LECTURE 10

SPRING 2021

APPLIED MACHINE LEARNING

CIHANG XIE

SLIDE CREDIT:

DHRUV BATRA

ANDREW ZISSERMAN

TODAY

- Support Vector Machine

  -- review

  -- Lagrangian duality

  -- kernel trick

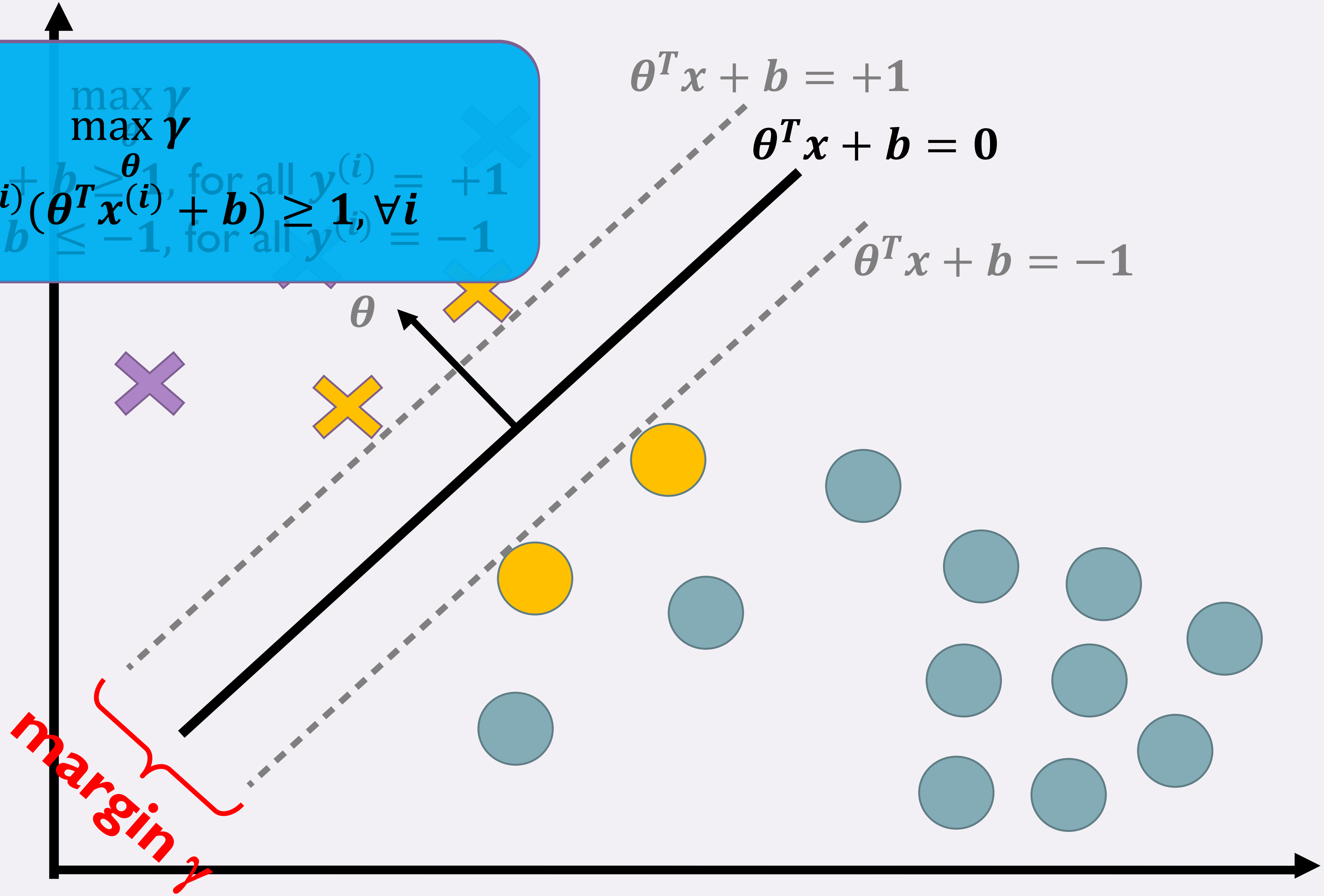
$$\max_{\theta} \gamma$$

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

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

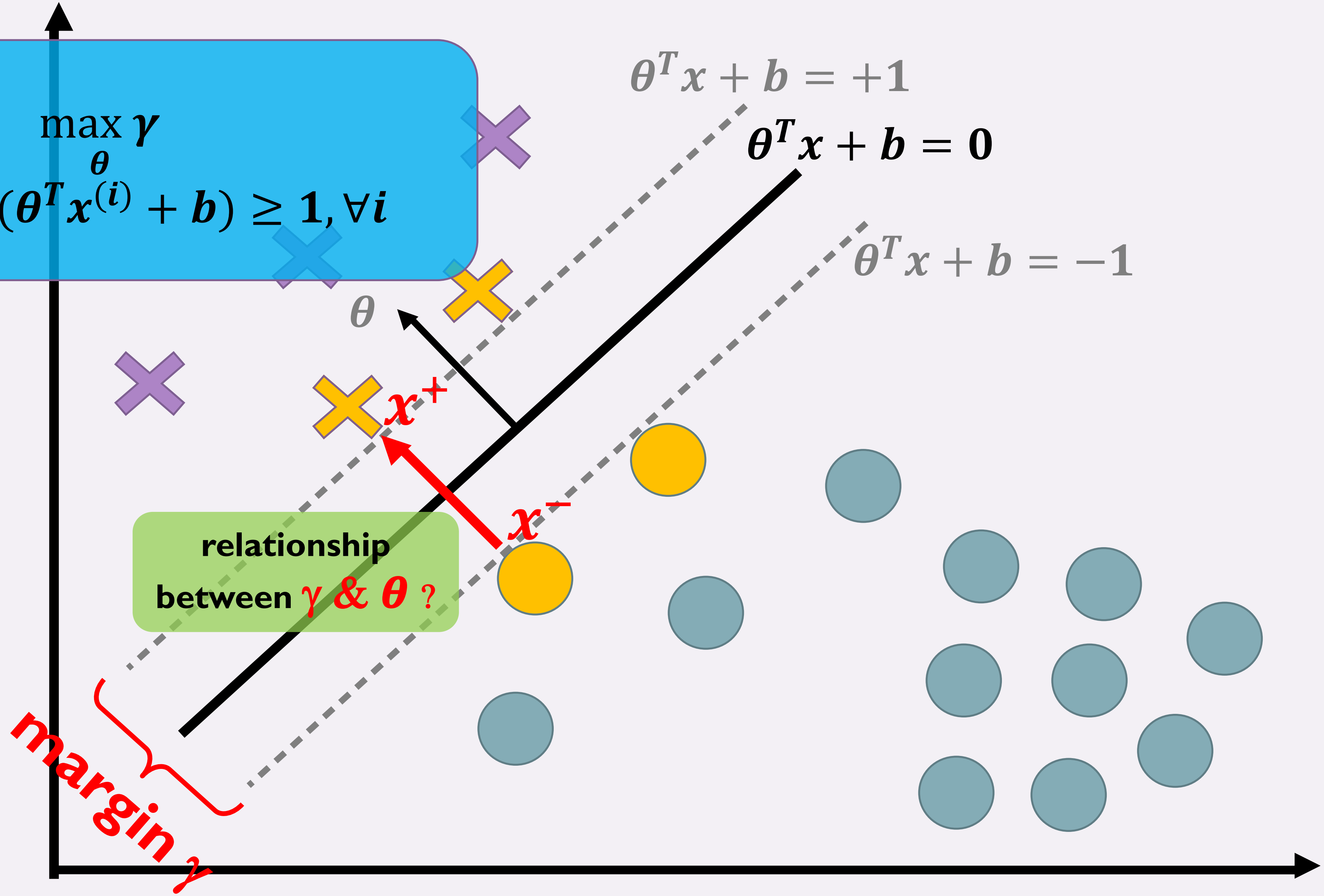
$$\text{s.t. } \theta^T x^{(i)} + b \geq 1, \text{ for all } y^{(i)} = +1$$

$$\theta^T x^{(i)} + b \leq -1, \text{ for all } y^{(i)} = -1$$



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

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

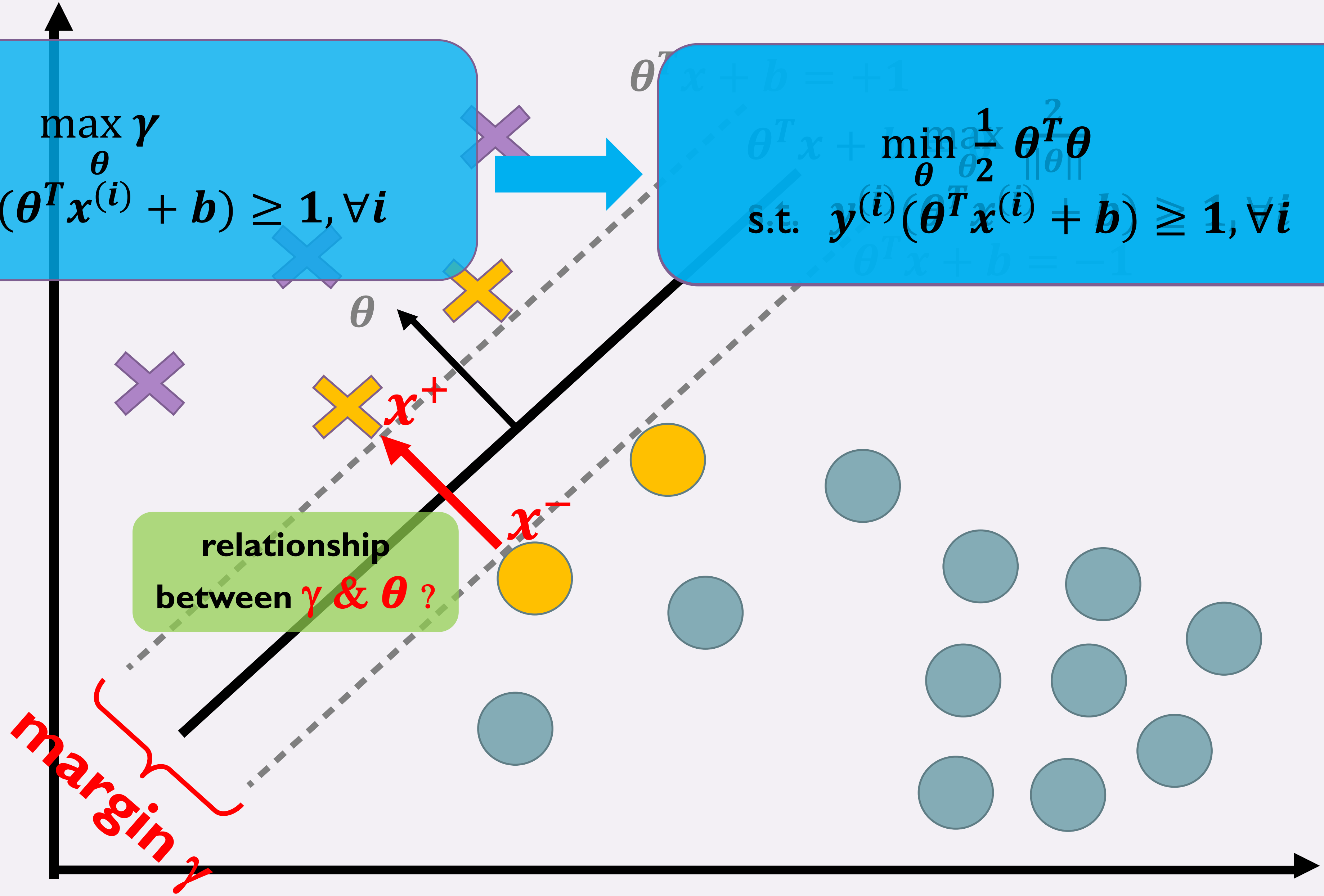


$$\max_{\theta} \gamma$$

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

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

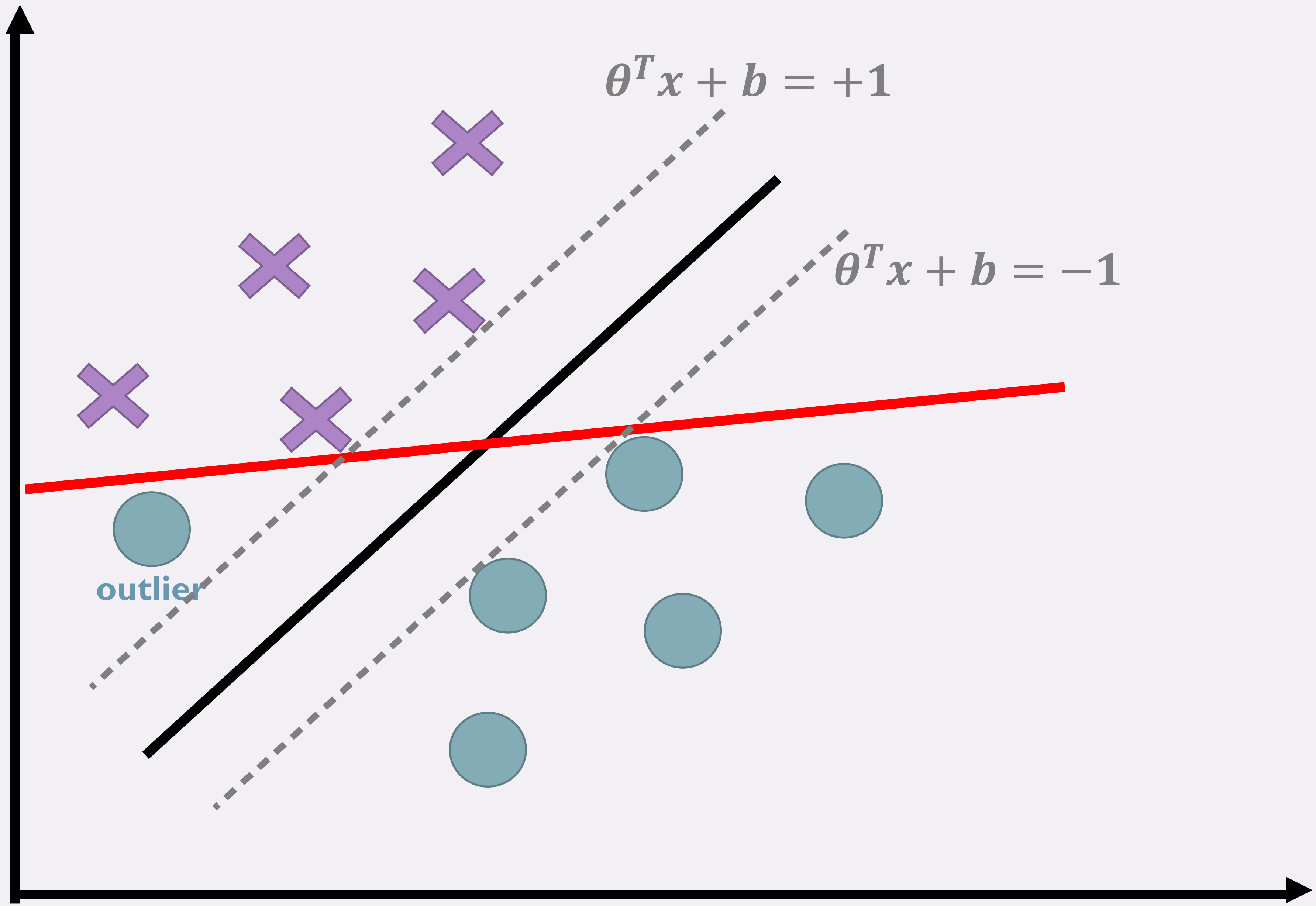
$$\text{s.t.} \quad y^{(i)}(\theta^T x^{(i)} + b) \geq 1, \forall i$$

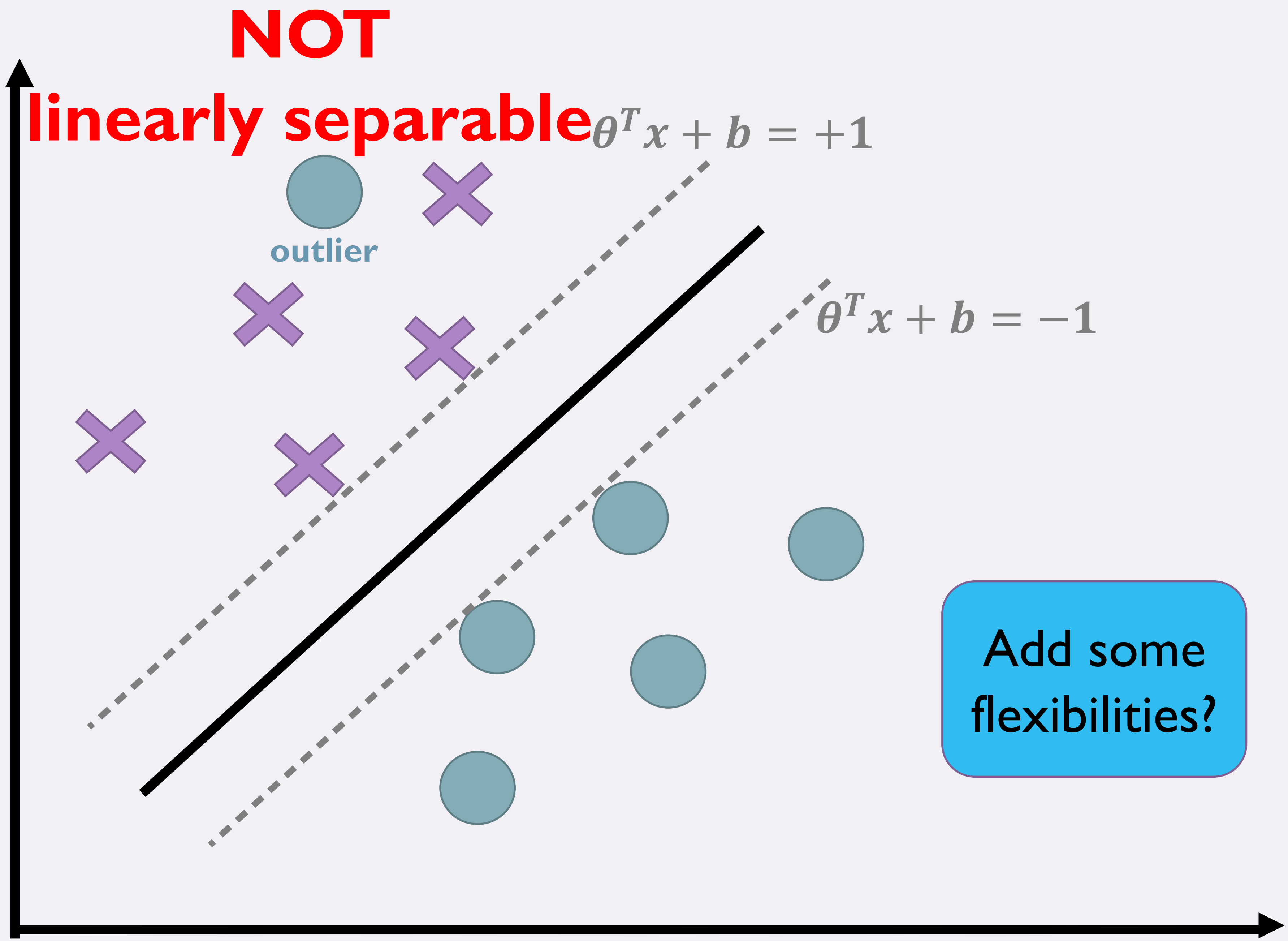


SUPPORT VECTOR MACHINE (SVM)

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

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


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


NOT
linearly separable
$\theta^T x + b = +1$
outlier
$\theta^T x + b = -1$
Add some flexibilities?

SOFT MARGIN SVM

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

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

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

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

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

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

$$\boldsymbol{\xi}_{\boldsymbol{i}} \geq \mathbf{0}, \forall \boldsymbol{i}$$

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

GRADIENT DESCENT FOR SVM

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

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

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

GRADIENT DESCENT FOR SVM

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

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

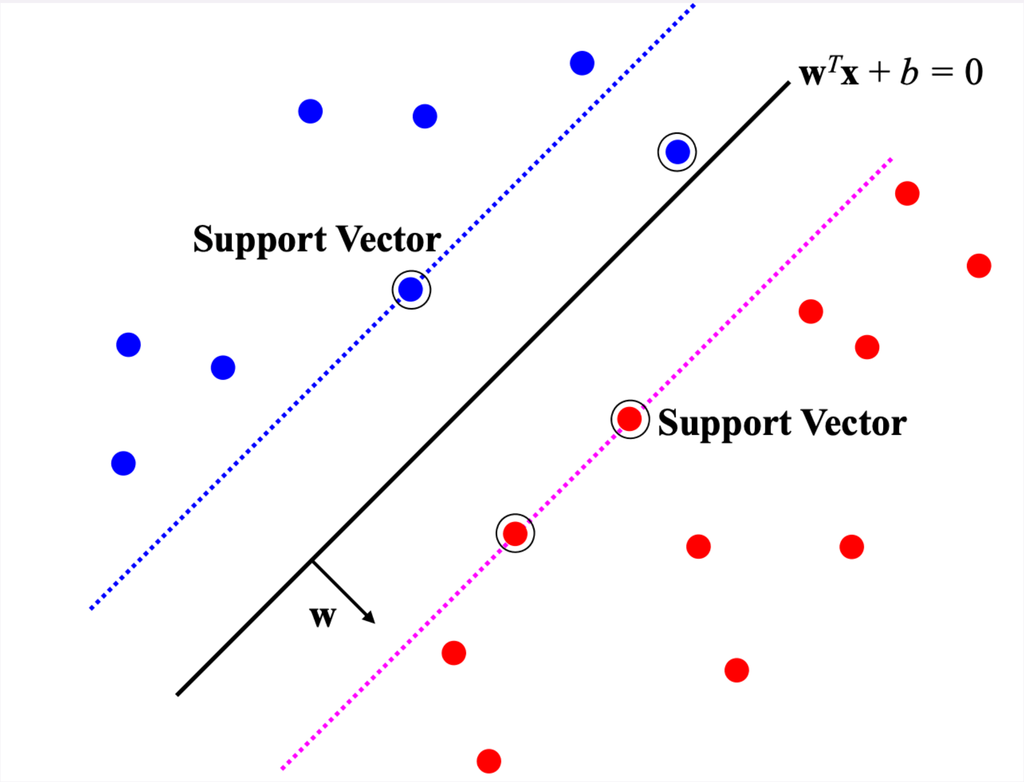
For each data point $x^{(i)}$

$$\frac{\partial Cost(\theta, b)}{\partial \theta_j} = \begin{cases} \frac{1}{N}\theta_j - \mathrm{C}\, y^{(i)} x_j^{(i)}, & \text{if } 1 - y^{(i)}(\theta^T x^{(i)} + b) > 0 \\ \frac{1}{N}\theta_j, & \text{otherwise} \end{cases}$$

$$\frac{\partial Cost(\theta, b)}{\partial b} = \begin{cases} -\mathrm{C}\, y^{(i)}, & \text{if } 1 - y^{(i)}(\theta^T x^{(i)} + b) > 0 \\ 0, & \text{otherwise} \end{cases}$$

OPTIMIZATION

$$\min_{\theta,b} \underbrace{\left\{\frac{1}{2}\theta^T\theta\right\}}_{\text{Regularization}} + \underbrace{\left\{C\sum_{i=1}^{N} \max\{0, 1 - y^{(i)}(\theta^T x^{(i)} + b)\}\right\}}_{\text{Model fit to data}}$$



1. $y^{(i)}(\theta^T x^{(i)} + b) > 1$ => Point is outside margin. No contribution to loss

2. $y^{(i)}(\theta^T x^{(i)} + b) = 1$ => Point is on margin. No contribution to loss.

3. $y^{(i)}(\theta^T x^{(i)} + b) < 1$ => Point violates margin constraint. Contributes to loss

RECALL: LOGISTIC REGRESSION

- Maximum likelihood estimation:

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

with:

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

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

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

RECALL: LOGISTIC REGRESSION

$$\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{i}} \log \boldsymbol{P}\left(\boldsymbol{y}^{(\boldsymbol{i})} \middle| \boldsymbol{x}^{(\boldsymbol{i})}; \boldsymbol{\theta}\right) = \max_{\boldsymbol{\theta}} \sum_{\boldsymbol{i}} \log \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{e}^{-\boldsymbol{y}^{(i)}(\boldsymbol{\theta}^{T}\boldsymbol{x}^{(i)}+\boldsymbol{b})}}$$

$$= \max_{\boldsymbol{\theta}} \sum\nolimits_{\boldsymbol{i}} (\log 1 - \log(1 + \boldsymbol{e}^{-\boldsymbol{y}^{(i)}(\boldsymbol{\theta}^{T}\boldsymbol{x}^{(i)}+\boldsymbol{b})}))$$

$$= \max_{\boldsymbol{\theta}} \sum\nolimits_{\boldsymbol{i}} - \log(1 + \boldsymbol{e}^{-\boldsymbol{y}^{(i)}(\boldsymbol{\theta}^{T}\boldsymbol{x}^{(i)}+\boldsymbol{b})})$$

$$= \min_{\boldsymbol{\theta}} \sum\nolimits_{\boldsymbol{i}} \log(1 + \boldsymbol{e}^{-\boldsymbol{y}^{(i)}(\boldsymbol{\theta}^{T}\boldsymbol{x}^{(i)}+\boldsymbol{b})})$$

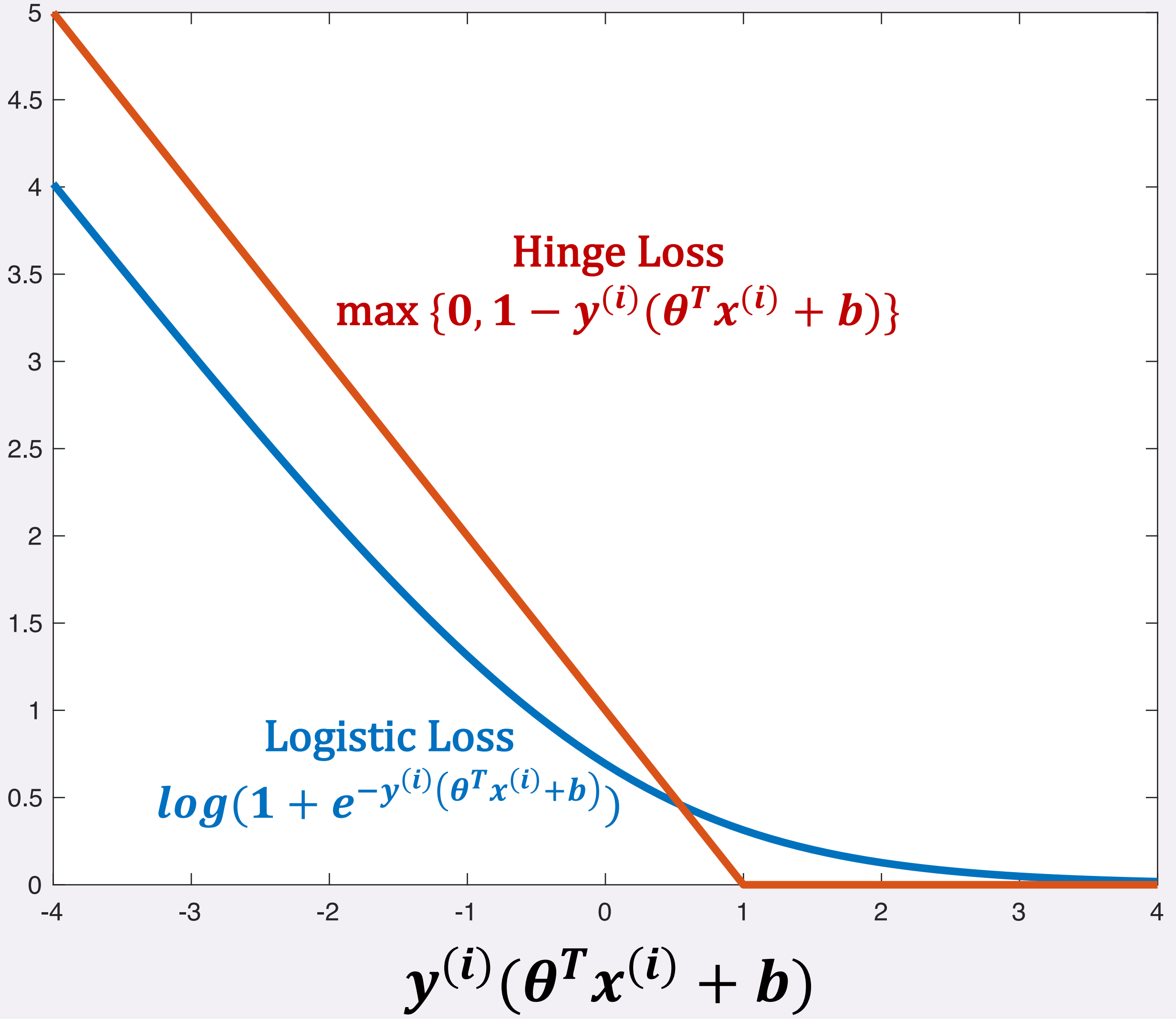
RELATIONSHIP TO LOGISTIC REGRESSION

$$\min_{\boldsymbol{\theta},\boldsymbol{b}} \boldsymbol{\lambda}\boldsymbol{\theta}^{T}\boldsymbol{\theta} - \sum_{\boldsymbol{i}} \log \boldsymbol{P}\left(\boldsymbol{y}^{(i)} \middle| \boldsymbol{x}^{(i)}; \boldsymbol{\theta}, \boldsymbol{b}\right)$$

$$\min_{\boldsymbol{\theta},\boldsymbol{b}} \underbrace{\boldsymbol{\lambda}\boldsymbol{\theta}^{T}\boldsymbol{\theta}}_{\text{Regularization}} + \sum_{\boldsymbol{i}} \underbrace{\log\left(1 + \boldsymbol{e}^{-\boldsymbol{y}^{(i)}\left(\boldsymbol{\theta}^{T}\boldsymbol{x}^{(i)} + \boldsymbol{b}\right)}\right)}_{\text{Logistics Loss}}$$

$$\min_{\boldsymbol{\theta},\boldsymbol{b}} \underbrace{\frac{1}{2}\boldsymbol{\theta}^{T}\boldsymbol{\theta}}_{\text{Regularization}} + \mathrm{C}\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} \underbrace{\max\left\{\mathbf{0}, \mathbf{1} - \boldsymbol{y}^{(i)}\left(\boldsymbol{\theta}^{T}\boldsymbol{x}^{(i)} + \boldsymbol{b}\right)\right\}}_{\text{Hinge Loss}}$$

RELATIONSHIP TO LOGISTIC REGRESSION



Logistic loss is sometime viewed as the **smooth version** of the Hinge loss.

HW2 (DUE MAY 9)

QUESTIONS?