CS M146: Introduction to Machine Learning
Support Vector Machines

Aditya Grover

https://aditya-grover.github.io/ @adityagrover_

Recap: Perceptrons

- Training instances

$$\mathbf{x} \in \mathbb{R}^{d+1}, x_0 = 1$$

$$y \in \{-1, 1\}$$

- Model parameters

$$\boldsymbol{\theta} \in \mathbb{R}^{d+1}$$

- Hyperplane separator

$$\boldsymbol{\theta}^\intercal \mathbf{x} = \langle \boldsymbol{\theta}, \mathbf{x} \rangle = 0$$

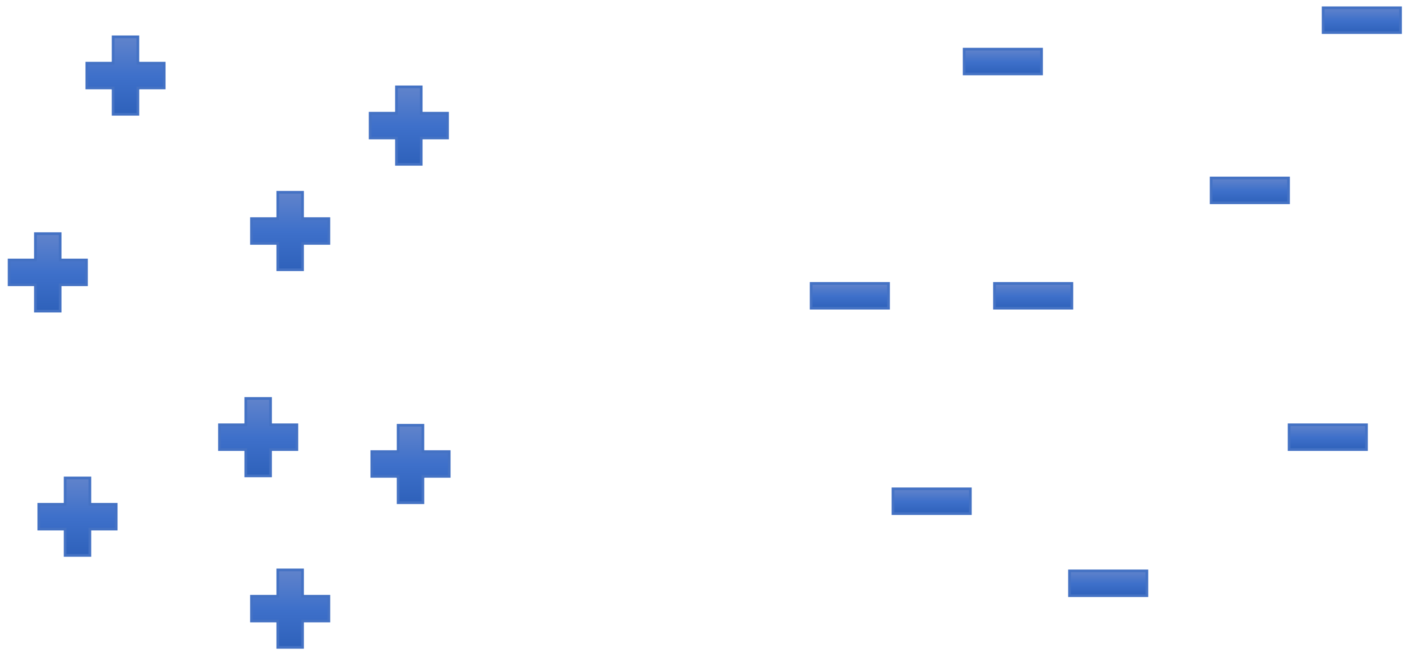
- Decision function

$$h(\mathbf{x}) = \mathrm{sign}(\boldsymbol{\theta}^\intercal \mathbf{x}) = \mathrm{sign}(\langle \boldsymbol{\theta}, \mathbf{x} \rangle)$$

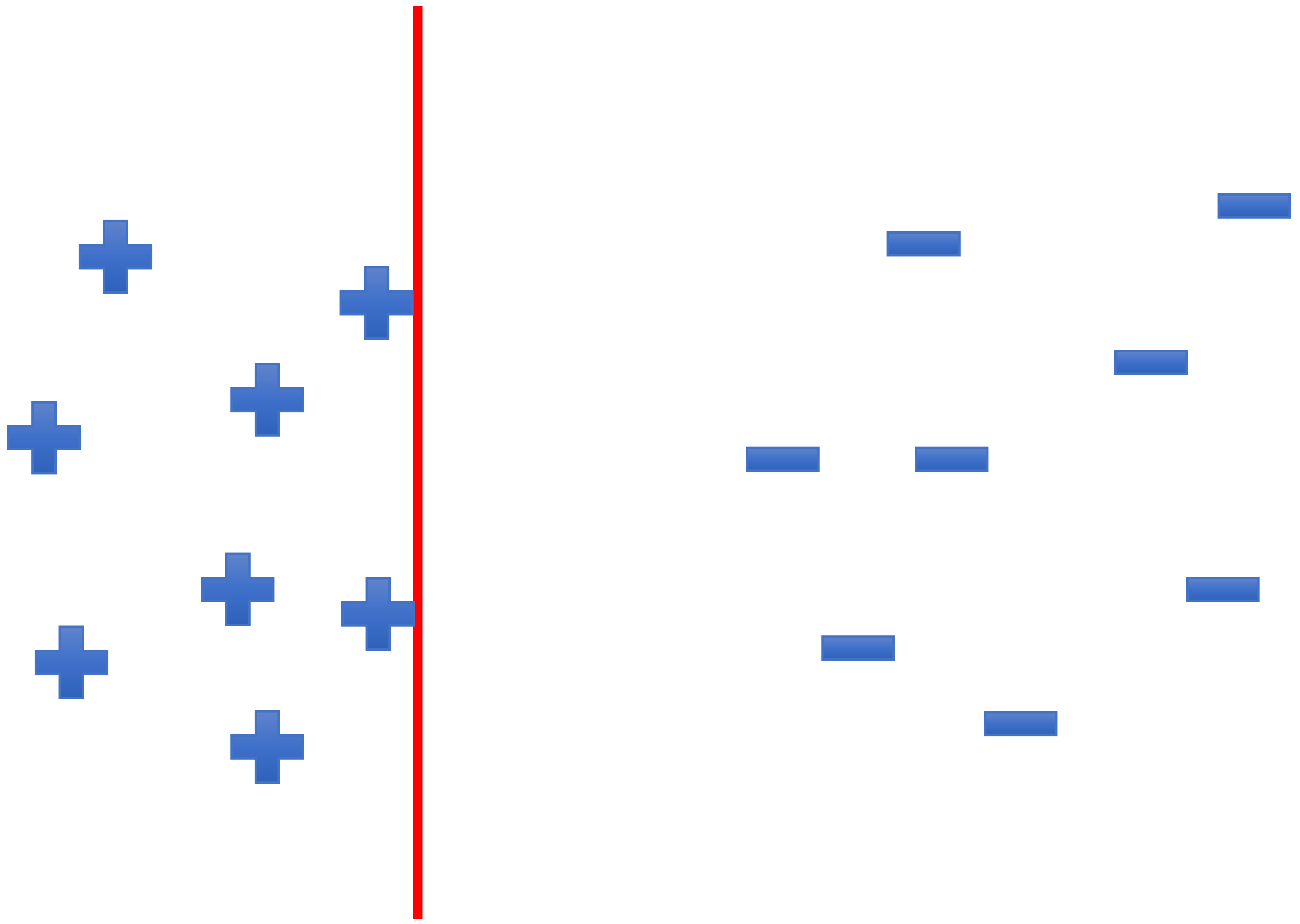
<u>Recall:</u>
Inner (dot) product:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u} \cdot \mathbf{v} = \mathbf{u}^\intercal \mathbf{v} = \sum_i u_i v_i$$

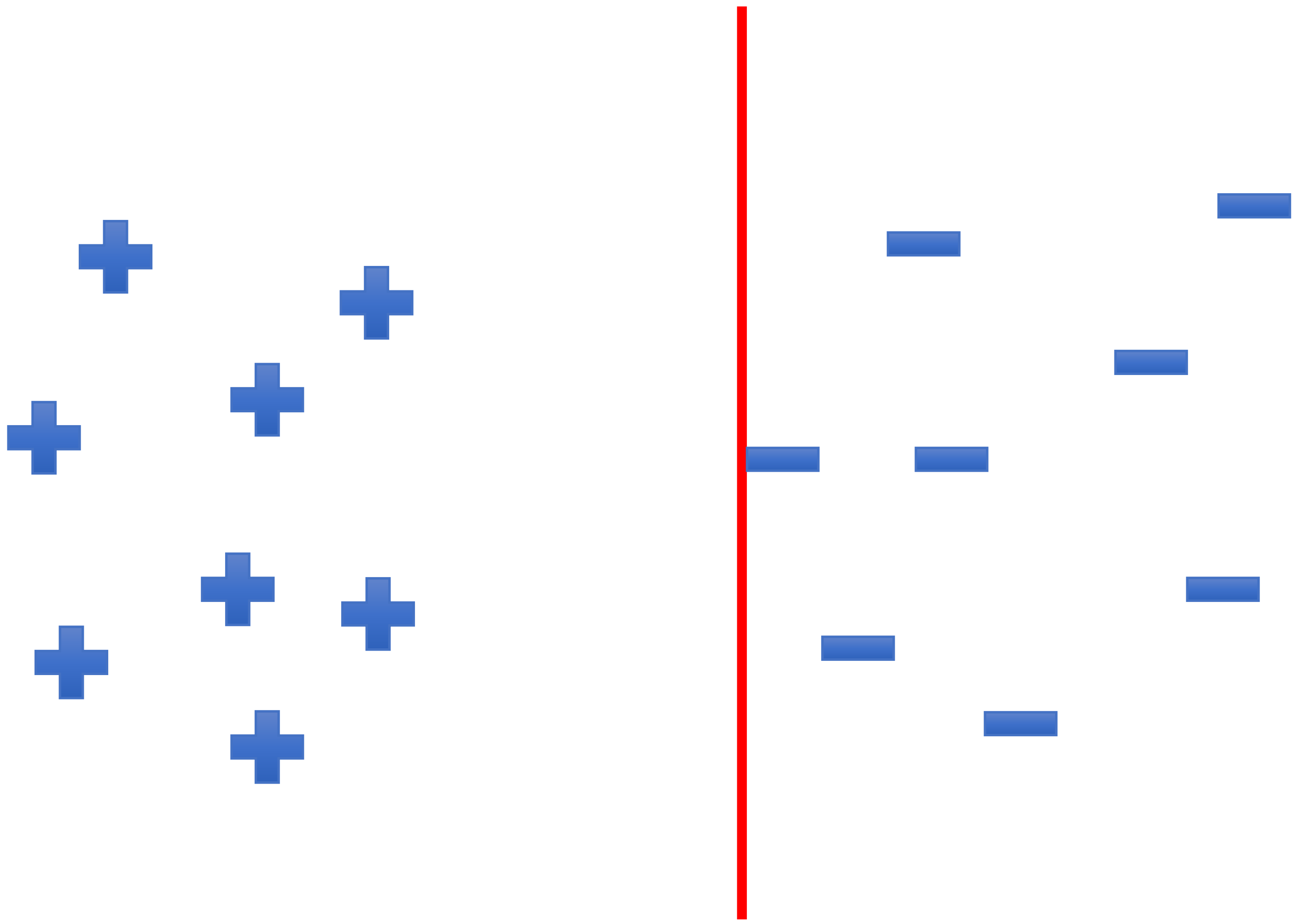
Which Separator to Pick?



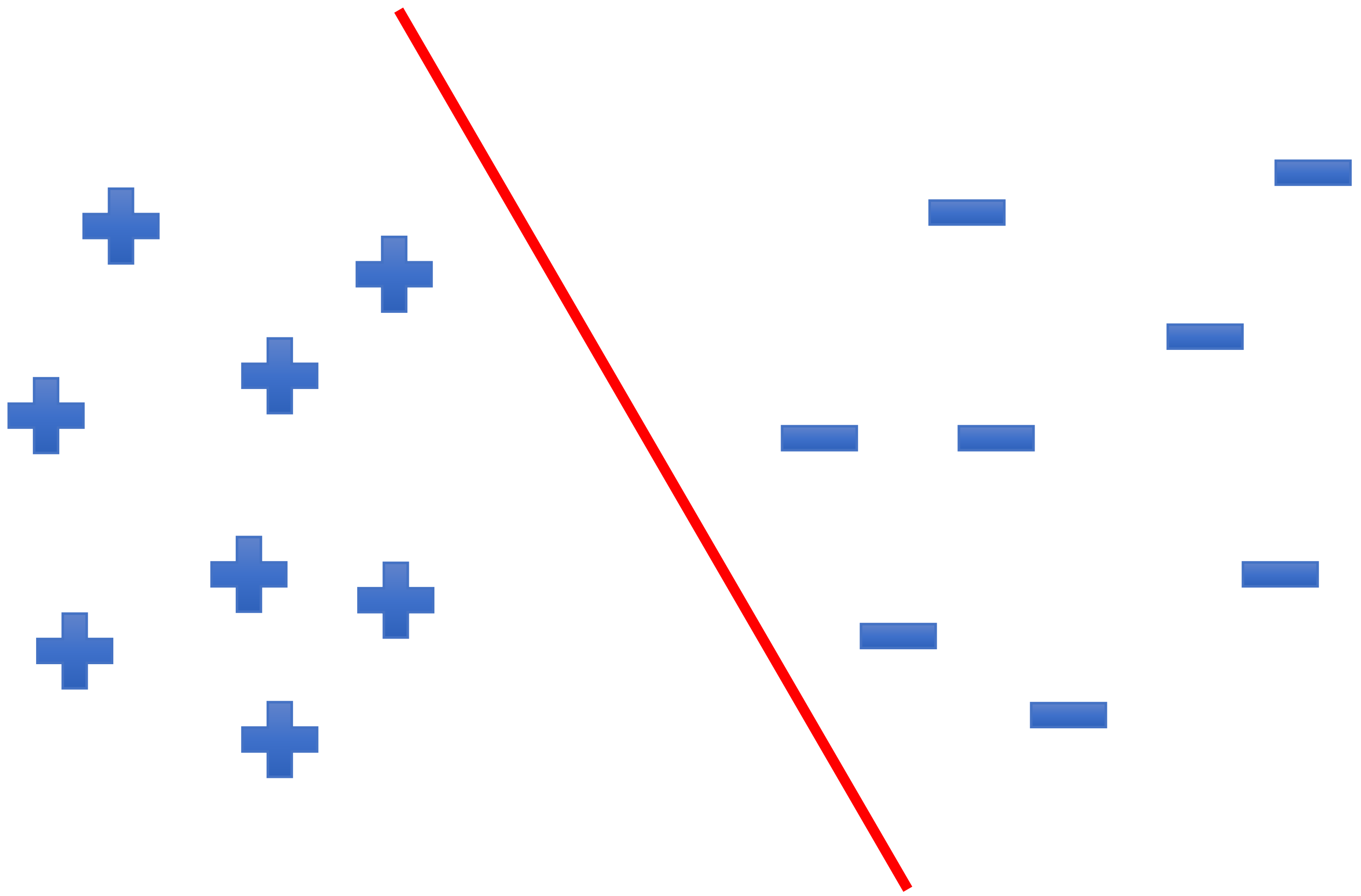
Which Separator to Pick?



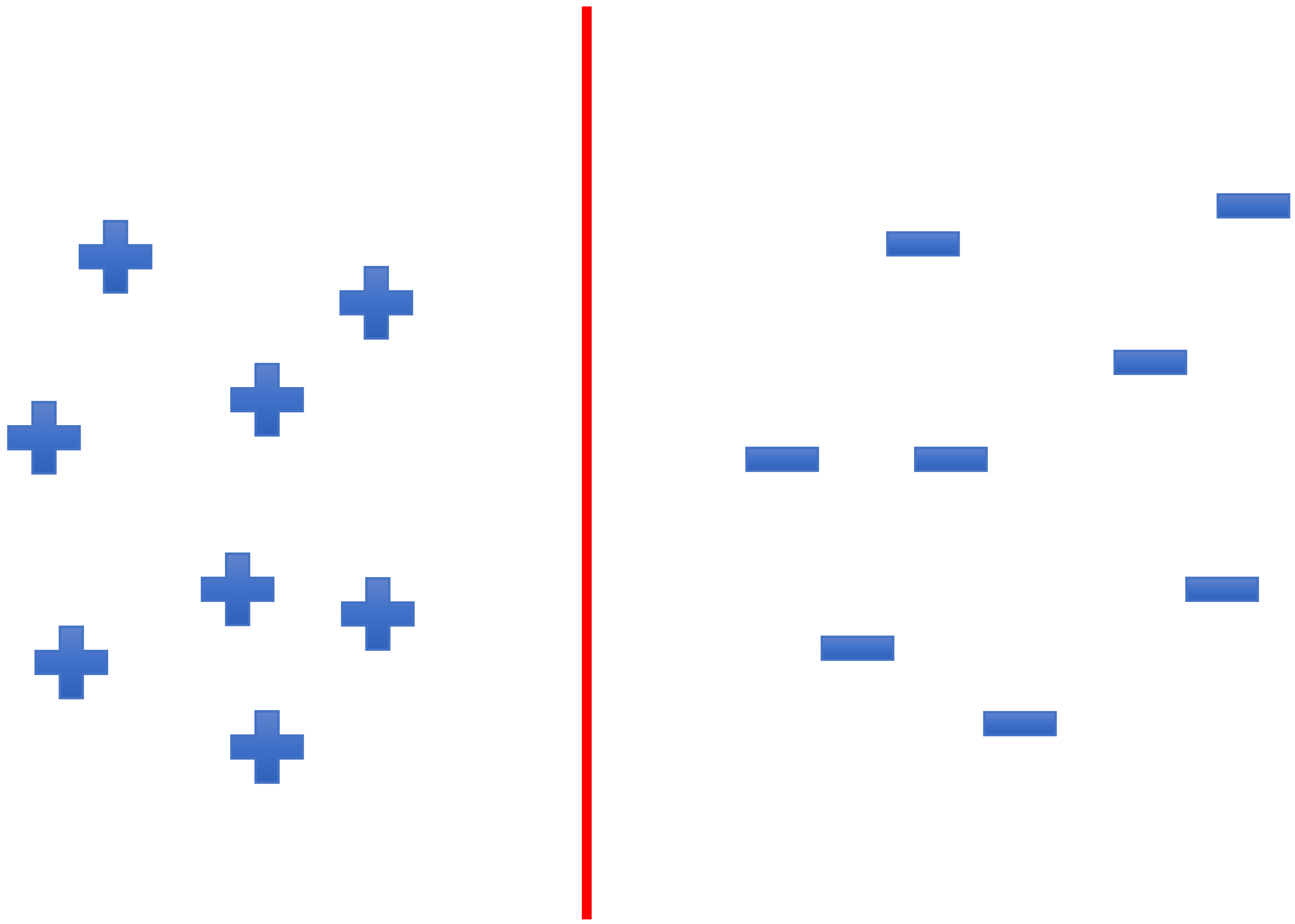
Which Separator to Pick?



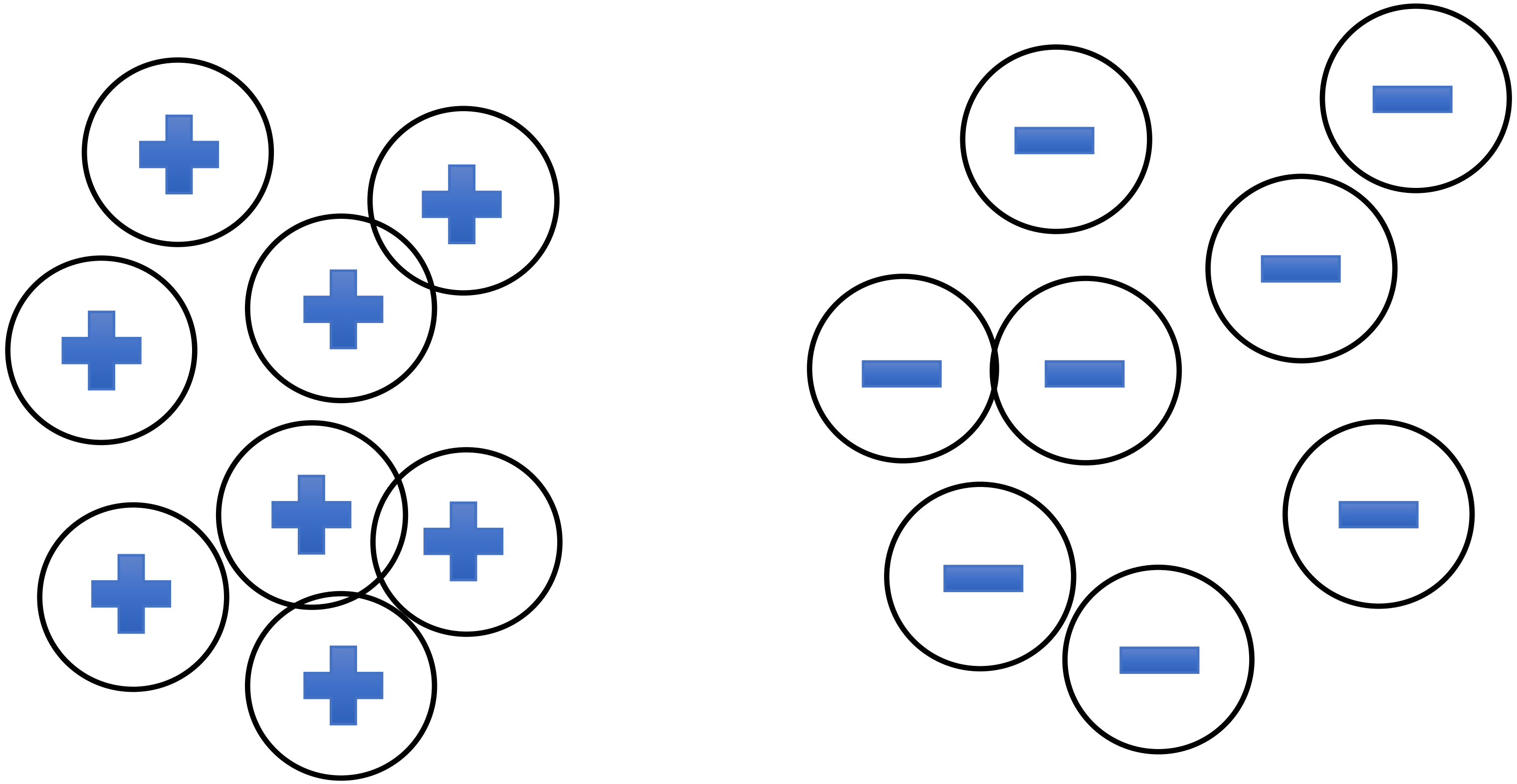
Which Separator to Pick?



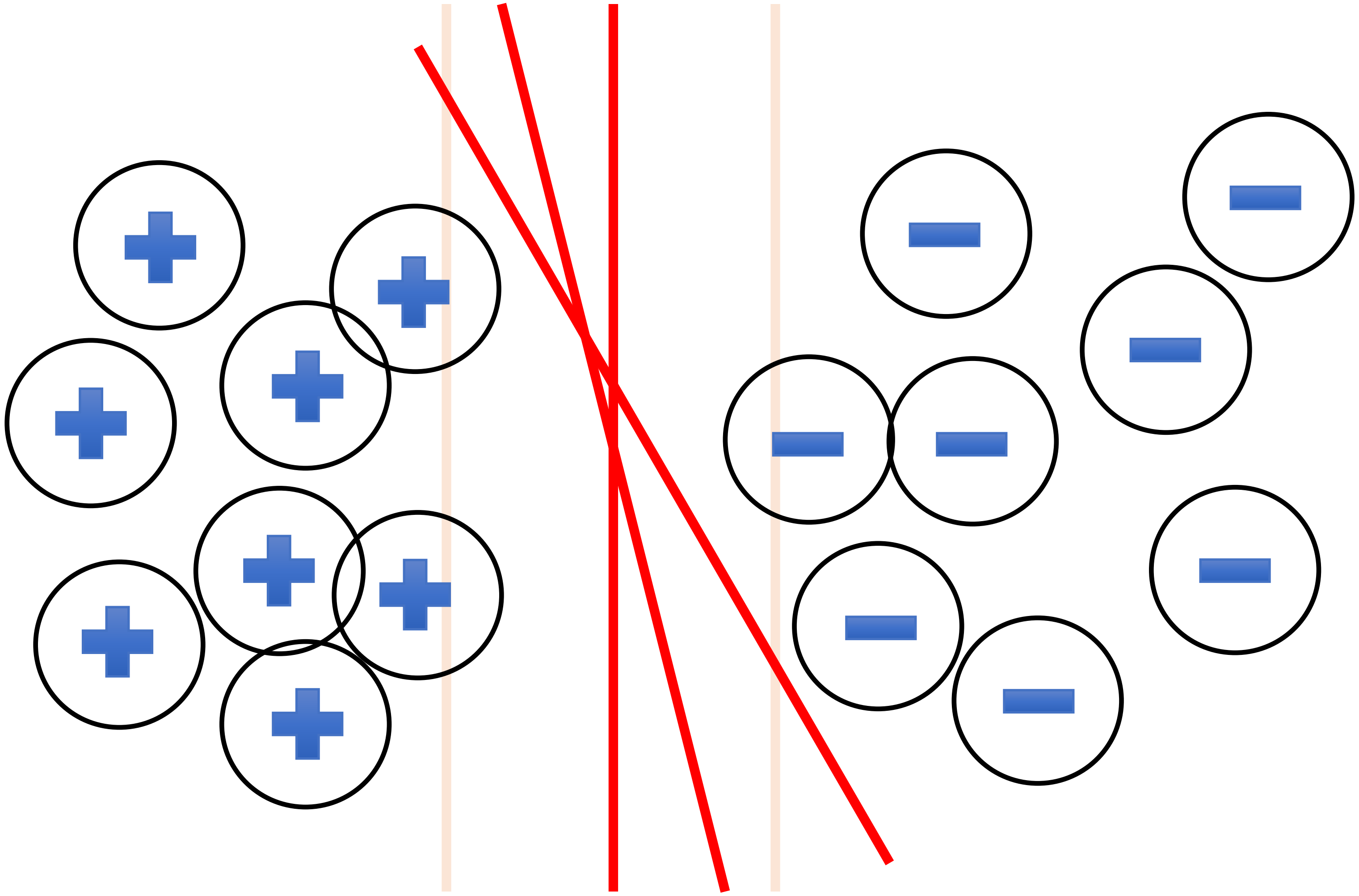
A “Good” Separator



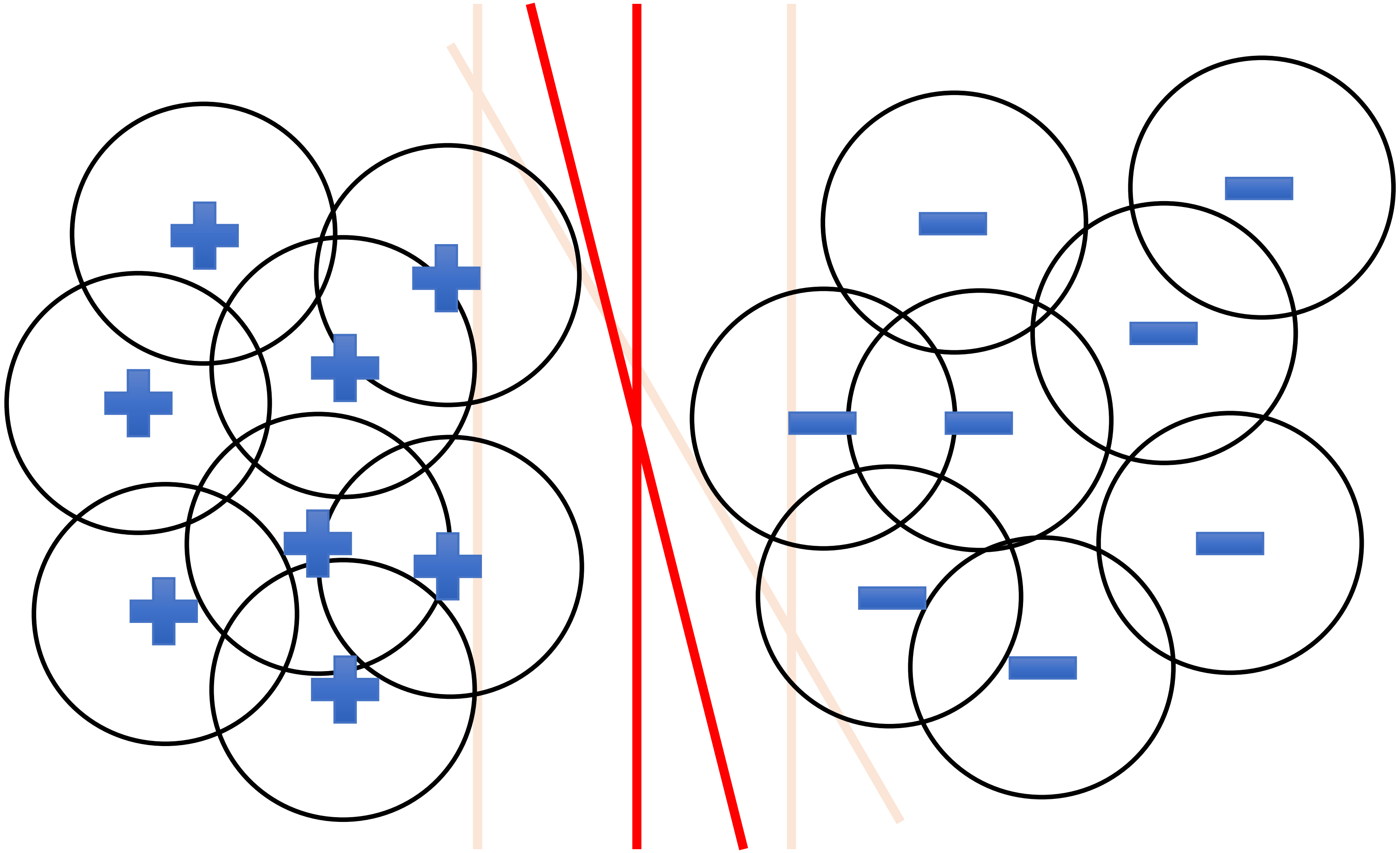
Noise in the Observations



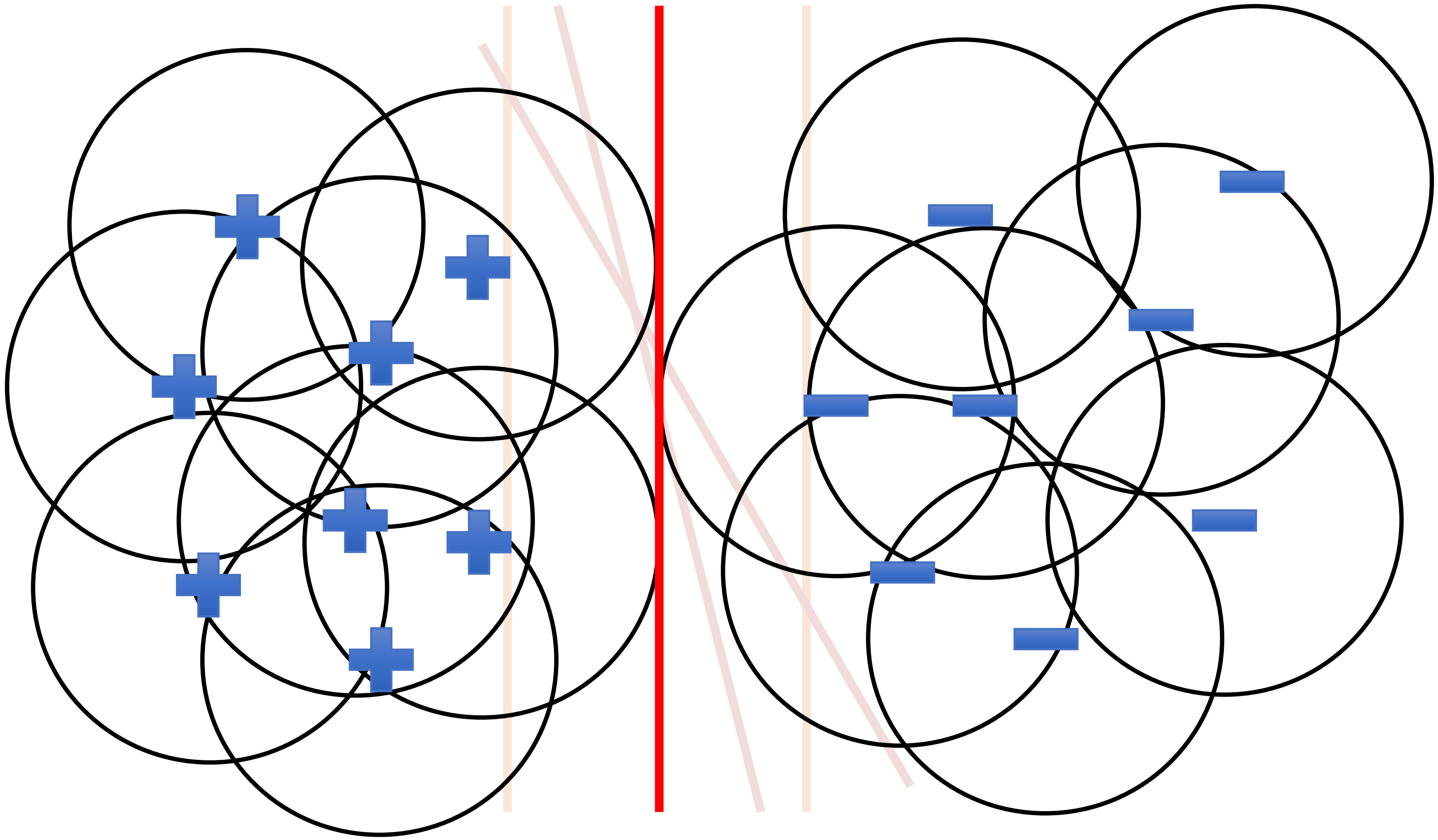
Ruling Out Non-Robust Separators



Lots of Noise



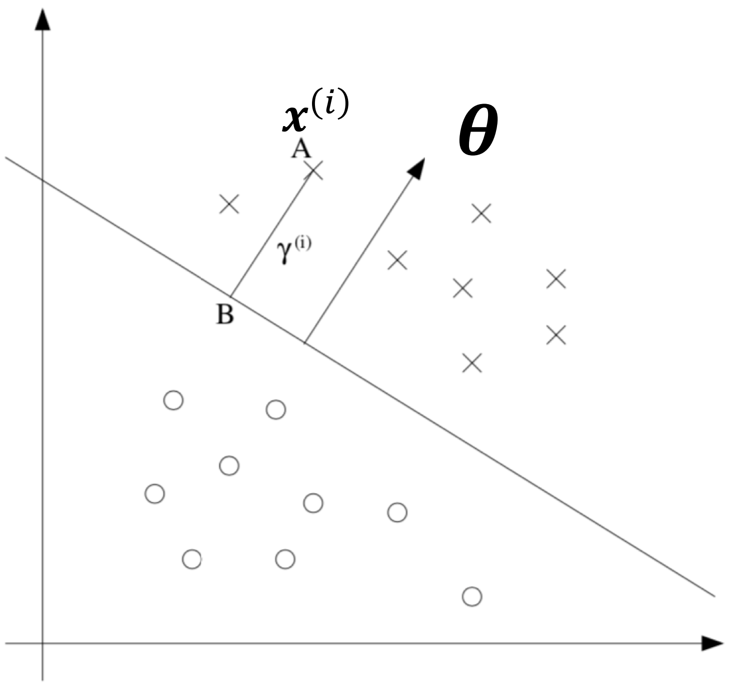
Only One Robust Separator Remains



Margin of a Linear Separator

- Consider binary classification with $\{1, -1\}$ labels and a hypothesis class of linear separators
- Further, we assume data is linearly separable (relaxed later)
- **Definition:** The margin of a point $\boldsymbol{x}^{(i)}$ w.r.t. a hyperplane is the perpendicular distance between $\boldsymbol{x}^{(i)}$ and the hyperplane

$$\gamma^{(i)} = \text{length(AB)}$$



Computing Margin

$$\gamma^{(i)} = \text{length(AB)}$$

- Assume that $\boldsymbol{\theta}$ is a perfect linear separator for the data with no bias (i.e., $\boldsymbol{x}^{(i)}$ does **not** have a dummy dimension $x_0^{(i)} = 1$)
- Without loss of generality, we also assume that $\boldsymbol{x}^{(i)}$ (denoted as A) has a **positive label** $y^{(i)} = 1$. Hence, $\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} > 0$
- From geometry, we know that $\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2}$ is a unit normal vector to the hyperplane. Hence,

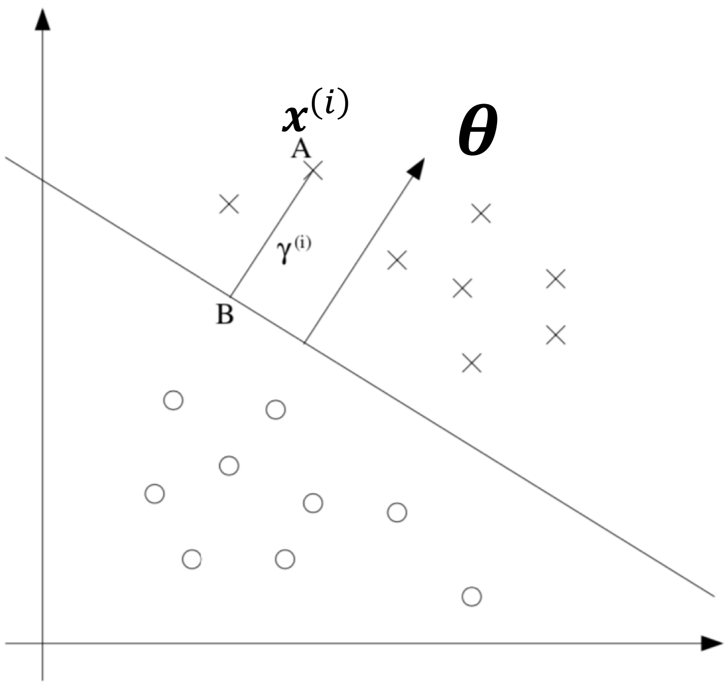
$$\text{B} = \boldsymbol{x}^{(i)} - \gamma^{(i)} \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2}$$

- Since B lies on the hyperplane, it satisfies

$$\boldsymbol{\theta}^T \boldsymbol{x} = 0$$

- Hence, $\boldsymbol{\theta}^T \left(\boldsymbol{x}^{(i)} - \gamma^{(i)} \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|} \right) = 0$ which gives us

$$\gamma^{(i)} = \frac{\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}{\|\boldsymbol{\theta}\|_2}$$



Computing Margin

$$\gamma^{(i)} = \text{length(AB)}$$

- If we assume that $\boldsymbol{x}^{(i)}$ (denoted as A) has a **negative label** $y^{(i)} = -1$, then $\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} < 0$
- From geometry, we know that $\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2}$ is a unit normal vector to the hyperplane. Hence,

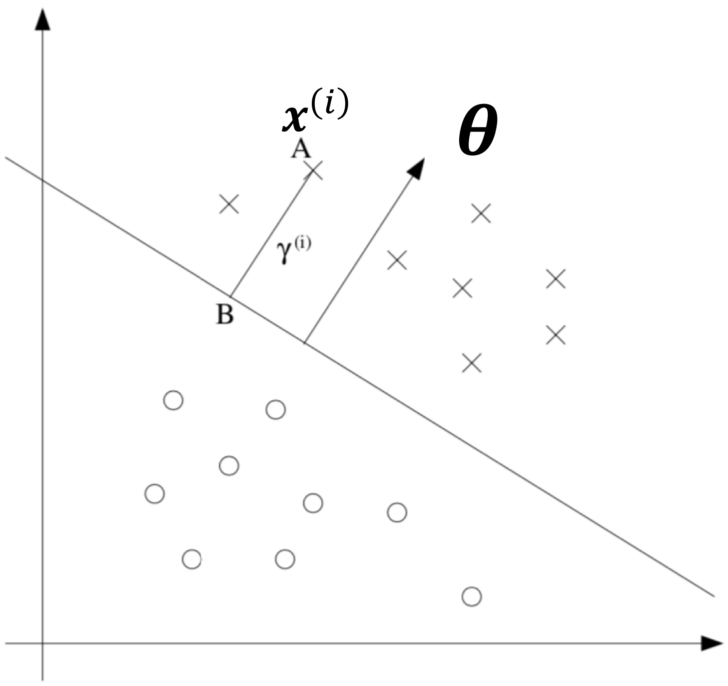
$$\text{B} = \boldsymbol{x}^{(i)} + \gamma^{(i)} \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2}$$

- Since B lies on the hyperplane, it satisfies

$$\boldsymbol{\theta}^T \boldsymbol{x} = 0$$

- Hence, $\boldsymbol{\theta}^T \left(\boldsymbol{x}^{(i)} + \gamma^{(i)} \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|} \right) = 0$ which gives us

$$\gamma^{(i)} = -\frac{\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}{\|\boldsymbol{\theta}\|_2}$$



Computing Margin

$$\gamma^{(i)} = \text{length(AB)}$$

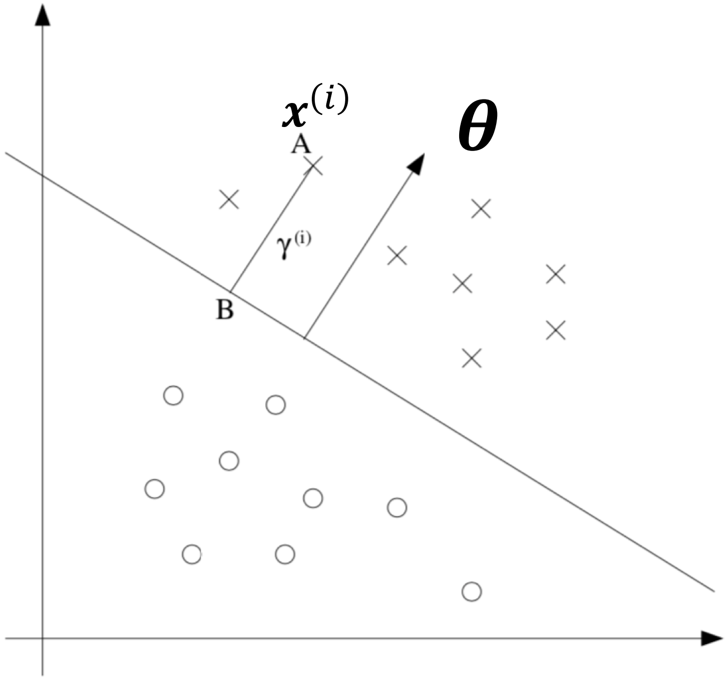
- If $y^{(i)} = +1$, then $\gamma^{(i)} = \frac{\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}{\|\boldsymbol{\theta}\|_2}$
- If $y^{(i)} = -1$, then $\gamma^{(i)} = -\frac{\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}{\|\boldsymbol{\theta}\|_2}$
- We can combine the two cases as:

$$\gamma^{(i)} = y^{(i)} \frac{\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}{\|\boldsymbol{\theta}\|_2}$$

- (DIY) If there is bias ($\theta_0 \neq 0$, add $x_0^{(i)} = 1$), then:

$$\gamma^{(i)} = y^{(i)} \frac{\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}}{\|\boldsymbol{w}\|_2}$$

(recall weights $\boldsymbol{w} = \boldsymbol{\theta}_{1:d}$)



Max Margin Classification

- **Definition:** The margin of a dataset $\left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{n}$ w.r.t. a hyperplane is the minimum of the margins of all datapoints

$$\gamma = \min_{i=1,2,\cdots,n} \gamma^{(i)}$$

- **Max Margin Classification:** Optimize for a hyperplane that maximizes the margin of the training dataset

$$\max_{\boldsymbol{\theta}} \min_{i=1,2,\cdots,n} \gamma^{(i)}$$

Today's Lecture

- 3 Different Views of SVMs

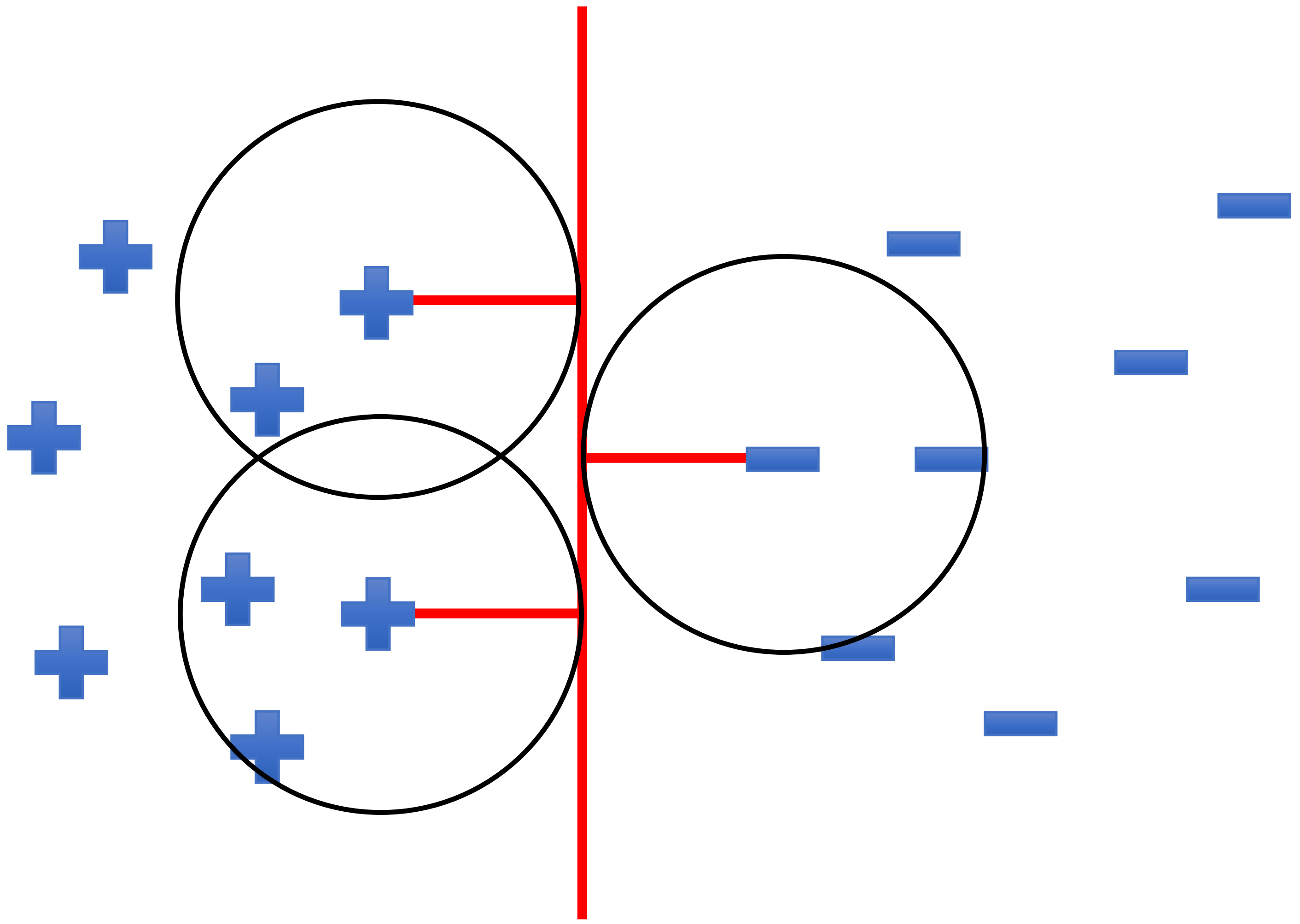
Constrained Optimization

- Hard-margin SVM
- Soft-margin SVM

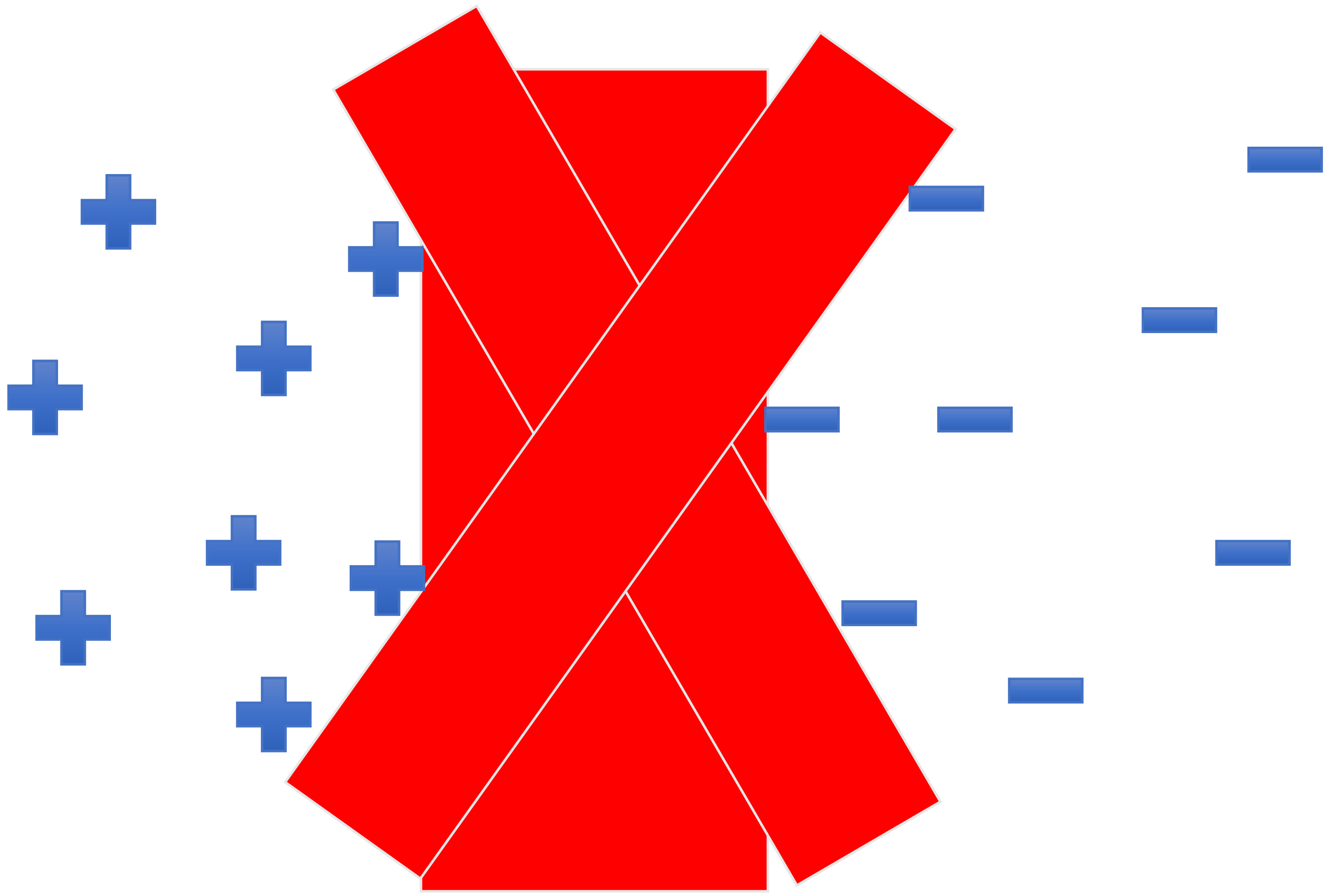
Unconstrained Optimization

- Hinge Loss

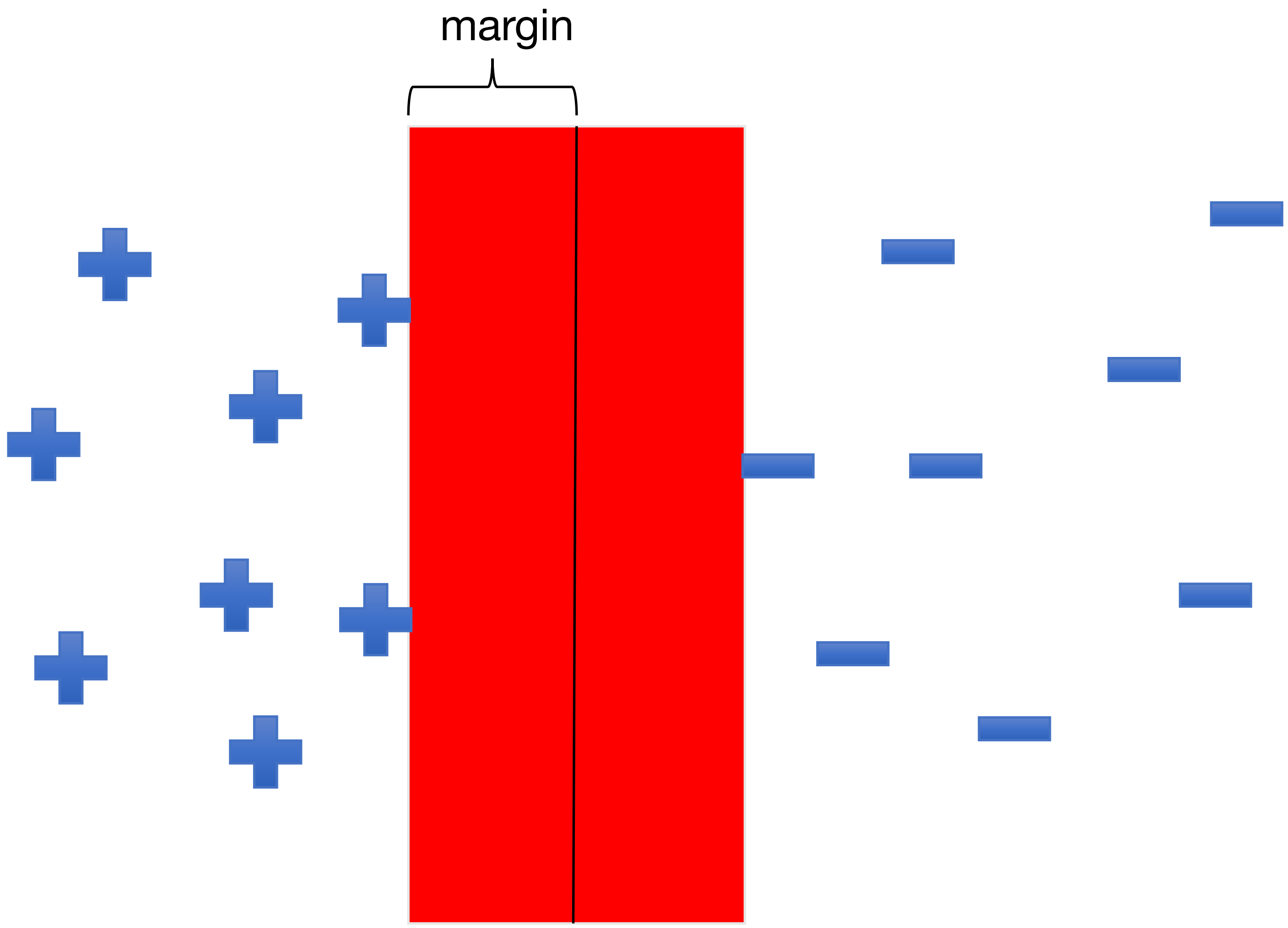
Maximizing the Margin



High Margin Separators



High Margin Separators



Maximizing Margins

- **Principle:** Optimize for a hyperplane that maximizes the margin of the training dataset

$$\max_{\boldsymbol{\theta}} \gamma = \max_{\boldsymbol{\theta}} \min_{i=1,2,\cdots,n} \gamma^{(i)}$$
$$= \max_{\boldsymbol{\theta}} \frac{1}{\|\boldsymbol{w}\|_2} \min_{i=1,2,\cdots,n} y^{(i)} \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}$$

- Note $\boldsymbol{w} = \boldsymbol{\theta}_{1:d}$ includes variables being optimized
- (Stated without proof): Objective is non-convex
  - Hard optimization problem
- Can we transform it to a simpler problem?
  - Intuition: separate $\frac{1}{\|\boldsymbol{w}\|_2}$ (nice) from $\min_{i=1,2,\cdots,n} y^{(i)} \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}$ (nasty) using **constraints**

Background: Constrained Optimization

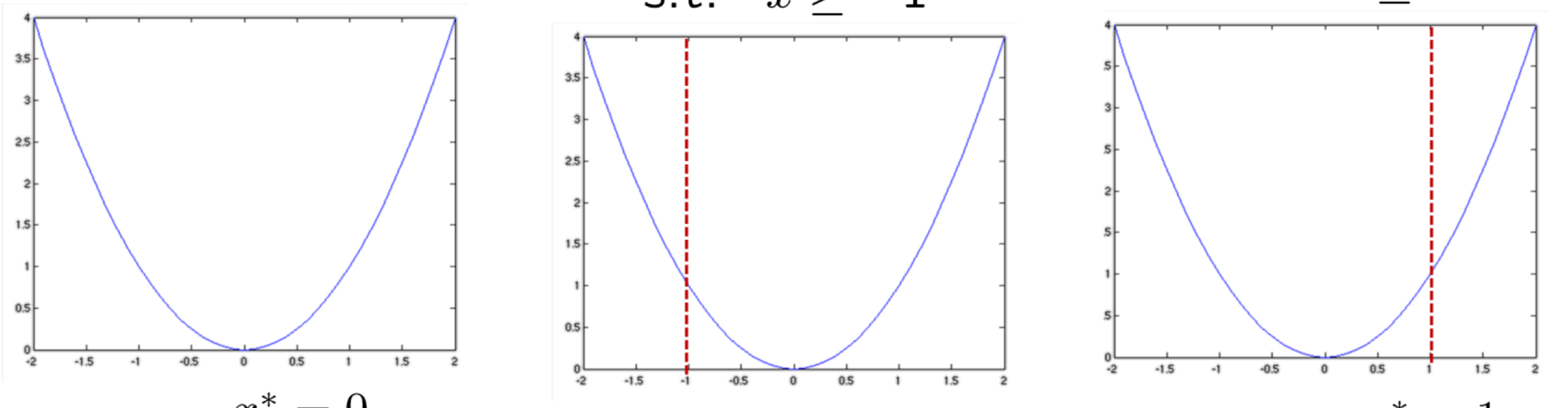
$\boxed{\text{min}_x \;\; x^2}$ **Objective function**

$\boxed{\text{s.t.} \;\; x \geq b}$ **Constraints**

$\text{min}_x \;\; x^2$

$\text{min}_x \;\; x^2$
$\text{s.t.} \;\; x \geq -1$

$\text{min}_x \;\; x^2$
$\text{s.t.} \;\; x \geq 1$



$x^* = 0$ $\quad$ $x^* = 0$ $\quad$ $x^* = 1$

Hard-Margin SVMs

- **Hard-Margin Support Vector Machines (SVM)**

$$\boxed{\min_{\boldsymbol{\theta}} \frac{1}{2} ||\boldsymbol{w}||_2^2}$$ **Objective function**

such that

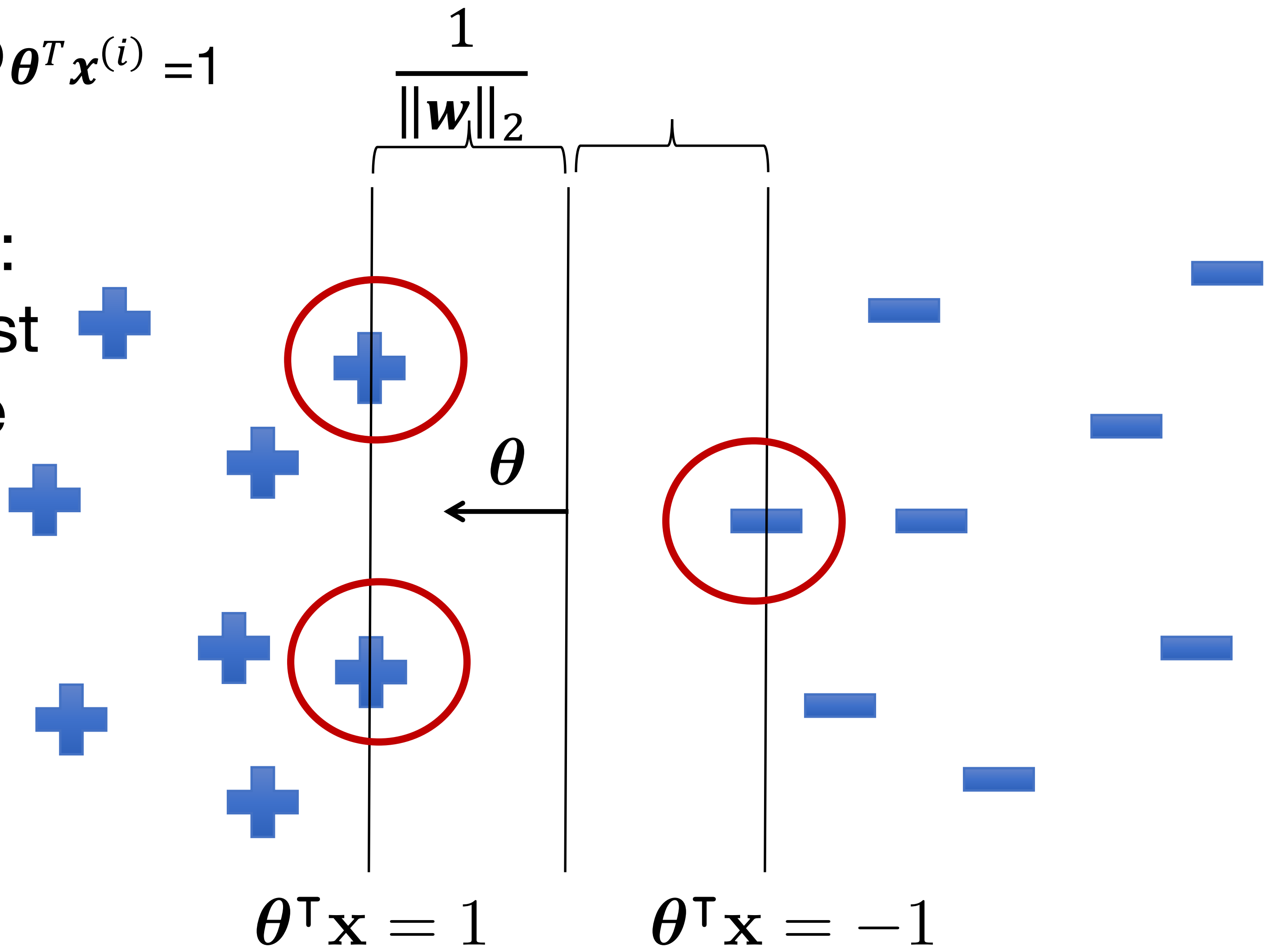
$$\boxed{y^{(i)} \boldsymbol{\theta}^T \boldsymbol{x}^{(i)} \geq 1 \text{ for all } i = 1, 2, \cdots, n}$$ **Constraints**

- (Stated without proof): Same solution for $\boldsymbol{\theta}$ as previous slide
- Convex quadratic objective, n linear constraints
  - Can be solved using off-the-shelf quadratic programming solvers

Intuition: Hard Margin

Assume $\min_{i=1,2,\cdots,n} y^{(i)}\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}$ =1

Support Vectors: Datapoints closest to the hyperplane



Support Vector Machines

- **Hard-Margin Support Vector Machines (SVM)**

$$\boxed{\max_{\boldsymbol{\theta}} \frac{1}{\|\boldsymbol{w}\|_2}}$$ **Objective function**

such that

$$\boxed{y^{(i)}\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} \geq 1 \text{ for all } i = 1,2,\cdots,n}$$ **Constraint**

- (Stated without proof): This is equivalent to **max-margin classification**
- SVMs are also called max-margin classifiers
- **Hard-margin:** Every training point has a margin

Support Vector Machines

- Note: $\arg\max_{\boldsymbol{\theta}} \frac{1}{\|\boldsymbol{w}\|_2} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{w}\|_2^2$
    - RHS is convex (similar to MSE)
- **Hard-Margin Support Vector Machines (SVM) -**

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{w}\|_2^2$$

such that

$$y^{(i)}\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} \geq 1 \text{ for all } i = 1,2,\cdots,n$$

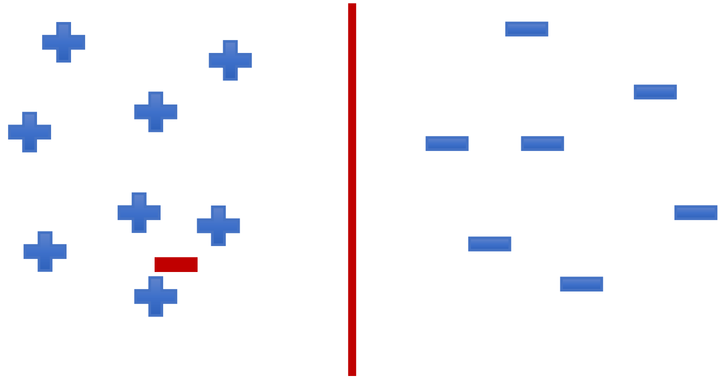
- Same solution for $\boldsymbol{\theta}$ as previous slide; just written in standard form as a minimization problem
- Convex quadratic objective, linear constraints
- Can be solved using off-the-shelf quadratic programming solvers

Non Separable Data

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|\boldsymbol{\theta}\|_2^2$$

subject to

$$y^{(i)} \boldsymbol{\theta}^T \boldsymbol{x}^{(i)} \geq 1 \; \forall i = 1,2,\cdots,n$$



- So far, we have assumed our data is linearly separable
- For any linear separator here, at least one constraint is violated and hence, the problem is infeasible
- Can we relax the constraints?

Soft-margin SVM with Slack Variables

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \|\boldsymbol{\theta}\|_2^2$$
subject to
$$y^{(i)} \boldsymbol{\theta}^T \boldsymbol{x}^{(i)} \geq 1 \ \forall i = 1,2,\cdots,n$$

Hard Margin SVM

$$\min_{\boldsymbol{\theta},\boldsymbol{\varepsilon}} \frac{1}{2} \|\boldsymbol{\theta}\|_2^2 + \mathrm{C} \sum_{i=1}^{n} \varepsilon_i$$
subject to
$$y^{(i)} \boldsymbol{\theta}^T \boldsymbol{x}^{(i)} \geq 1 - \varepsilon_i \ \forall i = 1,2,\cdots,n$$
$$\varepsilon_i \geq 0 \quad \forall i = 1,2,\cdots,n$$

Soft Margin SVM

- **Slack variables:** Introduce an additional non-negative optimization variable for each training point $\varepsilon_i \geq 0, i = 1, \dots, n$
- Set margin threshold to $1 - \varepsilon_i$ [i.e., the permitted slack]
- Incorporate another loss term to minimize overall slack
- New optimization problem (objective, constraints) is still convex

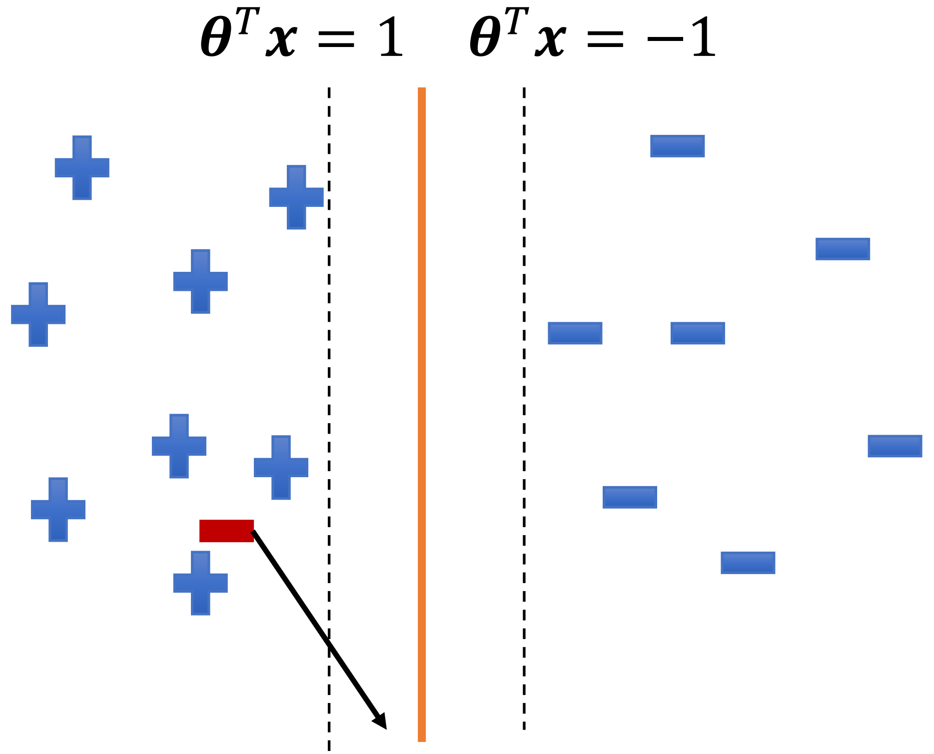
Interpreting Slack Constraints

$$\min_{\boldsymbol{\theta},\boldsymbol{\varepsilon}} \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 + \mathrm{C}\sum_{i=1}^{n} \varepsilon_i$$

subject to

$$y^{(i)}\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} \geq 1 - \varepsilon_i \;\forall i = 1,2,\cdots,n$$

$$\varepsilon_i \geq 0 \quad \forall i = 1,2,\cdots,n$$



$\varepsilon_i > 0$ for points **violating** $y^{(i)}\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} \geq 1$
and $\varepsilon_i = 0$ **otherwise**

Interpreting C

How to choose C?
Treat as hyperparameter and validate

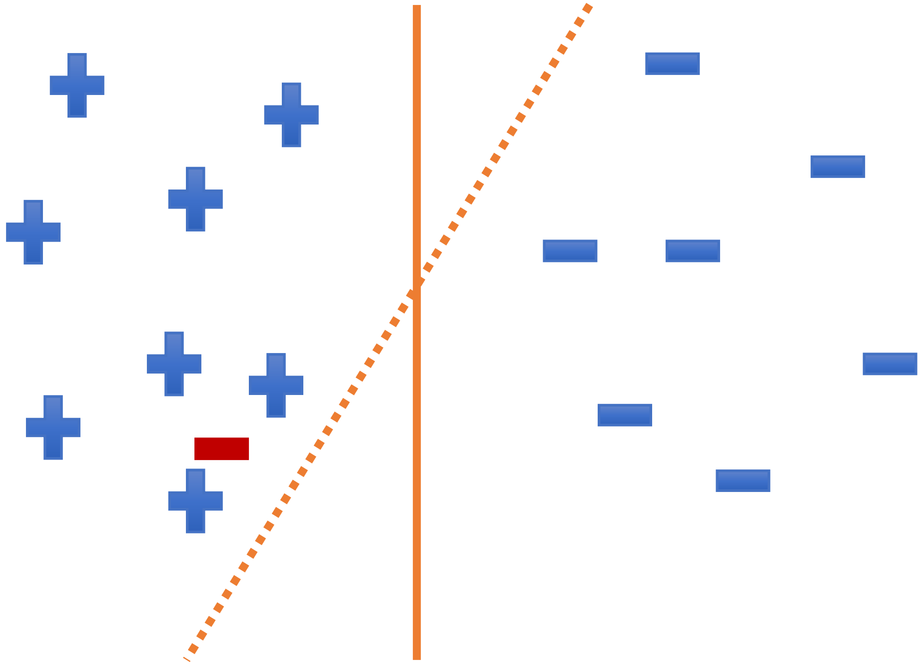
$$\min_{\boldsymbol{\theta},\boldsymbol{\varepsilon}} \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 + C\sum_{i=1}^{n} \varepsilon_i$$

subject to

$$y^{(i)}\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} \geq 1 - \varepsilon_i \;\forall i = 1,2,\cdots,n$$

$$\varepsilon_i \geq 0 \quad \forall i = 1,2,\cdots,n$$

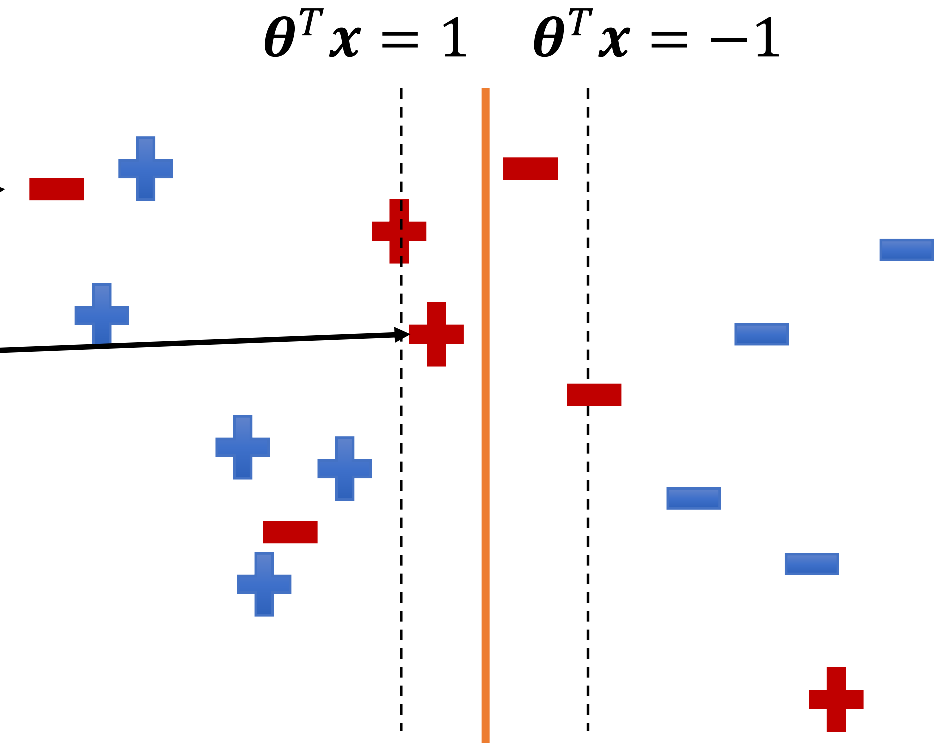
High slack preferred
when C **is low**



Low slack preferred
when C **is high**

General Definition of Support Vectors

- The SVM solution only depends on a subset of the points called **support vectors**
  - Misclassified points i.e. $y^{(i)}\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} \le 0$
  - Points within default margin i.e., $0 < y^{(i)}\ \boldsymbol{\theta}^T\boldsymbol{x}^{(i)} \le 1$
- Removing any other point will not change the SVM solution
- All **red** points are support vectors



Hinge Loss Perspective for SVMs

- Recall Regularized Linear models

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = g(\boldsymbol{\theta}^T \boldsymbol{x})$$

$$J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} L(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}, y^{(i)}) + \lambda \|\boldsymbol{\theta}_{1:d}\|_2^2$$

where

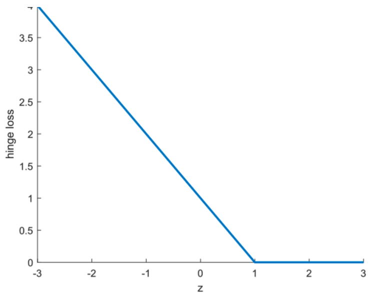
  - **Linear Regression:** $g(z) = z$ and $L\left(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}, y^{(i)}\right) = \left(y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right)^2$
  - **Perceptron:** $g(z) = \mathrm{sign}(z)$ and $L\left(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}, y^{(i)}\right) = \max(0, -y^{(i)} \boldsymbol{\theta}^T \boldsymbol{x}^{(i)})$
  - **Logistic Regression:** $g(z) = \mathrm{sigmoid}(z)$ and $L\left(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}, y^{(i)}\right) = -y^{(i)} \log\left(\mathrm{sigmoid}\left(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right)\right) - \left(1 - y^{(i)}\right) \log\left(1 - \mathrm{sigmoid}\left(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right)\right)$

- For **SVMs**, $g(z) = \mathrm{sign}(z)$. Can we find an $L(\cdot)$ to fit SVMs in this template?

Hinge Loss

- Given an input pair $(\boldsymbol{x}^{(i)}, y^{(i)})$ and a linear separator $\boldsymbol{\theta}$, the **hinge loss** is:

$$L\left(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}, y^{(i)}\right) = \max\left(0, 1 - y^{(i)} \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right) = \left(1 - y^{(i)} \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right)_+$$



- No penalty if raw output, $\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}$ has same sign and is far enough from decision boundary
- Otherwise pay a growing penalty, between 0 and 1 if signs match, and greater than one otherwise

Summary

Margin

Formalizing distance of a separating hyperplane to a dataset

Support Vector Machines

New class of ML models that based on maximizing margins

Can be transformed into a constrained optimization problem

- Convex quadratic objective, linear constraints