# Today

- Last time
  - Dual formulation of SVMs

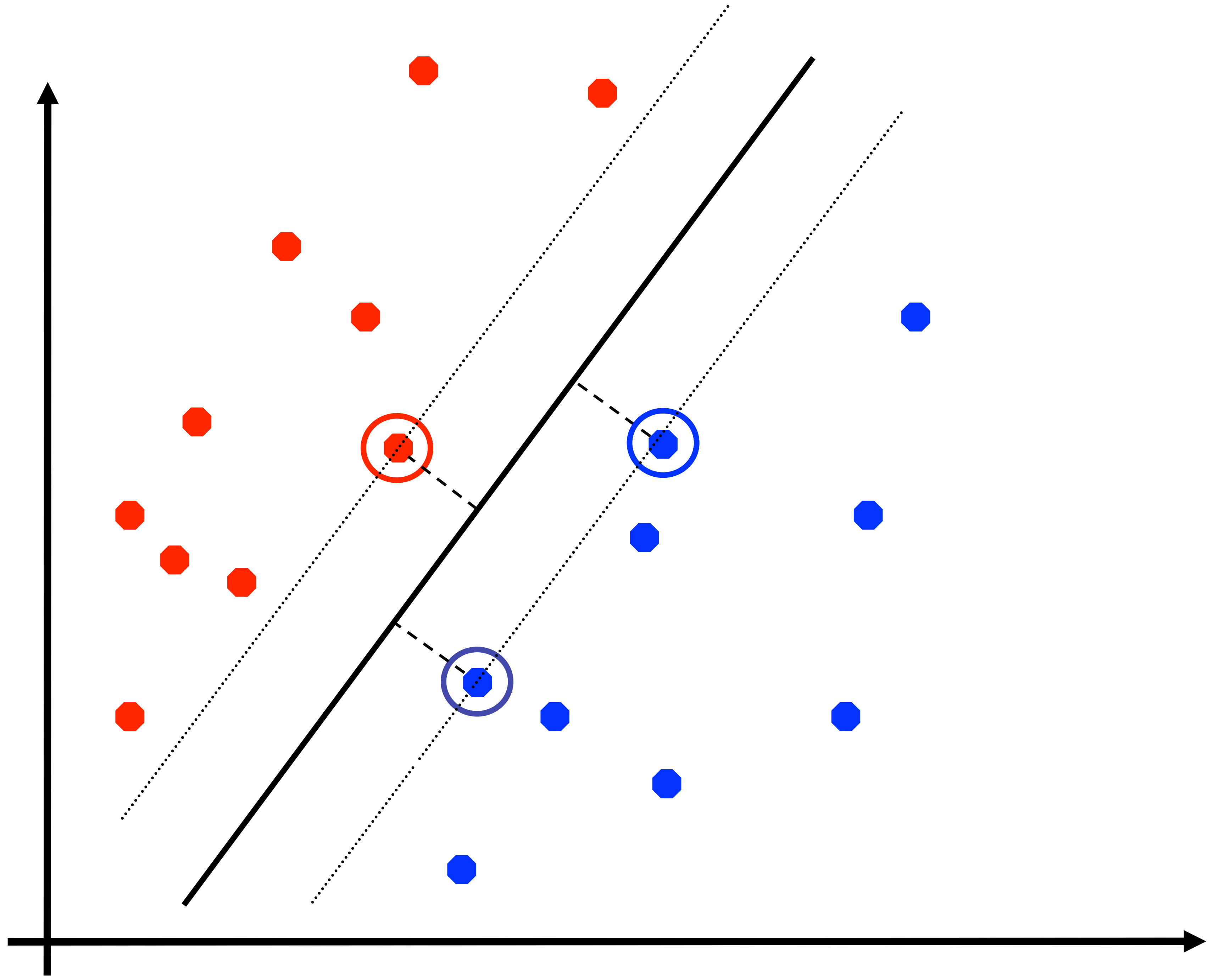
- Today
  - SVMs with slack (reading: Bishop, Sec 7.1.1, Notes by Andrew Ng http://cs229.stanford.edu/notes/cs229-notes3.pdf )

- Announcements
  - HW 2 due on Wed, Oct 14

  - Midterm exam is coming up (Wed, Oct 21)
    - Time for Q&A on Mon, Oct 19 in lecture
    - Office hours: Mondays from 9-10am, Wednesdays 12.15-1.15pm
    - Current plan: Midterm online via GradScope (more details to come)
    - Tell me asap if you cannot take the exam on Wed, Oct 21 from 11am-12.15pm because of timezone conflicts

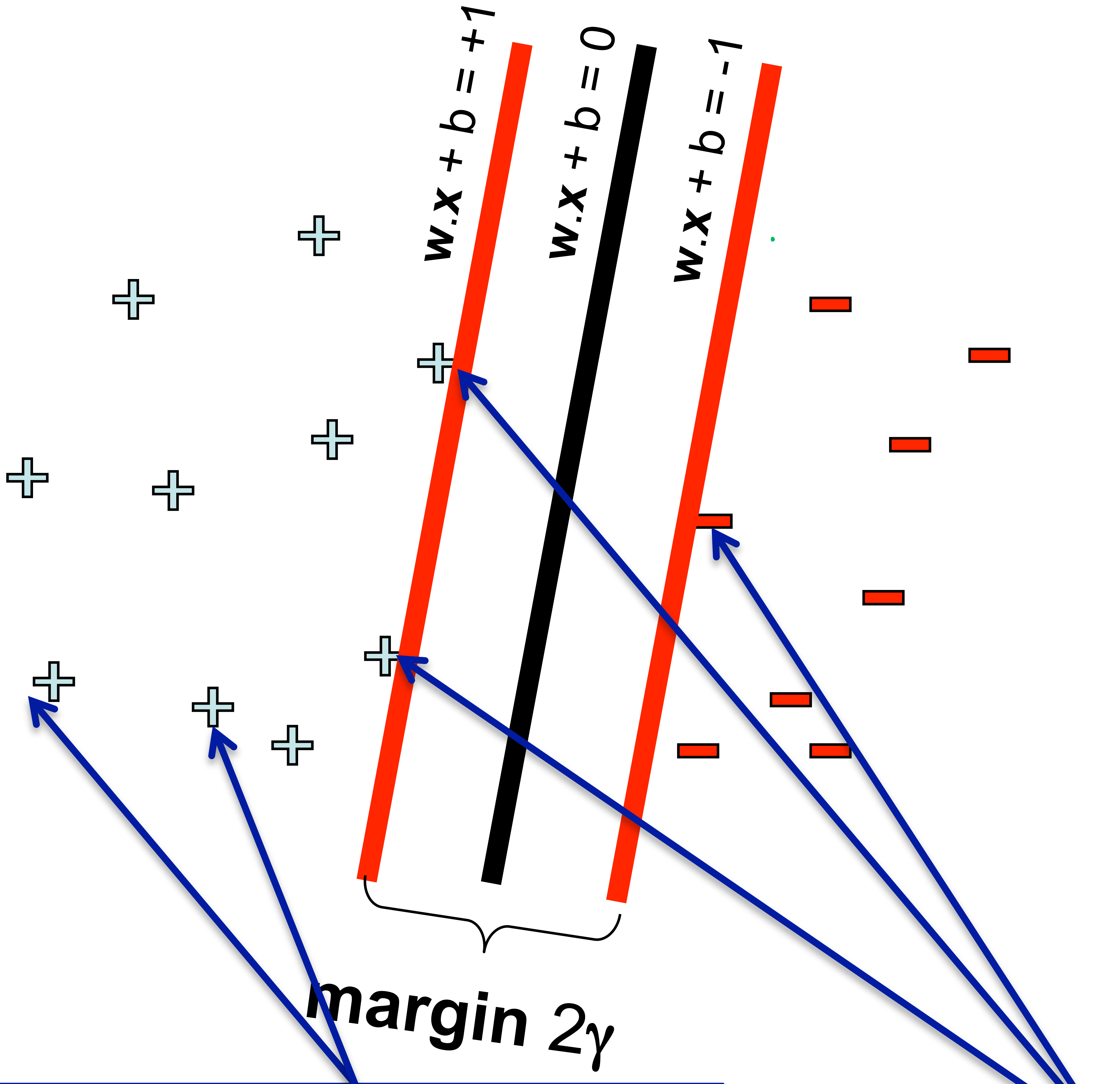
# Support vector machines

- SVMs (Vapnik, 1990’s) choose the linear separator with the **largest margin**



- Good according to intuition, theory, practice

[David Sontag]




**Non-support Vectors:**

- everything else
- moving them will not change **w**

**Support Vectors:**

- data points on the canonical lines

# Optimal margin classifier (cont'd)

- Invoke that functional margin $\hat{\gamma}$ depends on scaling
  - Multiplying $w, b$ by constant, multiplies $\hat{\gamma}$ by that constant
- Introducing constraint $\hat{\gamma} = 1$, which indeed is a scaling constraint on $w, b$ and obtain

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2$$

$$\textbf{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1\,, \quad i = 1, \ldots, N$$

- Note: maximizing $\hat{\gamma}/\|w\|$ (with $\hat{\gamma} = 1$) is same as minimizing $\|w\|^2$
- Convex quadratic objective, linear constraints
- The solution is the optimal margin classifier

# Dual formulation of SVMs

Solve:

$$\max_{\alpha} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y^{(i)} y^{(j)} \alpha_i \alpha_j < x^{(i)}, x^{(j)} >$$

s.t. $\quad \alpha_i \geq 0 \, , i = 1, \ldots, N$

$$\sum_{i=1}^{N} \alpha_i y^{(i)} = 0$$

Predict:

$$\sum_{i=1}^{N} \alpha_i^* y^{(i)} < (x^{(i)})^T, x > + b$$

- Solve and predict “touch” training data only via inner products
- This is key for using SVMs with kernels

# SVMs with slack

# Assumption of linear separability

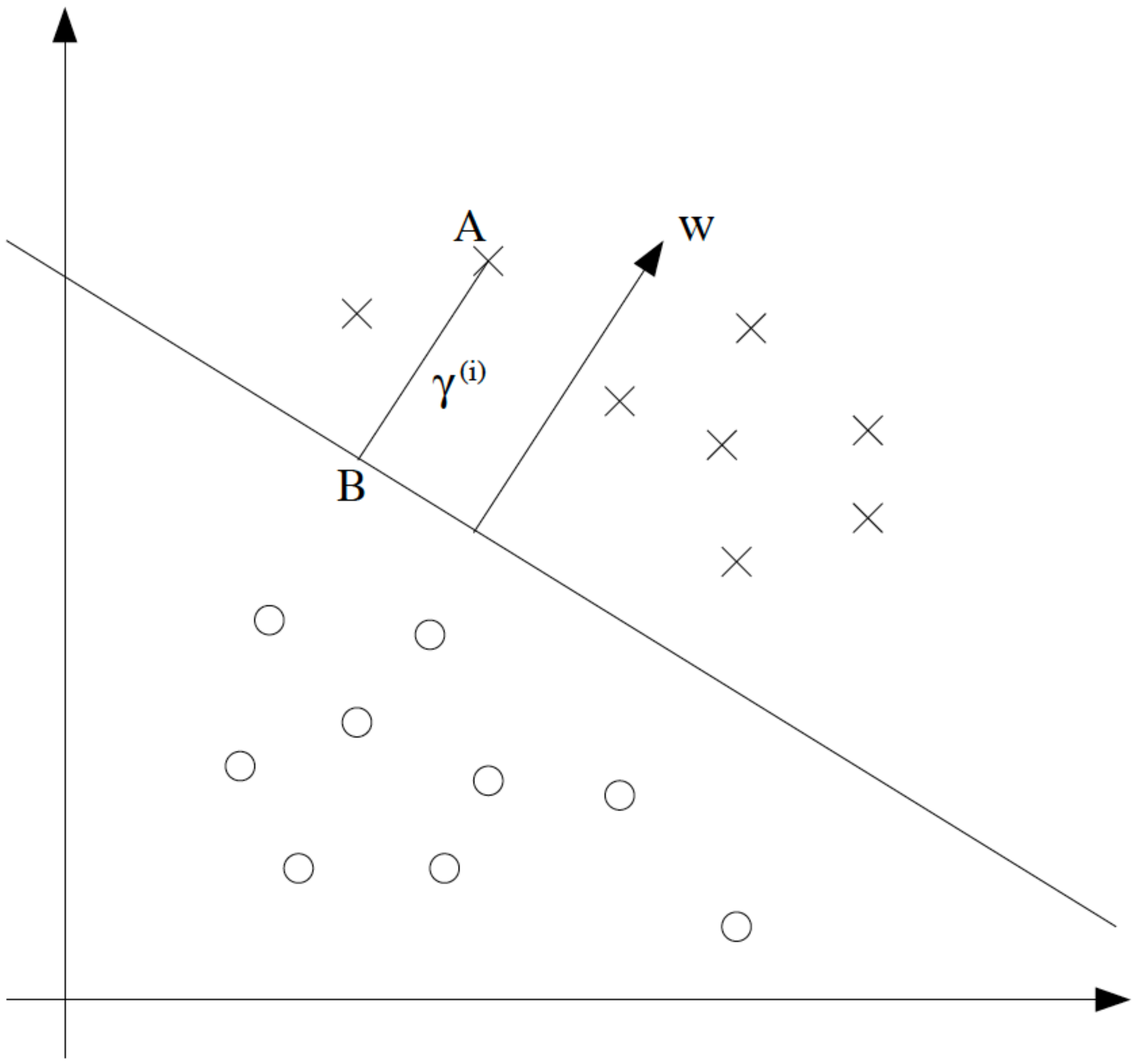
- So far, operated under the assumption that data set is linearly separable

$$\max_{\gamma, w, b} \quad \gamma$$

$$\textbf{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \quad i = 1, \ldots, N$$

$$\|w\| = 1$$

- Why is linear separability a key assumption?



# Assumption of linear separability

- So far, operated under the assumption that data set is linearly separable

$$\max_{\gamma, w, b} \quad \gamma$$

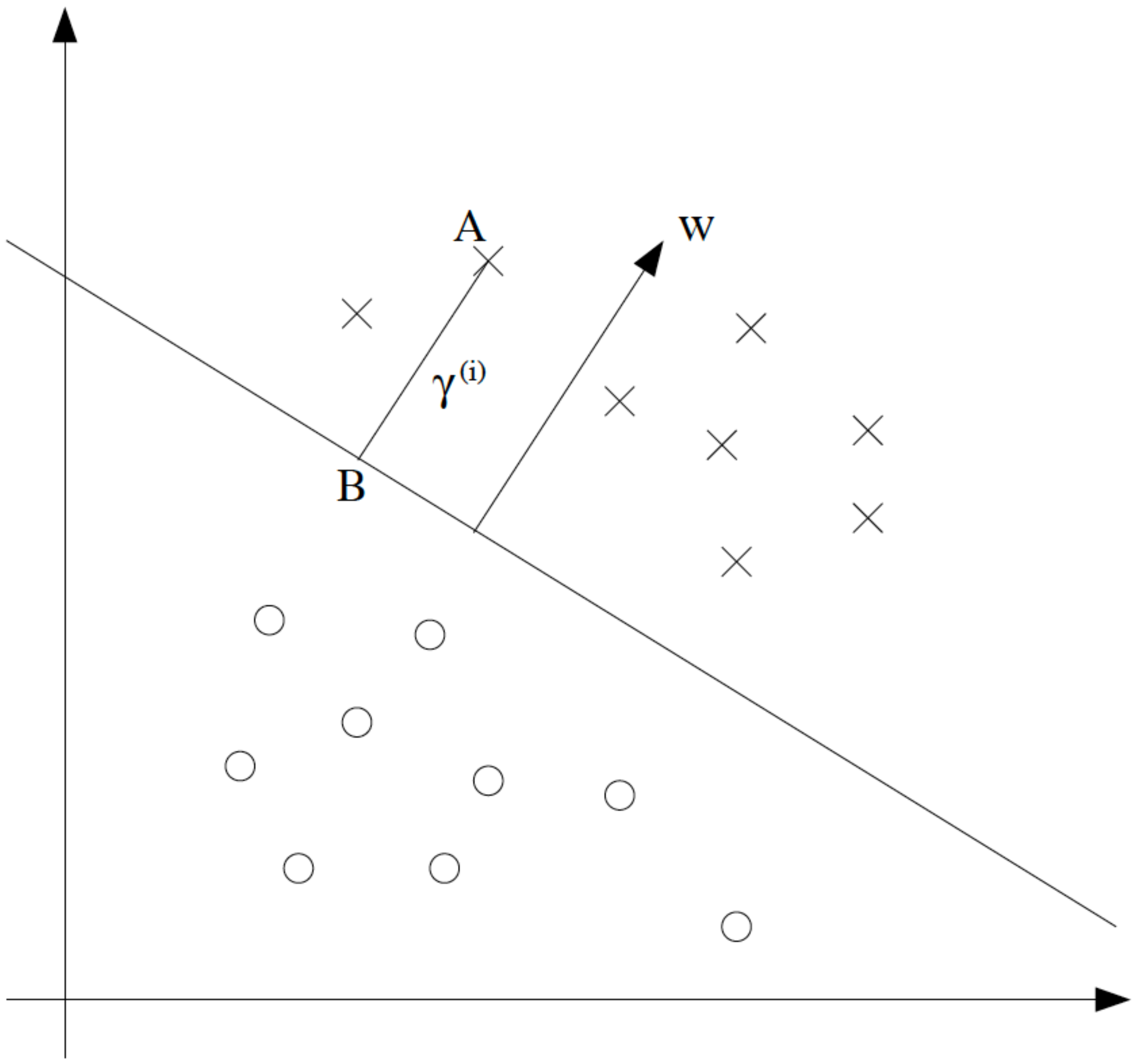
$$\textbf{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \quad i = 1, \ldots, N$$

$$\|w\| = 1$$

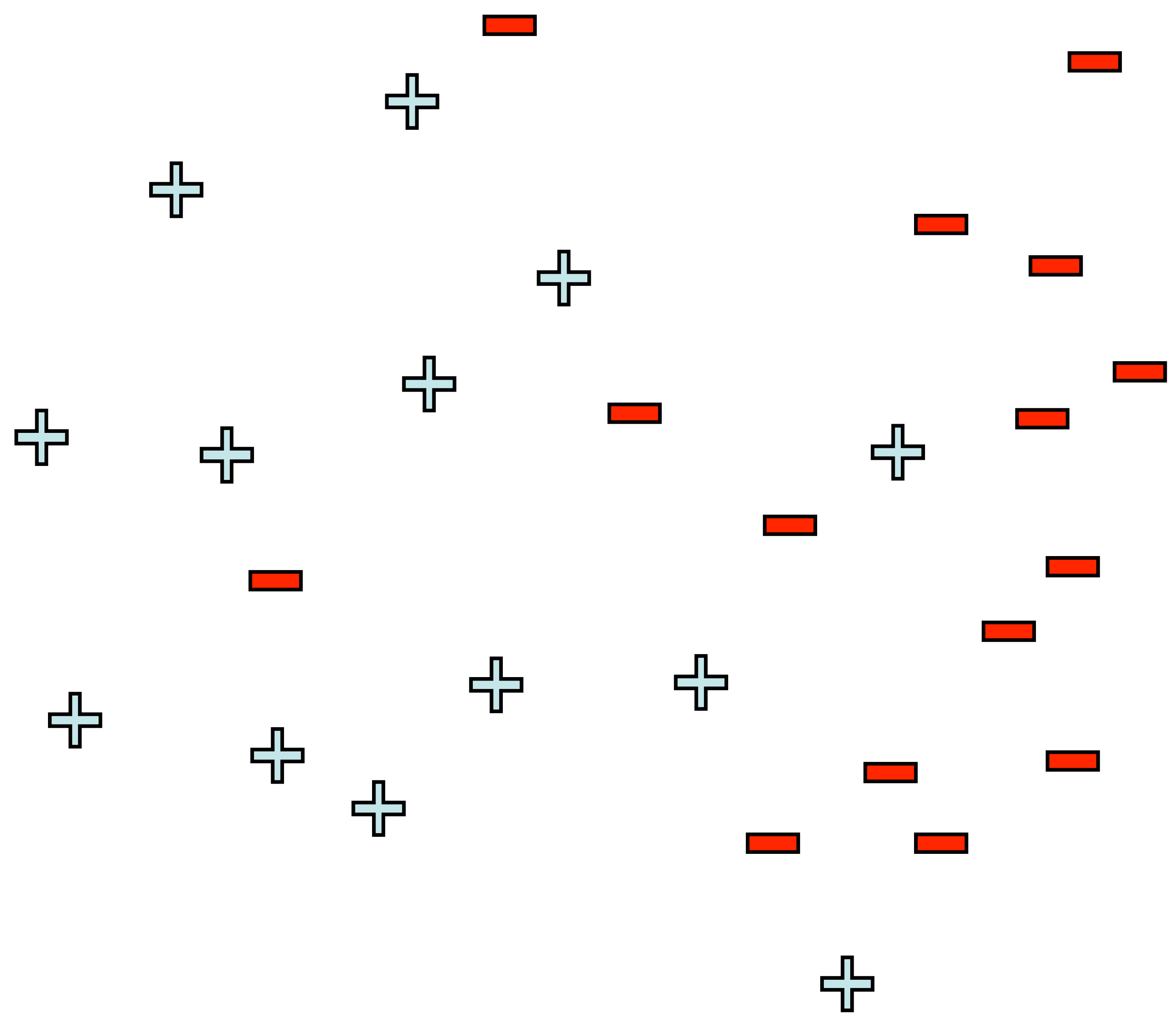
- Why is linear separability a key assumption?
- Otherwise, all $w, b, \gamma$ violate constraint

$$y^{(i)}(w^T x^{(i)} + b) \geq \gamma$$

Geometric margin $\gamma$ losses meaning



# Minimizing number of errors (0-1 loss)



- Try to find weights that violate as few constraints as possible?

$$\text{minimize}_{\mathbf{w},b} \quad \#(\text{mistakes})$$
$$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1 \qquad , \forall j$$

- Formalize this using the 0-1 loss:

$$\min_{\mathbf{w},b} \sum_j \ell_{0,1}(y_j, \, w \cdot x_j + b)$$
where $\ell_{0,1}(y, \hat{y}) = 1[y \neq \text{sign}(\hat{y})]$

- Unfortunately, minimizing 0-1 loss is NP-hard in the worst-case
  - Non-starter. We need another approach.

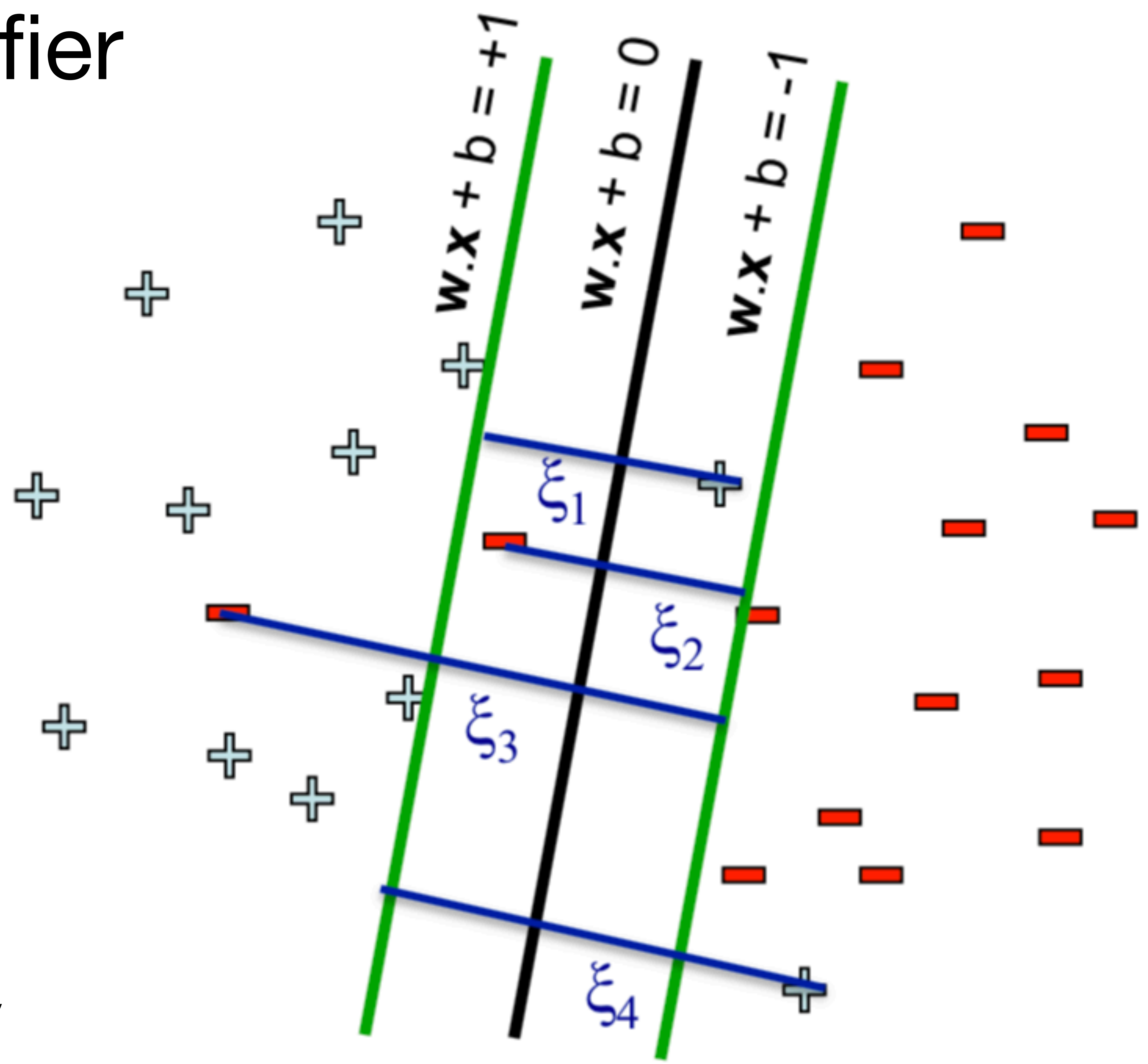
# Allow slack

- Let us ignore for a moment that we want to find the largest margin classifier
- Let us just find ***a*** classifier with minimal slack

$$\min_{w,b,\xi_i} \quad \sum_{i=1}^{N} \xi_i$$

$$\textbf{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i \, , i = 1, \ldots, N$$

$$\xi_i \geq 0, i = 1, \ldots, N$$



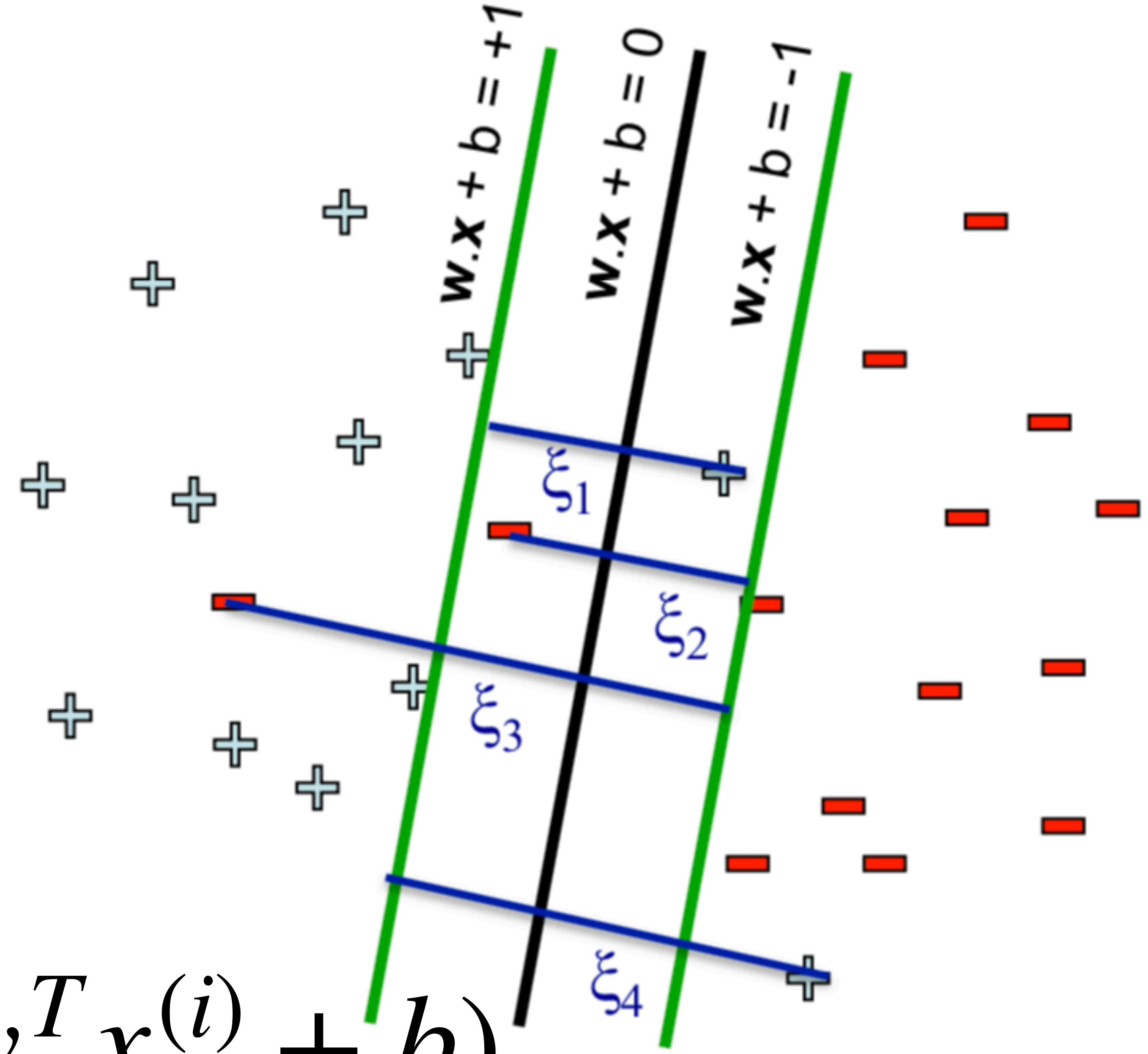

- If functional margin $\geq 1$, no penalty
- If functional margin $< 1$, pay linear penalty

# Optimal value of slack variables

$$\min_{w,b,\xi_i} \quad \sum_{i=1}^{N} \xi_i$$

**s.t.** $\quad y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, i = 1,\ldots,N$

$\quad \xi_i \geq 0, i = 1,\ldots,N$

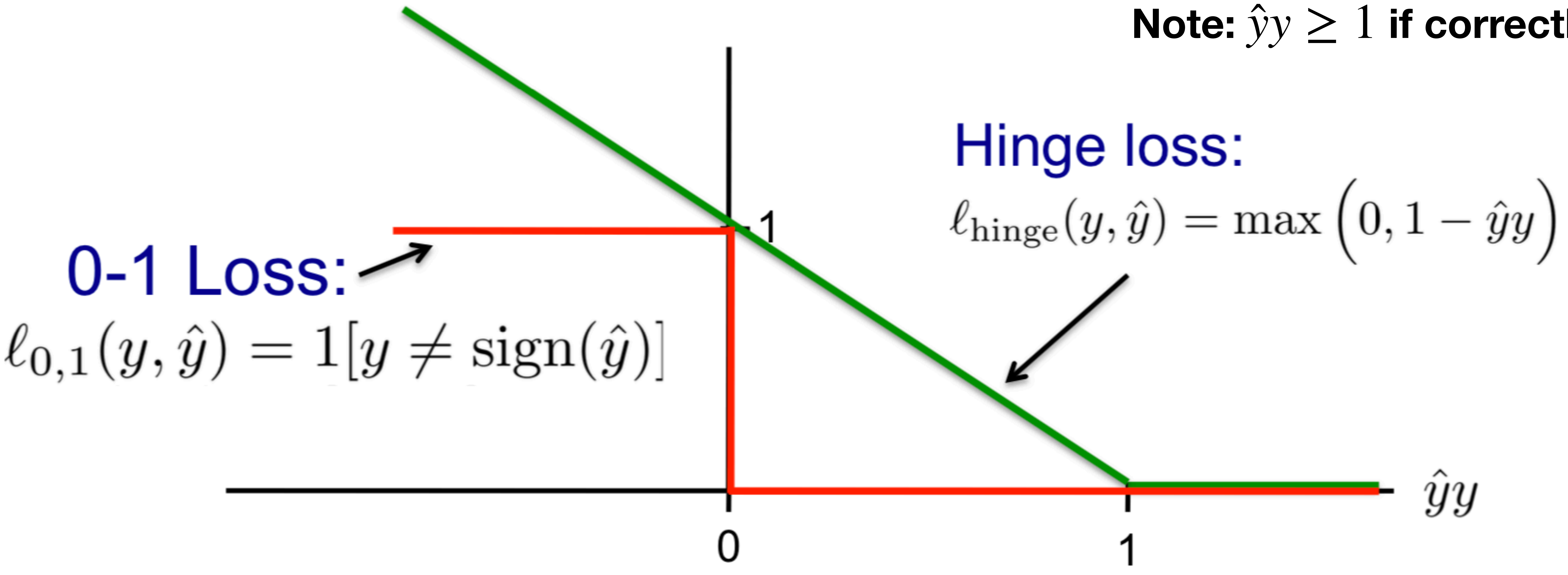



- Optimal value of slack variables
  - If $y^{(i)}(w^T x^{(i)} + b) \geq 1 \Rightarrow \xi_i = 0$
  - If $y^{(i)}(w^T x^{(i)} + b) < 1 \Rightarrow \xi_i = 1 - y^{(i)}(w^T x^{(i)} + b)$
- Write as

$$\xi_i = \max(0, 1 - y^{(i)}(w^T x^{(i)} + b))$$

# Hinge loss

**Note:** $\hat{y}y \geq 1$ **if correctly classified**




Tightest convex upper bound of the 0-1 loss

# Equivalent formulation of slack SVM

- With

$$\xi_i = \max(0, 1 - y^{(i)}(w^T x^{(i)} + b))$$

obtain equivalent optimization problem

$$\min_{w,b} \quad \sum_{i=1}^{N} \max(0, 1 - y^{(i)}\underbrace{(w^T x^{(i)} + b)}_{\hat{y}^{(i)}})$$

- The hinge loss is defined as

$$\ell_{\text{hinge}}(\hat{y}, y) = \max(0, 1 - \hat{y}y)$$

and the above corresponds to the empirical risk minimization

$$\min_{w,b} \sum_{i=1}^{N} \ell_{\text{hinge}}(w^T x^{(i)} + b, y^{(i)})$$

# SVM with slack

- Include that we want to find largest margin classifier with slack

$$\min_{w,b,\xi_i} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\textbf{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i\,, i = 1,\ldots,N$$

$$\xi_i \geq 0, i = 1,\ldots,N$$

- Have 2 terms in objective that are balanced by slack penalty $C$
  - If $C = \infty$, have to separate data
  - If $C = 0$, completely ignore data
  - Serves as a regularization parameter

# Equivalent formulation via hinge loss

$$
\begin{aligned}
\min_{w,b,\xi_i} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\
\textbf{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i\,, i = 1,\ldots,N \\
& \xi_i \geq 0, i = 1,\ldots,N
\end{aligned}
$$

$$\Downarrow$$

$$\min_{w,b}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\ell_{\textbf{hinge}}(y^{(i)}, w^T x^{(i)} + b)$$

- The term $\|w\|^2$ is a regularization to prevent overfitting to data
- The hinge loss ensure closeness to data

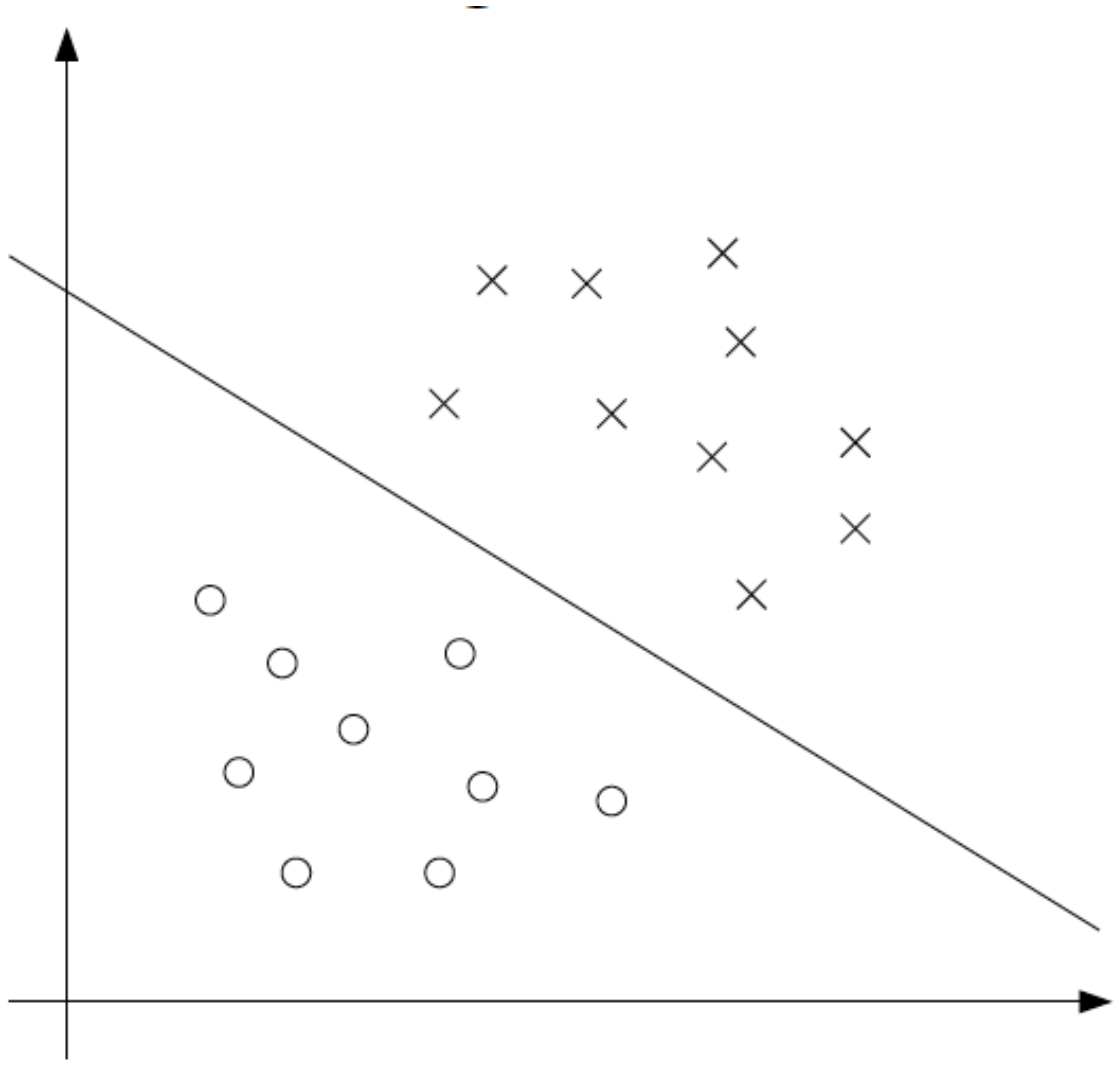
# Regularization

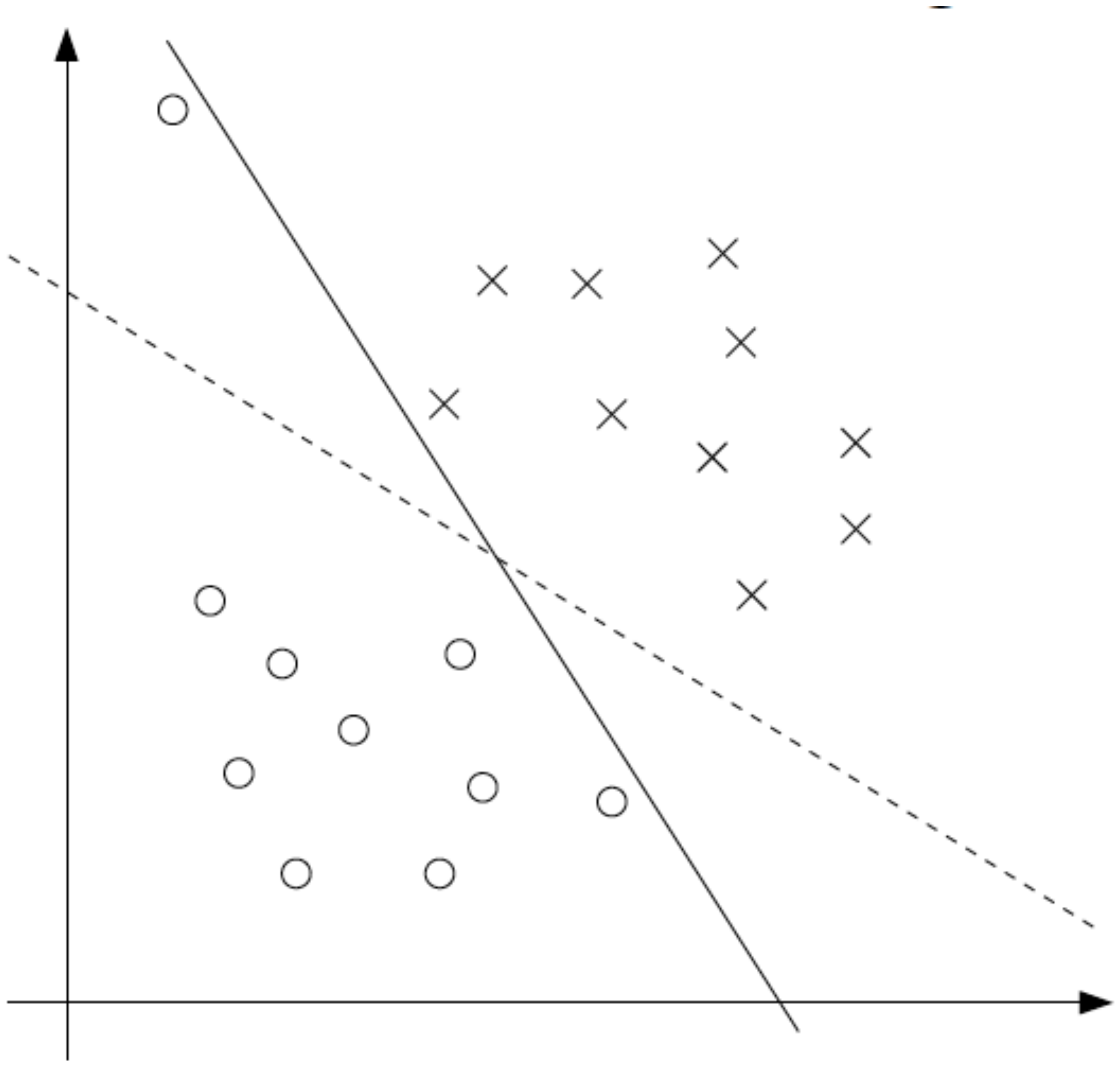
- Combine term that data ensures closeness to data with term that avoids overfitting by imposing structure such as smoothness

$$\min_{\theta} \sum_{i=1}^{N} \ell(y^{(i)}, h_\theta(x^{(i)})) + \alpha \mathscr{R}(\theta)$$

- Empirical risk based on loss $\ell$ ensures closeness to data
- Regularization term $\mathscr{R}(\theta)$ imposes structure
- Regularization parameter $\alpha$ trades off both objectives
- One way to find regularization parameter is via cross-validation

# Helps against outliers





- One outlier can lead to dramatic change in decision boundary
- Slack variables (hinge loss) helps to prevent this

# Dual formulation of slack SVM

$$
\begin{aligned}
\max_\alpha \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\
\text{s.t.} \quad & 0 \le \alpha_i \le C, \quad i = 1, \ldots, m \\
& \sum_{i=1}^{m} \alpha_i y^{(i)} = 0,
\end{aligned}
$$

$$
\begin{aligned}
\alpha_i = 0 \quad & \Rightarrow \quad y^{(i)}(w^T x^{(i)} + b) \ge 1 \\
\alpha_i = C \quad & \Rightarrow \quad y^{(i)}(w^T x^{(i)} + b) \le 1 \\
0 < \alpha_i < C \quad & \Rightarrow \quad y^{(i)}(w^T x^{(i)} + b) = 1.
\end{aligned}
$$