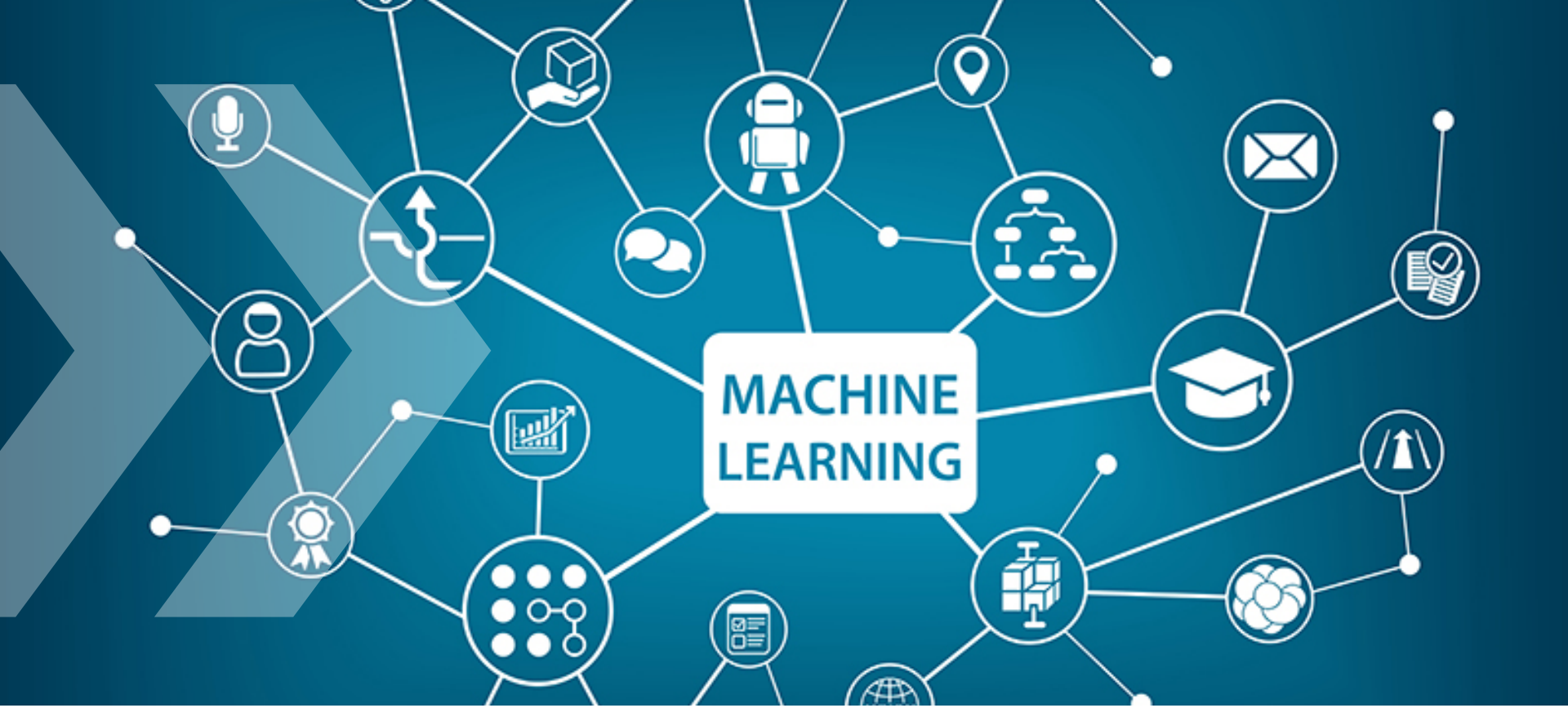


CS480/680: Intro to ML

Lecture 08: Soft-margin SVM

UNIVERSITY OF WATERLOO

Outline

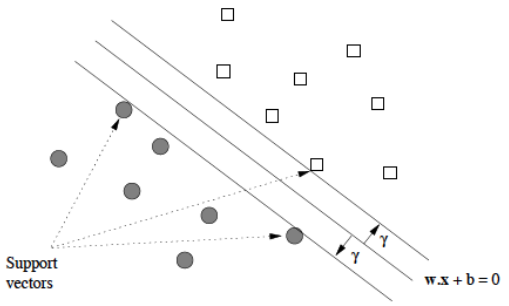
- Formulation
- Dual
- Optimization
- Extension

Hard-margin SVM

Primal

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

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



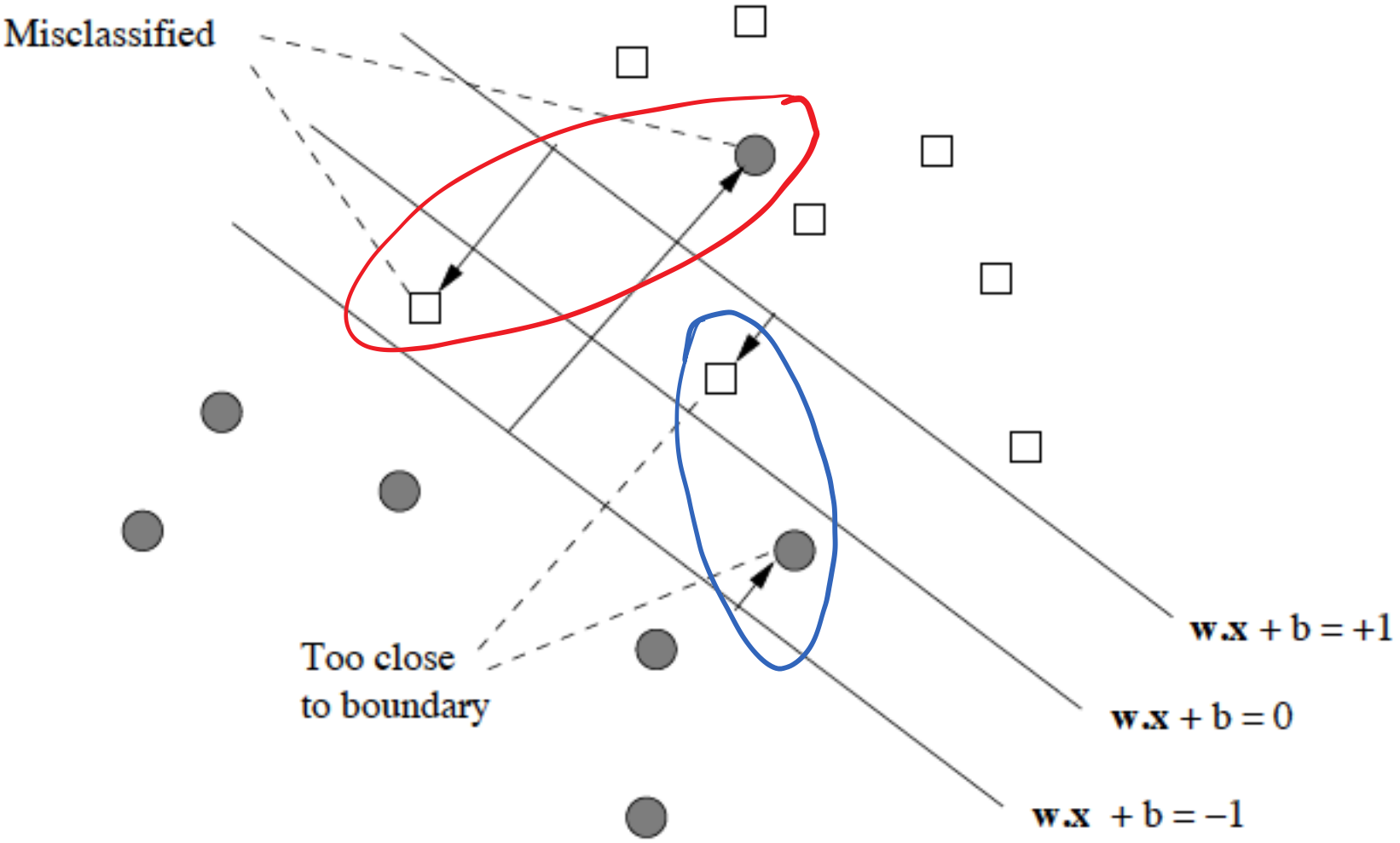
Dual

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

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

hard constraint

What if inseparable?



Soft-margin (Cortes & Vapnik'95)

Primal

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

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

hard constraint

$y_i\hat{y}_i \geq 1$

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

propto 1/margin

hyper-parameter

training error

$$= \begin{cases} 0, & 1 - y_i\hat{y}_i \leq 0 \\ 1 - y_i\hat{y}_i, & \text{o.w.} \end{cases}$$

Primal

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

soft constraint

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

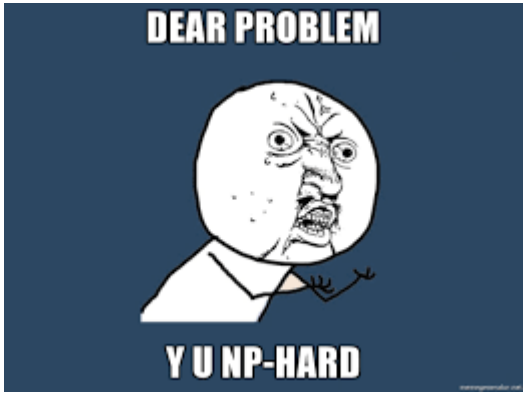
prediction (no sign)

Zero-one loss

$\hat{Y}$

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

your prediction



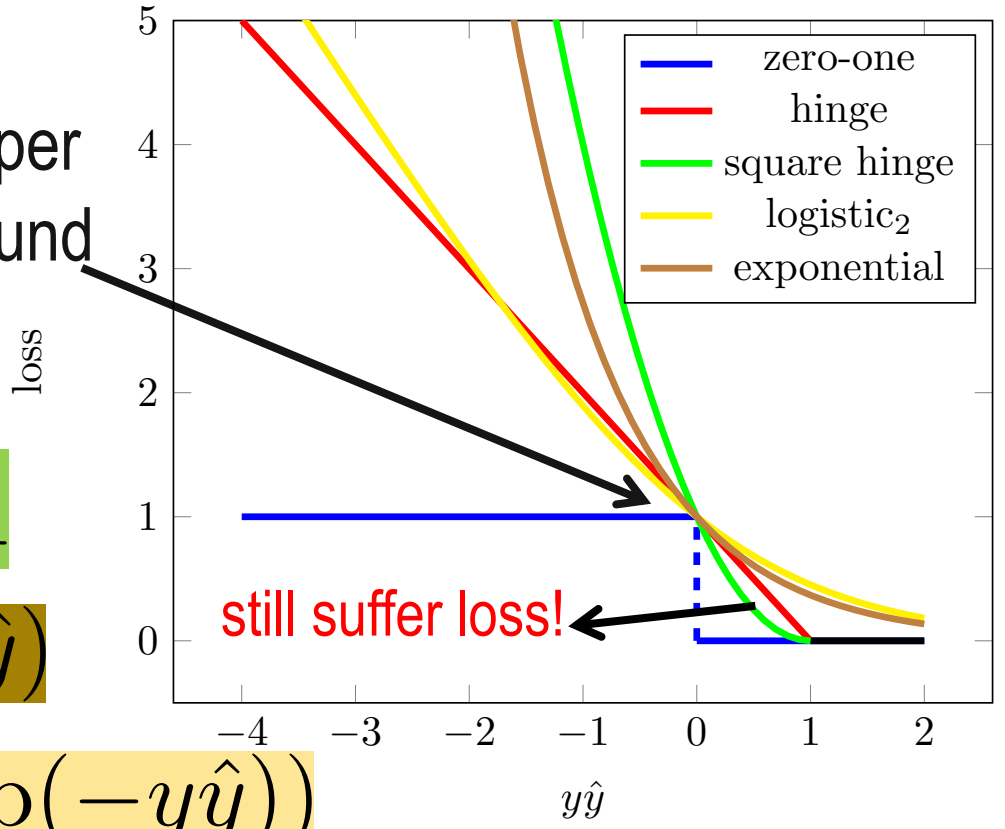
$$\frac{1}{n} \sum_{i=1}^{n} 1_{-Y_i \hat{Y}_i \geq 0}$$

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

The hinge loss

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

upper bound



zero-one $1_{-y\hat{y} \geq 0}$

squared hinge $(1 - y\hat{y})_+^2$

exponential loss $\exp(-y\hat{y})$

logistic loss $\log_2(1 + \exp(-y\hat{y}))$

Classification-calibration

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

Theorem (Bartlett, Jordan, McAuLiffe'06). Any convex margin loss ℓ is classification-calibrated iff ℓ is differentiable at 0 and $\ell'(0) < 0$.

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*

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

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

slack

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

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

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

Slack for "wrong" prediction

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

UNIVERSITY OF WATERLOO

Lagrangian

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

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

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

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

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

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

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

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

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

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

- C → 0?
- C → inf?

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

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

$\mathbf{R}^n$

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

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

Karush-Kuhn-Tucker conditions

- Primal constraints on **w**, b and ξ: $(1 - y_i \hat{y}_i)_+ \leq \xi_i$
- Dual constraints on α and β: $\boldsymbol{\alpha} \geq 0 \quad \boldsymbol{\beta} \leq 0$
- Complementary slackness

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

- Stationarity

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

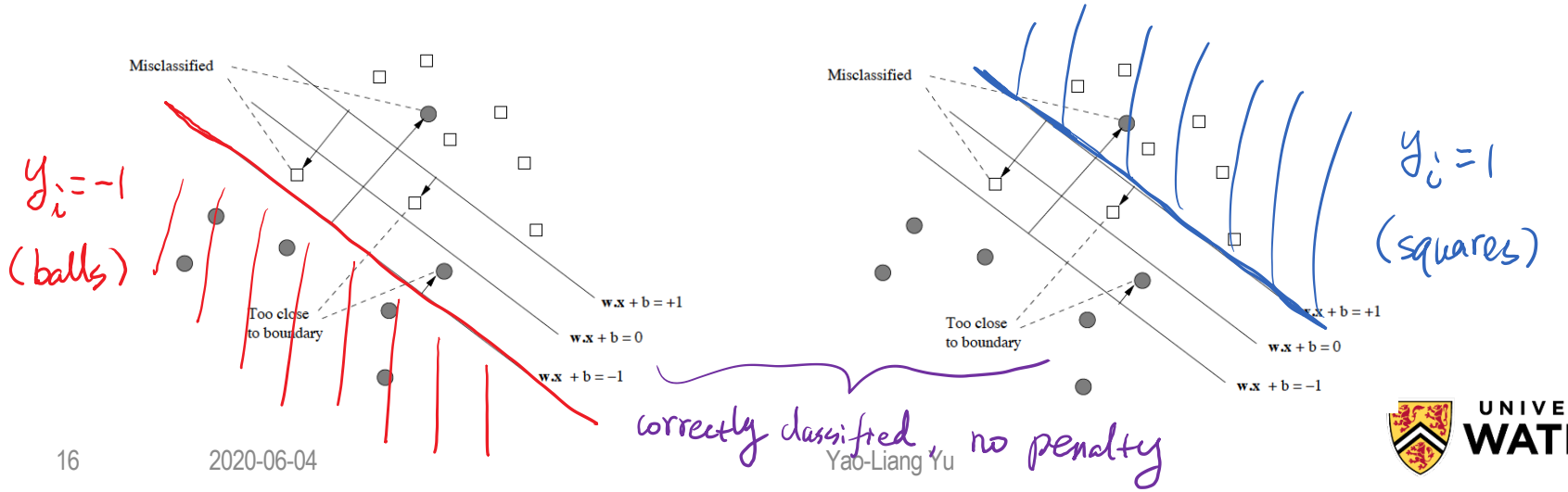
$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \max_{\alpha \geq 0, \beta \leq 0} \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_i C\xi_i + \alpha_i(1 - y_i \hat{y}_i - \xi_i) + \beta_i \xi_i$$

Parsing the equations

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

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

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



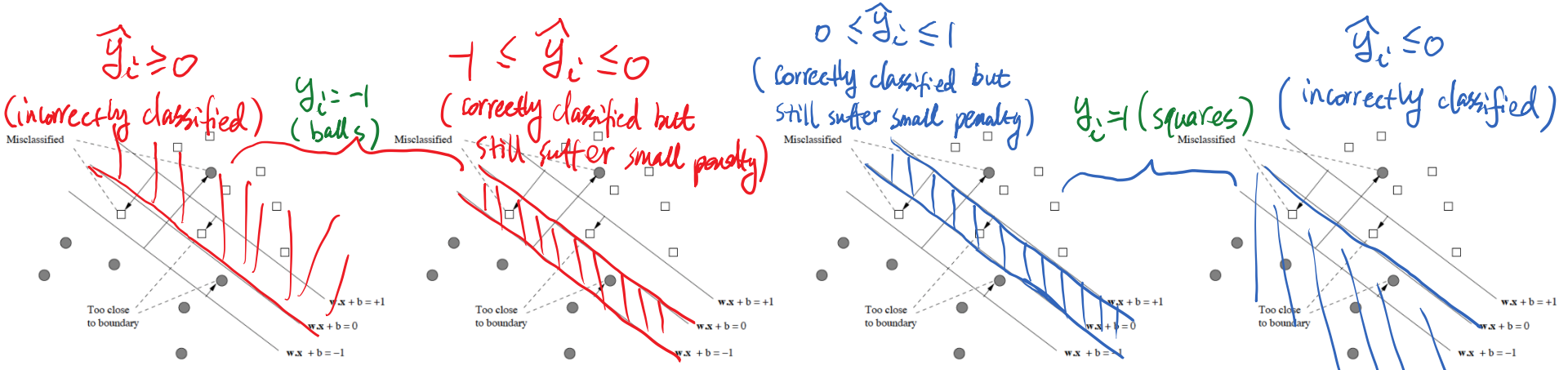
UNIVERSITY OF WATERLOO

Parsing the equations

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

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

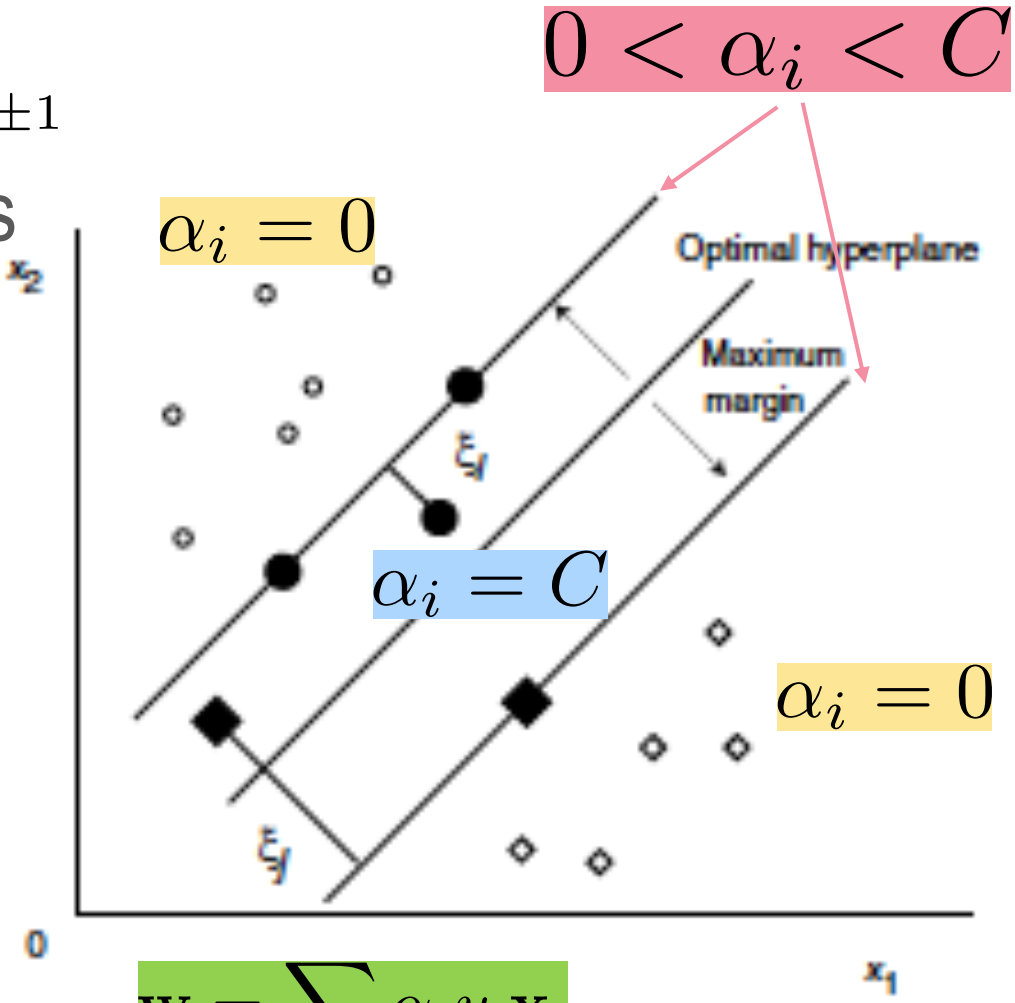
- $\alpha_i > 0 \implies 1 - y_i\hat{y}_i - \xi_i = 0 \implies y_i\hat{y}_i \leq 1$



WATERLOO

Support Vectors

- $\alpha_i = 0$: on or beyond $H_{\pm 1}$
- $\alpha_i = C$: "incorrect" points
- $0 < \alpha_i < C$: on $H_{\pm 1}$
- x_i is on $H_{\pm 1}$: what do we know about its α_i?
- x_i is strictly beyond $H_{\pm 1}$: what is α_i?
- x_i is strictly within $H_{\pm 1}$: what is α_i?



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

UNIVERSITY OF WATERLOO

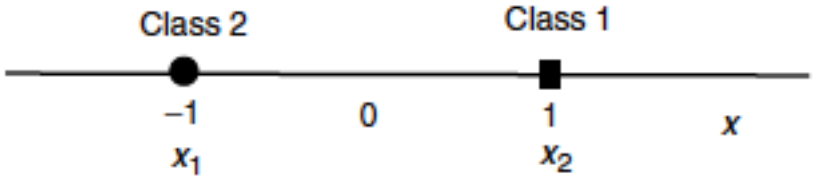
Recover b

- Take any i such that $C > \alpha_i > 0$
- Then $\mathbf{x}_i$ is on the hyperplane:

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

- How to recover ξ ?

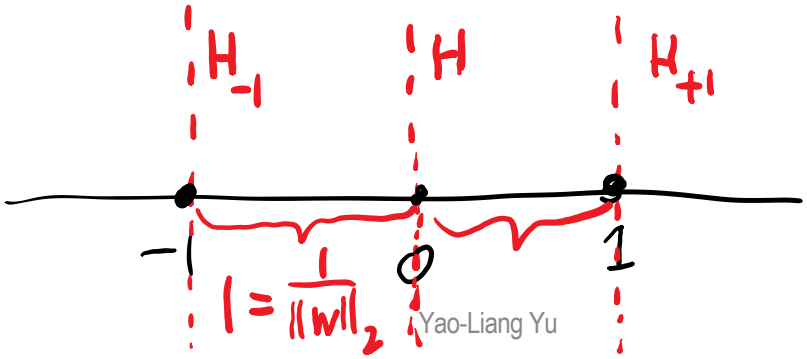
A simple example



$$\min_{C \geq \alpha_1 \geq 0} 2\alpha_1^2 - 2\alpha_1 \Rightarrow \alpha_1 = \alpha_2 = \begin{cases} 1/2, & C \geq 1/2 \\ C, & C \leq 1/2 \end{cases}$$

- C > ½ → $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i = 2\alpha_1 = 1$

  $0 < \alpha_1 = \frac{1}{2} < C$ → x_2 is on H_1 → b=0 → hard-margin

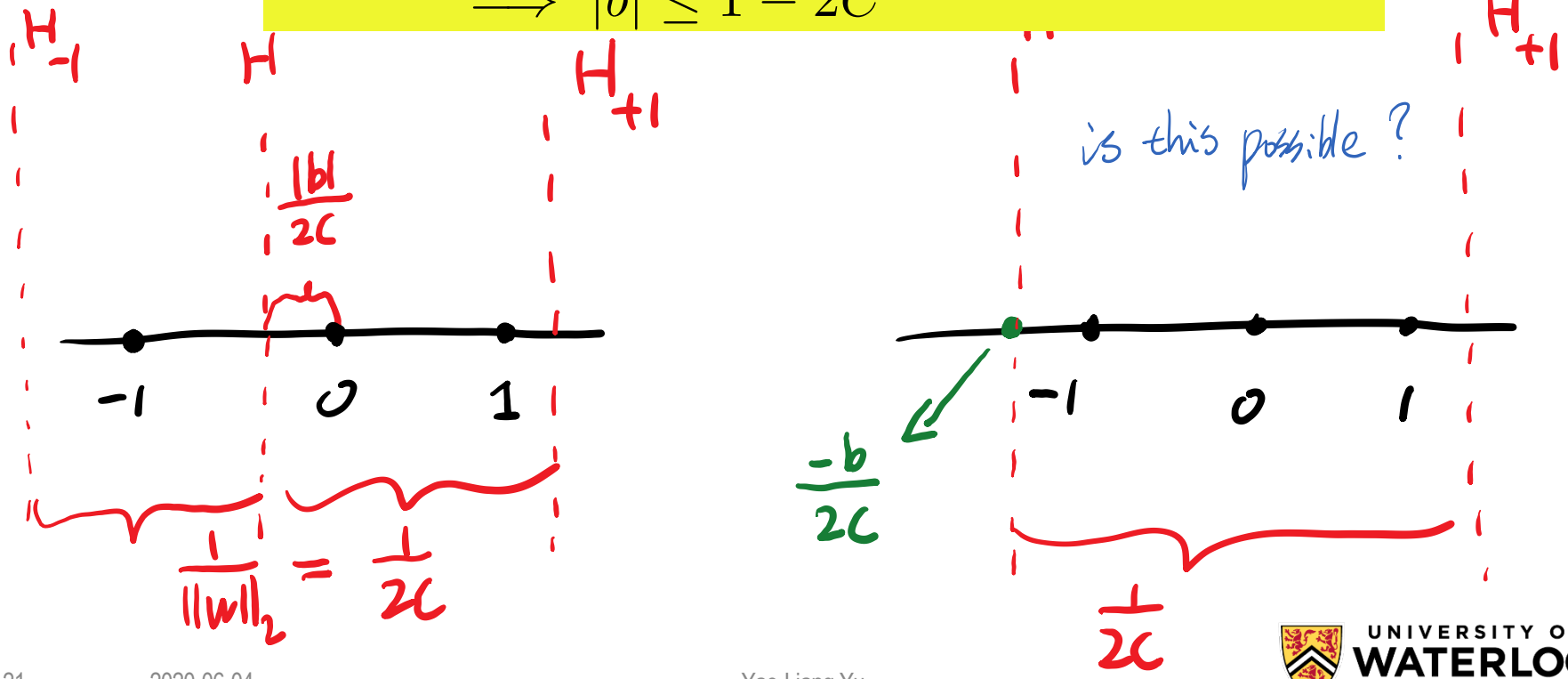


A simple example cont'

- C <= ½ → w=2C,

$$H_t = \{\mathbf{x} : \mathbf{w}^\top \mathbf{x} + b = t\}$$

$$\begin{aligned} \alpha_1 = C &\implies (\mathbf{x}_1, y_1 = -1) \text{ is on or above } H_{-1} \\ &\implies y_i(\mathbf{w}^\top \mathbf{x}_i + b) \le 1 \\ &\implies |b| \le 1 - 2C \end{aligned}$$



UNIVERSITY OF WATERLOO

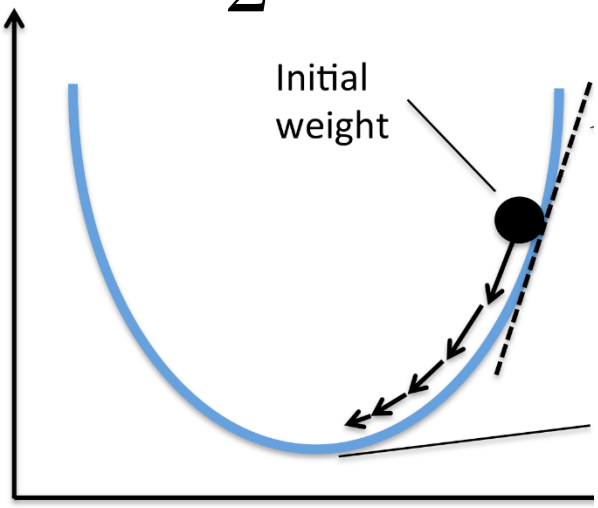
Outline

- Formulation
- Dual
- Optimization
- Extension

Gradient Descent

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

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



Step size (learning rate)

- const., if *L* is smooth
- diminishing, otherwise

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

(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

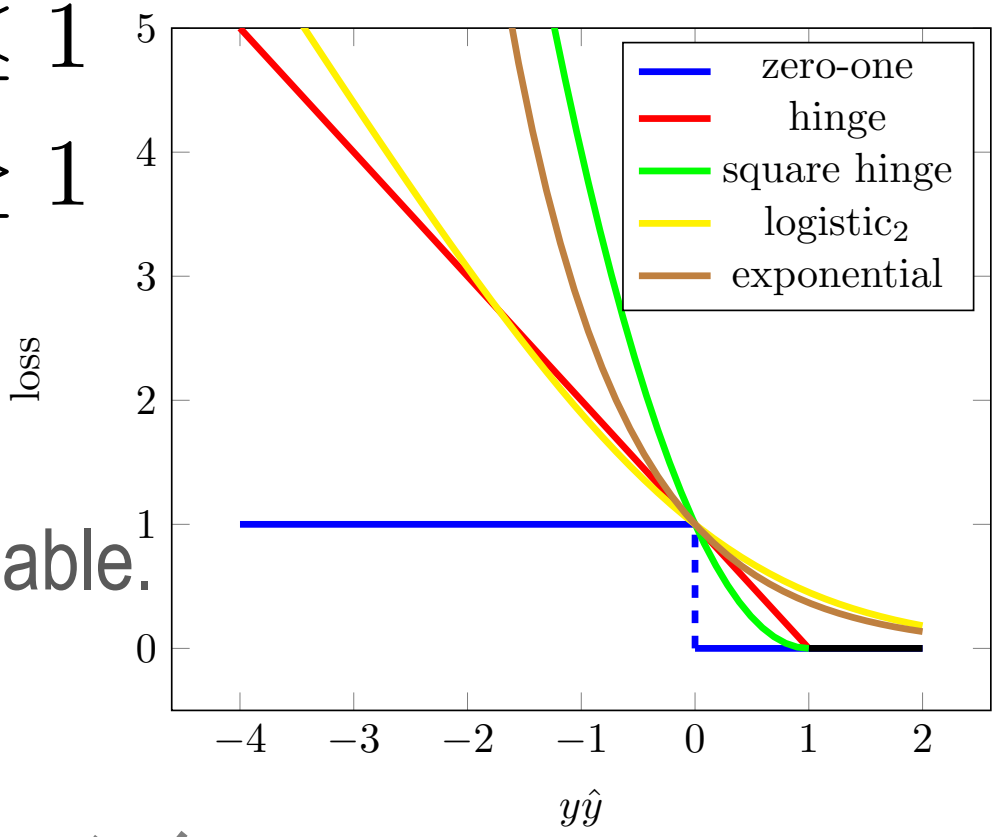
$$\mathbf{w}_{t+1} = (1 - \eta_t)\mathbf{w}_t - \eta_t C \ell'(y_{i_t} \hat{y}_{i_t}) y_{i_t} \mathbf{x}_{i_t}$$

O(d)

- 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

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 differentiable.

What about perceptron?

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

Solving the dual

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

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

$$\boldsymbol{\alpha}_{t+1} \leftarrow \boldsymbol{\alpha}_t - \eta_t (K \odot \mathbf{y}\mathbf{y}^\top) \boldsymbol{\alpha}_t + \eta_t \mathbf{1}$$

$$\boldsymbol{\alpha}_{t+1} \leftarrow \text{prox}(\alpha_{t+1})$$

O(n*n)

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

Outline

- Formulation
- Dual
- Optimization
- Extension

Multiclass (Crammer & Singer'01)

$$\min_{W} \ \frac{1}{2}\|W\|_{\mathsf{F}}^2$$

$$\text{s.t. } \forall i, \forall k \neq y_i, \ \mathbf{x}_i^\top \mathbf{w}_{y_i} \geq 1 + \mathbf{x}_i^\top \mathbf{w}_k$$

separate by a "safety margin"

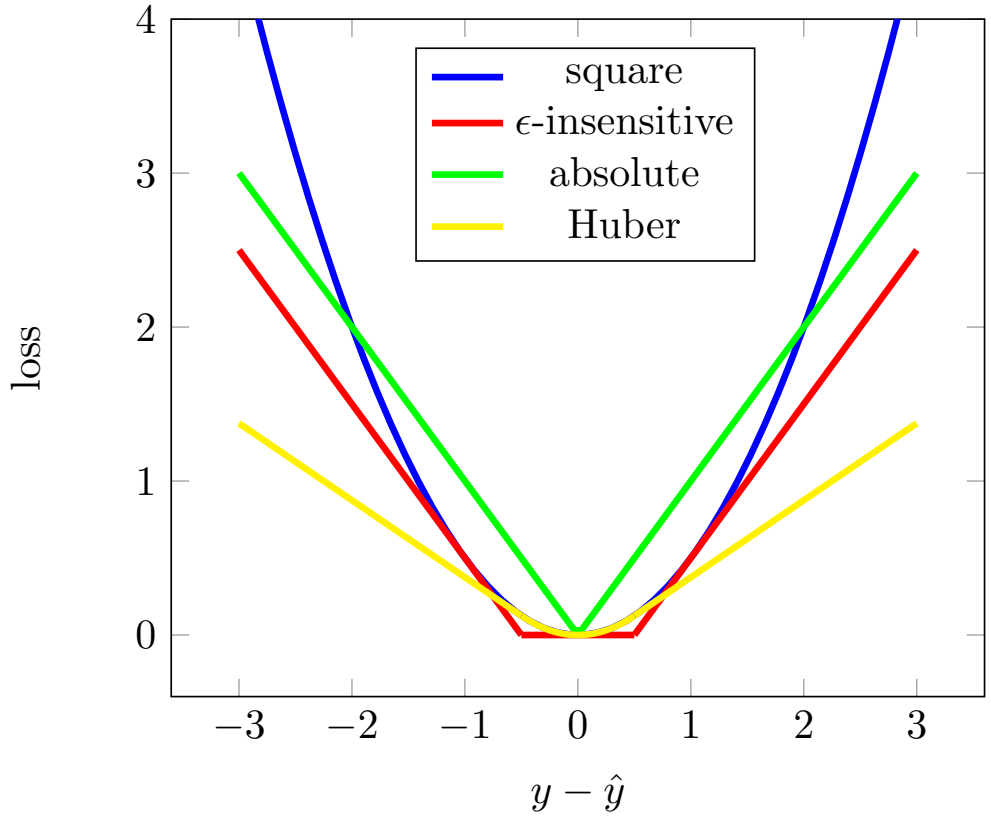
Prediction for correct class

Prediction for wrong classes

- Soft-margin is similar
- Many other variants

Regression (Drucker et al.'97)

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



Questions?