Vector-valued Manifold Regularization

Hà Quang Minh MINH.HAQUANG@IIT.IT
Italian Institute of Technology, Via Morego 30, Genoa 16163, Italy

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

Abstract

We consider the general problem of learning an unknown functional dependency, $f : \mathcal{X} \mapsto \mathcal{Y}$, between a structured input space $\mathcal{X}$ and a structured output space $\mathcal{Y}$, from labeled and unlabeled examples. We formulate this problem in terms of data-dependent regularization in Vector-valued Reproducing Kernel Hilbert Spaces (Micchelli & Pontil, 2005) which elegantly extend familiar scalar-valued kernel methods to the general setting where $\mathcal{Y}$ has a Hilbert space structure. Our methods provide a natural extension of Manifold Regularization (Belkin et al., 2006) algorithms to also exploit output inter-dependencies while enforcing smoothness with respect to input data geometry. We propose a class of matrix-valued kernels which allow efficient implementations of our algorithms via the use of numerical solvers for Sylvester matrix equations. On multi-label image annotation and text classification problems, we find favorable empirical comparisons against several competing alternatives.

1. Introduction

The statistical and algorithmic study of regression and binary classification problems has formed the bedrock of modern machine learning. Motivated by new applications, data characteristics, and scalability requirements, several generalizations and extensions of these canonical settings have been vigorously pursued in recent years. We point out two particularly dominant threads of research: (1) *semi-supervised learning*, i.e., learning from unlabeled examples by exploiting the geometric structure of the marginal probability distribution over the input space, and (2) *structured multi-output prediction*, i.e., learning to simultaneously predict a collection of output variables by exploiting their inter-dependencies. We point the reader to Chapelle et al. (2006) and Bakir et al. (2007) for several representative papers on semi-supervised learning and structured prediction respectively. In this paper, we consider a problem at the intersection of these threads: non-parametric estimation of a *vector-valued function*, $f : \mathcal{X} \mapsto \mathcal{Y}$, from labeled and unlabeled examples.

Our starting point is multivariate regression in a regularized least squares (RLS) framework (see, e.g., Brown & Zidek (1980)), which is arguably the classical precursor of much of the modern literature on structured prediction, multi-task learning, multi-label classification and related themes that attempt to exploit output structure. We adopt the formalism of Vector-valued Reproducing Kernel Hilbert Spaces (Micchelli & Pontil, 2005) to pose function estimation problems naturally in an RKHS of $\mathcal{Y}$-valued functions, where $\mathcal{Y}$ in general can be an infinite-dimensional (Hilbert) space. We derive an abstract system of functional linear equations that gives the solution to a generalized Manifold Regularization (Belkin et al., 2006) framework for vector-valued semi-supervised learning. For multivariate problems with n output variables, the kernel $K(\cdot, \cdot)$ associated with a vector-valued RKHS is matrix-valued, i.e., for any $x, z \in \mathcal{X}$, $K(x, z) \in \mathbb{R}^{n \times n}$. We show that a natural choice for a matrix-valued kernel leads to a Sylvester Equation, whose solution can be obtained relatively efficiently using techniques in numerical linear algebra. This leads to a vector-valued Laplacian Regularized Least Squares (Laplacian RLS) model that learns not only from the geometry of unlabeled data Belkin et al. (2006) but also from dependencies among output variables estimated using an output graph Laplacian. We

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

find encouraging empirical results with this approach on semi-supervised multi-label classification problems, in comparison to several recently proposed alternatives. We begin this paper with relevant background material on Manifold Regularization and multivariate RLS. Throughout the paper, we draw attention to mathematical correspondences between scalar and vector-valued settings.

2. Background

Let us recall the familiar regression and classification setting where $\mathcal{Y} = \mathbb{R}$. Let $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a standard kernel with an associated RKHS family of functions $\mathcal{H}_k$. Given a collection of labeled examples, $\{x_i, y_i\}_{i=1}^l$, kernel-based prediction methods set up a Tikhonov regularization problem,

$$f^\star = \arg\min_{f \in \mathcal{H}_k} \frac{1}{l} \sum_{i=1}^{l} V(y_i, f(x_i)) + \gamma \|f\|_k^2 \qquad (1)$$

where the choice $V(t, y) = (t - y)^2$ leads to Regularized Least Squares (RLS) while $V(t, y) = \max(0, 1 - yt)$ leads to the SVM algorithm. By the classical Representer theorem (Schölkopf & Smola, 2002), this family of algorithms reduces to estimation of finite-dimensional coefficients, $\mathbf{a} = [\alpha_1, \ldots, \alpha_l]^T$, for a minimizer that can be shown to have the form $f^\star(x) = \sum_{i=1}^l \alpha_i k(x, x_i)$. In particular, RLS reduces to solving the linear system, $[G_k^l + \gamma l I_l]\mathbf{a} = \mathbf{y}_l$ where $\mathbf{y}_l = [y_1 \ldots y_l]^T$, I_l is the $l \times l$ identity matrix and G_k^l denotes the Gram matrix of the kernel over the labeled data, i.e., $(G_k^l)_{ij} = k(x_i, x_j)$. Let us now review two extensions of this algorithm: first for semi-supervised learning, and then for multivariate problems where $\mathcal{Y} = \mathbb{R}^n$.

Semi-supervised learning typically proceeds by making assumptions such as smoothness of the prediction function with respect to an underlying low-dimensional data manifold or presence of clusters as detected using a relatively large set of u unlabeled examples, $\{x_i\}_{i=l+1}^{N=l+u}$. We will use the notation $N = l + u$. In Manifold Regularization (Belkin et al., 2006), a nearest neighbor graph, W, is constructed, which serves as a discrete probe for the geometric structure of the data. The Laplacian L of this graph provides a natural intrinsic measure of data-dependent smoothness:

$$\mathbf{f}^T L \mathbf{f} = \frac{1}{2} \sum_{i,j=1}^{N} W_{ij} (f(x_i) - f(x_j))^2$$

where $\mathbf{f} = [f(x_1) \ldots f(x_N)]$. Thus, it is natural to extend (1) as follows,

$$f^\star = \arg\min_{f \in \mathcal{H}_k} \frac{1}{l} \sum_{i=1}^{l} (y_i - f(x_i))^2 + \gamma_A \|f\|_k^2 + \gamma_I \mathbf{f}^T L \mathbf{f} \qquad (2)$$

where γ_A, γ_I are referred to as ambient and intrinsic regularization parameters. By using reproducing properties of $\mathcal{H}_k$, the Representer theorem can correspondingly be extended to show that the minimizer has the form, $f^\star(x) = \sum_{i=1}^N \alpha_i k(x, x_i)$ involving both labeled and unlabeled data. The Laplacian RLS algorithm estimates $\mathbf{a} = [\alpha_1 \ldots \alpha_N]^T$ by solving the linear system $[J_l^N G_k^N + l\gamma_I L G_k^N + l\gamma_A I_N]\mathbf{a} = \mathbf{y}$ where G_k^N is the Gram matrix of k with respect to both labeled and unlabeled examples, I_N is the $N \times N$ identity matrix, J_l^N is an $N \times N$ diagonal matrix with first l diagonal entries equaling 1 and the rest being 0 valued, and $\mathbf{y}$ is the $N \times 1$ label vector with $y_i = 0, i > l$. Laplacian RLS and Laplacian SVM tend to give similar empirical performance (Sindhwani et al., 2005).

Consider now two natural approaches to extending Laplacian RLS for the multivariate case $\mathcal{Y} = \mathbb{R}^n$. Let $f = (f_1 \ldots f_n)$ be components of a vector-valued function where each $f_i \in \mathcal{H}_k$. Let the j^{th} output label of the x_i be denoted as y_{ij}. Then, one formulation for multivariate LapRLS is to solve,

$$f^\star = \arg\min_{\substack{f_j \in \mathcal{H}_k \\ 1 \le j \le n}} \frac{1}{l} \sum_{i=1}^{l} \sum_{j=1}^{n} (y_{ij} - f_j(x_i))^2 + \gamma_A \|f_j\|_k^2 + \gamma_I trace[\mathbf{F}^T L \mathbf{F}] \qquad (3)$$

where $\mathbf{F}_{ij} = f_j(x_i)$, $1 \le i \le N, 1 \le j \le n$. Let $\boldsymbol{\alpha}$ be an $N \times n$ matrix of expansion coefficients, i.e., the minimizers have the form $f_j(x) = \sum_{i=1}^N \alpha_{ij} k(x_i, x)$. It is easily seen that the solution is given by,

$$[J_l^N G_k^N + l\gamma_I L G_k^N + l\gamma_A l I_N]\boldsymbol{\alpha} = \mathbf{Y} \qquad (4)$$

where $\mathbf{Y}$ is the label matrix with $Y_{ij} = 0$ for $i > l$ and all j. It is clear that this multivariate solution is equivalent to learning each output independently – ignoring prior knowledge such as the availability of a similarity graph W_{out} over output variables. Such prior knowledge can naturally be incorporated by adding a smoothing term to (3) which, for example, enforces f_i to be close to f_j in the RKHS norm $\|\cdot\|_k$ if output i is similar to output j, i.e., $(W_{out})_{ij}$ is sufficiently large. We defer this development to later in the paper as both these two solutions are special cases of a broader vector-valued RKHS framework for Laplacian RLS where they correspond to certain choices of a matrix-valued kernel. We first give a self-contained review of the language of vector-valued RKHS in the following section.

3. Vector-Valued RKHS

The study of RKHS has been extended to vector-valued functions and further developed and applied in machine learning (see (Carmeli et al., 2006; Micchelli & Pontil, 2005; Caponnetto et al., 2008) and references therein). In the following, denote by $\mathcal{X}$ a nonempty set, $\mathcal{Y}$ a real Hilbert space with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$, $\mathcal{L}(\mathcal{Y})$ the Banach space of bounded linear operators on $\mathcal{Y}$.

Let $\mathcal{Y}^{\mathcal{X}}$ denote the vector space of all functions $f : \mathcal{X} \to \mathcal{Y}$. A function $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is said to be an **operator-valued positive definite kernel** if for each pair $(x, z) \in \mathcal{X} \times \mathcal{X}$, $K(x, z) \in \mathcal{L}(\mathcal{Y})$ is a self-adjoint operator and

$$\sum_{i,j=1}^{N} \langle y_i, K(x_i, x_j) y_j \rangle_{\mathcal{Y}} \geq 0 \tag{5}$$

for every finite set of points $\{x_i\}_{i=1}^N$ in $\mathcal{X}$ and $\{y_i\}_{i=1}^N$ in $\mathcal{Y}$. As in the scalar case, given such a K, there exists a unique $\mathcal{Y}$-valued RKHS $\mathcal{H}_K$ with reproducing kernel K. The construction of the space $\mathcal{H}_K$ proceeds as follows. For each $x \in X$ and $y \in \mathcal{Y}$, we form a function $K_x y = K(., x)y \in \mathcal{Y}^X$ defined by

$$(K_x y)(z) = K(z, x)y \quad \text{for all} \quad z \in X.$$

Consider the set $\mathcal{H}_0 = \text{span}\{\mathrm{K_x y} \mid \mathrm{x} \in \mathrm{X},\ \mathrm{y} \in \mathcal{Y}\} \subset \mathcal{Y}^{\mathrm{X}}$. For $f = \sum_{i=1}^N K_{x_i} w_i$, $g = \sum_{i=1}^N K_{z_i} y_i \in \mathcal{H}_0$, we define

$$\langle\, f, g \,\rangle_{\mathcal{H}_K} = \sum_{i,j=1}^{N} \langle w_i, K(x_i, z_j) y_j \rangle_{\mathcal{Y}}.$$

Taking the closure of $\mathcal{H}_0$ gives us the Hilbert space $\mathcal{H}_K$. The **reproducing property** is

$$\langle f(x), y \rangle_{\mathcal{Y}} = \langle f, K_x y \rangle_{\mathcal{H}_K} \quad \text{for all} \quad f \in \mathcal{H}_K. \tag{6}$$

As in the scalar case, applying Cauchy-Schwarz inequality gives

$$|\langle f(x), y \rangle_{\mathcal{Y}}| \leq \sqrt{||K(x, x)||}\, ||f||_{\mathcal{H}_K} ||y||_{\mathcal{Y}}.$$

Thus for each $x \in X$, each $y \in \mathcal{Y}$, the evaluation operator $E_{x|y} : f \to \langle f(x), y \rangle_{\mathcal{Y}}$ is bounded as a linear operator from $\mathcal{H}_K$ to $\mathbb{R}$. As in the scalar case, the converse is true by the Riesz Representation Theorem.

Let $K_x : \mathcal{Y} \to \mathcal{H}_K$ be the linear operator with $K_x w$ defined as above, then

$$||K_x y||^2_{\mathcal{H}_K} = \langle K(x, x)y, y \rangle_{\mathcal{Y}} \leq ||K(x, x)||\ ||y||^2_{\mathcal{Y}},$$

which implies that

$$||K_x : \mathcal{Y} \to \mathcal{H}_K|| \leq \sqrt{||K(x, x)||},$$

so that K_x is a bounded operator for each $x \in X$. Let $K_x^* : \mathcal{H}_K \to \mathcal{Y}$ be the adjoint operator of K_x, then from (6), we have

$$f(x) = K_x^* f \qquad \text{for all} \quad x \in X, f \in \mathcal{H}_K. \tag{7}$$

From this we deduce that for all $x \in X$ and all $f \in \mathcal{H}_K$,

$$||f(x)||_{\mathcal{Y}} \leq ||K_x^*||\ ||f||_{\mathcal{H}_K} \leq \sqrt{||K(x, x)||}\ ||f||_{\mathcal{H}_K},$$

that is for each $x \in X$, the evaluation operator $E_x : \mathcal{H}_K \to \mathcal{Y}$ defined by $E_x f = K_x^* f$ is a bounded linear operator. In particular, if $\kappa = \sup_{x \in X} \sqrt{||K(x, x)||} < \infty$, then $||f||_\infty = \sup_{x \in X} ||f(x)||_{\mathcal{Y}} \leq \kappa ||f||_{\mathcal{H}_K}$ for all $f \in \mathcal{H}_K$. In this paper, we will be concerned with kernels for which $\kappa < \infty$.

3.1. Vector-valued Regularized Least Squares

Let $\mathcal{Y}$ be a separable Hilbert space. Let $\mathbf{z} = \{(x_i, y_i)_{i=1}^m\}$ be the given labeled data. Consider now the vector-valued version of regularized least square algorithm in $\mathcal{H}_K$, with $\gamma > 0$:

$$f_{\mathbf{z},\gamma} = \arg \min_{f \in \mathcal{H}_K} \frac{1}{m} \sum_{i=1}^{m} ||f(x_i) - y_i||^2_{\mathcal{Y}} + \gamma ||f||^2_{\mathcal{H}_K}. \tag{8}$$

From Inverse Problems literature, recall the familiar Tikhonov Regularization form in Hilbert spaces: $\arg\min_{x \in \mathcal{H}_1} \|Ax - b\|^2_{\mathcal{H}_2} + \gamma \|x\|^2_{\mathcal{H}_1}$, where $\mathcal{H}_1, \mathcal{H}_2$ are Hilbert Spaces and $A : \mathcal{H}_1 \mapsto \mathcal{H}_2$ is a bounded linear operator. To cast (8) into this familiar form, we can consider an approach as in Caponnetto & Vito (2007). First, we set up the sampling operator $S_{\mathbf{x}} : \mathcal{H}_K \to \mathcal{Y}^m$ defined by $S_{\mathbf{x}}(f) = (f(x_1), \ldots, f(x_m))$. By definition, we have for any $f \in \mathcal{H}_K$ and $\mathbf{y} = (y_1, \ldots, y_m) \in \mathcal{Y}^m$, $\langle S_{\mathbf{x}} f, \mathbf{y} \rangle_{\mathcal{Y}^m} = \sum_{i=1}^m \langle S_{x_i} f, y_i \rangle_{\mathcal{Y}} = \sum_{i=1}^m \langle f, S_{x_i}^* y_i \rangle_{\mathcal{H}_K} = \sum_{i=1}^m \langle f, K_{x_i} y_i \rangle_{\mathcal{H}_K} = \langle f, \sum_{i=1}^m K_{x_i} y_i \rangle_{\mathcal{H}_K}$. It follows that the adjoint operator $S_{\mathbf{x}}^* : \mathcal{Y}^m \to \mathcal{H}_K$ is given by $S_{\mathbf{x}}^* \mathbf{y} = S_{\mathbf{x}}^*(y_1, \ldots, y_m) = \sum_{i=1}^m K_{x_i} y_i$, and the operator $S_{\mathbf{x}}^* S_{\mathbf{x}} : \mathcal{H}_K \to \mathcal{H}_K$ is given by $S_{\mathbf{x}}^* S_{\mathbf{x}} f = \sum_{i=1}^m K_{x_i} f(x_i)$. We can now cast expression (8) into the familiar Tikhonov form,

$$f_{\mathbf{z},\gamma} = \arg \min_{f \in \mathcal{H}_K} \frac{1}{m} ||S_{\mathbf{x}} f - \mathbf{y}||^2_{\mathcal{Y}^m} + \gamma ||f||^2_{\mathcal{H}_K}.$$

This problem has a unique solution, given by

$$f_{\mathbf{z},\gamma} = (S_{\mathbf{x}}^* S_{\mathbf{x}} + m\gamma I)^{-1} S_{\mathbf{x}}^* \mathbf{y} \tag{9}$$

This function has the explicit form $f_{\mathbf{z},\gamma} = \sum_{i=1}^m K_{x_i} a_i$ with $f_{\mathbf{z},\gamma}(x) = \sum_{i=1}^m K(x, x_i) a_i$, where the vectors

$a_i \in \mathcal{Y}$ satisfy the m linear equations

$$\sum_{j=1}^{m} K(x_i, x_j)a_j + m\gamma a_i = y_i. \tag{10}$$

for $1 \le i \le m$. This result was first reported in Micchelli & Pontil (2005). It is a special case of our Proposition 1 below. In the scalar case $\mathcal{Y} = \mathbb{R}$ it specializes to the well known Regularized Least squares solution reviewed in Section 2. Note that in the finite-dimensional case $\mathcal{Y} = \mathbb{R}^n$, the kernel function is matrix-valued; and unless it is diagonal, outputs are not treated independently.

4. Vector-valued Manifold Regularization

Let $\mathbf{z} = \{(x_i, y_i)_{i=1}^{l}\} \cup \{(x_i)_{i=l+1}^{u+l}\}$ be the given set of labeled and unlabeled data. Let $M : \mathcal{Y}^{u+l} \to \mathcal{Y}^{u+l} \in \mathcal{L}(\mathcal{Y}^{u+l})$ be a symmetric, positive operator, that is $\langle y, My\rangle_{\mathcal{Y}^{u+l}} \ge 0$ for all $y \in \mathcal{Y}^{u+l}$. Here $\mathcal{Y}^{u+l}$ is the usual $u+l$-direct product of $\mathcal{Y}$, with inner product

$$\langle (y_1, \ldots, y_{u+l}), (w_1, \ldots, w_{u+l})\rangle_{\mathcal{Y}^{u+l}} = \sum_{i=1}^{u+l} \langle y_i, w_i\rangle_{\mathcal{Y}}.$$

For $f \in \mathcal{H}_K$, let $\mathbf{f} = (f(x_1), \ldots, f(x_{u+l})) \in \mathcal{Y}^{u+l}$. Consider the following optimization problem

$$f_{\mathbf{z},\gamma} = \arg\min_{f\in\mathcal{H}_K} \frac{1}{l}\sum_{i=1}^{l} V(f(x_i), y_i) + \gamma_A ||f||_K^2 + \gamma_I \langle \mathbf{f}, M\mathbf{f}\rangle_{\mathcal{Y}^{u+l}}, \tag{11}$$

where V is some convex error function, and $\gamma_A, \gamma_I > 0$ are regularization parameters.

4.1. Representer Theorem

Theorem 1. *The minimization problem (11) has a unique solution, given by $f_{\mathbf{z},\gamma} = \sum_{i=1}^{u+l} K_{x_i} a_i$ for some vectors $a_i \in \mathcal{Y}$, $1 \le i \le u+l$.*

Proof. Denote the right handside of (11) by $I_l(f)$. Then $I_l(f)$ is coercive and strictly convex in f, and thus has a unique minimizer. Let $\mathcal{H}_{K,\mathbf{x}} = \{\sum_{i=1}^{u+l} K_{x_i} y_i : \mathbf{y} \in \mathcal{Y}^{u+l}\}$. For $f \in \mathcal{H}_{K,\mathbf{x}}^{\perp}$, by the reproducing property, the sampling operator $S_{\mathbf{x}}$ satisfies

$$\langle S_{\mathbf{x}} f, \mathbf{y}\rangle_{\mathcal{Y}^{u+l}} = \langle f, \sum_{i=1}^{u+l} K_{x_i} y_i\rangle_{\mathcal{H}_K} = 0.$$

This holds true for all $\mathbf{y} \in \mathcal{Y}^{u+l}$, thus

$$S_{\mathbf{x}} f = (f(x_1), \ldots, f(x_{u+l})) = 0.$$

For an arbitrary $f \in \mathcal{H}_K$, consider the orthogonal decomposition $f = f_0 + f_1$, with $f_0 \in \mathcal{H}_{K,\mathbf{x}}$, $f_1 \in \mathcal{H}_{K,\mathbf{x}}^{\perp}$. Then, because $||f_0 + f_1||_{\mathcal{H}_K}^2 = ||f_0||_{\mathcal{H}_K}^2 + ||f_1||_{\mathcal{H}_K}^2$, the result just obtained shows that

$$I_l(f) = I_l(f_0 + f_1) \ge I_l(f_0)$$

with equality if and only if $||f_1||_{\mathcal{H}_K} = 0$, that is $f_1 = 0$. Thus the minimizer of (11) must lie in $\mathcal{H}_{K,\mathbf{x}}$. □

4.2. Vector-valued Laplacian RLS

For manifold regularization of vector-valued functions using least square error, the minimization problem is

$$f_{\mathbf{z},\gamma} = \arg\min \frac{1}{l}\sum_{i=1}^{l} ||f(x_i) - y_i||_{\mathcal{Y}}^2 + \gamma_A ||f||_K^2 + \gamma_I \langle \mathbf{f}, M\mathbf{f}\rangle_{\mathcal{Y}^{u+l}}, \tag{12}$$

for some $\gamma_A > 0, \gamma_I > 0$.

Proposition 1. *The minimization problem (12) has a unique solution $f_{\mathbf{z},\gamma} = \sum_{=1}^{u+l} K_{x_i} a_i$, where the vectors $a_i \in \mathcal{Y}$ satisfy the $u+l$ linear equations:*

$$(1 + l\gamma_I M)\sum_{j=1}^{u+l} K(x_i, x_j)a_j + l\gamma_A a_i = w_i \tag{13}$$

for $1 \le i \le l$, and

$$\gamma_I M \sum_{j=1}^{u+l} K(x_i, x_j)a_j + \gamma_A a_i = 0 \tag{14}$$

for $l+1 \le i \le u+l$.

Remark 1. We emphasize that in general, the vectors a_i's can be infinite dimensional, being elements of $\mathcal{Y}$. Thus (13) and (14) are *generally linear systems of functional equations*. The finite dimensional case, where we have a system of linear equations, is discussed below.

Proof. The fact that problem (12) has a unique solution comes from the observation that the functional being minimized is coercive and strictly convex. Using the sampling operator $S_{\mathbf{x}}$, we write

$$f_{\mathbf{z},\gamma} = \arg\min \frac{1}{l}||S_{\mathbf{x}_l} f - \mathbf{y}||_{\mathcal{Y}^l}^2 + \gamma_A ||f||_K^2 + \gamma_I \langle S_{\mathbf{x},u+l} f, M S_{\mathbf{x},u+l} f\rangle_{\mathcal{Y}^{u+l}}.$$

Differentiating gives

$$f_{\mathbf{z},\gamma} = (S_{\mathbf{x}_l}^* S_{\mathbf{x}_l} + l\gamma_A I + l\gamma_I S_{\mathbf{x},u+l}^* M S_{\mathbf{x},u+l})^{-1} S_{\mathbf{x}_l}^* \mathbf{y}.$$

This is equivalent to

$$(S_{\mathbf{x}_l}^* S_{\mathbf{x}_l} + l\gamma_A I + l\gamma_I S_{\mathbf{x},u+l}^* M S_{\mathbf{x},u+l}) f_{\mathbf{z},\gamma} = S_{\mathbf{x}_l}^* \mathbf{y}.$$

By definition of the sampling operators, this is

$$\sum_{i=1}^{l} K_{x_i} f_{\mathbf{z},\gamma}(x_i) + l\gamma_A f_{\mathbf{z},\gamma} + l\gamma_I \sum_{i=1}^{u+l} K_{x_i}(M\mathbf{f}_{\mathbf{z},\gamma})_i = \sum_{i=1}^{l} K_{x_i} y_i,$$

which we will rewrite as

$$f_{\mathbf{z},\gamma} = -\frac{\gamma_I}{\gamma_A} \sum_{i=1}^{u+l} K_{x_i}(M\mathbf{f}_{\mathbf{z},\gamma})_i + \sum_{i=1}^{l} K_{x_i} \frac{y_i - f_{\mathbf{z},\gamma}(x_i)}{l\gamma_A}.$$

This shows that there are vectors a_i's in $\mathcal{Y}$ such that $f_{\mathbf{z},\gamma} = \sum_{i=1}^{u+l} K_{x_i} a_i$.. Using the formulas $\mathbf{f}_{\mathbf{z},\gamma}(x_i) = \sum_{j=1}^{u+l} K(x_i, x_j) a_j$, $(M\mathbf{f}_{\mathbf{z},\gamma})_i = M \sum_{j=1}^{u+l} K(x_i, x_j) a_j$, we have for $1 \le i \le l$:

$$a_i = -\frac{\gamma_I}{\gamma_A} M \sum_{j=1}^{u+l} K(x_i, x_j) a_j + \frac{y_i - \sum_{j=1}^{u+l} K(x_i, x_j) a_j}{l\gamma_A},$$

which gives the formula

$$(1 + l\gamma_I M) \sum_{j=1}^{u+l} K(x_i, x_j) a_j + l\gamma_A a_i = y_i.$$

Similarly, $a_i = -\frac{\gamma_I}{\gamma_A} M \sum_{j=1}^{u+l} K(x_i, x_j) a_j$, for $l+1 \le i \le u+l$, implying $\gamma_I M \sum_{j=1}^{u+l} K(x_i, x_j) a_j + \gamma_A a_i = 0$. This completes the proof. □

Example 1. For $\mathcal{Y} = \mathbb{R}$, and choosing M to be the graph Laplacian, it can be easily verified that (13,14) specializes to (2). This gives the scalar case for manifold regularization as obtained in Belkin et al. (2006).

Example 2. For $\mathcal{Y} = \mathbb{R}^n$ (or any n-dimensional inner product space), the kernel $K(x, z)$ is an $n \times n$ matrix. Let G be the $Nn \times Nn$ block matrix, whose (i, j) block is the $n \times n$ matrix $K(x_i, x_j)$, where $N = l + u$. Let $\mathbf{a} = (a_1, \ldots, a_{u+l})$ and $\mathbf{y} = (y_1, \ldots, y_{u+l})$ be column vectors in $\mathbb{R}^{n(u+l)}$, with each $a_i, y_i \in \mathbb{R}^n$, and $y_{l+1} = \ldots = y_{u+l} = 0$. Then the system of linear equations (13), (14) becomes

$$\left(J_{nl}^{Nn} G + l\gamma_I M G + l\gamma_A I_{Nn}\right) \mathbf{a} = \mathbf{y}, \qquad (15)$$

where M now is a positive semidefinite matrix of size $n(u+l) \times n(u+l)$, $J_{nl}^{Nn} = \mathrm{diag}(1, \ldots, 1, 0, \ldots, 0)$, the $Nn \times Nn$ diagonal matrix where the first nl diagonal positions are 1, and the rest 0.

4.3. Vector-valued graph Laplacian

Let $G = (V, E)$ be an undirected graph of size $|V| = N$, with symmetric, nonnegative weight matrix W. For an $\mathbb{R}^n$-valued function $f = (f_1, \ldots, f_n)$, we define $\Delta f = (Lf_1, \ldots, Lf_n)$, where L is the scalar Laplacian. Thus one can set Δ to be a block diagonal matrix of size $nN \times nN$, with each block (i, i) being the scalar graph Laplacian of size $N \times N$. For $\mathbf{f} = (\mathbf{f}_k)_{k=1}^n$, where $\mathbf{f}_k = (f_k(x_1), \ldots, f_k(x_N))$, we have $\langle \mathbf{f}, \Delta\mathbf{f} \rangle_{\mathbb{R}^{nN}} = \sum_{k=1}^n \langle \mathbf{f}_k, L\mathbf{f}_k \rangle_{\mathbb{R}^N} = \frac{1}{2} \sum_{i,j=1}^N W_{ij} ||f(x_i) - f(x_j)||^2_{\mathbb{R}^n}$. It is clear that here we have $\Delta = L \otimes I_n$, where $\otimes$ denotes the Kronecker (tensor) matrix product (compare with the examples of the last section). It is straightforward to generalize to $\mathcal{Y}$-valued function spaces, where $\mathcal{Y}$ is any separable Hilbert space, by defining $(\Delta\mathbf{f})_i = \sum_{j=1}^N W_{ij}(f(x_i) - f(x_j))$, and

$$\langle \mathbf{f}, \Delta\mathbf{f} \rangle_{\mathcal{Y}^N} = \frac{1}{2} \sum_{i,j=1}^{N} W_{ij} ||f(x_i) - f(x_j)||^2_{\mathcal{Y}}.$$

Example 3. Let $M = L \otimes I_n$ and $K(x, z) = k(x, z) I_n$ where $k(x, z)$ is a standard scalar kernel and I_n is the $n \times n$ identity matrix. Then (15) reduces to Multivariate Laplacian RLS (3) which is equivalent to n independent scalar Laplacian RLS solutions.

5. Matrix-valued Kernels and Numerical Implementation

Proposition 1 provides general functional solutions to Vector-valued Manifold Regularization. We now focus our attention to finite dimensional output spaces $\mathcal{Y} = \mathbb{R}^n$ and in particular to solving Equation (15). We address two questions: what is an appropriate choice for a matrix-valued kernel, and how can (15) be solved more efficiently than the prohibitive $O(N^3 n^3)$ complexity of a general dense linear system solver. For a list of vector-valued kernels considered so far in the literature, see Micchelli & Pontil (2005); Caponnetto et al. (2008); Baldassarre et al. (2010). For practical applications, in this paper we consider matrix-valued kernels of the form,

$$K(x_i, x_j) = k(x_i, x_j) \left(\gamma_O L_{out}^{\dagger} + (1 - \gamma_O) I_n\right) \qquad (16)$$

where $k(\cdot, \cdot)$ is a scalar-valued kernel, $0 \le \gamma_O < 1$ is a real-valued parameter, I_n is the $n \times n$ identity matrix and $L_{out}^{\dagger}$ is the pseudo-inverse of the normalized Laplacian matrix, L_{out}, of a similarity graph W_{out} over outputs. We assume that W_{out} is either available as prior knowledge or, as in Section 6, estimated from nearest neighbor graphs over the available labels. We note the following properties of this kernel that justifies its choice in practical problems.

Universality: Let $\mathcal{X}$ be a Hausdorff topological space, $0 \leq \gamma_O < 1$ and $k(\cdot,\cdot)$ be a universal scalar valued kernel, e.g., the Gaussian kernel $k(x,z) = e^{-\frac{\|x-z\|^2}{2\sigma^2}}$. Then, $\mathcal{H}_K$ is universal, i.e., any continuous function from a compact subset of $\mathcal{X}$ to $\mathcal{Y}$ can be uniformly approximated by functions in $\mathcal{H}_K$. This observation follows from Theorem 12 of Caponnetto et al. (2008).

Regularization: Let $f = (f_1, \ldots, f_n)$ be a vector-valued function such that each $f_i \in \mathcal{H}_k$. When $\gamma_O = 1$, the norm of a vector-valued function in the RKHS induced by (16) is given by Baldassarre et al. (2010),

$$\|f\|_K^2 = \frac{1}{2}\sum_{i,j=1}^{n} \|f_i - f_j\|_k^2 \frac{(W_{out})_{ij}}{\sqrt{d_i d_j}} + \sum_{i=1}^{n} \|f_i\|_k^2$$

where W is the adjacency matrix of a similarity graph over outputs, $d_i = \sum_j (W_{out})_{ij}$, and $\|\cdot\|_k$ is the norm in the scalar-valued RKHS induced by $k(\cdot,\cdot)$. When $\gamma_O = 0$, K becomes diagonal and reduces to multivariate Laplacian RLS, Eq. (3), where each output is estimated independently of others. Thus, γ_O parameterizes the degree to which output inter-dependencies are enforced in the model.

Computational Consequences: The Gram matrix of the vector-valued kernel is the Kronecker product, $G = G_k^N \otimes Q$, where G_k^N is the Gram matrix of the scalar kernel over labeled and unlabeled data and $Q = \left(\gamma_O L_{out}^{\dagger} + (1-\gamma_O) I_n\right)$. Using $M = L \otimes I_n$ as in Section 4.3, the linear system (15) becomes the following,

$$\left[J_{nl}^{Nn}(G_k^N \otimes Q) + l\gamma_I (L \otimes I_n)(G_k^N \otimes Q) + l\gamma_A I_N\right] \mathbf{a} = \mathbf{y}.$$

Applying only basic properties of Kronecker products, it can be shown that this is identical to solving the following *Sylvester Equation* for the matrix A, where $\mathbf{a} = vec(A^T)$,

$$-\frac{1}{l\gamma_A}(J_l^N G_k^N + l\gamma_I L G_k^N) A Q - A + \frac{1}{l\gamma_A} Y = 0. \quad (17)$$

This Sylvester equation can be solved much more efficiently than directly calling a dense linear solver for (15). The number of floating point operations of a popular implementation[1] (Golub et al., 1979) is $\frac{5}{2}N^3 + 10n^3 + 5N^2 n + 2.5Nn^2$. Note that this complexity is of the same order as that of (non-linear) RLS with N (or n) labeled datapoints. Larger-scale iterative solvers are also available for such problems.

6. Empirical Studies

We consider semantic scene annotation and hierarchical text categorization as empirical testbeds for Vector-valued Manifold Regularization (abbreviated VVMR) implemented by solving (17). To the best of our knowledge, these are the first applications of vector-valued RKHS methods to multi-label learning problems. We empirically explore the value of incorporating unlabeled data in conjunction with dependencies across multiple output variables. A study is conducted to evaluate the quality of out-of-sample extension to unseen test data in comparison to transductive performance on the unlabeled set, and a comparison is presented against several recent state of the art multi-label learning methods. For our performance evaluation metric, we compute mean of the area under the ROC curve over all labels, as also used in Sun et al. (2011).

Semantic Scene Annotation: The data for this problem domain is drawn from Boutell et al. (2004). The task is to annotate images of natural scenes with 6 semantic categories: *Beach, Sunset, Fall foliage, Field, Mountain* and *Urban* given a 294-dimensional feature representation of color images. Thus each image $\mathbf{x}$ may be associated with a multi-label $\mathbf{y} \in \mathbb{R}^6$ where $y_i = 1$ indicates presence and $y_i = -1$ indicates absence of the i^{th} semantic category. We split the 2407 images available in this dataset into a training set (labeled + unlabeled) of 1204 images and a test set of 1203 images. All results reported in section are averaged over 10 random choices of 100 labels, treating the remaining training images as unlabeled data.

RCV1 - Hierarchical Text Categorization: The data for this problem domain is drawn from Lewis et al. (2004). The task is to place Reuters news-stories into a hierarchy of categories spanning various Corporate, Government, Markets and Economics related sub-topics. We used a subset of 6000 documents split evenly between a training (labeled+unlabeled) and a test set, available at `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel.html`. We eliminated words with small document and category frequencies resulting in a 2900-dimensional dataset with 25 categories. All results reported in this section are averaged over 10 random choices of 50 labels, treating the rest of the training set as unlabeled data.

Hyperparameters: Note that model selection in semi-supervised settings is a challenging problem in its own right, and not the focus of this paper. For scene annotation, we set $\gamma_A = 0.0001$ and used RBF scalar kernels with $\sigma = 4.3$, a default value corresponding to the median of pairwise distances amongst training datapoints. We use 5 and 2 nearest neighbor graphs to construct the input and the output normalized Graph Laplacians respectively. For RCV1, we use linear scalar kernels which are the most popular

[1] In Matlab $X = dlyap(A, B, C)$ solves $AXB - X + C = 0$

choice for text datasets. We set $\gamma_A = 1$ and used 50 and 20 nearest neighbor graphs to construct the input (iterated to degree 5) and the output normalized Graph Laplacians respectively. The fixed choices of some of these hyperparameters are based on values reported in Sindhwani et al. (2005) for image and text datasets. No further optimization of these values was attempted.

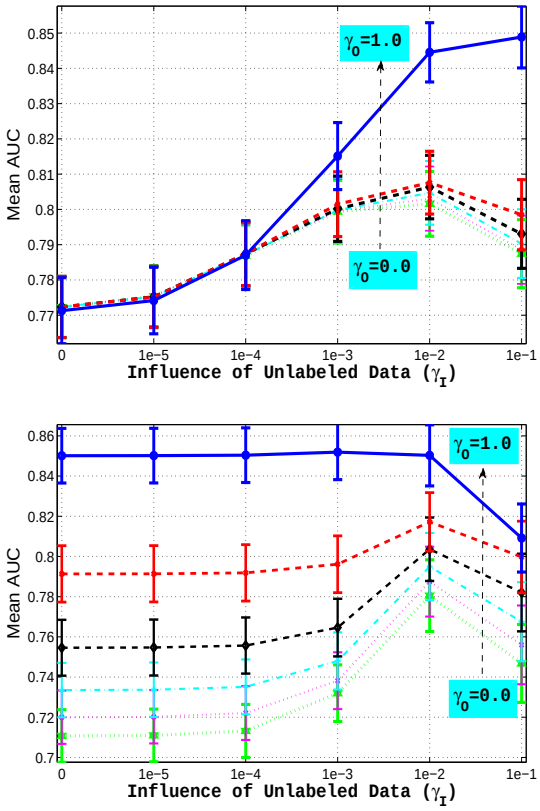

Figure 1. Influence of unlabeled data (γ_I) and output dependencies (γ_O): scene (top) and text (bottom) datasets.

Performance Enhancements with Unlabeled Data and Output Dependencies: In Figure 6, we evaluate Vector-valued Manifold Regularization in predicting the unknown multi-labels of the unlabeled set as a function of γ_I and γ_O. As the influence of unlabeled data is increased (γ_I taken steadily from 0 to 0.1), on both datasets we see a significant improvement in performance for any fixed value of γ_O. This entire performance curve shows consistent lift as γ_O is steadily increased from 0 to 1. These results show the benefit of incorporating both the data geometry and the output dependencies in a single model. On RCV1, unlabeled data shows greater relative benefit when output dependencies are weakly enforced sug-

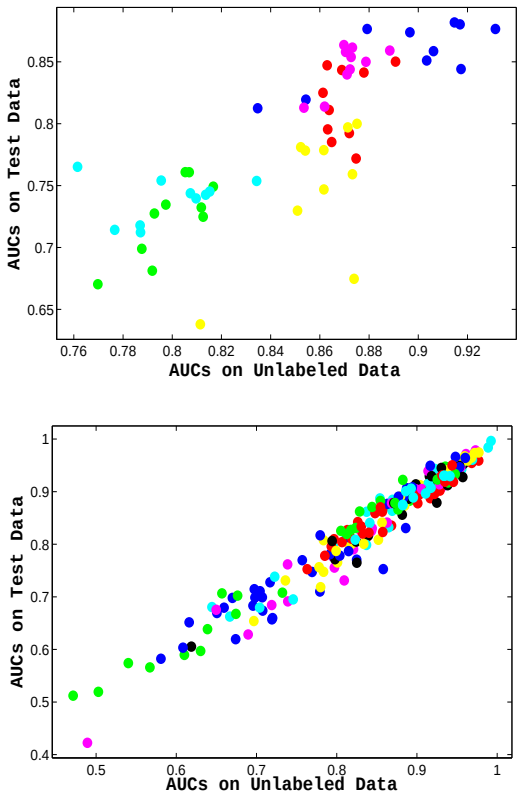

Figure 2. Transductive and Inductive AUCs across outputs (colored) on scene (top) and text (bottom) datasets.

gesting that it also provides robustness against imprecise model selection. Note that the special case, $\gamma_I = 0, \gamma_O > 0$ corresponds the vector-valued RLS, while $\gamma_I > 0, \gamma_O = 0$ corresponds to multivariate Laplacian RLS. In general, vector-valued manifold regularization outperforms these extremes.

Out-of-sample Generalization: In Figure 6, we show a scatter plot of performance on the unseen test set (induction) versus the performance on unlabeled data (transduction) for each of the different labels across the 10 random choices of labeled data. In nearly all cases, we see high-quality out-of-sample generalization from the unlabeled set to the test set. Purely transductive semi-supervised multilabel methods (Chen et al., 2008; Liu et al., 2006) do not naturally provide an out-of-sample extension.

Comparisons: We compared $VVMR$, i.e., solving (17), with several recently proposed supervised and semi-supervised multilabel learning methods: (1) $KCCA$ (Sun et al., 2011): a kernelized version of Canonical correlation analysis applied to labeled input and output data; its ridge parameter was optimized over $\{0, 10^k, -3 \leq k \leq 3\}$, (2) $M3L$ (Hariharan et al.,

2010): a large-scale margin based method for multi-label learning; its C parameter was optimized in the range $10^k, -3 \leq k \leq 3$, (3) SMSE2 (Chen et al., 2008): a transductive method that enforces smoothness with respect to input and output graphs; its parameters (analogous to γ_A, γ_O) were optimized over the range reported in Chen et al. (2008) and (4) CNMF (Liu et al., 2006): a constrained non-negative matrix factorization approach that also transductively exploits a given input and output similarity structure; its parameters were also optimized over the ranges reported in Liu et al. (2006). Note that these semi-supervised methods do not provide an out of sample extension as our method naturally can. From Table 1, we see that $VVMR$ provides statistically significant improvements (paired t-test at $p = 0.05$) over each of these alternatives in predicting the labels of unlabeled data. Exactly the same choices of input-output similarity graphs, and scalar kernel functions were used across different methods in this comparison.

Table 1. Comparison with competing algorithms (bold indicates statistically significant improvements)

Dataset	Scene	RCV1
KCCA	84.0(0.5)	80.0(2.3)
M3L	81.7(0.8)	84.3(1.5)
SMSE2	55.4(0.5)	52.5(0.7)
CNMF	52.7(5.5)	50.8(1.4)
VVMR	$\mathbf{84.9^\star(2.3)}$	$\mathbf{85.2^\star(1.4)}$

7. Conclusion

We consider this paper as a preliminary foray into more widespread adoption of vector-valued RKHS formalism for structured semi-supervised problems. Our results confirm that output dependencies and data geometry can both be exploited in a complementary fashion. Our choice of the kernel was dictated by certain natural regularization properties and computational tractability. The theoretical, algorithmic and empirical study of a wider class of operator-valued kernels offers a rich avenue for future work.

References

Bakir, G.H., Hofmann, T., Schlkopf, B., Smola, A.J., Taskar, B., and Vishwanathan, S.V.N (Ed). *Predicting Structured Data*. MIT Press, 2007.

Baldassarre, L., Rosasco, Lorenzo, Barla, Annalisa, and Verri, Alessandro. Vector-field learning with spectral filtering. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, volume 7, pp. 2399–2434, 2010.

Belkin, M., Niyogi, P., and Sindhwani, V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

Boutell, M.R., Luo, J., Shen, X., and Brown, C.M. Learning multi-label scene classification. *Pattern Recognition*, 2004.

Brown, P.J. and Zidek, J.V. On adaptive multivariate regression. *Annals of Statistics*, 1980.

Caponnetto, A. and Vito, E. De. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Caponnetto, A., Pontil, M., C.Micchelli, and Ying, Y. Universal multi-task kernels. *Journal of Machine Learning Research*, 9:1615–1646, 2008.

Carmeli, C., Vito, E. De, and Toigo, A. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4:377–408, 2006.

Chapelle, O., Schlkopf, B., and Zien, A. (Ed). *Semi-supervised Learning*. MIT Press, 2006.

Chen, G., Song, Y., Wang, F., and Zhang, C. Semi-supervised multi-label learning by solving a sylvester equation. In *SIAM Conference on Data Mining (SDM)*, 2008.

Golub, G.H., Nash, S., and Van Loan, C.F. A hessenberg-schur method for the problem A X + X B = C. *IEEE Transactions on Automatic Control*, AC-24:909–913, 1979.

Hariharan, B., Zelnik-Manor, L., Vishwanathan, S. V. N., and Varma, M. Large scale max-margin multi-label classification with priors. In *Proceedings of the International Conference on Machine Learning*, 2010.

Lewis, D. D., Yang, Y., and Li, F. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

Liu, Y., Jin, R., and Yang, L. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *The Twentieth Conference on Artificial Intelligence (AAAI)*, 2006.

Micchelli, C. A. and Pontil, M. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005.

Schölkopf, B. and Smola, A. *Learning with kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, 2002.

Sindhwani, V., Niyogi, P., and M.Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *Proceedings of the International Conference on Machine Learning*, 2005.

Sun, Liang, Ji, Shuiwang, and Ye, Jieping. Canonical correlation analysis for multilabel classification: A least squares formulation, extensions and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:194–200, 2011.