Support Vector Machine

Mahdi Roozbahani
Georgia Tech

Some of the slides are inspired based on slides from Andrew Zisserman, Yaser S. Abu-Mostafa.

Math moment

$$7 \times 0 = 0$$
$$3 \times 0 = 0$$
$$0 = 0$$
$$7 \times 0 = 3 \times 0$$
$$7 = 3$$

Outline

- Precursor: Linear Classifier and Perceptron
- Support Vector Machine
- Parameter Learning

$X =$ x_1, x_2 ... x_d $\quad y_a =$

$\theta_1 x_1 + \theta_2 x_2 + \dots + \theta_d x_d = f(x) = \hat{y}_p$

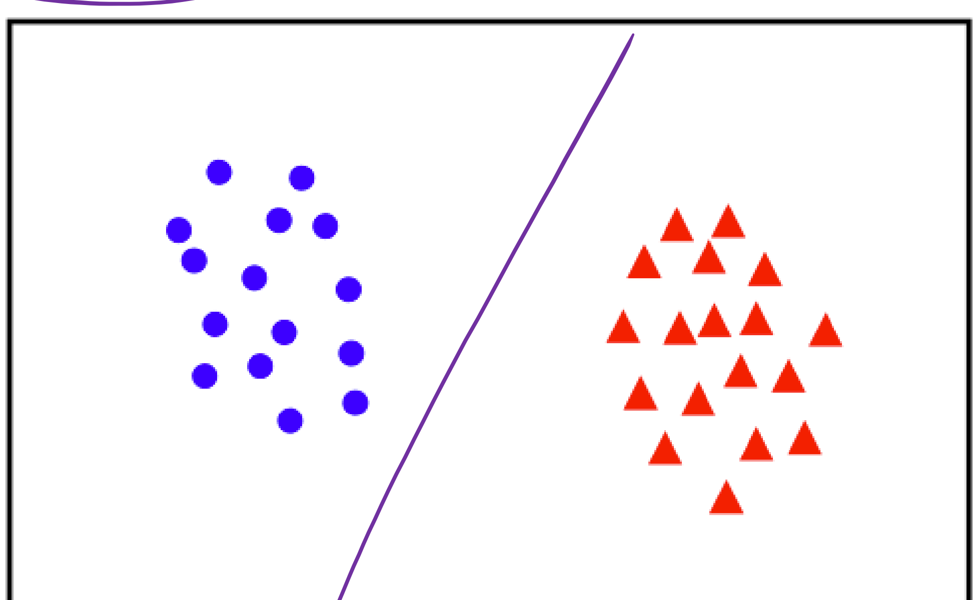
Binary Classification

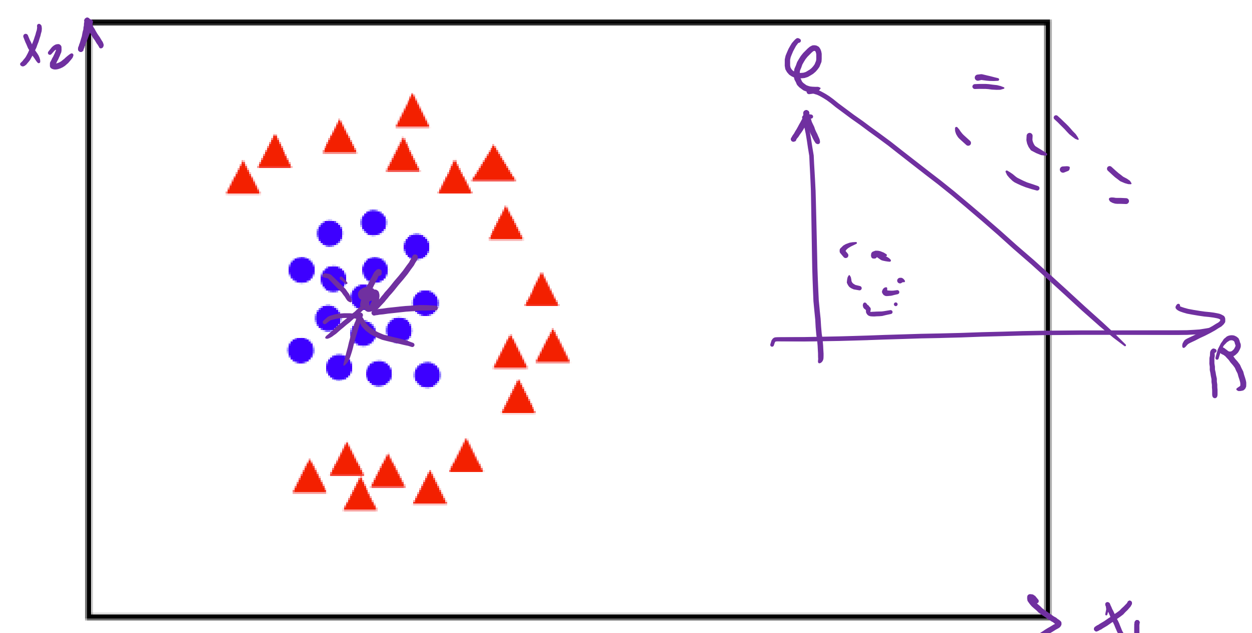
Given training data $(\overset{1 \times d}{x^{\{i\}}}, y^{\{i\}})$ for $i = 1 \dots N$, with $x^{\{i\}} \in \mathbb{R}^d$ and $y^{\{i\}} \in \{-1, 1\}$, learn a classifier $f(\mathbf{x})$ such that

$f(x^{\{i\}}) = \hat{y}_p$

$$f(x^{\{i\}}) \begin{cases} \geq 0 & \rightarrow +1 \\ < 0 & \rightarrow -1 \end{cases}$$

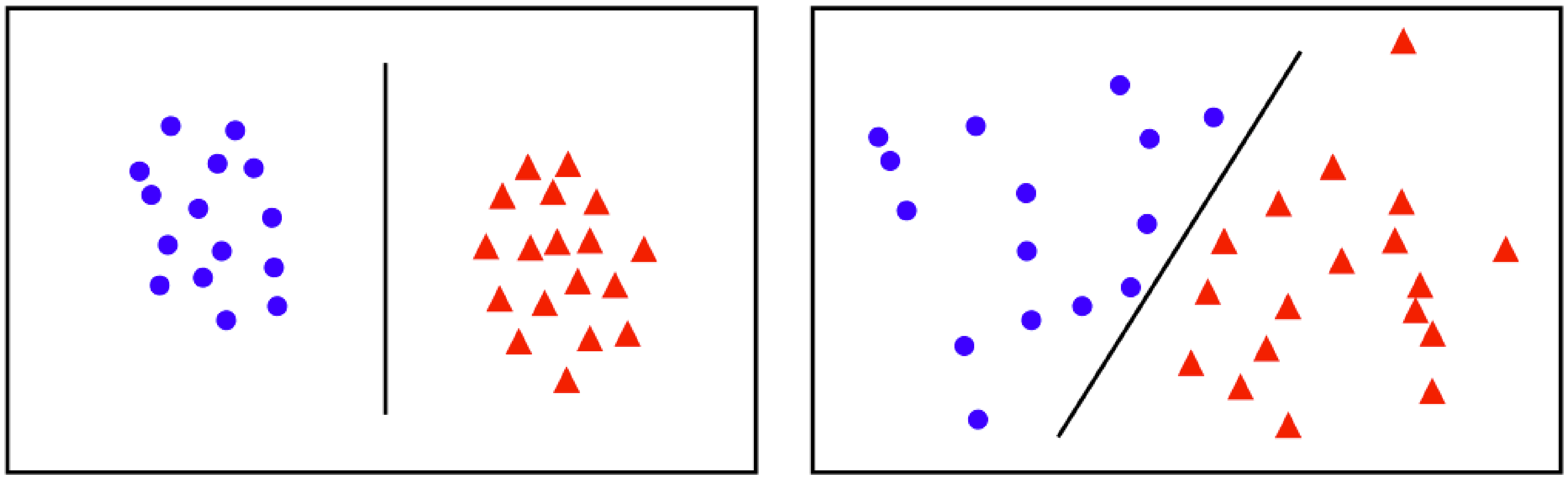
i.e. $\underset{\text{actual}}{\overset{+\;-}{y^{\{i\}}}}\;\underset{\text{predicted}}{\overset{+\;-}{f(x^{\{i\}})}} > 0$ for a $\boxed{\text{correct classification}}$.



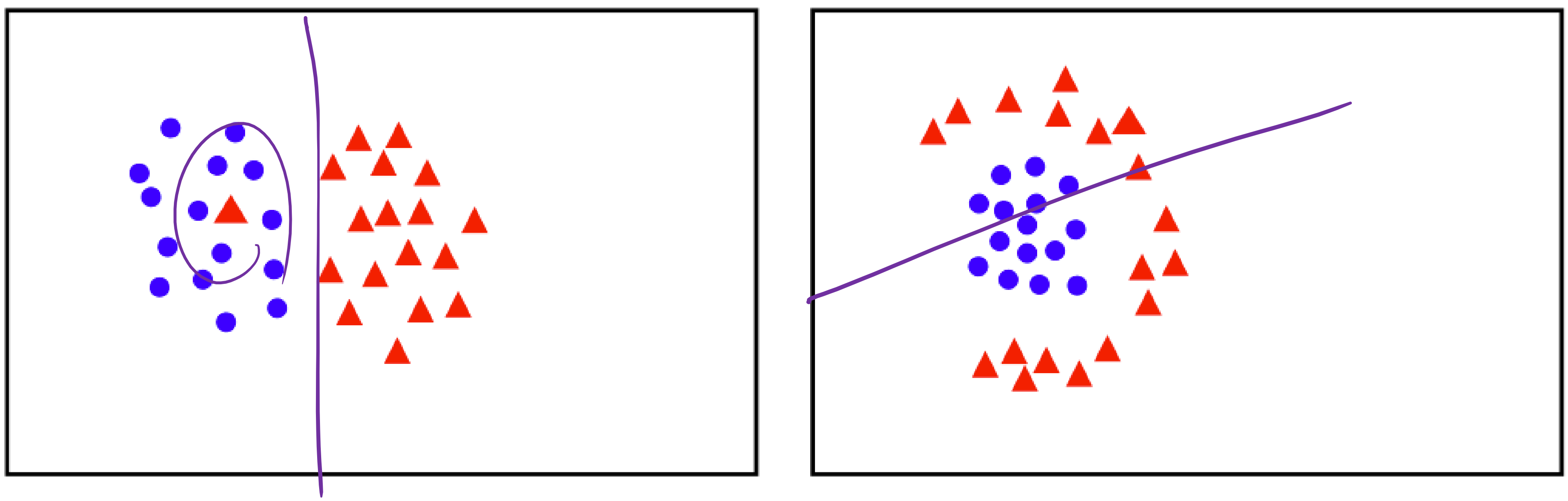


Linear Separability

linearly separable



not linearly separable



$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}_{d+1 \times 1}$$

Linear Classifier

vector

$x \perp \theta$ $\qquad f(x) = 0 = \underbrace{\overset{\text{vector}}{x}\overset{\text{vector}}{\theta}}_{1\times1} = 0 \in \mathbb{R}$

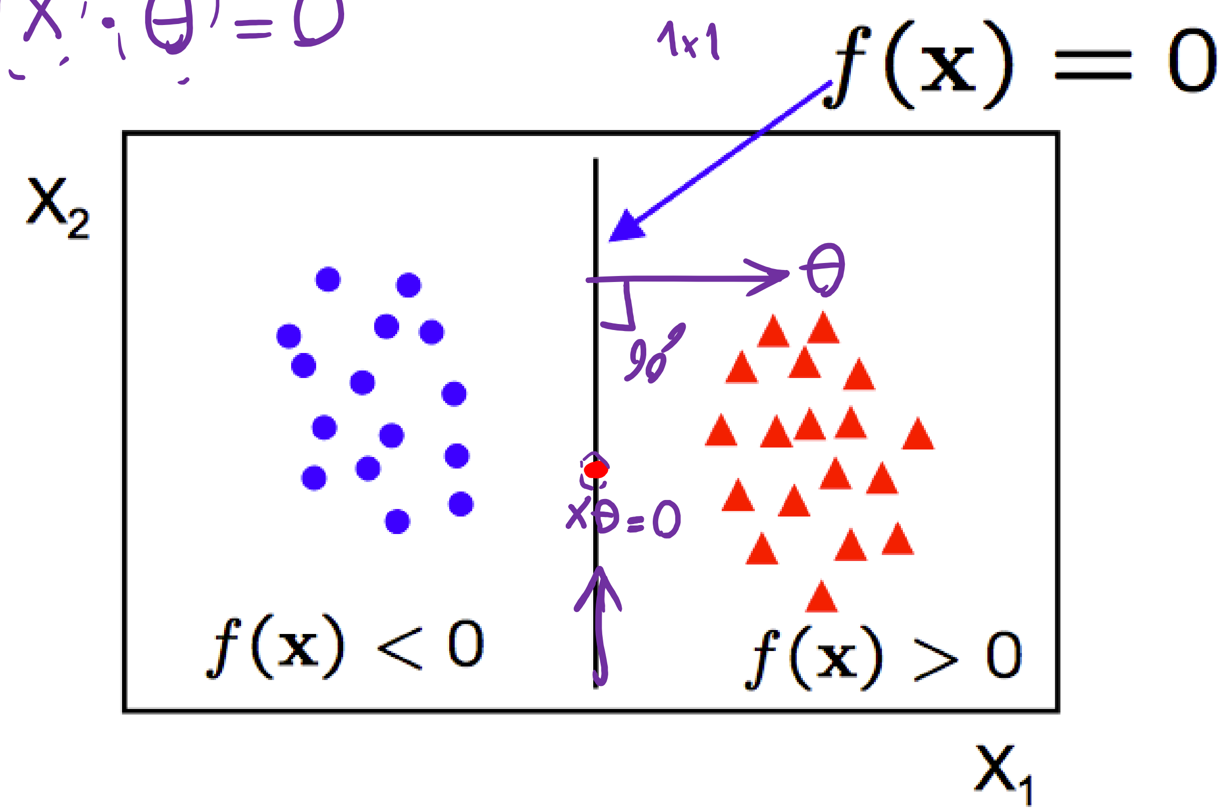
$1\times d \quad d\times 1$

$x \cdot \theta = 0$

A linear classifier has the form

rotation $\qquad$ bias term

$$f(x) = x\theta + \theta_0$$

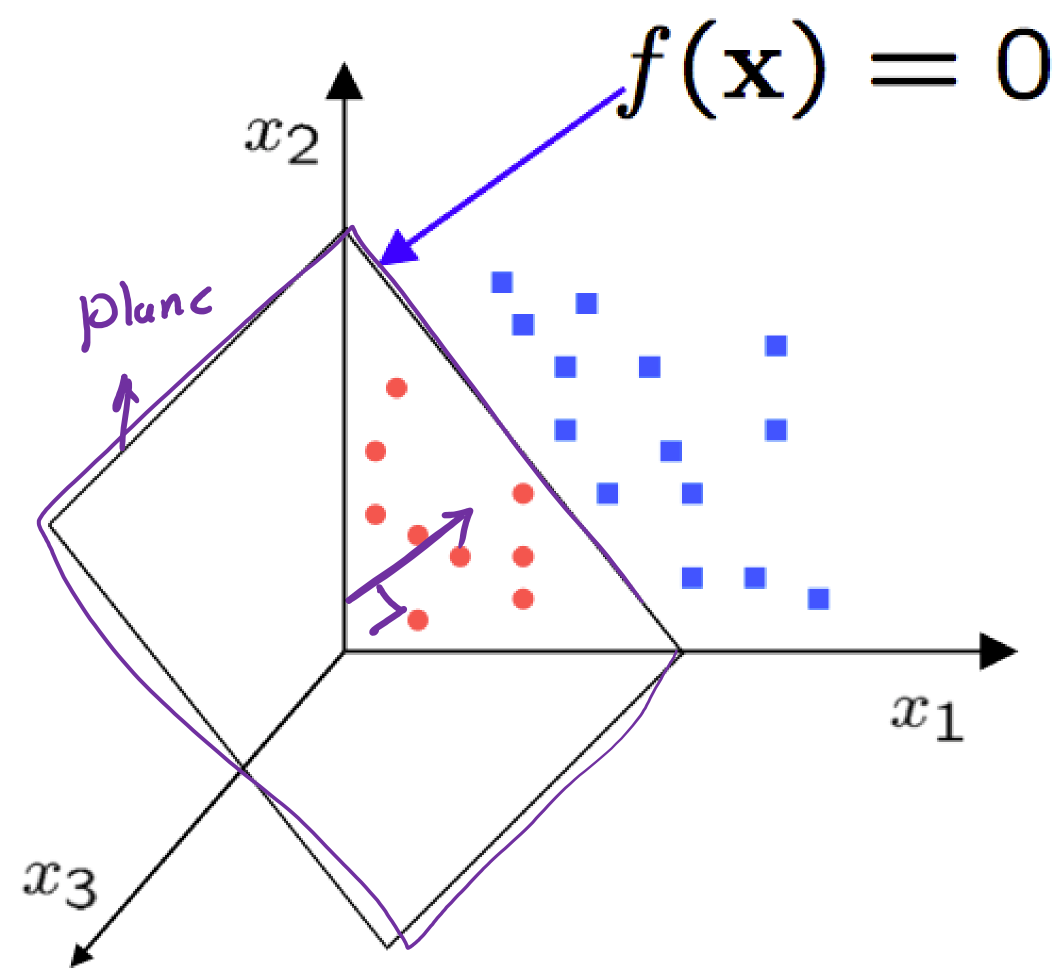


- in 2D the discriminant is a line
- θ is the normal to the line, and θ_0 the bias
- θ is known as the weight vector

Linear Classifier (higher dimension)

A linear classifier has the form

$$f(x) = x\theta + \theta_0$$



- in 3D the discriminant is a plane, and in nD it is a hyperplane

$x_1 \; x_2 \; \dots \; x_d$

$X = \begin{bmatrix} \; \end{bmatrix}$

$f(x^{\{i\}}) = \theta_0 + \theta_1 x_1 + \dots + \theta_d x_d \in \mathbb{R}$

The Perceptron Classifier

Considering $\boldsymbol{x}$ is linearly separable and $\boldsymbol{y}$ has two labels of $\{-1,1\}$

predicted $f(x^{\{i\}})$

$y_a \in \{+1, -1\}$

actual

$$f(x^{\{i\}}) = x^{\{i\}}\theta$$ Bias is inside θ now

How can we separate datapoints with label 1 from datapoints with label -1 using a line?

$X^{\{i\}} \rightarrow$ misclassified

$(-)373.5$

$f(x^{\{i\}})\, y_a < 0$ $\quad (+1)$

$f(x^{\{i\}}) = x^{\{i\}}\theta$

Misclassified

Ex. $y^{\{i\}} f\left(x^{\{i\}}\right) < 0$

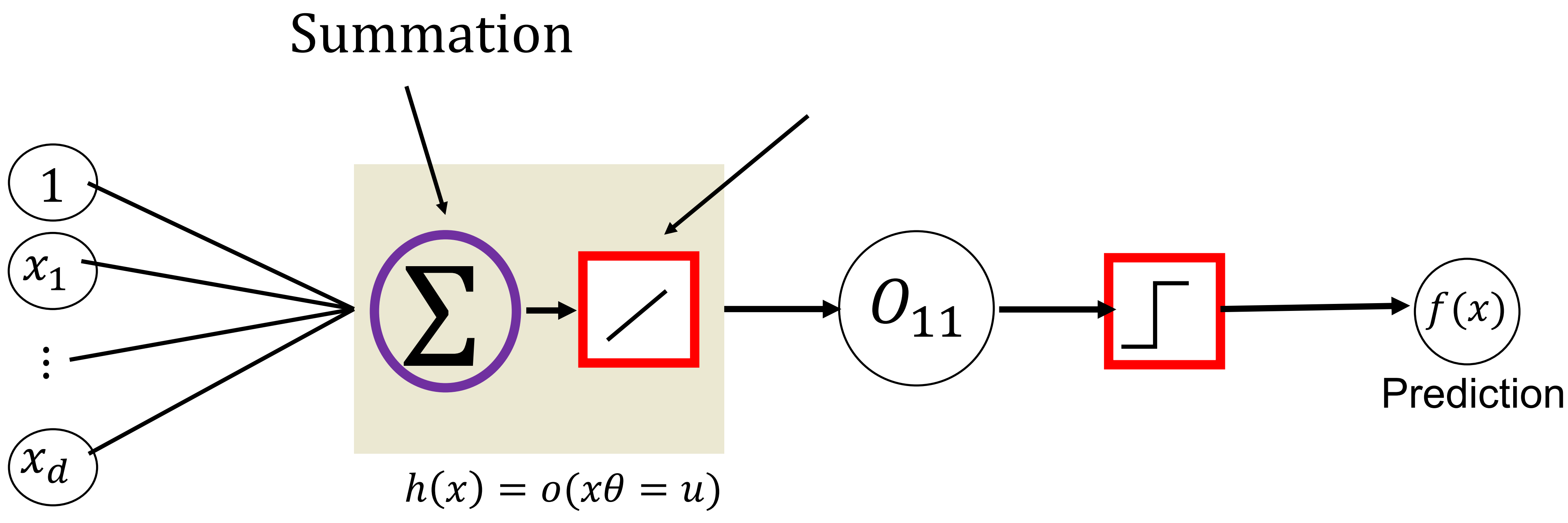
actual predicted

Perceptron Algorithm:

random

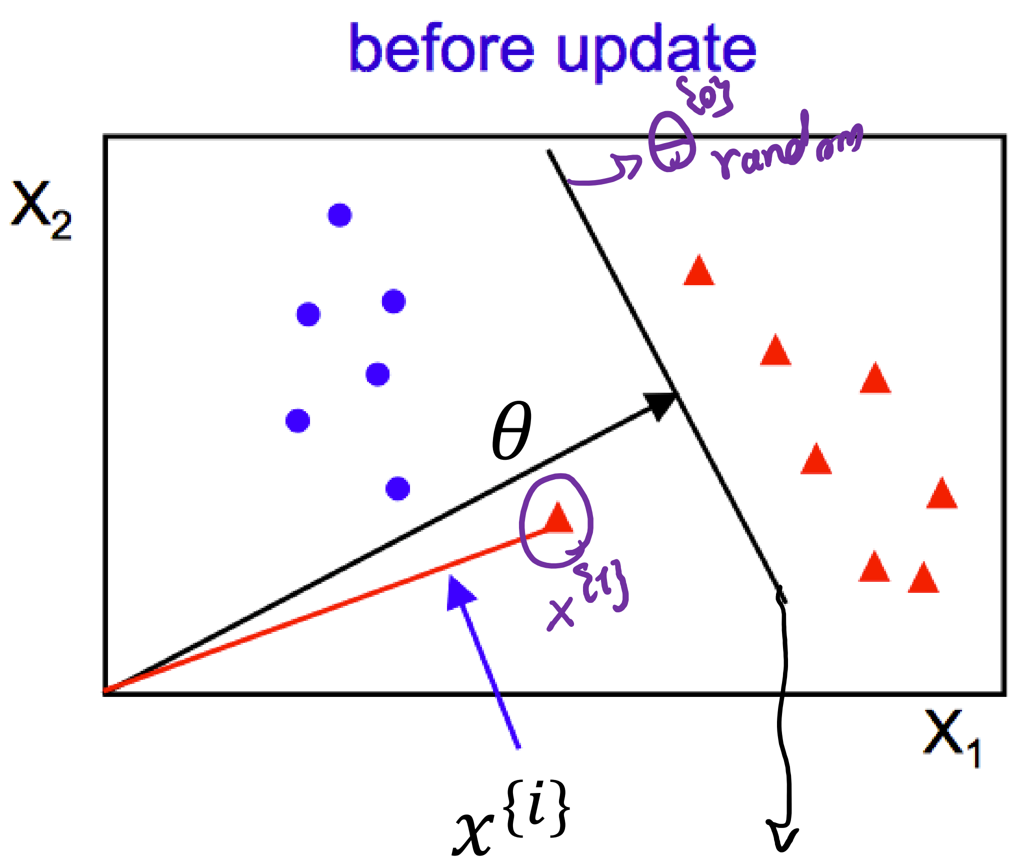
- Initialize θ =0
- Go through each datapoint $\{x^{\{i\}}, y^{\{i\}}\}$
  - If $x^{\{i\}}$ is misclassified then $\theta^{t+1} \leftarrow \theta^t + \alpha y^{\{i\}} x^{\{i\}}$

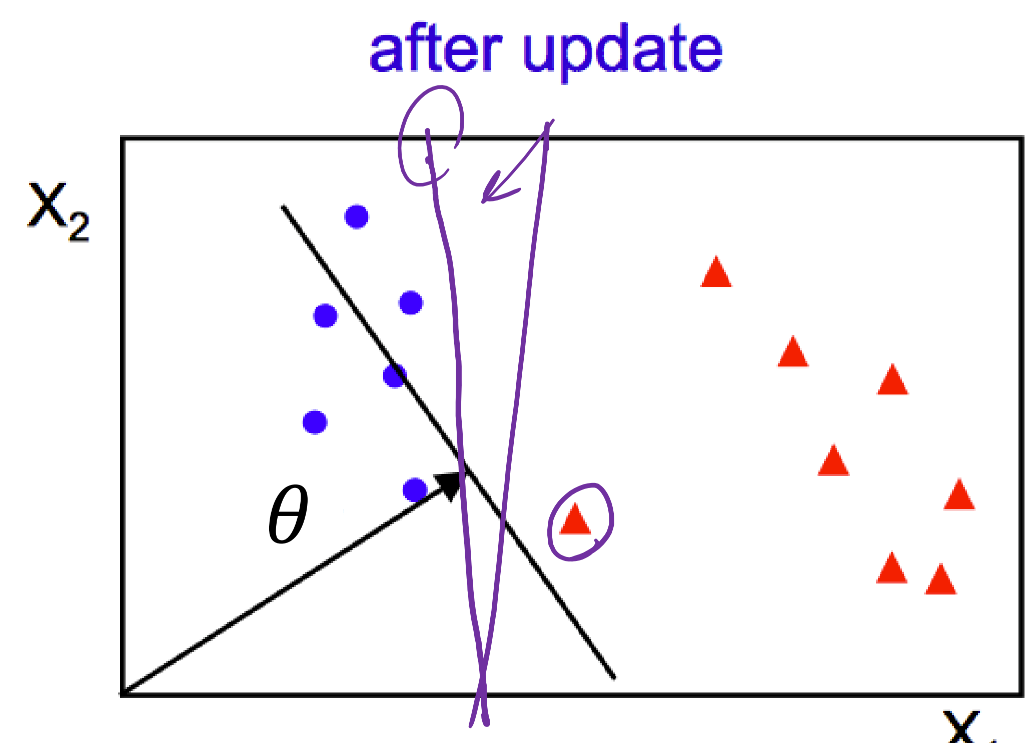
    learning rate
- Until all datapoints are correctly classified


Summation
1
x_1
x_d
O_{11}
$f(x)$
Prediction
$h(x) = o(x\theta = u)$

- Initialize θ =0
- Go through each datapoint $\{x^{\{i\}}, y^{\{i\}}\}$
    - If $x^{\{i\}}$ is misclassified then $\theta^{t+1} \leftarrow \theta^{t} + \alpha y^{\{i\}} x^{\{i\}}$

- Until all datapoints are correctly classified



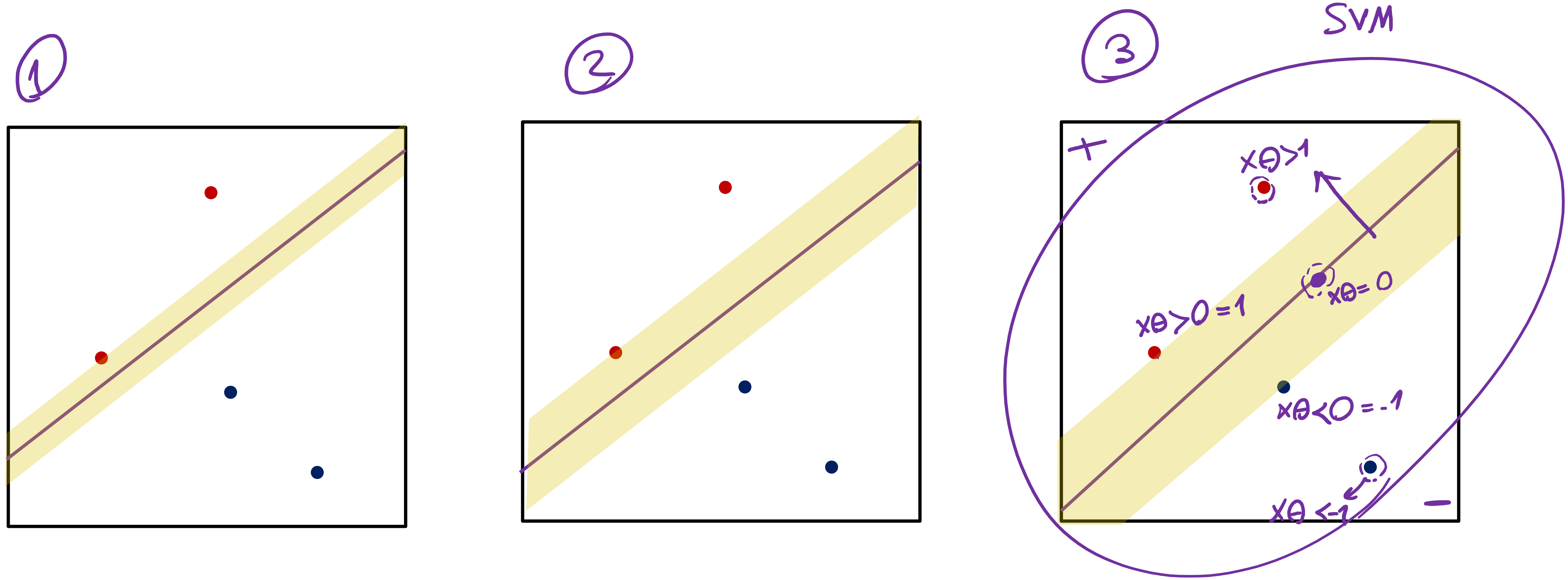


$\theta^{t+1} \leftarrow \theta^{t} - \alpha x^{\{i\}}$

$f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

Linear separation

We can have different separating lines



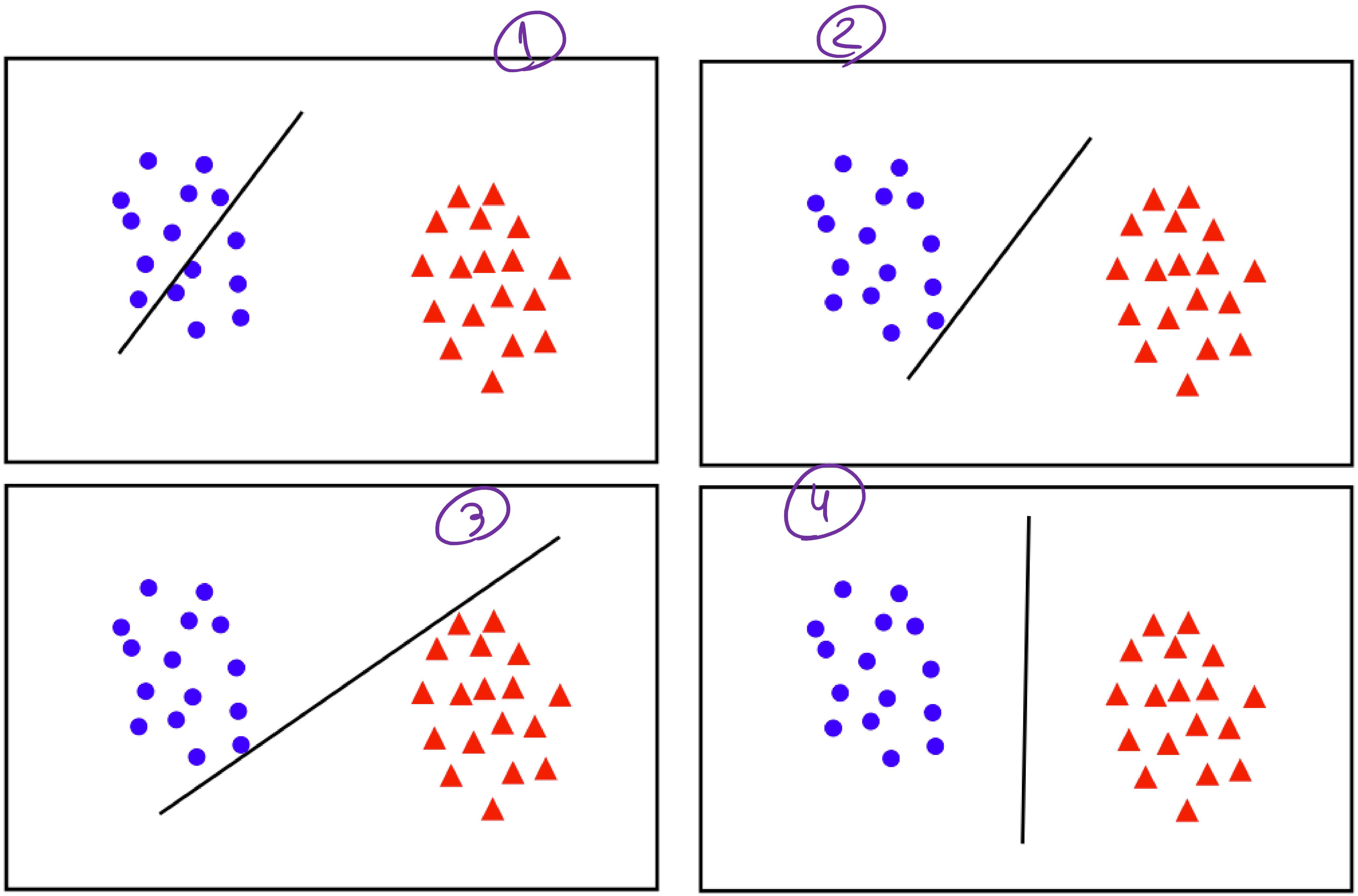
Which line is the best?

Why is the bigger margin better?

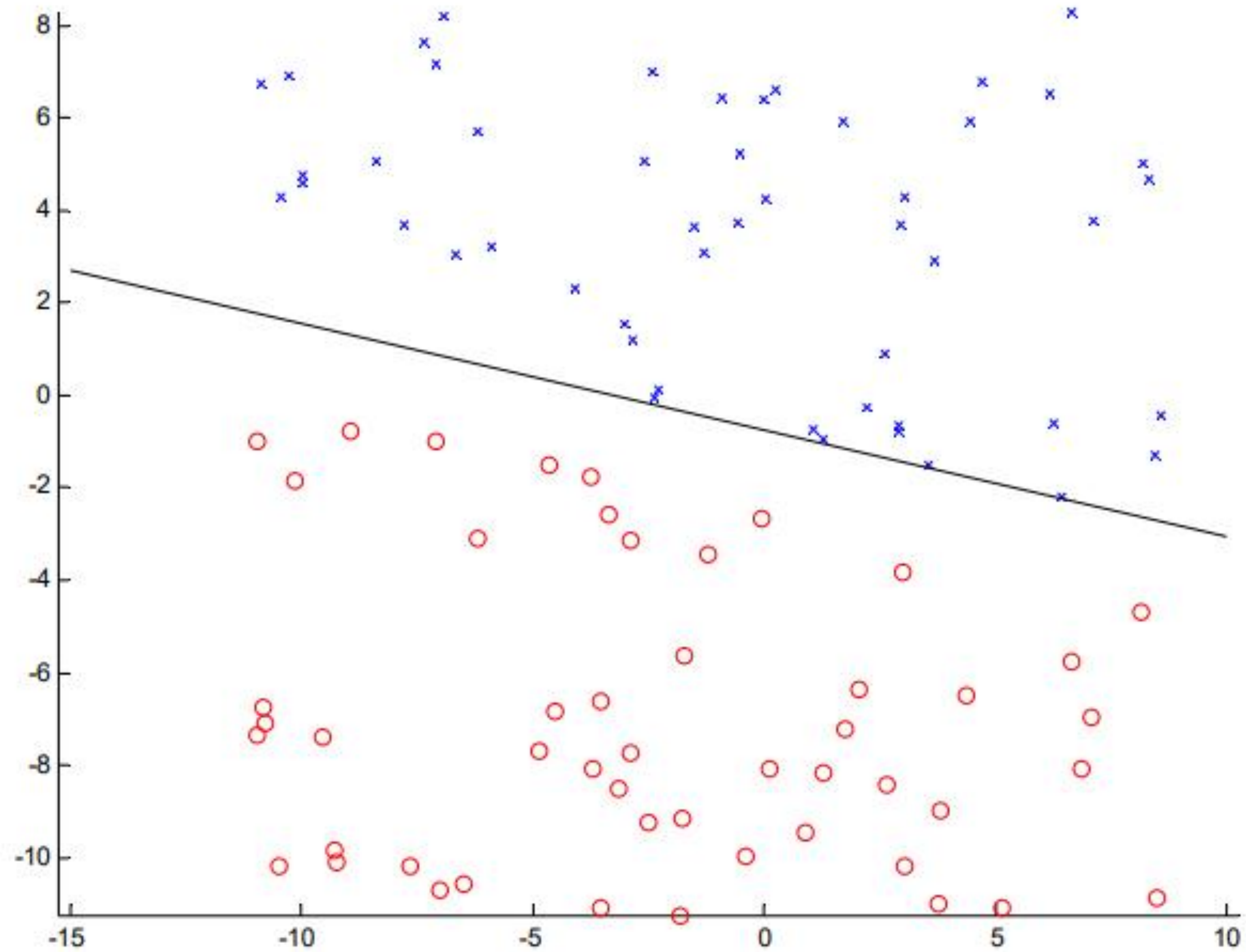
What θ maximizes the margin?

All cases, error is zero and they are linear, so they are all good for generalization.

What is the Best θ?



Perceptron example



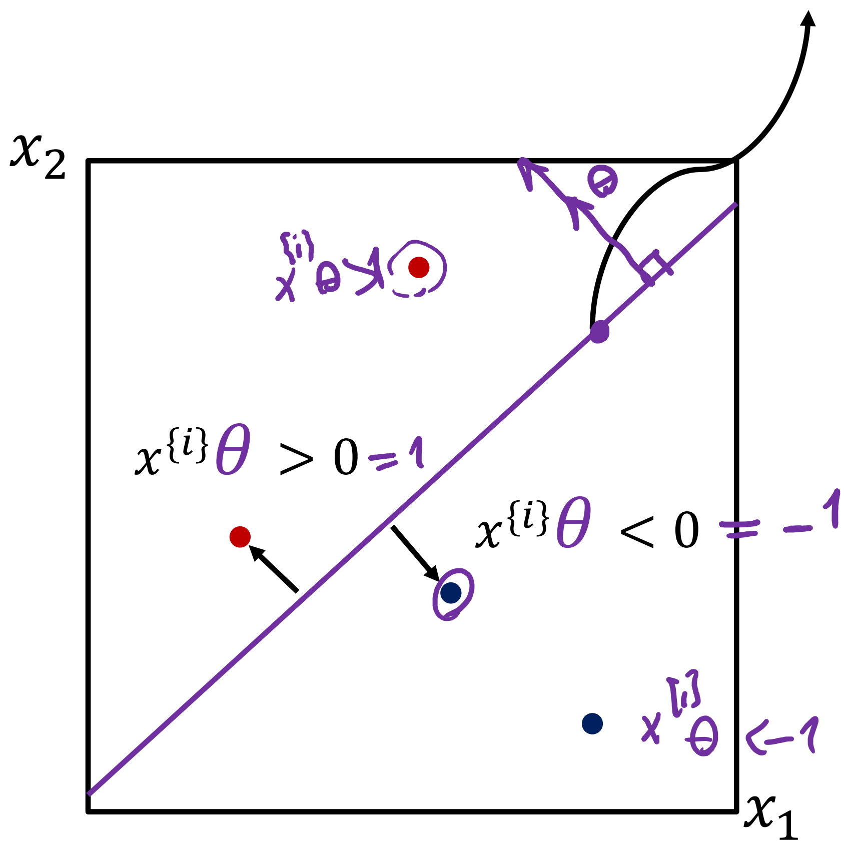
- if the data is linearly separable, then the algorithm will converge
- convergence can be slow …
- separating line close to training data
- we would prefer a larger margin for generalization (better generalization)

Outline

- Precursor: Linear Classifier and Perceptron
- Support Vector Machine
- Parameter Learning

Finding θ with a **fat** margin

Solution (decision boundary) of the line: $x\theta = 0$



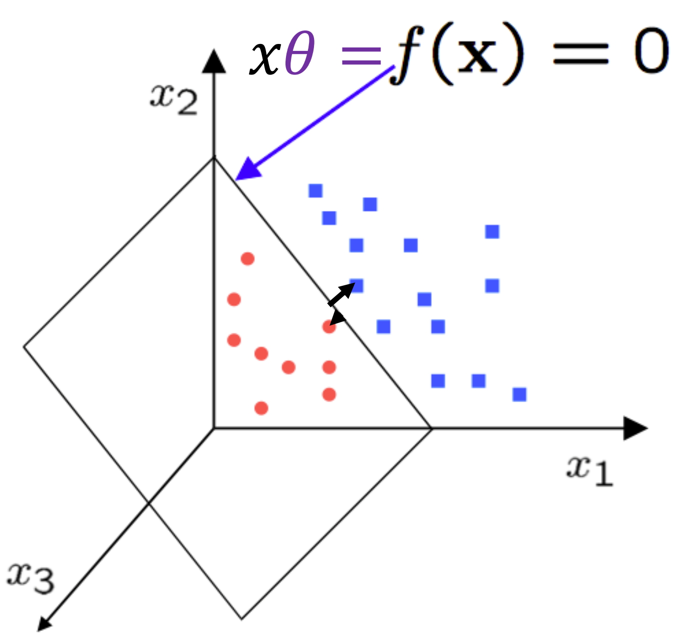
Let $\boldsymbol{x}^{\{i\}}$ to be the nearest data point to the line (plane):

$$\left|x^{\{i\}}\theta\right| > 0$$

Our line solution is $x\theta = 0$

Does it matter if I scale up or down θ for the decision boundary?

$$\left|x^{\{i\}}\theta\right| = 1 \rightarrow \text{normalization}$$



Let's pull out θ_0 from $\theta = (\theta_1, \ldots, \theta_d)$ and call it be b

Decision boundary would be: $\boxed{x\theta + b = 0}$

Computing the distance

The distance between $\boldsymbol{x}^{\{i\}}$ and the line $x\theta + b = 0$ where $\left|x^{\{i\}}\theta + b\right| = 1$

The vector θ is perpendicular to the decision line.

You should ask me why?

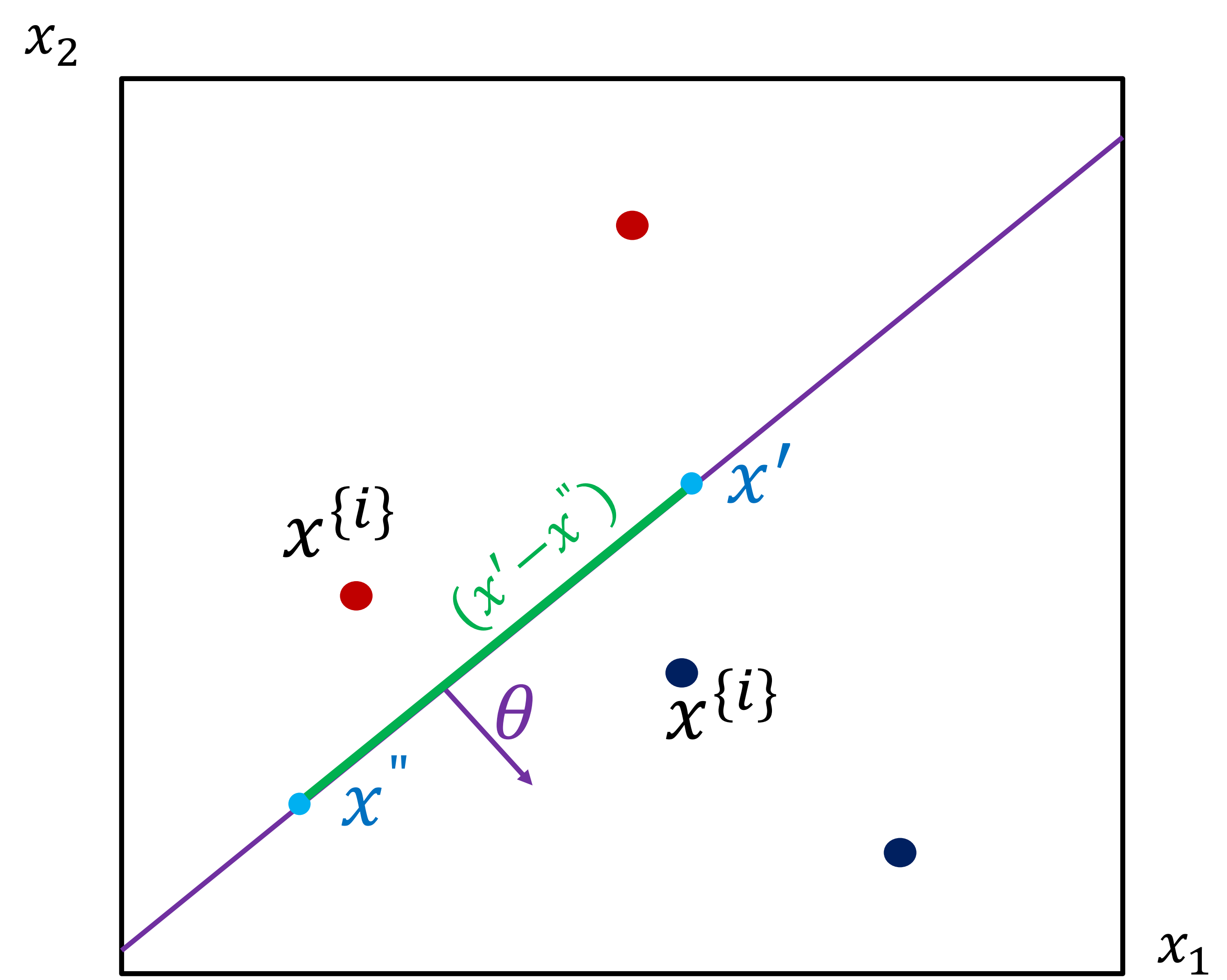
Consider x' and x'' on the plane

$$x'\theta + b = 0 \quad \text{and} \quad x''\theta + b = 0$$

$$\Downarrow$$

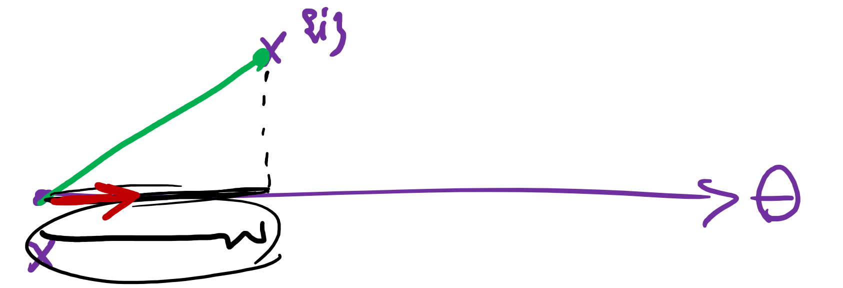
$$x'\theta + b = x''\theta + b$$

$$\Rightarrow \quad (x' - x'')\theta = 0$$



What is the distance of my fat margin?

What is the distance between $x^{\{i\}}$ and the plane?



Let's take any point x on the line:

$(x^{\{i\}} - x) \cdot \frac{\theta}{||\theta||} = (x^{\{i\}} - x)\hat{\theta}$

Distance would be projection of $(x^{\{i\}} - x)$ vector on θ.

To project the vector, we need to normalize θ to get the unit vector. $\frac{1}{||\theta||}$ $\frac{2}{||\theta||}$

$\hat{\theta} = \frac{\theta}{||\theta||} \Rightarrow \text{distance} = \left| (x^{\{i\}} - x)\hat{\theta} \right|$ which is the dot product x^2 $2x^2$

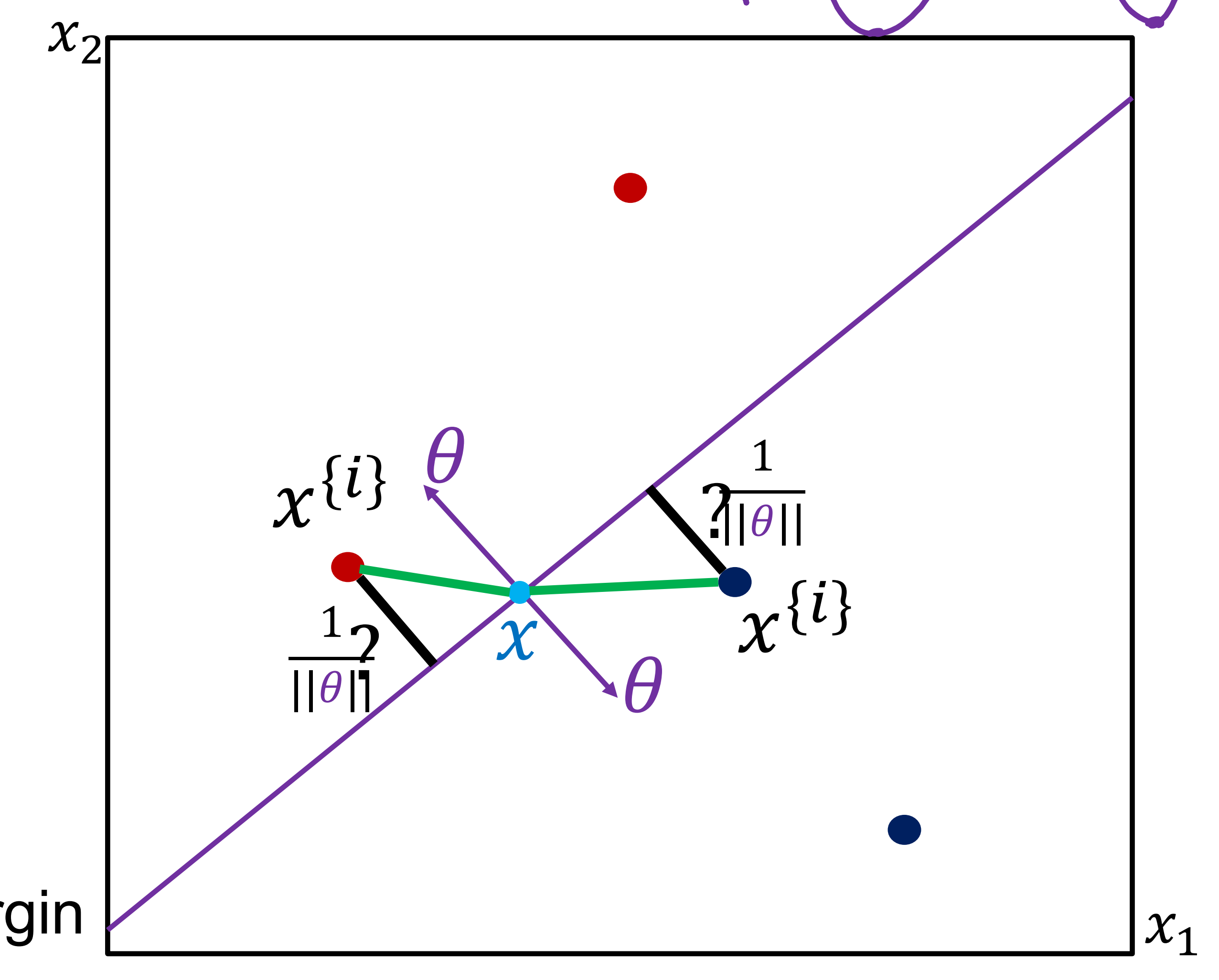
$$\text{distance} = \frac{1}{||\theta||} \left| (x^{\{i\}}\theta - x\theta) \right|$$

$$= \frac{1}{||\theta||} \left| (\underbrace{\overset{1}{x^{\{i\}}\theta + b}}_{\text{My constraint}} \underbrace{\overbrace{- x\theta - b}^{0}}_{\text{A point on the decision line}}) \right| = \frac{1}{||\theta||}$$

My constraint $\left| x^{\{i\}}\theta + b \right| = 1$

A point on the decision line $x\theta + b = 0$

The margin



Now we need to maximize the margin

$f(x) = |x|$

Maximize $\dfrac{1}{||\theta||}$

Subject to $\underset{i=1,2,\dots,N}{\text{Min value of}} \left|x^{\{i\}}\theta + b\right| = 1 \Rightarrow nearest\ neighbour$

There is a "min" in our constraining; it can be hard to optimize this problem(non-convex form)

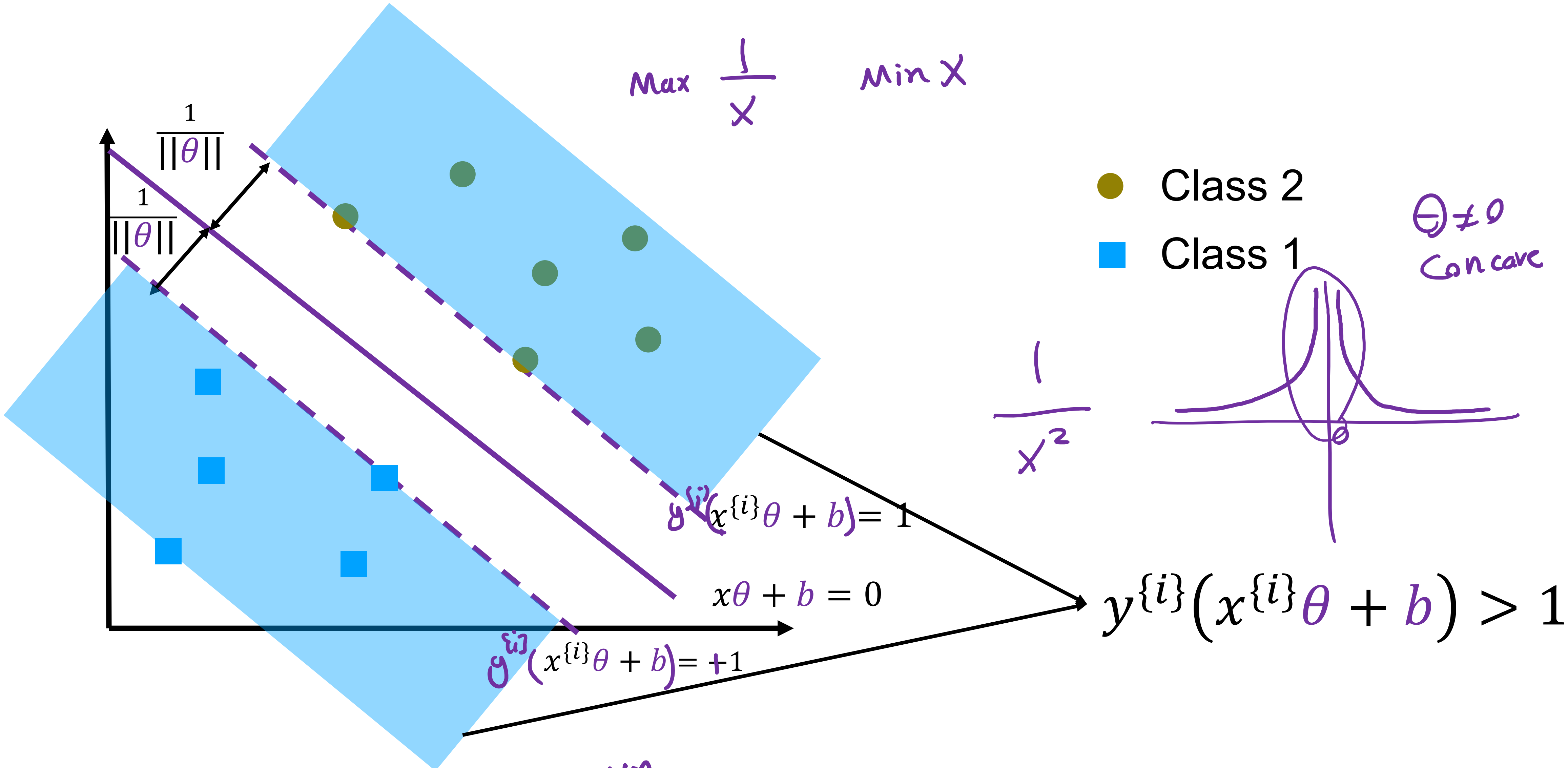
Can I write the following term to get rid of absolute value?

$$\left|x^{\{i\}}\theta + b\right| = y^{\{i\}}\left(x^{\{i\}}\theta + b\right) \Rightarrow \text{for a correct classification}$$

If min $\left|x^{\{i\}}\theta + b\right| = 1 \Rightarrow so\ it\ can\ be\ at\ least\ 1$

Maximize $\dfrac{1}{||\theta||}$

Subject to $y^{\{i\}}\left(x^{\{i\}}\theta + b\right) \geq 1$ for $i = 1,2,\dots,N$



Maximize $\frac{2}{||\theta||}$ $\quad \frac{1}{2}||\theta||$ ← min $= \theta\theta^T$

If $\theta \neq 0,$ there exists a max value

Subject to $y^{\{i\}}(x^{\{i\}}\theta + b) \geq 1$ for $i = 1,2,\ldots,N$

$\theta^{2?}$ LP

Minimize $\frac{1}{2}\theta\theta^T$ QP

linear

Subject to $y^{\{i\}}(x^{\{i\}}\theta + b) \geq 1$ for $i = 1,2,\ldots,N$

Constrained optimization

Minimize $\frac{1}{2}\theta\theta^T$ $S_1, S_2, ..., S_n$

$\alpha_1 \quad \alpha_2 \quad \alpha_n$

$\beta_1 \quad \beta_2 \quad \beta_n$

Subject to $y^{\{i\}}(x^{\{i\}}\theta + b) \geq 1$ for $i = 1,2,\ldots,N$

$\theta \in \mathbb{R}^d, b \in \mathbb{R}$

Using Lagrange method: But wait, there is an **inequality** in our constraints

We use **Karush-Kuhn-Tucker (KKT)** condition to deal with this problem

$g(x) = y^{\{i\}}(x^{\{i\}}\theta + b) - 1 \geq 0$ $\alpha = lagrange\ multiplier$

$y^{\{i\}}(x^{\{i\}}\theta + b) \geq 1$

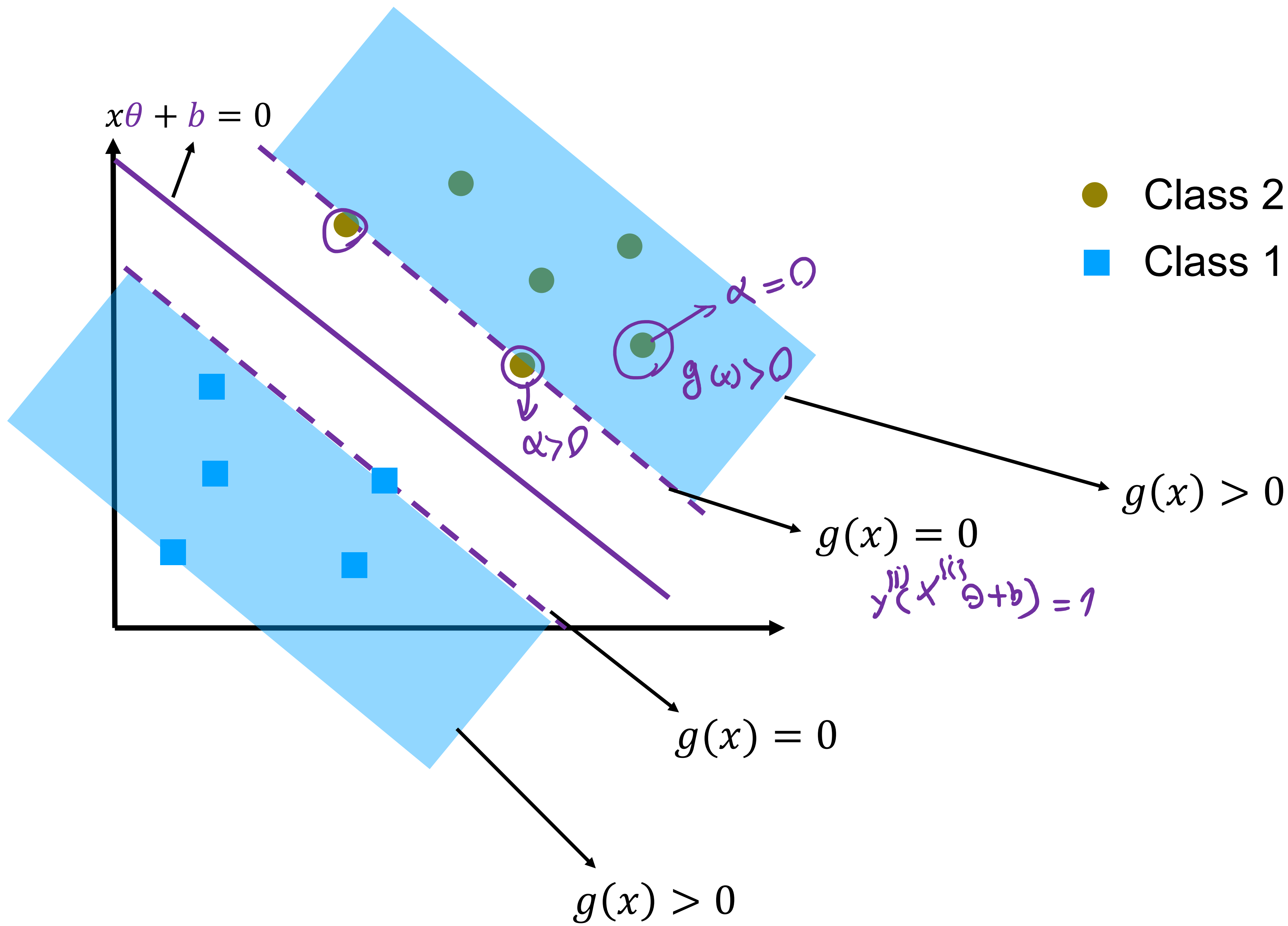
We need to optimize $\theta, b, and\ \alpha$

1) Stationary

2) $g(x) \geq 0$ Primal feasibility

3) $\alpha \geq 0$ Dual feasibility

4) $g(x)\alpha = 0$ Complementary slackness $\Rightarrow \begin{cases} g(x) > 0, & \alpha = 0 \\ \alpha > 0, & g(x) = 0 \end{cases}$



$$g(x) = y^{\{i\}}\left(x^{\{i\}}\theta + b\right) - 1$$

3) $g(x)\alpha = 0$ Complementary slackness $\Rightarrow \begin{cases} g(x) > 0, & \alpha = 0 \\ \alpha > 0, & g(x) = 0 \end{cases}$

Lagrange formulation

$L(\theta, S) = f(x) \pm S\, g(x)$

Minimize $\underbrace{\frac{1}{2}\theta\theta^T}_{f(x)}$ s.t. $\underbrace{y^{\{i\}}(x^{\{i\}}\theta + b) - 1 \geq 0}_{g(x)}$

Max, dual $L(\alpha)$

$$\underset{\text{Min}}{\mathcal{L}(\theta, b, \alpha)} = \frac{1}{2}\theta\theta^T - \sum_{i=1}^{N} \overset{+}{\alpha^{\{i\}}} \overbrace{(y^{\{i\}}(x^{\{i\}}\theta + b) - 1)}^{+}$$

primal

Minimize w.r.t θ and b and maximize *w.r.t* each $\alpha^{\{i\}} \geq 0$

$$\theta = \sum_{i=1}^{N} \alpha^{\{i\}} y^{\{i\}} x^{\{i\}}$$

$$\nabla_\theta \mathcal{L}(\theta, b, \alpha) = \theta - \sum_{i=1}^{N} \alpha^{\{i\}} y^{\{i\}} x^{\{i\}} = 0$$

$$\nabla_b \mathcal{L}(\theta, b, \alpha) = -\sum_{i=1}^{N} \alpha^{\{i\}} y^{\{i\}} = 0$$

constraint

$$\boxed{\theta = \sum_{i=1}^{N} \alpha^{\{i\}} y^{\{i\}} x^{\{i\}}}$$

$$\sum_{i=1}^{N} \alpha^{\{i\}} y^{\{i\}} = 0$$

$$\theta^T_{d\times 1}$$

$$-\underset{1\times d}{\theta}\sum_{i=1}^{N}\alpha^{\{i\}}y^{\{i\}}x^{\{i\}} - b\sum_{i=1}^{N}\alpha^{\{i\}}y^{\{i\}} \to 0$$

Let’s substitute these in the Lagrangian:

$$-\left(\sum\alpha^{\{i\}}y^{\{i\}}x^{\{i\}}\theta + \sum\alpha^{\{i\}}y^{\{i\}}b - \sum\alpha^{\{i\}}\right)$$

$\in \mathbb{R}$

$$\mathcal{L}(\theta, b, \alpha) = \frac{1}{2}\theta\theta^T - \sum_{i=1}^{N} \alpha^{\{i\}}\left(y^{\{i\}}\left(x^{\{i\}}\theta + b\right) \cancel{-1}\right)$$

$$\mathcal{L}(\theta, b, \alpha) = \sum_{i=1}^{N} \alpha^{\{i\}} + \frac{1}{2}\theta\theta^T - \sum_{i=1}^{N} \alpha^{\{i\}}\left(y^{\{i\}}\left(x^{\{i\}}\theta \cancel{+b}\right)\right)$$

$$\mathcal{L}(\theta, b, \alpha) = \sum_{i=1}^{N} \alpha^{\{i\}} + \frac{1}{2}\theta\theta^T - \sum_{i=1}^{N} \alpha^{\{i\}}\left(y^{\{i\}}\left(x^{\{i\}}\theta\right)\right) = \sum_{i=1}^{N} \alpha^{\{i\}} + \frac{1}{2}\theta\theta^T - \theta\theta^T =$$

$$L(\theta, b, \alpha) = \sum_{i=1}^{N} \alpha^{\{i\}} - \frac{1}{2}\theta\theta^T$$

primal form $\to L(\alpha)$

$$\theta = \sum_{i=1}^{N} \alpha^{\{i\}} y^{\{i\}} x^{\{i\}}$$

$$\sum_{i=1}^{N} \alpha^{\{i\}} y^{\{i\}} = 0$$

$$\mathcal{L}(\theta, b, \alpha) = \sum_{i=1}^{N} \alpha^{\{i\}} - \frac{1}{2} \theta \theta^{T}$$

Max

$$\mathcal{L}(\cancel{\theta, b}, \alpha) = \sum_{i=1}^{N} \alpha^{\{i\}} - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \overbrace{\overset{1\times1}{y^{\{i\}}} \overset{1\times1}{y^{\{j\}}} \overset{1\times1}{\alpha^{\{i\}}} \overset{1\times1}{\alpha^{\{j\}}} \overset{1\times d}{x^{\{i\}}} \overset{d\times1}{x^{\{j\}T}}}^{1\times1}$$

maximize $w.r.t$ each $\alpha^{\{i\}} \geq 0$ for $i = 1, \ldots, N$

and

$\nabla_b = 0$

$$\sum_{i=1}^{N} \alpha^{\{i\}} y^{\{i\}} = 0$$

The solution – quadratic programming

$X = \begin{bmatrix} \quad \end{bmatrix} \qquad y = \begin{bmatrix} \quad \end{bmatrix}$

$$\max_{\alpha} \sum_{i=1}^{N} \alpha^{\{i\}} - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y^{\{i\}} y^{\{j\}} \alpha^{\{i\}} \alpha^{\{j\}} x^{\{i\}} x^{\{j\}^T}$$

Quadratic programming packages usually use “min”

$$\alpha = \underset{n\times 1}{\begin{bmatrix} \alpha^{\{1\}} \\ \alpha^{\{2\}} \\ \vdots \\ \alpha^{\{n\}} \end{bmatrix}}$$

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y^{\{i\}} y^{\{j\}} \alpha^{\{i\}} \alpha^{\{j\}} x^{\{i\}} x^{\{j\}^T} - \sum_{i=1}^{N} \alpha^{\{i\}} \in \mathbb{R}$$

$\alpha^{\{1\}} + \dots + \alpha^{\{n\}} \in \mathbb{R}$

1×1

$$\underbrace{\min_{\alpha} \frac{1}{2} \underset{1\times N}{\alpha^T} \underset{N\times N}{\begin{bmatrix} \overbrace{\overset{1\times 1}{y^{\{1\}}} \overset{1\times 1}{y^{\{1\}}} \overset{1\times d}{x^{\{1\}}} \overset{d\times 1}{x^{\{1\}^T}}}^{1\times 1} & \cdots & y^{\{1\}} y^{\{N\}} x^{\{1\}} x^{\{N\}^T} \\ \vdots & \ddots & \vdots \\ y^{\{N\}} y^{\{1\}} x^{\{N\}} x^{\{1\}^T} & \cdots & y^{\{N\}} y^{\{N\}} x^{\{N\}} x^{\{N\}^T} \end{bmatrix}} \underset{N\times 1}{\alpha} + (-I^T)\alpha}_{1\times 1}$$

$$\min_{\alpha} \frac{1}{2}\alpha^T \underbrace{\begin{bmatrix} y^{\{1\}}y^{\{1\}}x^{\{1\}}x^{\{1\}^T} & \cdots & y^{\{1\}}y^{\{N\}}x^{\{1\}}x^{\{N\}^T} \\ \vdots & \ddots & \vdots \\ y^{\{N\}}y^{\{1\}}x^{\{N\}}x^{\{1\}^T} & \cdots & y^{\{N\}}y^{\{N\}}x^{\{N\}}x^{\{N\}^T} \end{bmatrix}}_{\text{Quadratic coefficients}} \alpha + \underbrace{(-I^T)}_{\text{Linear term}}\alpha$$

QP

$f(x,y) = x^2 + y^2$

Linear

Subject to $\underbrace{\sum_{i=1}^{N} \alpha^{\{i\}} y^{\{i\}} = y^T\alpha = 0}_{\text{Linear equality constraint}}$

$\nabla b = 0$

Pass these to a quadratic programming package

$\alpha \geq 0 \rightsquigarrow$ Dual feasibility

$$lower\ bound(0) \leq \alpha \leq \text{upper bound}(\infty)$$

$$\min_{\alpha} \frac{1}{2}\alpha^T Q\alpha - 1^T\alpha \quad \text{subject to} \quad y^T\alpha = 0; \alpha \geq 0$$

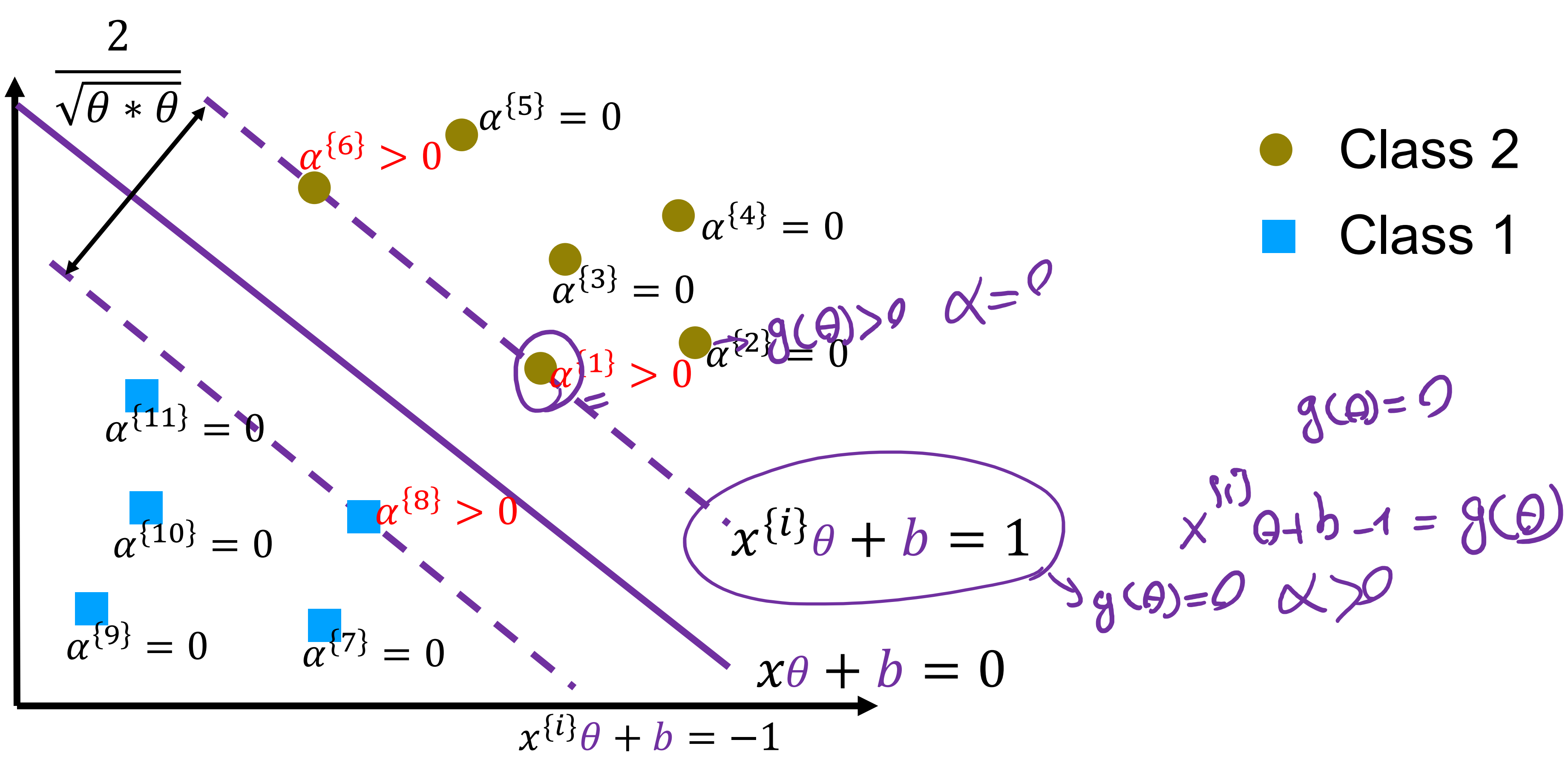
Quadratic programming will give us α

Solution: $\alpha = \alpha^{\{1\}}, \dots, \alpha^{\{N\}}$

KKT condition ($\alpha_i g_i(\theta) = 0$): $\alpha^{\{i\}}\left(y^{\{i\}}\left(x^{\{i\}}\theta + b\right) - 1\right) = 0$

$$\left(y^{\{i\}}\left(x^{\{i\}}\theta + b\right) - 1\right) > 0 \quad \Rightarrow \quad \alpha^{\{i\}} = 0$$

$$\left(y^{\{i\}}\left(x^{\{i\}}\theta + b\right) - 1\right) = 0 \quad \Rightarrow \quad \alpha^{\{i\}} > 0 \Rightarrow x_i \text{ is a support vector}$$



Training

$$\theta = \sum_{i=1}^{N} \alpha^{\{i\}} y^{\{i\}} x^{\{i\}}$$

No need to go over all datapoints

$$\rightarrow \theta = \sum_{x^{\{i\}} in\ SV} \alpha^{\{i\}} y^{\{i\}} x^{\{i\}}$$

and for b pick any support vector and calculate:

$$y^{\{i\}}\left(x^{\{i\}}\theta + b\right) = 1$$

Testing

For a new test point s

Compute:

$-$ -173.3 $+$ 272.3

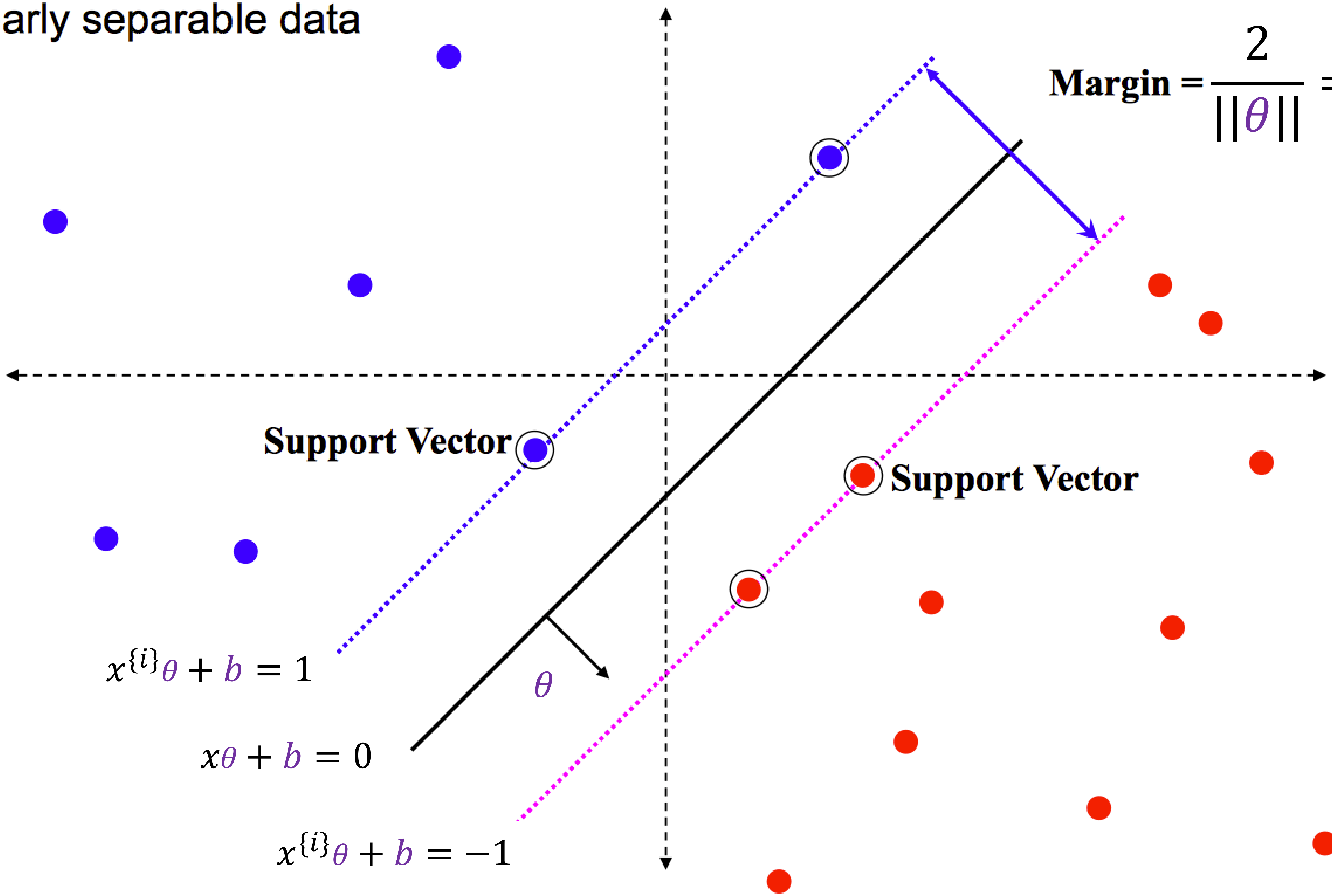
$$\underbrace{\text{s}\theta + \text{b}}_{\text{primal}} = \underbrace{\sum_{x^{\{i\}} in\ SV} \alpha^{\{i\}} y^{\{i\}} x^{\{i\}} s^T + b}_{\text{dual}}$$

Classify s as class 1 if the result is positive, and class 2 otherwise

Geometric Interpretation

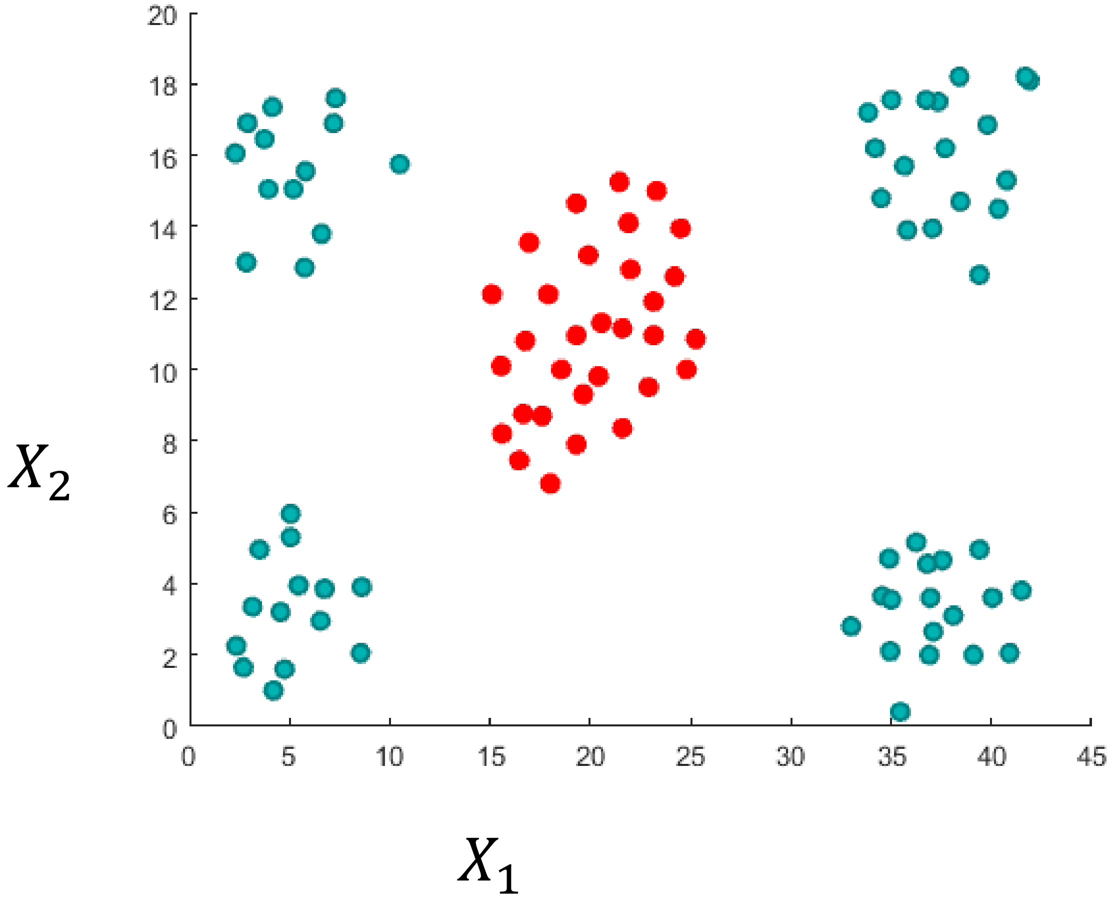
Hard svm

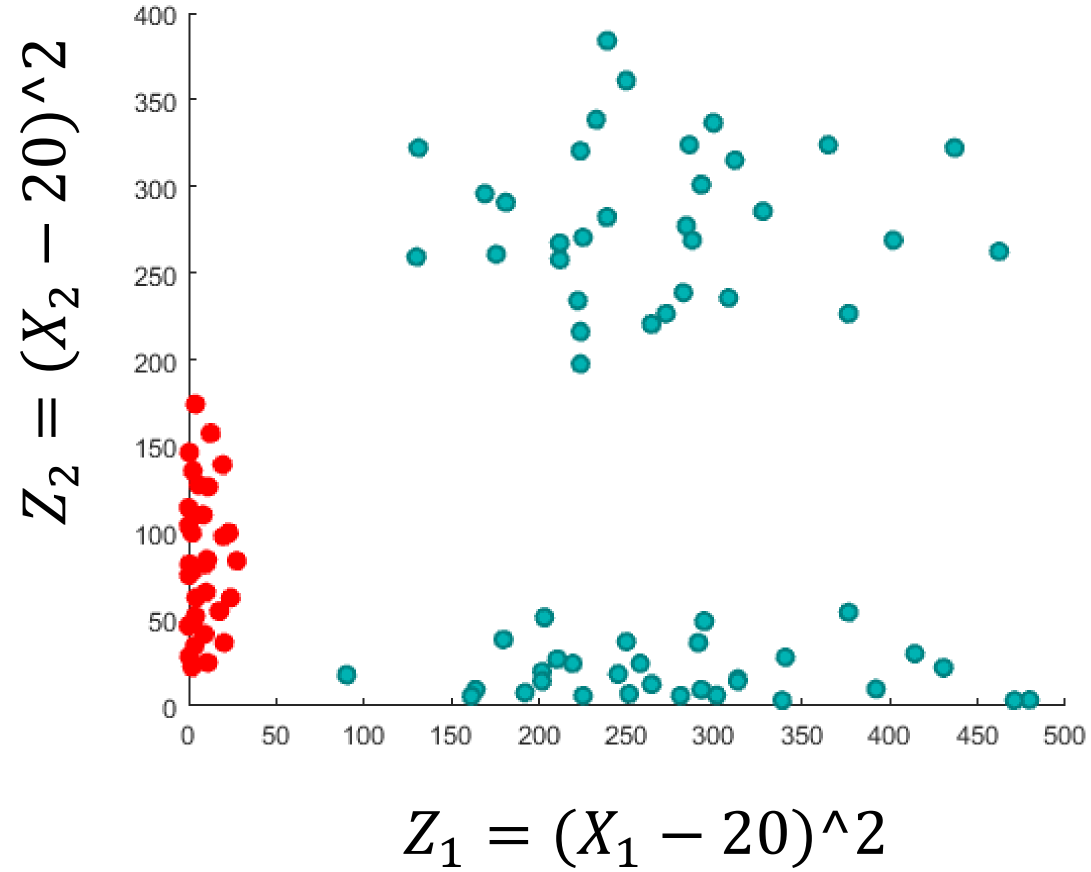
linearly separable data



$$\textbf{Margin} = \frac{2}{||\theta||} = \frac{2}{\sqrt{\theta * \theta}}$$

From x to z space





$$X \xrightarrow{\qquad\qquad} Z$$

$$(X_1^2, X_2^2)$$

$X_{n \times d}$ $\quad D >>> d$

$Z_{n \times D}$

In x space

$ZZ^T \rightarrow N \times N$

$XX^T \rightarrow N \times N$

$$\max_{\alpha} \sum_{i=1}^{N} \alpha^{\{i\}} - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y^{\{i\}} y^{\{j\}} \alpha^{\{i\}} \alpha^{\{j\}} x^{\{i\}} x^{\{j\}^T}$$

X_2

$$\max_{\alpha} \sum_{i=1}^{N} \alpha^{\{i\}} - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y^{\{i\}} y^{\{j\}} \alpha^{\{i\}} \alpha^{\{j\}} x^{\{i\}} x^{\{j\}^T}$$

X_1

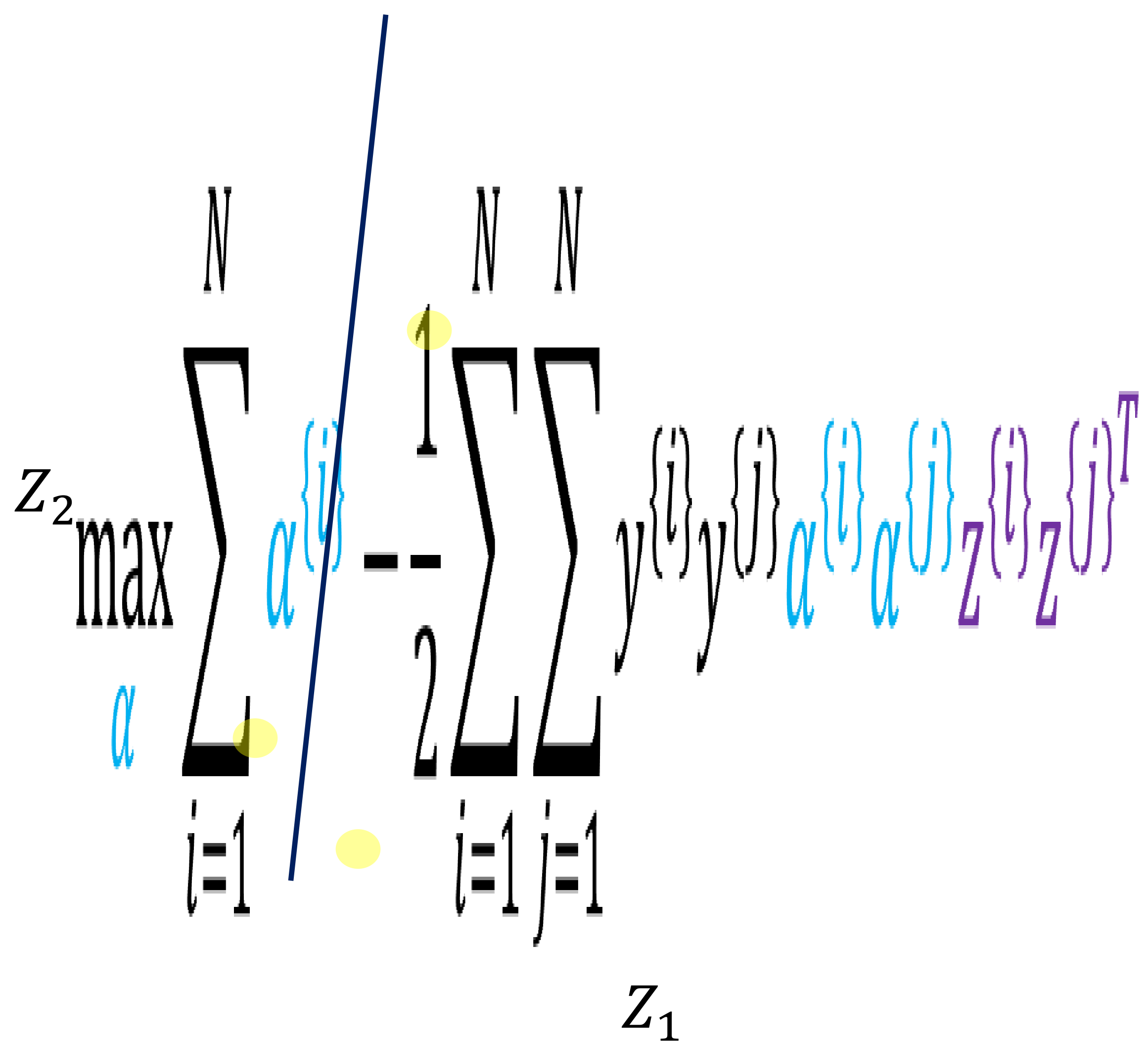
$let's\ say\ x\ is\ n \times d$

$x\mathrm{x}^{\mathrm{T}}$ will be $\mathrm{n} \times n$

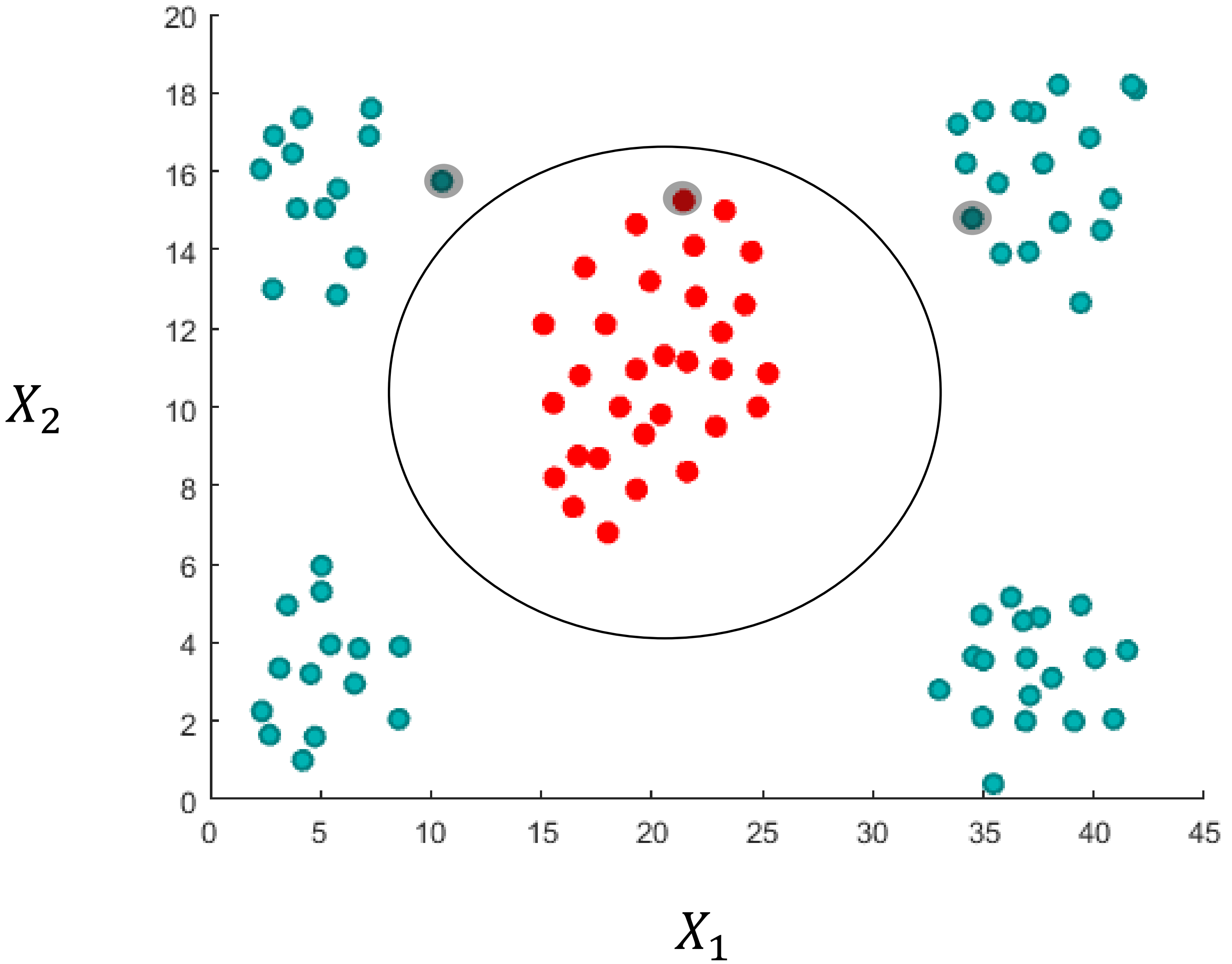
If I add millions of dimensions to x, would it affect the final size of $x\mathrm{x}^{\mathrm{T}}$?

In z space

$$\max_{\alpha} \sum_{i=1}^{N} \alpha^{\{i\}} - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y^{\{i\}} y^{\{j\}} \alpha^{\{i\}} \alpha^{\{j\}} z^{\{i\}} z^{\{j\}^T}$$



In x space, they are called pre-images of support vectors



Take-Home Messages

- Linear Separability
- Perceptron
- SVM: Geometric Intuition and Formulation