Introduction to Machine Learning (CSCI-UA 473): Fall 2021

Lecture 7: Support Vector Machines - 2

Sumit Chopra
Courant Institute of Mathematical Sciences
Department of Radiology - Grossman School of Medicine
NYU

Slides derived from materials from Benjamin Peherstorfer, Kyunghyun Cho, Andrew Gordon Wilson

Lecture Outline

Primal Formation of SVMs

Overview of Optimization

Dual Formulation of SVMs

SVMs for Non-Linearly Separable Case

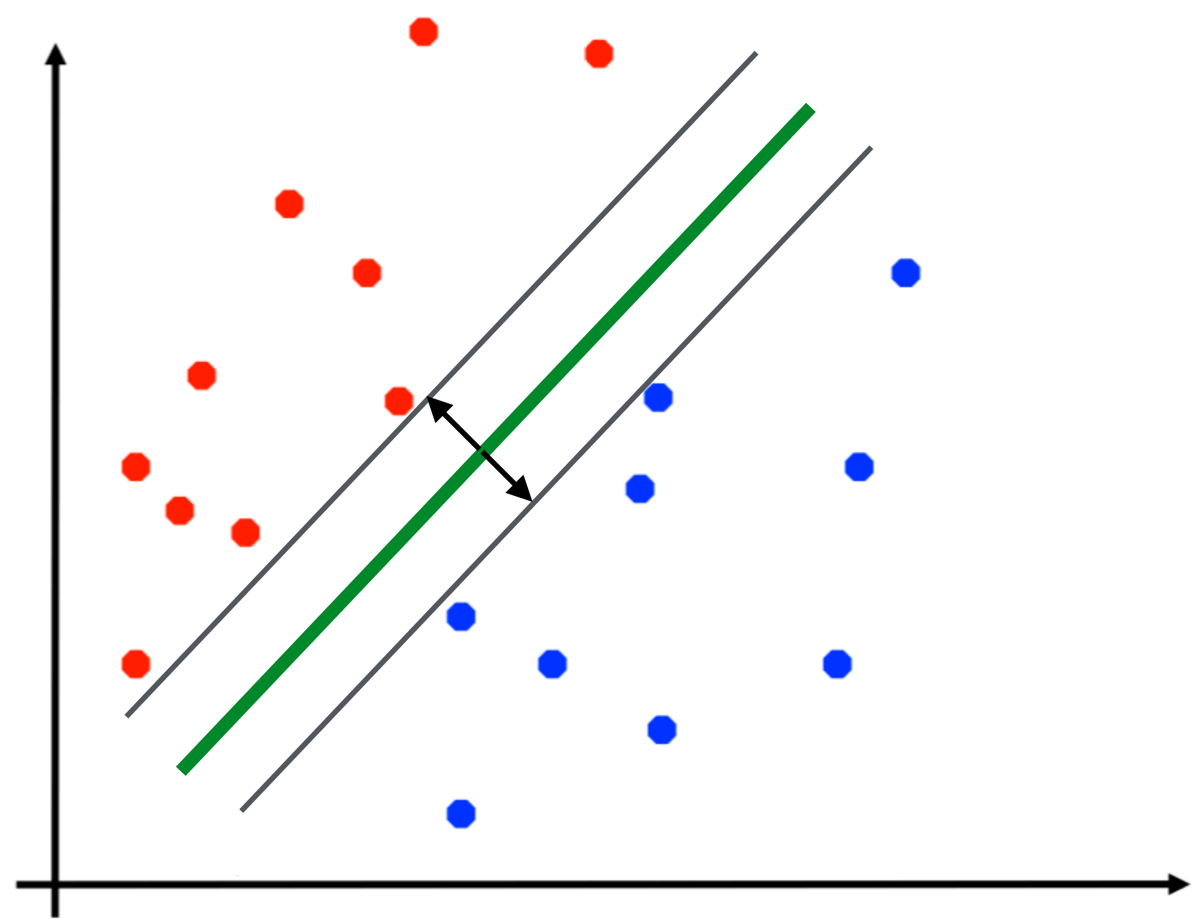
Maximum Margin Classifiers

SVMs solve the following optimization problem to compute a hyper-plane which has the maximum margin

Convex quadratic optimization with linear constraints

$$\min_{w,b} \quad \frac{1}{2}||w||^2$$

$$\mathbf{s.t.} \quad y^i(w^T x^i + b) \geq 1, \qquad \forall i = 1,\ldots,n$$



Maximum Margin Classifiers

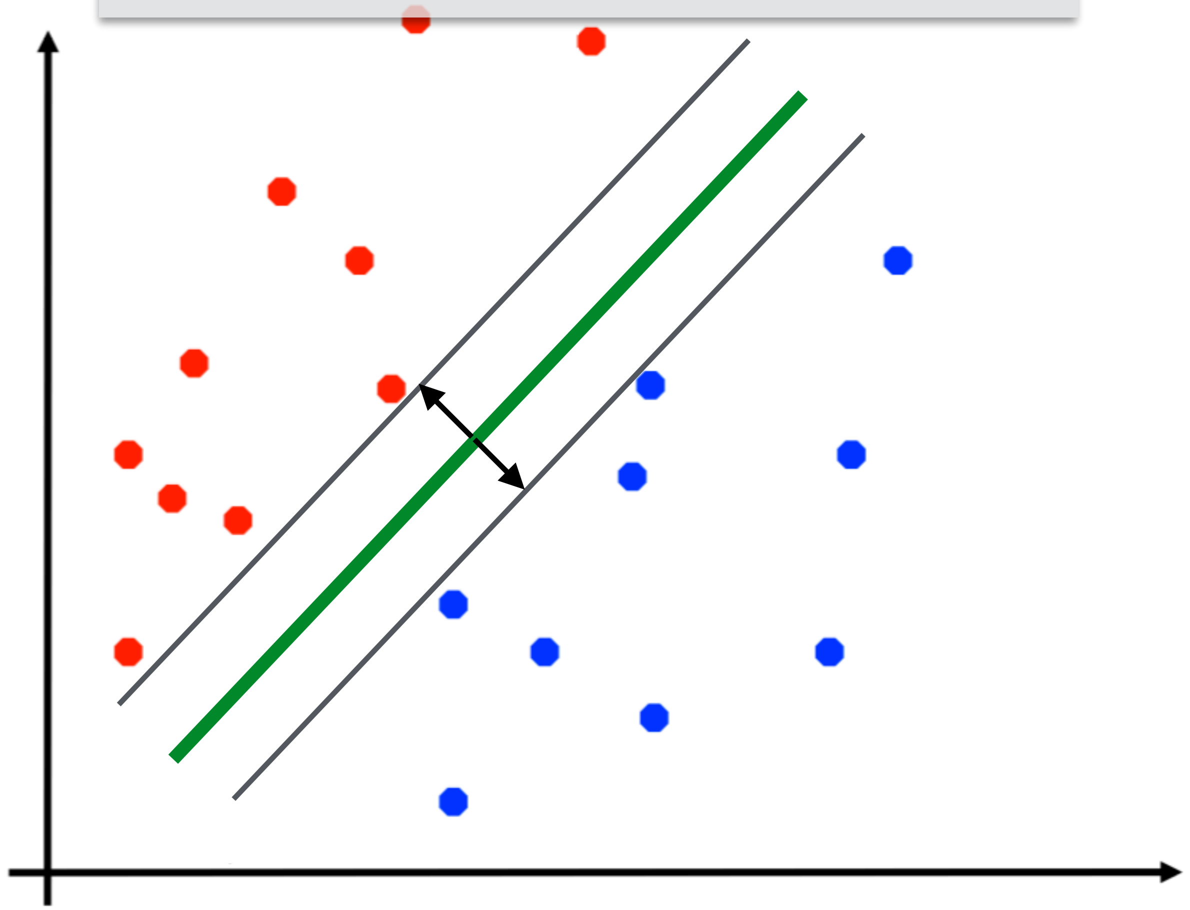
SVMs solve the following optimization problem to compute a hyper-plane which has the maximum margin

Convex quadratic optimization with linear constraints

$$\min_{w,b} \quad \frac{1}{2}||w||^2$$

$$\textbf{s.t.} \quad y^i(w^T x^i + b) \geq 1, \qquad \forall i = 1,\ldots,n$$

How do we solve this?

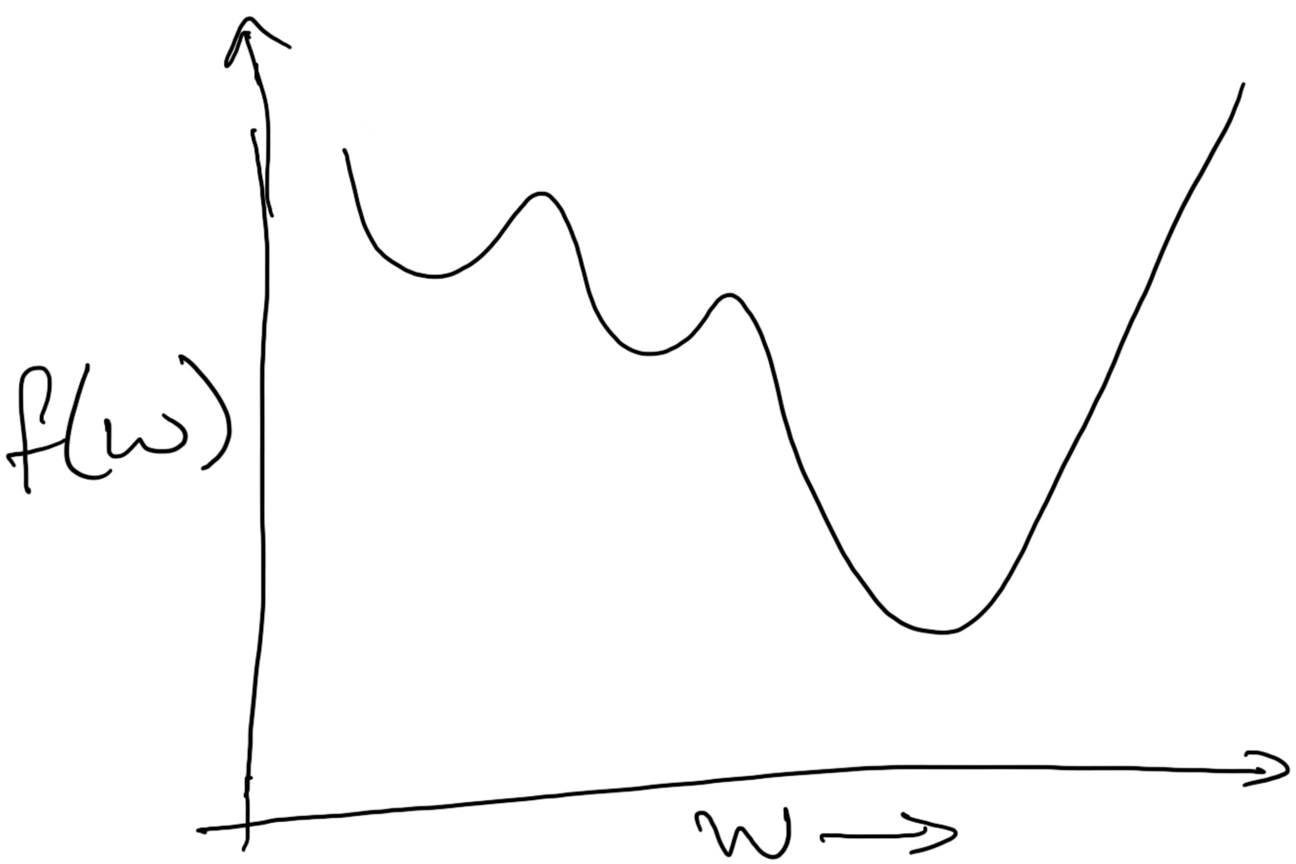


60,000 View of Function Optimization

Overview of Optimization

$$\min(or\ \max) \quad f(w) \qquad\qquad w \in \Re^d$$

Unconstrained Optimization Problem



Overview of Optimization

Constrained Optimization Problem

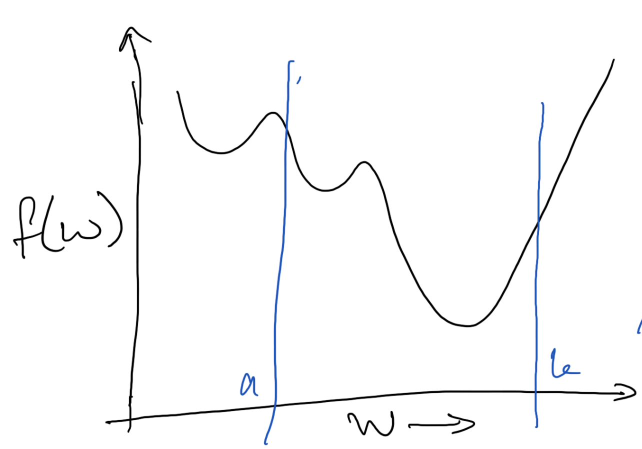
$$\min(or \max)_w \quad f(w) \qquad w \in \Re^d$$

$$g_1(w) \leq 0 \qquad h_1(w) = 0$$
$$g_1(w) \leq 0 \qquad h_1(w) = 0$$
$$\vdots \qquad \vdots$$
$$g_m(w) \leq 0 \qquad h_n(w) = 0$$

Inequality constrains | Equality constraints

Any equality constraint can be represented as two inequality constraint

$$g(w) = 0 \quad \rightarrow \quad g(w) \leq 0 \quad -g(w) \leq 0$$



$w \leq b$
$w \geq a$
In GENERAL.
$g_1(w) \leq b$
$g_2(w) \geq a$
$\vdots$

Constrained Optimization Problem: Lagrangian Multipliers

Equality Constraints

$$\min_{w} \quad f(w) \qquad \mathbf{s.t.} \quad \begin{aligned} h_1(w) &= 0 \\ h_1(w) &= 0 \\ &\vdots \\ h_e(w) &= 0 \end{aligned} \qquad w \in \Re^d$$

Reduces the problem with d variables with e constraints into an unconstrained problem of $d + e$ variables

Introduces a scalar variable (called the Lagrangian multiplier) for each constraint and forms a linear combination involving the multipliers as coefficients

$$L(w, \beta) = f(w) + \sum_{i=1}^{e} \beta_i h_i(w)$$

$$\frac{\partial L}{\partial w_j} = 0, \qquad \frac{\partial L}{\partial \beta_i} = 0 \qquad \forall j \in \{1,\ldots,d\} \qquad \forall i \in \{1,\ldots,e\}$$

Generalized Lagrangian

Inequality
Constraints

$$\min_{w} \quad f(w) \qquad \mathbf{s.t.} \quad \begin{array}{ll} h_1(w) = 0 & g_1(w) \leq 0 \\ h_1(w) = 0 & g_1(w) \leq 0 \\ \vdots & \vdots \\ h_e(w) = 0 & g_k(w) \leq 0 \end{array} \qquad w \in \Re^d$$

$$L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{e} \beta_i h_i(w)$$

Generalized Lagrangian

$$O_p(w) = \max_{\alpha,\beta:\alpha_i \geq 0} L(w, \alpha, \beta) = \max_{\alpha,\beta:\alpha_i \geq 0} \left[f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{e} \beta_i h_i(w) \right]$$

Generalized Lagrangian

$$O_p(w) = \max_{\alpha,\beta:\alpha_i \geq 0} L(w,\alpha,\beta) = \max_{\alpha,\beta:\alpha_i \geq 0} \left[f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{e} \beta_i h_i(w) \right]$$

If w violates any of the constraints, either: $g_i(w) > 0$ or $h_i(w) \neq 0$ then

$$\begin{aligned} O_p(w) &= \max_{\alpha=\{\alpha_1,\ldots,\alpha_k\},\beta=\{\beta_1,\ldots,\beta_e\}:\alpha_i \geq 0} \left[f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{e} \beta_i h_i(w) \right] \\ &= \infty \end{aligned}$$

Generalized Lagrangian

$$O_p(w) = \max_{\alpha,\beta:\alpha_i \geq 0} L(w,\alpha,\beta) = \max_{\alpha,\beta:\alpha_i \geq 0} \left[f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{e} \beta_i h_i(w) \right]$$

If w violates any of the constraints, either: $g_i(w) > 0$ or $h_i(w) \neq 0$ then

$$\begin{aligned} O_p(w) &= \max_{\alpha=\{\alpha_1,\ldots,\alpha_k\},\beta=\{\beta_1,\ldots,\beta_e\}:\alpha_i \geq 0} \left[f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{e} \beta_i h_i(w) \right] \\ &= \infty \end{aligned}$$

If w satisfies all the constraints

$g_i(w) \leq 0$ $\quad$ $h_i(w) = 0$

$$\begin{aligned} O_p(w) &= \max_{\alpha=\{\alpha_1,\ldots,\alpha_k\},\beta=\{\beta_1,\ldots,\beta_e\}:\alpha_i \geq 0} \left[f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{e} \beta_i h_i(w) \right] \\ &= f(w) \end{aligned}$$

Generalized Lagrangian

If w satisfies all the constraints

$$O_p(w) = \max_{\alpha=\{\alpha_1,\ldots,\alpha_k\},\beta=\{\beta_1,\ldots,\beta_e\}:\alpha_i\geq 0} \left[f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{e} \beta_i h_i(w) \right]$$
$$= f(w)$$

So long as the constraints are satisfied solution to $O_p(w)$ is the same as $f(w)$

$$p^* = \min_w O_p(w) = \min_w \max_{\alpha,\beta:\alpha_i\geq 0} L(w,\alpha,\beta) = \min_w f(x)$$

Generalized Lagrangian

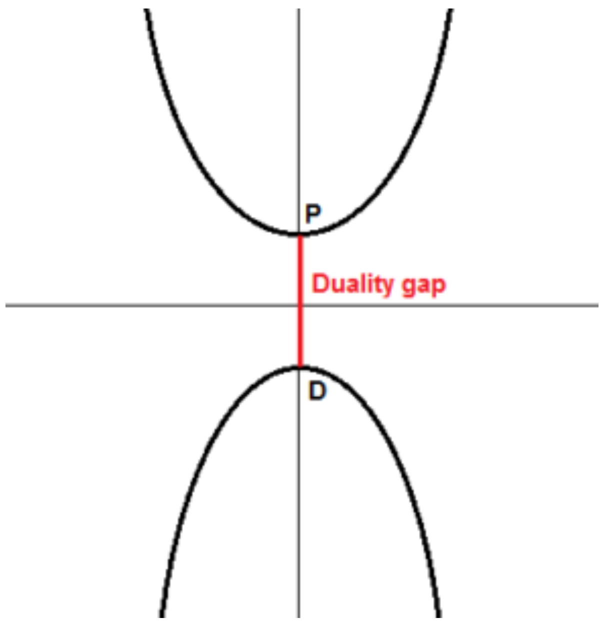
Define a Dual problem

$$\max_{\alpha,\beta:\alpha_i\geq 0} O_D(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i\geq 0}\left[\min_{w} L(w,\alpha,\beta)\right]$$

$$d^* = \max_{\alpha,\beta:\alpha_i\geq 0}\left[\min_{w} L(w,\alpha,\beta)\right]$$

Generalized Lagrangian

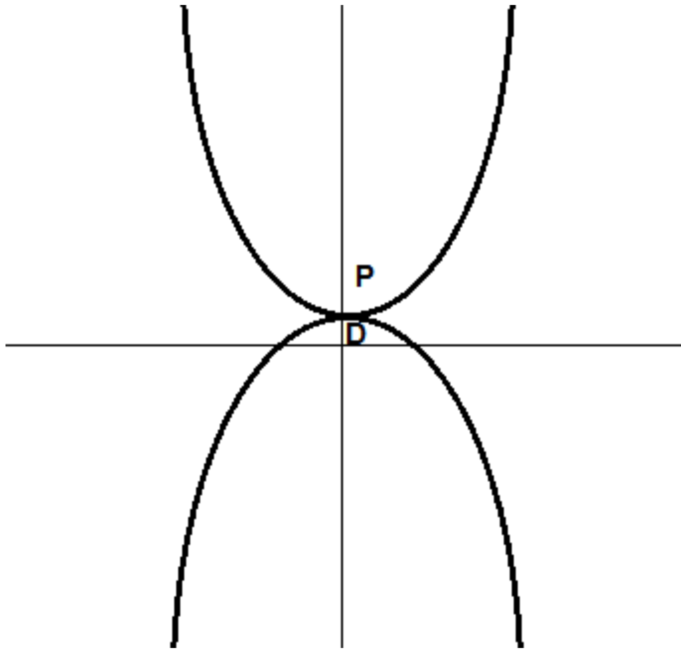
Define a Dual problem



$$\max_{\alpha,\beta:\alpha_i\geq 0} O_D(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i\geq 0}\left[\min_w L(w,\alpha,\beta)\right]$$

$$d^* = \max_{\alpha,\beta:\alpha_i\geq 0}\left[\min_w L(w,\alpha,\beta)\right]$$

$d^* \leq p^*$: always true

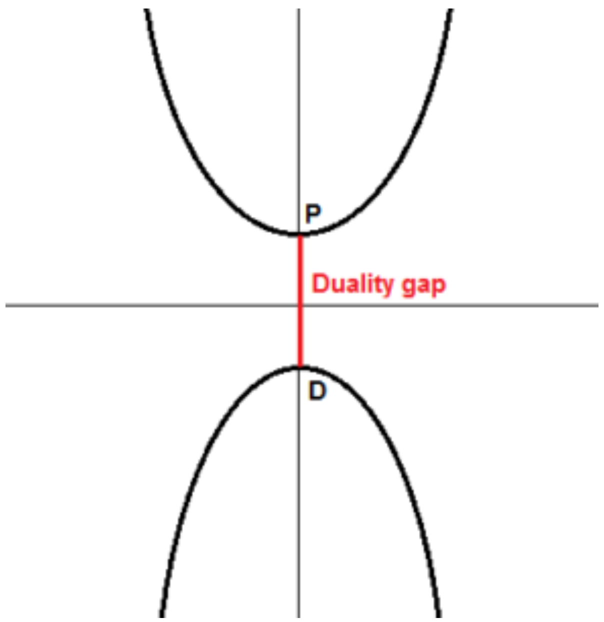


$d^* = p^*$: sometimes true. Under certain condition

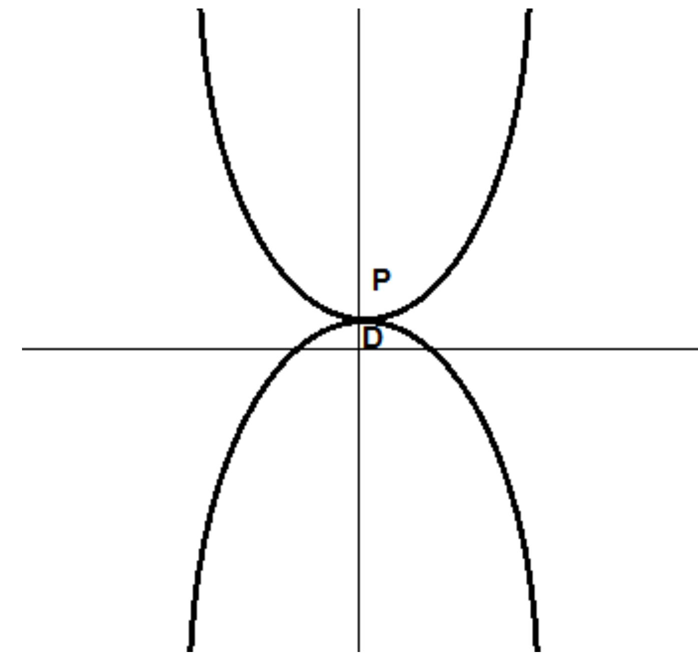
Generalized Lagrangian

Define a Dual problem

$$\max_{\alpha,\beta:\alpha_i\geq 0} O_D(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i\geq 0}\left[\min_w L(w,\alpha,\beta)\right]$$



$$d^* = \max_{\alpha,\beta:\alpha_i\geq 0}\left[\min_w L(w,\alpha,\beta)\right]$$



$d^* \leq p^*$: always true

$d^* = p^*$: sometimes true. Under certain condition

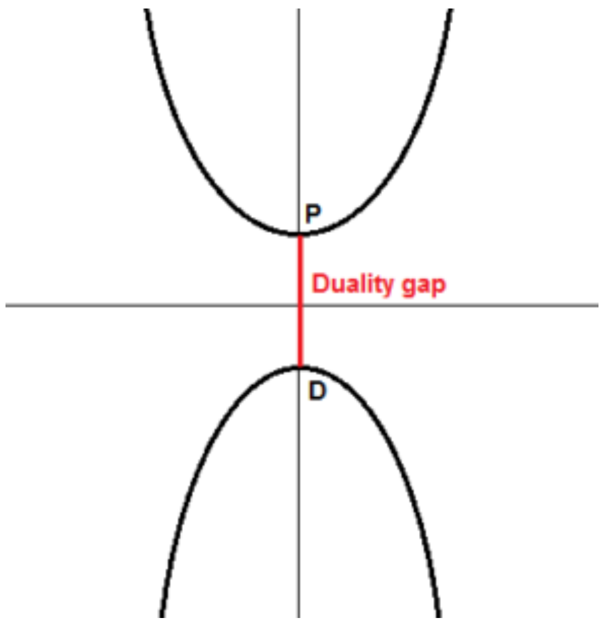
f, g_i: are convex

h_i :affine ($h_i(w) = a_i^T w + b_i$)

g_i are strictly possible (i.e., there exists some w such that $g_i(w) \leq 0$

Generalized Lagrangian

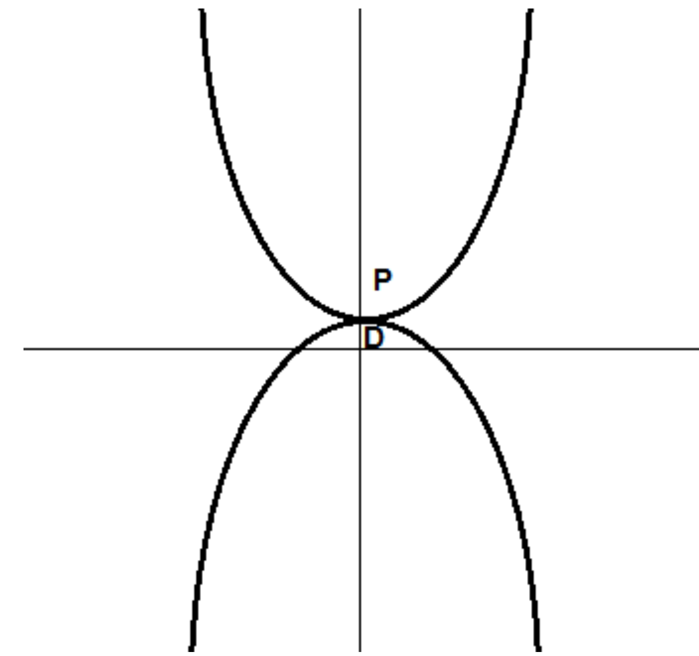
Define a Dual problem



$$\max_{\alpha,\beta:\alpha_i\geq 0} O_D(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i\geq 0}\left[\min_w L(w,\alpha,\beta)\right]$$

$$d^* = \max_{\alpha,\beta:\alpha_i\geq 0}\left[\min_w L(w,\alpha,\beta)\right]$$

$d^* \leq p^*$: always true



$d^* = p^*$: sometimes true. Under certain condition

f, g_i: are convex

h_i :affine ($h_i(w) = a_i^T w + b_i$)

g_i are strictly possible (i.e., there exists some w such that $g_i(w) \leq 0$

Then there exists w^*, α^*, β^*, such that

w^* solves primal problem

α^*, β^* solves the dual problem

$d^* = p^*$

Generalized Lagrangian

The solution $w*, \alpha*, \beta*$ satisfies the KKT (Karush-Kuhn-Tucker) Conditions

$$\frac{\partial}{\partial w_i} L(w*, \alpha*, \beta*) = 0 \qquad i = 1, \ldots, d$$

$$\frac{\partial}{\partial \beta_i} L(w*, \alpha*, \beta*) = 0 \qquad i = 1, \ldots, e$$

$$\alpha_i^* g_i(w*) = 0 \qquad i = 1, \ldots, k$$

$$g_i(w*) \leq 0 \qquad i = 1, \ldots, k$$

$$\alpha_i^* \geq 0 \qquad i = 1, \ldots, k$$

Generalized Lagrangian

The solution w^*, α^*, β^* satisfies the KKT (Karush-Kuhn-Tucker) Conditions

$$
\begin{aligned}
\frac{\partial}{\partial w_i} L(w^*, \alpha^*, \beta^*) &= 0 \qquad && i = 1, \ldots, d \\
\frac{\partial}{\partial \beta_i} L(w^*, \alpha^*, \beta^*) &= 0 \qquad && i = 1, \ldots, e \\
\alpha_i^* g_i(w^*) &= 0 \qquad && i = 1, \ldots, k \\
g_i(w^*) &\leq 0 \qquad && i = 1, \ldots, k \\
\alpha_i^* &\geq 0 \qquad && i = 1, \ldots, k
\end{aligned}
$$

Complementarity Condition

$$\alpha_i > 0 \implies g_i(w^*) = 0$$

We say that constraint $g_i(w^*) \leq 0$ is **active**

Support Vector Machines

Support Vector Machines

$$\min_{w,b} \quad \frac{1}{2}||w||^2$$
$$\text{s.t.} \quad y^i(w^T x^i + b) \geq 1, \qquad \forall i = 1,\ldots,n$$

The Lagrangian is given by

$$L(\alpha, w, b) = \frac{1}{2}||w||^2 - \sum_{i=1}^{N} \alpha_i \left(y^i(w^T x^i + b) - 1 \right)$$

The Dual of the problem is given by

$$O_D(\alpha) = \min_{w,b} L(\alpha, w, b)$$

Complementarity Condition

$$\nabla_w L(\alpha, w, b) = w - \sum_{i=1}^{N} \alpha_i y^i x^i = 0 \implies w = \sum_{i=1}^{N} \alpha_i y^i x^i$$

$$\nabla_b L(\alpha, w, b) = -\sum_{i=1}^{N} \alpha_i y^i = 0$$

$$\alpha_i[y^i(w^T x^i + b) - 1] = 0$$

Plug the value of w back into the Lagrangian

Support Vector Machines

The Lagrangian is given by

$$L(\alpha, w, b) = \frac{1}{2}||w||^2 - \sum_{i=1}^{N} \alpha_i \left(y^i(w^T x^i + b) - 1 \right)$$

$$L(\alpha, w, b) = \underbrace{\frac{1}{2}\left|\left| \sum_{i=1}^{N} \alpha_i y^i x^i w \right|\right|^2 - \sum_{i,j=1}^{N} \alpha_i \alpha_j y^i y^j (x^i \cdot x^j)}_{-\frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y^i y^j (x^i \cdot x^j)} - \sum_{i=1}^{N} \alpha_i y^i b + \sum_{i=1}^{N} \alpha_i$$

$$\nabla_w L(\alpha, w, b) = w - \sum_{i=1}^{N} \alpha_i y^i x^i = 0 \quad \Longrightarrow \quad w = \sum_{i=1}^{N} \alpha_i y^i x^i$$

$$\nabla_b L(\alpha, w, b) = -\sum_{i=1}^{N} \alpha_i y^i = 0$$

$$\alpha_i [y^i(w^T x^i + b) - 1] = 0$$

Support Vector Machines

$$\max_{\alpha} L(\alpha, w, b) = \max_{\alpha} \sum_{i=1}^{N} \alpha_i - \sum_{i,j=1}^{N} \alpha_i \alpha_j y^i y^j (x^i \cdot x^j)$$

$$\text{s.t.} \quad \begin{aligned} &\alpha_i \geq 0 \\ &\sum_{i=1}^{N} \alpha_i y^i = 0 \end{aligned}$$

This is Dual formulation of SVMs

We can solve the Dual problem and find α^* instead of the Primal problem

$$\min_{w,b} \quad \frac{1}{2} ||w||^2 \longrightarrow \quad w^* = \sum_{i=1}^{N} \alpha_i^* y^i x^i$$

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

$$b^* = -\frac{\max_{i, y^i = -1} (w^*)^T x^i + \min_{i, y^i = 1} (w^*)^T x^i}{2}$$

Support Vectors

Complementarity Condition

KKT conditions at the optimal solution

$$\nabla_w L(\alpha, w, b) = w - \sum_{i=1}^{N} \alpha_i y^i x^i = 0 \implies w = \sum_{i=1}^{N} \alpha_i y^i x^i$$

$$\nabla_b L(\alpha, w, b) = - \sum_{i=1}^{N} \alpha_i y^i = 0$$

$$\alpha_i [y^i (w^T x^i + b) - 1] = 0$$

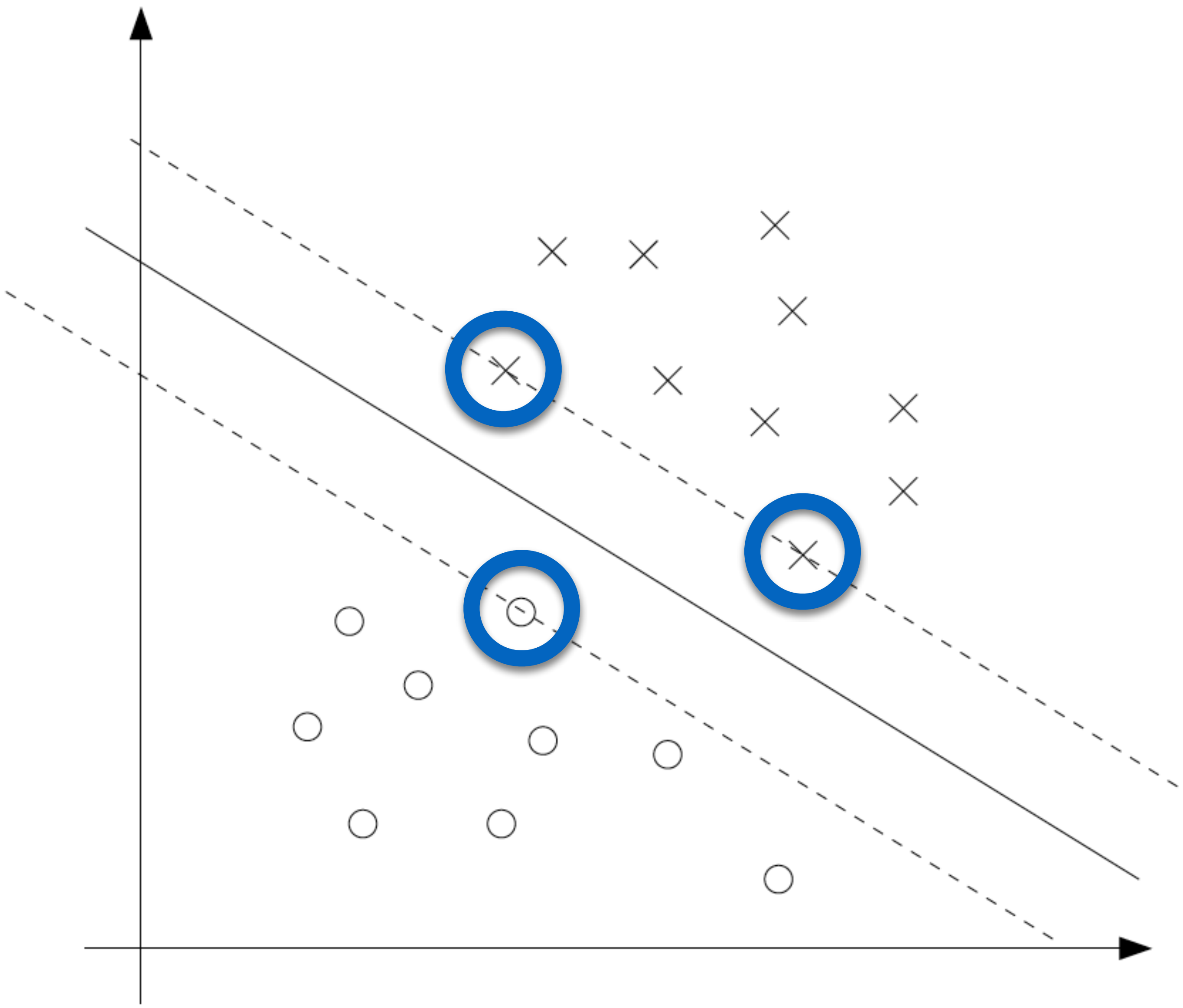
Either $\alpha_i = 0$ or $y^i (w^T x^i + b) - 1 = 0$

Thus for all i for which $\alpha_i \neq 0$ we have $y^i (w^T x^i + b) = 1$

The constraint is active

These are the points for which the Geometric Margin is equal to 1

Support Vectors



Constraint is active for only a few training points

These points are called the support vectors

Figure credits Benjamin Peherstorfer

Predictions with SVMs

We have solved the optimization problem and found the solution α^* and hence w^*

What is the class for a new data point x^0?

Naive approach

Compute $(w^*)^T \cdot x^0 + b^*$ and assign class +1 if positive otherwise assign class -1

There is a better way

$$
\begin{aligned}
(w^*)^T x^0 + b^* &= \left(\sum_{i=1}^{N} \alpha_i^* y^i x^i \right)^T \cdot x + b^* \\
&= \sum_{i=1}^{N} \alpha_i^* y^i ({x^i}^T \cdot x^0) + b^*
\end{aligned}
$$

You only need the inner products with the training samples!

Furthermore, α_i^* is non-zero for only a few training points (the support vectors)

Summary

Solve

$$\max_{\alpha} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} y^{(i)} y^{(j)} \alpha_i \alpha_j < x^{(i)}, x^{(j)} >$$

s.t. $\quad \alpha_i \geq 0\,, i = 1, \ldots, N$

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

Predict

$$\sum_{i=1}^{N} \alpha_i^* y^{(i)} < (x^{(i)})^T, x > + b$$

Inner products in SVMs is the key

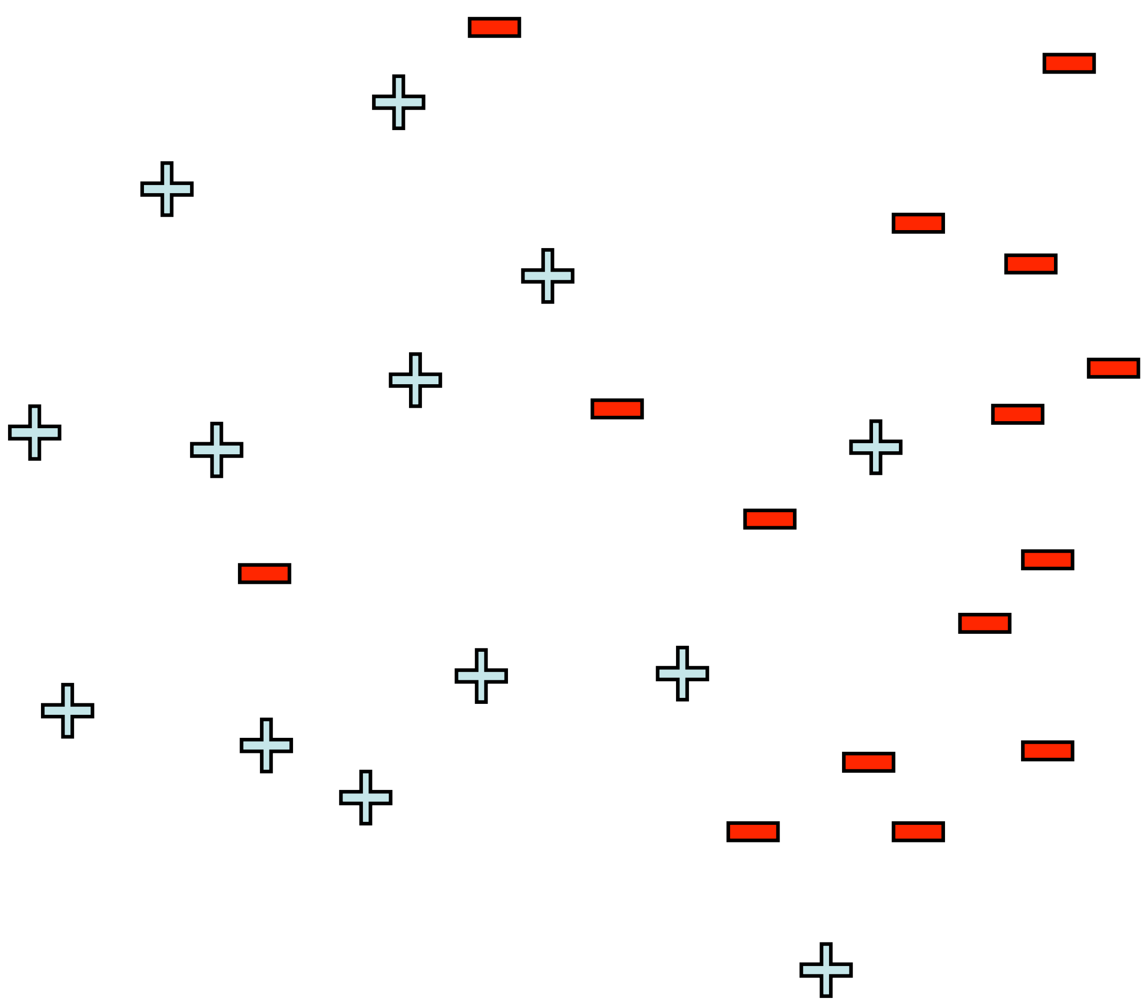
Also forms the basis of the Kernels in SVMs

When Linear Separability Does Not Exist

What happens when there is no linear separability?

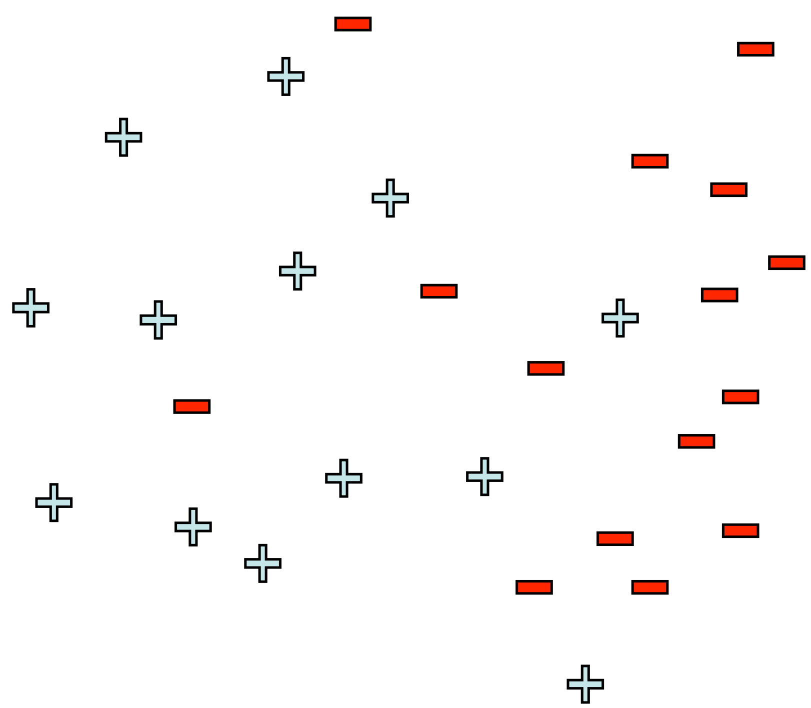
All the constraints get violated by w, b, γ

The geometric margin loses its meaning



When Linear Separability Does Not Exist

One possible solution is to find w, b such that the minimum number of constraints are violated



$$\min_{w,b} \quad \#mistakes$$

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

$$\min_{w,b} \sum_{i=1}^{N} l_{0,1}(y^i, w^T x^i + b)$$

$$where \quad l_{0,1}(y, \hat{y}) = 1 \quad if \quad \hat{y} \neq y$$

This is an NP-Hard Problem

When Linear Separability Does Not Exist

Another solution: Allow some slack!

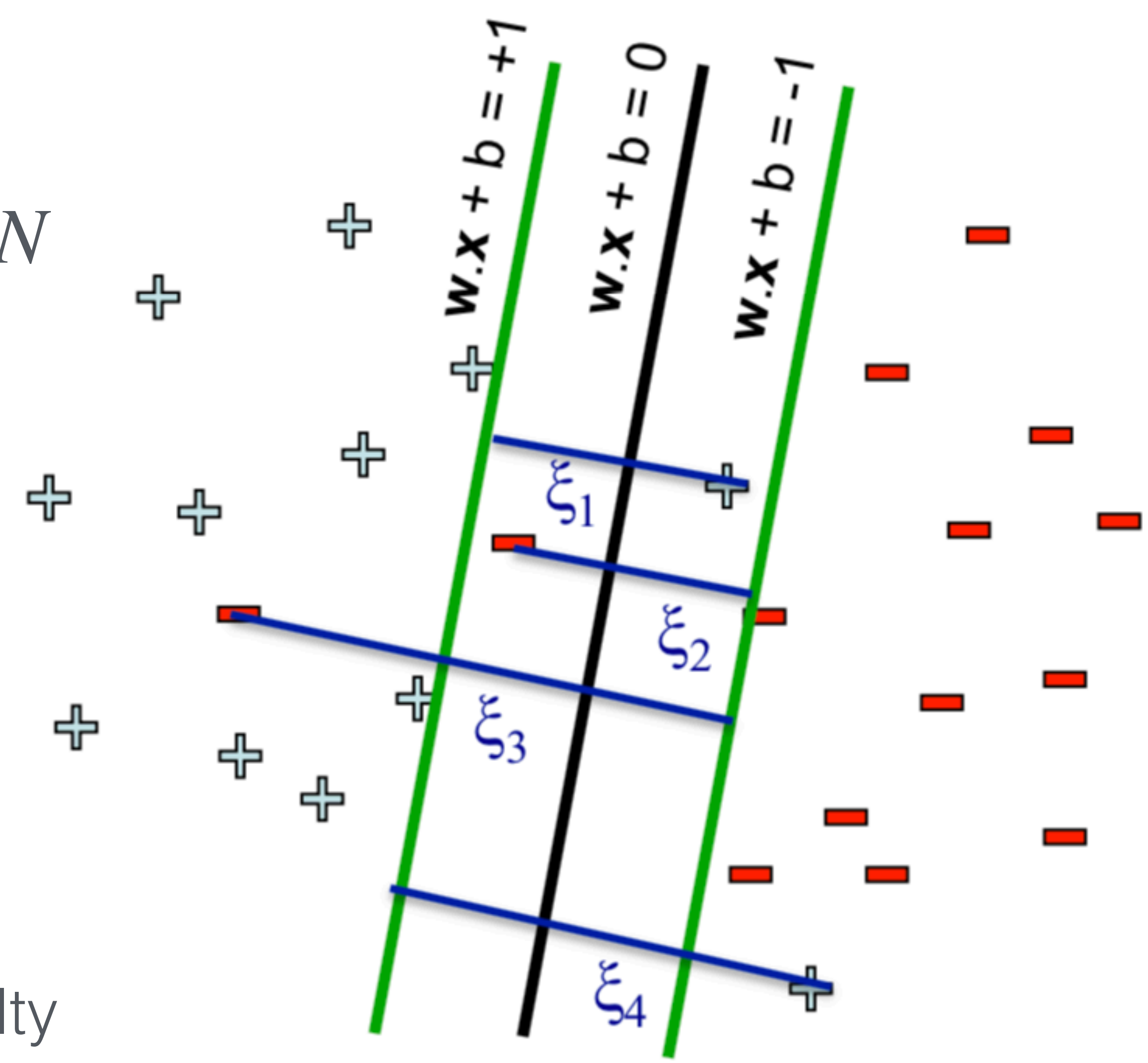
Let's ignore that we are looking for a largest margin classifier

Instead look for a classifier with the minimal slack

$$\min_{w,b,\xi_i} \quad \sum_{i=1}^{N} \xi_i$$

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

$$\xi_i \geq 0, i = 1,\ldots,N$$



If functional margin is ≥ 1 the no penalty

If functional margin is < 1 then pay linear penalty

When Linear Separability Does Not Exist

Another solution: Classifier with minimal slack

Optimal value of slack variables

$$\min_{w,b,\xi_i} \quad \sum_{i=1}^{N} \xi_i$$

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

$$\xi_i \geq 0, i = 1,\ldots,N$$

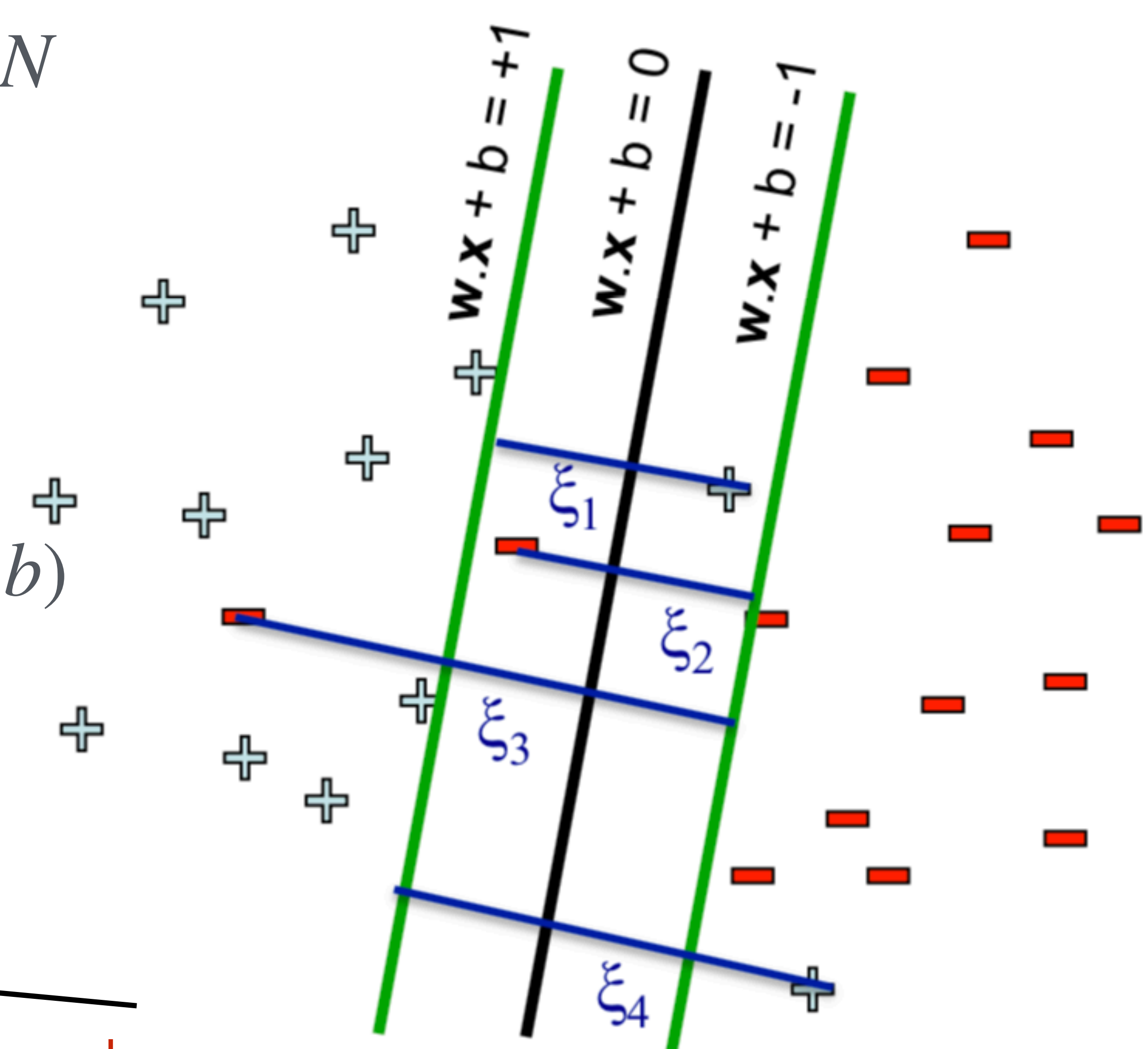
If $y^i(w^T x^i + b) \geq 1 \implies \xi_i = 0$

If $y^i(w^T x^i + b) < 1 \implies \xi_i = 1 - y^i(w^T x^i + b)$

These two conditions can be written as

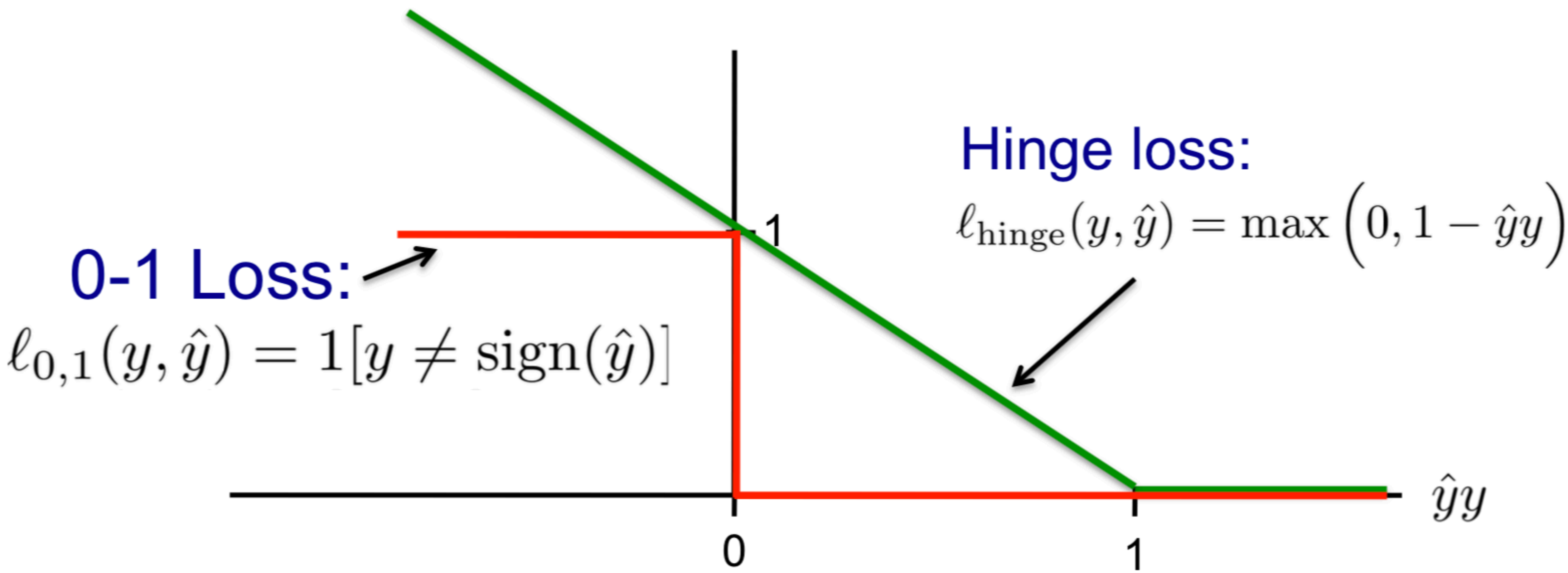
$$\xi_i = \max[0, 1 - y^i(w^T x^i + b)]$$

Hinge Loss



When Linear Separability Does Not Exist

This is the tightest convex upper bound of the intractable 0/1 loss



When Linear Separability Does Not Exist

With $\xi_i = \max(0,1 - y^i(w^T x^i + b))$ we can write the optimization problem as

$$\min_{w,b} \quad \sum_{i=1}^{N} \max(0,1 - y^i(\underbrace{w^T x^i + b}_{\hat{y}^i}))$$

$$\min_{w,b} \quad \sum_{i=1}^{N} L_{hinge}(y^i, w^T x^i + b)$$

SVMs Under No Linear Separability

We find the largest margin classifier with some slack

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

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

$$\xi_i \geq 0, \qquad i = 1,\ldots,N$$

Thus there are two terms in the objective function that are balanced by the slack penalty C

If $C = \infty$ the you have to separate the data

If $C = 0$ then completely ignore the data

This also servers as another regularizer parameter

SVMs Under No Linear Separability

Equivalent formulation via Hinge loss

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

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

$$\xi_i \geq 0, \qquad i = 1,\ldots,N$$

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} L_{hinge}(y^i, w^T x^i + b)$$

Regularizer to prevent over fitting

End of Lecture 07