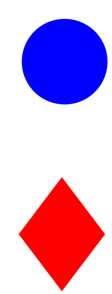
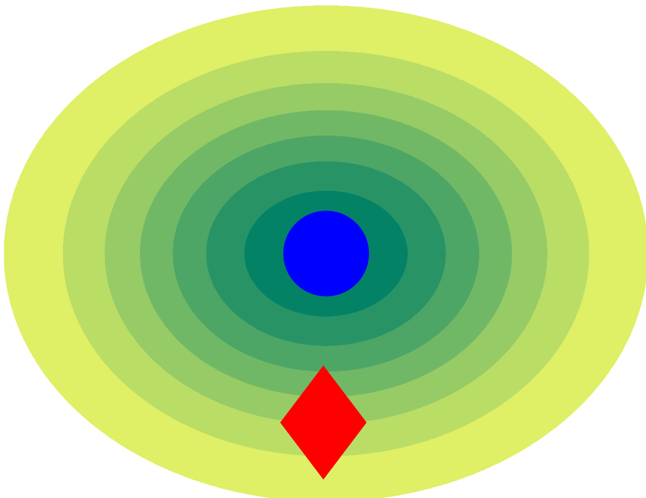
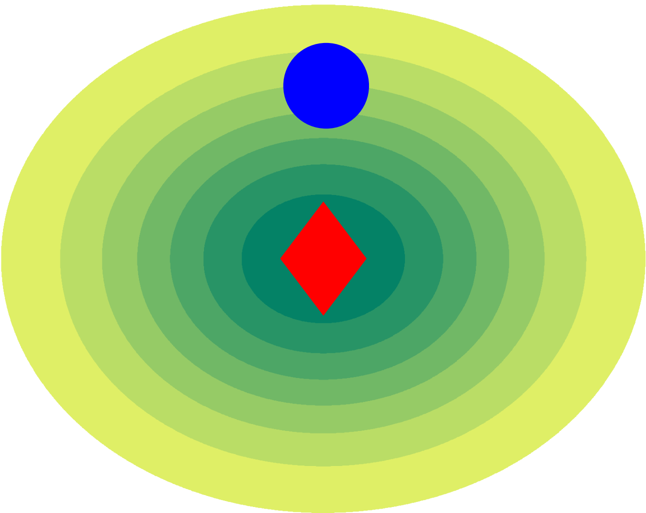
Classify these points using SVMs



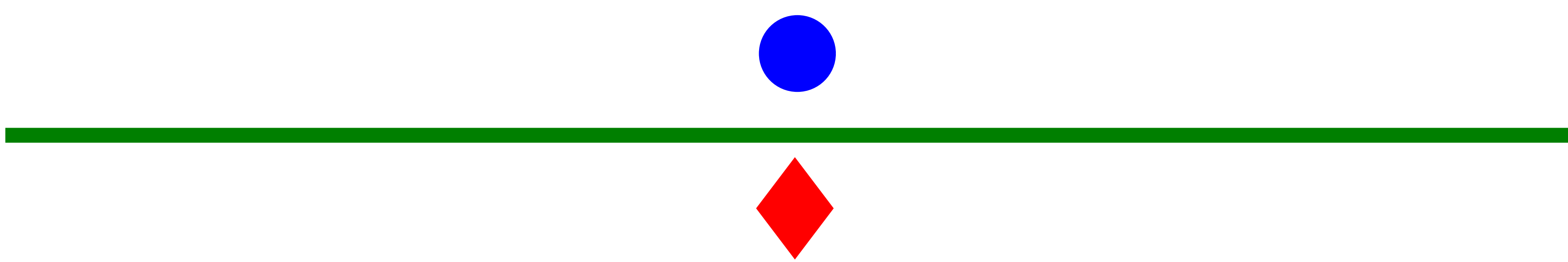
Step 1: Place a Gaussian bump on point 1



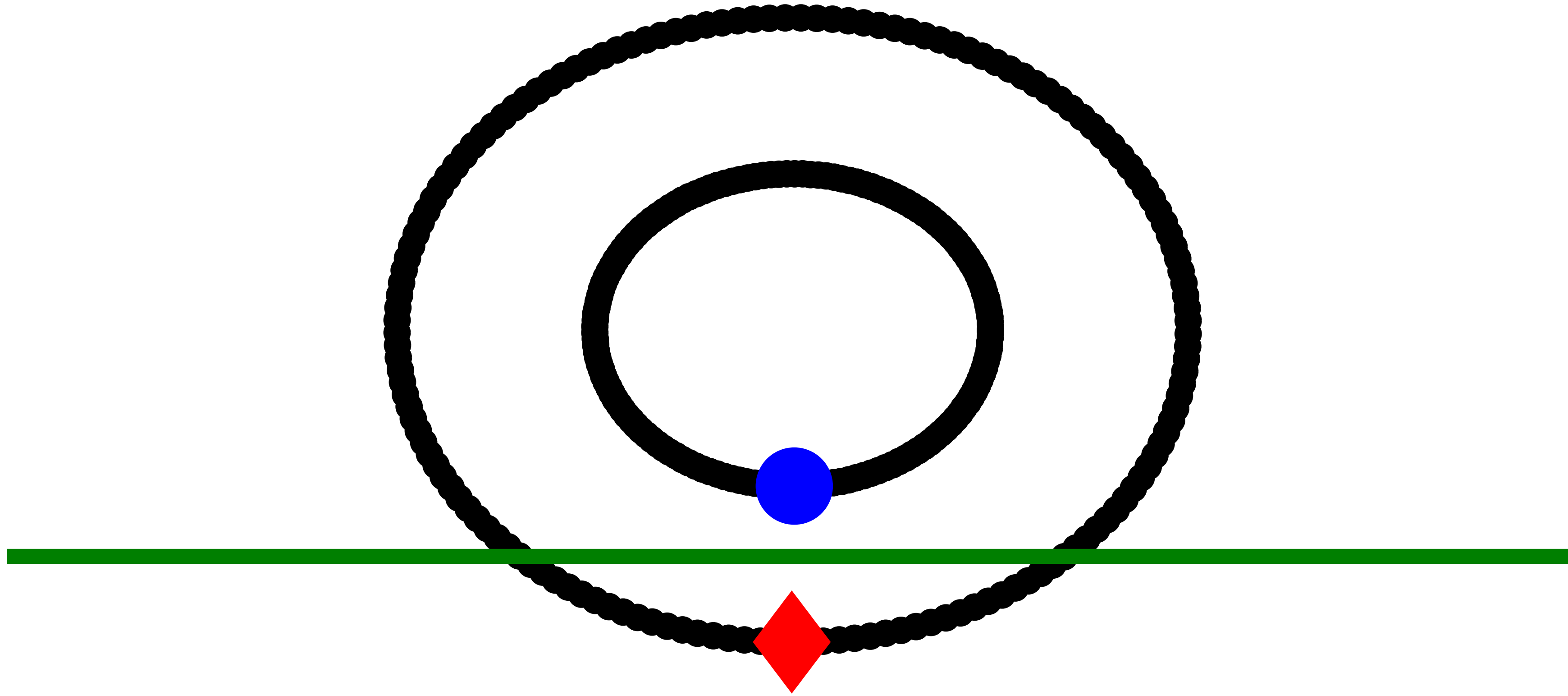
Step 2: Place a Gaussian bump on point 2



Step 3: Run an SVM: $f^*(x) = \alpha_1 e^{-\gamma\|x_1 - x\|^2} + \alpha_2 e^{-\gamma\|x_2 - x\|^2}$

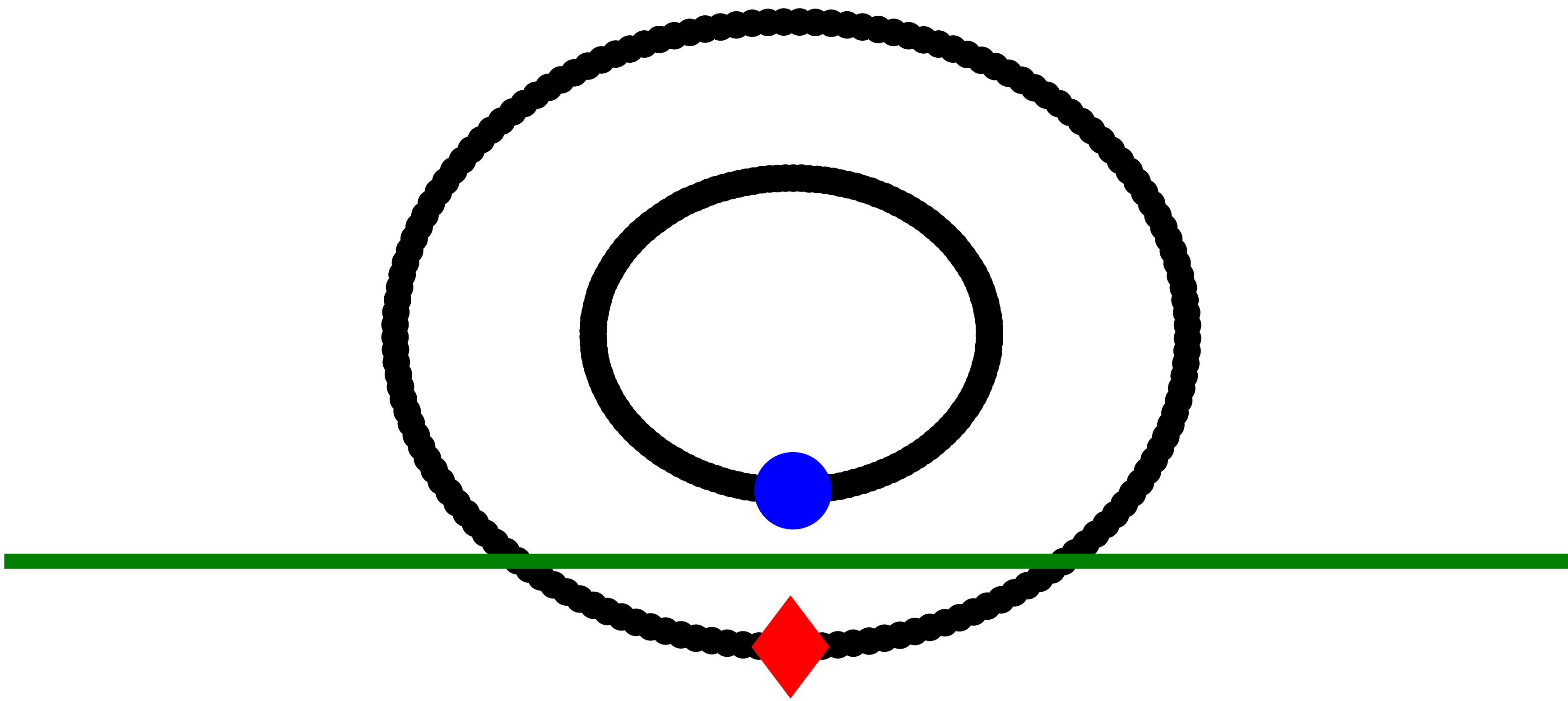


Semi-supervised Learning



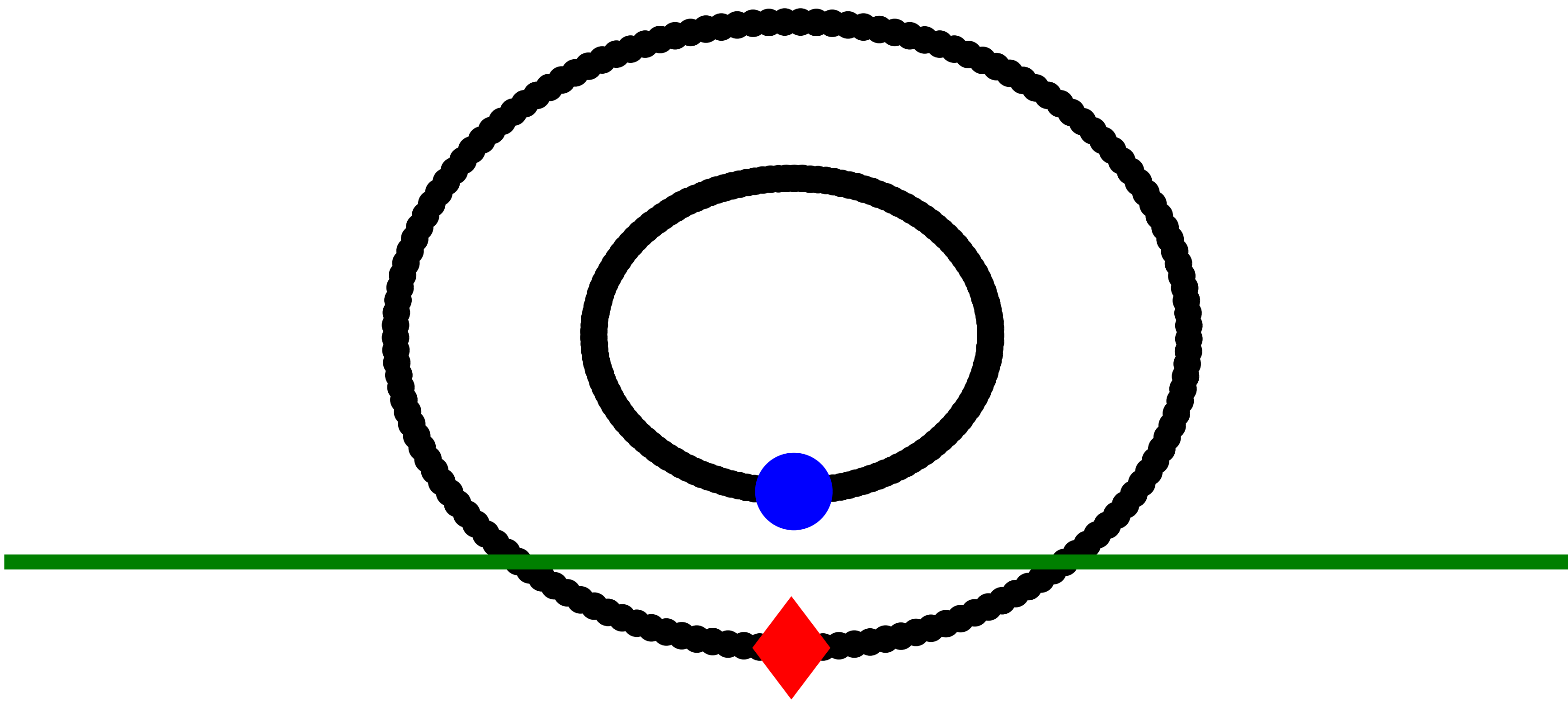
Semi-supervised Learning

- unlabeled data changes our belief.



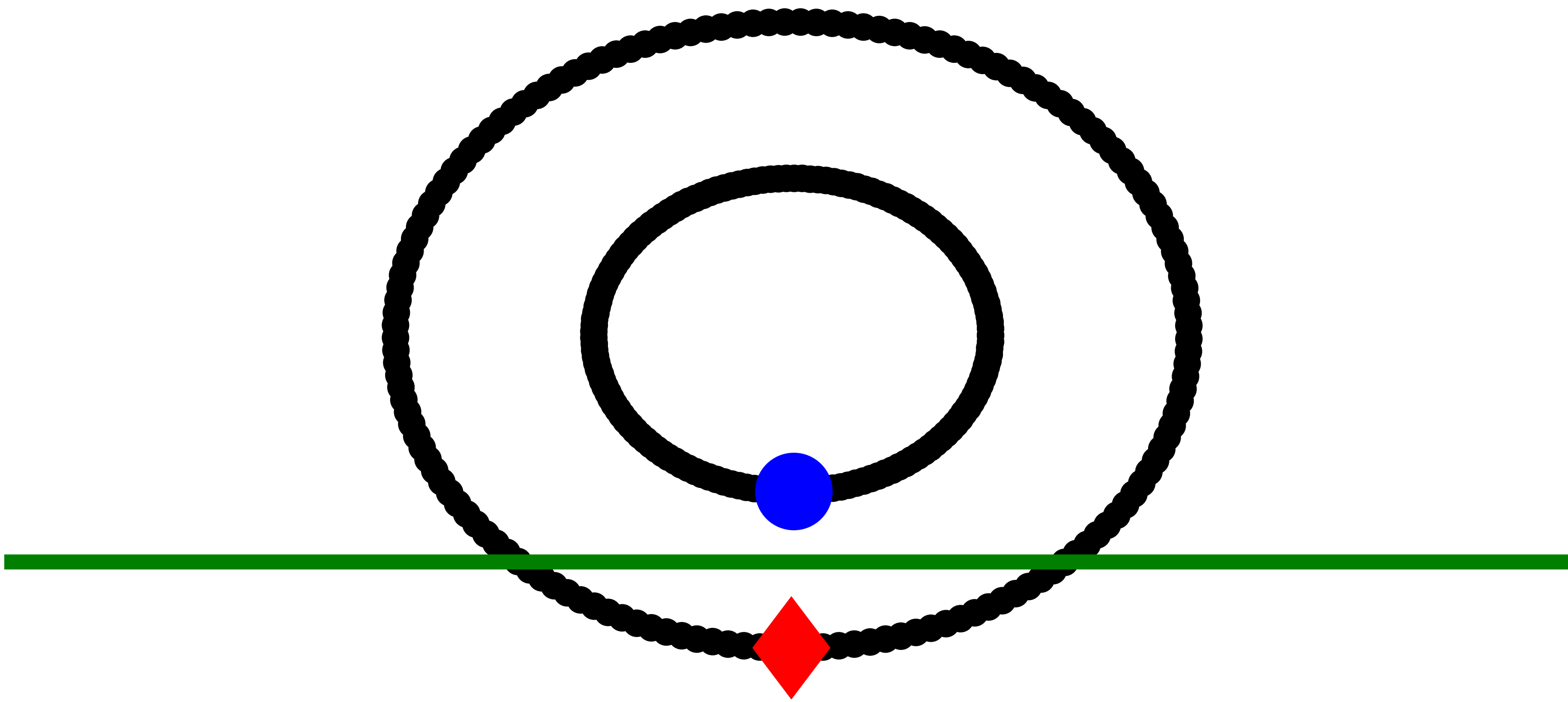
Semi-supervised Learning

- unlabeled data changes our belief.
- operating assumptions: Cluster/Manifold.



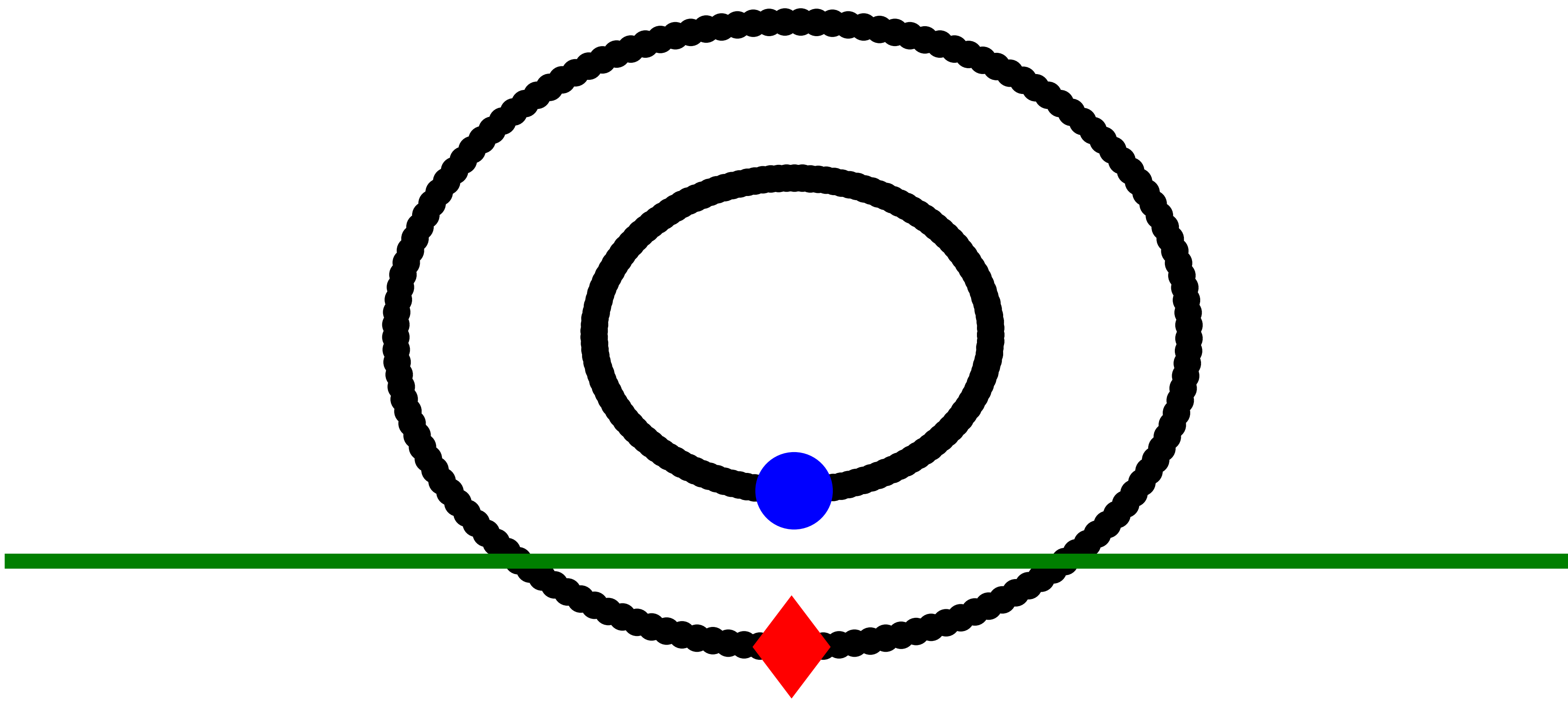
Semi-supervised Learning

- unlabeled data changes our belief.
- operating assumptions: Cluster/Manifold.
- what kind of kernel will work ?



Semi-supervised Learning

- unlabeled data changes our belief.
- what kind of kernel will work ?
- operating assumptions: Cluster/Manifold.
- original space rich enough. complexity ?



Question

Can we define a kernel $\tilde{k}$ that is adapted to the geometry of the data?

Question

Can we define a kernel $\tilde{k}$ that is adapted to the geometry of the data?

Think of a continuous symmetric, positive-definite function $\tilde{k}$ so that

$$f^*(x) = \beta_1 \tilde{k}(x_1, x) + \beta_2 \tilde{k}(x_2, x)$$

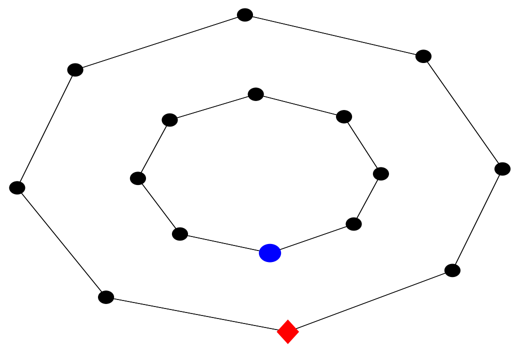
gives a circular decision surface

Transductive Versus Semi-supervised Learning

Transductive Learning:

- Data is a Point Cloud V. • Model Geometry as a graph.
- Defi ne a Graph kernel $k_G : V \times V \mapsto \mathcal{R}$. • Learn a function $f : V \mapsto \mathcal{R}$.

Problem: Out-of-Sample Extension

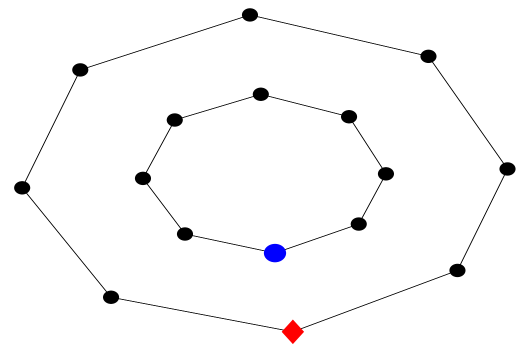


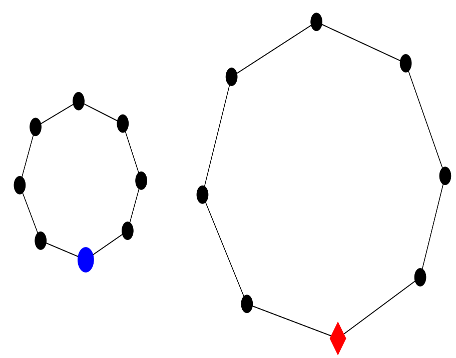
Transductive Versus Semi-supervised Learning

Transductive Learning:

- Data is a Point Cloud V. • Model Geometry as a graph.
- Define a Graph kernel $k_G : V \times V \mapsto \mathcal{R}$. • Learn a function $f : V \mapsto \mathcal{R}$.

Problem: Out-of-Sample Extension



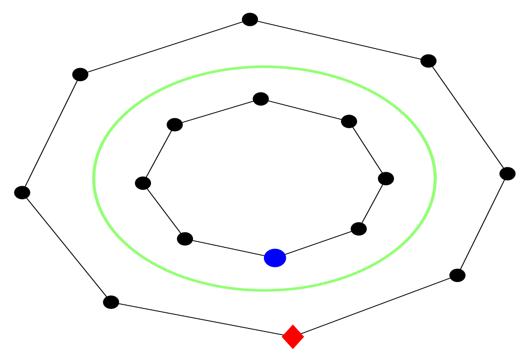


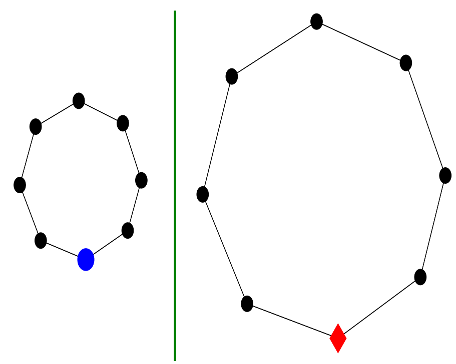
Same Picture for Transduction.

Transductive Versus Semi-supervised Learning

Semi-supervised Learning:

- Data is a Point Cloud V in an *ambient space*. • Model Geometry as a graph.
- Defi ne an ambient kernel $\tilde{k} : X \times X \mapsto \mathcal{R}$. • Learn a function $f : X \mapsto \mathcal{R}$.





Beyond the Point Cloud: from Transductive to Semi-supervised Learning

Vikas Sindhwani, Partha Niyogi, Mikhail Belkin

Department of Computer Science

University of Chicago

Program

Warping an RKHS

for Semi-supervised Learning.

Before Observing Unlabeled Data

Choose a Kernel (encodes some form of prior knowledge)

$$k(x,z) : X \times X \mapsto \mathcal{R}$$

- space of functions: $f \in \mathcal{H} : X \mapsto \mathcal{R}$
- inner product on that space: $\langle f, g \rangle_{\mathcal{H}}$
- complexity measure: $\|f\|_{\mathcal{H}}$

After Observing Unlabeled Data

Data $\{x_i\}_{i=1}^{n}$ (drawn from some unknown distribution) alters our complexity beliefs.

- Data dependent map $S : \mathcal{H} \mapsto \mathcal{V}$
- inner product on $\mathcal{V}$: $\langle ., . \rangle_{\mathcal{V}}$
- complexity measure: $\|Sf\|_{\mathcal{V}}$

Warping an RKHS

Construct $\tilde{\mathcal{H}}$ by warping $\mathcal{H}$:

$\tilde{\mathcal{H}}$ has same functions: $\tilde{\mathcal{H}} = \{f \in \mathcal{H}\}$

But modified inner product:

$$\langle f, g \rangle_{\tilde{\mathcal{H}}} = \langle f, g \rangle_{\mathcal{H}} + \langle Sf, Sg \rangle_{\mathcal{V}}$$

And a data-refined notion of complexity:

$$\|f\|^2_{\tilde{\mathcal{H}}} = \|f\|^2_{\mathcal{H}} + \|Sf\|^2_{\mathcal{V}}$$

Warping an RKHS

If S is a bdd linear operator $\|Sf\|_{\mathcal{V}} \leq M\|f\|_{\mathcal{H}}$

- The two norms are compatible since:

$$\|f\|^2_{\mathcal{H}} \leq \|f\|^2_{\tilde{\mathcal{H}}} \leq (M+1)\|f\|^2_{\mathcal{H}}$$

$\therefore \tilde{\mathcal{H}}$ is a Hilbert Space.

- Evaluation functionals on $\tilde{\mathcal{H}}$ are bounded:

$$f(x) \leq C\|f\|_{\mathcal{H}} \implies f(x) \leq C\|f\|_{\tilde{\mathcal{H}}}$$

$\therefore \tilde{\mathcal{H}}$ is a (random) RKHS; kernel $\tilde{k} : X \times X \mapsto \mathcal{R}$

Warping an RKHS

What is the kernel $\tilde{k}(x, y)$ associated with the warped RKHS $\tilde{\mathcal{H}}$?

Can explicitly compute, for evaluation maps:

$$S : \mathcal{H} \mapsto \mathcal{R}^n \qquad Sf = [f(x_1) \ldots f(x_n)] = \mathbf{f}$$

with a Point Cloud semi-norm:

$$\|f\|^2_{\mathcal{V}} = \mathbf{f}^T M \mathbf{f} \qquad M \succeq 0 \quad \text{symmetric}$$

Data-deformed Kernel

Reproducing Property in $\mathcal{H}$: $f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}}$

Reproducing Property in $\tilde{\mathcal{H}}$: $f(x) = \langle f, \tilde{k}(x,\cdot)\rangle_{\tilde{\mathcal{H}}}$

Decompose: $\mathcal{H} = span\,\{k(x_i,\cdot)\}_{i=1}^{n} \oplus \mathcal{H}^{\perp}$

$\forall f \in \mathcal{H}^{\perp}\ \ \langle f, k_{x_i}\rangle_{\mathcal{H}} = f(x_i) = 0$

S only deforms the span

$\therefore Sf = 0 \implies f(x) = \langle f, \tilde{k}(x,\cdot)\rangle_{\mathcal{H}}$

$\therefore \langle f, k(x,\cdot) - \tilde{k}(x,\cdot)\rangle_{\mathcal{H}} = 0$

$\implies k(x,\cdot) - \tilde{k}(x,\cdot) \in span\,\{k(x_i,\cdot)\}_{i=1}^{n}$

Data-deformed Kernel

So, $$\tilde{k}(x,\cdot) = k(x,\cdot) + \sum_{j=1}^{n} \beta_j(x) k(x_j,\cdot)$$

Find $\beta(x) = [\beta_1(x) \ldots \beta_n(x)]$ by solving a linear system: $k(x_i, x) = \langle k(x_i, .), \tilde{k}(x,\cdot)\rangle_{\tilde{\mathcal{H}}}$

$$= \langle k(x_i, .), k(x,\cdot) + \textstyle\sum_j \beta_j(x) k(x_j,\cdot)\rangle_{\tilde{\mathcal{H}}}$$

$$= \langle k(x_i, .), k(x,\cdot) + \textstyle\sum_j \beta_j(x) k(x_j,\cdot)\rangle_{\mathcal{H}} + \mathbf{k}_{\mathbf{x_i}}{}^t M \mathbf{g}$$

$\mathbf{k}_{\mathbf{x_i}k} = k(x_i, x_k)$

$\mathbf{g}_k = k(x, x_k) + \sum_j \beta_j(x) k(x_j, x_k)$

Kernel of the Warped RKHS

Solve:

$$(K + KMK)\beta(x) = -KM\mathbf{k}_x$$

Kernel of $\tilde{H}$:

$$\tilde{k}(x,z) = k(x,z) - \mathbf{k}_x^t(I + MK)^{-1}M\mathbf{k}_z$$

where $\mathbf{k}_x = [k(x_1,x)...k(x_n,x)]$
and K is the gram matrix of $k(.,.)$ over $\{x_i\}_{i=1}^n$.

Choosing M for SSL

- Construct a Graph W.
- Compute Laplacian of the Point Cloud.

$$L = D - W \quad where \quad D_{ii} = \sum_j W_{ij}$$

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

Other Choices: L^p , $r(L) = \sum_{i=1}^{n} r(\lambda_i) v_i v_i^T$

Algorithms

Laplacian RLS, Laplacian SVM:

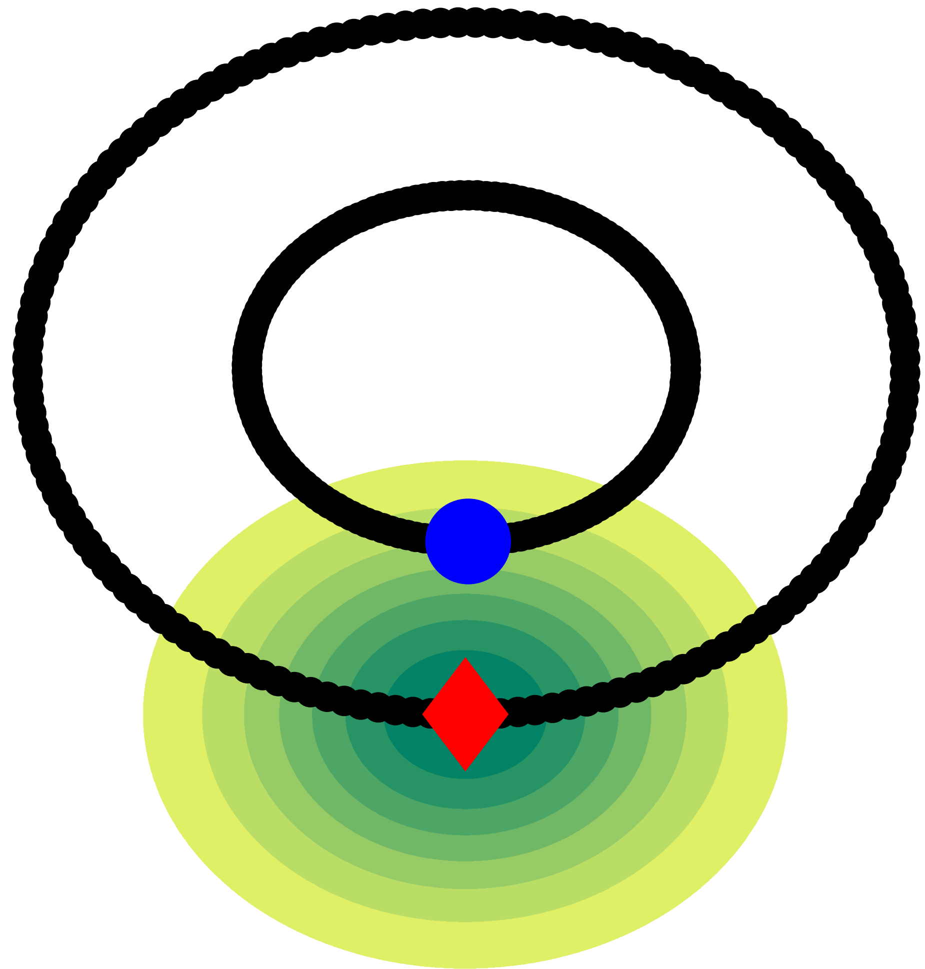
$$f^* = \underset{\tilde{\mathcal{H}}}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_{\tilde{k}}^2$$

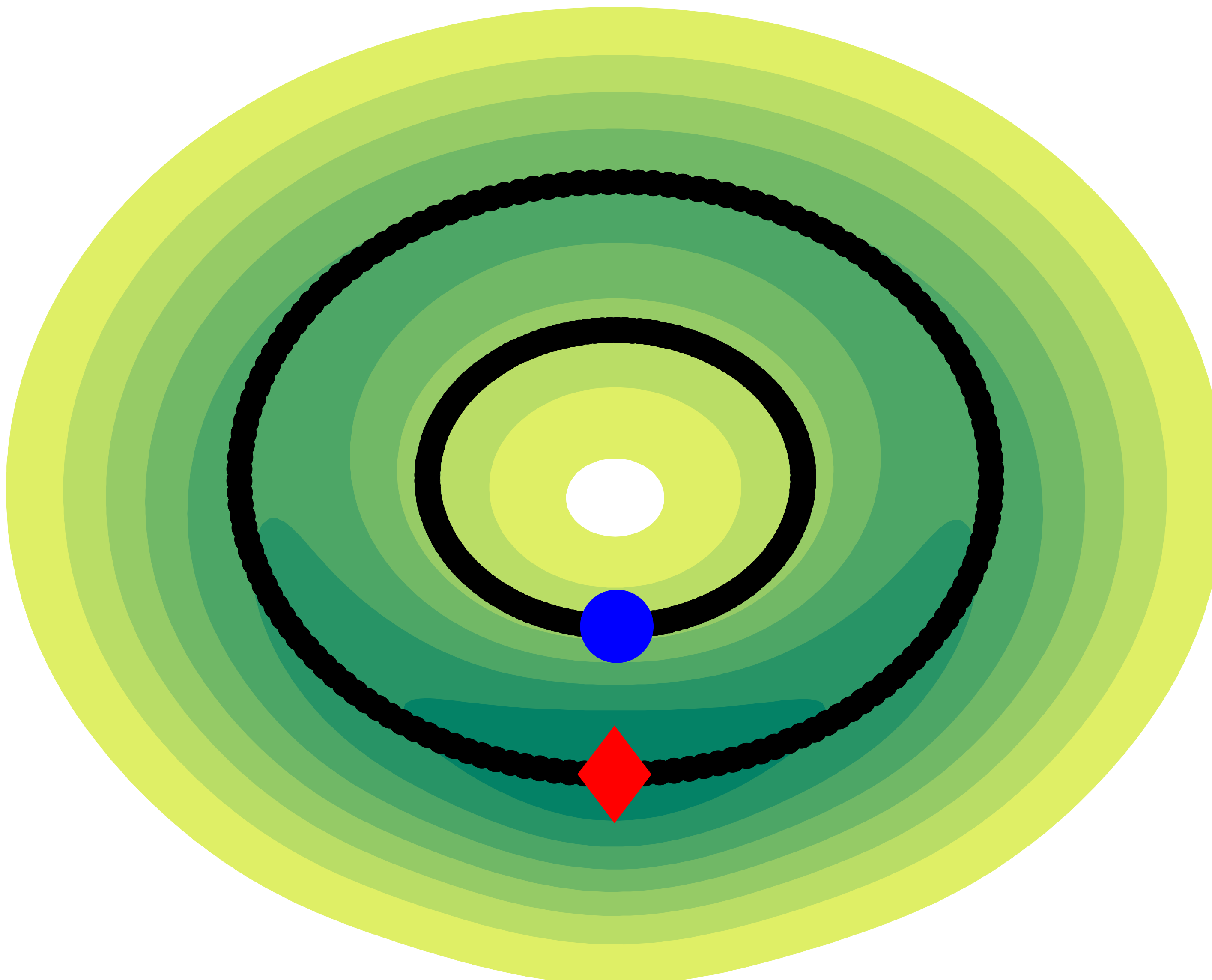
Setting $M = \frac{\gamma_I}{\gamma_A} L$ we can re-interpret Manifold Regularization (Belkin, Niyogi, Sindhwani 2004) algorithms as standard kernel methods in this warped, random RKHS $\tilde{\mathcal{H}}$. Think of the ratio $\frac{\gamma_I}{\gamma_A}$ as the strength of deformation

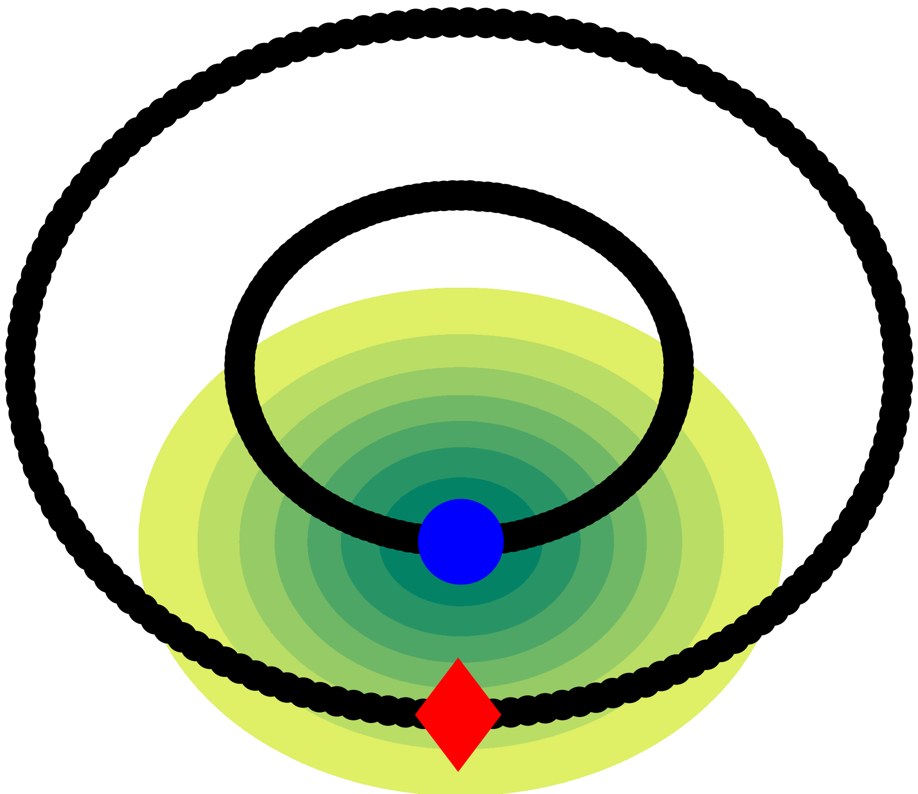
Algorithms

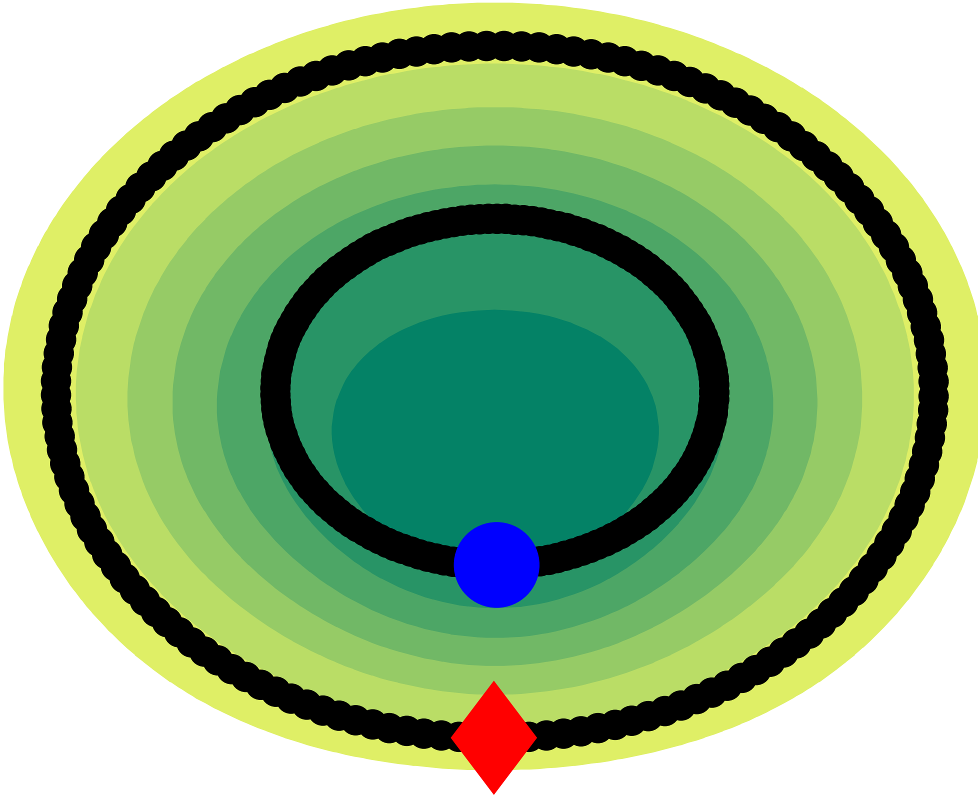
Other possibilities? :

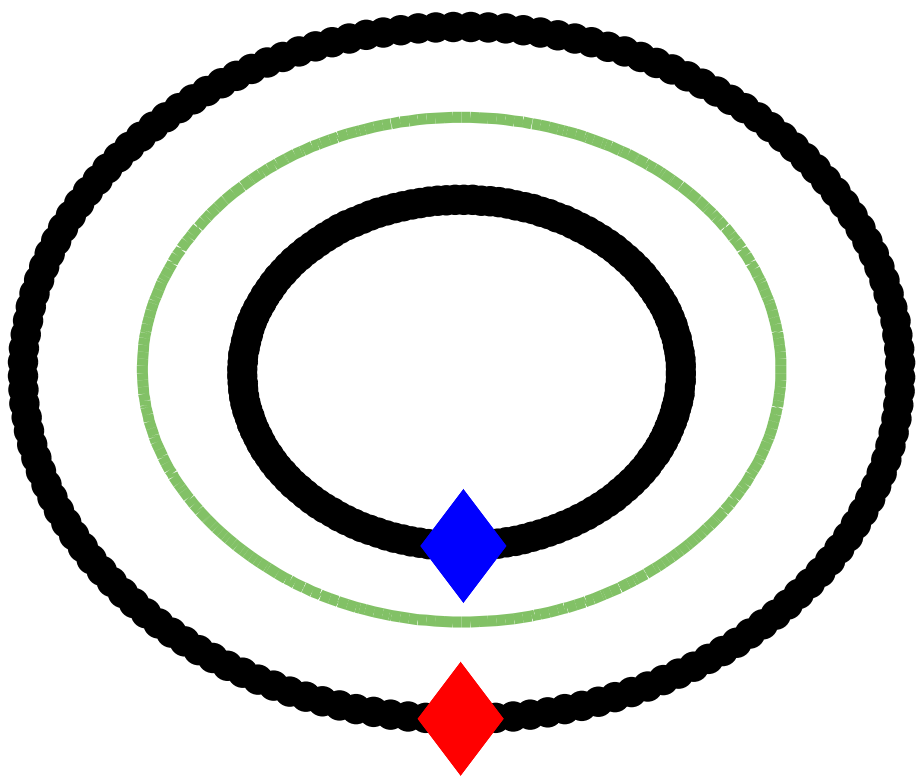
Laplacian SVR, One-class SVMs ...











Datasets

Dataset	c	d	l	n	k
g50c	2	50	50	550	Gaussian
Coil20	20	1024	40	1440	Gaussian
Uspst	10	256	50	2007	Gaussian
mac-windows	2	7511	50	1946	Gaussian
Webkb (page)	2	3000	12	1051	linear
Webkb (link)	2	1840	12	1051	linear
Webkb (page+link)	2	4840	12	1051	linear

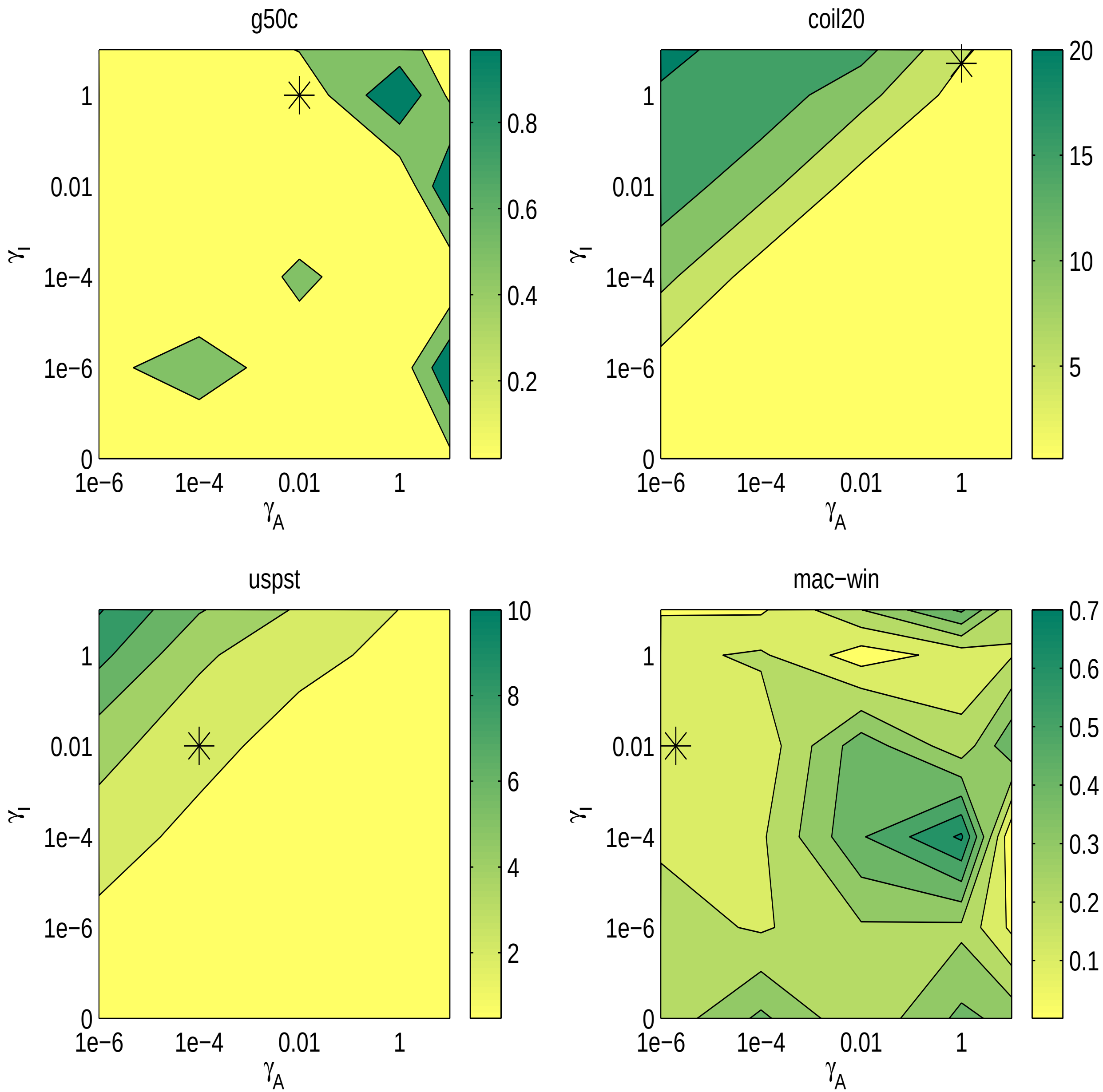
Parameters:
All datasets except Webkb: $\gamma_A = 10^{-6}, \gamma_I = 0.01$ held fixed,
$nn = \{10, 50, 100, 200\}, \sigma = \sigma_0 * [0.25\ 0.5\ 1\ 2\ 4], p = \{1, 2, ..., 5\}$ tuned by 5-fold CV.
For WebKB, $nn = 200, p = 5$, γ_A, γ_I optimized for best performance over the unlabeled data.

Experiments: Performance on $n-l$ unlabeled examples

Dataset → Algorithm ↓	g50c	Coil20	Uspst	mac-win	WebKB (link)	W ebKB (page)	WebKB (page+link)
SVM (n)	4.0	0.0	2.8	2.4	5.1	5.3	0.7
RLS (n)	4.0	0.0	2.5	2.8	5.6	6.4	2.2
SVM (l)	9.7	24.6	23.6	18.9	28.1	24.3	18.2
RLS (l)	8.5	26.0	23.6	18.8	30.3	30.2	23.9
Graph-Trans	17.3	6.2	21.3	11.7	22.0	10.7	6.6
TSVM	6.9	26.3	26.5	7.4	14.5	8.6	7.8
Graph-density	8.3	6.4	16.9	10.5	-	-	-
∇TSVM	5.8	17.6	17.6	5.7	-	-	-
LDS	5.6	4.9	15.8	5.1	-	-	-
LapSVM	5.4	4.0	12.7	10.4	17.2	10.9	6.4
LapRLS	5.2	4.3	12.7	10.0	19.2	11.2	7.5

http://www.cs.uchicago.edu/∼vikass/research.html

Experiments: Out of Sample Extension (4-fold CV variant)



Contribution

- Discussed a procedure for warping an RKHS for Semi-supervised Learning.

- Derived a Kernel for SSL.
Turns transductive and supervised methods into Semi-supervised Learners.

- Demonstrates good performance in both Transductive and Semi-supervised settings (out of sample extension).