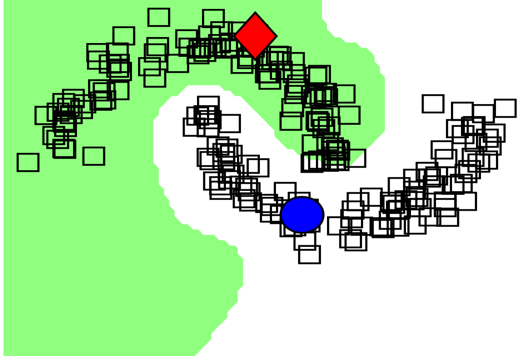
Manifold Regularization



Vikas Sindhwani

Department of Computer Science

University of Chicago

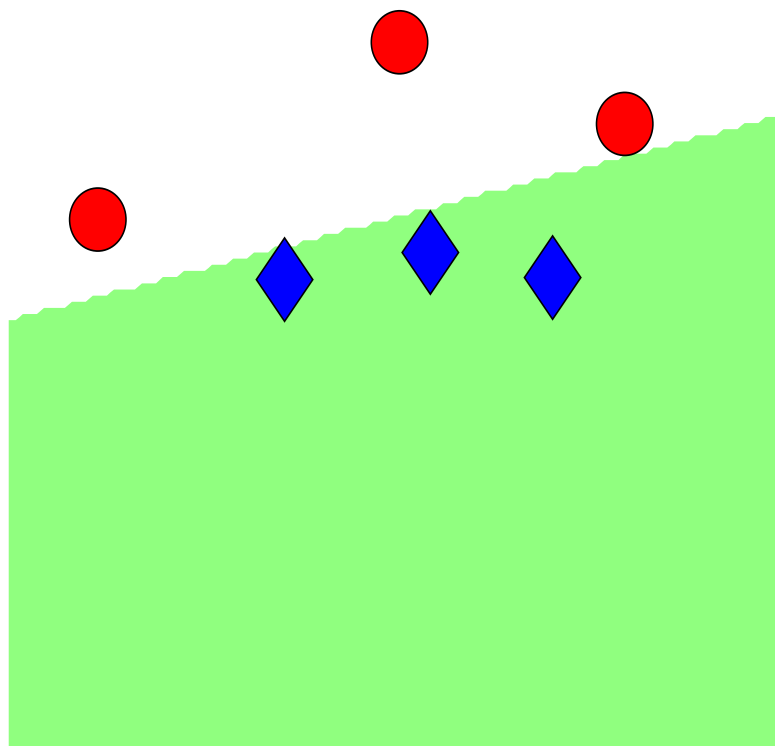
Joint Work with Mikhail Belkin and Partha Niyogi

The Problem of Learning

- $S = \{x_i, y_i\}_{i=1}^{l}$ is drawn from an unknown probability distribution $P_{X \times Y}$.
- A Learning Algorithm maps S to an element f_S of a hypothesis space $\mathcal{H}$ of functions mapping $X \to Y$.
- f_S should provide good labels for future examples.
- Regularization : Choose a simple function that agrees with data.

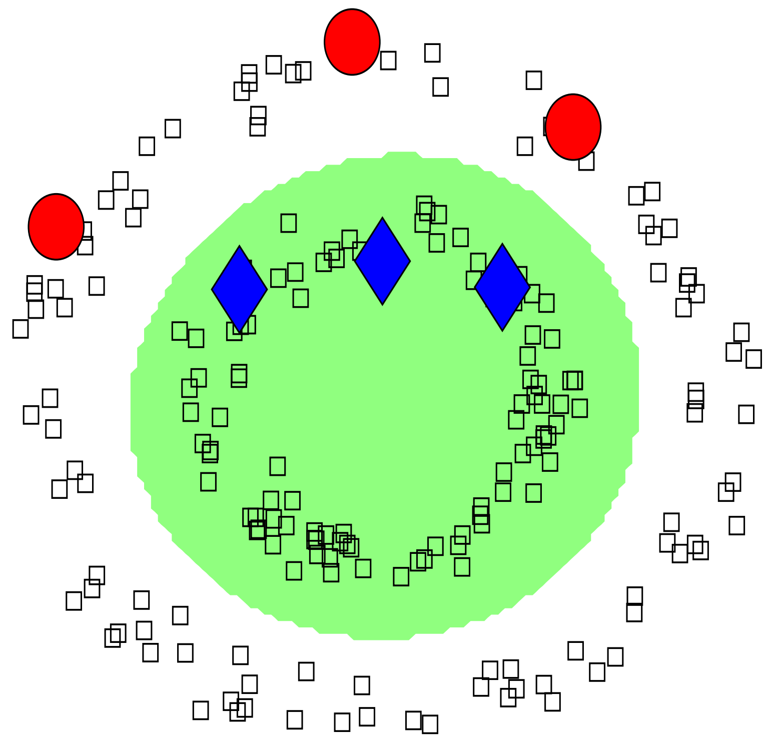
The Problem of Learning

Notions of simplicity are the key to successful learning. Here's a simple function that agrees with data.



Learning and Prior Knowledge

But Simplicity is a Relative Concept. Prior Knowledge of the Marginal can modify our notions of simplicity.



Motivation

- How can we exploit prior knowledge of the marginal distribution $\mathcal{P}_X$?
- More practically, how can we use unlabeled examples drawn from $\mathcal{P}_X$
- Why is this important ?
  - Natural Data has structure to exploit.
  - Natural Learning is largely semi-supervised.
  - Labels are Expensive, Unlabeled data is cheap and plenty.

Contributions

- A data-dependent, Geometric Regularization Framework for Learning from examples.
- Representer Theorems provide solutions.
- Extensions of SVM and RLS for Semi-supervised Learning.
- Regularized Spectral Clustering and Dimensionality Reduction.
- The problem of Out-of-sample extensions in graph methods is resolved.
- Good Empirical Performance.

Regularization with RKHS

- Learning in Reproducing Kernel Hilbert Spaces :

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

- Regularized Least Squares (RLS) :
  $V(f(x_i), y_i) = (y_i - f(x_i))^2$
  Support Vector Machine (SVM) :
  $V(f(x_i), y_i) = \max\left[0, 1 - y_i f(x_i)\right]$

What are RKHS ?

- Hilbert Spaces with a nice property :
  - If two functions $f, g \in \mathcal{H}$ are close in the distance derived from the inner product, their values $f(x), g(x)$ are close $\forall x \in X$.
- Reproducing Property :
  - $\mathcal{E}_x : f \mapsto f(x)$ is linear, continuous. By Reisz's Representation theorem, $\exists K(x, .) \in \mathcal{H} : \mathcal{E}_x(f) = \langle f, K_x \rangle_{\mathcal{H}} = f(x)$.
- Kernel Function $\leftrightarrow$ RKHS :
  - $K(x, t) = K_x(t) = \langle K_x, K_t \rangle$

Why RKHS ?

- Rich Function Spaces with complexity control
  e.g Gaussian Kernel $K(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$:
  $\|f^*\|_K^2 = \int |\tilde{f}(\omega)|^2 r(\|w\|^2) d\omega$
- Representer Theorems show that the minimizer has the form :
  $f^*(.) = \sum_{i=1}^{l} \alpha_i K(x_i, .)$ and therefore,
  $\|f^*\|_K^2 = \langle f^*, f^* \rangle_{\mathcal{H}_K} = \sum_{i,j=1}^{l} \alpha_i \alpha_j K(x_i, x_j)$
- Motivates kernelization (KPCA, KFD, etc).
- Good empirical performance.

Known Marginal

- If $\mathcal{P}_X$ is known, solve :

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

- Extrinsic and Intrinsic Regularization
  - γ_A controls complexity in ambient space.
  - γ_I controls complexity in the intrinsic geometry of $\mathcal{P}_X$

Continuous Representer Theorem

Assume that the penalty term $\|f\|_I$ is sufficiently smooth with respect to the RKHS norm $\|f\|_K$. Then the solution f^* to the optimization problem exists and admits the following representation

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

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

A Manifold Regularizer

If $\mathcal{M}$, the support of the marginal is a compact submanifold $\mathcal{M} \subset X = R^n$, it seems natural to choose :

$$\|f\|_I^2 = \int_{\mathcal{M}} \langle \nabla_{\mathcal{M}} f, \nabla_{\mathcal{M}} f \rangle$$

and to find $f^* \in \mathcal{H}_K$ that minimizes :

$$\frac{1}{l} \sum_{i=1}^{l} V(f(x_i), y_i) + \gamma_A \|f\|_K^2 + \gamma_I \int_{\mathcal{M}} \langle \nabla_{\mathcal{M}} f, \nabla_{\mathcal{M}} f \rangle$$

Laplace Beltrami Operator

The intrinsic regularizer is a quadratic form involving the Laplace-Beltrami operator on the manifold $\mathcal{L}f \stackrel{def}{=} -div\nabla_{\mathcal{M}}f$:

$$\|f\|_I = \int_{\mathcal{M}} \langle \nabla_{\mathcal{M}} f, \nabla_{\mathcal{M}} f \rangle = \int_{\mathcal{M}} f \mathcal{L} f$$

because some calculus on manifolds establishes that for any vector field $V (= \nabla_{\mathcal{M}} f)$,

$$\int_{\mathcal{M}} \langle V, \nabla_{\mathcal{M}} f \rangle = - \int_{\mathcal{M}} div(V) f$$

Passage to the Discrete

- In reality, $\mathcal{M}$ is unknown and sampled only via examples $\{x_i\}_{i=1}^{l+u}$. Labels are not required for empirical estimates of $\|f\|_I^2$.
- Manifold $\mathcal{M} \leftrightarrow$ Graph $\mathcal{G}(V, E)$
  $V = \{x_i\}_{i=1}^{l+u} \;\; E = \{(x_i, x_j) : x_i \;\sim_{W_{ij}} x_j\}$.
- Laplace Beltrami $\mathcal{L} \leftrightarrow$ Graph Laplacian L
  $L \stackrel{def}{=} D - W \quad D = diag\{D_{ii} = \sum_j W_{ij}\}$.
- $\|f\|_I^2 = \int_{\mathcal{M}} f \mathcal{L} f \leftrightarrow \widehat{\|f\|_I^2} = \mathbf{f}^T L \mathbf{f} = \sum (f(x_i) - f(x_j))^2 W_{ij}$

Algorithms

- We have motivated the following optimization problem : Find a function $f^* \in \mathcal{H}_K$ that minimizes :

$$\frac{1}{l}\sum_{i=1}^{l} V(f(x_i), y_i) + \gamma_A \|f\|_K^2 + \frac{\gamma_I}{(l+u)^2}\mathbf{f}^T L\mathbf{f}$$

- Laplacian RLS
  $V(f(x_i), y_i) = (y_i - f(x_i))^2$
  Laplacian SVM
  $V(f(x_i), y_i) = \max\left[0, 1 - y_i f(x_i)\right]$

Empirical Representer Theorem

The minimizer admits an expansion

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

Proof :
Write any $f \in \mathcal{H}_K$ as $\sum_{i=1}^{l+u} \alpha_i K(x_i, x) + f_\perp$

- $f(x_j) = \langle f, K_{x_j} \rangle = \sum_{i=1}^{l+u} \alpha_i K(x_i, x_j)$
- $f_\perp$ increases the norm. So $f^*_\perp = 0$.

Laplacian RLS

By the Representer Theorem, the problem becomes finite dimensional. For Laplacian RLS, we find $\alpha^* \in \mathcal{R}^{l+u}$ that minimizes :

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

where K : Gram Matrix ; $Y = [y_1, \dots, y_l, 0 \dots, 0]$ and $J = diag(1, \dots, 1, 0, \dots, 0)$. The solution is :

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

Laplacian SVM

For Laplacian SVMs, we solve a QP :

$$\beta^* = \operatorname{argmax}_{\beta \in \mathcal{R}^l} \sum_{i=1}^{l} \beta_i - \frac{1}{2}\beta^T Q \beta$$

subject to :

$$\sum_{i=1}^{l} \beta_i y_i = 0$$

$$0 \le \beta_i \le \frac{1}{l}$$

where $Q = YJK(2\gamma_A I + 2\frac{\gamma_I}{(l+u)^2}LK)^{-1}J^T Y$, and then invert a linear system :

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

Manifold Regularization

- Input : l labeled and u unlabeled examples
- Output : $f : \mathcal{R}^n \mapsto \mathcal{R}$
- Algorithm :
  - Contruct adjacency Graph. Compute Laplacian.
  - Choose Kernel $K(x, y)$. Compute Gram matrix K.
  - Choose γ_A, γ_I. (?)
  - Compute α^*.
  - Output $f^*(x) = \sum_{i=1}^{l+u} \alpha_i^* K(x_i, x)$

Unity of Learning

Supervised	Partially Supervised	Unsupervised
SVM/RLS $\text{argmin}_{f \in \mathcal{H}_K}$ $\frac{1}{l}\sum_{i=1}^{l} V(y_i, f(x_i)) +$ $\gamma \|f\|_K^2$	*Graph Regularization* $\text{argmin}_{\mathbf{f} \in \mathcal{R}^{(l+u)}}$ $\frac{1}{l}\sum_{i=1}^{l} V(y_i, \mathbf{f_i}) + \gamma \mathbf{f}^T L \mathbf{f}$ *Out-of-sample Extn.* $\text{argmin}_{f \in \mathcal{H}_K}$ $\frac{1}{l}\sum_{i=1}^{l} V(y_i, \mathbf{f_i}) + \gamma \mathbf{f}^T L \mathbf{f}$	*Graph Mincut* $\text{argmin}_{\mathbf{f} \in \{-1,+1\}^u}$ $\frac{1}{4}\sum_{i,j=1}^{u} w_{ij}(\mathbf{f}_i - \mathbf{f}_j)^2$ *Spectral Clustering* $\text{argmin}_{\mathbf{f} \in \mathcal{R}^u} \frac{1}{2}\mathbf{f}^T L \mathbf{f}$ *Out-of-sample Extn.* $\text{argmin}_{f \in \mathcal{H}_K} \frac{1}{2}\mathbf{f}^T L \mathbf{f}$ *Reg. Spectral Clust.* $\text{argmin}_{f \in \mathcal{H}_K}$ $\frac{1}{2}\mathbf{f}^T L \mathbf{f} + \gamma \|f\|_K^2$

Regularized Spectral Clustering

Unsupervised Manifold Regularization :

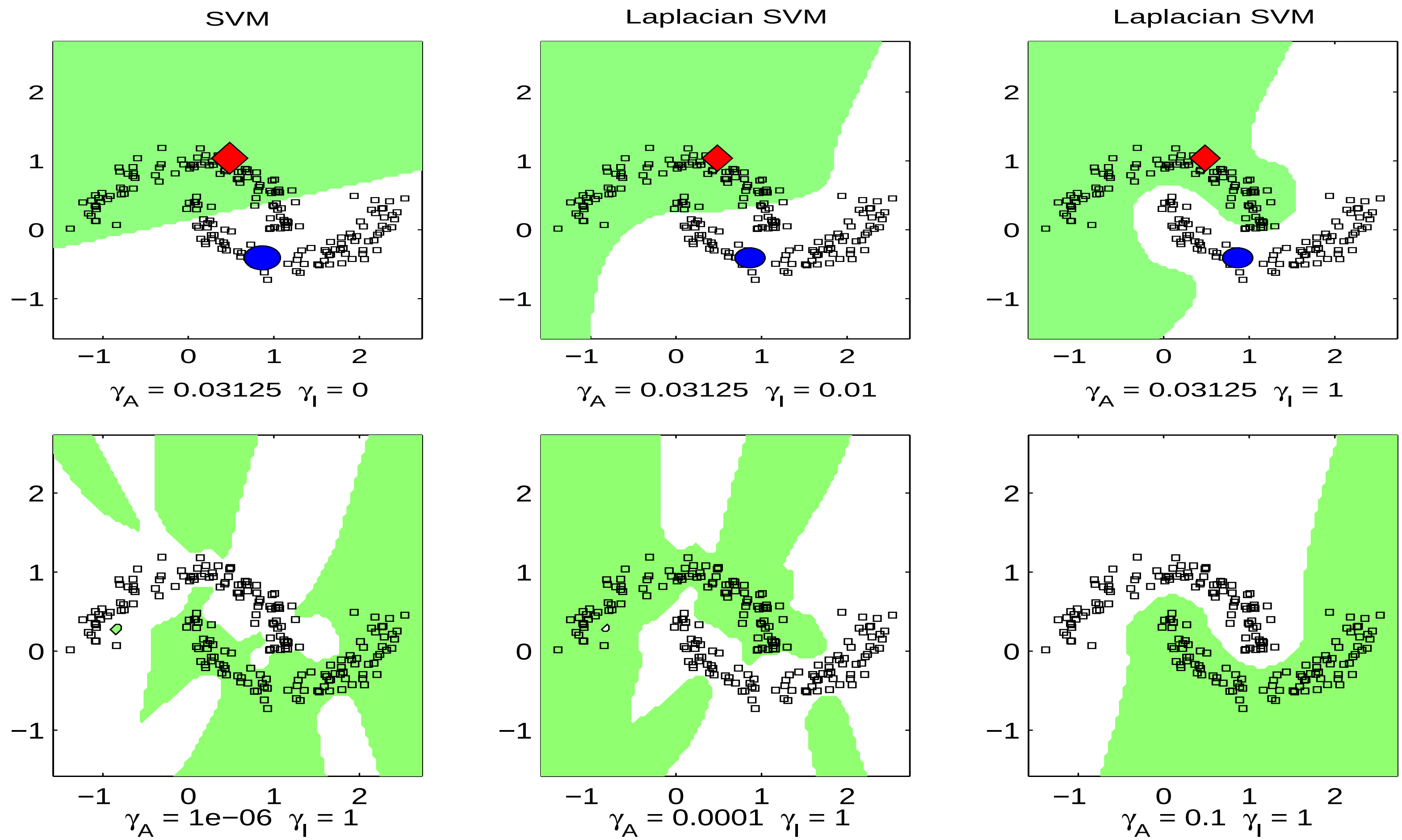
$$f^* = \underset{\substack{\mathbf{1^T f = 0};\ \|\mathbf{f}\|_2^2 = 1 \\ f \in \mathcal{H}_K}}{\operatorname{argmin}} \ \gamma \|f\|_K^2 + \mathbf{f}^T L \mathbf{f}$$

Representer Theorem : $f^*(x) = \sum_{i=1}^{u} \alpha_i^* K(x_i, x)$ leads to an eigenvalue problem :

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

and $\alpha^* = P v^*$. v^* is the smallest-eigenvalue eigenvector; P projects orthogonal to $K\mathbf{1}$.

Experiments : Synthetic



Related Algorithms

- Transductive SVMs [Joachims, Vapnik]

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

- Semi-supervised SVMs [Bennet,Fung et al]

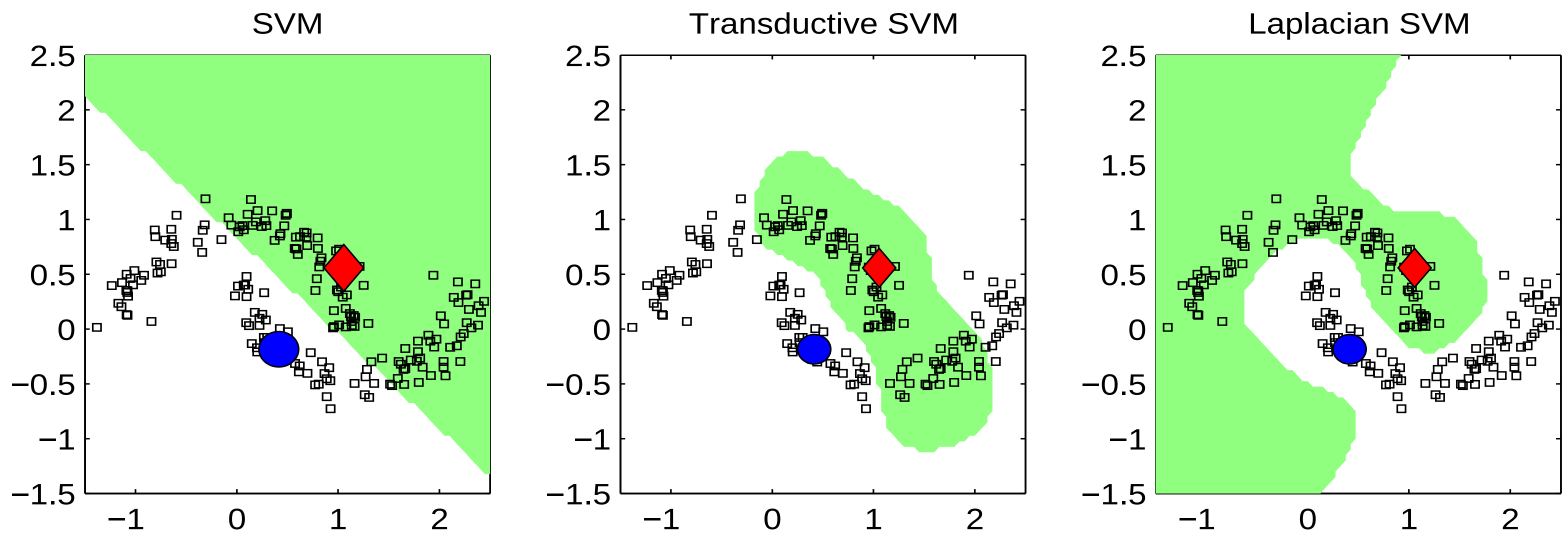
$$f^* = \underset{f \in \mathcal{H}_K, y_{l+1}, \ldots y_{l+u}}{\operatorname{argmin}} C \sum_{i=0}^{l} (1 - y_i f(x_i))_+ +$$

$$\mathrm{C} \sum_{i=l+1}^{l+u} \min \{(1 - f(x_i))_+, (1 + f(x_i))_+\} + \|f\|_K^2$$

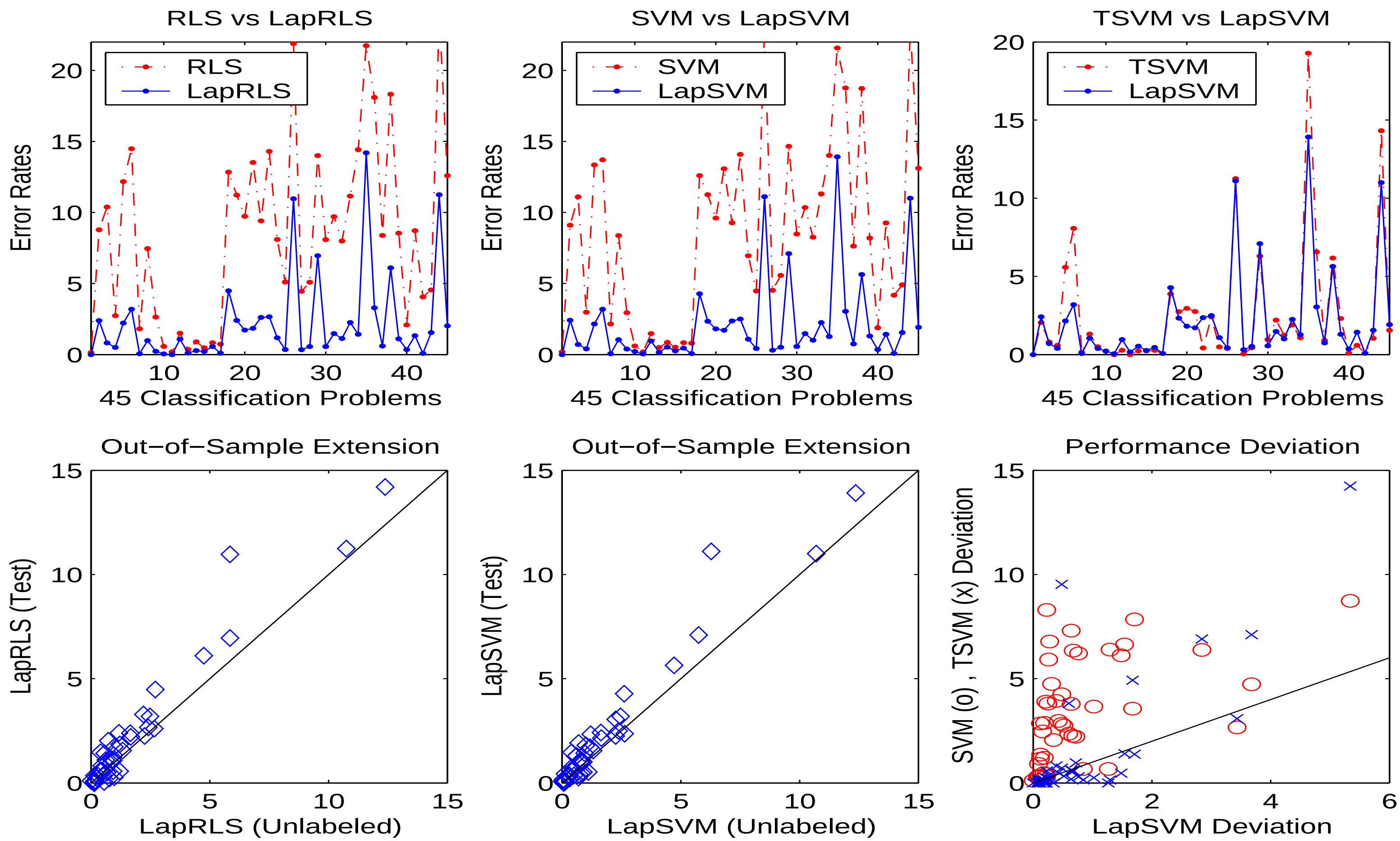
- Measure-based Reg. [Bousquet et al]

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

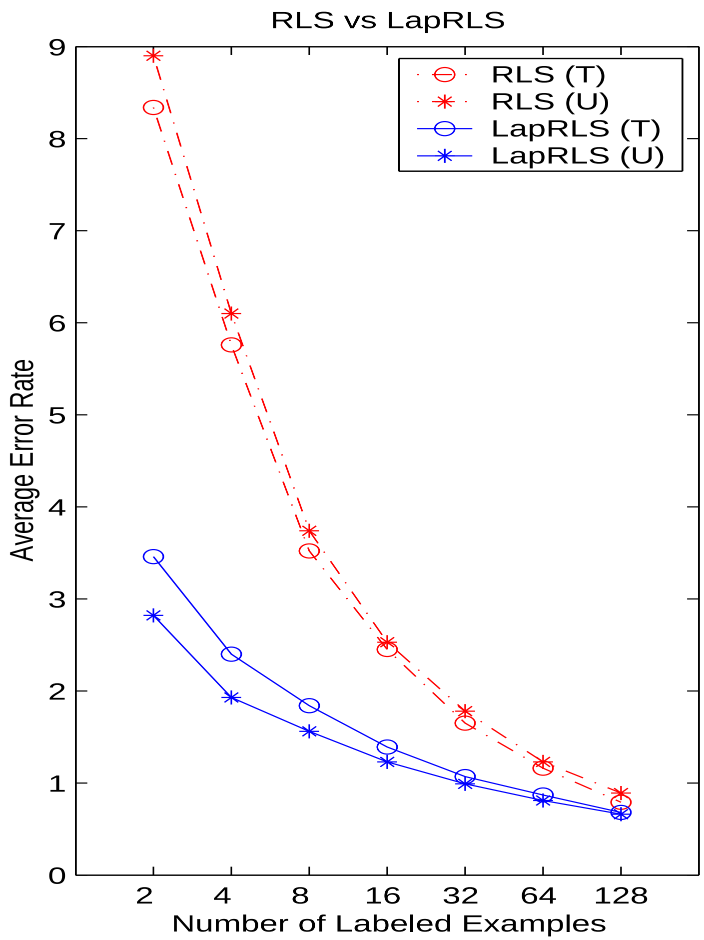
Experiments : Synthetic Data

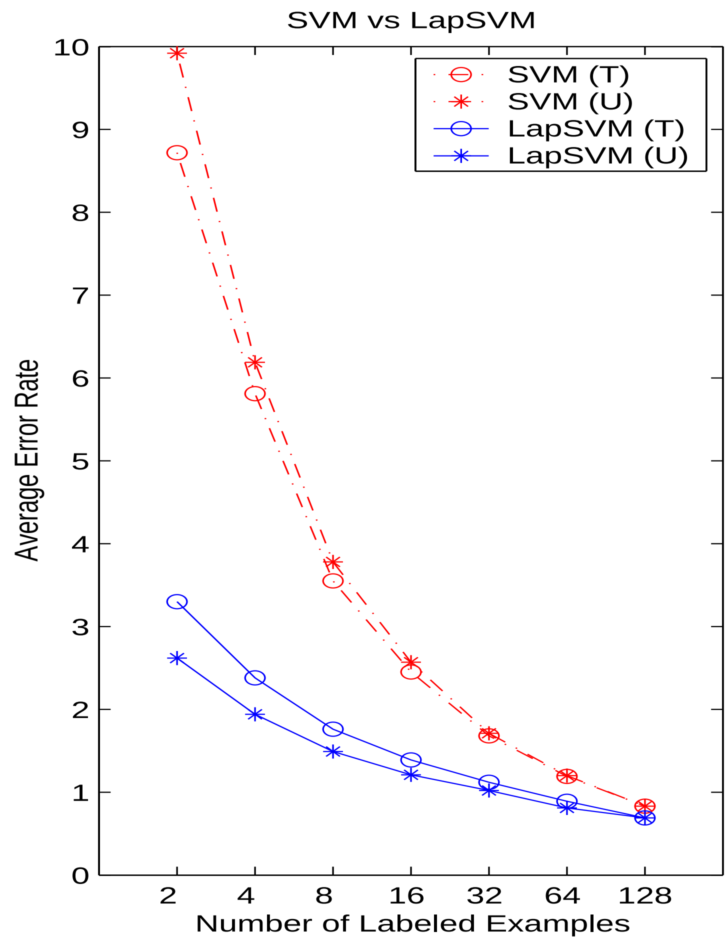


Experiments : Digits

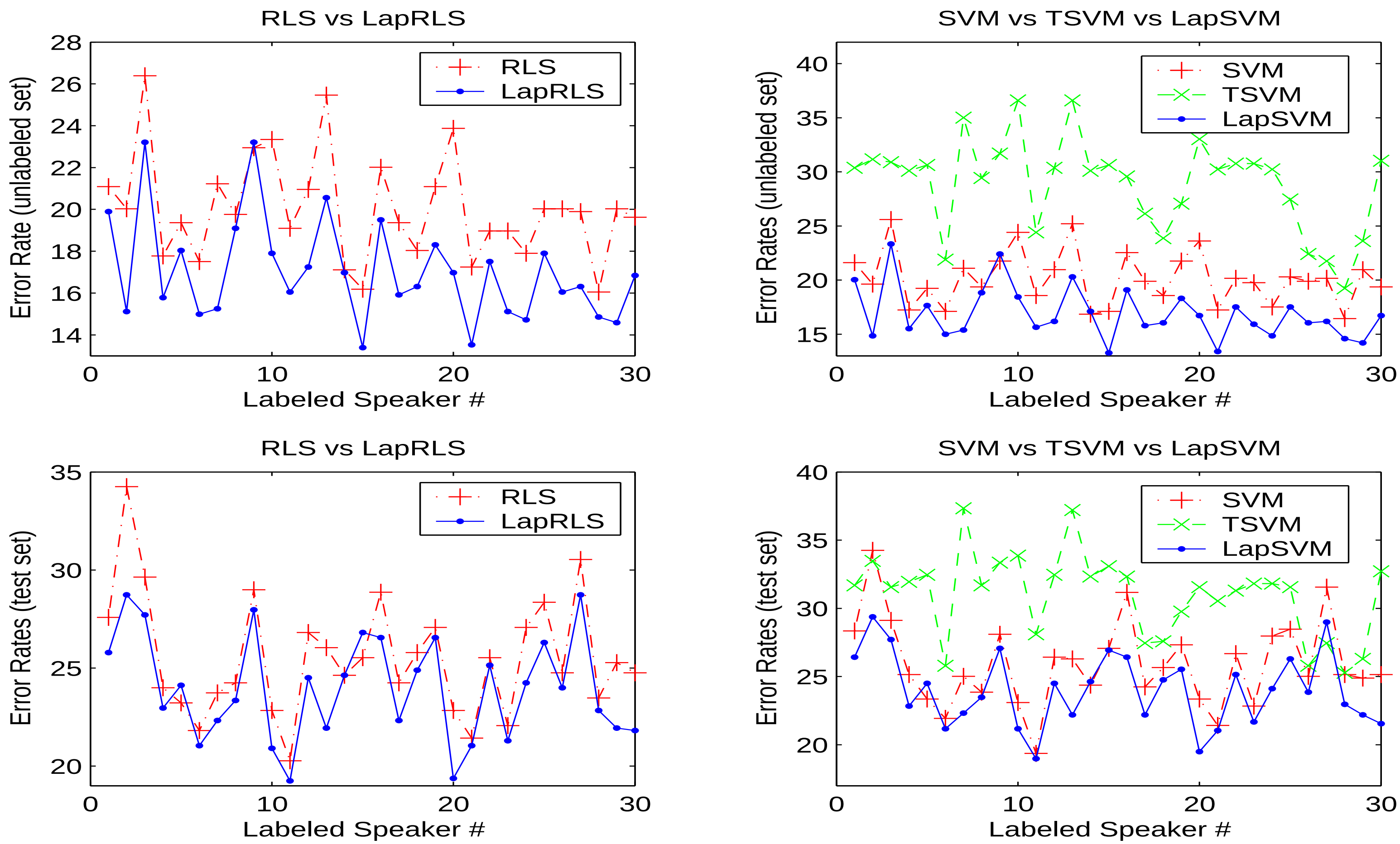


Experiments : Digits

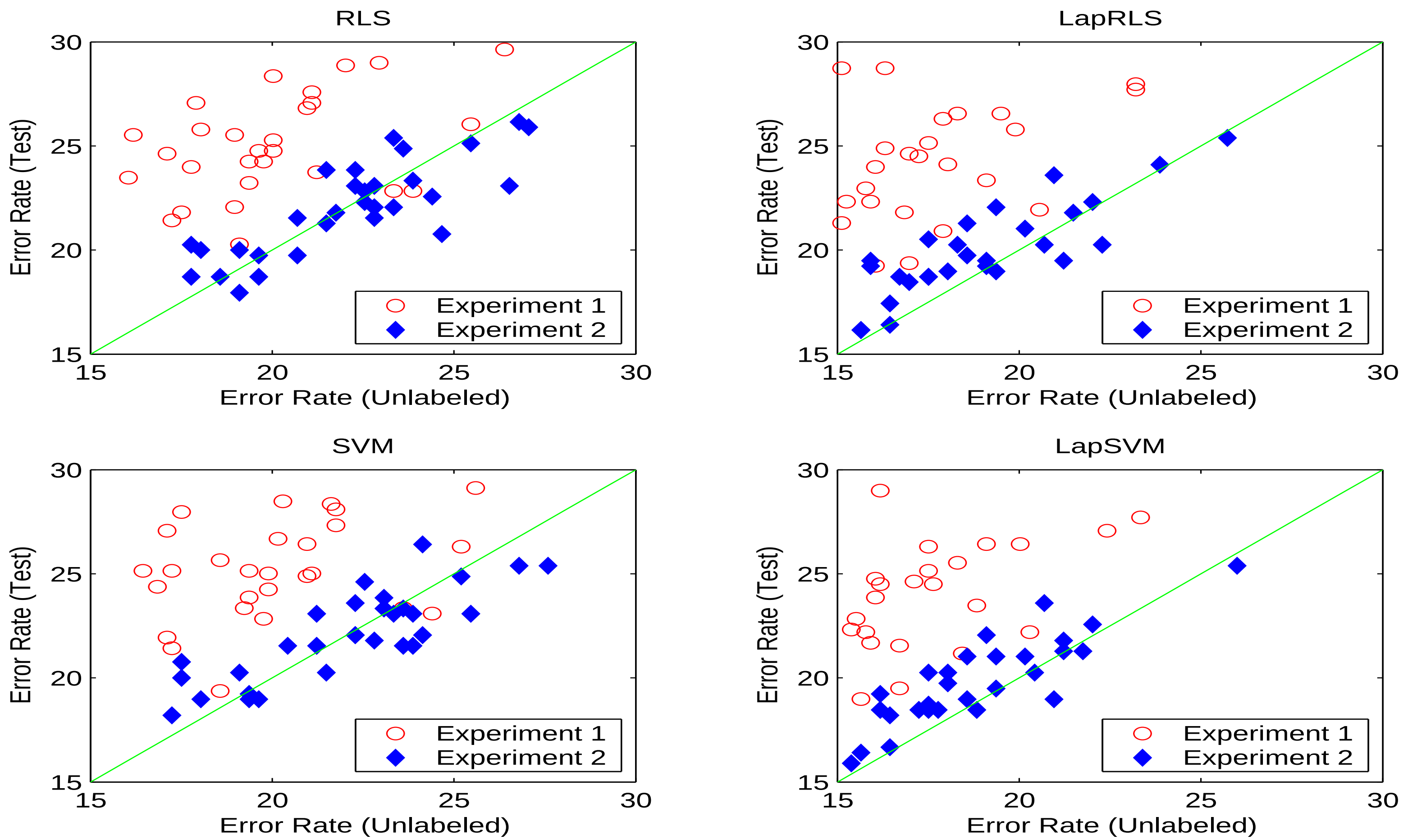




Experiments : Speech



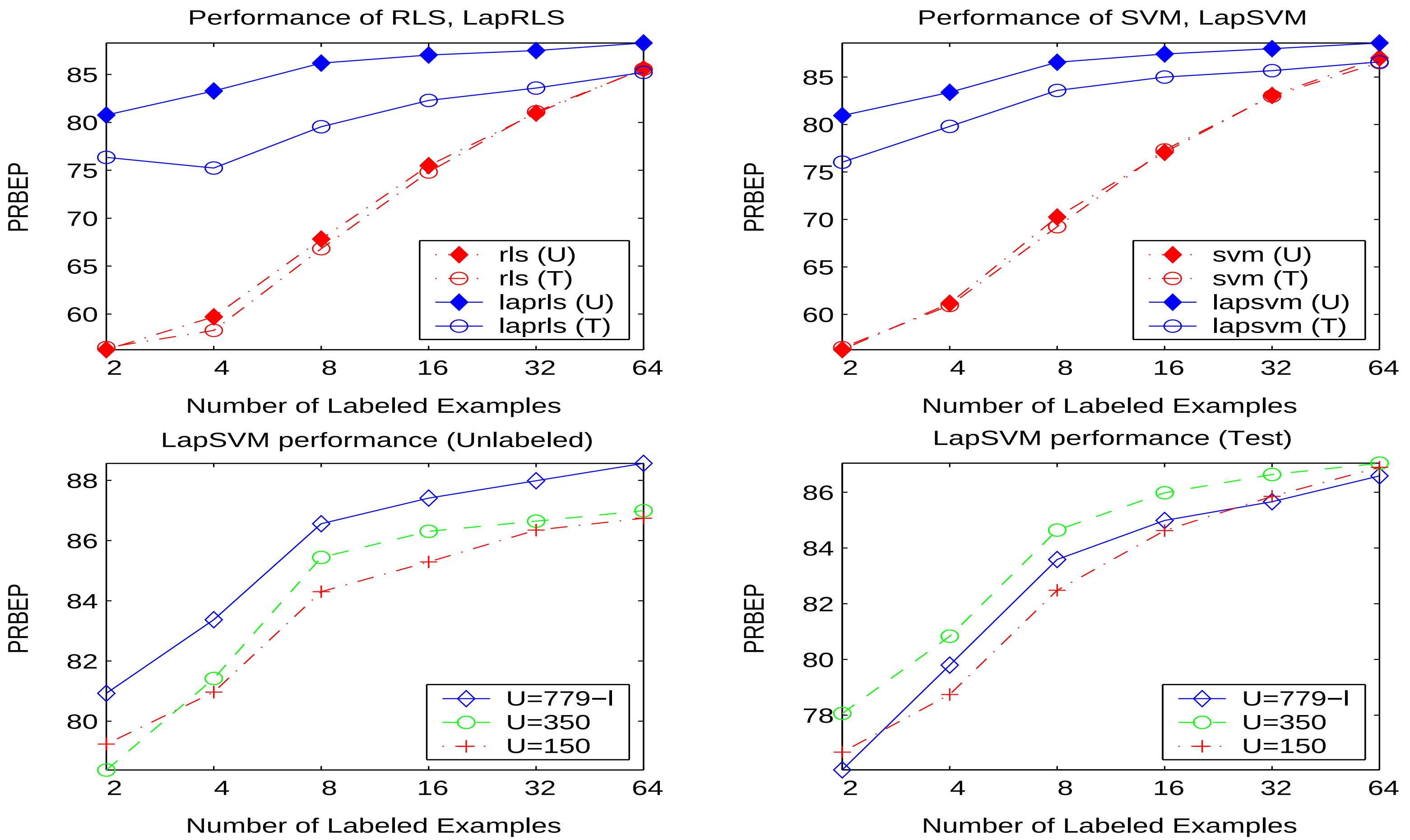
Experiments : Speech



Experiments : Text

Method	PRBEP	Error
k-NN	73.2	13.3
SGT	86.2	6.2
Naive-Bayes	—	12.9
Cotraining	—	6.20
SVM	76.39 (5.6)	10.41 (2.5)
TSVM	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)

Experiments : Text



Future Work

- Generalization as a function of labeled and unlabeled examples.
- Additional Structure : Structured Outputs, Invariances
- Active Learning , Feature Selection
- Efficient Algorithms : Linear Methods, Sparse Solutions
- Applications : Bioinformatics, Text, Speech, Vision, ...