Multi-Manifold Semi-Supervised Learning

Andrew B. Goldberg[†] **Xiaojin Zhu**[†] **Aarti Singh**[‡] **Zhiting Xu**[†] **Robert Nowak**[*]

[†]Computer Sciences Dept.
University of Wisconsin-Madison
Madison, WI 53706, USA
{goldberg,jerryzhu,zhiting}
@cs.wisc.edu

[‡]Applied and Computational Math
Princeton University
Princeton, NJ 08544, USA
asingh@princeton.edu

[*]Elec. and Computer Engineering
University of Wisconsin-Madison
Madison, WI 53706, USA
nowak@ece.wisc.edu

Abstract

We study semi-supervised learning when the data consists of multiple intersecting manifolds. We give a finite sample analysis to quantify the potential gain of using unlabeled data in this multi-manifold setting. We then propose a semi-supervised learning algorithm that separates different manifolds into decision sets, and performs supervised learning within each set. Our algorithm involves a novel application of Hellinger distance and size-constrained spectral clustering. Experiments demonstrate the benefit of our multi-manifold semi-supervised learning approach.

1 INTRODUCTION

The promising empirical success of semi-supervised learning algorithms in favorable situations has triggered several recent attempts (Balcan & Blum 2005, Ben-David, Lu & Pal 2008, Kaariainen 2005, Lafferty & Wasserman 2007, Niyogi 2008, Rigollet 2007) at developing a theoretical understanding of semi-supervised learning. In a recent paper (Singh, Nowak & Zhu 2008), it was established using a finite sample analysis that if the complexity of the distributions under consideration is too high to be learnt using n labeled data points, but is small enough to be learnt using $m \gg n$ unlabeled data points, then semi-supervised learning (SSL) can improve the performance of a supervised learning (SL) task. There have also been many successful practical SSL algorithms as summarized in (Chapelle, Zien & Schölkopf 2006, Zhu 2005). These theoretical analyses and practical algorithms often assume that the data forms clusters or resides on a single manifold.

However, both a theory and an algorithm are lacking when the data is supported on a mixture of manifolds. Such data occurs naturally in practice. For instance, in handwritten digit recognition each digit forms its own manifold in the feature space; in computer vision motion segmentation, moving objects trace different trajectories which are low dimensional manifolds (Tron & Vidal 2007). These manifolds may intersect or partially overlap, while having different dimensionality, orientation, and density. Existing SSL approaches cannot be directly applied to multi-manifold data. For instance, traditional graph-based SSL algorithms may create a graph that connects points on different manifolds near a manifold intersection, thus diffusing information across the wrong manifolds.

In this paper, we generalize the theoretical analysis of (Singh et al. 2008) to the case where the data is supported on a mixture of manifolds. Guided by the theory, we propose an SSL algorithm that handles multiple manifolds as well as clusters. The algorithm builds upon novel Hellinger-distance-based graphs and size-constrained manifold clustering. Experiments show that our algorithm can perform SSL on multiple intersecting, overlapping, and noisy manifolds.

2 THEORETIC PERSPECTIVES

In this section, we first review the conclusions of (Singh et al. 2008), which are based on the cluster assumption, and then give conjectured bounds in the single manifold and multi-manifold case.

The cluster assumption, as formulated in (Singh et al. 2008), states that the target regression function or class label is locally smooth over certain subsets of the D-dimensional feature space that are delineated by changes in the marginal density—throughout this paper, we assume the marginal density is bounded above

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

and below (away from zero). We refer to these delineated subsets as *decision sets*; i.e., all non-empty sets formed by intersections between the cluster support sets and their complements. If these decision sets, denoted by C, can be learnt using unlabeled data, the learning task on each decision set is simplified. The results of (Singh et al. 2008) suggest that if the decision sets can be resolved using unlabeled data, but not using labeled data, then semi-supervised learning can help. However, this simple argument, and hence the distinctions between SSL and SL, are not always captured by standard asymptotic arguments based on rates of convergence. (Singh et al. 2008) used finite sample bounds to characterize both the SSL gains and the relative value of unlabeled data.

To derive the finite sample bounds, the first step is to understand when the decision sets are resolvable using data. This depends on the interplay between the complexity of the class of distributions under consideration and the number of unlabeled points m and labeled points n. For the cluster case, the complexity of the distributions is determined by the margin γ, defined as the minimum separation between clusters or the minimum width of a decision set (Singh et al. 2008). If the margin γ is larger than the typical distance between the data points ($m^{-1/D}$ if using unlabeled data, or $n^{-1/D}$ if using only labeled data), then with high probability the decision sets can be learnt up to a high accuracy (which depends on m or n, respectively) (Singh et al. 2008). This implies that if $\gamma > m^{-1/D}$ (margin exists with respect to density of *unlabeled data*), then the finite sample performance (the expected excess error $\mathcal{E}rr$) of a semi-supervised learner $\widehat{f}_{m,n}$ relative to the performance of a clairvoyant supervised learner $\widehat{f}_{C,n}$, which has perfect knowledge of the decision sets C, can be characterized as follows:

$$\sup_{\mathcal{P}_{XY}(\gamma)} \mathcal{E}rr(\widehat{f}_{m,n}) \leq \sup_{\mathcal{P}_{XY}(\gamma)} \mathcal{E}rr(\widehat{f}_{C,n}) + \delta(m,n). \quad (1)$$

Here $\mathcal{P}_{XY}(\gamma)$ denotes the cluster-based class of distributions with complexity γ, and $\delta(m,n)$ is the error incurred due to inaccuracies in learning the decision sets using unlabeled data. Comparing this upper bound on the semi-supervised learning performance to a finite sample minimax lower bound on the performance of any supervised learner provides a sense of the relative performance of SL vs. SSL. Thus, SSL helps if complexity of the class of distributions $\gamma > m^{-1/D}$ and *both* of the following conditions hold: **(i)** knowledge of decision sets simplifies the supervised learning task, that is, the error of the clairvoyant learner $\sup_{\mathcal{P}_{XY}(\gamma)} \mathcal{E}rr(\widehat{f}_{C,n}) < \inf_{f_n} \sup_{\mathcal{P}_{XY}(\gamma)} \mathcal{E}rr(f_n)$, the smallest error that can be achieved by any supervised learner based on n labeled data. The difference quantifies the SSL performance gain. **(ii)** m is large enough so that the error incurred due to using a finite amount of unlabeled data to learn the decision sets is negligible: $\delta(m,n) = O\left(\sup_{\mathcal{P}_{XY}(\gamma)} \mathcal{E}rr(\widehat{f}_{C,n})\right)$. This quantifies the relative value of labeled and unlabeled data.

Table 1: Conjectured finite sample performance of SSL and SL for regression of a Hölder-α, $\alpha > 1$, smooth function (with respect to geodesic distance in the manifold cases). These bounds hold for $D \geq 2$, $d < D$, $m \gg n$, and suppress constants and log factors.

Complexity range	SSL upper bound	SL lower bound	SSL helps
Cluster Assumption			
$\gamma \geq n^{-\frac{1}{D}}$	$n^{-\frac{2\alpha}{2\alpha+D}}$	$n^{-\frac{2\alpha}{2\alpha+D}}$	No
$n^{-\frac{1}{D}} > \gamma \geq m^{-\frac{1}{D}}$	$n^{-\frac{2\alpha}{2\alpha+D}}$	$n^{-\frac{1}{D}}$	Yes
$m^{-\frac{1}{D}} > \gamma \geq -m^{-\frac{1}{D}}$	$n^{-\frac{1}{D}}$	$n^{-\frac{1}{D}}$	No
$-m^{-\frac{1}{D}} > \gamma$	$n^{-\frac{2\alpha}{2\alpha+D}}$	$n^{-\frac{1}{D}}$	Yes
Single Manifold $\kappa_{\mathrm{SM}} := \min(r_0, s_0)$			
$\kappa_{\mathrm{SM}} \geq n^{-\frac{1}{D}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	No
$n^{-\frac{1}{D}} > \kappa_{\mathrm{SM}} \geq m^{-\frac{1}{D}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$\Omega(1)$	Yes
$m^{-\frac{1}{D}} > \kappa_{\mathrm{SM}} \geq 0$	$O(1)$	$\Omega(1)$	No
Multi-Manifold $\kappa_{\mathrm{MM}} := \mathrm{sgn}(\gamma) \cdot \min(\lvert\gamma\rvert, r_0, s_0)$			
$\kappa_{\mathrm{MM}} \geq n^{-\frac{1}{D}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	No
$n^{-\frac{1}{D}} > \kappa_{\mathrm{MM}} \geq m^{-\frac{1}{D}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$\Omega(1)$	Yes
$m^{-\frac{1}{D}} > \kappa_{\mathrm{MM}} \geq -m^{-\frac{1}{D}}$	$O(1)$	$\Omega(1)$	No
$-m^{-\frac{1}{D}} > \kappa_{\mathrm{MM}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$\Omega(1)$	Yes

The finite sample performance bounds on SSL and SL performance as derived in (Singh et al. 2008) for the cluster assumption are summarized in Table 1 for the regression setting, where the target function is a Hölder-α smooth function on each decision set and $\alpha > 1$. We can see that SSL provides improved performance, by capitalizing on the local smoothness of the function on each decision set, when the separation between the clusters is large compared to the typical distance between unlabeled data $m^{-1/D}$ but less than the typical distance between labeled data $n^{-1/D}$. Negative γ refers to the case where the clusters are not separated, but can overlap and give rise to decision sets that are adjacent (see (Singh et al. 2008)). In this case, SSL always outperforms SL provided the width of the resulting decision sets is detectable using unlabeled data. Thus, the interplay between the margin and the number of labeled and unlabeled data characterizes the relative performance of SL vs. SSL under the cluster assumption. Similar results can be derived in the classification setting where an exponential improvement (from $n^{-1/D}$ to e^{-n}) is possible provided the number of unlabeled data m grows exponentially with n (Singh et al. 2008).

2.1 SINGLE MANIFOLD CASE

In the single manifold case, the assumption is that the target function lies on a lower d-dimensional manifold, where $d < D$, and is Hölder-α smooth ($\alpha > 1$) with respect to the geodesic distance on the manifold. Hence knowledge of the manifold, or equivalently the geodesic distances between all pairs of data points, can be gleaned using unlabeled data and reduces the dimensionality of the learning task.

In the case of distributions supported on a single manifold, the ability to learn the geodesic distances well, and hence the complexity κ_{SM} of the distributions, depends on two geometric properties of the manifold—its minimum radius of curvature r_0 and proximity to self-intersection s_0 (also known as branch separation) (Bernstein, de Silva, Langford & Tenenbaum 2000). If $\kappa_{\mathrm{SM}} := \min(r_0, s_0)$ is larger than the typical distance between the data points ($m^{-1/D}$ with unlabeled data, or $n^{-1/D}$ with only labeled data), then with high probability the manifold structure is resolvable and geodesic distances can be learnt up to a high accuracy (which depends on m or n, respectively). This can be achieved by using shortest distance paths on an ϵ- or k-nearest neighbor graph to approximate the geodesic distances (Bernstein et al. 2000). The use of approximate geodesic distances to learn the target function gives rise to an error-in-variable problem. Though the overall learning problem is now reduced to a lower-dimensional problem, we are now faced with two types of errors—the label noise and the error in the estimated distances. However, the error incurred in the final estimation due to errors in geodesic distances depends on m which is assumed to be much greater than n. Thus, the effect of the geodesic distance errors is negligible, compared to the error due to label noise, for m sufficiently large. This suggests that for the manifold case, if $\kappa_{\mathrm{SM}} > m^{-1/D}$, then finite sample performance of semi-supervised learning can again be related to the performance of a clairvoyant supervised learner $\widehat{f}_{C,n}$ as in (1) above, since $\delta(m, n)$ is negligible for m sufficiently large.

Comparing this SSL performance bound to a finite sample minimax lower bound on the performance of any supervised learner indicates SSL's gain in the single manifold case and is summarized in Table 1. These are conjectured bounds based on the arguments above and similar arguments in (Niyogi 2008). The SSL upper bound can be achieved using a learning procedure adaptive to both α and d, such as the method proposed in (Bickel & Li 2007)[1]. The SL lower bounds follow from the results in (Tsybakov 2004) and (Niyogi 2008). SSL provides improved performance by capitalizing on the lower-dimensional structure of the manifold when the minimum radius of curvature and branch separation are large compared to the typical distance between unlabeled data $m^{-1/D}$, but small compared to the typical distance between labeled data $n^{-1/D}$.

2.2 MULTI-MANIFOLD CASE

The multi-manifold case addresses the generic setting where the clusters are low-dimensional manifolds that possibly intersect or overlap. In this case, the target function is supported on multiple manifolds and can be piecewise smooth on each manifold. Thus, it is of interest to resolve the manifolds, as well as the subsets of each manifold where the decision label varies smoothly (that are characterized by changes in the marginal density). The analysis for this case is a combination of the cluster and single manifold case. The complexity of the multi-manifold class of distributions, denoted κ_{MM}, is governed by the minimum of the manifold curvatures, branch separations, and the separations and overlaps between distinct manifolds. For the regression setting, the conjectured finite sample minimax analysis is presented in Table 1.

These results indicate that when there is enough unlabeled data, but not enough labeled data, to handle the complexity of the class, then semi-supervised learning can help by adapting to both the intrinsic dimensionality and smoothness of the target function. Extensions of these results to the classification setting are straightforward, as discussed under the cluster assumption.

3 AN ALGORITHM

Guided by the theoretical analysis in the previous section, we propose a "cluster-then-label" type of SSL algorithm, see Figure 1. It consists of three main steps: (1) It uses the unlabeled data to form a small number of *decision sets*, on which the target function is assumed to be smooth. The decision sets are defined in the ambient space, not just on the labeled and unlabeled points. (2) The target function within a particular decision set is estimated using only labeled data that fall in that decision set, and using a supervised learner specified by the user. (3) a new test point is predicted by the target function in the decision set it falls into.

There have been several cluster-then-label approaches in the SSL literature. For example, the early work of Demiriz et al. modifies the objective of standard

[1] Note, however, that the analysis in (Bickel & Li 2007) considers the asymptotic performance of SL, whereas here we are studying the finite-sample performance of SSL.

k-means clustering algorithms to include a class impurity term (Demiriz, Bennett & Embrechts 1999). El-Yaniv and Gerzon enumerate all spectral clusterings of the unlabeled data with varying number of clusters, which together with labeled data induce a hypothesis space. They then select the best hypothesis based on an Occam's razor-type transductive bound (El-Yaniv & Gerzon 2005). Some work in "constrained clustering" is also closely related to cluster-then-label from an SSL perspective (Basu, Davidson & Wagstaff 2008). Compared to these approaches, our algorithm has two advantages: i) it is supported by our SSL minimax theory; ii) it handles both overlapping clusters and intersecting manifolds by detecting changes in support, density, dimensionality or orientation.

Our algorithm is also different from the family of graph-regularized SSL approaches, such as manifold regularization (Belkin, Sindhwani & Niyogi 2006) and earlier variants (Joachims 2003, Zhou, Bousquet, Lal, Weston & Schölkopf 2004, Zhu, Ghahramani & Lafferty 2003). Those approaches essentially add a graph-regularization term in the objective. They also depend on the "manifold assumption" that the target function *indeed* varies smoothly on the manifold. In contrast, i) our algorithm is a *wrapper* method, which uses any user-specified supervised learner SL as a subroutine. This allows us to directly take advantage of advances in supervised learning without the need to derive new algorithms. ii) Our theory ensures that, even when the manifold assumption is wrong, our SSL performance bound is the same as that of the supervised learner (up to a log factor).

Finally, step 1 of our algorithm is an instance of manifold clustering. Recent advances on this topic include Generalized Principal Component Analysis (Vidal, Ma & Sastry 2008) and lossy coding (Ma, Derksen, Hong & Wright 2007) for mixtures of linear subspaces, multiscale manifold identification with algebraic multigrid (Kushnir, Galun & Brandt 2006), tensor voting (Mordohai & Medioni 2005), spectral curvature clustering (Chen & Lerman 2008), and translated Poisson mixture model (Haro, Randall & Sapiro 2008) for mixtures of nonlinear manifolds. Our algorithm is unique in two ways: i) its use of Hellinger distance offers a new approach to detecting overlapping clusters and intersecting manifolds; ii) our decision sets have minimum size constraints, which we enforce by constrained k-means.

3.1 HELLINGER DISTANCE GRAPH

Let the labeled data be $\{(x_i, y_i)\}_{i=1}^n$, and the unlabeled data be $\{x_j\}_{j=1}^M$, where $M \gg n$. The building block of our algorithm is a *local sample covariance matrix*. For a point x, define $N(x)$ to be a small neighborhood around x in Euclidean space. Let Σ_x be the local sample covariance matrix at x: $\Sigma_x = \sum_{x' \in N(x)} (x' - \mu_x)(x' - \mu_x)^\top / (|N(x)| - 1)$, where $\mu_x = \sum_{x' \in N(x)} x' / |N(x)|$ is the neighborhood mean. In our experiments, we let $|N(x)| \sim O(\log(M))$ so that the neighborhood size grows with unlabeled data size M. The covariance Σ_x captures the local geometry around x.

Our intuition is that points x_i, x_j on different manifolds or in regions with different density should be considered dissimilar. This intuition is captured by the Hellinger distance between their local sample covariance matrices Σ_i, Σ_j. The squared Hellinger distance is defined between two pdf's p, q: $H^2(p, q) = \frac{1}{2} \int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 \mathrm{d}x$. By setting $p(x) = \mathcal{N}(x; 0, \Sigma_i)$, i.e., a Gaussian with zero mean and covariance Σ_i, and similarly $q(x) = \mathcal{N}(x; 0, \Sigma_j)$, we extend the definition of Hellinger distance to covariance matrices: $H(\Sigma_i, \Sigma_j) \equiv H\left(\mathcal{N}(x; 0, \Sigma_i), \mathcal{N}(x; 0, \Sigma_i)\right) = \sqrt{1 - 2^{D/2} |\Sigma_i|^{1/4} |\Sigma_j|^{1/4} / |\Sigma_i + \Sigma_j|^{1/2}}$, where D is the dimensionality of the ambient feature space. We will also call $H(\Sigma_i, \Sigma_j)$ the Hellinger distance between the two points x_i, x_j. When $\Sigma_i + \Sigma_j$ is rank deficient, H is computed in the subspace occupied by $\Sigma_i + \Sigma_j$. The Hellinger distance H is symmetric and in $[0, 1]$. H is small when the local geometry is similar, and large when there is significant difference in density, manifold dimensionality or orientation. Example 3D covariance matrices and their H values are shown in Figure 2.

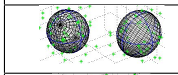
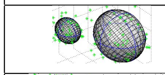
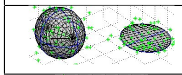
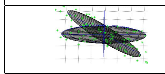
Cov. matrices	Comment	$H(\Sigma_1, \Sigma_2)$
	similar	0.02
	density	0.28
	dimension	1
	orientation	1

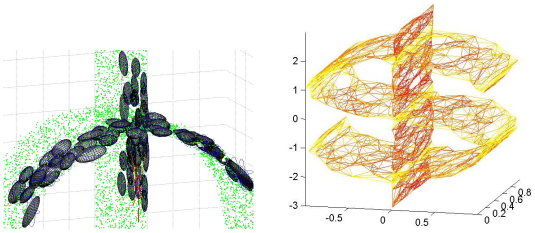
Figure 2: Hellinger distance

It would seem natural to compute all pairwise Hellinger distances between our dataset of $n + M$ points to form a graph, and apply a graph-cut algorithm to separate multiple manifolds or clusters. However, if x_i and x_j are very close to each other, their local neighborhoods $N(x_i), N(x_j)$ will strongly overlap. Then, even if the two points are on different manifolds the Hellinger distance will be small, because their covariance matrices Σ_i, Σ_j will be similar. Therefore, we select a subset of $m \sim O\left(M / \log(M)\right)$ unlabeled points so that they are farther apart while still covering the whole dataset. This is done using a greedy procedure, Given n labeled examples and M unlabeled examples, and a supervised learner SL,

1. Use the unlabeled data to infer $k \sim O(\log(n))$ decision sets $\widehat{C_i}$:
   (a) Select a subset of $m < M$ unlabeled points
   (b) Form a graph on the $n + m$ labeled and unlabeled points, where the edge weights are computed from the Hellinger distance between local sample covariance matrices
   (c) Perform size-constrained spectral clustering to cut the graph into k parts, while keeping enough labeled and unlabeled points in each part
2. Use the labeled data in $\widehat{C_i}$ and the supervised learning SL to train $\widehat{f_i}$
3. For test point $x^* \in \widehat{C_i}$, predict $\widehat{f_i}(x^*)$.

Figure 1: The Multi-Manifold Semi-Supervised Learning Algorithm

where we first select an arbitrary unlabeled point $x^{(0)}$. We then remove its unlabeled neighbors $N(x^{(0)})$, including itself. We select $x^{(1)}$ to be the next nearest neighbor, and repeat. This procedure thus approximately selects a cover of the dataset. We focus on the subset of m unlabeled and n labeled points. Each of these $n + m$ points has its local covariance Σ computed from the original full dataset. We then discard the $M - m$ unselected unlabeled points. Notice, however, that the number m of effective unlabeled data points is polynomially of the same order as the total number M of available unlabeled data points.



(a) (b)

Figure 3: The graph on the dollar sign dataset.

We can now define a sparse graph on the $n+m$ points. Each point x is connected by a weighted, undirected edge to $O(\log(n+m))$ of its nearest Mahalanobis neighbors chosen from the the set of $n + m$ points too. The choice of $O(\log(n + m))$ allows neighborhood size to grow with dataset size. Since we know the local geometry around x (captured by Σ_x), we "follow the manifold" by using the Mahalanobis distance as the local distance metric at x: $d_M^2(x, x') = (x-x')^\top \Sigma_x^{-1}(x-x')$. For example, a somewhat flat Σ_x will preferentially connect to neighbors in or near the same flat subspace. The graph edges are weighted using the standard RBF scheme, but with Hellinger distance: $w_{ij} = \exp\left(-H^2(\Sigma_i, \Sigma_j)/(2\sigma^2)\right)$. Figure 3(a) shows a small part of a synthetic "dollar sign" dataset, consisting of two intersecting manifolds: "S" and "|". The green dots are the original unlabeled points, and the ellipsoids are the contours of covariance matrices around the subset of selected unlabeled points within a small region. Figure 3(b) shows the graph on the complete dollar sign dataset, where red edges have large weights and yellow edges have small weights. Thus the graph combines locality and geometry: an edge has large weight when the two nodes are close in Mahalanobis distance, and have similar covariance structure.

3.2 SIZE-CONSTRAINED SPECTRAL CLUSTERING

We perform spectral clustering on this graph of $n + m$ nodes. We hope each resulting cluster represents a separate manifold, from which we will define a decision set. Of the many spectral clustering algorithms, we chose ratio cut for its simplicity, though others can be similarly adapted for use here. The standard ratio cut algorithm for k clusters has four steps (von Luxburg 2007): 1. Compute the unnormalized graph Laplacian $L = Deg - W$, where $W = [w_{ij}]$ is the weight matrix, and $Deg_{ii} = \sum_j w_{ij}$ form the diagonal degree matrix. 2. Compute the k eigenvectors $v_1 \dots v_k$ of L with the smallest eigenvalues. 3. Form matrix V with $v_1 \dots v_k$ as columns. Use the ith row of V as the new representation of x_i. 4. Cluster all x under the new representation into k clusters using k-means.

Our ultimate goal of semi-supervised learning poses new challenges; we want our SSL algorithm to degrade gracefully, even when the manifold assumption does not hold. The SSL algorithm should not break the problem into too many subproblems and increase the complexity of the supervised learning task. This is achieved by requiring that the algorithm does not generate too many clusters and that each cluster contains "enough" labeled points. Because we will simply do supervised learning within each decision set, as long as the number of sets does not grow polynomially with n, the performance of our algorithm is guaranteed to be polynomially no worse than the performance of the supervised learner when the manifold assumption fails. Thus, we automatically revert to the supervised learning performance. One way to achieve this is to have three requirements: i) the number of clusters grows as $k \sim O(\log(n))$; ii) each cluster must have at least $a \sim O(n/\log^2(n))$ labeled points; iii) each spectral cluster must have at least $b \sim O(m/\log^2(n))$ unlabeled points. The first sets the number of clusters k, allowing more clusters and thus handling more complex problems as labeled data size grows, while suffering only a logarithmic performance loss compared to a supervised learner if the manifold assumption fails. The second requirement ensures that each decision set has $O(n)$ labeled points up to log factor[2]. The third is similar, and makes spectral clustering more robust.

Spectral clustering with minimum size constraints a, b on each cluster is an open problem. Directly enforcing these constraints in graph partitioning leads to difficult integer programs. Instead, we enforce the constraints in k-means (step 4) of spectral clustering. Our approach is a straightforward extension to the constrained k-means algorithm of Bradley et al. (Bradley, Bennett & Demiriz 2000). For point x_i, let $T_{i1} \dots T_{ik} \in \mathbb{R}$ be its cluster indicators: ideally, $T_{ih} = 1$ if x_i is in cluster h, and 0 otherwise. Let $c_1 \dots c_k \in \mathbb{R}^d$ denote the cluster centers. Constrained k-means is the iterative minimization over T and c of the following problem:

$$\begin{aligned}\min_{T,c} \quad & \textstyle\sum_{i=1}^{n+m} \sum_{h=1}^{k} T_{ih} \|x_i - c_h\|^2 \\ \text{s.t.} \quad & \textstyle\sum_{h=1}^{k} T_{ih} = 1,\ T \geq 0 \\ & \textstyle\sum_{i=1}^{n} T_{ih} \geq a,\ \sum_{i=n+1}^{n+m} T_{ih} \geq b,\ h = 1 \dots k, \end{aligned} \quad (2)$$

where we assume the points are ordered so that the first n points are labeled. Fixing T, optimizing over c is trivial, and amounts to moving the centers to the cluster means.

Bradley et al. showed that fixing c and optimizing T can be converted into a Minimum Cost Flow problem, which can be exactly solved. In a Minimum Cost Flow problem, there is a directed graph where each node is either a "supply node" with a number $r > 0$, or a "demand node" with $r < 0$. The arcs from $i \to j$ is associated with cost s_{ij}, and flow t_{ij}. The goal is to find the flow t such that supply meets demand at all nodes, while the cost is minimized:

$$\min_t \sum_{i \to j} s_{ij} t_{ij} \quad \text{s.t.} \sum_j t_{ij} - \sum_j t_{ji} = r_i,\ \forall i. \qquad (3)$$

For our problem (2), the corresponding Minimum Cost Flow problem is shown in Figure 4. The supply nodes are $x_1 \dots x_{n+m}$ with $r = 1$. There are two sets of cluster center nodes. One set $c_1^\ell \dots c_k^\ell$, each with demand $r = -a$, is due to the labeled data size constraint. The other set $c_1^u \dots c_k^u$, each with demand $r = -b$, is due to the unlabeled data size constraint. Finally, a sink demand node with $r = -(n + m - ak - bk)$ catches all the remaining flow. The cost from x_i to c_h is $s_{ih} = \|x_i - c_h\|^2$, and from c_h to the sink is 0. It is then clear that the Minimum Cost Flow problem (3) is equivalent to (2) with $T_{ih} = t_{ih}$ and c fixed. Interestingly, (3) is proven to have integer solutions which correspond exactly to the desired cluster indicators.

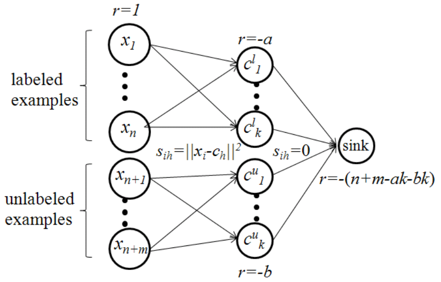


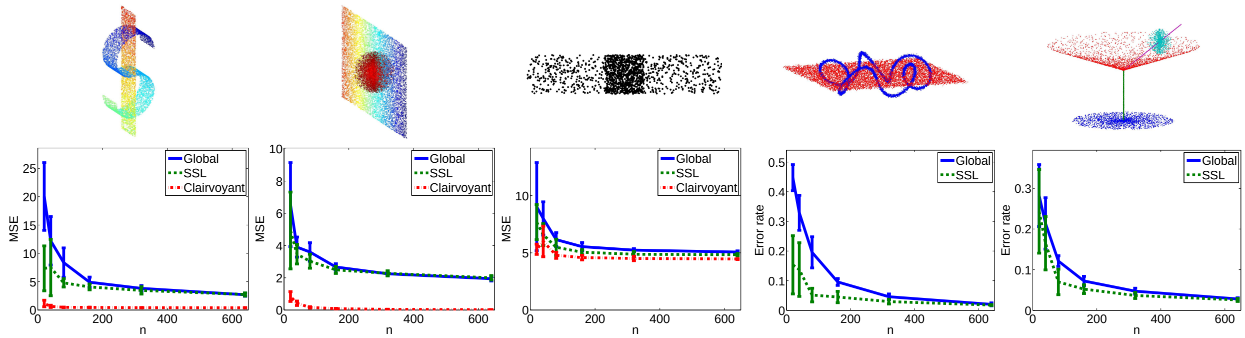
Figure 4: The Minimum Cost Flow problem

Once size-constrained spectral clustering is completed, the $n+m$ points will each have a cluster index in $1 \dots k$. We define k decision sets $\{\widehat{C_i}\}_{i=1}^k$ by the Voronoi cells around these points: $\widehat{C_i} = \{x \in \mathbb{R}^D \mid x$'s Euclidean nearest neighbor among the $n + m$ points has cluster index $i\}$. We train a separate predictor $\widehat{f_i}$ for each decision set using the labeled points in that decision set, and a user-specified supervised learner. During test time, an unseen point $x^* \in \widehat{C_i}$ is predicted as $\widehat{f_i}(x^*)$. Therefore, the unlabeled data in our algorithm is used merely to determine the decision sets.

4 EXPERIMENTS

Data Sets. We experimented with five synthetic (Figure 5) and one real data sets. Data sets 1–3 are for regression, and 4–6 are for classification: **(1). Dollar sign** contains two intersecting manifolds. The "S" manifold has target y varying from 0 to 3π. The "|" manifold has target function $y = x_{\cdot 3} + 13$, where $x_{\cdot 3}$ is the vertical dimension. White noise $\epsilon \sim \mathcal{N}(0, 0.01^2)$ is added to y. **(2) Surface-sphere** slices a 2D surface through a solid ball. The ball has target function $y = ||x||$, and the surface has $y = x_{\cdot 2} - 5$. **(3) Density change** contains two overlapping rectangles. One rectangle is wide and sparse with $y = x_{\cdot 1}$, the other is narrow and five times as dense with $y = 10 - 5x_{\cdot 1}$. Together they produce three decision sets. **(4) Surface-helix** has a 1D toroidal helix intersecting a surface. Each manifold is a separate class. **(5) Martini** is a composition of five manifolds (classes) to form the shape of a martini glass with an olive on a toothpick, as shown in Figure 5(e). **(6) MNIST** digits. We scaled

[2] The square allows the size ratio between two clusters to be arbitrarily skewed as n grows. We do not want to fix the relative sizes of the decision sets *a priori*.



(a) Dollar sign (b) Surface-sphere (c) Density change (d) Surface-helix (e) Martini

Figure 5: Regression MSE (a-c) and classification error (d-e) for synthetic data sets. All curves are based on $M = 20000$, 10-trial averages, and error bars plot ±1 standard deviation. Clairvoyant classification error is 0.

down the images to 16 x 16 pixels and used the official train/test split, with different numbers of labeled and unlabeled examples sampled from the training set.

Methodology & Implementation Details. In all experiments, we report results that are the average of 10 trials over random draws of M unlabeled and n labeled points. We compare three learners: [**Global**]: supervised learner trained on all of the labeled data, ignoring unlabeled data. [**Clairvoyant**]: with the knowledge of the true decision sets, trains one supervised learner per decision set. [**SSL**]: our semi-supervised learner that discovers the decision sets using unlabeled data, then trains one supervised learner per decision set. After training, each classifier is evaluated on a massive test set, also sampled from the underlying distribution, to estimate generalization error. We implemented the algorithms in MATLAB, with Minimum Cost Flow solved by the network simplex method in CPLEX. We used the same set of parameters for all experiments and all data sets: We chose the number of decision sets to be $k = \lceil 0.5\ log(n) \rceil$. To obtain the subset of m unlabeled points, we let the neighborhood size $|N(x)| = \lfloor 3\ log(M) \rfloor$. When creating the graph W, we used $\lfloor 1.5\ log(m+n) \rfloor$ nearest Mahalanobis neighbors, and an RBF bandwidth $\sigma = 0.2$ to convert Hellinger distances to edge weights. The size constraints were $a = \lfloor 1.25n/log^2(n) \rfloor, b = \lfloor 1.25m/log^2(n) \rfloor$. Finally, to avoid poor local optima in spectral clustering, we ran 10 random restarts for constrained k-means, and chose the result with the lowest objective. For the regression tasks, we used kernel regression with an RBF kernel, and tuned the bandwidth parameter with 5-fold cross validation using only labeled data in each decision set (or globally for "Global"). For classification, we used a support vector machine (LIBSVM) with an RBF kernel, and tuned its bandwidth and regularization parameter with 5-fold cross validation. We used Euclidean distance in each decision region for the supervised learner, but we expect performance with geodesic distance would be even better.

Results of Large M: Figure 5 reports the results for the five synthetic data sets. In all cases, we used $M = 20000$, $n \in \{20, 40, 80, 160, 320, 640\}$, and the resulting regressors/classifiers are evaluated in terms of MSE or error rate using a test set of 20000 points. These results show that our SSL algorithm can discover multiple manifolds and changes in density well enough to consistently outperform SL in both regression and classification settings of varying complexity[3]. We also observed that even under- or over-partitioning into fewer or more decision sets than manifolds can still improve SSL performance[4].

We performed three experiments with the digit recognition data: binary classification of the digits 2 vs 3, and three-way classification of $1, 2, 3$ and $7, 8, 9$. Here, we fixed $n = 20$, $M = 5000$, 10 random training trials, each tested on the official test set. Table 2 contains results averaged over these trials. SSL outperforms Global in all three digit tasks, and all differences are statistically significant ($\alpha = 0.05$). Note that we used the same parameters as the synthetic data experiments, which results in $k = 2$ decision sets for $n = 20$; again, the algorithm performs well even when there are fewer decision sets than classes. Close inspection revealed that our clustering step creates relatively pure decision sets. For the binary task, this leads to two

[3] Though not shown in Figure 5, we found that a standard graph-based SSL algorithm manifold regularization (Belkin et al. 2006), using Euclidean kNN graphs with RBF weights and all parameters tuned using cross validation, performs worse than Global on these datasets due to the strong connections across manifolds.

[4] We compared Global and SSL's 10 trials at each n using two-tailed paired t-tests. SSL was statistically significantly better ($\alpha = 0.05$) in the following cases: dollar sign at $n = 20$–80, density at $n = 40$–640, surface-helix at $n = 20$–320, and martini at $n = 40$–320. The two methods were statistically indistinguishable in other cases.

Table 2: 10-trial average test set error rates ± one standard deviation for handwritten digit recognition.

Method	2 vs 3	1, 2, 3	7, 8, 9
Global	0.17 ± 0.12	0.20 ± 0.10	0.33 ± 0.20
SSL	0.05 ± 0.01	0.10 ± 0.04	0.20 ± 0.10

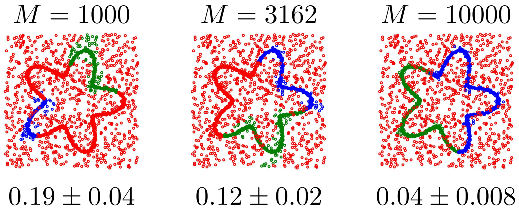


Figure 6: Effect of varying M for the surface-helix data set ($n = 80$). See text for details.

trivial classification problems, and errors are due only to incorrect assignments of test points to decision sets. For the 3-way tasks, the algorithm creates 1+2 and 3 clusters, and 7+9 and 8 clusters.

Effect of Too Small an M: Finally, we examine our SSL algorithm's performance with less unlabeled data. For the surface-helix data set, we now fix $n = 80$ (which leads to $k = 3$ decision sets) and reduce M. Figure 6 depicts example partitionings for three M values, along with 10-trial average error rates (± one standard deviation) in each setting. Note these are top-down views of the data in Figure 5(d). When M is small, the resulting subset of m unlabeled points is too small, and the partition boundaries cannot be reliably estimated. Segments of the helix shown in red and areas of the surface in blue or green correspond to such partitioning errors. Nevertheless, even when M is as small as 1000, SSL's performance is no worse than Global supervised learning, which achieves an error rate of 0.20 ± 0.05 when $n = 80$ (see Figure 5(d)).

Conclusions: We have extended SSL theory and practice to multi-manifolds. A detailed analysis of geodesic distances, automatic parameter selection, and large scale empirical study remains as future work.

Acknowledgements We thank the Wisconsin Alumni Research Foundation. AG is supported in part by a Yahoo! Key Technical Challenges Grant.

References

Balcan, M.-F. & Blum, A. (2005), A PAC-style model for learning from labeled and unlabeled data, *in* 'COLT'.

Basu, S., Davidson, I. & Wagstaff, K., eds (2008), *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, Chapman & Hall/CRC Press.

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

Ben-David, S., Lu, T. & Pal, D. (2008), Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning, *in* 'COLT'.

Bernstein, M., de Silva, V., Langford, J. & Tenenbaum, J. (2000), Graph approximations to geodesics on embedded manifolds, Technical report, Stanford.

Bickel, P. & Li, B. (2007), 'Local polynomial regression on unknown manifolds', *Complex datasets and inverse problems: Tomography, Networks and Beyond, IMS Lecture Notes-Monograph Series* **54**, 177–186.

Bradley, P., Bennett, K. & Demiriz, A. (2000), Constrained k-means clustering, Technical Report MSR-TR-2000-65, Microsoft Research.

Chapelle, O., Zien, A. & Schölkopf, B., eds (2006), *Semi-supervised learning*, MIT Press.

Chen, G. & Lerman, G. (2008), Spectral curvature clustering, *in* 'IJCV'.

Demiriz, A., Bennett, K. & Embrechts, M. (1999), 'Semi-supervised clustering using genetic algorithms', *Artificial Neural Networks in Engineering* .

El-Yaniv, R. & Gerzon, L. (2005), 'Effective transductive learning via objective model selection', *Pattern Recognition Letters* **26**(13), 2104–2115.

Haro, G., Randall, G. & Sapiro, G. (2008), 'Translated poisson mixture model for stratification learning', *IJCV* **80**, 358–374.

Joachims, T. (2003), Transductive learning via spectral graph partitioning, *in* 'ICML'.

Kaariainen, M. (2005), Generalization error bounds using unlabeled data, *in* 'COLT'.

Kushnir, D., Galun, M. & Brandt, A. (2006), 'Fast multi-scale clustering and manifold identification', *Pattern Recognition* **39**, 1876–1891.

Lafferty, J. & Wasserman, L. (2007), Statistical analysis of semi-supervised regression, *in* 'NIPS'.

Ma, Y., Derksen, H., Hong, W. & Wright, J. (2007), 'Segmentation of multivariate mixed data via lossy coding and compression', *PAMI* **29**(9), 1546–1562.

Mordohai, P. & Medioni, G. (2005), Unsupervised dimensionality estimation and manifold learning in high-dimensional spaces by tensor voting, *in* 'IJCAI'.

Niyogi, P. (2008), Manifold regularization and semi-supervised learning: Some theoretical analyses, Technical Report TR-2008-01, CS Dept, U. of Chicago.

Rigollet, P. (2007), 'Generalization error bounds in semi-supervised classification under the cluster assumption', *JMLR* **8**, 1369–1392.

Singh, A., Nowak, R. & Zhu, X. (2008), Unlabeled data: Now it helps, now it doesn't, *in* 'NIPS'.

Tron, R. & Vidal, R. (2007), A benchmark for the comparison of 3-d motion segmentation algorithms, *in* 'CVPR'.

Tsybakov, A. B. (2004), *Introduction a l'estimation non-parametrique*, Springer, Berlin Heidelberg.

Vidal, R., Ma, Y. & Sastry, S. (2008), *Generalized Principal Component Analysis (GPCA)*, Springer Verlag.

von Luxburg, U. (2007), 'A tutorial on spectral clustering', *Statistics and Computing* **17**(4), 395–416.

Zhou, D., Bousquet, O., Lal, T., Weston, J. & Schölkopf, B. (2004), Learning with local and global consistency, *in* 'NIPS'.

Zhu, X. (2005), Semi-supervised learning literature survey, Technical Report 1530, Department of Computer Sciences, University of Wisconsin, Madison.

Zhu, X., Ghahramani, Z. & Lafferty, J. (2003), Semi-supervised learning using Gaussian fields and harmonic functions, *in* 'ICML'.