Support Vector Machines

Somak Aditya Sudeshna Sarkar
CSE Department, IIT Kharagpur
Sep 1, 2023

These slides were assembled by Byron Boots, with only minor modifications from Eric Eaton's slides and grateful acknowledgement to the many others who made their course materials freely available online. Feel free to reuse or adapt these slides for your own academic purposes, provided that you include proper attribution.

Adapted from slides by Tim Oates

Strengths of SVMs

- Good generalization
  - in theory
  - in practice
- Works well with few training instances
- Find globally best model
- Efficient algorithms
- Amenable to the kernel trick

Notational Conventions

To better match notation used in SVMs

...and to make matrix formulas simpler

We will use the following for the i th instance

i th instance → $\mathbf{x}_i$

Bold denotes vector

i th instance label → y_i

Non-bold denotes scalar

j th feature of i th instance → x_{ij}

Linear Separators

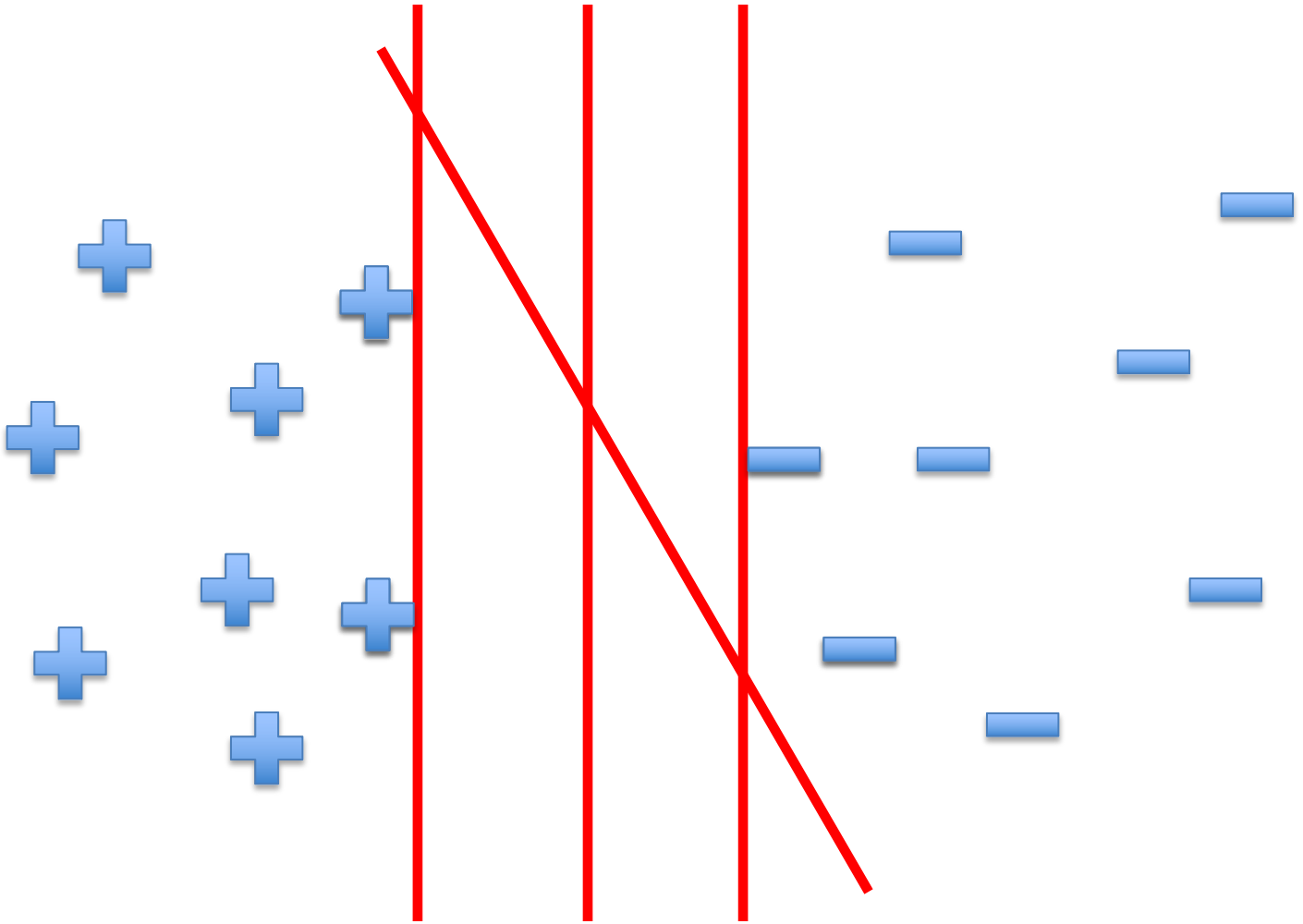
- Training instances
  $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{d+1}}, \boldsymbol{x_0} = \mathbf{1}$
  $y \in \{-1, 1\}$
- Model parameters
  $\theta \in \mathbb{R}^{d+1}$
- Hyperplane
  $\theta^\top x = \langle \theta, x \rangle = 0$
- Decision function
  $h(x) = sign(\theta^\top x)$
  $= sign(\langle \theta, x \rangle)$

Recall:

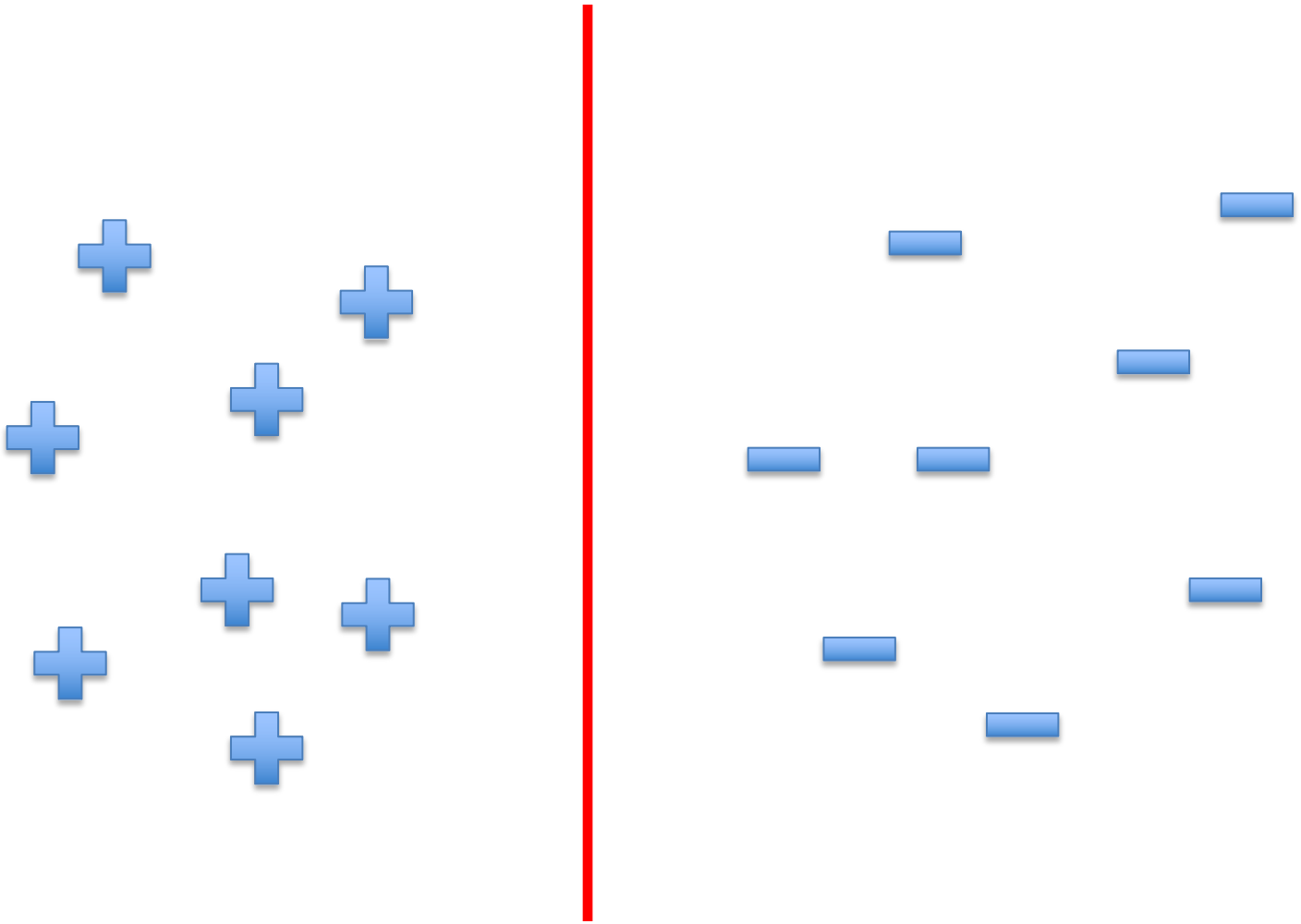
Inner (dot) product:

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u} \cdot \boldsymbol{v}$$
$$= \boldsymbol{u}^\top v$$
$$= \sum_i u_i v_i$$

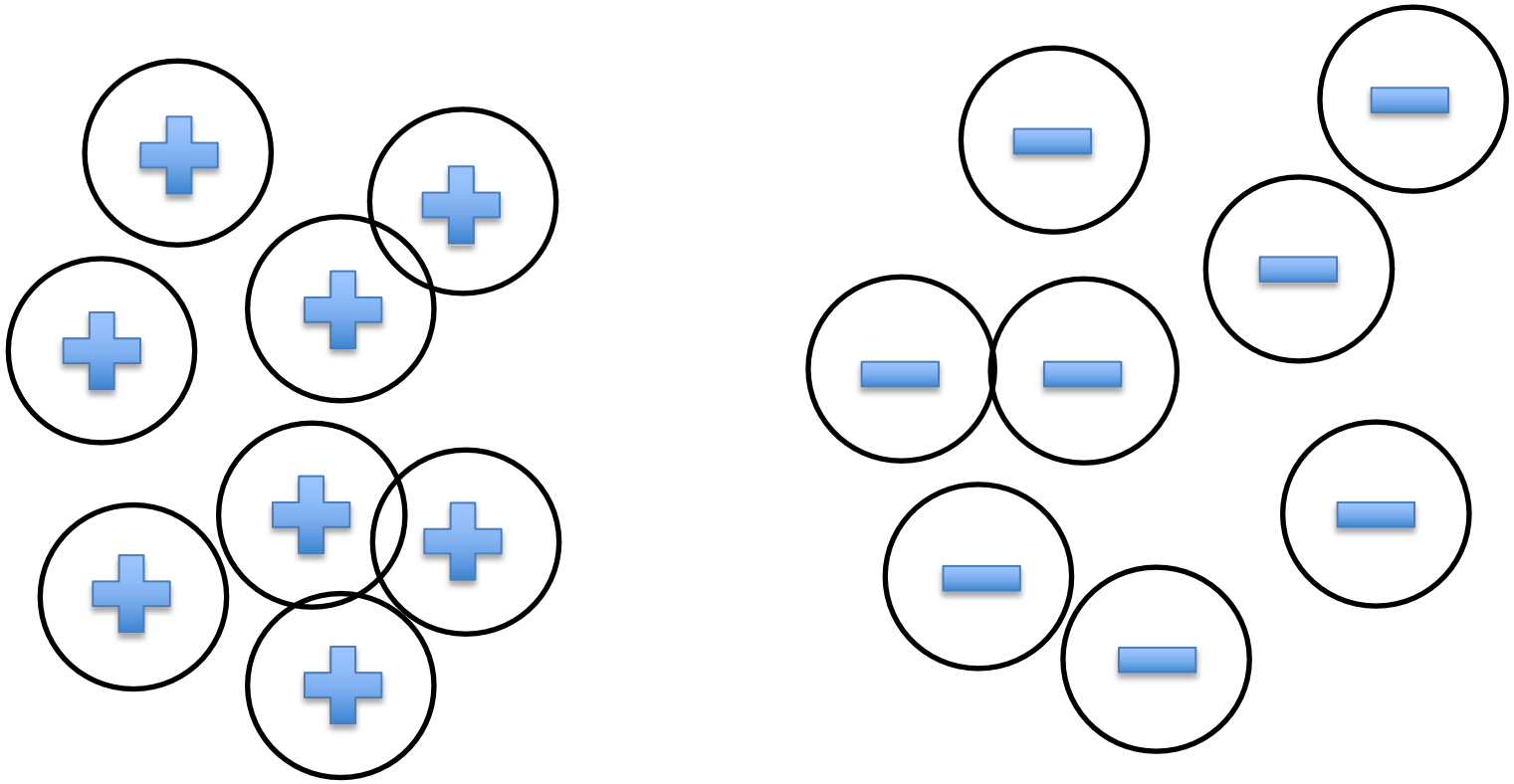
Intuitions



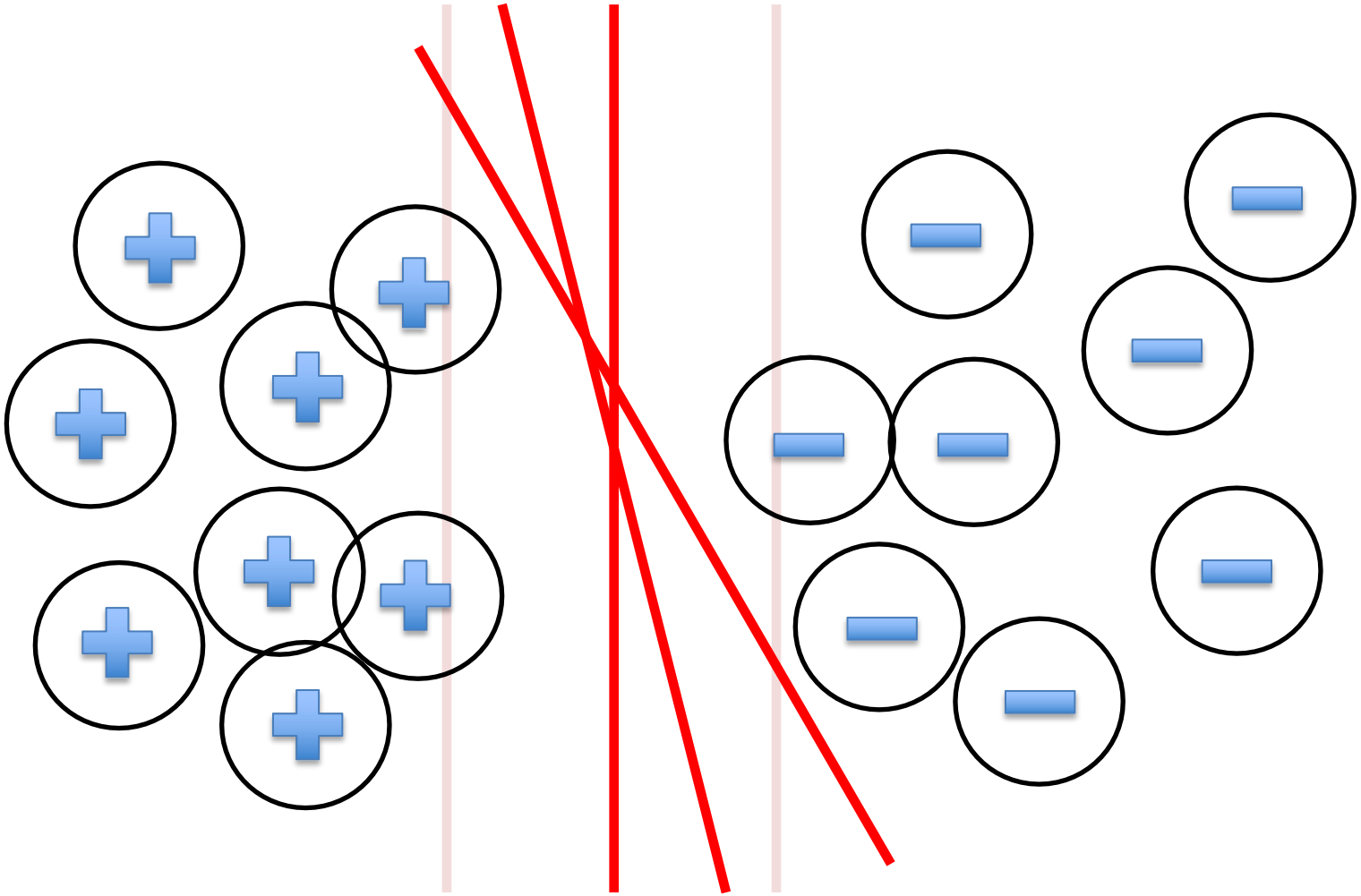
A ″Good″ Separator



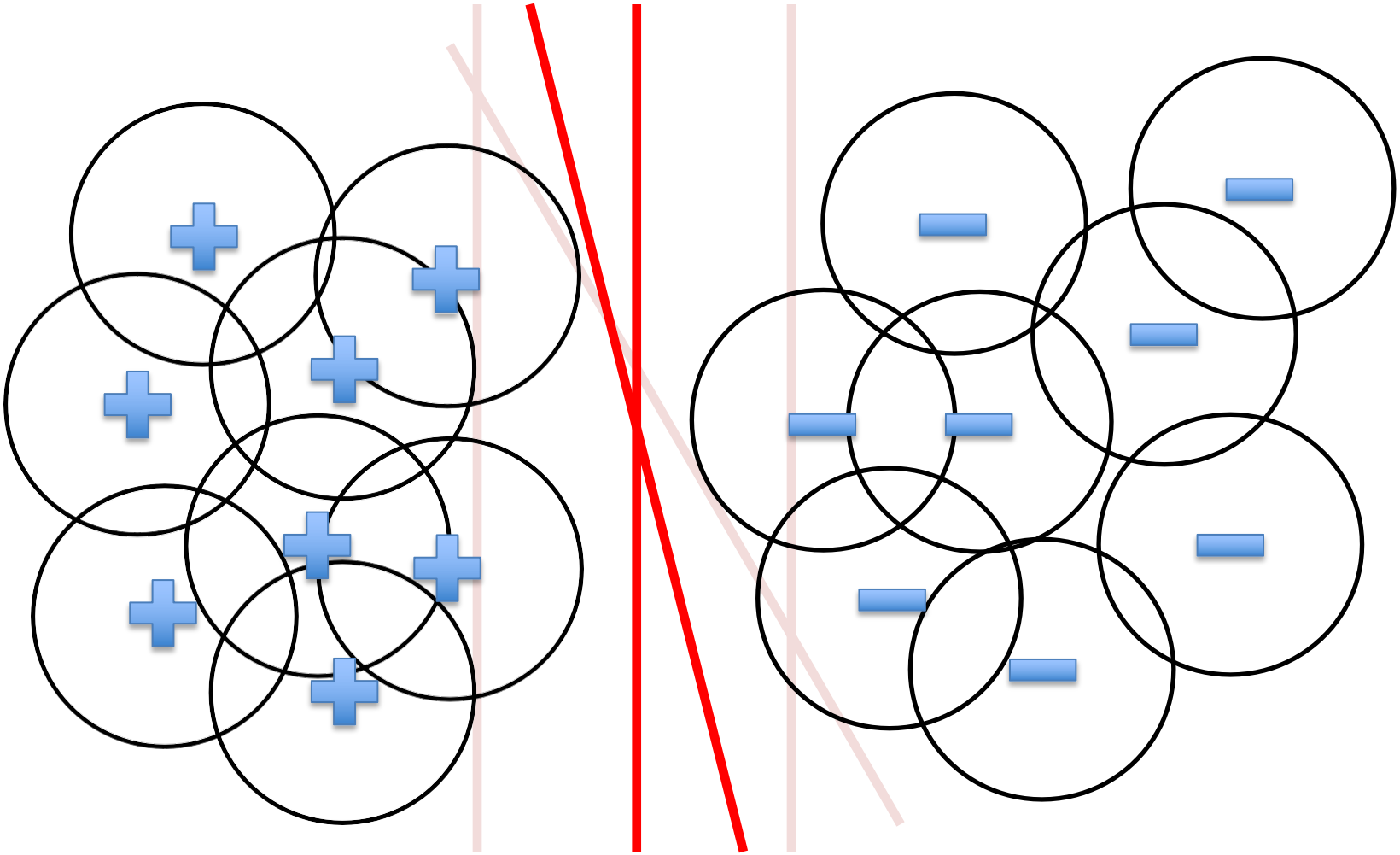
Noise in the Observations



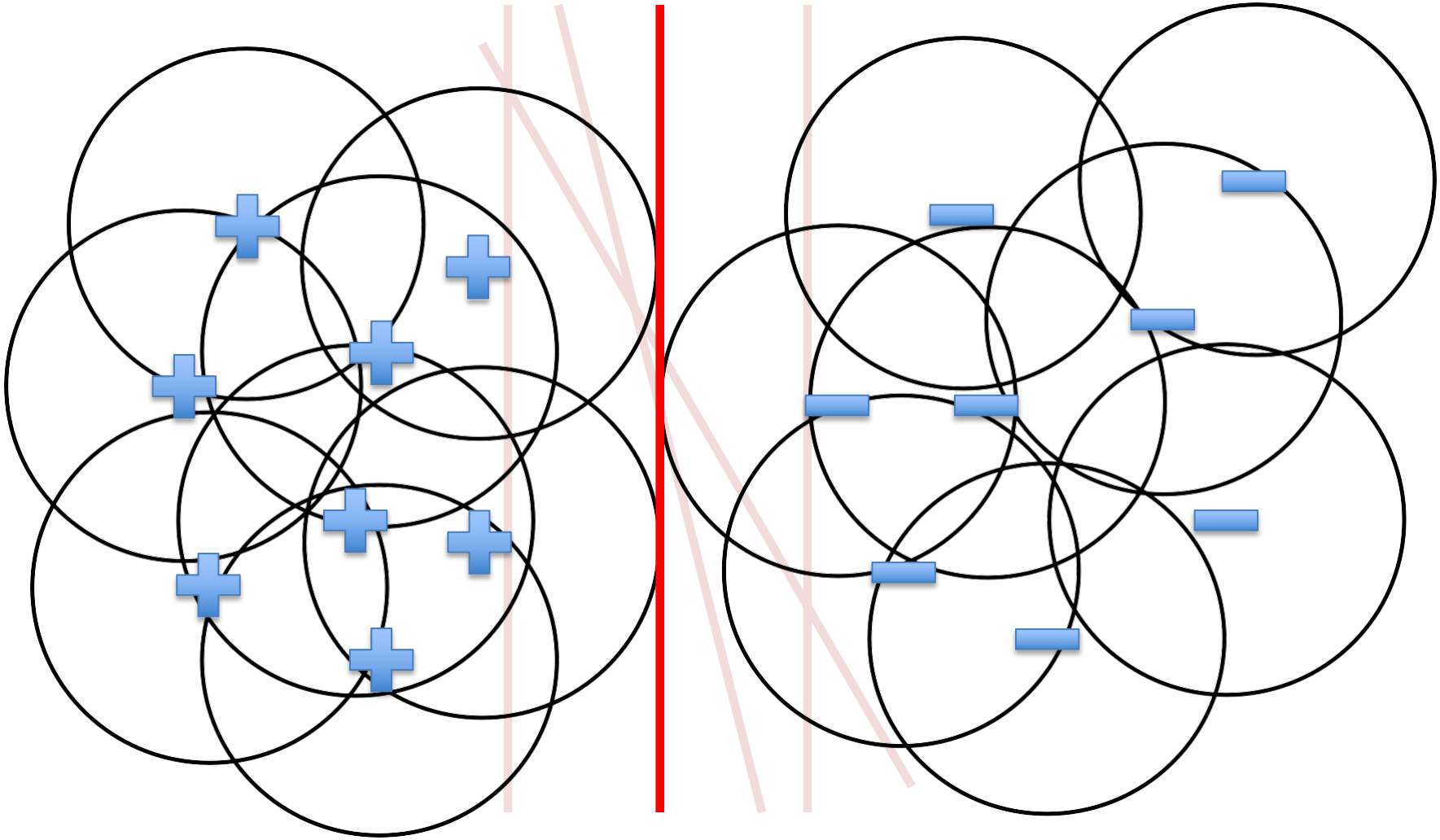
Ruling Out Some Separators



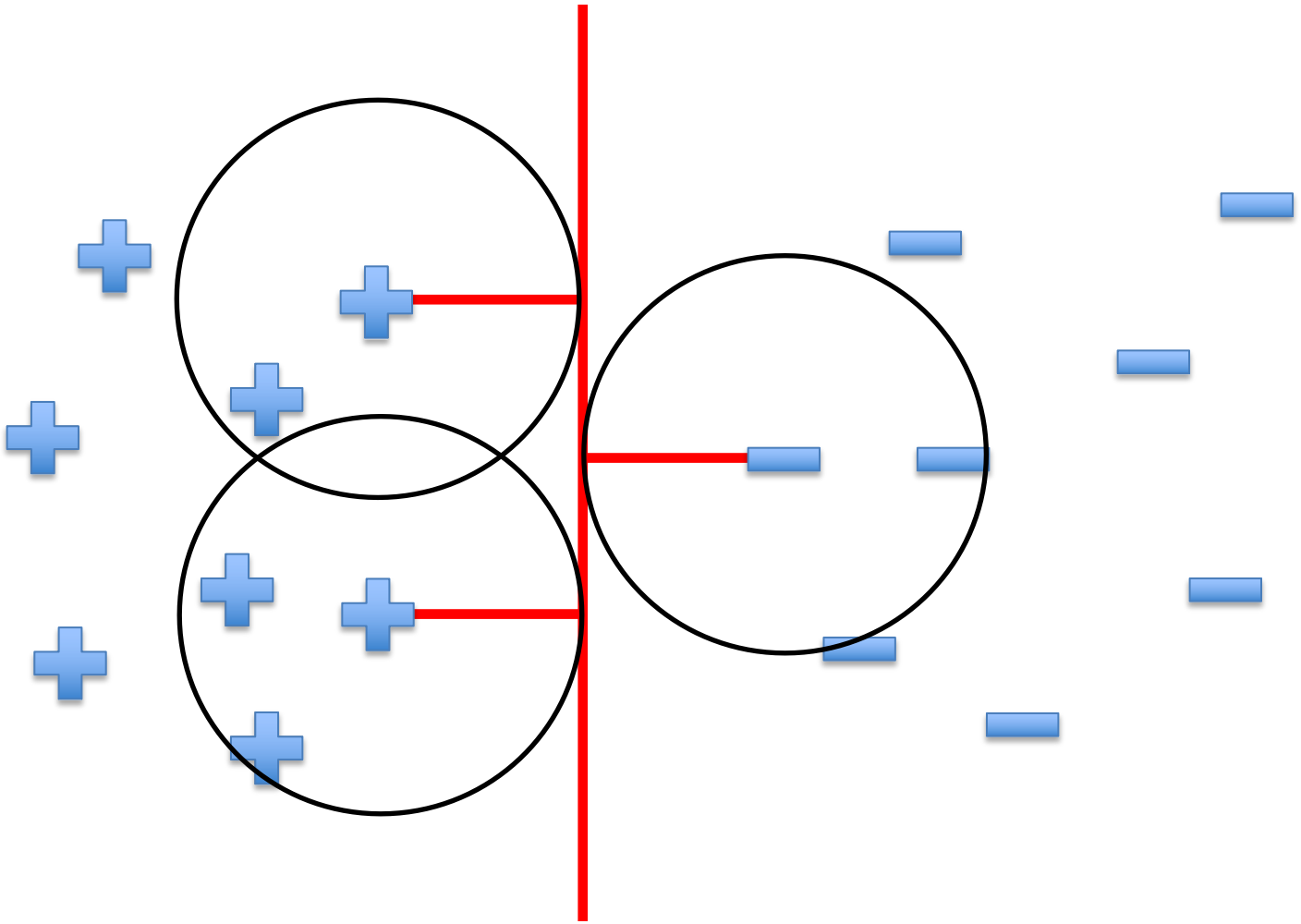
Lots of Noise



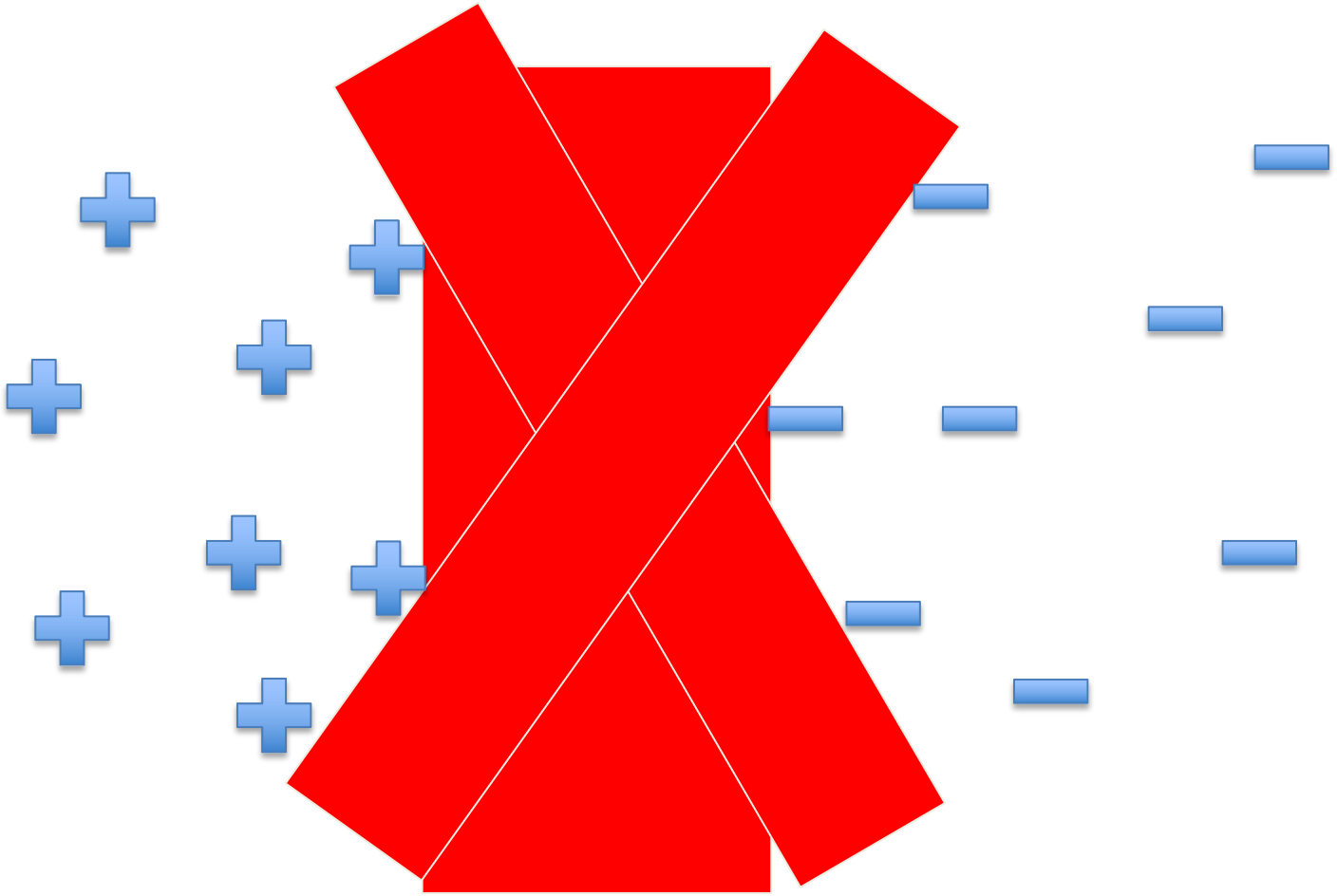
Only One Separator Remains



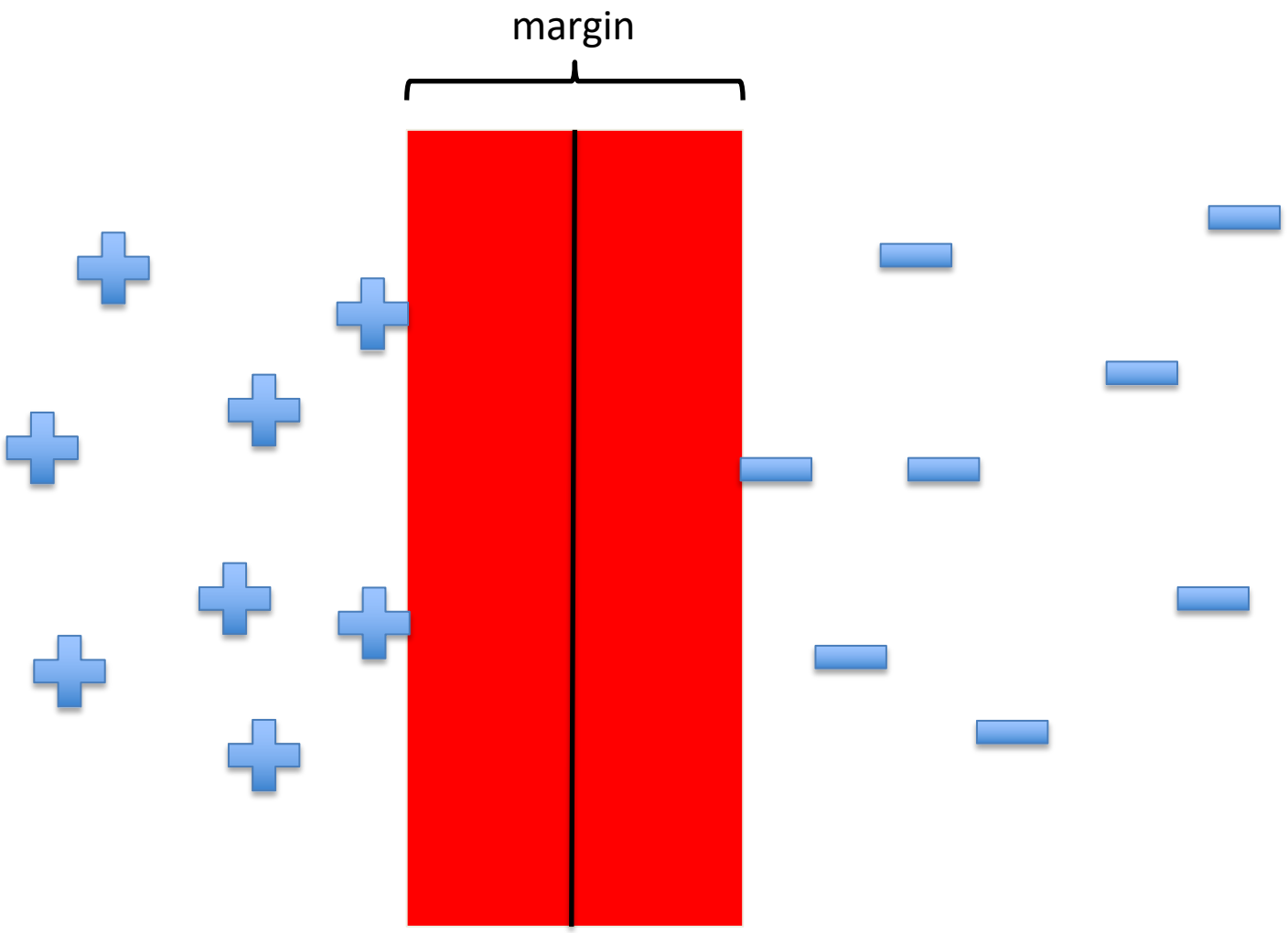
Maximizing the Margin



″Fat″ Separators



″Fat″ Separators



Why Maximize Margin

Increasing margin reduces *capacity*

- i.e., fewer possible models
- *What about bias? Variance?*

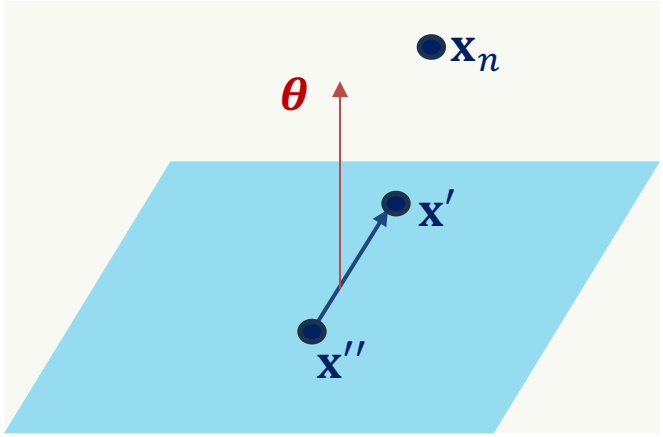
Lesson from Learning Theory:

- If the following holds:
  - H is sufficiently constrained in size
  - and/or the size of the training data set n is large,

  then low training error is likely to be evidence of low generalization error

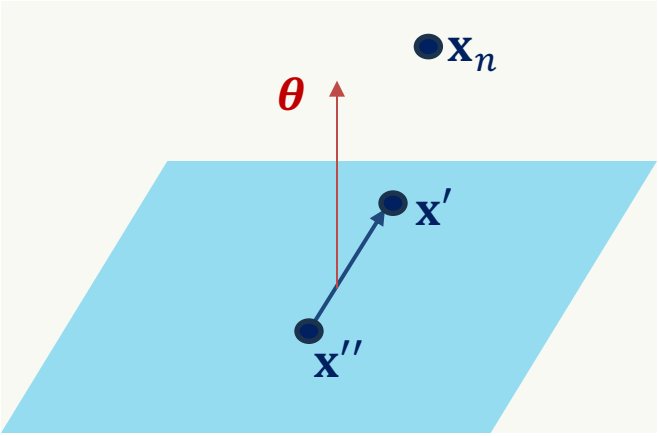
Computing the margin

Margin = The distance between $\mathbf{x}_n$ and the plane $\theta^{\mathsf{T}}\mathbf{x} = 0$, where $|\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}_n| = 1$



Computing the margin

Proposition:
The vector $\boldsymbol{\theta}$ is orthogonal to the plane in the *X* space



Take any two points $\mathbf{x}'$ and $\mathbf{x}''$ on the plane.

$\theta^{\mathsf{T}}\mathbf{x}' = 0$ and $\theta^{\mathsf{T}}\mathbf{x}'' = 0$
=> $\theta^{\mathsf{T}}(\mathbf{x}' - \mathbf{x}'') = 0$

Hence $\boldsymbol{\theta}$ is orthogonal to any vector that lies on the plane => $\boldsymbol{\theta}$ is orthogonal to the plane

Margin: distance between x_n and the plane

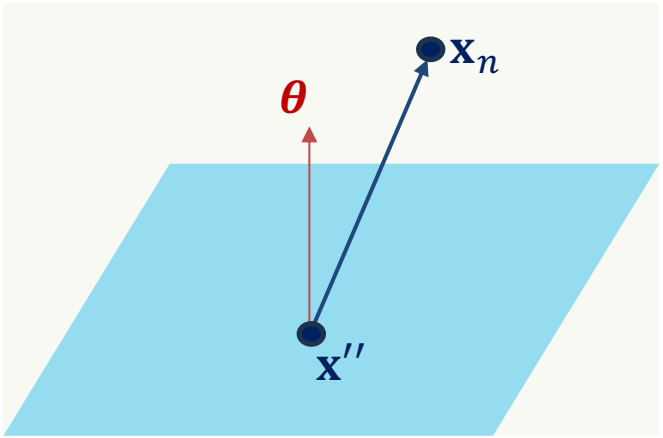
Take any point $\mathbf{x}$ on the plane

Projection of $\mathbf{x}_n - \mathbf{x}$ on $\boldsymbol{\theta}$ (direction orthogonal to the plane)

$$\widehat{\boldsymbol{\theta}} = \frac{\boldsymbol{\theta}}{||\boldsymbol{\theta}||} \Rightarrow \textbf{distance} = |\,\widehat{\boldsymbol{\theta}}^{\mathsf{T}}(\mathbf{x}_n - \mathbf{x})|$$

Projection of the vector $(\mathbf{x}_n - \mathbf{x})$ along $\boldsymbol{\theta}$
computed by taking the vector product of $(\mathbf{x}_n - \mathbf{x})$ with the unit vector in the direction of $\boldsymbol{\theta}$

$||\boldsymbol{\theta}||$ is the norm of $\boldsymbol{\theta}$



Margin: distance between x_n and the plane

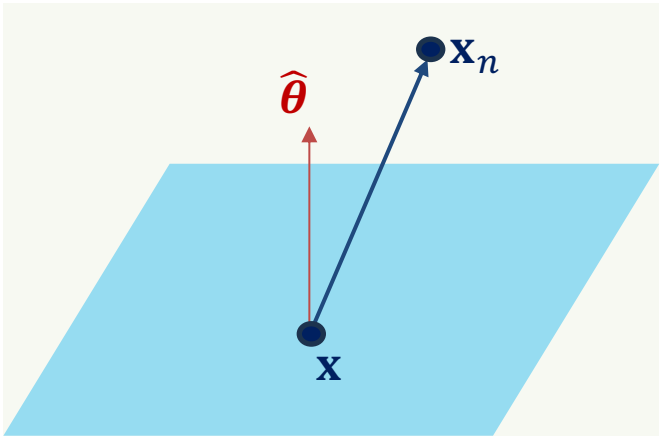
$$\text{distance} = \frac{1}{||\boldsymbol{\theta}||}|\boldsymbol{\theta}^\top \mathbf{x}_n - \boldsymbol{\theta}^\top \mathbf{x}|$$

$$\frac{1}{||\boldsymbol{\theta}||}|\boldsymbol{\theta}^\top \mathbf{x}_n - \boldsymbol{\theta}^\top \mathbf{x}| = \frac{1}{||\boldsymbol{\theta}||}$$

$\mathbf{x}$ is a point on the plane.

Hence $\Rightarrow \boldsymbol{\theta}^\top \mathbf{x} = 0$

$|\boldsymbol{\theta}^\top \mathbf{x}_n| = 1$ for the nearest point $\mathbf{x}_n$ (due to our normalization)



The optimization problem

$$\text{Maximize } \frac{1}{||\boldsymbol{\theta}||}$$
$$\text{subject to } \min_{n=1,2,\ldots,N} \left|\left|\boldsymbol{\theta}^T \mathbf{x}_n\right|\right| = 1$$

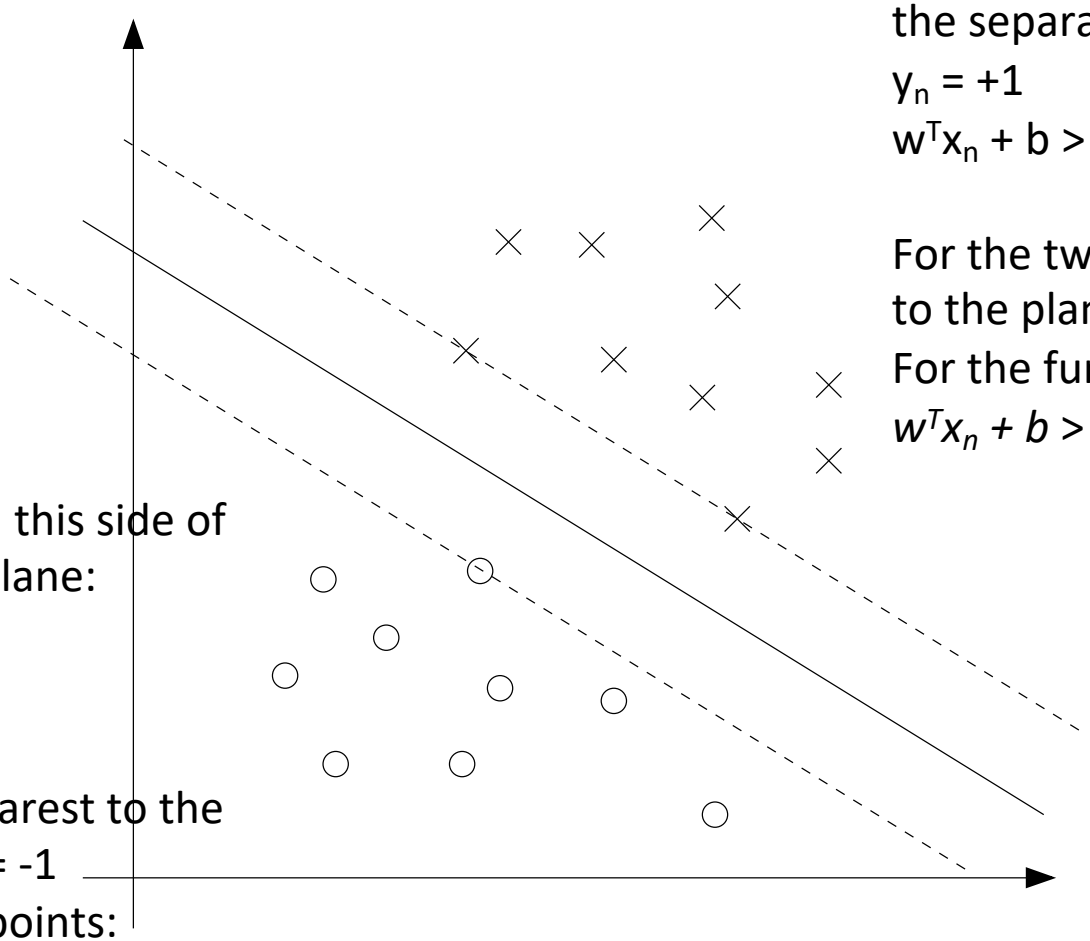
This optimization problem is too complex, because of

(i) the norm in the objective function, and

(ii) the minimum term in the constraints

Can we find an equivalent optimization problem that is easier to tackle?

The geometry



For any point on this side of the separating plane:
$y_n = +1$
$w^T x_n + b > 0$

For the two points nearest to the plane: $w^T x_n + b = 1$
For the further points:
$w^T x_n + b > 1$

For any point on this side of the separating plane:
$y_n = -1$
$w^T x_n + b < 0$

For the point nearest to the plane: $w^T x_n + b = -1$
For the further points:
$w^T x_n + b < -1$

Notice: $|\boldsymbol{\theta}^\top \mathbf{x}_n| = y_n(\boldsymbol{\theta}^\top \mathbf{x}_n)$

Equivalent optimization problem

Maximize $\frac{1}{||\boldsymbol{\theta}||}$

subject to $\min_{n=1,2,\ldots,N} |\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}_n| = 1$

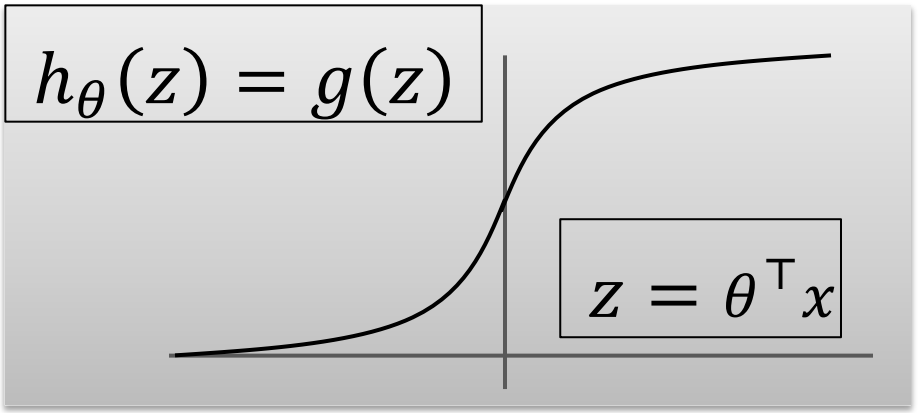
Notice: $|\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}_n| = y_n(\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}_n)$

Minimize $\frac{1}{2}\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\theta}$

subject to $y_n(\boldsymbol{\theta}^{\mathsf{T}}x_n) \geq 1$, for $n = 1,2, \ldots N$

Alternative View of Logistic Regression

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^\top x}}$$



If $y = 1$, we want $h_\theta(x) \approx 1, \theta^\top x \gg 0$

If $y = 1$, we want $h_\theta(x) \approx 0, \theta^\top x \ll 0$

$$J(\theta) = -\sum_{i=1}^{n} [y_i \underbrace{\log h_\theta(x_i)}_{cost_1(\theta^\top x_i)} + (1 - y_i) \underbrace{\log(1 - h_\theta(x_i))}_{cost_0(\theta^\top x_i)}]$$

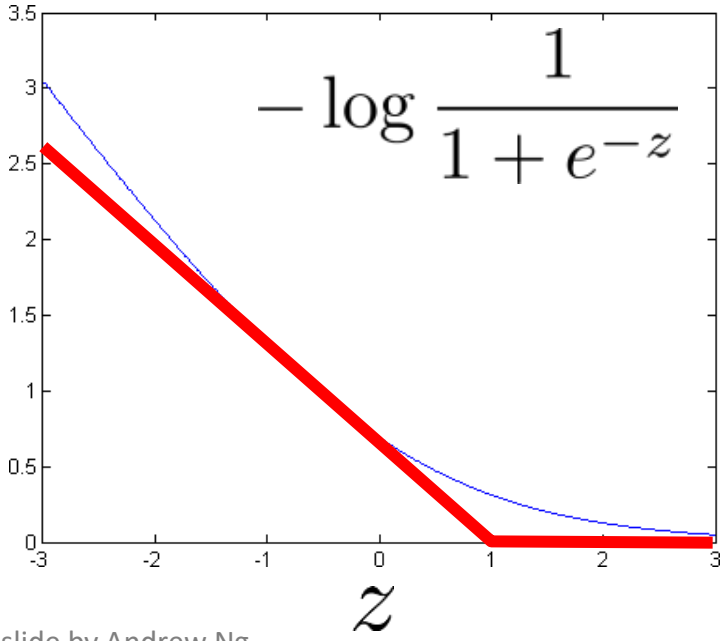
$$\min_\theta J(\theta)$$

Alternative View of Logistic Regression

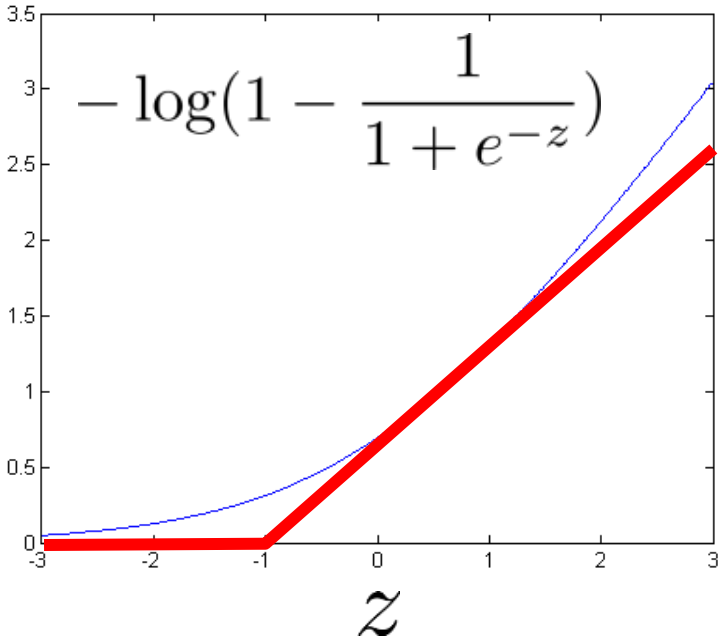
Cost of example: $-y_i \log h_{\boldsymbol{\theta}}(\mathbf{x}_i) - (1 - y_i) \log (1 - h_{\boldsymbol{\theta}}(\mathbf{x}_i))$

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}}} \qquad z = \boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}$$

If $y = 1$ (want $\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x} \gg 0$):

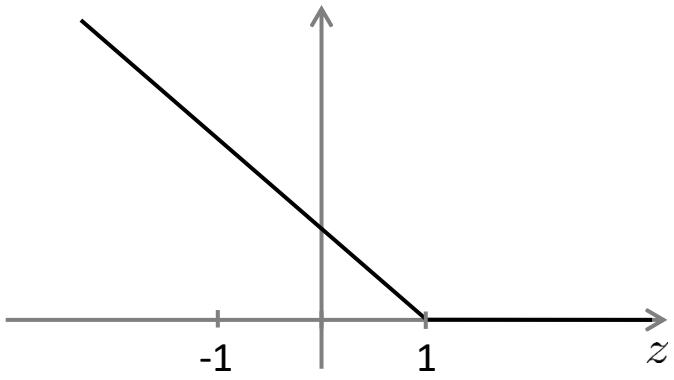


If $y = 0$ (want $\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x} \ll 0$):

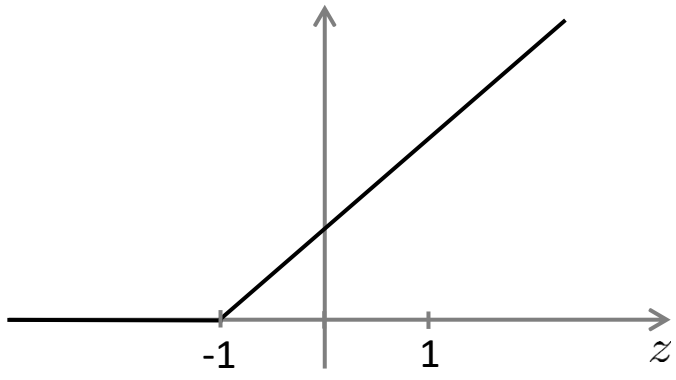


Support Vector Machine

If $y = 1$ (want $\theta^T x \geq 1$)



If $y = -1$ (want $\theta^T x \leq -1$)



$$l_{hinge}(h(\boldsymbol{x})) = \max(0, 1 - y \cdot h(\boldsymbol{x}))$$

Based on slide by Andrew Ng

Support Vector Machine

$$\min_{\theta} C \sum_{i=1}^{n} [y_i cost_1(\theta^\top x_i) + (1 - y_i) cost_0(\theta^\top x_i)] + \frac{1}{2} \sum_{j=1}^{d} \theta_j^2$$

y = 1 / 0

y = +1 / -1

with C = 1

$$\min_{\theta} \frac{1}{2} \sum_{j=1}^{d} \theta_j^2$$

$$\text{s.t. } \boldsymbol{\theta}^\top \mathbf{x}_i \geq 1 \quad \text{if } y_i = 1$$

$$\boldsymbol{\theta}^\top \mathbf{x}_i \leq -1 \quad \text{if } y_i = -1$$

$$\min_{\theta} \frac{1}{2} \sum_{j=1}^{d} \theta_j^2$$

$$\text{s.t. } y_i(\boldsymbol{\theta}^\top \mathbf{x}_i) \geq 1$$