. Hence, $x^{(1)}$ is an element of a real linear space $X^{(1)}$ and $x^{(2)}$ is in a real linear space $X^{(2)}$ (and both $x^{(1)}$ and $x^{(2)}$ are in ℓ_2). Conceptually, we should think of these spaces as being high dimensional (or countably infinite dimensional).

We also have outputs y that are in $\mathbb{R}$, along with a joint distribution $D(x, y)$ over $X \times \mathbb{R}$. We assume that the second moment of the output is bounded by 1, i.e. $\mathbb{E}[y^2|x] \leq 1$ — it is not required that y itself be bounded. No boundedness assumptions on $x \in X$ are made, since these assumptions would have no impact on our analysis as it is only the subspace defined by X that is relevant.

We also assume that our algorithm has knowledge of the second order statistics of $D(x)$, i.e. we assume that the covariance matrix of x is known. In both the transductive setting and the fixed design setting, such an assumption holds. This is discussed in more detail in Section 3.

The loss function considered for $g : X \to \mathbb{R}$ is the average squared error. More formally,

$$\text{loss}(g) = \mathbb{E}\left[(g(x) - y)^2\right]$$

where the expectation is with respect to (x, y) sampled from D. We are also interested in the losses for predictors, $g^{(1)} : X^{(1)} \to \mathbb{R}$ and $g^{(2)} : X^{(2)} \to \mathbb{R}$, based on the different views, which are just $\text{loss}(g^{(\nu)})$ for $\nu \in \{1, 2\}$.

The following assumption is made throughout the paper.

Assumption 1 *(Multi-View Assumption) Define $L(Z)$ to be the space of linear mappings from a linear space Z to the reals and define:*

$$\begin{aligned} f^{(1)} &= \text{argmin}_{g \in L(X^{(1)})} \text{loss}(g) \\ f^{(2)} &= \text{argmin}_{g \in L(X^{(2)})} \text{loss}(g) \\ f &= \text{argmin}_{g \in L(X)} \text{loss}(g) \end{aligned}$$

which exist since X is a subset of ℓ_2. The multi-view assumption is that

$$\text{loss}(f^{(\nu)}) - \text{loss}(f) \leq \epsilon$$

for $\nu \in \{1, 2\}$.

Note that this assumption makes no reference to any coordinate system or norm over the linear functions. Also, it is not necessarily assumed that the losses, themselves are small. However, if $\mathrm{loss}(f^{(\nu)})$ is small for $\nu \in \{1,2\}$, say less than ϵ, then it is clear that the above assumption is satisfied.

The following Lemma shows that the above assumption implies that $f^{(1)}$ and $f^{(2)}$ tend to agree on average.

Lemma 1. *Assumption 1 implies that:*

$$\mathbb{E}\left(f^{(1)}(x^{(1)}) - f^{(2)}(x^{(2)})\right)^2 \leq 4\epsilon$$

where the expectation is with respect to x sampled from D.

The proof is provided in the Appendix. As mentioned in the Introduction, this agreement is explicitly used in the co-regularized least squares algorithms of Sindhwani et al. [2005], Brefeld et al. [2006].

2.2 CCA and the Canonical Basis

A useful basis is that provided by CCA, which we define as the *canonical basis.*

Definition 1. *Let $B^{(1)}$ be a basis of $X^{(1)}$ and $B^{(2)}$ be a basis of $X^{(2)}$. Let $x_1^{(\nu)}, x_2^{(\nu)}, \ldots$ be the coordinates of $x^{(\nu)}$ in $B^{(\nu)}$. The pair of bases $B^{(1)}$ and $B^{(2)}$ are the* canonical bases *if the following holds (where the expectation is with respect to D):*

1. *Orthogonality Conditions:*
$$\mathbb{E}[x_i^{(\nu)} x_j^{(\nu)}] = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases}$$
2. *Correlation Conditions:*
$$\mathbb{E}\left[x_i^{(1)} x_j^{(2)}\right] = \begin{cases} \lambda_i & \text{if } i = j \\ 0 & \text{else} \end{cases}$$
where, without loss of generality, it is assumed that $1 \geq \lambda_i \geq 0$ and that
$$1 \geq \lambda_1 \geq \lambda_2 \geq \ldots$$

The i-th canonical correlation coefficient *is defined as λ_i.*

Roughly speaking, the joint covariance matrix of $x = (x^{(1)}, x^{(2)})$ in the canonical basis has a particular structured form: the individual covariance matrices of $x^{(1)}$ and $x^{(2)}$ are just identity matrices and the cross covariance matrix between $x^{(1)}$ and $x^{(2)}$ is diagonal. CCA can also be specified as an eigenvalue problem [3] (see Hardoon et al. [2004] for review).

[3] CCA finds such a basis is as follows. The correlation coefficient between two real values (jointly distributed) is defined as $\mathrm{corr}(z, z') = \frac{\mathbb{E}[zz']}{\sqrt{\mathbb{E}[z^2]\mathbb{E}[z'^2]}}$ Let $\mathit{\Pi}_a x$ be the projection operator, which projects x onto direction a. The first canonical basis vectors $b_1^{(1)} \in B^{(1)}$ and $b_1^{(2)} \in B^{(2)}$ are the unit length directions a and b which maximize $\mathrm{corr}(\mathit{\Pi}_a x^{(1)}, \mathit{\Pi}_b x^{(2)})$ and the corresponding canonical correlation coefficient λ_1 is this maximal correlation. Inductively, the next pair of directions can be found which maximize the correlation subject to the pair being orthogonal to the previously found pairs.

3 Learning

Now let us assume we have observed a training sample $T = \{(x_m^{(\nu)}, y_m)\}_{m=1}^n$ of size n from a view ν, where the samples drawn independently from D. We also assume that our algorithm has access to the covariance matrix of x, so that the algorithm can construct the canonical basis.

Our goal is to construct an estimator $\widehat{f}^{(\nu)}$ of $f^{(\nu)}$ — recall $f^{(\nu)}$ is the best linear predictor using only view ν — such that the regret

$$\text{loss}(\widehat{f}^{(\nu)}) - \text{loss}(f^{(\nu)})$$

is small.

Remark 1. (The Transductive and Fixed Design Setting) There are two natural settings where this assumption of knowledge about the second order statistics of x holds — the *random transductive* case and the *fixed design* case. In both cases, X is a known finite set. In the random transductive case, the distribution D is assumed to be uniform over X, so each x_m is sampled uniformly from X and each y_m is sampled from $D(y|x_m)$. In the fixed design case, assume that each $x \in X$ appears exactly once in T and again y_m is sampled from $D(y|x_m)$. The fixed design case is commonly studied in statistics and is also referred to as signal reconstruction.[4] The covariance matrix of x is clearly known in both cases.

3.1 A Shrinkage Estimator (via Ridge Regression)

Let the representation of our estimator $\widehat{f}^{(\nu)}$ in the canonical basis $B^{(\nu)}$ be

$$\widehat{f}^{(\nu)}(x^{(\nu)}) = \sum_i \widehat{\beta}_i^{(\nu)} x_i^{(\nu)} \tag{1}$$

where $x_i^{(\nu)}$ is the i-th coordinate in $B^{(\nu)}$. Define the canonical shrinkage estimator of $\widehat{\beta}^{(\nu)}$ as:

$$\widehat{\beta}_i^{(\nu)} = \lambda_i \widehat{\mathbb{E}}[x_i y] \equiv \frac{\lambda_i}{n} \sum_m x_{m,i}^{(\nu)} y_m \tag{2}$$

Intuitively, the shrinkage by λ_i down-weights directions that are less correlated with the other view. In the extreme case, this estimator ignores the uncorrelated coordinates, those where $\lambda_i = 0$. The following remark shows how this estimator has a natural interpretation in the fixed design setting — it is the result of ridge regression with a specific norm (induced by CCA) over functions in $L(X^{(\nu)})$.

[4] In the fixed design case, one can view each $y_m = f(x_m) + \eta$, where η is 0 mean noise so $f(x_m)$ is the conditional mean. After observing a sample $\{(x_m^{(\nu)}, y_m)\}_{m=1}^{|X|}$ for *all* $x \in X$ (so $n = |X|$), the goal is to reconstruct $f(\cdot)$ accurately.

Remark 2. (Canonical Ridge Regression). We now specify a ridge regression algorithm for which the shrinkage estimator is the solution. Define the *canonical norm* for a linear function in $L(X^{(\nu)})$ as follows: using the representation of $\widehat{f}^{(\nu)}$ in $B^{(\nu)}$ as defined in Equation 1, the canonical norm of $\widehat{f}^{(\nu)}$ is defined as:

$$||\widehat{f}^{(\nu)}||_{\text{CCA}} = \sqrt{\sum_i \frac{1-\lambda_i}{\lambda_i} \left(\widehat{\beta}_i^{(\nu)}\right)^2} \tag{3}$$

where we overload notation and write $||\widehat{f}^{(\nu)}||_{\text{CCA}} = ||\widehat{\beta}^{(\nu)}||_{\text{CCA}}$. Hence, functions which have large weights in the less correlated directions (those with small λ_i) have larger norms. Equipped with this norm, the functions in $L(X^{(\nu)})$ define a Banach space. In the fixed design setting, the ridge regression algorithm with this norm chooses the $\widehat{\beta}^{(\nu)}$ which minimizes:

$$\frac{1}{|X|}\sum_{m=1}^{|X|}\left(y_m - \widehat{\beta}^{(\nu)} \cdot x_m^{(\nu)}\right)^2 + ||\widehat{\beta}^{(\nu)}||_{\text{CCA}}^2$$

Recall, that in the fixed design setting, we have a training example for each $x \in X$, so the sum is over all $x \in X$.

It is straightforward to show (by using orthogonality) that the estimator which minimizes this loss is the canonical shrinkage estimator defined above. In the more general transductive case, it is not quite this estimator, since the sampled points $\{x_m^{(\nu)}\}_m$ may not be orthogonal in the training sample (they are only orthogonal when summed over all $x \in X$). However, in this case, we expect that the estimator provided by ridge regression is approximately equal to the shrinkage estimator.

We now state the first main theorem.

Theorem 1. *Assume that $\mathbb{E}[y^2|x] \leq 1$ and that Assumption 1 holds. Let $\widehat{f}^{(\nu)}$ be the estimator constructed with the canonical shrinkage estimator (Equation 2) on training set T. For $\nu \in 1, 2$, then*

$$\mathbb{E}_T[\text{loss}(\widehat{f}^{(\nu)})] - \text{loss}(f^{(\nu)}) \leq 4\epsilon + \frac{\sum_i \lambda_i^2}{n}$$

where expectation is with respect to the training set T sampled according to D^n.

We comment on obtaining high probability bounds in the Discussion. The proof (presented in Section 3.3) shows that the 4ϵ results from the bias in the algorithm and $\frac{\sum_i \lambda_i^2}{n}$ results from the variance. It is natural to interpret $\sum_i \lambda_i^2$ as the intrinsic dimensionality.

Note that Assumption 1 implies that:

$$\mathbb{E}_T[\text{loss}(\widehat{f}^{(\nu)})] - \text{loss}(f) \leq 5\epsilon + \frac{\sum_i \lambda_i^2}{n}$$

where the comparison is to the best linear predictor f over both views.

Remark 3. (Intrinsic Dimensionality) Let $\widehat{\beta}^{(\nu)}$ be a linear estimator in the vector of sampled outputs, $Y = (y_1, y_2, \ldots y_m)$. Note that the previous thresholded estimator is such a linear estimator (in the fixed design case). We can write $\widehat{\beta}^{(\nu)} = PY$ where P is a linear operator. Zhang [2005] defines $tr(P^T P)$ as the intrinsic dimensionality, where $tr(\cdot)$ is the trace operator. This was motivated by the fact that in the fixed design setting the error drops as $\frac{tr(P^T P)}{n}$, which is bounded by $\frac{d}{n}$ in a finite dimensional space. Zhang [2005] then goes on to analyze the intrinsic dimensionality of kernel methods in the random design setting (obtaining high probability bounds). In our setting, the sum $\sum_i \lambda_i^2$ is precisely this trace, as P is a diagonal matrix with entries λ_i.

3.2 A (Possibly) Lower Dimensional Estimator

Consider the thresholded estimator:

$$\widehat{\beta}_i^{(\nu)} = \begin{cases} \widehat{\mathbb{E}}[x_i y] & \text{if } \lambda_i \geq 1 - \sqrt{\epsilon} \\ 0 & \text{else} \end{cases} \tag{4}$$

where again $\widehat{\mathbb{E}}[x_i y]$ is the empirical expectation $\frac{1}{n} \sum_m x_{m,i}^{(\nu)} y_m$. This estimator uses an unbiased estimator of $\beta_i^{(\nu)}$ for those i with large λ_i and thresholds to 0 for those i with small λ_i. Hence, the estimator lives in a finite dimensional space (determined by the number of λ_i which are greater than $1 - \sqrt{\epsilon}$).

Theorem 2. *Assume that $\mathbb{E}[y^2|x] \leq 1$ and that Assumption 1 holds. Let d be the number of λ_i for which $\lambda_i \geq 1 - \sqrt{\epsilon}$. Let $\widehat{f}^{(\nu)}$ be the estimator constructed with the threshold estimator (Equation 4) on training set T. For $\nu \in 1, 2$, then*

$$\mathbb{E}_T[\text{loss}(\widehat{f}^{(\nu)})] - \text{loss}(f^{(\nu)}) \leq 4\sqrt{\epsilon} + \frac{d}{n}$$

where expectation is with respect to the training set T sampled according to D^n.

Essentially, the above increases the bias to $4\sqrt{\epsilon}$ and (potentially) decreases the variance. Such a bound may be useful if we desire to explicitly keep $\widehat{\beta}^{(\nu)}$ in a lower dimensional space — in contrast, the explicit dimensionality of the shrinkage estimator could be as large as $|X|$.

3.3 The Bias-Variance Tradeoff

This section provides lemmas for the proofs of the previous theorems. We characterize the bias-variance tradeoff in this error analysis. First, a key technical lemma is useful, for which the proof is provided in the Appendix.

Lemma 2. *Let the representation of the best linear predictor $f^{(\nu)}$ (defined in Assumption 1) in the canonical basis $B^{(\nu)}$ be*

$$f^{(\nu)}(x^{(\nu)}) = \sum_i \beta_i^{(\nu)} x_i^{(\nu)} \tag{5}$$

Assumption 1 implies that

$$\sum_i (1-\lambda_i)\left(\beta_i^{(\nu)}\right)^2 \leq 4\epsilon$$

for $\nu \in \{1,2\}$.

This lemma shows how the weights (of an optimal linear predictor) cannot be too large in coordinates with small canonical correlation coefficients. This is because for those coordinates with small λ_i, the corresponding β_i must be small enough so that the bound is not violated. This lemma provides the technical motivation for our algorithms.

Now let us review some useful properties of the square loss. Using the representations of $f^{(\nu)}$ and f defined in Equations 1 and 5, a basic fact for the square loss with linear predictors is that

$$\text{loss}(\widehat{f}^{(\nu)}) - \text{loss}(f^{(\nu)}) = ||\widehat{\beta}^{(\nu)} - \beta^{(\nu)}||_2^2$$

where $||x||_2 = \sqrt{\sum_i x_i^2}$. The expected regret can be decomposed as follows:

$$\mathbb{E}_T\left[||\widehat{\beta}^{(\nu)} - \beta^{(\nu)}||_2^2\right] = ||\mathbb{E}_T[\widehat{\beta}^{(\nu)}] - \beta^{(\nu)}||_2^2 + \mathbb{E}_T\left[||\widehat{\beta}^{(\nu)} - \mathbb{E}_T[\widehat{\beta}^{(\nu)}]||_2^2\right] \tag{6}$$

$$= ||\mathbb{E}_T[\widehat{\beta}^{(\nu)}] - \beta^{(\nu)}||_2^2 + \sum_i \text{Var}(\widehat{\beta}_i^{(\nu)}) \tag{7}$$

where the first term is the bias and the second is the variance.

The proof of Theorems 1 and 2 follow directly from the next two lemmas.

Lemma 3. *(Bias-Variance for the Shrinkage Estimator) Under the preconditions of Theorem 1, the bias is bounded as:*

$$||\mathbb{E}_T[\widehat{\beta}^{(\nu)}] - \beta^{(\nu)}||_2^2 \leq 4\epsilon$$

and the variance is bounded as:

$$\sum_i Var(\widehat{\beta}_i^{(\nu)}) \leq \frac{\sum_i \lambda_i^2}{n}$$

Proof. It is straightforward to see that:

$$\beta_i^{(\nu)} = \mathbb{E}[x_i y]$$

which implies that

$$\mathbb{E}_T[\widehat{\beta}_i^{(\nu)}] = \lambda_i \beta_i^{(\nu)}$$

Hence, for the bias term, we have:

$$\begin{aligned}
||\mathbb{E}_T[\widehat{\beta}^{(\nu)}] - \beta^{(\nu)}||_2^2 &= \sum_i (1-\lambda_i)^2 (\beta_i^{(\nu)})^2 \\
&\leq \sum_i (1-\lambda_i)(\beta_i^{(\nu)})^2 \\
&\leq 4\epsilon
\end{aligned}$$

We have for the variance

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta}_i^{(\nu)}) &= \frac{\lambda_i^2}{n} \mathrm{Var}(x_i^{(\nu)} y) \\
&\leq \frac{\lambda_i^2}{n} \mathbb{E}[(x_i^{(\nu)} y)^2] \\
&= \frac{\lambda_i^2}{n} \mathbb{E}[(x_i^{(\nu)})^2 \mathbb{E}[y^2|x]] \\
&\leq \frac{\lambda_i^2}{n} \mathbb{E}[(x_i^{(\nu)})^2] \\
&= \frac{\lambda_i^2}{n}
\end{aligned}$$

The proof is completed by summing over i. □

Lemma 4. *(Bias-Variance for the Thresholded Estimator) Under the preconditions of Theorem 2, the bias is bounded as:*

$$||\mathbb{E}_T[\widehat{\beta}^{(\nu)}] - \beta^{(\nu)}||_2^2 \leq 4\sqrt{\epsilon}$$

and the variance is bounded as:

$$\sum_i \mathit{Var}(\widehat{\beta}_i^{(\nu)}) \leq \frac{d}{n}$$

Proof. For those i such that $\lambda_i \geq 1 - \sqrt{\epsilon}$,

$$\mathbb{E}_T[\widehat{\beta}_i^{(\nu)}] = \beta_i^{(\nu)}$$

Let j be the index at which the thresholding begins to occur, i.e. it is the smallest integer such that $\lambda_j < 1 - \sqrt{\epsilon}$. Using that for $i \geq j$, we have $1 < (1-\lambda_j)/\sqrt{\epsilon} \leq$

$(1-\lambda_i)/\sqrt{\epsilon}$, so the bias can be bounded as follows:

$$\begin{aligned}
||\mathbb{E}_T[\widehat{\beta}^{(\nu)}] - \beta^{(\nu)}||_2^2 &= \sum_i \left(\mathbb{E}_T[\widehat{\beta}_i^{(\nu)}] - \beta_i^{(\nu)}\right)^2 \\
&= \sum_{i \geq j} (\beta_i^{(\nu)})^2 \\
&\leq \sum_{i \geq j} \frac{1-\lambda_i}{\sqrt{\epsilon}} (\beta_i^{(\nu)})^2 \\
&\leq \frac{1}{\sqrt{\epsilon}} \sum_i (1-\lambda_i)(\beta_i^{(\nu)})^2 \\
&\leq 4\sqrt{\epsilon}
\end{aligned}$$

where the last step uses Lemma 2.

Analogous to the previous proof, for each $i < j$, we have:

$$\mathrm{Var}(\widehat{\beta}_i^{(\nu)}) \leq 1$$

and there are d such i. □

4 Discussion

Why does unlabeled data help? Theorem 1 shows that the regret drops at a uniform rate (down to ϵ). This rate is the intrinsic dimensionality, $\sum_i \lambda_i^2$, divided by the sample size n. Note that this intrinsic dimensionality is only a property of the input distribution. Without the multi-view assumption (or working in the single view case), the rate at which our error drops is governed by the extrinsic dimensionality of x, which could be large (or countably infinite), making this rate very slow without further assumptions. It is straightforward to see that the intrinsic dimensionality is no greater than the extrinsic dimensionality (since λ_i is bounded by 1), though it could be much less. The knowledge of the covariance matrix of x allows us to compute the CCA basis and construct the shrinkage estimator which has the improved converge rate based on the intrinsic dimensionality. Such second order statistical knowledge can be provided by the unlabeled data, such as in the transductive and fixed design settings.

Let us compare to a ridge regression algorithm (in the single view case), where one apriori chooses a norm for regularization (such as an RKHS norm imposed by a kernel). As discussed in Zhang [2005], this regularization governs the bias-variance tradeoff. The regularization can significantly decrease the variance — the variance drops as $\frac{d}{n}$ where d is a notion of intrinsic dimensionality defined in Zhang [2005]. However, the regularization also biases the algorithm to predictors with small norm — there is no apriori reason that there exists a good predictor with a bounded norm (under the pre-specified norm). In order to obtain a reasonable convergence rate, it must also be the case that the best predictor (or a good one) has a small norm under our pre-specified norm.

In contrast, in the multi-view case, the multi-view assumption implies that the bias is bounded — recall that Lemma 3 showed that the bias was bounded by 4ϵ. Essentially, our proof shows that the bias induced by using the special norm induced by CCA (in Equation 3) is small.

Now it may be the case that we have apriori knowledge of what a good norm is. However, learning the norm (or learning the kernel) is an important open question. The multi-view setting provides one solution to this problem.

Can the bias be decreased to 0 asymptotically? Theorem 1 shows that the error drops down to 4ϵ for large n. It turns out that we can not drive this bias to 0 asymptotically without further assumptions, as the input space could be infinite dimensional.

On obtaining high probability bounds. Clearly, stronger assumptions are needed than just a bounded second moment to obtain high probability bounds with concentration properties. For the fixed design setting, if y is bounded, then it is straightforward to obtain high probability bounds through standard Chernoff arguments. For the random transductive case, this assumption is not sufficient — this is due to the additional randomness from x. Note that we cannot artificially impose a bound on x as the algorithm only depends on the subspace spanned by X, so upper bounds have no meaning — note the algorithm scales X such that it has an identity covariance matrix (e.g. $E[x_i^2] = 1$). However, if we have a higher moment bound, say on the ratio of $E[x_i^4]/E[x_i^2]$, then one could use the Bennett bound can be used to obtain data dependent high probability bounds, though providing these is beyond the scope of this paper.

Related work. The most closely related multi-view learning algorithms are the SVM-2K algorithm of Farquhar et al. [2005] and the co-regularized least squares regression algorithm of Sindhwani et al. [2005]. Roughly speaking, both of these algorithms try to find two hypothesis — $h^{(1)}(\cdot)$, based on view one, and $h^{(2)}(\cdot)$, based on view two — which both have low training error and which tend to agree with each other on unlabeled error, where the latter condition is enforced by constraining $h^{(1)}(x^{(1)})$ to usually equal $h^{(2)}(x^{(2)})$ on an unlabeled data set.

The SVM-2K algorithm considers a classification setting and the algorithm attempts to force agreement between the two hypothesis with slack variable style constraints, common to SVM algorithms. While this algorithm is motivated by kernel CCA and SVMs, the algorithm does not directly use kernel CCA, in contrast to our algorithm, where CCA naturally provides a coordinate system. The theoretical analysis in [Farquhar et al., 2005] argues that the Rademacher complexity of the hypothesis space is reduced due to the agreement constraint between the two views.

The multi-view approach to regression has been previously considered in Sindhwani et al. [2005]. Here, they specify a co-regularized least squares regression algorithm, which is a ridge regression algorithm with an additional penalty term which forces the two predictions, from both views, to agree. A theoretical analysis of this algorithm is provided in Rosenberg and Bartlett [2007], which shows that the Rademacher complexity of the hypothesis class is reduced by forcing agreement.

Both of these previous analysis do not explicitly state a multi-view assumption, so it hard to directly compare the results. In our setting, the multi-view regret is explicitly characterized by ϵ. In a rather straightforward manner (without appealing to Rademacher complexities), we have shown that the rate at which the regret drops to 4ϵ is determined by the intrinsic dimensionality. Furthermore, both of these previous algorithms use an apriori specified norm over their class of functions (induced by an apriori specified kernel), and the Rademacher complexities (which are used to bound the convergence rates) depend on this norm. In contrast, our framework assumes no norm — the norm over functions is imposed by the correlation structure between the two views.

We should also note that their are close connections to those unsupervised learning algorithms which attempt to maximize relevant information. The Imax framework of Becker and Hinton [1992], Becker [1996] attempts to maximize information between two views $x^{(1)}$ and $x^{(2)}$, for which CCA is a special case (in a continuous version). Subsequently, the information bottleneck provided a framework for capturing the mutual information between two signals [Tishby et al., 1999]. Here, the goal is to compress a signal $x^{(1)}$ such that it captures relevant information about another signal $x^{(2)}$. The framework here is unsupervised as there is no specific supervised task at hand. For the case in which the joint distribution of $x^{(1)}$ and $x^{(2)}$ is Gaussian, Chechik et al. [2003] completely characterizes the compression tradeoffs for capturing the mutual information between these two signals — CCA provides the coordinate system for this compression.

In our setting, we do not explicitly care about the mutual information between $x^{(1)}$ and $x^{(2)}$ — performance is judged only by performance at the task at hand, namely our loss when predicting some other variable y. However, as we show, it turns out that these unsupervised mutual information maximizing algorithms provide appropriate intuition for multi-view regression, as they result in CCA as a basis.

Acknowledgements

We thank the anonymous reviewers for their helpful comments.

References

Steven Abney. Understanding the yarowsky algorithm. *Comput. Linguist.*, 30(3):365–395, 2004. ISSN 0891-2017.

Maria-Florina Balcan and Avrim Blum. A pac-style model for learning from labeled and unlabeled data. In *Semi-Supervised Learning*, pages 111–126. MIT Press, 2006.

S. Becker. Mutual information maximization: Models of cortical self-organization. *Network: Computation in Neural Systems*, 1996.

Suzanna Becker and Geoffrey E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, January 1992. doi: 10.1038/355161a0.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, New York, NY, USA, 1998. ACM Press. ISBN 1-58113-057-0.

Ulf Brefeld, Thomas Gartner, Tobias Scheffer, and Stefan Wrobel. Efficient co-regularised least squares regression. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 137–144, New York, NY, USA, 2006. ACM Press. ISBN 1-59593-383-2.

G. Chechik, A. Globerson, N. Tishby, and Y. Weiss. Information bottleneck for gaussian variables, 2003. URL citeseer.ist.psu.edu/article/chechik03information.html.

Sanjoy Dasgupta, Michael L. Littman, and David Mcallester. Pac generalization bounds for co-training, 2001.

Jason D. R. Farquhar, David R. Hardoon, Hongying Meng, John Shawe-Taylor, and Sándor Szedmák. Two view learning: Svm-2k, theory and practice. In *NIPS*, 2005.

David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664, 2004. ISSN 0899-7667.

H. Hotelling. The most predictable criterion. *Journal of Educational Psychology*, 1935.

D. Rosenberg and P Bartlett. The rademacher complexity of co-regularized kernel classes. submitted. *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, 2007.

V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularized approach to semi-supervised learning with multiple views. *Proceedings of the ICML Workshop on Learning with Multiple Views*, 2005.

N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999. URL citeseer.ist.psu.edu/tishby99information.html.

David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196, Morristown, NJ, USA, 1995. Association for Computational Linguistics.

Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.*, 17(9):2077–2098, 2005. ISSN 0899-7667.

5 Appendix

We now provide the proof of Lemma 1

Proof. (of Lemma 1). Let $\beta^{(\nu)}$ be the weights for $f^{(\nu)}$ and let β be the weights of f in some basis. Let $\beta^{(\nu)} \cdot x^{(\nu)}$ and $\beta \cdot x$ be the representation of $f^{(\nu)}$ and f in this basis. By Assumption 1

$$\begin{aligned}\epsilon &\geq \mathbb{E}(\beta^{(\nu)} \cdot x^{(\nu)} - y)^2 - \mathbb{E}(\beta \cdot x - y)^2 \\ &= \mathbb{E}(\beta^{(\nu)} \cdot x^{(\nu)} - \beta \cdot x + \beta \cdot x - y)^2 - \mathbb{E}(\beta \cdot x - y)^2 \\ &= \mathbb{E}(\beta^{(\nu)} \cdot x^{(\nu)} - \beta \cdot x)^2 - 2\mathbb{E}[(\beta^{(\nu)} \cdot x^{(\nu)} - \beta \cdot x)(\beta \cdot x - y)]\end{aligned}$$

Now the "normal equations" for β (the first derivative conditions for the optimal linear predictor β) states that for each i:

$$\mathbb{E}[x_i(\beta \cdot x - y)] = 0$$

where x_i is the i component of x. This implies that both

$$\begin{aligned}\mathbb{E}[\beta \cdot x(\beta \cdot x - y)] &= 0\\ \mathbb{E}[\beta^{(\nu)} \cdot x^{(\nu)}(\beta \cdot x - y)] &= 0\end{aligned}$$

where the last equation follows since $x^{(\nu)}$ has components in x.

Hence,

$$\mathbb{E}[(\beta^{(\nu)} \cdot x^{(\nu)} - \beta \cdot x)(\beta \cdot x - y)] = 0$$

and we have shown that:

$$\begin{aligned}\mathbb{E}(\beta^{(1)} \cdot x^{(1)} - \beta \cdot x)^2 &\le \epsilon\\ \mathbb{E}(\beta^{(2)} \cdot x^{(2)} - \beta \cdot x)^2 &\le \epsilon\end{aligned}$$

The triangle inequality states that:

$$\begin{aligned}&\mathbb{E}(\beta^{(1)} \cdot x^{(1)} - \beta^{(2)} \cdot x)^2\\ &\le \left(\sqrt{\mathbb{E}(\beta^{(1)} \cdot x^{(1)} - \beta \cdot x)^2} + \sqrt{\mathbb{E}(\beta^{(2)} \cdot x^{(2)} - \beta \cdot x)^2}\right)^2\\ &\le (2\sqrt{\epsilon})^2\end{aligned}$$

which completes the proof. □

Below is the proof of Lemma 2.

Proof. (of Lemma 2) From Lemma 1, we have:

$$\begin{aligned}4\epsilon &\ge \mathbb{E}\left[(\beta^{(1)} \cdot x^{(1)} - \beta^{(2)} \cdot x^{(2)})^2\right]\\ &= \sum_i \left((\beta_i^{(1)})^2 + (\beta_i^{(2)})^2 - 2\lambda_i \beta_i^{(1)} \beta_i^{(2)}\right)\\ &= \sum_i \left((1-\lambda_i)(\beta_i^{(1)})^2 + (1-\lambda_i)(\beta_i^{(2)})^2 + \lambda_i((\beta_i^{(1)})^2 + (\beta_i^{(2)})^2 - 2\beta_i^{(1)}\beta_i^{(2)})\right)\\ &= \sum_i \left((1-\lambda_i)(\beta_i^{(1)})^2 + (1-\lambda_i)(\beta_i^{(2)})^2 + \lambda_i(\beta_i^{(1)} - \beta_i^{(2)})^2\right)\\ &\ge \sum_i \left((1-\lambda_i)(\beta_i^{(1)})^2 + (1-\lambda_i)(\beta_i^{(2)})^2\right)\\ &\ge \sum_i (1-\lambda_i)(\beta_i^{(\nu)})^2\end{aligned}$$

where the last step holds for either $\nu = 1$ or $\nu = 2$. □