Machine Learning Basics
Lecture 4: SVM I

Princeton University COS 495

Instructor: Yingyu Liang

Review: machine learning basics

Math formulation

- Given training data $\{(x_i, y_i): 1 \le i \le n\}$ i.i.d. from distribution D
- Find $y = f(x) \in \boldsymbol{\mathcal{H}}$ that minimizes $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} l(f, x_i, y_i)$
- s.t. the expected loss is small

$$L(f) = \mathbb{E}_{(x,y)\sim D}[l(f, x, y)]$$

Machine learning 1-2-3

- Collect data and extract features
- Build model: choose hypothesis class $\mathcal{H}$ and loss function l
- Optimization: minimize the empirical loss

Loss function

- l_2 loss: linear regression
- Cross-entropy: logistic regression
- Hinge loss: Perceptron

- General principle: maximum likelihood estimation (MLE)
  - l_2 loss: corresponds to Normal distribution
  - logistic regression: corresponds to sigmoid conditional distribution

Optimization

- Linear regression: closed form solution
- Logistic regression: gradient descent
- Perceptron: stochastic gradient descent

- General principle: local improvement
    - SGD: Perceptron; can also be applied to linear regression/logistic regression

Principle for hypothesis class?

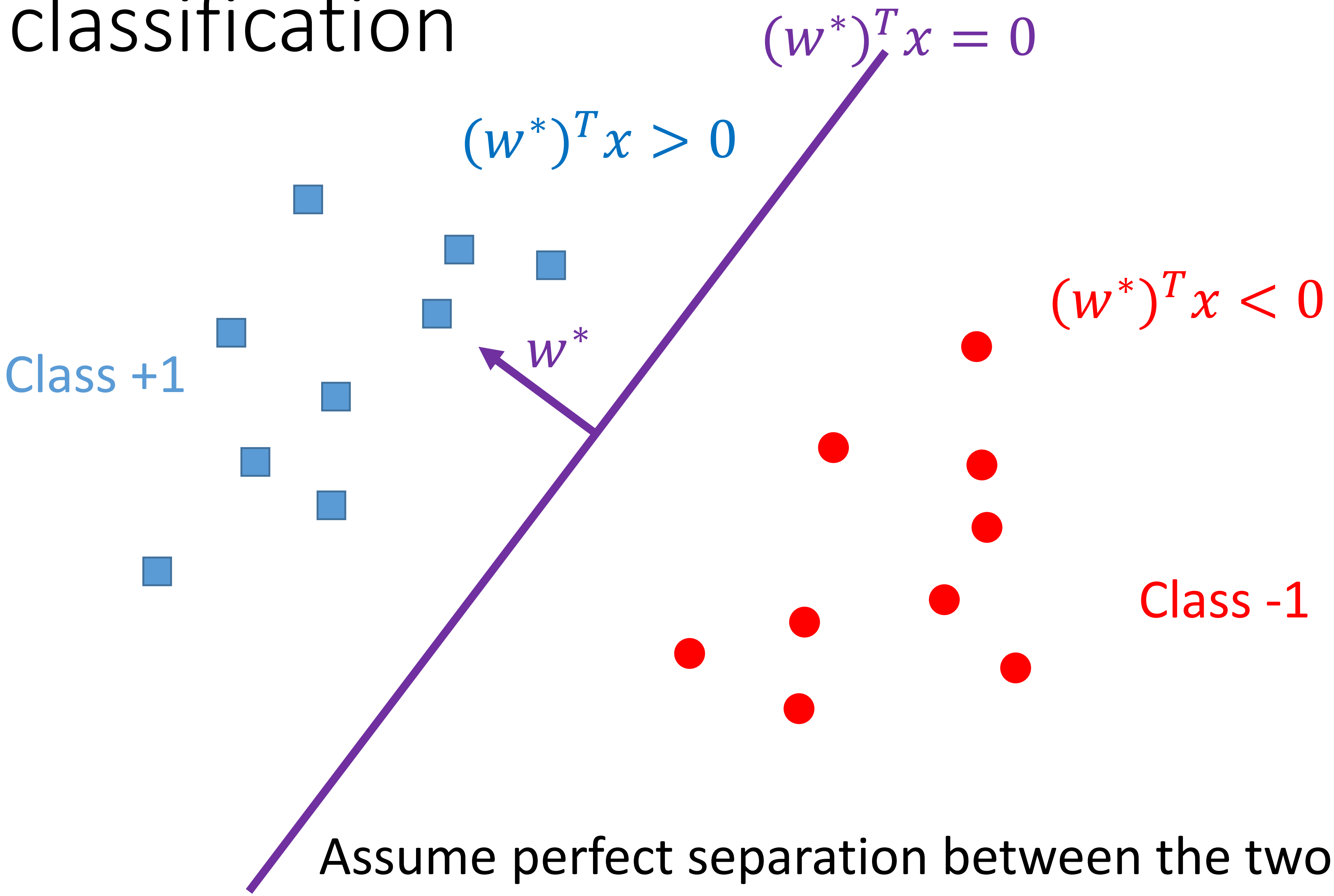
- Yes, there exists a general principle (at least philosophically)

- Different names/faces/connections
  - Occam's razor
  - VC dimension theory
  - Minimum description length
  - Tradeoff between Bias and variance; uniform convergence
  - The curse of dimensionality

- Running example: Support Vector Machine (SVM)

Motivation

Linear classification



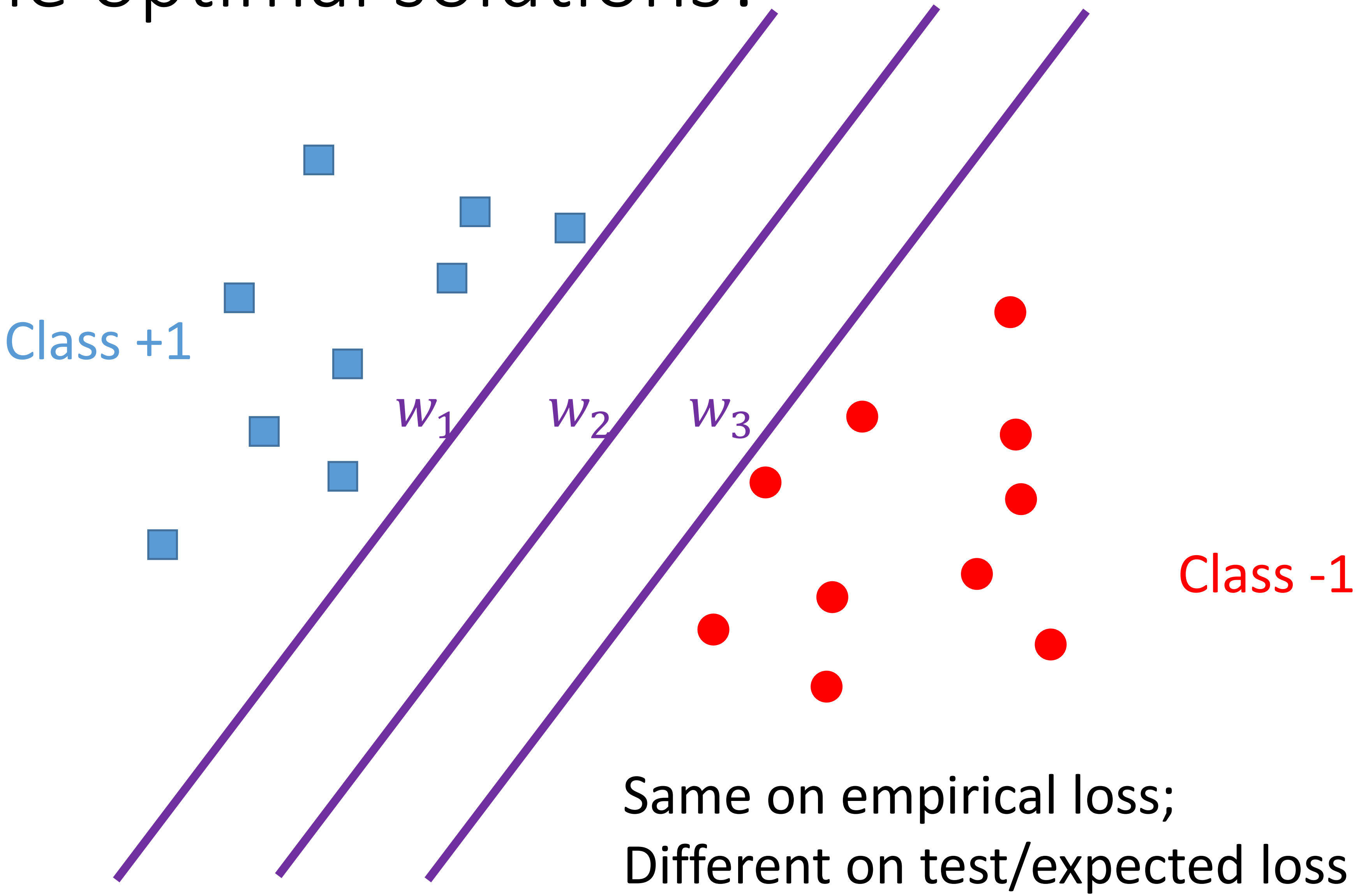
Assume perfect separation between the two classes

Attempt

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$ i.i.d. from distribution D
- Hypothesis $y = \text{sign}(f_w(x)) = \text{sign}(w^T x)$
  - $y = +1$ if $w^T x > 0$
  - $y = -1$ if $w^T x < 0$

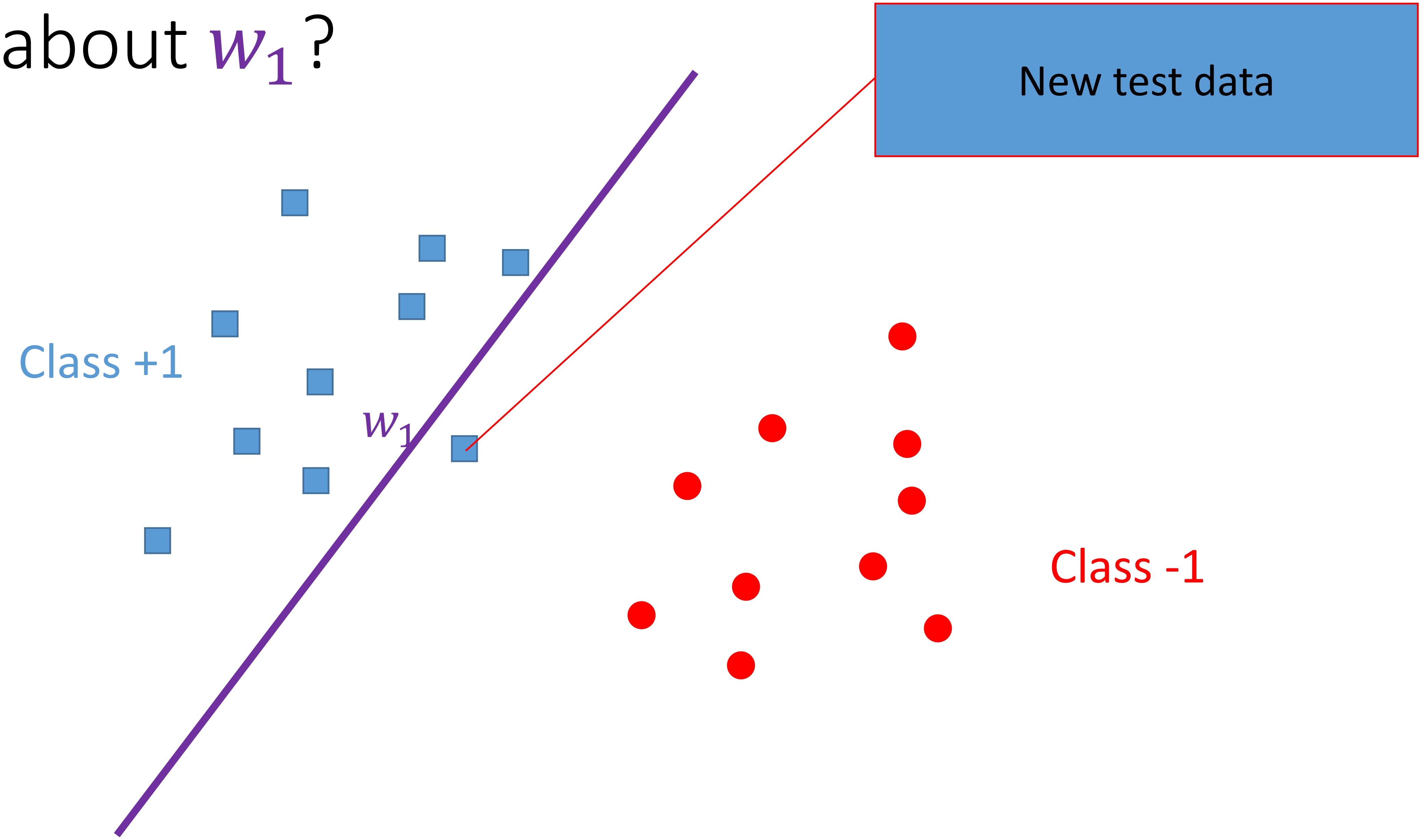
- Let's assume that we can optimize to find w

Multiple optimal solutions?

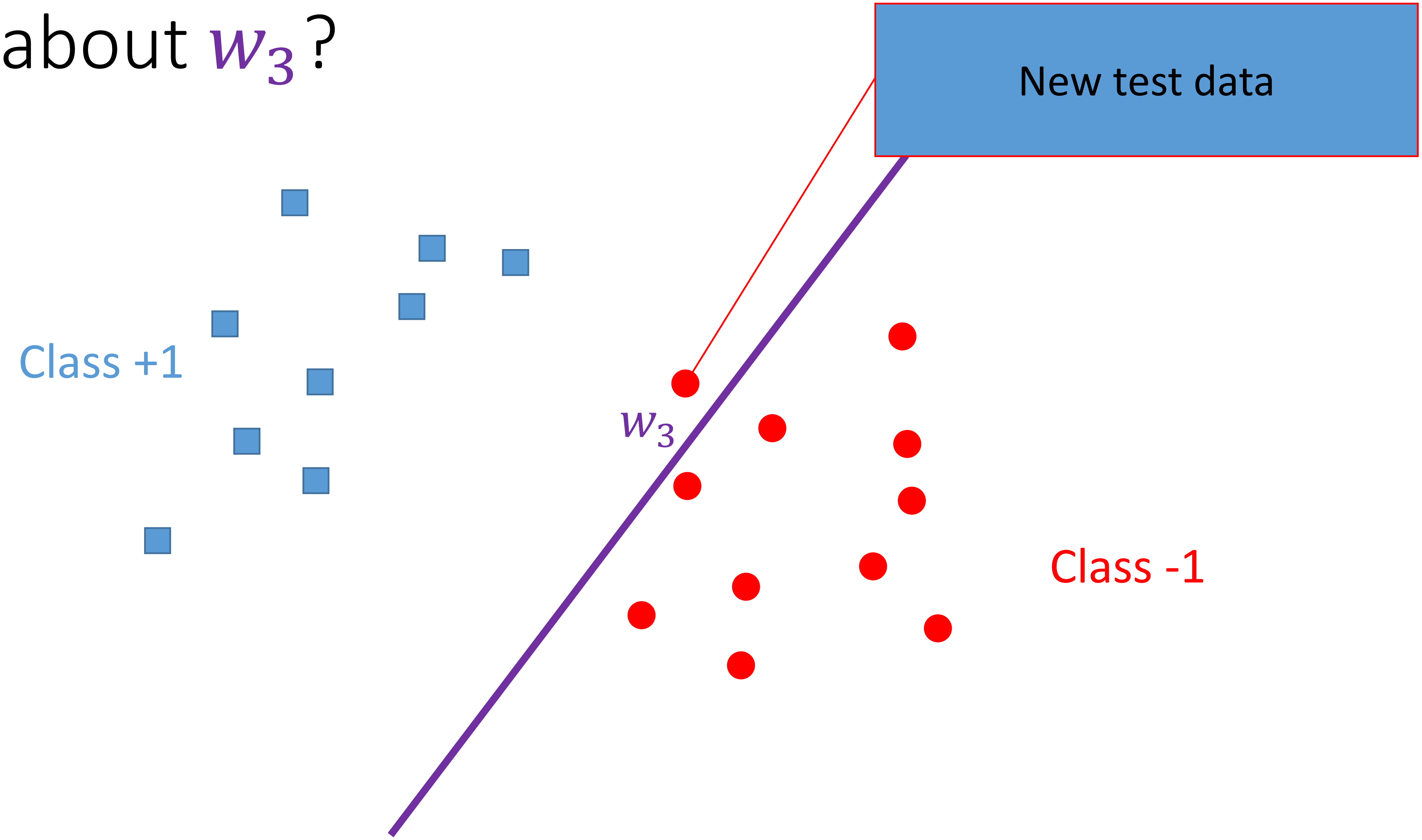


Same on empirical loss;
Different on test/expected loss

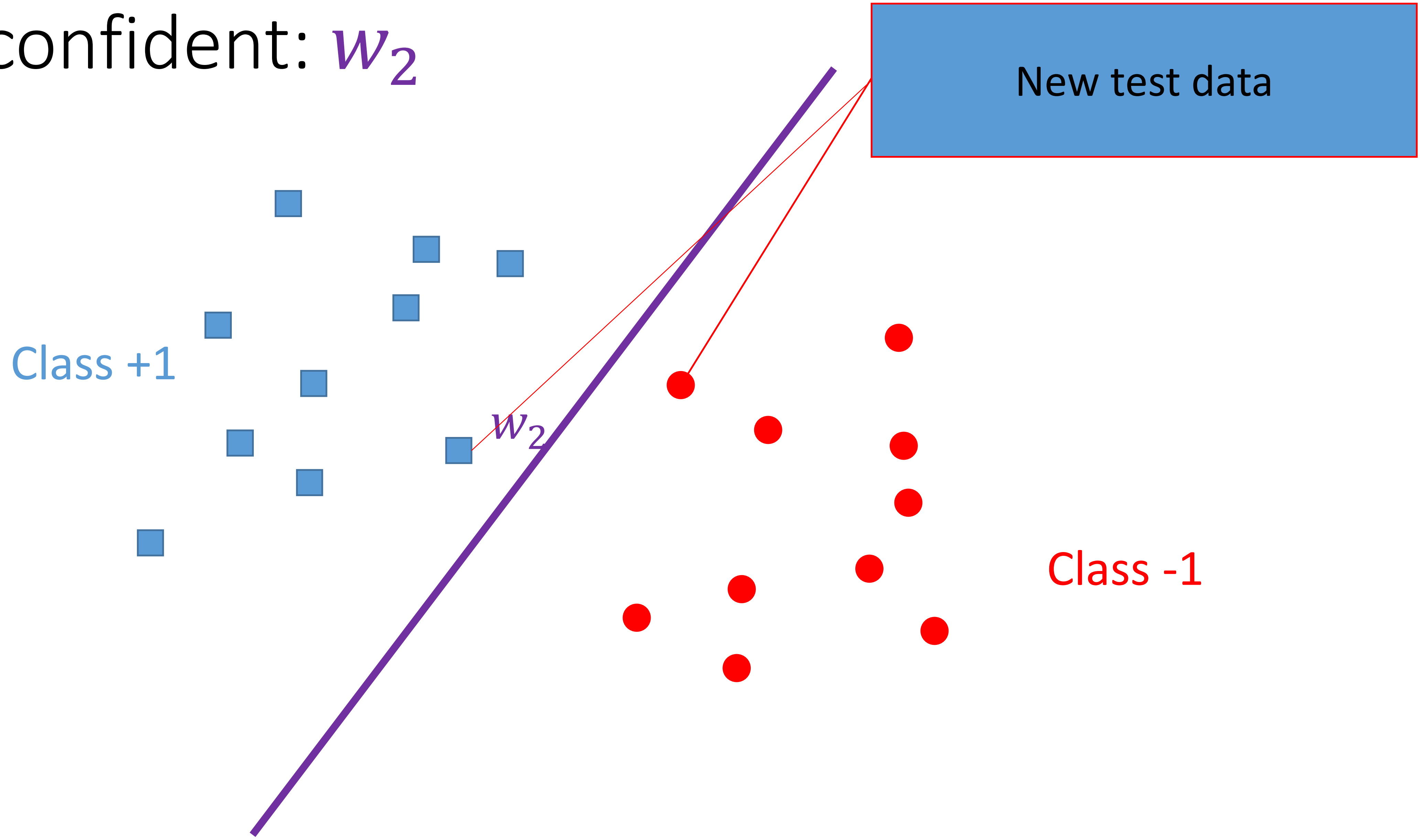
What about w_1?



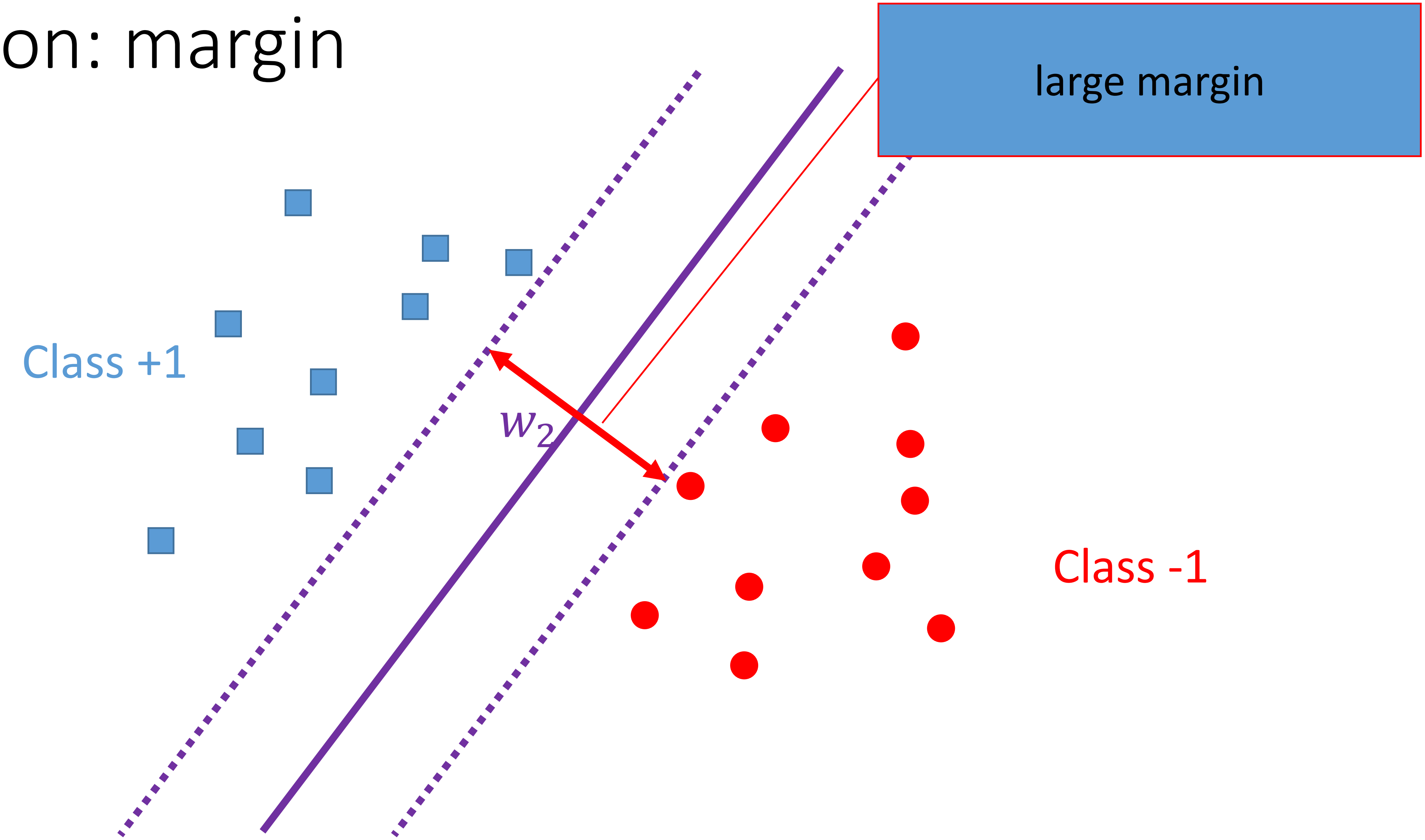
What about w_3?



Most confident: w_2



Intuition: margin



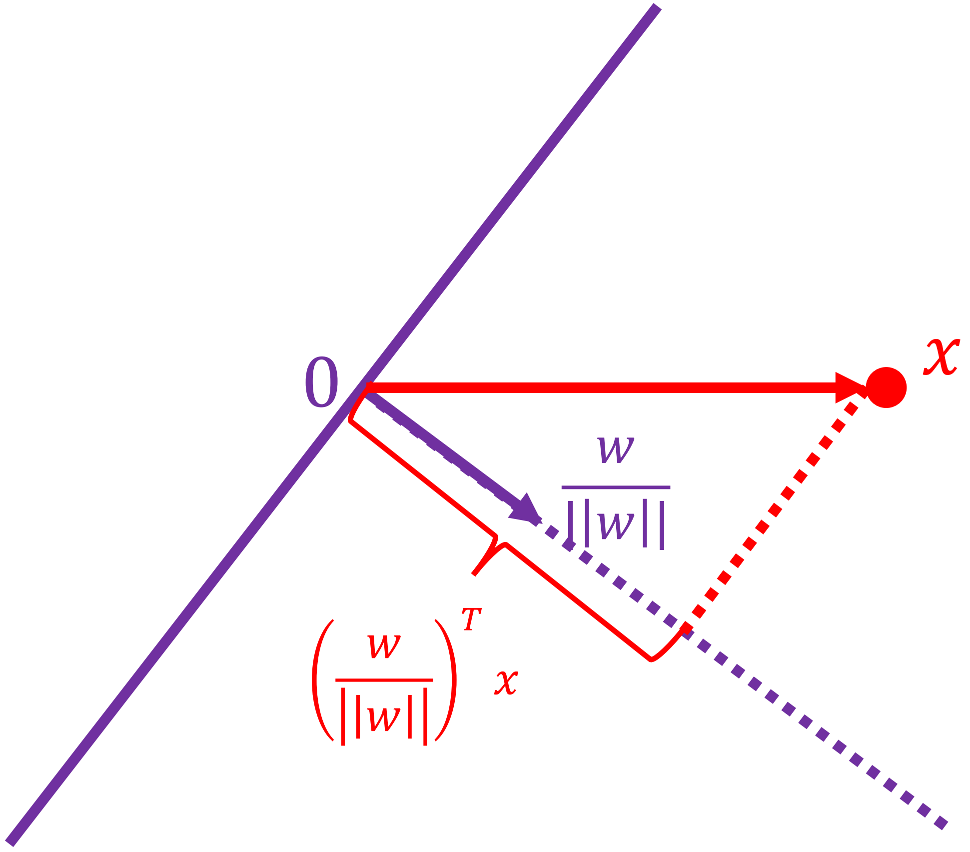
Margin

Margin

- Lemma 1: x has distance $\frac{|f_w(x)|}{||w||}$ to the hyperplane $f_w(x) = w^T x = 0$

Proof:

- w is orthogonal to the hyperplane
- The unit direction is $\frac{w}{||w||}$
- The projection of x is $\left(\frac{w}{||w||}\right)^T x = \frac{f_w(x)}{||w||}$



Margin: with bias

- Claim 1: w is orthogonal to the hyperplane $f_{w,b}(x) = w^T x + b = 0$

Proof:

- pick any x_1 and x_2 on the hyperplane
- $w^T x_1 + b = 0$
- $w^T x_2 + b = 0$

- So $w^T (x_1 - x_2) = 0$

Margin: with bias

- Claim 2: 0 has distance $\frac{-b}{||w||}$ to the hyperplane $w^T x + b = 0$

Proof:

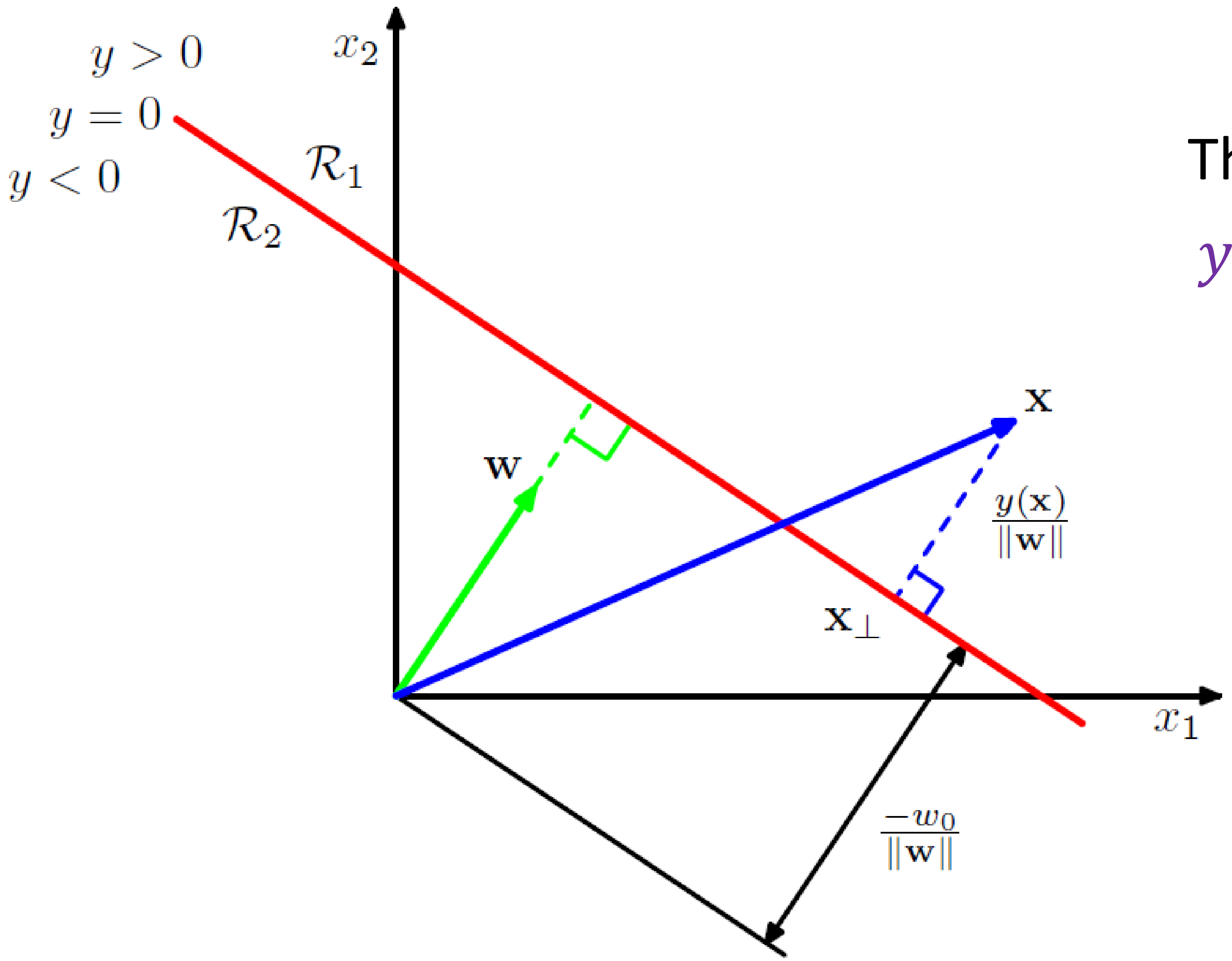
- pick any x_1 the hyperplane
- Project x_1 to the unit direction $\frac{w}{||w||}$ to get the distance
- $\left(\frac{w}{||w||}\right)^T x_1 = \frac{-b}{||w||}$ since $w^T x_1 + b = 0$

Margin: with bias

- Lemma 2: x has distance $\frac{|f_{w,b}(x)|}{||w||}$ to the hyperplane $f_{w,b}(x) = w^T x + b = 0$

Proof:

- Let $x = x_\perp + r\frac{w}{||w||}$, then $|r|$ is the distance
- Multiply both sides by w^T and add b
- Left hand side: $w^T x + b = \ f_{w,b}(x)$
- Right hand side: $w^T x_\perp + r\frac{w^T w}{||w||} + b = 0 + r||w||$



The notation here is:

$$y(x) = w^T x + w_0$$

Figure from *Pattern Recognition and Machine Learning*, Bishop

Support Vector Machine (SVM)

SVM: objective

- Margin over all training data points:

$$\gamma = \min_i \frac{|f_{w,b}(x_i)|}{||w||}$$

- Since only want correct $f_{w,b}$, and recall $y_i \in \{+1, -1\}$, we have

$$\gamma = \min_i \frac{y_i f_{w,b}(x_i)}{||w||}$$

- If $f_{w,b}$ incorrect on some x_i, the margin is negative

SVM: objective

- Maximize margin over all training data points:

$$\max_{w,b} \gamma = \max_{w,b} \min_{i} \frac{y_i f_{w,b}(x_i)}{||w||} = \max_{w,b} \min_{i} \frac{y_i (w^T x_i + b)}{||w||}$$

- A bit complicated …

SVM: simplified objective

- Observation: when (w, b) scaled by a factor c, the margin unchanged

$$\frac{y_i(cw^T x_i + cb)}{||cw||} = \frac{y_i(w^T x_i + b)}{||w||}$$

- Let's consider a fixed scale such that

$$y_{i^*}(w^T x_{i^*} + b) = 1$$

where x_{i^*} is the point closest to the hyperplane

SVM: simplified objective

- Let’s consider a fixed scale such that

$$y_{i^*}(w^T x_{i^*} + b) = 1$$

  where x_{i^*} is the point closet to the hyperplane

- Now we have for all data

$$y_i(w^T x_i + b) \geq 1$$

  and at least for one i the equality holds

- Then the margin is $\frac{1}{||w||}$

SVM: simplified objective

- Optimization simplified to

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

$$y_i(w^T x_i + b) \geq 1, \forall i$$

- How to find the optimum $\widehat{w}^*$?

SVM: principle for hypothesis class

Thought experiment

- Suppose pick an R, and suppose can decide if exists w satisfying

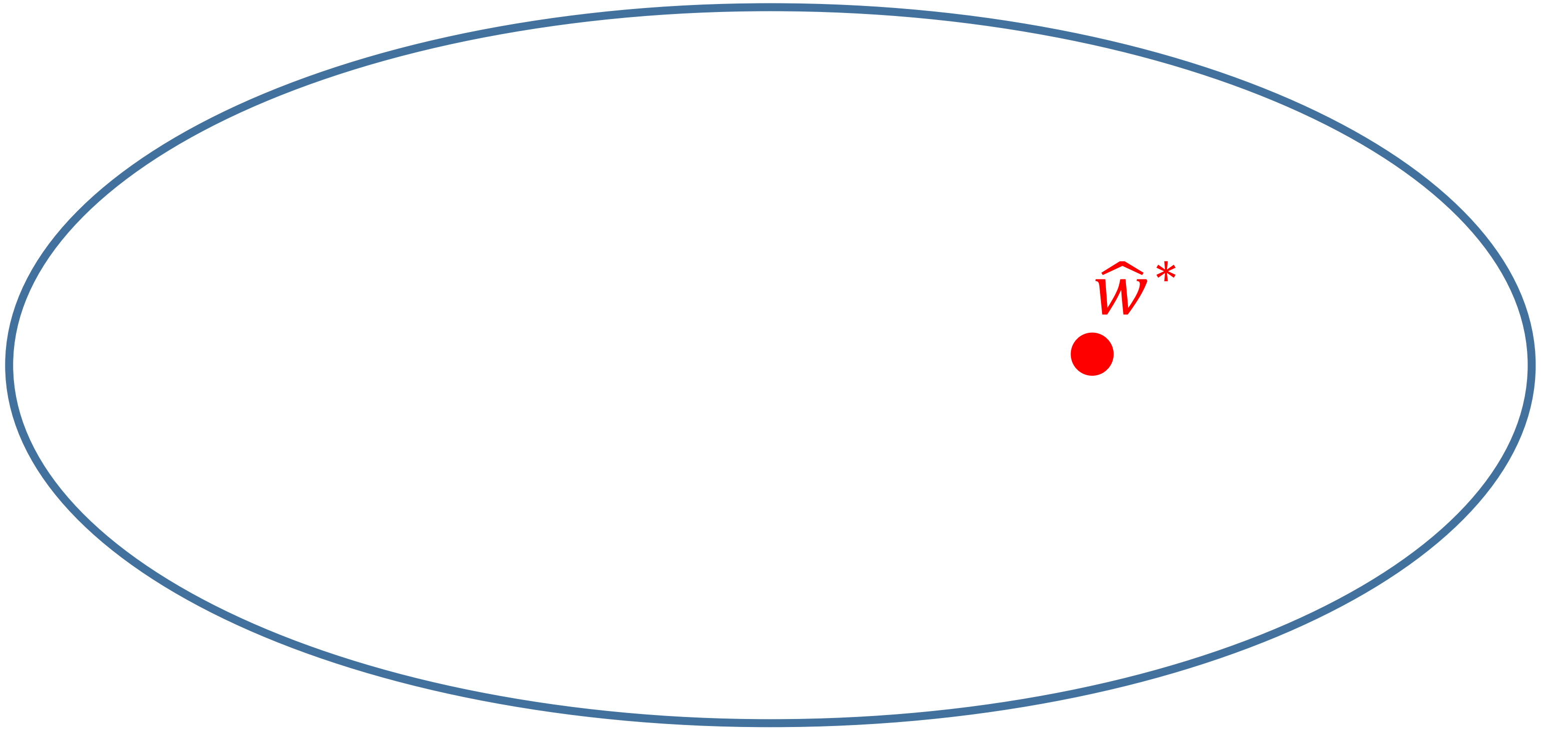
$$\frac{1}{2}||w||^2 \leq R$$

$$y_i(w^T x_i + b) \geq 1, \forall i$$

- Decrease R until cannot find w satisfying the inequalities

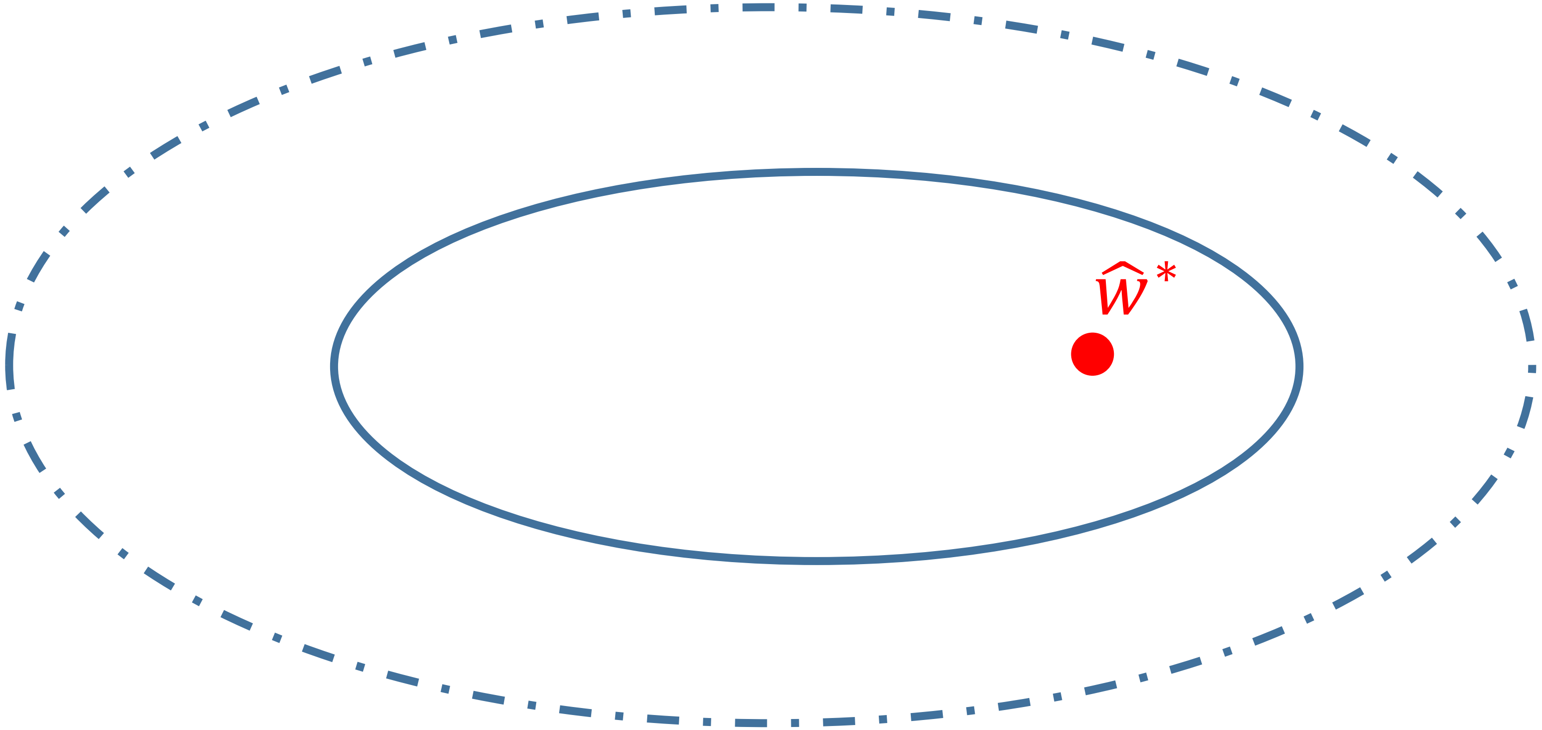
Thought experiment

- $\hat{w}^*$ is the best weight (i.e., satisfying the smallest R)



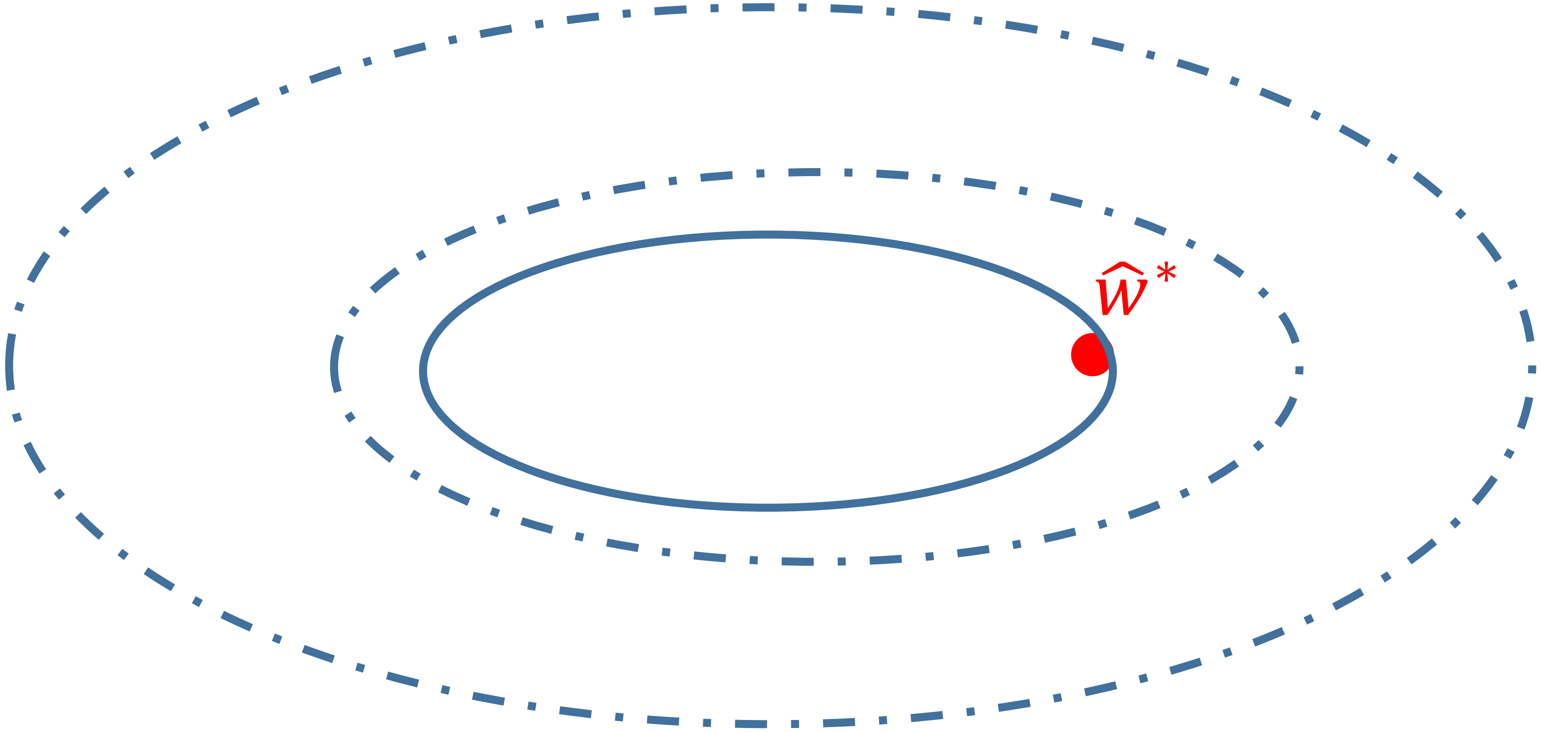
Thought experiment

- $\widehat{w}^*$ is the best weight (i.e., satisfying the smallest R)



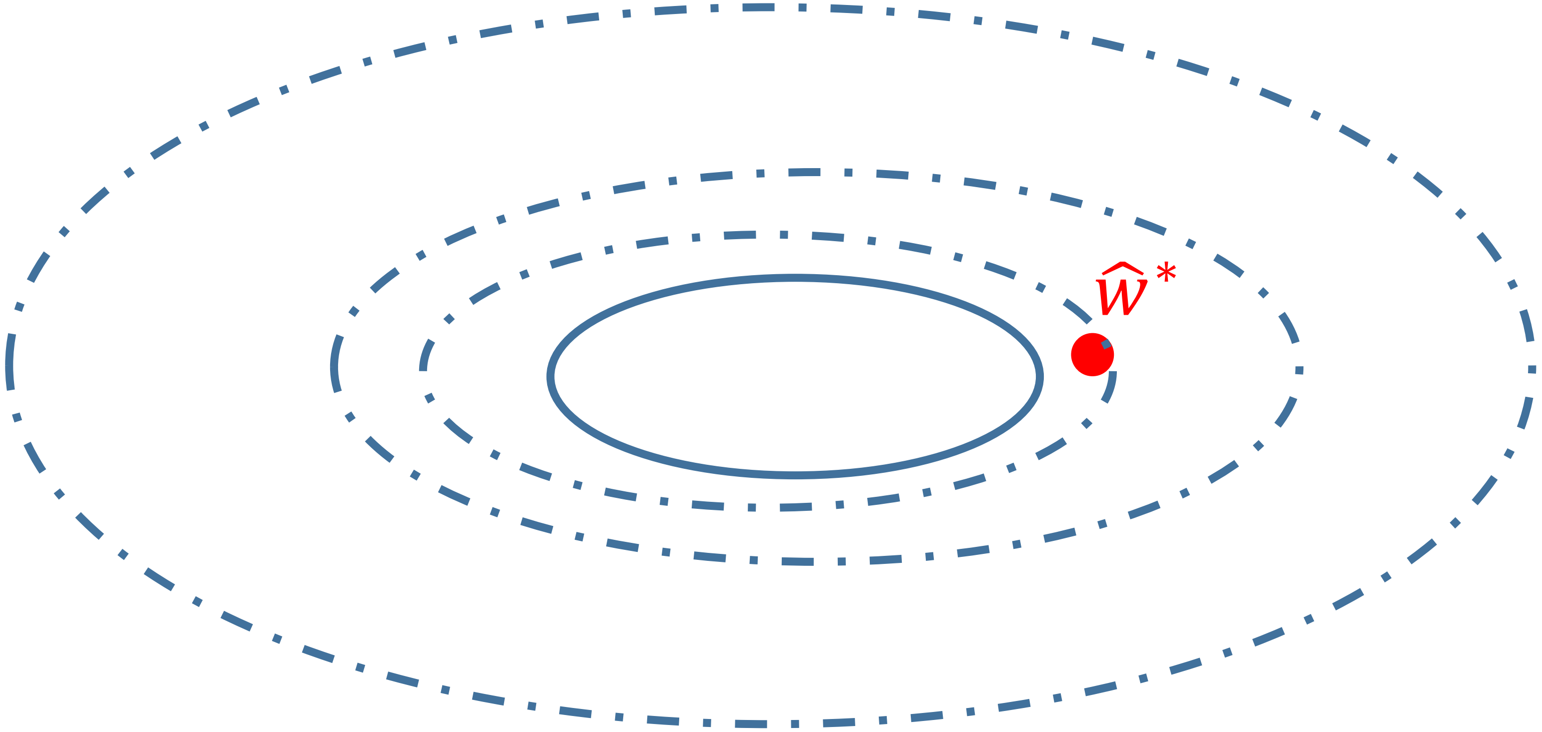
Thought experiment

- $\hat{w}^*$ is the best weight (i.e., satisfying the smallest R)



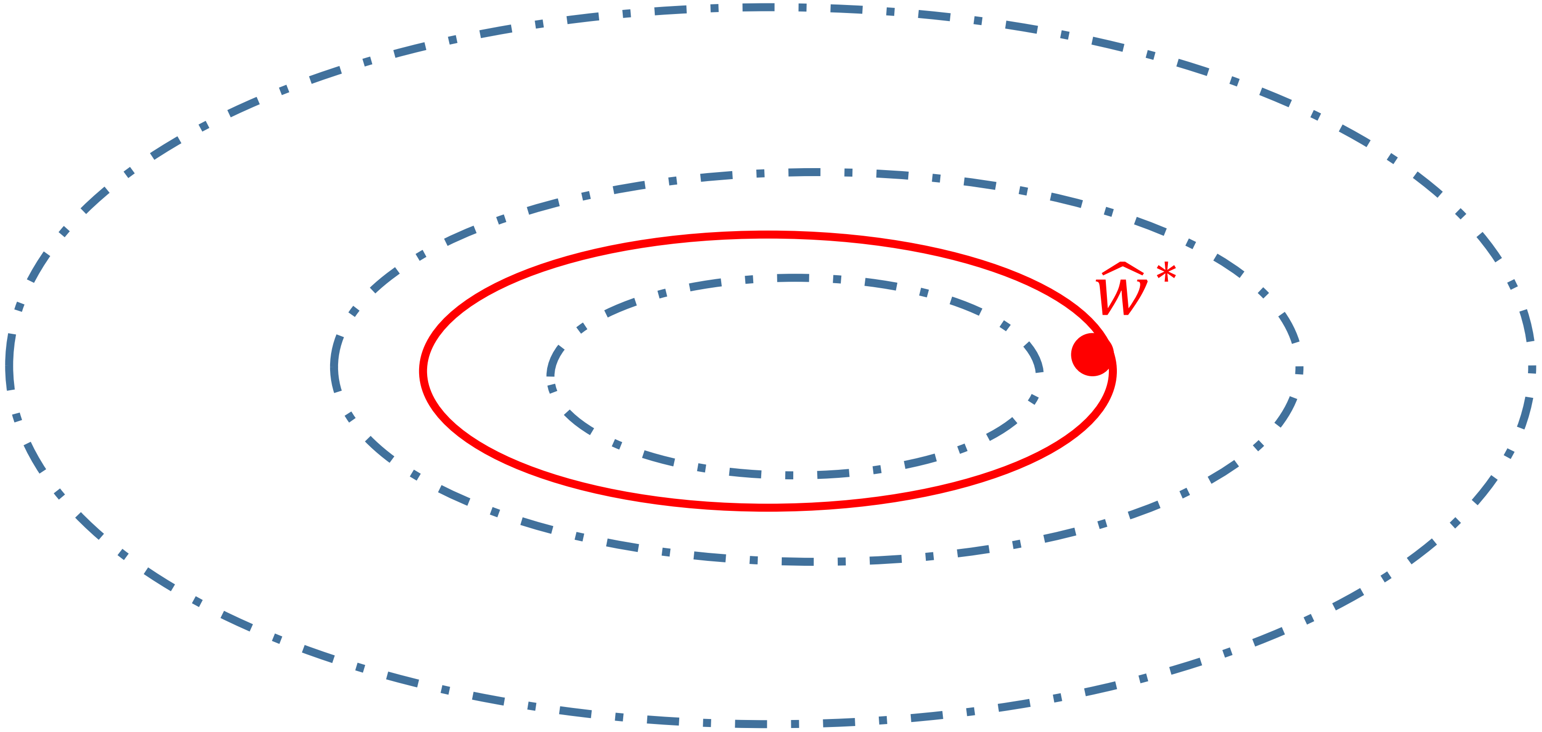
Thought experiment

- $\hat{w}^*$ is the best weight (i.e., satisfying the smallest R)



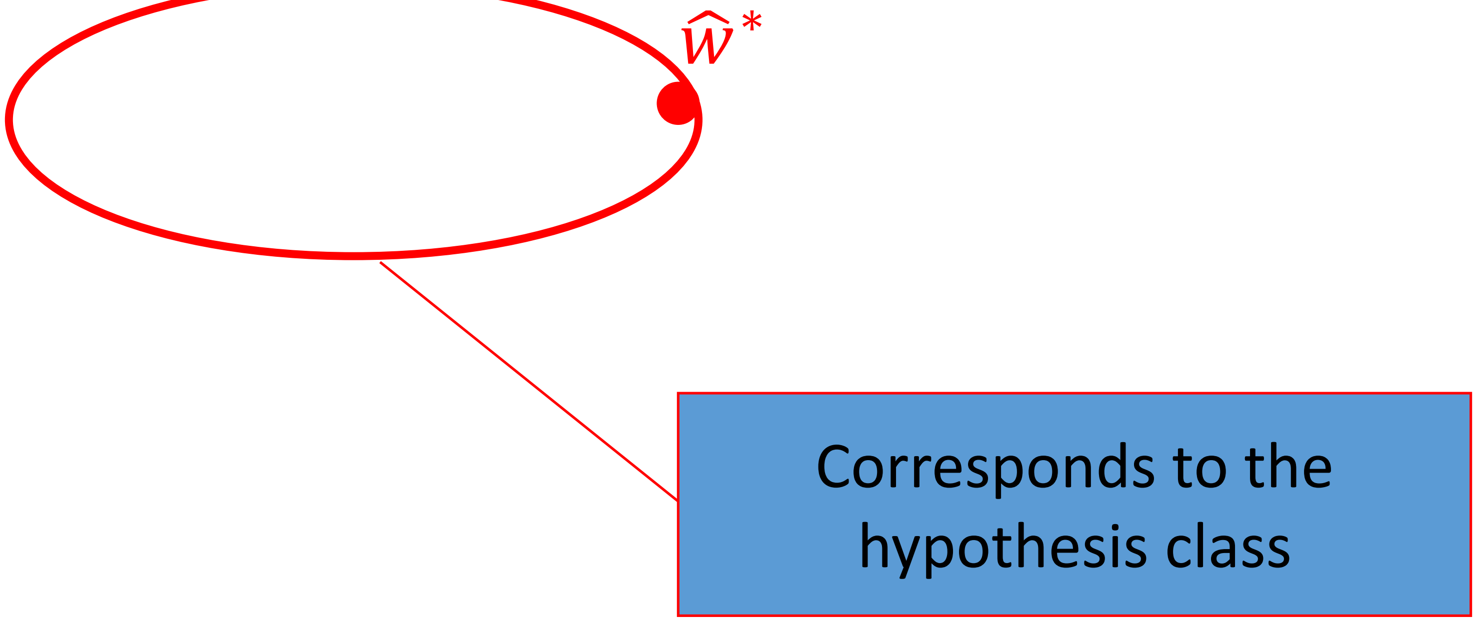
Thought experiment

- $\widehat{w}^*$ is the best weight (i.e., satisfying the smallest R)



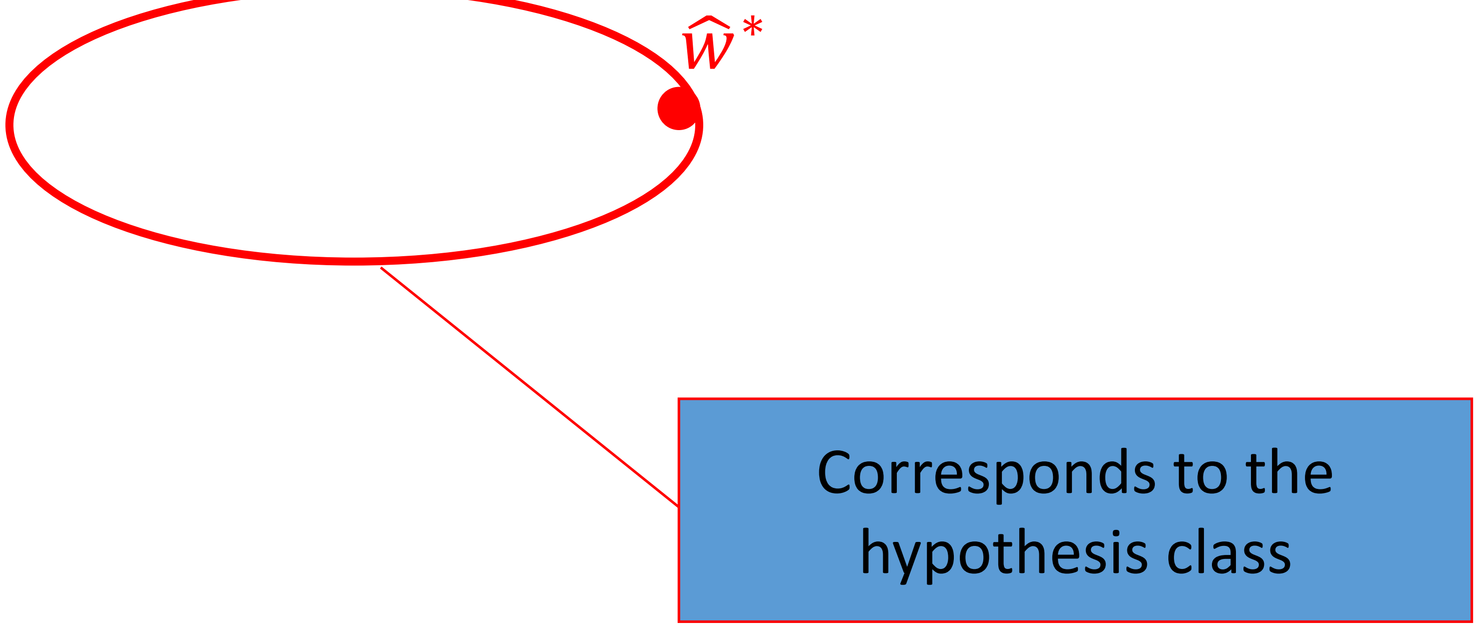
Thought experiment

- To handle the difference between empirical and expected losses →
- Choose large margin hypothesis (high confidence) →
- Choose a small hypothesis class



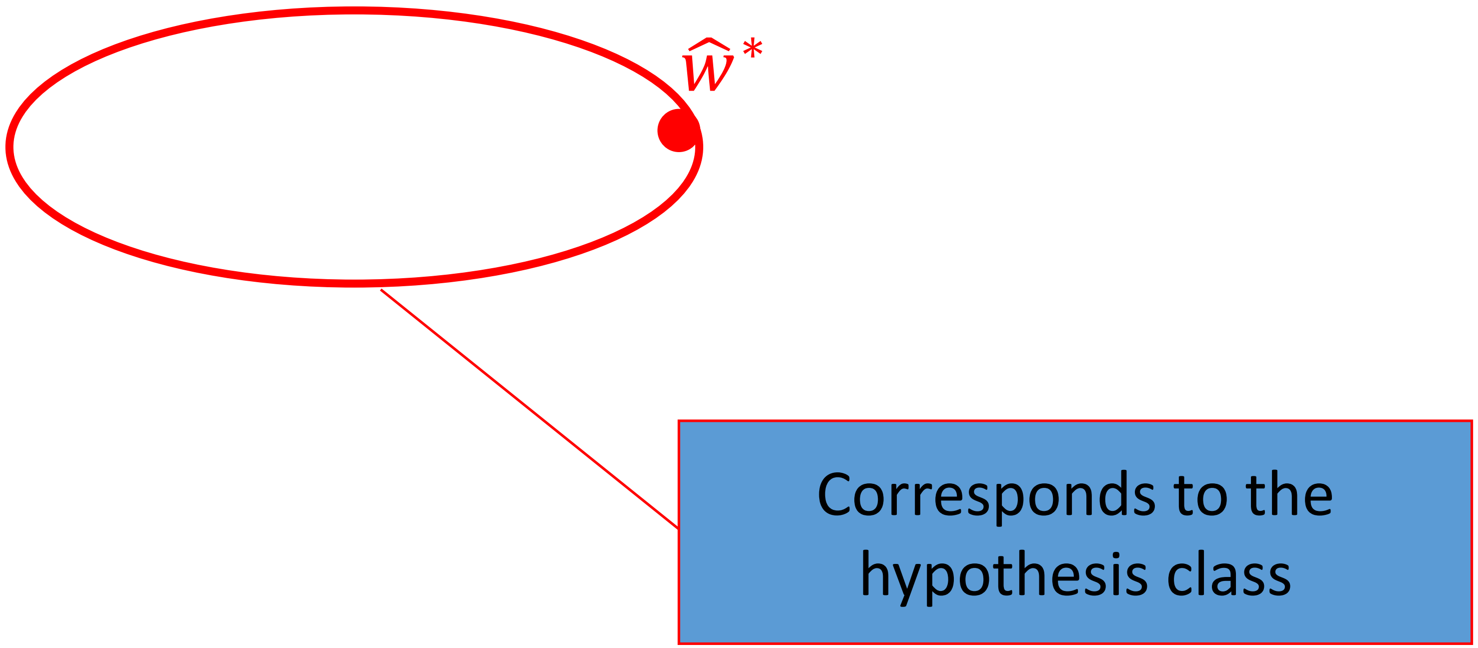
Thought experiment

- Principle: use smallest hypothesis class still with a correct/good one
  - Also true beyond SVM
  - Also true for the case without perfect separation between the two classes
  - Math formulation: VC-dim theory, etc.



Thought experiment

- Principle: use smallest hypothesis class still with a correct/good one
  - Whatever you know about the ground truth, add it as constraint/regularizer



SVM: optimization

- Optimization (Quadratic Programming):

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

$$y_i(w^T x_i + b) \geq 1, \forall i$$

- Solved by Lagrange multiplier method:

$$\mathcal{L}(w, b, \boldsymbol{\alpha}) = \frac{1}{2}||w||^2 - \sum_i \alpha_i [y_i(w^T x_i + b) - 1]$$

  where $\boldsymbol{\alpha}$ is the Lagrange multiplier
- Details in next lecture

Reading

- Review Lagrange multiplier method
- E.g. Section 5 in Andrew Ng's note on SVM
  - posted on the course website: http://www.cs.princeton.edu/courses/archive/spring16/cos495/