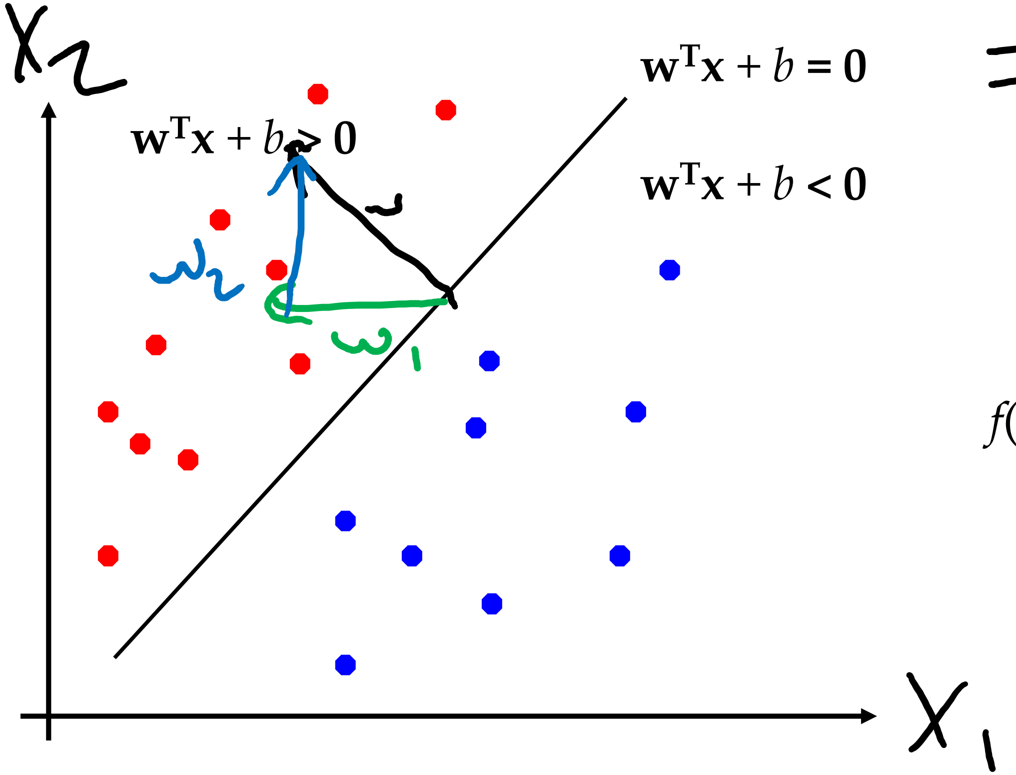
CS 412

FEB 13TH SVM

HTF – CHAPTER 12

Linear Separators

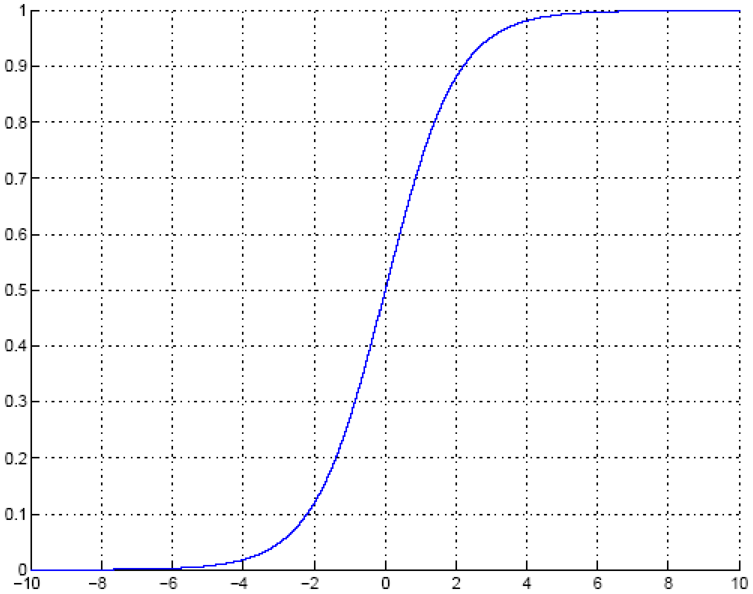
Binary classification can be viewed as the task of separating classes in feature space:



$= X_1 \cdot w_1 + X_2 \cdot w_2 + b_0$

$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$ Red are the positive points

Sigmoid (Logistic) Function



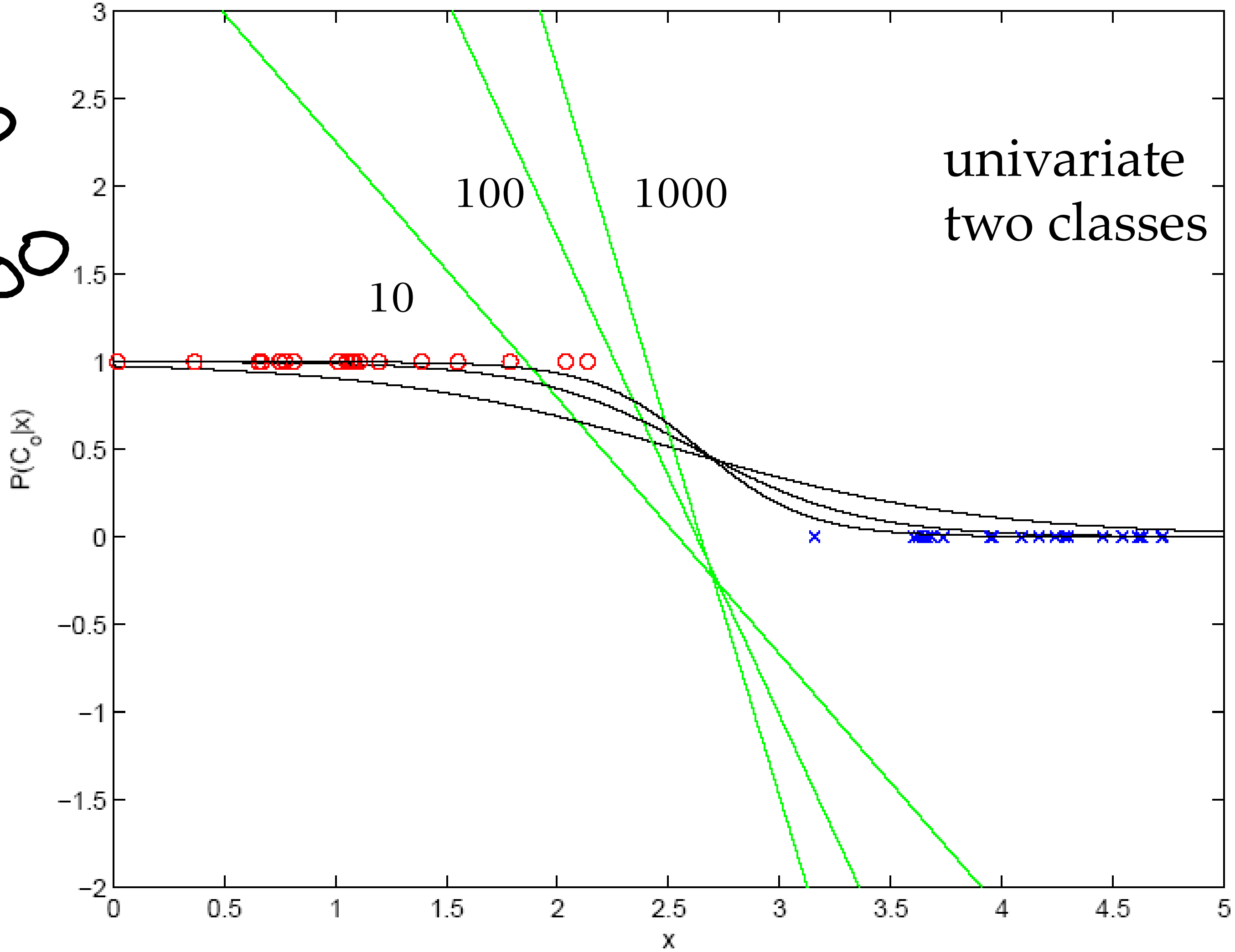
Calculate $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ and choose C_1 if $g(\mathbf{x}) > 0$, or

Calculate $y = \text{sigmoid}\left(\mathbf{w}^T\mathbf{x} + w_0\right)$ and choose C_1 if $y > 0.5$

$$= \text{sigmoid}(a), \text{where } a = \mathbf{w}^T\mathbf{x} + w_0 \qquad \frac{dy}{da} = y(1-y)$$

after 10, 100, 1000 iterations

$w_1 = 600$
$w_2 = 400$
$w_0 = 1000$



$w_1 = 6$
$w_2 = 4$
$w_0 = 10$

lowest error for linear regression

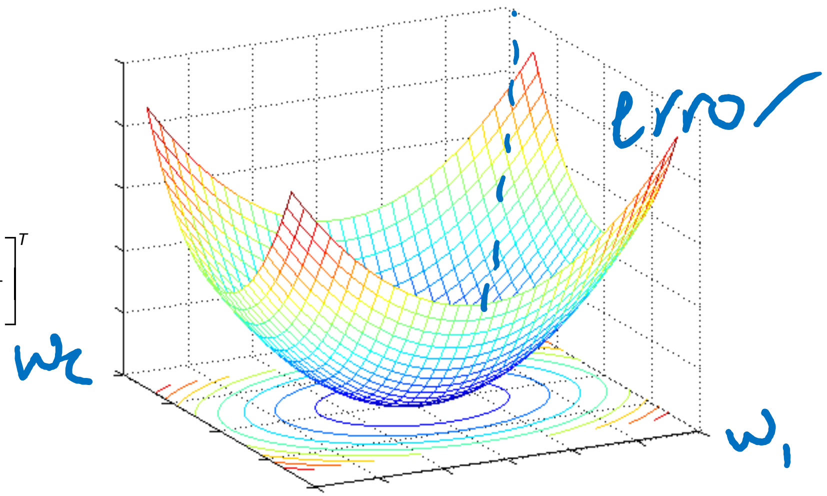
Gradient-Descent

E($\boldsymbol{w}$|X) is error with parameters $\boldsymbol{w}$ on sample X

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} E(\boldsymbol{w} \mid X)$$

Gradient

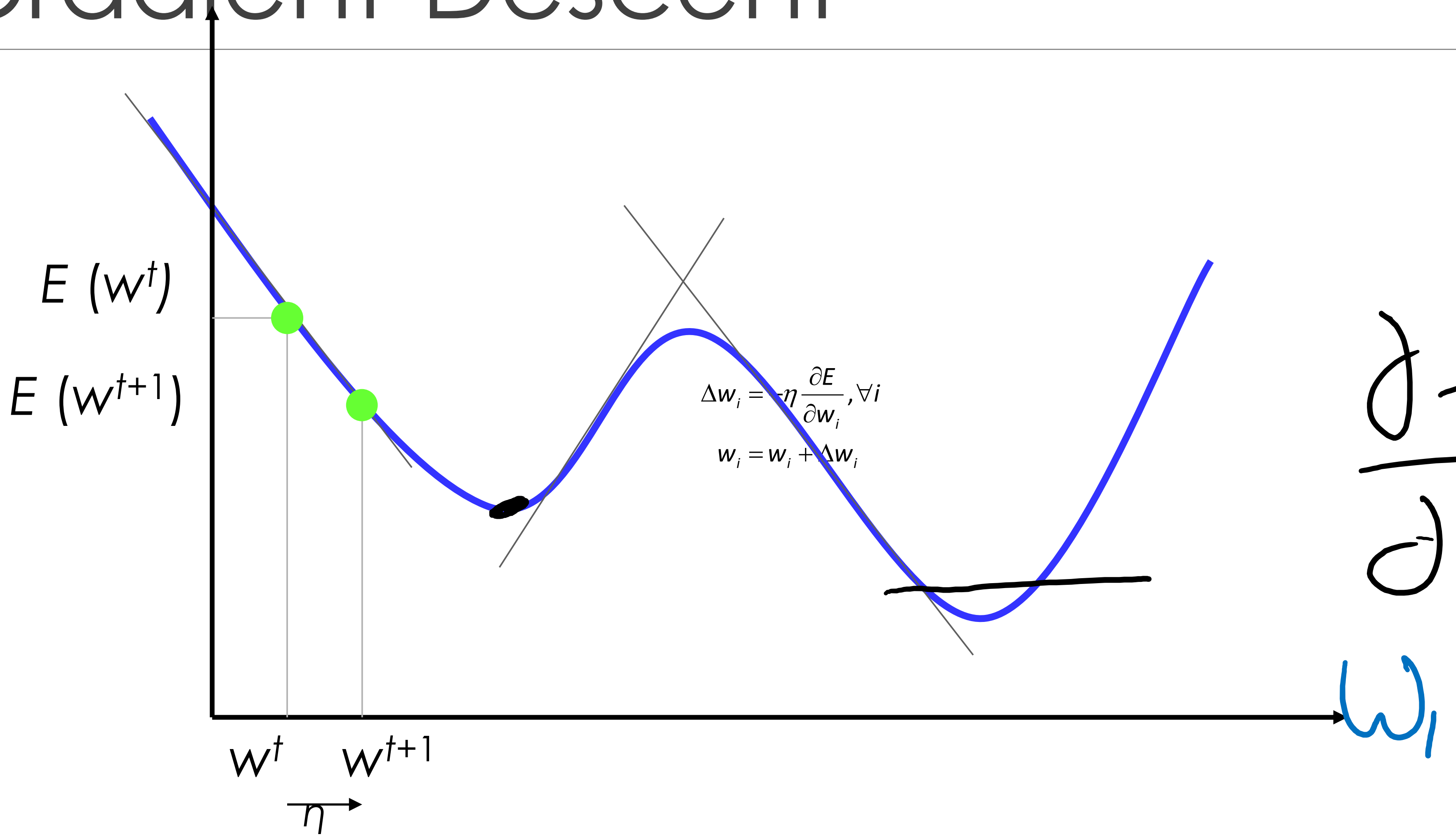
$$\nabla_w E = \left[\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \ldots, \frac{\partial E}{\partial w_d}\right]^T$$



Gradient-descent:

Starts from random $\boldsymbol{w}$ and updates $\boldsymbol{w}$ iteratively in the negative direction of gradient

Gradient-Descent



$$\frac{\partial E}{\partial w_i} < \varepsilon$$

Logistic regression and overfitting

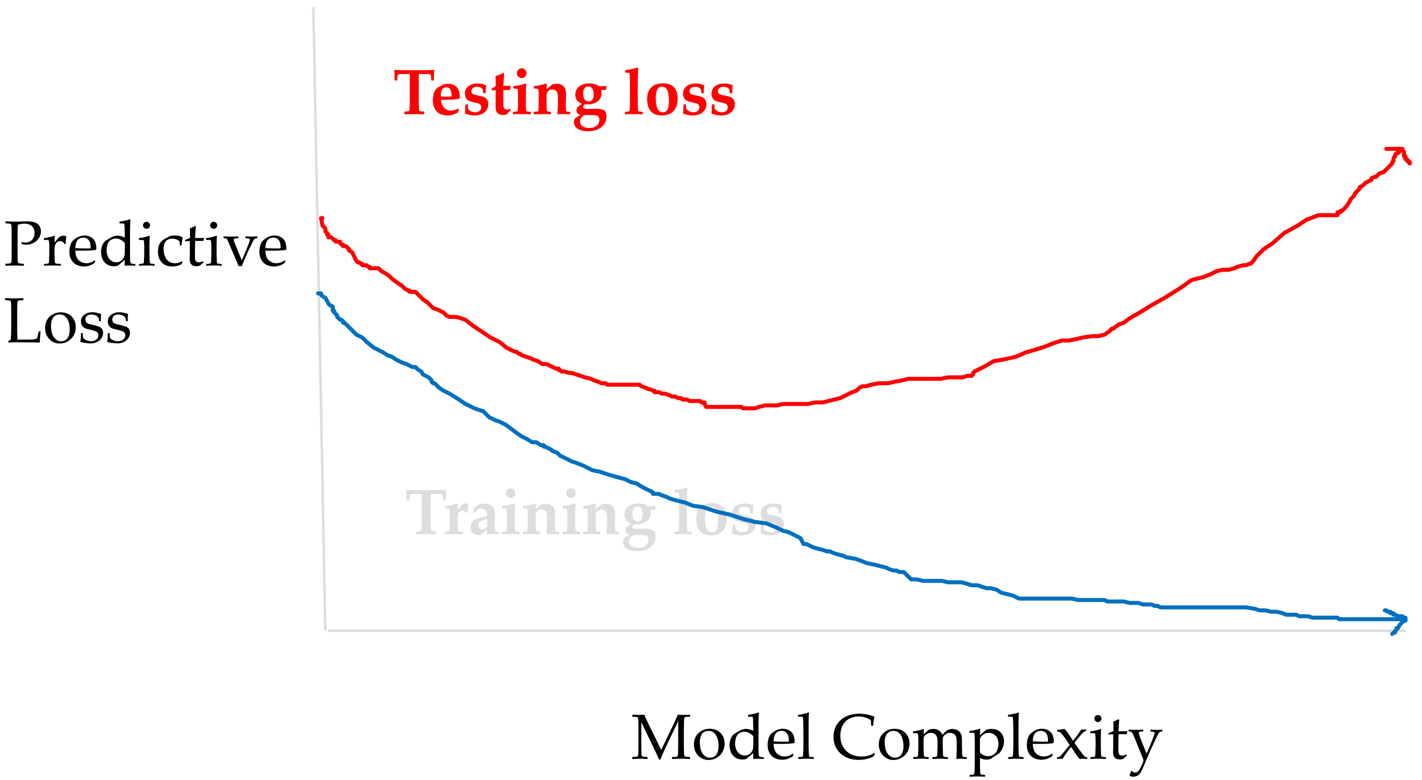
Overfitting

- Occurs when very few instances and feature space is high dimensional

To avoid, a common approach is defining a prior on w

- Corresponds to Regularization
- Helps with avoiding large weights
- “Pushes” parameters to zero

Overfitting



Need to prevent complex hypotheses

Overfitting

- Occurs when very few instances and feature space is high dimensional

Idea #1: Restrict the number of features considered

- Cross-validation

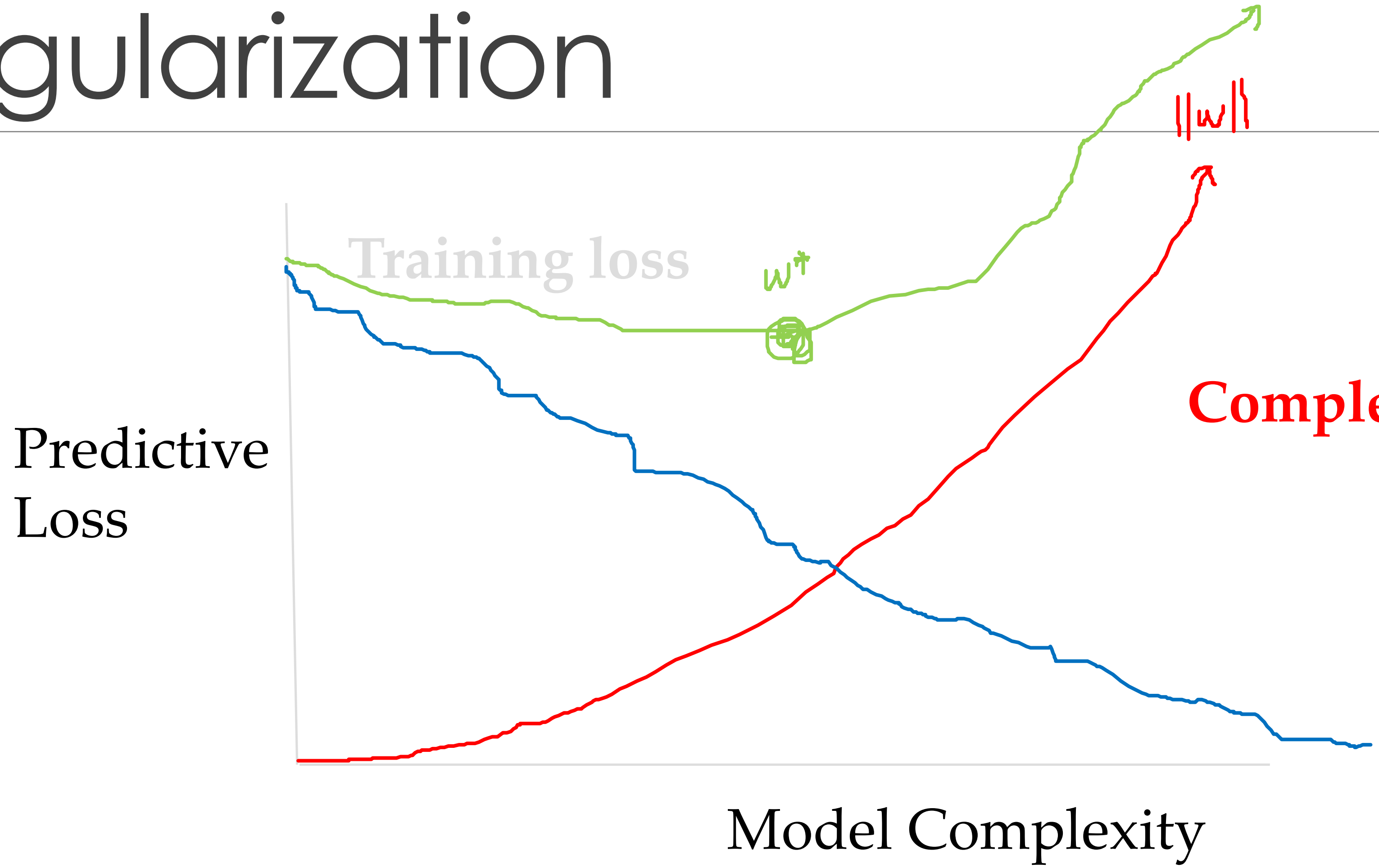
Idea #2: Penalize complex hypotheses in the model search

- Regularization!

Subset selection

Feature extraction

Regularization



Regularization

Recall the objective of logistic regression:

$$E(\mathbf{w}, w_0 \mid \mathcal{X}) = -\sum_t r^t \log y^t + (1 - r^t) \log (1 - y^t)$$

L2 regularization

$$\underset{\mathbf{w}}{argmin} \quad E(\mathbf{w}, w_0 | X) + \lambda \sum_i w_i^2$$

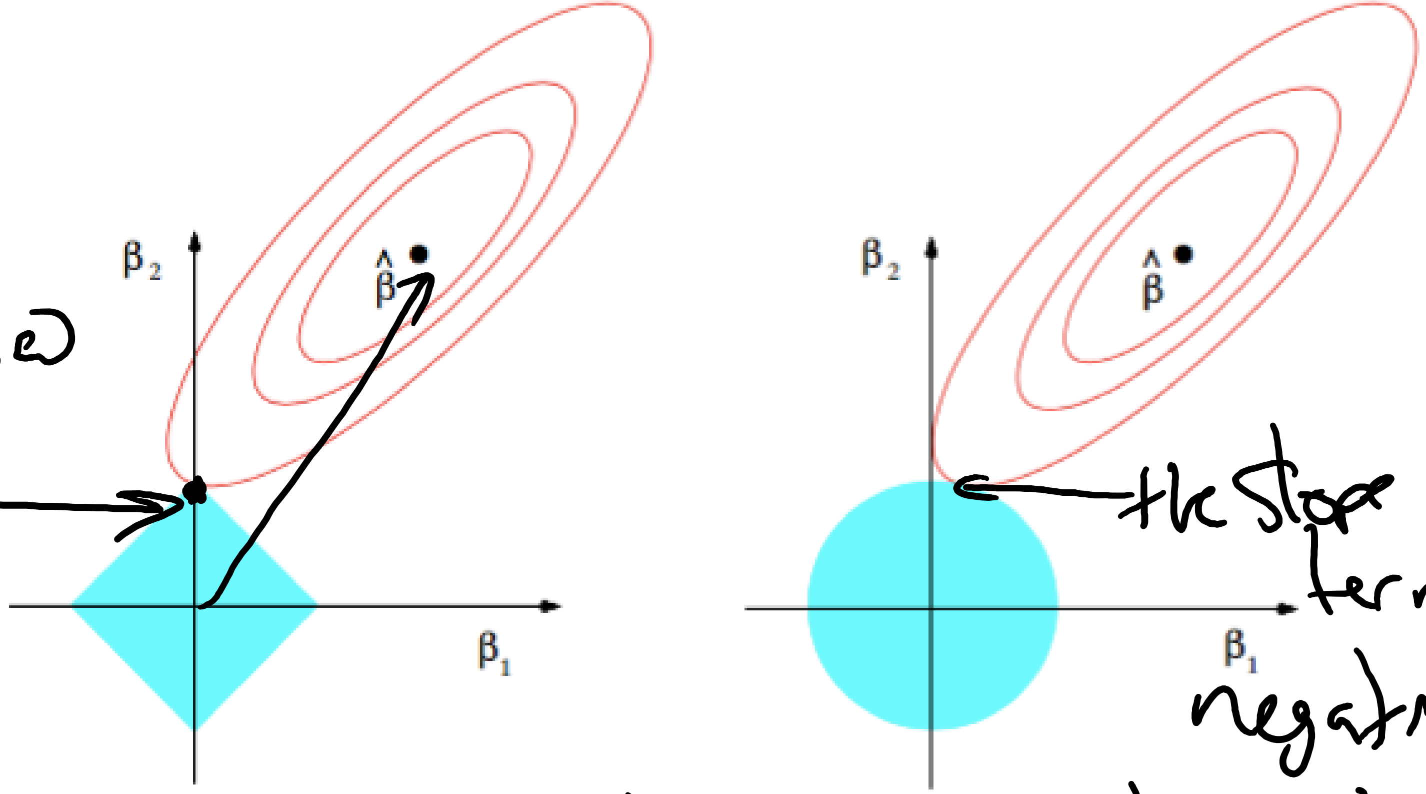
L1 regularization

$$\underset{\mathbf{w}}{argmin} \quad E(\mathbf{w}, w_0 | X) + \lambda \sum_i |w_i|$$

$\lambda > 0$ is a weight, chosen by, e.g., cross validation

Regularization

regularized point



the slope of the Error term = negative regularization

gives many zero weights

Penalizes large ones

Kernel Machines

Discriminant-based: No need to estimate densities first

Define the discriminant in terms of support vectors

The use of kernel functions, application-specific measures of similarity

No need to represent instances as vectors

Convex optimization problems with a unique solution

Hyperplane that correctly separates

$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_t \text{ where } r^t = \begin{cases} +1 & \text{if } \mathbf{x}^t \in C_1 \\ -1 & \text{if } \mathbf{x}^t \in C_2 \end{cases}$$

find $\mathbf{w}$ and w_0 such that

$\mathbf{w}^T\mathbf{x}^t + w_0 \geq 0$ for $r^t = +1$

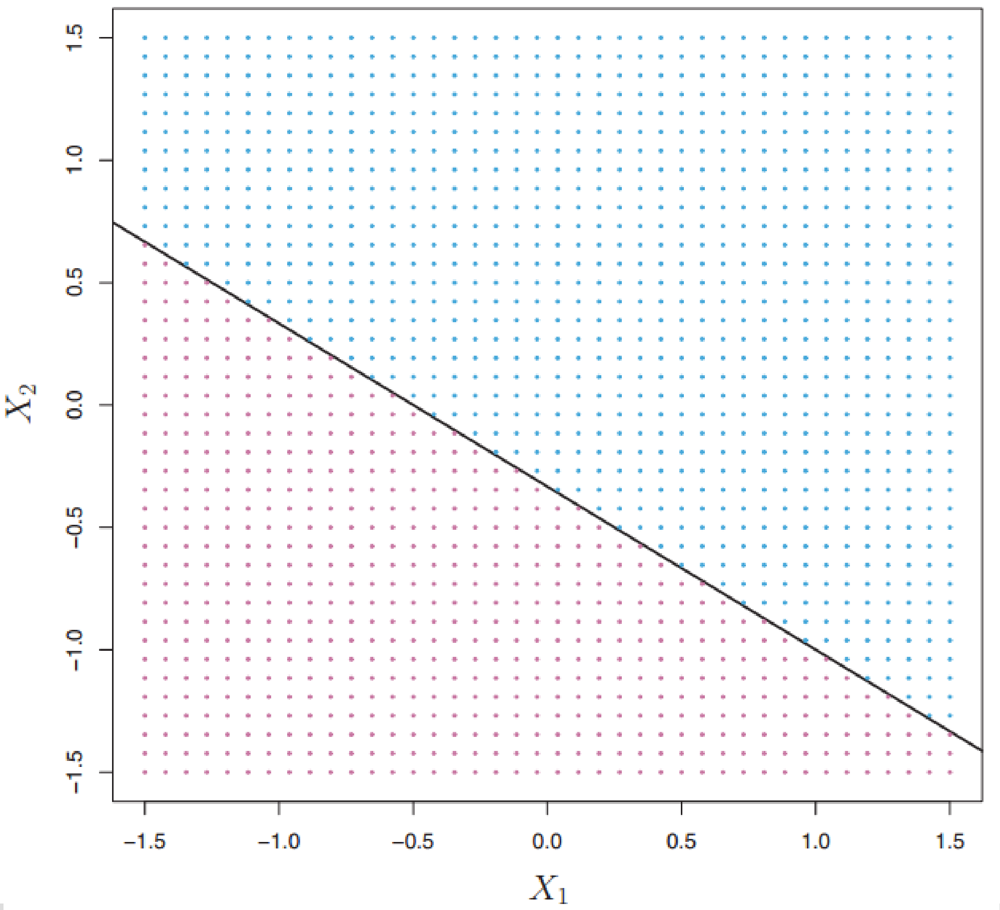
$\mathbf{w}^T\mathbf{x}^t + w_0 \leq 0$ for $r^t = -1$

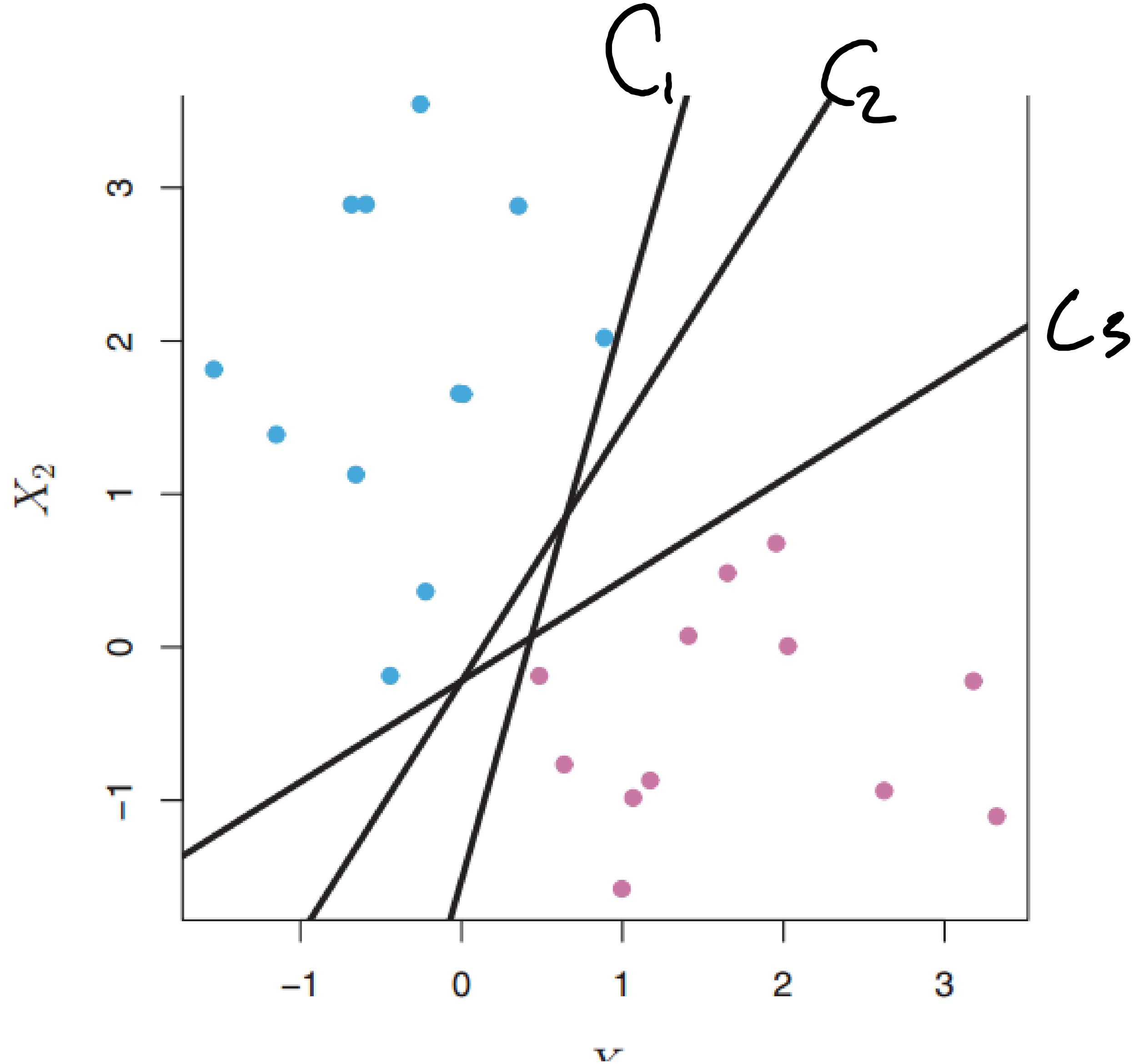
0

which can be rewritten as

$r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) \geq +1$

- Usually no solutions (not linearly separable)
- But…assume there is a solution, then what?

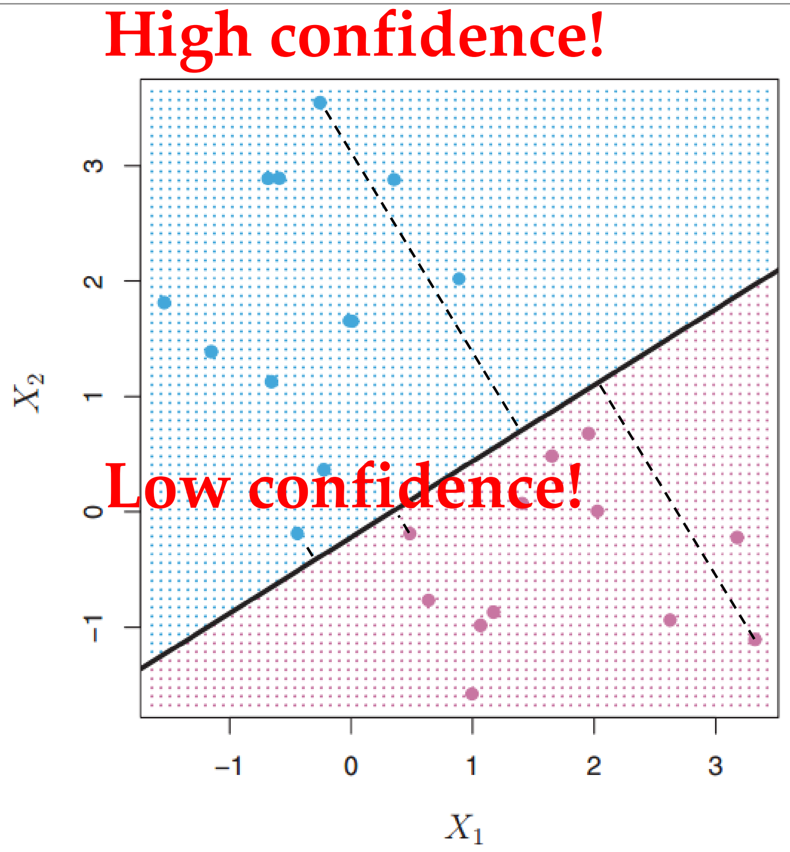




Linear classifiers: Which hyperplane is best?

"Confidence" of Predictions

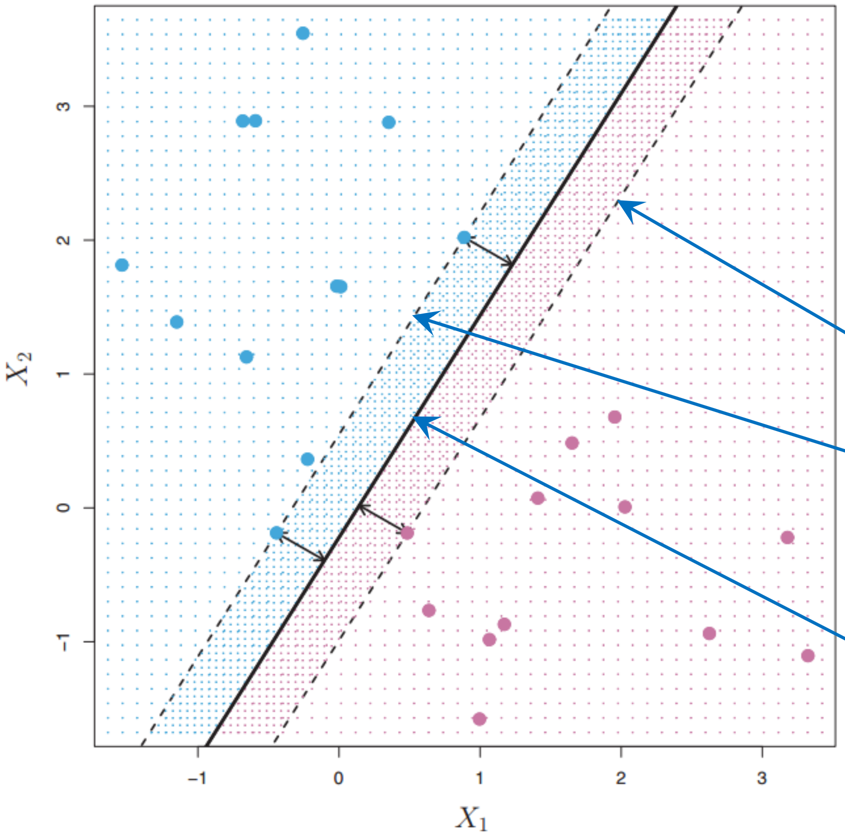
$$\begin{cases} C_1 & \text{if } t > 0 \\ C_2 & \text{if } t < 0 \end{cases}$$



"Confidence" = $r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right)$

What about multiplying **w** and w_0 by 2 or 100?

Pick the one with the largest margin!

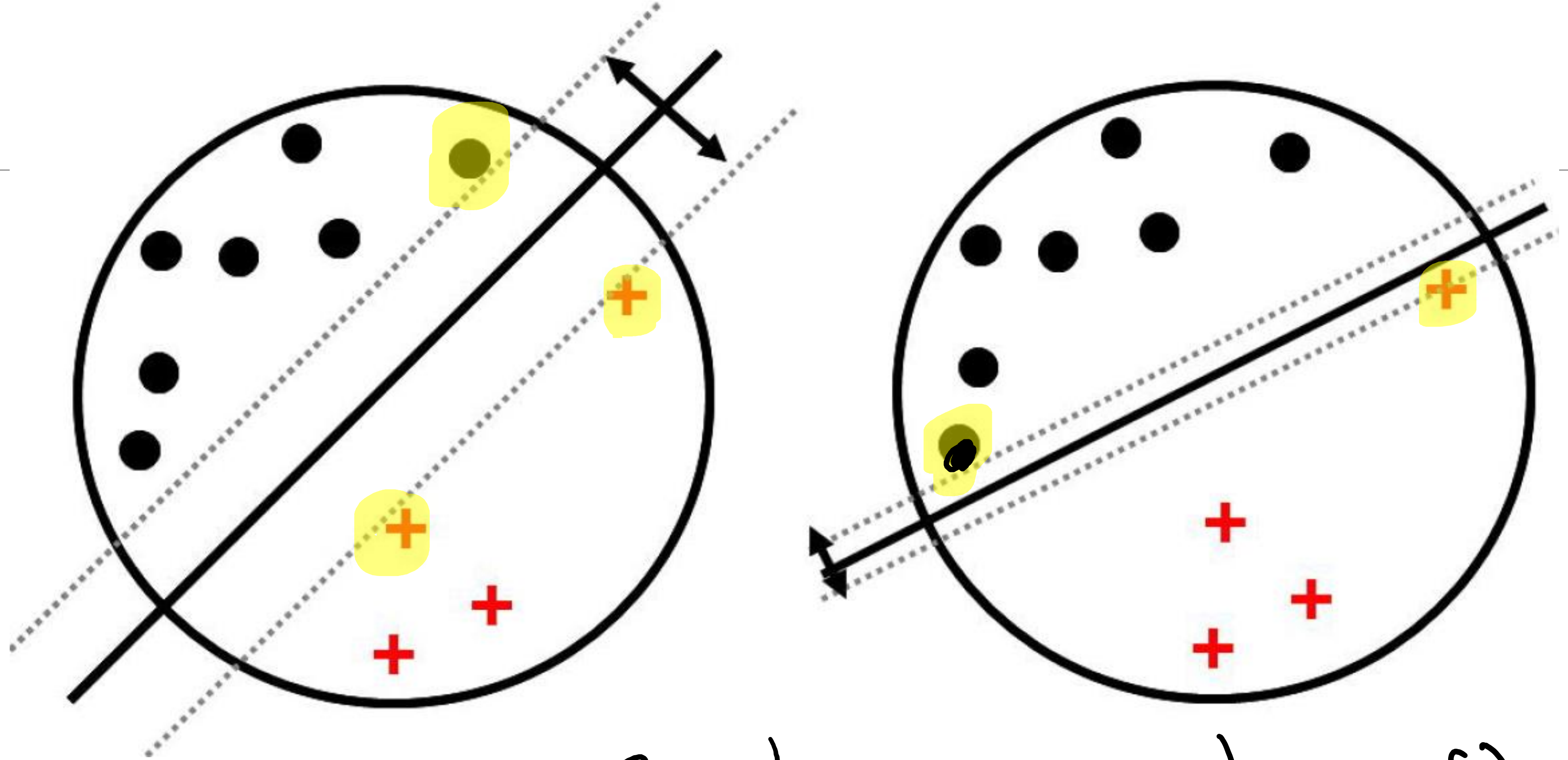


Points on the margin boundary have the lowest "confidence" over all points

Let's maximize this!

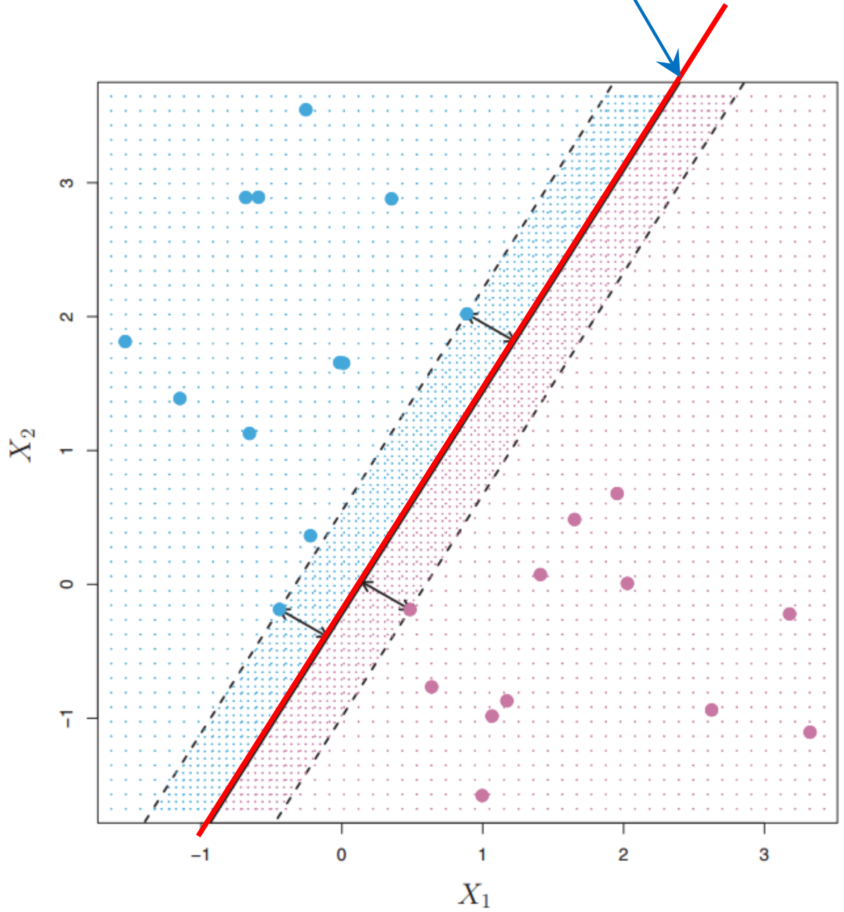
margin boundaries

$\mathbf{w}^T\mathbf{x}^t + w_0 = 0$ separation boundary



Points of minimal confidence

Pick the one with the largest margin! $\mathbf{w}^T\mathbf{x}^t + w_0 = 0$



Points on the margin boundary have the lowest "confidence" over all points

Let's maximize this!

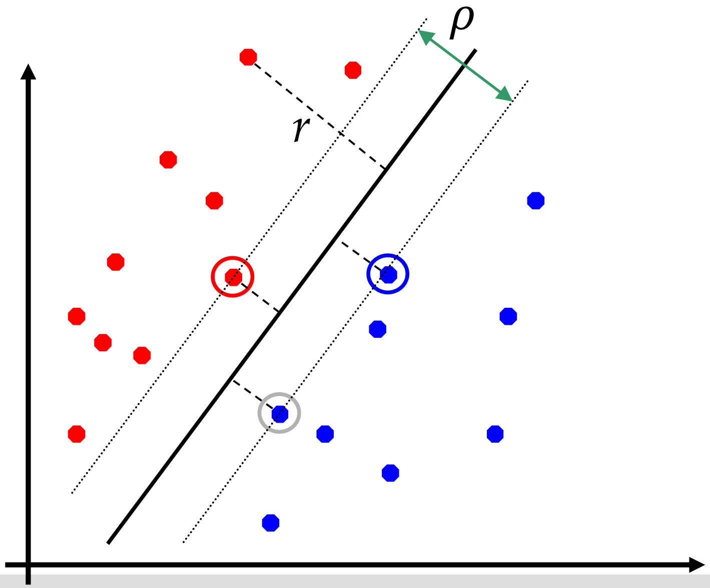
Naturally, we want the margin to be the same for pos and neg

Classification Margin

Distance from example $\mathbf{x}_i$ to the separator is

Examples closest to the hyperplane are ***support vectors***.

Margin ρ of the separator is the distance between support vectors.

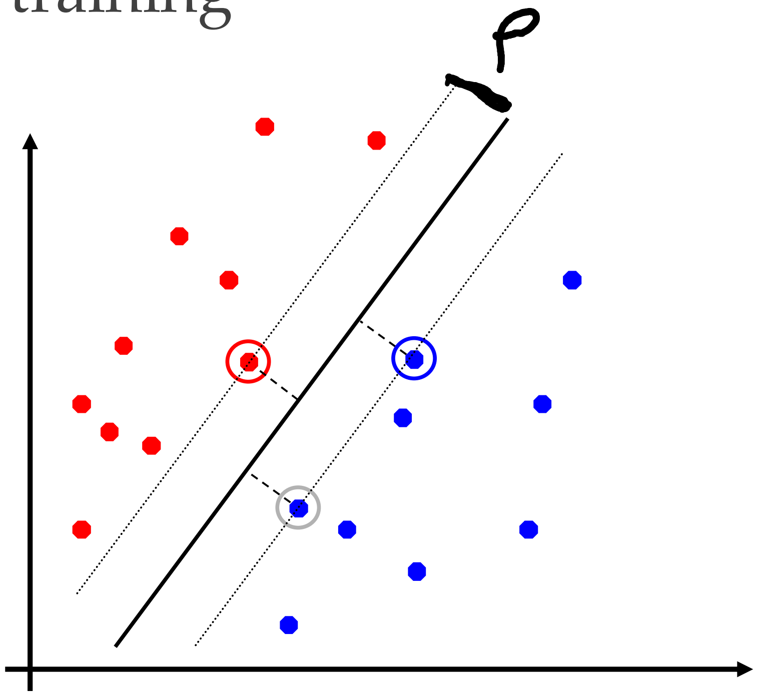


$$r = \frac{\mathbf{w}^T \mathbf{x}_i + b}{\|\mathbf{w}\|}$$

Maximum Margin Classification

Maximizing the margin is good according to intuition and PAC theory.

Implies that only support vectors matter; other training examples are ignorable.



Linear SVM Mathematically

Let training set $\{(\mathbf{x}_i, y_i)\}_{i=1..n}$, $\mathbf{x}_i \in \mathbf{R}^d$, $y_i \in \{-1, 1\}$ be separated by a hyperplane with margin ρ. Then for each training example $(\mathbf{x}_i, y_i)$:

$$\begin{aligned} \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b &\leq -\rho/2 \quad \text{if } y_i = -1 \\ \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b &\geq \rho/2 \quad \text{if } y_i = 1 \end{aligned} \quad \Leftrightarrow \quad y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq \rho/2$$

For every support vector $\mathbf{x}_s$ the above inequality is an equality. After rescaling **w** and b by $\rho/2$ in the equality, we obtain that distance between each $\mathbf{x}_s$ and the hyperplane is

$$r = \frac{\mathrm{y}_s(\mathbf{w}^T\mathbf{x}_s + b)}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}$$

Then the margin can be expressed through (rescaled) **w** and b as:

$$\rho = 2r = \frac{2}{\|\mathbf{w}\|}$$

Linear SVMs Mathematically

Then we can formulate the *quadratic optimization problem:*

Find **w** such that

$\rho = \frac{2}{\|\mathbf{w}\|}$ is maximized ⟵ biggest boundary

and for all $(\mathbf{x}_i, y_i)$, i=1..n : $y_i(\mathbf{w}^T\mathbf{x}_i) \geq 1$ ⟵ all points are correctly classified

Supposes a linear separator

Which can be reformulated as:

Find **w** such that

$\mathbf{\Phi}(\mathbf{w}) = ||\mathbf{w}||^2 = \mathbf{w}^T\mathbf{w}$ is minimized

and for all $(\mathbf{x}_i, y_i)$, i=1..n : $y_i(\mathbf{w}^T\mathbf{x}_i) \geq 1$

The Optimization Problem Solution

Given a solution $\alpha_1 \ldots \alpha_n$ to the dual problem, solution to the primal is:

$$\boxed{\mathbf{w} = \Sigma\alpha_i y_i \mathbf{x}_i \qquad b = y_k - \Sigma\alpha_i y_i \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_k \quad \text{for any } \alpha_k > 0}$$

α_i is a "mask" for non-relevant points not on the margin

Each non-zero α_i indicates that corresponding $\mathbf{x}_i$ is a support vector.

Then the classifying function is (note that we don't need **w** explicitly):

$$\boxed{f(\mathbf{x}) = \Sigma\alpha_i y_i \mathbf{x}_i^{\mathrm{T}}\mathbf{x} + b}$$

Notice that it relies on an *inner product* between the test point **x** and the support vectors $\mathbf{x}_i$ – we will return to this later.

Also keep in mind that solving the optimization problem involved computing the inner products $\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j$ between all training points.

Hard margin SVM (linearly separable)

- Distance from the discriminant to the closest instances on either side
- Distance of x to the hyperplane is $\frac{\left|\mathbf{w}^T\mathbf{x}^t + w_0\right|}{\|\mathbf{w}\|}$
- We require $\frac{r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right)}{\|\mathbf{w}\|} \geq \rho, \forall t$.
  - ρ: margin of the dataset (invariant to scaling of w)
- For a unique sol'n, fix $\rho \, ||w||=1$
  - Maximize margin ρ $\Leftrightarrow$ minimize $||w||$

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) \geq +1, \forall t$$

Margin and support vector

- Support vectors: points lying on the marginal hyperplanes
- NO change of solution does if: remove all other points and retrain
- Margin

$$\min_t \frac{r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right)}{\|\mathbf{w}\|} = \frac{1}{||\mathbf{w}||}$$

- Marginal hyperplanes

$$\mathbf{w}^T\mathbf{x}^t + w_0 = -1$$

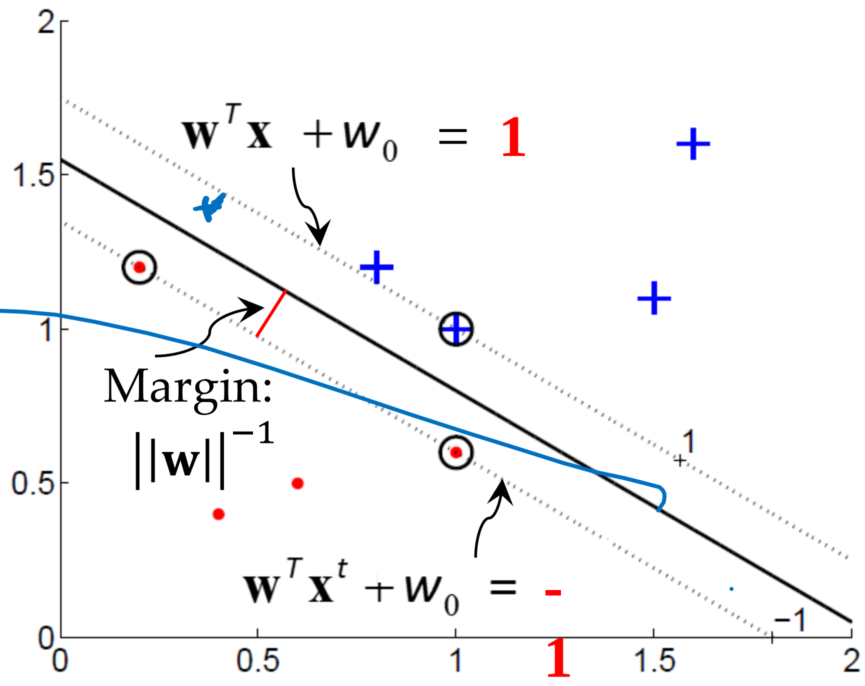
$$\mathbf{w}^T\mathbf{x}^t + w_0 = 1$$

unique solution

- Separating hyperplane

$$\mathbf{w}^T\mathbf{x}^t + w_0 = 0$$

↳ however classify



Soft Margin Hyperplane

- Linear separable:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) \geq +1, \forall t$$

- Not linearly separable

$$r^t\left(\mathbf{w}^T x^t + w_0\right) \geq 1 - \xi^t$$

- Soft error $\sum_t \xi^t$
- New (primal) objective is

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \xi^t \quad \text{subject to } r^t\left(\mathbf{w}^T x^t + w_0\right) \geq 1 - \xi^t \qquad \xi^t \geq 0$$

Soft Margin Classification Mathematically

The old formulation:

hard margin

Find $\mathbf{w}$ such that
$\mathbf{\Phi}(\mathbf{w}) = \mathbf{w}^T\mathbf{w}$ is minimized
and for all $(\mathbf{x}_i, y_i)$, $i=1..n$: $\quad y_i(\mathbf{w}^T\mathbf{x}_i) \geq 1$

Modified formulation incorporates slack variables:

larger margins

less hingeloss

Find $\mathbf{w}$ such that
$\mathbf{\Phi}(\mathbf{w}) = \mathbf{w}^T\mathbf{w} + C\Sigma\xi_i$ is minimized
and for all $(\mathbf{x}_i, y_i)$, $i=1..n$: $\quad y_i(\mathbf{w}^T\mathbf{x}_i) \geq 1 - \xi_i, \quad \xi_i \geq 0$

Parameter C can be viewed as a way to control overfitting: it "trades off" the relative importance of maximizing the margin and fitting the training data.

- Support vectors: $r^t\left(w^T x^t + w_0\right) \le 1$
  - Positive points lying on the side of $w^T x^t + w_0 \le 1$
  - Negative points lying on the side of $w^T x^t + w_0 \ge -1$
  - NO change of solution if:

    remove all other points

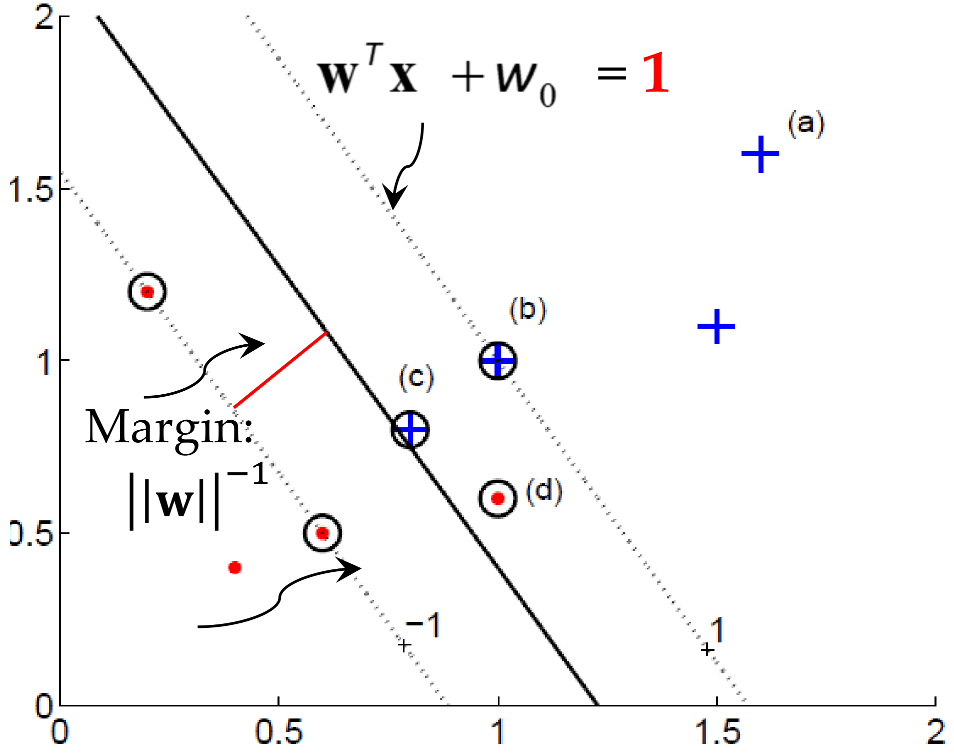
    and retrain

- Margin? $\frac{1}{||\mathbf{w}||} \neq \min_t \frac{r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right)}{\|\mathbf{w}\|}$

- Marginal hyperplanes

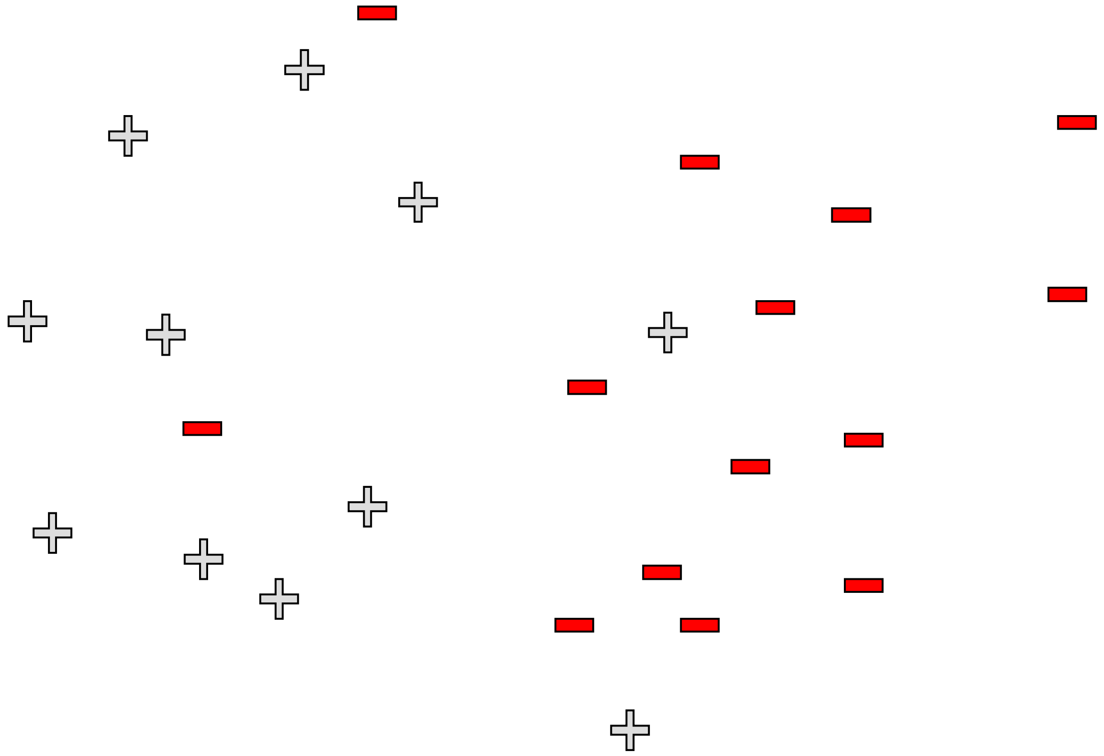
$$\mathbf{w}^T\mathbf{x} + w_0 = -1 \text{ or } 1$$

$$\mathbf{w}^T\mathbf{x}^t + w_0 = -1$$



Support vectors of SVMs

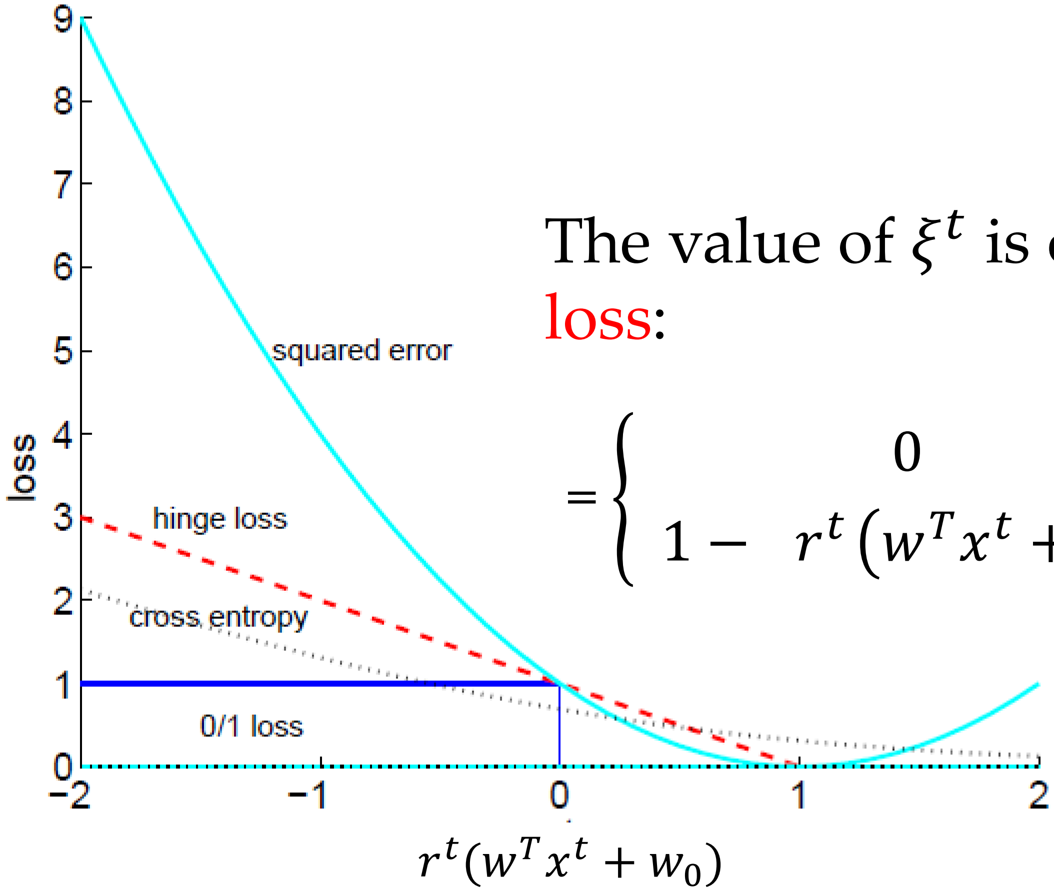
Which examples influence the margin and decision boundaries?



Hinge Loss

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \xi^t \quad \text{subject to } r^t\left(\mathbf{w}^T x^t + w_0\right) \geq 1 - \xi^t$$

$$\xi^t \geq 0$$



The value of ξ^t is called hinge loss:

$$= \begin{cases} 0 & \text{if } r^t\left(w^T x^t + w_0\right) \geq 1 \\ 1 - r^t\left(w^T x^t + w_0\right) & \text{otherwise} \end{cases}$$

Linear SVMs: Overview

The classifier is a *separating hyperplane.*

Most "important" training points are support vectors; they define the hyperplane.

Quadratic optimization algorithms can identify which training points $\mathbf{x}_i$ are support vectors with non-zero Lagrangian multipliers α_i.

Both in the dual formulation of the problem and in the solution training points appear only inside inner products:

Find $\alpha_1 \ldots \alpha_N$ such that
$\mathbf{Q}(\boldsymbol{\alpha}) = \Sigma\alpha_i - \frac{1}{2}\Sigma\Sigma\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$ is maximized
and
(1) $\Sigma\alpha_i y_i = 0$
(2) $0 \leq \alpha_i \leq C$ for all α_i

$f(\mathbf{x}) = \Sigma\alpha_i y_i \mathbf{x}_i^T\mathbf{x} + b$