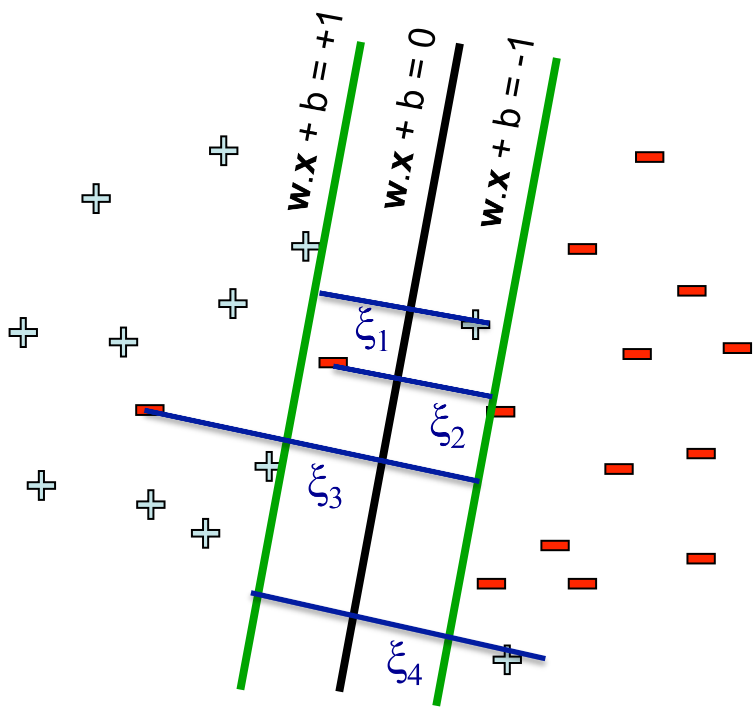
Support vector machines (SVMs)
Lecture 4

David Sontag
New York University

Slides adapted from Luke Zettlemoyer, Vibhav Gogate, and Carlos Guestrin

Key idea #1: Allow for *slack*



$$\text{minimize}_{\mathbf{w},b,\xi} \quad \Sigma_j \, \xi_j$$

$$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1 - \xi_j \quad , \forall j \quad \xi_j \geq 0$$

"slack variables"

Solve for the optimal value ξ_j^* as a function of **w** and b:

If $(w \cdot x_j + b)\, y_j \geq 1$, then $\xi_j^*(w,b) = 0$

If $(w \cdot x_j + b)\, y_j < 1$, then $\xi_j^*(w,b) = 1 - (w \cdot x_j + b)\, y_j$

$$\xi_j^* = \max\left(0, 1 - (w \cdot x_j + b)\, y_j\right)$$

Equivalent hinge loss formulation

$$\text{minimize}_{\mathbf{w},b,\xi} \quad \Sigma_j \, \xi_j$$

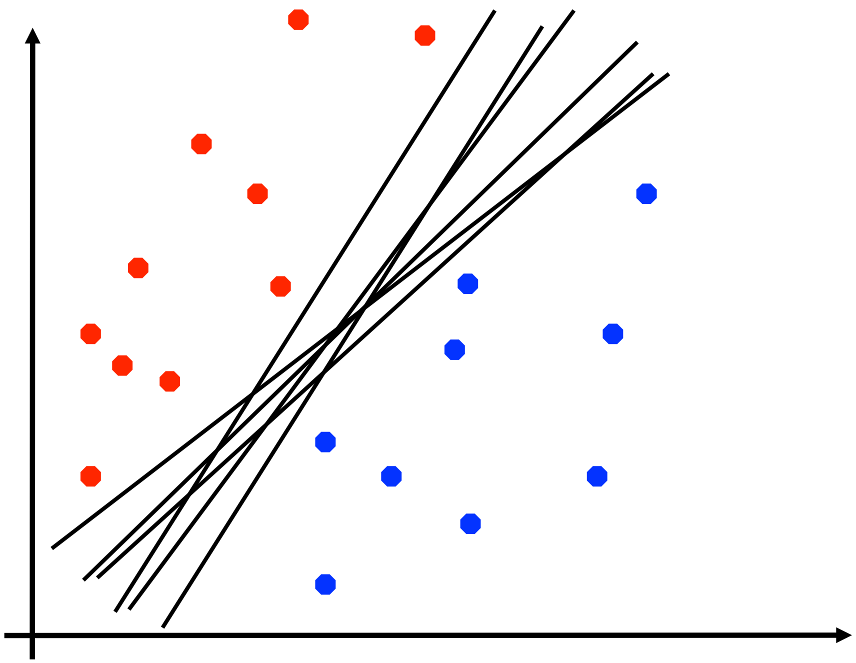
$$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1 - \xi_j \quad , \forall j \quad \xi_j \geq 0$$

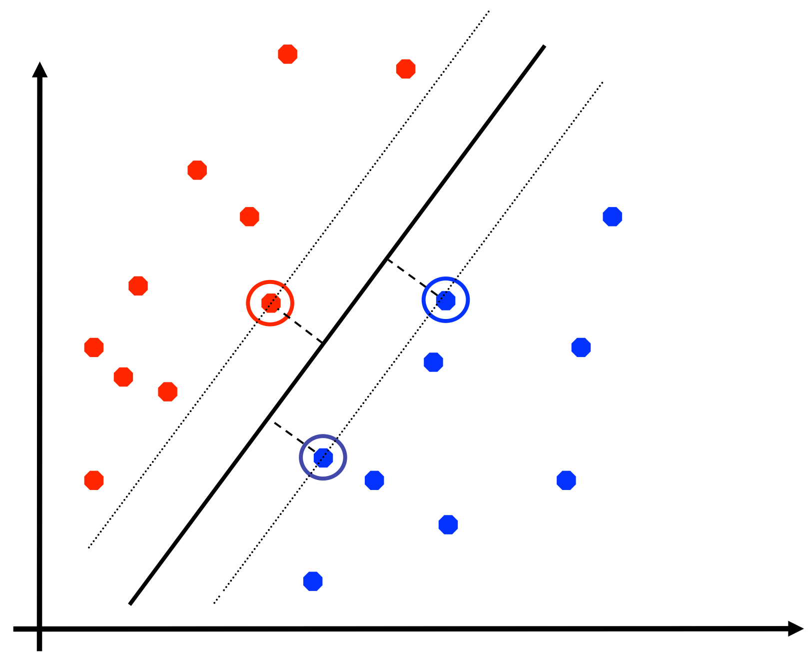
Substituting $\xi_j = \max\left(0, 1 - (w \cdot x_j + b)\, y_j\right)$ into the objective, we get:

$$\min_{w,b} \sum_j \max\left(0, 1 - (w \cdot x_j + b)\, y_j\right)$$

Now an *unconstrained* optimization problem. No longer a *linear* objective, but it is *convex.*

Key idea #2: seek large margin



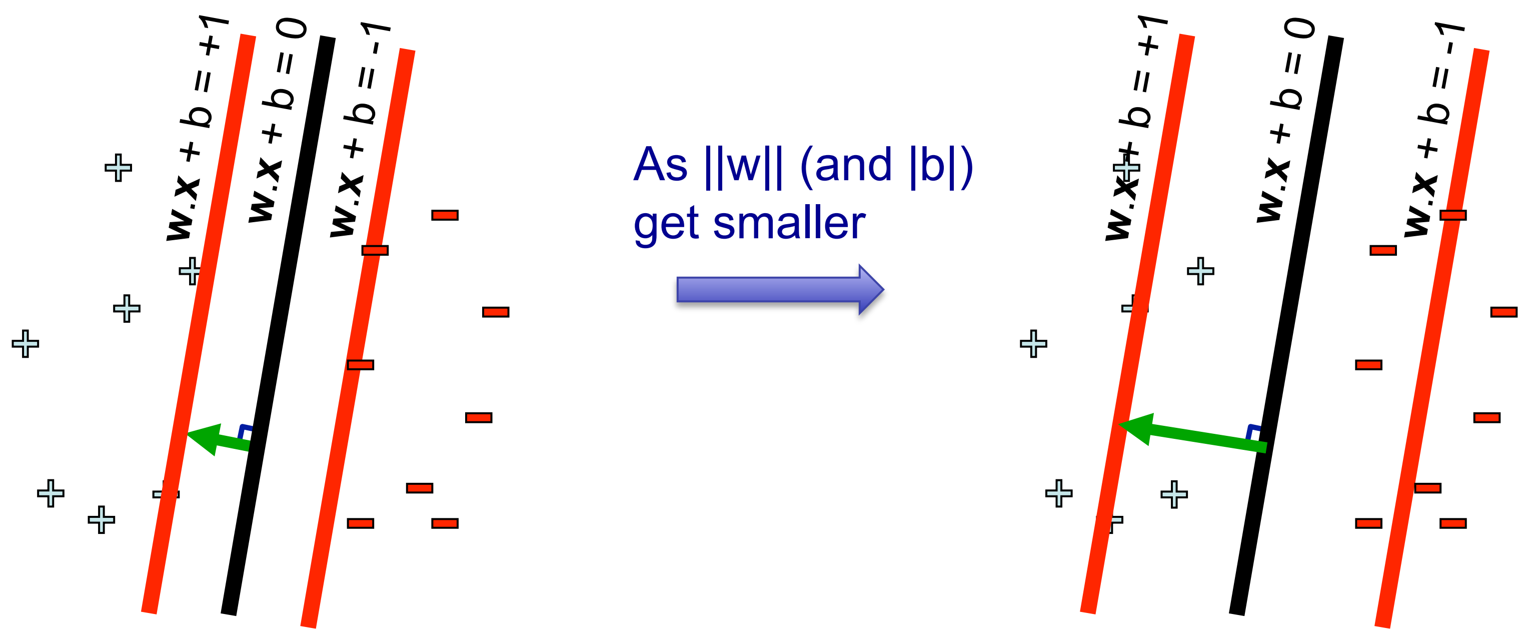


Key idea #2: seek large margin

- Consider the constraints:

$$y_t \left(w \cdot x_t + b \right) \geq 1 \quad \forall t$$

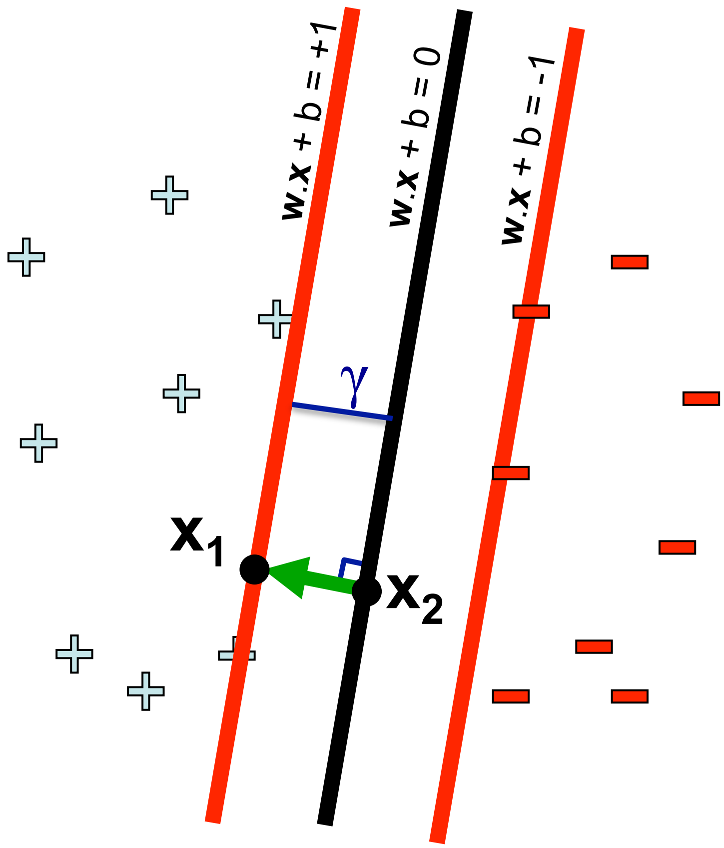
- As the norm of the weight vector ||w|| and b get **smaller**, the optimization problem becomes infeasible:



What is γ (geometric margin) as a function of **w**?

$$\gamma_i = \text{Distance to } i\text{'th data point}$$

$$\gamma = \min_i \gamma_i$$



$$w \cdot x_1 + b = 1$$

-

$$w \cdot x_2 + b = 0$$

$$w \cdot (x_1 - x_2) = 1$$

Plug in

We also know that:

$$x_1 - x_2 = \gamma \frac{w}{||w||}$$

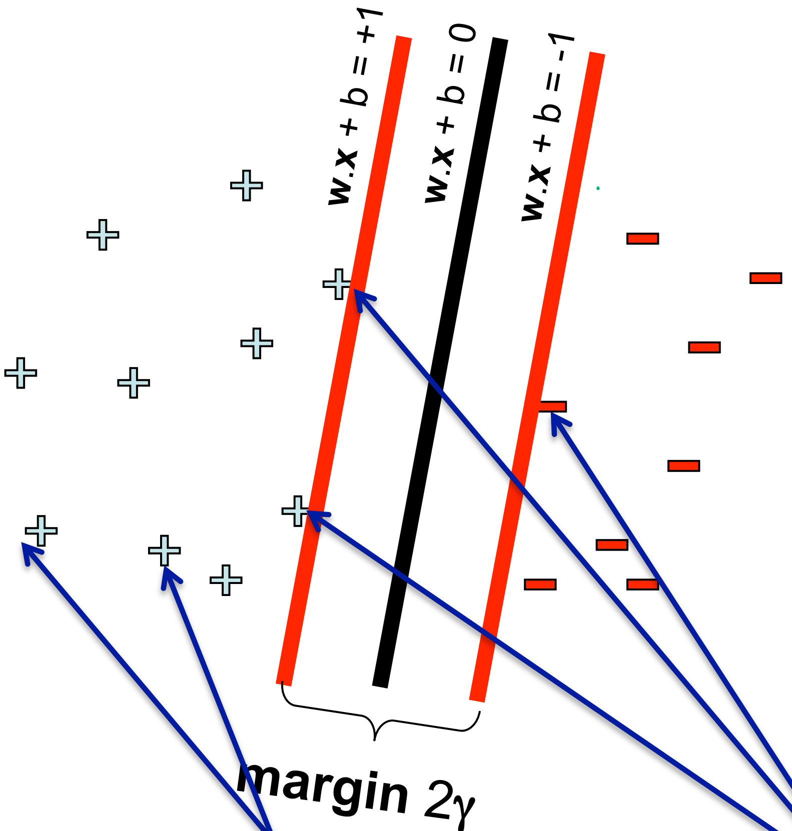
$$1 = w \cdot \left(\gamma \frac{w}{||w||}\right) = \frac{\gamma}{||w||} w \cdot w = \gamma ||w||$$

So, $\gamma = \dfrac{1}{||w||}$

(assuming there is a data point on the **w.x** + b = +1 or -1 line)

Final result: can maximize γ by minimizing $||w||_2$!!!

(Hard margin) support vector machines



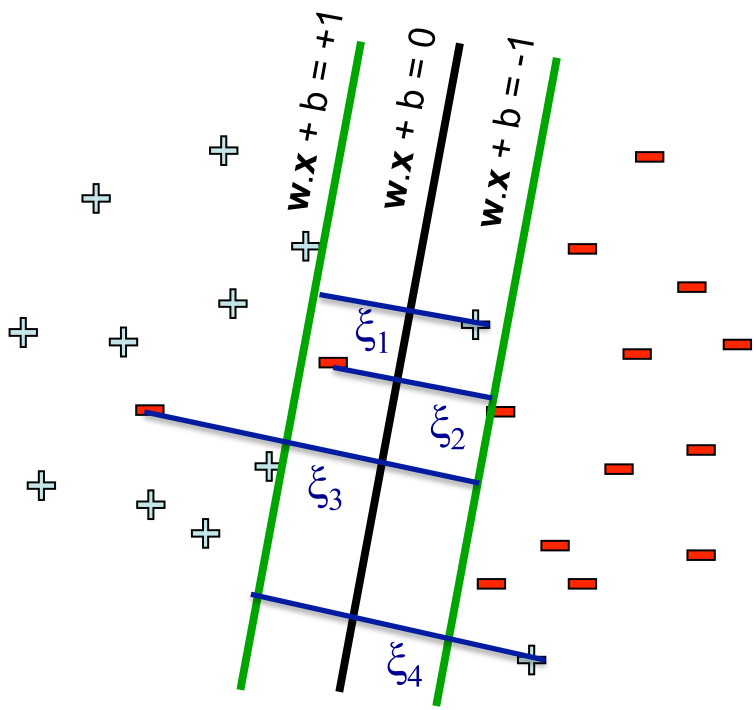
$$\text{minimize}_{\mathbf{w},b} \quad \mathbf{w}.\mathbf{w}$$
$$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1, \ \ \forall j$$

- Example of a **convex optimization** problem
  - A quadratic program
  - Polynomial-time algorithms to solve!
- Hyperplane defined by **support vectors**
  - Could use them as a lower-dimension basis to write down line, although we haven't seen how yet
- More on these later

Non-support Vectors:
- everything else
- moving them will not change **w**

Support Vectors:
- data points on the canonical lines

Allowing for slack: "Soft margin SVM"



$$\text{minimize}_{\mathbf{w},b} \quad \mathbf{w}.\mathbf{w} + C\,\Sigma_j\,\xi_j$$

$$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1 - \xi_j \quad, \forall j \quad \xi_j \geq 0$$

"slack variables"

Slack penalty $C > 0$:

- $C=\infty$ → have to separate the data!
- $C=0$ → ignores the data entirely!
- **Select using cross-validation**

For each data point:

- •If margin ≥ 1, don't care
- •If margin < 1, pay linear penalty

Equivalent formulation using hinge loss

$$\text{minimize}_{\mathbf{w},b} \quad \mathbf{w}.\mathbf{w} + C\,\Sigma_j\,\xi_j$$

$$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1 - \xi_j \quad ,\forall j \quad \xi_j \geq 0$$

Substituting $\xi_j = \max\left(0, 1 - (w \cdot x_j + b)\, y_j\right)$ into the objective, we get:

$$\boxed{\min ||w||^2 + C \sum_j \max\left(0, 1 - (w \cdot x_j + b)\, y_j\right)}$$

Recall, the **hinge loss** is $\ell_{\text{hinge}}(y, \hat{y}) = \max\Big(0, 1 - \hat{y}y\Big)$

$$\min_{\mathbf{w},b} ||w||_2^2 + C \underbrace{\sum_j \ell_{\text{hinge}}(y_j,\, w \cdot x_j + b)}$$

This is called **regularization**; used to prevent overfitting!

This part is empirical risk minimization, using the hinge loss