Semi-Supervised Learning with Max-Margin Graph Cuts

Branislav Kveton
Intel Labs Santa Clara

Michal Valko
University of Pittsburgh

Ali Rahimi and **Ling Huang**
Intel Labs Berkeley

Abstract

This paper proposes a novel algorithm for semi-supervised learning. This algorithm learns graph cuts that maximize the margin with respect to the labels induced by the harmonic function solution. We motivate the approach, compare it to existing work, and prove a bound on its generalization error. The quality of our solutions is evaluated on a synthetic problem and three UCI ML repository datasets. In most cases, we outperform manifold regularization of support vector machines, which is a state-of-the-art approach to semi-supervised max-margin learning.

1 INTRODUCTION

Semi-supervised learning is a field of machine learning that studies learning from both labeled and unlabeled examples. This learning paradigm is suitable for real-world problems, where data is often abundant but the resources to label them are limited. As a result, many semi-supervised learning algorithms have been proposed in the past years (Zhu, 2008). The closest to this work are semi-supervised support vector machines (S3VMs) (Bennett & Demiriz, 1999), manifold regularization of support vector machines (SVMs) (Belkin et al., 2006), and harmonic function solutions on data adjacency graphs (Zhu et al., 2003). Manifold regularization of SVMs essentially combines the ideas of harmonic function solutions and semi-supervised SVMs in a single convex objective.

This paper proposes a different way of combining these two ideas. First, we compute the harmonic function solution on the data adjacency graph and then, we learn a discriminator, which is conditioned on the labels induced by this solution. We refer to our method as *max-margin graph cuts* because the discriminator maximizes the margin with respect to the inferred labels. The method has many favorable properties. For instance, it incorporates the kernel trick (Wahba, 1999), it takes advantage of sparse data adjacency matrices, and its generalization error can be bounded. Moreover, it typically yields better results than manifold regularization of SVMs, especially for linear and cubic decision boundaries.

In addition to proposing a new algorithm, this paper makes two contributions. First, we show how manifold regularization of linear and cubic SVMs fails on almost a trivial problem. Second, we show how to make the harmonic function solution with soft labeling constraints (Cortes et al., 2008) stable.

The paper is organized as follows. In Section 2, we review the harmonic function solution and discuss how to regularize it to interpolate between supervised learning on labeled examples and semi-supervised learning on all data. In Section 3, we introduce our learning algorithm. The algorithm is compared to existing work in Section 4 and we bound its generalization error in Section 5. In Section 6, we evaluate the quality of our solutions on UCI ML repository datasets, and show that they usually outperform manifold regularization of SVMs.

The following notation is used in the paper. The symbols $\mathbf{x}_i$ and y_i refer to the i-th data point and its label, respectively. The data points are divided into labeled and unlabeled sets, l and u, and labels $y_i \in \{-1, 1\}$ are observed for the labeled data only. The cardinality of the labeled and unlabeled sets is $n_l = |l|$ and $n_u = |u|$, respectively, and the total number of training examples is $n = n_l + n_u$.

2 REGULARIZED HARMONIC FUNCTION SOLUTION

In this section, we review the harmonic function solution of Zhu et al. (2003). Moreover, we show how to regularize it to interpolate between semi-supervised learning on all data and supervised learning on labeled examples.

A standard approach to semi-supervised learning on graphs is to minimize the quadratic objective function:

$$\begin{aligned} \min_{\boldsymbol{\ell} \in \mathbb{R}^n} \quad & \boldsymbol{\ell}^\mathsf{T} L \boldsymbol{\ell} \\ \text{s.t.} \quad & \ell_i = y_i \text{ for all } i \in l; \end{aligned} \qquad (1)$$

where $\boldsymbol{\ell}$ denotes the vector of predictions, $L = D - W$ is the Laplacian of the data adjacency graph, which is represented by a matrix W of pairwise similarities w_{ij}, and D is

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.

a diagonal matrix whose entries are given by $d_i = \sum_j w_{ij}$. This problem has a closed-form solution:

$$\boldsymbol{\ell}_u = (D_{uu} - W_{uu})^{-1} W_{ul} \boldsymbol{\ell}_l, \tag{2}$$

which satisfies the *harmonic property* $\ell_i = \frac{1}{d_i} \sum_{j \sim i} w_{ij} \ell_j$, and therefore is commonly known as the *harmonic function solution*. Since the solution can be also computed as:

$$\boldsymbol{\ell}_u = (I - P_{uu})^{-1} P_{ul} \boldsymbol{\ell}_l, \tag{3}$$

it can be viewed as a product of a random walk on the graph W with the transition matrix $P = D^{-1}W$. The probability of moving between two arbitrary vertices i and j is w_{ij}/d_i, and the walk terminates when the reached vertex is labeled. Each element of the solution is given by:

$$\begin{aligned} \ell_i &= (I - P_{uu})^{-1}_{iu} P_{ul} \boldsymbol{\ell}_l \\ &= \underbrace{\sum_{j : y_j = 1} (I - P_{uu})^{-1}_{iu} P_{uj}}_{p_i^1} - \underbrace{\sum_{j : y_j = -1} (I - P_{uu})^{-1}_{iu} P_{uj}}_{p_i^{-1}} \\ &= p_i^1 - p_i^{-1}, \end{aligned} \tag{4}$$

where p_i^1 and p_i^{-1} are probabilities by which the walk starting from the vertex i ends at vertices with labels 1 and -1, respectively. Therefore, when ℓ_i is rewritten as $|\ell_i| \operatorname{sgn}(\ell_i)$, $|\ell_i|$ can be interpreted as a *confidence* of assigning the label $\operatorname{sgn}(\ell_i)$ to the vertex i. The maximum value of $|\ell_i|$ is 1, and it is achieved when either $p_i^1 = 1$ or $p_i^{-1} = 1$. The closer the confidence $|\ell_i|$ to 0, the closer the probabilities p_i^1 and p_i^{-1} to 0.5, and the more *uncertain* the label $\operatorname{sgn}(\ell_i)$.

To control the confidence of labeling unlabeled examples, we suggest regularizing the Laplacian L as $L + \gamma_g I$, where γ_g is a scalar and I is the identity matrix. Similarly to our original problem (1), the corresponding harmonic function solution:

$$\begin{aligned} \min_{\boldsymbol{\ell} \in \mathbb{R}^n} \quad & \boldsymbol{\ell}^\mathsf{T} (L + \gamma_g I) \boldsymbol{\ell} \\ \text{s.t.} \quad & \ell_i = y_i \text{ for all } i \in l \end{aligned} \tag{5}$$

can be computed in a closed form:

$$\boldsymbol{\ell}_u = (L_{uu} + \gamma_g I)^{-1} W_{ul} \boldsymbol{\ell}_l. \tag{6}$$

It can be also interpreted as a random walk on the graph W with an extra sink. At each step, this walk may terminate at the sink with probability $\gamma_g/(d_i + \gamma_g)$. Therefore, the scalar γ_g essentially controls how the confidence $|\ell_i|$ of labeling unlabeled vertices decreases with the number of hops from labeled vertices.

Several examples of how γ_g affects the regularized solution are shown in Figure 1. When $\gamma_g = 0$, the solution turns into the ordinary harmonic function solution. When $\gamma_g = \infty$, the confidence of labeling unlabeled vertices decreases to zero. Finally, note that our regularization corresponds to increasing all eigenvalues of the Laplacian L by γ_g (Smola & Kondor, 2003). In Section 5, we use this property to bound the generalization error of our solutions.

3 MAX-MARGIN GRAPH CUTS

Our semi-supervised learning algorithm involves two steps. First, we obtain the regularized harmonic function solution $\boldsymbol{\ell}^*$ (Equation 6). The solution is computed from the system of linear equations $(L_{uu} + \gamma_g I)\boldsymbol{\ell}_u = W_{ul}\boldsymbol{\ell}_l$. This system of linear equations is sparse when the data adjacency graph W is sparse. Second, we learn a max-margin discriminator, which is conditioned on the labels induced by the harmonic solution. The optimization problem is given by:

$$\begin{aligned} \min_{f \in \mathcal{H}_K} \quad & \sum_{i : |\ell_i^*| \geq \varepsilon} V(f, \mathbf{x}_i, \operatorname{sgn}(\ell_i^*)) + \gamma \|f\|_K^2 \\ \text{s.t.} \quad & \boldsymbol{\ell}^* = \arg\min_{\boldsymbol{\ell} \in \mathbb{R}^n} \boldsymbol{\ell}^\mathsf{T} (L + \gamma_g I) \boldsymbol{\ell} \\ & \quad \text{s.t. } \ell_i = y_i \text{ for all } i \in l; \end{aligned} \tag{7}$$

where $V(f, \mathbf{x}, y) = \max\{1 - yf(\mathbf{x}), 0\}$ denotes the *hinge loss*, f is a function from some *reproducing kernel Hilbert space (RKHS)* $\mathcal{H}_K$, and $\|\cdot\|_K$ is the norm that measures the complexity of f.

Training examples $\mathbf{x}_i$ in our problem are selected based on our confidence into their labels. When the labels are highly *uncertain*, which means that $|\ell_i^*| < \varepsilon$ for some small $\varepsilon \geq 0$, the examples are excluded from learning. Note that as the regularizer γ_g increases, the values $|\ell_i^*|$ decrease towards 0 (Figure 1), and the ε thresholding allows for smooth interpolations between supervised learning on labeled examples and semi-supervised learning on all data. The tradeoff between the regularization of f and the minimization of hinge losses $V(f, \mathbf{x}_i, \operatorname{sgn}(\ell_i^*))$ is controlled by the parameter γ.

Due to the representer theorem (Wahba, 1999), the optimal solution f^* to our problem has a special form:

$$f^*(\mathbf{x}) = \sum_{i : |\ell_i^*| \geq \varepsilon} \alpha_i^* k(\mathbf{x}_i, \mathbf{x}), \tag{8}$$

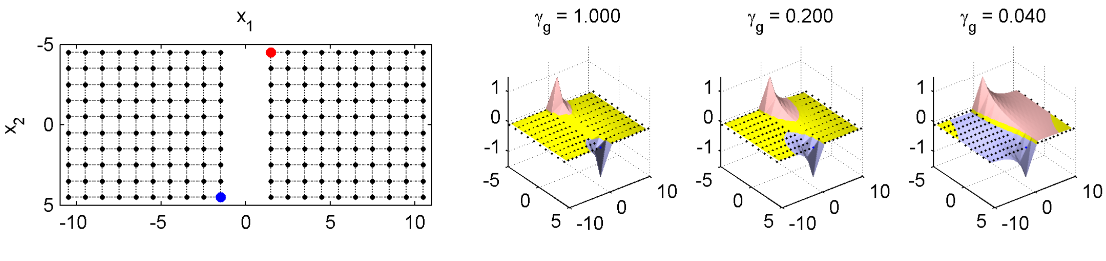
where $k(\cdot, \cdot)$ is a Mercer kernel associated with the RKHS $\mathcal{H}_K$. Therefore, we can apply the kernel trick and optimize rich classes of discriminators in a finite-dimensional space of $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)$. Finally, note that when $\gamma_g = \infty$, our solution f^* corresponds to supervised learning with SVMs.

4 EXISTING WORK

Most of the existing work on semi-supervised max-margin learning can be viewed as manifold regularization of SVMs (Belkin et al., 2006) or semi-supervised SVMs with the hat loss on unlabeled data (Bennett & Demiriz, 1999). The two approaches are reviewed in the rest of the section.

4.1 SEMI-SUPERVISED SVMS

Semi-supervised support vector machines with the *hat loss* $\widehat{V}(f, \mathbf{x}) = \max\{1 - |f(\mathbf{x})|, 0\}$ on unlabeled data (Bennett



(a) **(b)**

Figure 1: **a.** An example of a simple data adjacency graph. The vertices of the graph are depicted as dots. The red and blue dots are labeled vertices. The edges of the graph are shown as dotted lines and weighted as $w_{ij} = \exp[- \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 / 2]$. **b.** Three regularized harmonic function solutions on the data adjacency graph from Figure 1a. The plots are cubic interpolations of the solutions. The pink and blue colors denote parts of the feature space $\mathbf{x}$ where $\ell_i > 0$ and $\ell_i < 0$, respectively. The yellow color marks regions where the confidence $|\ell_i|$ is less than 0.05.

& Demiriz, 1999):

$$\min_f \sum_{i \in l} V(f, \mathbf{x}_i, y_i) + \gamma \|f\|_K^2 + \gamma_u \sum_{i \in u} \widehat{V}(f, \mathbf{x}_i) \quad (9)$$

compute max-margin decision boundaries that avoid dense regions of data. The hat loss makes the optimization problem non-convex. As a result, it is hard to solve the problem optimally and most of the work in this field has focused on approximations. A comprehensive review of these methods was done by Zhu (2008).

In comparison to semi-supervised SVMs, learning of max-margin graph cuts (7) is a convex problem. The convexity is achieved by having a two-stage learning algorithm. First, we infer labels of unlabeled examples using the regularized harmonic function solution, and then, we minimize the corresponding convex losses.

4.2 MANIFOLD REGULARIZATION OF SVMS

Manifold regularization of SVMs (Belkin et al., 2006):

$$\min_{f \in \mathcal{H}_K} \sum_{i \in l} V(f, \mathbf{x}_i, y_i) + \gamma \|f\|_K^2 + \gamma_u \mathbf{f}^\mathsf{T} L \mathbf{f}, \quad (10)$$

where $\mathbf{f} = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))$, computes max-margin decision boundaries that are smooth in the feature space. The smoothness is achieved by the minimization of the regularization term $\mathbf{f}^\mathsf{T} L \mathbf{f}$. Intuitively, when two examples are close on a manifold, the minimization of $\mathbf{f}^\mathsf{T} L \mathbf{f}$ leads to assigning the same label to both examples.

In some aspects, manifold regularization is similar to max-margin graph cuts. In particular, note that its objective (10) is similar to the regularized harmonic function solution (5). Both objectives involve regularization by a manifold, $\mathbf{f}^\mathsf{T} L \mathbf{f}$ and $\boldsymbol{\ell}^\mathsf{T} L \boldsymbol{\ell}$, regularization in the space of learned parameters, $\|f\|_K^2$ and $\boldsymbol{\ell}^\mathsf{T} I \boldsymbol{\ell}$, and some labeling constraints $V(f, \mathbf{x}_i, y_i)$ and $\ell_i = y_i$. Since max-margin graph cuts are learned conditionally on the harmonic function solution, the problems (7) and (10) may sometimes have similar solutions. A necessary condition is that the regularization terms in both objectives are weighted in the same proportions, for instance, by setting $\gamma_g = \gamma / \gamma_u$. We adopt this setting when manifold regularization of SVMs is compared to max-margin graph cuts in Section 6.

4.3 MANIFOLD REGULARIZATION FAILS

The major difference between manifold regularization (10) and the regularized harmonic function solution (5) is in the space of optimized parameters. In particular, manifold regularization is performed on a class of functions $\mathcal{H}_K$. When this class is severely restricted, such as linear functions, the minimization of $\mathbf{f}^\mathsf{T} L \mathbf{f}$ may lead to results, which are significantly worse than the harmonic function solution.

This issue can be illustrated on the problem from Figure 1, where we learn a linear decision boundary $f(\mathbf{x}) = \alpha_1 x_1 + \alpha_2 x_2$ through manifold regularization of linear SVMs:

$$\min_{\alpha_1, \alpha_2} \sum_{i \in l} V(f, \mathbf{x}_i, y_i) + \gamma[\alpha_1^2 + \alpha_2^2] + \gamma_u \mathbf{f}^\mathsf{T} L \mathbf{f}. \quad (11)$$

The structure of our problem simplifies the computation of the regularization term $\mathbf{f}^\mathsf{T} L \mathbf{f}$. In particular, since all edges in the data adjacency graph are either horizontal or vertical, the term $\mathbf{f}^\mathsf{T} L \mathbf{f}$ can be expressed as a function of α_1^2 and α_2^2:

$$\begin{aligned} \mathbf{f}^\mathsf{T} L \mathbf{f} &= \frac{1}{2} \sum_{i,j} w_{ij} (f(\mathbf{x}_i) - f(\mathbf{x}_j))^2 \\ &= \frac{1}{2} \sum_{i,j} w_{ij} (\alpha_1(\mathbf{x}_{i1} - \mathbf{x}_{j1}) + \alpha_2(\mathbf{x}_{i2} - \mathbf{x}_{j2}))^2 \\ &= \frac{\alpha_1^2}{2} \underbrace{\sum_{i,j} w_{ij} (\mathbf{x}_{i1} - \mathbf{x}_{j1})^2}_{\Delta = 218.351} + \\ &\quad \frac{\alpha_2^2}{2} \underbrace{\sum_{i,j} w_{ij} (\mathbf{x}_{i2} - \mathbf{x}_{j2})^2}_{\Delta = 218.351}, \end{aligned} \quad (12)$$

and incorporated in our objective function as an additional weight at the regularizer $[\alpha_1^2 + \alpha_2^2]$:

$$\min_{\alpha_1,\alpha_2} \sum_{i \in l} V(f, \mathbf{x}_i, y_i) + \left(\gamma + \frac{\gamma_u \Delta}{2}\right)[\alpha_1^2 + \alpha_2^2]. \quad (13)$$

Thus, manifold regularization of linear SVMs on our problem can be viewed as supervised learning with linear SVMs with a varying weight at the regularizer. Since the problem involves only two labeled examples, changes in the weight $\left(\gamma + \frac{\gamma_u \Delta}{2}\right)$ do not affect the direction of the discriminator $f^*(\mathbf{x}) = 0$ and only change the slope of f^* (Figure 2).

The above analysis shows that the discriminator $f^*(\mathbf{x}) = 0$ does not change with γ_u. As a result, all discriminators are equal to the discriminator for $\gamma_u = 0$, which can be learned by linear SVMs, and none of them solves our problem optimally. Max-margin graph cuts solve the problem optimally for small values of γ_g (Figure 2).

A similar line of reasoning can be used to extend our results to polynomial kernels. Figure 2 indicates that max-margin learning with the cubic kernel exhibits similar trends to the linear case.

5 THEORETICAL ANALYSIS

The notion of algorithmic stability can be used to bound the generalization error of many learning algorithms (Bousquet & Elisseeff, 2002). In this section, we discuss how to make the harmonic function solution stable and prove a bound on the generalization error of max-margin cuts (7). Our bound combines existing transductive (Belkin et al., 2004; Cortes et al., 2008) and inductive (Vapnik, 1995) bounds.

5.1 GENERALIZATION ERROR

Our objective is to show that the *risk* of our solutions f:

$$R_P(f) = \mathbb{E}_{P(\mathbf{x})}[\mathcal{L}(f(\mathbf{x}), y(\mathbf{x}))] \quad (14)$$

is bounded by the *empirical risk* on graph-induced labels:

$$\frac{1}{n}\sum_i \mathcal{L}(f(\mathbf{x}_i), \mathrm{sgn}(\ell_i^*)) \quad (15)$$

and error terms, which can be computed from training data. The function $\mathcal{L}(y', y) = \mathbb{1}\{\mathrm{sgn}(y') \neq y\}$ computes the zero-one loss of the prediction $\mathrm{sgn}(y')$ given the ground truth y, and $P(\mathbf{x})$ is the distribution of our data. For simplicity, we assume that the label y is a deterministic function of $\mathbf{x}$. Our proof starts by relating $R_P(f)$ and graph-induced labels ℓ_i^*.

Lemma 1. *Let f be from a function class with the VC dimension h and $\mathbf{x}_i$ be n examples, which are sampled i.i.d. with respect to the distribution $P(\mathbf{x})$. Then the inequality:*

$$\begin{aligned} R_P(f) \ \le\ & \frac{1}{n}\sum_i \mathcal{L}(f(\mathbf{x}_i), \mathrm{sgn}(\ell_i^*)) + \\ & \frac{1}{n}\sum_i (\ell_i^* - y_i)^2 + \\ & \underbrace{\sqrt{\frac{h(\ln(2n/h) + 1) - \ln(\eta/4)}{n}}}_{\text{inductive error } \Delta_I(h,n,\eta)} \end{aligned}$$

holds with probability $1 - \eta$, where y_i and ℓ_i^* represent the true and graph-induced soft labels, respectively.

Proof: Based on Equations 3.15 and 3.24 (Vapnik, 1995), the inequality:

$$R_P(f) \le \frac{1}{n}\sum_i \mathcal{L}(f(\mathbf{x}_i), y_i) + \Delta_I(h, n, \eta)$$

holds with probability $1 - \eta$. Our final claim follows from bounding all terms $\mathcal{L}(f(\mathbf{x}_i), y_i)$ as:

$$\mathcal{L}(f(\mathbf{x}_i), y_i) \le \mathcal{L}(f(\mathbf{x}_i), \mathrm{sgn}(\ell_i^*)) + (\ell_i^* - y_i)^2.$$

The above bound holds for any $y_i \in \{-1, 1\}$ and ℓ_i^*. ■

It is hard to bound the error term $\frac{1}{n}\sum_i (\ell_i^* - y_i)^2$ when the constraints $\ell_i = y_i$ (5) are enforced in a hard manner. Thus, in the rest of our analysis, we consider a relaxed version of the harmonic function solution (Cortes et al., 2008):

$$\min_{\boldsymbol{\ell} \in \mathbb{R}^n} (\boldsymbol{\ell} - \mathbf{y})^\mathsf{T} C (\boldsymbol{\ell} - \mathbf{y}) + \boldsymbol{\ell}^\mathsf{T} L \boldsymbol{\ell}, \quad (16)$$

where L is the Laplacian of the data adjacency graph, C is a diagonal matrix such that $C_{ii} = c_l$ for all labeled examples, and $C_{ii} = c_u$ otherwise, and $\mathbf{y}$ is a vector of pseudo-targets such that y_i is the label of the i-th example when the example is labeled, and $y_i = 0$ otherwise.

The generalization error of the solution to the problem (16) is bounded in Lemma 2. To simplify the proof, we assume that $c_l = 1$ and $c_l > c_u$.

Lemma 2. *Let $\boldsymbol{\ell}^*$ be a solution to the problem:*

$$\min_{\boldsymbol{\ell} \in \mathbb{R}^n} (\boldsymbol{\ell} - \mathbf{y})^\mathsf{T} C (\boldsymbol{\ell} - \mathbf{y}) + \boldsymbol{\ell}^\mathsf{T} Q \boldsymbol{\ell},$$

where $Q = L + \gamma_g I$ and all labeled examples l are selected i.i.d. Then the inequality:

$$\begin{aligned} R_P^W(\boldsymbol{\ell}^*) \ &\le\ \widehat{R}_P^W(\boldsymbol{\ell}^*) + \underbrace{\beta + \sqrt{\frac{2\ln(2/\delta)}{n_l}}(n_l\beta + 4)}_{\text{transductive error } \Delta_T(\beta, n_l, \delta)} \\ \beta \ &\le\ 2\left[\frac{\sqrt{2}}{\gamma_g + 1} + \sqrt{2n_l}\frac{1 - \sqrt{c_u}}{\sqrt{c_u}}\frac{\lambda_M(L) + \gamma_g}{\gamma_g^2 + 1}\right] \end{aligned}$$

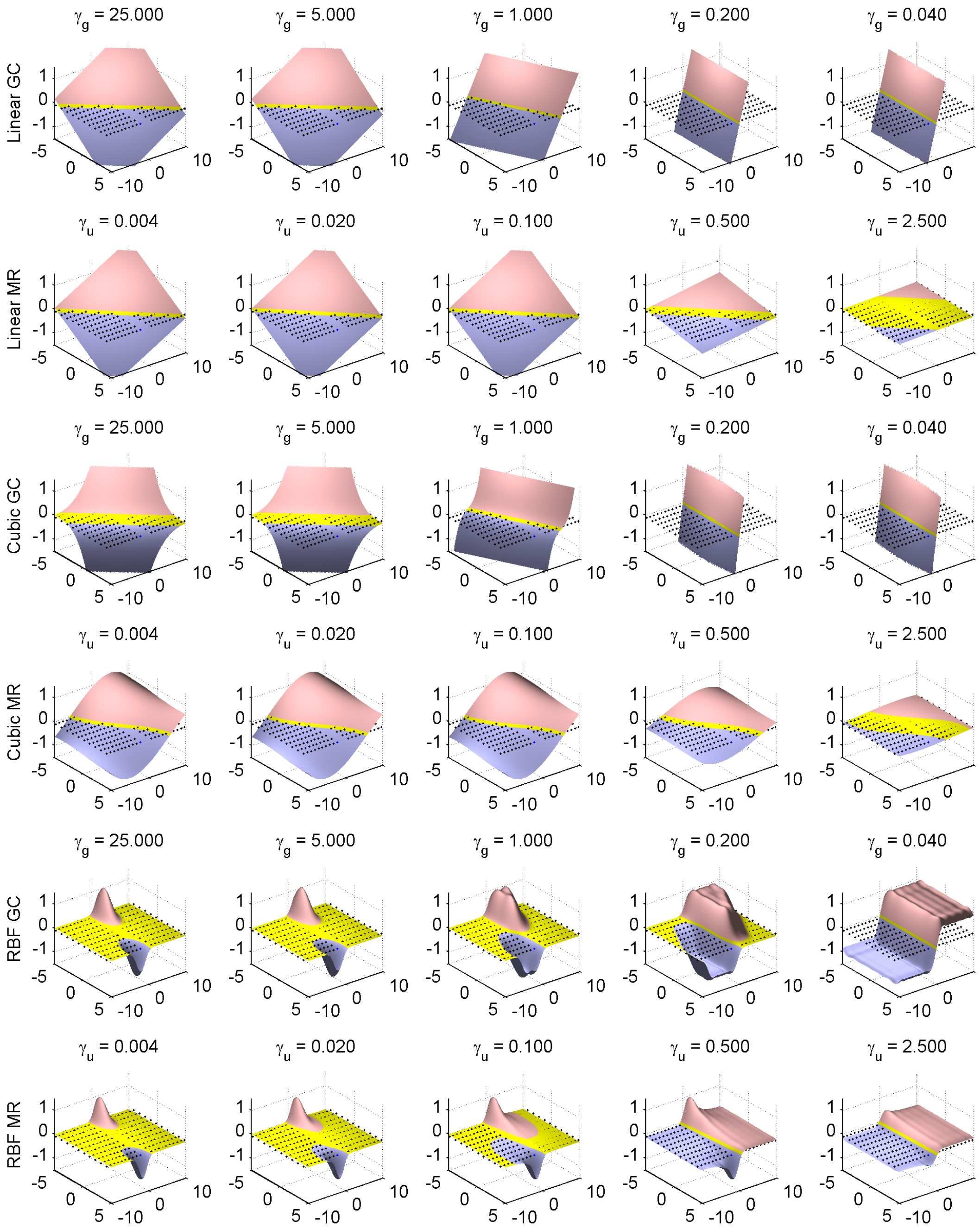


Figure 2: Linear, cubic, and RBF decision boundaries obtained by manifold regularization of SVMs (MR) and max-margin graph cuts (GC) on the problem from Figure 1. The regularization parameter $\gamma_g = \gamma/\gamma_u$ is set as suggested in Section 4.2, $\gamma = 0.1$, and $\varepsilon = 0.01$. The pink and blue colors denote parts of the feature space $\mathbf{x}$ where the discriminators f are positive and negative, respectively. The yellow color marks regions where $|f(\mathbf{x})| < 0.05$.

holds with probability $1 - \delta$*, where:*

$$\begin{aligned} R_P^W(\boldsymbol{\ell}^*) &= \frac{1}{n} \sum_i (\ell_i^* - y_i)^2 \\ \widehat{R}_P^W(\boldsymbol{\ell}^*) &= \frac{1}{n_l} \sum_{i \in l} (\ell_i^* - y_i)^2 \end{aligned}$$

are risk terms for all and labeled vertices, respectively, and β *is the stability coefficient of the solution* $\boldsymbol{\ell}^*$*.*

Proof: Our risk bound follows from combining Theorem 1 of Belkin et al. (2004) with the assumptions $|y_i| \leq 1$ and $|\ell_i^*| \leq 1$. The coefficient β is derived based on Section 5 of Cortes et al. (2008). In particular, based on the properties of the matrix C and Proposition 1 (Cortes et al., 2008), we conclude:

$$\beta = 2 \left[\frac{\sqrt{2}}{\lambda_m(Q) + 1} + \sqrt{2 n_l} \frac{1 - \sqrt{c_u}}{\sqrt{c_u}} \frac{\lambda_M(Q)}{(\lambda_m(Q) + 1)^2} \right],$$

where $\lambda_m(Q)$ and $\lambda_M(Q)$ refer to the smallest and largest eigenvalues of Q, respectively, and can be further rewritten as $\lambda_m(Q) = \lambda_m(L) + \gamma_g$ and $\lambda_M(Q) = \lambda_M(L) + \gamma_g$. Our final claim directly follows from applying the lower bounds $\lambda_m(L) \geq 0$ and $(\lambda_m(L) + \gamma_g + 1)^2 \geq \gamma_g^2 + 1$. ■

Lemma 2 is practical when the error $\Delta_T(\beta, n_l, \delta)$ decreases at the rate of $O(n_l^{-\frac{1}{2}})$. This is achieved when $\beta = O(1/n_l)$, which corresponds to $\gamma_g = \Omega(n_l^{\frac{3}{2}})$. Thus, when the problem (16) is sufficiently regularized, its solution is stable, and the generalization error of the solution is bounded.

Lemmas 1 and 2 can be combined using the union bound.

Proposition 1. *Let* f *be from a function class with the VC dimension* h*. Then the inequality:*

$$\begin{aligned} R_P(f) \;\leq\; & \frac{1}{n} \sum_i \mathcal{L}(f(\mathbf{x}_i), \operatorname{sgn}(\ell_i^*)) + \\ & \widehat{R}_P^W(\boldsymbol{\ell}^*) + \Delta_T(\beta, n_l, \delta) + \Delta_I(h, n, \eta) \end{aligned}$$

holds with probability $1 - (\eta + \delta)$*.*

The above result can be viewed as follows. If both n and n_l are large, the sum of $\frac{1}{n} \sum_i \mathcal{L}(f(\mathbf{x}_i), \operatorname{sgn}(\ell_i^*))$ and $\widehat{R}_P^W(\boldsymbol{\ell}^*)$ provides a good estimate of the risk $R_P(f)$. Unfortunately, our bound is not practical for setting γ_g because it is hard to find γ_g that minimizes both $\widehat{R}_P^W(\boldsymbol{\ell}^*)$ and $\Delta_T(\beta, n_l, \delta)$. The same phenomenon was observed by Belkin et al. (2004) in a similar context. To solve our problem, we suggest setting γ_g based on the validation set. This methodology is used in the experimental section.

5.2 THRESHOLD ε

Finally, note that when $|\ell_i^*| < \varepsilon$, where ε is a small number, $|\ell_i^* - y_i|$ is close to 1 irrespective of y_i, and a trivial upper bound $\mathcal{L}(f(\mathbf{x}_i), y_i) \leq 1$ is almost as good as $\mathcal{L}(f(\mathbf{x}_i), y_i) \leq \mathcal{L}(f(\mathbf{x}_i), \operatorname{sgn}(\ell_i^*)) + (\ell_i^* - y_i)^2$ for any f. This allows us to justify the ε threshold in the problem (7). In particular, note that $\mathcal{L}(f(\mathbf{x}_i), y_i)$ is bounded by $1 - (\ell_i^* - y_i)^2 + (\ell_i^* - y_i)^2$. When $|\ell_i^*| < \varepsilon$, $1 - (\ell_i^* - y_i)^2 < 2\varepsilon - \varepsilon^2$, and we conclude the following.

Proposition 2. *Let* f *be from a function class with the VC dimension* h *and* n_ε *be the number of examples such that* $|\ell_i^*| < \varepsilon$*. Then the inequality:*

$$\begin{aligned} R_P(f) \;\leq\; & \frac{1}{n} \sum_{i : |\ell_i^*| \geq \varepsilon} \mathcal{L}(f(\mathbf{x}_i), \operatorname{sgn}(\ell_i^*)) + \frac{2 \varepsilon n_\varepsilon}{n} + \\ & \widehat{R}_P^W(\boldsymbol{\ell}^*) + \Delta_T(\beta, n_l, \delta) + \Delta_I(h, n, \eta) \end{aligned}$$

holds with probability $1 - (\eta + \delta)$*.*

Proof: The generalization bound is proved as:

$$\begin{aligned} R_P(f) &\leq \widehat{R}_P(f) + \Delta_I(h, n, \eta) \\ &= \frac{1}{n} \sum_{i : |\ell_i^*| \geq \varepsilon} \mathcal{L}(f(\mathbf{x}_i), y_i) + \frac{1}{n} \sum_{i : |\ell_i^*| < \varepsilon} \mathcal{L}(f(\mathbf{x}_i), y_i) + \\ &\quad \Delta_I(h, n, \eta) \\ &\leq \frac{1}{n} \sum_{i : |\ell_i^*| \geq \varepsilon} \left[\mathcal{L}(f(\mathbf{x}_i), \operatorname{sgn}(\ell_i^*)) + (\ell_i^* - y_i)^2 \right] + \\ &\quad \frac{1}{n} \sum_{i : |\ell_i^*| < \varepsilon} \left[1 - (\ell_i^* - y_i)^2 + (\ell_i^* - y_i)^2 \right] + \\ &\quad \Delta_I(h, n, \eta) \\ &= \frac{1}{n} \sum_{i : |\ell_i^*| \geq \varepsilon} \mathcal{L}(f(\mathbf{x}_i), \operatorname{sgn}(\ell_i^*)) + \\ &\quad \frac{1}{n} \sum_{i : |\ell_i^*| < \varepsilon} \left[1 - (\ell_i^* - y_i)^2 \right] + \frac{1}{n} \sum_i (\ell_i^* - y_i)^2 + \\ &\quad \Delta_I(h, n, \eta) \\ &\leq \frac{1}{n} \sum_{i : |\ell_i^*| \geq \varepsilon} \mathcal{L}(f(\mathbf{x}_i), \operatorname{sgn}(\ell_i^*)) + \frac{2 \varepsilon n_\varepsilon}{n} + \\ &\quad \widehat{R}_P^W(\boldsymbol{\ell}^*) + \Delta_T(\beta, n_l, \delta) + \Delta_I(h, n, \eta). \end{aligned}$$

The last step follows from the inequality $1 - (\ell_i^* - y_i)^2 < 2\varepsilon$ and Lemma 2. ■

When $\varepsilon \leq n_l^{-\frac{1}{2}}$, the new upper bound is asymptotically as good as the bound in Proposition 1. As a result, we get the same convergence guarantees although highly-uncertain labels $|\ell_i^*| < \varepsilon$ are excluded from our optimization.

In practice, optimization of the thresholded objective often yields a lower risk $\frac{1}{n} \sum_{i : |\ell_i^*| \geq \varepsilon} \mathcal{L}(f^*(\mathbf{x}_i), \operatorname{sgn}(\ell_i^*)) + \frac{2 \varepsilon n_\varepsilon}{n}$, and also lower training and test errors. This is a result of excluding the most uncertain examples $|\ell_i^*| < \varepsilon$ from learning. Figure 3 illustrates these trends on three learning problems.

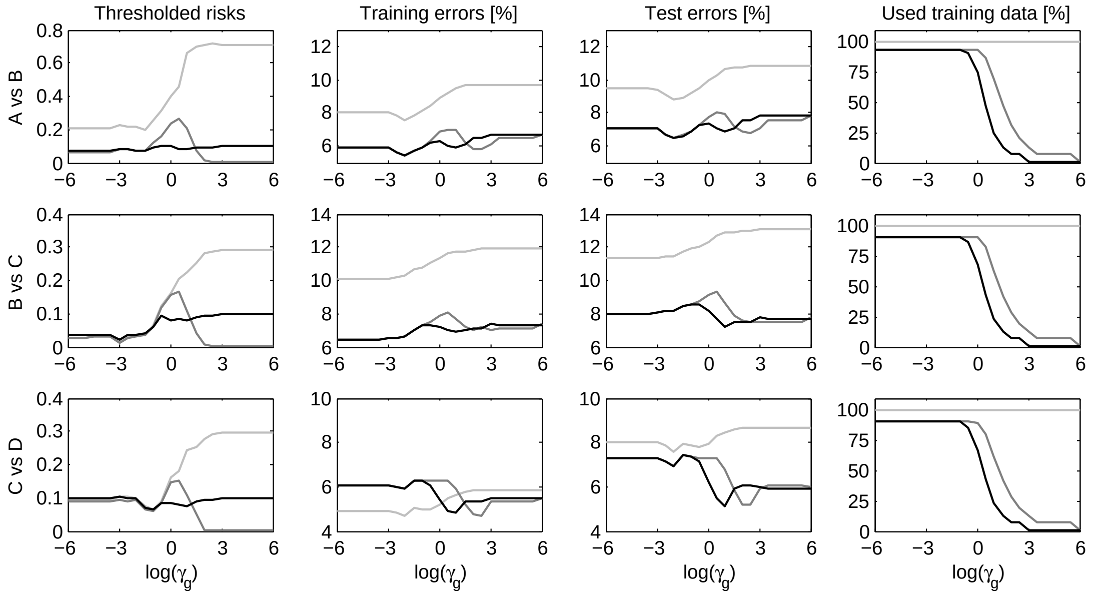


Figure 3: The thresholded empirical risk $\frac{1}{n}\sum_{i:|\ell_i^*|\geq\varepsilon}\mathcal{L}(f^*(\mathbf{x}_i), \mathrm{sgn}(\ell_i^*)) + \frac{2\varepsilon n_\varepsilon}{n}$ of the optimal max-margin graph cut f^* (7), its training and test errors, and the percentage of training examples such that $|\ell_i^*| \geq \varepsilon$, on 3 letter recognition problems from the UCI ML repository. The plots are shown as functions of the parameter γ_g, and correspond to the thresholds $\varepsilon = 0$ (light gray lines), $\varepsilon = 10^{-6}$ (dark gray lines), and $\varepsilon = 10^{-3}$ (black lines). All results are averaged over 50 random choices of 1 percent of labeled examples.

Note that the parameters γ_g and ε are redundant in the sense that the same result is often achieved by different combinations of parameter values. This problem is addressed in the experimental section by fixing ε and optimizing γ_g only.

6 EXPERIMENTS

The experimental section is divided into two parts. The first part compares max-margin graph cuts to manifold regularization of SVMs on the problem from Figure 1. The second part compares max-margin graph cuts, manifold regularization of SVMs, and supervised learning with SVMs on three UCI ML repository datasets (Asuncion & Newman, 2007).

Manifold regularization of SVMs is evaluated based on the implementation of Belkin et al. (2006). Max-margin graph cuts and SVMs are implemented using LIBSVM (Chang & Lin, 2001).

6.1 SYNTHETIC PROBLEM

The first experiment (Figure 2) illustrates linear, cubic, and RBF graph cuts (7) on the synthetic problem from Figure 1. The cuts are shown for various settings of the regularization parameter γ_g. As γ_g decreases, note that the cuts gradually interpolate between supervised learning on just two labeled examples and semi-supervised learning on all data. The resulting discriminators are max-margin decision boundaries that separate the corresponding colored regions in Figure 1. Figure 2 also shows that manifold regularization of SVMs (10) with linear and cubic kernels cannot perfectly separate the two clusters in Figure 1 for any setting of the parameter γ_u. The reason for this problem is discussed in Section 4.3. Finally, note the similarity between max-margin graph cuts and manifold regularization of SVMs with the RBF kernel. This similarity was suggested in Section 4.2.

6.2 UCI ML REPOSITORY DATASETS

The second experiment (Figure 4) shows that max-margin graph cuts (7) typically outperform manifold regularization of SVMs (10) and supervised learning with SVMs. The experiment is done on three UCI ML repository datasets: letter recognition, digit recognition, and image segmentation. The datasets are multi-class and thus, we transform each of them into a set of binary classification problems. The digit recognition and image segmentation datasets are converted into 45 and 15 problems, respectively, where all classes are discriminated against every other class. The letter recognition dataset is turned into 25 problems that involve pairs of consecutive letters. Each dataset is divided into three folds. The first fold is used for training, the second one for selecting the parameters $\gamma \in [0.01, 0.1]n_l$, $\gamma_u \in [10^{-3}, 10^3]\gamma$, and $\gamma_g = \gamma/\gamma_u$, and the last fold is used for testing.[1] The fraction of labeled examples in the training set is varied from 1 to 10 percent. All examples in the validation set are labeled

[1] Alternatively, the regularization parameters γ, γ_u, and γ_g can be set using leave-one-out cross-validation on labeled examples.

Dataset	L	Misclassification errors [%]								
		Linear kernel			Cubic kernel			RBF kernel		
		SVM	MR	GC	SVM	MR	GC	SVM	MR	GC
Letter recognition	1	18.90	30.94	**15.79**	20.54	25.96	**17.45**	20.06	17.61	**16.01**
	2	12.92	28.45	**10.79**	12.18	18.34	**10.90**	13.52	13.10	**11.83**
	5	8.21	27.13	**5.65**	5.49	18.77	**4.80**	6.81	8.06	**5.65**
	10	6.51	25.45	**3.96**	4.17	14.03	**2.96**	4.95	6.14	**3.32**
Digit recognition	1	7.06	9.59	**6.88**	9.62	**5.29**	8.55	8.22	**6.36**	7.65
	2	4.87	7.97	**4.60**	6.06	**5.06**	5.09	6.17	**4.21**	5.61
	5	2.97	3.68	**2.29**	3.04	**2.27**	2.36	2.74	2.29	**2.19**
	10	1.70	2.86	**1.59**	1.87	**1.60**	1.74	1.68	1.75	**1.35**
Image segmentation	1	14.02	11.81	**10.27**	23.30	**12.02**	14.10	14.02	11.60	**9.51**
	2	8.54	10.87	**7.69**	14.28	13.07	**7.73**	9.06	8.93	**7.34**
	5	4.73	7.83	**4.49**	8.32	8.79	**7.17**	5.87	5.43	**5.31**
	10	3.30	6.26	**3.28**	3.65	6.64	**3.60**	3.84	4.81	**3.73**

Figure 4: Comparison of SVMs, max-margin graph cuts (GC), and manifold regularization of SVMs (MR) on three datasets from the UCI ML repository. The fraction of labeled examples L varies from 1 to 10 percent.

and its size is limited to the number of labeled examples in the training set.

In all experiments, we use 5-nearest neighbor graphs whose edges are weighted as $w_{ij} = \exp[-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 / (2K\sigma^2)]$, where K is the number of features and σ denotes the mean of their standard deviations. The width of radial basis functions (RBFs) is set accordingly to $\sqrt{K}\sigma$, and the threshold ε for choosing training examples (7) is 10^{-6}.

Test errors of all compared algorithms are averaged over all binary problems within each dataset and shown in Figure 4. Max-margin graph cuts outperform manifold regularization of SVMs in 29 out of 36 experiments. Note that the lowest errors are usually obtained for linear and cubic kernels, and our method improves the most over manifold regularization of SVMs in these settings.

7 CONCLUSIONS

This paper proposes a novel algorithm for semi-supervised learning. The algorithm learns max-margin graph cuts that are conditioned on the labels induced by the harmonic function solution. We motivate the approach, prove its generalization bound, and compare it to state-of-the-art algorithms for semi-supervised max-margin learning. The approach is evaluated on a synthetic problem and three UCI ML repository datasets, and we show that it usually outperforms manifold regularization of SVMs.

In our future work, we plan to investigate some of the shortcomings of this paper. For instance, note that the theoretical analysis of max-margin graph cuts (Section 5) assumes soft labels but our solutions (7) are computed using the hard labels $\ell_i = y_i$. Whether the theoretically sound setting yields better results in practice is an open question.

References

Asuncion, A., & Newman, D. (2007). UCI machine learning repository..

Belkin, M., Matveeva, I., & Niyogi, P. (2004). Regularization and semi-supervised learning on large graphs. In *Proceeding of the 17th Annual Conference on Learning Theory*, pp. 624–638.

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

Bennett, K., & Demiriz, A. (1999). Semi-supervised support vector machines. In *Advances in Neural Information Processing Systems 11*, pp. 368–374.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, *2*, 499–526.

Chang, C.-C., & Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Cortes, C., Mohri, M., Pechyony, D., & Rastogi, A. (2008). Stability of transductive regression algorithms. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 176–183.

Smola, A., & Kondor, R. (2003). Kernels and regularization on graphs. In *Proceedings of the 16th Annual Conference on Learning Theory and the 7th Workshop on Kernel Machines*.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY.

Wahba, G. (1999). *Support Vector Machines, Reproducing Kernel Hilbert Spaces, and Randomized GACV*, pp. 69–88. MIT Press, Cambridge, MA.

Zhu, X. (2008). Semi-supervised learning literature survey. Tech. rep. 1530, University of Wisconsin-Madison.

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 912–919.