Announcements

Assignments

- HW6 (written + programming)
  - Due Thu 3/26, 11:59 pm
- HW7 (online)
  - Out later tonight
  - Due Tue 3/31, 11:59 pm

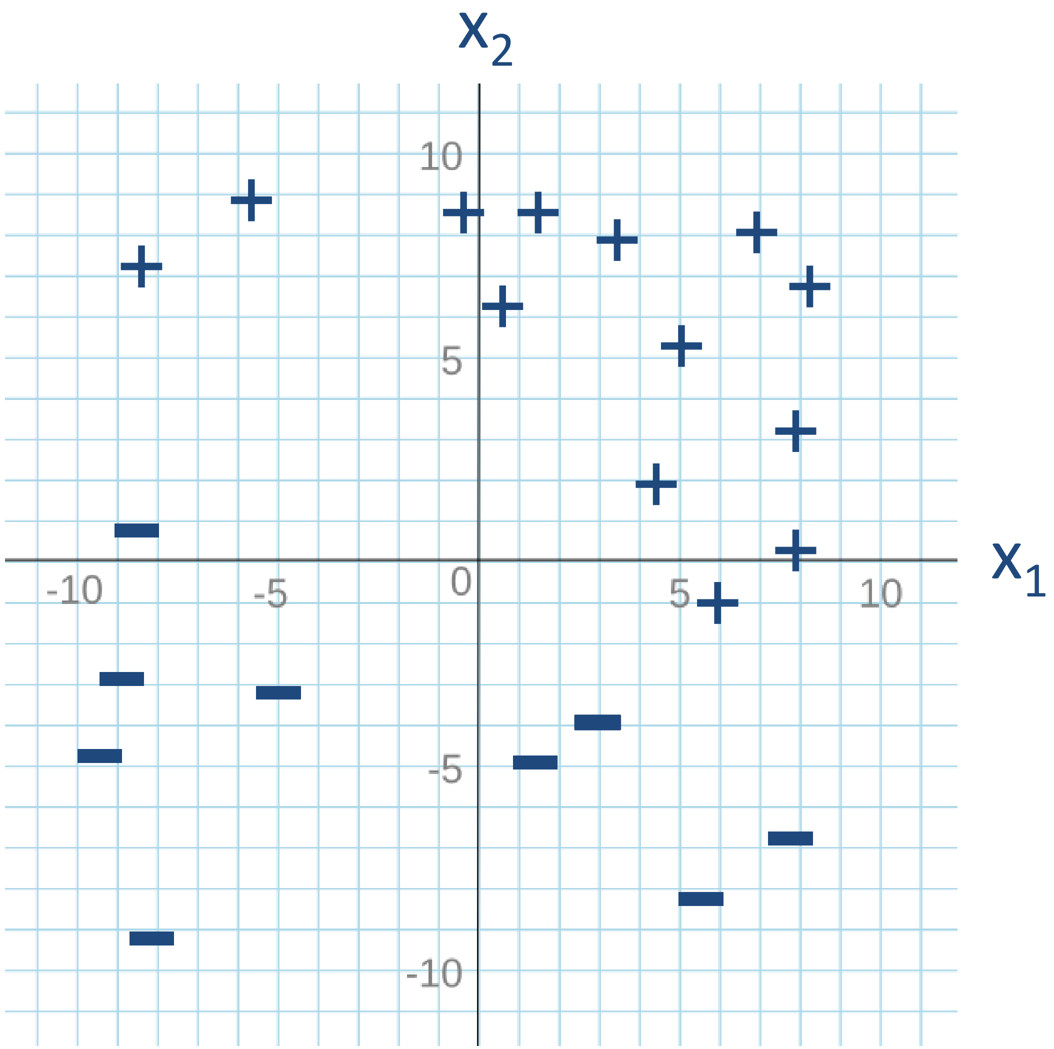
Introduction to Machine Learning

Support Vector Machines

Instructor: Pat Virtue

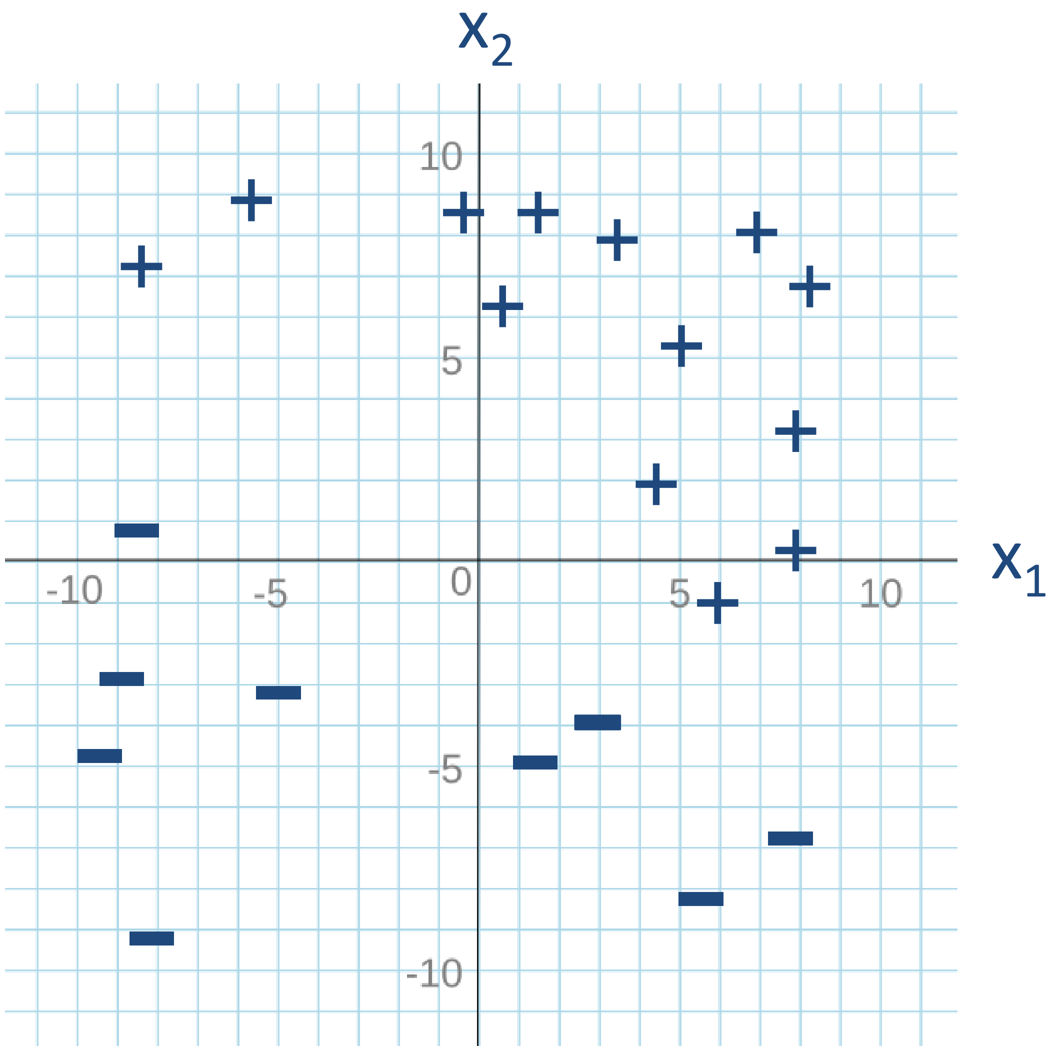
Support Vector Machines

Linear Classification



Support Vector Machines

Max Margin



Constrained Optimization

Linear Program

$$\min_{\boldsymbol{x}} \quad \boldsymbol{c}^T\boldsymbol{x}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} \preceq \boldsymbol{b}$$

Constrained Optimization

Linear Program

$$\min_{\boldsymbol{x}} \quad \boldsymbol{c}^T \boldsymbol{x}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} \preceq \boldsymbol{b}$$

Solvers

- Simplex
- Interior point methods

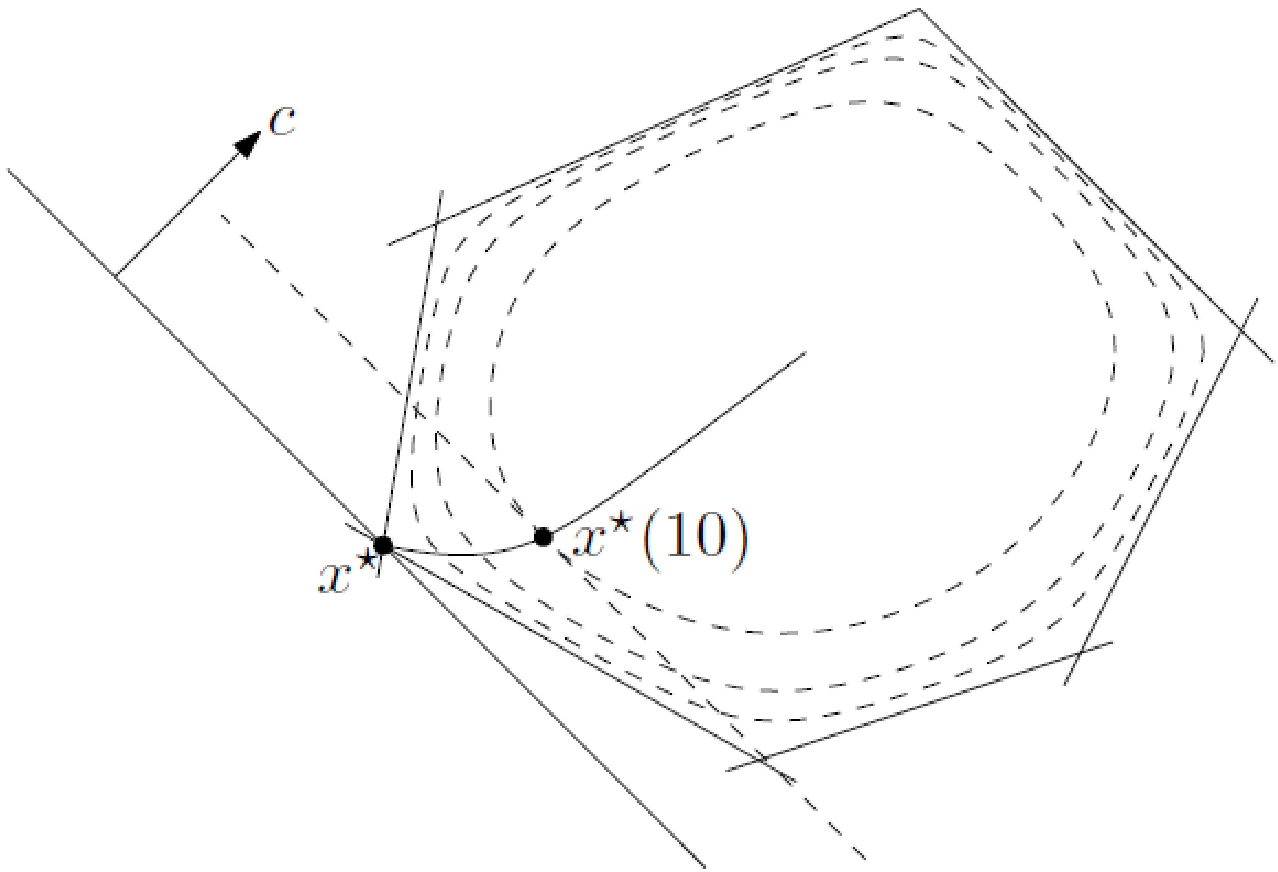


Figure: Fig 11.2 from Boyd and Vandenberghe, *Convex Optimization*

Constrained Optimization

Linear Program

$$\min_{\boldsymbol{x}} \quad \boldsymbol{c}^T \boldsymbol{x}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} \preceq \boldsymbol{b}$$

Quadratic Program

$$\min_{\boldsymbol{x}} \quad \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{c}^T \boldsymbol{x}$$
$$\text{s.t.} \quad A\boldsymbol{x} \preceq \boldsymbol{b}$$

Solvers

- Simplex
- Interior point methods

Constrained Optimization

Linear Program

$$\min_{\boldsymbol{x}} \quad \boldsymbol{c}^T\boldsymbol{x}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} \preceq \boldsymbol{b}$$

Solvers

- Simplex
- Interior point methods

Quadratic Program

$$\min_{\boldsymbol{x}} \quad \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{c}^T\boldsymbol{x}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} \preceq \boldsymbol{b}$$

Solvers

- Conjugate gradient
- Ellipsoid method
- Interior point methods

Constrained Optimization

Linear Program

$$\min_{\boldsymbol{x}} \quad \boldsymbol{c}^T \boldsymbol{x}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} \preceq \boldsymbol{b}$$

Quadratic Program

$$\min_{\boldsymbol{x}} \quad \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{c}^T \boldsymbol{x}$$
$$\text{s.t.} \quad A\boldsymbol{x} \preceq \boldsymbol{b}$$

Solvers

- Simplex
- Interior point methods

Special Case

- If $\boldsymbol{Q}$ is positive-definite, the problem is convex
- $\boldsymbol{Q}$ is positive-definite if:
  $\boldsymbol{v}^T \boldsymbol{Q} \boldsymbol{v} > 0 \quad \forall\, \boldsymbol{v} \in \mathbb{R}^M \backslash \boldsymbol{0}$

Convex Optimization

Linear function

If $f(\boldsymbol{x})$ is linear, then:

- $f(\boldsymbol{x} + \boldsymbol{z}) = f(\boldsymbol{x}) + f(\boldsymbol{z})$
- $f(\theta \boldsymbol{x}) = \theta f(\boldsymbol{x}) \quad \forall\, \theta$
- $f(\theta \boldsymbol{x} + (1 - \theta)\boldsymbol{z}) = \theta f(\boldsymbol{x}) + (1 - \theta) f(\boldsymbol{z}) \quad \forall\, \theta$

Convex Optimization

Convex function

If $f(\boldsymbol{x})$ is convex, then:

- $f(\theta \boldsymbol{x} + (1-\theta)\boldsymbol{z}) \leq \theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{z}) \quad \forall\, 0 \leq \theta \leq 1$

Convex optimization

If $f(\boldsymbol{x})$ is convex, then:

- Every local minimum is also a global minimum ☺

Constrained Optimization

Linear Program

$$\min_{\boldsymbol{x}} \quad \boldsymbol{c}^T \boldsymbol{x}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} \preceq \boldsymbol{b}$$

Quadratic Program

$$\min_{\boldsymbol{x}} \quad \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{c}^T \boldsymbol{x}$$
$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} \preceq \boldsymbol{b}$$

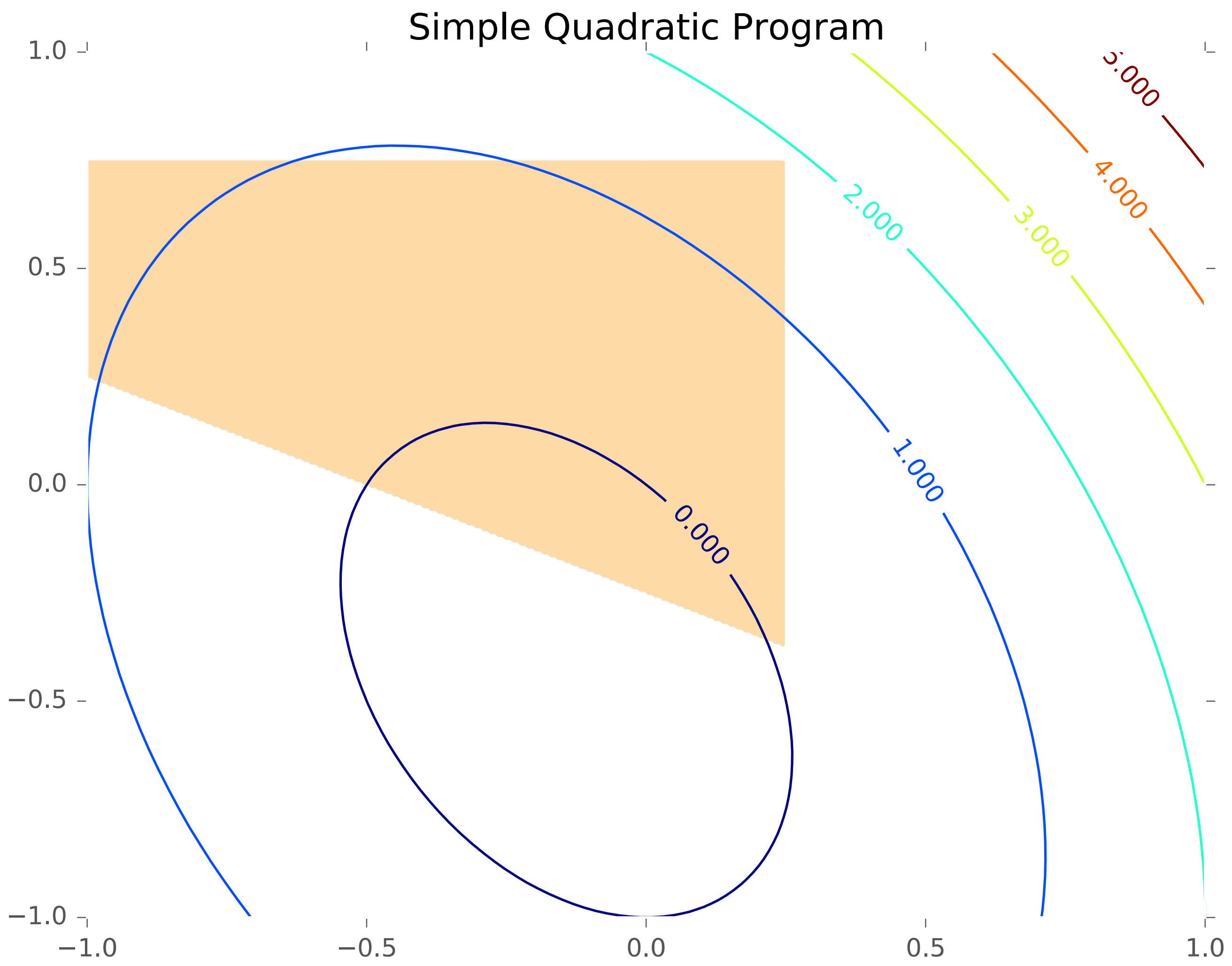
Solvers

- Simplex
- Interior point methods

Special Case

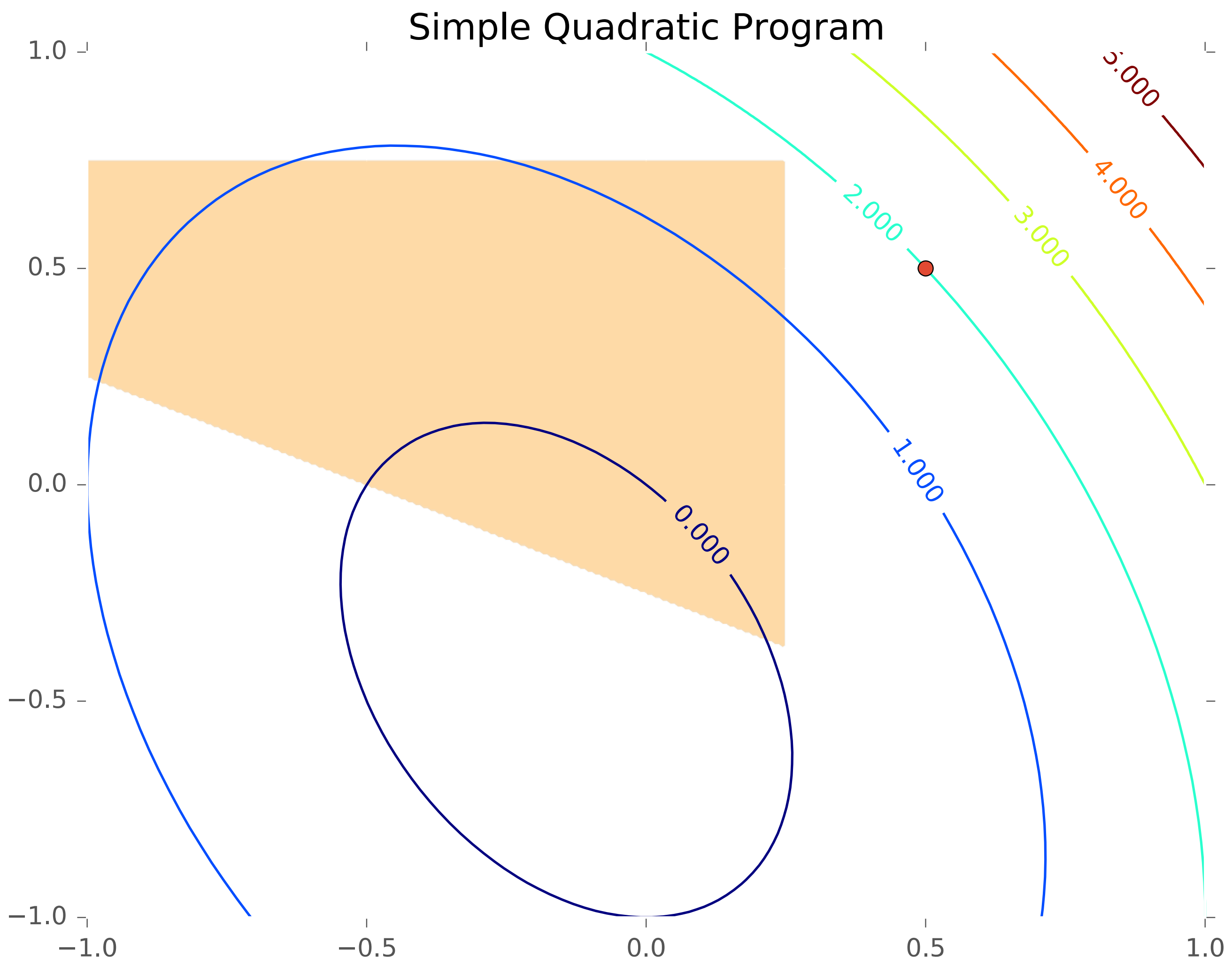
- If $\boldsymbol{Q}$ is positive-definite, the problem is convex
- $\boldsymbol{Q}$ is positive-definite if:
  $\boldsymbol{v}^T \boldsymbol{Q} \boldsymbol{v} > 0 \quad \forall\, \boldsymbol{v} \in \mathbb{R}^M \backslash \boldsymbol{0}$

Quadratic Program



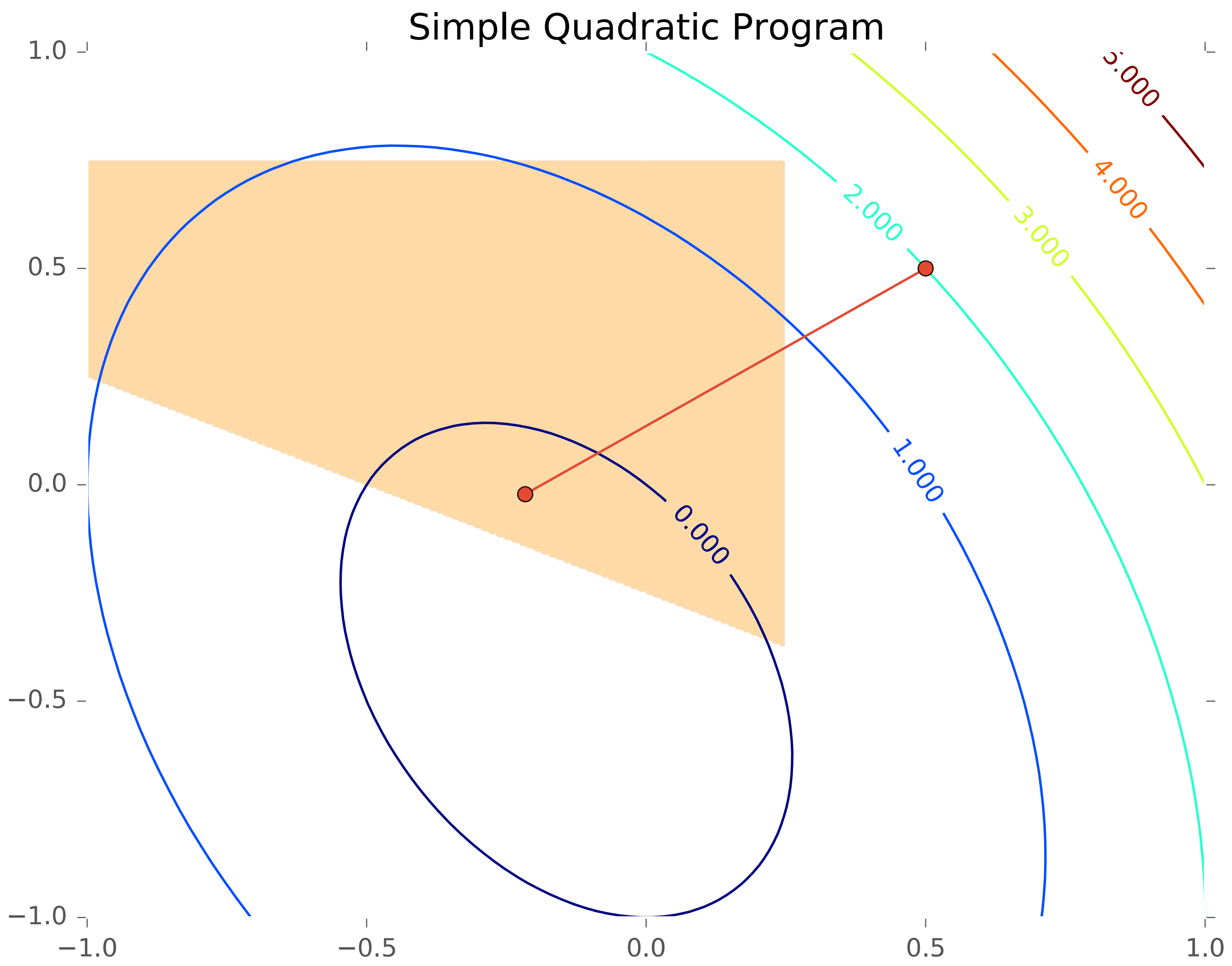
Slide credit: CMU MLD Matt Gormley

Quadratic Program



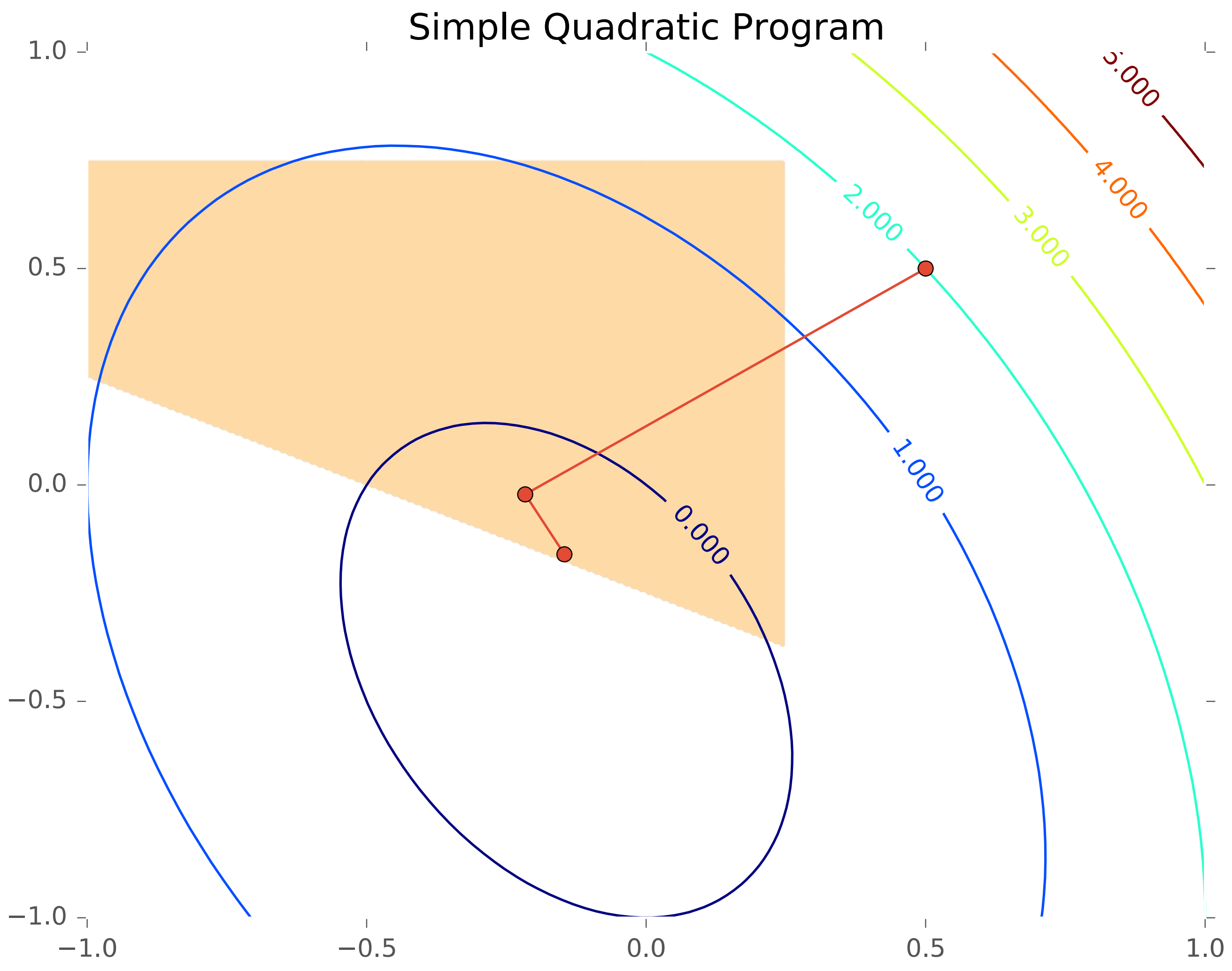
Slide credit: CMU MLD Matt Gormley

Quadratic Program



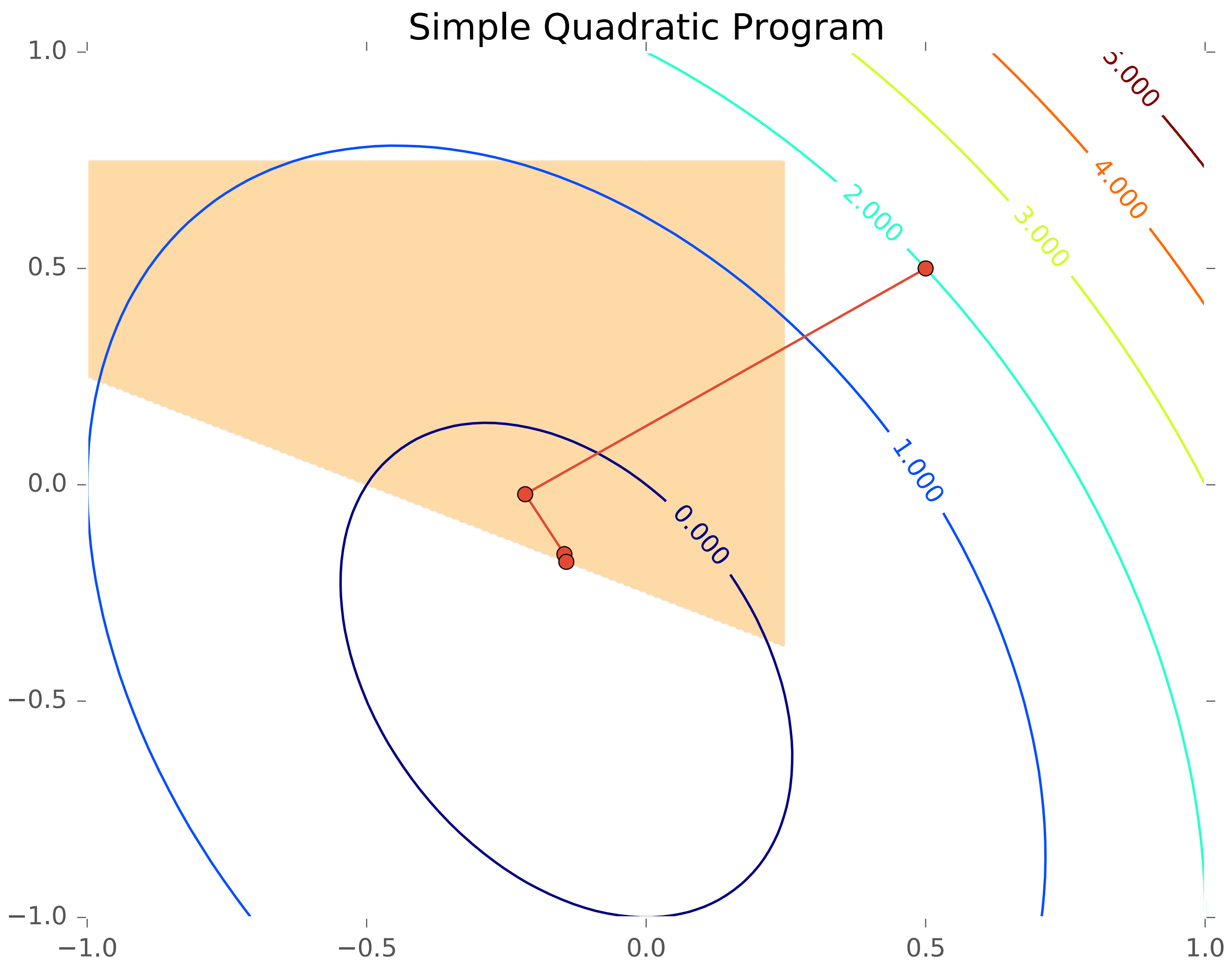
Slide credit: CMU MLD Matt Gormley

Quadratic Program



Slide credit: CMU MLD Matt Gormley

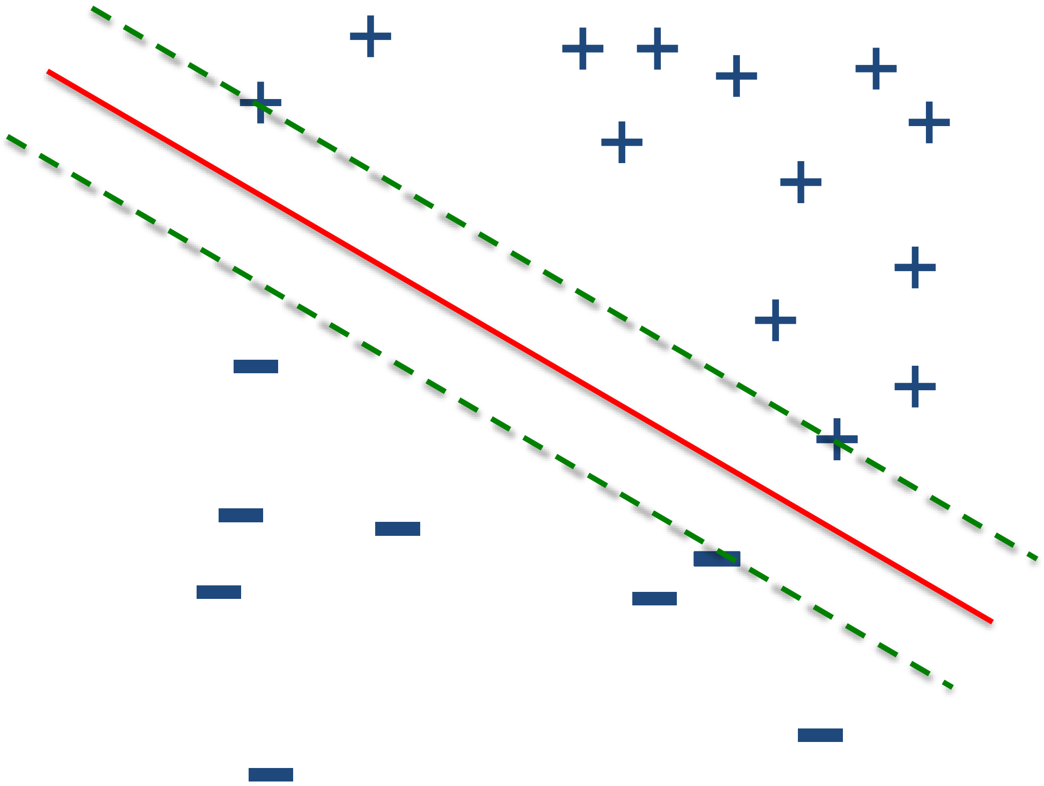
Quadratic Program



Slide credit: CMU MLD Matt Gormley

Support Vector Machines

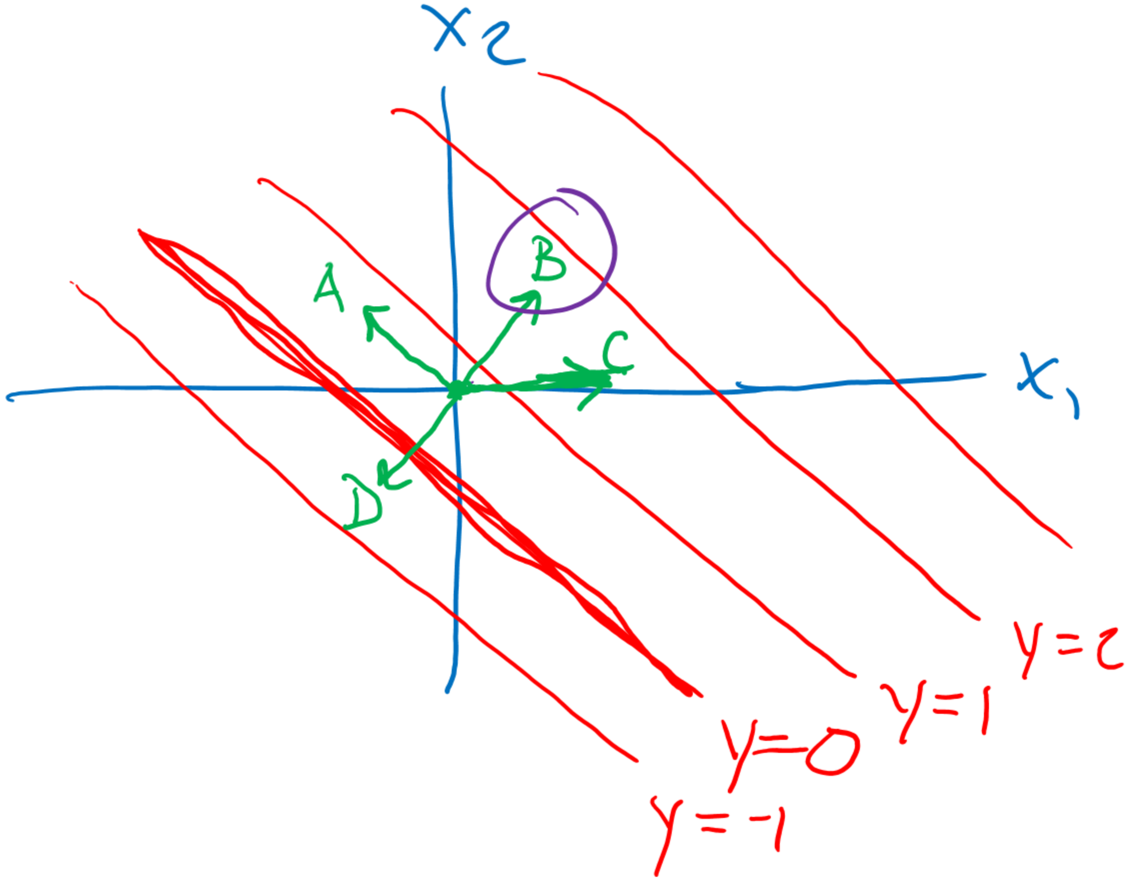
Find linear separator with maximum margin



(Lecture 5) Poll

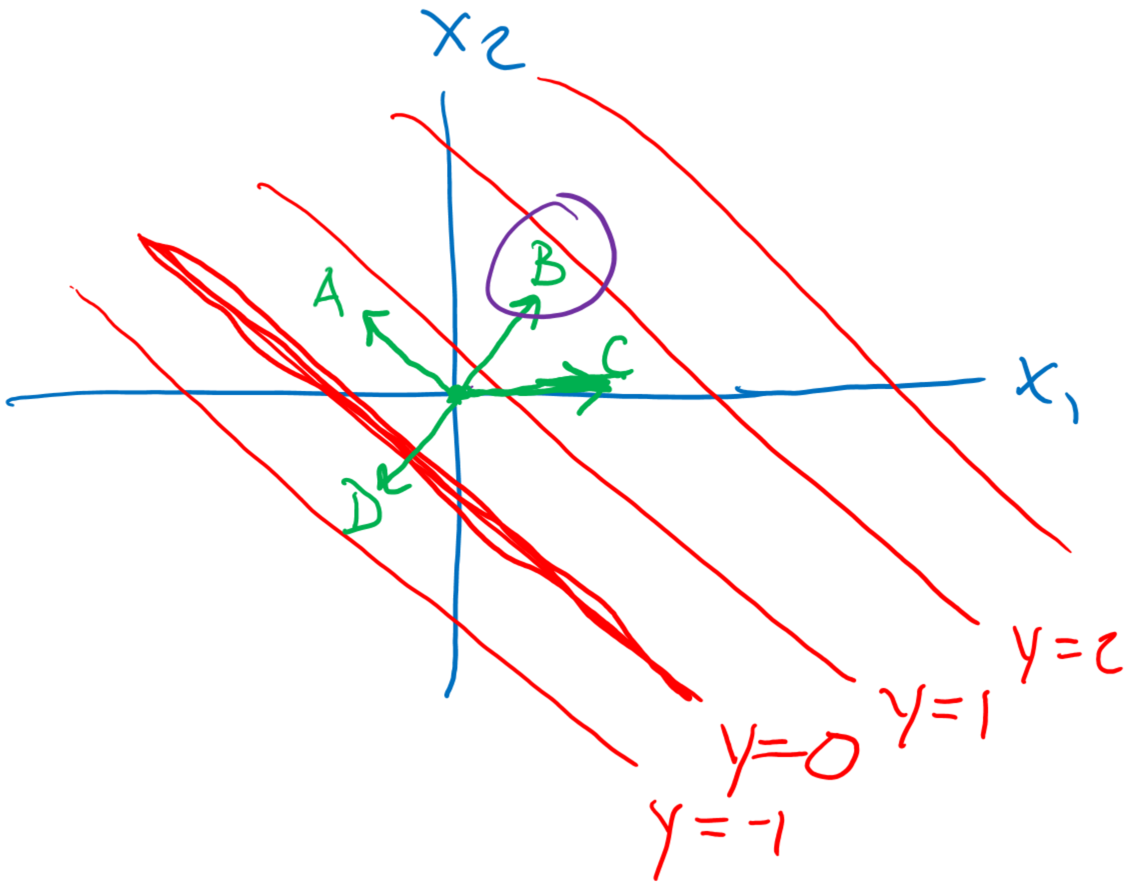
Which vector is the correct $\boldsymbol{w}$?

$y = \boldsymbol{w}^T \boldsymbol{x} + b$



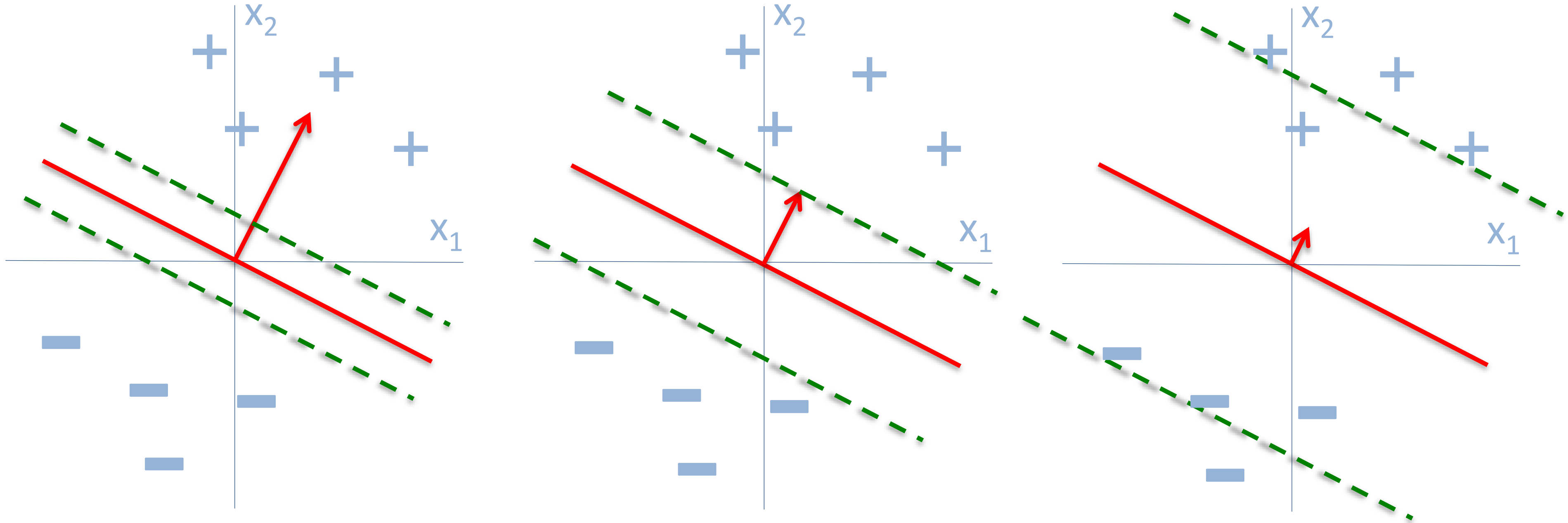
Piazza Poll 1

As the magnitude of w increases, will the distance between the contour lines of $y = \boldsymbol{w}^T\boldsymbol{x} + b$ increase or decrease?



Support Vector Machines

Find linear separator with maximum margin



Linear Separability

Data

$$\mathcal{D} = \{\boldsymbol{x}^{(i)}, y^{(i)}\}_{i=1}^{N} \quad \boldsymbol{x} \in \mathbb{R}^M, \quad y \in \{-1, +1\}$$

Linearly separable iff:

$$\exists\, \boldsymbol{w}, b \qquad s.t. \quad \boldsymbol{w}^T \boldsymbol{x}^{(i)} + b > 0 \quad \text{if} \quad y^{(i)} = +1 \quad \text{and}$$

$$\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b < 0 \quad \text{if} \quad y^{(i)} = -1$$

Linear Separability

Data

$$\mathcal{D} = \{\boldsymbol{x}^{(i)}, y^{(i)}\}_{i=1}^{N} \quad \boldsymbol{x} \in \mathbb{R}^M, \quad y \in \{-1, +1\}$$

Linearly separable iff:

$$\exists\, \boldsymbol{w}, b \quad s.t. \quad \boldsymbol{w}^T \boldsymbol{x}^{(i)} + b > 0 \text{ if } y^{(i)} = +1 \text{ and}$$

$$\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b < 0 \text{ if } y^{(i)} = -1$$

$$\Leftrightarrow \exists\, \boldsymbol{w}, b \quad s.t. \quad y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) > 0$$

$$\Leftrightarrow \exists\, \boldsymbol{w}, b, c \quad s.t. \quad y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) \geq c \text{ and } c > 0$$

$$\Leftrightarrow \exists\, \boldsymbol{w}, b \quad s.t. \quad y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) \geq 1$$

Slide credit: CMU MLD Matt Gormley

Piazza Poll 2

Are these to statements equivalent?

$$\exists\, \boldsymbol{w}, b, c \quad s.t. \quad y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) \geq c \quad \text{and} \quad c > 0$$

$$\exists\, \boldsymbol{w}, b \quad s.t. \quad y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) \geq 1$$

Linear Separability

Data

$$\mathcal{D} = \{\boldsymbol{x}^{(i)}, y^{(i)}\}_{i=1}^{N} \quad \boldsymbol{x} \in \mathbb{R}^M, \quad y \in \{-1, +1\}$$

Linearly separable iff:

$$\exists\, \boldsymbol{w}, b \quad s.t. \quad \boldsymbol{w}^T \boldsymbol{x}^{(i)} + b > 0 \;\text{ if }\; y^{(i)} = +1 \quad \text{and}$$

$$\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b < 0 \;\text{ if }\; y^{(i)} = -1$$

$$\Leftrightarrow\; \exists\, \boldsymbol{w}, b \quad s.t. \quad y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) > 0$$

$$\Leftrightarrow\; \exists\, \boldsymbol{w}, b, c \quad s.t. \quad y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) \geq c \quad \text{and} \quad c > 0$$

$$\Leftrightarrow\; \exists\, \boldsymbol{w}, b \quad s.t. \quad y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) \geq 1$$

Support Vector Machines

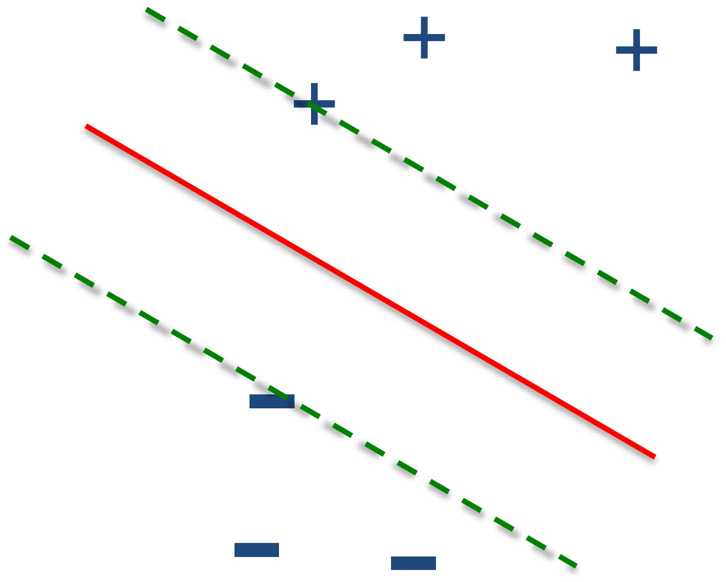
Find linear separator with maximum margin

Let $\boldsymbol{x}_+$ and $\boldsymbol{x}_-$ be hypothetical points on the +/- margin from the decision boundary

$$\exists\, \boldsymbol{w}, b \quad s.t. \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1$$

$$\Leftrightarrow \exists\, \boldsymbol{w}, b \quad s.t. \quad \boldsymbol{w}^T\boldsymbol{x}_+ + b \geq +1 \quad \text{and}$$

$$\boldsymbol{w}^T\boldsymbol{x}_- + b \leq -1$$



Consider the vector from $\boldsymbol{x}_-$ to $\boldsymbol{x}_+$ and its projection onto the vector $\boldsymbol{w}$:

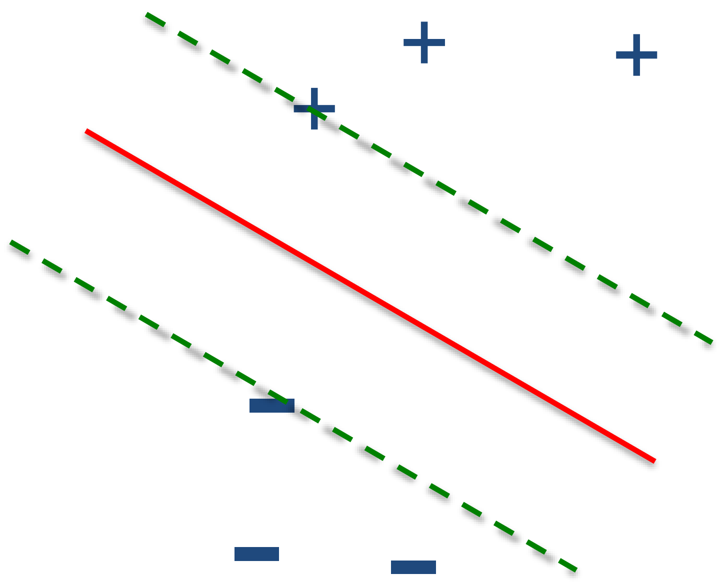
Support Vector Machines

$$width = \frac{\boldsymbol{w}^T}{\|\boldsymbol{w}\|_2}(\boldsymbol{x}_{+} - \boldsymbol{x}_{-})$$

Find linear separator with maximum margin

$$\max_{\boldsymbol{w},\boldsymbol{b}} \quad \text{"width"}$$

$$\text{s.t.} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 \quad \forall\, i$$

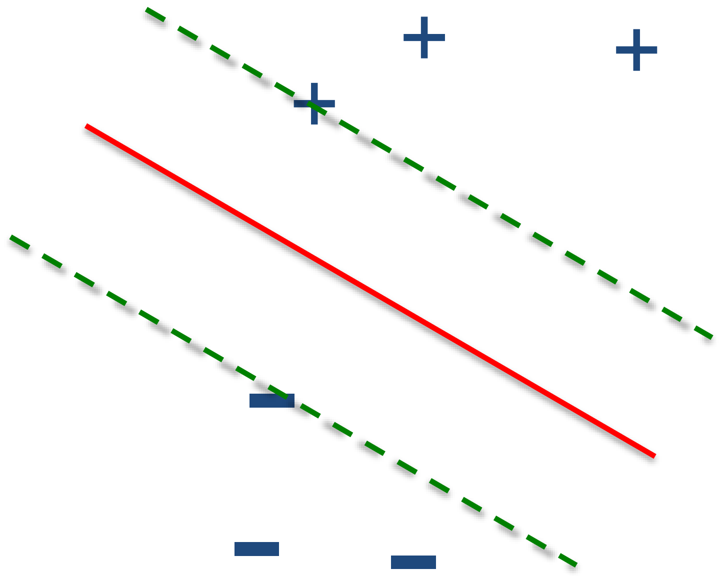


Support Vector Machines

$$\text{width} = \frac{2}{\|w\|_2}$$

Find linear separator with maximum margin

$$\underset{\boldsymbol{w},\boldsymbol{b}}{\operatorname{argmax}} \quad \text{width}$$



Support Vector Machines

$$\text{width} = \frac{2}{\|\boldsymbol{w}\|_2}$$

Find linear separator with maximum margin

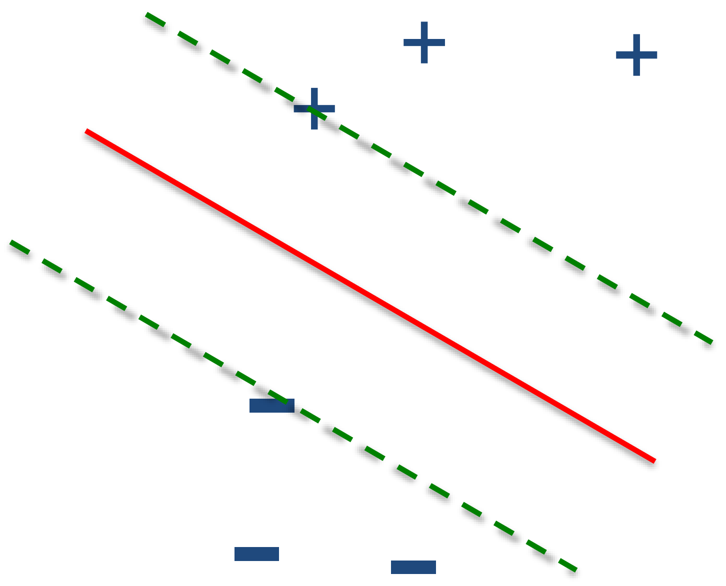
$$\underset{\boldsymbol{w},b}{\operatorname{argmax}} \quad \text{width}$$

$$\Leftrightarrow \underset{\boldsymbol{w},b}{\operatorname{argmax}} \quad \frac{2}{\|\boldsymbol{w}\|_2}$$

$$\Leftrightarrow \underset{\boldsymbol{w},b}{\operatorname{argmin}} \quad \frac{1}{2}\|\boldsymbol{w}\|_2$$

$$\Leftrightarrow \underset{\boldsymbol{w},b}{\operatorname{argmin}} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2$$

$$\Leftrightarrow \underset{\boldsymbol{w},b}{\operatorname{argmin}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$



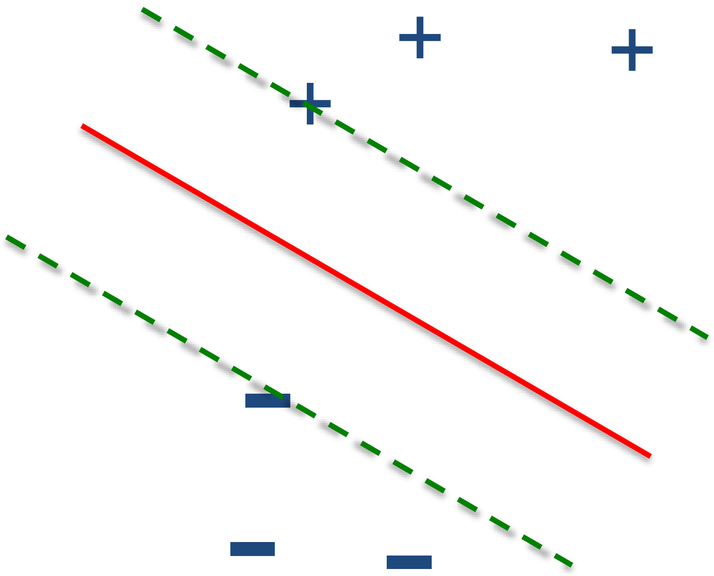
SVM Optimization

Quadratic Program

$$\min_{\boldsymbol{x}} \quad \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{c}^T \boldsymbol{x}$$
$$\text{s.t.} \quad A\boldsymbol{x} \preceq \boldsymbol{b}$$

Quadratic program!

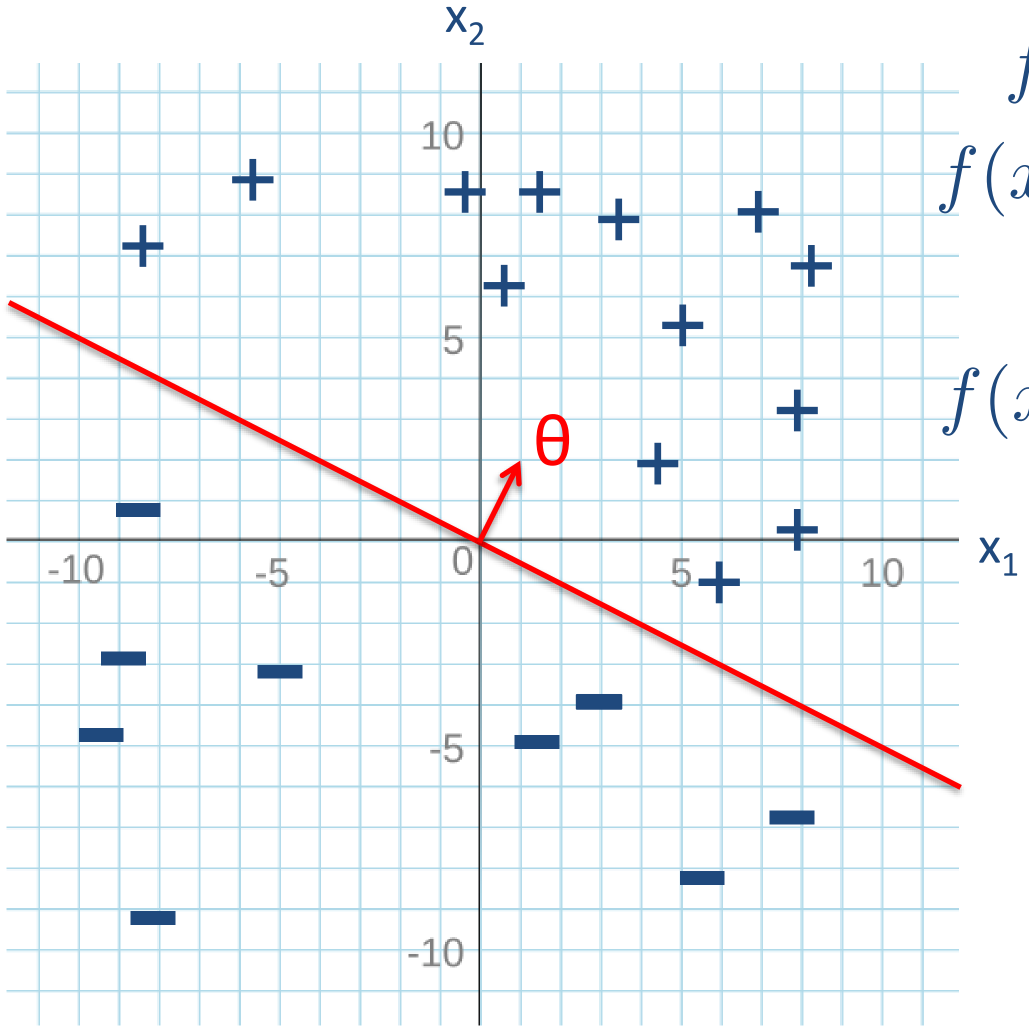
$$\min_{\boldsymbol{w},\boldsymbol{b}} \quad \frac{1}{2} \boldsymbol{w}^T \boldsymbol{w}$$
$$\text{s.t.} \quad y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) \geq 1 \quad \forall\, i$$



SVM Optimization

How did we go from maximizing margin to minimizing $||\boldsymbol{w}||_2$?

Find a θ vector that causes f(**x**) to separate the data in to positive and negative based on class



$$f(x_1, x_2) = \theta_1 x_1 + \theta_2 x_2$$

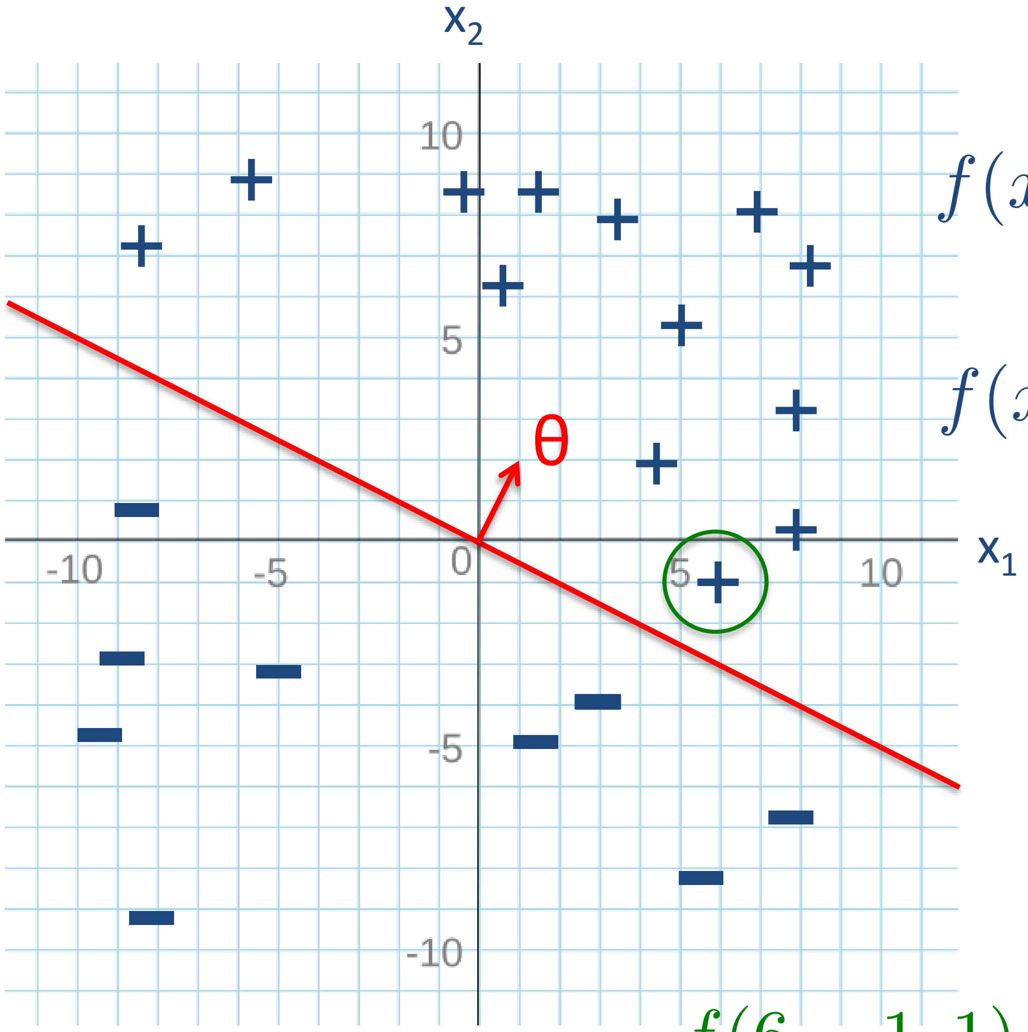
$$f(x_1, x_2, y) = y(\theta_1 x_1 + \theta_2 x_2)$$

$$\theta = [1; 2]$$

$$f(x_1, x_2, y) = y(x_1 + 2x_2)$$

Note: We added y = ±1 to the function to make any correct classification have an f(**x**, y) > 0

Given θ, what is the value of f(**x**, y) for the points closest to the hyperplane?



$$f(x_1, x_2, y) = y(\theta_1 x_1 + \theta_2 x_2)$$

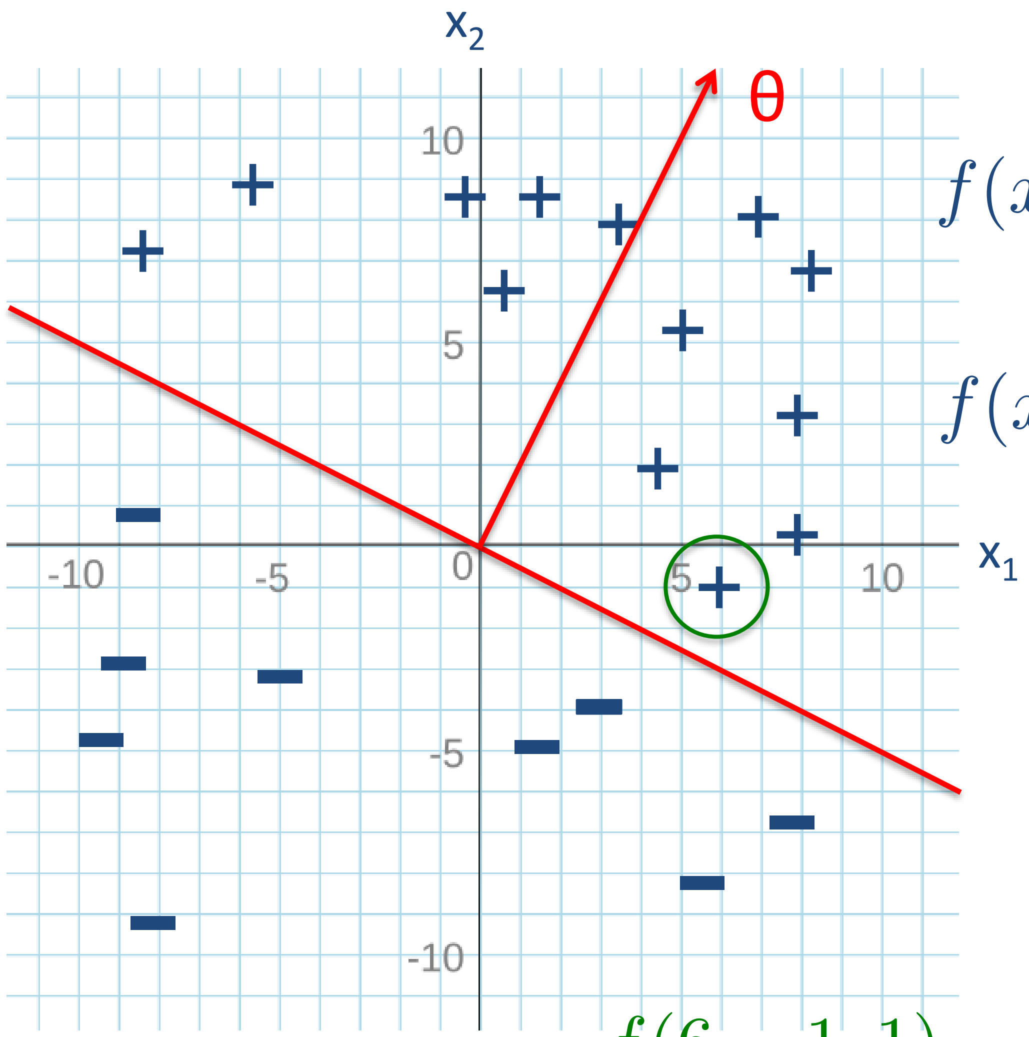
$$\theta = [1; 2]$$

$$f(x_1, x_2, y) = y(x_1 + 2x_2)$$

Closest positive point
(a support vector)
x = [6; -1], y = 1

$$f(6, -1, 1) = 1(1 \cdot 6 + 2 \cdot -1) = 4$$

What happens to the value of f(**x**, y) if we scale θ?



$$f(x_1, x_2, y) = y(\theta_1 x_1 + \theta_2 x_2)$$

$$\theta = [5; 10]$$

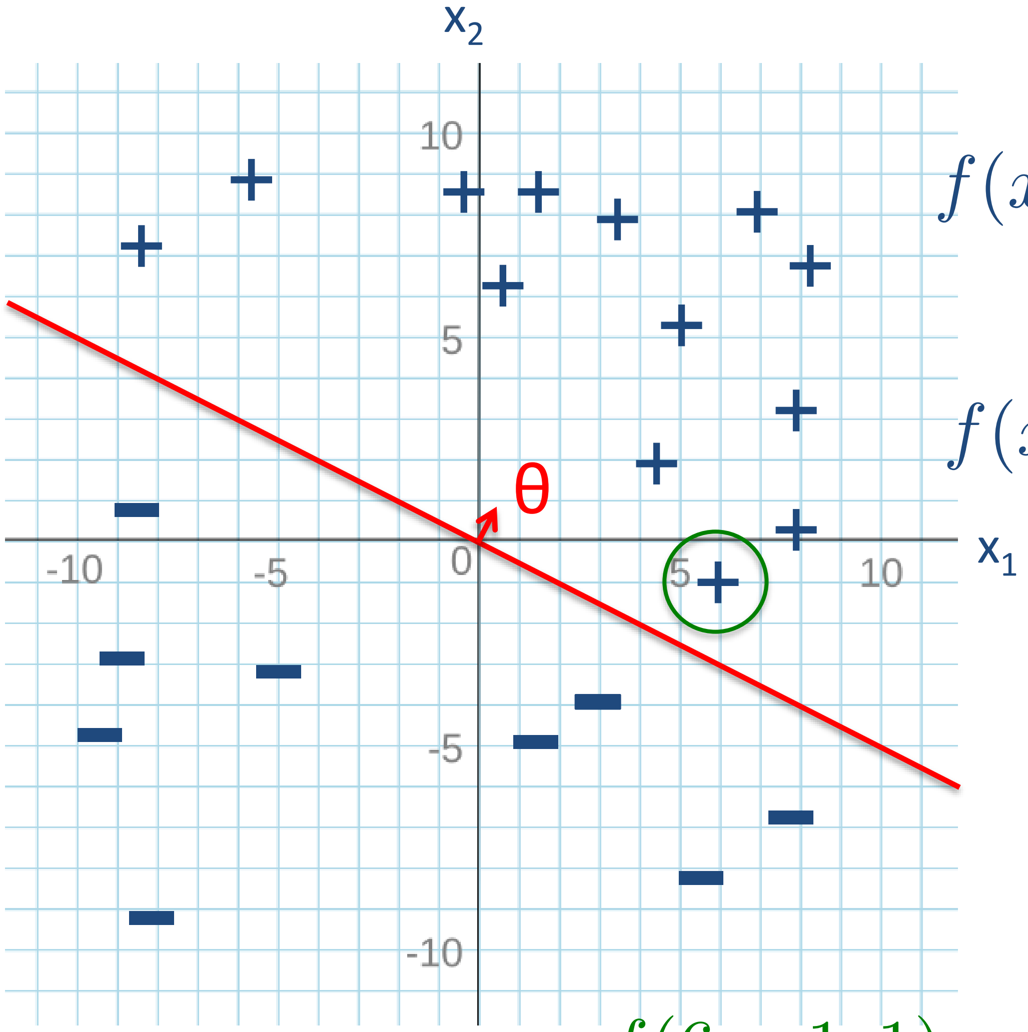
$$f(x_1, x_2, y) = y(5x_1 + 10x_2)$$

Closest positive point
(a support vector)
x = [6; -1], y = 1

$$f(6, -1, 1) = 1(5 \cdot 6 + 10 \cdot -1) = 20$$

The value of f(**x**, y) increases even though the hyperplane does not change

We can scale θ, such that f(**x**, y) = 1 for the support vector points



$$f(x_1, x_2, y) = y(\theta_1 x_1 + \theta_2 x_2)$$

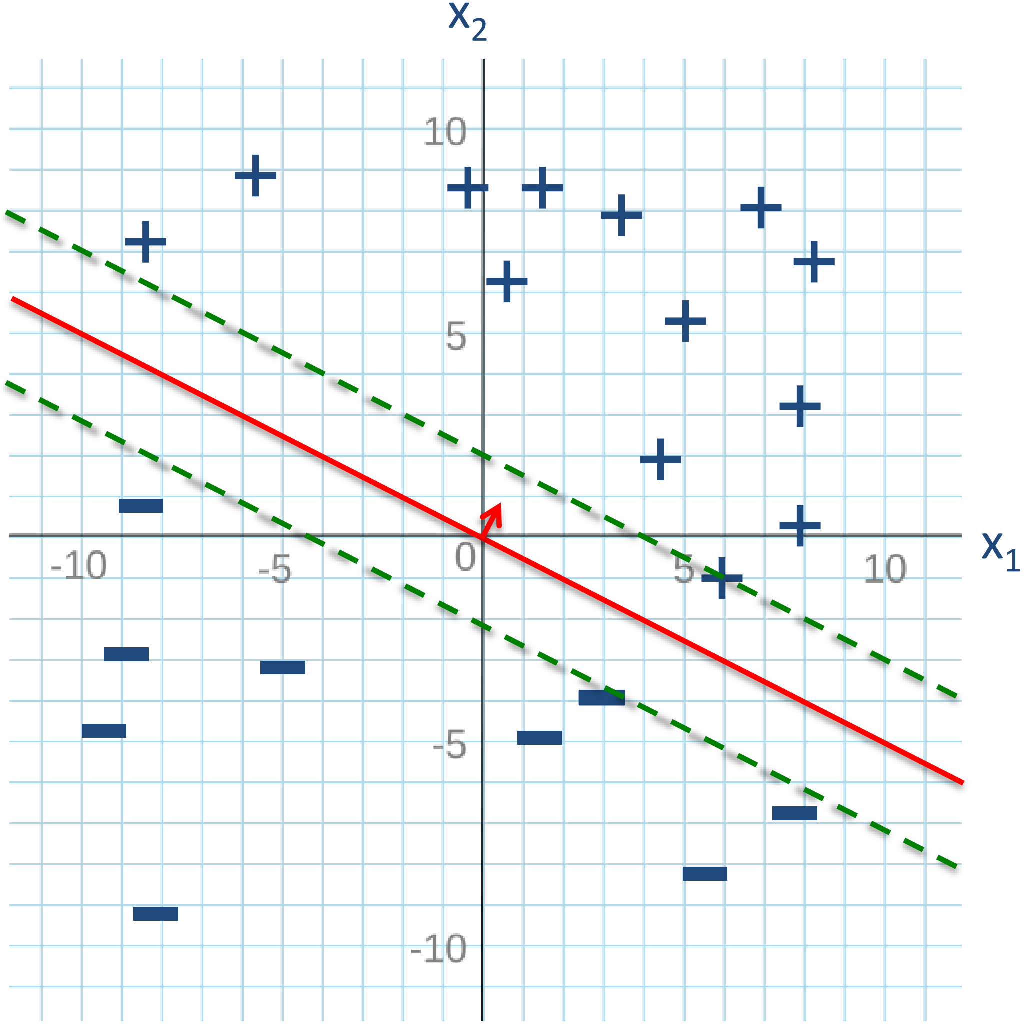
$$\theta = [1/4; 1/2]$$

$$f(x_1, x_2, y) = y(\frac{1}{4}x_1 + \frac{1}{2}x_2)$$

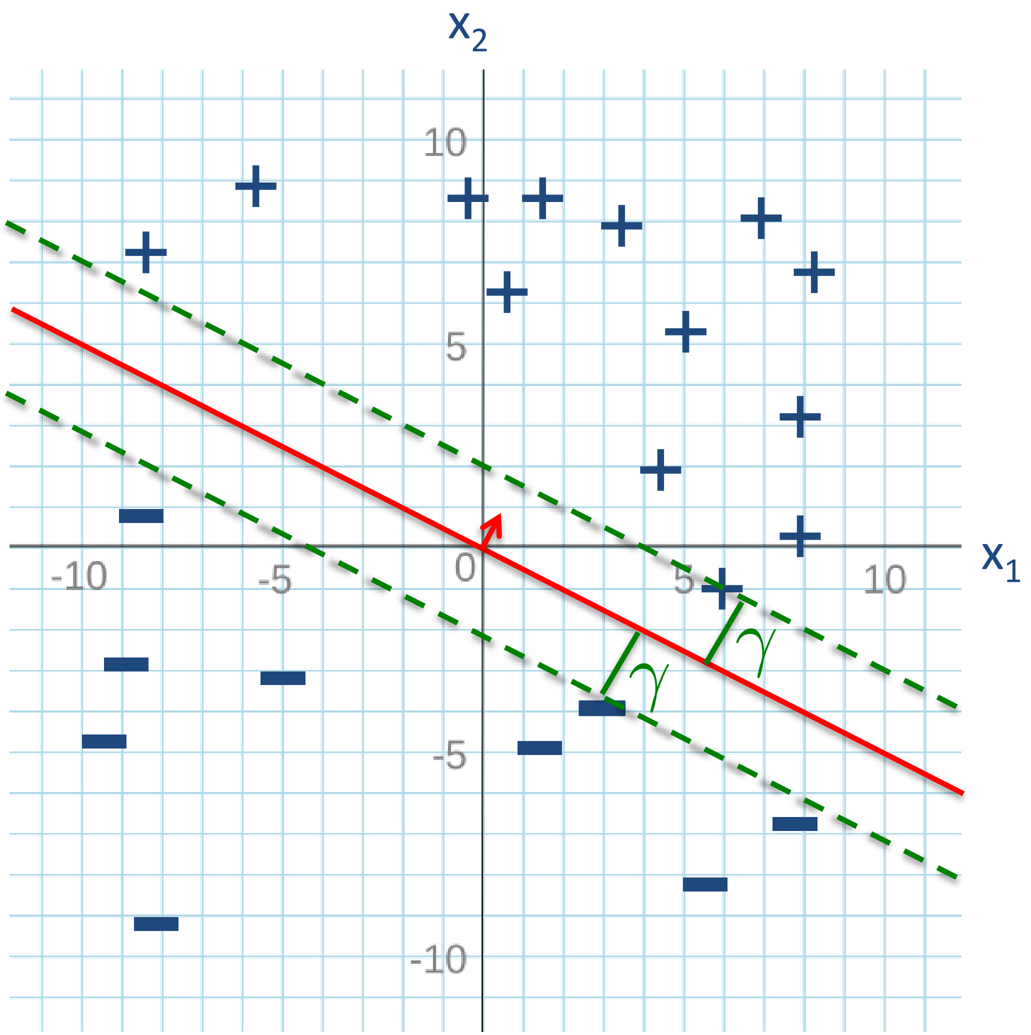
Closest positive point
(a support vector)
x = [6; -1], y = 1

$$f(6, -1, 1) = 1(\frac{1}{4} \cdot 6 + \frac{1}{2} \cdot -1) = 1$$

With f($\mathbf{x}$, y) = 1 for the support vector points, f($\mathbf{x}^i$, y^i) ≥ 1 for all points



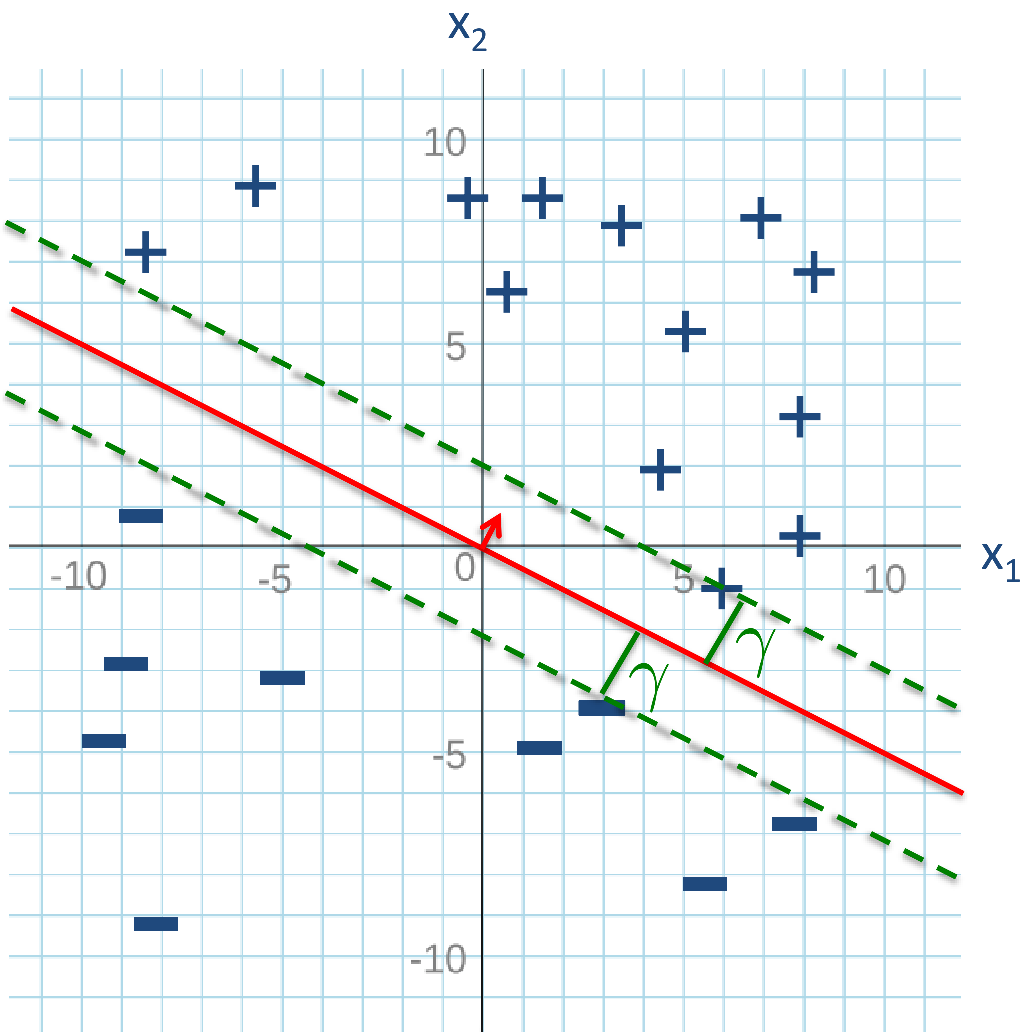
Although f(**x**, y) is dependent on the magnitude of θ, we define the margin by the Euclidean distance (normalized by the magnitude of θ)



Margin γ is the distance from the hyperplane to the support vector points

$$\gamma = \min_i \frac{y^i\ \theta \cdot x^i}{||\theta||}$$

Our decision to make f(**x**, y) = 1 for the support vector points, helps to simplify the definition of margin



For support vectors,

$$f(x, y) = y\ \theta \cdot x = 1$$

The minimum in the margin equation is realized by the support vectors. Thus:

$$\gamma = \min_i \frac{y^i\ \theta \cdot x^i}{||\theta||}$$

becomes

$$\gamma = \frac{1}{||\theta||}$$

Now, we can finally get to optimization

We want to maximize the margin, while still making sure that all points are on the correct side of the margin: f($\mathbf{x}^i$, y^i) ≥ 1

$$\theta^* = \operatorname*{argmax}_{\theta} \gamma \quad \text{s.t. } y^i(\theta \cdot x^i) \geq 1, \forall i$$

From the previous slide

$$\gamma = \frac{1}{||\theta||}$$

$$= \operatorname*{argmax}_{\theta} \frac{1}{||\theta||} \quad \text{s.t. } y^i(\theta \cdot x^i) \geq 1, \forall i$$

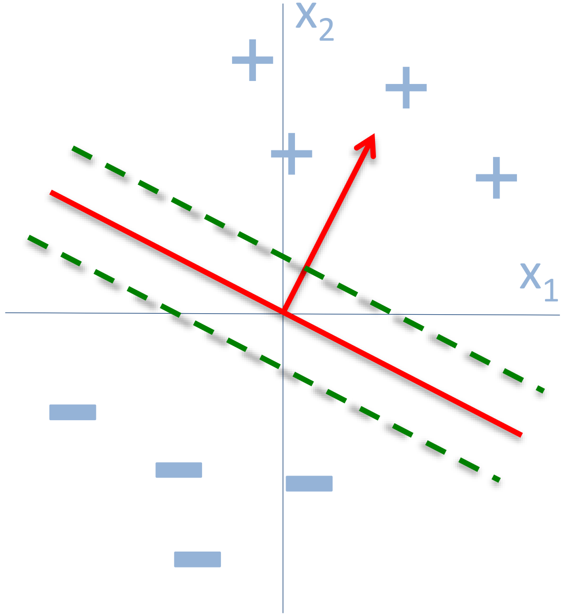
Invert the objective to switch from max to min

$$= \operatorname*{argmin}_{\theta} ||\theta|| \quad \text{s.t. } y^i(\theta \cdot x^i) \geq 1, \forall i$$

The margin is now fixed to the locations where

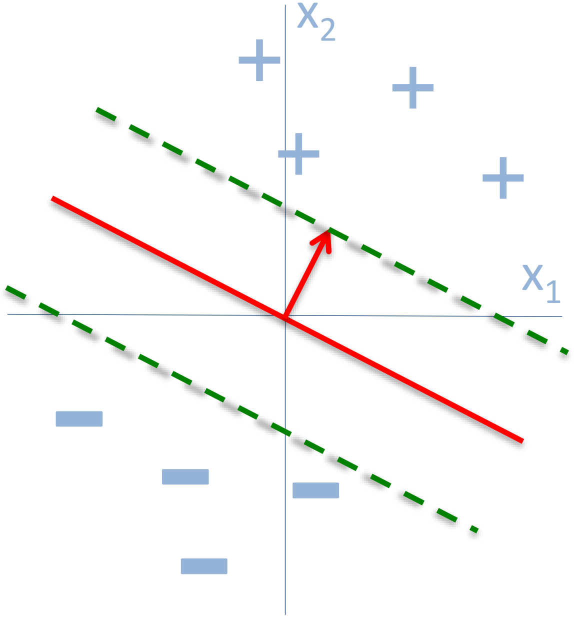
$$f(x,y) = y(\theta \cdot x) = 1$$

Now, let's see what happens when we scale θ



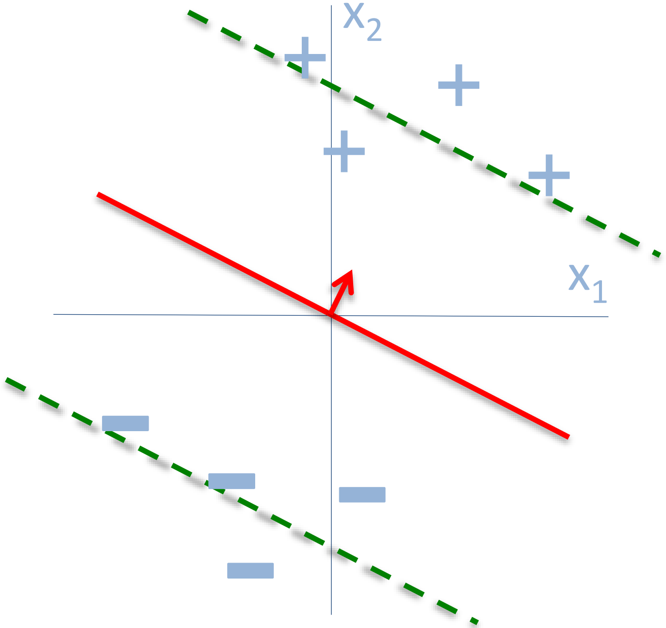
$$||\theta|| = 2$$

$$\frac{1}{||\theta||} = \frac{1}{2}$$



$$||\theta|| = 1$$

$$\frac{1}{||\theta||} = 1$$



$$||\theta|| = \frac{1}{2}$$

$$\frac{1}{||\theta||} = 2$$