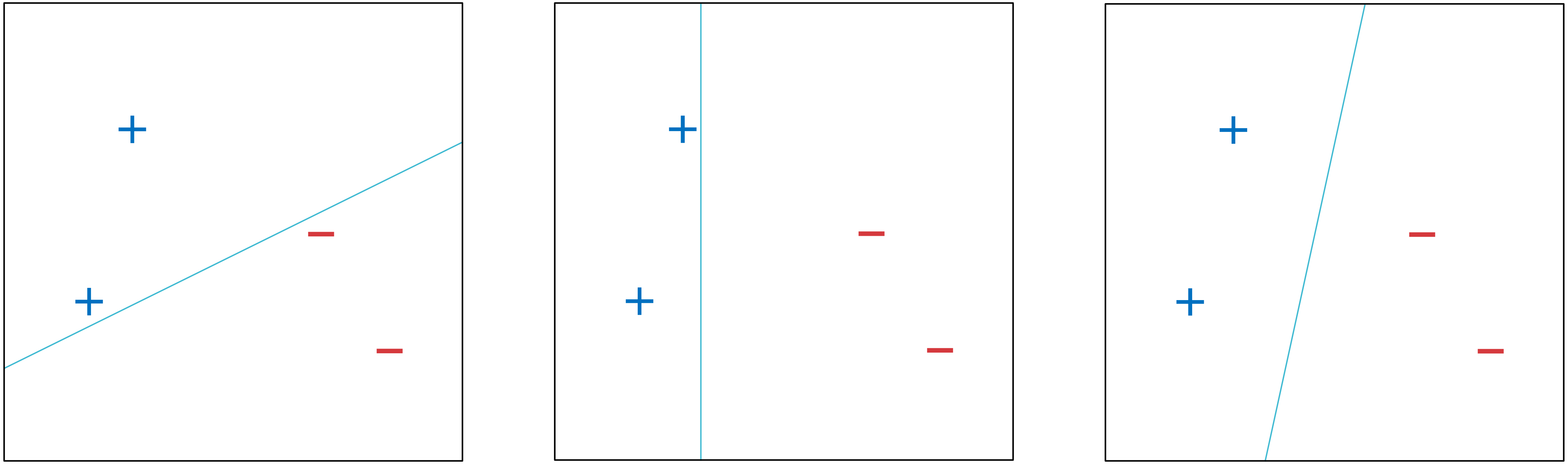
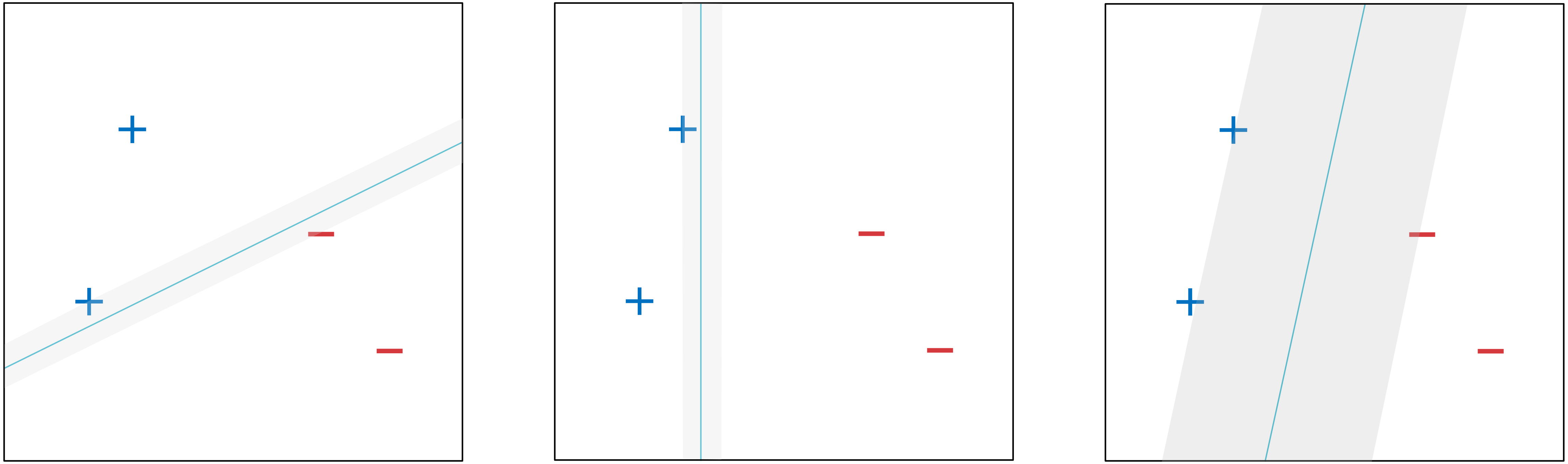
10-315: Introduction to Machine Learning Lecture 15– Support Vector Machines

Henry Chai

3/21/22



Which linear separator is best?



Which linear separator is best?

Maximal Margin Linear Separators

- The margin of a linear separator is the distance between it and the nearest training data point
- Questions:
  1. How can we efficiently find a maximal-margin linear separator?
  2. Why are linear separators with larger margins better?
  3. What can we do if the data is not linearly separable?

Hyperplanes

- For linear models, decision boundaries are D-dimensional ***hyperplanes*** defined by a weight vector, $[b, \boldsymbol{w}]$

$$\boldsymbol{w}^T \boldsymbol{x} + b = 0$$

- Problem: there are infinitely many weight vectors that describe the same hyperplane
  - $x_1 + 2x_2 + 2 = 0$ is the same line as $2x_1 + 4x_2 + 4 = 0$, which is the same line as $1000000x_1 + 2000000x_2 + 2000000 = 0$

Solution: normalize the weight vectors w.r.t. training data set

Normalizing Hyperplanes

- Given a dataset $\mathcal{D} = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ where $y \in \{-1, +1\}$, $\hat{y} = \text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b)$ is a valid linear separator if

$$y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) > 0 \;\forall\; (\boldsymbol{x}^{(i)}, y^{(i)}) \in \mathcal{D}$$

For SVMs, only consider

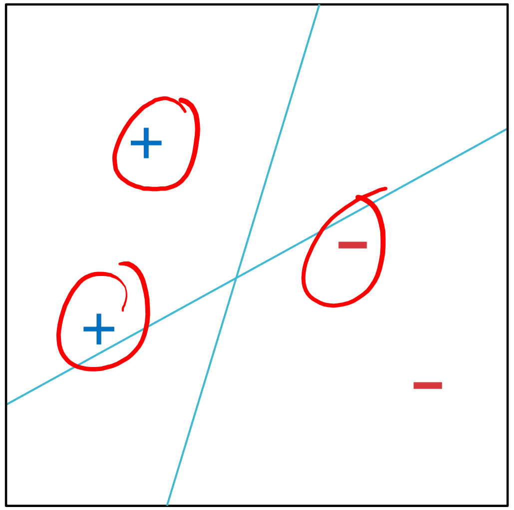
$$\mathcal{H} = \{\hat{y} = \text{sign}(w^T x^{(i)} + b) : \underline{\min_{(x^{(i)}, y^{(i)})} y^{(i)}(w^T x^{(i)} + b) = 1}\}$$

if $\hat{y} = \text{sign}(w^T x^{(i)} + b)$, then $\exists\, \rho$ s.t.

$$\hat{y} = \text{sign}\left(\frac{w^T}{\rho} x^{(i)} + \frac{b}{\rho}\right) \in \mathcal{H}$$

Normalizing Hyperplanes: Example

b	w_1	w_2	
-0.2	-0.6	1	$\notin \mathcal{H}$
-0.4	-1.2	2	$\notin \mathcal{H}$
-2	-6	10	$\notin \mathcal{H}$
-10	-30	50	$\in \mathcal{H}$
0.2	-0.6	0.2	$\notin \mathcal{H}$
0.1	-0.3	0.1	$\notin \mathcal{H}$
1	-3	1	$\notin \mathcal{H}$
2	-6	2	$\in \mathcal{H}$



x_1	x_2	y	$y(\boldsymbol{w}^T\boldsymbol{x} + b)$
0.2	0.4	+1	1.6
0.3	0.8	+1	1.8
0.7	0.6	-1	1
0.8	0.3	-1	2.2

$-1(0.7(-6) + 0.6(10) - 2)$

$-1(-4.2 + 6 - 2) = 0.2$

Computing the Margin

- Claim: $\boldsymbol{w}$ is orthogonal to the hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ (the decision boundary)
- A vector is orthogonal to a hyperplane if it is orthogonal to every vector in that hyperplane
- Vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are orthogonal if $\boldsymbol{\alpha}^T\boldsymbol{\beta} = 0$
- Proof:

Let x' and x'' be two arbitrary points on

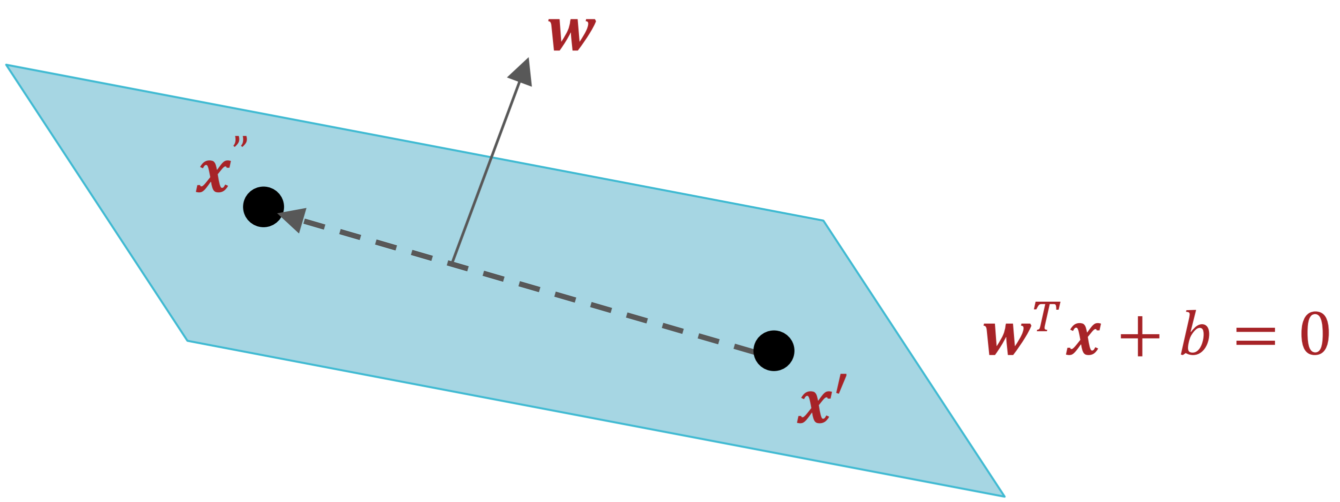
$w^Tx = -b \leftarrow w^Tx + b = 0 \Rightarrow x' - x''$ is a vector on $w^Tx + b = 0$

$w^T(x' - x'') = w^Tx' - w^Tx''$

$= (-b) - (-b) = 0$ ∎

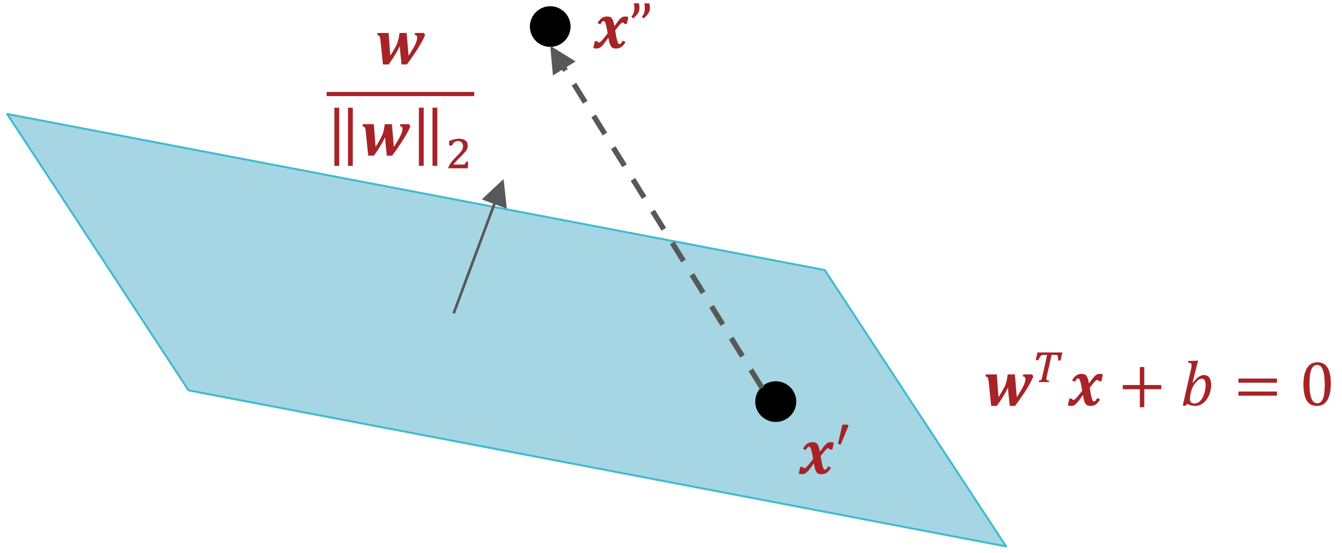
Computing the Margin

- Claim: $\boldsymbol{w}$ is orthogonal to the hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ (the decision boundary)
- A vector is orthogonal to a hyperplane if it is orthogonal to every vector in that hyperplane
- Vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are orthogonal if $\boldsymbol{\alpha}^T\boldsymbol{\beta} = 0$



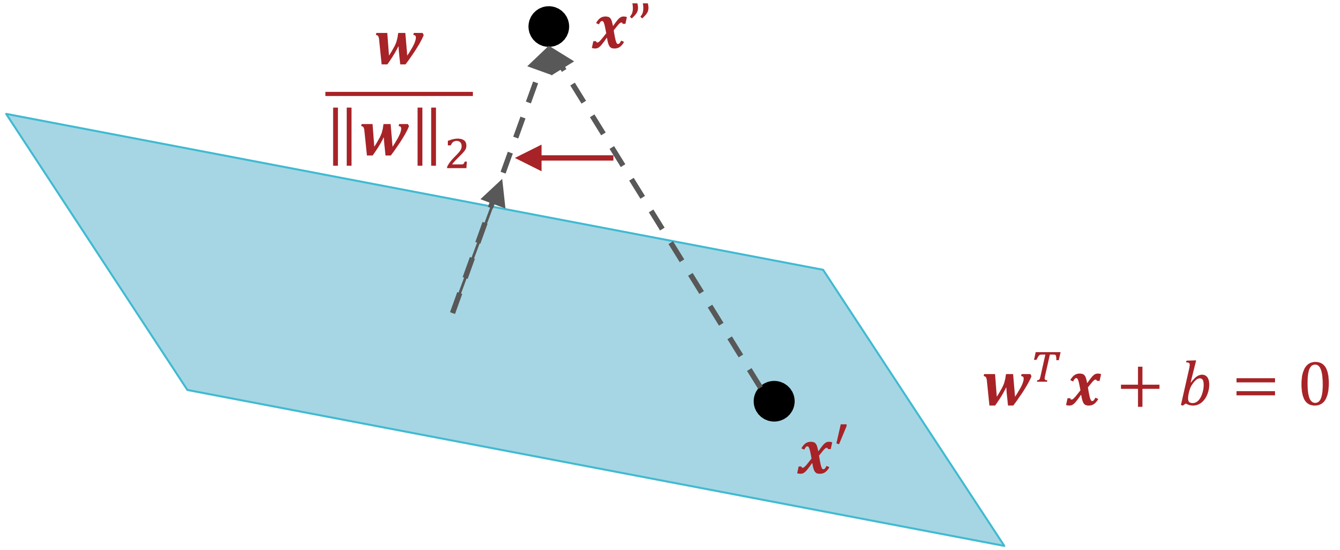
Computing the Margin

- Let $\boldsymbol{x}'$ be an arbitrary point on the hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ and let $\boldsymbol{x}''$ be an arbitrary point
- The distance between $\boldsymbol{x}''$ and $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ is equal to the magnitude of the projection of $\boldsymbol{x}'' - \boldsymbol{x}'$ onto $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, the unit vector orthogonal to the hyperplane



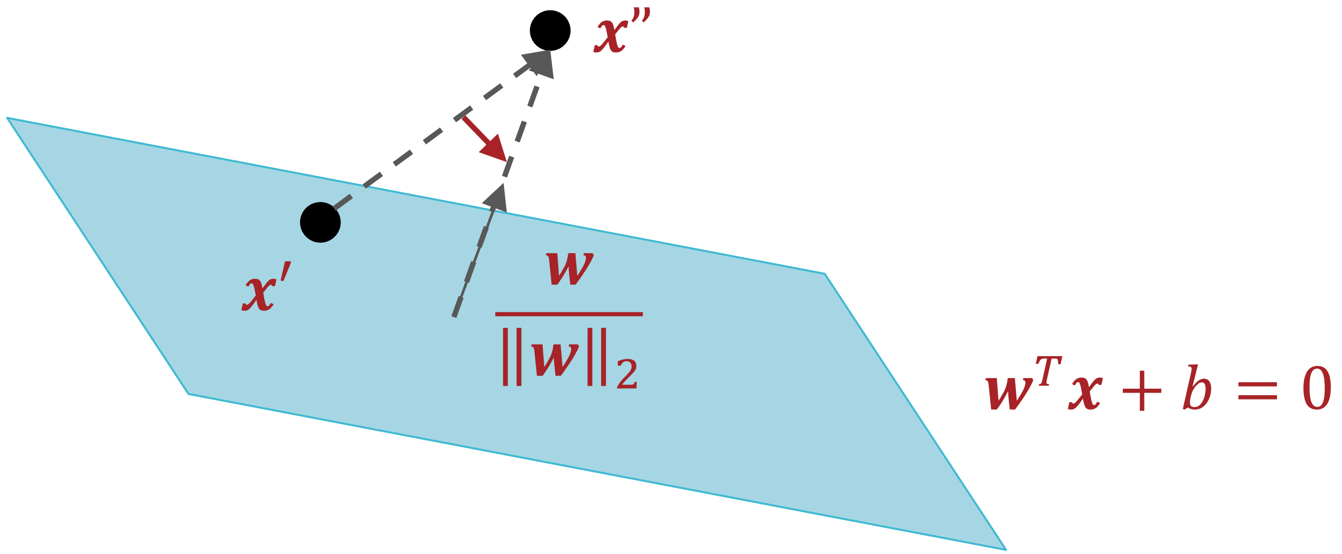
Computing the Margin

- Let $\boldsymbol{x}'$ be an arbitrary point on the hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ and let $\boldsymbol{x}''$ be an arbitrary point
- The distance between $\boldsymbol{x}''$ and $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ is equal to the magnitude of the projection of $\boldsymbol{x}'' - \boldsymbol{x}'$ onto $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, the unit vector orthogonal to the hyperplane



Computing the Margin

- Let $\boldsymbol{x}'$ be an arbitrary point on the hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ and let $\boldsymbol{x}''$ be an arbitrary point
- The distance between $\boldsymbol{x}''$ and $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ is equal to the magnitude of the projection of $\boldsymbol{x}'' - \boldsymbol{x}'$ onto $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, the unit vector orthogonal to the hyperplane



Computing the Margin

- Let $\boldsymbol{x}'$ be an arbitrary point on the hyperplane $h(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b = 0$ and let $\boldsymbol{x}''$ be an arbitrary point
- The distance between $\boldsymbol{x}''$ and $h(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b = 0$ is equal to the magnitude of the projection of $\boldsymbol{x}'' - \boldsymbol{x}'$ onto $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, the unit vector orthogonal to the hyperplane

$$d(x'', h) = \left|\frac{w^T(x'' - x')}{\|w\|_2}\right| = \left|\frac{w^Tx'' - w^Tx'}{\|w\|_2}\right|$$

$$= \left|\frac{w^Tx'' - (-b)}{\|w\|_2}\right| = \left|\frac{w^Tx'' + b}{\|w\|_2}\right|$$

Computing the Margin

- The margin of a linear separator is the distance between it and the nearest training data point

$$\min_{(x^{(i)},y^{(i)})\in D} d(x^{(i)},h) = \min_{(x^{(i)},y^{(i)})\in D} \left|\frac{w^T x^{(i)}+b}{\|w\|_2}\right|$$

$$= \frac{1}{\|w\|_2} \min_{(x^{(i)},y^{(i)})\in D} |w^T x^{(i)}+b|$$

$y^{(i)}$

$$= \frac{1}{\|w\|_2} \underbrace{\min_{(x^{(i)},y^{(i)})\in D} y^{(i)}(w^T x^{(i)}+b)}_{(1)}$$

$$= \frac{1}{\|w\|_2}$$

Maximizing the Margin

$$\text{maximize} \quad \frac{1}{\|w\|_2} = \frac{1}{\sqrt{w^T w}}$$

$$\text{s.t.} \quad \min_{(x^{(i)}, y^{(i)}) \in D} y^{(i)}\left(w^T x^{(i)} + b\right) = 1$$

$$\text{minimize} \quad w^T w$$

$$\text{s.t.} \quad \min_{(x^{(i)}, y^{(i)}) \in D} y^{(i)}\left(w^T x^{(i)} + b\right) = 1$$

$$\text{minimize} \quad w^T w$$

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

Maximizing the Margin

minimize $\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$

subject to $y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 \;\forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$

- If $\left[\hat{b}, \widehat{\boldsymbol{w}}\right]$ is the optimal solution, then $\exists$ at least one training data point $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$ s.t $y^{(i)}\left(\widehat{\boldsymbol{w}}^T\boldsymbol{x}^{(i)} + \hat{b}\right) = 1$
  - All training data points $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$ where $y^{(i)}\left(\widehat{\boldsymbol{w}}^T\boldsymbol{x}^{(i)} + \hat{b}\right) = 1$ are known as *support vectors*
- Converting the non-linear constraint (involving the $\min$) to N linear constraints means we can use quadratic programming (QP) to solve this problem in $O\left(D^3\right)$ time

Recipe for SVMs

- Define a model and model parameters
  - Assume a linear decision boundary (with normalized weights)

$$h(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b = 0$$

  - Parameters: $\boldsymbol{w} = [w_1, \ldots, w_D]$ and b
- Write down an objective function (with constraints)

$$\text{minimize} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{subject to} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

- Optimize the objective w.r.t. the model parameters
  - Solve using quadratic programming

Why Maximal Margins?

- Consider three binary data points in a **bounded** 2-D space
- Let $\mathcal{H}$ = {all linear separators} and

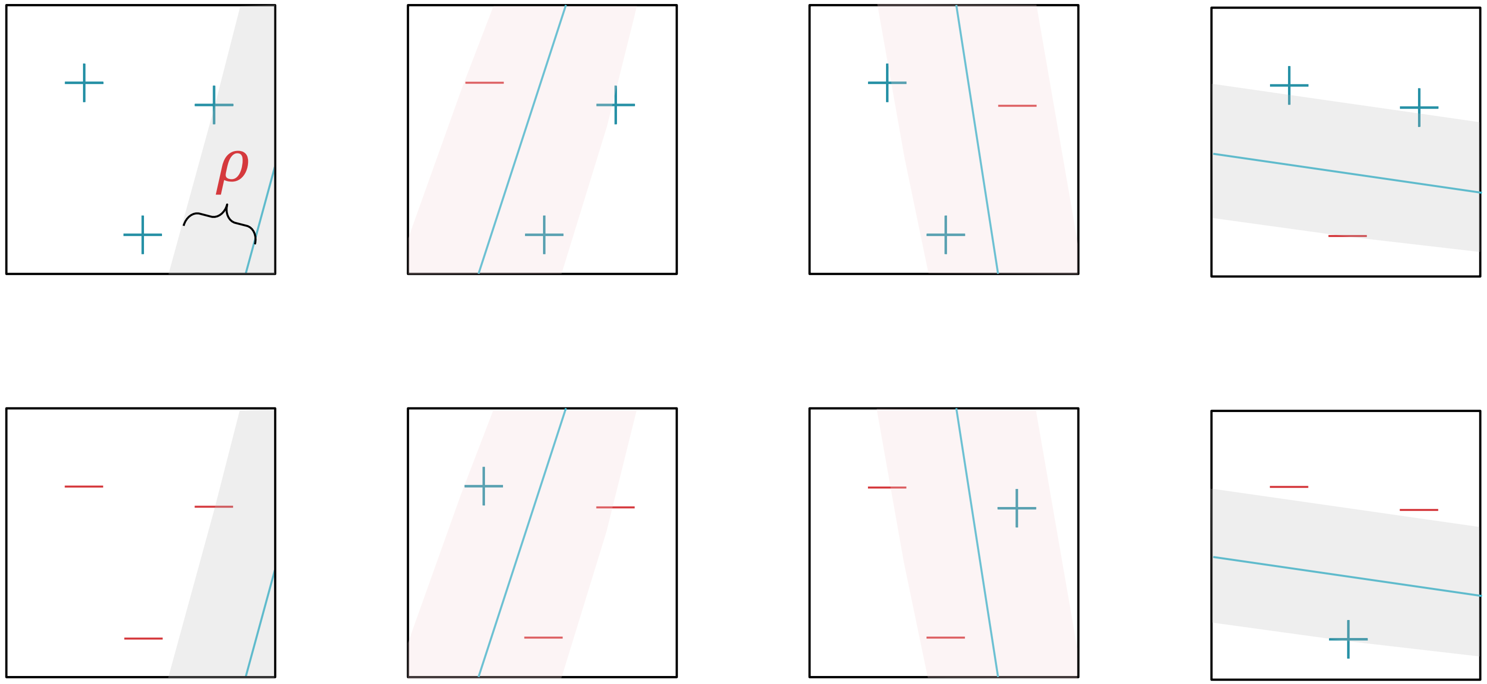
  $\mathcal{H}_\rho$ = {all linear separators with minimum margin ρ}

Why Maximal Margins?

- Consider three binary data points in a **<u>bounded</u>** 2-D space
- $\mathcal{H}$ = {all linear separators} can always correctly classify any three (non-colinear) data points in this space

Why Maximal Margins?

- Consider three binary data points in a **bounded** 2-D space
- $\mathcal{H}_\rho$ = {all linear separators with minimum margin ρ} cannot always correctly classify three non-colinear data points



Summary Thus Far

- The margin of a linear separator is the distance between it and the nearest training data point
- Questions:
  1. How can we efficiently find a maximal-margin linear separator? By solving a constrained quadratic optimization problem using quadratic programming
  2. Why are linear separators with larger margins better? They're simpler *waves hands*
  3. What can we do if the data is not linearly separable? Next!

Linearly Inseparable Data

- What can we do if the data is not linearly separable?

today → 1. Accept some non-zero training error

later → 2. Apply some feature transformation that makes your data linearly separable

SVMs

$$\text{minimize } \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{subject to } y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

- When $\mathcal{D}$ is not linearly separable, there are no feasible solutions to this optimization problem

Hard-margin SVMs

$$\text{minimize } \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{subject to } y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

- When $\mathcal{D}$ is not linearly separable, there are no feasible solutions to this optimization problem

Soft-margin SVMs

$$\text{minimize } \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi^{(i)}$$

$$\text{subject to } y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 - \xi^{(i)} \; \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

$$\xi^{(i)} \geq 0 \qquad \forall\, i \in \{1, \ldots, N\}$$

Soft-margin SVMs

minimize $\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi^{(i)}$ $\quad \lambda \geq 0$

subject to $y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 - \xi^{(i)}\ \forall\left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$

$\xi^{(i)} \geq 0 \qquad \forall\, i \in \{1, \dots, N\}$

- $\xi^{(i)}$ is the "soft" error on the i^{th} training data point
  - If $\xi^{(i)} > 1$, then $y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) < 0 \Rightarrow$ $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)$ is incorrectly classified
  - If $0 < \xi^{(i)} < 1$, then $y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) > 0 \Rightarrow$ $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)$ is correctly classified but inside the margin
- $\sum_{i=1}^{N}\xi^{(i)}$ is the "soft" training error

Soft-margin SVMs

$$\text{minimize} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi^{(i)}$$

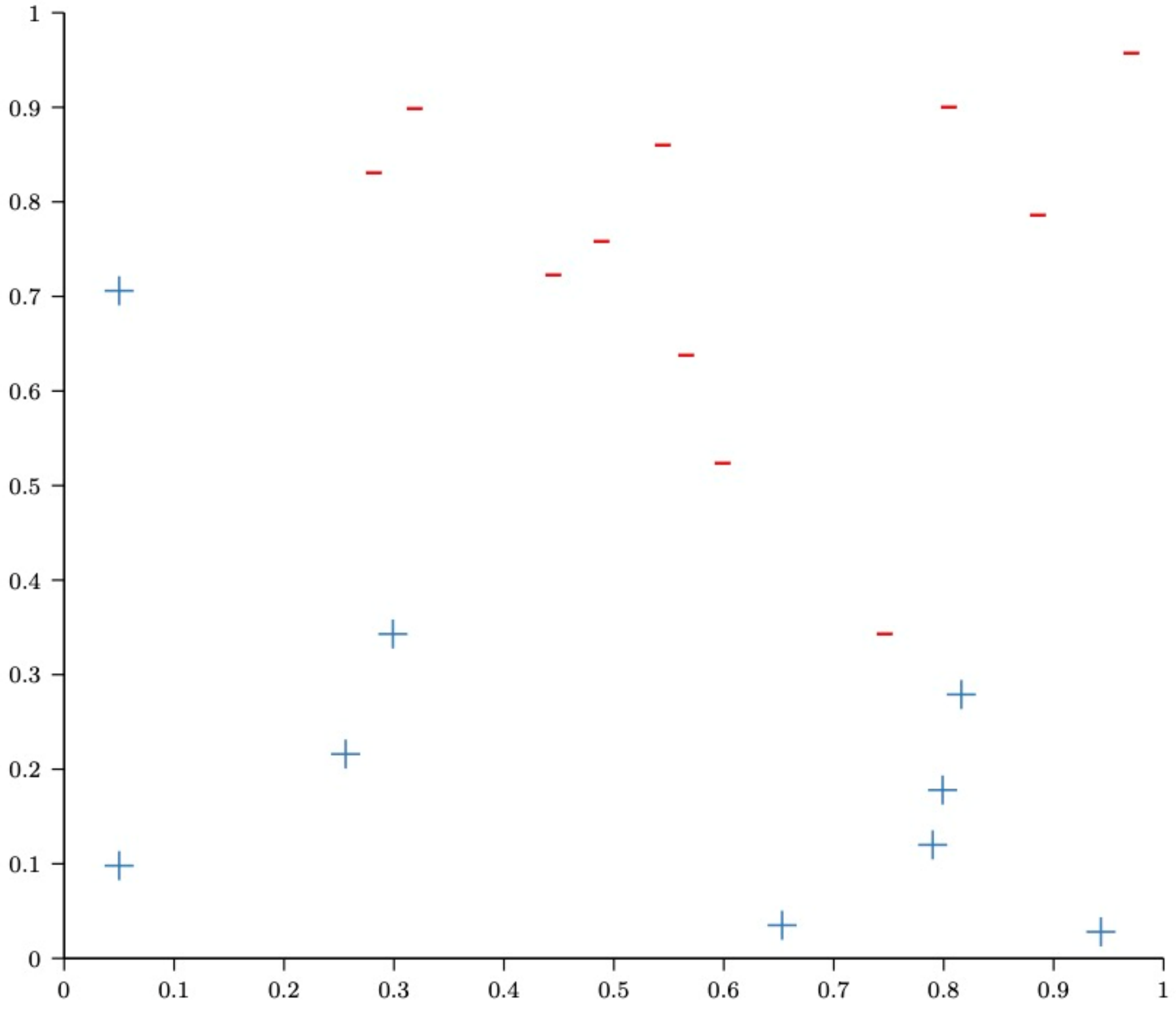
= 0

$$\text{subject to} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 - \xi^{(i)} \ \forall\left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

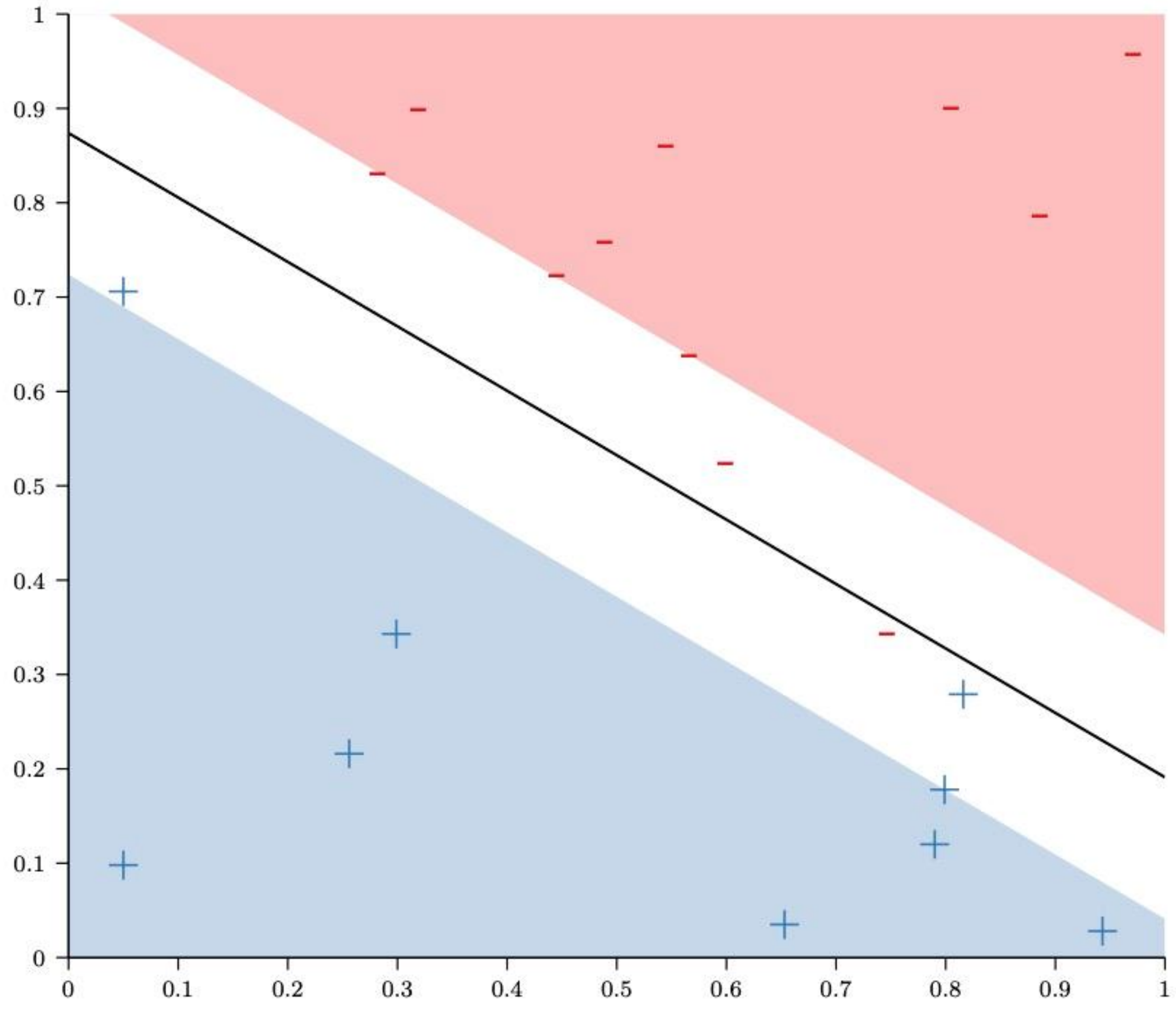
$$\xi^{(i)} \geq 0 \quad \forall\, i \in \{1, \dots, N\}$$

- Still solvable using quadratic programming
- All training data points $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$ where $\underbrace{y^{(i)}\left(\widehat{\boldsymbol{w}}^T\boldsymbol{x}^{(i)} + \hat{b}\right)}_{= \text{ for hard-margin}} \leq 1$ are known as *support vectors*

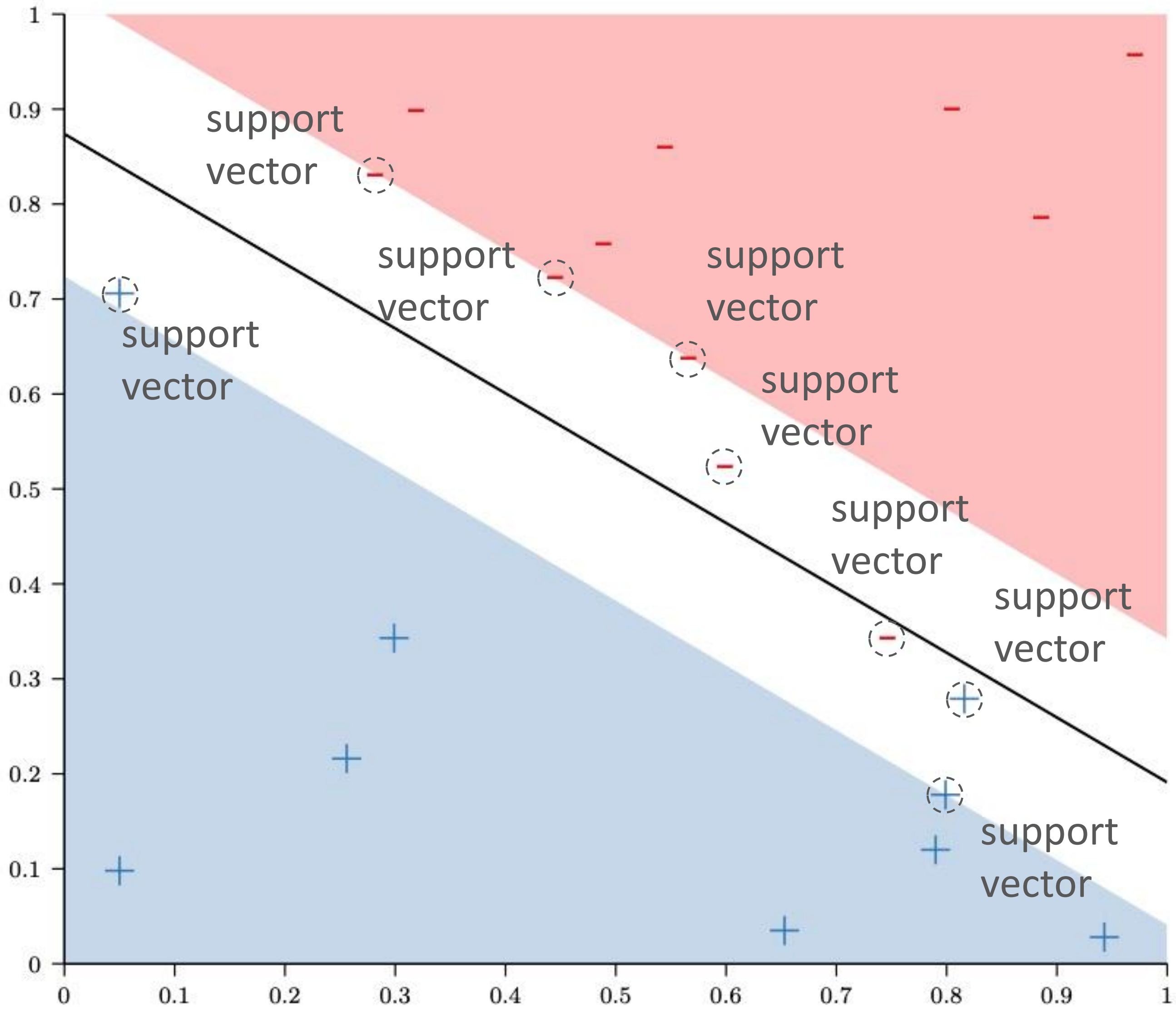
Interpreting $\xi^{(i)}$



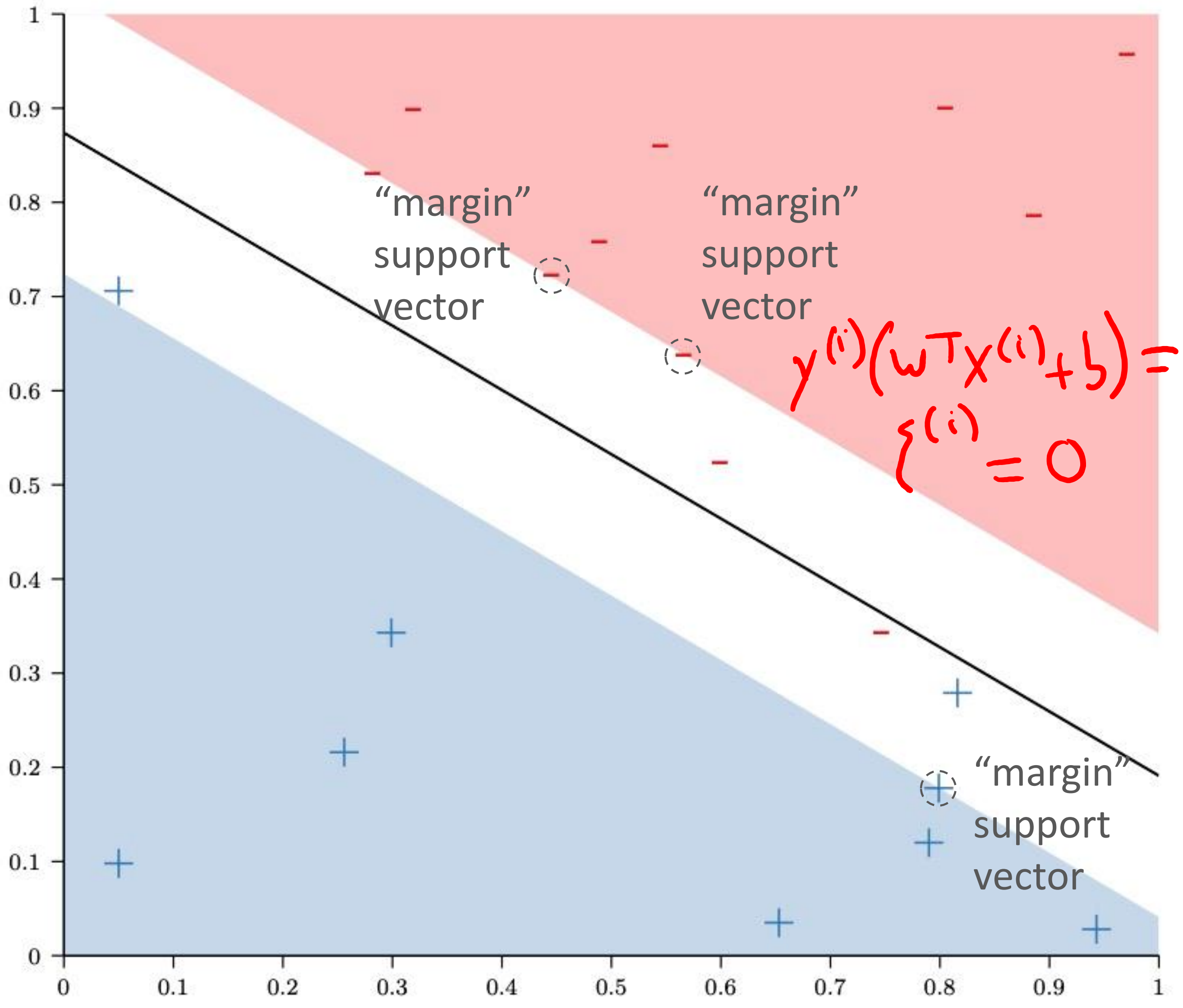
Interpreting $\xi^{(i)}$



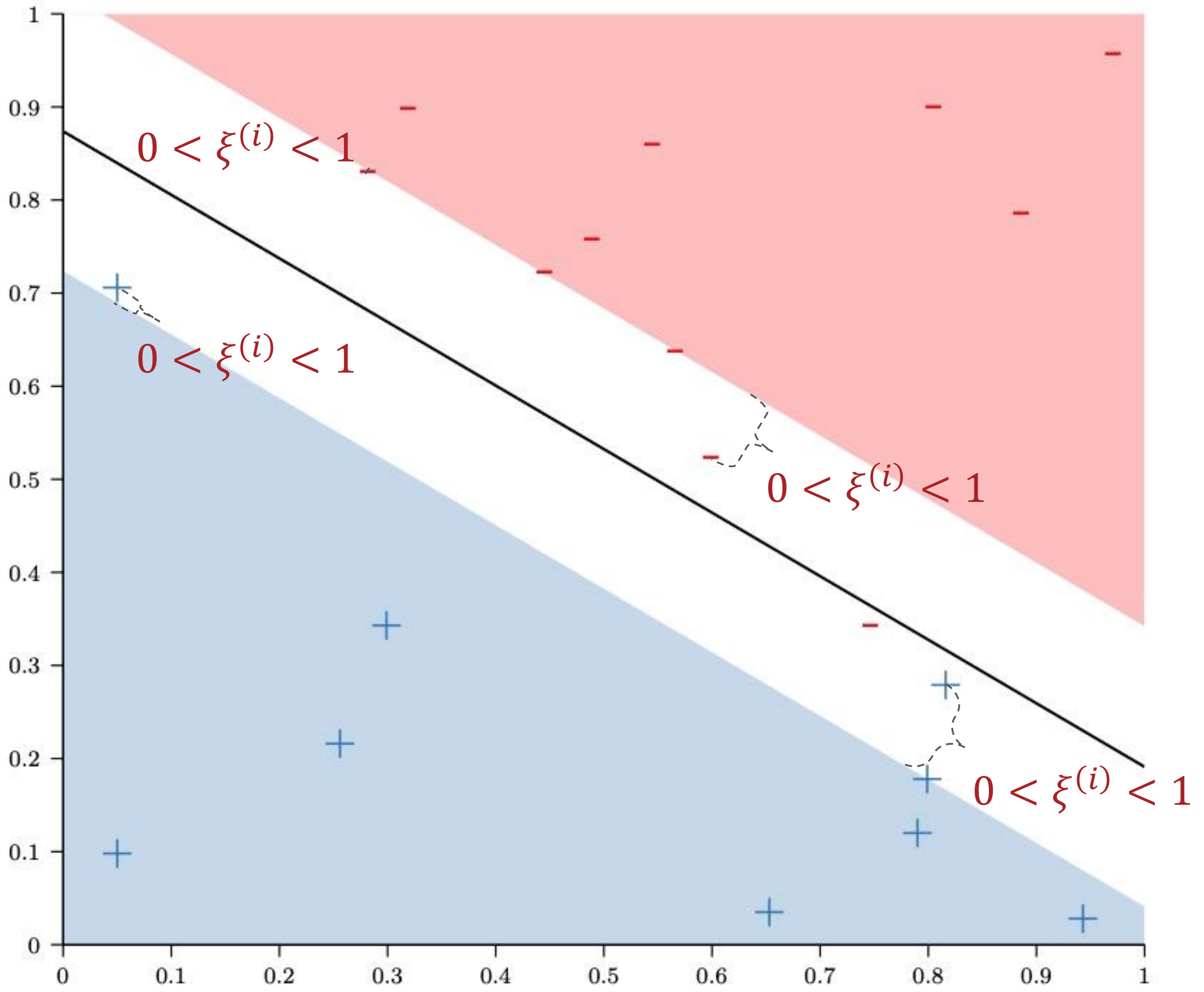
Interpreting $\xi^{(i)}$



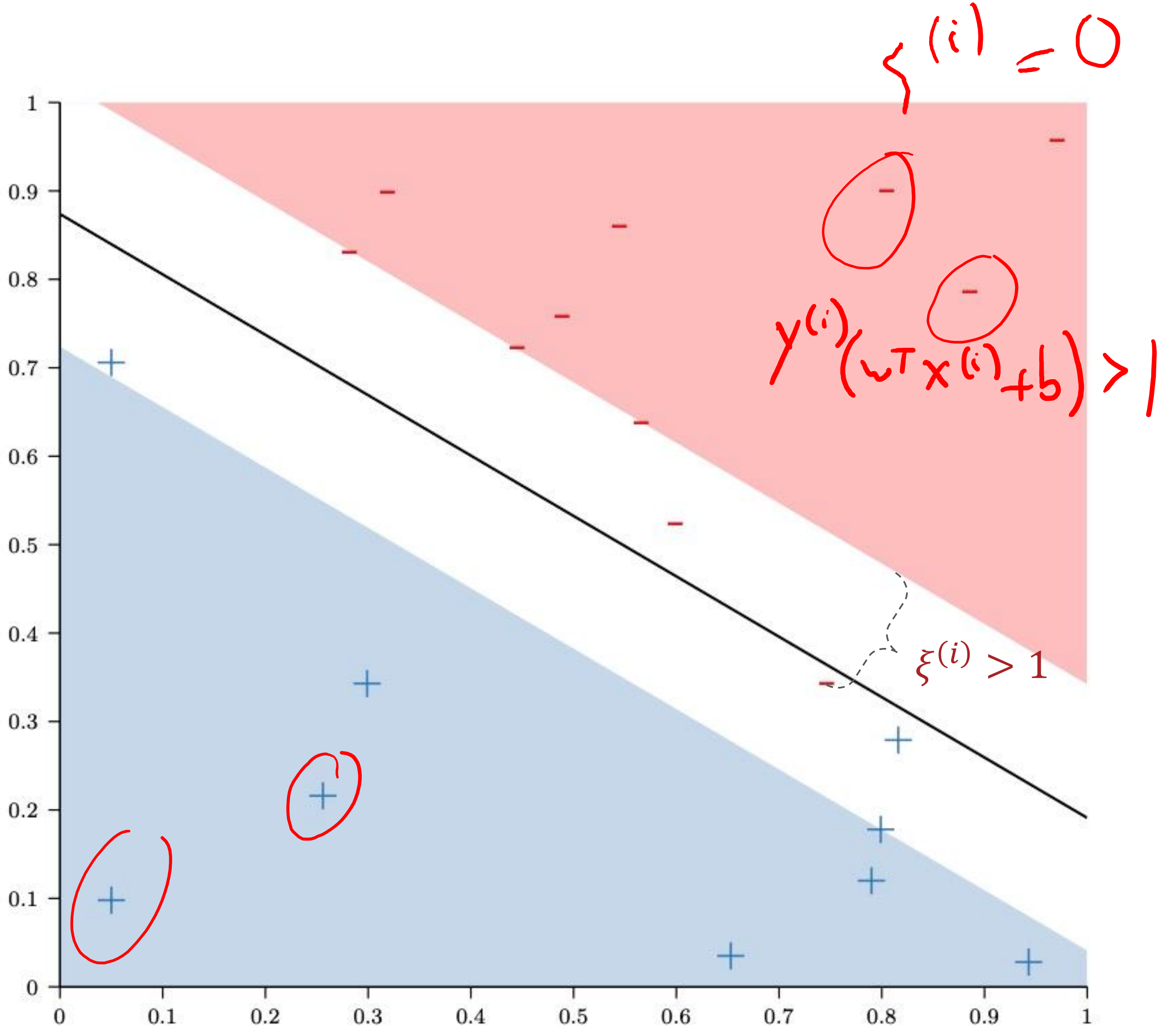
Interpreting $\xi^{(i)}$

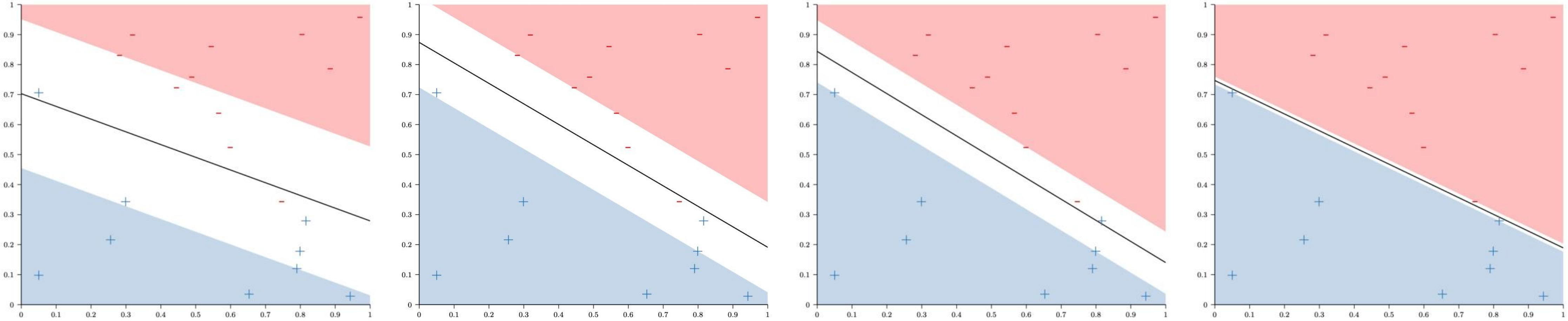


Interpreting $\xi^{(i)}$



Interpreting $\xi^{(i)}$





Smaller C

$$\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi^{(i)}$$

Larger C

Hard Margin

Setting C

C is a tradeoff parameter (much like the tradeoff parameter in regularization)