

# CS489/698: Intro to ML

## Lecture 07: Soft-margin SVM

# Announcement

- A2 due **tonight**
- Proposal due next **Tuesday**
- **ICLR 2018 Reproducibility Challenge**
  - select a paper from 2018 ICLR submissions, and aim to replicate the experiments described in the paper
  - produce a Reproducibility report, describing the target questions, experimental methodology, implementation details, analysis and discussion of findings, conclusions on reproducibility of the paper

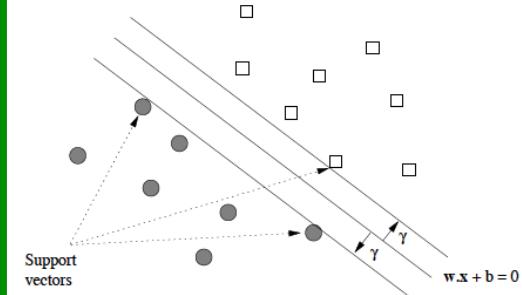
# Outline

- Formulation
- Dual
- Optimization
- Extension

# Hard-margin SVM

Primal

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 \\ \text{s.t. } \forall i, \quad & y_i (\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \end{aligned}$$

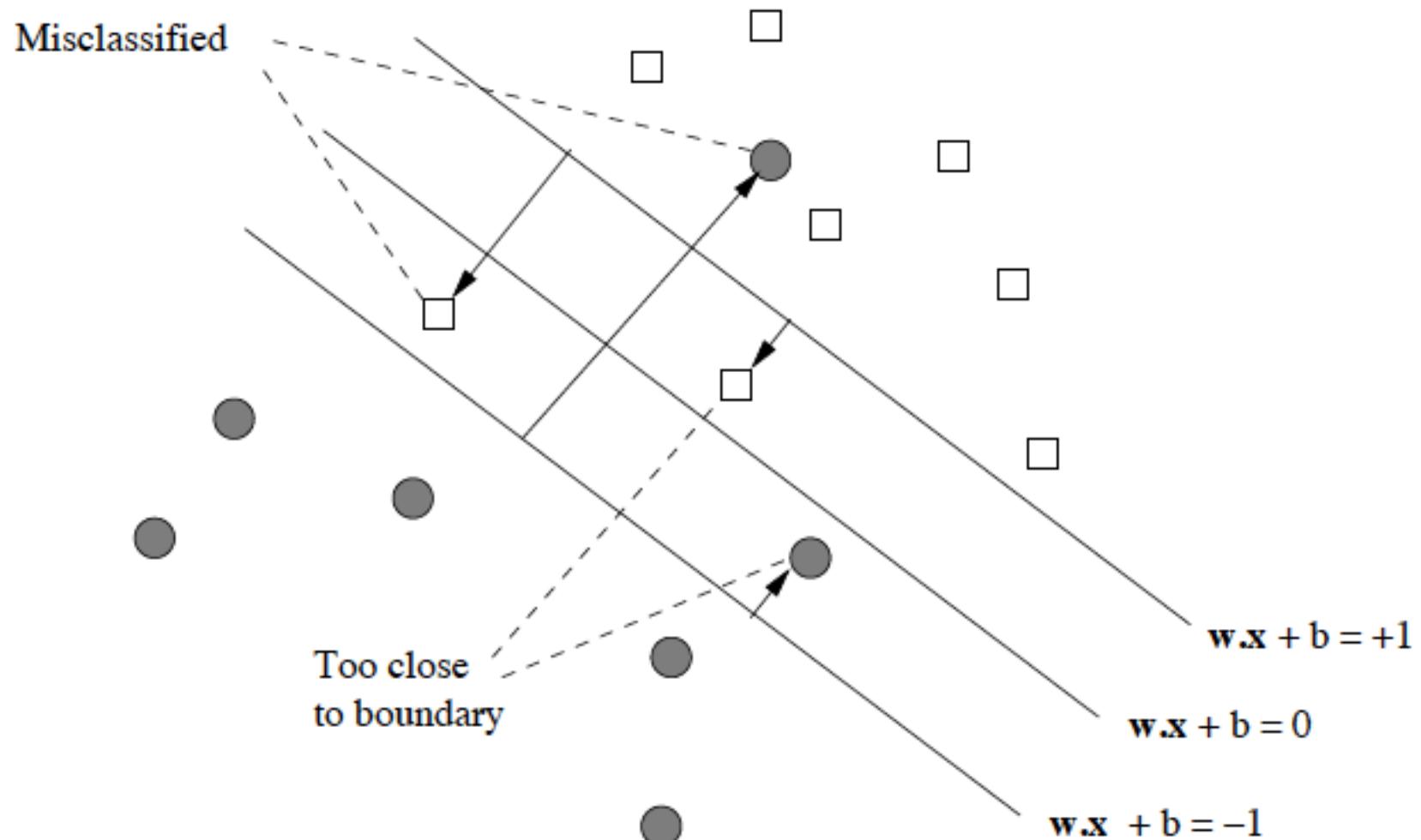


Dual

$$\begin{aligned} \min_{\alpha \geq 0} \quad & \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j - \sum_k \alpha_k \\ \text{s.t.} \quad & \sum_i \alpha_i y_i = 0 \end{aligned}$$

hard constraint

# What if **inseparable**?



# Soft-margin (Cortes & Vapnik'95)

Primal

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2$$

$$\text{s.t. } \forall i, y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$$

hard constraint

proto 1/margin

hyper-parameter

training error

Primal

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^n (1 - y_i \hat{y}_i)_+$$

soft constraint

$$\forall i, \hat{y}_i = \mathbf{w}^\top \mathbf{x}_i + b$$

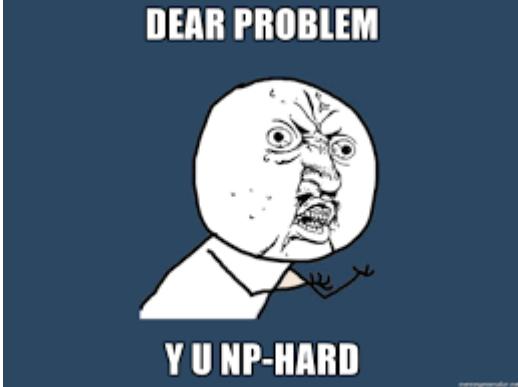
prediction  
(no sign)

# Zero-one loss

$$\Pr(\text{sign}(f(X)) \neq Y) = E[1_{Y f(X) \geq 0}]$$

$\hat{Y}$

your prediction



$w^T x_i + b$

$$\frac{1}{n} \sum_{i=1}^n 1_{Y_i \hat{Y}_i \geq 0}$$

- Find prediction rule  $f$  so that on an unseen random  $X$ , our prediction  $\text{sign}(f(X))$  has small chance to be different from the true label  $Y$

# The hinge loss

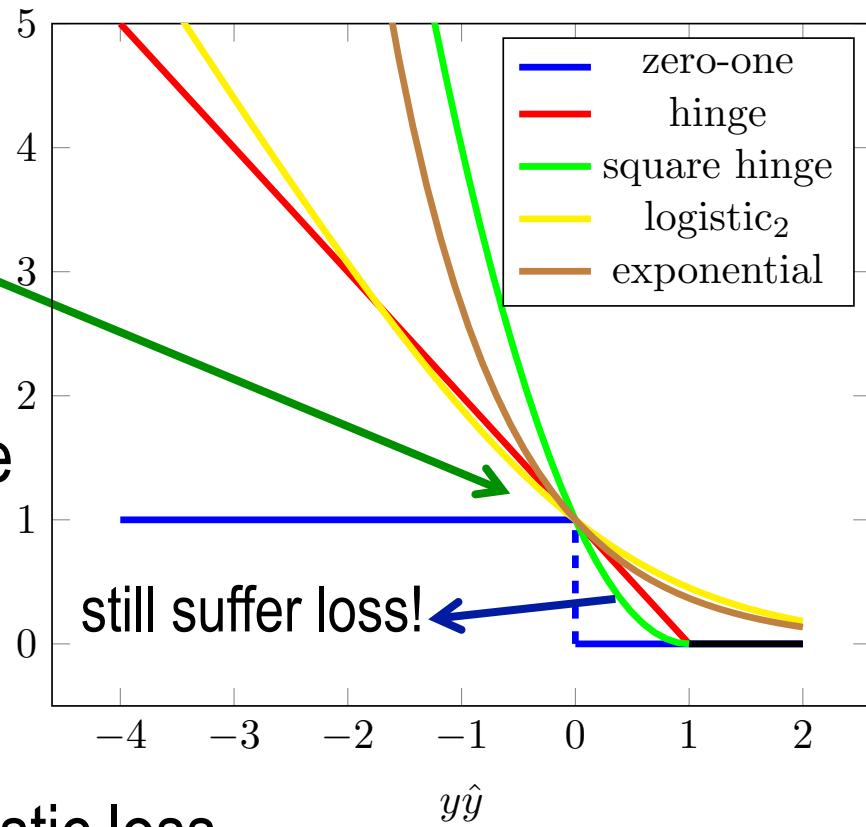
$$(1 - y\hat{y})_+ = \max\{1 - y\hat{y}, 0\}$$



- $1 - y\hat{y} \geq 0$  zero-one  
Squared hinge  $(1 - y\hat{y})_+^2$   
 $\exp(-y\hat{y})$  exponential loss  
 $\log_2(1 + \exp(-y\hat{y}))$  logistic loss

upper bound

loss



# Classification-calibration

- Want to minimize zero-one loss
- End up with minimizing some **other** loss

**Theorem (Bartlett, Jordan, McAuleiffe'06).** Any convex margin loss  $\ell$  is classification-calibrated **iff**  $\ell$  is differentiable at 0 and  $\ell'(0) < 0$ .  $\eta(x) = p(Y=1|x) = \omega^T x + b$

**Classification calibration.**  $\arg \min_a \mathbf{E}[\ell(Ya)|X=x]$   
has the same sign as  $2\eta(x) - 1$ , i.e., the Bayes rule.

$$\eta(x)\ell(a) + (1 - \eta(x))\ell(-a)$$

# Outline

- Formulation
- Dual
- Optimization
- Extension

# Important optimization trick

$$\min_x f(x)$$



joint over  
 $x$  and  $t$

$$\min_{x,t} t$$

$$\text{s.t. } f(x) \leq t$$



$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^n (1 - y_i \hat{y}_i)_+$$

$$\forall i, \hat{y}_i = \mathbf{w}^\top \mathbf{x}_i + b$$



↓

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^n \xi_i$$

Slack for  
“wrong”  
prediction

s.t.  $\forall i, (1 - y_i \hat{y}_i)_+ \leq \xi_i \rightarrow \begin{cases} 1 - y_i \hat{y}_i \leq \xi_i \\ 0 \leq \xi_i \end{cases}$

# Lagrangian

$$\min_{\mathbf{w}, b, \xi} \max_{\alpha \geq 0, \beta \leq 0} \frac{1}{2} \|\mathbf{w}\|_2^2 + \sum_i C\xi_i + \alpha_i(1 - y_i \hat{y}_i - \xi_i) + \beta_i \xi_i$$
$$\max_{\alpha \geq 0, \beta \leq 0} \min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|_2^2 + \sum_i C\xi_i + \alpha_i(1 - y_i \hat{y}_i - \xi_i) + \beta_i \xi_i$$


$$\frac{\partial}{\partial b} = \sum_i \alpha_i y_i = 0$$

$$\frac{\partial}{\partial \xi_i} = C - \alpha_i + \beta_i = 0$$

$$\frac{\partial}{\partial \mathbf{w}} = \mathbf{w} - \sum_i \alpha_i y_i \mathbf{x}_i = 0$$

# Dual problem

$$\max_{C \geq \alpha \geq 0} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$$

$$\text{s.t. } \sum_i \alpha_i y_i = 0$$

only dot product is  
needed!

$$\min_{C \geq \alpha \geq 0} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j - \sum_{i=1}^n \alpha_i$$

# The effect of C

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^n (1 - y_i \hat{y}_i)_+$$

**R<sup>d</sup>xR**  $\forall i, \hat{y}_i = \mathbf{w}^\top \mathbf{x}_i + b$

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2$$

s.t.  $\forall i, y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$

- $C \rightarrow 0?$
- $C \rightarrow \infty?$

$$\min_{\alpha \geq 0} \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j - \sum_i \alpha_i$$

s.t.  $\sum_i \alpha_i y_i = 0$

**R<sup>n</sup>**

$$\min_{C \geq \alpha \geq 0} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j - \sum_{i=1}^n \alpha_i$$

s.t.  $\sum_i \alpha_i y_i = 0$

# Karush-Kuhn-Tucker conditions

- Primal constraints on  $\mathbf{w}$ ,  $b$  and  $\xi$ :  $(1 - y_i \hat{y}_i)_+ \leq \xi_i$
- Dual constraints on  $\alpha$  and  $\beta$ :  $\alpha \geq 0$   $\beta \leq 0$   $c \geq \alpha_i$
- Complementary slackness

$$\alpha_i(1 - y_i \hat{y}_i - \xi_i) = 0$$

$$(c - \alpha_i) \beta_i \xi_i = 0$$

- Stationarity

$$\frac{\partial}{\partial b} = \sum_i \alpha_i y_i = 0$$
$$\frac{\partial}{\partial \xi_i} = C - \alpha_i + \beta_i = 0$$
$$\frac{\partial}{\partial \mathbf{w}} = \mathbf{w} - \sum_i \alpha_i y_i \mathbf{x}_i = 0$$

$$\min_{\mathbf{w}, b, \xi} \max_{\alpha \geq 0, \beta \leq 0} \frac{1}{2} \|\mathbf{w}\|_2^2 + \sum_i C \xi_i + \alpha_i(1 - y_i \hat{y}_i - \xi_i) + \beta_i \xi_i$$

# Parsing the equations

$$\alpha_i(1 - y_i \hat{y}_i - \xi_i) = 0$$

$$(C - \alpha_i)\xi_i = 0$$

$$(1 - y_i \hat{y}_i)_+ \leq \xi_i$$

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$$

- $\alpha_i < C \implies \xi_i = 0 \implies y_i \hat{y}_i \geq 1$
- $\alpha_i > 0 \implies 1 - y_i \hat{y}_i - \xi_i = 0 \Rightarrow y_i \hat{y}_i = 1 - \xi_i \leq 1$



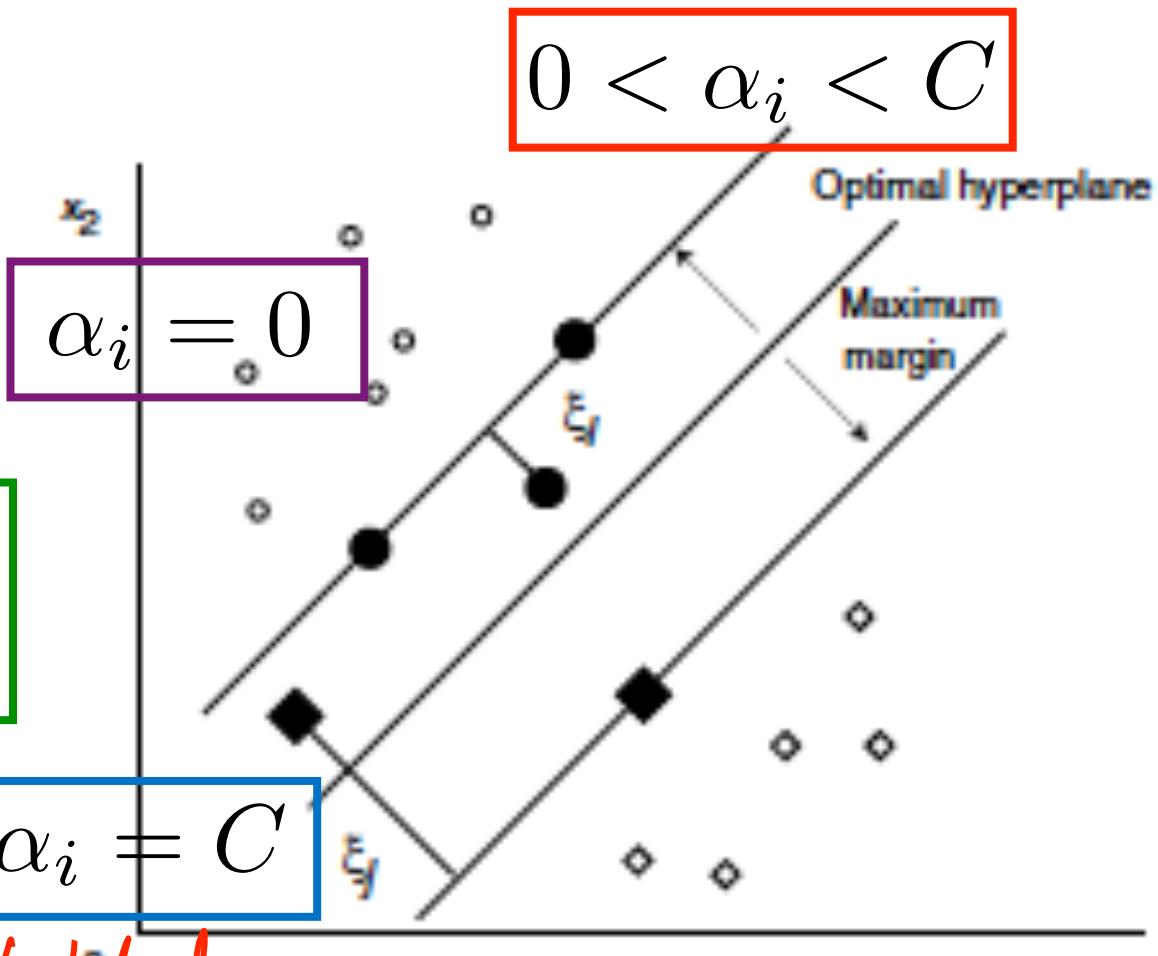
# Support Vectors

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$$

if  $\mathbf{x}_i$  is strictly beyond  
the sep. hyperplane

then  $\alpha_i = 0$

if  $\alpha_i = 0$  then  $\mathbf{x}_i$  is {strictly beyond  
on the hyperplane}



# Recover b

- Take any i such that  $C > \alpha_i > 0$
- Then  $x_i$  is on the hyperplane:

$$1 - y_i \hat{y}_i = 0$$

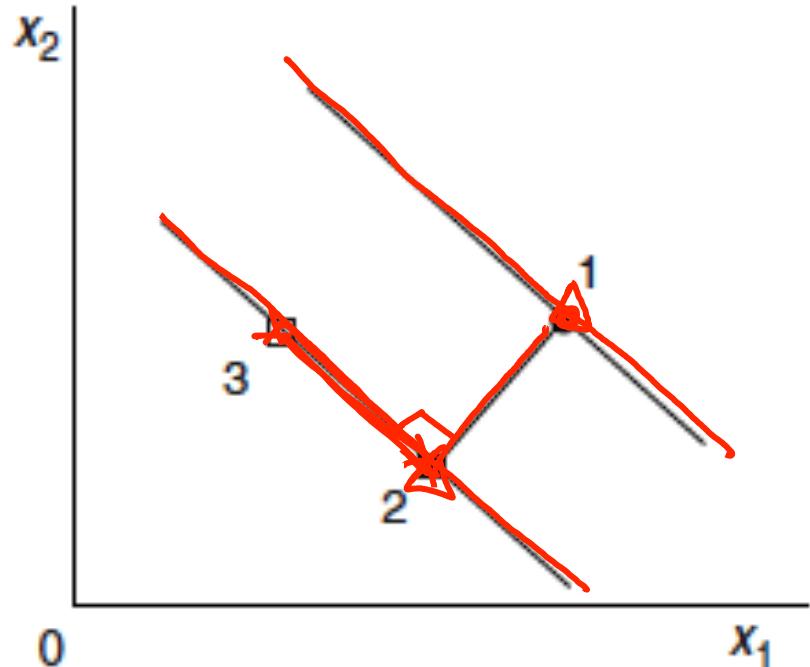
$$\mathbf{w}^T x_i + b$$

$$(1 - y_i \hat{y}_i)_+ \neq \xi_i$$

- How to recover  $\xi$  ?

- What if there is no such i ?

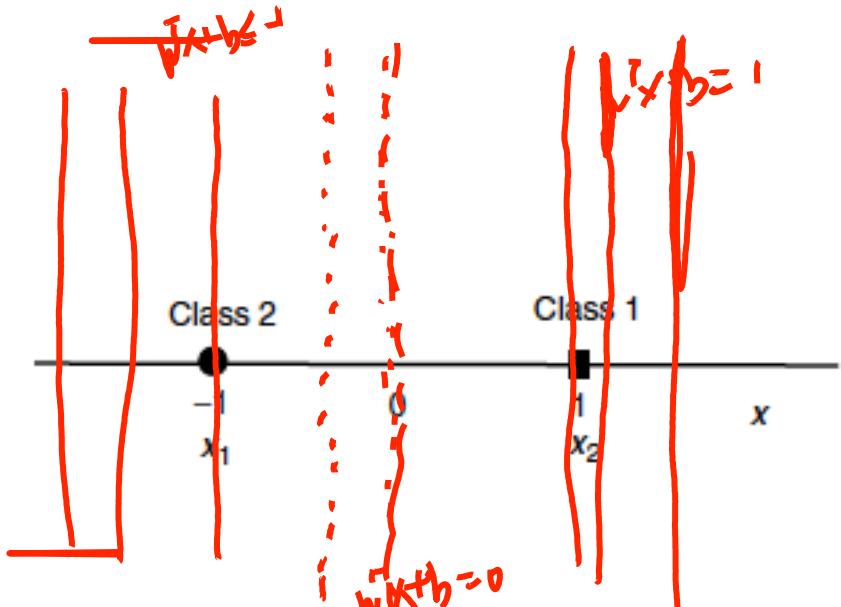
# More examples



$$w = \sum_i \alpha_i y_i \vec{x}_i$$

$$\alpha_i > 0 \implies 1 - y_i \hat{y}_i - \xi_i = 0$$

$$C < \frac{1}{2} \quad \alpha_1^* = \alpha_2^* = C$$



$$\min_{\substack{C \geq \alpha_1 \geq 0}} 2\alpha_1^2 - 2\alpha_1$$

$$\alpha_i^* = \min \left\{ \frac{1}{2}, C \right\}$$

# Just in case

# Outline

- Formulation
- Dual
- Optimization
- Extension

# Gradient Descent

$$\min_{\mathbf{w}, b}$$

$$L(\mathbf{w}) := \frac{C}{n} \sum_{i=1}^n \ell(y_i \hat{y}_i) + \frac{1}{2} \|\mathbf{w}\|_2^2$$

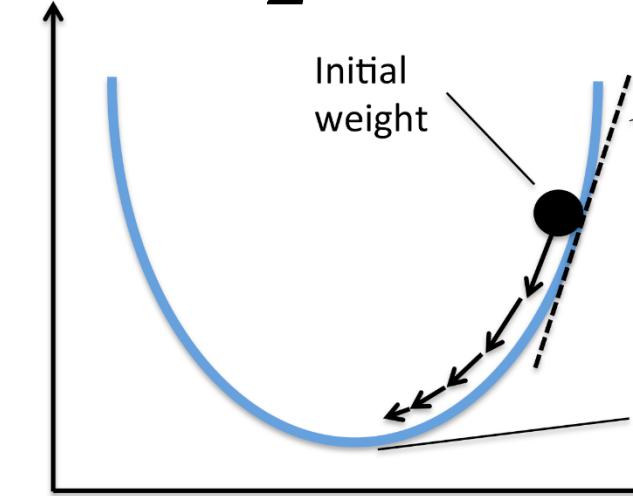
$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \nabla L(\mathbf{w}_t)$$

- Step size (learning rate)
- const., if  $L$  is smooth
  - diminishing, otherwise

$$\frac{C}{n} \sum_i \ell'(y_i \hat{y}_i) y_i \mathbf{x}_i + \mathbf{w}_t$$

(Generalized) gradient

$O(nd)$  !



# Stochastic Gradient Descent (SGD)

$$\mathbf{w}_{t+1} = (1 - \eta_t) \mathbf{w}_t - \eta_t C \frac{1}{n} \sum_i \ell'(y_i \hat{y}_i) y_i \mathbf{x}_i$$

↳ average over  $n$  samples

a random sample suffices



$$\mathbf{w}_{t+1} = (1 - \eta_t) \mathbf{w}_t - \eta_t C \ell'(y_{i_t} \hat{y}_{i_t}) y_{i_t} \mathbf{x}_{i_t}$$

- diminishing step size, e.g.,  $1/\sqrt{t}$  or  $1/t$
- averaging, momentum, variance-reduction, etc.
- sample w/o replacement; cycle; permute in each pass

$O(d)$

# The derivative

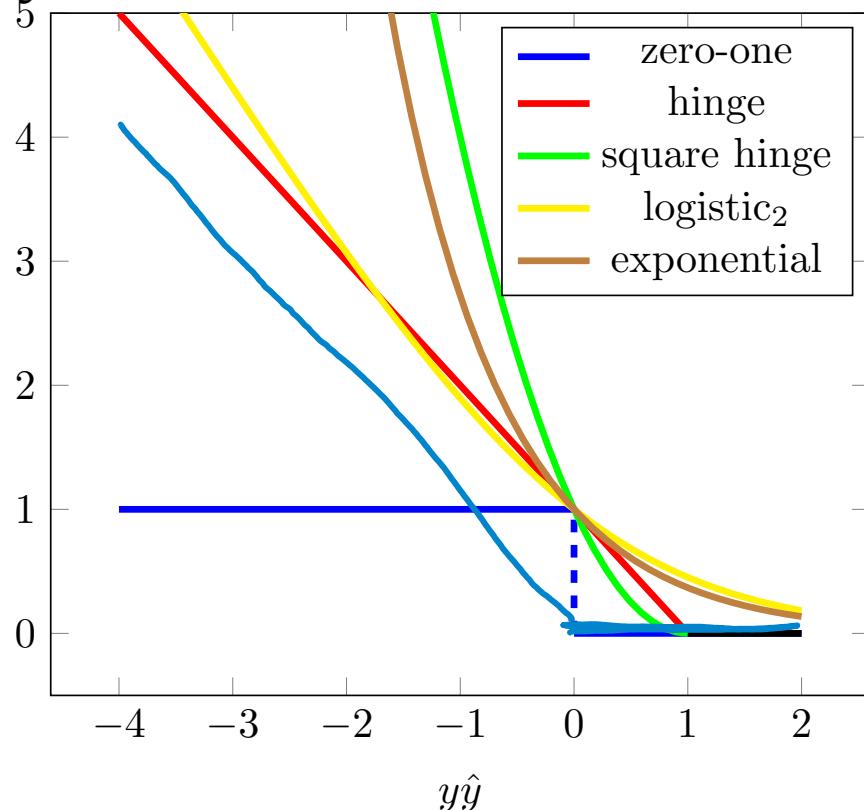
$$\ell'_{\text{hinge}}(t) = \begin{cases} -1, & t \leq 1 \\ 0, & t \geq 1 \end{cases}$$

What about zero-one loss?

All other losses are diff.

What about perceptron?

$$\mathbf{w}_{t+1} = (1 - \eta_t) \mathbf{w}_t - \eta_t C \ell'(y_i \hat{y}_i) y_i \mathbf{x}_i$$



# Solving the dual

$$\begin{array}{ll} \min_{\substack{\alpha \geq 0}} & \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j - \sum_{i=1}^n \alpha_i \\ \text{s.t.} & \sum_i \alpha_i y_i = 0 \end{array}$$

$$\alpha_{t+1} \leftarrow \alpha_t - \eta_t (K \odot \mathbf{y} \mathbf{y}^\top) \alpha_t + \eta_t \mathbf{1}$$

$\longrightarrow$

$$\alpha_{t+1} \leftarrow \text{prox}(\alpha_{t+1}) \quad O(n^*n)$$

- Can choose constant step size  $\eta_t = \eta$
- Faster algorithms exist: e.g., choose a pair of  $\alpha_p$  and  $\alpha_q$  and derive a closed-form update

# A little history on optimization

- Gradient descent mentioned first in (Cauchy, 1847)

ANALYSE MATHÉMATIQUE. — *Méthode générale pour la résolution des systèmes d'équations simultanées; par M. AUGUSTIN CAUCHY.*

- First rigorous convergence proof (Curry, 1944)

THE METHOD OF STEEPEST DESCENT FOR NON-LINEAR  
MINIMIZATION PROBLEMS\*

By HASKELL B. CURRY (*Frankford Arsenal*)

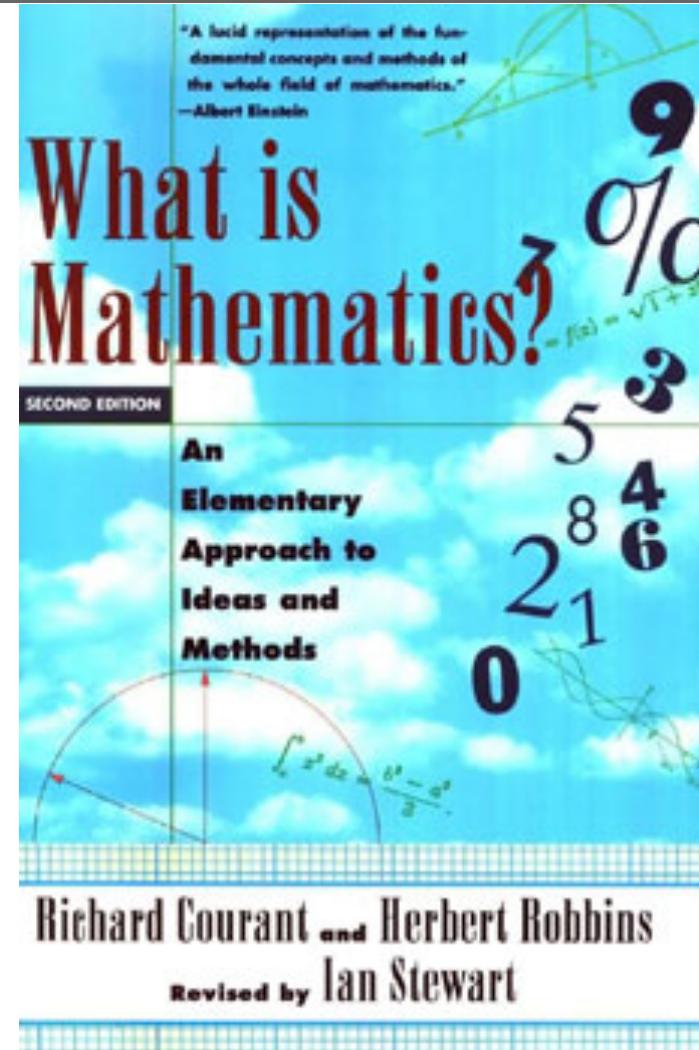
- SGD proposed and analyzed (Robbins & Monro, 1951)

A STOCHASTIC APPROXIMATION METHOD<sup>1</sup>

By HERBERT ROBBINS AND SUTTON MONRO

*University of North Carolina*

# Herbert Robbins (1915 – 2001)



# Outline

- Formulation
- Dual
- Optimization
- Extension

# Multiclass (Crammer & Singer'01)

$$\min_W \frac{1}{2} \|W\|_F^2$$

s.t.  $\forall i, \forall k \neq y_i,$

$$\boxed{\mathbf{x}_i^\top \mathbf{w}_{y_i}}$$

separate by a “safety margin”

$$\geq 1 + \boxed{\mathbf{x}_i^\top \mathbf{w}_k}$$

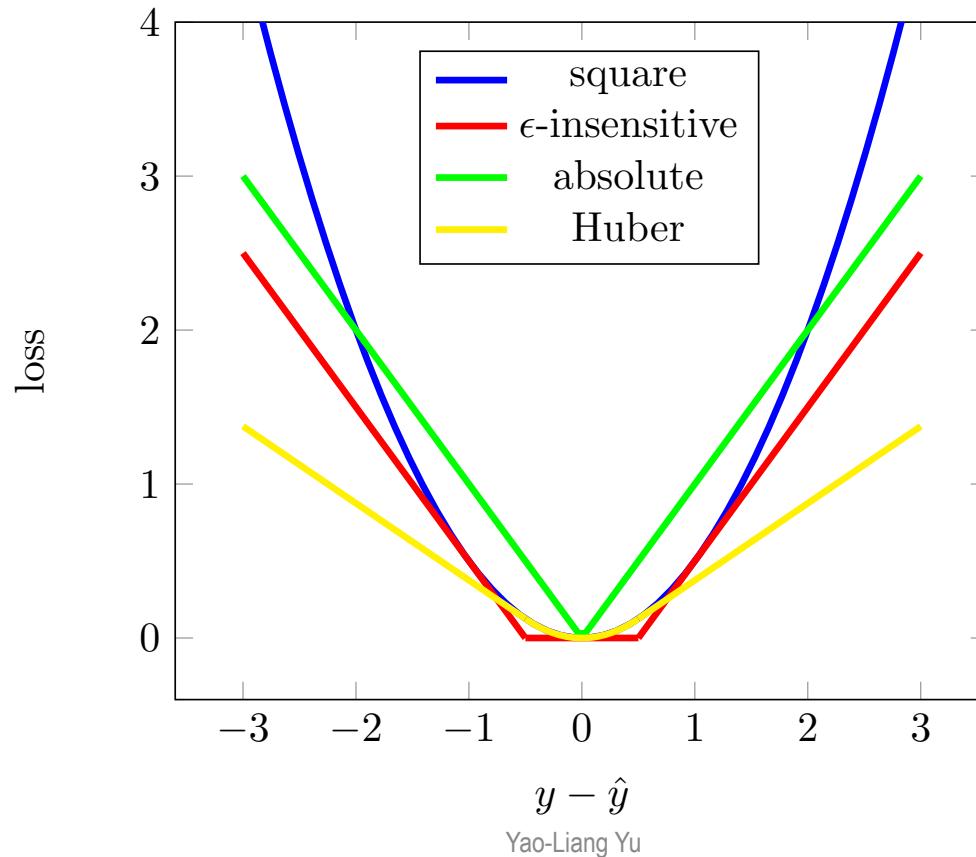
Prediction for  
correct class

Prediction for  
wrong classes

- Soft-margin is similar
- Many other variants
- Calibration theory is more involved

# Regression (Drucker et al.'97)

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^n (|y - \hat{y}| - \epsilon)_+$$



# Large-scale training (You, Demmel, et al.'17)

- Randomly partition training data evenly into  $p$  nodes
- Train SVM independently on each node
- Compute center on each node
- For a test sample
  - Find the **nearest** center (node / SVM)
  - Predict using the corresponding node / SVM

# Questions?



# “Universal” rule

Hard constraint: you must satisfy



Soft constraint: OK to violate, if you pay!

