Machine Learning from Data

Lecture 23: Spring 2021

Today's Lecture

- Support Vector Machines (SVMs)
  - Maximizing the margins

RECAP: **Linear Models, RBFs, Neural Networks**

<u>Linear Model with Nonlinear Transform</u>

$$h(\mathbf{x}) = \theta\left(w_0 + \sum_{j=1}^{\tilde{d}} w_j \Phi_j(\mathbf{x})\right)$$

<u>Neural Network</u>

$$h(\mathbf{x}) = \theta\left(w_0 + \sum_{j=1}^{m} w_j \theta\left(\mathbf{v}_j{}^{\mathrm{T}}\mathbf{x}\right)\right)$$

gradient descent

<u>k-RBF-Network</u>

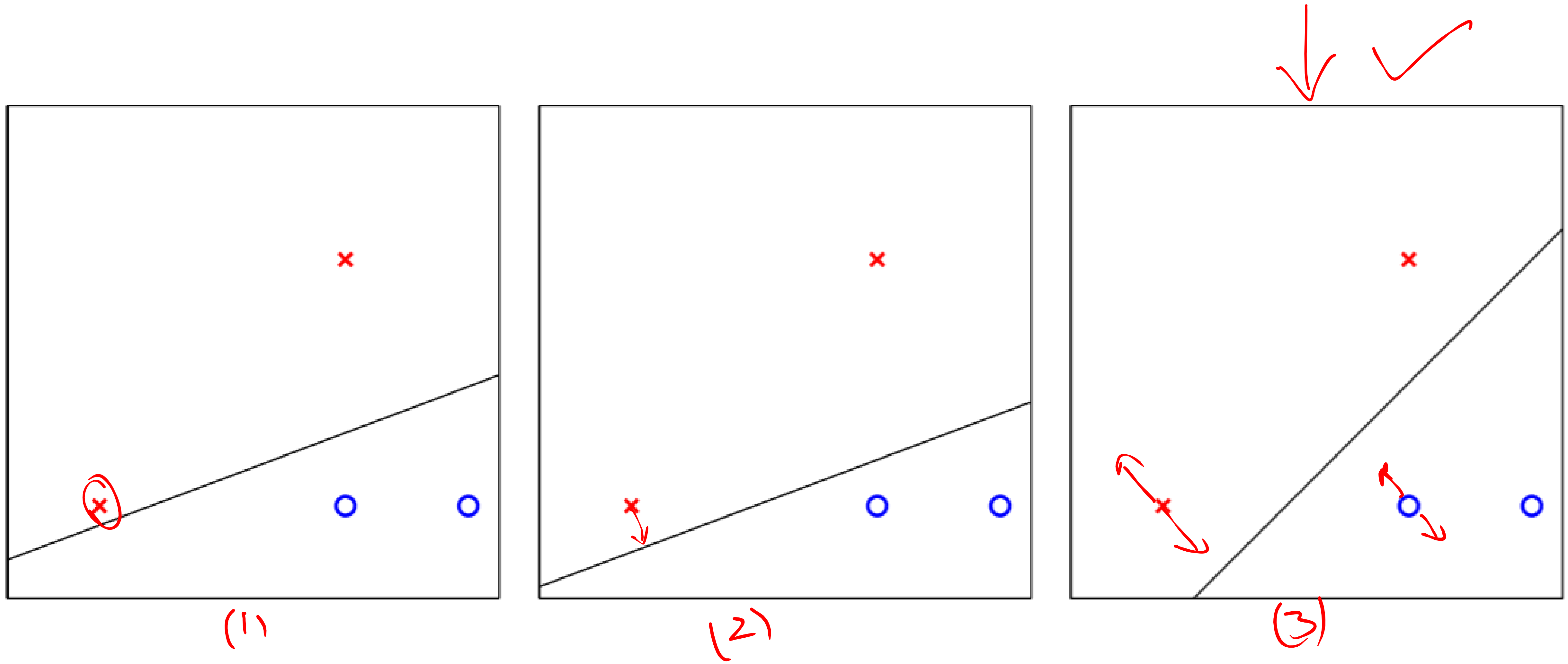
$$h(\mathbf{x}) = \theta\left(w_0 + \sum_{j=1}^{k} w_j \phi\left(\| \mathbf{x} - \boldsymbol{\mu}_j \|\right)\right)$$

k-means

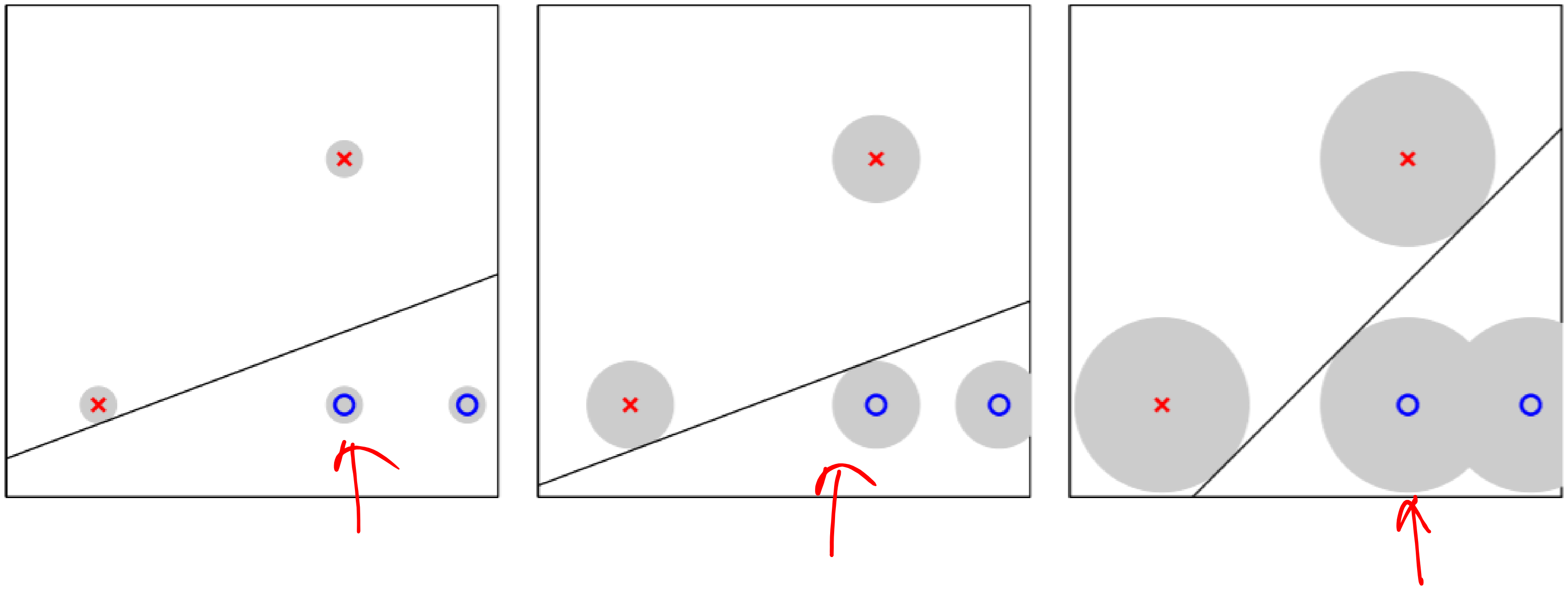
Neural Network: generalization of linear model by adding layers.

Support Vector Machine: more 'robust' linear model

Which Separator Do You Pick?

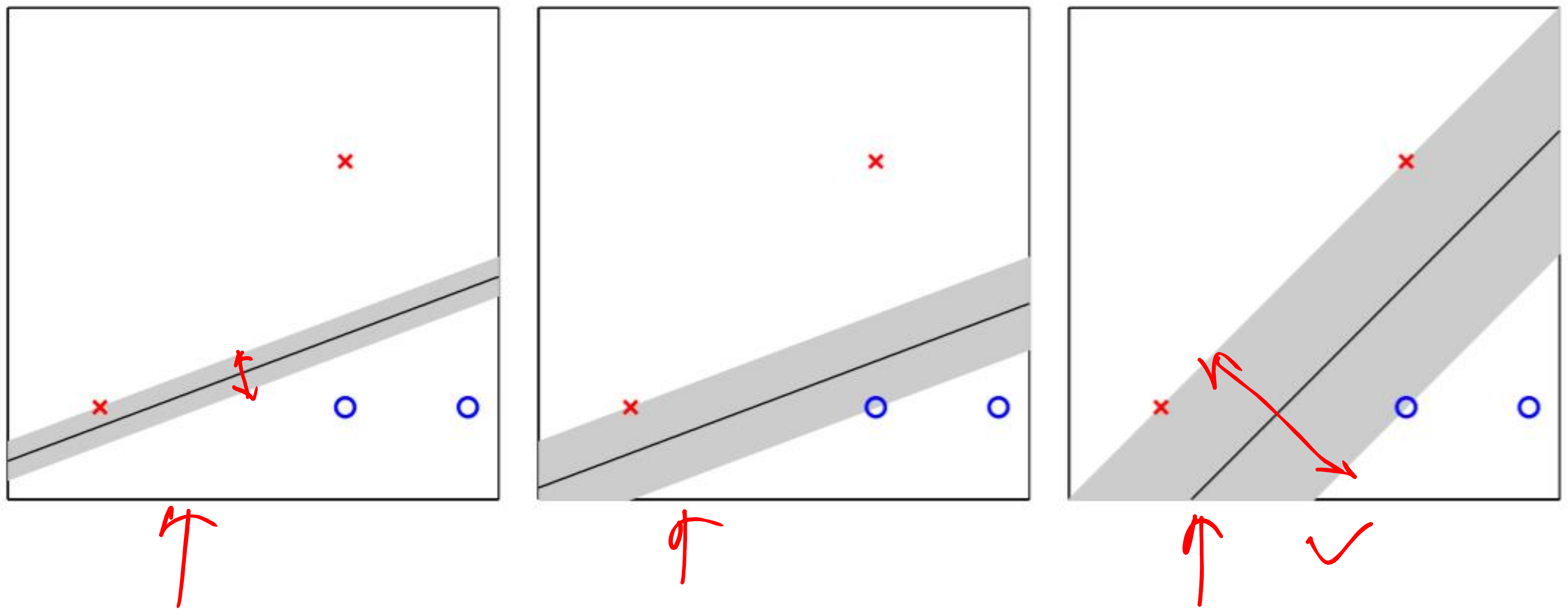


Robustness to Noisy Data



Being robust to noise (measurement error) is good (remember regularization).

Thicker Cushion Means More Robustness



We call such hyperplanes **fat**

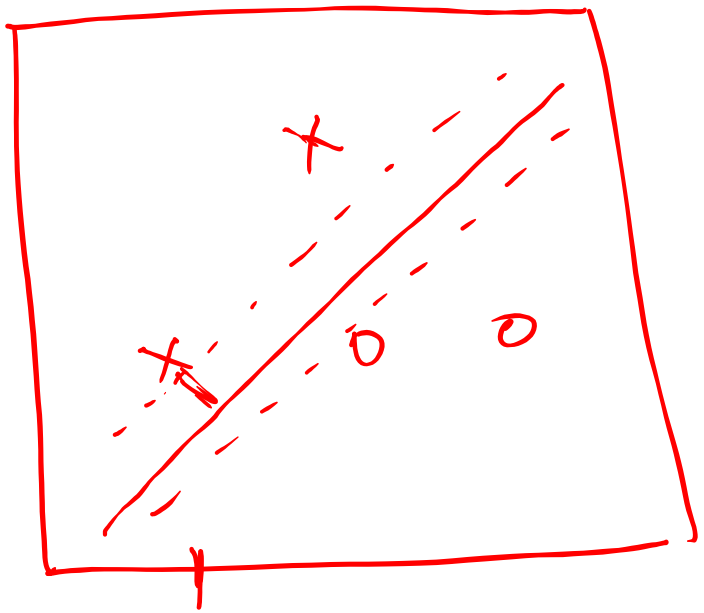
Two Crucial Questions

1. Can we efficiently find the fattest separating hyperplane?

2. Is a fatter hyperplane better than a thin one?

Optimal Hyperplane. $\longrightarrow$ Separates the data.

Thickness: Distance to nearest datapoint



$$x = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad y = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}$$

Hyperplane → set of weights.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \quad w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

w_0 (bias) → b

Earlier $h(x) = \text{sign}(w^T x)$

$h(x) = \text{sign}(w^T x + b)$

3-d space

$+$ $w^T x + b > 0$

$w^T x + b = 0$

$-$ $w^T x + b < 0$

Separating

$$\text{sign}(w^T x_n + b) = y_n$$

$$y_n(\underbrace{w^T x_n + b}) > 0 \quad ; \quad n = 1, 2, \dots N$$

$$\min_n y_n(w^T x_n + b) = \rho > 0$$

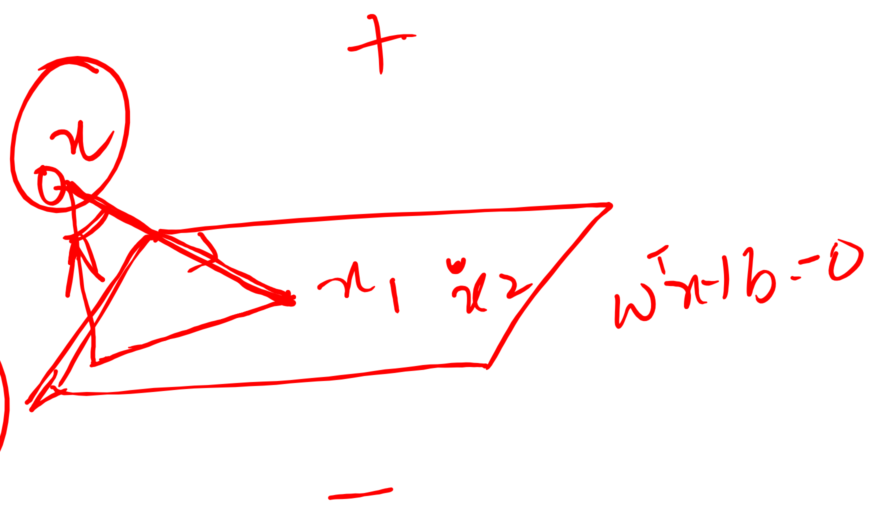
OR $$\min_n y_n\left(\frac{w^T x_n}{\rho} + \frac{b}{\rho}\right) = 1$$

Definition: $h = (w, b)$ separates the data iff

$$\min_n \underbrace{y_n(w^T x_n + b)} = 1$$

$$h = (\underline{w}, b)$$

$$\text{dist}(x, h) = |U \cdot (x - x_1)|$$



$w \rightarrow$ normal (in the direction of U)

or $U = \dfrac{w}{\|w\|}$

$x_2 - x_1 \longrightarrow$ vector , $\left.\begin{array}{l} w^T x_1 + b = 0 \\ w^T x_2 + b = 0 \end{array}\right\}$

$$w^T x_1 - w^T x_2 = 0 \implies w \cdot (x_1 - x_2) = 0$$

$$w / \|w\|$$

$$\mathrm{dist}(x,h) = \left| \frac{w}{\|w\|} \cdot (x - x_1) \right|$$

$$= \frac{1}{\|w\|} \left| w^T x - w^T x_1 \right|$$

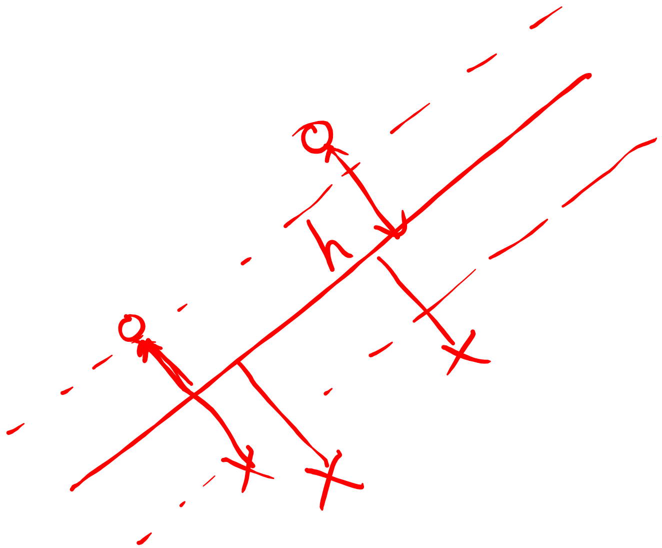
$$= \frac{1}{\|w\|} \left| w^T x + b - \underbrace{(w^T x_1 - b)}_{0} \right|$$

simpler

$$\boxed{\underline{\mathrm{dist}(x,h)} = \frac{1}{\|w\|} \left| w^T x + b \right|}$$

Determine Thickness of a Separating Hyperplane

The thickness – margin $(\underline{r(h)})$



$$r(h) = \min_{n} \operatorname{dist}(x_n, h)$$

$$= \min_{n} \frac{|w^T x_n + b|}{\|w\|}$$

$$r(h) = \frac{1}{\|w\|} \min_{n} |w^T x_n + b|$$

$$\boxed{\min_n \; y_n(w^T x_n + b) = 1}$$

Separating hyperplane

$$\min_n \; |y_n(w^T x_n + b)| = 1$$

Substitute

$$\boxed{\min_n \; |w^T x_n + b| = 1}$$

largest.

$$\boxed{r(h) = \frac{1}{\|w\|}}$$

Thickness or Margin

$$\mathrm{dist}(x, h) = \frac{|w^T x + b|}{\|w\|}$$

$$\left.\begin{aligned} &\underline{\text{Maximize}} \quad \frac{1}{\|w\|} \\ &\text{s.t.} \ \min_n \ y_n(w^T x_n + b) = 1 \end{aligned}\right\}$$

$$\min \|w\| \Leftrightarrow \frac{1}{2}\|w\|^2 \Rightarrow \min \frac{1}{2} w^T w$$

$$\text{s.t.} \ \min_n \ y_n(w^T x_n + b) = 1$$

$$\Downarrow$$

$$y_n(w^T x_n + b) \geq 1$$

$$\Leftarrow$$

$$\boxed{\begin{aligned} &\min \frac{1}{2} w^T w \\ &\text{s.t.} \ y_n(w^T x_n + b) \geq 1 \end{aligned}}$$

i) Variables in optimization (w, b)

ii) Objective function is quadratic.

iii) Linear Inequality constraints

Quadratic Program (QP).

$$x = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad y = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix} \Longrightarrow w_1, w_2, b$$

$$\min \frac{1}{2}(w_1^2 + w_2^2)$$

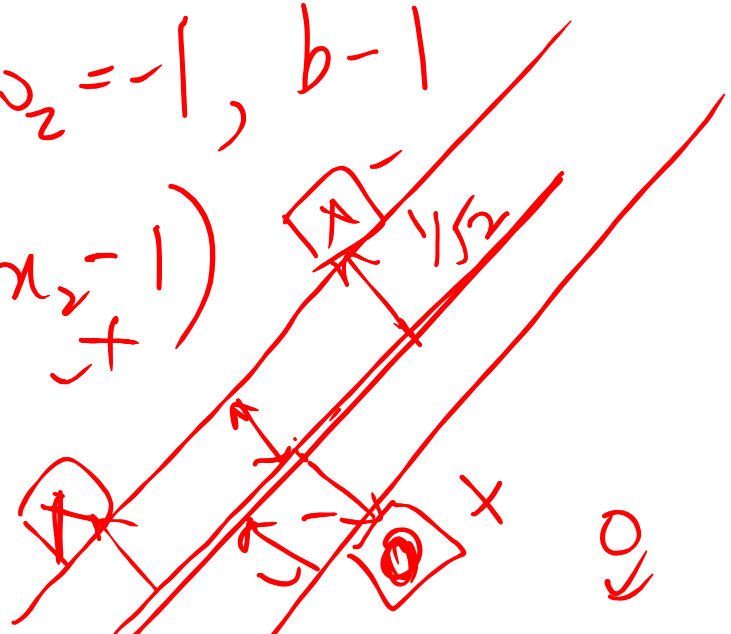
Support Vectors

$$\begin{cases} -b \geq 1 \\ -1(2w_1 + 2w_2 + b) \geq 1 \\ 1(2w_1 + b) \geq 1 \\ 1(3w_1 + b) \geq 1 \end{cases} \Longrightarrow w_1 \geq 1, \; w_2 \leq -1$$

$$w_1 = 1 \;\&\; w_2 = -1, \; b - 1$$

$$\therefore h_{thickest} = \text{sign}(x_1 - x_2 - 1)$$

$$\text{Thickness} = \frac{1}{\|w\|} = \frac{1}{\sqrt{2}}$$



Supporting.

Quadratic Program in standard form

$$\min_{u} \ \frac{1}{2} u^T Q u + p^T u$$

$$\text{s.t.} \quad Au \geqslant c$$

Q: is a $q \times q$ matrix

A: is a $N \times q$ matrix

p: is a vector $q \times 1$

c: is a vector $N \times 1$

$$u \in \mathbb{R}^q$$

$$\boxed{\text{QPAC}} \rightarrow QP \rightarrow u^*$$

find $\rightarrow$ b, w

$\rightarrow u = \begin{bmatrix} b \\ w \end{bmatrix} \in \mathbb{R}^{d+1}$, $\min \frac{1}{2} w^T w$

I_d: $d \times d$ identity matrix

0_d - d dimensional zero vector.

$$\underbrace{w^T w} = \begin{bmatrix} b & w^T \end{bmatrix} \begin{bmatrix} 0 & 0_d^T \\ 0_d & I_d \end{bmatrix} \begin{bmatrix} b \\ w^T \end{bmatrix}$$

$$= u^T \begin{bmatrix} 0 & 0_d^T \\ 0_d & I_d \end{bmatrix} u \rightarrow Q, \quad p = 0_{d+1}$$

✓ $C \leftarrow l_N$

A $\quad y_n (w^T x_n + b) \geq 1 \Rightarrow \begin{bmatrix} y_n & y_n x_n^T \end{bmatrix} u \geq 1$

$\rightarrow a_n^T = y_n \begin{bmatrix} 1 & x_n^T \end{bmatrix} \rightarrow c_n = 1$

We have identified QPAC $\rightarrow$ QP $\rightarrow$ $u^* = \begin{bmatrix} b^* \\ w^* \end{bmatrix}$

Linear Hard-Margin SVM with QP

1: Let $\mathbf{p} = \mathbf{0}_{d+1}$ (($d+1$)-dimensional zero vector) and $\mathbf{c} = \mathbf{1}_N$ (N-dimensional vector of ones). Construct matrices Q and A, where

$$\mathrm{Q} = \begin{bmatrix} 0 & \mathbf{0}_d^T \\ \mathbf{0}_d & \mathrm{I}_d \end{bmatrix}, \qquad \mathrm{A} = \underbrace{\begin{bmatrix} y_1 & -y_1\mathbf{x}_1^T- \\ \vdots & \vdots \\ y_N & -y_N\mathbf{x}_N^T- \end{bmatrix}}_{\text{signed data matrix}}.$$

2: Calculate $\begin{bmatrix} b^* \\ \mathbf{w}^* \end{bmatrix} = \mathbf{u}^* \leftarrow \mathrm{QP}(\mathrm{Q}, \mathbf{p}, \mathrm{A}, \mathbf{c})$.

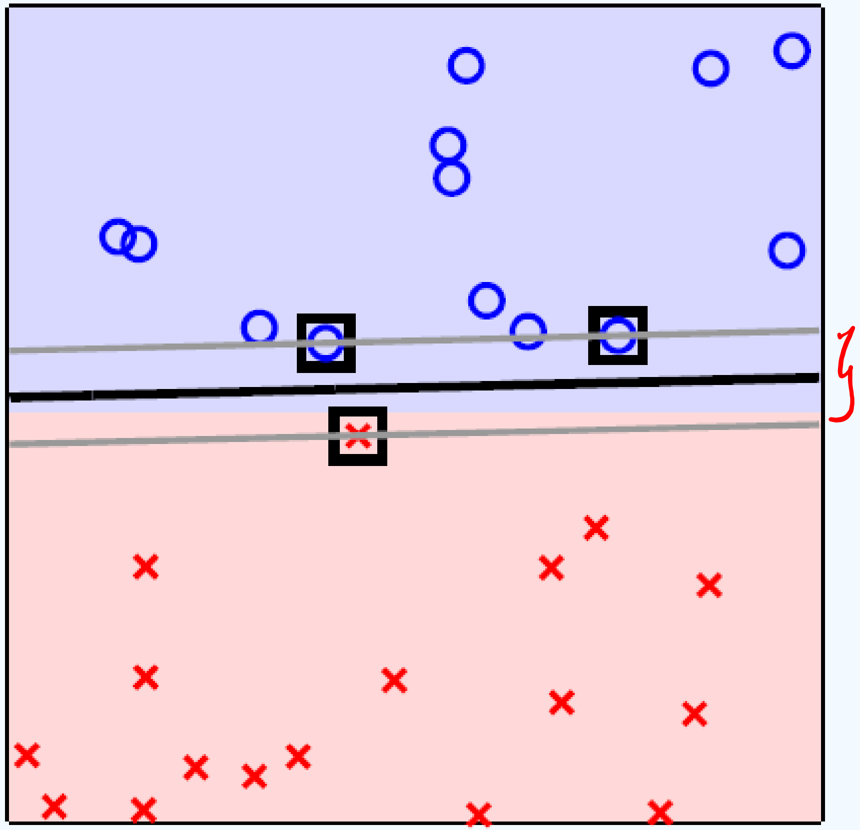
3: Return the hypothesis $g(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^{*T}\mathbf{x} + b^*)$.

Toy Example

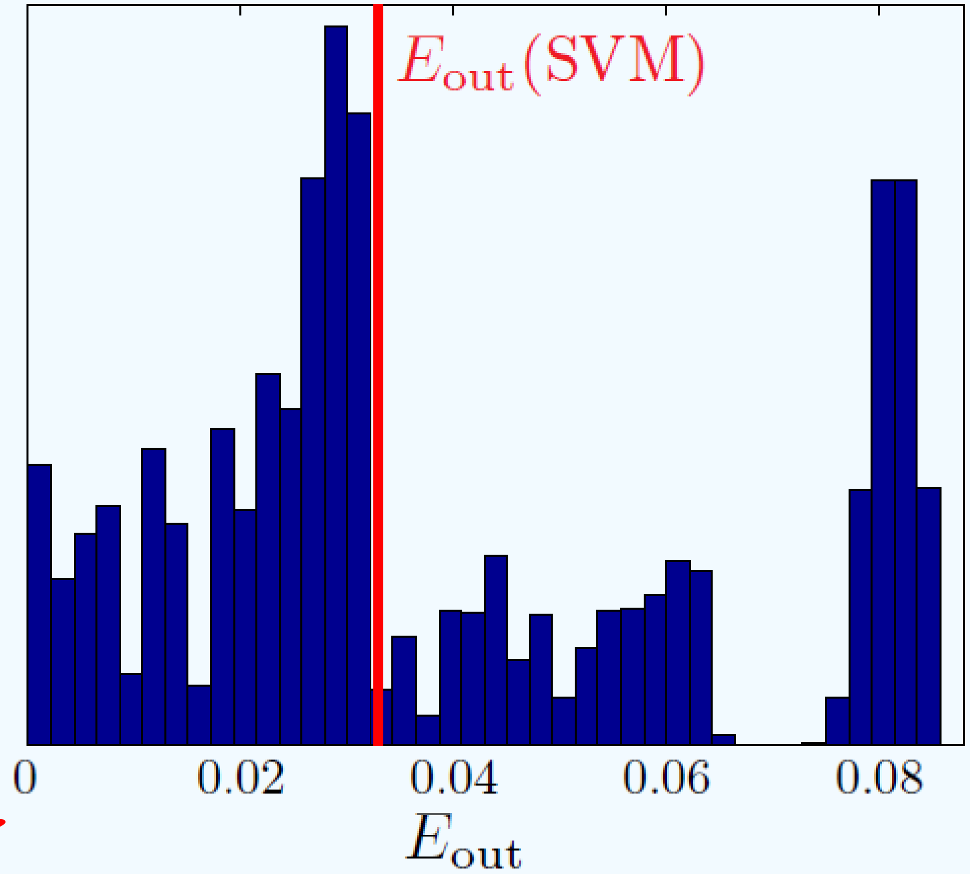
+1 -1

$$Q = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{p} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad A = \begin{bmatrix} -1 & 0 & 0 \\ -1 & -2 & -2 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \end{bmatrix} \quad \mathbf{c} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

A standard QP-solver gives $(b^*, w_1^*, w_2^*) = (-1, 1, -1)$, the same solution we computed manually, but obtained in less than a millisecond. $\square$



(a) Data and SVM separator



(b) Histogram of E_{out}(PLA)

	optimal hyperplane	regularization
minimize:	$\mathbf{w}^{\mathsf{T}}\mathbf{w}$	E_{in}
subject to:	$E_{\text{in}} = 0$	$\mathbf{w}^{\mathsf{T}}\mathbf{w} \le C$

Thanks!