

# CS480/680: Introduction to Machine Learning

## Lec 05: Soft-margin Support Vector Machines

Yaoliang Yu

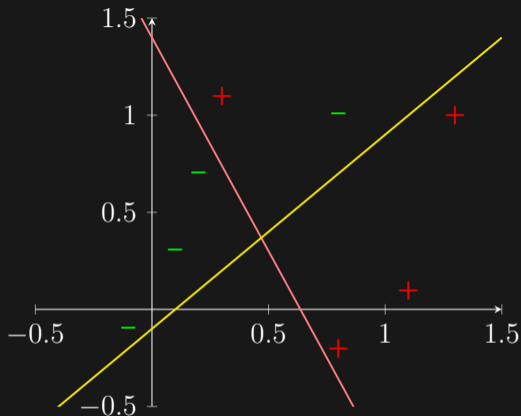


UNIVERSITY OF  
**WATERLOO**

FACULTY OF MATHEMATICS  
DAVID R. CHERITON SCHOOL  
OF COMPUTER SCIENCE

Jan 23, 2025

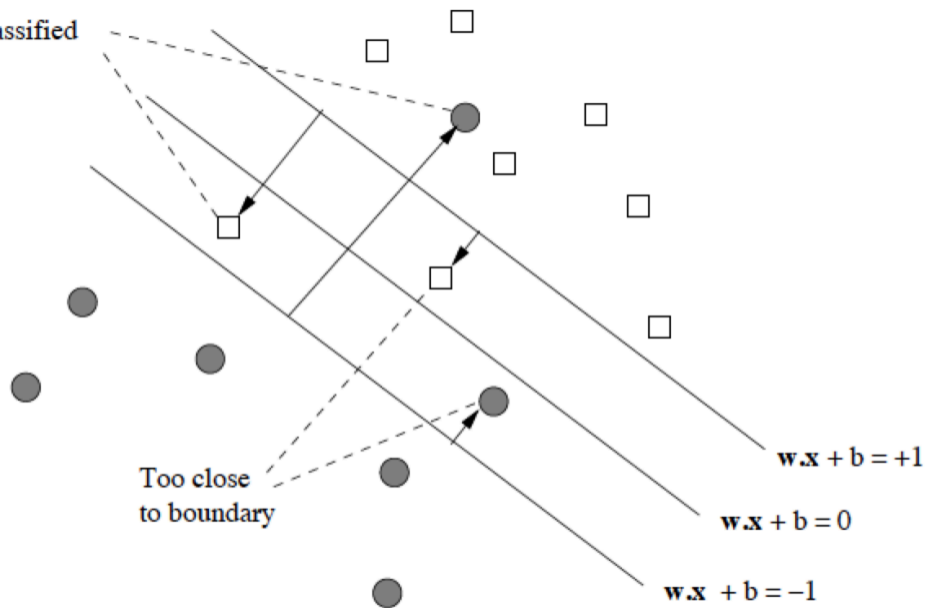
# Beyond Separability



- Balancing between margin maximization and the **soft-margin** loss:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \cdot \sum_i (1 - y_i \hat{y}_i)^+, \quad \text{s.t.} \quad \hat{y}_i := \langle \mathbf{x}_i, \mathbf{w} \rangle + b$$

Misclassified



# Soft-margin SVM

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

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

- Hard constraint: must respect; “live or die”

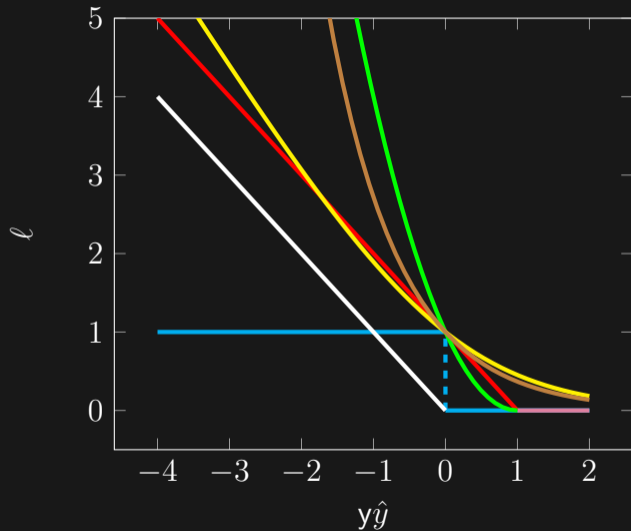
$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \cdot \sum_{i=1}^n (1 - y_i \hat{y}_i)^+$$

$$\text{s.t. } \hat{y}_i = \langle \mathbf{x}_i, \mathbf{w} \rangle + b, \forall i$$

- Soft penalty: the more you deviate the heavier the penalty

- $\frac{1}{2} \|\mathbf{w}\|_2^2$ : margin maximization
- $(1 - y_i \hat{y}_i)^+$ :  $i$ -th training error, 0 if  $y_i \hat{y}_i \geq 1$  and  $1 - y_i \hat{y}_i$  (grow linearly) otherwise
- $C$ : hyper-parameter to control tradeoff

# The Hinge Loss



- zero-one:  $\mathbb{I}[-y\hat{y} \geq 0]$
- hinge:  $(1 - y\hat{y})^+$
- square hinge:  $(1 - y\hat{y})_+^2$
- logistic<sub>2</sub>:  $\log_2(1 + \exp(-y\hat{y}))$
- exponential:  $\exp(-y\hat{y})$
- Perceptron:  $(-y\hat{y})^+$



# Zero-one Loss and Generalization Error

$$\Pr(\hat{Y} \neq Y) = \mathbb{E} \llbracket -Y f(\mathbf{X}) \geq 0 \rrbracket, \quad \text{where } \hat{Y} = \text{sign}(f(\mathbf{X}))$$

- $f : \mathcal{X} \rightarrow \mathbb{R}$  is our real-valued predictor, e.g.,  $f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle$
- Training error after sampling

$$\frac{1}{n} \sum_{i=1}^n \llbracket -Y_i f(\mathbf{X}_i) \geq 0 \rrbracket$$

- Even with linear predictors, minimizing the above training error is NP-hard

---

A. L. Blum and R. L. Rivest. "Training a 3-node neural network is NP-complete". *Neural Networks*, vol. 5, no. 1 (1992), pp. 117–127.  
S. Ben-David, N. Eiron, and P. M. Long. "On the difficulty of approximately maximizing agreements". *Journal of Computer and System Sciences*, vol. 66, no. 3 (2003), pp. 496–514.

# Classification Calibration

- Want to minimize the 0-1 loss, but often end up with minimizing something else
- *Is this sensible?*

## Definition: Bayes rule

Let  $\eta(\mathbf{x}) := \Pr(Y = 1|X = \mathbf{x})$ . The optimal Bayes classifier is  $\text{sign}(2\eta(\mathbf{x}) - 1)$ .

## Definition: Classification calibrated

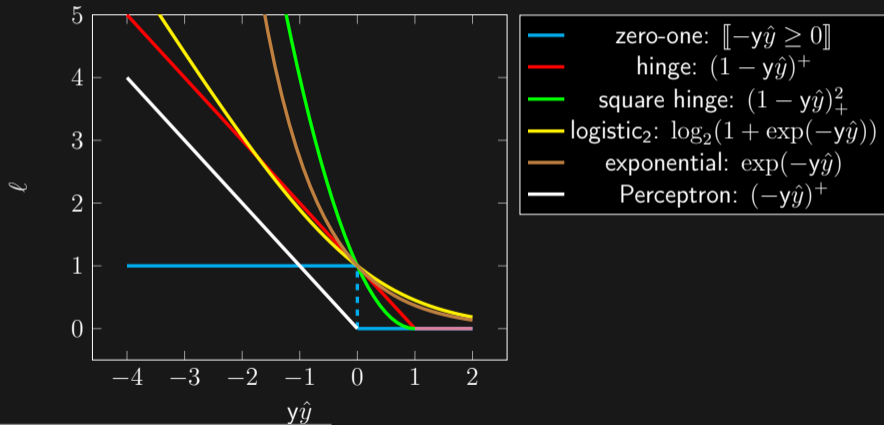
We say a (margin) loss  $\ell(y\hat{y})$  is classification calibrated iff

$$\hat{y} = \hat{y}(\mathbf{x}) := \underset{y \in \mathbb{R}}{\text{argmin}} \eta(\mathbf{x})\ell(y) + [1 - \eta(\mathbf{x})]\ell(-y) \quad \backslash\backslash = \mathbb{E}[\ell(yY)|X = \mathbf{x}]$$

has the same sign as the Bayes rule.

## Theorem: Characterization under convexity

Any **convex** (margin) loss  $\ell$  is classification calibrated iff  $\ell$  is differentiable at 0 and  $\ell'(0) < 0$ .





# A Simpler Way to Derive Lagrangian Dual

$$C \cdot (t)^+ := \max\{Ct, 0\} = \max_{0 \leq \alpha \leq C} \alpha t$$

- Apply above to each term:

$$\min_{\mathbf{w}, b} \max_{0 \leq \alpha \leq C} \left[ \frac{1}{2} \|\mathbf{w}\|_2^2 + \sum_i \alpha_i [1 - y_i (\langle \mathbf{x}_i, \mathbf{w} \rangle + b)] \right]$$

- Swap min with max:

$$\boxed{\max_{0 \leq \alpha \leq C} \min_{\mathbf{w}, b} \left[ \frac{1}{2} \|\mathbf{w}\|_2^2 + \sum_i \alpha_i [1 - y_i (\langle \mathbf{x}_i, \mathbf{w} \rangle + b)] \right]}$$

- Solving the inner unconstrained problem by setting derivative to 0:

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

# Lagrangian Dual Cont'

- Plug in back to eliminate the inner problem (of  $\mathbf{w}$  and  $b$ ):

$$\max_{0 \leq \alpha \leq C} \sum_i \alpha_i - \frac{1}{2} \left\| \sum_i \alpha_i y_i \mathbf{x}_i \right\|_2^2$$

- Changing max to min and expanding the norm:

$$\min_{0 \leq \alpha \leq C} \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \boxed{\langle \mathbf{x}_i, \mathbf{x}_j \rangle} - \sum_i \alpha_i$$

- What happens if  $C \rightarrow \infty$ ?
- What happens if  $C \rightarrow 0$ ?

# Comparison

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 + \sum_{i=1}^n \ell_{1-y_i \hat{y}_i} \leq 0$$

s.t.  $\hat{y}_i = \langle \mathbf{x}_i, \mathbf{w} \rangle + b, \forall i$

$$\min_{\alpha \geq 0} - \sum_i \alpha_i + \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

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

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

s.t.  $\hat{y}_i = \langle \mathbf{x}_i, \mathbf{w} \rangle + b, \forall i$

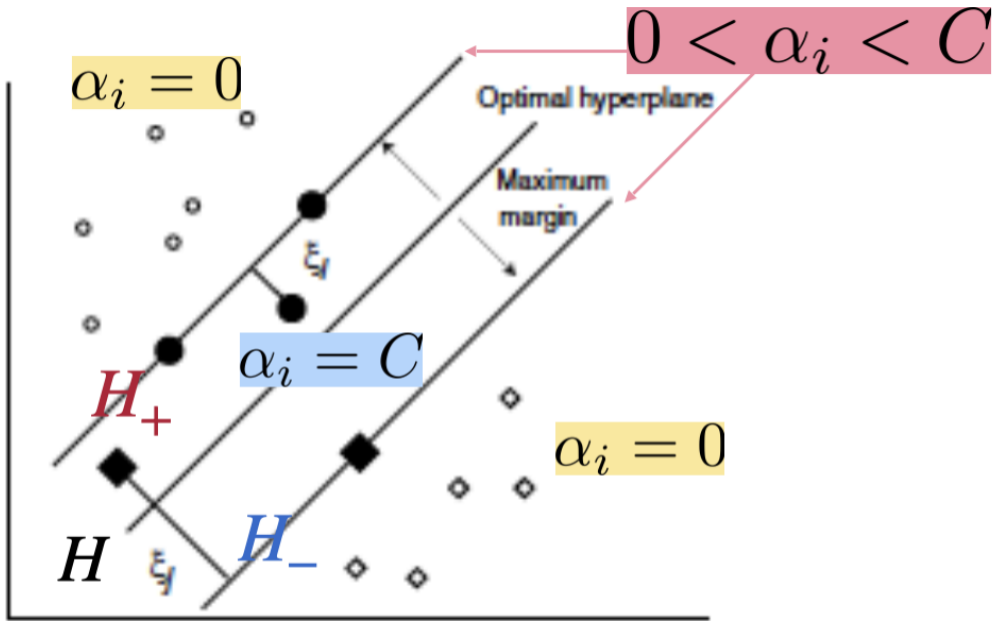
$$\min_{C \geq \alpha \geq 0} - \sum_i \alpha_i + \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

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

# Complementarity Slackness

$$C \cdot (t)^+ := \max\{Ct, 0\} = \max_{0 \leq \alpha \leq C} \alpha t$$

- $t > 0 \implies \alpha = C$  and  $\alpha = C \implies t \geq 0$
- $t < 0 \implies \alpha = 0$  and  $\alpha = 0 \implies t \leq 0$
- Apply to each term in soft-margin SVM:
  - $1 > y_i \hat{y}_i \implies \alpha_i = C$  and  $\alpha_i = C \implies 1 \geq y_i \hat{y}_i$  (wrong side of  $H_{\pm 1}$ , correct/incorrect)
  - $1 < y_i \hat{y}_i \implies \alpha_i = 0$  and  $\alpha_i = 0 \implies 1 \leq y_i \hat{y}_i$  (correctly classified, on/beyond  $H_{\pm 1}$ )
  - $1 = y_i \hat{y}_i \implies 0 \geq \alpha_i \geq C$  and  $0 < \alpha_i < C \implies 1 = y_i \hat{y}_i$  (correctly classified, on  $H_{\pm 1}$ )

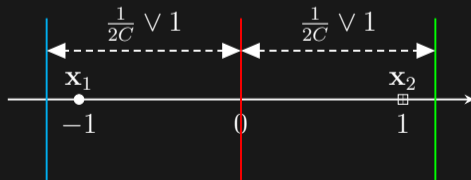


# A Simple Example

$$\min_{w,b} \frac{1}{2}w^2 + C(1-w+b)^+ + C(1-w-b)^+$$

$$\begin{aligned} \min_{C \geq \alpha \geq 0} & \frac{1}{2}(\alpha_1 + \alpha_2)^2 - \alpha_1 - \alpha_2 \\ \text{s.t.} & \alpha_1 - \alpha_2 = 0 \end{aligned}$$

$$\alpha_1 = \alpha_2 = \frac{1}{2} \wedge C, \quad w = 1 \wedge (2C), \quad |b| \leq 1 - w$$



# Recovering $b$

- W.l.o.g., there is always (at least) one data point sitting at one of  $H_{\pm 1}$ 
  - suppose not, move the hyperplanes to the left / right until touching a data point
  - one of the directions must not increase the soft-margin loss
- This point can be used to recover  $b$ :  $y(\langle \mathbf{x}, \mathbf{w} \rangle + b) = 1$ 
  - can average if multiple points are (close to be) on  $H_{\pm 1}$

# A Word About Stochastic Gradient

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

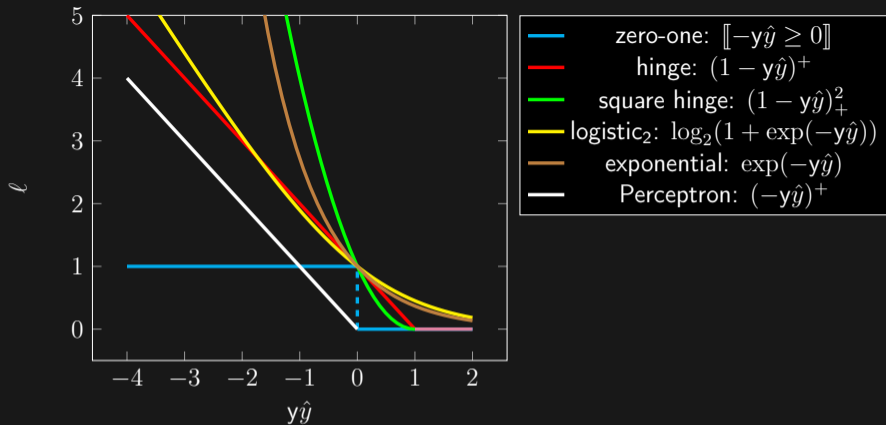
- Gradient descent costs  $O(nd)$ :

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \left[ \frac{1}{n} \sum_{i=1}^n \ell'(y_i \hat{y}_i) y_i \mathbf{x}_i + \frac{\mathbf{w}}{\lambda} \right]$$

- A random sample suffices:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \left[ \frac{1}{n} \sum_{i=1}^n \ell'(y_i \hat{y}_i) y_i \mathbf{x}_i + \frac{\mathbf{w}}{\lambda} \right]$$





- $$\ell'_{\text{hinge}}(t) = \begin{cases} -1, & t < 1 \\ 0, & t > 1 \\ [-1, 0], & t = 1 \end{cases}$$
 while we choose 
$$\ell'_{\text{Perceptron}}(t) = \begin{cases} -1, & t \leq 0 \\ 0, & t > 0 \end{cases}$$

- What about the zero-one loss? Other losses?

# Multi-class

$$\forall i, \hat{\mathbf{y}}_i = W\mathbf{x}_i + \mathbf{b} \in \mathbb{R}^c,$$

$$\min_{W, \mathbf{b}} \frac{1}{2} \|W\|_F^2$$

$$\text{s.t. } \hat{y}_{y_i, i} \geq \mathbb{I}[k \neq y_i] + \hat{y}_{k, i}, \quad \forall i, \forall k = 1, \dots, c$$

$$\min_{W, \mathbf{b}} \frac{1}{2} \|W\|_F^2 + C \sum_{i=1}^n \max_{k=1, \dots, c} \{ \mathbb{I}[k \neq y_i] + \hat{y}_{k, i} - \hat{y}_{y_i, i} \}$$

# Regression

$$\min_W \frac{1}{2} \|W\|_F^2 + C \sum_{i=1}^n (\|y - \hat{y}_i\| - \epsilon)^+$$

