An RKHS for Multi-View Learning and Manifold Co-Regularization

Vikas Sindhwani VSINDHW@US.IBM.COM
Mathematical Sciences, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598 USA

David S. Rosenberg DROSEN@STAT.BERKELEY.EDU
Department of Statistics, University of California Berkeley, CA 94720 USA

Abstract

Inspired by co-training, many multi-view semi-supervised kernel methods implement the following idea: find a function in each of multiple Reproducing Kernel Hilbert Spaces (RKHSs) such that (a) the chosen functions make similar predictions on unlabeled examples, and (b) the average prediction given by the chosen functions performs well on labeled examples. In this paper, we construct a single RKHS with a data-dependent "co-regularization" norm that reduces these approaches to standard supervised learning. The reproducing kernel for this RKHS can be explicitly derived and plugged into any kernel method, greatly extending the theoretical and algorithmic scope of co-regularization. In particular, with this development, the Rademacher complexity bound for co-regularization given in (Rosenberg & Bartlett, 2007) follows easily from well-known results. Furthermore, more refined bounds given by localized Rademacher complexity can also be easily applied. We propose a co-regularization based algorithmic alternative to manifold regularization (Belkin et al., 2006; Sindhwani et al., 2005a) that leads to major empirical improvements on semi-supervised tasks. Unlike the recently proposed transductive approach of (Yu et al., 2008), our RKHS formulation is truly semi-supervised and naturally extends to unseen test data.

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

1. Introduction

In semi-supervised learning, we are given a few labeled examples together with a large collection of unlabeled data from which to estimate an unknown target function. Suppose we have two hypothesis spaces, $\mathcal{H}^1$ and $\mathcal{H}^2$, each of which contains a predictor that well-approximates the target function. We know that *predictors that agree with the target function also agree with each other on unlabeled examples.* Thus, any predictor in one hypothesis space that does not have an "agreeing predictor" in the other can be safely eliminated from consideration. Due to the resulting reduction in the complexity of the joint learning problem, one can expect improved generalization performance.

These conceptual intuitions and their algorithmic instantiations together constitute a major line of work in semi-supervised learning. One of the earliest approaches in this area was "co-training" (Blum & Mitchell, 1998), in which $\mathcal{H}^1$ and $\mathcal{H}^2$ are defined over different representations, or "views", of the data, and trained alternately to maximize mutual agreement on unlabeled examples. More recently, several papers have formulated these intuitions as joint complexity regularization, or *co-regularization*, between $\mathcal{H}^1$ and $\mathcal{H}^2$ which are taken to be Reproducing Kernel Hilbert Spaces (RKHSs) of functions defined on the input space $\mathcal{X}$. Given a few labeled examples $\{(\boldsymbol{x}_i, y_i)\}_{i\in L}$ and a collection of unlabeled data $\{\boldsymbol{x}_i\}_{i\in U}$, co-regularization learns a prediction function,

$$f_\star(\boldsymbol{x}) = \frac{1}{2}\big(f^1_\star(\boldsymbol{x}) + f^2_\star(\boldsymbol{x})\big) \tag{1}$$

where $f^1_\star \in \mathcal{H}^1$ and $f^2_\star \in \mathcal{H}^2$ are obtained by solving the following optimization problem,

$$(f^1_\star, f^2_\star) = \operatorname*{argmin}_{f^1\in\mathcal{H}^1, f^2\in\mathcal{H}^2} \gamma_1 ||f^1||^2_{\mathcal{H}^1} + \gamma_2 ||f^2||^2_{\mathcal{H}^2}$$
$$+\mu \sum_{i\in U} [f^1(\boldsymbol{x}_i) - f^2(\boldsymbol{x}_i)]^2 + \sum_{i\in L} V\left(y_i, f(\boldsymbol{x}_i)\right) \tag{2}$$

In this objective function, the first two terms measure complexity by the RKHS norms $\|\cdot\|^2_{\mathcal{H}_1}$ and $\|\cdot\|^2_{\mathcal{H}_2}$ in $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively, the third term enforces agreement among predictors on unlabeled examples, and the final term evaluates the empirical loss of the mean function $f = (f^1 + f^2)/2$ on the labeled data with respect to a loss function $V(\cdot,\cdot)$. The real-valued parameters γ_1, γ_2, and μ allow different tradeoffs between the regularization terms. L and U are index sets over labeled and unlabeled examples respectively.

Several variants of this formulation have been proposed independently and explored in different contexts: linear logistic regression (Krishnapuram et al., 2005), regularized least squares classification (Sindhwani et al., 2005b), regression (Brefeld et al., 2006), support vector classification (Farquhar et al., 2005), Bayesian co-training (Yu et al., 2008), and generalization theory (Rosenberg & Bartlett, 2007).

The main theoretical contribution of this paper is the construction of a new "co-regularization RKHS," in which standard supervised learning recovers the solution to the co-regularization problem of Eqn. 2. Theorem 2.2 presents the RKHS and gives an explicit formula for its reproducing kernel. This "co-regularization kernel" can be plugged into any standard kernel method giving convenient and immediate access to two-view semi-supervised techniques for a wide variety of learning problems. Utilizing this kernel, in Section 3 we give much simpler proofs of the results of (Rosenberg & Bartlett, 2007) concerning bounds on the Rademacher complexity and generalization performance of co-regularization. As a more algorithmic application, in Section 4 we consider the semi-supervised learning setting where examples live near a low-dimensional manifold embedded in a high dimensional ambient euclidean space. Our approach, manifold co-regularization (CoMR), gives major empirical improvements over the manifold regularization (MR) framework of (Belkin et al., 2006; Sindhwani et al., 2005a).

The recent work of (Yu et al., 2008) considers a similar reduction. However, this reduction is strictly transductive and does not allow prediction on unseen test examples. By contrast, our formulation is truly semi-supervised and provides a principled out-of-sample extension.

2. An RKHS for Co-Regularization

We start by reformulating the co-regularization optimization problem, given in Eqn. 1 and Eqn. 2, in the following equivalent form where we directly solve for the final prediction function $f_\star$:

$$f_\star = \operatorname*{argmin}_{f} \min_{\substack{f=f^1+f^2 \\ f^1\in\mathcal{H}^1, f^2\in\mathcal{H}^2}} \frac{\gamma_1}{2}||f^1||^2_{\mathcal{H}^1} + \frac{\gamma_2}{2}||f^2||^2_{\mathcal{H}^2} + \frac{\mu}{2}\sum_{i\in U}[f^1(\boldsymbol{x}_i) - f^2(\boldsymbol{x}_i)]^2 + \frac{1}{2}\sum_{i\in L} V\left(y_i, \frac{1}{2}f(\boldsymbol{x}_i)\right) \quad (3)$$

Consider the sum space of functions, $\tilde{\mathcal{H}}$, given by,

$$\begin{aligned}\tilde{\mathcal{H}} &= \mathcal{H}^1 \oplus \mathcal{H}^2 \quad (4)\\ &= \{f | f(\boldsymbol{x}) = f^1(\boldsymbol{x}) + f^2(\boldsymbol{x}), f^1 \in \mathcal{H}^1, f^2 \in \mathcal{H}^2\}\end{aligned}$$

and impose on it a data-dependent norm,

$$\|f\|^2_{\tilde{\mathcal{H}}} = \min_{\substack{f=f^1+f^2 \\ f^1\in\mathcal{H}^1, f^2\in\mathcal{H}^2}} \gamma_1\|f^1\|^2_{\mathcal{H}^1} + \gamma_2\|f^2\|^2_{\mathcal{H}^2} + \mu\sum_{i\in U}\left[f^1(\boldsymbol{x}_i) - f^2(\boldsymbol{x}_i)\right]^2 \quad (5)$$

The minimization problem in Eqn. 3 can then be posed as standard supervised learning in $\tilde{\mathcal{H}}$ as follows,

$$f_\star = \operatorname*{argmin}_{f\in\tilde{\mathcal{H}}} \gamma\|f\|^2_{\tilde{\mathcal{H}}} + \frac{1}{2}\sum_{i\in L} V\left(y_i, \frac{1}{2}f(\boldsymbol{x}_i)\right) \quad (6)$$

where $\gamma = \frac{1}{2}$. Of course, this reformulation is not really useful unless $\tilde{\mathcal{H}}$ itself is a valid new RKHS. Let us recall the definition of an RKHS.

Definition 2.1 (RKHS)**.** *A reproducing kernel Hilbert space (RKHS) is a Hilbert Space $\mathcal{F}$ that possesses a reproducing kernel,* i.e., *a function $k : \mathcal{X}\times\mathcal{X} \to \mathcal{R}$ for which the following hold: (a) $k(\boldsymbol{x}, .) \in \mathcal{F}$ for all $\boldsymbol{x} \in \mathcal{X}$, and (b) $\langle f, k(\boldsymbol{x}, .)\rangle_{\mathcal{F}} = f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$ and $f \in \mathcal{F}$, where $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ denotes inner product in $\mathcal{F}$.*

In Theorem 2.2, we show that $\tilde{\mathcal{H}}$ is indeed an RKHS, and moreover we give an explicit expression for its reproducing kernel. Thus, it follows that although the domain of optimization in Eqn. 6 is nominally a function space, by the Representer Theorem we can express it as a finite-dimensional optimization problem.

2.1. Co-Regularization Kernels

Let $\mathcal{H}^1, \mathcal{H}^2$ be RKHSs with kernels given by k^1, k^2 respectively, and let $\tilde{\mathcal{H}} = \mathcal{H}^1 \oplus \mathcal{H}^2$ as defined in Eqn. 4. We have the following result.

Theorem 2.2. *There exists an inner product on $\tilde{\mathcal{H}}$ for which $\tilde{\mathcal{H}}$ is a RKHS with norm defined by Eqn. 5 and reproducing kernel $\tilde{k} : \mathcal{X}\times\mathcal{X} \to \mathcal{R}$ given by,*

$$\tilde{k}(\boldsymbol{x}, \boldsymbol{z}) = s(\boldsymbol{x}, \boldsymbol{z}) - \mu\mathbf{d}_{\boldsymbol{x}}^T H \mathbf{d}_{\boldsymbol{z}} \quad (7)$$

where $s(\boldsymbol{x},\boldsymbol{z})$ is the (scaled) sum of kernels given by,

$$s(\boldsymbol{x},\boldsymbol{z}) = \gamma_1^{-1} k^1(\boldsymbol{x},\boldsymbol{z}) + \gamma_2^{-1} k^2(\boldsymbol{x},\boldsymbol{z}),$$

and $\mathbf{d}_{\boldsymbol{x}}$ is a vector-valued function that depends on the difference in views measured as,

$$\mathbf{d}_{\boldsymbol{x}} = \gamma_1^{-1}\mathbf{k}^1_{U\boldsymbol{x}} - \gamma_2^{-1}\mathbf{k}^2_{U\boldsymbol{x}},$$

where $\mathbf{k}^i_{U\boldsymbol{x}} = \left[k^i(\boldsymbol{x},\boldsymbol{x}_j), j \in U\right]^T$, and H is a positive-definite matrix given by $H = (I+\mu S)^{-1}$. Here, S is the gram matrix of $s(\cdot,\cdot)$, i.e., *$S = \left(\gamma_1^{-1}K^1_{UU} + \gamma_2^{-1}K^2_{UU}\right)$ where $K^i_{UU} = k^i(U,U)$ denotes the Gram matrices of k^i over unlabeled examples.*

We give a rigorous proof in Appendix A.

2.2. Representer Theorem

Theorem 2.2 says that $\tilde{\mathcal{H}}$ is a valid RKHS with kernel $\tilde{k}$. By the Representer Theorem, the solution to Eqn.6 is given by

$$f_\star(\boldsymbol{x}) = \sum_{i\in L}\alpha_i \tilde{k}(\boldsymbol{x}_i,\boldsymbol{x}) \tag{8}$$

The corresponding components in $\mathcal{H}^1, \mathcal{H}^2$ can also be retrieved as,

$$f^1_\star(\boldsymbol{x}) = \sum_{i\in L}\alpha_i\gamma_1^{-1}\left(k^1(\boldsymbol{x}_i,\boldsymbol{x}) - \mu\mathbf{d}^T_{\boldsymbol{x}_i}Hk^1_{U\boldsymbol{x}}\right) \tag{9}$$

$$f^2_\star(\boldsymbol{x}) = \sum_{i\in L}\alpha_i\gamma_2^{-1}\left(k^2(\boldsymbol{x}_i,\boldsymbol{x}) + \mu\mathbf{d}^T_{\boldsymbol{x}_i}Hk^2_{U\boldsymbol{x}}\right) \tag{10}$$

Note that $\mathcal{H}^1$ and $\mathcal{H}^2$ are defined on the same domain $\mathcal{X}$ so that taking the mean prediction is meaningful. In a two-view problem one may begin by defining $\mathcal{H}^1, \mathcal{H}^2$ on different view spaces $\mathcal{X}^1, \mathcal{X}^2$ respectively. Such a problem can be mapped to our framework by extending $\mathcal{H}^1, \mathcal{H}^2$ to $\mathcal{X} = \mathcal{X}^1 \times \mathcal{X}^2$ by re-defining $f^1(\boldsymbol{x}^1,\boldsymbol{x}^2) = f^1(\boldsymbol{x}^1), f^1 \in \mathcal{H}^1$; similarly for $\mathcal{H}^2$. While we omit these technical details, it is important to note that in such cases, Eqns. 9 and 10 can be reinterpreted as predictors defined on $\mathcal{X}^1, \mathcal{X}^2$ respectively.

3. Bounds on Complexity and Generalization

By eliminating all predictors that do not collectively agree on unlabeled examples, co-regularization intuitively reduces the complexity of the learning problem. It is reasonable then to expect better test performance for the same amount of labeled training data. In (Rosenberg & Bartlett, 2007), the size of the co-regularized function class is measured by its empirical Rademacher complexity, and tight upper and lower bounds are given on the Rademacher complexity of the co-regularized hypothesis space. This leads to generalization bounds in terms of the Rademacher complexity. In this section, we derive these complexity bounds in a few lines using Theorem 2.2 and a well-known result on RKHS balls. Furthermore, we present improved generalization bounds based on the theory of localized Rademacher complexity.

3.1. Rademacher Complexity Bounds

Definition 3.1. *The empirical Rademacher complexity of a function class $\mathcal{A} = \{f : \mathcal{X} \to \mathcal{R}\}$ on a sample $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_\ell \in \mathcal{X}$ is defined as*

$$\hat{R}_\ell(\mathcal{A}) = E^\sigma\left[\sup_{f\in\mathcal{A}}\left|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i f(\boldsymbol{x}_i)\right|\right],$$

where the expectation is with respect to $\sigma = \{\sigma_1,\ldots,\sigma_\ell\}$, and the σ_i are i.i.d. Rademacher random variables, that is, $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$.

Let $\mathcal{H}$ be an arbitrary RKHS with kernel $k(\cdot,\cdot)$, and denote the standard RKHS supervised learning objective function by $Q(f) = \sum_{i\in L} V(y_i, f(\boldsymbol{x}_i)) + \lambda||f||^2_{\mathcal{H}}$. Let $f_\star = \operatorname{argmin}_{f\in\mathcal{H}} Q(f)$. Then $Q(f_\star) \le Q(0) = \sum_{i\in L} V(y_i,0)$. It follows that $\|f_\star\|^2_{\mathcal{H}} \le Q(0)/\lambda$. Thus if we have some control *a priori* on $Q(0)$, then we can restrict the search for $f_\star$ to a ball in $\mathcal{H}$ of radius $r = \sqrt{Q(0)/\lambda}$.

We now cite a well-known result about the Rademacher complexity of a ball in an RKHS (see *e.g.* (Boucheron et al., 2005)). Let $\mathcal{H}_r := \{f \in \mathcal{H} : ||f||_{\mathcal{H}} \le r\}$ denote the ball of radius r in $\mathcal{H}$. Then we have the following:

Lemma 3.2. *The empirical Rademacher complexity on the sample $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_\ell \in \mathcal{X}$ for the RKHS ball $\mathcal{H}_r$ is bounded as follows: $\frac{1}{\sqrt[4]{2}}\frac{2r}{\ell}\sqrt{\mathrm{tr}K} \le \hat{R}_\ell(\mathcal{H}_r) \le \frac{2r}{\ell}\sqrt{\mathrm{tr}K}$ where $K = \left(k(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)^{\ell}_{i,j=1}$ is the kernel matrix.*

For the co-regularization problem described in Eqns. 3 and 6, we have $f_\star \in \tilde{\mathcal{H}}_r$ where $r^2 = \ell\sup_y V(0,y)$, where ℓ is number of labeled examples. We now state and prove bounds on the Rademacher complexity of $\tilde{\mathcal{H}}_r$. The bounds here are exactly the same as those given in (Rosenberg & Bartlett, 2007). However, while they have a lengthy "bare-hands" approach, here we get the result as a simple corollary of Theorem 2.2 and Lemma 3.2.

Theorem 3.3. *The empirical Rademacher complexity on the labeled sample $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_\ell \in \mathcal{X}$ for the RKHS ball $\tilde{\mathcal{H}}_r$ is bounded as follows:*

$$\frac{1}{\sqrt[4]{2}}\frac{2r}{\ell}\sqrt{\mathrm{tr}\tilde{K}} \le \hat{R}_\ell(\tilde{\mathcal{H}}_r) \le \frac{2r}{\ell}\sqrt{\mathrm{tr}\tilde{K}},$$

where
$\tilde{K} = \gamma_1^{-1} K_{LL}^1 + \gamma_2^{-1} K_{LL}^2 - \mu D_{UL}^T (I + \mu S)^{-1} D_{UL}$ *and*
$D_{UL} = \left(\gamma_1^{-1} K_{UL}^1 - \gamma_2^{-1} K_{UL}^2\right)$

Proof. Note that $\tilde{K}$ is just the kernel matrix for the co-regularization kernel $\tilde{k}(\cdot,\cdot)$ on the labeled data. Then the bound follows immediately from Lemma 3.2. □

3.2. Co-Regularization Reduces Complexity

The co-regularization parameter μ controls the extent to which we enforce agreement between f^1 and f^2. Let $\tilde{\mathcal{H}}(\mu)$ denote the co-regularization RKHS for a particular value of μ. From Theorem 3.3, we see that the Rademacher complexity for a ball of radius r in $\tilde{\mathcal{H}}(\mu)$ decreases with μ by an amount determined by

$$\Delta(\mu) = \operatorname{tr}\left[\mu D_{UL}^T (I + \mu S)^{-1} D_{UL}\right] \quad (11)$$

$$= \sum_{i=1}^{\ell} \rho^2 \left(\mathbf{k}_{U\boldsymbol{x}_i}^1 \mathbf{k}_{U\boldsymbol{x}_i}^2\right) \quad (12)$$

where $\rho(\cdot,\cdot)$ is a metric on $\mathcal{R}^{|U|}$ defined by
$\rho^2(\mathbf{s},\mathbf{t}) = \mu(\gamma_1^{-1}\mathbf{s} - \gamma_2^{-1}\mathbf{t})' (I + \mu S)^{-1} (\gamma_1^{-1}\mathbf{s} - \gamma_2^{-1}\mathbf{t})$

We see that the complexity reduction, $\Delta(\mu)$, grows with the ρ-distance between the two different representations of the labeled points. Note that the metric ρ is determined by S, which is the weighted sum of the gram matrices of the two kernels on unlabeled data.

3.3. Generalization Bounds

With Theorem 2.2 allowing us to express multi-view co-regularization problems as supervised learning in a data-dependent RKHS, we can now bring a large body of theory to bear on the generalization performance of co-regularization methods. We start by quoting the theorem proved in (Rosenberg & Bartlett, 2007). Next, we state an improved bound based on localized Rademacher complexity. Below, we denote the unit ball in $\tilde{\mathcal{H}}$ by $\tilde{\mathcal{H}}_1$.

Condition 1. The loss $V(\cdot,\cdot)$ is Lipschitz in its first argument, i.e., there exists a constant A such that $\forall y, \hat{y}_1, \hat{y}_2$: $|V(\hat{y}_1, y) - V(\hat{y}_2, y)| \leq A\,|\hat{y}_1 - \hat{y}_2|$

Theorem 3.4. *Suppose $V : \mathcal{Y}^2 \to [0,1]$ satisfies Condition 1. Then conditioned on the unlabeled data, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the sample of labeled points $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell)$ drawn i.i.d. from P, we have for any predictor $f \in \tilde{\mathcal{H}}_1$ that*

$$P[V(\varphi(\boldsymbol{x}), y)] \leq \frac{1}{\ell}\sum_{i=1}^{\ell} V(\varphi(\boldsymbol{x}_i), y_i) + 2B\hat{R}_\ell(\tilde{\mathcal{H}}_1) + \frac{1}{\sqrt{\ell}}\left(2 + 3\sqrt{\ln(2/\delta)/2}\right)$$

We need two more conditions for the localized bound:

Condition 2. For any probability distribution P, there exists $f_\star \in \tilde{\mathcal{H}}_1$ satisfying $P[V(f_\star(\boldsymbol{x}), y)] = \inf_{f \in \tilde{\mathcal{H}}_1} P[V(f(\boldsymbol{x}), y)]$

Condition 3. There is a constant $B \geq 1$ such that for every probability distribution P and every $f \in \tilde{\mathcal{H}}_1$ we have, $P(f - f_*)^2 \leq BP(V[f(\boldsymbol{x}), y] - V[f_\star(\boldsymbol{x}), y])$

In the following theorem, let P_ℓ denote the empirical probability measure for the labeled sample of size ℓ.

Theorem 3.5. *[Cor. 6.7 from (Bartlett et al., 2002)] Assume that $\sup_{\boldsymbol{x}\in\mathcal{X}} k(\boldsymbol{x},\boldsymbol{x}) \leq 1$ and that V satisfies the 3 conditions above. Let $\hat{f}$ be any element of $\tilde{\mathcal{H}}_1$ satisfying $P_\ell V[\hat{f}(\boldsymbol{x}), y] = \inf_{f\in\tilde{\mathcal{H}}_1} P_\ell V[f(\boldsymbol{x}), y]$. There exist a constant c depending only on A and B s.t. with probability at least $1 - 6e^{-\nu}$,*

$$P\left(V[\hat{f}(\boldsymbol{x}), y] - V[f_\star(\boldsymbol{x}), y]\right) \leq c\left(\hat{r}^* + \frac{\nu}{\ell}\right),$$

where $\hat{r}^* \leq \min_{0\leq h\leq \ell}\left(\frac{h}{\ell} + \frac{1}{\ell}\sqrt{\sum_{i>h}\lambda_i}\right)$ and where $\lambda_1, \ldots, \lambda_\ell$ are the eigenvalues of the labeled-data kernel matrix $\tilde{K}_{LL}$ in decreasing order.

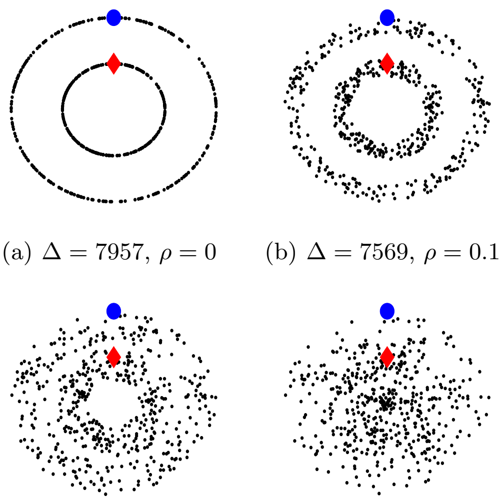
Note that while Theorem 3.4 bounds the gap between expected and empirical performance of an arbitrary $f \in \tilde{\mathcal{H}}_1$, Theorem 3.5 bounds the gap between the empirical loss minimizer over $\tilde{\mathcal{H}}_1$ and true risk minimizer in $\tilde{\mathcal{H}}_1$. Since the localized bound only needs to account for the capacity of the function class in the neighborhood of f_*, the bounds are potentially tighter. Indeed, while the bound in Theorem 3.4 is in terms of the trace of the kernel matrix, the bound in Theorem 3.5 involves the tail sum of kernel eigenvalues. If the eigenvalues decay very quickly, the latter is potentially much smaller.

4. Manifold Co-Regularization

Consider the picture shown in Figure 1(a) where there are two classes of data points in the plane ($\mathcal{R}^2$) lying on one of two concentric circles. The large, colored points are labeled while the smaller, black points are unlabeled. The picture immediately suggests two notions of distance that are very natural but radically different. For example, the two labeled points are close in the *ambient* euclidean distance on $\mathcal{R}^2$, but infinitely apart in terms of *intrinsic* geodesic distance measured along the circles.

Suppose for this picture one had access to two kernel functions, k^1, k^2 that assign high similarity to nearby points according to euclidean and geodesic distance respectively. Because of the difference in ambient and intrinsic representations, by co-regularizing the associated RKHSs one can hope to get good reductions in complexity, as suggested in section 3.2. In Figure 1, we report the value of complexity reduction (Eqn. 12) for four point clouds generated at increasing levels of noise off the two concentric circles. When noise becomes large, the ambient and intrinsic notions of distance converge and the amount of complexity reduction decreases.

Figure 1. Complexity Reduction $\Delta(\mu = 1)$ (Eqn. 12) with respect to noise level ρ. The choice of k^1, k^2 is discussed in the following subsections.



(a) $\Delta = 7957$, $\rho = 0$ (b) $\Delta = 7569$, $\rho = 0.1$

(c) $\Delta = 3720$, $\rho = 0.2$ (d) $\Delta = 3502$, $\rho = 0.5$

The setting where data lies on a low-dimensional submanifold $\mathcal{M}$ embedded in a higher dimensional ambient space $\mathcal{X}$, as in the concentric circles case above, has attracted considerable research interest recently, almost orthogonal to multi-view efforts. The main assumption underlying manifold-motivated semi-supervised learning algorithms is the following: *two points that are close with respect to geodesic distances on $\mathcal{M}$ should have similar labels.* Such an assumption may be enforced by an *intrinsic* regularizer that emphasizes complexity along the manifold.

Since $\mathcal{M}$ is truly unknown, the intrinsic regularizer is empirically estimated from the point cloud of labeled and unlabeled data. In the graph transduction approach, an nn-nearest neighbor graph $\mathcal{G}$ is constructed which serves as an empirical substitute for $\mathcal{M}$. The vertices $\mathcal{V}$ of this graph are the set of labeled and unlabeled examples. Let $\mathcal{H}_\mathcal{I}$ denote the space of all functions mapping $\mathcal{V}$ to $\mathcal{R}$, where the subscript $\mathcal{I}$ implies "intrinsic." Any function $f \in \mathcal{H}_\mathcal{I}$ can be identified with the vector $\boldsymbol{f} = [f(\boldsymbol{x}_i), \boldsymbol{x}_i \in \mathcal{V}]^T$. One can impose a norm $\|\boldsymbol{f}\|_\mathcal{I}^2 = \sum_{ij} W_{ij} \left[f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\right]^2$ on $\mathcal{H}_\mathcal{I}$ that provides a natural measure of smoothness for $\boldsymbol{f}$ with respect to the graph. Here, W denotes the adjacency matrix of the graph. When $\mathcal{X}$ is a euclidean space, a typical W is given by $W_{ij} = \exp(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2})$ if i and j are nearest neighbors and 0 otherwise. In practice, one may use a problem dependent similarity matrix to set these edge weights. This norm can be conveniently written as a quadratic form $\boldsymbol{f}^T M \boldsymbol{f}$, where M is the graph Laplacian matrix defined as $M = D - W$, and D is a diagonal degree matrix with entrees $D_{ii} = \sum_j W_{ij}$.

It turns out that $\mathcal{H}_\mathcal{I}$ with the norm $\|\cdot\|_\mathcal{I}$ is an RKHS whose reproducing kernel $k_\mathcal{I} : \mathcal{V} \times \mathcal{V} \to \mathcal{R}$ is given by $k_\mathcal{I}(\boldsymbol{x}_i, \boldsymbol{x}_j) = M_{ij}^\dagger$, where $M^\dagger$ denotes the pseudo-inverse of the Laplacian. Given $\mathcal{H}_\mathcal{I}$ with its reproducing kernel, graph transduction solves the standard RKHS regularization problem, $f_\star = \mathrm{argmin}_{f \in \mathcal{H}_I} \gamma \|f\|_\mathcal{I}^2 + \sum_{i \in L} V(y_i, f(\boldsymbol{x}_i))$, where y_i is the label associated with the node $\boldsymbol{x}_i$. Note that the solution $f_\star$ is only defined over $\mathcal{V}$, the set of labeled and unlabeled examples. Since graph transduction does not provide a function whose domain is the ambient space $\mathcal{X}$, it is not clear how to make predictions on unseen test points $\boldsymbol{x} \in \mathcal{X}$. Possessing a principled "out-of-sample extension" distinguishes semi-supervised methods from transductive procedures.

4.1. Ambient and Intrinsic Co-Regularization

We propose a co-regularization solution for out-of-sample prediction. Conceptually, one may interpret the manifold setting as a multi-view problem where each labeled or unlabeled example appears in two "views": (a) an ambient view in $\mathcal{X}$ in terms of euclidean co-ordinates $\boldsymbol{x}$ and (b) an intrinsic view in $\mathcal{G}$ as a vertex index i. Let $\mathcal{H}_\mathcal{A} : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$ be an RKHS defined over the ambient domain with an associated kernel $k_\mathcal{A} : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$. We can now combine ambient and intrinsic views by co-regularizing $\mathcal{H}_\mathcal{A}, \mathcal{H}_\mathcal{I}$. This can be done by setting $k^1 = k_\mathcal{A}$ and $k^2 = k_\mathcal{I}$ in Eqn. 7 and solving Eqn. 6. The combined prediction function $f_\star$ given by Eqn. 8 is the mean of an ambient component $f_\star^1$ given by Eqn. 9 and an intrinsic component $f_\star^2$ given by Eqn. 10. Even though $f_\star$ is transductive and only defined on labeled and unlabeled examples, the ambient component $f_\star^1$ can be used for out-of-sample prediction. Due to co-regularization, this ambient component is (a) smooth in $\mathcal{H}_\mathcal{X}$ and (b) in agreement with a smooth function on the data manifold. We call this approach manifold co-regularization, and abbreviate it as CoMR.

4.2. Manifold Regularization

In the manifold regularization (MR) approach of (Belkin et al., 2006; Sindhwani et al., 2005a), the following optimization problem is solved:

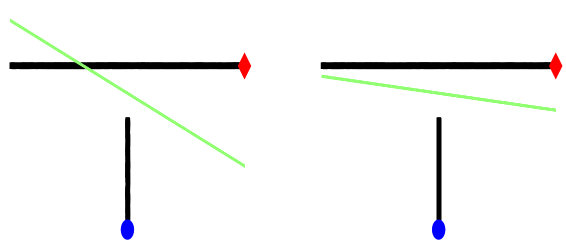
$$f^\star = \underset{f\in\mathcal{H}_\mathcal{A}}{\operatorname{argmin}}\, \gamma_1\|f\|^2_{\mathcal{H}_\mathcal{A}} + \gamma_2 \boldsymbol{f}^T M \boldsymbol{f} + \sum_{i\in L} V(y_i, f(\boldsymbol{x}_i)) \qquad (13)$$

where $\boldsymbol{f} = [f(\boldsymbol{x}_i), i \in L, U]^T$. The solution, $f^\star$ is defined on $\mathcal{X}$, and can therefore be used for out-of-sample prediction.

In Figure 2, we show a simple two-dimensional dataset where MR provably fails when $\mathcal{H}_\mathcal{A}$ is the space of linear functions on $\mathcal{R}^2$. The LINES dataset consists of two classes spread along perpendicular lines. In MR the intrinsic regularizer is enforced directly on $\mathcal{H}_\mathcal{A}$. It can be easily shown that the intrinsic norm of a linear function $f(\boldsymbol{x}) = \mathbf{w}^T\boldsymbol{x}$ along the perpendicular lines is exactly the same as the ambient norm, *i.e.*, $\|f\|^2_{\mathcal{H}_\mathcal{I}} = \|f\|^2_{\mathcal{H}_\mathcal{A}} = \mathbf{w}^T\mathbf{w}$. Due to this, MR simply ignores unlabeled data and reduces to supervised training with the regularization parameter $\gamma_1 + \gamma_2$.

The linear function that gives maximally smooth predictions on one line also gives the maximally non-smooth predictions on the other line. One way to remedy this restrictive situation is to introduce slack variables $\boldsymbol{\xi} = (\xi_i)_{i\in L\cup U}$ in Eqn. 13 with an ℓ_2 penalty, and instead solve: $f^\star = \operatorname{argmin}_{\boldsymbol{f}\in\mathcal{H}_\mathcal{A},\boldsymbol{\xi}}\, \gamma_1\|f\|^2_{\mathcal{H}_\mathcal{A}} + \gamma_2(\boldsymbol{f}+\boldsymbol{\xi})^T M(\boldsymbol{f}+\boldsymbol{\xi}) + \mu\|\boldsymbol{\xi}\|^2 + \sum_{i\in L} V(y_i, f(\boldsymbol{x}_i))$. Re-parameterizing $\mathbf{g} = \boldsymbol{f} + \boldsymbol{\xi}$, we can re-write the above problem as, $f^\star = \operatorname{argmin}_{f\in\mathcal{H}_\mathcal{A}, g\in\mathcal{H}_\mathcal{I}}\, \gamma_1\|f\|^2_{\mathcal{H}_\mathcal{A}} + \gamma_2\|g\|^2_{\mathcal{H}_\mathcal{I}} + \mu\|\boldsymbol{f} - \mathbf{g}\|^2 + \sum_{i\in L} V(y_i, f(\boldsymbol{x}_i))$, which may be viewed as a variant of the co-regularization problem in Eqn. 2 where empirical loss is measured for f alone. Thus, this motivates the view that CoMR adds extra slack variables in the MR objective function to better fit the intrinsic regularizer. Figure 2 shows that CoMR achieves better separation between classes on the LINES dataset.

Figure 2. Decision boundaries of MR and CoMR (using the quadratic hinge loss) on the LINES dataset



(a) MR (b) CoMR

4.3. Experiments

In this section, we compare MR and CoMR. Similar to our construction of the co-regularization kernel, (Sindhwani et al., 2005a) provide a data-dependent kernel that reduces manifold regularization to standard supervised learning in an associated RKHS. We write the manifold regularization kernel in the following form,

$$\tilde{k}^{mr}(\boldsymbol{x}, \boldsymbol{z}) = \bar{s}(\boldsymbol{x}, \boldsymbol{z}) - \bar{\mathbf{d}}^T_{\boldsymbol{x}} \bar{H} \bar{\mathbf{d}}_{\boldsymbol{z}} \qquad (14)$$

where we have, $\bar{s} = \gamma_1^{-1}k^1(\boldsymbol{x}, \boldsymbol{z})$, $\bar{\mathbf{d}}_{\boldsymbol{x}} = \gamma_1^{-1}k^1_{U\boldsymbol{x}}$ and $\bar{H} = \left(\gamma_1^{-1}\bar{K}^1 + \gamma_2^{-1}\bar{K}^2\right)^{-1}$, where $\bar{K}^1$ is the Gram Matrix of $k^1 = k_\mathcal{A}$ over labeled and unlabeled examples, and $\bar{K}^2 = M^\dagger$. We use the notation $\bar{s}, \bar{\mathbf{d}}, \bar{H}$ so that the kernel can be easily compared with corresponding quantities in the co-regularization kernel Eqn. 7. In this section we empirically compare this kernel with the co-regularization kernel of Eqn. 7 for exactly the same choice of k^1, k^2. Semi-supervised classification experiments were performed on 5 datasets described in table 1.

Table 1. Datasets with d features and c classes. 10 random data splits were created with l labeled, u unlabeled, t test, and v validation examples.

Dataset	d	c	l	u	t	v
LINES	2	2	2	500	250	250
G50C	50	2	50	338	112	50
USPST	256	10	50	1430	477	50
COIL20	241	20	40	1320	40	40
PCMAC	7511	2	50	1385	461	50

The LINES dataset is a variant of the two-dimensional problem shown in Figure 2 where we added random noise around the two perpendicular lines. The G50C, USPST, COIL20, and PCMAC datasets are well known and have frequently been used for empirical studies in semi-supervised learning literature. They were used for benchmarking manifold regularization in (Sindhwani et al., 2005a) against a number of competing methods. G50C is an artificial dataset generated from two unit covariance normal distributions with equal probabilities. The class means are adjusted so that the Bayes error is 5%. COIL20 consists of 32×32 gray scale images of 20 objects viewed from varying angles. USPST is taken from the test subset of the USPS dataset of images containing 10 classes of handwritten digits. PCMAC is used to setup binary text categorization problems drawn from the 20-newsgroups dataset.

For each of the 5 datasets, we constructed random splits into labeled, unlabeled, test and validation sets. The sizes of these sets are given in table 1. For all datasets except LINES, we used Gaussian am-

Table 2. Error Rates (in percentage) on Test Data

DATASET	MR	CoMR
LINES	7.7 (1.2)	**1.0** (1.5)
G50C	5.8 (2.8)	5.5 (2.3)
USPST	18.2 (1.5)	**14.1** (1.6)
COIL20	23.8 (11.1)	**14.8** (8.8)
PCMAC	11.9 (3.4)	**8.9** (2.6)

Table 3. Error Rates (in percentage) on Unlabeled Data

DATASET	MR	CoMR
LINES	7.5 (1.0)	**1.3** (2.0)
G50C	6.6 (0.8)	6.9 (1.0)
USPST	18.6 (1.4)	**13.3** (1.0)
COIL20	37.5 (6.0)	**14.8** (3.3)
PCMAC	11.0 (2.4)	**9.4** (1.9)

bient kernels $k^1(\boldsymbol{x}, \boldsymbol{z}) = \exp(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2\sigma^2})$, and intrinsic graph kernel whose gram matrix is of the form $K_2 = (M^p + 10^{-6}I)^{-1}$. Here, M is the normalized Graph Laplacian constructed using nn nearest neighbors and p is an integer. These parameters are tabulated in Table 4 for reproducibility. For more details on these parameters see (Sindhwani et al., 2005a).

We chose squared loss for $V(\cdot,\cdot)$. Manifold regularization with this choice is also referred to as Laplacian RLS and empirically performs as well as Laplacian SVMs. For multi-class problems, we used the one-versus-rest strategy. γ_1, γ_2 were varied on a grid of values: $10^{-6}, 10^{-4}, 10^{-2}, 1, 10, 100$ and chosen with respect to validation error. The chosen parameters are also reported in Table 4. Finally, we evaluated the MR solution and the ambient component of CoMR on an unseen test set. In Tables 2 and 3 we report the mean and standard deviation of error rates on test and unlabeled examples with respect to 10 random splits. We performed a paired t-test at 5% significance level to assess the statistical significance of the results. Results shown in bold are statistically significant.

Our experimental protocol makes MR and CoMR exactly comparable. We find that CoMR gives major empirical improvements over MR on all datasets except G50C where both methods approach the Bayes error rate.

5. Conclusion

In this paper, we have constructed a single, new RKHS in which standard supervised algorithms are immediately turned into multi-view semi-supervised learners. This construction brings about significant theoretical simplifications and algorithmic benefits, which we have demonstrated in the context of generalization analysis and manifold regularization respectively.

Table 4. Parameters Used. Note $\mu = 1$ for CoMR. Linear kernel was used for LINES dataset.

DATASET	nn	σ	p	MR γ_1, γ_2	CoMR γ_1, γ_2
LINES	10	–	1	$0.01, 10^{-6}$	$10^{-4}, 100$
G50C	50	17.5	5	$1, 100$	$10, 10$
USPST	10	9.4	2	$0.01, 0.01$	$10^{-6}, 10^{-4}$
COIL20	2	0.6	1	$10^{-4}, 10^{-6}$	$10^{-6}, 10^{-6}$
PCMAC	50	2.7	5	$10, 100$	$1, 10$

References

Bartlett, P. L., Bousquet, O., & Mendelson, S. (2002). Localized rademacher complexities. *COLT 02* (pp. 44–58). Springer-Verlag.

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, *7*, 2399–2434.

Bertinet, A., & Thomas-Agnan, C. (2004). *Reproducing kernel hilbert spaces in probability and statistics*. Kluwer Academic Publishers.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *COLT*.

Boucheron, S., Bousquet, O., & Lugosi, G. (2005). Theory of classification: a survey of some recent advances. *ESAIM: P&S*, *9*, 323–375.

Brefeld, U., Gärtner, T., Scheffer, T., & Wrobel, S. (2006). Efficient co-regularised least squares regression. *ICML* (pp. 137–144).

Farquhar, J. D. R., Hardoon, D. R., Meng, H., Taylor, J. S., & Szedmák, S. (2005). Two view learning: SVM-2K, theory and practice. *NIPS*.

Krishnapuram, B., Williams, D., Xue, Y., & A. Hartemink, L. C. (2005). On semi-supervised classification. *NIPS*.

Rosenberg, D., & Bartlett, P. L. (2007). The Rademacher complexity of co-regularized kernel classes. *AISTATS*.

Sindhwani, V., Niyogi, P., & Belkin, M. (2005a). Beyond the point cloud: From transductive to semi-supervised learning. *ICML*.

Sindhwani, V., Niyogi, P., & Belkin, M. (2005b). A co-regularization approach to semi-supervised learning with multiple views. *ICML Workshop on Learning in Multiple Views* (pp. 824–831).

Yu, S., Krishnapuram, B., Rosales, R., Steck, H., & Rao, R. (2008). Bayesian co-training. *NIPS*.

A. Proof of Theorem 2.2

This theorem generalizes Theorem 5 in (Bertinet & Thomas-Agnan, 2004).

The Product Hilbert Space We begin by introducing the product space,

$$\mathcal{F} = \mathcal{H}^1 \times \mathcal{H}^2 = \{(f^1, f^2) : f^1 \in \mathcal{H}^1, f^2 \in \mathcal{H}^2\},$$

and an inner product on $\mathcal{F}$ defined by,

$$\langle (f^1, f^2), (g^1, g^2) \rangle_{\mathcal{F}} = \gamma_1 \langle f^1, g^1 \rangle_{\mathcal{H}^1} + \gamma_2 \langle f^2, g^2 \rangle_{\mathcal{H}^2} + \mu \sum_{i \in U} \left(f^1(\boldsymbol{x}_i) - f^2(\boldsymbol{x}_i)\right)\left(g^1(\boldsymbol{x}_i) - g^2(\boldsymbol{x}_i)\right) \quad (15)$$

It's straightforward to check that $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ is a valid inner product. Moreover, we have the following:

Lemma A.1. *$\mathcal{F}$ is a Hilbert space.*

Proof. (Sketch.) We need to show that $\mathcal{F}$ is complete. Let (f_n^1, f_n^2) be a Cauchy sequence in $\mathcal{F}$. Then f_n^1 is Cauchy in $\mathcal{H}^1$ and f_n^2 is Cauchy in $\mathcal{H}^2$. By the completeness of $\mathcal{H}^1$ and $\mathcal{H}^2$, we have $f_n^1 \to f^1$ in $\mathcal{H}^1$ and $f_n^2 \to f^2$ in $\mathcal{H}^2$, for some $(f^1, f^2) \in \mathcal{F}$. Since $\mathcal{H}^1$ and $\mathcal{H}^2$ are RKHSs, convergence in norm implies pointwise convergence, and thus the co-regularization part of the distance also goes to zero. □

$\tilde{\mathcal{H}}$ is a Hilbert Space Recall the definition of $\tilde{\mathcal{H}}$ in Eqn. 4. Define the map $u : \mathcal{F} \to \tilde{\mathcal{H}}$ by $u(f^1, f^2) = \frac{1}{2}\left(f^1 + f^2\right)$. The map's kernel $N := u^{-1}(0)$ is a closed subspace of $\mathcal{F}$, and thus its orthogonal complement $N^\perp$ is also a closed subspace. We can consider $N^\perp$ as a Hilbert space with the inner product that is the natural restriction of $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ to $N^\perp$. Define $v : N^\perp \to \tilde{\mathcal{H}}$ as the restriction of u to $N^\perp$. Then v is a bijection, and we define an inner product on $\tilde{\mathcal{H}}$ by

$$\langle f, g \rangle_{\tilde{\mathcal{H}}} = \langle v^{-1}(f), v^{-1}(g) \rangle_{\mathcal{F}}. \quad (16)$$

We conclude that $\tilde{\mathcal{H}}$ is a Hilbert space isomorphic to $N^\perp$.

The Co-Regularization Norm Fix any $f \in \tilde{\mathcal{H}}$, and note that $u^{-1}(f) = \left\{v^{-1}(f) + n \mid n \in N\right\}$. Since $v^{-1}(f)$ and N are orthogonal, it's clear by the Pythagorean theorem that $v^{-1}(f)$ is the element of $u^{-1}(f)$ with minimum norm. Thus

$$||f||^2_{\tilde{\mathcal{H}}} = ||v^{-1}(f)||^2_{\mathcal{F}} = \min_{(f^1, f^2) \in u^{-1}(f)} ||(f^1, f^2)||^2_{\mathcal{F}}$$

We see that the inner product on $\tilde{\mathcal{H}}$ induces the norm claimed in the theorem statement.

We next check the two conditions for validity of an RKHS (see Definition 2.1).

(a) $\tilde{k}(\boldsymbol{z}, \cdot) \in \tilde{\mathcal{H}} \quad \forall \boldsymbol{z} \in \mathcal{X}$ Recall from Eqn. 7 that the co-regularization kernel is defined as

$$\begin{aligned}\tilde{k}(\boldsymbol{x}, \boldsymbol{z}) &= \gamma_1^{-1} k^1(\boldsymbol{x}, \boldsymbol{z}) + \gamma_2^{-1} k^2(\boldsymbol{x}, \boldsymbol{z}) \\ &\quad -\mu \left(\gamma_1^{-1} \mathbf{k}^1_{U\boldsymbol{x}} - \gamma_2^{-1} \mathbf{k}^2_{U\boldsymbol{x}}\right)^T \beta_{\boldsymbol{z}}\end{aligned}$$

where $\beta_{\boldsymbol{z}} = H\mathbf{d}_z = (I + \mu S)^{-1}\left(\gamma_1^{-1}\mathbf{k}^1_{U\boldsymbol{z}} - \gamma_2^{-1}\mathbf{k}^2_{U\boldsymbol{z}}\right)$. Define $h^1(\boldsymbol{x}) = \gamma_1^{-1} k^1(\boldsymbol{x}, \boldsymbol{z}) - \mu\gamma_1^{-1}\mathbf{k}^1_{U\boldsymbol{x}}\beta_{\boldsymbol{z}}$ and $h^2(\boldsymbol{x}) = \gamma_2^{-1} k^2(\boldsymbol{x}, \boldsymbol{z}) + \mu\gamma_2^{-1} k^2(\boldsymbol{x}, U)\beta_{\boldsymbol{z}}$. Note that, $h^1 \in \operatorname{span}\left\{k^1(\boldsymbol{z}, \cdot), k^1(\boldsymbol{x}_1, \cdot), \ldots, k^1(\boldsymbol{x}_u, \cdot)\right\} \subset \mathcal{H}^1$, and similarly, $h^2 \in \mathcal{H}^2$. It's clear that $\tilde{k}(\boldsymbol{z}, \cdot) = \left[h^1(\cdot) + h^2(\cdot)\right]$, and thus $\tilde{k}(\boldsymbol{z}, \cdot) \in \tilde{\mathcal{H}}$.

(b) Reproducing Property For convenience, we collect some basic properties of h^1 and h^2 in the following lemma:

Lemma A.2 (Properties of h^1 and h^2)**.** *Writing $h^1(U)$ for the column vector with entries $h^1(\boldsymbol{x}_i)\ \forall i \in U$, and similarly for other functions on $\mathcal{X}$, we have the following:*

$$\left\langle f^1, h^1 \right\rangle_{\mathcal{H}^1} = \gamma_1^{-1} f^1(\boldsymbol{z}) - \mu\gamma_1^{-1} f^1(U)^T \beta_{\boldsymbol{z}} \quad (17)$$

$$\left\langle f^2, h^2 \right\rangle_{\mathcal{H}^2} = \gamma_2^{-1} f^2(\boldsymbol{z}) + \mu\gamma_2^{-1} f^2(U)^T \beta_{\boldsymbol{z}} \quad (18)$$

$$h^1(U) - h^2(U) = \beta_{\boldsymbol{z}} \quad (19)$$

Proof. The first two equations follow from the definitions of h^1 and h^2 and the reproducing kernel property. The last equation is derived as follows:

$$\begin{aligned}h^1(U) - h^2(U) =& \gamma_1^{-1} k^1(U, \boldsymbol{z}) - \mu\gamma_1^{-1} k^1(U, U)\beta_{\boldsymbol{z}} \\ & - \gamma_2^{-1} k^2(U, \boldsymbol{z}) - \mu\gamma_2^{-1} k^2(U, U)\beta_{\boldsymbol{z}} \\ =& \mathbf{d}_z - \mu S(I + \mu S)^{-1}\mathbf{d}_z \\ =& \left(I - \mu S(I + \mu S)^{-1}\right)\mathbf{d}_z \\ =& (I + \mu S)^{-1}\mathbf{d}_z = \beta_{\boldsymbol{z}}\end{aligned}$$

where the last line follows from the Sherman-Morrison-Woodbury inversion formula. □

Since $\tilde{k}(\boldsymbol{z}, \cdot) = h_1(\cdot) + h_2(\cdot)$, it is clear that $(h_1, h_2) = v^{-1}\left(\tilde{k}(\boldsymbol{z}, \cdot)\right) + n$, for some $n \in N$. Since $v^{-1}(f) \in N^\perp$, we have

$$\begin{aligned}\left\langle f, \tilde{k}(\boldsymbol{z}, \cdot)\right\rangle_{\tilde{\mathcal{H}}} =& \left\langle v^{-1}(f), v^{-1}(\tilde{k}(\boldsymbol{z}, \cdot))\right\rangle_{\mathcal{F}} \\ =& \left\langle v^{-1}(f), (h_1, h_2) - n\right\rangle_{\mathcal{F}} \\ =& \left\langle v^{-1}(f), (h_1, h_2)\right\rangle_{\mathcal{F}} \\ =& \gamma_1 \left\langle f^1, h^1\right\rangle_{\mathcal{H}_1} + \gamma_2 \left\langle f_2, h_2\right\rangle_{\mathcal{H}_2} \\ & + \mu\left[h^1(U) - h^2(U)\right]^T\left[f^1(U) - f^2(U)\right] \\ =& f^1(\boldsymbol{z}) + f^2(\boldsymbol{z}) - \mu\left[f^1(U) - f^2(U)\right]^T \beta_{\boldsymbol{z}} \\ & + \mu\left[f^1(U) - f^2(U)\right]^T \beta_{\boldsymbol{z}} \text{ (from A.2)} \\ & = f(\boldsymbol{z})\end{aligned}$$

□