An Information Theoretic Framework for Multi-view Learning

Karthik Sridharan and **Sham M. Kakade**
Toyota Technological Institute at Chicago
{karthik, sham}@tti-c.org

Abstract

In the multi-view learning paradigm, the input variable is partitioned into two different views X_1 and X_2 and there is a target variable Y of interest. The underlying assumption is that either view *alone* is sufficient to predict the target Y accurately. This provides a natural semi-supervised learning setting in which unlabeled data can be used to eliminate hypothesis from either view, whose predictions tend to disagree with predictions based on the other view.

This work explicitly formalizes an information theoretic, multi-view assumption and studies the multi-view paradigm in the PAC style semi-supervised framework of Balcan and Blum [2006]. Underlying the PAC style framework is that an *incompatibility function* is assumed to be known — roughly speaking, this incompatibility function is a means to score how good a function is based on the unlabeled data alone. Here, we show how to derive incompatibility functions for certain loss functions of interest, so that minimizing this incompatibility over unlabeled data helps reduce expected loss on future test cases. In particular, we show how the class of empirically successful co-regularization algorithms fall into our framework and provide performance bounds (using the results in Rosenberg and Bartlett [2007], Farquhar et al. [2005]).

We also provide a normative justification for canonical correlation analysis (CCA) as a dimensionality reduction technique. In particular, we show (for strictly convex loss functions of the form $\ell(w \cdot x, y)$) that we can first use CCA as dimensionality reduction technique and (if the multi-view assumption is satisfied) this projection does not throw away much predictive information about the target Y — the benefit being that subsequent learning with a labeled set need only work in this lower dimensional space.

1 Introduction

The "multi-view" approach to learning has been receiving increasing attention as a paradigm for semi-supervised learning. The implicit assumption is that either view *alone* has sufficient information about the target Y. The basic intuition as to why this assumption is helpful is that the complexity of the learning problem could be reduced by eliminating hypothesis from each view that tend not to agree with each other, which, crucially, can be done using unlabeled data.

There are many natural applications for which this assumption is applicable. For example, consider a setting where it is easy to obtain pictures of objects from different camera angles and say our supervised task is one of object recognition. Intuitively, we can think of unlabeled data as providing examples of viewpoint invariance. 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, and the supervised task could be predicting some high level property of the document. 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 hyper-link structure.

This work explicitly formalizes a general information theoretic multi-view assumption. Based on this assumption, we seek to understand the reduction in label complexity from using unlabeled data. There are two natural classes of algorithms in the literature which can be considered multi-view algorithms. These classes are the co-regularization algorithms and algorithms based on CCA. For the former, we analyze the co-regularization algorithms of Sindhwani et al. [2005], Brefeld et al. [2006] (and the related SVM-2K algorithm of Farquhar et al. [2005]) in a generalization of the PAC style semi-supervised framework of Balcan and Blum [2006]. Technically, this PAC model is for the $0/1$ loss, but we generalize the framework to arbitrary loss functions. For the latter class of algorithms, we generalize the CCA results in Kakade and Foster [2007] to show how CCA can be used for dimensionality reduction, when dealing with convex loss functions (under linear prediction). In the Discussion, we present a practical answer to the open problem presented in Balcan and Blum [2007] (presented at COLT 2007) using

co-regularization algorithms, under the theory of surrogate loss functions [Bartlett et al., 2006], and we also discuss the connection to the Information Bottleneck method of Tishby et al. [1999].

In the remainder of the Introduction, we present our setting and main information theoretic assumption, and then summarize our contributions and related work.

1.1 A Multi-View Assumption

In the (multi-view) semi-supervised setting, we assume that we have n labeled examples $S = \{(x_1^i, x_2^i, y^i)\}_{i=1}^n$ and m unlabeled examples $U = \{(x_1^i, x_2^i)\}_{i=n+1}^{n+m}$, where $y_i \in \mathcal{Y}$ and $x_v^i \in \mathcal{X}_v$ for $v \in \{1, 2\}$, which are both sampled in an i.i.d. manner from some unknown underlying joint distribution (typically $m >> n$). In particular, the joint underlying distribution is over $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}$. As usual, the goal is to predict Y, as measured with respect to some known loss function.

Information theory provides the natural language to state an assumption for multi-view learning. In particular, the conditional mutual information $I(A : B|C)$ (between random outcomes A and B conditioned on C) measures how much information is shared between A and B conditioned on already knowing C, which can be viewed as how much knowing A reduces our uncertainty of B, conditioned on already knowing C. We now state our first main assumption.

Assumption 1 (Multi-View Assumption) *There exists an* $\epsilon_{\text{info}} > 0$ *such that*

$$I(Y : X_2|X_1) \leq \epsilon_{\text{info}}$$

and

$$I(Y : X_1|X_2) \leq \epsilon_{\text{info}}$$

Let us try to get an intuitive feel for this assumption. The assumption states that (on average) if we already knew X_1 then there is little more information that we could gain about Y from learning X_2 (and vice-versa) — this small potential gain is quantified by ϵ_{info}. Hence, we can think of this assumption as stating that both X_1 and X_2 are (approximately) redundant with regards to their information about Y.

Let us examine how the compatibility assumption made in the co-training case [Blum and Mitchell, 1998], where $Y \in \{0, 1\}$, is related to this assumption. Here, it was assumed that a perfect prediction of Y is possible using the knowledge of either view alone. This implies the above conditions are satisfied with $\epsilon_{\text{info}} = 0$, since conditioned on either view, the target Y is already known (so there is no possible reduction in uncertainty with knowledge from the remaining view).

However, note that under this assumption, neither view need accurately predict the target, just that they both carry (roughly) the same information about the target. Hence, the assumption is well suited to situations with noise. In fact, even if $\epsilon_{\text{info}} = 0$, there need not exist perfect predictions of the target — though for this case we would expect that the optimal predictions should perfectly agree (as they carry the same information about Y), a point which we return to.

The work in Blum and Mitchell [1998] also introduced a further conditional independence assumption, which states that X_1 and X_2 are independent conditioned on the knowledge of Y. The work of Dasgupta et al. [2001], Abney [2004] shows how unreasonably strong this extra assumption is, with regards to classification. In our work, we make no further assumptions on the underlying data distribution.

1.2 Co-Regularization

There is a recent class of algorithms which control model complexity in the two view setting by *co-regularizing* [Sindhwani et al., 2005, Brefeld et al., 2006]. A related algorithm is the two view SVM-2K algorithm of Farquhar et al. [2005]. These class of algorithms all have demonstrated empirical successes. The question we seek to understand is how unlabeled data improves the performance of these algorithms.

These co-regularization algorithms add an additional regularizer which penalizes using functions from either view which tend to disagree. The (kernelized) algorithm of Sindhwani et al. [2005], Brefeld et al. [2006] is to find two predictors f_1 and f_2 (where $f_1 : \mathcal{X}_1 \to \mathcal{Y}$ and $f_2 : \mathcal{X}_2 \to \mathcal{Y}$) which minimize the following co-regularized loss:

$$\frac{1}{2}(\widehat{E}_S[\ell(f_1(x_1), y)] + \widehat{E}_S[\ell(f_2(x_2), y)])$$
$$\lambda\|f_1\|_K^2 + \lambda\|f_2\|_K^2 + \lambda_{co}\widehat{E}_U(f_1(x_1) - f_2(x_2))^2 \qquad (1)$$

where $\|\cdot\|_K$ is a pre-specified norm over functions; $\widehat{E}_S$ and $\widehat{E}_U$ are empirical averages with respect to the labeled and unlabeled sets S and U, respectively; and ℓ is some convex loss (such as the hinge loss or squared loss). The last term is the co-regularizer. Note that if $\lambda_{co} = 0$ then this problem just reduces to solving two independent (regularized) problems. The SVM-2K algorithm of Farquhar et al. [2005] is similar — it essentially imposes an agreement constraint into the SVM objective function, based on the L_1 norm (which allows for an efficient implementation).

Rosenberg and Bartlett [2007] provide generalization bounds for co-regularization (using a co-regularizer that is the square loss) in terms of Rademacher complexities. Farquhar et al. [2005] also provide generalization bounds (again using Rademacher complexities) for the SVM-2K algorithm. These bounds characterize how much the complexity class of the hypothesis space decreases with the co-regularization. We can view these bounds as characterizing how much the variance of the algorithm decreases. In particular, as λ_{co} increases, this has the effect of decreasing the variance (as a harder constraint is being imposed). While these are valid generalization bounds (which compare the empirical expectation of a predictor to the true expectation), they do not address the bias issue of how performance could decrease as λ_{co} is increased too much. In particular, as λ_{co} is increased, the algorithm is not as free to use certain hypothesis (which we can think of as the bias). Roughly speaking, these previous multi-view results quantify how model complexity is reduced, but they do not specify *why* this is reasonable to do. Hence, to understand how unlabeled data could improve performance, we must characterize how much the co-regularization effects this bias-variance trade-off.

We address these issues under the recent PAC framework for semi-supervised learning of Balcan and Blum [2006] — though we generalize the setting for arbitrary loss functions (Balcan and Blum [2006] only considered the $0/1$ loss).

Their framework assumes an *incompatibility* function — a function which scores how good hypothesis are just based on the underlying data distribution. They provide a general framework for characterizing how such an incompatibility function can reduce the need for labeled samples. Intuitively, one can view the co-regularizer as an incompatibility function, as it is scoring hypothesis based on unlabeled data — if a pair of hypothesis disagree strongly under the co-regularizer it is unlikely that they would be good predictors.

One of our main contributions for analyzing these co-regularization algorithms is that we show how the incompatibility function is really a derived property of the loss function — the incompatibility function needs to satisfy a rather mild inverse Lipschitz condition. Under relatively general conditions, incompatibility functions can be derived for many loss functions of interest — we provide examples for the (regularized) hinge loss, the square loss, for the $0/1$ loss, and for strictly convex losses. Interestingly (and rather subtly), our incompatibility function for the $0/1$ loss makes use of Tsybakov's noise condition.

We then explicitly use the Rademacher bounds in Rosenberg and Bartlett [2007], Farquhar et al. [2005] to provide performance bounds under the multi-view assumption. These bounds characterize the bias-variance trade-off. We explicitly quantify how to set the co-regularization parameter λ_{co} in terms of ϵ_{info}, showing that an appropriate setting of λ_{co} is $O(1/\sqrt{\epsilon_{\text{info}}})$. In particular, this shows it is appropriate for $\lambda_{co} \to \infty$ as $\epsilon_{\text{info}} \to 0$, i.e. when the information theoretic assumption is as sharp as possible, we are permitted to co-regularize as hard as possible (without introducing any bias). For this case, the co-regularization algorithms obtain their maximal reduction in variance.

1.3 Dimensionality Reduction

While PCA is the time-honoured and simplest dimensionality reduction technique, there are few normative reasons as to why this technique is appropriate. The typical justification is that the top k principal directions are those which best reconstruct the data, in a mean squared sense. One common criticism of this oft used justification is that a rescaling of the data could change the outcome of PCA.

Canonical Correlation Analysis (CCA) [Hotelling, 1935] — like PCA but for the two view setting — also serves as a rather general and widely used dimensionality reduction technique. Roughly speaking, it uses the cross-correlation matrix between the two views to find the canonical directions — those directions which are most correlated (in a normalized sense) between the views. As a dimensionality reduction procedure, one can take the top k CCA directions which, roughly speaking, preserves the most correlated coordinates. However, unlike PCA, CCA is invariant to linear transformations of the data. (Under the linear transformation $x_1 \to Lx_1$ and $x_2 \to L'x_2$, the result of CCA does not change. This is because CCA works in terms of normalized correlation coefficients.) We define CCA more precisely in Section 3.

In certain special cases, there are normative justifications for CCA as a dimensionality reduction technique. When x_1 and x_2 are jointly distributed as a Gaussian, the Gaussian Information Bottleneck method [Chechik et al., 2005] shows that CCA provides an appropriate compression scheme (under the Information Bottleneck criterion [Tishby et al., 1999]). In a semi-supervised multi-view setting, Kakade and Foster [2007] show that CCA provides the natural dimensionality reduction technique by which one can project x onto a lower dimensional space (using CCA) and yet still retain predictive information about y. However, this work was rather specific to the square loss and used a multi-view assumption tailored to the square loss.

This work provides a normative justification of CCA in a rather broad sense — we generalize the work of Kakade and Foster [2007]. We consider a setting where have a convex loss function of the form $\ell(w \cdot x, y)$, where either the loss function is strictly convex (e.g. log loss, square loss) or we use a strictly convex regularizer (e.g. hinge loss with L_2 regularization). We show that, under the multi-view assumption above, if we perform CCA and project the data onto to the top k canonical directions (where k is determined by the canonical eigenspectrum), then this projection loses little predictive information about Y. Hence, our subsequent supervised learning problem is simpler as we can work with a lower dimensional space (with the knowledge that we have not thrown away useful predictive information in working with this lower dimensional space). We state this precisely in Section 3.

2 Co-Regularization and Compatibility

We now consider the PAC style semi-supervised framework introduced in Balcan and Blum [2006] and generalize the framework to general loss functions. We work with a prediction space $\widehat{\mathcal{Y}}$ that need not be equal to $\mathcal{Y}$. The goal is to learn a pair of predictors (f_1, f_2), where $f_1 : \mathcal{X}_1 \to \widehat{\mathcal{Y}}$ and $f_2 : \mathcal{X}_2 \to \widehat{\mathcal{Y}}$, based on the labeled and unlabeled data such that the expected loss of any one of these predictors is small. We work with loss functions (bounded in $[0, 1]$) of the form $\ell(f; (x_1, x_2, y))$ (usually the loss functions are of the more restricted form $\ell(f(x), y)$ though in some cases, e.g. Example 4, this more general form is appropriate). Denote by $L(f_1)$ the expected loss of f_1, i.e. $L(f_1) = E\ell(f_1; (x_1, y))$, and $L(f_2)$ is similarly defined. Let $\mathcal{F}_1$ and $\mathcal{F}_2$ denote the hypothesis classes of interest, consisting of functions from $\mathcal{X}_1$ (and, respectively, $\mathcal{X}_2$) to the prediction space $\widehat{\mathcal{Y}}$. Let a Bayes optimal predictor with respect to loss L based on input X_1, X_2 be denoted by $y^*(X_1, X_2)$. So $y^* \in \operatorname{argmin}_f L(f)$, where the argmin is over all functions. Similarly, let y_v^* for $v \in \{1, 2\}$ be Bayes optimal predictors with respect to loss function L based on input X_v.

2.1 Compatible Function Classes

As discussed in the Introduction, to leverage our information theoretic assumption, we would like to say that a near optimal predictor using information from one view tends to agree with a near optimal predictor from another view. If this were the case, then the intuitive basis for an algorithm would be to find predictors from either view which tend to agree. However, quantifying this statement depends on the details of the loss function and the prediction space, since we need to specify a relationship between a measure of "closeness"

of the loss function and a measure of agreement between hypothesis. We do this in the following assumption, which can be considered an *inverse* Lipschitz condition, which bounds how close two functions are in terms of how close their loss is.

Assumption 2 (Inverse Lipschitz Condition) *There exists a symmetric function $d : \widehat{\mathcal{Y}} \times \widehat{\mathcal{Y}} \to \mathbb{R}^+$ and a monotonically increasing non-negative function Φ on the reals (with $\Phi(0) = 0$) such that for all f,*

$$E[d(f(x), y^*(x))] \leq \Phi(L(f) - L(y^*))$$

where the expectation is with respect to $x = (x_1, x_2)$, and y^* is some Bayes optimal predictor with respect to loss L. Furthermore, for $v \in 1, 2$ and all f_v,

$$E[d(f_v(x), y_v^*(x))] \leq \Phi(L(f_v) - L(y_v^*))$$

where y_v^* is a Bayes optimal predictor using only knowledge of x_v.

While we this assumption seems natural enough, we should note that there some subtleties. For example, if we are dealing with binary prediction and the $0/1$ loss function (the binary classification loss), consider the case where the target function is complete noise. Here, all predictors are Bayes optimal and have the maximal error rate of 0.5. Hence, predictors can be far from agreeing yet they are all optimal. In general, for the $0/1$ loss, the higher the noise, the less near-optimal predictors need to agree. In the next Subsection, we consider this case in more detail (in Example 2), and we also consider other commonly used loss functions.

While it is natural to assume that d satisfies the triangle inequality, there are some natural choices of d which do not satisfy this. In particular, in some cases we would like to use $d(y, y') = (y - y')^2$, which does not satisfy the triangle inequality. Hence, we only assume a relaxed version of the triangle inequality.

Assumption 3 (Relaxed Triangle Inequality) *For the function d, there exists a $c_d \geq 1$ such that*

$$\forall \widehat{y}_1, \widehat{y}_2, \widehat{y}_3 \in \widehat{\mathcal{Y}}, \;\; d(\widehat{y}_1, \widehat{y}_2) \leq c_d(d(\widehat{y}_1, \widehat{y}_3) + d(\widehat{y}_3, \widehat{y}_2))$$

We now introduce the incompatibility framework of Balcan and Blum [2006] for the multi-view setting. Here, we have a function $\chi : \widehat{\mathcal{Y}} \times \widehat{\mathcal{Y}} \to \mathbb{R}^+$, which we think of as scoring how incompatible two functions are. In particular, in this framework, they desire to use functions which are highly compatible. To formalize this, define the compatible function class with respect to incompatibility function χ and some $t \geq 0$ as those pairs of functions which are compatible to the tune of t, more precisely:

$$\mathcal{C}^\chi(t) = \{(f_1, f_2) : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2 \, and \, E[\chi(f_1, f_2)] \leq t\}$$

where we are slightly abusing notation by referring to $\chi(f, f')$ as meaning $\chi(f(x_1, x_2), f'(x_1, x_2))$, which we do throughout.

In order to characterize how good this compatibility class is, in terms of our multi-view assumption, we need to also define the Bayes regret:

$$\epsilon_{\text{bayes}} = \max\{L(f_1^*) - L(y_1^*), L(f_2^*) - L(y_2^*)\}$$

where $f_v^* \in \mathcal{F}_v$ is the optimal predictor for view v within the hypothesis class $\mathcal{F}_v$.

Our first result shows that for a particular choice of t, the incompatibility class contains a good pair of hypothesis.

Theorem 1 (Bias) *If Assumptions 1, 2, and 3 are satisfied, then given a loss function ℓ bounded by 1 and if we set the incompatibility function to be d, i.e. $\chi = d$, then for $t = 2c_d^2(\Phi(\sqrt{\epsilon_{\text{info}}}) + \Phi(\epsilon_{\text{bayes}}))$, we have:*

$$\inf_{(f_1, f_2) \in \mathcal{C}^\chi(t)} \frac{L(f_1) + L(f_2)}{2} \leq L(y^*) + \epsilon_{\text{bayes}} + \sqrt{\epsilon_{\text{info}}}$$

(The proof is provided in the Appendix).

Of course, for convex loss functions we have $L(\frac{f_1+f_2}{2}) \leq \frac{L(f_1)+L(f_2)}{2}$.

The need for stating the bound in terms of the Bayes regret ϵ_{bayes} is due to our information theoretic Assumption 1 not explicitly referring to any hypothesis classes $\mathcal{F}_1$ and $\mathcal{F}_2$. The square root dependence on ϵ_{info} is a result of using Pinsker's equality in the proof, which relates the L_1 distance to the KL-distance (see Cover and Thomas [1991]).

Note that in Balcan and Blum [2006] they did *not* explicitly characterize the quality of the incompatibility class — they assumed that χ was known and that a setting of t was known such that $\mathcal{C}^\chi(t)$ contained a 'good' predictor. Here, we derive our incompatibility function and we specify a value t. Intuitively, this lemma characterizes the bias — the reduction in performance — by using $\mathcal{C}^\chi(t)$ instead of the full hypothesis classes $\mathcal{F}_1$ and $\mathcal{F}_2$, in terms of the error ϵ_{info}.

We now provide examples of pairs χ and Φ for commonly used loss functions, showing that our multi-view framework is quite general.

2.2 Examples of Loss/Incompatibility Pairs

Example 1 (Squared Loss) *Let $\mathcal{Y}, \widehat{\mathcal{Y}} = \mathbb{R}$. Consider the loss function $\ell(\widehat{y}, y) = (y - \widehat{y})^2$. Here, we can choose the incompatibility function $\chi(\widehat{y}_1, \widehat{y}_2) = d(\widehat{y}_1, \widehat{y}_2) = (\widehat{y}_1 - \widehat{y}_2)^2$ and $\Phi(x) = x$. To see that this satisfies all the requisite assumptions, note that since $(a - b)^2 \leq 2(a^2 + b^2)$, we have that χ satisfies the relaxed triangle inequality with $c_d = 2$. Also, since that $y_v^* = E[Y|X_v]$ and $y^* = E[Y|X_1, X_2]$, we have:*

$$E(f_v - y_v^*)^2 = E(f_v - y)^2 - E(y_v^* - y)^2 \;\; ,$$
$$E(f - y^*)^2 = E(f - y)^2 - E(y^* - y)^2$$

so our inverse Lipschitz condition is satisfied with equality.

Example 2 (Zero-one Loss) *Here, we have $\mathcal{Y}, \widehat{\mathcal{Y}} = \{1, -1\}$ with $\ell(\widehat{y}, y) = \mathbb{1}_{\{y \neq \widehat{y}\}}$. As discussed in the previous Subsection, there is no natural choice of d and Φ for this loss function, without further restrictions on the noise. Hence, let us assume that Tsybakov's noise condition [Tsybakov, 2004] holds for each view independently and for both views together for some noise exponent $\alpha \in (0, 1]$, which we define below. Now we can choose the incompatibility function $\chi(\widehat{y}_1, \widehat{y}_2) = \mathbb{1}_{\{\widehat{y}_1 \neq \widehat{y}_2\}}$ with $\Phi(x) = cx^\alpha$ where $c > 0$ (defined below). Here, χ is in fact a metric and hence satisfies the triangle inequality.*

To see that the choice of Φ is appropriate, first note that by definition of Tsybakov's noise condition, for all $f_1 : \mathcal{X}_1 \to \widehat{\mathcal{Y}}$, $f_2 : \mathcal{X}_2 \to \widehat{\mathcal{Y}}$ and $f : \mathcal{X}_1 \times \mathcal{X}_2 \to \widehat{\mathcal{Y}}$ there exists $c > 0$ such that for $v \in \{1, 2\}$

$$Pr(f(X_v)(\eta_v(X_v) - \frac{1}{2}) \leq 0) \leq c(L(f_v) - L(y_v^*))^\alpha$$

and

$$Pr(f(X_1, X_2)(\eta(X) - \frac{1}{2}) \leq 0) \leq c(L(f) - L(y^*))^\alpha$$

where η_v and η stand for $P(Y = 1|X_v)$ and $P(Y = 1|X_1, X_2)$ respectively. Now since $\text{sign}(\eta(X) - \frac{1}{2})$ is the Bayes optimal predictor, $\mathbb{1}_{\{f(X)(\eta(X)-\frac{1}{2})\leq 0\}} = \mathbb{1}_{\{f(X)\neq y^*(X)\}} = \chi(f, y^*)$ and thus, under Tsybakov's noise condition, Assumption 2 is satisfied.

Example 3 (Strictly Convex Losses) *Consider a loss function $\ell(\widehat{y}, y)$ where, for each y, $\ell(\cdot, y)$ is strictly convex with respect to pseudo-metric d with modulus of convexity δ (defined below). Let the prediction space $\widehat{\mathcal{Y}}$ and output space $\mathcal{Y}$ be bounded a subset of $\mathbb{R}$. Here, $\chi(\widehat{y}_1, \widehat{y}_2) = \delta(d(\widehat{y}_1, \widehat{y}_2))$ satisfies Assumption 2 with $\Phi(x) = \frac{x}{2}$ (provided the modulus of convexity function $\delta(\epsilon) \leq \epsilon^p$ for some $p > 0$). In this case it is easy to check that $c_d = 1$ if $p < 1$ and $c_d = 2^{p-1}$ otherwise.*

To see this, we first define modulus of convexity of the loss function ℓ with respect to pseudometric d (in its first parameter). We say that for a given y, $\ell(\cdot, y)$ has modulus of convexity δ if,

$$\delta_y(\epsilon) = \inf\{\frac{\ell(\widehat{y}, y) + \ell(\widehat{y}', y)}{2} - \ell(\frac{\widehat{y} + \widehat{y}'}{2}, y) : d(\widehat{y}, \widehat{y}') \geq \epsilon\}$$

where the inf is over $\widehat{y}, \widehat{y}' \in \widehat{\mathcal{Y}}$. We actually want to work with a uniform bound on this function and so we define δ to be any function satisfying,

$$\delta(\epsilon) \leq \inf_{y\in\mathcal{Y}} \delta_y(\epsilon)$$

Now note that

$$\frac{L(f_v) + L(y_v^*)}{2} - L(\frac{f_v + y_v^*}{2}) \geq E\delta(d(f_v, y_v^*))$$

and

$$\frac{L(f) + L(y^*)}{2} - L(\frac{f + y^*}{2}) \geq E\delta(d(f, y^*))$$

Since $L(\frac{f_v+y_v^*}{2}) \geq L(y_v^*)$ and $L(\frac{f+y^*}{2}) \geq L(y^*)$ we have that,

$$E[\chi(f_v, y_v^*)] = E\delta(d(f_v, y_v^*)) \leq \frac{L(f_v) - L(y_v^*)}{2}$$

and

$$E[\chi(f, y^*)] = E\delta(d(f, y^*)) \leq \frac{L(f) - L(y^*)}{2}$$

which shows our choice of χ and Φ is appropriate.

Remark 1 *It is worth noting that whenever Assumption 2 is satisfied with $\chi(\widehat{y}_1, \widehat{y}_2) = g(d(\widehat{y}_1, \widehat{y}_2))$ where d is some pseudo-metric and g is an invertible convex function then Assumption 2 is also with $\chi' = d$ as the incompatibility function and $\Phi_{\chi'} = g^{-1}(\Phi)$. This is a simple consequence of Jensen's inequality.*

Example 4 (L_2 Regularized Losses) *Say we have some loss function ℓ that is convex and $\widehat{\mathcal{Y}} = \mathbb{R}$. Now consider the regularized loss functional for a certain RKHS function class $\mathcal{F}$,*

$$\ell_\lambda(f; x, y) := \ell(f(x), y) + \lambda\|f\|_K^2 \tag{2}$$

Taking $\chi(\widehat{y}_1, \widehat{y}_2) = (\widehat{y}_1 - \widehat{y}_2)^2$ we can show that Assumption 2 is satisfied for the regularized loss with $\Phi(x) = \frac{(K+\lambda)^2}{2\lambda}x$, where $K := \sup_{x\in\mathcal{X}} \sqrt{K(x, x)}$ (note that here we overload the notation K, but it is clear from context).

To see this, define for $f, f' \in \mathcal{F}$ the metric

$$d_{\lambda,x}(f, f') = |f(x) - f'(x)| + \lambda\|f - f'\|_K$$

One can show that $E[\ell_\lambda(f)]$ is strictly convex with respect to $d_{\lambda,x}$ (Steinwart and Scovel [2006], Lemma 6.4) with modulus of convexity $\delta(\epsilon) = \frac{\lambda\epsilon^2}{(K+\lambda)^2}$. From this we see that

$$\begin{aligned}
&\frac{E[\ell_\lambda(f; x, y)] - E[\ell_\lambda(f^*; x, y)]}{2} \\
&\geq E[\frac{\ell_\lambda(f; x, y) + \ell_\lambda(f^*; x, y)}{2} - \ell_\lambda(\frac{f + f^*}{2}; x, y)] \\
&\geq E\delta(d'_{\lambda,x}(f, f^*)) \\
&\geq E\delta(|f(x) - f^*(x)| + \lambda\|f - f^*\|) \\
&\geq E\delta(|f(x) - f^*(x)|) \\
&\geq \frac{\lambda}{(K+\lambda)^2} E(f(x) - f^*(x))^2
\end{aligned}$$

Thus we see that for the regularized loss functional ℓ_λ the squared incompatibility satisfies Assumption 2, with our choice of $\Phi(x) = \frac{(K+\lambda)^2}{2\lambda}x$.

2.3 Convergence Bounds

We now characterize the sample complexity of an algorithm which uses a labeled and unlabeled data set, sampled from the underlying distribution. Our framework again parallels that of Balcan and Blum [2006] — broadened to include more general loss functions.

The basic algorithm we consider is identical to that in Balcan and Blum [2006]. Given an unlabeled data set U, we define the empirical compatibility class as:

$$\widehat{\mathcal{C}^\chi}(t) = \{(f_1, f_2) : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2 \text{ and } \widehat{E}_U[\chi(f_1, f_2)] \leq t\}$$

where the empirical expectation is:

$$\widehat{E}_U[\chi(f_1, f_2)] = \frac{1}{m} \sum_{(x_1,x_2)\in U} \chi(f_1(x_1), f_2(x_2)) \ .$$

The algorithm simply minimizes the average loss of predictions over labeled data subject to the constraint of choosing

hypothesis from $\widehat{\mathcal{C}^\chi}(t)$. More precisely, for a given t, the algorithm simply chooses the best pair in this class:

$$(\widehat{f}_1, \widehat{f}_2) = \operatorname*{argmin}_{f_1, f_2 \in \widehat{\mathcal{C}^\chi}(t)} \widehat{E}_S[\ell(f_1(x_1), y) + \ell(f_2(x_2), y)] \quad (3)$$

The co-regularization algorithm can viewed as a dual version of this algorithm, which we consider in the following Subsection.

As we are dealing with abstract hypothesis classes, as in Balcan and Blum [2006], we make an assumption about the learning complexity with respect to these abstract hypothesis class — we give examples shortly. This assumption is stated in terms of both S and U, which allows us to use data-dependent sample complexity bounds (such as the Rademacher bounds), which is important in the next Subsection (for the analysis of the co-regularization algorithms and SVM-2K).

Assumption 4 (Sample Complexity) *For the hypothesis classes $\mathcal{F}_1$ and $\mathcal{F}_2$,*

Unlabeled: *With probability greater than $1-\delta$ over the i.i.d. sampling of unlabeled data set U we have that $\forall (f_1, f_2) \in \mathcal{F}_1 \times \mathcal{F}_2$*

$$\widehat{E}[\chi(f_1, f_2)] \leq E[\chi(f_1, f_2)] + G_\chi(\mathcal{F}_1 \times \mathcal{F}_2, U, \delta)$$

where G_χ is some notion of the generalization of the function class.

Labelled Case: *For any given unlabeled data set U, with probability greater than $1-\delta$ over i.i.d sampling of labeled data set S we have that for all pairs $(f_1, f_2) \in \widehat{\mathcal{C}^\chi}(t)$,*

$$|L(f_1) + L(f_2) - (\widehat{L}(f_1) + \widehat{L}(f_2))| \leq G_\ell(\widehat{\mathcal{C}^\chi}(t), S, \delta)$$

where G_ℓ is some notion of the generalization of the function class.

We now provide some standard sample complexity bounds.

Remark 2 (Examples of G_χ and G_ℓ) *Assumption 4 is satisfied in the following standard examples.*
Finite Hypothesis Class: *If the hypothesis classes are finite, then using Chernoff and union bounds we have*

$$G_\chi(\mathcal{H}, U, \delta) = O\left(\sqrt{\frac{log(|\mathcal{H}|) + log(\frac{1}{\delta})}{m}}\right)$$

and $G_\chi = G_\ell$.

Finite VC Class: *If the hypotheses map to $[0,1]$ and the VC dimension is finite, then*

$$G_\chi(\mathcal{H}, U, \delta) = O\left(\sqrt{\frac{VCdim(\mathcal{H}) + log(\frac{1}{\delta})}{m}}\right)$$

and $G_\chi = G_\ell$.

Rademacher Bounds : *For bounded loss and incompatibility functions, Rademacher bounds give us:*

$$G_\chi(\mathcal{H}, U, \delta) = O\left(\widehat{R}_m(\mathcal{H}) + 3\sqrt{\frac{ln(2/\delta)}{2m}}\right)$$

and $G_\chi = G_\ell$. Here, $\widehat{R}_n(\mathcal{H}) = \frac{1}{n} E_\sigma \sup_{f \in \mathcal{H}} \sum_{i=1}^n \sigma_i f(x_i)$ where σ_i are Rademacher variables.

We are now ready to state our main result on the complexity of our multi-view algorithm.

Theorem 2 *Assume that the function ℓ is bounded by 1, the incompatibility function $\chi = d$ and that Assumptions 1, 2, 3 and 4 hold. Set*

$$t = 2c_d^2(\Phi(\sqrt{\epsilon_{\text{info}}}) + \Phi(\epsilon_{\text{bayes}})) + G_\chi(\mathcal{F}_1 \times \mathcal{F}_2, U, \delta)$$

and let the pair $(\widehat{f}_1, \widehat{f}_2)$ be the output of the algorithm (as defined by Equation 3) with this setting of t. Then with probability greater than $1-\delta$ over an i.i.d sample of both the labeled dataset S and unlabeled dataset U, we have

$$\frac{L(\widehat{f}_1) + L(\widehat{f}_2)}{2} \leq L(y^*) + G_\ell(\widehat{\mathcal{C}^\chi}(t), S, \delta/3) + \epsilon_{\text{bayes}} + \sqrt{\epsilon_{\text{info}}}$$

(The proof is provided in the Appendix).

This statement is analogous to the main complexity statements in the semi-supervised PAC framework of Balcan and Blum [2006]. In particular, the unlabeled complexity G_χ only alters the setting of t, just as in Balcan and Blum [2006]. The labeled complexity term, G_ℓ, appears as a penalization to the bound, again as in the semi-supervised PAC framework.

The main difference is that we now specify the value of t to be used and compare ourselves to the Bayes optimal. Note that in Balcan and Blum [2006], there is no explicit characterization as to how much bias is introduced by using $\mathcal{C}^\chi(t)$ as opposed to using the unconstrained hypothesis space. The information theoretic assumption is what allows us to make this explicit characterization. The term $\sqrt{\epsilon_{\text{info}}}$ is the bias introduced by using the constrained hypothesis space rather than the unconstrained hypothesis space. The benefit is that we could substantially reduce the variance. In particular, this variance reduction is reflected by that the labeled complexity term, G_ℓ, only depends on the restricted hypothesis space, $\widehat{\mathcal{C}^\chi}(t)$, rather than the full hypothesis space — the former of which could have significantly less complexity.

We now show specific algorithms and analyses fit into this framework.

2.4 Algorithms

We now provide bounds for co-regularization algorithms and the SVM-2K algorithm of Farquhar et al. [2005]. For $v \in \{1, 2\}$ let $\mathcal{F}_v$ be some RKHS with respect to norm $\|\cdot\|_K$. Define ℓ_λ as in Example 4, i.e.

$$\ell_\lambda(f; x, y) := \ell(f(x), y) + \lambda \|f\|_K^2 \quad (4)$$

where $\ell(f(x), y)$ is convex. Define

$$L_\lambda(f) := E\ell_\lambda(f; (x_1, x_2, y)) \ .$$

Also let

$$f^* = \operatorname*{argmin}_f E[L_\lambda(f)]$$

where the argmin is over all functions (so f_* is the Bayes optimal predictor). By the Representer Theorem, f^* lives in the RKHS. This implies that $\epsilon_{\text{bayes}} = 0$.

Throughout this section we overload notation by using $K := \sup_{x \in \mathcal{X}} \sqrt{K(x,x)}$ (when it is clear from context).

Co-Regularization (with squared incompatibility)

The original co-regularization algorithm introduced in Sindhwani et al. [2005] and also the co-regularized least squares regression Brefeld et al. [2006] both minimize the objective in Equation 1. Recall that for the regularized convex loss functions in Example 4, we already showed that $\chi(f_1(x_1), f_2(x_2)) = (f_1(x_1) - f_2(x_2)^2$ satisfies Assumption 2. Therefore we see that Theorem 2 justifies these co-regularization algorithms under the information theoretic Assumption 1.

Rosenberg and Bartlett [2007] provide an estimate for the Rademacher complexity of kernel class for co-regularization in a transductive type setting (i.e. conditioned on the unlabeled data). The bound given is exactly of the form needed in Assumption 4. The subtlety in using these complexity bounds is that the co-regularization algorithms are a dual formulation of our Algorithm (see Equation 3), the latter of which imposes a hard agreement constraint. Hence, to provide a bound we need find an appropriate setting of the parameter λ_{co}. The following theorem does this.

Corollary 3 *Assume we are working in the transductive setting (where U is known and the underlying data distribution is uniform over U). Let C_{lip} be the Lipschitz constant for the loss. Let $K^v_{S\times S}$, $K^v_{S\times U}$ and $K^v_{U\times U}$ stand for the kernel matrix between labeled examples, between labeled and unlabeled examples, and unlabeled and unlabeled samples for view $v \in \{1,2\}$ respectively.*

Given $\lambda > 0$, if we set $\lambda_{co} = \frac{\lambda}{4(K+\lambda)^2\sqrt{\epsilon_{\text{info}}}}$ then for the pair of functions $(\widehat{f}_1, \widehat{f}_2) \in \mathcal{F}_1 \times \mathcal{F}_2$ returned by the co-regularization algorithm (Equation 1), with probability at least $1-\delta$ over labeled samples,

$$L_\lambda(\tfrac{\widehat{f}_1+\widehat{f}_2}{2}) \leq L_\lambda(f^*) + \frac{1}{\sqrt{n}}\left(2 + 3\sqrt{\frac{ln(\frac{2}{\delta})}{2}}\right) + 2C_{Lip}\widehat{R}_n(\widehat{\mathcal{C}}^\chi(\tfrac{1}{\lambda_{co}})) + \sqrt{\epsilon_{\text{info}}}$$

Where,

$$\widehat{R}_n(\widehat{\mathcal{C}}^\chi(\tfrac{1}{\lambda_{co}})) \leq \frac{R}{n}$$

$$R^2 = \lambda^{-1}tr(K^1_{S\times S}) + \lambda^{-1}tr(K^2_{S\times S}) - \frac{\lambda}{4(K+\lambda)^2\sqrt{\epsilon_{\text{info}}}}tr(J^T(I+\lambda M)^{-1}J)$$

$$J = \lambda^{-1}K^1_{U\times S} - \lambda^{-1}K^2_{U\times S}\,,\; M = \lambda^{-1}K^1_{U\times U} - \lambda^{-1}K^2_{U\times U}$$

(The proof is provided in the Appendix).

An important difference between our bounds and that in Rosenberg and Bartlett [2007] is that the above bound compares to the Bayes optimal predictor f^*, while Rosenberg and Bartlett [2007] only compare to the best function in $\widehat{\mathcal{C}}^\chi(t)$ (without any normative justification for how to set the parameter t). Our comparison to f^* leads to the additional penalty of $\sqrt{\epsilon_{\text{info}}}$ (and we specify a value of λ_{co} in the bound).

Note that the appropriate setting of λ_{co} is $O(1/\sqrt{\epsilon_{\text{info}}})$. In particular, this shows it is appropriate for $\lambda_{co} \to \infty$ as $\epsilon_{\text{info}} \to 0$, i.e. when the information theoretic assumption is as sharp as possible, we are permitted to co-regularize as hard as possible (without introducing any bias). For this case, the co-regularization algorithms obtain their maximal reduction in variance.

To convert the above corollary to an inductive bound (where U is a random sample) we need to establish an unlabeled complexity statement of the kind in Assumption 4. Note that if the prediction space is bounded then it can be shown using covering number arguments (Zhang [2002]) that $G_\chi(\mathcal{F}_1 \times \mathcal{F}_2, U, \delta)$ will be $c\sqrt{\frac{log(1/\delta)}{m}}$ where c is some constant (which depends of λ_{co} and K). Hence by setting $t = 2c_d^2(\Phi(\epsilon_{\text{bayes}}) + \Phi(\sqrt{\epsilon_{\text{info}}})) + c\sqrt{\frac{log(1/\delta)}{m}}$ we can get the inductive statement required.

Two View SVM

The SVM-2K approach proposed by Farquhar et al. [2005] can be formulated as the following optimization problem:

$$\operatorname*{argmin}_{(f_1,f_2)\in\mathcal{F}_1\times\mathcal{F}_2} \frac{1}{2}(\widehat{E}_S[\ell(f_1(x_1),y)] + \widehat{E}_S[\ell(f_2(x_2),y)]) + \lambda\|f_1\|_K^2 + \lambda\|f_2\|_K^2 + \lambda_{co}\widehat{E}_U[|f_1(x_1) - f_2(x_2)|] \quad (5)$$

where ℓ is the hinge loss. Technically, the formulation in Farquhar et al. [2005] uses slack variables (more in line with the usual SVM formulation), but the above formulation is identical. [1]

SVM-2K can be viewed as using the incompatibility function $\chi(\widehat{y}_1, \widehat{y}_2) = |\widehat{y}_1 - \widehat{y}_2|$. Recall that for regularized convex loss functions in Example 4, we already showed that $(f_1(x_1) - f_2(x_2))^2$ satisfies Assumption 2. Hence using Remark 1 we see that this incompatibility function for SVM-2K also satisfies Assumption 3 and 2 with $c_d = 1$ and $\phi(x) = \sqrt{\frac{(K+\lambda)^2}{2\lambda}}x$. Hence, we get the following Corollary.

Corollary 4 *Assume we are working in the transductive setting (where U is known and the underlying data distribution is uniform over U). Given $\lambda > 0$, if we set and $\lambda_{co} = \sqrt{\frac{\lambda}{2(K+\lambda)^2\sqrt{\epsilon_{\text{info}}}}}$ then with probability at least $1-\delta$ over labeled samples, for the pair of functions $(\widehat{f}_1, \widehat{f}_2) \in \mathcal{F}_1 \times \mathcal{F}_2$ returned by SVM-2K algorithm (Equation 5),*

$$L_\lambda(\tfrac{\widehat{f}_1+\widehat{f}_2}{2}) \leq L_\lambda(f^*) + 2\widehat{R}_n(\widehat{\mathcal{C}}^\chi(\tfrac{1}{\lambda_{co}})) + 3\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} + \sqrt{\epsilon_{\text{info}}}$$

where $\widehat{R}_n(\widehat{\mathcal{C}}^\chi(\frac{1}{\lambda_{co}}))$ is the data-dependent Rademacher complexity.

[1] Technically, the SVM-2K algorithm has a parameter ϵ which allows a little more disagreement, but the algorithm we specify is equivalent to the SVM-2K algorithm with $\epsilon = 0$.

In particular, Farquhar et al. [2005] show how to upper bound $\widehat{R}_n(\widehat{\mathcal{C}}^\chi(t))$ as a solution to a particular optimization problem. The proof is essentially identical to the previous Corollary, and is not provided.

Again, the main extension in our work is that we compare the algorithm's performance to the loss of the Bayes optimal predictor f^*, while Farquhar et al. [2005] only compares to the best function in $\widehat{\mathcal{C}}^\chi(t)$. Our comparison to f^* leads to the additional penalty of $\sqrt{\epsilon_{\text{info}}}$ (and we specify a value of t in the bound).

The appropriate setting of λ_{co} is $O(1/\sqrt{\epsilon_{\text{info}}})$ which again shows that smaller ϵ_{info} gets, the harder we can co-regularize.

3 Dimensionality Reduction and CCA

Consider a setting where $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ is a real vector space (of finite or countably infinite dimension). Here, we work with linear predictors of the form $w^T x$ and convex losses of form $\ell(w^T x, y)$ that satisfy Assumptions 2 and 3 with respect to the squared incompatibility function. For example, most strictly convex loss functions can be used with the squared incompatibility function, including the square loss, log loss, exponential loss, and L_2 regularized losses. Let $L(w) = E[\ell(w^T x, y)]$. For simplicity, we work in the transductive setting — in particular, we only assume knowledge of the second order statistics of the underlying data distribution (i.e. we know the covariance matrix of $\mathcal{X}$).

Assume that the loss function is twice differentiable and that the second derivative of the loss function is bounded from above by some constant C, that is

$$\forall z \frac{d^2 \ell(z, y)}{dz^2} \leq C \tag{6}$$

Note that this assumption is satisfied for common strictly convex losses.

Define canonical correlation analysis (CCA) as follows:

Definition 5 *The bases B_1, B_2 for $\mathcal{X}_1$ and $\mathcal{X}_2$ is the* canonical basis *for the two views if for (x_1, x_2) in this basis the following holds:*

1. *Orthogonality Conditions: For $v \in \{1, 2\}$*

$$E[(x_v)_i (x_v)_j] = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

2. *Correlation Conditions:*

$$E[(x_1)_i (x_2)_j] = \begin{cases} \gamma_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

where γ_i is the i^{th} correlation coefficient. We assume without loss of generality that $1 \geq \gamma_1 \geq \gamma_2 \geq ... \geq 0$.

Now we present the main algorithm, which uses CCA as a dimensionality reduction technique. Consider some threshold, $0 < \gamma_{\text{thresh}} < 1$. Let i_{thresh} be the smallest i such that

$$\gamma_i < \gamma_{\text{thresh}}$$

First, project x_v to the subspace spanned by the first $1, ..., i_{thresh}$ canonical coordinates. Denote this projection by $\Pi_{cca}(x_v)$. Let $\beta^{(v)}_{proj}$ be the optimal linear predictor for view v using only the projected $\Pi_{cca}(x_v)$ as input.

We now show that the loss of performance due to this projection is small if ϵ_{info} is small.

Theorem 6 *Assume that Equation 6 holds, that Assumption 1 is satisfied, and that Assumptions 2 and 3 hold with respect to the squared incompatibility function. Then*

$$L(\beta^{(v)}_{proj}) - L(y^*_v) \leq \frac{4C\left(\Phi(\sqrt{\epsilon_{\text{info}}}) + \Phi(\epsilon_{\text{bayes}})\right)}{1 - \gamma_{\text{thresh}}} + \epsilon_{\text{bayes}}$$

where C satisfies Equation 6.

(The proof is provided in the Appendix).

In particular, if the cutoff, γ_{thresh}, is $\frac{1}{2}$, then makes the $\frac{1}{1-\gamma_{\text{thresh}}}$ factor in the bound into 2.

Let us consider the implications for learning with a random labeled data set S using $\Pi_{cca}(x_v)$. Here, the a learning algorithm only needs to work with the coordinates which have sufficiently large γ_i. Hence, the supervised learning problem is simpler as we can work with a lower dimensional space. This Theorem is analogous to the dimensionality reduction statements in Kakade and Foster [2007] — though there the statements were restricted to the square loss (and a multi-view assumption based on the square loss).

4 Discussion

An Open Problem from Balcan and Blum [2007]

This problem (presented at COLT 2007) is where we have the $0/1$ loss, and it is assumed that classifiers from either view can perfectly predict the data (so the best classifiers agree completely on the unlabeled data). Furthermore, they assume that the classifiers are linearly separable. The question posed is can an efficient algorithm be found? A more general and practically relevant question is this case but with noise, which of course makes the problem harder. Here, the optimal predictors (from either view) may not agree perfectly on the unlabeled data. However, under Example 2, we know that choosing d to be the $0/1$ loss is a suitable discrepancy function (with Φ being defined in terms of the Tsybakov noise exponent).

In practice, even in the single view case, one is rarely able to directly minimize the $0/1$ loss. Instead, what one actually does is minimize a *surrogate* loss function, such as the hinge loss, logistic loss, or exponential loss. Furthermore, through the work of Bartlett et al. [2006], we have an understanding of how minimizing these surrogate losses relate to the $0/1$ loss.

In our framework, we are able to choose a discrepancy functions tailored to our loss (as long as the discrepancy satisfies Assumption 2). Hence, if we are using a surrogate loss (for the $0/1$ loss) then we should choose a incompatibility function that satisfies Assumption 2 with respect to this surrogate loss. We view both the co-regulation algorithms and the SVM-2K algorithm as the solution to this problem, under the theory of surrogate losses (where both these algorithms are using the surrogate hinge loss).

4.1 Relations to the Information Bottleneck

We end with a note on the connection to the Information Bottleneck method. In this method, the goal is to compress X_1 to Z such that Z has maximum information about X_2 — in particular, Z is a compression of X_1 that retains all the information that X_1 has about X_2, that is,

$$Z = \operatorname*{argmin}_A I(A : X_1)$$
$$\text{s.t. } I(A : X_2) = I(X_1 : X_2)$$

where the argmin is over compression functions A of X_1.

In the multi-view setting, if we find such a Z (with respect to to X_1 and X_2), it can be shown that

$$I(Z : Y) \geq I(X_1 : Y) - \epsilon_{\text{info}}$$

This shows that Z looses little predictive information about Y. In this sense, the Information Bottleneck is not throwing much relevant information with regards to Y and can be used as a semi-supervised algorithm.

In fact, using Lemma 7, one can show that for any loss bounded by 1, the Bayes optimal predictor which uses only knowledge of Z has a regret of at most $\sqrt{\epsilon_{\text{info}}}$ with respect to the Bayes optimal predictor y^*. An interesting direction to pursue is to learn with Z as inputs to our learning algorithm rather than X_v, since Z has lower entropy. Two issues to consider are: 1) the mapping Z has an abstract range (so one needs to take care in how to learn a function from $Z \to Y$) and 2) it is not clear how to implement the Information Bottleneck without knowledge of the underlying distribution.

Acknowledgements

We thank Gilles Blanchard for a number of helpful suggestions.

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.

Maria-Florina Balcan and Avrim Blum. Open problems in efficient semi-supervised pac learning. In *Conference on Computational Learning Theory (COLT)*, 2007.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. In *Journal of the American Statistical Association*, volume 101, No. 473, pages 138–156, 2006.

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.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. *J. Mach. Learn. Res.*, 6:165–188, 2005. ISSN 1533-7928.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, 1991.

Sanjoy Dasgupta, Michael L. Littman, and David A. McAllester. Pac generalization bounds for co-training. In *NIPS*, pages 375–382, 2001.

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

H. Hotelling. The most predictable criterion. *Journal of Educational Psychology*, 26:139–142, 1935.

Sham M. Kakade and Dean P. Foster. Multi-view regression via canonical correlation analysis. In Nader H. Bshouty and Claudio Gentile, editors, *COLT*, volume 4539 of *Lecture Notes in Computer Science*, pages 82–96. Springer, 2007.

David Rosenberg and Peter L. Bartlett. The rademacher complexity of co-regularized kernel classes. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, 2007.

V. Sindhwani, P. Niyogi, and M. Belkin. A Co-Regularization Approach to Semi-supervised Learning with Multiple Views. In *Workshop on Learning with Multiple Views, Proceedings of International Conference on Machine Learning*, 2005.

Ingo Steinwart and C. Scovel. Fast rates for support vector machines using gaussian kernels. In Los Alamos National Laboratory Technical Report LA-UR-04-8796, editor, *Annals of Statistics*, 2006.

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.

A. Tsybakov. Optimal aggregation of classifiers in statistical learning. In *Annals of Statistics*, volume 32 No. 1, 2004.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

A Proofs

First we state two Lemmas that will be used in proving the theorem.

Lemma 7 *For $v \in \{1, 2\}$, if the loss function ℓ is bounded by 1 then we have that*

$$|L(y^*) - L(y_v^*)| \leq \sqrt{\epsilon_{\text{info}}} \quad \textit{and} \quad |L(y_1^*) - L(y_2^*)| \leq 2\sqrt{\epsilon_{\text{info}}}$$

Proof: Consider some function $g : \mathcal{X} \to [0,1]$ and some two probability measures P and Q. We have that

$$\begin{aligned} |\int g(x)dQ - \int g(x)dP| &= |\int (1-\beta)g(x)dQ| \\ &\leq \int |1-\beta|dQ \\ &\leq \sqrt{DK(Q\|P)} \end{aligned} \tag{7}$$

where $\beta = \frac{dP}{dQ}$ and the last step is because the L_1 variational distance is bounded by square root of the KL divergence (Pinsker's Inequality). Now using this we get that for a fixed x_1, x_2 we have that

$$\begin{aligned} |E_{Y|X_1=x_1}\ell(y^*(x_1,x_2),y) - E_{Y|X=(x_1,x_2)}\ell(y^*(x_1,x_2),y)| \\ \leq \sqrt{DK(P_{Y|X=(x_1,x_2)}\|P_{Y|X_1=x_1})} \end{aligned}$$

Taking expectation with respect to $X = (X_1, X_2)$ and using Jensen's inequality twice (once on the left for convex function $|x|$ and once on the right for concave function $\sqrt{x}$) we get that

$$\begin{aligned} &|E_X E_{Y|X_1=x_1}\ell(y^*(x_1,x_2),y) - L(y^*)| \\ &\leq \sqrt{E_X DK(P_{Y|X=(x_1,x_2)}\|P_{Y|X_1=x_1})} \end{aligned}$$

Now note that since

$$L(y_1^*) \leq E_X E_{Y|X_1=x_1}\ell(y^*(x_1,x_2),y)$$

and $L(y_1^*) \geq L(y^*)$, we get

$$\begin{aligned} &|L(y_1^*) - L(y^*)| \\ &\leq \sqrt{E_X DK(P_{Y|X=(x_1,x_2)}\|P_{Y|X_1=x_1})} \end{aligned}$$

Also,

$$E_X DK(P_{Y|X=(x_1,x_2)}\|P_{Y|X_1=x_1}) = I_{Y:X_2|X_1}$$

and so we have that

$$|L(y_1^*) - L(y^*)| \leq \sqrt{\epsilon_{\text{info}}}$$

similarly we have

$$|L(y_2^*) - L(y^*)| \leq \sqrt{\epsilon_{\text{info}}}$$

Also the above two inequalities together imply that

$$|L(y_1^*) - L(y_2^*)| \leq 2\sqrt{\epsilon_{\text{info}}}$$

■

Lemma 8 *For any f_1, f_2 assume*

$$L(f_1) - L(y_1^*) \leq \epsilon', \qquad L(f_2) - L(y_1^*) \leq \epsilon'$$

then given Assumptions 1, 2 and 3 and that the loss function is bounded by B, we have that

$$E[\chi(f_1,f_2)] \leq 2c_d^2(\Phi(\epsilon') + \Phi(\sqrt{\epsilon_{\text{info}}}))$$

Proof: First note that by Assumptions 2 and 3 we have that for f_1 and f_2 there exists y_1^* and y_2^* such that

$$\begin{aligned} E[\chi(f_1,y_1^*)] \leq \Phi(L(f_1) - L(y_1^*)) \quad and \\ E[\chi(f_2,y_2^*)] \leq \Phi(L(f_2) - L(y_2^*)) \end{aligned}$$

and since Φ is monotonically increasing we have that

$$\begin{aligned} E[\chi(f_1,y_1^*)] \leq \Phi(\epsilon') \quad and \\ E[\chi(f_2,y_2^*)] \leq \Phi(\epsilon') \end{aligned}$$

Again by Assumptions 2 and 3 we have that for some specific y^*,

$$E[\chi(y_1^*,y^*)] \leq \Phi(L(y_1^*) - L(y^*)) \leq \Phi(\sqrt{\epsilon_{\text{info}}})$$

and

$$E[\chi(y_2^*,y^*)] \leq \Phi(L(y_2^*) - L(y^*)) \leq \Phi(\sqrt{\epsilon_{\text{info}}})$$

Since χ satisfies the relaxed triangle inequality Assumption 3, we get that

$$E[\chi(y_2^*,y_1^*)] \leq c_d\Phi(\sqrt{\epsilon_{\text{info}}})$$

Again using relaxed triangle inequality Assumption 3, we get the required result that

$$\begin{aligned} E[\chi(f_1,f_2)] &\leq c_d^2(E[\chi(f_1,y_1^*)] + E[\chi(y_1^*,y_2^*)] + E[\chi(f_2,y_2^*)]) \\ &\leq 2c_d^2(\Phi(\epsilon') + \Phi(\sqrt{\epsilon_{\text{info}}})) \end{aligned}$$

■

Proof:[of Theorem 1]
Using Lemma 8 we see that

$$E[\chi(f_1^*,f_2^*)] \leq 2c_d^2(\Phi(\epsilon_{\text{bayes}}) + \Phi(\sqrt{\epsilon_{\text{info}}}))$$

Therefore setting $t = 2c_d^2(\Phi(\epsilon_{\text{bayes}}) + \Phi(\sqrt{\epsilon_{\text{info}}}))$ we find that $(f_1^*, f_2^*) \in \mathcal{C}^\chi(t)$ and thus,

$$(f_1^*,f_2^*) = \operatorname*{argmin}_{(f_1,f_2)\in\mathcal{C}^\chi(t)} \frac{L(f_1)+L(f_2)}{2}$$

Now by definition of ϵ_{bayes} we have that

$$\min_{f_v\in\mathcal{F}_v} L(f_v) - L(y_v^*) \leq \epsilon_{\text{bayes}}$$

Therefore,

$$\min_{(f_1,f_2)\in\mathcal{C}^\chi(t)} \frac{L(f_1)+L(f_2)}{2} \leq \frac{L(y_1^*)+L(y_2^*)}{2} + \epsilon_{\text{bayes}} \tag{8}$$

Now by Lemma 7 we see that for each $v \in \{1,2\}$, $L(y_v^*) - L(y^*) \leq \sqrt{\epsilon_{\text{info}}}$. Hence using this in Equation (8) we conclude that

$$\min_{(f_1,f_2)\in\mathcal{C}^\chi(t)} \frac{L(f_1)+L(f_2)}{2} \leq L(y^*) + \epsilon_{\text{bayes}} + \sqrt{\epsilon_{\text{info}}}$$

■

Proof:[of Theorem 2] Let $(f_{1\widehat{Co}}^*, f_{2\widehat{Co}}^*) \in \widehat{\mathcal{C}^\chi}(t)$ be the minimizer of $L(f_1) + L(f_2)$ in the class $\widehat{\mathcal{C}^\chi}(t)$. Using statement Assumption 4 (labeled) we have that with probability at least $1-\delta$ over the sample S,

$$\begin{aligned} \widehat{L}(f_{1\widehat{Co}}^*) + \widehat{L}(f_{2\widehat{Co}}^*) - L(f_{1\widehat{Co}}^*) - L(f_{2\widehat{Co}}^*) \\ \leq G_\ell(\widehat{\mathcal{C}^\chi}(t), S, \delta) \end{aligned}$$

Also for any $(f_1, f_2) \in \widehat{\mathcal{C}^\chi}(t)$ we have that with probability at least $1-\delta$ over the sample S,

$$L(f_1) + L(f_2) - \widehat{L}(f_1) - \widehat{L}(f_2) \leq G_\ell(\widehat{\mathcal{C}^\chi}(t), S, \delta)$$

Hence combining the two, for the pair $(\widehat{f}_1, \widehat{f}_2) \in \widehat{\mathcal{C}^\chi}(t)$ that minimizes $\widehat{L}(f_1) + \widehat{L}(f_1)$ we have that with probability at least $1-2\delta$ over the sample S,

$$L(\widehat{f}_1) + L(\widehat{f}_2) - L(f^*_{1\,\widehat{co}^t}) - L(f^*_{2\,\widehat{co}^t}) \leq 2G_\ell(\widehat{\mathcal{C}^\chi}(t), S, \delta)$$

Now Let $t' = 2c_d^2(\Phi(\sqrt{\epsilon_{\text{info}}}) + \Phi(\epsilon_{\text{bayes}}))$ then we see that if $(f_1, f_2) \in \mathcal{C}^\chi(t')$ then,

$$E[\chi(f_1, f_2)] \leq t'$$

However applying Assumption 4 (unlabeled) we find that with probability greater than$1-\delta$ over the unlabeled dataset U we have that

$$\widehat{E}\chi(f_1, f_2) \leq E[\chi(f_1, f_2)] + G_\chi(\mathcal{F}_1 \times \mathcal{F}_2, U, \delta)$$

Thus we can conclude that with probability greater than $1-\delta$ over the i.i.d. unlabeled sample we have that $(f_1, f_2) \in \widehat{\mathcal{C}^\chi}(t)$. Now using the above we see that with probability $1-\delta$ over unlabeled data

$$\min_{(f_1,f_2)\in\widehat{\mathcal{C}^\chi}(t)} L(f_1) + L(f_2) = \min_{(f_1,f_2)\in\mathcal{C}^\chi(t')} L(f_1) + L(f_2)$$

Hence using the result of Theorem 1 we can conclude that with probability $1-3\delta$ over both labeled and unlabeled data we have that

$$L(\widehat{f}_1) + L(\widehat{f}_2) \leq 2L(y^*) + 2G_\ell(\widehat{\mathcal{C}^\chi}(t), S, \delta) + 2\epsilon_{\text{bayes}} + 2\sqrt{\epsilon_{\text{info}}}$$

■

Proof:[Proof of Corollary 3] First note that we can write $f_1 \in \mathcal{F}_1$ as $(f_1, 0) \in \mathcal{F}_1 \times \mathcal{F}_2$ and similarly we can define any $f_2 \in \mathcal{F}_2$ as $(0, f_2) \in \mathcal{F}_1 \times \mathcal{F}_2$ so that we can consider only the joint RKHS defined by sum of f_1 and f_2. From Example 4 we first of all have that for the regularized loss Assumption 2 is satisfied by the squared incompatibility (i.e.. $\chi(\widehat{y}_1, \widehat{y}_2) = (\widehat{y}_1 - \widehat{y}_2)^2$) function with $\Phi(x) = \frac{(K+\lambda)^2}{2\lambda}x$. Also note that in this case $\epsilon_{\text{bayes}} = 0$ since f^* is in the RKHS (in fact for the regularized loss to even be applicable the function needs to live in the RKHS). Hence if we restrict ourselves to the class $\mathcal{C}^\chi(t)$ where $t = \frac{8(\lambda+K)^2\sqrt{\epsilon_{\text{info}}}}{\lambda}$ then using Theorem 2, we see that we can get a low regularized regret with respect to f^*. Now without loss of generality assume that for the given loss ℓ we have that $\ell(0, y) = 1$. Then using this in Equation 1 we see that,

$$\lambda_{co}\, \hat{E}_U[f_1(x_1) - f_2(x_2)]^2 \leq 1$$

and so using $\lambda_{co} = \frac{1}{t}$ we see that for any function pairs (f_1, f_2) returned by the algorithm $\widehat{E}_U[\chi(f_1, f_2)] \leq t$. However since we are in the transductive setting $\widehat{E}_U[\chi(f_1, f_2)] = E[\chi(f_1, f_2)]$. Now we use the result from Rosenberg and Bartlett [2007] to establish a statement of the form Assumption 4 (labeled).

To this end define,

$$\mathcal{H}(t) = \{(f_1, f_2) : \lambda\|f_1\|^2 + \lambda\|f_2\|^2 + \lambda_{co}\hat{E}_U(f_1(x_1) - f_2(x_2))^2 \leq 1\}$$

Notice that the solution of the co-regularization algorithm is contained in this class. Further as in Rosenberg and Bartlett [2007] define $\mathcal{J}(t) = \{x \to \frac{f_1(x_1)+f_2(x_2)}{2} : (f_1, f_2) \in \mathcal{H}\}$. Now we can directly use Theorem 2 of their paper (assuming ℓ is bounded by 1) to get that with probability at least $1-\delta$ over labeled samples, for all $(f_1, f_2) \in \widehat{\mathcal{C}^\chi}(t)$

$$L(f_1)+L(f_2) \leq \hat{L}(f_1) + \hat{L}(f_2) + 2C_{Lip}\widehat{R}_n(\mathcal{J}(t)) + \frac{1}{\sqrt{n}}(2 + 3\sqrt{\frac{ln(2/\delta)}{2}}) \quad (9)$$

Where by Theorem 3 of Rosenberg and Bartlett [2007] we find that

$$\widehat{R}_n(\mathcal{J}(t)) \leq \frac{R}{n}$$

where

$$R^2 = \lambda^{-1}tr(K^1_{S\times S}) + \lambda^{-1}tr(K^2_{S\times S}) - \frac{\lambda}{(\lambda+K)^2 t}tr(J^T(I+\lambda M)^{-1}J)$$

and

$$J = \lambda^{-1}K^1_{U\times S} - \lambda^{-1}K^2_{U\times S} \qquad M = \lambda^{-1}K^1_{U\times U} - \lambda^{-1}K^2_{U\times U}$$

Now this establishes the Assumption 4 , labeled statement we were aiming for.

Now putting the regularization term on both sides of the inequality in Equation 9 we get that

$$E[\ell_\lambda(f_1, x_1, y) + \ell_\lambda(f_2, x_2, y)] \leq \hat{E}[\ell_\lambda(f_1, x_1, y) + \ell_\lambda(f_2, x_2, y)] + 4C_{Lip}\widehat{R}_n(\mathcal{J}(t)) + \frac{1}{\sqrt{n}}(2 + 3\sqrt{\frac{ln(2/\delta)}{2}})$$

Now this is essentially the labeled statement in Assumption 4 and since we are in the transductive case we do not need the unlabeled part of the assumption. Hence using Theorem 2 we see that with probability at least $1-\delta$ over labeled samples for the pair $\widehat{f}_1, \widehat{f}_2^A$ returned by co-regularization algorithm,

$$E\frac{\ell(\widehat{f}_1 x_1, y) + \ell(\widehat{f}_2, x_2, y)}{2}] \leq E[\ell_\lambda(f^*, x_1, x_2, y)] + 2C_{Lip}\widehat{R}_n(\mathcal{J}(t)) + \frac{1}{\sqrt{n}}(2 + 3\sqrt{\frac{ln(2/\delta)}{2}}) + \sqrt{\epsilon_{\text{info}}}$$

Now using Jensen's Inequality we see that the regularized loss of the average predictor is bounded by average of regularized loss of the predictors and hence the result.

■

Proof:[of Theorem 6] Without loss of generality we assume we are in the CCA basis. For each $v \in \{1, 2\}$ let $\beta^{(v)}$ be the

minimizer with respect to β of $E[\ell(\beta^T x_v, y)]$. From the result of Lemma 8 using the the squared incompatibility function ($c_d = 2$ in this case) we have that

$$\begin{aligned}8\Phi(\sqrt{\epsilon_{\text{info}}}) + 8\Phi(\epsilon_{\text{bayes}}) &\geq E[(x_1^T\beta^{(1)} - x_2^T\beta^{(2)})^2] \\ &= \sum_i [(\beta_i^{(1)})^2 + (\beta_i^{(2)})^2 - 2\gamma_i\beta_i^{(1)}\beta_i^{(2)}] \\ &\geq \sum_i [(1-\gamma_i)(\beta_i^{(1)})^2 + (1-\gamma_i)(\beta_i^{(2)})^2]\end{aligned}$$

(the last step is due to the identity $2ab \leq a^2 + b^2$). Hence we conclude that

$$\sum_i (1-\gamma_i)(\beta_i^{(v)})^2 \leq 8\Phi(\sqrt{\epsilon_{\text{info}}}) + 8\Phi(\epsilon_{\text{bayes}}) \quad (10)$$

Let $\beta_P^{(v)}$ be the projection of $\beta^{(v)}$ on to the first i_{thresh} coordinates. Consider a twice differentiable loss function. By Taylor's theorem (second order) we have that there exists some $\tilde{\beta}$ such that

$$\begin{aligned}\ell(x_v^T\beta_P^{(v)}, y) = \ell(x_v^T\beta^{(v)}, y) + (\beta_P^{(v)} - \beta^{(v)})^T\nabla\ell(\beta^v) \\ + \frac{1}{2}(\beta_P^{(v)} - \beta^{(v)})^T\nabla^2\ell(\tilde{\beta}^T x_v, y)(\beta_P^{(v)} - \beta^{(v)})\end{aligned}$$

Taking expectation and noting that since $\beta^{(v)}$ is the minimizer of the expected loss we find that

$$\begin{aligned}&L(\beta_P^{(v)}) - L(\beta^{(v)}) = \\ &\frac{1}{2}(\beta^{(v)} - \beta_P^{(v)})^T E[\nabla^2\ell(\tilde{\beta}^T x_v, y)](\beta^{(v)} - \beta_P^{(v)})\end{aligned}$$

Let $\beta_{res}^{(v)} = \beta^{(v)} - \beta_P^{(v)}$. Note that since $(\beta_P^{(v)})_i = (\beta^{(v)})_i$ for all i's corresponding to correlation values greater than the threshold we see that $\beta_{res}^{(v)}$ is zero in the first i_{thresh} coordinates and is equal to $\beta^{(v)}$ on the rest. Now note that for a loss function that is twice differentiable and a function of $\tilde{\beta}^T x_v$ we have that by chain rule

$$\nabla^2\ell(\beta.x_v, y) = \frac{d^2\ell(\tilde{\beta}^T x_v, y)}{d(\tilde{\beta}^T x_v)^2} x_v x_v^T$$

Now using the assumption that the second derivative of the loss function is bounded by some C we then see that

$$L(\beta_P^{(v)}) - L(\beta^{(v)}) \leq \frac{C}{2}(\beta_{res}^{(v)})^T E[x_v x_v^T](\beta_{res}^{(v)})$$

Note that since we are in the CCA basis we have that $E[(x_v)_i(x_v)_j] = 0$ when $i \neq j$ and is 1 otherwise. Now note that for all $i > i_{\text{tresh}}$ we have that $1 - \gamma_i > 1 - \gamma_{\text{thresh}}$ and so,

$$\begin{aligned}L(\beta_P^{(v)}) - L(\beta^{(v)}) &\leq \frac{C}{2}\|\beta_{res}^{(v)}\|^2 \\ &= \frac{C}{2}\sum_{i > i_{thresh}} (\beta_i^{(v)})^2 \\ &\leq \frac{C}{2}\sum_{i > i_{thresh}} \frac{1-\gamma_i}{1-\gamma_{\text{thresh}}}(\beta_i^{(v)})^2 \\ &\leq \frac{C}{2(1-\gamma_{\text{thresh}})}\sum_{i > i_{thresh}} (1-\gamma_i)(\beta_i^{(v)})^2 \\ &\leq \frac{C}{2(1-\gamma_{\text{thresh}})}\sum_i (1-\gamma_i)(\beta_i^{(v)})^2\end{aligned}$$

Hence using Equation 10 we can conclude that

$$L(\beta_P^{(v)}) - L(\beta^{(v)}) \leq \frac{4C\left(\Phi(\sqrt{\epsilon_{\text{info}}}) + \Phi(\epsilon_{\text{bayes}})\right)}{(1-\gamma_{\text{thresh}})}$$

Now since $L(\beta_P^{(v)}) \geq L(\beta_{proj}^{(v)})$ we conclude that

$$L(\beta_{proj}^{(v)}) - L(\beta^{(v)}) \leq \frac{4C\left(\Phi(\sqrt{\epsilon_{\text{info}}}) + \Phi(\epsilon_{\text{bayes}})\right)}{(1-\gamma_{\text{thresh}})}$$

Finally since $L(\beta^{(v)}) - L(y_v^*) \leq \epsilon_{\text{bayes}}$ we have the required result. ■