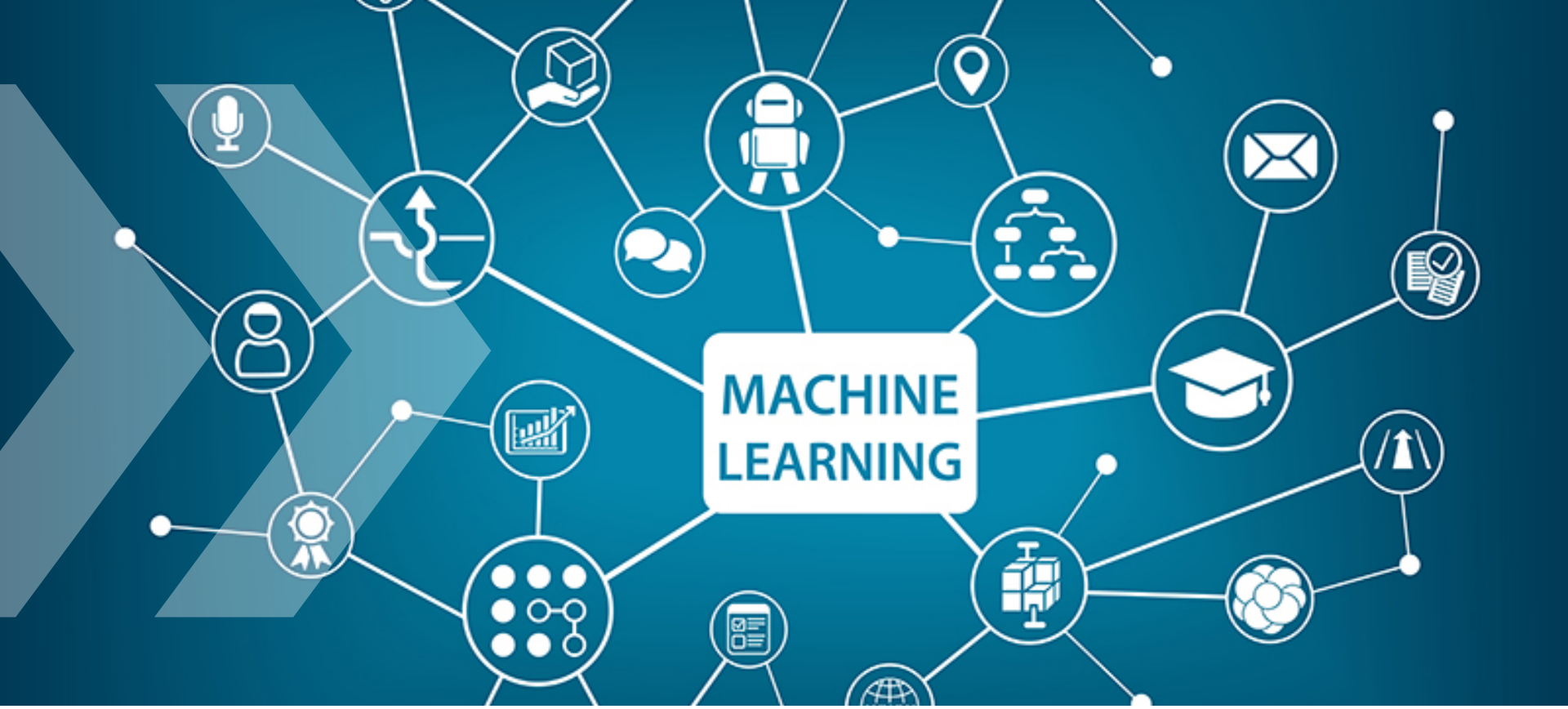



# CS489/698: Intro to ML

## Lecture 07: Soft-margin SVM

UNIVERSITY OF WATERLOO

# 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

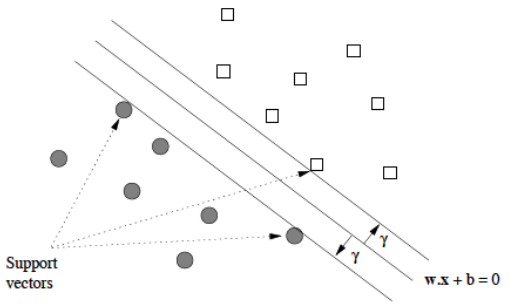
- Formulation
- Dual
- Optimization
- Extension

# Hard-margin SVM

Primal

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

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



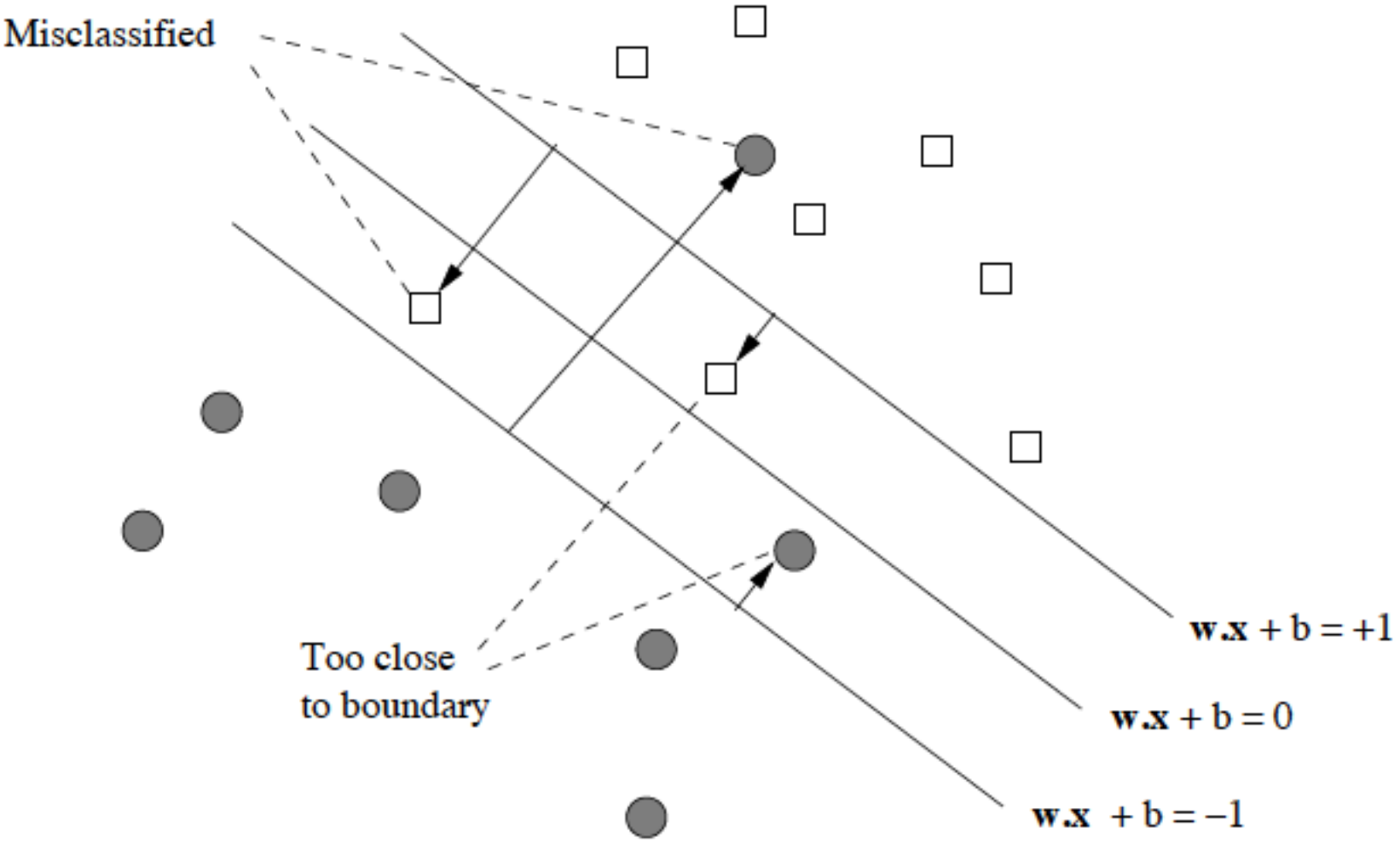

Dual

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

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

hard constraint

# What if inseparable?



# Soft-margin (Cortes & Vapnik'95)

Primal

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

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

hard constraint

propto 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

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

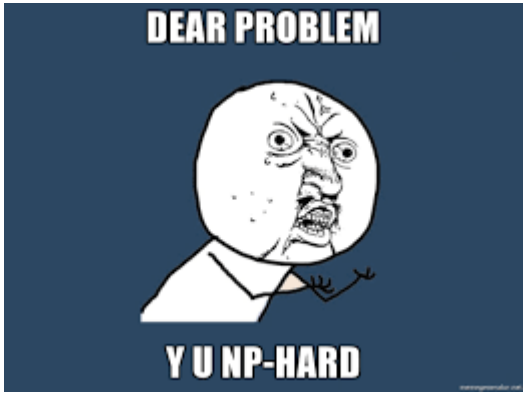
prediction (no sign)

# Zero-one loss

$\hat{Y}$

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

your prediction



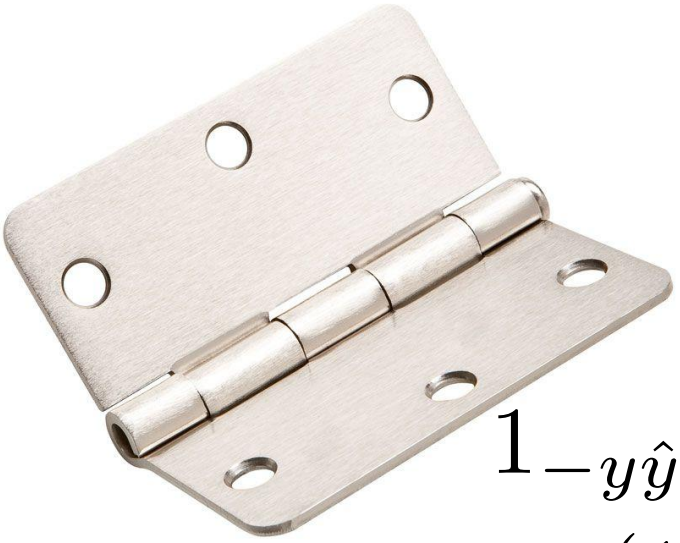

$$\frac{1}{n} \sum_{i=1}^{n} 1_{-Y_i \hat{Y}_i \geq 0}$$

$\hat{Y}_i = w^T X_i + b$

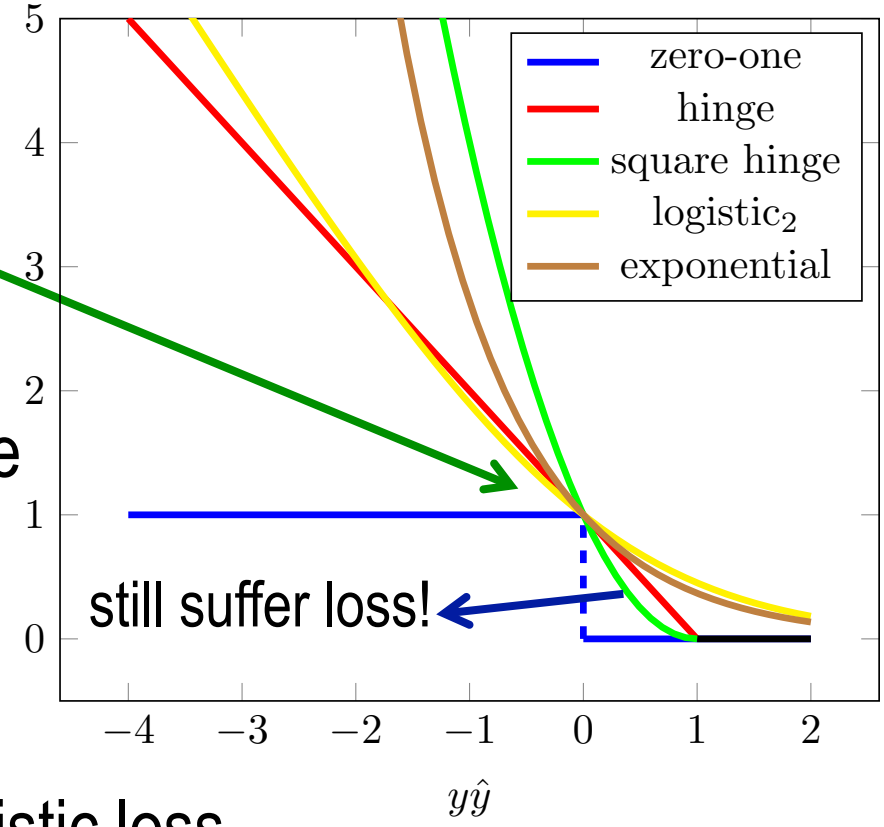
- Find prediction rule *f* so that on an unseen random X, our prediction sign(f(X)) has small chance to be different from the true label Y

# The hinge loss

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



upper bound




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

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

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

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



UNIVERSITY OF WATERLOO

# Classification-calibration

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

Theorem (Bartlett, Jordan, McAuLiffe'06). Any convex margin loss $\ell$ is classification-calibrated iff $\ell$ 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*

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

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


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

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


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

Slack for "wrong" prediction

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

# 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

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

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

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

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

- C → 0? we care about margin
- C → inf? we care about classification

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

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

$\mathbf{R}^n$

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

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

# 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: $\boldsymbol{\alpha} \geq 0 \quad \boldsymbol{\beta} \leq 0$ $\alpha \leq C$
- 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

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

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

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




UNIVERSITY OF WATERLOO

# Support Vectors

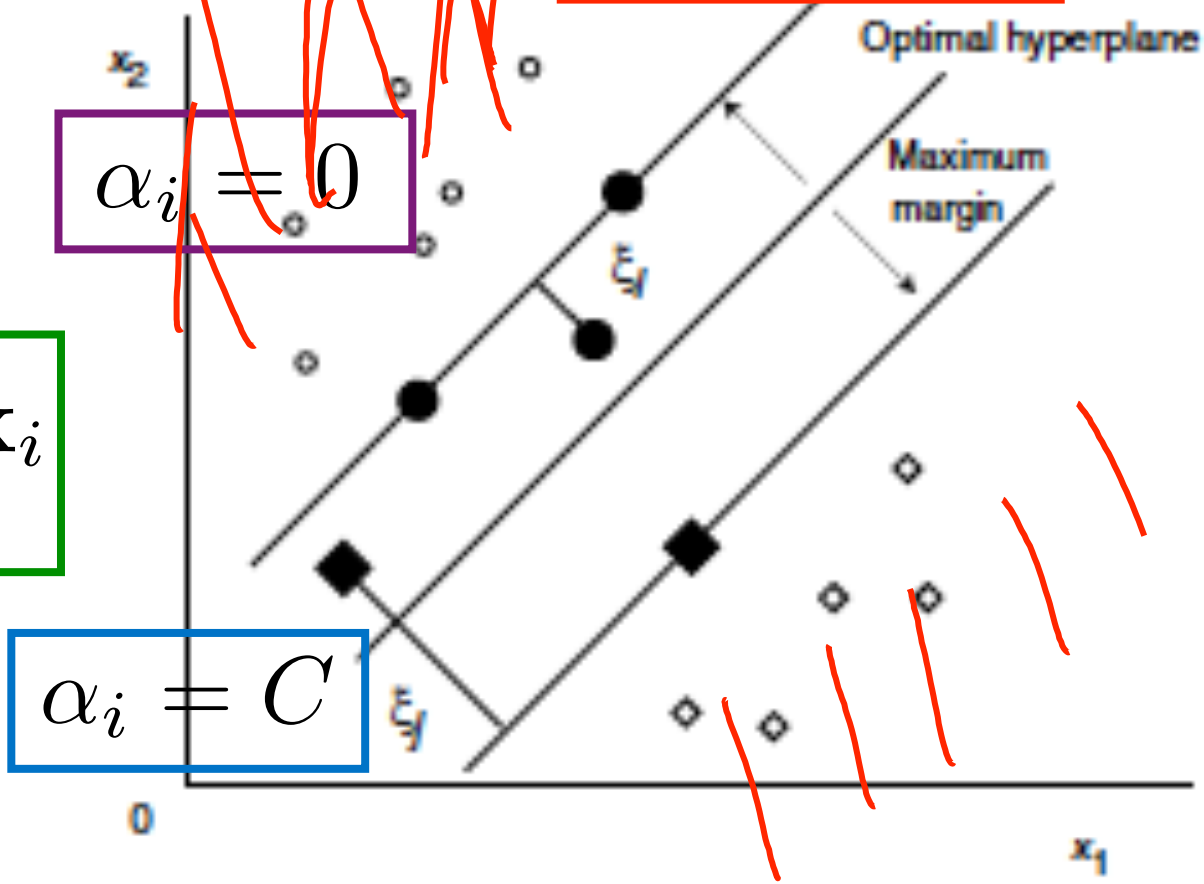
Text

$$0 < \alpha_i < C$$

$$\alpha_i = 0$$

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

$$\alpha_i = C$$



# Recover b

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

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

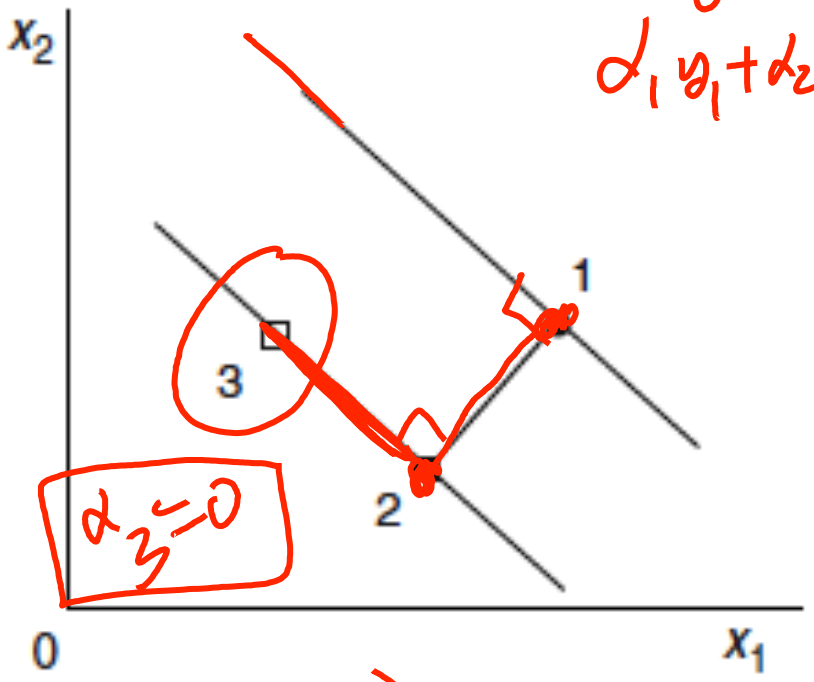
$= w^T x_i + b$

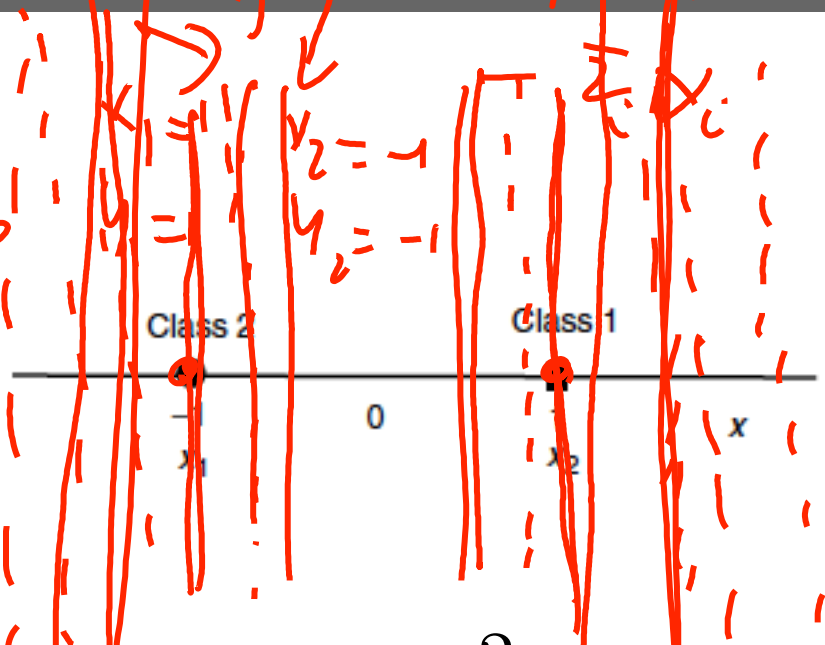
- How to recover $\xi$ ?

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

- What if there is no such i ?

# More examples




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

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

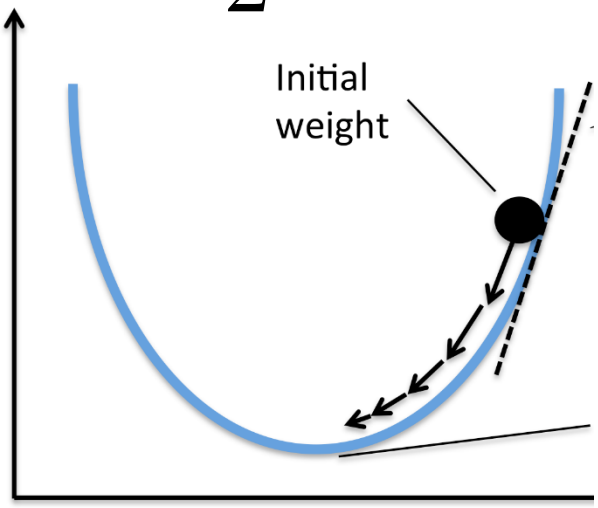
# Outline

- 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

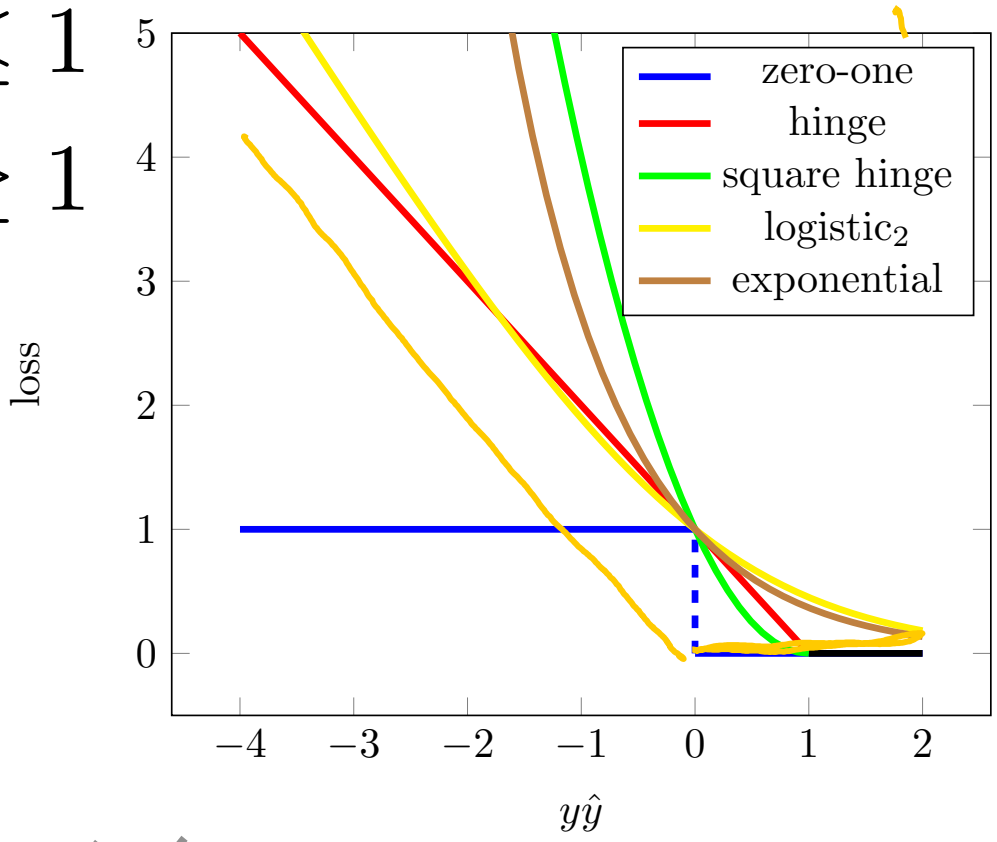
$$\mathbf{w}_{t+1} = (1 - \eta_t)\mathbf{w}_t - \eta_t C \ell'(y_{i_t} \hat{y}_{i_t}) y_{i_t} \mathbf{x}_{i_t}$$

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 diff.

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

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

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

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

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

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

# Outline

- Formulation
- Dual
- Optimization
- Extension

# Multiclass (Crammer & Singer'01)

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

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

separate by a "safety margin"

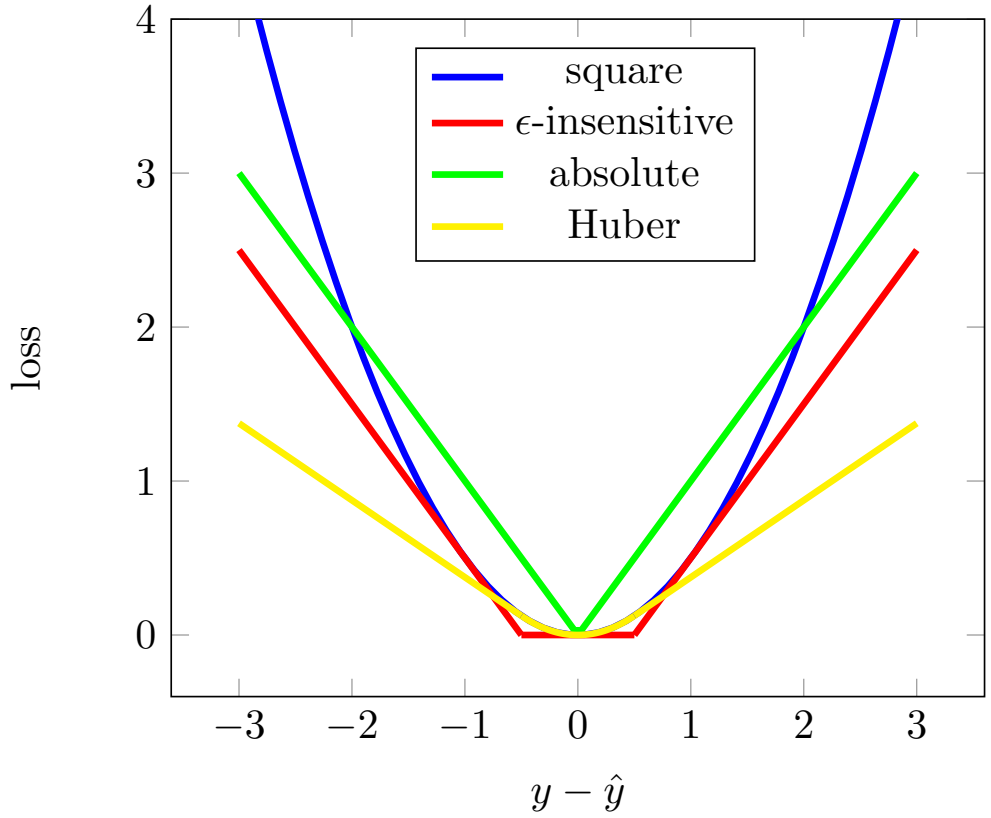
Prediction for correct class

Prediction for wrong classes

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

# Regression (Drucker et al.'97)

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






UNIVERSITY OF WATERLOO

# 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