LECTURE 9

SPRING 2021
APPLIED MACHINE LEARNING
CIHANG XIE

SLIDE CREDIT:
DHRUV BATRA
ANDREW ZISSERMAN

TODAY

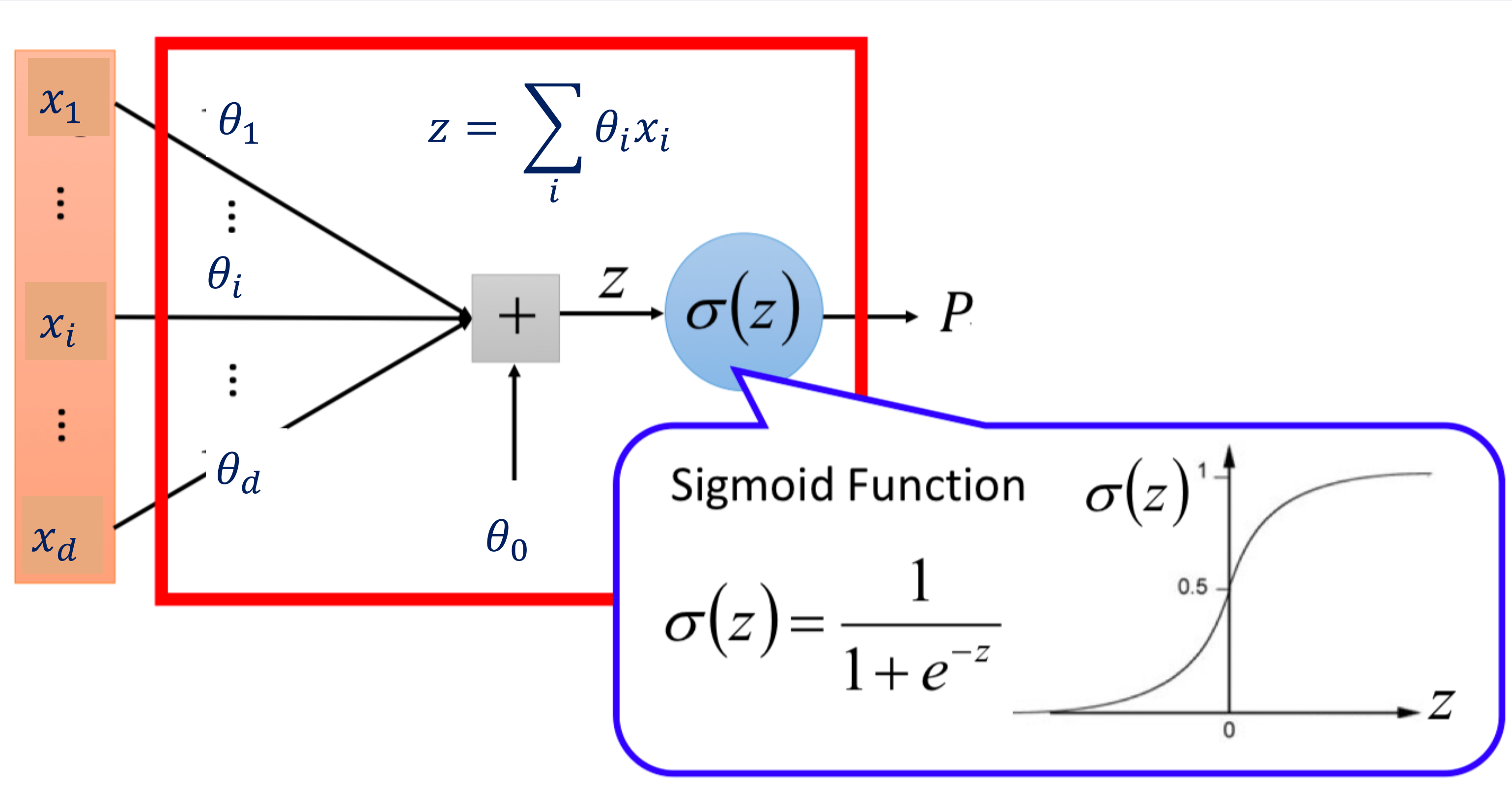
- Review of Logistic Regression
- Support Vector Machine

PROBABILITY REVIEW

- Odds
- Logit function
- Logistic function

	notation	Range Equivalents		
Standard probability	p	0	0.5	1
Odds	$\frac{p}{q}$	0	1	$+\infty$
Log odds (logit)	$\log(\frac{p}{q})$	$-\infty$	0	$+\infty$

LOGISTIC FUNCTION



https://walkccc.github.io/CS/ML/5/

PROPERTIES OF LR

- One parameter per data dimension (feature) and a bias
- Features can be discrete or continuous
- Output of the model $y \in [0, 1]$
- Allows for gradient-based learning of parameters

BEST $\boldsymbol{\theta}$?

- Maximum likelihood estimation:

$$\max_{\theta} \quad ll(w) = \max_{\theta} \quad \sum_{i} \log P(y^{(i)} | x^{(i)}; \boldsymbol{\theta})$$

with:

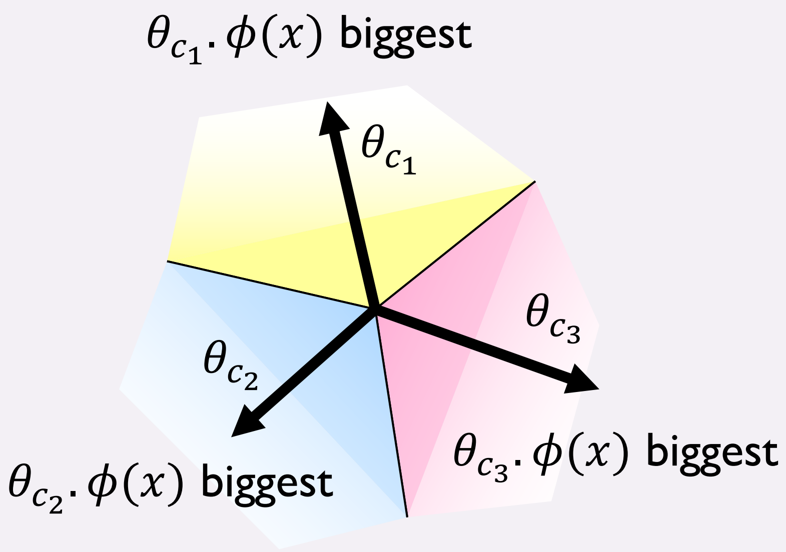
$$P(y^{(i)} = +1 | x^{(i)}; \boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T x^{(i)}}}$$

$$P(y^{(i)} = -1 | x^{(i)}; \boldsymbol{\theta}) = 1 - \frac{1}{1 + e^{-\boldsymbol{\theta}^T x^{(i)}}}$$

MULTICLASS PROBABILISTIC REGRESSION

- Recall:
  - A weight vector for each class: θ_c
  - Score (activation) of a class c: $z_c = \theta_c . \phi(x)$
  - Prediction highest score wins

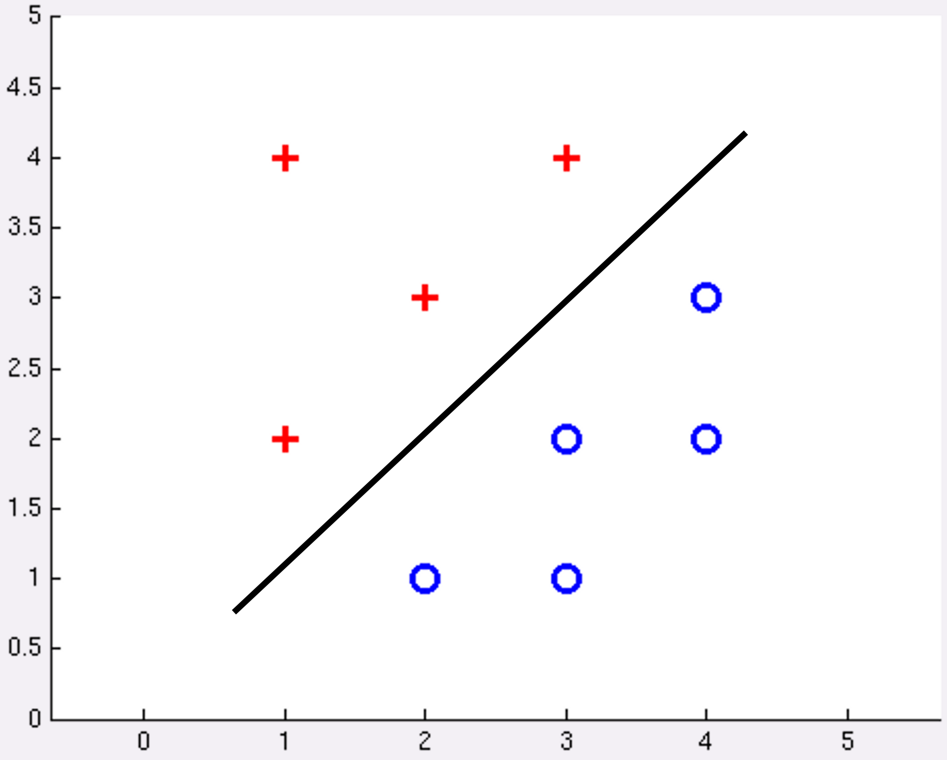
$$y = \underset{c}{\operatorname{argmax}}\, \theta_c . \phi(x)$$

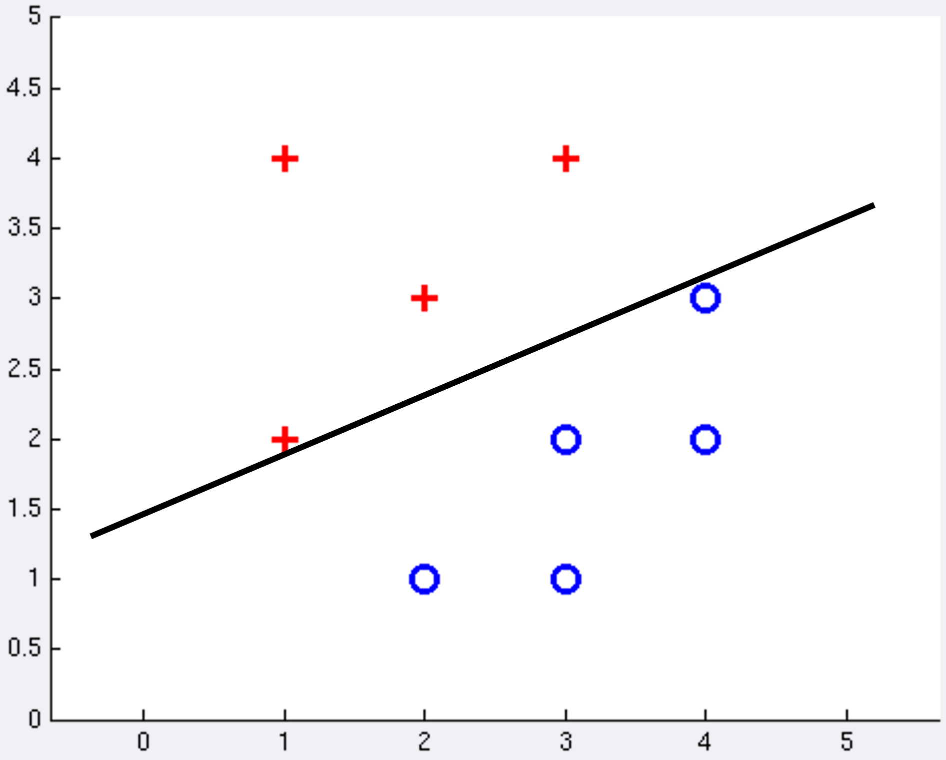


- How to make the scores into probabilities?

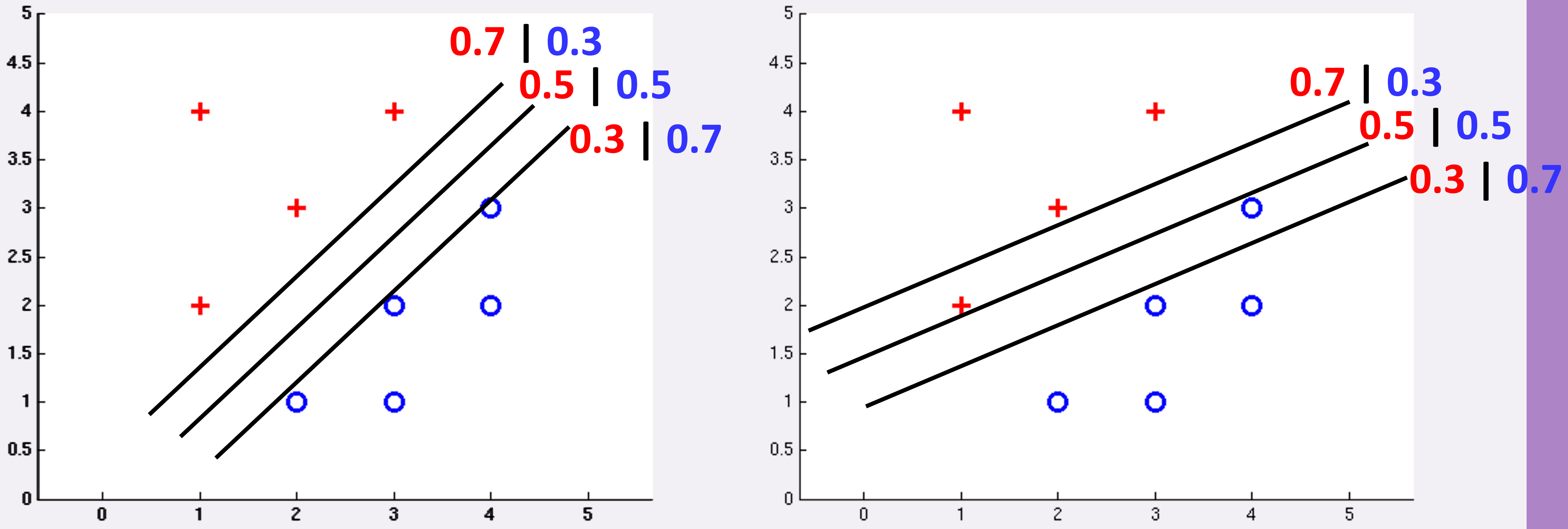
$$\underbrace{z_1, z_2, z_3}_{\text{original activations}} \rightarrow \underbrace{\frac{e^{z_1}}{e^{z_1}+e^{z_2}+e^{z_3}}, \frac{e^{z_2}}{e^{z_1}+e^{z_2}+e^{z_3}}, \frac{e^{z_3}}{e^{z_1}+e^{z_2}+e^{z_3}}}_{\text{softmax activations}}$$

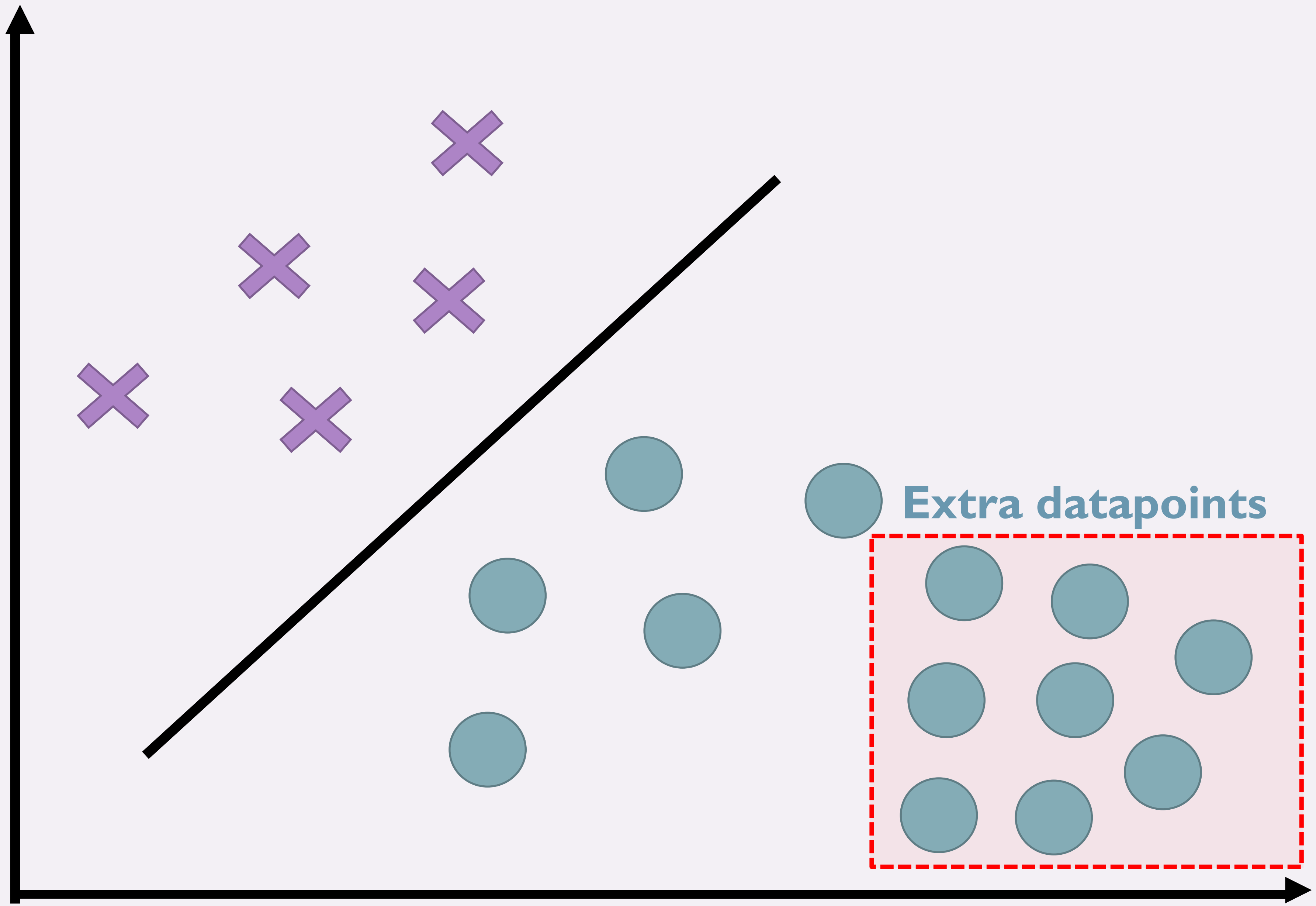
SEPARABLE CASE: DETERMINISTIC DECISION – MANY OPTIONS

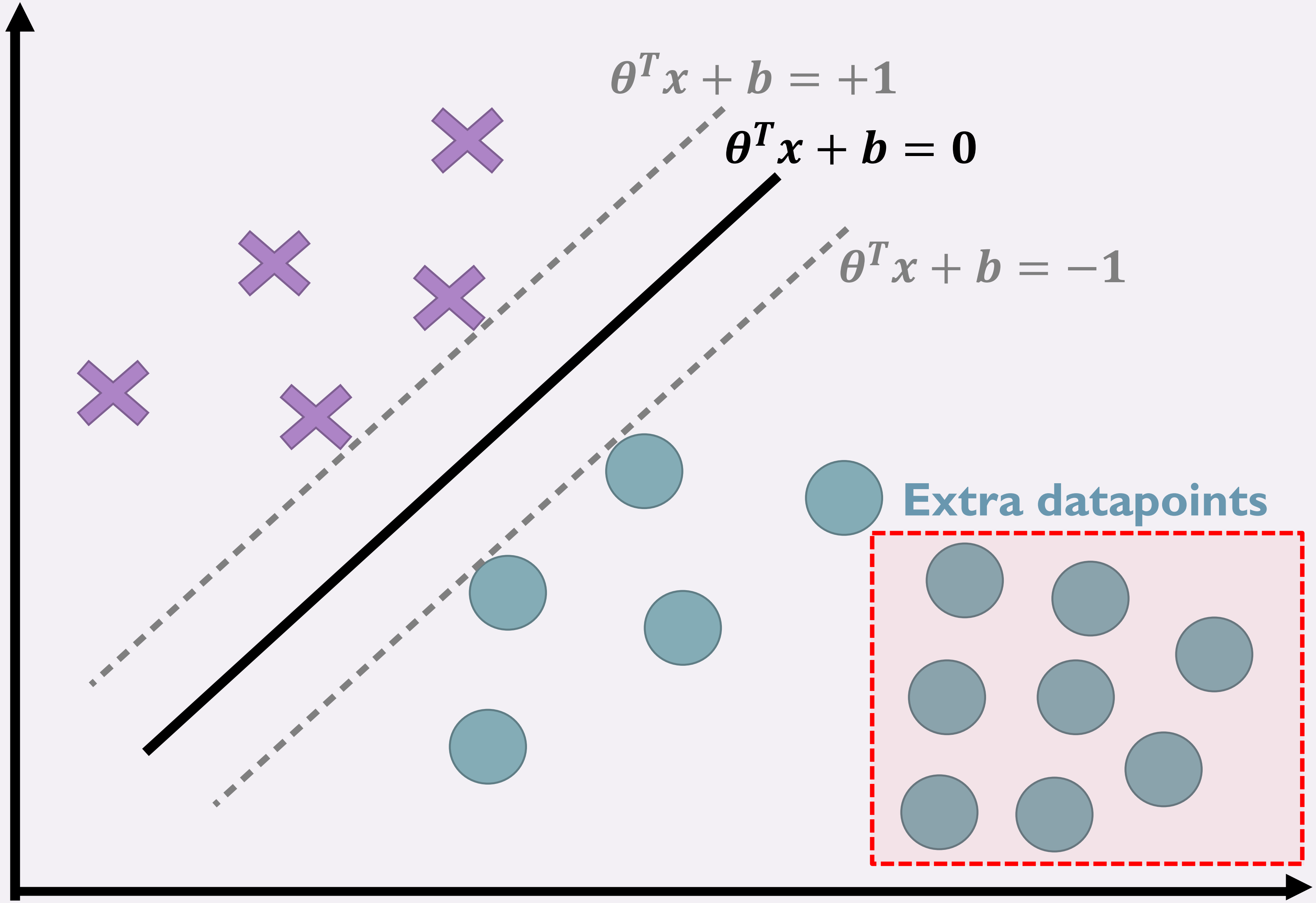




SEPARABLE CASE: PROBABILISTIC DECISION – CLEAR PREFERENCE

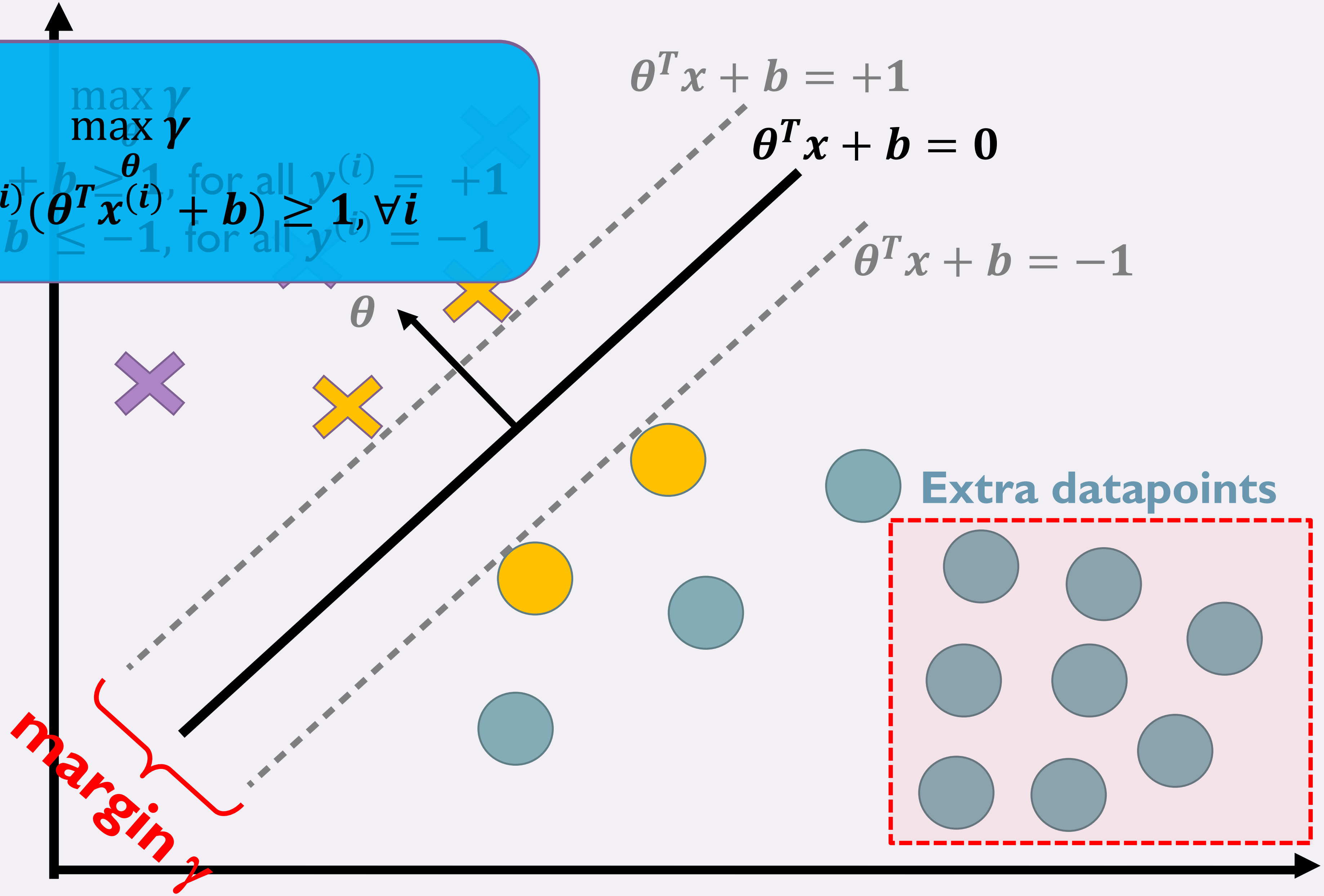



Extra datapoints


$\theta^T x + b = +1$
$\theta^T x + b = 0$
$\theta^T x + b = -1$
Extra datapoints

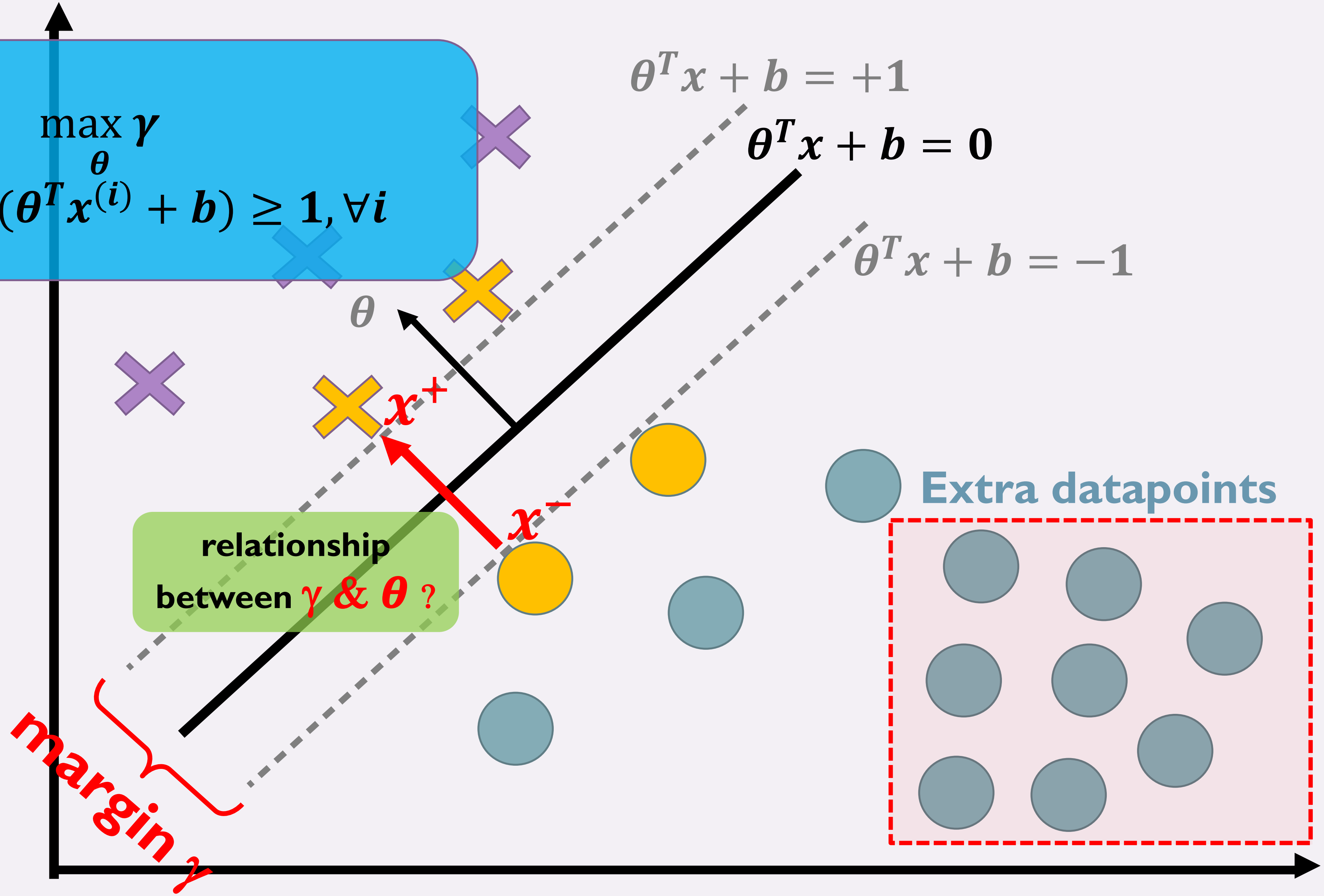
$$\max_{\theta} \gamma$$

$$\text{s.t.} \quad y^{(i)}(\theta^T x^{(i)} + b) \geq 1, \forall i$$



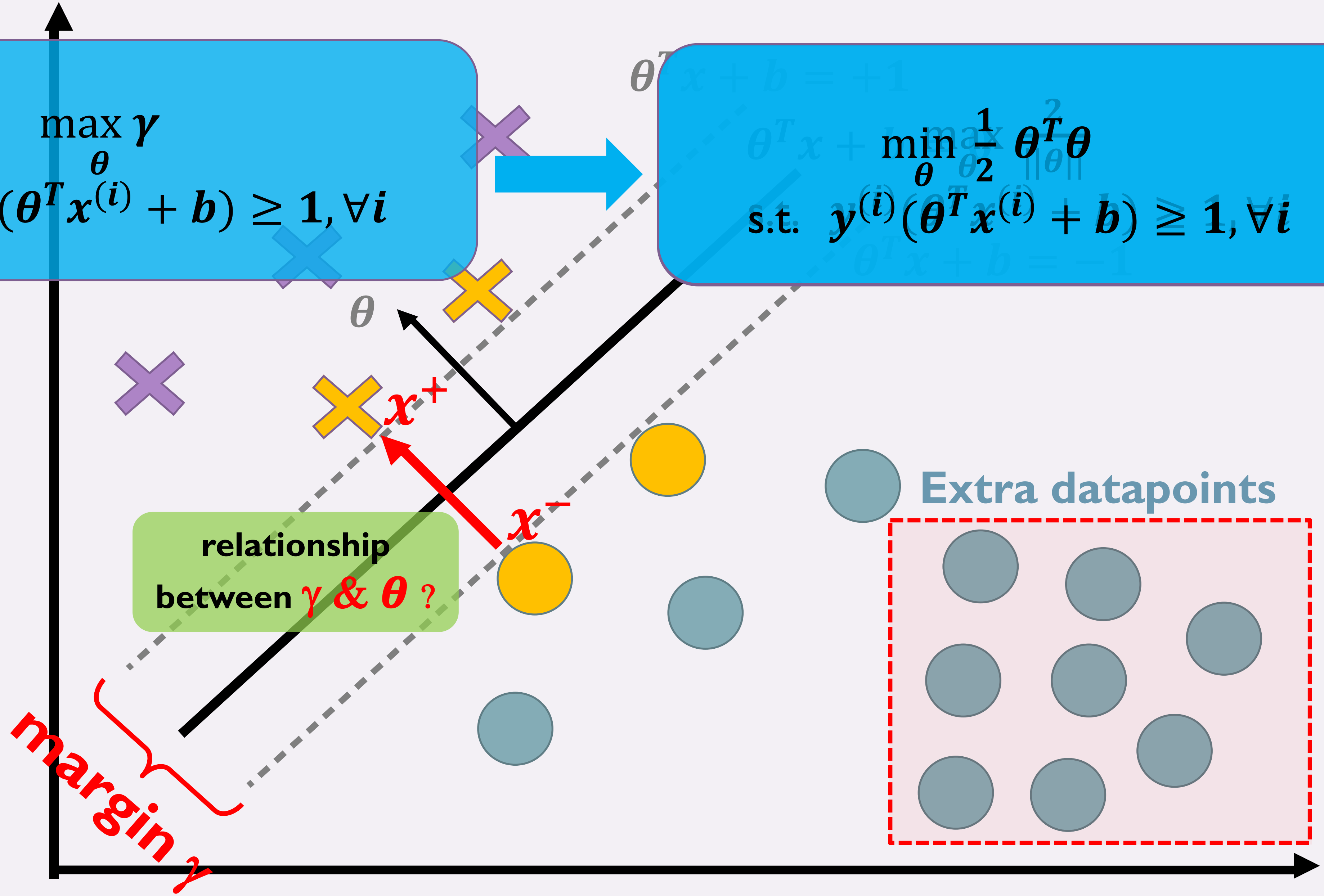
$$\max_{\boldsymbol{\theta}} \boldsymbol{\gamma}$$

$$\text{s.t.} \quad \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + \boldsymbol{b}) \geq \mathbf{1}, \forall \boldsymbol{i}$$



$$\max_{\boldsymbol{\theta}} \boldsymbol{\gamma}$$
$$\text{s.t. } \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + \boldsymbol{b}) \geq \mathbf{1}, \forall \boldsymbol{i}$$

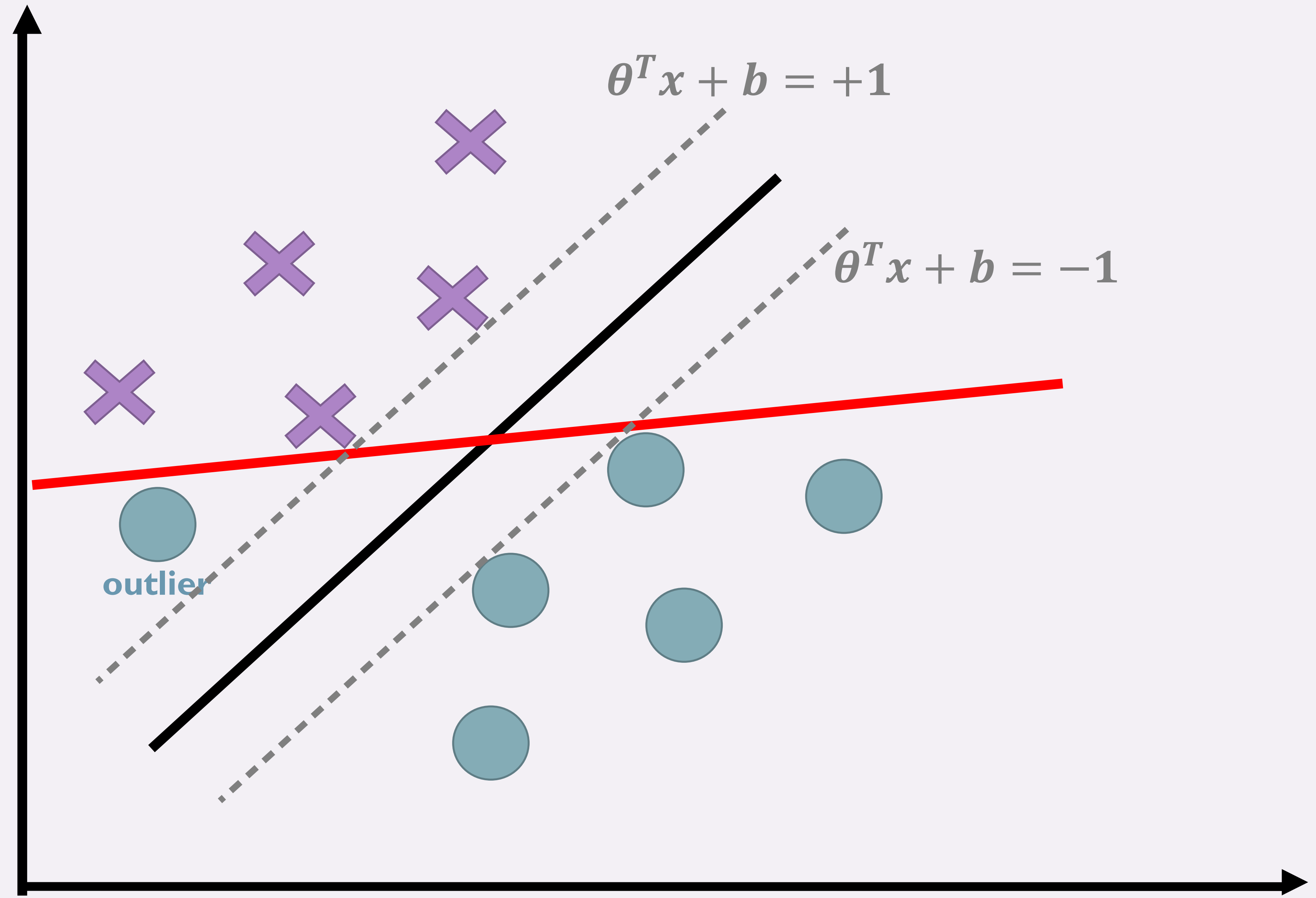
$$\min_{\boldsymbol{\theta}} \frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\theta}$$
$$\text{s.t. } \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + \boldsymbol{b}) \geq \mathbf{1}, \forall \boldsymbol{i}$$

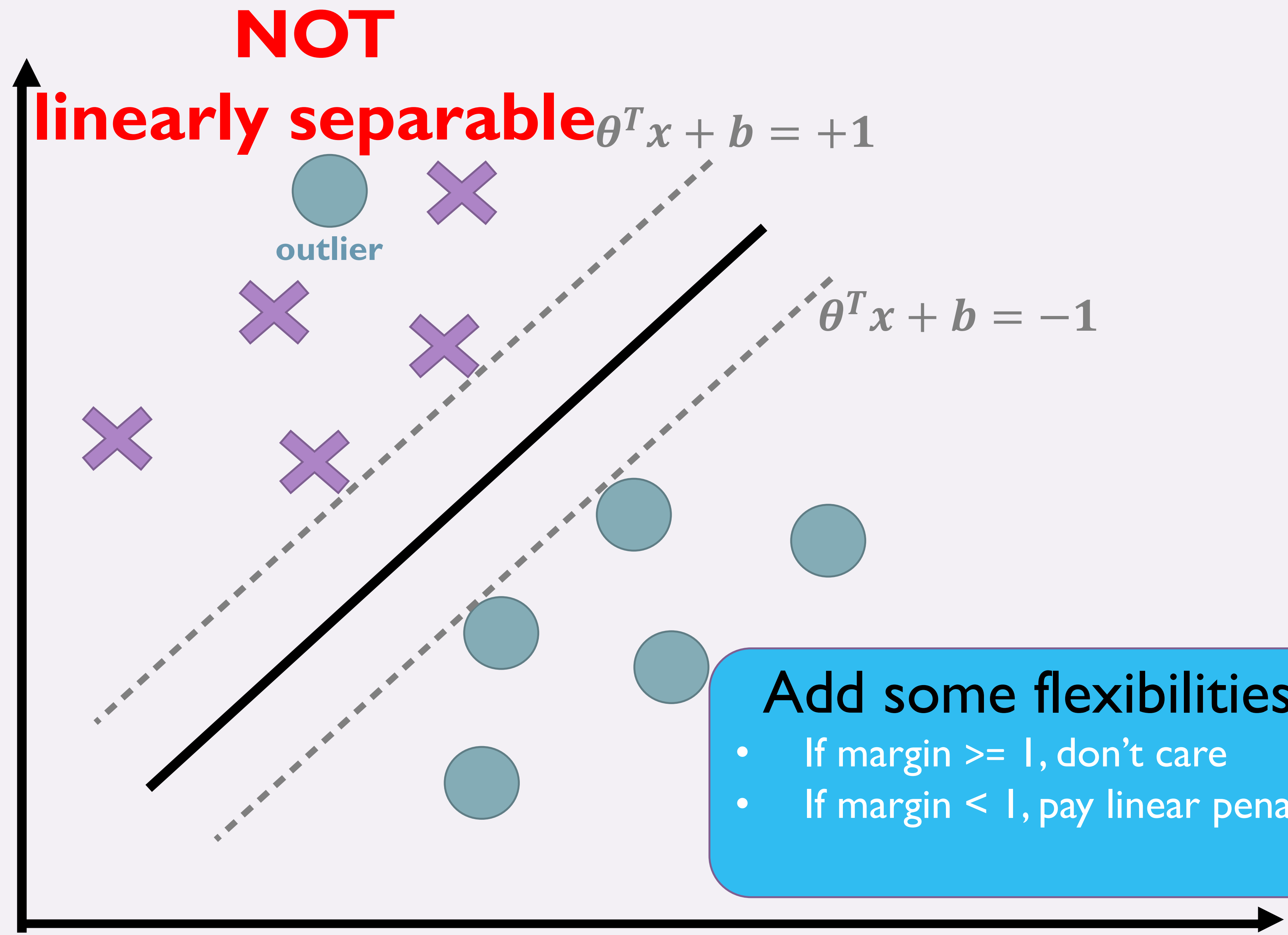


SUPPORT VECTOR MACHINE (SVM)

$$\min_{\boldsymbol{\theta}} \frac{\mathbf{1}}{\mathbf{2}} \boldsymbol{\theta}^{\boldsymbol{T}} \boldsymbol{\theta}$$

$$\text{s.t.} \quad \boldsymbol{y}^{(\boldsymbol{i})}(\boldsymbol{\theta}^{\boldsymbol{T}} \boldsymbol{x}^{(\boldsymbol{i})} + \boldsymbol{b}) \geq \mathbf{1}, \forall \boldsymbol{i}$$


$\theta^T x + b = +1$
$\theta^T x + b = -1$
outlier



Add some flexibilities?

- If margin >= 1, don't care
- If margin < 1, pay linear penalty

SOFT MARGIN SVM

$$\min_{\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{b}} \frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\theta} + \mathrm{C} \sum_{i=1}^{N} \xi_i$$

$$\text{s.t.} \quad \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + \boldsymbol{b}) + \xi_i \geq 1,$$

$$\xi_i \geq 0, \forall i$$

ξ_i is the "slack" variable

- for $0 < \xi_i \leq 1$ point is between margin and correct side of hyperplane. This is a margin violation
- for $\xi_i > 1$ point is misclassified

SOFT MARGIN SVM

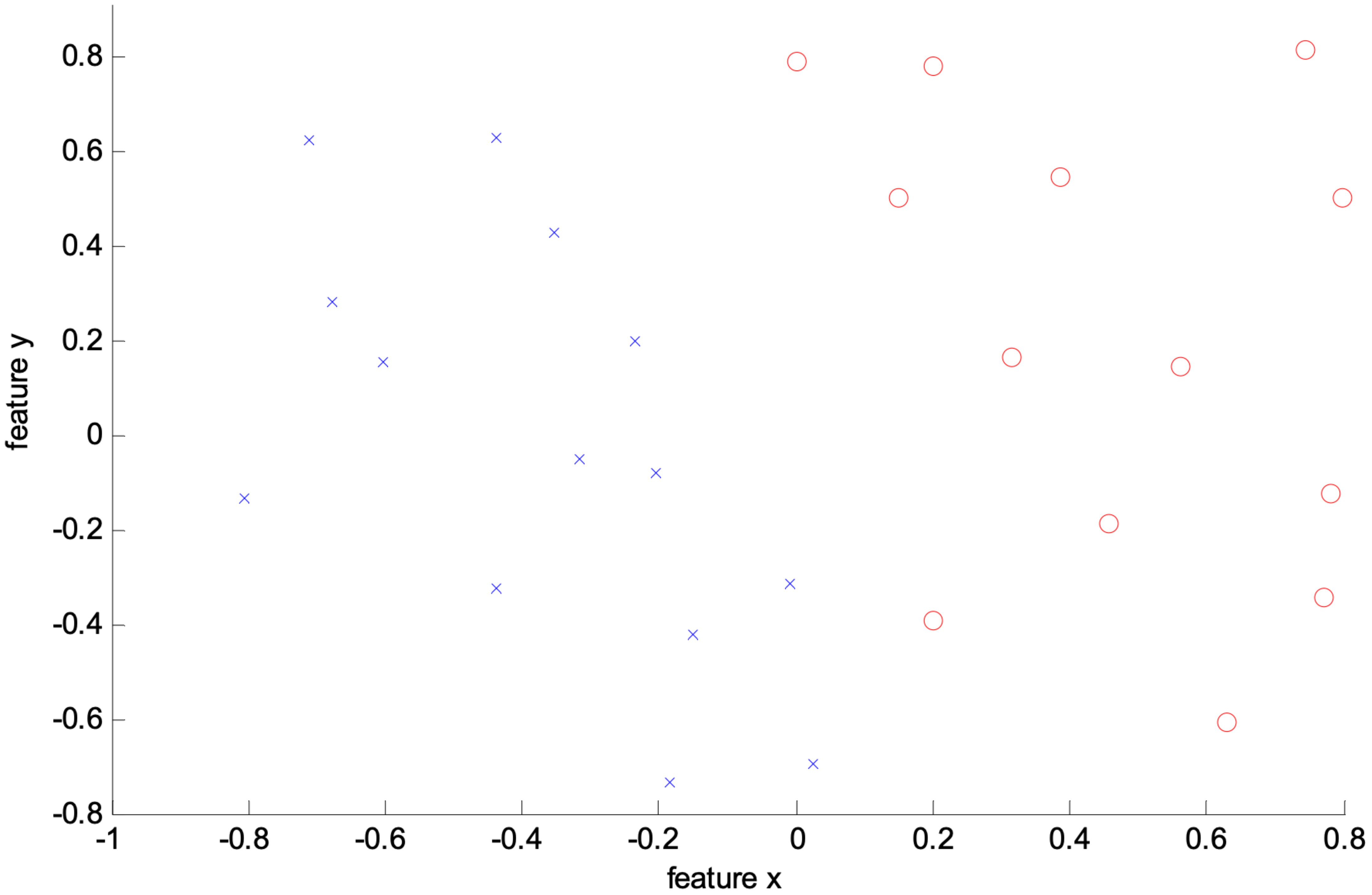
$$\min_{\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{b}} \frac{1}{2}\,\boldsymbol{\theta}^T\boldsymbol{\theta} + \mathrm{C}\,\sum_{i=1}^{N}\xi_i$$

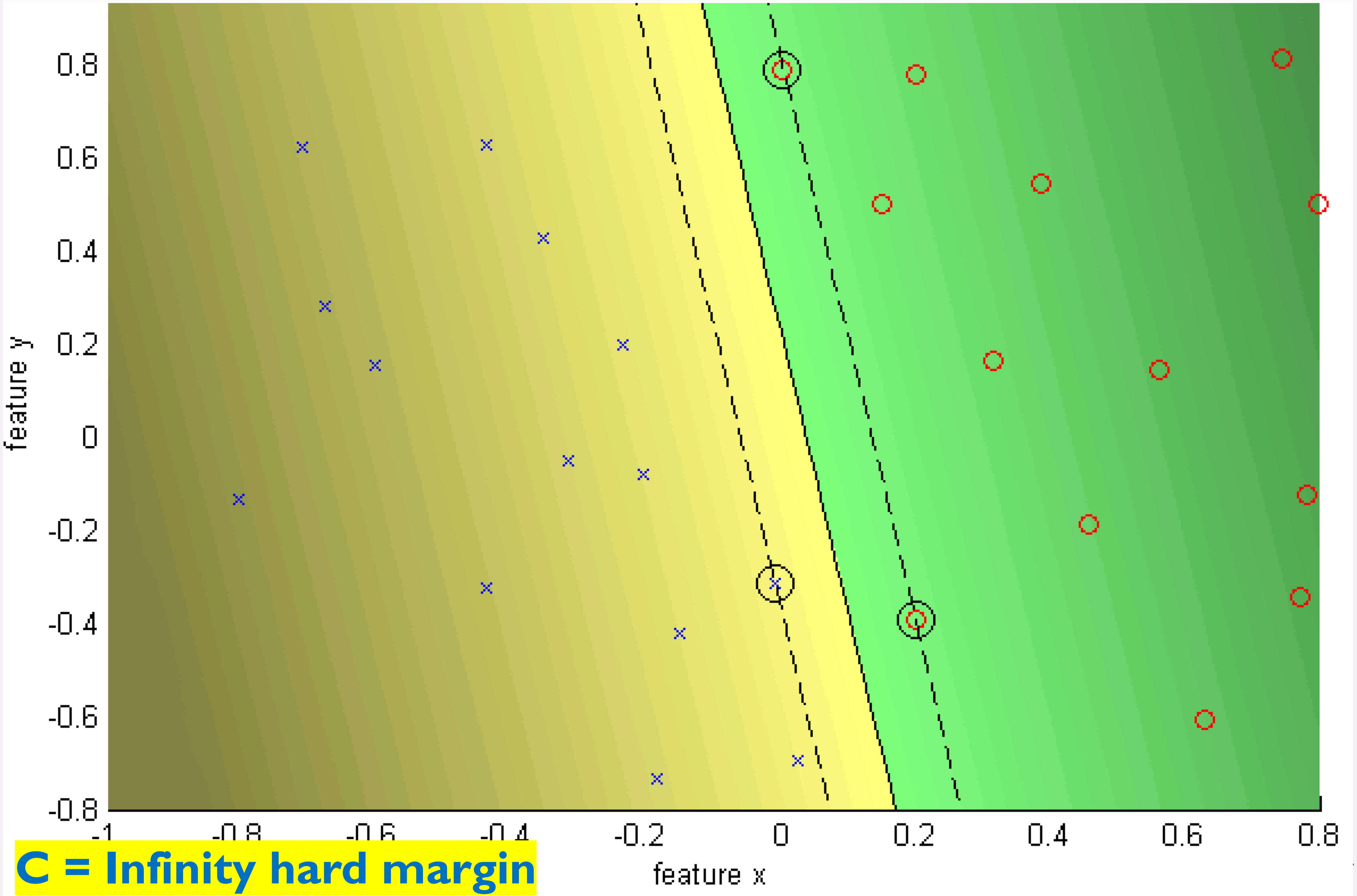
$$\text{s.t.}\quad \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + \boldsymbol{b}) + \xi_i \geq \mathbf{1},$$

$$\xi_i \geq \mathbf{0}, \forall i$$

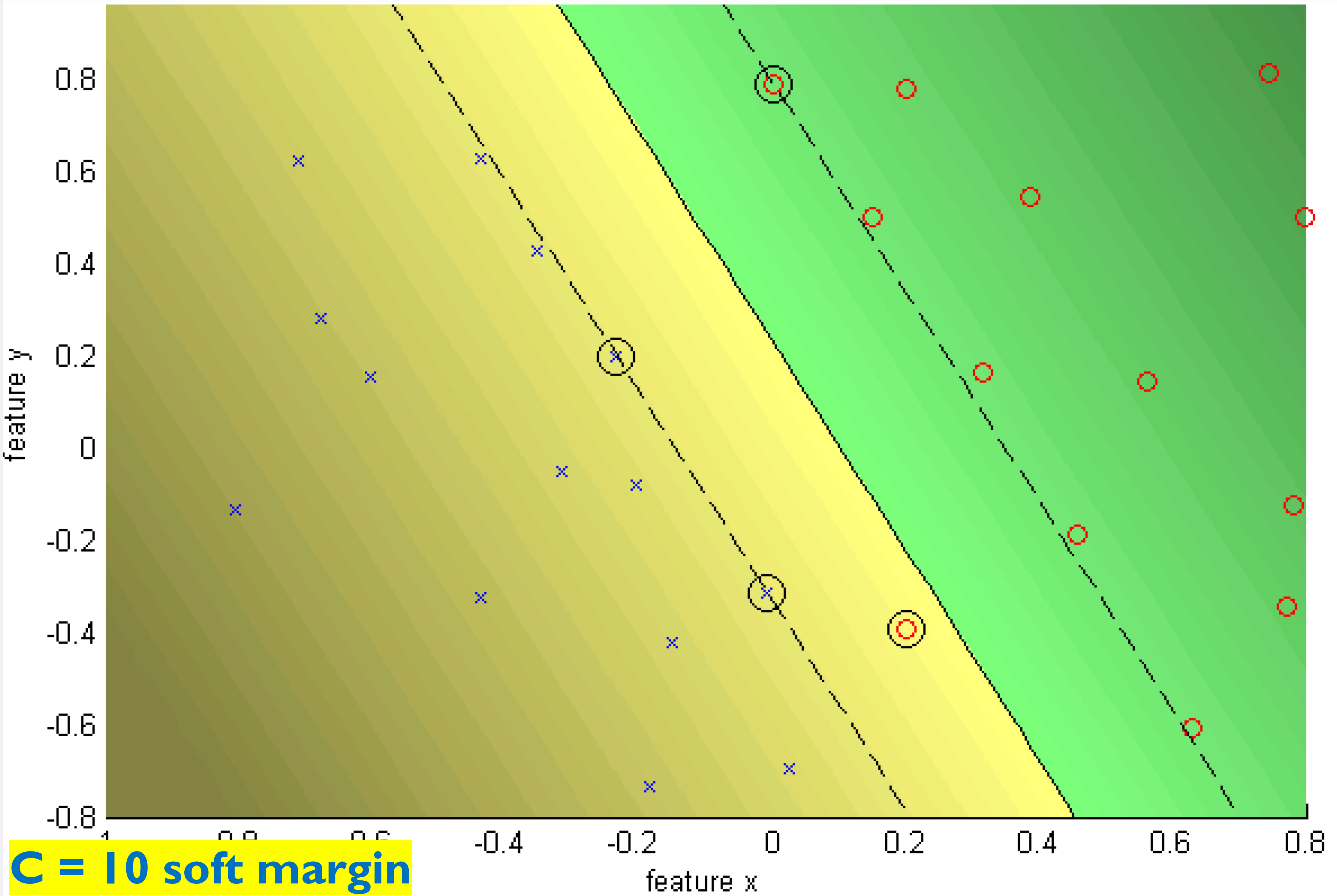
C is a regularization parameter:

- small C allows constraints to be easily ignored → large margin
- large C makes constraints hard to ignore → narrow margin
- C = ∞ enforces all constraints: hard margin


0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
feature y
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
feature x



C = Infinity hard margin



C = 10 soft margin

GRADIENT DESCENT FOR SVM

$$\boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + \boldsymbol{b}) + \boldsymbol{\xi}_i \geq \mathbf{1} \;\&\; \boldsymbol{\xi}_i \geq \mathbf{0}$$

$$\boldsymbol{\xi}_i = \max\{\mathbf{0}, \mathbf{1} - \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + \boldsymbol{b})\}$$

$$\min_{\boldsymbol{\theta},\boldsymbol{b}} \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{\theta}^T\boldsymbol{\theta} + \mathrm{C}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}} \max\{\mathbf{0}, \mathbf{1} - \boldsymbol{y}^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + \boldsymbol{b})\}$$

EXERCISE

HTTPS://BIT.LY/3AHLBZC

READING MATERIAL: LAGRANGIAN DUALITY

HTTPS://BIT.LY/3GIFCFJ

QUESTIONS?