CMSC 478
Machine Learning

KMA Solaiman
ksolaima@umbc.edu

(originally prepared by Tommi Jaakkola, MIT CSAIL)

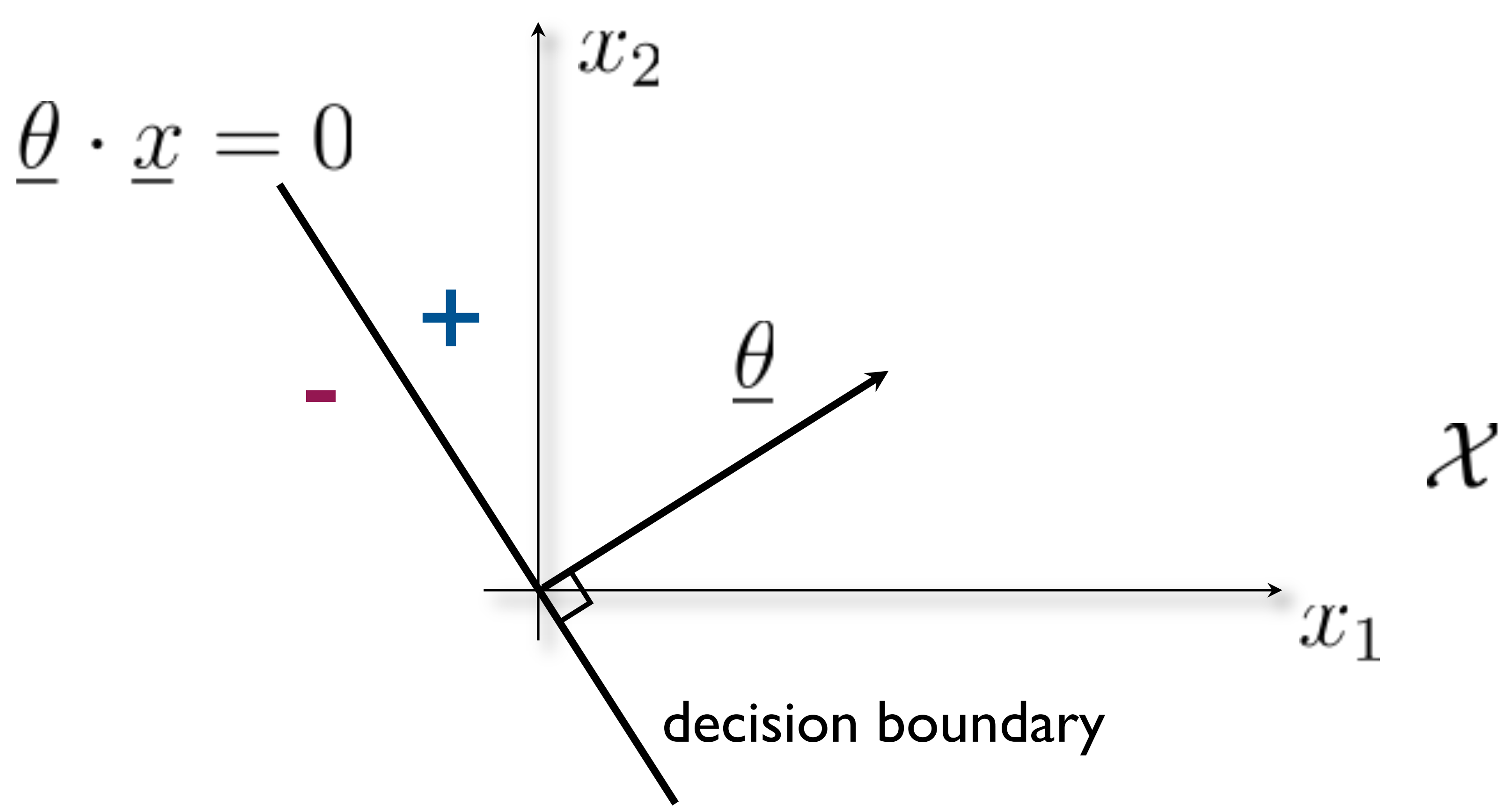
Today's topics

- Perceptron, convergence
  - the prediction game
  - mistakes, margin, and generalization
- Maximum margin classifier -- support vector machine
  - estimation, properties
  - allowing misclassified points

Recall: linear classifiers

- A linear classifier (through origin) with parameters $\underline{\theta}$ divides the space into positive and negative halves

$$
\begin{aligned}
f(\underline{x};\underline{\theta}) &= \text{sign}\big(\,\underline{\theta}\cdot\underline{x}\,\big) = \text{sign}\big(\,\underbrace{\theta_1 x_1 + \ldots + \theta_d x_d}_{\text{discriminant function}}\,\big) \\
&= \begin{cases} +1, & \text{if } \underline{\theta}\cdot\underline{x} > 0 \\ -1, & \text{if } \underline{\theta}\cdot\underline{x} \le 0 \end{cases}
\end{aligned}
$$



The perceptron algorithm

- A sequence of examples and labels

$$(\underline{x}_t, y_t), \quad t = 1, 2, \ldots$$

- The perceptron algorithm applied to the sequence

$$\text{Initialize: } \underline{\theta} = 0$$

$$\text{For } t = 1, 2, \ldots$$

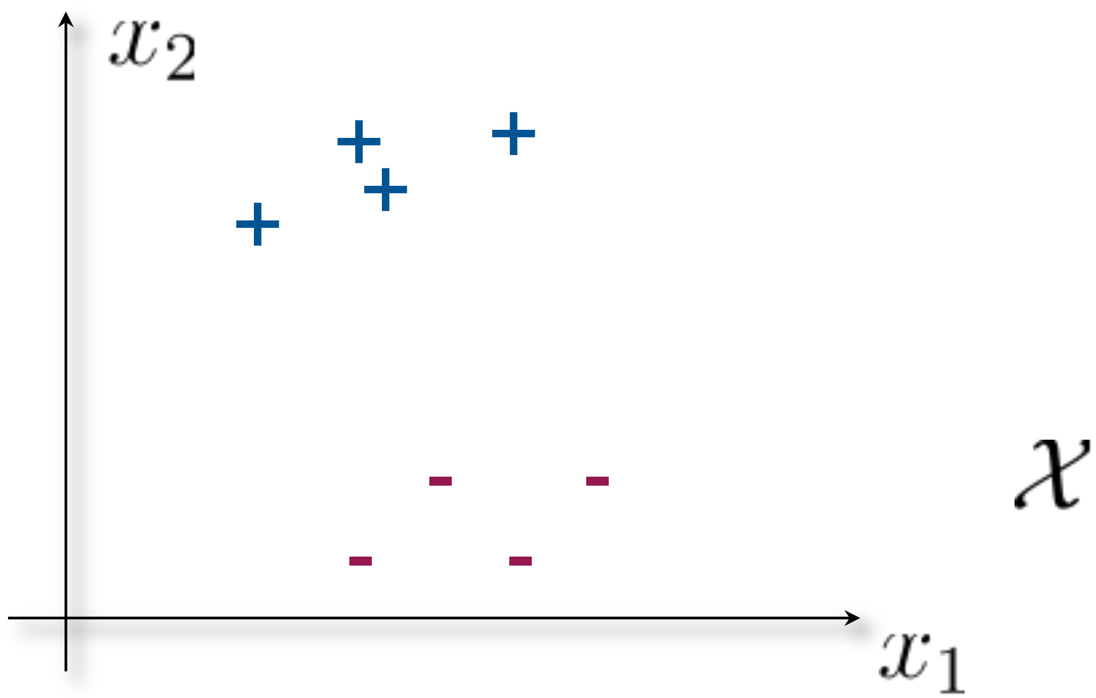
$$\text{if } y_t(\underline{\theta} \cdot \underline{x}_t) \leq 0 \text{ (mistake)}$$

$$\underline{\theta} \leftarrow \underline{\theta} + y_t \underline{x}_t$$

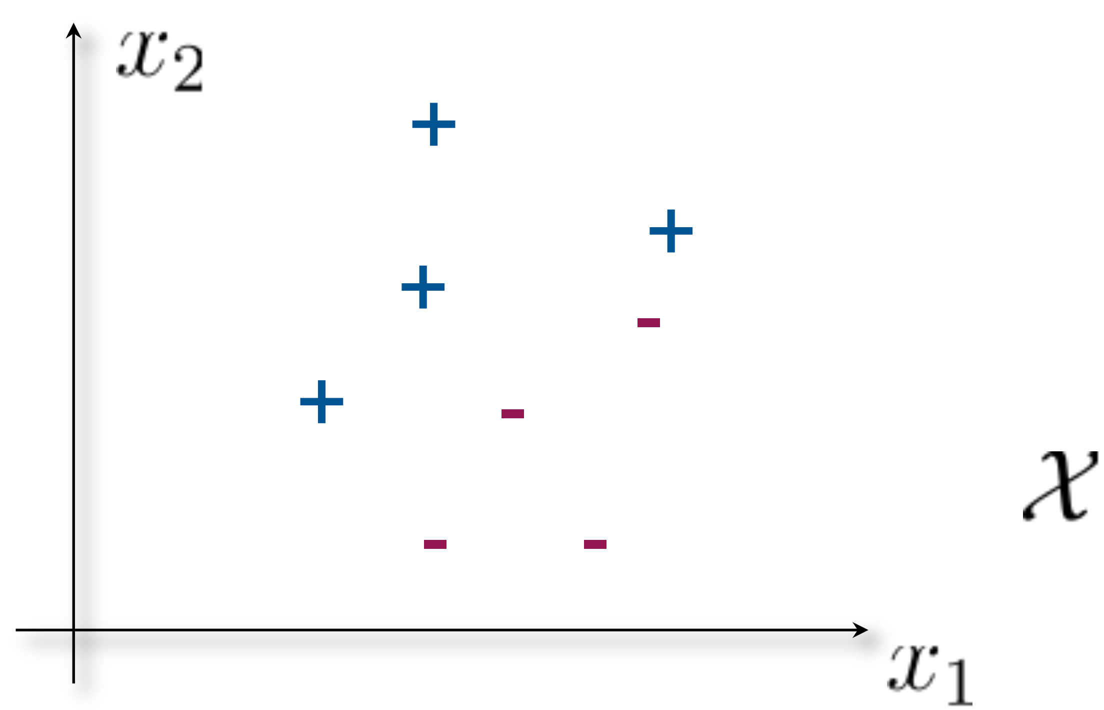
- We would like to bound the number of mistakes that the algorithm makes

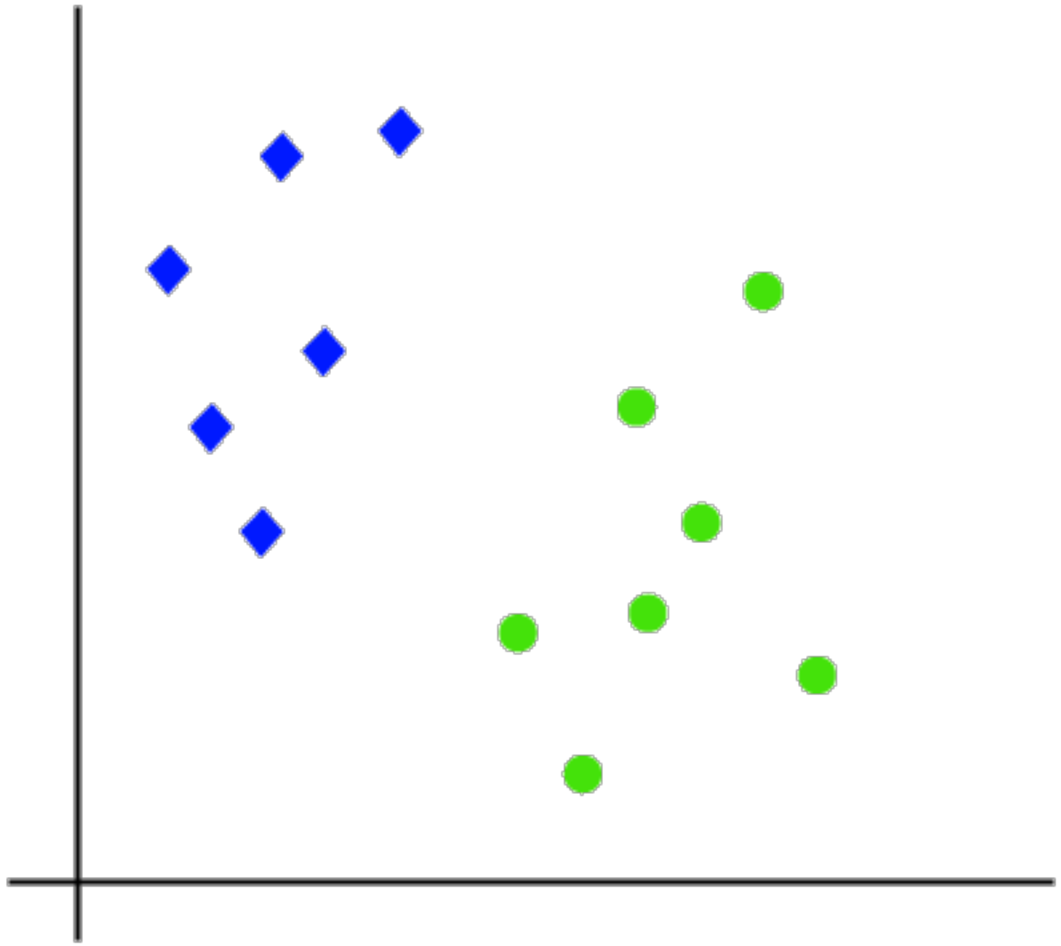
Mistakes and margin

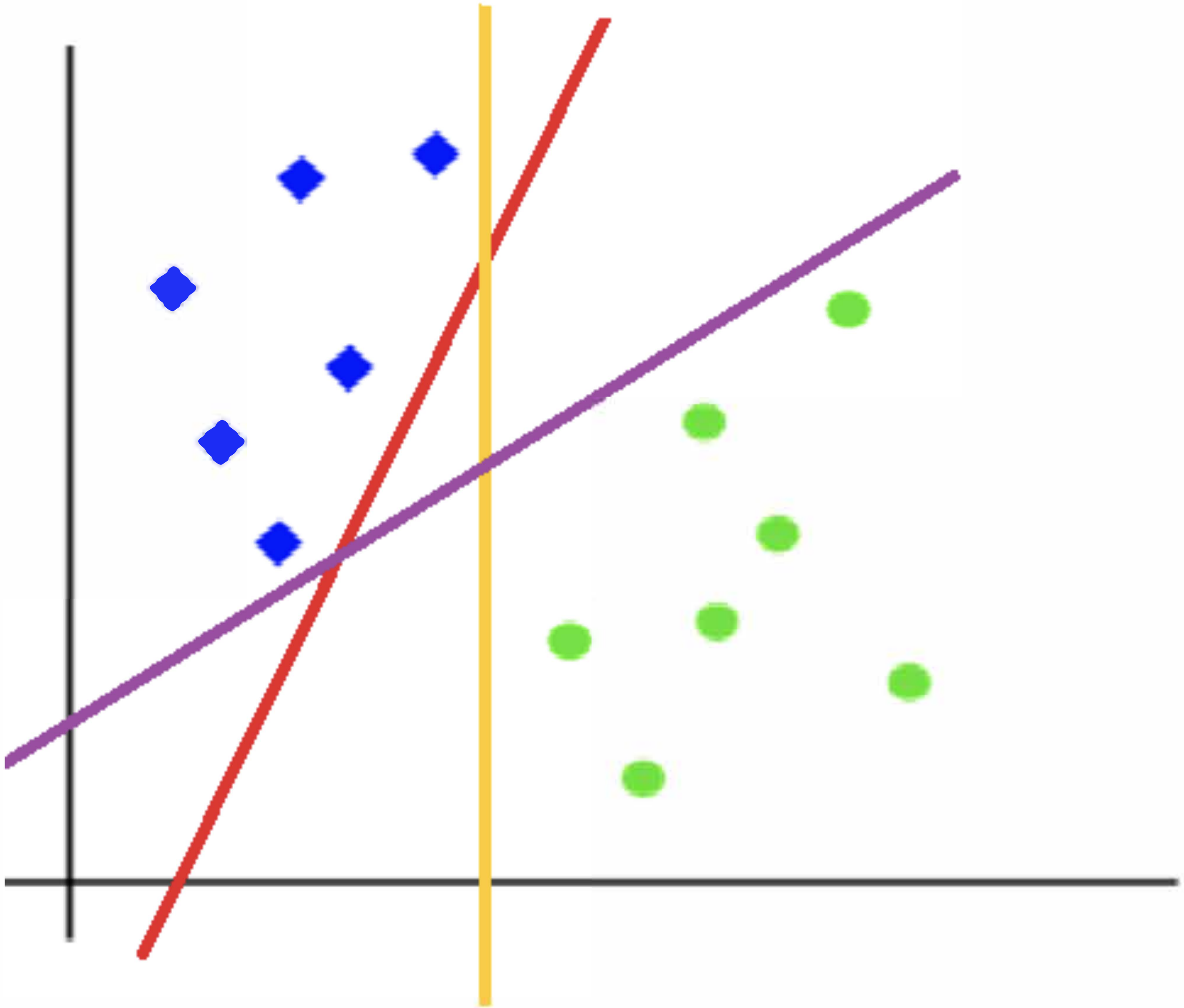
Easy problem
- large margin
- few mistakes

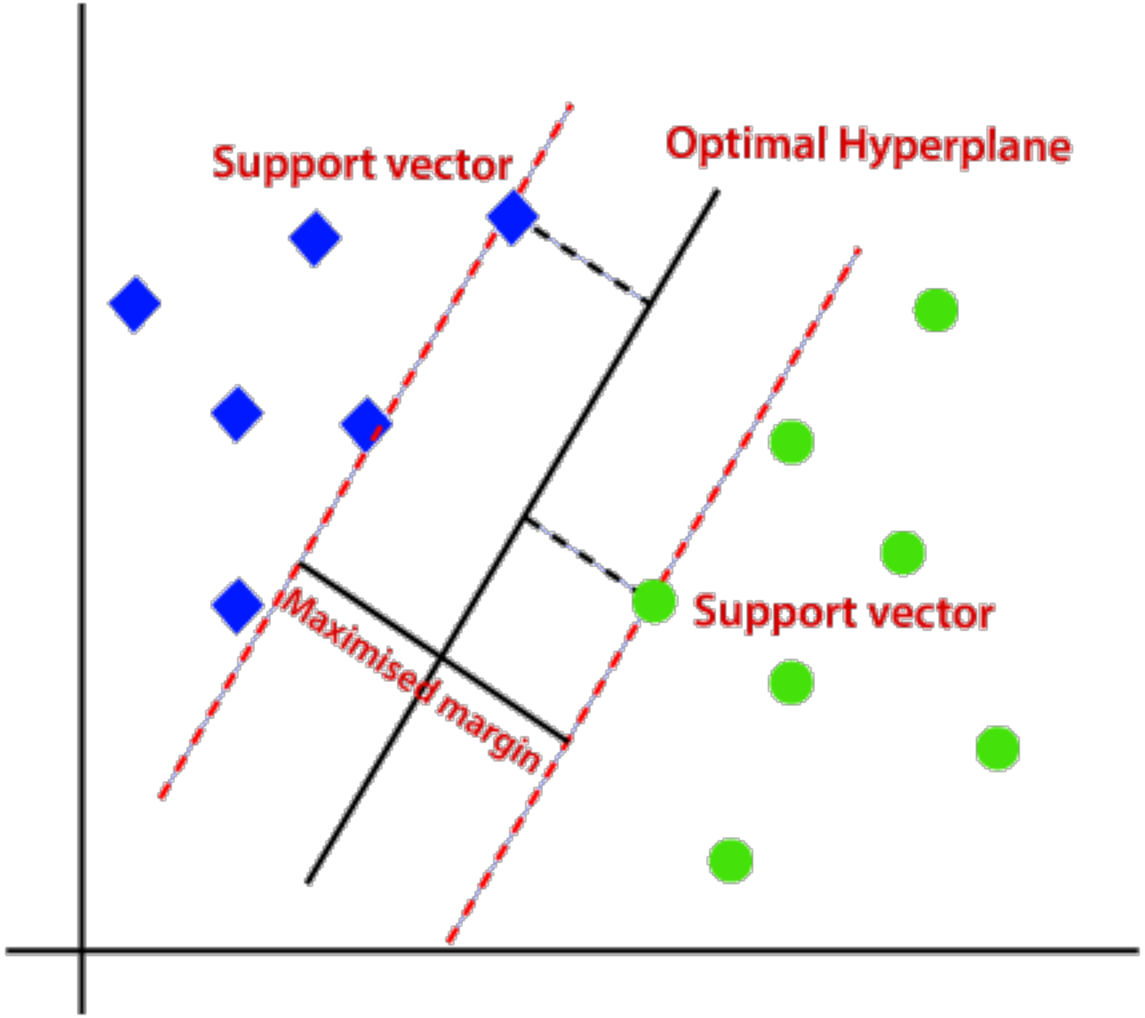


Harder problem
- small margin
- many mistakes



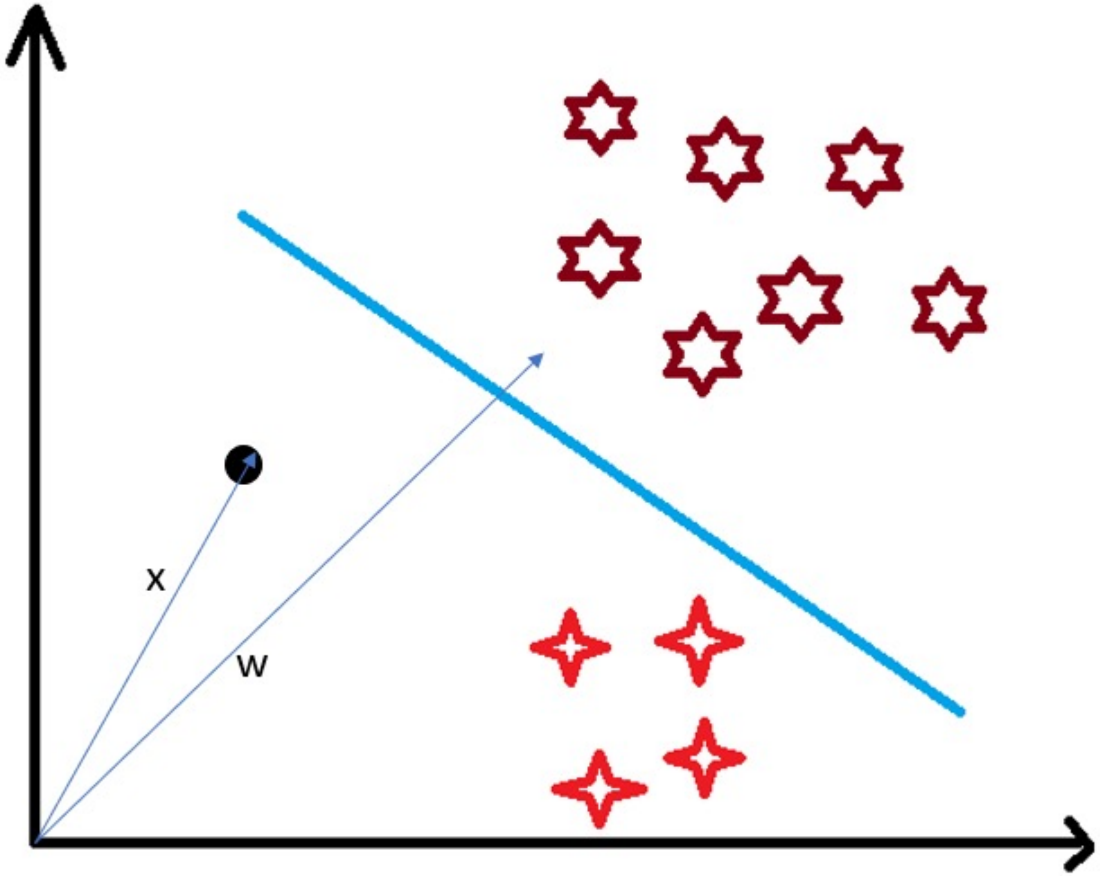






Support vector
Optimal Hyperplane
Support vector
Maximised margin

- A random point X
  - right side of the hyper plane or
  - left side of the hyper plane

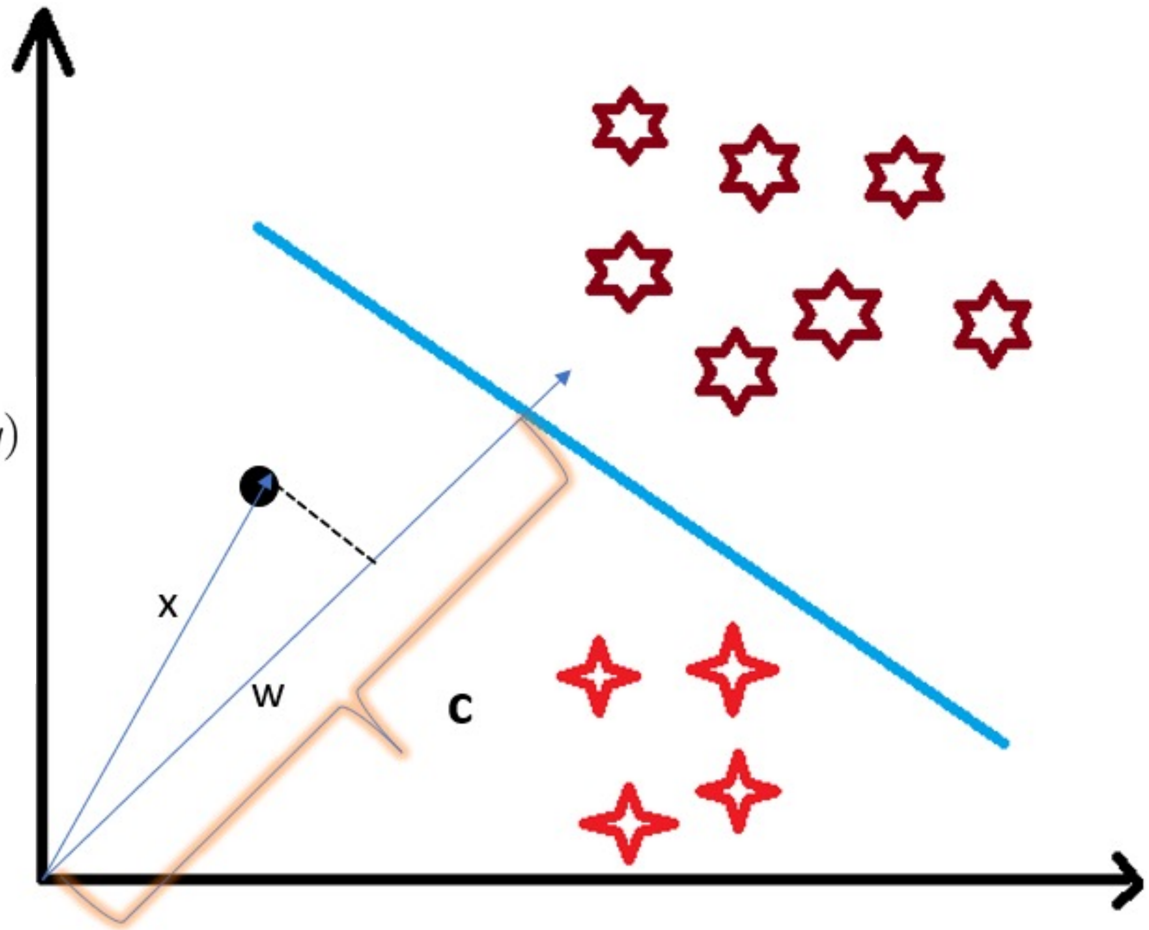


- $\vec{w}$ is perpendicular to the hyperplane
- distance of $\vec{w}$ from origin to decision boundary is c

$$\vec{X}.\vec{w} = c \quad (the\ point\ lies\ on\ the\ decision\ boundary)$$

$$\vec{X}.\vec{w} > c \ (positive\ samples)$$

$$\vec{X}.\vec{w} < c \ (negative\ samples)$$



Margin in SVM

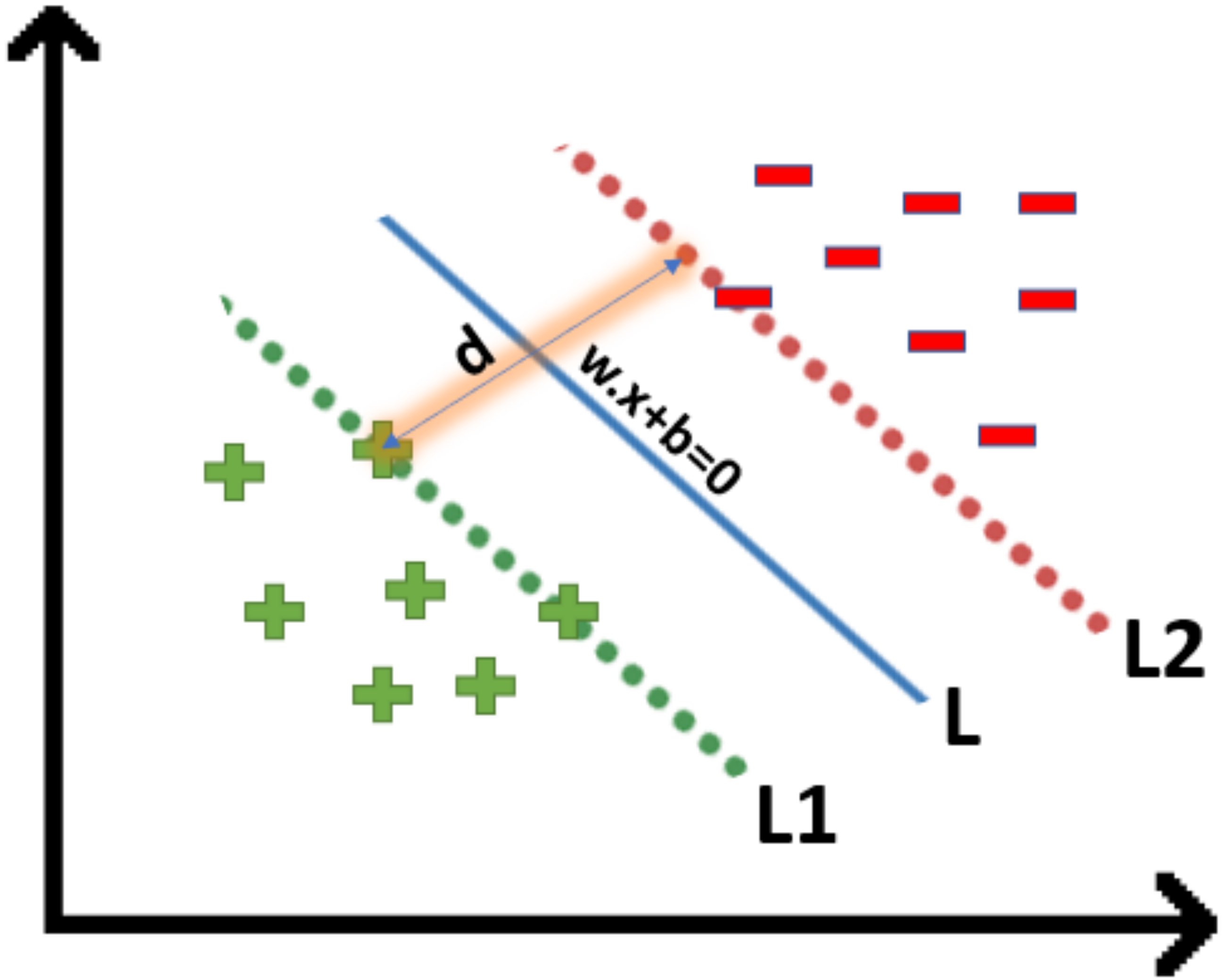
Without offset

$$y = \begin{cases} +1, & if\ \underline{w}\,.\,\underline{x} > 0 \\ -1, & if\ \underline{w}\,.\,\underline{x} \leq 0 \end{cases}$$

- b = 0
- Hyperplane through origin

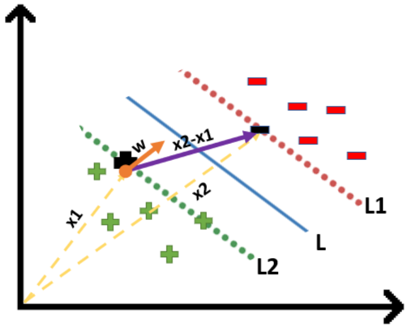
With offset

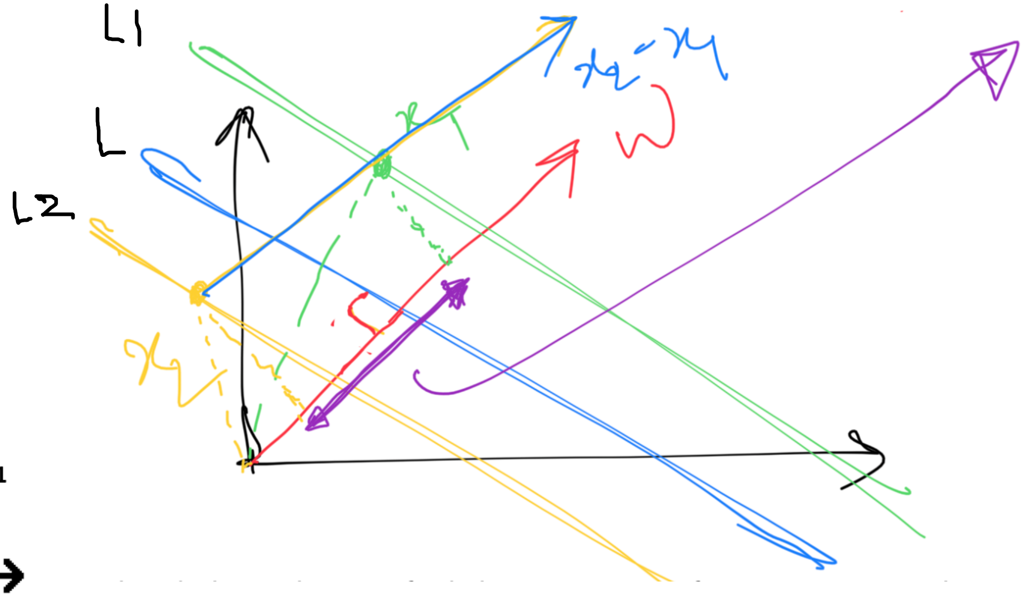
$$y = \begin{cases} +1, & if\ \underline{w}\,.\,\underline{x} + b > 0 \\ -1, & if\ \underline{w}\,.\,\underline{x} + b \leq 0 \end{cases}$$



Shortest distance betn

x_1 & x_2



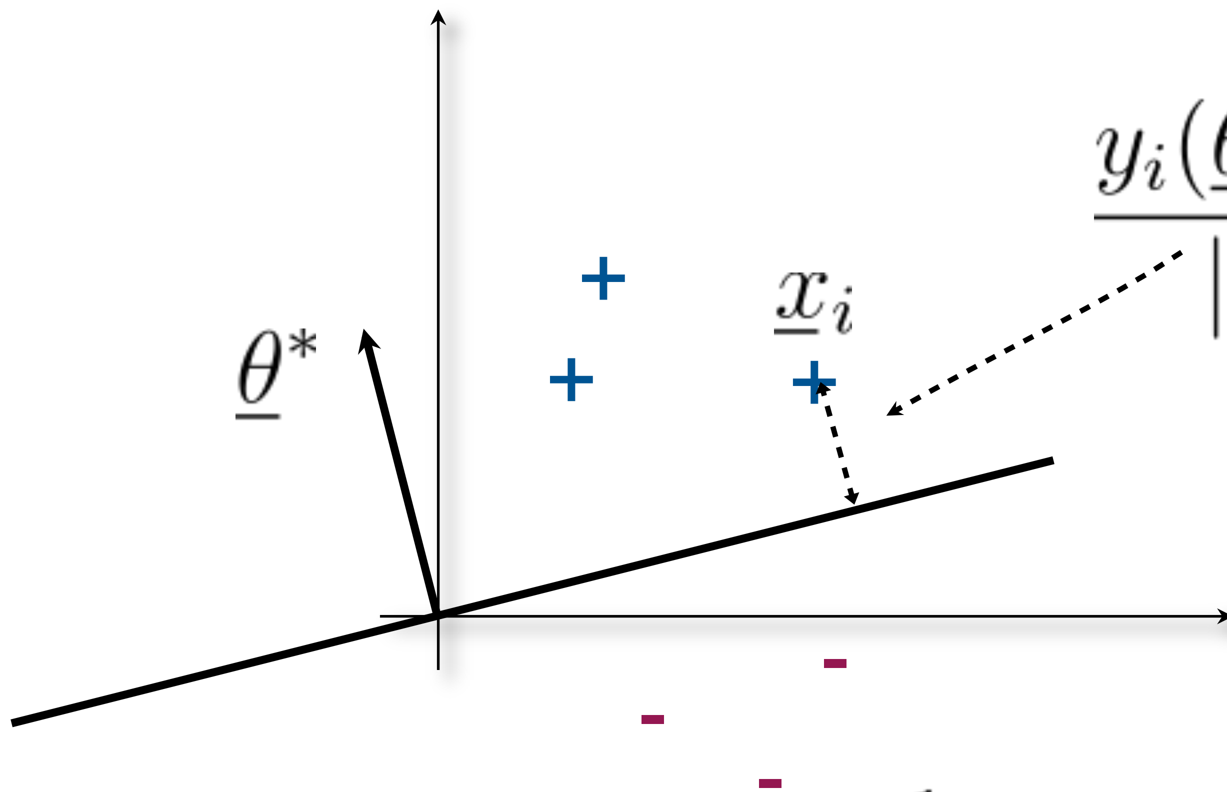


$$|\vec{x_1}|\cos\theta_{\hat{w}} - |\vec{x_2}|\cos\theta_{\hat{w}}$$

$$= (x_1 - x_2)\cos\theta_{\hat{w}}$$

$$= |x_2 - x_1|\cos\theta \cdot \hat{w}$$

Maximum margin classifier



$$\frac{y_i(\underline{\theta}^* \cdot \underline{x}_i)}{\|\underline{\theta}^*\|} \geq \gamma_g$$

geometric margin

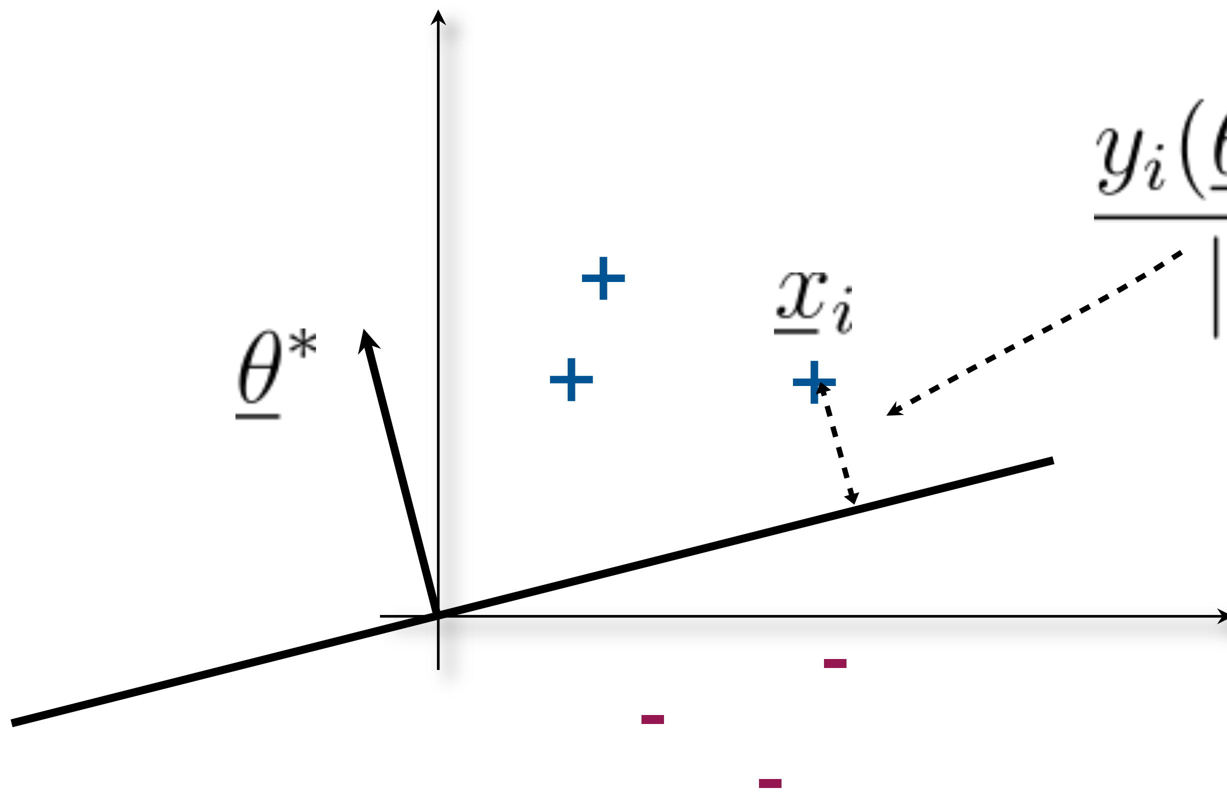
$$\gamma_g = \frac{1}{\|\underline{\theta}^*\|}$$

To find $\underline{\theta}^*$:

$$\text{maximize } \frac{1}{\|\underline{\theta}\|} \text{ subject to}$$

$$y_i(\underline{\theta} \cdot \underline{x}_i) \geq 1, \quad i = 1, \ldots, n$$

Maximum margin classifier



$$\frac{y_i(\underline{\theta}^* \cdot \underline{x}_i)}{\|\underline{\theta}^*\|} \geq \gamma_g$$

geometric margin

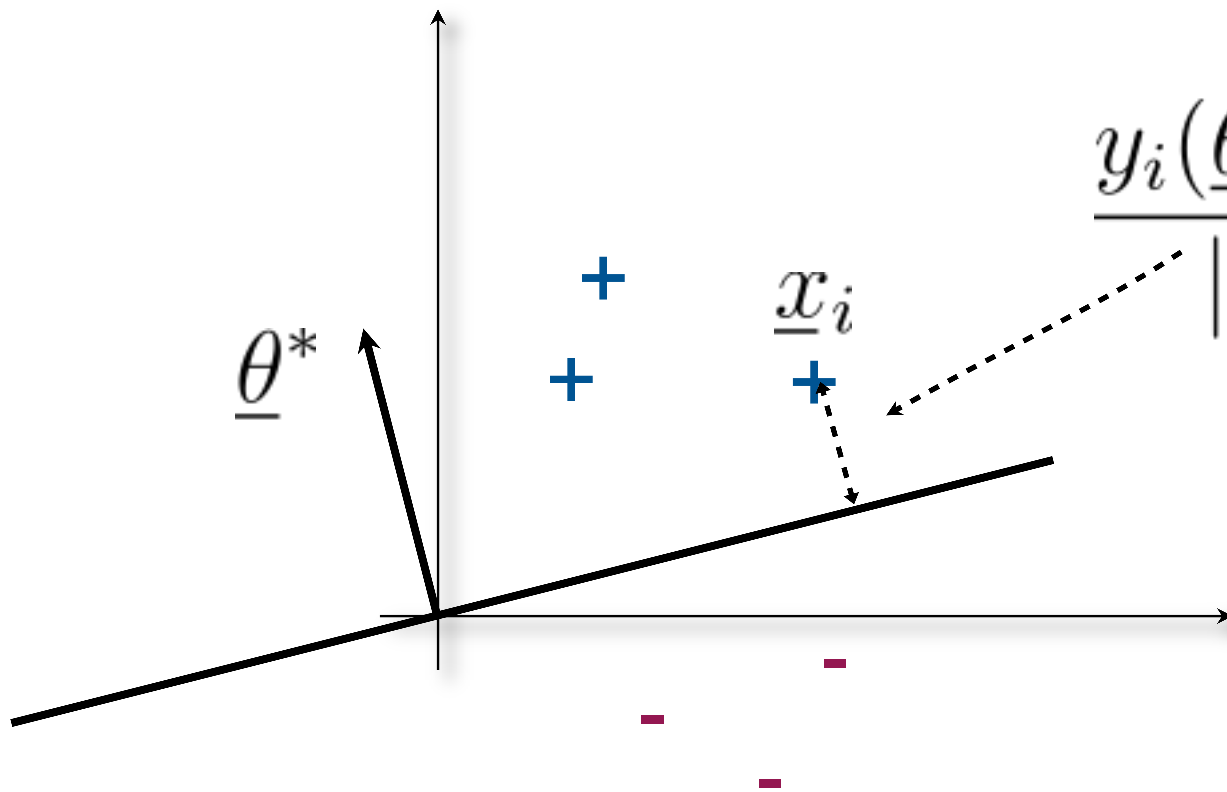
$$\gamma_g = \frac{1}{\|\underline{\theta}^*\|}$$

To find $\underline{\theta}^*$:

minimize $\|\underline{\theta}\|$ subject to

$$y_i(\underline{\theta} \cdot \underline{x}_i) \geq 1, \quad i = 1, \ldots, n$$

Support vector machine



$$\frac{y_i(\underline{\theta}^* \cdot \underline{x}_i)}{\|\underline{\theta}^*\|} \geq \gamma_g$$

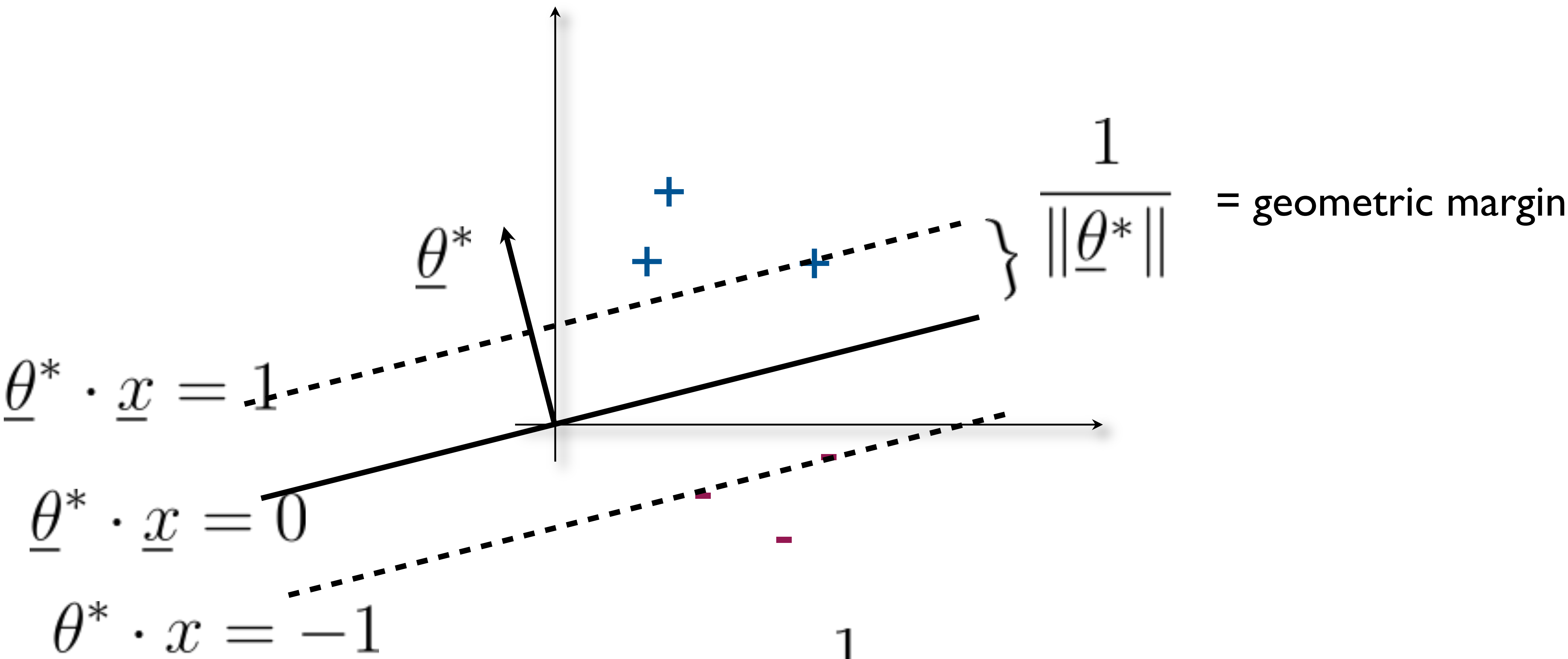
geometric margin

$$\gamma_g = \frac{1}{\|\underline{\theta}^*\|}$$

$$\text{To find } \underline{\theta}^* : \quad \begin{array}{l} \text{minimize } \frac{1}{2}\|\underline{\theta}\|^2 \text{ subject to} \\ y_i(\underline{\theta} \cdot \underline{x}_i) \geq 1, \quad i = 1, \ldots, n \end{array}$$

- This is a quadratic programming problem (quadratic objective, linear constraints)
- The solution is unique, typically obtained in the dual

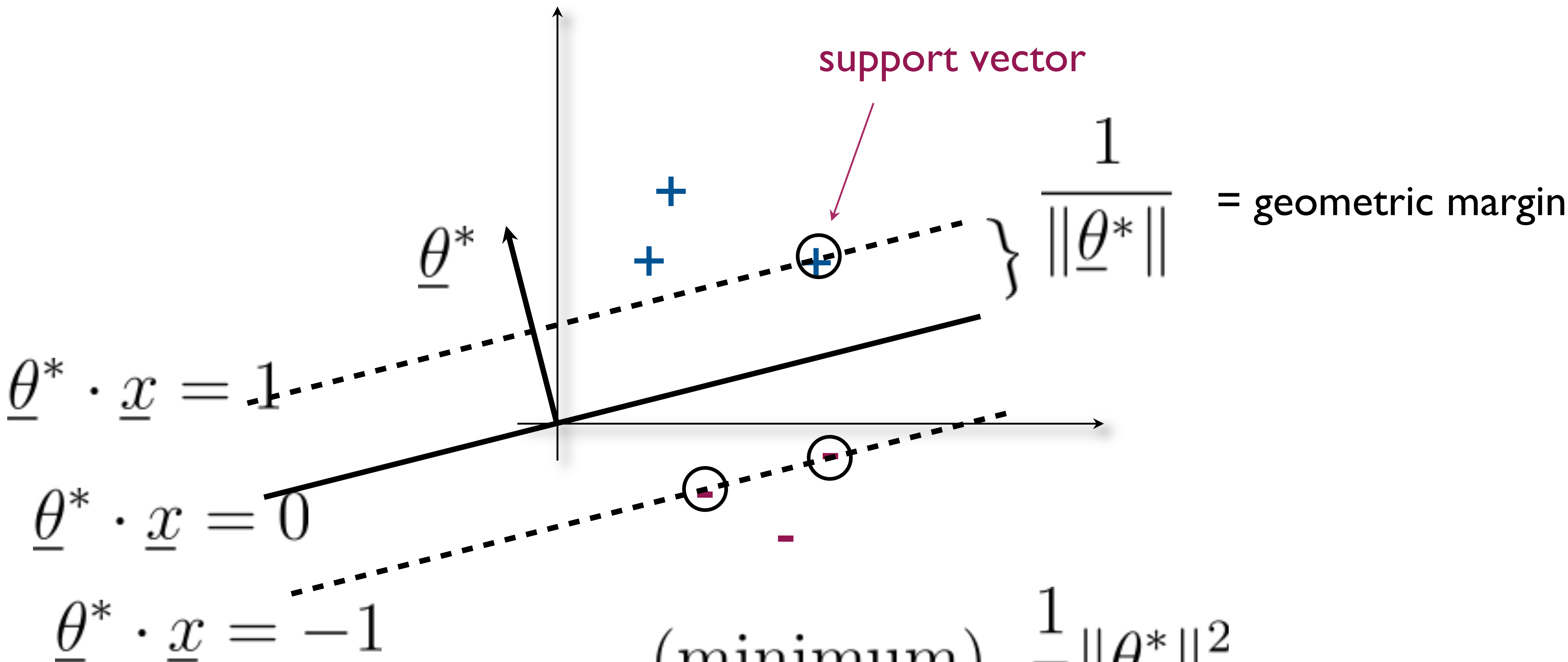
Support vector machine



To find $\underline{\theta}^*$:

$$\text{minimize } \frac{1}{2}\|\underline{\theta}\|^2 \text{ subject to}$$
$$y_i(\underline{\theta} \cdot \underline{x}_i) \geq 1, \quad i = 1, \dots, n$$

Support vector machine



$$(\text{minimum}) \quad \frac{1}{2}\|\underline{\theta}^*\|^2$$

$$y_1(\underline{\theta}^* \cdot \underline{x}_1) = 1$$

$$y_2(\underline{\theta}^* \cdot \underline{x}_2) > 1$$

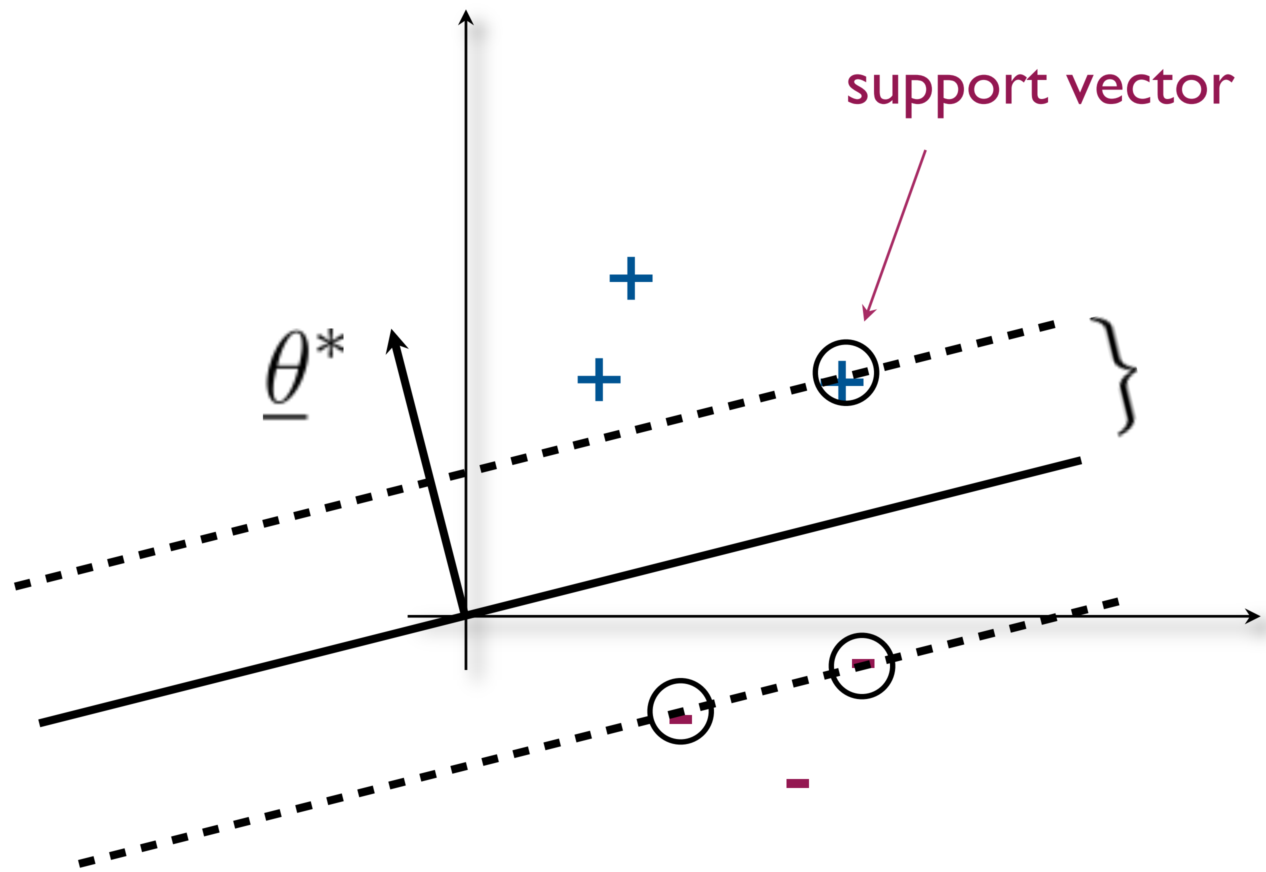
$$y_3(\underline{\theta}^* \cdot \underline{x}_3) = 1$$

$$\ldots$$

active constraints
= support vectors

The solution is **sparse**

Is sparse solution good?



- We can simulate test performance by evaluating Leave-One-Out Cross-Validation error

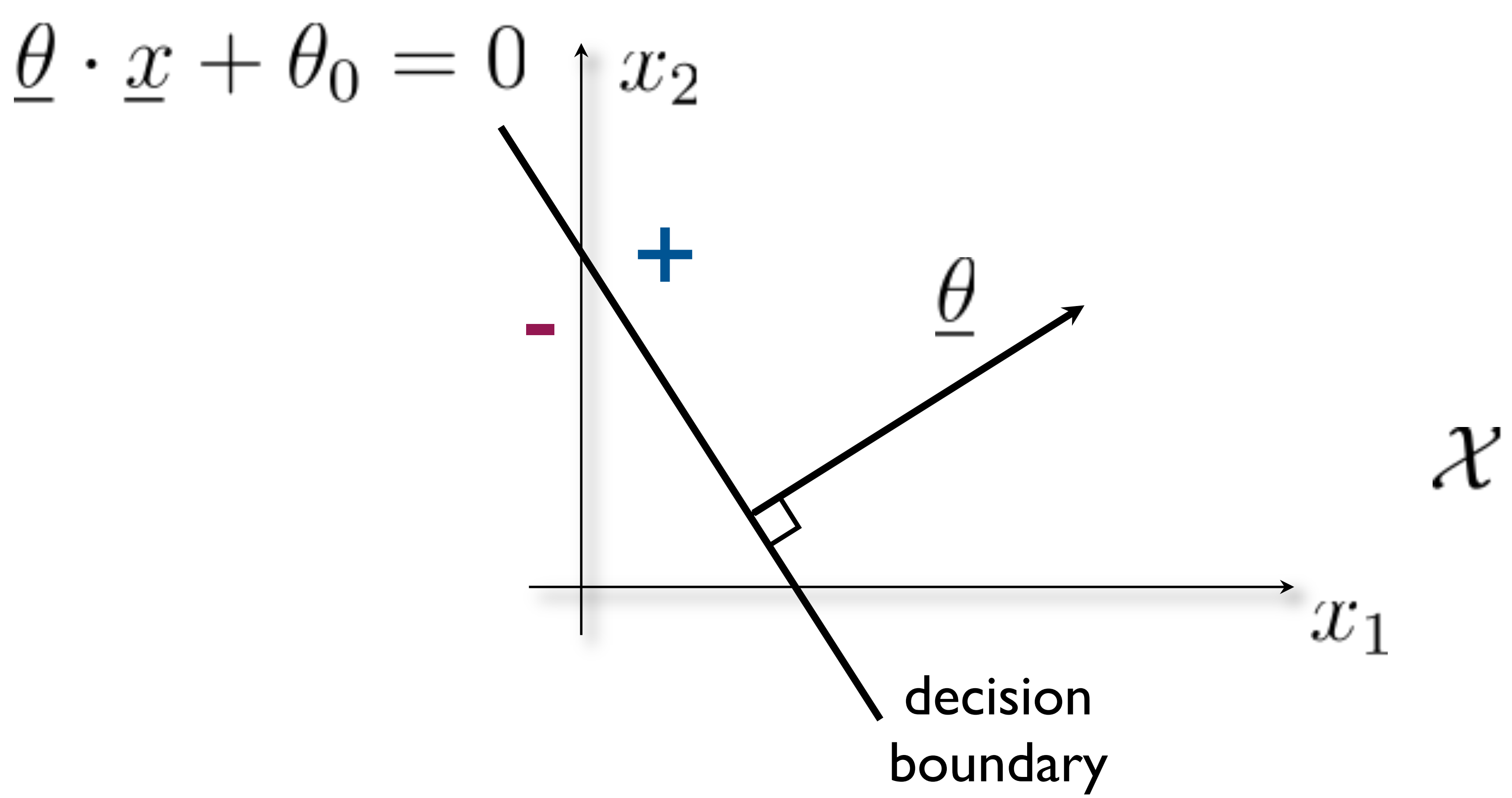
$$\mathrm{LOOCV}(\underline{\theta}^*) \leq \frac{\#\text{ of support vectors}}{n}$$

Intuitively:
if you remove the support vector from the training set, and you receive the support vector as a test point, then you would make a mistake

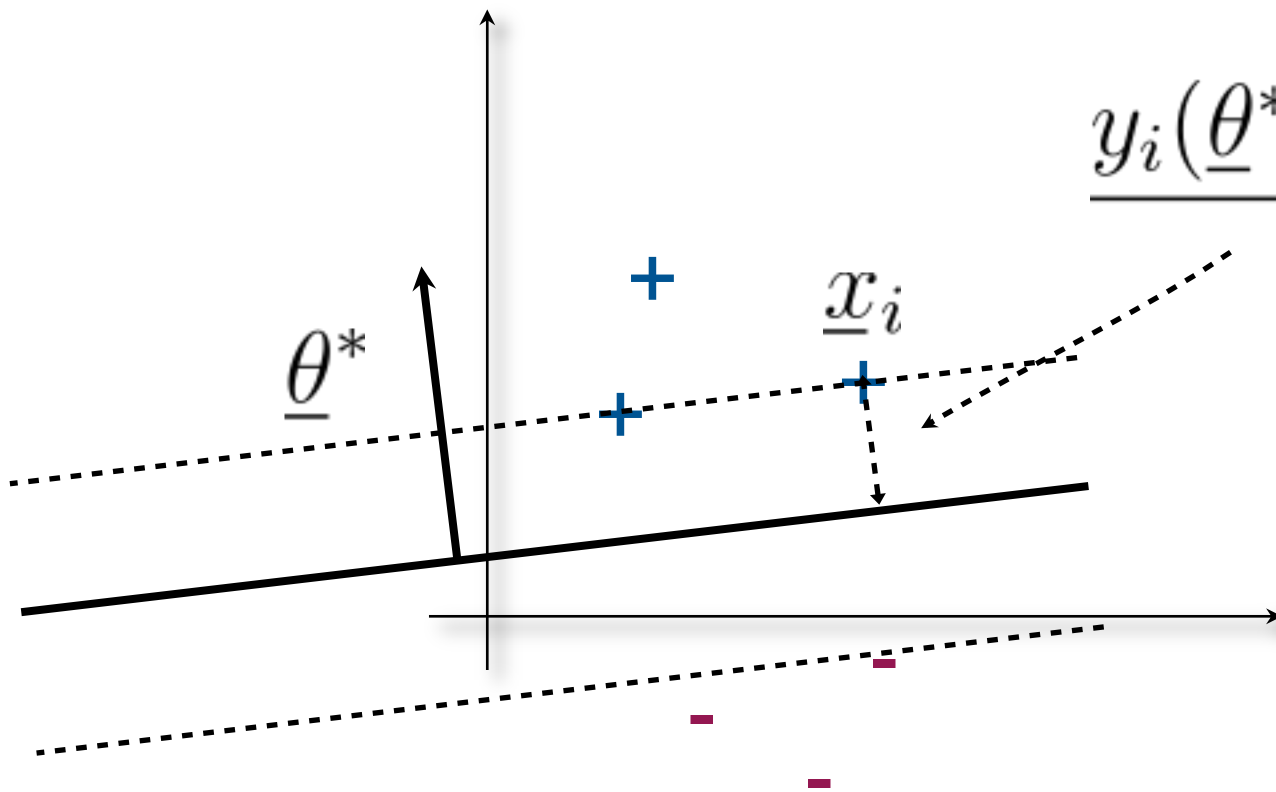
Linear classifiers (with offset)

- A linear classifier with parameters $(\underline{\theta}, \theta_0)$

$$
\begin{aligned}
f(\underline{x}; \underline{\theta}, \theta_0) &= \text{sign}\big(\underline{\theta} \cdot \underline{x} + \theta_0\big) \\
&= \begin{cases} +1, & \text{if } \underline{\theta} \cdot \underline{x} + \theta_0 > 0 \\ -1, & \text{if } \underline{\theta} \cdot \underline{x} + \theta_0 \leq 0 \end{cases}
\end{aligned}
$$



Support vector machine



$$\frac{y_i(\underline{\theta}^* \cdot \underline{x}_i + \theta_0^*)}{\|\underline{\theta}^*\|} \geq \gamma_g$$

geometric margin

$$\gamma_g = \frac{1}{\|\underline{\theta}^*\|}$$

To find $\underline{\theta}^*, \theta_0^*$:

$$\text{minimize } \frac{1}{2}\|\underline{\theta}\|^2 \text{ subject to}$$

$$y_i(\underline{\theta} \cdot \underline{x}_i + \theta_0) \geq 1, \quad i = 1, \ldots, n$$

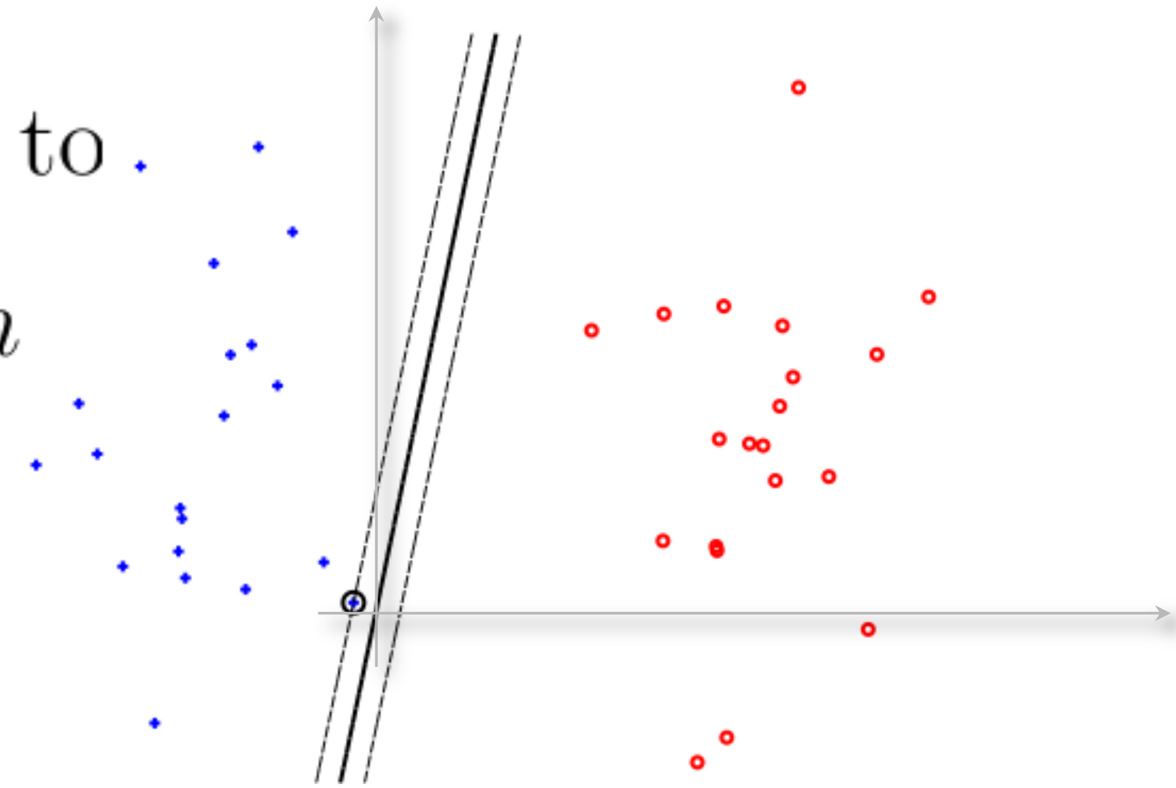
- Still a quadratic programming problem (quadratic objective, linear constraints)

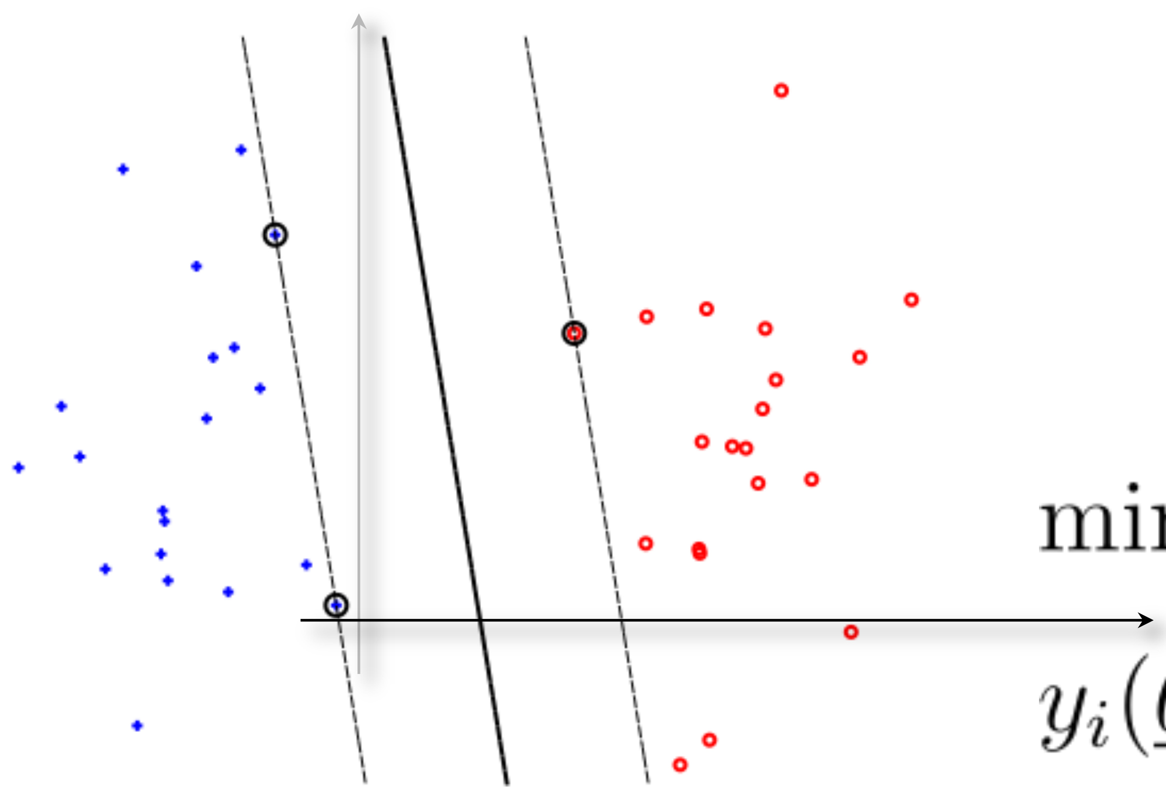
The impact of offset

- Adding the offset parameter to the linear classifier can substantially increase the margin

$$\text{minimize} \quad \frac{1}{2}\|\underline{\theta}\|^2 \quad \text{subject to}$$

$$y_i(\underline{\theta} \cdot \underline{x}_i) \geq 1, \quad i = 1, \ldots, n$$





$$\text{minimize} \quad \frac{1}{2}\|\underline{\theta}\|^2 \quad \text{subject to}$$

$$y_i(\underline{\theta} \cdot \underline{x}_i + \theta_0) \geq 1, \quad i = 1, \ldots, n$$

Support vector machine

- Several desirable properties
  - maximizes the margin on the training set ($\approx$ good generalization)
  - the solution is unique and sparse ($\approx$ good generalization)
- But...
  - the solution is sensitive to outliers, labeling errors, as they may drastically change the resulting max-margin boundary
  - if the training set is not linearly separable, there's no solution!

Support vector machine

- Relaxed quadratic optimization problem

penalty for constraint violation

$$\text{minimize} \quad \frac{1}{2}\|\underline{\theta}\|^2 \quad + \quad \overbrace{C\sum_{i=1}^{n}\xi_i} \quad \text{subject to}$$

$$\begin{aligned} y_i(\underline{\theta}\cdot\underline{x}_i+\theta_0) \quad &\geq \quad 1-\xi_i, \quad i=1,\ldots,n \\ \xi_i \quad &\geq \quad 0, \quad i=1,\ldots,n \end{aligned}$$

slack variables
permit us to violate
some of the margin
constraints

Support vector machine

- Relaxed quadratic optimization problem

penalty for constraint violation

$$\text{minimize} \quad \frac{1}{2}\|\underline{\theta}\|^2 \quad + \quad \overline{C \sum_{i=1}^{n} \xi_i} \quad \text{subject to}$$

$$\begin{aligned} y_i(\underline{\theta} \cdot \underline{x}_i + \theta_0) \quad &\geq \quad 1 - \xi_i, \quad i = 1, \ldots, n \\ \xi_i \quad &\geq \quad 0, \quad i = 1, \ldots, n \end{aligned}$$

slack variables permit us to violate some of the margin constraints

$$\text{large } C \Rightarrow \text{few (if any) violations}$$

$$\text{small } C \Rightarrow \text{many violations}$$

Support vector machine

- Relaxed quadratic optimization problem

penalty for constraint violation

$$\text{minimize} \quad \frac{1}{2}\|\underline{\theta}\|^2 \quad + \quad \overline{C\sum_{i=1}^{n}\xi_i} \quad \text{subject to}$$

$$\begin{aligned} y_i(\underline{\theta}\cdot\underline{x}_i+\theta_0) &\geq 1-\xi_i, \quad i=1,\ldots,n \\ \xi_i &\geq 0, \quad i=1,\ldots,n \end{aligned}$$

slack variables permit us to violate some of the margin constraints

large $C \Rightarrow$ few (if any) violations

small $C \Rightarrow$ many violations

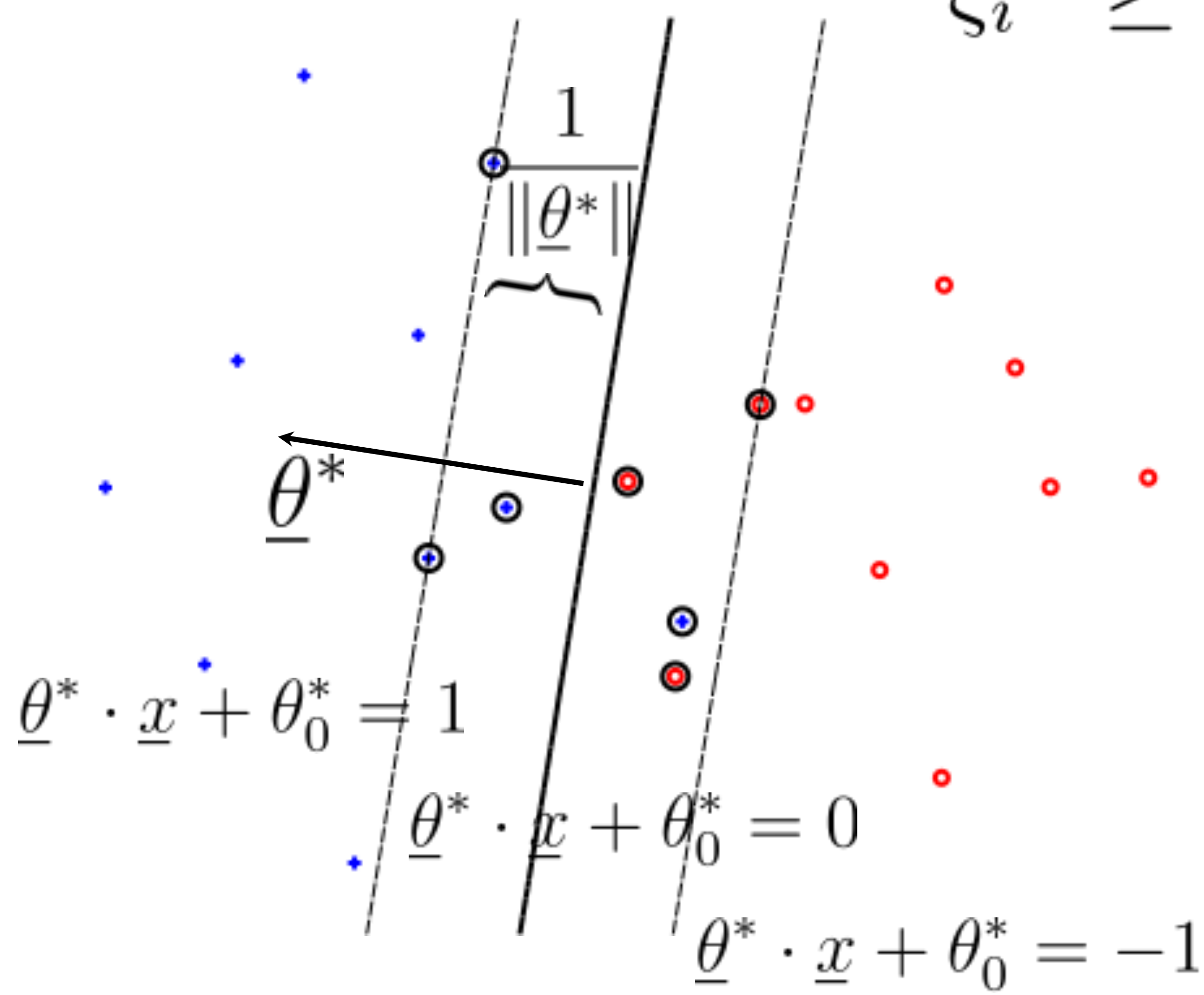
we can still interpret the margin as $1/\|\underline{\theta}^*\|$

Support vector machine

- Relaxed quadratic optimization problem

$$\text{minimize} \quad \frac{1}{2}\|\underline{\theta}\|^2 \quad + \quad C\sum_{i=1}^{n}\xi_i \quad \text{subject to}$$

$$\begin{aligned} y_i(\underline{\theta}\cdot\underline{x}_i+\theta_0) &\geq 1-\xi_i, \quad i=1,\ldots,n \\ \xi_i &\geq 0, \quad i=1,\ldots,n \end{aligned}$$



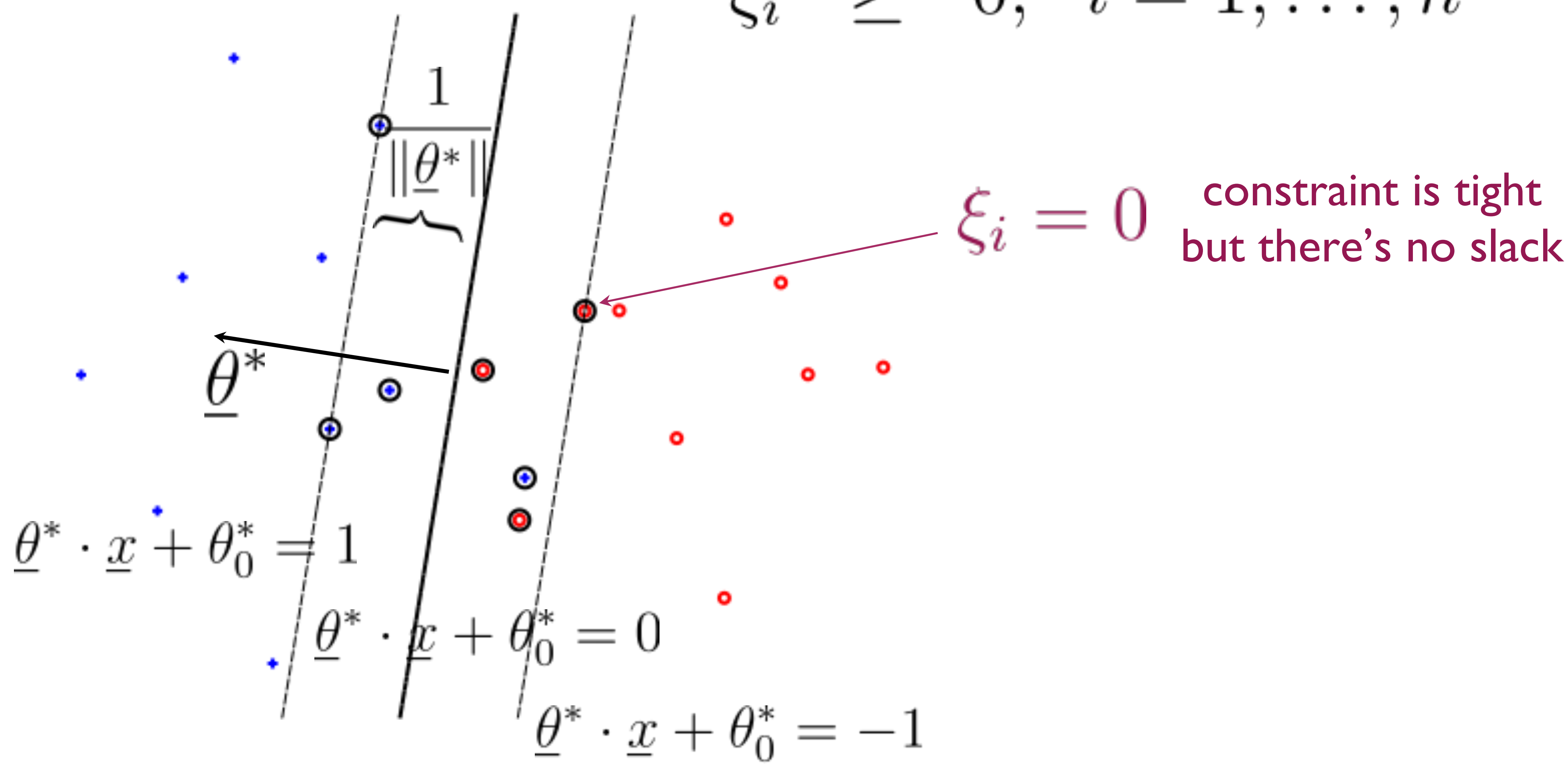
Support vectors and slack

- The solution now has three types of support vectors

$$\text{minimize} \quad \frac{1}{2}\|\underline{\theta}\|^2 \quad + \quad C\sum_{i=1}^{n}\xi_i \quad \text{subject to}$$

$$y_i(\underline{\theta}\cdot\underline{x}_i+\theta_0) \quad \geq \quad 1-\xi_i, \quad i=1,\ldots,n$$

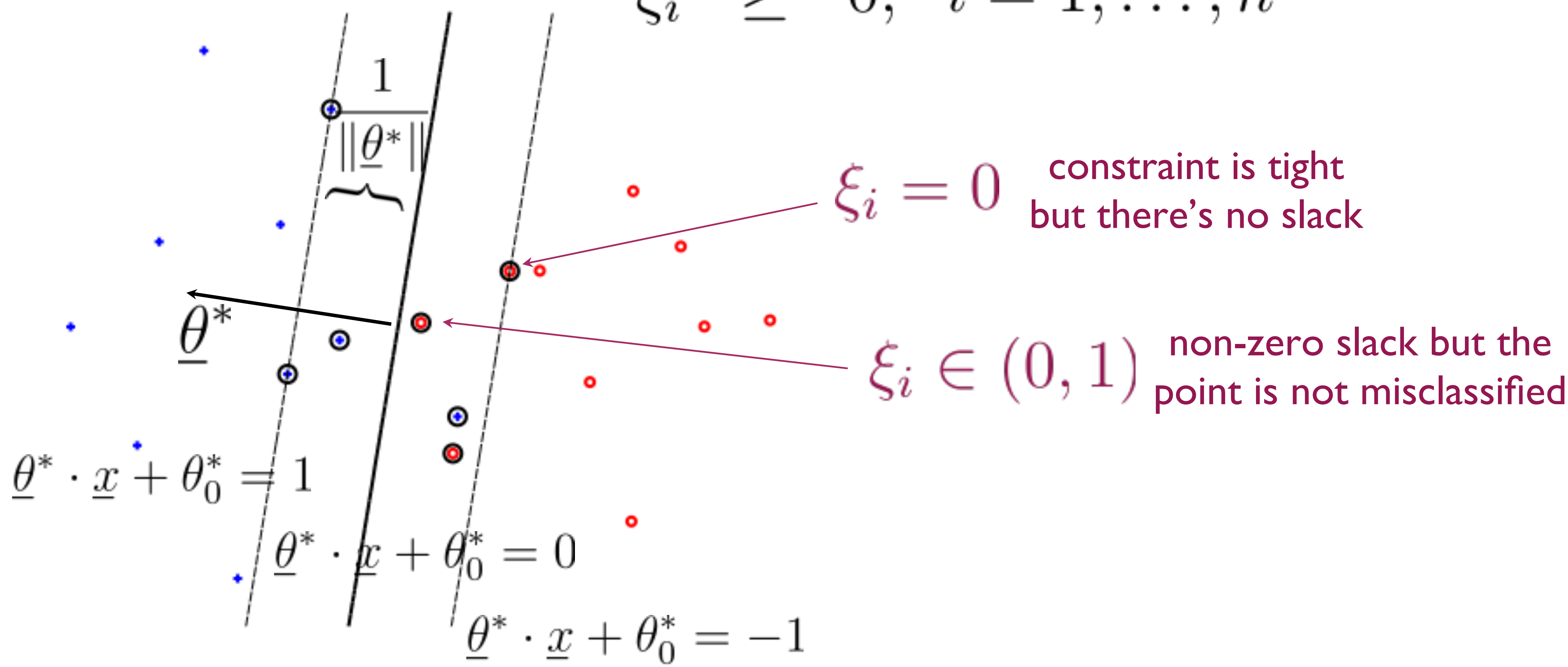
$$\xi_i \quad \geq \quad 0, \quad i=1,\ldots,n$$



Support vectors and slack

- The solution now has three types of support vectors

$$
\begin{aligned}
\text{minimize} \quad \frac{1}{2}\|\underline{\theta}\|^2 \quad + \quad C\sum_{i=1}^{n}\xi_i \quad \text{subject to} \\
y_i(\underline{\theta}\cdot\underline{x}_i+\theta_0) \quad &\geq \quad 1-\xi_i, \quad i=1,\ldots,n \\
\xi_i \quad &\geq \quad 0, \quad i=1,\ldots,n
\end{aligned}
$$



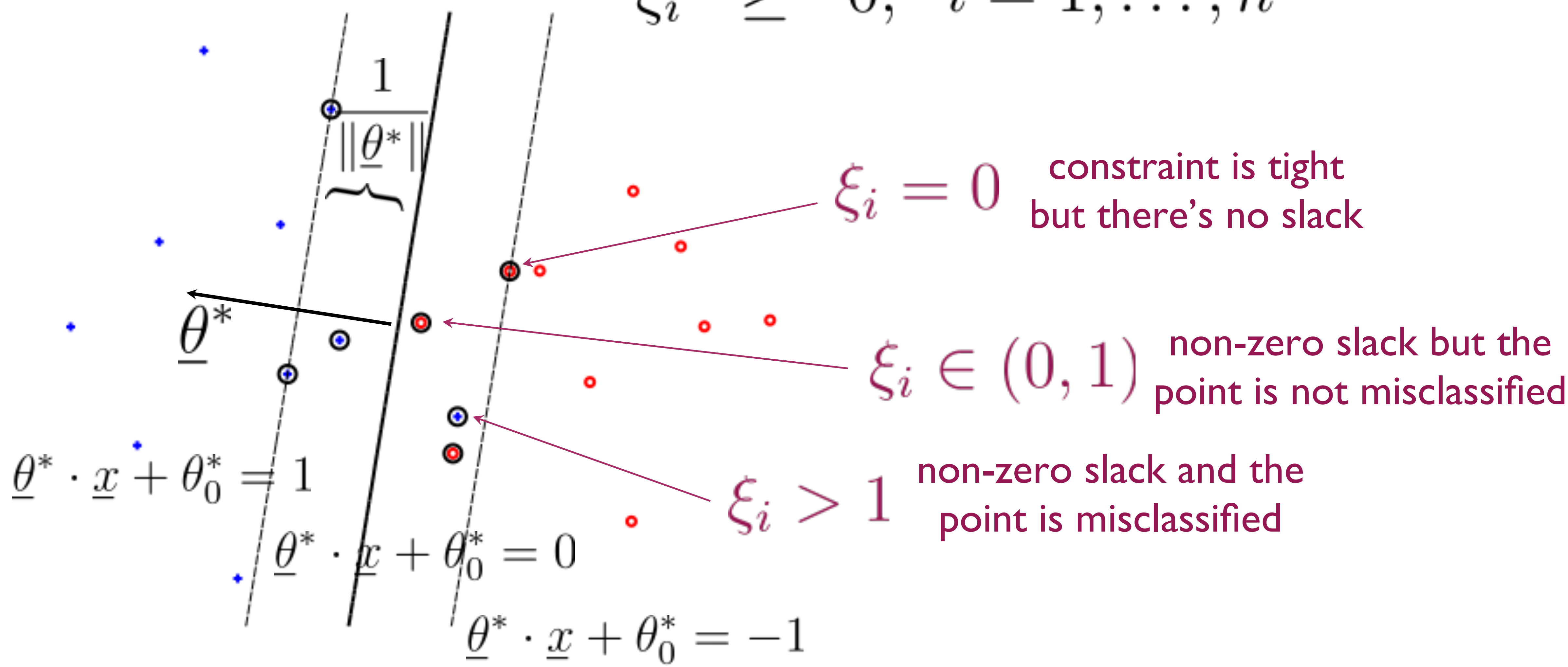
Support vectors and slack

- The solution now has three types of support vectors

$$\text{minimize} \quad \frac{1}{2}\|\underline{\theta}\|^2 \;+\; C\sum_{i=1}^{n}\xi_i \quad \text{subject to}$$

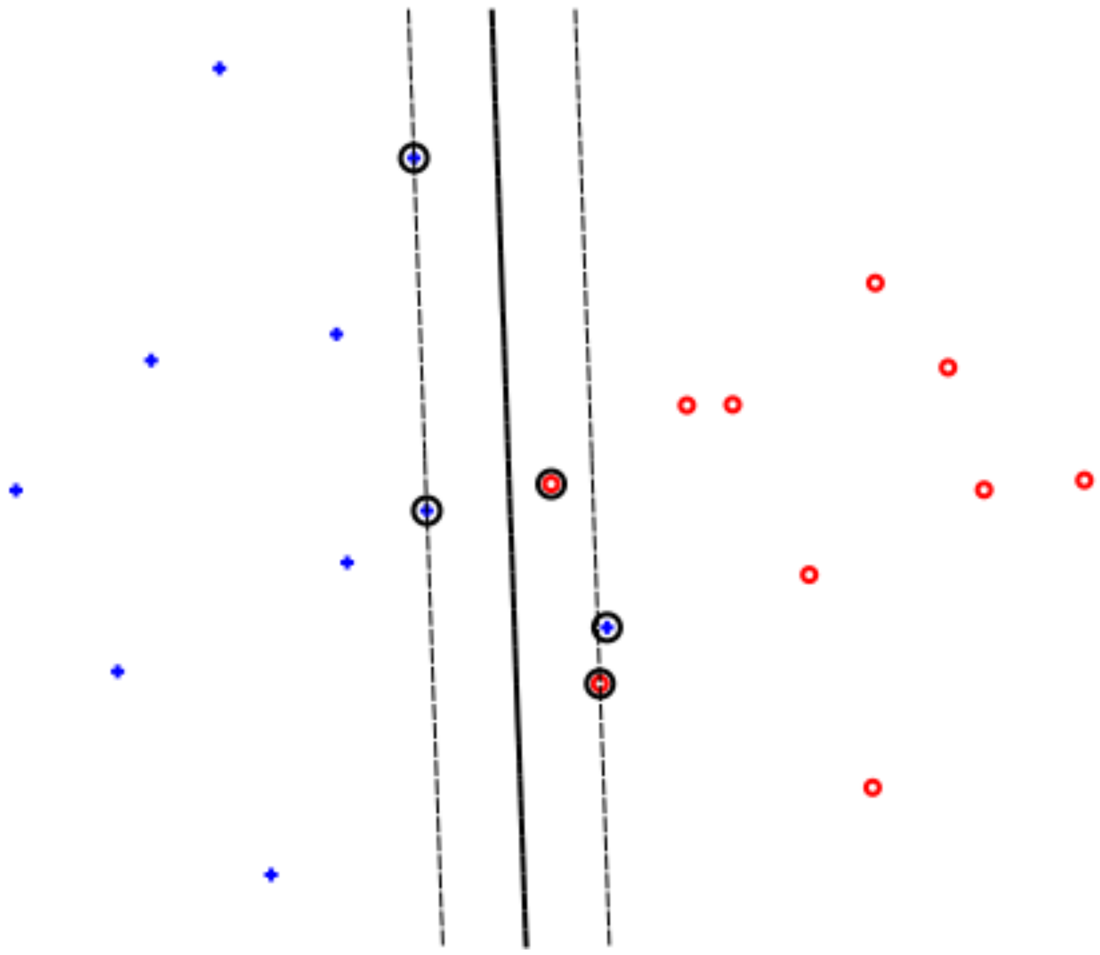
$$y_i(\underline{\theta}\cdot\underline{x}_i+\theta_0) \;\geq\; 1-\xi_i,\quad i=1,\ldots,n$$

$$\xi_i \;\geq\; 0,\quad i=1,\ldots,n$$



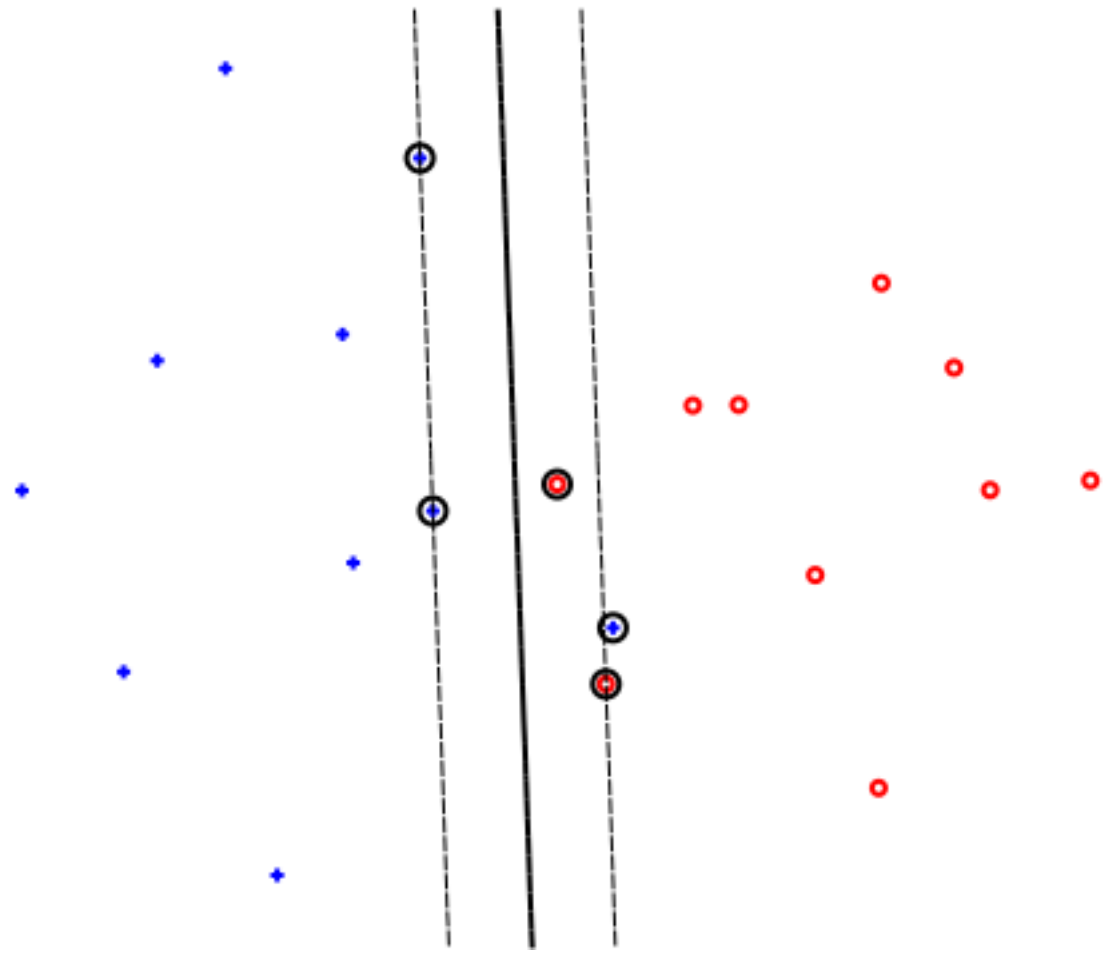
Examples

- C=100



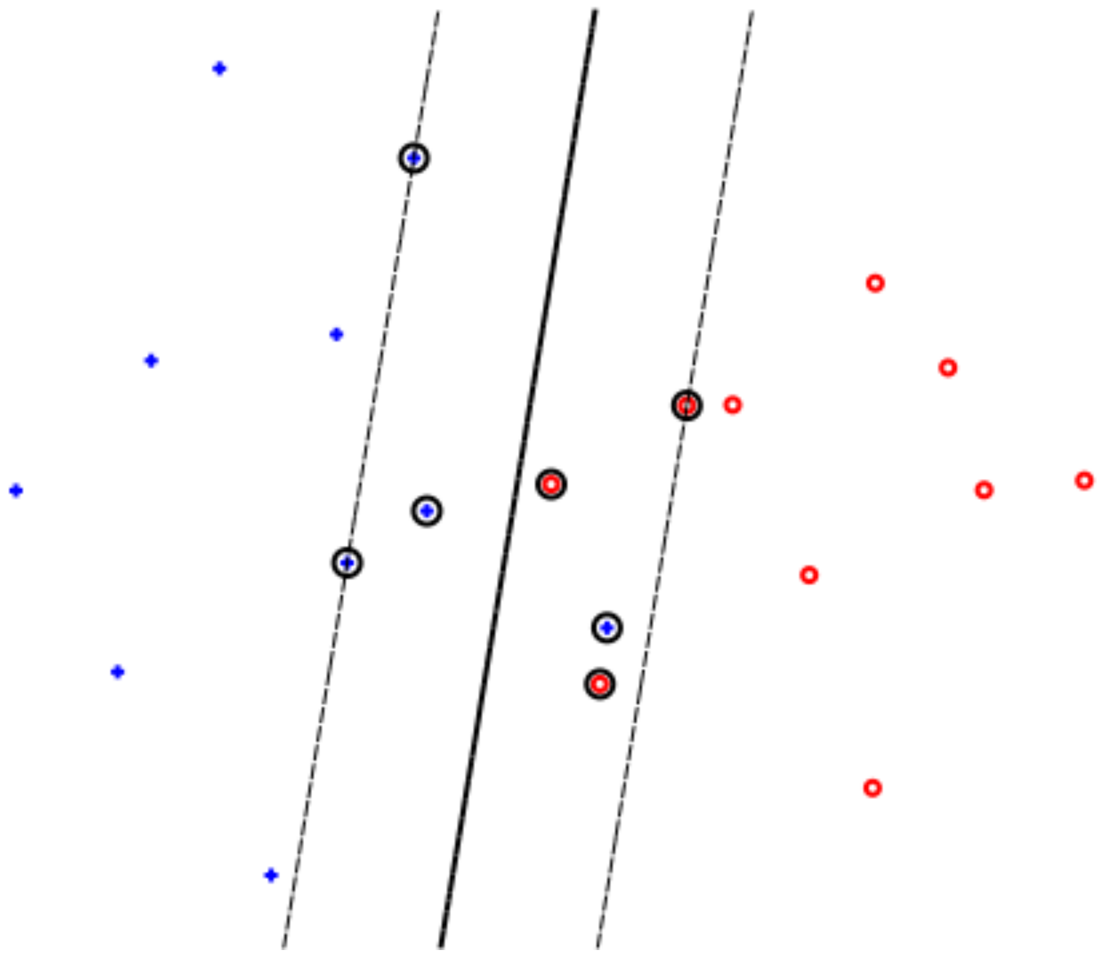
Examples

- C=10



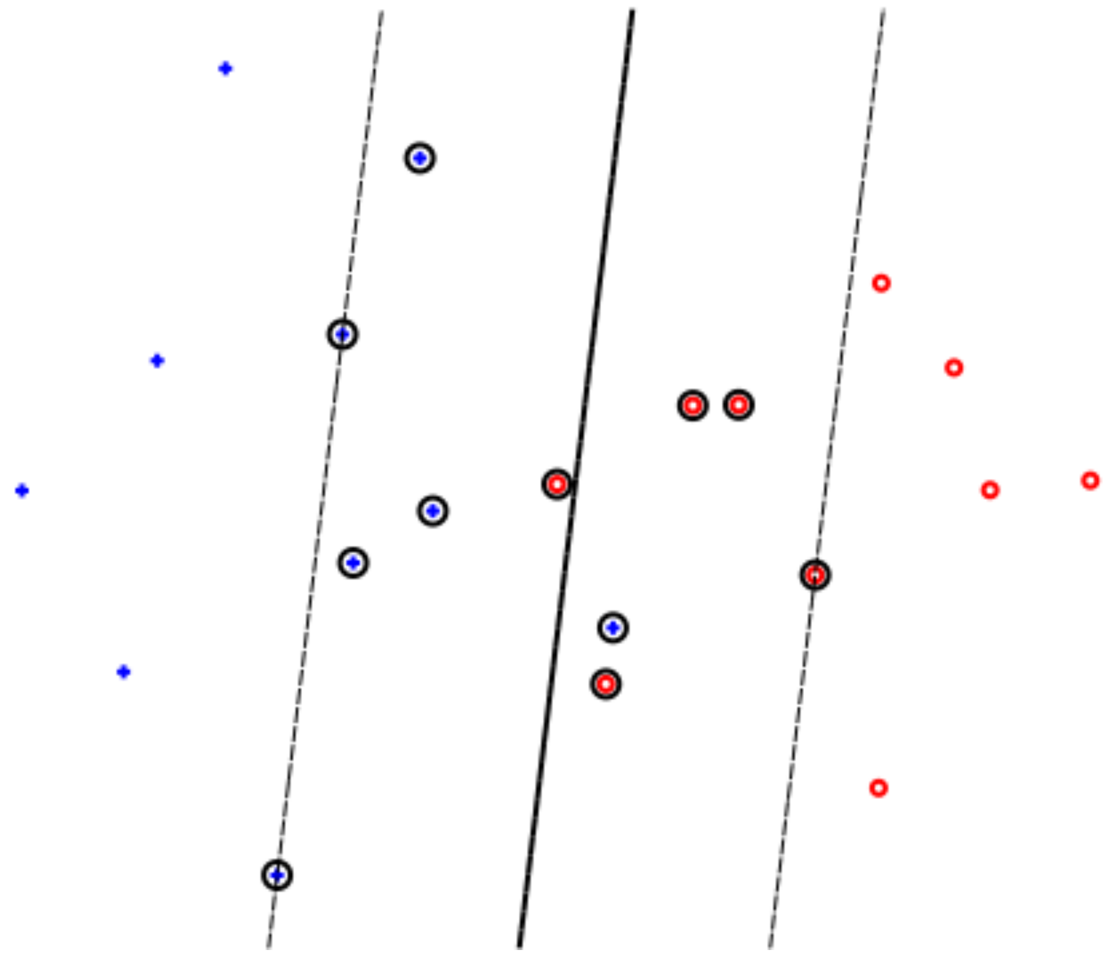
Examples

- C=1



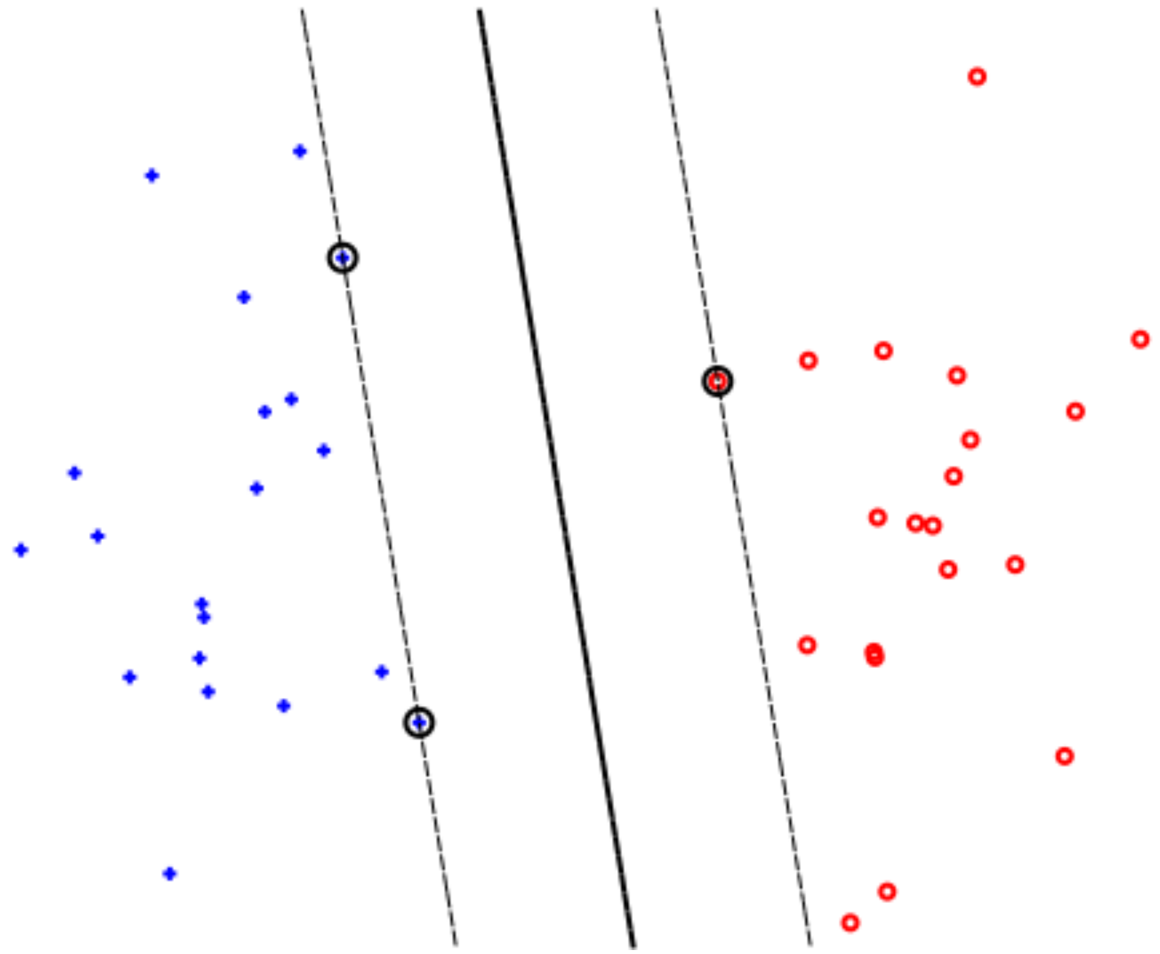
Examples

- C=0.1



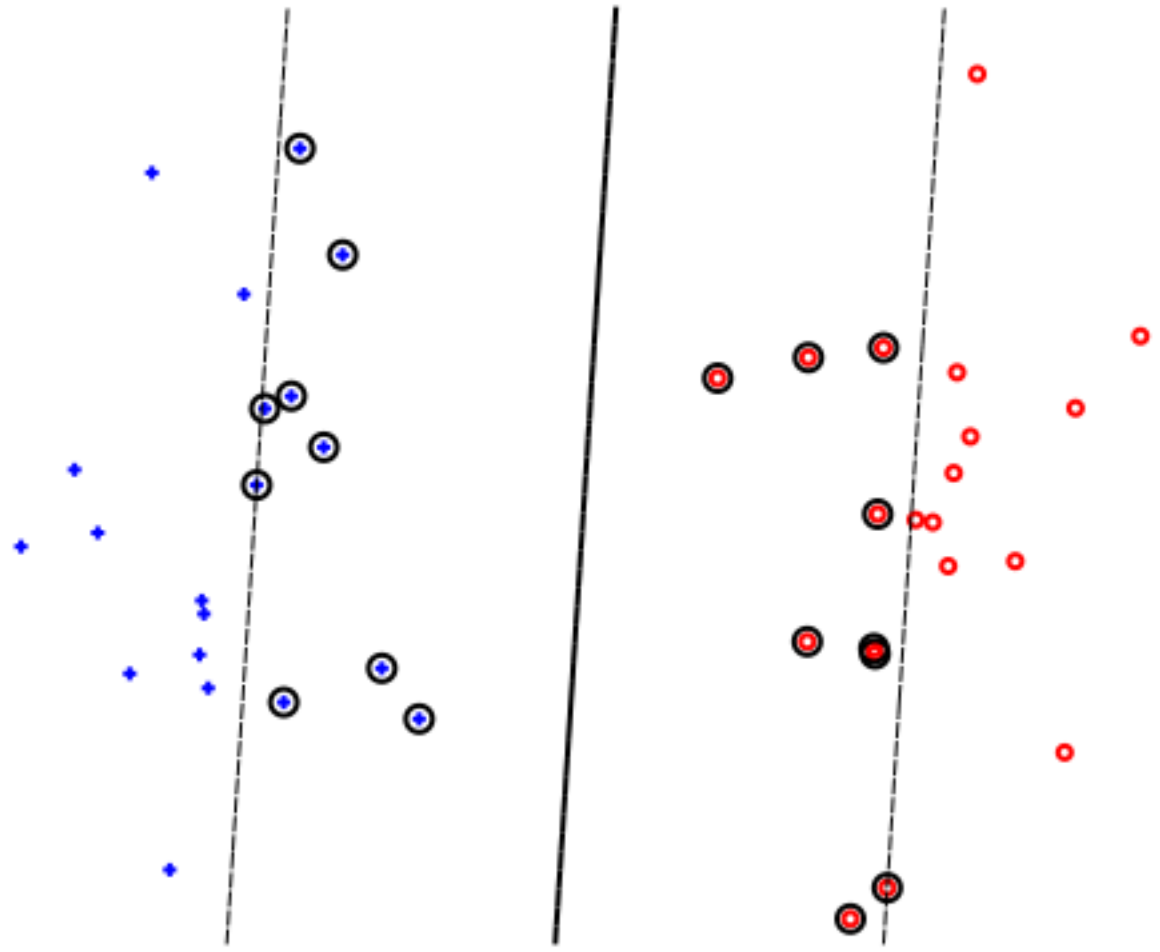
Examples

- C potentially affects the solution even in the separable case
- C = 1



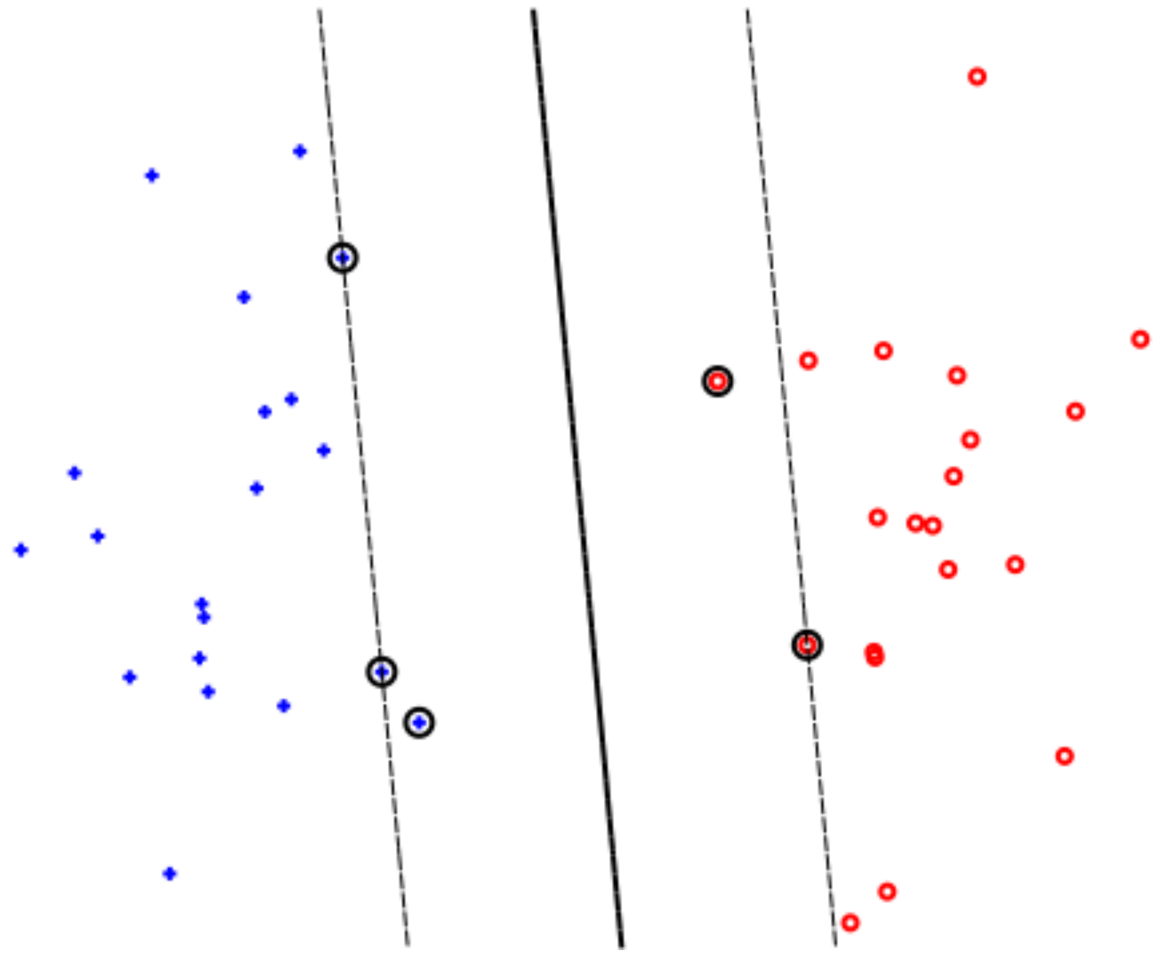
Examples

- C potentially affects the solution even in the separable case
- C = 0.01

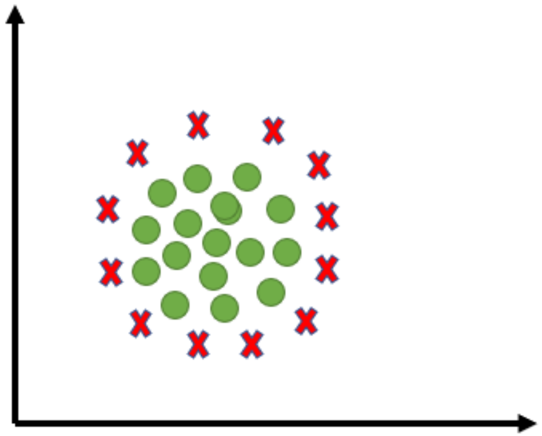


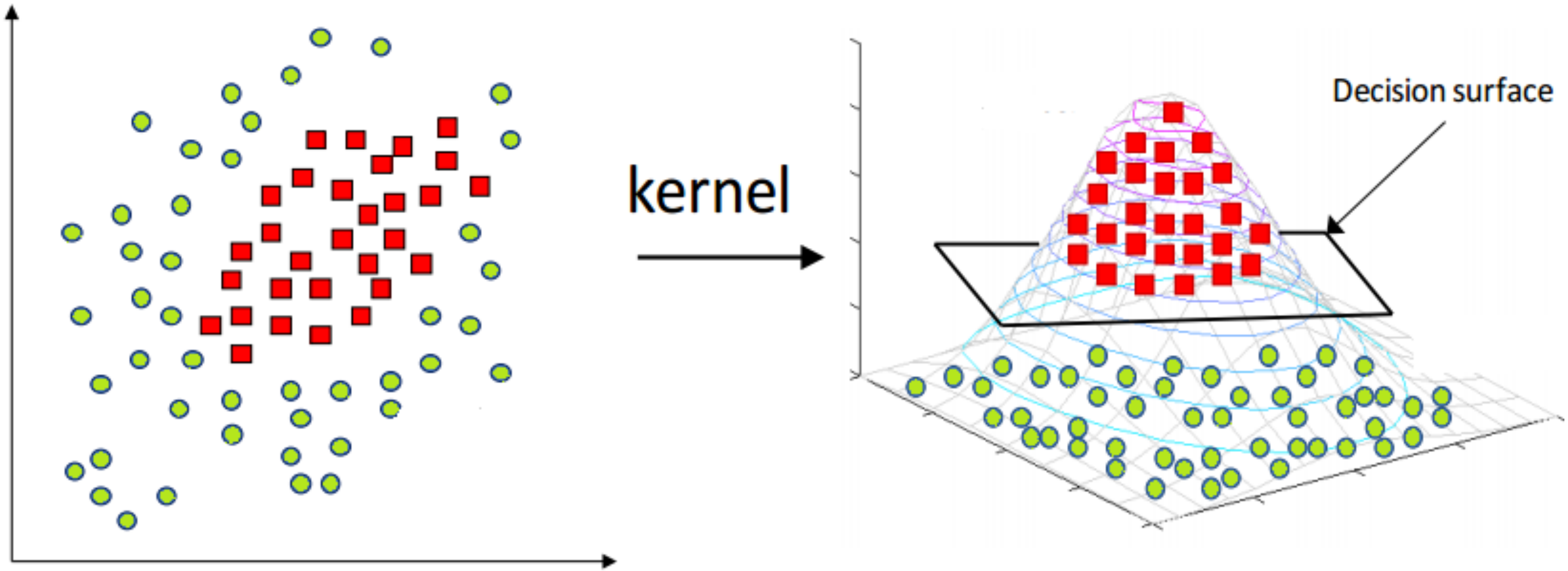
Examples

- C potentially affects the solution even in the separable case
- C = 0.1



Non-linear dataset




kernel
Decision surface

Different Types of kernel

Polynomial
Sigmoid
RBF

$$K(X1, X2) = (X1^T . X2 + 1)^d$$

$$K(x1, x2) = \tanh(\alpha x^T y + x)$$

$$K(x1, x2) = e^{\dfrac{-||(x1 - x2)||^2}{2\sigma^2}}$$

Polynomial Kernel

- $K(X1, X2) = \ \phi(X1).\phi(X2)$

$$X1^T.X2 = \begin{bmatrix} X1 \\ X2 \end{bmatrix} . \begin{bmatrix} X1 & X2 \end{bmatrix}$$

$$= \begin{bmatrix} X1^2 & X1.X2 \\ X1.X2 & X2^2 \end{bmatrix}$$