Journal of Machine Learning Research 1 (2006) 1-48 Submitted 4/00; Published 10/00

Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples

Mikhail Belkin MBELKIN@CSE.OHIO-STATE.EDU
Department of Computer Science and Engineering
The Ohio State University
2015 Neil Avenue, Dreese Labs 597
Columbus, OH 43210, USA

Partha Niyogi NIYOGI@CS.UCHICAGO.EDU
Departments of Computer Science and Statistics
University of Chicago
1100 E. 58th Street
Chicago, IL 60637, USA

Vikas Sindhwani VIKASS@CS.UCHICAGO.EDU
Department of Computer Science
University of Chicago
1100 E. 58th Street
Chicago, IL 60637, USA

Editor: Peter Bartlett

Abstract

We propose a family of learning algorithms based on a new form of regularization that allows us to exploit the geometry of the marginal distribution. We focus on a semi-supervised framework that incorporates labeled and unlabeled data in a general-purpose learner. Some transductive graph learning algorithms and standard methods including Support Vector Machines and Regularized Least Squares can be obtained as special cases. We utilize properties of Reproducing Kernel Hilbert spaces to prove new Representer theorems that provide theoretical basis for the algorithms. As a result (in contrast to purely graph-based approaches) we obtain a natural out-of-sample extension to novel examples and so are able to handle both transductive and truly semi-supervised settings. We present experimental evidence suggesting that our semi-supervised algorithms are able to use unlabeled data effectively. Finally we have a brief discussion of unsupervised and fully supervised learning within our general framework.

1. Introduction

In this paper, we introduce a framework for data-dependent regularization that exploits the geometry of the probability distribution. While this framework allows us to approach the full range of learning problems from unsupervised to supervised (discussed in Sections 6.1 and 6.2 respectively), we focus on the problem of semi-supervised learning.

The problem of learning from labeled and unlabeled data (*semi-supervised* and *transductive* learning) has attracted considerable attention in recent years. Some recently proposed methods include Transductive SVM (Vapnik, 1998; Joachims, 1999), Cotraining (Blum and Mitchell, 1998), and a variety of graph based methods (Blum and Chawla, 2001; Chapelle et al., 2003; Szummer and Jaakkola, 2002; Kondor and Lafferty, 2002; Smola and Kondor, 2003; Zhou et al., 2004; Zhu et al., 2003, 2005; Kemp et al., 2004; Joachims, 1999; Belkin and Niyogi, 2003a). We also note the regularization based techniques of Corduneanu and Jaakkola (2003) and Bousquet et al. (2004). The latter reference is closest in spirit to the intuitions of our paper. We postpone the discussion of related algorithms and various connections until Section 4.5.

The idea of regularization has a rich mathematical history going back to Tikhonov (1963), where it is used for solving ill-posed inverse problems. Regularization is a key idea in the theory of splines (e.g., Wahba (1990)) and is widely used in machine learning (e.g., Evgeniou et al. (2000)). Many machine learning algorithms, including Support Vector Machines, can be interpreted as instances of regularization.

Our framework exploits the geometry of the probability distribution that generates the data and incorporates it as an additional regularization term. Hence, there are two regularization terms — one controlling the complexity of the classifier in the *ambient space* and the other controlling the complexity as measured by the *geometry* of the distribution. We consider in some detail the special case where this probability distribution is supported on a submanifold of the ambient space.

The points below highlight several aspects of the current paper:

1. Our general framework brings together three distinct concepts that have received some independent recent attention in machine learning:
   i. The first of these is the technology of *spectral graph theory* (e.g., see Chung (1997)) that has been applied to a wide range of clustering and classification tasks over the last two decades. Such methods typically reduce to certain eigenvalue problems.
   ii. The second is the geometric point of view embodied in a class of algorithms that can be termed as *manifold learning*[1]. These methods attempt to use the geometry of the probability distribution by assuming that its support has the geometric structure of a Riemannian manifold.
   iii. The third important conceptual framework is the set of ideas surrounding regularization in Reproducing Kernel Hilbert Spaces (RKHS). This leads to the class of *kernel based algorithms* for classification and regression (e.g., see Scholkopf and Smola (2002); Wahba (1990); Evgeniou et al. (2000)).

   We show how these ideas can be brought together in a coherent and natural way to incorporate geometric structure in a kernel based regularization framework. As far as we know, these ideas have not been unified in a similar fashion before.

2. This general framework allows us to develop algorithms spanning the range from unsupervised to fully supervised learning.

   In this paper we primarily focus on the semi-supervised setting and present two families of algorithms: the Laplacian Regularized Least Squares (hereafter LapRLS) and the Laplacian Support Vector Machines (hereafter LapSVM). These are natural extensions of RLS and SVM respectively. In addition, several recently proposed transductive methods (e.g., Zhu et al. (2003); Belkin and Niyogi (2003a)) are also seen to be special cases of this general approach.

   In the absence of labeled examples our framework results in new algorithms for unsupervised learning, which can be used both for data representation and clustering. These algorithms are related to Spectral Clustering and Laplacian Eigenmaps (Belkin and Niyogi, 2003b).

1. see `http://www.cse.msu.edu/~lawhiu/manifold/` for a list of references

3. We elaborate on the RKHS foundations of our algorithms and show how geometric knowledge of the probability distribution may be incorporated in such a setting through an additional regularization penalty. In particular, a new Representer theorem provides a functional form of the solution when the distribution is known; its empirical version involves an expansion over labeled and unlabeled points when the distribution is unknown. These Representer theorems provide the basis for our algorithms.

4. Our framework with an ambiently defined RKHS and the associated Representer theorems result in a natural out-of-sample extension from the data set (labeled and unlabeled) to novel examples. This is in contrast to the variety of purely graph based approaches that have been considered in the last few years. Such graph based approaches work in a transductive setting and do not naturally extend to the semi-supervised case where novel test examples need to be classified (predicted). Also see Bengio et al. (2004); Brand (2003) for some recent related work on out-of-sample extensions. We also note that a method similar to our regularized spectral clustering algorithm has been independently proposed in the context of graph inference in Vert and Yamanishi (2005).

The work presented here is based on the University of Chicago Technical Report TR-2004-05, a short version in the Proceedings of AI & Statistics 2005, Belkin et al. (2005) and Sindhwani (2004).

1.1 The Significance of Semi-Supervised Learning

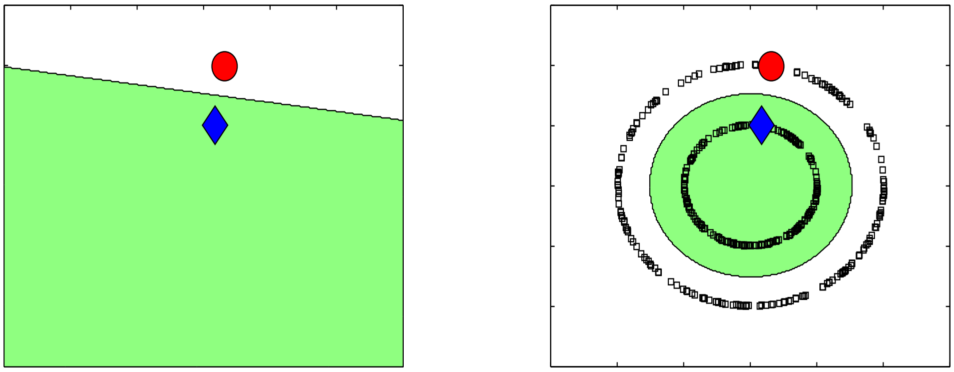
From an engineering standpoint, it is clear that collecting labeled data is generally more involved than collecting unlabeled data. As a result, an approach to pattern recognition that is able to make better use of unlabeled data to improve recognition performance is of potentially great practical significance.

However, the significance of semi-supervised learning extends beyond purely utilitarian considerations. Arguably, most natural (human or animal) learning occurs in the semi-supervised regime. We live in a world where we are constantly exposed to a stream of natural stimuli. These stimuli comprise the unlabeled data that we have easy access to. For example, in phonological acquisition contexts, a child is exposed to many acoustic utterances. These utterances do not come with identifiable phonological markers. Corrective feedback is the main source of directly labeled examples. In many cases, a small amount of feedback is sufficient to allow the child to master the acoustic-to-phonetic mapping of any language.

The ability of humans to learn unsupervised concepts (e.g. learning clusters and categories of objects) suggests that unlabeled data can be usefully processed to learn natural invariances, to form categories, and to develop classifiers. In most pattern recognition tasks, humans have access only to a small number of labeled examples. Therefore the success of human learning in this "small sample" regime is plausibly due to effective utilization of the large amounts of unlabeled data to extract information that is useful for generalization.

Consequently, if we are to make progress in understanding how natural learning comes about, we need to think about the basis of semi-supervised learning. Figure 1 illustrates how unlabeled examples may force us to restructure our hypotheses during learning. Imagine a situation where one is given two labeled examples — one positive and one negative — as shown in the left panel. If one is to induce a classifier on the basis of this, a natural choice would seem to be the linear separator as shown. Indeed, a variety of theoretical formalisms (Bayesian paradigms, Regularization, Minimum

Figure 1: Unlabeled Data and Prior Beliefs



Description Length or Structural Risk Minimization principles, and the like) have been constructed to rationalize such a choice. In most of these formalisms, one structures the set of one's hypothesis functions by a prior notion of simplicity and one may then justify why the linear separator is the simplest structure consistent with the data.

Now consider the situation where one is given additional unlabeled examples as shown in the right panel. We argue that it is self-evident that in the light of this new unlabeled set, one must re-evaluate one's prior notion of simplicity. The particular geometric structure of the marginal distribution suggests that the most natural classifier is now the circular one indicated in the right panel. Thus the geometry of the marginal distribution must be incorporated in our regularization principle to impose structure on the space of functions in nonparametric classification or regression. This is the intuition we formalize in the rest of the paper. The success of our approach depends on whether we can extract structure from the marginal distribution, and on the extent to which such structure may reveal the underlying truth.

1.2 Outline of the Paper

The paper is organized as follows: in Sec. 2, we develop the basic framework for semi-supervised learning where we ultimately formulate an objective function that can utilize both labeled and unlabeled data. The framework is developed in an RKHS setting and we state two kinds of Representer theorems describing the functional form of the solutions. In Sec. 3, we elaborate on the theoretical underpinnings of this framework and prove the Representer theorems of Sec. 2. While the Representer theorem for the finite sample case can be proved using standard orthogonality arguments, the Representer theorem for the known marginal distribution requires more subtle considerations. In Sec. 4, we derive the different algorithms for semi-supervised learning that arise out of our framework. Connections to related algorithms are stated. In Sec. 5, we describe experiments that evaluate the algorithms and demonstrate the usefulness of unlabeled data. In Sec. 6, we consider the cases of fully supervised and unsupervised learning. In Sec. 7 we conclude this paper.

2. The Semi-Supervised Learning Framework

2.1 Background

Recall the standard framework of learning from examples. There is a probability distribution P on $X \times \mathbb{R}$ according to which examples are generated for function learning. Labeled examples are (x, y) pairs generated according to P. Unlabeled examples are simply $x \in X$ drawn according to the marginal distribution $\mathcal{P}_X$ of P.

One might hope that knowledge of the marginal $\mathcal{P}_X$ can be exploited for better function learning (e.g. in classification or regression tasks). Of course, if there is no identifiable relation between $\mathcal{P}_X$ and the conditional $\mathcal{P}(y|x)$, the knowledge of $\mathcal{P}_X$ is unlikely to be of much use.

Therefore, we will make a specific assumption about the connection between the marginal and the conditional distributions. We will assume that if two points $x_1, x_2 \in X$ are *close* in the *intrinsic* geometry of $\mathcal{P}_X$, then the conditional distributions $\mathcal{P}(y|x_1)$ and $\mathcal{P}(y|x_2)$ are similar. In other words, the conditional probability distribution $\mathcal{P}(y|x)$ varies smoothly along the geodesics in the intrinsic geometry of $\mathcal{P}_X$.

We utilize these geometric intuitions to extend an established framework for function learning. A number of popular algorithms such as SVM, Ridge regression, splines, Radial Basis Functions may be broadly interpreted as regularization algorithms with different empirical cost functions and complexity measures in an appropriately chosen Reproducing Kernel Hilbert Space (RKHS).

For a Mercer kernel $K : X \times X \to \mathbb{R}$, there is an associated RKHS $\mathcal{H}_K$ of functions $X \to \mathbb{R}$ with the corresponding norm $\| \, \|_K$. Given a set of labeled examples (x_i, y_i), $i = 1, \ldots, l$ the standard framework estimates an unknown function by minimizing

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma \|f\|_K^2 \tag{1}$$

where V is some loss function, such as squared loss $(y_i - f(x_i))^2$ for RLS or the hinge loss function $\max[0, 1 - y_i f(x_i)]$ for SVM. Penalizing the RKHS norm imposes smoothness conditions on possible solutions. The classical Representer Theorem states that the solution to this minimization problem exists in $\mathcal{H}_K$ and can be written as

$$f^*(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x) \tag{2}$$

Therefore, the problem is reduced to optimizing over the finite dimensional space of coefficients α_i, which is the algorithmic basis for SVM, Regularized Least Squares and other regression and classification schemes.

We first consider the case when the marginal distribution is already known.

2.2 Marginal $\mathcal{P}_X$ is known

Our goal is to extend this framework by incorporating additional information about the geometric structure of the marginal $\mathcal{P}_X$. We would like to ensure that the solution is smooth with respect to both the ambient space and the marginal distribution $\mathcal{P}_X$. To achieve that, we introduce an additional regularizer:

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2 \tag{3}$$

where $\|f\|_I^2$ is an appropriate penalty term that should reflect the intrinsic structure of $\mathcal{P}_X$. Intuitively, $\|f\|_I^2$ is a smoothness penalty corresponding to the probability distribution. For example, if the probability distribution is supported on a low-dimensional manifold, $\|f\|_I^2$ may penalize f along that manifold. γ_A controls the complexity of the function in the *ambient* space while γ_I controls the complexity of the function in the *intrinsic* geometry of $\mathcal{P}_X$. It turns out that one can derive an explicit functional form for the solution f^* as shown in the following theorem.

Theorem 1 *Assume that the penalty term $\|f\|_I$ is sufficiently smooth with respect to the RKHS norm $\|f\|_K$ (see Section 3.2 for the exact statement). Then the solution f^* to the optimization problem in Eqn. 3 above exists and admits the following representation*

$$f^*(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x) + \int_{\mathcal{M}} \alpha(z) K(x, z)\, d\mathcal{P}_X(z) \tag{4}$$

where $\mathcal{M} = \operatorname{supp}\{\mathcal{P}_X\}$ is the support of the marginal $\mathcal{P}_X$.

We postpone the proof and the formulation of smoothness conditions on the norm $\| \ \|_I$ until the next section.

The Representer Theorem above allows us to express the solution f^* directly in terms of the labeled data, the (ambient) kernel K, and the marginal $\mathcal{P}_X$. If $\mathcal{P}_X$ is unknown, we see that the solution may be expressed in terms of an empirical estimate of $\mathcal{P}_X$. Depending on the nature of this estimate, different approximations to the solution may be developed. In the next section, we consider a particular approximation scheme that leads to a simple algorithmic framework for learning from labeled and unlabeled data.

2.3 Marginal $\mathcal{P}_X$ Unknown

In most applications the marginal $\mathcal{P}_X$ is not known. Therefore we must attempt to get empirical estimates of $\mathcal{P}_X$ and $\| \ \|_I$. Note that in order to get such empirical estimates it is sufficient to have *unlabeled* examples.

A case of particular recent interest (e.g., see Roweis and Saul (2000); Tenenbaum et al. (2000); Belkin and Niyogi (2003b); Donoho and Grimes (2003); **?** for a discussion on dimensionality reduction) is when the support of $\mathcal{P}_X$ is a compact submanifold $\mathcal{M} \subset \mathbb{R}^n$. In that case, one natural choice for $\|f\|_I$ is $\int_{x\in\mathcal{M}} \|\nabla_{\mathcal{M}} f\|^2\, d\mathcal{P}_X(x)$, where $\nabla_{\mathcal{M}}$ is the *gradient* (see, e.g., Do Carmo (1992) for an introduction to differential geometry) of f along the manifold $\mathcal{M}$ and the integral is taken over the marginal distribution.

The optimization problem becomes

$$f^* = \operatorname*{argmin}_{f\in\mathcal{H}_K} \frac{1}{l}\sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \gamma_I \int_{x\in\mathcal{M}} \|\nabla_{\mathcal{M}} f\|^2\, d\mathcal{P}_X(x)$$

The term $\int_{x\in\mathcal{M}} \|\nabla_{\mathcal{M}} f\|^2\, d\mathcal{P}_X(x)$ may be approximated on the basis of labeled and unlabeled data using the graph Laplacian associated to the data. While an extended discussion of these issues goes beyond the scope of this paper, it can be shown that under certain conditions choosing exponential weights for the adjacency graph leads to convergence of the graph Laplacian to the Laplace-Beltrami operator $\Delta_{\mathcal{M}}$ (or its weighted version) on the manifold. See the Remarks below and (Belkin, 2003; Lafon, 2004; Belkin and Niyogi, 2005; Coifman and Lafon, 2005; Hein et al., 2005) for details.

Thus, given a set of l labeled examples $\{(x_i, y_i)\}_{i=1}^{l}$ and a set of u unlabeled examples $\{x_j\}_{j=l+1}^{j=l+u}$, we consider the following optimization problem :

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \frac{\gamma_I}{(u+l)^2} \sum_{i,j=1}^{l+u} (f(x_i) - f(x_j))^2 W_{ij}$$

$$= \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \frac{\gamma_I}{(u+l)^2} \mathbf{f}^T L \mathbf{f} \tag{5}$$

where W_{ij} are edge weights in the data adjacency graph, $\mathbf{f} = [f(x_1), \ldots, f(x_{l+u})]^T$, and L is the graph Laplacian given by $L = D - W$. Here, the diagonal matrix D is given by $D_{ii} = \sum_{j=1}^{l+u} W_{ij}$. The normalizing coefficient $\frac{1}{(u+l)^2}$ is the natural scale factor for the empirical estimate of the Laplace operator. We note than on a sparse adjacency graph it may be replaced by $\sum_{i,j=1}^{l+u} W_{ij}$.

The following version of the Representer Theorem shows that the minimizer has an expansion in terms of both labeled and unlabeled examples and is a key to our algorithms.

Theorem 2 *The minimizer of optimization problem 5 admits an expansion*

$$f^*(x) = \sum_{i=1}^{l+u} \alpha_i K(x_i, x) \tag{6}$$

in terms of the labeled and unlabeled examples.

The proof is a variation of the standard orthogonality argument and is presented in Section 3.4.

Remark 1: Several natural choices of $\| \ \|_I$ exist. Some examples are:

1. Iterated Laplacians $(\Delta_{\mathcal{M}})^k$. Differential operators $(\Delta_{\mathcal{M}})^k$ and their linear combinations provide a natural family of smoothness penalties.

   Recall that the Laplace-Beltrami operator $\Delta_{\mathcal{M}}$ can be defined as the divergence of the gradient vector field $\Delta_{\mathcal{M}} f = \operatorname{div}(\nabla_{\mathcal{M}} f)$ and is characterized by the equality

   $$\int_{x \in \mathcal{M}} f(x) \Delta_{\mathcal{M}} f(x) d\mu = \int_{x \in \mathcal{M}} \|\nabla_{\mathcal{M}} f(x)\|^2 \, d\mu$$

   where μ is the standard measure (uniform distribution) on the Riemannian manifold. If μ is taken to be non-uniform, then the corresponding notion is the weighted Laplace-Beltrami operator (e.g., Grigor'yan (2006)).

2. Heat semigroup $e^{-t\Delta_{\mathcal{M}}}$ is a family of smoothing operators corresponding to the process of diffusion (Brownian motion) on the manifold. One can take $\|f\|_I^2 = \int_{\mathcal{M}} f \, e^{t\Delta_{\mathcal{M}}}(f) d\mathcal{P}_X$. We note that for small values of t the corresponding Green's function (the heat kernel of $\mathcal{M}$), which is close to a Gaussian in the geodesic coordinates, can also be approximated by a sharp Gaussian in the ambient space.

3. Squared norm of the Hessian (cf. Donoho and Grimes (2003)). While the Hessian $\mathbf{H}(f)$ (the matrix of second derivatives of f) generally depends on the coordinate system, it can

be shown that the Frobenius norm (the sum of squared eigenvalues) of $\mathbf{H}$ is the same in any geodesic coordinate system and hence is invariantly defined for a Riemannian manifold $\mathcal{M}$. Using the Frobenius norm of $\mathbf{H}$ as a regularizer presents an intriguing generalization of thin-plate splines. We also note that $\Delta_{\mathcal{M}}(f) = \mathrm{tr}(\mathbf{H}(f))$.

Remark 2: Why not just use the intrinsic regularizer? Using ambient and intrinsic regularizers jointly is important for the following reasons:

1. We do not usually have access to $\mathcal{M}$ or the true underlying marginal distribution, just to data points sampled from it. Therefore regularization with respect only to the sampled manifold is ill-posed. By including an ambient term, the problem becomes well-posed.

2. There may be situations when regularization with respect to the ambient space yields a better solution, e.g., when the manifold assumption does not hold (or holds to a lesser degree). Being able to trade off these two regularizers may be important in practice.

Remark 3: While we use the graph Laplacian for simplicity, the *normalized Laplacian*

$$\tilde{L} = D^{-1/2} L D^{-1/2}$$

can be used interchangeably in all our formulas. Using $\tilde{L}$ instead of L provides certain theoretical guarantees (see von Luxburg et al. (2004)) and seems to perform as well or better in many practical tasks. In fact, we use $\tilde{L}$ in all our empirical studies in Section 5. The relation of $\tilde{L}$ to the weighted Laplace-Beltrami operator was discussed in Lafon (2004).

Remark 4: Note that a global kernel K restricted to $\mathcal{M}$ (denoted by $K_{\mathcal{M}}$) is also a kernel defined on $\mathcal{M}$ with an associated RKHS $\mathcal{H}_{\mathcal{M}}$ of functions $\mathcal{M} \to \mathbb{R}$. While this might suggest

$$\|f\|_I = \|f_{\mathcal{M}}\|_{K_{\mathcal{M}}}$$

($f_{\mathcal{M}}$ is f restricted to $\mathcal{M}$) as a reasonable choice for $\|f\|_I$, it turns out, that for the minimizer f^* of the corresponding optimization problem we get $\|f^*\|_I = \|f^*\|_K$, yielding the same solution as standard regularization, although with a different parameter γ. This observation follows from the restriction properties of RKHS discussed in the next section and is formally stated as Proposition 6. Therefore it is impossible to have an out-of-sample extension without two *different* measures of smoothness. On the other hand, a different ambient kernel restricted to $\mathcal{M}$ can potentially serve as the intrinsic regularization term. For example, a sharp Gaussian kernel can be used as an approximation to the heat kernel on $\mathcal{M}$. Thus one (sharper) kernel may be used in conjunction with unlabeled data to estimate the heat kernel on $\mathcal{M}$ and a wider kernel for inference.

3. Theoretical Underpinnings and Results

In this section we briefly review the theory of Reproducing Kernel Hilbert Spaces and their connection to integral operators. We proceed to establish the Representer theorems from the previous section.

3.1 General Theory of RKHS

We start by recalling some basic properties of Reproducing Kernel Hilbert Spaces (see the original work Aronszajn (1950) and also Cucker and Smale (2002) for a nice discussion in the context of learning theory) and their connections to integral operators. We say that a Hilbert space $\mathcal{H}$ of functions $X \to \mathbb{R}$ has the *reproducing property*, if $\forall x \in X$ the evaluation functional $f \to f(x)$ is continuous. For the purposes of this discussion we will assume that X is compact. By the Riesz representation theorem it follows that for a given $x \in X$, there is a function $h_x \in \mathcal{H}$, s.t.

$$\forall f \in \mathcal{H} \quad \langle h_x, f \rangle_{\mathcal{H}} = f(x)$$

We can therefore define the corresponding *kernel function*

$$K(x, y) = \langle h_x, h_y \rangle_{\mathcal{H}}$$

It follows that $h_x(y) = \langle h_x, h_y \rangle_{\mathcal{H}} = K(x, y)$ and thus $\langle K(x, \cdot), f \rangle = f(x)$. It is clear that $K(x, \cdot) \in \mathcal{H}$.

It is easy to see that $K(x, y)$ is a positive semi-definite kernel as defined below:

Definition: We say that $K(x, y)$, satisfying $K(x, y) = K(y, x)$, is a positive semi-definite kernel if given an arbitrary finite set of points $x_1, \ldots, x_n$, the corresponding $n \times n$ matrix K with $K_{ij} = K(x_i, x_j)$ is positive semi-definite.

Importantly, the converse is also true. Any positive semi-definite kernel $K(x, y)$ gives rise to an RKHS $\mathcal{H}_K$, which can be constructed by considering the space of finite linear combinations of kernels $\sum \alpha_i K(x_i, \cdot)$ and taking completion with respect to the inner product given by $\langle K(x, \cdot), K(y, \cdot) \rangle_{\mathcal{H}_K} = K(x, y)$. See Aronszajn (1950) for details.

We therefore see that Reproducing Kernel Hilbert Spaces of functions on a space X are in *one-to-one correspondence* with positive semidefinite kernels on X.

It can be shown that if the space $\mathcal{H}_K$ is sufficiently rich, that is if for any distinct point $x_1, \ldots, x_n$ there is a function f, s.t. $f(x_1) = 1, f(x_i) = 0, i > 1$, then the corresponding matrix $K_{ij} = K(x_i, x_j)$ is strictly positive definite. For simplicity we will sometimes assume that our RKHS are rich (the corresponding kernels are sometimes called *universal*).

Notation: In what follows, we will use kernel K to denote inner products and norms in the corresponding Hilbert space $\mathcal{H}_K$, that is, we will write $\langle\ ,\ \rangle_K$, $\|\ \|_K$, instead of the more cumbersome $\langle\ ,\ \rangle_{\mathcal{H}_K}$, $\|\ \|_{\mathcal{H}_K}$.

We proceed to endow X with a measure μ (supported on all of X). The corresponding $\mathcal{L}^2_\mu$ Hilbert space inner product is given by

$$\langle f, g \rangle_\mu = \int_X f(x) g(x) d\mu$$

We can now consider the integral operator L_K corresponding to the kernel K:

$$(L_K f)(x) = \int_X f(y) K(x, y)\, d\mu \tag{7}$$

It is well-known that if X is a compact space, L_K is a compact operator and is self-adjoint with respect to $\mathcal{L}^2_\mu$. By the Spectral Theorem, its eigenfunctions $e_1(x), e_2(x), \ldots$, (scaled to norm 1) form an orthonormal basis of $\mathcal{L}^2_\mu$. The spectrum of the operator is discrete and the corresponding eigenvalues $\lambda_1, \lambda_2, \ldots$ are of finite multiplicity, $\lim_{i \to \infty} \lambda_i = 0$.

We see that

$$\langle K(x, \cdot), e_i(\cdot) \rangle_\mu = \lambda_i e_i(x) \tag{8}$$

and therefore $K(x,y) = \sum_i \lambda_i e_i(x)e_i(y)$. Writing a function f in that basis, we have $f = \sum a_i e_i(x)$ and $\langle K(x,\cdot), f(\cdot)\rangle_\mu = \sum_i \lambda_i a_i e_i(x)$.

It is not hard to show that the eigenfunctions e_i are in $\mathcal{H}_K$ (e.g., see the argument below). Thus we see that

$$e_j(x) = \langle K(x,\cdot), e_j(\cdot)\rangle_K = \sum_i \lambda_i e_i(x)\langle e_i, e_j\rangle_K \tag{9}$$

Therefore $\langle e_i, e_j\rangle_K = 0$, if $i \neq j$, and $\langle e_i, e_i\rangle_K = \frac{1}{\lambda_i}$. On the other hand $\langle e_i, e_j\rangle_\mu = 0$, if $i \neq j$, and $\langle e_i, e_i\rangle_\mu = 1$.

This observation establishes a simple relationship between the Hilbert norms in $\mathcal{H}_K$ and $\mathcal{L}^2_\mu$. We also see that $f = \sum a_i e_i(x) \in \mathcal{H}_K$ if and only if $\sum \frac{a_i^2}{\lambda_i} < \infty$.

Consider now the operator $L_K^{1/2}$. It can be defined as the only positive definite self-adjoint operator, s.t. $L_K = L_K^{1/2} \circ L_K^{1/2}$. Assuming that the series $\tilde{K}(x,y) = \sum_i \sqrt{\lambda_i} e_i(x)e_i(y)$ converges, we can write

$$(L_K^{1/2} f)(x) = \int_X f(y)\tilde{K}(x,y)\, d\mu \tag{10}$$

It is easy to check that $L_K^{1/2}$ is an isomorphism between $\mathcal{H}$ and $\mathcal{L}^2_\mu$, that is

$$\forall f,g \in \mathcal{H}_K \quad \langle f,g\rangle_\mu = \langle L_K^{1/2} f, L_K^{1/2} g\rangle_K \tag{11}$$

Therefore $\mathcal{H}_K$ is the image of $L_K^{1/2}$ acting on $\mathcal{L}^2_\mu$.

Lemma 3 *A function $f(x) = \sum_i a_i e_i(x)$ can be represented as $f = L_K g$ for some g if and only if*

$$\sum_{i=1}^{\infty} \frac{a_i^2}{\lambda_i^2} < \infty \tag{12}$$

Proof Suppose $f = L_K g$. Write $g(x) = \sum_i b_i e_i(x)$. We know that $g \in L^2_\mu$ if and only if $\sum_i b_i^2 < \infty$. Since $L_K(\sum_i b_i e_i) = \sum_i b_i \lambda_i e_i = \sum_i a_i e_i$, we obtain $a_i = b_i \lambda_i$. Therefore $\sum_{i=1}^{\infty} \frac{a_i^2}{\lambda_i^2} < \infty$.

Conversely, if the condition in the inequality 12 is satisfied, $f = L_k g$, where $g = \sum \frac{a_i}{\lambda_i} e_i$. ■

3.2 Proof of Theorems

Now let us recall the Eqn. 3:

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2 \tag{13}$$

We have an RKHS $\mathcal{H}_K$ and the probability distribution μ which is supported on $\mathcal{M} \subset X$. We denote by $\mathcal{S}$ the linear space, which is the closure with respect to the RKHS norm of $\mathcal{H}_K$, of the linear span of kernels centered at points of $\mathcal{M}$:

$$\mathcal{S} = \overline{\operatorname{span}\{K(x,\cdot) \mid x \in \mathcal{M}\}} \tag{14}$$

Notation. By the subscript $\mathcal{M}$ we will denote the restriction to $\mathcal{M}$. For example, by $\mathcal{S}_{\mathcal{M}}$ we denote functions in $\mathcal{S}$ restricted to the manifold $\mathcal{M}$. It can be shown (Aronszajn (1950), p. 350) that the space $(\mathcal{H}_K)_{\mathcal{M}}$ of functions from $\mathcal{H}_K$ restricted to $\mathcal{M}$ is an RKHS with the kernel $K_{\mathcal{M}}$, in other words $(\mathcal{H}_K)_{\mathcal{M}} = \mathcal{H}_{K_{\mathcal{M}}}$.

Lemma 4 *The following properties of $\mathcal{S}$ hold:*

1. *$\mathcal{S}$ with the inner product induced by $\mathcal{H}_K$ is a Hilbert space.*
2. *$\mathcal{S}_{\mathcal{M}} = (\mathcal{H}_K)_{\mathcal{M}}$.*
3. *The orthogonal complement $\mathcal{S}^{\perp}$ to $\mathcal{S}$ in $\mathcal{H}_K$ consists of all functions vanishing on $\mathcal{M}$.*

Proof

1. From the definition of $\mathcal{S}$ it is clear by that $\mathcal{S}$ is a complete subspace of $\mathcal{H}_K$.

2. We give a convergence argument similar to the one found in Aronszajn (1950). Since $(\mathcal{H}_K)_{\mathcal{M}} = \mathcal{H}_{K_{\mathcal{M}}}$ any function $f_{\mathcal{M}}$ in it can be written as $f_{\mathcal{M}} = \lim_{n\to\infty} f_{\mathcal{M},n}$, where $f_{\mathcal{M},n} = \sum_i \alpha_{in} K_{\mathcal{M}}(x_{in}, \cdot)$ is a sum of kernel functions.

Consider the corresponding sum $f_n = \sum_i \alpha_{in} K(x_{in}, \cdot)$. From the definition of the norm we see that $\|f_n - f_k\|_K = \|f_{\mathcal{M},n} - f_{\mathcal{M},k}\|_{K_{\mathcal{M}}}$ and therefore f_n is a Cauchy sequence. Thus $f = \lim_{n\to\infty} f_n$ exists and its restriction to $\mathcal{M}$ must equal $f_{\mathcal{M}}$. This shows that $(\mathcal{H}_K)_{\mathcal{M}} \subset \mathcal{S}_{\mathcal{M}}$. The other direction follows by a similar argument.

3. Let $g \in \mathcal{S}^{\perp}$. By the reproducing property for any $x \in \mathcal{M}$, $g(x) = \langle K(x,\cdot), g(\cdot)\rangle_K = 0$ and therefore any function in $\mathcal{S}^{\perp}$ vanishes on $\mathcal{M}$. On the other hand, if g vanishes on $\mathcal{M}$ it is perpendicular to each $K(x,\cdot), x \in \mathcal{M}$ and is therefore perpendicular to the closure of their span $\mathcal{S}$. ■

Lemma 5 *Assume that the intrinsic norm is such that for any $f, g \in \mathcal{H}_K$, $(f-g)|_{\mathcal{M}} \equiv 0$ implies that $\|f\|_I = \|g\|_I$. Then assuming that the solution f^* of the optimization problem in Eqn. 13 exists, $f^* \in \mathcal{S}$.*

Proof Any $f \in \mathcal{H}_K$ can be written as $f = f_{\mathcal{S}} + f_{\mathcal{S}}^{\perp}$, where $f_{\mathcal{S}}$ is the projection of f to $\mathcal{S}$ and $f_{\mathcal{S}}^{\perp}$ is its orthogonal complement.

For any $x \in M$ we have $K(x,\cdot) \in \mathcal{S}$. By the previous Lemma $f_{\mathcal{S}}^{\perp}$ vanishes on $\mathcal{M}$. We have $f(x_i) = f_{\mathcal{S}}(x_i)\ \forall_i$ and by assumption $\|f_{\mathcal{S}}\|_I = \|f\|_I$.

On the other hand, $\|f\|_K^2 = \|f_{\mathcal{S}}\|_K^2 + \|f_{\mathcal{S}}^{\perp}\|_K^2$ and therefore $\|f\|_K \geq \|f_{\mathcal{S}}\|_K$. It follows that the minimizer f^* is in $\mathcal{S}$. ■

As a direct corollary of these consideration, we obtain the following

Proposition 6 *If $\|f\|_I = \|f\|_{K_{\mathcal{M}}}$ then the minimizer of Eqn. 13 is identical to that of the usual regularization problem (Eqn. 1) although with a different regularization parameter ($\lambda_A + \lambda_I$).*

We can now restrict our attention to the study of $\mathcal{S}$. While it is clear that the right-hand side of Eqn. 4 lies in $\mathcal{S}$, not every element in $\mathcal{S}$ can be written in that form. For example, $K(x,\cdot)$, where x

is not one of the data points x_i cannot generally be written as

$$\sum_{i=1}^{l} \alpha_i K(x_i, x) + \int_{\mathcal{M}} \alpha(y) K(x, y)\, d\mu \tag{15}$$

We will now assume that for $f \in \mathcal{S}$

$$\|f\|_I^2 = \langle f, Df \rangle_{\mathcal{L}^2_\mu}$$

We usually assume that D is an appropriate smoothness penalty, such as an inverse integral operator or a differential operator, e.g., $Df = \Delta_{\mathcal{M}} f$. The Representer theorem, however, holds under quite mild conditions on D:

Theorem 7 *Let D be a bounded operator $D : \mathcal{S} \to \mathcal{L}^2_{\mathcal{P}_X}$. Then the solution f^* of the optimization problem in Eqn. 3 exists and can be written*

$$f^*(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x) + \int_{\mathcal{M}} \alpha(y) K(x, y)\, d\mathcal{P}_X(y) \tag{16}$$

Proof

For simplicity we will assume that the loss function V is differentiable. This condition can ultimately be eliminated by approximating a non-differentiable function appropriately and passing to the limit.

Put

$$H(f) = \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f(x_i)) + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2 \tag{17}$$

We first show that the solution to Eqn. 3 f^* exists and by Lemma 5 belongs to $\mathcal{S}$. It follows easily from Cor. 10 and standard results about compact embeddings of Sobolev spaces (e.g., Adams (1975)) that a ball $\mathcal{B}_r \subset \mathcal{H}_K$, $\mathcal{B}_r = \{f \in \mathcal{S}, s.t.\ \|f\|_K \le r\}$ is compact in $\mathcal{L}_X^\infty$. Therefore for any such ball the minimizer in that ball f_r^* must exist and belong to $\mathcal{B}_r$. On the other hand, by substituting the zero function

$$H(f_r^*) \le H(0) = \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, 0) \tag{18}$$

If the loss is actually zero, then zero function is a solution, otherwise

$$\gamma_A \|f_r^*\|_K^2 < \sum_{i=1}^{l} V(x_i, y_i, 0) \tag{19}$$

and hence $f_r^* \in \mathcal{B}_r$, where

$$r = \sqrt{\frac{\sum_{i=1}^{l} V(x_i, y_i, 0)}{\gamma_A}}$$

Therefore we cannot decrease $H(f^*)$ by increasing r beyond a certain point, which shows that $f^* = f_r^*$ with r as above, which completes the proof of existence. If V is convex, such solution will also be unique.

We proceed to derive the Eqn. 16. As before, let $e_1, e_2, \dots$ be the basis associated to the integral operator $(L_K f)(x) = \int_{\mathcal{M}} f(y)K(x,y)\, d\mathcal{P}_X(y)$. Write $f^* = \sum_i a_i e_i(x)$. By substituting f^* into $H(f)$ we obtain:

$$H(f^*) = \frac{1}{l}\sum_{j=1}^{l} V(x_j, y_j, \sum_i a_i e_i(x_i)) + \gamma_A \|f^*\|_K^2 + \gamma_I \|f^*\|_I^2 \tag{20}$$

Assume that V is differentiable with respect to each a_k. We have $\|\sum_i a_i e_i(x)\|_K^2 = \sum_i \frac{a_i^2}{\lambda_i}$. Differentiating with respect to the coefficients a_i yields the following set of equations:

$$0 = \frac{\partial H(f^*)}{\partial a_k} = \frac{1}{l}\sum_{j=1}^{l} e_k(x_j)\partial_3 V\,(x_j, y_j, \sum_i a_i e_i) + 2\gamma_A \frac{a_k}{\lambda_k} + \gamma_I \langle Df, e_k\rangle + \gamma_I \langle f, De_k\rangle \tag{21}$$

where $\partial_3 V$ denotes the derivative with respect to the third argument of V.

$\langle Df, e_k\rangle + \langle f, De_k\rangle = \langle (D+D^*)f, e_k\rangle$ and hence

$$a_k = -\frac{\lambda_k}{2\gamma_A l}\sum_{j=1}^{l} e_k(x_j)\partial_3 V\,(x_j, y_j, f^*) - \frac{\gamma_I}{2\gamma_A}\lambda_k \langle Df^* + D^* f^*, e_k\rangle \tag{22}$$

Since $f^*(x) = \sum_k a_k e_k(x)$ and recalling that $K(x,y) = \sum_i \lambda_i e_i(x) e_i(y)$

$$f^*(x) = -\frac{1}{2\gamma_A l}\sum_k \sum_{j=1}^{l} \lambda_k e_k(x) e_k(x_j)\partial_3 V\,(x_j, y_j, f^*) - \frac{\gamma_I}{2\gamma_A}\sum_k \lambda_k \langle Df^* + D^* f^*, e_k\rangle e_k$$

$$= -\frac{1}{2\gamma_A l}\sum_{j=1}^{l} K(x, x_j)\partial_3 V\,(x_j, y_j, f^*) - \frac{\gamma_I}{2\gamma_A}\sum_k \lambda_k \langle Df^* + D^* f^*, e_k\rangle e_k \tag{23}$$

We see that the first summand is a sum of the kernel functions centered at data points. It remains to show that the second summand has an integral representation, i.e. can be written as $\int_{\mathcal{M}} \alpha(y) K(x,y)\, d\mathcal{P}_X(y)$, which is equivalent to being in the image of L_K. To verify this we apply Lemma 3. We need that

$$\sum_k \frac{\lambda_k^2 \langle Df^* + D^* f^*, e_k\rangle^2}{\lambda_k^2} = \sum_k \langle Df^* + D^* f^*, e_k\rangle^2 < \infty \tag{24}$$

Since D, its adjoint operator D^* and hence their sum are bounded the inequality above is satisfied for any function in $\mathcal{S}$. ■

3.3 Manifold Setting[2]

We now show that for the case when $\mathcal{M}$ is a manifold and D is a differential operator, such as the Laplace-Beltrami operator $\Delta_{\mathcal{M}}$, the boundedness condition of Theorem 7 is satisfied. While we

2. We thank Peter Constantin and Todd Dupont for help with this section.

consider the case when the manifold has no boundary, the same argument goes through for manifold with boundary, with, for example, Dirichlet's boundary conditions (vanishing at the boundary). Thus the setting of Theorem 7 is very general, applying, among other things, to arbitrary differential operators on compact domains in Euclidean space.

Let $\mathcal{M}$ be a $\mathcal{C}^\infty$ manifold without boundary with an infinitely differentiable embedding in some ambient space X, D a differential operator with $\mathcal{C}^\infty$ coefficients and let μ, be the measure corresponding to some $\mathcal{C}^\infty$ nowhere vanishing volume form on $\mathcal{M}$. We assume that the kernel $K(x,y)$ is also infinitely differentiable.[3] As before for an operator A, A^* denotes the adjoint operator.

Theorem 8 *Under the conditions above D is a bounded operator $\mathcal{S} \to \mathcal{L}^2_\mu$.*

Proof First note that it is enough to show that D is bounded on $\mathcal{H}_{K_\mathcal{M}}$, since D only depends on the restriction $f_\mathcal{M}$. As before, let $L_{K_\mathcal{M}}(f)(x) = \int_\mathcal{M} f(y) K_\mathcal{M}(x,y)\, d\mu$ is the integral operator associated to K_M. Note that D^* is also a differential operator of the same degree as D. The integral operator $L_{K_\mathcal{M}}$ is bounded (compact) from L^2_μ to any Sobolev space H^{sob}. Therefore the operator $L_{K_\mathcal{M}} D$ is also bounded. We therefore see that $DL_{K_\mathcal{M}}D^*$ is bounded $L^2_\mu \to L^2_\mu$. Therefore there is a constant C, s.t. $\langle DL_{K_\mathcal{M}}D^* f, f\rangle_{L^2_\mu} \le C\|f\|_{L^2_\mu}$.

The square root $T = L^{1/2}_{K_\mathcal{M}}$ of the self-adjoint positive definite operator $L_{K_\mathcal{M}}$ is a self-adjoint positive definite operator as well. Thus $(DT)^* = TD^*$. By definition of the operator norm, for any $\epsilon > 0$ there exists $f \in L^2_\mu$, $\|f\|_{L^2_\mu} \le 1+\epsilon$, such that

$$\|DT\|^2_{L^2_\mu} = \|TD^*\|^2_{L^2_\mu} \le \langle TD^* f, TD^* f\rangle_{L^2_\mu} = \tag{25}$$

$$= \langle DLD^* f, f\rangle_{L^2_\mu} \le \|DLD^*\|_{L^2_\mu} \|f\|^2_{L^2_\mu} \le C(1+\epsilon)^2$$

Therefore the operator $DT : L^2_\mu \to L^2_\mu$ is bounded (and also $\|DT\|_{\mathcal{L}^2_\mu} \le C$, since ϵ is arbitrary).

Now recall that T provides an isometry between L^2_μ and $\mathcal{H}_{K_\mathcal{M}}$. That means that for any $g \in \mathcal{H}_{K_\mathcal{M}}$ there is $f \in L^2_\mu$, such that $Tf = g$ and $\|f\|_{L^2_\mu} = \|g\|_{K_\mathcal{M}}$. Thus $\|Dg\|_{\mathcal{L}^2_\mu} = \|DTf\|_{\mathcal{L}^2_\mu} \le C\|g\|_{K_\mathcal{M}}$, which shows that $T : \mathcal{H}_{K_\mathcal{M}} \to L^2_\mu$ is bounded and concludes the proof. ■

Since $\mathcal{S}$ is a subspace of $\mathcal{H}_K$ the main result follows immediately:

Corollary 9 *D is a bounded operator $\mathcal{S} \to L^2_\mu$ and the conditions of Theorem 7 hold.*

Before finishing the theoretical discussion we obtain a useful

Corollary 10 *The operator $T = L^{1/2}_K$ on L^2_μ is a bounded (and in fact compact) operator $L^2_\mu \to H^{sob}$, where H^{sob} is an arbitrary Sobolev space.*

Proof Follows from the fact that DT is bounded operator $L^2_\mu \to L^2_\mu$ for an arbitrary differential operator D and standard results on compact embeddings of Sobolev spaces (see e.g., Adams (1975)). ■

3. While we have assumed that all objects are infinitely differentiable, it is not hard to specify the precise differentiability conditions. Roughly speaking, a degree k differential operator D is bounded as an operator $\mathcal{H}_K \to L^2_\mu$, if the kernel $K(x,y)$ has $2k$ derivatives.

3.4 The Representer Theorem for the Empirical Case

In the case when $\mathcal{M}$ is unknown and sampled via labeled and unlabeled examples, the Laplace-Beltrami operator on $\mathcal{M}$ may be approximated by the Laplacian of the data adjacency graph (see Belkin (2003); Bousquet et al. (2004) for some discussion). A regularizer based on the graph Laplacian leads to the optimization problem posed in Eqn. 5. We now provide a proof of Theorem 2 which states that the solution to this problem admits a representation in terms of an expansion over labeled and unlabeled points. The proof is based on a simple orthogonality argument (e.g., Scholkopf and Smola (2002)).

Proof *(Theorem 2)* Any function $f \in \mathcal{H}_K$ can be uniquely decomposed into a component $f_{||}$ in the linear subspace spanned by the kernel functions $\{K(x_i, \cdot)\}_{i=1}^{l+u}$, and a component $f_\perp$ orthogonal to it. Thus,

$$f = f_{||} + f_\perp = \sum_{i=1}^{l+u} \alpha_i K(x_i, \cdot) + f_\perp \tag{26}$$

By the reproducing property, as the following arguments show, the evaluation of f on any data point x_j, $1 \leq j \leq l+u$ is independent of the orthogonal component $f_\perp$:

$$f(x_j) = \langle f, K(x_j, \cdot)\rangle = \langle \sum_{i=1}^{l+u} \alpha_i K(x_i, \cdot), K(x_j, \cdot)\rangle + \langle f_\perp, K(x_j, \cdot)\rangle \tag{27}$$

Since the second term vanishes, and $\langle K(x_i, \cdot), K(x_j, \cdot)\rangle = K(x_i, x_j)$, it follows that $f(x_j) = \sum_{i=1}^{l+u} \alpha_i K(x_i, x_j)$. Thus, the empirical terms involving the loss function and the intrinsic norm in the optimization problem in Eqn. 5 depend only on the value of the coefficients $\{\alpha_i\}_{i=1}^{l+u}$ and the gram matrix of the kernel function.

Indeed, since the orthogonal component only increases the norm of f in $\mathcal{H}_K$:

$$\|f\|_K^2 = \|\sum_{i=1}^{l+u} \alpha_i K(x_i, \cdot)\|_K^2 + \|f_\perp\|_K^2 \geq \|\sum_{i=1}^{l+u} \alpha_i K(x_i, \cdot)\|_K^2$$

It follows that the minimizer of problem 5 must have $f_\perp = 0$, and therefore admits a representation $f^*(\cdot) = \sum_{i=1}^{l+u} \alpha_i K(x_i, \cdot)$. ■

The simple form of the minimizer, given by this theorem, allows us to translate our extrinsic and intrinsic regularization framework into optimization problems over the finite dimensional space of coefficients $\{\alpha_i\}_{i=1}^{l+u}$, and invoke the machinery of kernel based algorithms. In the next section, we derive these algorithms, and explore their connections to other related work.

4. Algorithms

We now discuss standard regularization algorithms (RLS and SVM) and present their extensions (LapRLS and LapSVM respectively). These are obtained by solving the optimization problems posed in Eqn. (5) for different choices of cost function V and regularization parameters γ_A, γ_I. To fix notation, we assume we have l labeled examples $\{(x_i, y_i)\}_{i=1}^{l}$ and u unlabeled examples $\{x_j\}_{j=l+1}^{j=l+u}$. We use K interchangeably to denote the kernel function or the Gram matrix.

4.1 Regularized Least Squares

The Regularized Least Squares algorithm is a fully supervised method where we solve :

$$\min_{f\in\mathcal{H}_K} \frac{1}{l}\sum_{i=1}^{l}(y_i - f(x_i))^2 + \gamma\|f\|_K^2 \tag{28}$$

The classical Representer Theorem can be used to show that the solution is of the following form:

$$f^\star(x) = \sum_{i=1}^{l}\alpha_i^\star K(x, x_i) \tag{29}$$

Substituting this form in the problem above, we arrive at following convex differentiable objective function of the l-dimensional variable $\alpha = [\alpha_1 \dots \alpha_l]^T$:

$$\alpha^* = \operatorname{argmin} \frac{1}{l}(Y - K\alpha)^T(Y - K\alpha) + \gamma\alpha^T K\alpha \tag{30}$$

where K is the $l \times l$ gram matrix $K_{ij} = K(x_i, x_j)$ and Y is the label vector $Y = [y_1 \dots y_l]^T$.

The derivative of the objective function vanishes at the minimizer :

$$\frac{1}{l}(Y - K\alpha^*)^T(-K) + \gamma K\alpha^* = 0$$

which leads to the following solution.

$$\alpha^* = (K + \gamma l I)^{-1}Y \tag{31}$$

4.2 Laplacian Regularized Least Squares (LapRLS)

The Laplacian Regularized Least Squares algorithm solves the optimization problem in Eqn. (5) with the squared loss function:

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

As before, the Representer Theorem can be used to show that the solution is an expansion of kernel functions over both the labeled and the unlabeled data :

$$f^\star(x) = \sum_{i=1}^{l+u}\alpha_i^\star K(x, x_i) \tag{32}$$

Substituting this form in the equation above, as before, we arrive at a convex differentiable objective function of the $l + u$-dimensional variable $\alpha = [\alpha_1 \dots \alpha_{l+u}]^T$:

$$\alpha^* = \operatorname*{argmin}_{\alpha\in\mathbb{R}^{l+u}} \frac{1}{l}(Y - JK\alpha)^T(Y - JK\alpha) + \gamma_A\alpha^T K\alpha + \frac{\gamma_I}{(u+l)^2}\alpha^T KLK\alpha \tag{33}$$

where K is the $(l + u) \times (l + u)$ Gram matrix over labeled and unlabeled points; Y is an $(l + u)$ dimensional label vector given by : $Y = [y_1, \dots, y_l, 0, \dots, 0]$ and J is an $(l+u) \times (l+u)$ diagonal matrix given by $J = \mathrm{diag}(1, \dots, 1, 0, \dots, 0)$ with the first l diagonal entries as 1 and the rest 0.

The derivative of the objective function vanishes at the minimizer :

$$\frac{1}{l}(Y - JK\alpha)^T(-JK) + (\gamma_A K + \frac{\gamma_I l}{(u+l)^2} KLK)\alpha = 0 \tag{34}$$

which leads to the following solution.

$$\alpha^* = (JK + \gamma_A l I + \frac{\gamma_I l}{(u+l)^2} LK)^{-1} Y \tag{35}$$

Note that when $\gamma_I = 0$, Eqn. (35) gives zero coefficients over unlabeled data, and the coefficients over the labeled data are exactly those for standard RLS.

4.3 Support Vector Machine Classification

Here we outline the SVM approach to binary classification problems. For SVMs, the following problem is solved :

$$\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} (1 - y_i f(x_i))_+ + \gamma \|f\|_K^2$$

where the hinge loss is defined as: $(1 - yf(x))_+ = \max(0, 1 - yf(x))$ and the labels $y_i \in \{-1, +1\}$.

Again, the solution is given by:

$$f^\star(x) = \sum_{i=1}^{l} \alpha_i^\star K(x, x_i) \tag{36}$$

Following SVM expositions, the above problem can be equivalently written as:

$$\begin{aligned} \min_{f \in \mathcal{H}_K, \xi_i \in \mathbb{R}} & \frac{1}{l} \sum_{i=1}^{l} \xi_i + \gamma \|f\|_K^2 \\ \text{subject to : } & y_i f(x_i) \geq 1 - \xi_i \quad i = 1, \ldots, l \\ & \xi_i \geq 0 \quad i = 1, \ldots, l \end{aligned} \tag{37}$$

Using the Lagrange multipliers technique, and benefiting from strong duality, the above problem has a simpler quadratic dual program in the Lagrange multipliers $\beta = [\beta_1, \ldots, \beta_l]^T \in \mathbb{R}^l$:

$$\begin{aligned} \beta^\star &= \max_{\beta \in \mathbb{R}^l} \sum_{i=1}^{l} \beta_i - \frac{1}{2}\beta^T Q \beta \\ \text{subject to : } & \sum_{i=1}^{l} y_i \beta_i = 0 \\ & 0 \leq \beta_i \leq \frac{1}{l} \quad i = 1, \ldots, l \end{aligned} \tag{38}$$

where the equality constraint arises due to an unregularized bias term that is often added to the sum in Eqn. (36), and the following notation is used :

$$\begin{aligned} Y &= \text{diag}(y_1, y_2, ..., y_l) \\ Q &= Y\left(\frac{K}{2\gamma}\right)Y \\ \alpha^\star &= \frac{Y\beta^\star}{2\gamma} \end{aligned} \tag{39}$$

Here again, K is the gram matrix over labeled points. SVM practitioners may be familiar with a slightly different parameterization involving the C parameter : $C = \frac{1}{2\gamma l}$ is the weight on the hinge loss term (instead of using a weight γ on the norm term in the optimization problem). The C parameter appears as the upper bound (instead of $\frac{1}{l}$) on the values of β in the quadratic program. For additional details on the derivation and alternative formulations of SVMs, see Scholkopf and Smola (2002), Rifkin (2002).

4.4 Laplacian Support Vector Machines

By including the intrinsic smoothness penalty term, we can extend SVMs by solving the following problem:

$$\min_{f \in \mathcal{H}_K} \frac{1}{l}\sum_{i=1}^{l}(1 - y_i f(x_i))_+ + \gamma_A \|f\|_K^2 + \frac{\gamma_I}{(u+l)^2}\mathbf{f}^T L \mathbf{f} \tag{40}$$

By the representer theorem,as before, the solution to the problem above is given by:

$$f^\star(x) = \sum_{i=1}^{l+u} \alpha_i^\star K(x, x_i) \tag{41}$$

Often in SVM formulations, an unregularized bias term b is added to the above form. Again, the primal problem can be easily seen to be the following:

$$\min_{\alpha \in \mathbb{R}^{l+u}, \xi \in \mathbb{R}^l} \frac{1}{l}\sum_{i=1}^{l}\xi_i + \gamma_A \alpha^T K \alpha + \frac{\gamma_I}{(u+l)^2}\alpha^T K L K \alpha \tag{42}$$

$$\text{subject to : } y_i(\sum_{j=1}^{l+u}\alpha_j K(x_i, x_j) + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l$$

$$\xi_i \geq 0 \quad i = 1, \ldots, l$$

Introducing the Lagrangian, with β_i, ζ_i as Lagrange multipliers:

$$\begin{aligned} L(\alpha, \xi, b, \beta, \zeta) &= \frac{1}{l}\sum_{i=1}^{l}\xi_i + \frac{1}{2}\alpha^T(2\gamma_A K + 2\frac{\gamma_A}{(l+u)^2}KLK)\alpha \\ &\quad - \sum_{i=1}^{l}\beta_i(y_i(\sum_{j=1}^{l+u}\alpha_j K(x_i, x_j) + b) - 1 + \xi_i) - \sum_{i=1}^{l}\zeta_i \xi_i \end{aligned} \tag{43}$$

Passing to the dual requires the following steps:

$$\begin{aligned}\frac{\partial L}{\partial b} = 0 &\implies \sum_{i=1}^{l} \beta_i y_i = 0 \qquad (44)\\ \frac{\partial L}{\partial \xi_i} = 0 &\implies \frac{1}{l} - \beta_i - \zeta_i = 0\\ &\implies 0 \le \beta_i \le \frac{1}{l} \quad (\xi_i, \zeta_i \text{ are non-negative})\end{aligned}$$

Using above identities, we formulate a reduced Lagrangian:

$$\begin{aligned}L^R(\alpha,\beta) &= \frac{1}{2}\alpha^T(2\gamma_A K + 2\frac{\gamma_I}{(u+l)^2}KLK)\alpha - \sum_{i=1}^{l}\beta_i(y_i\sum_{j=1}^{l+u}\alpha_j K(x_i,x_j) - 1)\\ &= \frac{1}{2}\alpha^T(2\gamma_A K + 2\frac{\gamma_I}{(u+l)^2}KLK)\alpha - \alpha^T K J^T Y \beta + \sum_{i=1}^{l}\beta_i \qquad (45)\end{aligned}$$

where $J = [I \;\; 0]$ is an $l \times (l+u)$ matrix with I as the $l \times l$ identity matrix (assuming the first l points are labeled) and $Y = \mathrm{diag}(y_1, y_2, ..., y_l)$.

Taking derivative of the reduced Lagrangian with respect to α:

$$\frac{\partial L^R}{\partial \alpha} = (2\gamma_A K + 2\frac{\gamma_I}{(u+l)^2}KLK)\alpha - KJ^T Y\beta$$

This implies:

$$\alpha = (2\gamma_A I + 2\frac{\gamma_I}{(u+l)^2}LK)^{-1}J^T Y\beta^\star \qquad (46)$$

Note that the relationship between α and β is no longer as simple as the SVM algorithm. In particular, the $(l+u)$ expansion coefficients are obtained by solving a linear system involving the l dual variables that will appear in the SVM dual problem.

Substituting back in the reduced Lagrangian we get:

$$\beta^* = \max_{\beta\in\mathbb{R}^l} \sum_{i=1}^{l}\beta_i - \frac{1}{2}\beta^T Q\beta \qquad (47)$$

$$\begin{aligned}\text{subject to :} \quad & \sum_{i=1}^{l}\beta_i y_i = 0\\ & 0 \le \beta_i \le \frac{1}{l} \;\; i = 1,\ldots,l \qquad (48)\end{aligned}$$

where

$$Q = YJK(2\gamma_A I + 2\frac{\gamma_I}{(l+u)^2}LK)^{-1}J^T Y$$

Laplacian SVMs can be implemented by using a standard SVM solver with the quadratic form induced by the above matrix, and using the solution to obtain the expansion coefficients by solving the linear system in Eqn. (46).

Note that when $\gamma_I = 0$, the SVM QP and Eqns. (47,46), give zero expansion coefficients over the unlabeled data. The expansion coefficients over the labeled data and the Q matrix are as in standard SVM, in this case.

The Manifold Regularization algorithms are summarized in the Table 1.

Table 1: A Summary of the algorithms

	Manifold Regularization algorithms
Input:	l labeled examples $\{(x_i, y_i)\}_{i=1}^{l}$, u unlabeled examples $\{x_j\}_{j=l+1}^{l+u}$
Output:	Estimated function $f : \mathbb{R}^n \to \mathbb{R}$
Step 1	► Construct data adjacency graph with $(l+u)$ nodes using, e.g, k nearest neighbors or a graph kernel. Choose edge weights W_{ij}, e.g. binary weights or heat kernel weights $W_{ij} = e^{-\|x_i - x_j\|^2/4t}$.
Step 2	► Choose a kernel function $K(x, y)$. Compute the Gram matrix $K_{ij} = K(x_i, x_j)$.
Step 3	► Compute graph Laplacian matrix : $L = D - W$ where D is a diagonal matrix given by $D_{ii} = \sum_{j=1}^{l+u} W_{ij}$.
Step 4	► Choose γ_A and γ_I.
Step 5	► Compute α^* using Eqn. (35) for squared loss (Laplacian RLS) or using Eqns. (47,46) together with the SVM QP solver for soft margin loss (Laplacian SVM).
Step 6	► Output function $f^*(x) = \sum_{i=1}^{l+u} \alpha_i^* K(x_i, x)$.

Efficiency Issues: It is worth noting that our algorithms compute the inverse of a dense Gram matrix which leads to $O((l+u)^3)$ complexity. This may be impractical for large datasets. In the case of linear kernels, instead of using Eqn. 6, we can directly write $f^\star(x) = w^T x$ and solve for the weight vector w using a primal optimization method. This is much more efficient when the data is low-dimensional. For highly sparse datasets, e.g. in text categorization problems, effective conjugate gradient schemes can be utilized in a large scale implementation, as outlined in Sindhwani et al. (2006). For the non-linear case, one may obtain approximate solutions (e.g. using greedy, matching pursuit techniques) where the optimization problem is solved over the span of a small set of basis functions instead of using the full representation in Eqn. 6. We note these directions for future work. In section 5, we evaluate the empirical performance of our algorithms with exact computations as outlined in Table 1 with non-linear kernels. For other recent work addressing scalability issues in semi-supervised learning, see, e.g., Tsang and Kwok. (2005); Bengio et al. (2004).

4.5 Related Work and Connections to Other Algorithms

In this section we survey various approaches to semi-supervised and transductive learning and highlight connections of Manifold Regularization to other algorithms.

Transductive SVM (TSVM) (Vapnik, 1998; Joachims, 1999): TSVMs are based on the following optimization principle :

$$f^* = \underset{\substack{f \in \mathcal{H}_K \\ y_{l+1}, \dots y_{l+u}}}{\operatorname{argmin}} \; C \sum_{i=1}^{l} (1 - y_i f(x_i))_+ + C^* \sum_{i=l+1}^{l+u} (1 - y_i f(x_i))_+ + \|f\|_K^2 \tag{49}$$

which proposes a joint optimization of the SVM objective function over binary-valued labels on the unlabeled data and functions in the RKHS. Here, C, C^* are parameters that control the relative hinge-loss over labeled and unlabeled sets. The joint optimization is implemented in Joachims (1999) by first using an inductive SVM to label the unlabeled data and then iteratively solving SVM quadratic programs, at each step switching labels to improve the objective function. However this procedure is susceptible to local minima and requires an unknown, possibly large number of label switches before converging. Note that even though TSVM were inspired by transductive inference, they do provide an out-of-sample extension.

Semi-Supervised SVMs (S^{3}VM) (Bennett and Demiriz, 1999; Fung and Mangasarian, 2001): S^{3}VM incorporate unlabeled data by including the minimum hinge-loss for the two choices of labels for each unlabeled example. This is formulated as a mixed-integer program for linear SVMs in Bennett and Demiriz (1999) and is found to be intractable for large amounts of unlabeled data. Fung and Mangasarian (2001) reformulate this approach as a concave minimization problem which is solved by a successive linear approximation algorithm. The presentation of these algorithms is restricted to the linear case.

Measure-Based Regularization (Bousquet et al., 2004): The conceptual framework of this work is closest to our approach. The authors consider a gradient based regularizer that penalizes variations of the function more in high density regions and less in low density regions leading to the following optimization principle:

$$f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{l} V(f(x_i), y_i) + \gamma \int_X \langle \nabla f(x), \nabla f(x) \rangle p(x) dx \tag{50}$$

where p is the density of the marginal distribution $\mathcal{P}_X$. The authors observe that it is not straightforward to find a kernel for arbitrary densities p, whose associated RKHS norm is

$$\int \langle \nabla f(x), \nabla f(x) \rangle p(x) dx$$

Thus, in the absence of a representer theorem, the authors propose to perform minimization of the regularized loss on a fixed set of basis functions chosen apriori, i.e, $\mathcal{F} = \{\sum_{i=1}^{q} \alpha_i \phi_i\}$. For the hinge loss, this paper derives an SVM quadratic program in the coefficients $\{\alpha_i\}_{i=1}^{q}$ whose Q matrix is calculated by computing q^2 integrals over gradients of the basis functions. However the algorithm does not demonstrate performance improvements in real world experiments. It is also worth noting that while Bousquet et al. (2004) use the gradient $\nabla f(x)$ in the ambient space, we use the gradient over a submanifold $\nabla_{\mathcal{M}} f$ for penalizing the function. In a situation where the data truly lies on or near a submanifold $\mathcal{M}$, the difference between these two penalizers can be significant since smoothness in the normal direction to the data manifold is irrelevant to classification or regression.

Graph Based Approaches See, e.g., Blum and Chawla (2001); Chapelle et al. (2003); Szummer and Jaakkola (2002); Zhou et al. (2004); Zhu et al. (2003, 2005); Kemp et al. (2004); Joachims

(2003); Belkin and Niyogi (2003a): A variety of graph based methods have been proposed for transductive inference. However, these methods do not provide an out-of-sample extension. In Zhu et al. (2003), nearest neighbor labeling for test examples is proposed once unlabeled examples have been labeled by transductive learning. In Chapelle et al. (2003), test points are approximately represented as a linear combination of training and unlabeled points in the feature space induced by the kernel. For Graph Regularization and Label Propagation see (Smola and Kondor, 2003; Belkin et al., 2004; Zhu et al., 2003). Smola and Kondor (2003) discusses the construction of a canonical family of graph regularizers based on the graph Laplacian. Zhu et al. (2005) presents a non-parametric construction of graph regularizers.

Manifold regularization provides natural out-of-sample extensions to several graph based approaches. These connections are summarized in Table 2.

We also note the recent work Delalleau et al. (2005) on out-of-sample extensions for semi-supervised learning where an induction formula is derived by assuming that the addition of a test point to the graph does not change the transductive solution over the unlabeled data.

Cotraining (Blum and Mitchell, 1998): The Co-training algorithm was developed to integrate abundance of unlabeled data with availability of multiple sources of information in domains like web-page classification. Weak learners are trained on labeled examples and their predictions on subsets of unlabeled examples are used to mutually expand the training set. Note that this setting may not be applicable in several cases of practical interest where one does not have access to multiple information sources.

Bayesian Techniques See e.g., Nigam et al. (2000); Seeger (2001); Corduneanu and Jaakkola (2003). An early application of semi-supervised learning to Text classification appeared in Nigam et al. (2000) where a combination of EM algorithm and Naive-Bayes classification is proposed to incorporate unlabeled data. Seeger (2001) provides a detailed overview of Bayesian frameworks for semi-supervised learning. The recent work in Corduneanu and Jaakkola (2003) formulates a new information-theoretic principle to develop a regularizer for conditional log-likelihood.

Table 2: Connections of Manifold Regularization to other algorithms

Parameters	Corresponding algorithms (square loss or hinge loss)
$\gamma_A \geq 0 \ \ \gamma_I \geq 0$	Manifold Regularization
$\gamma_A \geq 0 \ \ \gamma_I = 0$	Standard Regularization (RLS or SVM)
$\gamma_A \to 0 \ \ \gamma_I > 0$	Out-of-sample extension for Graph Regularization (RLS or SVM)
$\gamma_A \to 0 \ \ \gamma_I \to 0$ $\gamma_I \gg \gamma_A$	Out-of-sample extension for Label Propagation (RLS or SVM)
$\gamma_A \to 0 \ \ \gamma_I = 0$	Hard margin SVM or Interpolated RLS

5. Experiments

We performed experiments on a synthetic dataset and three real world classification problems arising in visual and speech recognition, and text categorization. Comparisons are made with inductive methods (SVM, RLS). We also compare Laplacian SVM with Transductive SVM. All software and

datasets used for these experiments will be made available at:
http://www.cs.uchicago.edu/∼vikass/manifoldregularization.html.

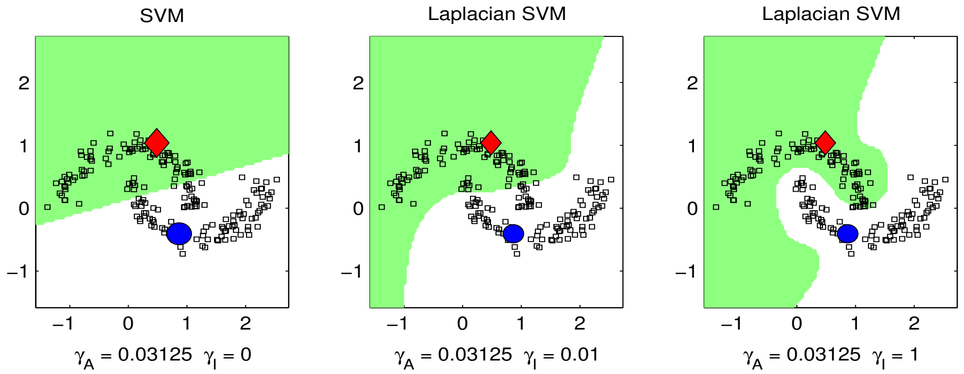
For further experimental benchmark studies and comparisons with numerous other methods, we refer the reader to Chapelle et al. (2006); Sindhwani et al. (2006, 2005).

5.1 Synthetic Data : Two Moons Dataset

The two moons dataset is shown in Figure 2. The dataset contains 200 examples with only 1 labeled example for each class. Also shown are the decision surfaces of Laplacian SVM for increasing values of the intrinsic regularization parameter γ_I. When $\gamma_I = 0$, Laplacian SVM disregards unlabeled data and returns the SVM decision boundary which is fixed by the location of the two labeled points. As γ_I is increased, the intrinsic regularizer incorporates unlabeled data and causes the decision surface to appropriately adjust according to the geometry of the two classes.

In Figure 3, the best decision surfaces across a wide range of parameter settings are also shown for SVM, Transductive SVM and Laplacian SVM. Figure 3 demonstrates how TSVM fails to find the optimal solution, probably since it gets stuck in a local minimum. The Laplacian SVM decision boundary seems to be intuitively most satisfying.

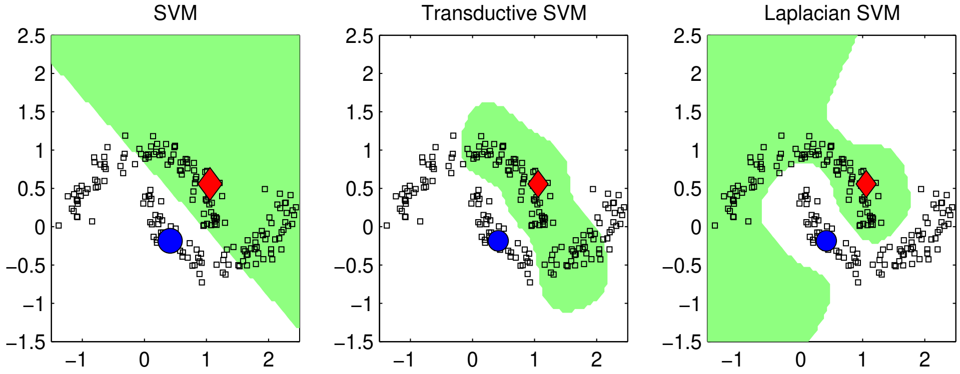
Figure 2: Laplacian SVM with RBF Kernels for various values of γ_I. Labeled points are shown in color, other points are unlabeled.



5.2 Handwritten Digit Recognition

In this set of experiments we applied Laplacian SVM and Laplacian RLS algorithms to 45 binary classification problems that arise in pairwise classification of handwritten digits. The first 400 images for each digit in the USPS training set (preprocessed using PCA to 100 dimensions) were taken to form the training set. The remaining images formed the test set. 2 images for each class were randomly labeled (l=2) and the rest were left unlabeled (u=398). Following Scholkopf et al. (1995), we chose to train classifiers with polynomial kernels of degree 3, and set the weight on the regularization term for inductive methods as $\gamma l = 0.05 (C = 10)$. For manifold regularization, we chose to split the same weight in the ratio $1 : 9$ so that $\gamma_A l = 0.005$, $\frac{\gamma_I l}{(u+l)^2} = 0.045$. The observations reported in this section hold consistently across a wide choice of parameters.

Figure 3: Two Moons Dataset: Best decision surfaces using RBF kernels for SVM, TSVM and Laplacian SVM. Labeled points are shown in color, other points are unlabeled.



In Figure 4, we compare the error rates of manifold regularization algorithms, inductive classifiers and TSVM, at the break-even points in the precision-recall curves for the 45 binary classification problems. These results are averaged over 10 random choices of labeled examples. The following comments can be made: (a) Manifold regularization results in significant improvements over inductive classification, for both RLS and SVM, and either compares well or significantly outperforms TSVM across the 45 classification problems. Note that TSVM solves multiple quadratic programs in the size of the labeled and unlabeled sets whereas LapSVM solves a single QP (Eqn. 47) in the size of the labeled set, followed by a linear system (Eqn. 46). This resulted in substantially faster training times for LapSVM in this experiment. (b) Scatter plots of performance on test and unlabeled data sets, in the bottom row of Figure 4, confirm that the out-of-sample extension is good for both LapRLS and LapSVM. (c) As shown, in the rightmost scatter plot in the bottom row of Figure 4, are standard deviation of error rates obtained by Laplacian SVM and TSVM. We found Laplacian algorithms to be significantly more stable than the inductive methods and TSVM, with respect to choice of the labeled data

In Figure 5, we demonstrate the benefit of unlabeled data as a function of the number of labeled examples.

Table 3: USPS: one-versus-rest multiclass error rates

Method	SVM	TSVM	LapSVM	RLS	LapRLS
Error	23.6	26.5	12.7	23.6	12.7

We also performed one-vs-rest multiclass experiments on the USPS test set with $l = 50$ and $u = 1957$ with 10 random splits as provided by Chapelle and Zien (2005). The mean error rates in predicting labels of unlabeled data are reported in Table 3. In this experiment, TSVM actually performs worse than the SVM baseline probably since local minima problems become severe in a one-vs-rest setting. For several other experimental observations and comparisons on this dataset, see Sindhwani et al. (2005).

Figure 4: USPS Experiment - (Top row) Error Rates at Precision-Recall Break-even points for 45 binary classification problems. (Bottom row) Scatter plots of error rates on test and unlabeled data for Laplacian RLS, Laplacian SVM; and standard deviations in test errors of Laplacian SVM and TSVM.

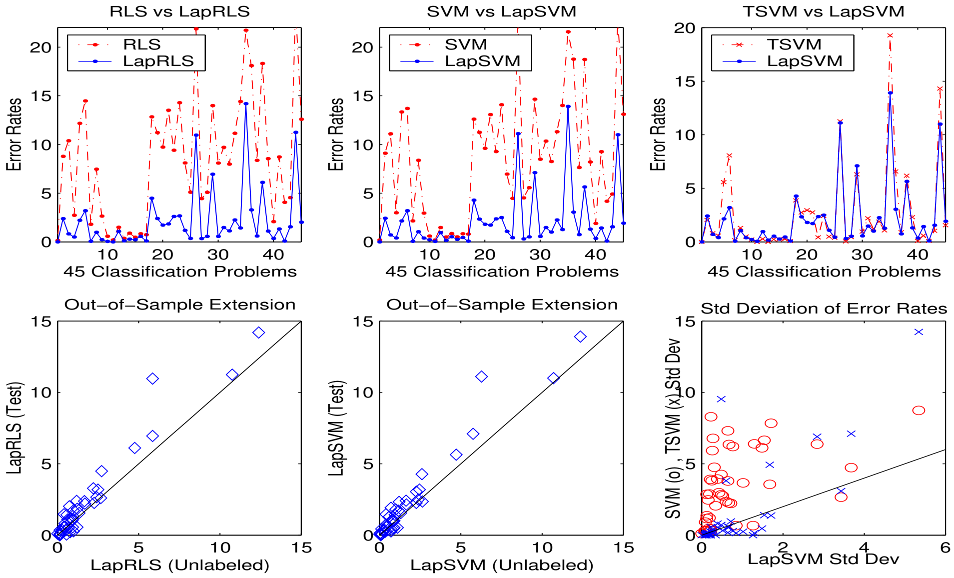
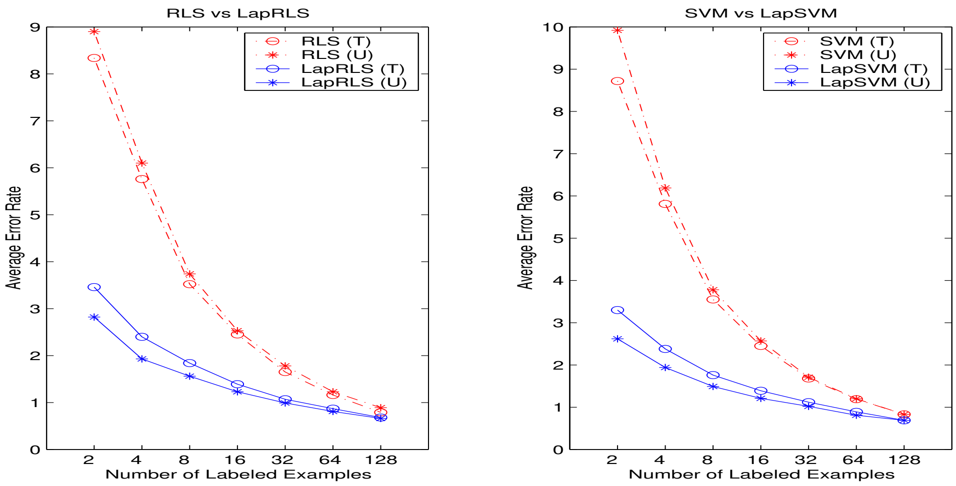


Figure 5: USPS Experiment - Mean Error Rate at Precision-Recall Break-even points as a function of number of labeled points (T: Test Set, U: Unlabeled Set)



5.3 Spoken Letter Recognition

This experiment was performed on the Isolet database of letters of the English alphabet spoken in isolation (available from the UCI machine learning repository). The data set contains utterances of

150 subjects who spoke the name of each letter of the English alphabet twice. The speakers are grouped into 5 sets of 30 speakers each, referred to as isolet1 through isolet5. For the purposes of this experiment, we chose to train on the first 30 speakers (isolet1) forming a training set of 1560 examples, and test on isolet5 containing 1559 examples (1 utterance is missing in the database due to poor recording). We considered the task of classifying the first 13 letters of the English alphabet from the last 13. We considered 30 binary classification problems corresponding to 30 splits of the training data where all 52 utterances of one speaker were labeled and all the rest were left unlabeled. The test set is composed of entirely new speakers, forming the separate group isolet5.

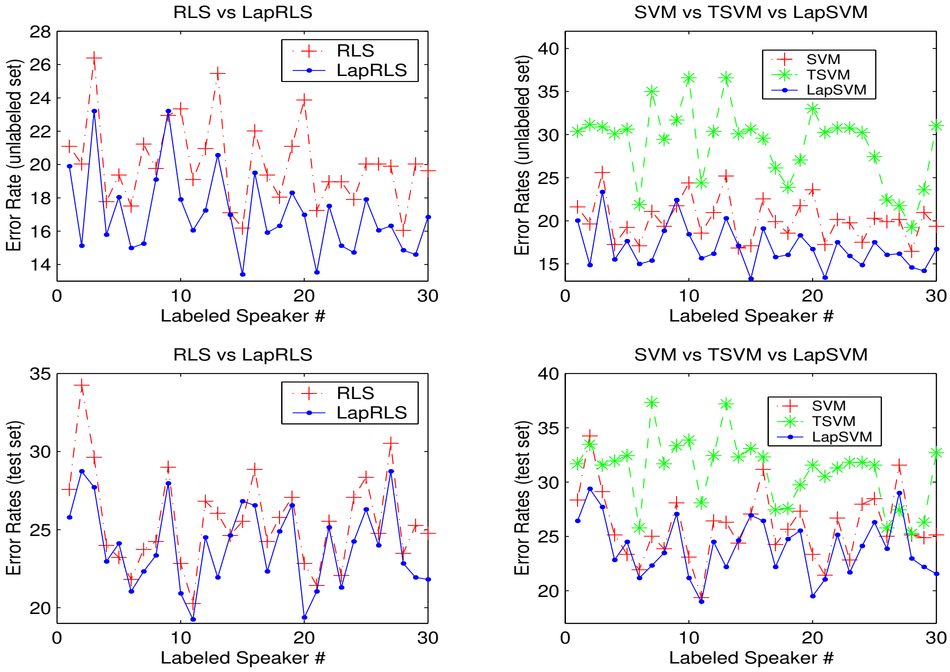


Figure 6: Isolet Experiment - Error Rates at precision-recall break-even points of 30 binary classification problems

We chose to train with RBF kernels of width $\sigma = 10$ (this was the best value among several settings with respect to 5-fold cross-validation error rates for the fully supervised problem using standard SVM). For SVM and RLS we set $\gamma l = 0.05$ ($C = 10$) (this was the best value among several settings with respect to mean error rates over the 30 splits). For Laplacian RLS and Laplacian SVM we set $\gamma_A l = \frac{\gamma_I l}{(u+l)^2} = 0.005$.

In Figure 6, we compare these algorithms. The following comments can be made: (a) LapSVM and LapRLS make significant performance improvements over inductive methods and TSVM, for predictions on unlabeled speakers that come from the same group as the labeled speaker, over all choices of the labeled speaker. (b) On Isolet5 which comprises of a separate group of speakers, performance improvements are smaller but consistent over the choice of the labeled speaker. This can be expected since there appears to be a systematic bias that affects all algorithms, in favor of same-group speakers. To test this hypothesis, we performed another experiment in which the

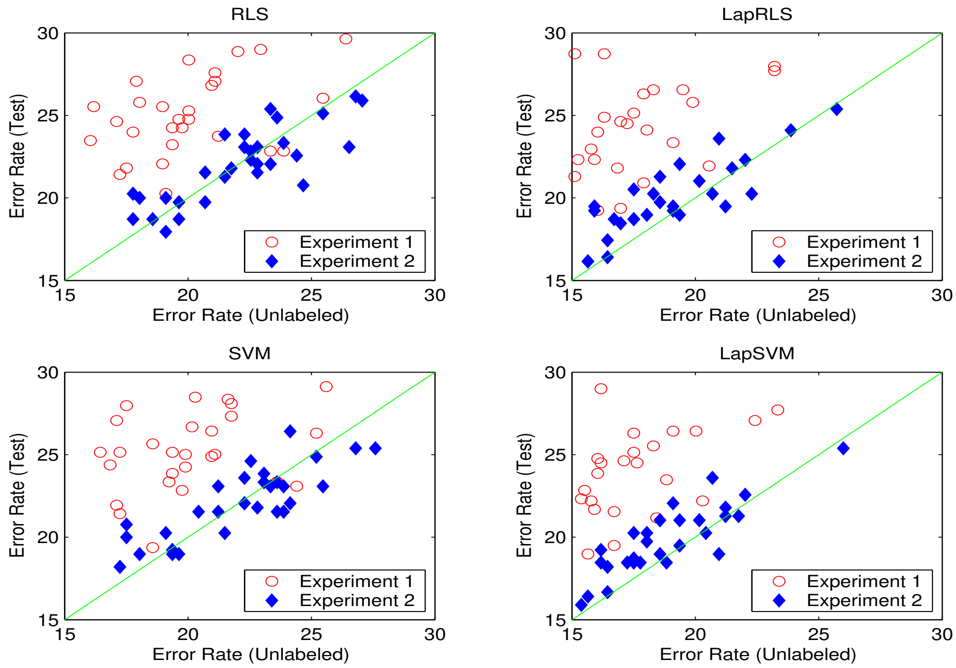


Figure 7: Isolet Experiment - Error Rates at precision-recall break-even points on Test set Versus Unlabeled Set. In Experiment 1, the training data comes from Isolet 1 and the test data comes from Isolet5; in Experiment 2, both training and test sets come from Isolet1.

training and test utterances are both drawn from Isolet1. Here, the second utterance of each letter for each of the 30 speakers in Isolet1 was taken away to form the test set containing 780 examples. The training set consisted of the first utterances for each letter. As before, we considered 30 binary classification problems arising when all utterances of one speaker are labeled and other training speakers are left unlabeled. The scatter plots in Figure 7 confirm our hypothesis, and show high correlation between in-sample and out-of-sample performance of our algorithms in this experiment. It is encouraging to note performance improvements with unlabeled data in Experiment 1 where the test data comes from a slightly different distribution. This robustness is often desirable in real-world applications.

Table 4: Isolet: one-versus-rest multiclass error rates

Method	SVM	TSVM	LapSVM	RLS	LapRLS
Error (unlabeled)	28.6	46.6	24.5	28.3	24.1
Error (test)	36.9	43.3	33.7	36.3	33.3

In Table 4 we report mean error rates over the 30 splits from one-vs-rest 26-class experiments on this dataset. The parameters were held fixed as in the 2-class setting. The failure of TSVM in producing reasonable results on this dataset has also been observed in Joachims (2003). With LapSVM and LapRLS we obtain around 3 to 4% improvement over their supervised counterparts.

5.4 Text Categorization

We performed Text Categorization experiments on the WebKB dataset which consists of 1051 web pages collected from Computer Science department web-sites of various universities. The task is to classify these web pages into two categories: *course* or *non-course*. We considered learning classifiers using only textual content of the web pages, ignoring link information. A bag-of-word vector space representation for documents is built using the the top 3000 words (skipping HTML headers) having highest mutual information with the class variable, followed by TFIDF mapping[4]. Feature vectors are normalized to unit length. 9 documents were found to contain none of these words and were removed from the dataset.

For the first experiment, we ran LapRLS and LapSVM in a transductive setting, with 12 randomly labeled examples (3 course and 9 non-course) and the rest unlabeled. In Table 4, we report the precision and error rates at the precision-recall break-even point averaged over 100 realizations of the data, and include results reported in Joachims (2003) for Spectral Graph Transduction, and the Cotraining algorithm (Blum and Mitchell, 1998) for comparison. We used 15 nearest neighbor graphs, weighted by cosine distances and used iterated Laplacians of degree 3. For inductive methods, $\gamma_A l$ was set to 0.01 for RLS and 1.00 for SVM. For LapRLS and LapSVM, γ_A was set as in inductive methods, with $\frac{\gamma_I l}{(l+u)^2} = 100\gamma_A l$. These parameters were chosen based on a simple grid search for best performance over the first 5 realizations of the data. Linear Kernels and cosine distances were used since these have found wide-spread applications in text classification problems, e.g., in (Dumais et al., 1998).

Table 5: Precision and Error Rates at the Precision-Recall Break-even Points of supervised and transductive algorithms.

Method	PRBEP	Error
k-NN Joachims (2003)	73.2	13.3
SGT Joachims (2003)	86.2	6.2
Naive-Bayes Blum and Mitchell (1998)	—	12.9
Cotraining Blum and Mitchell (1998)	—	6.20
SVM	76.39 (5.6)	10.41 (2.5)
TSVM[5]	88.15 (1.0)	5.22 (0.5)
LapSVM	87.73 (2.3)	5.41 (1.0)
RLS	73.49 (6.2)	11.68 (2.7)
LapRLS	86.37 (3.1)	5.99 (1.4)

Since the exact datasets on which these algorithms were run, somewhat differ in preprocessing, preparation and experimental protocol, these results are only meant to suggest that Manifold

4. TFIDF stands for Term Frequency Inverse Document Frequency. It is a common document preprocessing procedure, which combines the number of occurrences of a given term with the number of documents containing it.

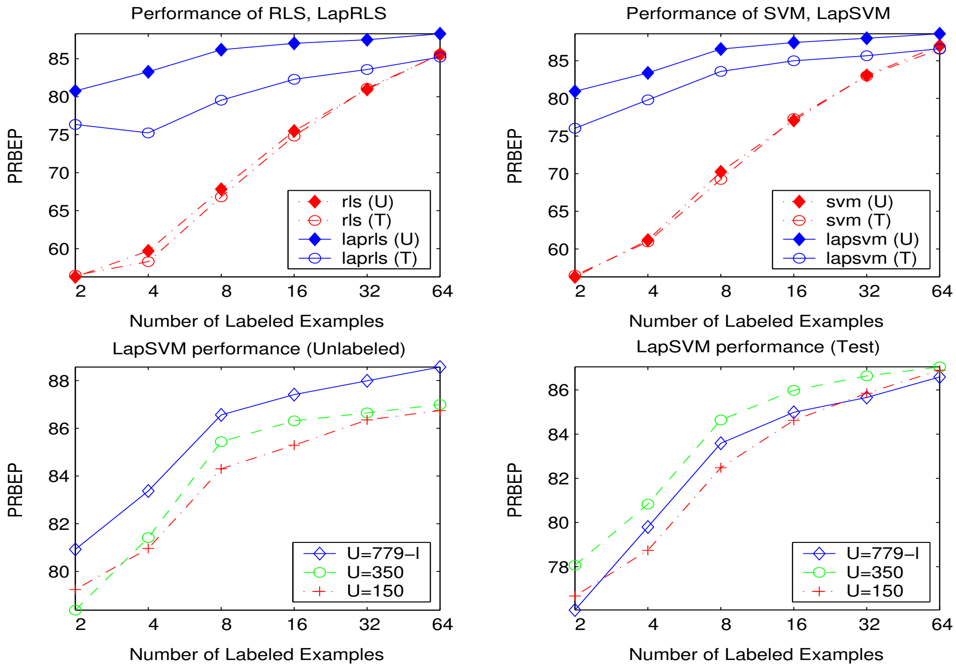


Figure 8: WebKb Text Classification Experiment : The top panel presents performance in terms of precision-recall break-even points (PRBEP) of RLS,SVM,Laplacian RLS and Laplacian SVM as a function of number of labeled examples, on Test (marked as T) set and Unlabeled set (marked as U and of size 779-number of labeled examples). The bottom panel presents performance curves of Laplacian SVM for different number of unlabeled points.

Regularization algorithms perform similar to state-of-the-art methods for transductive inference in text classification problems. The following comments can be made: (a) Transductive categorization with LapSVM and LapRLS leads to significant improvements over inductive categorization with SVM and RLS. (b) Joachims (2003) reports 91.4% precision-recall breakeven point, and 4.6% error rate for TSVM. Results for TSVM reported in the table were obtained when we ran the TSVM implementation using SVM-Light software on this particular dataset. The average training time for TSVM was found to be more than 10 times slower than for LapSVM (c) The Co-training results were obtained on unseen test datasets utilizing additional hyperlink information, which was excluded in our experiments. This additional information is known to improve performance, as demonstrated in Joachims (2003) and Blum and Mitchell (1998).

In the next experiment, we randomly split the WebKB data into a test set of 263 examples and a training set of 779 examples. We noted the performance of inductive and semi-supervised classifiers on unlabeled and test sets as a function of the number of labeled examples in the training set. The performance measure is the precision-recall break-even point (PRBEP), averaged over 100 random data splits. Results are presented in the top panel of Figure 8. The benefit of unlabeled data can be seen by comparing the performance curves of inductive and semi-supervised classifiers.

We also performed experiments with different sizes of the training set, keeping a randomly chosen test set of 263 examples. The bottom panel in Figure 8 presents the quality of transduction and semi-supervised learning with Laplacian SVM (Laplacian RLS performed similarly) as a function of the number of labeled examples for different amounts of unlabeled data. We find that transduction improves with increasing unlabeled data. We expect this to be true for test set performance as well, but do not observe this consistently possibly since we use a fixed set of parameters that become suboptimal as unlabeled data is increased. The optimal choice of the regularization parameters depends on the amount of labeled and unlabeled data, and should be adjusted by the model selection protocol accordingly.

6. Unsupervised and Fully Supervised Cases

While the previous discussion concentrated on the semi-supervised case, our framework covers both unsupervised and fully supervised cases as well. We briefly discuss each in turn.

6.1 Unsupervised Learning: Clustering and Data Representation

In the unsupervised case one is given a collection of unlabeled data points $x_1, \ldots, x_u$. Our basic algorithmic framework embodied in the optimization problem in Eqn. 3 has three terms: (i) fit to labeled data, (ii) extrinsic regularization and (iii) intrinsic regularization. Since no labeled data is available, the first term does not arise anymore. Therefore we are left with the following optimization problem:

$$\min_{f \in \mathcal{H}_K} \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2 \tag{51}$$

Of course, only the ratio $\gamma = \frac{\gamma_A}{\gamma_I}$ matters. As before $\|f\|_I^2$ can be approximated using the unlabeled data. Choosing $\|f\|_I^2 = \int_{\mathcal{M}} \langle \nabla_{\mathcal{M}} f, \nabla_{\mathcal{M}} f \rangle$ and approximating it by the empirical Laplacian, we are left with the following optimization problem:

$$f^* = \underset{\substack{\sum_i f(x_i)=0;\ \sum_i f(x_i)^2=1 \\ f \in \mathcal{H}_K}}{\operatorname{argmin}} \gamma \|f\|_K^2 + \sum_{i \sim j} (f(x_i) - f(x_j))^2 \tag{52}$$

Note that to avoid degenerate solutions we need to impose some additional conditions (cf. Belkin and Niyogi (2003b)). It turns out that a version of Representer theorem still holds showing that the solution to Eqn. 52 admits a representation of the form

$$f^* = \sum_{i=1}^{u} \alpha_i K(x_i, \cdot\,)$$

By substituting back in Eqn. 52, we come up with the following optimization problem:

$$\alpha = \underset{\substack{\mathbf{1}^T K\alpha=0 \\ \alpha^T K^2 \alpha=1}}{\operatorname{argmin}} \gamma \|f\|_K^2 + \sum_{i \sim j} (f(x_i) - f(x_j))^2 \tag{53}$$

where $\mathbf{1}$ is the vector of all ones and $\alpha = (\alpha_1, \ldots, \alpha_u)$ and K is the corresponding Gram matrix.

Letting P be the projection onto the subspace of $\mathbb{R}^u$ orthogonal to $K\mathbf{1}$, one obtains the solution for the constrained quadratic problem, which is given by the generalized eigenvalue problem

$$P(\gamma K + \ KLK)P\mathbf{v} = \lambda P K^2 P \mathbf{v} \tag{54}$$

The final solution is given by $\alpha = P\mathbf{v}$, where $\mathbf{v}$ is the eigenvector corresponding to the smallest eigenvalue.

Figure 9: Two Moons Dataset: Regularized Clustering

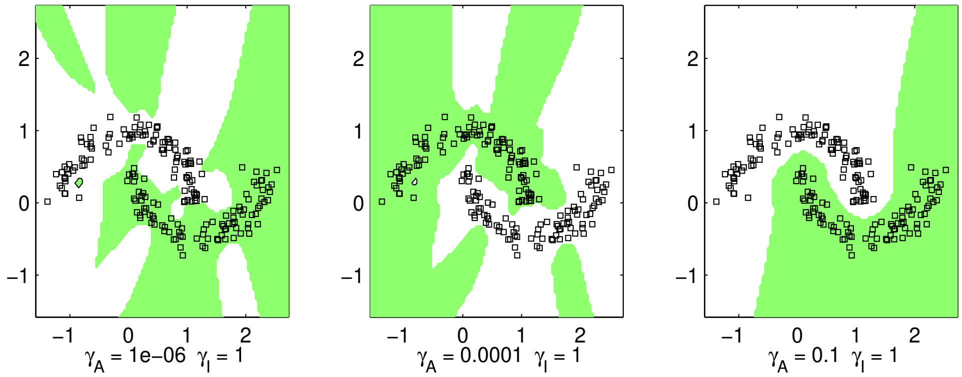
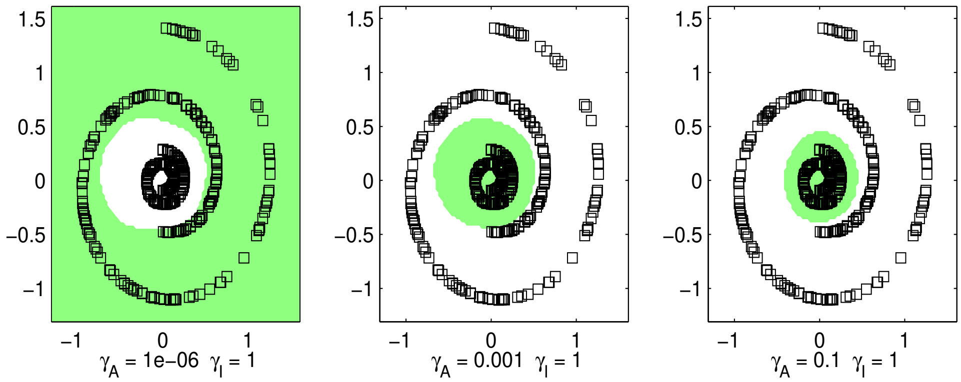


Figure 10: Two Spirals Dataset: Regularized Clustering



Remark 1: The framework for clustering sketched above provides a way of doing regularized spectral clustering, where γ controls the smoothness of the resulting function in the ambient space. We also obtain a natural out-of-sample extension for clustering points not in the original data set. Figures 9,10 show results of this method on two two-dimensional clustering problems. Unlike recent work (Bengio et al., 2004; Brand, 2003) on out-of-sample extensions, our method is based on a Representer theorem for RKHS.

Remark 2: By taking multiple eigenvectors of the system in Eqn. 54 we obtain a natural regularized out-of-sample extension of Laplacian Eigenmaps. This leads to new method for dimensionality reduction and data representation. Further study of this approach is a direction of future research. We note that a similar algorithm has been independently proposed in Vert and Yamanishi (2005) in the context of supervised graph inference. A relevant discussion is also presented in Ham et al. (2005) on the interpretation of several geometric dimensionality reduction techniques as kernel methods.

6.2 Fully Supervised Learning

The fully supervised case represents the other end of the spectrum of learning. Since standard supervised algorithms (SVM and RLS) are special cases of manifold regularization, our framework is also able to deal with a labeled dataset containing no unlabeled examples. Additionally, manifold regularization can augment supervised learning with intrinsic regularization, possibly in a class-dependent manner, which suggests the following algorithm :

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \frac{\gamma_I^+}{(u+l)^2} \mathbf{f}_+^T L_+ \mathbf{f}_+ + \frac{\gamma_I^-}{(u+l)^2} \mathbf{f}_-^T L_- \mathbf{f}_- \quad (55)$$

Here we introduce two intrinsic regularization parameters γ_I^+, γ_I^- and regularize separately for the two classes : $\mathbf{f}_+$, $\mathbf{f}_-$ are the vectors of evaluations of the function f, and L_+, L_- are the graph Laplacians, on positive and negative examples respectively. The solution to the above problem for RLS and SVM can be obtained by replacing $\gamma_I L$ by the block-diagonal matrix $\begin{pmatrix} \gamma_I^+ L_+ & 0 \\ 0 & \gamma_I^- L_- \end{pmatrix}$ in the manifold regularization formulas given in Section 4.

Detailed experimental study of this approach to supervised learning is left for future work.

7. Conclusions and Further Directions

We have a provided a novel framework for data-dependent geometric regularization. It is based on a new Representer theorem that provides a basis for several algorithms for unsupervised, semi-supervised and fully supervised learning. This framework brings together ideas from the theory of regularization in Reproducing Kernel Hilbert spaces, manifold learning and spectral methods.

There are several directions of future research:

1. Convergence and generalization error: The crucial issue of dependence of generalization error on the number of labeled and unlabeled examples is still very poorly understood. Some very preliminary steps in that direction have been taken in Belkin et al. (2004).

2. Model selection: Model selection involves choosing appropriate values for the extrinsic and intrinsic regularization parameters. We do not as yet have a good understanding of how to choose these parameters. More systematic procedures need to be developed.

3. Efficient algorithms: The naive implementations of our algorithms have cubic complexity in the number of labeled and unlabeled examples, which is restrictive for large scale real-world applications. Scalability issues need to be addressed.

4. Additional structure: In this paper we have shown how to incorporate the geometric structure of the marginal distribution into the regularization framework. We believe that this framework will extend to other structures that may constrain the learning task and bring about effective learnability. One important example of such structure is invariance under certain classes of natural transformations, such as invariance under lighting conditions in vision. Some ideas are presented in Sindhwani (2004).

Acknowledgments

We are grateful to Marc Coram, Steve Smale and Peter Bickel for intellectual support and to NSF funding for financial support. We would like to acknowledge the Toyota Technological Institute for its support for this work. We also thank the anonymous reviewers for helping to improve the paper.

References

R.A. Adams. *Sobolev spaces*. Academic Press New York, 1975.

N. Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

M. Belkin. *Problems of Learning on Manifolds*. PhD thesis, The University of Chicago, 2003.

M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. *COLT*, 2004.

M. Belkin and P. Niyogi. Using manifold structure for partially labeled classification. *Advances in Neural Information Processing Systems*, 15:929–936, 2003a.

M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. *Proc. of COLT*, 2005.

M. Belkin, P. Niyogi, and V. Sindhwani. On Manifold Regularization. *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTAT 2005)*, 2005.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003b.

Y. Bengio, J.F. Paiement, P. Vincent, and O. Delalleau. Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. *Advances in Neural Information Processing Systems*, 16, 2004.

K. Bennett and A. Demiriz. Semi-supervised support vector machines. *Advances in Neural Information Processing Systems*, 11:368–374, 1999.

A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. *Proc. 18th International Conf. on Machine Learning*, pages 19–26, 2001.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

O. Bousquet, O. Chapelle, and M. Hein. Measure based regularization. *Advances in Neural Information Processing Systems*, 16, 2004.

M. Brand. Continuous nonlinear dimensionality reduction by kernel eigenmaps. *Int. Joint Conf. Artif. Intel*, 2003.

O. Chapelle, B. Schölkopf, and A. Zien. *Semi-supervised Learning*. MIT Press, 2006.

O. Chapelle, J. Weston, and B. Scholkopf. Cluster kernels for semi-supervised learning. *Advances in Neural Information Processing Systems*, 15:585–592, 2003.

O. Chapelle and A. Zien. Semi-supervised classification by low density separation. *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 57–64, 2005.

F.R.K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

R.R. Coifman and S. Lafon. Diffusion maps. *Preprint, Jan*, 2005.

A. Corduneanu and T. Jaakkola. On information regularization. *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence*, 2003.

F. Cucker and S. Smale. On the mathematical foundations of learning. *AMERICAN MATHEMATICAL SOCIETY*, 39(1):1–49, 2002.

O. Delalleau, Y. Bengio, and N. Le Roux. Efficient non-parametric function induction in semi-supervised learning. *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTAT 2005)*, 2005.

M.P. Do Carmo. *Riemannian Geometry*. Birkhauser, 1992.

D.L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. *Proceedings of the Seventh International Conference on Information and Knowledge Management*, 11:16, 1998.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization Networks and Support Vector Machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.

G. Fung and O.L. Mangasarian. Semi-supervised support vector machines for unlabeled data classification. *Optimization Methods and Software*, 15(1):99–05, 2001.

A. Grigor'yan. Heat kernels on weighted manifolds and applications. *Cont. Math, 398, 93-191*, 2006.

J. Ham, D.D. Lee, and L.K. Saul. Semisupervised alignment of manifolds. *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence, Z. Ghahramani and R. Cowell, Eds*, 10:120–127, 2005.

M. Hein, J.Y. Audibert, and U. von Luxburg. From graphs to manifolds-weak and strong pointwise consistency of graph Laplacians. *Proceedings of the 18th Conference on Learning Theory (COLT)*, pages 470–485, 2005.

T. Joachims. Transductive inference for text classification using support vector machines. *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 200–209, 1999.

T. Joachims. Transductive learning via spectral graph partitioning. *Proceedings of the International Conference on Machine Learning*, pages 290–297, 2003.

C. Kemp, T.L. Griffiths, S. Stromsten, and J.B. Tenenbaum. Semi-supervised learning with trees. *Advances in Neural Information Processing Systems*, 16, 2004.

R.I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. *Proc. 19th International Conf. on Machine Learning*, 2002.

S. Lafon. *Diffusion Maps and Geometric Harmonics*. PhD thesis, Yale University, 2004.

K. Nigam, A.K. Mccallum, S. Thrun, and T. Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2):103–134, 2000.

R.M. Rifkin. *Everything Old Is New Again: A Fresh Look at Historical Approaches in Machine Learning*. PhD thesis, Massachusets Institute of Technology, 2002.

S.T. Roweis and L.K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323, 2000.

B. Scholkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. *Proceedings, First International Conference on Knowledge Discovery & Data Mining, Menlo Park*, 1995.

B. Scholkopf and A.J. Smola. *Learning with Kernels*. MIT Press Cambridge, Mass, 2002.

M. Seeger. Learning with labeled and unlabeled data. *Inst. for Adaptive and Neural Computation, technical report, Univ. of Edinburgh*, 2001.

V. Sindhwani. Kernel Machines for Semi-supervised Learning. Master's thesis, The University of Chicago, 2004.

V. Sindhwani, M. Belkin, and P. Niyogi. The geometric basis of semi-supervised learning. *Book Chapter in Semi-supervised Learning, Eds. Chapelle, O. and Schölkopf, B. and Zien, A.*, 2006.

V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. *ICML05, 22nd International Conference on Machine Learning*, 2005.

A. Smola and R. Kondor. Kernels and regularization on graphs. *Conference on Learning Theory, COLT/KW*, 2003.

M. Szummer and T. Jaakkola. Partially labeled classification with Markov random walks. *Advances in Neural Information Processing Systems*, 14:945–952, 2002.

J.B. Tenenbaum, V. Silva, and J.C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319, 2000.

A.N. Tikhonov. Regularization of incorrectly posed problems. *Sov. Math. Dokl*, 4:1624–1627, 1963.

I. W. Tsang and J. T. Kwok. Very large scale manifold regularization using core vector machines. *NIPS 2005 Workshop on Large Scale Kernel Machines*, 2005.

V.N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

J.P. Vert and Y. Yamanishi. Supervised graph inference. *Advances in Neural Information Processing Systems*, 17:1433–1440, 2005.

U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Max Planck Institute for Biological Cybernetics Technical Report TR*, 134, 2004.

G. Wahba. *Spline models for observational data*. Society for Industrial and Applied Mathematics Philadelphia, Pa, 1990.

D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 16:321–328, 2004.

X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. *Proceedings of the 20th International Conference on Machine Learning*, 2003.

X. Zhu, J. Kandola, Z. Ghahramani, and J. Lafferty. Nonparametric Transforms of Graph Kernels for Semi-Supervised Learning. *Advances in neural information processing systems*, 17, 2005.