Artificial Intelligence

Lecture 9

Bo Yuan, Ph.D.

Professor, Computer Science and Engineering

Shanghai Jiaotong University

Inspired by Eric Xing (CMU), Andrew Ng (Stanford), and Michael Jordan (UC Berkley)

Some of the slides were borrowed from Eric Xing (without Permission)

Review of Lecture 8

- Support Vector Machine (II)
  - SVM as a strong dual OPT problem
  - Implications
  - Nonlinear transformation of data
  - Kernel trick
  - Kernel function
  - Kernel matrix and implications
  - Examples

Today's Content

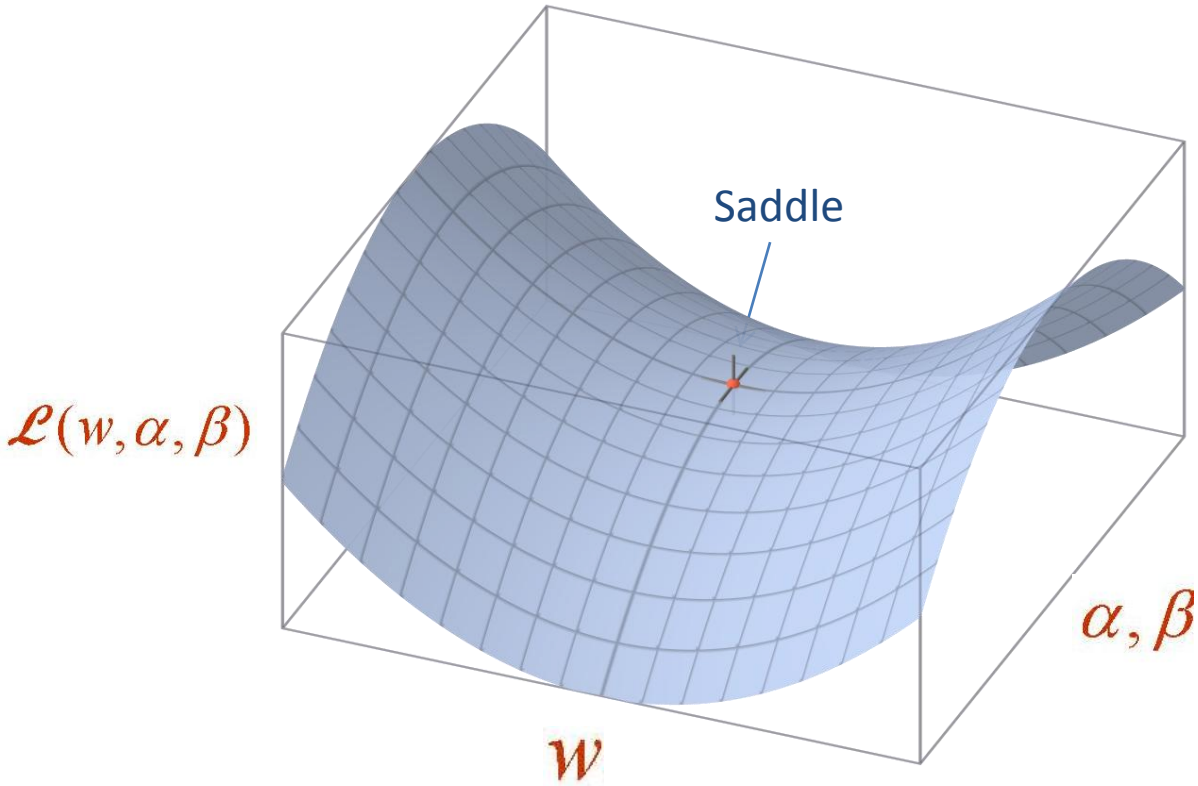
- Idea of Kernel
- Soft Margin
- Sequential Minimal Optimization (SMO)
- Some SVM Applications

Beauty of SVM

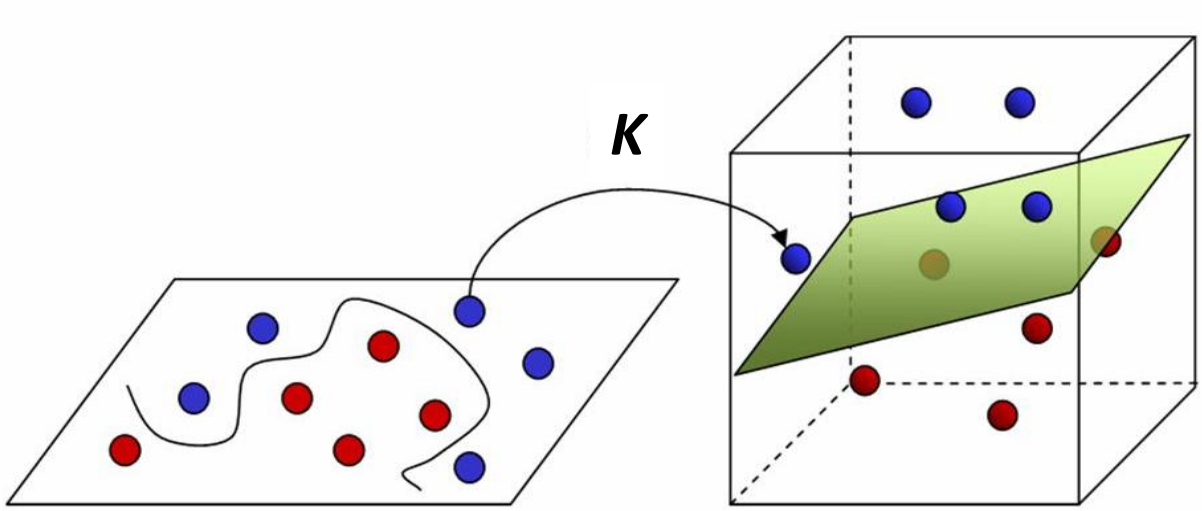
- A strong dual $\mathcal{L}$ problem;
- Only SVMs are considered in training;
- The Inner product as the kernel;
- Free Lunch for dimension expansion;
- SMO, a convex problem;

Strong Duality

Primal Problem Equivalent to the Dual Problem



$$\max_{\alpha,\beta,\alpha_i \geq 0} \min_{w} \mathcal{L}(w,\alpha,\beta) = \min_{w} \max_{\alpha,\beta,\alpha_i \geq 0} \mathcal{L}(w,\alpha,\beta)$$

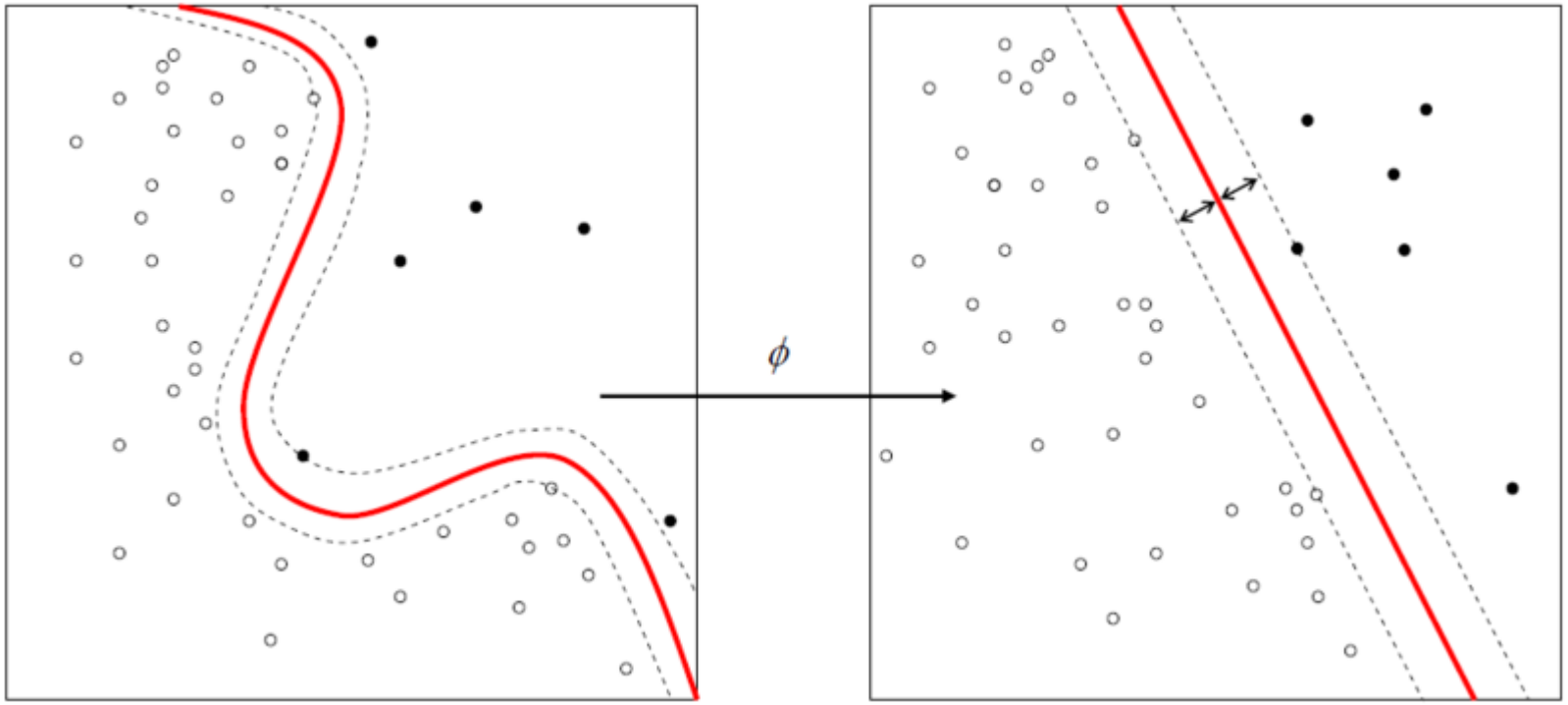


$$y^* = \text{sign}\left(\sum_{i \in SV} \alpha_i y_i \left(\mathbf{x}_i^T z\right) + b\right)$$

$$y^*(z) = \text{sign}\left(\sum_{i \in SV} \alpha_i y_i K\left(\mathbf{x}_i, z\right) + b\right)$$

Nonlinear operation
underlying the kernel

The Use of Kernel to Linearize A Similarity Measure



Kernel $k(\mathbf{x_i}, \mathbf{x_j}) = \varphi(\mathbf{x_i}) \cdot \varphi(\mathbf{x_j})$ $\mathbf{w} \cdot \varphi(\mathbf{x}) = \sum_i \alpha_i y_i k(\mathbf{x}_i, \mathbf{x})$

$\mathbf{w} = \sum_i \alpha_i y_i \varphi(\mathbf{x}_i)$

The General Idea of Kernel

- $X^TX' \rightarrow \phi(X)^T\phi(X') \rightarrow \kappa(X,X')$

- "Inner Product", a "Similarity Measure";

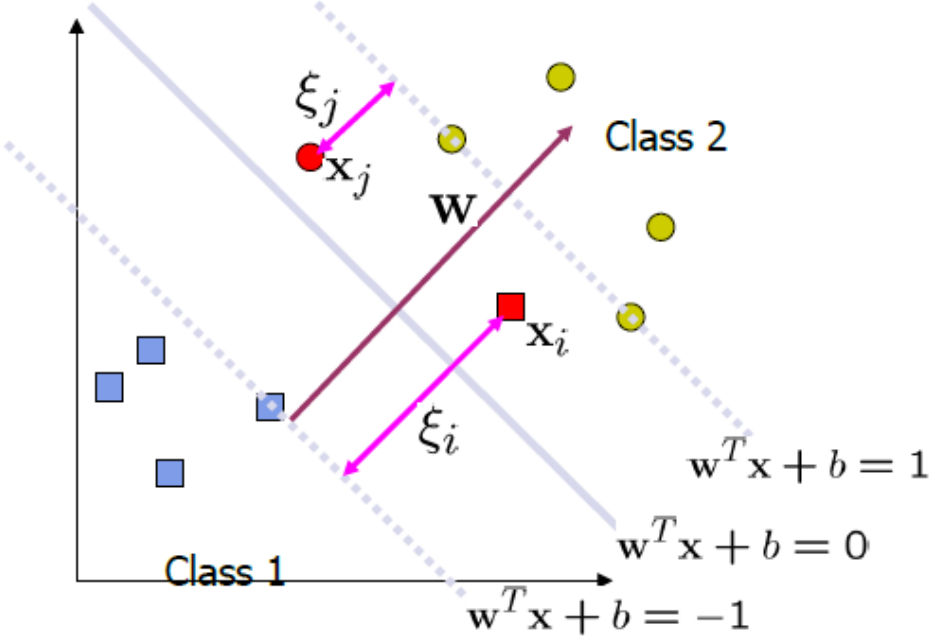
- "Kernelization", to project the original X^TX' to a "Feature Space" of higher dimensions ("Feature Vectors"), in order to better distinguish and measure the difference and similarity between the X and X'.

- Thus, the goal is to find a good "Distance Function";

Representative Kernel Functions

- Polynomial (homogeneous)
- Polynomial (inhomogeneous)
- Gaussian radial basis function
- Hyperbolic tangent

Regularization in Non-Separable Case



- We allow "error" ξ_i in classification; it is based on the output of the discriminant function $\boldsymbol{w}^T\boldsymbol{x}+b$
- ξ_i approximates the number of misclassified samples

Soft Margin

- Now we have a slightly different opt problem:

$$\min_{w,b} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i$$

$$\text{s.t} \quad y_i (w^T x_i + b) \geq 1 - \xi_i, \quad \forall i$$
$$\xi_i \geq 0, \quad \forall i$$

- ξ_i are "slack variables" in optimization
- Note that ξ_i=0 if there is no error for $\mathbf{x}_i$
- ξ_i is an upper bound of the number of errors
- C : tradeoff parameter between error and margin

$$\min_{\gamma,w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m$$

$$\xi_i \geq 0, \quad i = 1, \ldots, m.$$

A convex constrained OPT problem

Extra issue

Recall $y^{(i)}(w^T x^{(i)} + b) \geq 1$, correct classification

---

$$\mathcal{L}(w, b, \xi, \alpha, r) = \frac{1}{2}w^T w + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i\left[y^{(i)}(x^T w + b) - 1 + \xi_i\right] - \sum_{i=1}^{m} r_i\xi_i$$

Write down the $\mathcal{L}$

A new $\mathcal{L}$ multiplier

A new constrain

⇩ Convert it as a dual

$$\max_{\alpha} \quad W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m$$

$$\sum_{i=1}^{m}\alpha_i y^{(i)} = 0,$$

Now α's are constrained by a range.

⇩ Based on KKT

$$\begin{aligned}
\alpha_i = 0 \quad &\Rightarrow \quad y^{(i)}(w^T x^{(i)} + b) \geq 1 \\
\alpha_i = C \quad &\Rightarrow \quad y^{(i)}(w^T x^{(i)} + b) \leq 1 \\
0 < \alpha_i < C \quad &\Rightarrow \quad y^{(i)}(w^T x^{(i)} + b) = 1
\end{aligned}$$

A Similar Optimization Problem

- The dual of this new constrained optimization problem is

$$\max_{\alpha} \quad \mathcal{J}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j)$$

$$\text{s.t.} \quad 0 \le \alpha_i \le C, \quad i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0.$$

- This is very similar to the optimization problem in the linear separable case, except that there is an upper bound C on α_i now
- Once again, a QP solver can be used to find α_i

The SMO Algorithm

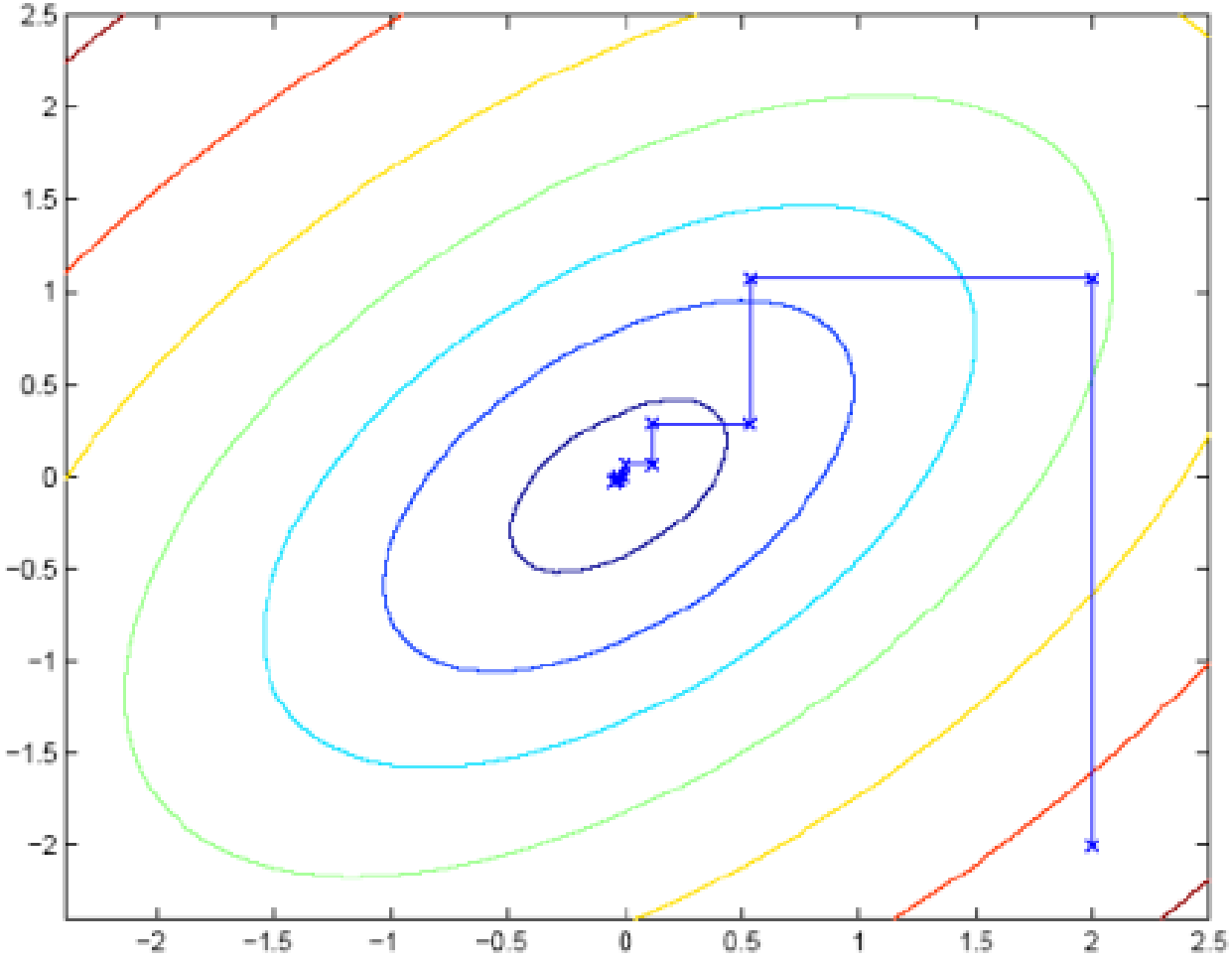
- To solve the following unconstrained OPT problem

$$\max_{\alpha} \boldsymbol{W}(\alpha_1,\alpha_2,\ldots,\alpha_m)$$

- We have already seen several opt algorithms!
  - Gradient (partial)
  - Newton (2nd partial)
  - Coordinate (sequential partial)

- Coordinate ascend

Coordinate Ascend



Keep the reamining wi, changing only wj, j \= j.

Sequential Minimal Optimization (SMO)

- Constrained optimization:

$$\max_{\alpha} \quad \mathcal{J}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j)$$

$$\text{s.t.} \quad 0 \le \alpha_i \le C, \quad i = 1, \ldots, m \qquad \text{Box Constrains}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0.$$

- Question: can we do coordinate along one direction at a time (i.e., hold all $\alpha_{[-i]}$ fixed, and update α_i?)

Can not only update one α_i, while keeping everything else fixed, because of the constrains $\sum_{i=1}^{m} \alpha_i y_i = 0.$ Given (m-1) α's fixed, no single α_i can be changed.

How about keep (m-2) α's fixed, that $\alpha_i + \alpha_j = C$?

The SMO Algorithm

Repeat till convergence

1. Select some pair α_i and α_j to update next (using a heuristic that tries to pick the two that will allow us to make the biggest progress towards the global maximum).

2. Re-optimize J(α) with respect to α_i and α_j, while holding all the other α_k's $(k \neq i; j)$ fixed.

Will this procedure converge?

Sequential w.r.t pairs! It is a convex problem!

Convergence of SMO

$$\max_{\alpha} \quad \mathcal{J}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j)$$

KKT:

$$\text{s.t.} \quad 0 \le \alpha_i \le C, \quad i = 1, \ldots, k$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0.$$

- Let's hold $\alpha_3, \ldots, \alpha_m$ fixed and reopt J w.r.t. α_1 and α_2

The SMO Algorithm

- The constrain: $\sum_{i=1}^{m} \alpha_i y_i = 0.$
- Select α_i, α_j (heuristic);
- Hold the remaining (m-2) α's fixed except α_i, α_j;
- Optimize **w**(α) w.r.t α_i, α_j (subject to the constrain);
- See below

Convergence of SMO

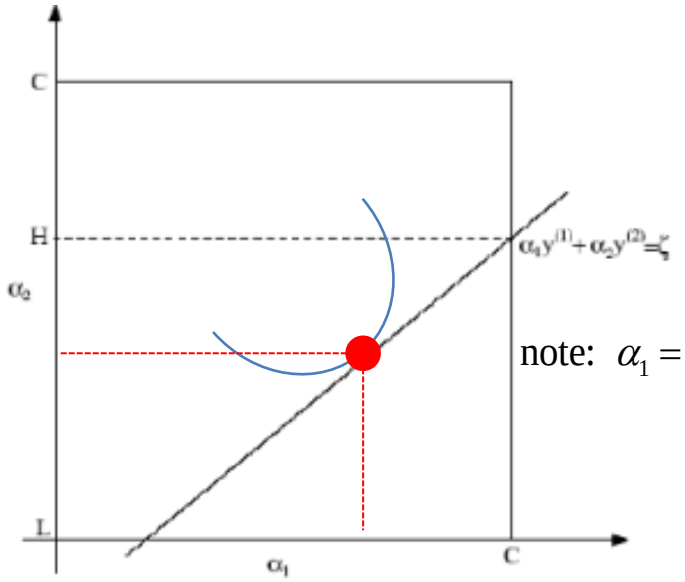
- The constraints:

$$\alpha_1 y_1 + \alpha_2 y_2 = \xi$$

Known because of the KKT condition!

$$0 \le \alpha_1 \le C$$

$$0 \le \alpha_2 \le C$$



note: $\alpha_1 = \dfrac{\xi - \alpha_2 y^{(2)}}{y^{(1)}}$

- The objective:

$$\mathcal{J}(\alpha_1, \alpha_2, \ldots, \alpha_m) = \mathcal{J}((\xi - \alpha_2 y_2) y_1, \alpha_2, \ldots, \alpha_m)$$

- Constrained opt:

$$= a\alpha_2^2 + b\alpha_2 + c$$

$$\alpha_2$$

What is SVM?

Primal problem (Maximal Margin)

Dual Problem (Support Vector, Kernel)

A Constrained OPT Problem

Solved by a SMO Algorithm

SVM - Summary

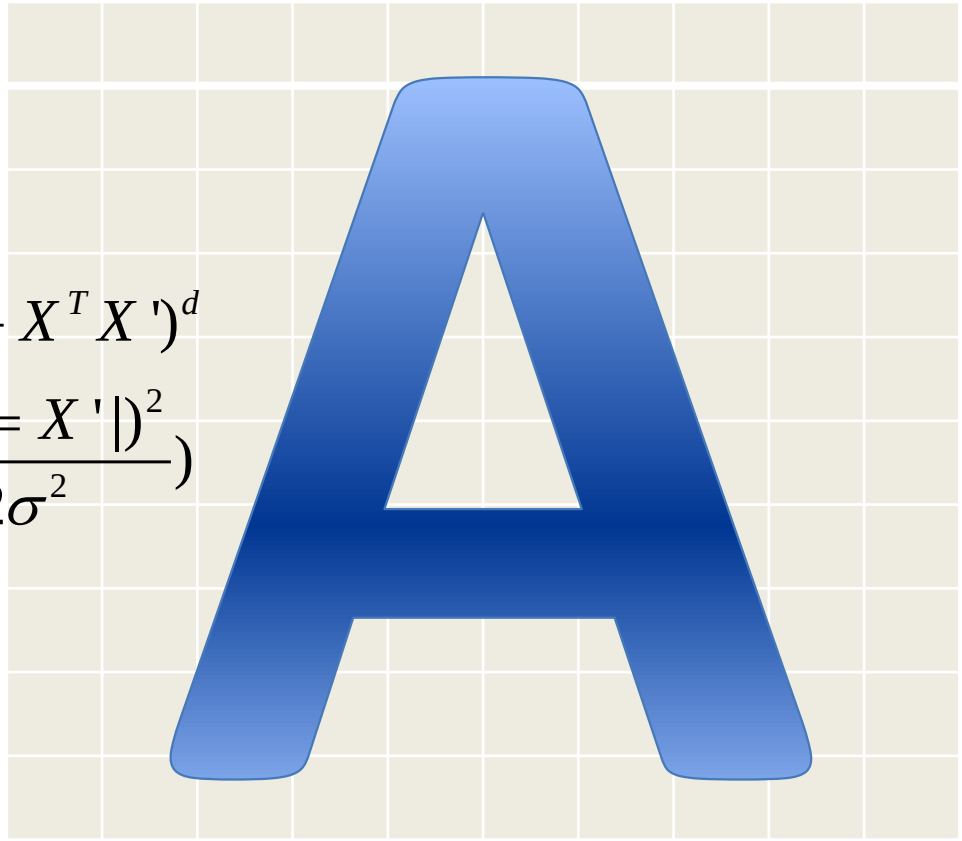
- Max-margin decision boundary
- Constrained convex optimization
  - Lagrangian Duality
  - KKT conditions and the support vectors
  - Kernel Function
  - Non-separable case and slack variables
- The SMO algorithm

Handwriting Recognition

$$X \in \mathbb{R}^{100}$$

$$K(X, X') = (c + X^T X')^d$$

$$\text{or} \quad \exp\left(-\frac{(|X = X'|)^2}{2\sigma^2}\right)$$



Protein Classification

- Composed of 20 amino acids (A,…,Z);
- What ϕ to chose?
- An example of a protein sequence:
  "BAJTSIAIBAJTAU…."
- Chose 4 tuples as the feature vectors →

$$\phi(x) \in \mathbb{R}^{(20^4)} = \mathbb{R}^{160000}$$

- Dynamical programming to compute $<\phi(x)^T \cdot \phi(x')>$
- This is a reasonable computing task, even though very high in dimension!
- A new kernel.

AAAA	0
AAAB	0
AAAC	0
…	
AAAZ	0
AABA	0
AACA	0
…	
..	
BAJT	2
..	
…	
TSIA	1
…	
…	
ZZZZ	0

$= \phi(x)$