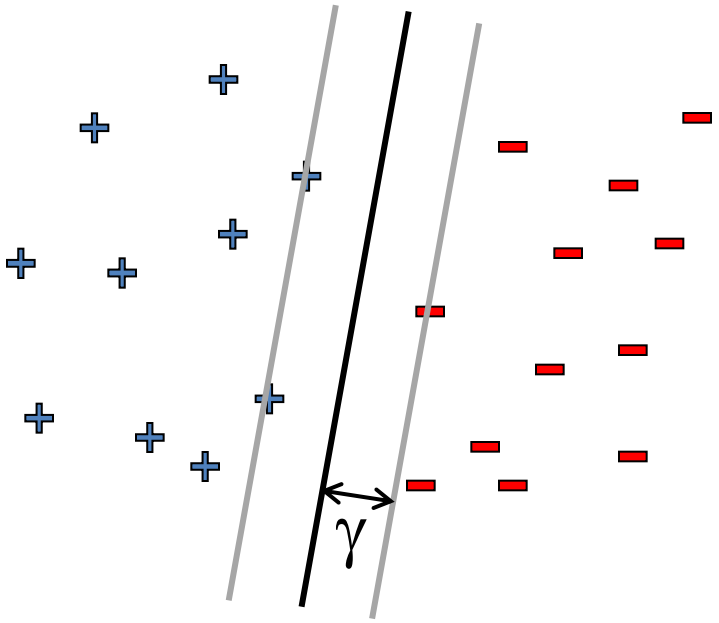
Support Vector Machines (SVMs) contd…

Aarti Singh

Machine Learning 10-315
Mar 23, 2022

Hard-margin SVM

Data perfectly separable by a linear decision boundary



Hard margin approach

$$\min_{\mathbf{w},b} \ \mathbf{w}.\mathbf{w}$$

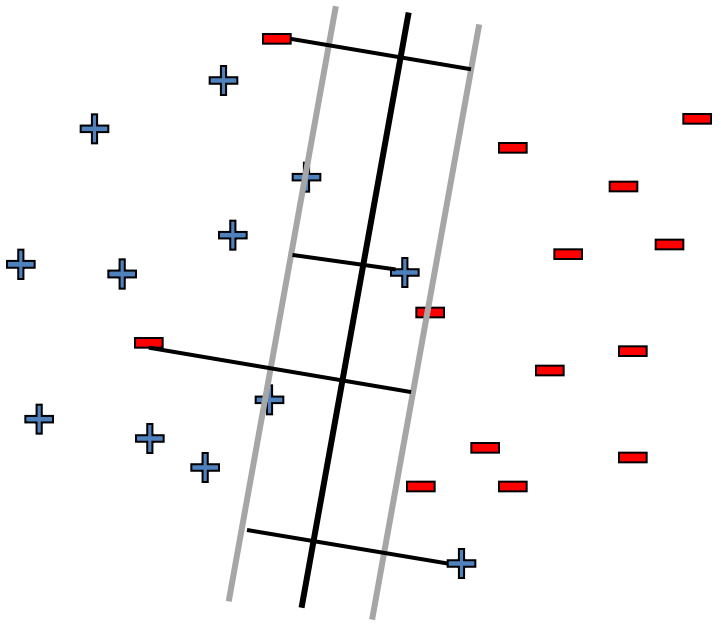
$$\text{s.t. } (\mathbf{w}.\mathbf{x}_j + b)\, y_j \geq 1 \quad \forall j$$

Solve using Quadratic Programming (QP)

Margin, $\gamma \ \alpha \ 1/\|w\|$

Soft-margin SVM

Allow “error” in classification



Soft margin approach

$$\min_{\mathbf{w},b,\{\xi_j\}} \mathbf{w}.\mathbf{w} + C \sum_j \xi_j$$

$$\text{s.t. } (\mathbf{w}.\mathbf{x}_j + b)\, y_j \geq 1 - \xi_j \quad \forall j$$

$$\xi_j \geq 0 \quad \forall j$$

ξ_j - “slack” variables
= (>1 if x_j misclassifed)
pay linear penalty if mistake

C - tradeoff parameter (chosen by cross-validation)

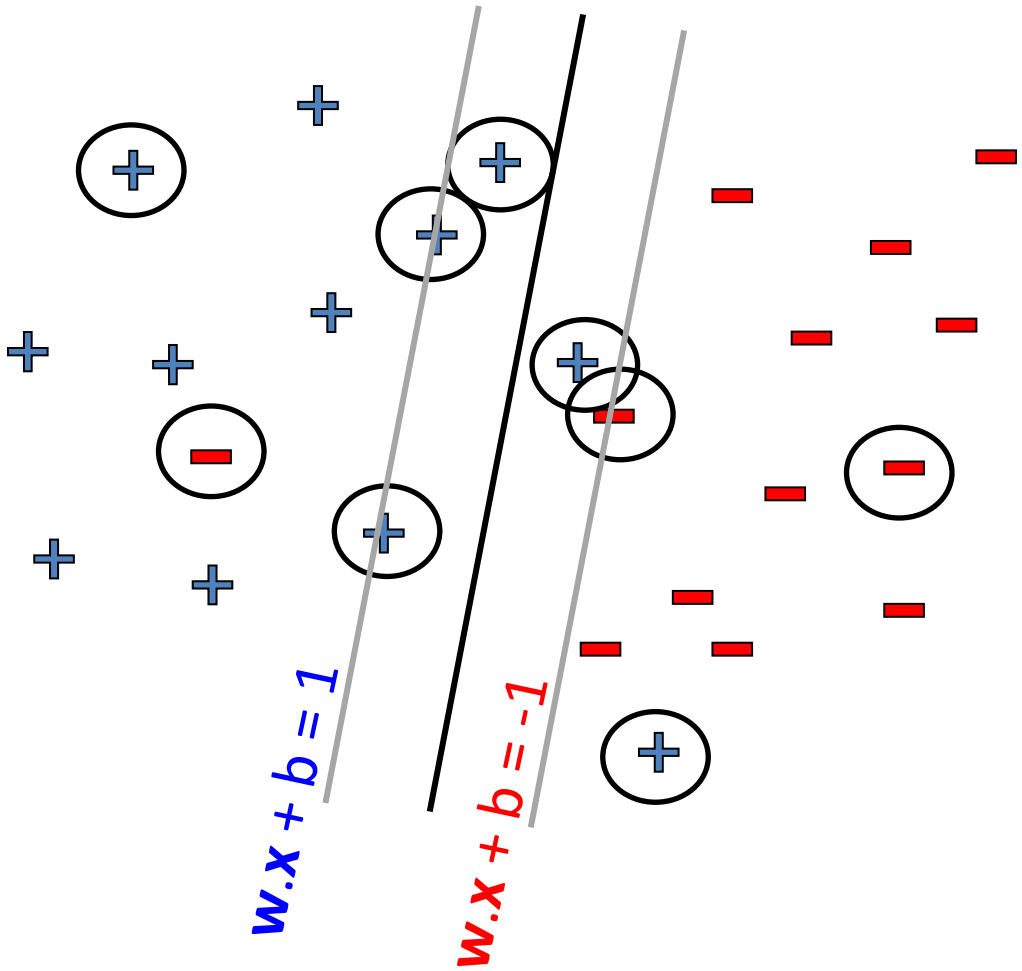
Still QP ☺

$$\min_{\mathbf{w},b,\{\xi_j\}} \quad \mathbf{w}.\mathbf{w} + C \Sigma \xi_j$$

$$\text{s.t. } (\mathbf{w}.\mathbf{x}_j+b)\, y_j \geq 1-\xi_j \quad \forall j$$

$$\xi_j \geq 0 \quad \forall j$$

Slack variables

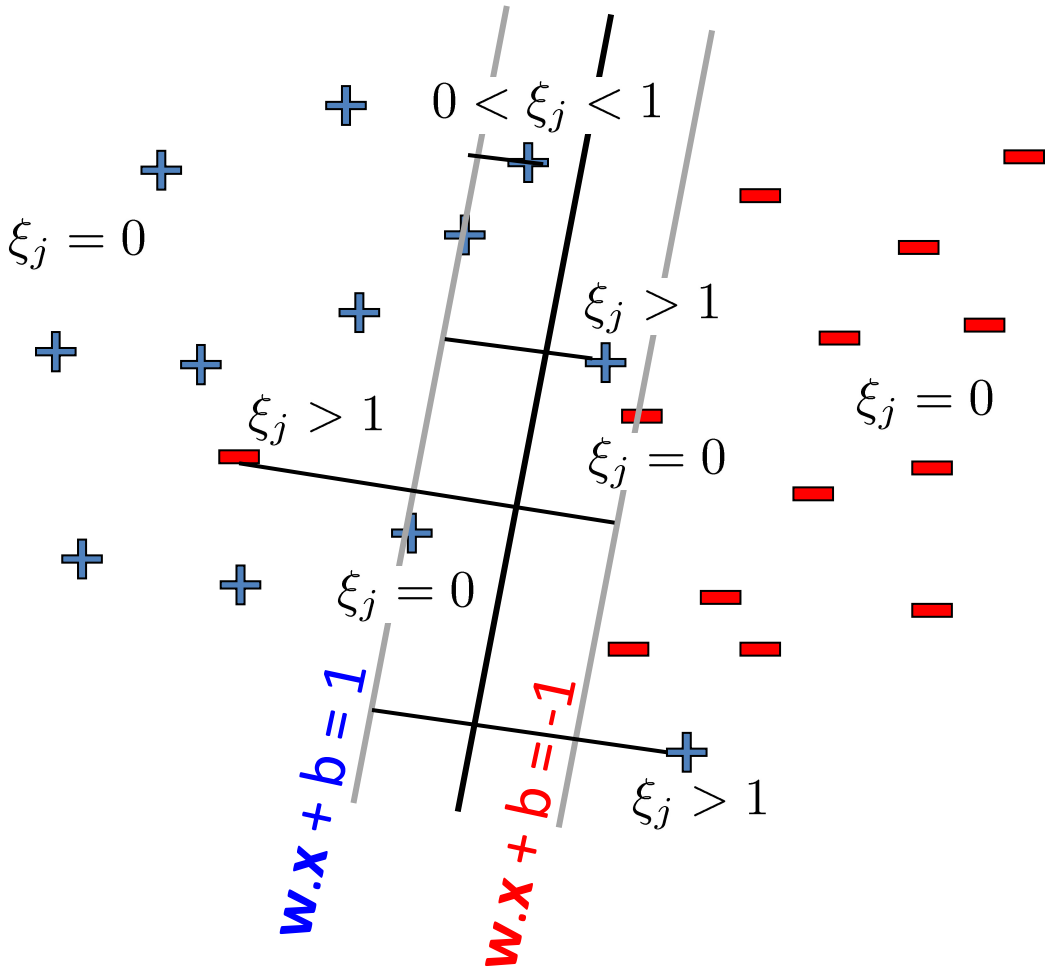


$$(\mathbf{w}.\mathbf{x}_j+b)\, y_j \geq 1-\xi_j \quad \forall j$$

What is the slack ξ_j for the following points?

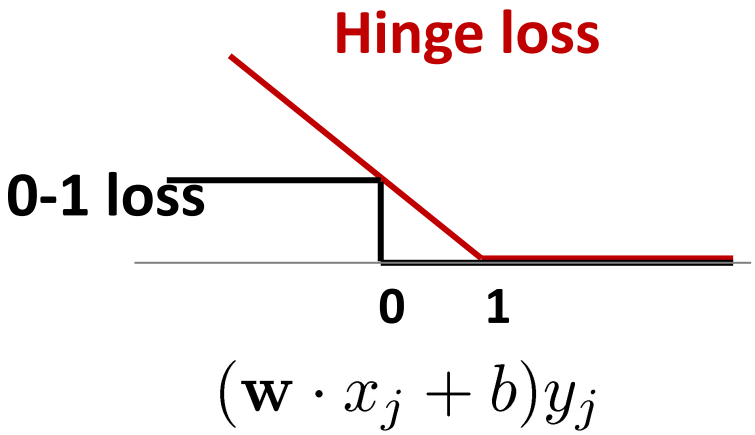
Confidence | Slack

Slack variables – Hinge loss



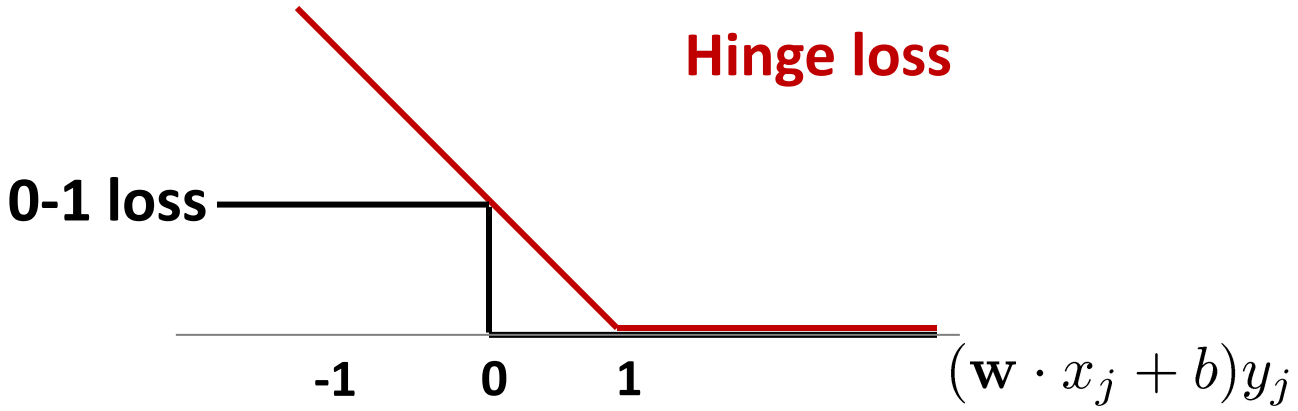
Notice that

$$\xi_j = (1 - (\mathbf{w} \cdot x_j + b)y_j))_+$$



Slack variables – Hinge loss

$$\xi_j = (1 - (\mathbf{w} \cdot x_j + b)y_j))_+$$



$$\min_{\mathbf{w},b,\{\xi_j\}} \mathbf{w}.\mathbf{w} + C \sum_j \xi_j$$

$$\text{s.t. } (\mathbf{w}.\mathbf{x}_j + b)\, y_j \geq 1 - \xi_j \quad \forall j$$

$$\xi_j \geq 0 \quad \forall j$$

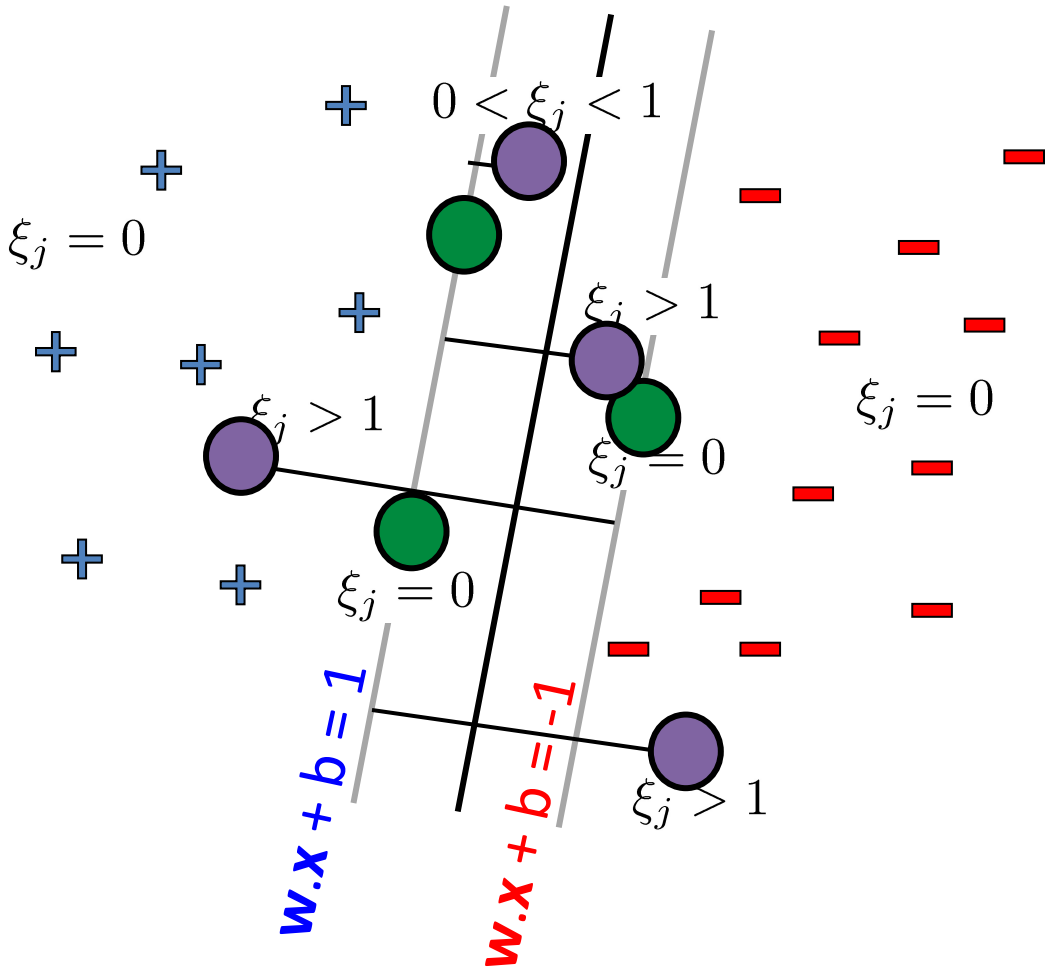
$\Leftrightarrow$

Regularized hinge loss

$$\min_{\mathbf{w},b} \mathbf{w}.\mathbf{w} + C \sum_j (1 - (\mathbf{w}.\mathbf{x}_j + b)y_j)_+$$

$$\min_{\mathbf{w},b,\{\xi_j\}} \quad \mathbf{w}.\mathbf{w} + C \Sigma \xi_j$$
$$\text{s.t. } (\mathbf{w}.\mathbf{x}_j + b)\, y_j \geq 1 - \xi_j \quad \forall j$$
$$\xi_j \geq 0 \quad \forall j$$

Support Vectors



Margin support vectors

$\xi_j = 0$, $(\mathbf{w}.\mathbf{x}_j + b)\, y_j = 1$
(don't contribute to objective but enforce constraints on solution)

Correctly classified but on margin

Non-margin support vectors

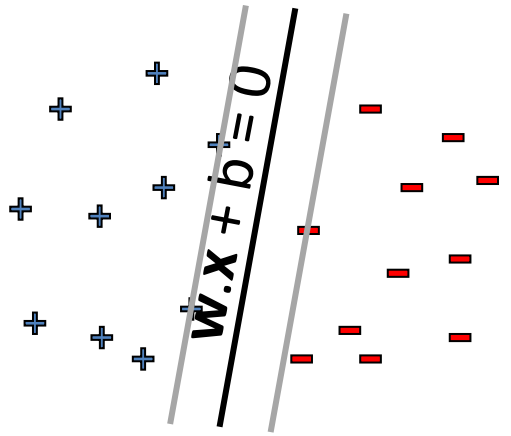
$\xi_j > 0$
(contribute to both objective and constraints)

$1 > \xi_j > 0$ Correctly classified but inside margin

$\xi_j > 1$ Incorrectly classified

SVM – linearly separable case

n training points ($\mathbf{x}_1, \ldots, \mathbf{x}_n$)
d features $\mathbf{x}_j$ is a d-dimensional vector



- Primal problem: $\text{minimize}_{\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}.\mathbf{w}$
  $\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1, \ \forall j$

w – weights on features (d-dim problem)

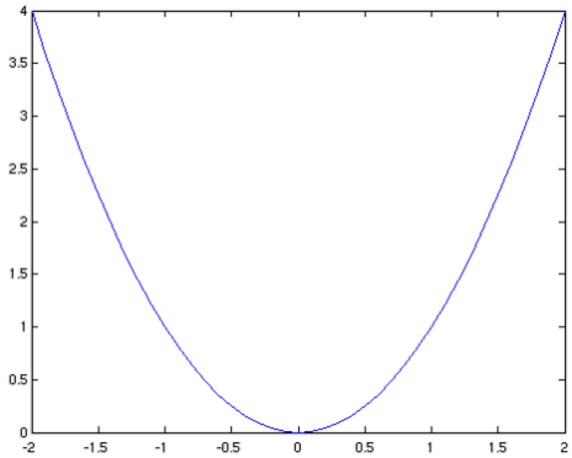
- Convex quadratic program – quadratic objective, linear constraints
- But expensive to solve if d is very large
- Often solved in dual form (n-dim problem)

Detour - Constrained Optimization

$$\min_x \; x^2 \quad \text{s.t.} \quad x \geq b$$

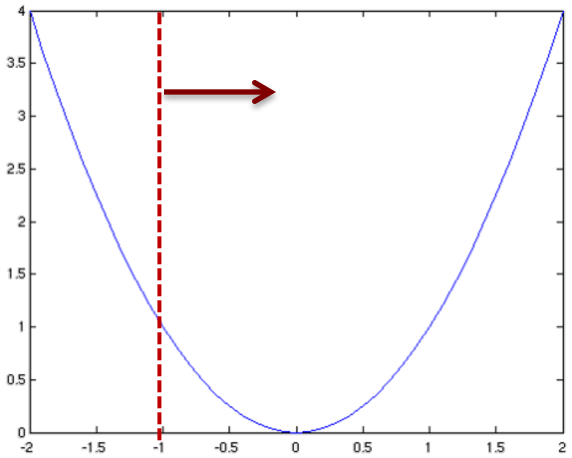
$$x^* = \max(b, 0)$$

$\min_x \; x^2$



$x^* = 0$

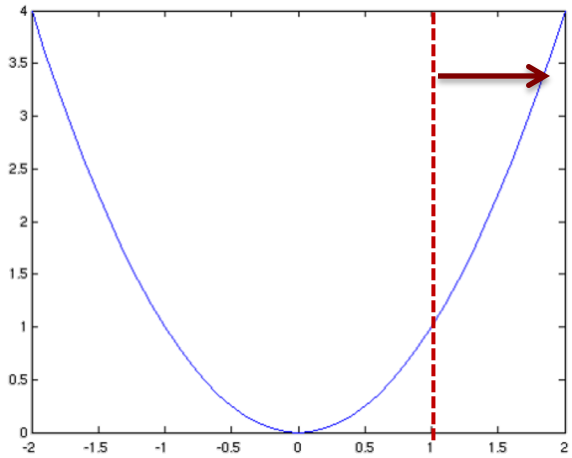
$\min_x \; x^2$ s.t. $x \geq -1$



$x^* = 0$

Constraint inactive

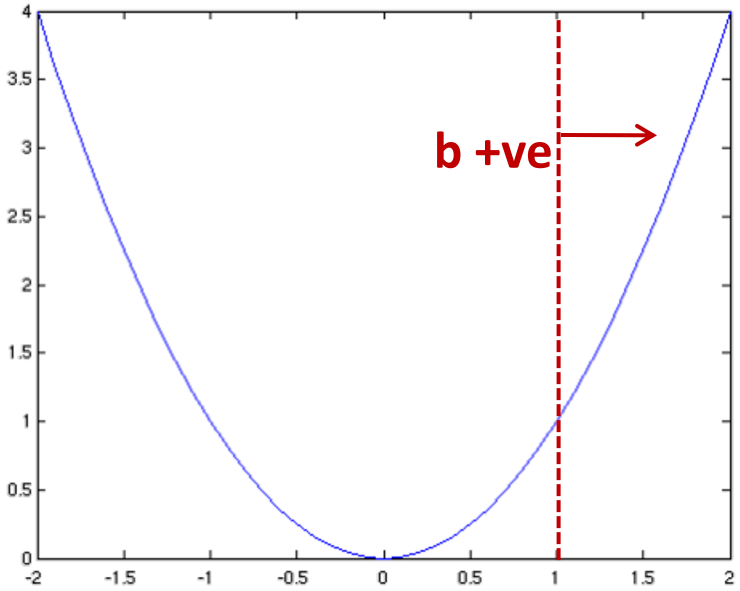
$\min_x \; x^2$ s.t. $x \geq 1$



$x^* = 1$

Constraint active (tight)

Constrained Optimization



$x^* = b$

$$\min_x \; x^2$$
$$\text{s.t.} \quad x \geq b$$

Equivalent unconstrained optimization:

$\min_x \; x^2 + I(x-b)$

Replace with lower bound ($\alpha >= 0$)

$x^2 + I(x-b) >= \underbrace{x^2 - \alpha(x-b)}_{L(x,\alpha)}$

Primal and Dual Problems

Primal problem: p* = $\min_x \; x^2$
s.t. $x \geq b$

Dual problem: d* = $\max_\alpha \; d(\alpha)$
s.t. $\alpha \geq 0$

= $\min_x \max_{\alpha \geq 0} L(x, \alpha)$

= $\max_\alpha \min_x L(x, \alpha)$
s.t. $\alpha \geq 0$

where Lagrangian $L(x, \alpha) = x^2 - \alpha(x - b)$

How to form the Lagrangian?

For each constraint, introduce a positive Lagrange multiplier

Fold constraints into objective

Why solve the Dual problem?

Primal problem: p* = $\min_x \; x^2$ s.t. $x \geq b$

Dual problem: d* = $\max_\alpha \; d(\alpha)$ s.t. $\alpha \geq 0$

$= \min_x \max_{\alpha \geq 0} L(x, \alpha)$

$= \max_\alpha \min_x L(x, \alpha)$ s.t. $\alpha \geq 0$

- **Dual problem (maximization) is always concave even if primal is not convex**

  Why? Pointwise infimum of concave functions is concave.
  [Pointwise supremum of convex functions is convex.]

$$L(x, \alpha) = x^2 - \alpha(x - b)$$

- **As many dual variables α as constraints, helpful if fewer constraints than dimension of primal variable x**

Connection between Primal and Dual

Primal problem: p* = $\min_x \; x^2$ s.t. $x \geq b$

Dual problem: d* = $\max_\alpha \; d(\alpha)$ s.t. $\alpha \geq 0$

- **Weak duality:** The dual solution d* lower bounds the primal solution p* i.e. d* ≤ p*

To see this, recall $L(x, \alpha) = x^2 - \alpha(x - b)$

For every feasible x' (i.e. x' ≥ b) and feasible α' (i.e. α' ≥ 0) , notice that

$$d(\alpha) = \min_x L(x, \alpha) \leq x'^2 - \alpha'(x'-b) \leq x'^2$$

Since above holds true for every feasible x', we have $d(\alpha) \leq x^{*2} = p^*$

Connection between Primal and Dual

Primal problem: p* = $\min_x \; x^2$ s.t. $x \geq b$

Dual problem: d* = $\max_\alpha \; d(\alpha)$ s.t. $\alpha \geq 0$

- **Weak duality:** The dual solution d* lower bounds the primal solution p* i.e. d* ≤ p*
- **Strong duality:** d* = p* holds often for many problems of interest e.g. if the primal is a feasible convex objective with linear constraints

Connection between Primal and Dual

What does strong duality say about α^* (the α that achieved optimal value of dual) and x^* (the x that achieves optimal value of primal problem)?

Whenever strong duality holds, the following conditions (known as KKT conditions) are true for α^* and x^*:

- 1. $\bigtriangledown L(x^*, \alpha^*) = 0$ i.e. Gradient of Lagrangian at x^* and α^* is zero.
- 2. $x^* \geq b$ i.e. x^* is primal feasible
- 3. $\alpha^* \geq 0$ i.e. α^* is dual feasible
- 4. $\alpha^*(x^* - b) = 0$ (called as complementary slackness)

We use the first one to relate x^* and α^*. We use the last one (complimentary slackness) to argue that $\alpha^* = 0$ if constraint is inactive and $\alpha^* > 0$ if constraint is active and tight.

Primal and Dual Problems

Primal problem: p* = $\min_x \; x^2$
s.t. $x \geq b$

Dual problem: d* = $\max_\alpha \; d(\alpha)$
s.t. $\alpha \geq 0$

= $\min_x \max_{\alpha \geq 0} L(x, \alpha)$

= $\max_\alpha \min_x L(x, \alpha)$
s.t. $\alpha \geq 0$

where Lagrangian $L(x, \alpha) = x^2 - \alpha(x - b)$

How to form the Lagrangian?

For each constraint, introduce a positive Lagrange multiplier

Fold constraints into objective

Dual SVM – linearly separable case

n training points, d features

($\mathbf{x}_1$, …, $\mathbf{x}_n$) where x_i is a d-dimensional vector

- Primal problem:

$$\text{minimize}_{\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}.\mathbf{w}$$
$$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1, \ \forall j$$

w – weights on features (d-dim problem)

- Dual problem (derivation):

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}.\mathbf{w} - \sum_j \alpha_j \left[\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j - 1\right]$$
$$\alpha_j \geq 0, \ \forall j$$

α – weights on training pts (n-dim problem)

Dual SVM – linearly separable case

- Dual problem:

$$\max_{\alpha} \min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}.\mathbf{w} - \sum_j \alpha_j \left[\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j - 1\right]$$

$$\alpha_j \geq 0, \ \forall j$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \qquad \Rightarrow \mathbf{w} = \sum_j \alpha_j y_j \mathbf{x}_j$$

$$\frac{\partial L}{\partial b} = 0 \qquad \Rightarrow \sum_j \alpha_j y_j = 0$$

If we can solve for αs (dual problem), then we have a solution for **w** (primal problem)

Dual SVM – linearly separable case

- Dual problem:

$$\max_{\alpha} \min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}.\mathbf{w} - \sum_j \alpha_j \left[\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j - 1\right]$$

$$\alpha_j \geq 0, \ \forall j$$

$$\Rightarrow \mathbf{w} = \sum_j \alpha_j y_j \mathbf{x}_j \qquad \Rightarrow \sum_j \alpha_j y_j = 0$$

Dual SVM – linearly separable case

$$\text{maximize}_{\alpha} \quad \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i . \mathbf{x}_j$$

$$\sum_i \alpha_i y_i = 0$$

$$\alpha_i \geq 0$$

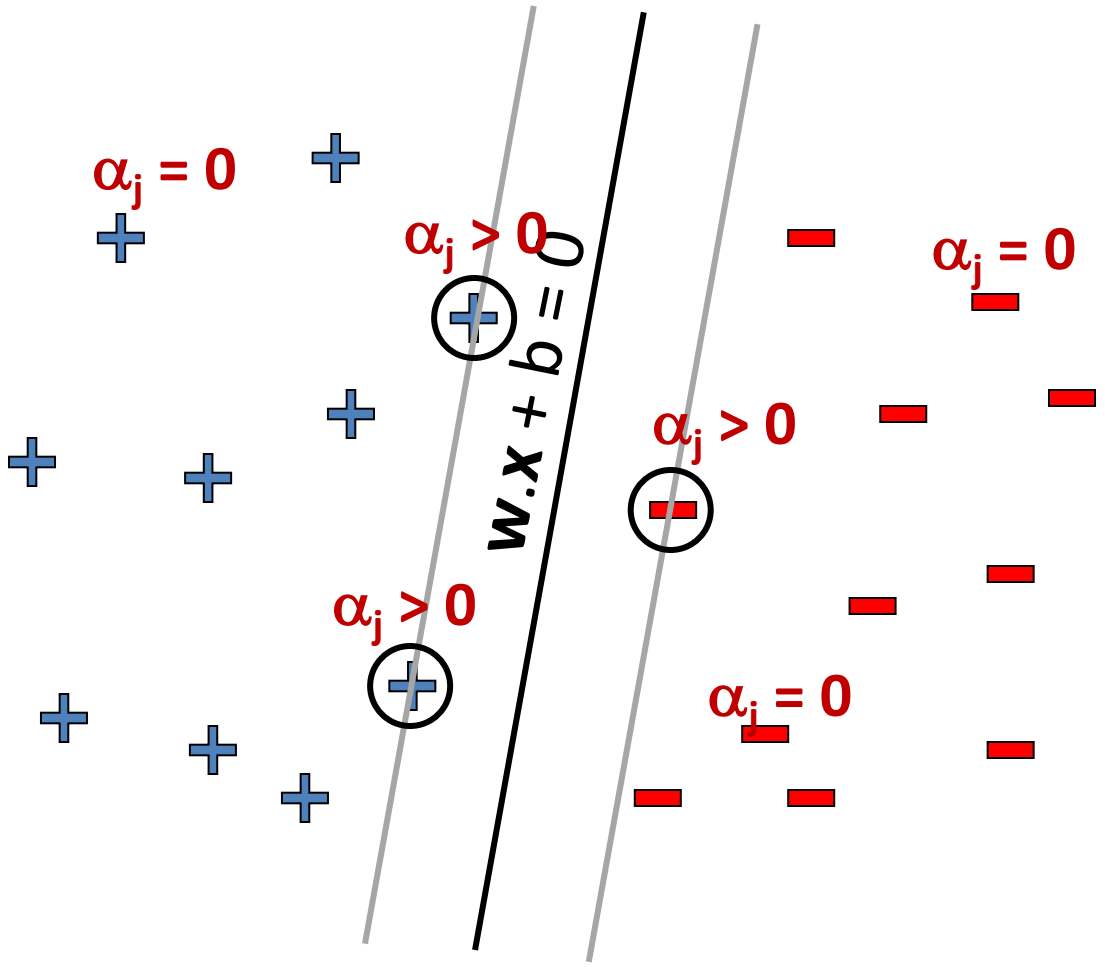
Dual problem is also QP

Solution gives α_js

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

What about b?

Dual SVM: Sparsity of dual solution



$$\mathbf{w} = \sum_j \alpha_j y_j \mathbf{x}_j$$

Complementary slackness implies

Only few α_js can be non-zero : where constraint is active and tight

$$(\mathbf{w}.\mathbf{x}_j + b)y_j = 1$$

Support vectors – training points j whose α_js are non-zero

Dual SVM – linearly separable case

$$\text{maximize}_{\alpha} \quad \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i . \mathbf{x}_j$$

$$\sum_i \alpha_i y_i = 0$$
$$\alpha_i \geq 0$$

Dual problem is also QP

Solution gives α_js ⟶

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

$$b = y_k - \mathbf{w}.\mathbf{x}_k$$

for any k where $\alpha_k > 0$

Use any one of support vectors with α_k>0 to compute b since constraint is tight $(\mathbf{w}.\mathbf{x}_k + b)y_k = 1$

So why solve the dual SVM?

- There are some quadratic programming algorithms that can solve the dual faster than the primal, (specially in high dimensions d>>n)

- But, more importantly, the "**kernel trick**"!!!