

Deterministic Annealing for Semi-supervised Kernel Machines

Vikas Sindhwani¹, Sathiya Keerthi², Olivier Chapelle³

¹University of Chicago

²Yahoo! Research

³Max Planck Institute, Tübingen

ICML 2006

V. Vapnik's *Transductive* SVM idea

Suppose, for a binary classification problem, we have

- l labeled examples $\{\mathbf{x}_i, y_i\}_{i=1}^l$, $\mathbf{x}_i \in \mathcal{X}, y_i \in \{-1, +1\}$
- u unlabeled examples $\{\mathbf{x}'_j\}_{j=1}^u$

Denote $\mathbf{y}' = (y'_1 \dots y'_u)$ as the unknown labels.

Train an SVM while optimizing unknown labels

Solve, over $f \in \mathcal{H}_K : \mathcal{X} \rightarrow \mathcal{R}$ and $\mathbf{y}' \in \{-1, +1\}^u$,

$$\min_{f, \mathbf{y}'} \underbrace{\frac{\lambda}{2} \|f\|_K^2}_{\text{regularizer}} + \underbrace{\frac{1}{l} \sum_{i=1}^l V(y_i, f(\mathbf{x}_i))}_{\text{labeled loss}} + \underbrace{\frac{\lambda'}{u} \sum_{j=1}^u V(y'_j, f(\mathbf{x}'_j))}_{\text{unlabeled loss}}$$

standard SVM

subject to: $\frac{1}{u} \sum_{j=1}^u \max(0, y'_j) = r$ (positive class ratio)

V. Vapnik's *Transductive* SVM idea

Suppose, for a binary classification problem, we have

- l labeled examples $\{\mathbf{x}_i, y_i\}_{i=1}^l$, $\mathbf{x}_i \in \mathcal{X}, y_i \in \{-1, +1\}$
- u unlabeled examples $\{\mathbf{x}'_j\}_{j=1}^u$

Denote $\mathbf{y}' = (y'_1 \dots y'_u)$ as the unknown labels.

Train an SVM while optimizing unknown labels

Solve, over $f \in \mathcal{H}_K : \mathcal{X} \rightarrow \mathcal{R}$ and $\mathbf{y}' \in \{-1, +1\}^u$,

$$\min_{f, \mathbf{y}'} \underbrace{\frac{\lambda}{2} \|f\|_K^2}_{\text{regularizer}} + \underbrace{\frac{1}{l} \sum_{i=1}^l V(y_i, f(\mathbf{x}_i))}_{\text{labeled loss}} + \underbrace{\frac{\lambda'}{u} \sum_{j=1}^u V(y'_j, f(\mathbf{x}'_j))}_{\text{unlabeled loss}}$$

standard SVM

subject to: $\frac{1}{u} \sum_{j=1}^u \max(0, y'_j) = r$ (positive class ratio)

Equivalent Continuous Optimization Problem

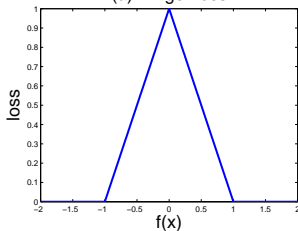
Optimization Problem

$$\min_{f, \mathbf{y}'} \mathcal{J}(f, \mathbf{y}') = \frac{\lambda}{2} \|f\|_K^2 + \frac{1}{l} \sum_{i=1}^l V(y_i, f(\mathbf{x}_i)) + \frac{\lambda'}{u} \sum_{j=1}^u V(\mathbf{y}'_j, f(\mathbf{x}'_j))$$

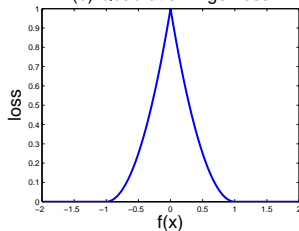
$$\min_f \mathcal{J}(f) = \frac{\lambda}{2} \|f\|_K^2 + \frac{1}{l} \sum_{i=1}^l V(y_i, f(\mathbf{x}_i)) + \frac{\lambda'}{u} \sum_{j=1}^u \underbrace{\min \left[V(+1, f(\mathbf{x}'_j)), V(-1, f(\mathbf{x}'_j)) \right]}_{\text{effective loss } V'(f(\mathbf{x}'_j))}$$

Effective Loss Function Over Unlabeled Examples

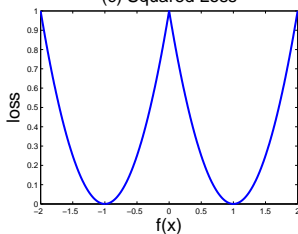
(a) Hinge Loss



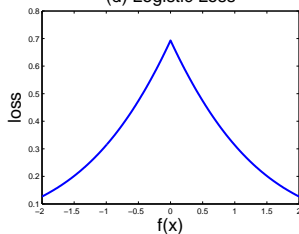
(b) Quadratic Hinge Loss



(c) Squared Loss

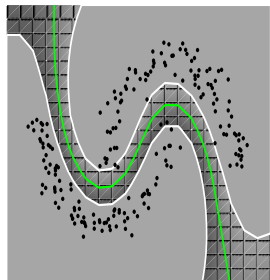


(d) Logistic Loss



- Non-convex
- Penalty if decision surface gets too close to unlabeled examples.

This idea implements a common assumption for SSL...



Low-Density Separation Assumption

The true decision boundary passes through a region containing low volumes of data. Implements the prior knowledge/assumption that

$$\int_{B(f)} P(\mathbf{x}) d\mathbf{x} \quad \text{is small}$$

where $B(f) = \{\mathbf{x} : |f(\mathbf{x})| < 1\}$

Cluster Assumption

Points in a data cluster belong to the same class.

Solution Strategies

JTSVM [Joachims, 98]

Label unlabeled data using supervised SVM. Alternate

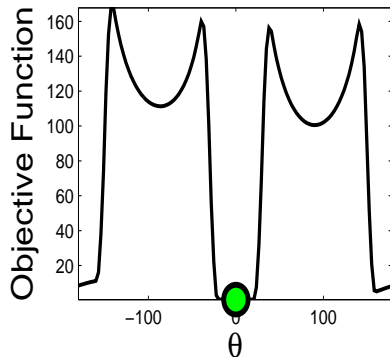
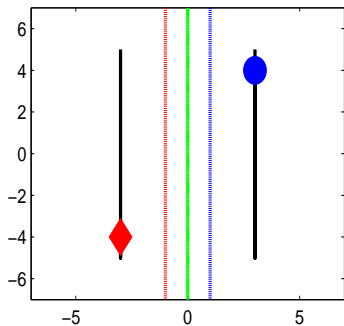
- Optimize f given current \mathbf{y}'
- Optimize \mathbf{y}' by switching a pair of labels

▽ TSVM [Chapelle and Zien, 05]

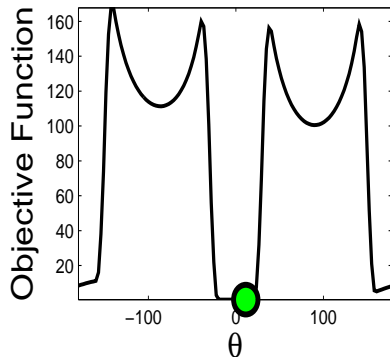
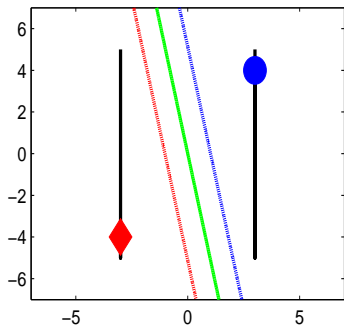
- Use differentiable losses – quadratic hinge loss over labels and a gaussian loss over unlabeled examples.
- apply gradient descent.

[Bennett & Demirez,98], [Fung & Mangasarian,01], [Collobert, Sinz,Weston,Bottou,05],
[Gartner,Le,Burton,Smola,Vishwanathan,05]

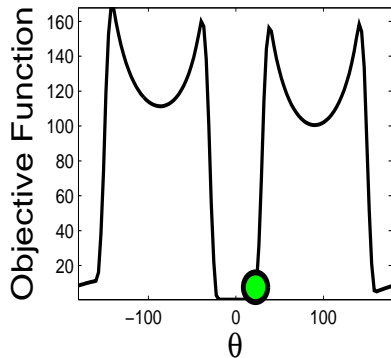
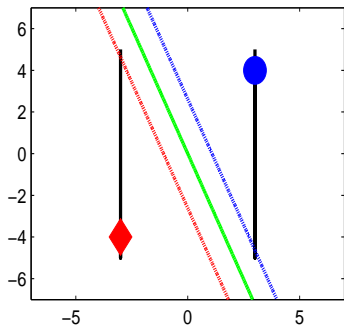
Non-convexity can hurt empirical performance



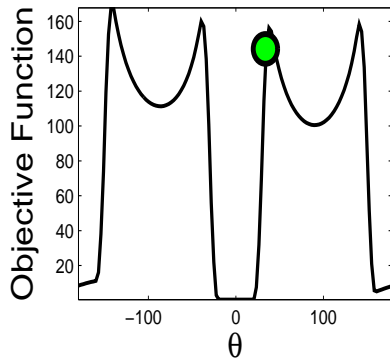
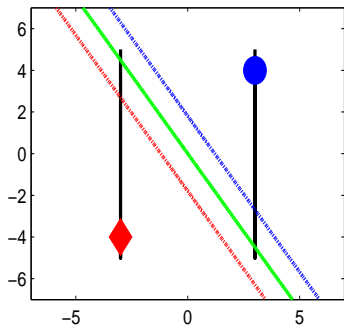
Non-convexity can hurt empirical performance



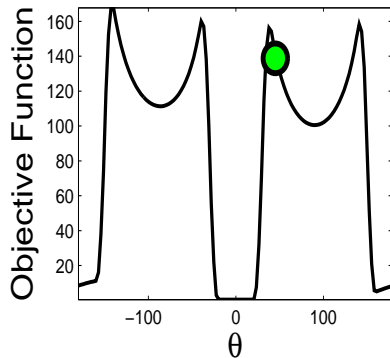
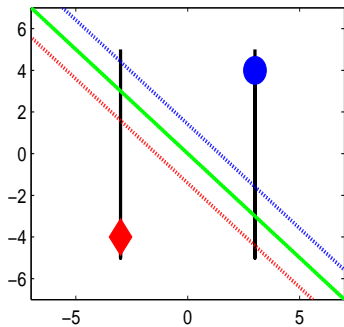
Non-convexity can hurt empirical performance



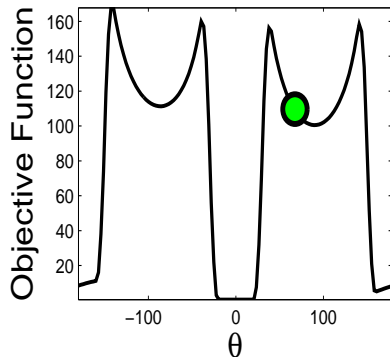
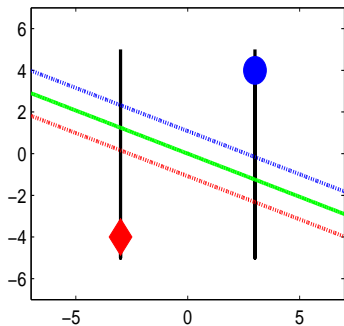
Non-convexity can hurt empirical performance



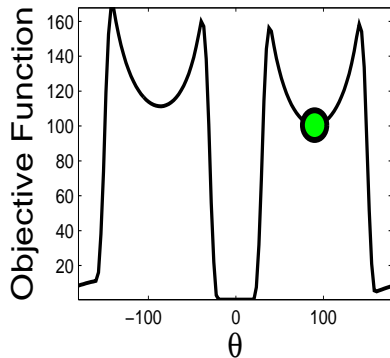
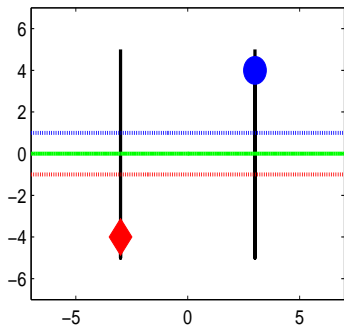
Non-convexity can hurt empirical performance



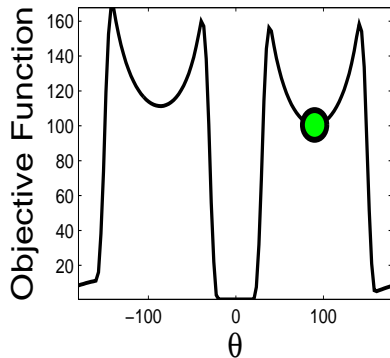
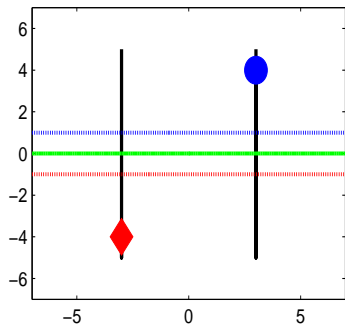
Non-convexity can hurt empirical performance



Non-convexity can hurt empirical performance



Non-convexity can hurt empirical performance

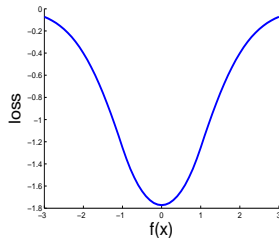
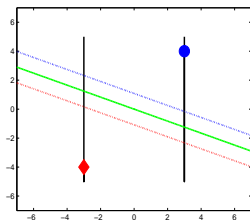


Error rates on COIL6: SVM 21.9, JTSVM 21.2, ∇ TSVM 21.6

Deterministic Annealing: Intuition

Question

What should the shape of the loss function be so that the decision boundary locally evolves in a desirable manner ?



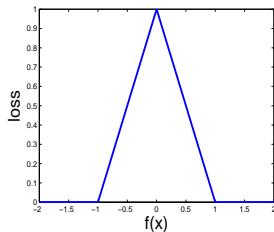
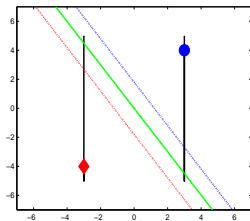
Key Idea

Deform the loss function (objective) as the optimization proceeds...somehow !

Deterministic Annealing: Intuition

Question

What should the shape of the loss function be so that the decision boundary locally evolves in a desirable manner ?



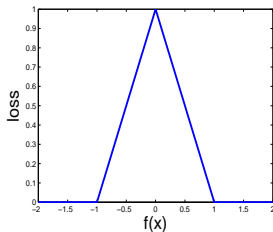
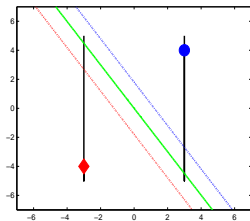
Key Idea

Deform the loss function (objective) as the optimization proceeds...somehow !

Deterministic Annealing: Intuition

Question

What should the shape of the loss function be so that the decision boundary locally evolves in a desirable manner ?



Key Idea

Deform the loss function (objective) as the optimization proceeds...somehow !

Deterministic Annealing as a Homotopy Method

- Work with a **family of objective functions** \mathcal{J}_T .
- Smoothly deform an “easy” (convex) function \mathcal{J}_{T_1} to the given “hard” function $\mathcal{J}_{T_2} = \mathcal{J}$ by varying T .
- Track minimizers along the deformation path.
- DA is a specific implementation of this idea.

Deterministic Annealing for Semi-supervised SVMs

Another Equivalent Continuous Optimization Problem

“Relax” \mathbf{y}' to $\mathbf{p} = (p_1 \dots p_u)$ where p_j is *like* the prob that $y'_j = 1$.

$$\begin{aligned} \mathcal{J}(f, \mathbf{p}) = E_{\mathbf{p}} \mathcal{J}(f, \mathbf{y}') &= \frac{\lambda}{2} \|f\|_K^2 + \frac{1}{l} \sum_{i=1}^l V(y_i, f(\mathbf{x}_i)) \\ &+ \frac{\lambda'}{u} \sum_{j=1}^u \left[p_j V(+1, f(\mathbf{x}'_j)) + (1 - p_j) V(-1, f(\mathbf{x}'_j)) \right] \end{aligned}$$

Family of Objective Functions: Avg Cost - T Entropy

$$\begin{aligned} \mathcal{J}_T(f, \mathbf{p}) = E_{\mathbf{p}} \mathcal{J}(f, \mathbf{y}') - &\underbrace{T H(\mathbf{p})}_{-\frac{T}{u} \sum_{j=1}^u [p_j \log p_j + (1-p_j) \log (1-p_j)]} \end{aligned}$$

Deterministic Annealing for Semi-supervised SVMs

Full Optimization problem at T

$$\min_{f, \mathbf{p}} \mathcal{J}_T(f, \mathbf{p}) = \frac{\lambda}{2} \|f\|_K^2 + \frac{1}{T} \sum_{i=1}^l V(y_i, f(\mathbf{x}_i)) + \frac{\lambda'}{u} \sum_{j=1}^u \left[p_j V(+1, f(\mathbf{x}'_j)) + (1 - p_j) V(-1, f(\mathbf{x}'_j)) \right] + \frac{T}{u} \sum_{j=1}^u [p_j \log p_j + (1 - p_j) \log p_j] \quad \text{s.t. } (1/u) \sum_{j=1}^u p_j = r$$

- **Deformation:** T controls non-convexity of $\mathcal{J}(f, \mathbf{p})$. At $T = 0$, reduces to the original non-convex objective function $\mathcal{J}(f, \mathbf{p})$.
- **Optimization at T** $(f_T^*, \mathbf{p}_T^*) = \operatorname{argmin}_{f, \mathbf{p}} \mathcal{J}_T(f, \mathbf{p})$
- **Annealing:** Return: $f^* = \lim_{T \rightarrow 0} f_T^*$
- **Balance constraint:** $\frac{1}{u} \sum_{j=1}^u p_j = r$

Alternating Convex Optimization

At any T , optimize f keeping \mathbf{p} fixed

- Representer theorem:

$$f(\mathbf{x}) = \sum_{i=1}^l \alpha_i \mathbf{K}(\mathbf{x}, \mathbf{x}_i) + \sum_{j=1}^u \alpha'_j \mathbf{K}(\mathbf{x}, \mathbf{x}'_j)$$

- Minimize weighted regularized loss using standard tricks.

At any T , optimize \mathbf{p} keeping f fixed

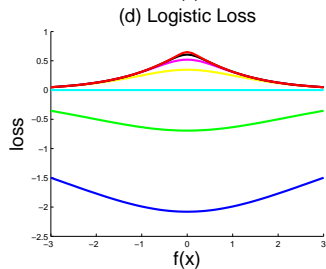
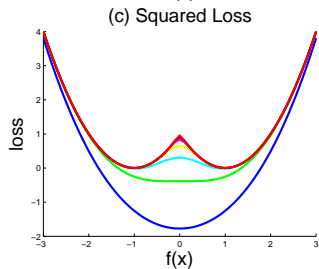
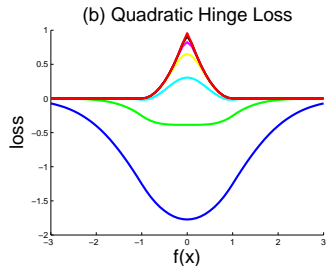
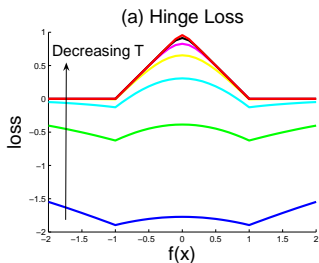
- $p_j^* = \frac{1}{1+e^{\frac{g_j - \nu}{T}}}$ $g_j = \lambda' \left[V(f(\mathbf{x}'_j)) - V(-f(\mathbf{x}'_j)) \right]$

- Obtain ν by solving $\frac{1}{u} \sum_{j=1}^u \frac{1}{1+e^{\frac{g_j - \nu}{T}}} = r$

Stopping Conditions

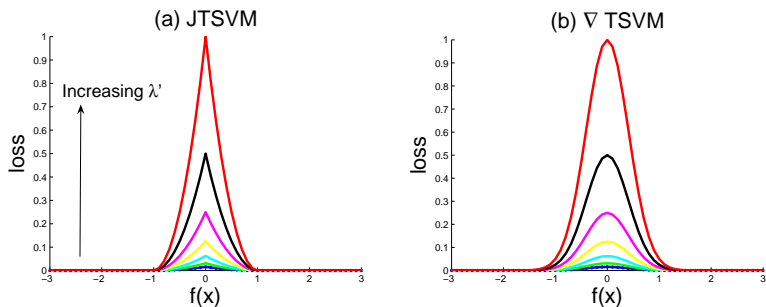
- At any T , alternate until $KL(\mathbf{p}_{new} | \mathbf{p}_{old}) < \epsilon$. Obtain p_T^* .
- Reduce T , Seed old p_T^* , until $H(\mathbf{p}_T^*) < \epsilon$.

How effective Loss deforms as a function of T



Effective Loss in JT SVM, ∇ TSVM wrt λ' .

$$J_{\lambda'}(f) = \frac{\lambda}{2} \|f\|_K^2 + \frac{1}{l} \sum_{i=1}^l V(y_i, f(\mathbf{x}_i)) + \frac{\lambda'}{u} \sum_{j=1}^u V'(f(\mathbf{x}'_j))$$



Unlabeled examples outside the margin do not influence the decision boundary!

Deterministic Annealing: Some Quick Comments

- **Smoothing:** At high T , spurious & shallow local min are smoothed away.
- **Simulated Annealing:** Stochastic search allowing “uphill” moves depending on T . Associated Markov process converges *slowly* to Gibbs distribution at equilibrium which minimizes $E_{\mathbf{p}}\mathcal{J} - TH(\mathbf{p})$ (*free energy*). As $T \rightarrow 0$ *very slowly*, global solution guaranteed (in prob). DA retains annealing but avoids stochastic search by directly optimizing $E_{\mathbf{p}}\mathcal{J} - TH(\mathbf{p})$ for \mathbf{p} .
- **Maximum Entropy:** $E_{\mathbf{p}}\mathcal{J} - TH(\mathbf{p})$ is the Lagrangian of: $\max_{\mathbf{p}} S(\mathbf{p})$ subject to $E_{\mathbf{p}}\mathcal{J} = \beta$.
- **Proven Heuristic:** Very strong record of empirical success, including in clustering, classification, compression problems. For SSL, has been applied with EM in [Nigam, 2001].

Deterministic Annealing: Some Quick Comments

- **Smoothing:** At high T , spurious & shallow local min are smoothed away.
- **Simulated Annealing:** Stochastic search allowing “uphill” moves depending on T . Associated Markov process converges *slowly* to Gibbs distribution at equilibrium which minimizes $E_{\mathbf{p}}\mathcal{J} - TH(\mathbf{p})$ (*free energy*). As $T \rightarrow 0$ *very slowly*, global solution guaranteed (in prob). DA retains annealing but avoids stochastic search by directly optimizing $E_{\mathbf{p}}\mathcal{J} - TH(\mathbf{p})$ for \mathbf{p} .
- **Maximum Entropy:** $E_{\mathbf{p}}\mathcal{J} - TH(\mathbf{p})$ is the Lagrangian of: $\max_{\mathbf{p}} S(\mathbf{p})$ subject to $E_{\mathbf{p}}\mathcal{J} = \beta$.
- **Proven Heuristic:** Very strong record of empirical success, including in clustering, classification, compression problems. For SSL, has been applied with EM in [Nigam, 2001].

Deterministic Annealing: Some Quick Comments

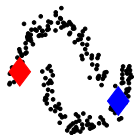
- **Smoothing:** At high T , spurious & shallow local min are smoothed away.
- **Simulated Annealing:** Stochastic search allowing “uphill” moves depending on T . Associated Markov process converges *slowly* to Gibbs distribution at equilibrium which minimizes $E_{\mathbf{p}}\mathcal{J} - TH(\mathbf{p})$ (*free energy*). As $T \rightarrow 0$ *very slowly*, global solution guaranteed (in prob). DA retains annealing but avoids stochastic search by directly optimizing $E_{\mathbf{p}}\mathcal{J} - TH(\mathbf{p})$ for \mathbf{p} .
- **Maximum Entropy:** $E_{\mathbf{p}}\mathcal{J} - TH(\mathbf{p})$ is the Lagrangian of: $\max_{\mathbf{p}} S(\mathbf{p})$ subject to $E_{\mathbf{p}}\mathcal{J} = \beta$.
- **Proven Heuristic:** Very strong record of empirical success, including in clustering, classification, compression problems. For SSL, has been applied with EM in [Nigam, 2001].

Deterministic Annealing: Some Quick Comments

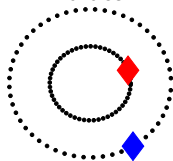
- **Smoothing:** At high T , spurious & shallow local min are smoothed away.
- **Simulated Annealing:** Stochastic search allowing “uphill” moves depending on T . Associated Markov process converges *slowly* to Gibbs distribution at equilibrium which minimizes $E_{\mathbf{p}}\mathcal{J} - TH(\mathbf{p})$ (*free energy*). As $T \rightarrow 0$ *very slowly*, global solution guaranteed (in prob). DA retains annealing but avoids stochastic search by directly optimizing $E_{\mathbf{p}}\mathcal{J} - TH(\mathbf{p})$ for \mathbf{p} .
- **Maximum Entropy:** $E_{\mathbf{p}}\mathcal{J} - TH(\mathbf{p})$ is the Lagrangian of: $\max_{\mathbf{p}} S(\mathbf{p})$ subject to $E_{\mathbf{p}}\mathcal{J} = \beta$.
- **Proven Heuristic:** Very strong record of empirical success, including in clustering, classification, compression problems. For SSL, has been applied with EM in [Nigam, 2001].

First Experiments

2moons



2circles



Successes in 10 trials. $l=2$

Dataset → Algorithm ↓	2MOONS	2CIRCLES
JTSVM (l_2)	0	1
JTSVM (l_1)	0	1
∇ T SVM	3	2
DA (l_2)	6	3
DA (l_1)	10	10

Used RBF Kernels, Optimal parameters.

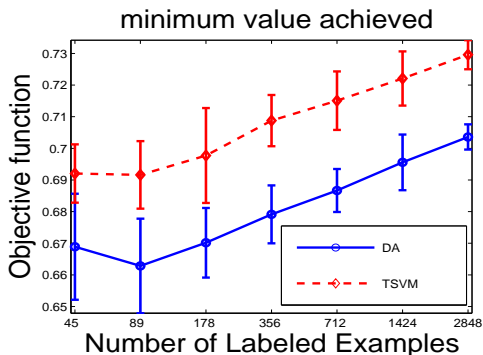
Real-world Datasets

$\lambda' = 1$; λ, σ optimized; avg. over 10 random splits. $T = \frac{10}{1.5^i}$ $i = 0, 1, \dots$

Unlab	USPS2	COIL6	PCMAC	ESET2
SVM	7.5	21.5	18.9	19.4
JTSVM	7.6	19.9	10.4	9.2
∇ TSM	6.9	21.4	5.4	8.7
DA(l_2)	6.4	13.6	5.3	8.1
DA(<i>sqr</i>)	5.7	13.8	5.4	9.0
Test				
SVM	7.8	21.9	17.9	19.7
JTSVM	7.2	21.2	7.0	8.9
∇ TSM	7.1	21.6	4.5	9.1
DA(l_2)	6.3	15.0	4.8	8.5
DA(<i>sqr</i>)	6.3	15.2	4.7	9.4

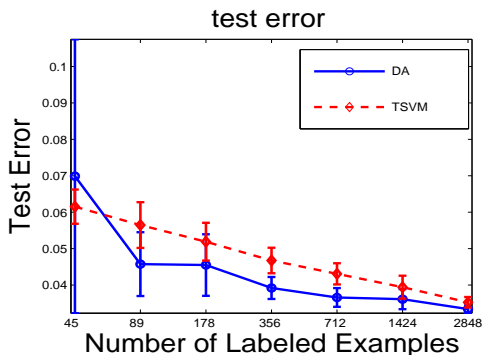
Large Scale Text Categorization

UseNet articles from two discussion groups: Auto-vs-Aviation.
Used special primal routines for linear kernels, [Keerthi and Decoste, 2005]. More results in [SK,SIGIR 06]
#features=20707, #training=35543, #test=35587.



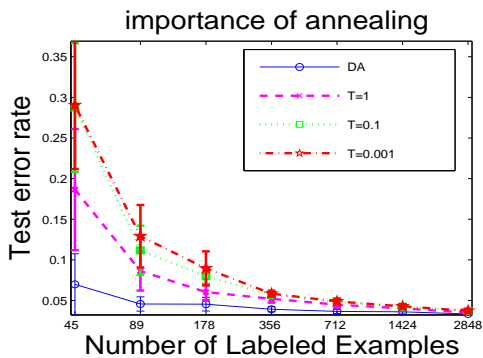
Large Scale Text Categorization

UseNet articles from two discussion groups: Auto-vs-Aviation.
Used special primal routines for linear kernels, [Keerthi and Decoste, 2005]. More results in [SK,SIGIR 06]
#features=20707, #training=35543, #test=35587.



Large Scale Text Categorization

UseNet articles from two discussion groups: Auto-vs-Aviation.
Used special primal routines for linear kernels, [Keerthi and Decoste, 2005]. More results in [SK,SIGIR 06]
#features=20707, #training=35543, #test=35587.



Importance of Annealing

Unlab	USPS2	COIL6	PC-MAC	ESET2
DA	6.4	13.6	5.3	8.1
T=0.1	6.6	20.0	5.7	7.8
T=0.01	7.6	20.1	7.1	8.1
T=0.001	7.9	20.3	9.1	8.8
Test				
DA	6.3	15.0	4.8	8.5
T=0.1	6.8	21.0	4.7	8.0
T=0.01	7.0	21.3	5.7	8.5
T=0.001	7.2	21.5	7.3	8.8

Summary and Open Questions

Summary

- New optimization method that better approaches global solution for TSVM-like SSL.
- “Easy” to “Hard” approach.
- Can use off-the-shelf optimization subroutines.

Open Questions

- Intriguing connections between annealing behaviour, loss function and regularization.
- Annealing sequence ? Detailed experimental studies.

Also see: *A Continuation method for Semi-supervised SVMs*,
O. Chapelle, M. Chi, A. Zien, ICML 2006.