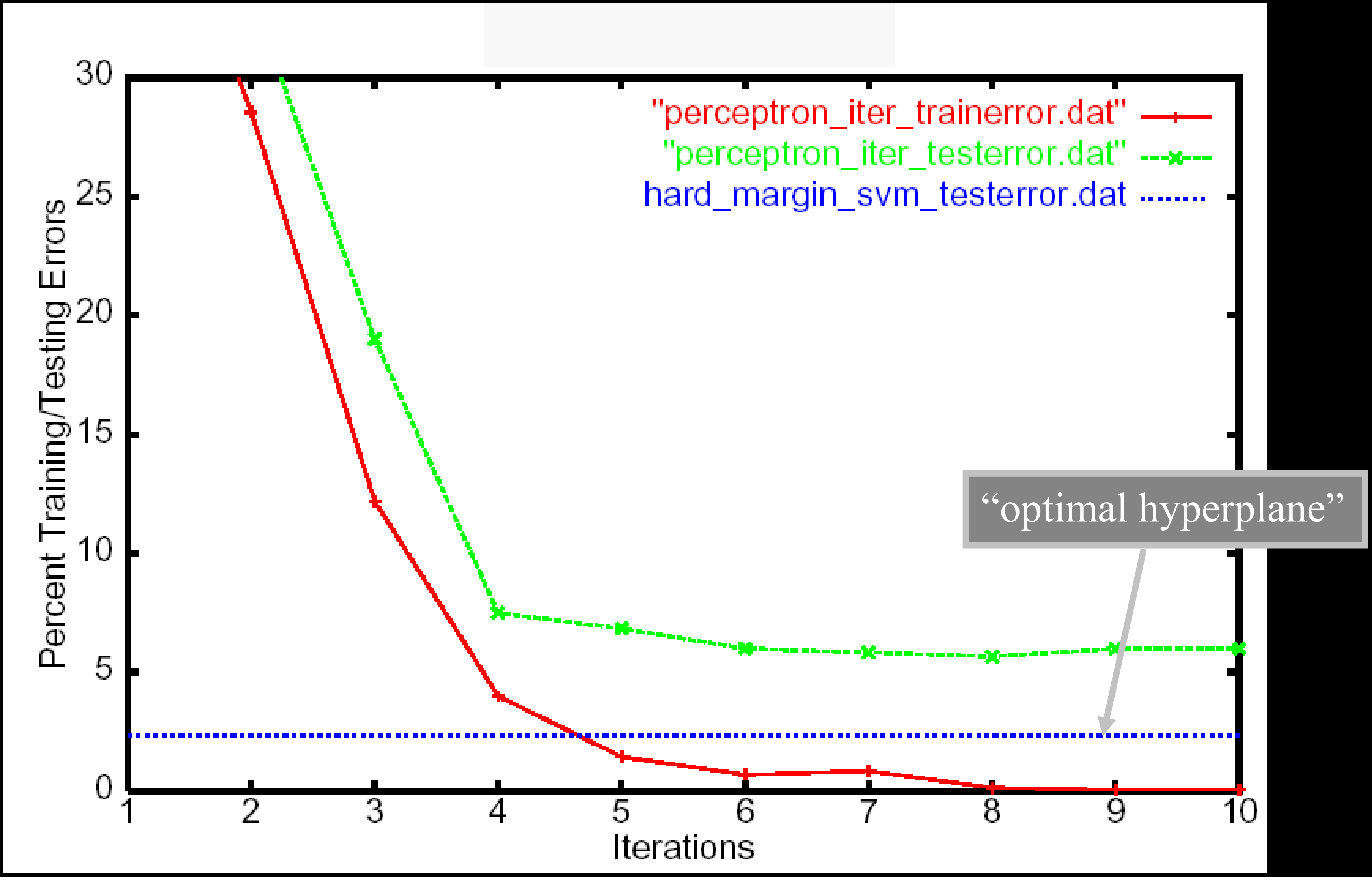
Support Vector Machines: Optimal Hyperplanes

CS6780 – Advanced Machine Learning
Spring 2015

Thorsten Joachims
Cornell University

Reading: Murphy 14.5
Schoelkopf/Smola Chapter 7.1-7.3, 7.5

Example: Reuters Text Classification



VC Dimension of Margin Hyperplanes

Theorem: Unbiased linear classifiers H_X with $\|w\| = 1/\delta$ and $\max_{\mathrm{i}}\|x_i\| \leq R$ and margin

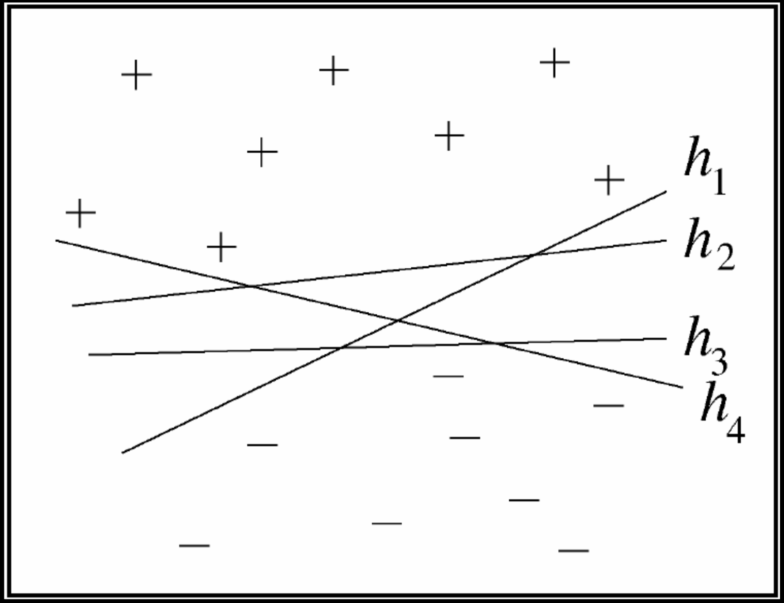
$$\min_i |w \cdot x_i| = 1$$

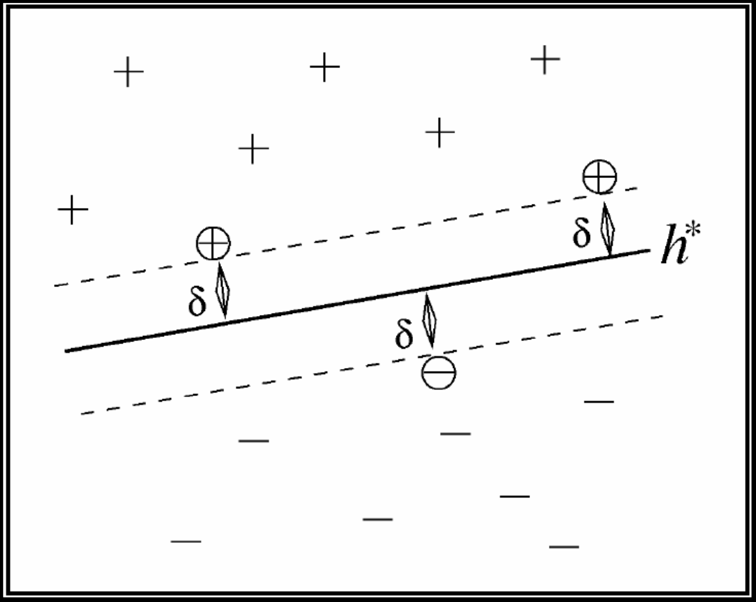
for a given set of instances $X = \{x_1, \dots, x_k\}$, have VC Dimension

$$VCDim(H_X) \leq \frac{R^2}{\delta^2}$$

Optimal Hyperplanes

- Assumption:
  - Training examples are linearly separable.





Margin of a Linear Classifier

Definition: *For a linear classifier h_w, the* **margin** δ *of an example* $(\vec{x}, y)$ *with* $\vec{x} \in \Re^N$ *and* $y \in \{-1, +1\}$ *is* $\delta = y(\vec{w} \cdot \vec{x})$.

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

Definition: *The (hard) margin of an unbiased linear classifier* $h_{\vec{w}}$ *on a sample* S *is* $\delta = min_{(\vec{x},y)\in S} y(\vec{w} \cdot \vec{x})$.

Definition: *The (hard) margin of an unbiased linear classifier* $h_{\vec{w}}$ *on a task* $P(X, Y)$ *is*

$$\delta = inf_{S \sim P(X,Y)} min_{(\vec{x},y)\in S} y(\vec{w} \cdot \vec{x}).$$

Hard-Margin Separation

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

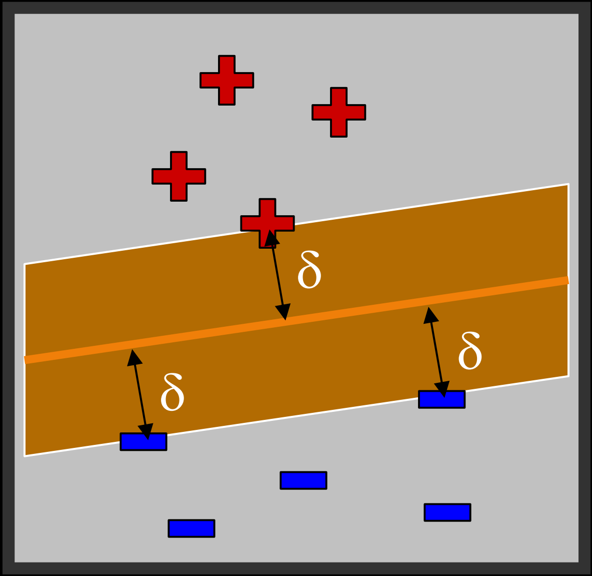
Optimization Problem (Primal):

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

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

$$\ldots$$

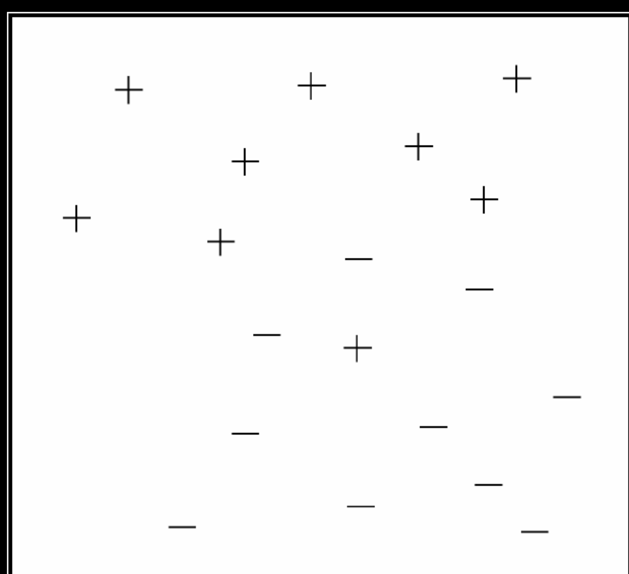
$$y_n(\vec{w}\cdot\vec{x}_n + b) \geq 1$$

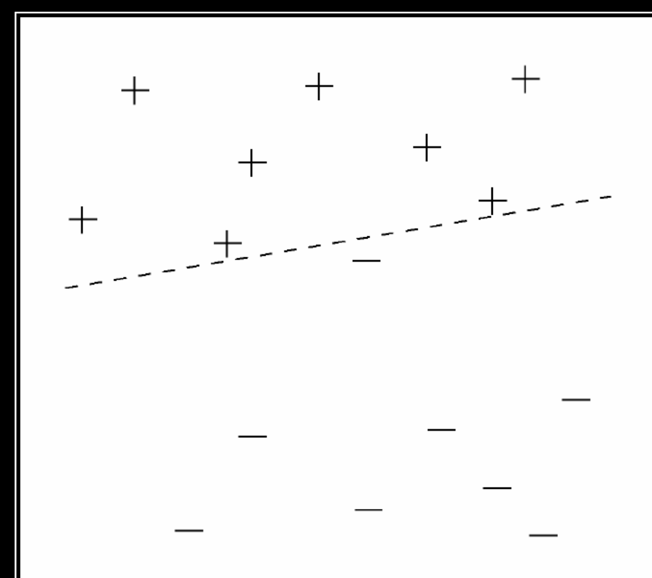


- Support Vectors:
  - Examples with minimal distance (i.e. margin).

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

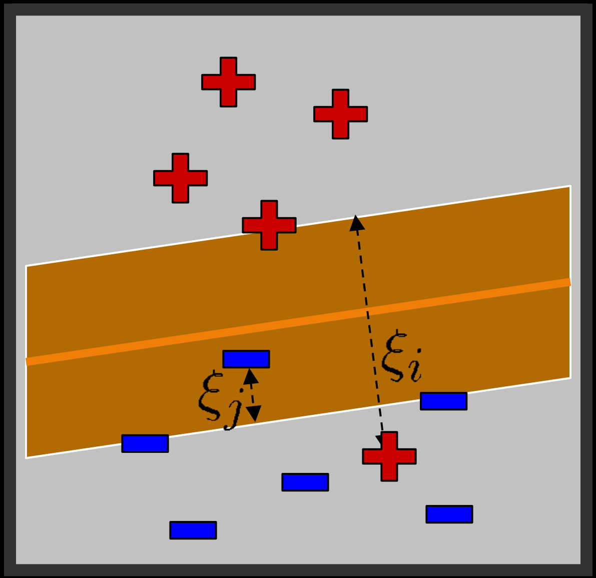
Hard-Margin OP (Primal):

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

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 δ
- $\Sigma\xi_i$ 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.

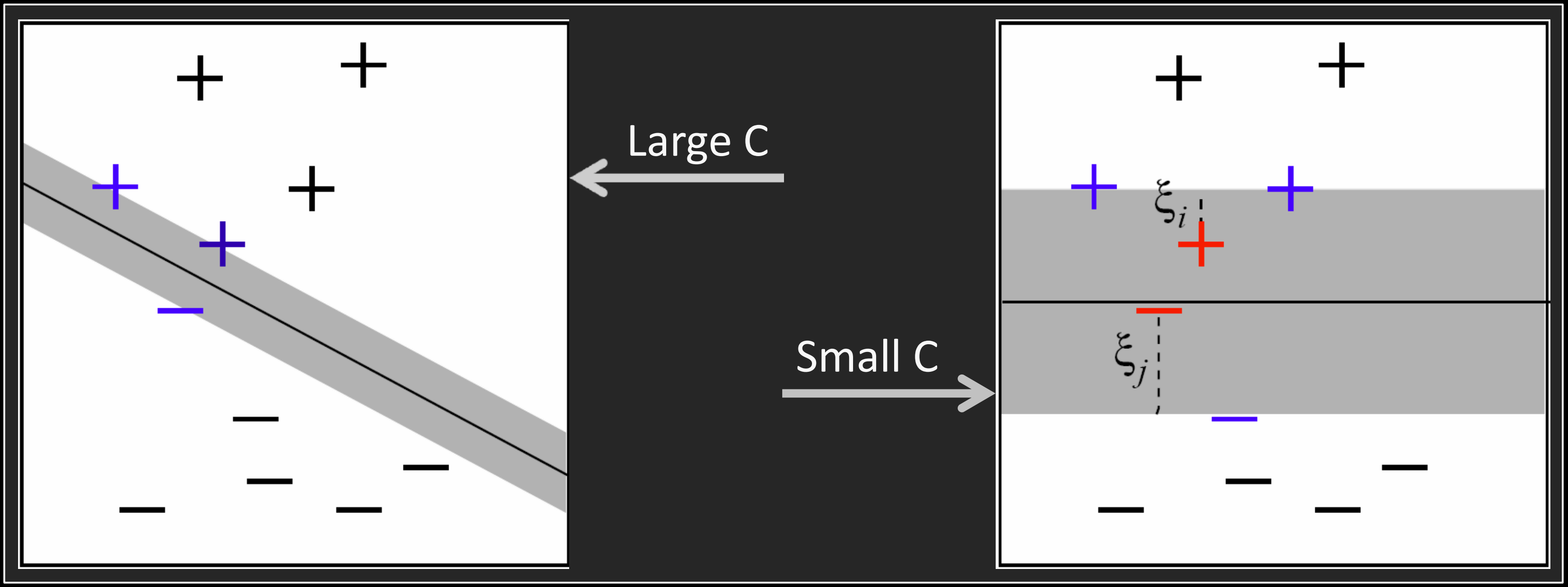
Soft-Margin OP (Primal):

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

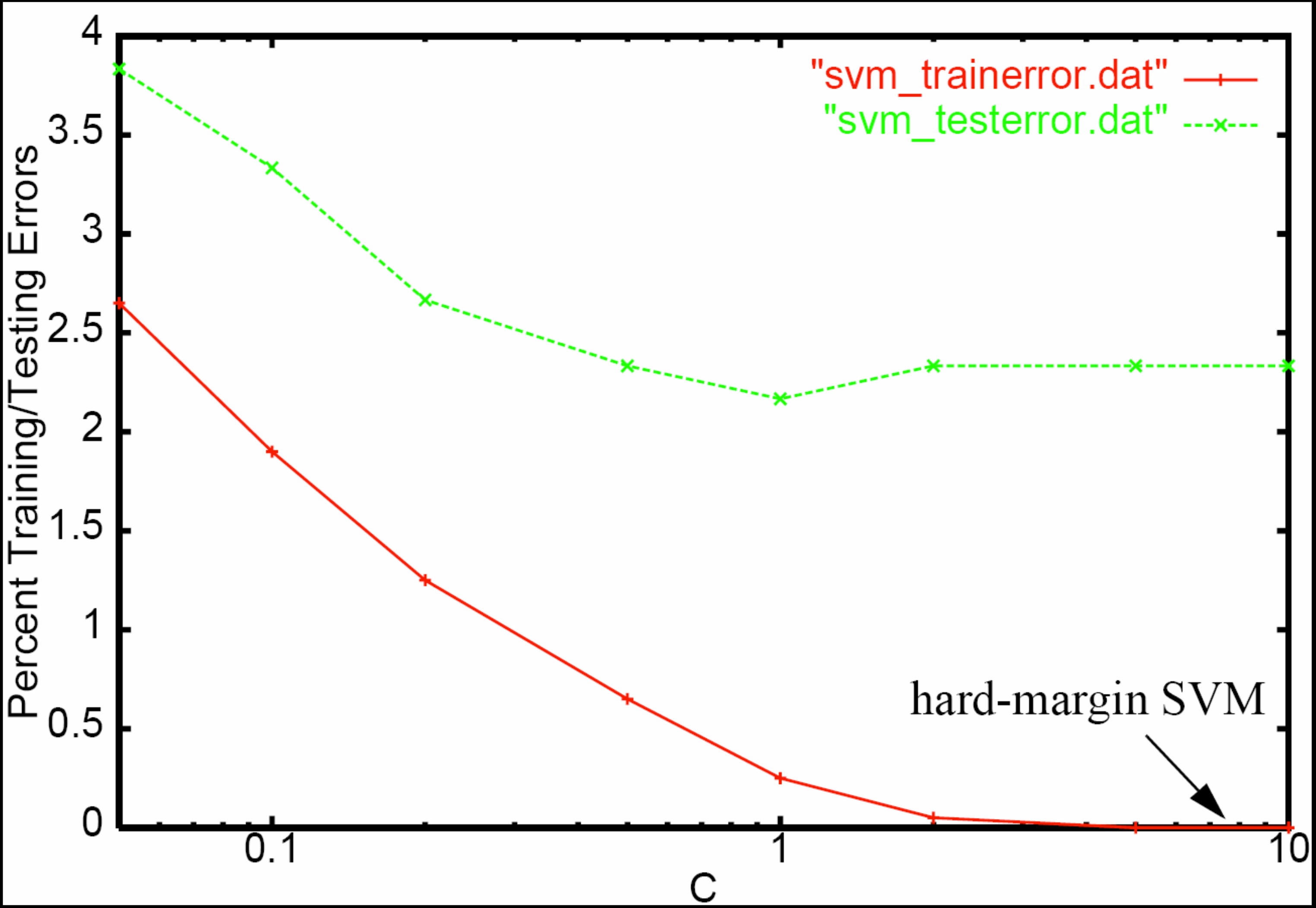
$$s.t.\ y_1(\vec{w}\cdot\vec{x}_1 + b) \geq 1-\xi_1 \wedge \xi_1 \geq 0$$

$$\dots$$

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



Example Reuters "acq": Varying C



Example: Margin in High-Dimension

Training Sample S_{train}	$\vec{x}$ x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
$(\vec{x}_1, y_1)$	1	0	0	1	0	0	0	1
$(\vec{x}_2, y_2)$	1	0	0	0	1	0	0	1
$(\vec{x}_3, y_3)$	0	1	0	0	0	1	0	-1
$(\vec{x}_4, y_4)$	0	1	0	0	0	0	1	-1
	$\vec{w}$ w_1	w_2	w_3	w_4	w_5	w_6	w_7	**b**
Hyperplane 1	1	1	0	0	0	0	0	2
Hyperplane 2	0	0	0	1	1	-1	-1	0
Hyperplane 3	1	-1	1	0	0	0	0	0
Hyperplane 4	0.5	-0.5	0	0	0	0	0	0
Hyperplane 5	1	-1	0	0	0	0	0	0
Hyperplane 6	0.95	-0.95	0	0.05	0.05	-0.05	-0.05	0
Hyperplane 7	0.67	-0.67	0	0.33	0.33	-0.33	-0.33	0