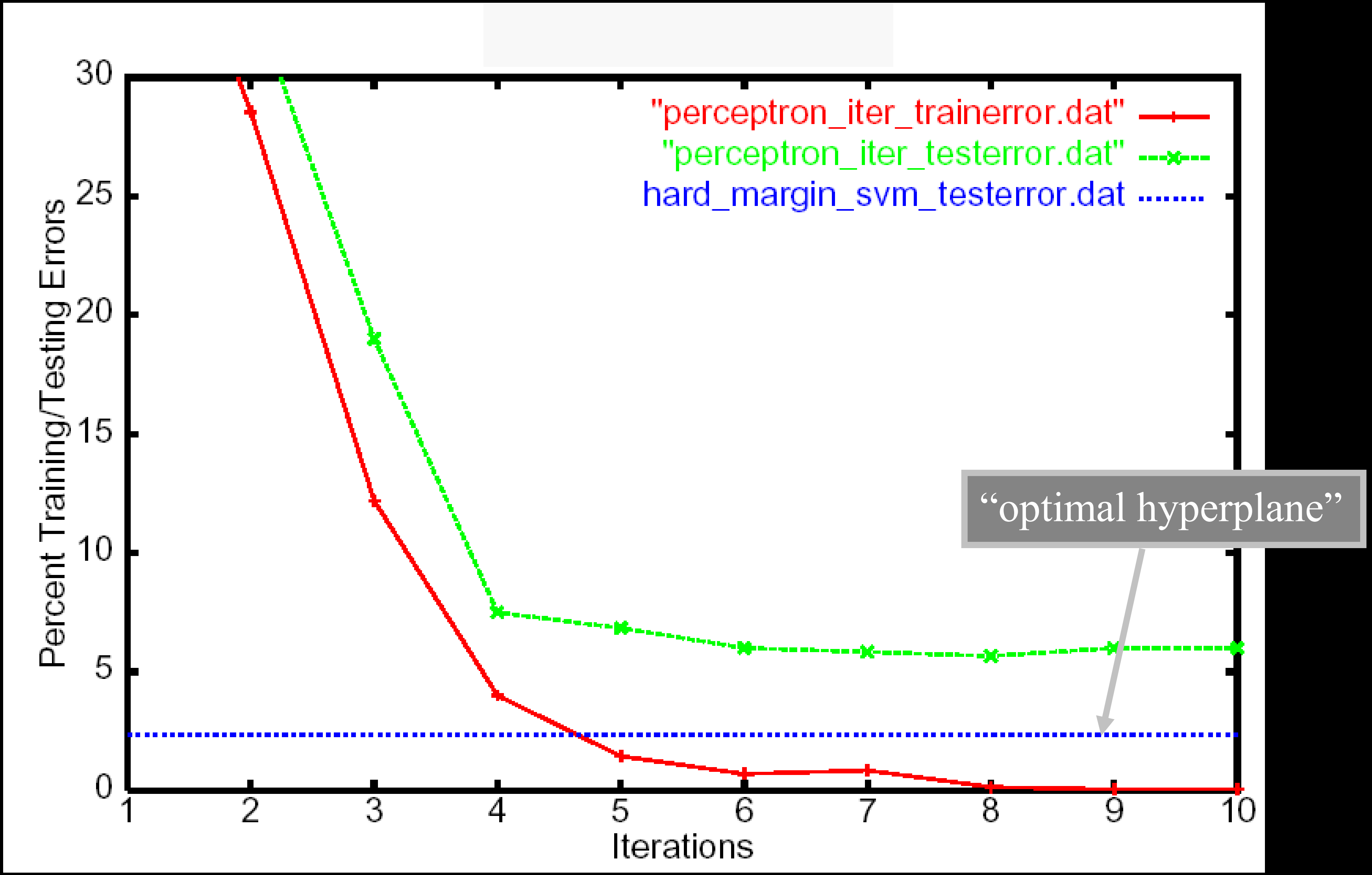
Optimal Hyperplanes and Support Vector Machines

CS4780/5780 – Machine Learning
Fall 2019

Nika Haghtalab & Thorsten Joachims
Cornell University

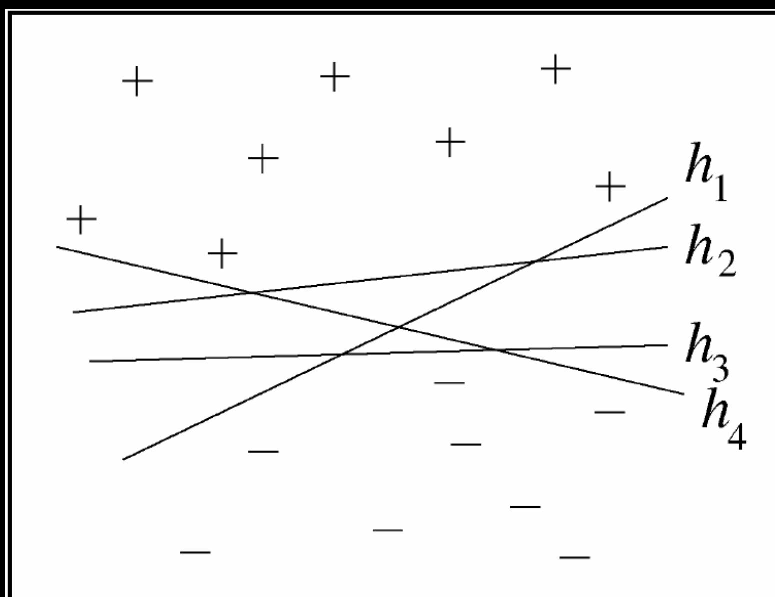
Reading: UML 15.1, 15.2

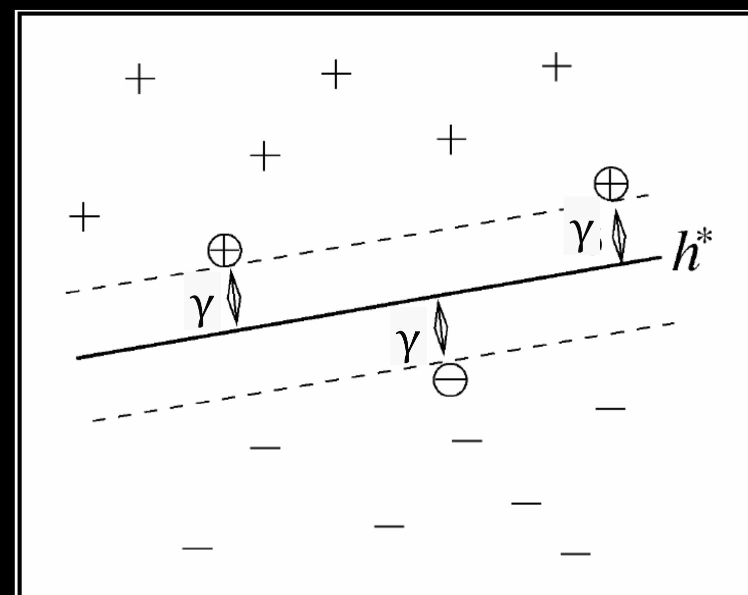
Example: Reuters Text Classification



Optimal Hyperplanes

- Assumption:
  - Training examples are linearly separable.





Margin of a Linear Classifier

- Definition: For a linear classifier h_w, the margin γ of an example (x, y) with $x \in \Re^N$ and $y \in \{-1, +1\}$ is

$$\gamma = y(w \cdot x + b)$$

- Definition: The margin is called geometric margin, if $||w|| = 1$. For general w, the term functional margin is used to indicate that the norm of w is not necessarily 1.

- Definition: The (hard) margin of a homogeneous linear classifier h_w on sample S is

$$\gamma = \min_{(x,y)\in S} y(w \cdot x + b)$$

Hard-Margin Separation

- Goal:
  - Find hyperplane with the largest distance to the closest training examples.

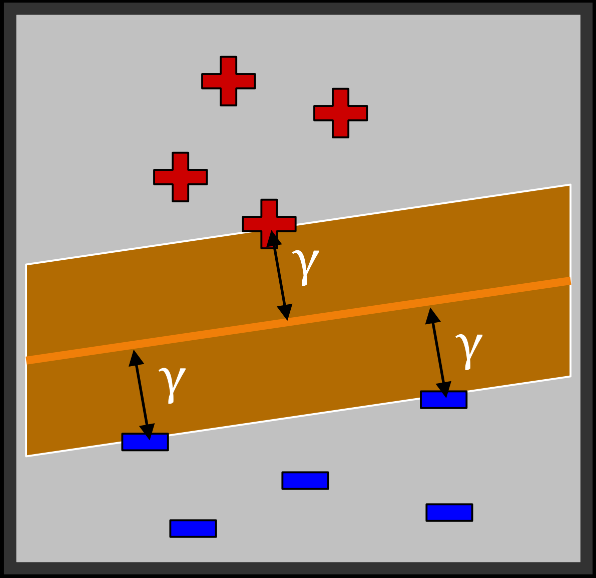
Optimization Problem (Primal):

$$\min_{\vec{w},b} \quad \frac{1}{2}\vec{w}\cdot\vec{w}$$

$$s.t. \quad y_1(\vec{w}\cdot\vec{x}_1 + b) \geq 1$$

$$\ldots$$

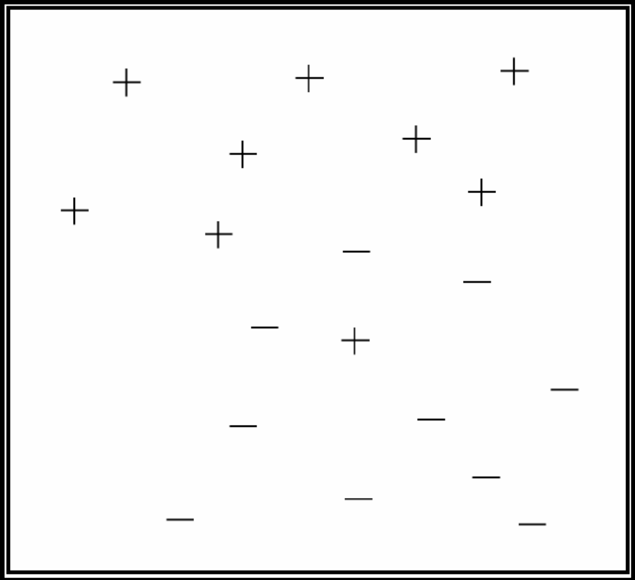
$$y_n(\vec{w}\cdot\vec{x}_n + b) \geq 1$$

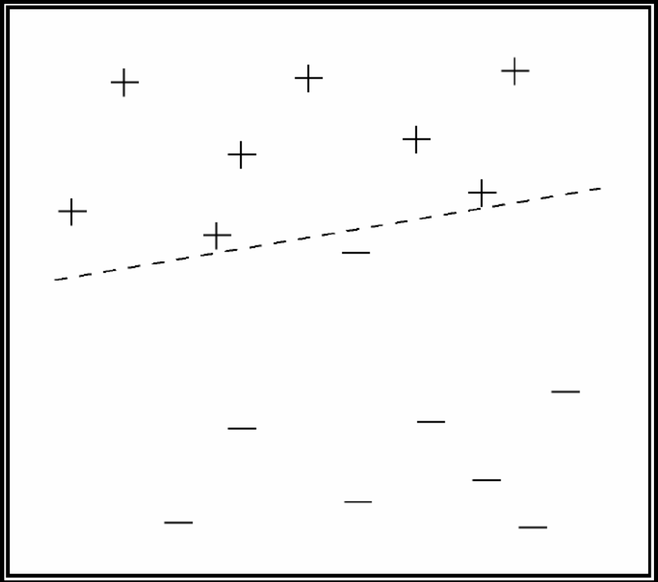


- Support Vectors:
  - Examples with minimal distance (i.e. margin).

Non-Separable Training Data

- Limitations of hard-margin formulation
  - For some training data, there is no separating hyperplane.
  - Complete separation (i.e. zero training error) can lead to suboptimal prediction error.





Soft-Margin Separation

Idea: Maximize margin and minimize training

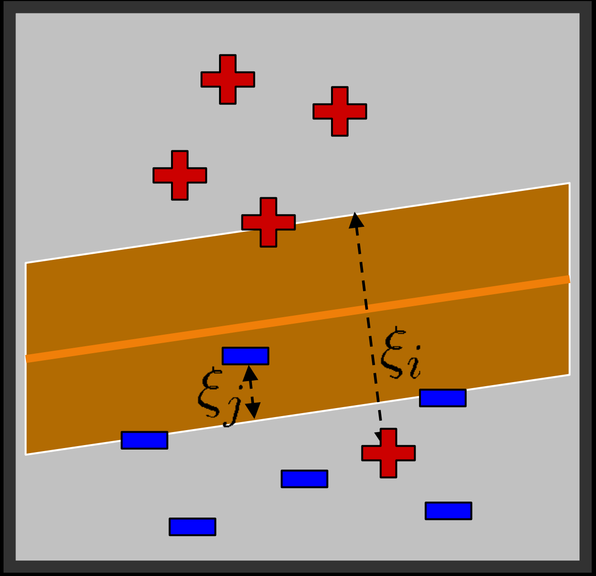
Hard-Margin OP (Primal):

$$\min_{\vec{w},b} \quad \frac{1}{2}\vec{w}\cdot\vec{w}$$
$$s.t. \quad y_1(\vec{w}\cdot\vec{x}_1+b)\geq 1$$
$$\ldots$$
$$y_n(\vec{w}\cdot\vec{x}_n+b)\geq 1$$

Soft-Margin OP (Primal):

$$\min_{\vec{w},\vec{\xi},b} \frac{1}{2}\vec{w}\cdot\vec{w}+C\sum_{i=1}^{n}\xi_i$$
$$s.t. \; y_1(\vec{w}\cdot\vec{x}_1+b)\geq 1-\xi_1 \wedge \xi_1 \geq 0$$
$$\ldots$$
$$y_n(\vec{w}\cdot\vec{x}_n+b)\geq 1-\xi_n \wedge \xi_n \geq 0$$

- Slack variable ξ_i measures by how much (x_i,y_i) fails to achieve margin γ
- $\Sigma\xi_i$ is upper bound on number of training errors
- C is a parameter that controls trade-off between margin and training error.



Controlling Soft-Margin Separation

- $\Sigma\xi_i$ is upper bound on number of training errors
- C is a parameter that controls trade-off between margin and training error.

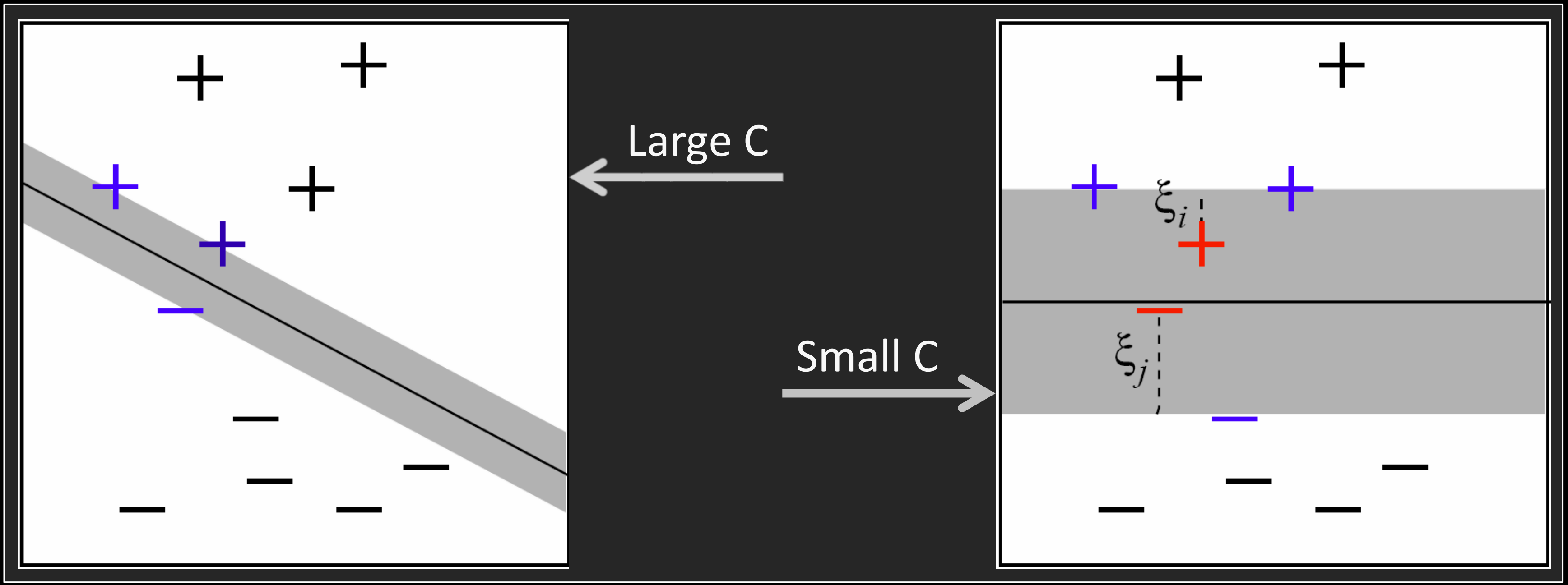
Soft-Margin OP (Primal):

$$\min_{\vec{w},\vec{\xi},b} \frac{1}{2}\vec{w}\cdot\vec{w} + C\sum_{i=1}^{n}\xi_i$$

$$s.t.\ y_1(\vec{w}\cdot\vec{x}_1 + b) \geq 1 - \xi_1 \wedge \xi_1 \geq 0$$

$$\ldots$$

$$y_n(\vec{w}\cdot\vec{x}_n + b) \geq 1 - \xi_n \wedge \xi_n \geq 0$$



Example Reuters “acq”: Varying C