

# Support Vector Machines

Machine Learning Course - CS-433

Nov 1, 2022

Nicolas Flammarion

**EPFL**

# Vapnik's invention

## A Training Algorithm for Optimal Margin Classifiers

**Bernhard E. Boser\***  
EECS Department  
University of California  
Berkeley, CA 94720  
boser@eecs.berkeley.edu

**Isabelle M. Guyon**  
AT&T Bell Laboratories  
50 Fremont Street, 6th Floor  
San Francisco, CA 94105  
isabelle@neural.att.com

**Vladimir N. Vapnik**  
AT&T Bell Laboratories  
Crawford Corner Road  
Holmdel, NJ 07733  
vlad@neural.att.com

### Abstract

A training algorithm that maximizes the margin between the training patterns and the decision boundary is presented. The technique is applicable to a wide variety of classification functions, including Perceptrons, polynomials, and Radial Basis Functions. The effective number of parameters is adjusted automatically to match the complexity of the problem. The solution is expressed as a linear combination of supporting patterns. These are the subset of training patterns that are closest to the decision boundary. Bounds on the generalization performance based on the leave-one-out method and the VC-dimension are given. Experimental results on optical character recognition problems demonstrate the good generalization obtained when compared with other COLT'92-7/92/PA,USA  
© 1992 ACM 0-89791-498-8/92/0007/0144...\$1.50

Machine Learning, 20, 273–297 (1995)

© 1995 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

## Support-Vector Networks

**CORINNA CORTES**  
**VLADIMIR VAPNIK**  
AT&T Bell Labs., Holmdel, NJ 07733, USA

corinna@neural.att.com  
vlad@neural.att.com

**Editor:** Lorenza Saitta

**Abstract.** The *support-vector network* is a new learning machine for two-group classification problems. The machine conceptually implements the following idea: input vectors are non-linearly mapped to a very high-dimension feature space. In this feature space a linear decision surface is constructed. Special properties of the decision surface ensures high generalization ability of the learning machine. The idea behind the support-vector network was previously implemented for the restricted case where the training data can be separated without errors. We here extend this result to non-separable training data.

High generalization ability of support-vector networks utilizing polynomial input transformations is demonstrated. We also compare the performance of the support-vector network to various classical learning algorithms that all took part in a benchmark study of Optical Character Recognition.



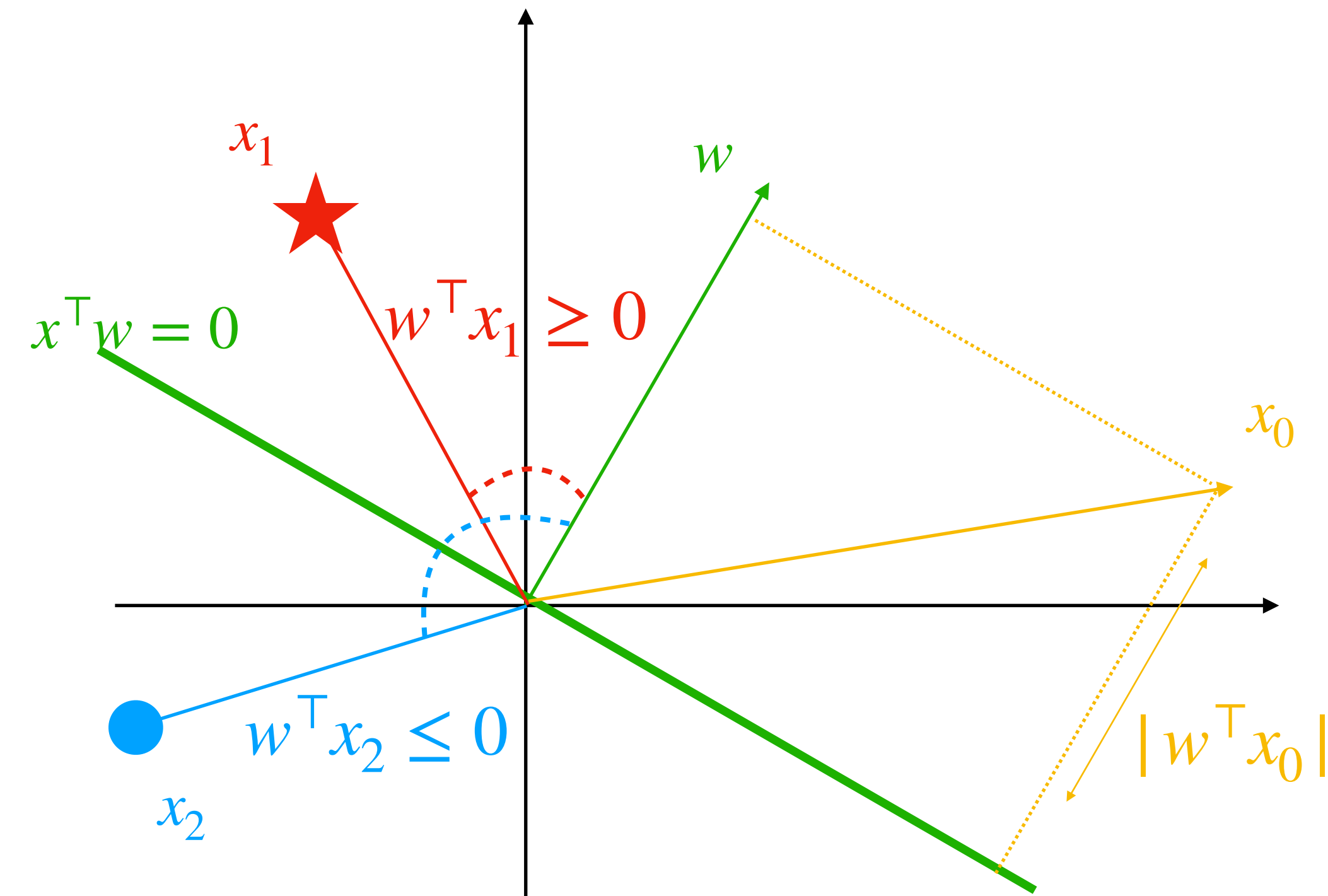
# Linear Classifier

Define a hyperplane by  $\{x : w^\top x = 0\}$   
where  $\|w\| = 1$

Prediction:

$$g(x) = \text{sign}(x^\top w)$$

Claim: The distance between a point  $x_0$  and the hyperplane defined by  $w$  is  $|w^\top x_0|$



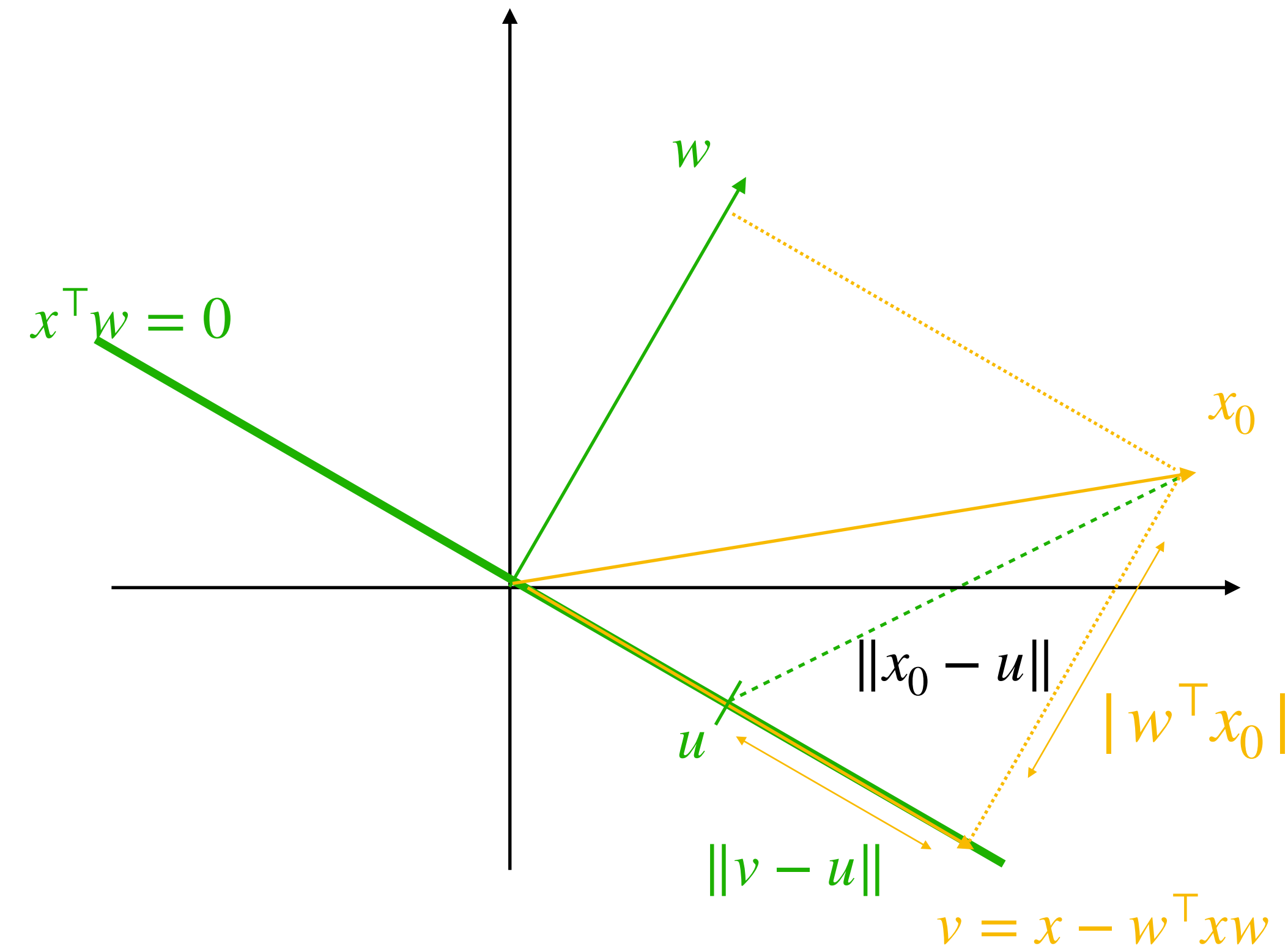
# Linear Classifier

Proof: distance between  $x_0$  and the hyperplane is defined by  $\min_{u:w^\top u=0} \|x_0 - u\|$

Let  $v = x_0 - w^\top x_0 w$  then by the Pythagorean theorem for any  $u$  s.t.  $w^\top u = 0$

$$\|x_0 - u\|^2 = (w^\top x_0)^2 + \|v - u\|^2 \geq (w^\top x_0)^2$$

Claim: The distance between a point  $x_0$  and the hyperplane defined by  $w$  is  $|w^\top x_0|$





# Hard-SVM rule: max-margin separating hyperplane

First assume the dataset  $(x_n, y_n)_{n=1}^N$  is linearly separable

Margin of a hyperplane:  $\min_{n \leq N} |w^\top x_n|$

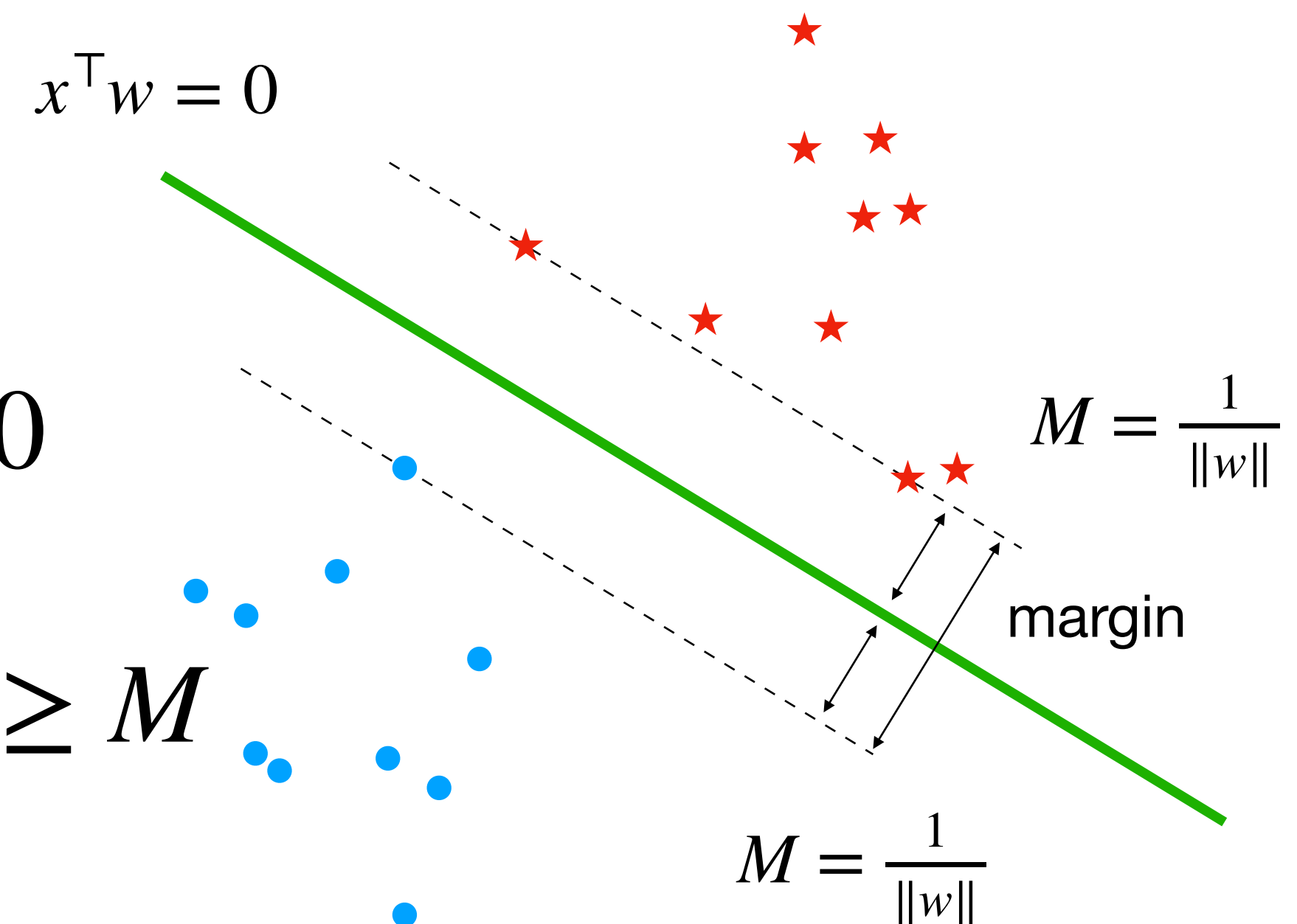
Max-margin separating hyperplane:

$$\max_{w, \|w\|=1} \min_{n \leq N} |w^\top x_n| \text{ such that } \forall n, y_n x_n^\top w \geq 0$$

Equivalent to  $\max_{M \in \mathbb{R}, w, \|w\|=1} M$  such that  $\forall n, y_n x_n^\top w \geq M$

also equivalent to:

$$\min_w \frac{1}{2} \|w\|^2 \text{ such that } \forall n, y_n x_n^\top w \geq 1$$



# Proof of the equivalent formulations

Claim: The following optimization problems are equivalent

$$\begin{aligned} & \max_{w, \|w\|=1} \min_{n \leq N} |w^\top x_n| \\ & \text{s.t. } \forall n, y_n x_n^\top w \geq 0 \end{aligned} \quad (\text{I})$$

$$\begin{aligned} & \max_{M \in \mathbb{R}, w, \|w\|=1} M \\ & \text{s.t. } \forall n, y_n x_n^\top w \geq M \end{aligned} \quad (\text{II})$$

Proof: let  $w_1$  be a solution of (I) and  $M_1 = \min_{n \leq N} |w_1^\top x_n|$  and let  $w_2$  and  $M_2$  be solutions of (II)

- $(w_1, M_1)$  is admissible for (II) so  $M_1 \leq M_2$
- $w_2$  is admissible for (I) so  $\min_{n \leq N} |w_2^\top x_n| \leq \min_{n \leq N} |w_1^\top x_n|$
- $\forall n, y_n x_n^\top w_2 \geq M_2$  implies that  $\forall n, |x_n^\top w_2| \geq M_2$  and  $\min_{n \leq N} |x_n^\top w_2| \geq M_2$

Therefore  $M_1 = \min_{n \leq N} |w_1^\top x_n| \geq \min_{n \leq N} |w_2^\top x_n| \geq M_2 \geq M_1$

And the two problems are equivalent

# Proof of the equivalent formulations

Claim: The following optimization problems are equivalent

$$\begin{array}{ll} \max_{M \in \mathbb{R}, w, \|w\|=1} M & \min_w \frac{1}{2} \|w\|^2 \\ \text{s.t. } \forall n, y_n x_n^\top w \geq M & \text{s.t. } \forall n, y_n x_n^\top w \geq 1 \end{array} \quad \begin{array}{l} \text{(II)} \\ \text{(III)} \end{array}$$

Proof:

$$\max_{M \in \mathbb{R}, w, \|w\|=1} M \text{ such that } \forall n, y_n x_n^\top w \geq M$$

$$\iff \max_{M \in \mathbb{R}, w} M \text{ such that } \forall n, y_n x_n^\top \frac{w}{\|w\|} \geq M$$

The constraints are independent of the scale of  $w$ . Set  $\|w\| = 1/M$ :

$$\iff \max_w 1/\|w\| \text{ such that } \forall n, y_n x_n^\top w \geq 1$$

$$\iff \min_w \frac{1}{2} \|w\|^2 \text{ such that } \forall n, y_n x_n^\top w \geq 1$$

# Proof of the equivalent formulations

Claim: The following optimization problems are equivalent

$$\begin{aligned} \max_{M \in \mathbb{R}, w, \|w\|=1} \quad & M \\ \text{s.t.} \quad & \forall n, y_n x_n^\top w \geq M \end{aligned} \quad (\text{II})$$

$$\begin{aligned} \min_w \quad & \frac{1}{2} \|w\|^2 \\ \text{s.t.} \quad & \forall n, y_n x_n^\top w \geq 1 \end{aligned} \quad (\text{III})$$

Proof bis: Let  $w_2$  and  $M_2$  be solutions of (II) and  $w_3$  a solution of (III)

- $w_3/\|w_3\|, 1/\|w_3\|$  is admissible for (II) thus  $M_2 \geq 1/\|w_3\|$
- $w_2/M_2$  is admissible for (III) thus  $\|w_3\| \leq \|w_2/M_2\| = 1/M_2$

Thus  $M_2 = 1/\|w_3\|$  and

- $w_3/\|w_3\|, 1/\|w_3\|$  is a solution of (II)
- $w_2/M_2$  is a solution of (I)



# Soft SVM: relaxation of the Hard-SVM rule that can be applied even if the training set is not linearly separable

Idea: still maximize the margin, but allow some of the constraints to be violated

How: by introducing positive slack variables  $\xi_1, \dots, \xi_N$  and replacing the constraints by  $y_n x_n^\top w \geq 1 - \xi_n$

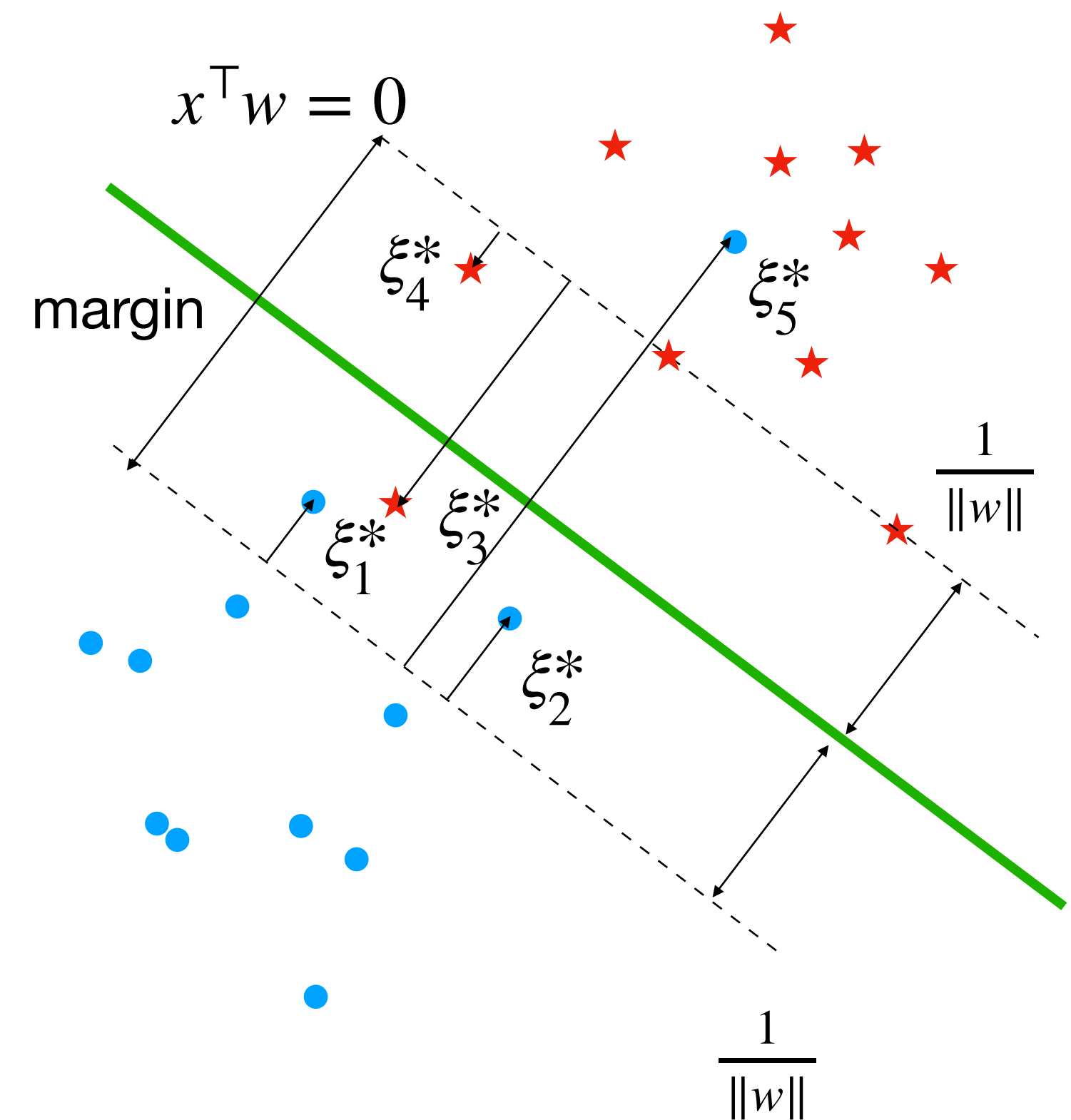
Soft SVM:

$$\begin{aligned} \min_{w, \xi} \quad & \frac{\lambda}{2} \|w\|^2 + \frac{1}{N} \sum_{n=1}^N \xi_n \\ \text{s.t. } \quad & \forall n, y_n x_n^\top w \geq 1 - \xi_n \quad \text{and} \quad \xi_n \geq 0 \end{aligned}$$

which is equivalent to

$$\min_w \frac{\lambda}{2} \|w\|^2 + \frac{1}{N} \sum_{n=1}^N [1 - y_n x_n^\top w]_+$$

$[\alpha]_+ = \max\{0, \alpha\}$



# Soft SVM: relaxation of the Hard-SVM rule that can be applied even if the training set is not linearly separable

Proof: Fix  $w$  and consider the minimization over  $\xi$ :

- If  $y_n x_n^\top w \geq 1$ , then  $\xi_n = 0$
- If  $y_n x_n^\top w < 1$ ,  $\xi_n = 1 - y_n x_n^\top w$

Therefore  $\xi_n = [1 - y_n x_n^\top w]_+$

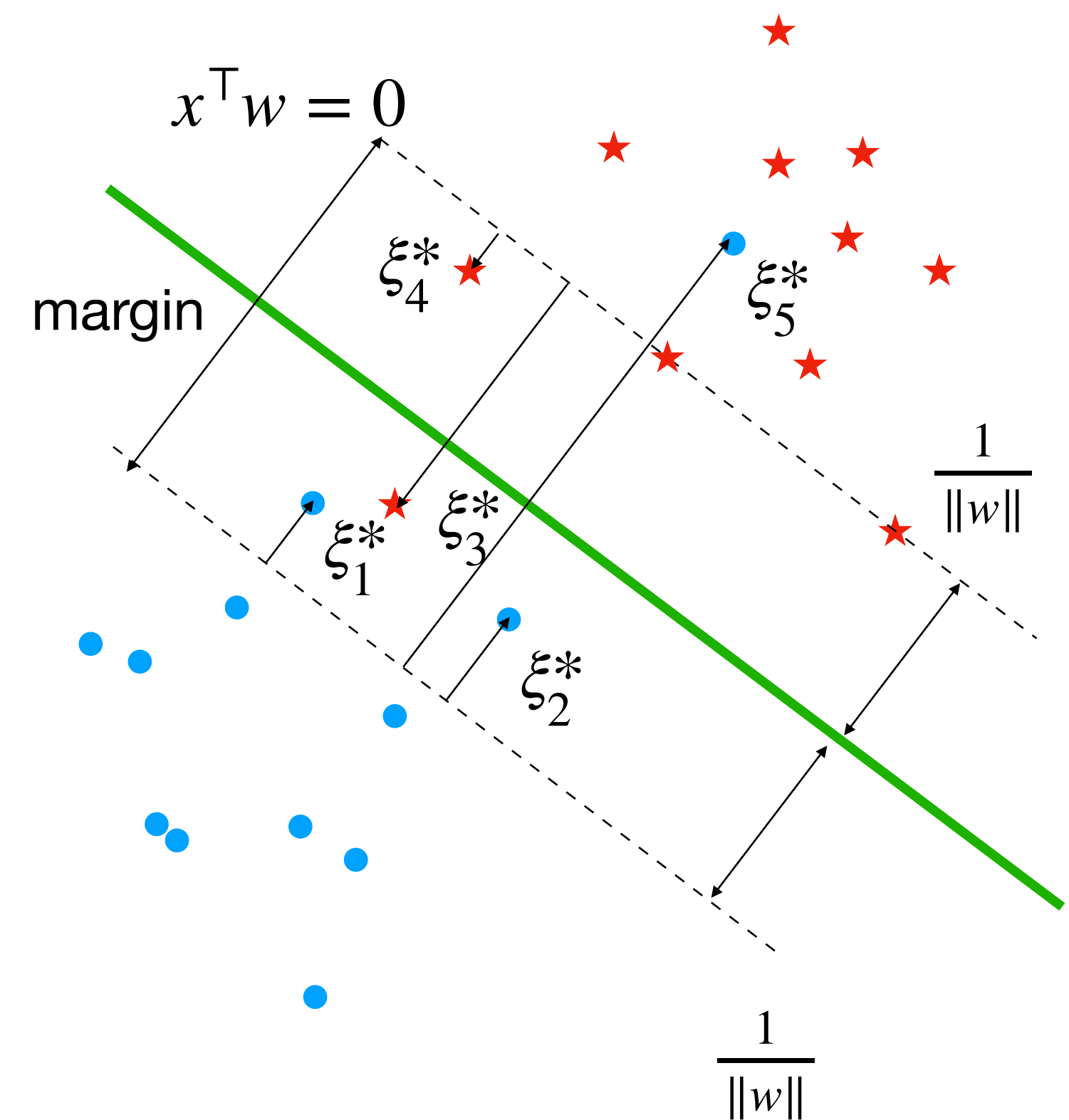
and

$$\begin{aligned} \min_{w, \xi} \quad & \frac{\lambda}{2} \|w\|^2 + \frac{1}{N} \sum_{n=1}^N \xi_n \\ \text{s.t. } \quad & \forall n, y_n x_n^\top w \geq 1 - \xi_n \quad \text{and} \quad \xi_n \geq 0 \end{aligned}$$

which is equivalent to

$$\min_w \frac{\lambda}{2} \|w\|^2 + \frac{1}{N} \sum_{n=1}^N [1 - y_n x_n^\top w]_+$$

$[\alpha]_+ = \max\{0, \alpha\}$



# Classification by risk minimization

Setting:  $(X, Y) \sim \mathcal{D}$  with ranges  $\mathcal{X}$  and  $\mathcal{Y} = \{-1, 1\}$

Goal: Predict with a classifier  $g : \mathcal{X} \rightarrow \mathcal{Y}$  with as low as possible true risk

$$L(g) = \mathbb{P}_{\mathcal{D}}(Y \neq g(X))$$

How: empirical risk minimization (ERM):

$$\min_{g: \mathcal{X} \rightarrow \mathcal{Y}} L_{\text{train}}(g) := \frac{1}{N} \sum_{n=1}^N 1_{g(x_n) \neq y_n} = \frac{1}{N} \sum_{n=1}^N 1_{-y_n g(x_n) \geq 0}$$

Problem:  $L_{\text{train}}$  is not convex:

1. The set of classifiers is not convex because  $\mathcal{Y}$  is discrete
2. The indicator function 1 is not convex because it is not continuous

# Convex relaxation of the classification risk

1. Consider the set of linear predictors  $w^\top x$  and then predict with  $g(x) = \text{sign}(w^\top x)$

$$1_{-yx^\top w > 0} \leq 1_{g(x) \neq y} \leq 1_{-yx^\top w \geq 0} \implies$$

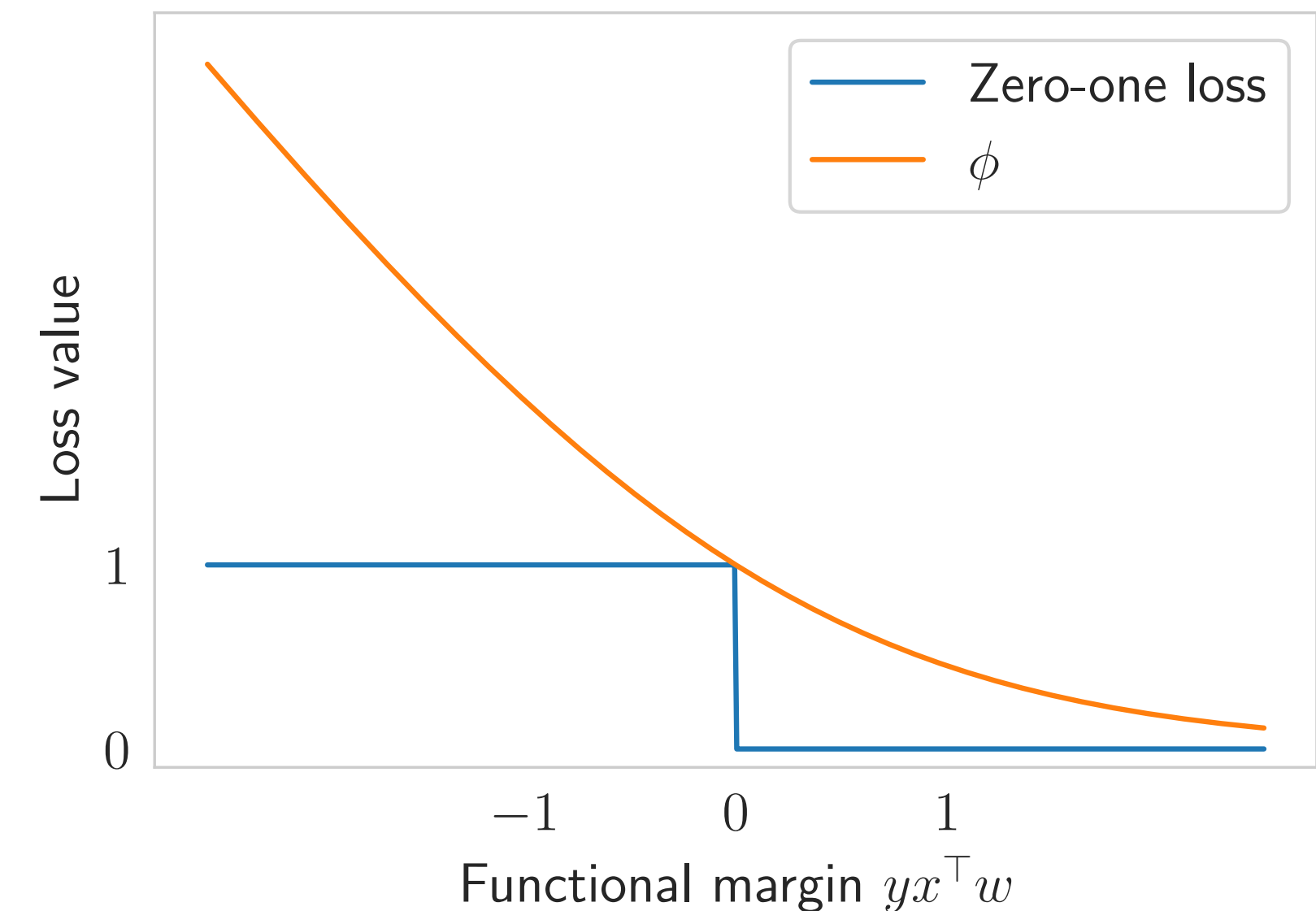
$$\min_w \frac{1}{N} \sum_{n=1}^N 1_{-y_n x_n^\top w \geq 0}$$

2. Replace the indicator function by a convex surrogate  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  and minimize

$$\min_w \frac{1}{N} \sum_{n=1}^N \phi(y_n x_n^\top w)$$

$\phi$  is a function of the functional margin  $y_n x_n^\top w$

Remark: possible to bound the zero-one risk  $L(g)$  by the  $\phi$  risk \*



\* Under technical assumptions on the function  $\phi$

# Losses for Classification

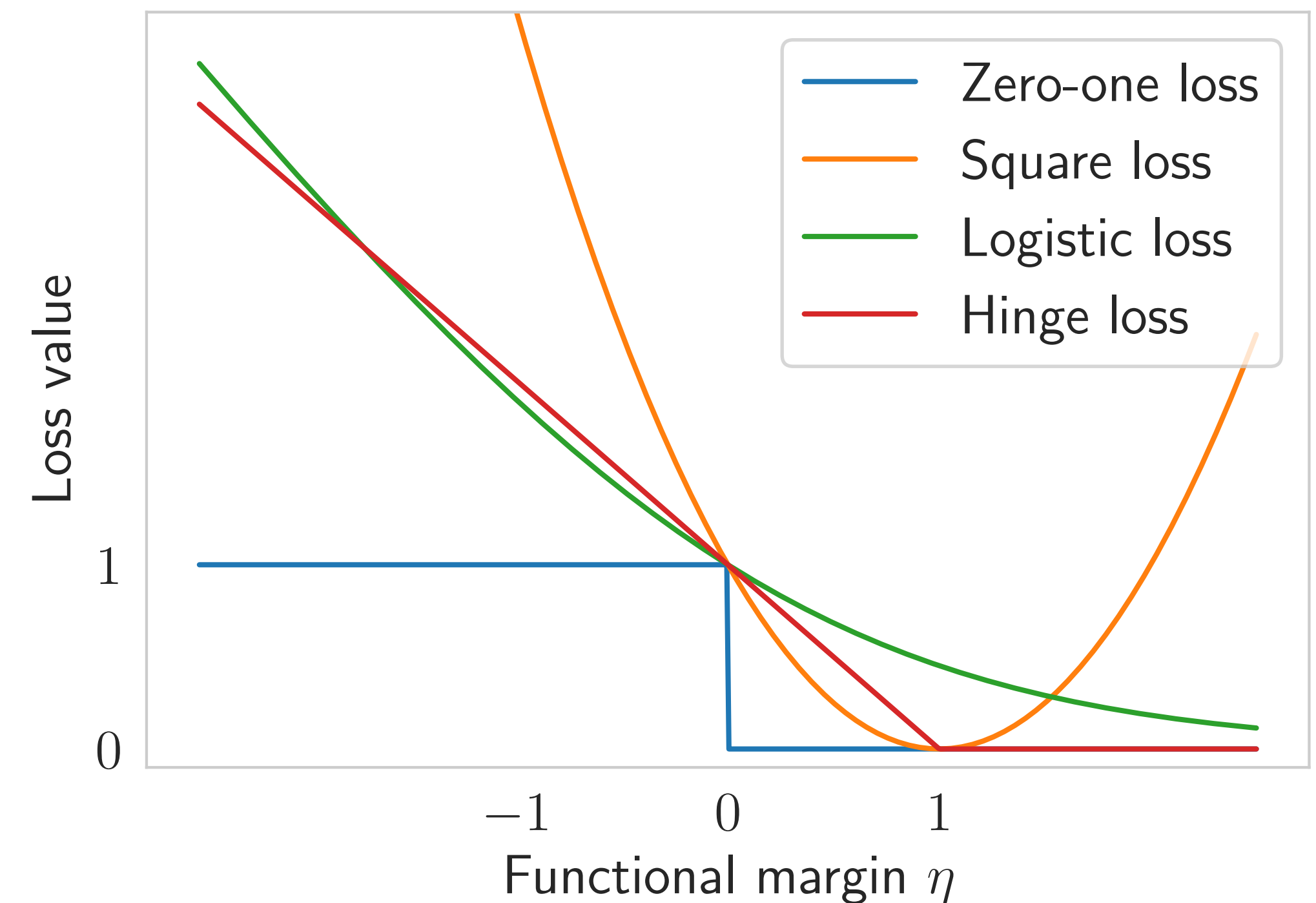
Examples of margin based losses ( $\eta = yx^\top w$ ):

- Quadratic loss:  $\text{MSE}(\eta) = (1 - \eta)^2$
- Logistic loss:  $\text{Logistic}(\eta) = \frac{\log(1 + \exp(-\eta))}{\log(2)}$
- Hinge loss:  $\text{Hinge}(\eta) = [1 - \eta]_+$

Common features: they are convex and upper bound the zero-one loss

Behavior differences:

- MSE punishes any deviation from 1
- The logistic cost is asymmetric – we always incur a cost
- Hinge loss: we incur a cost if the prediction is incorrect or not confident enough





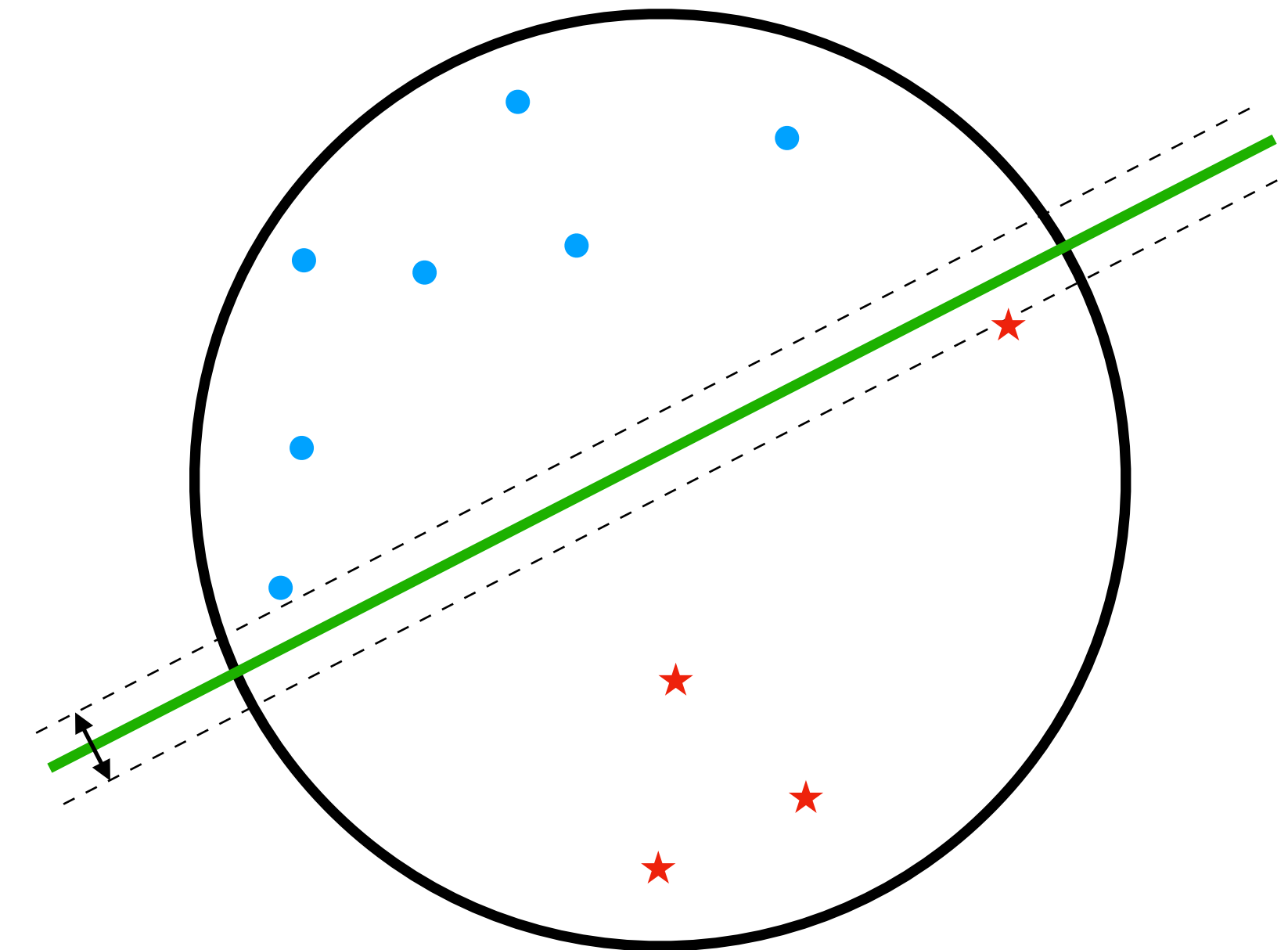
# Summary

$$\min_w \frac{\lambda}{2} \|w\|^2 + \frac{1}{N} \sum_{n=1}^N [1 - y_n x_n^\top w]_+$$

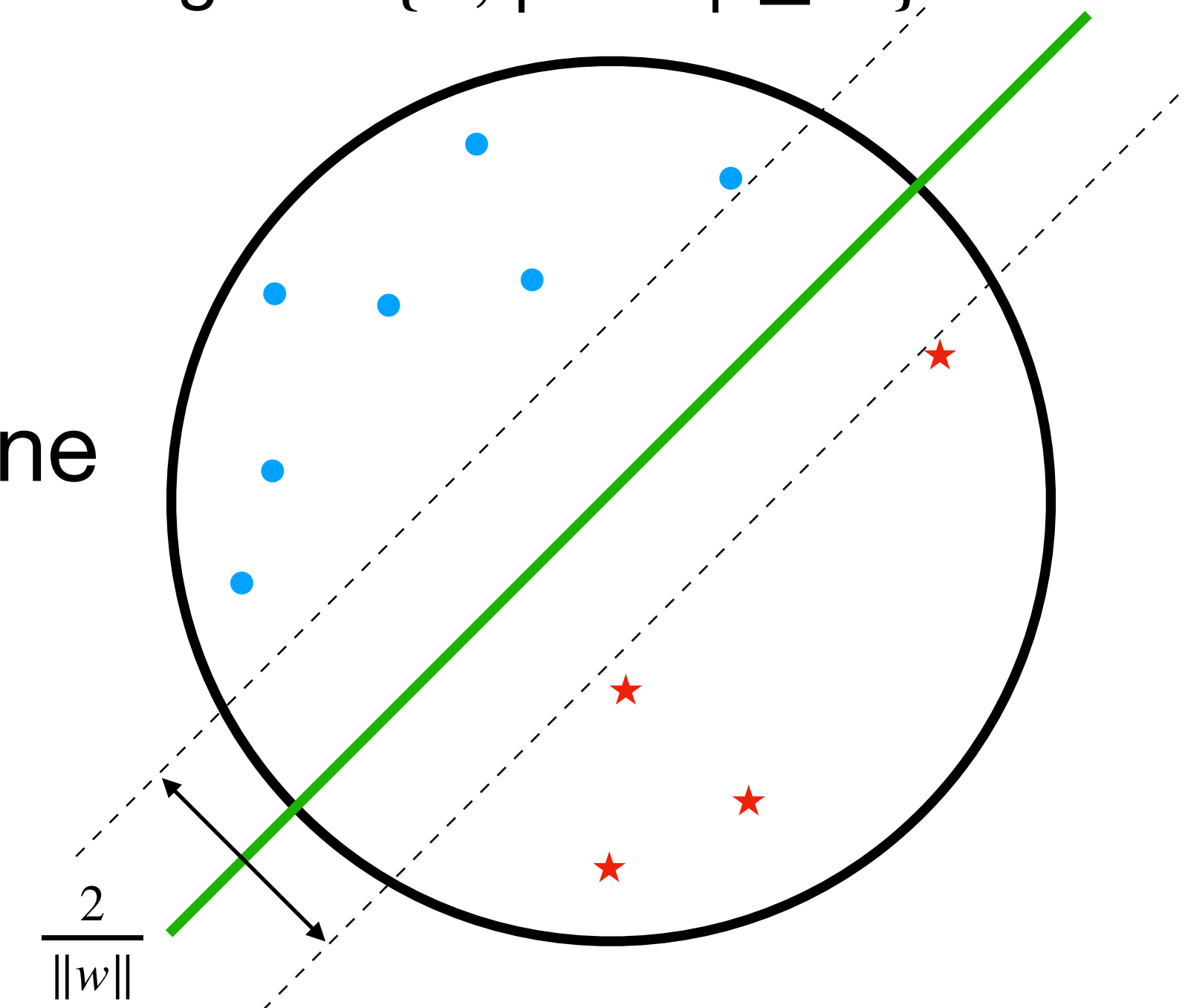
ERM for the hinge loss with ridge regularization

Interpretation for separable data and small  $\lambda$ : select

1. The direction of  $w$  so that  $w^\perp$  is a separating hyperplane
2. The scale of  $w$  so that no point is in the margin
3. Take the one for which the margin is the largest



Margin:  $= \{x; |x^\top w| \leq 1\}$



# Optimization: How to get $w$ ?

$$\min_w \frac{1}{N} \sum_{n=1}^N [1 - y_n x_n^\top w]_+ + \frac{\lambda}{2} \|w\|^2$$

Convex (but non smooth) objective which can be minimized with:

- Subgradient method
- Stochastic Subgradient method

# Convex duality

Assume you can define an auxiliary function  $G(w, \alpha)$  such that

$$\min_w L(w) = \min_w \max_{\alpha} G(w, \alpha)$$

Primal problem:  $\min_w \max_{\alpha} G(w, \alpha)$

Dual problem:  $\max_{\alpha} \min_w G(w, \alpha)$

➡ Sometimes the dual problem is simpler to solve than the primal one

Questions:

1. How do we find a suitable  $G(w, \alpha)$ ?
2. When can the min and the max be switched?
3. When is the dual problem easier to solve than the primal one?

# Q1: How do we find a suitable $G(w, \alpha)$ ?

$$[z]_+ = \max(0, z) = \max_{\alpha \in [0, 1]} \alpha z$$

$$\text{Therefore } [1 - y_n x_n^\top w]_+ = \max_{\alpha_n \in [0, 1]} \alpha_n (1 - y_n x_n^\top w)$$

The SVM problem is equivalent to:

$$\min_w L(w) = \min_w \max_{\alpha \in [0, 1]^n} \underbrace{\frac{1}{N} \sum_{n=1}^N \alpha_n (1 - y_n x_n^\top w)}_{G(w, \alpha)} + \frac{\lambda}{2} \|w\|_2^2$$

The function  $G$  is convex in  $w$  and concave in  $\alpha$

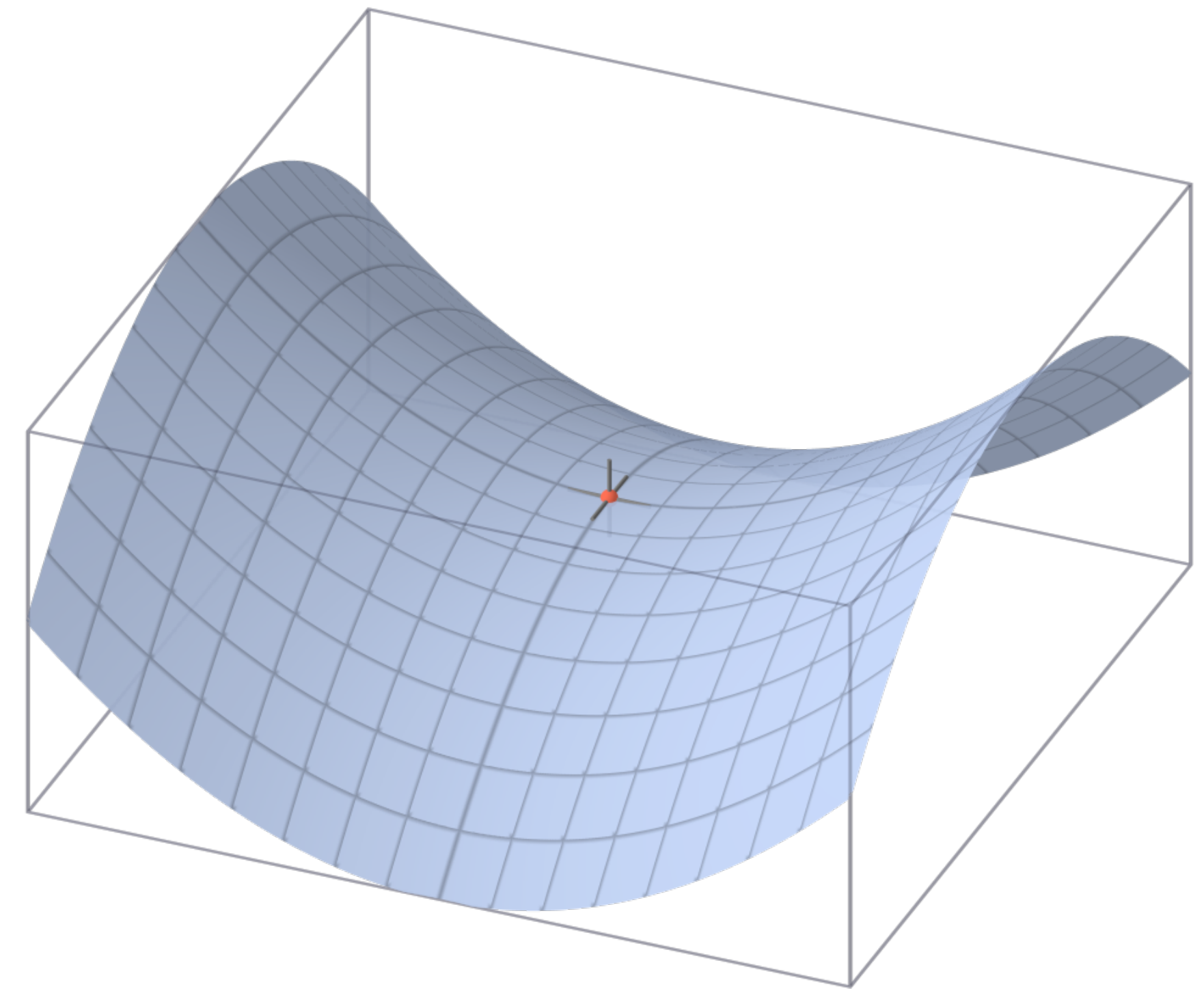
# Q2: Can we exchange the min and the max?

Always true:

$$\max_{\alpha} \min_w G(w, \alpha) \leq \min_w \max_{\alpha} G(w, \alpha)$$

Equality if  $G$  is convex in  $w$ , concave in  $\alpha$  and the domains of  $w$  and  $\alpha$  are convex and compact:

$$\max_{\alpha} \min_w G(w, \alpha) = \min_w \max_{\alpha} G(w, \alpha)$$





# Q2: Can we exchange the min and the max?

Always true:

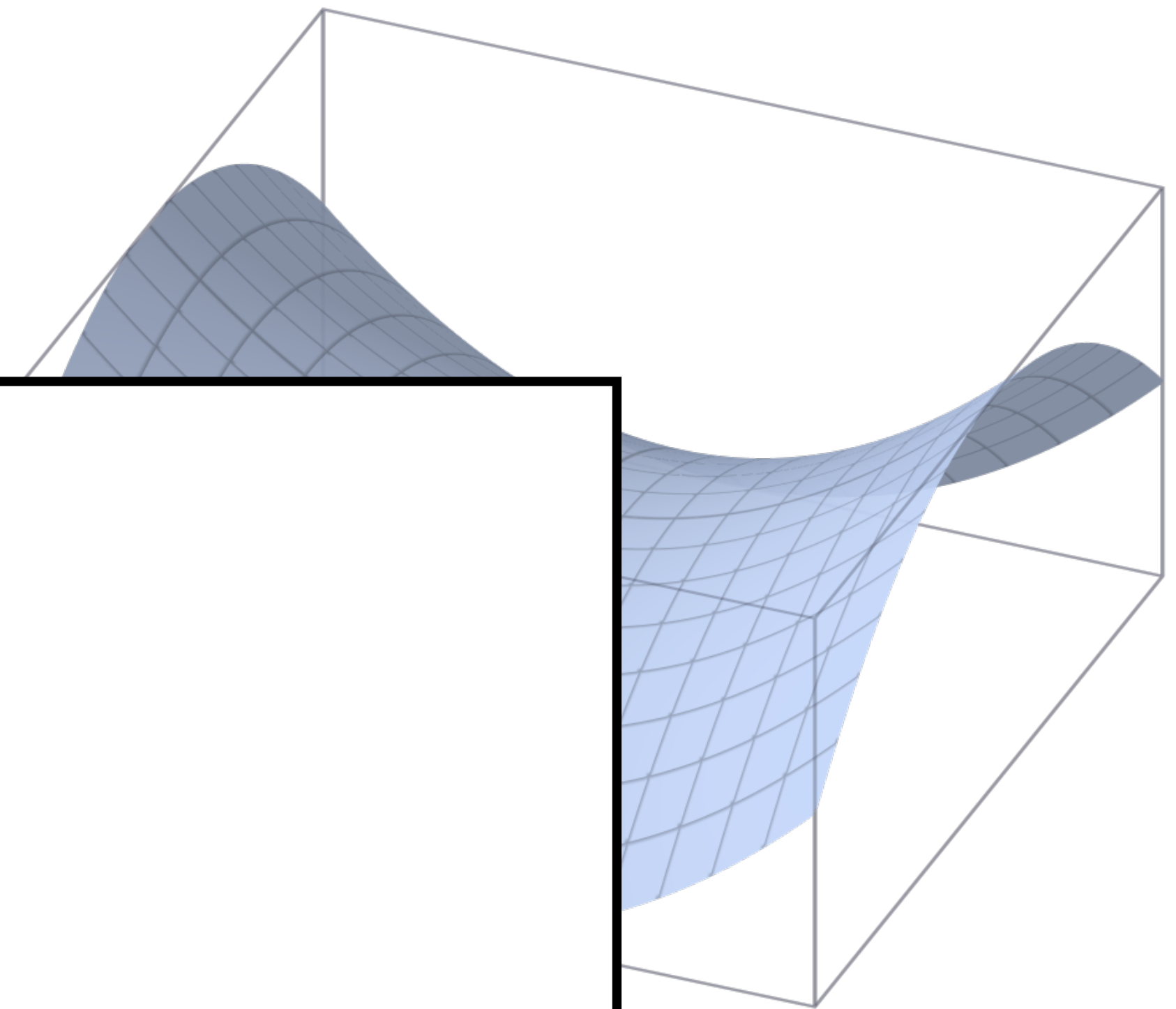
$$\max_{\alpha} \min_w G(w, \alpha) \leq \min_w \max_{\alpha} G(w, \alpha)$$

Proof:

$$\min_w G(\alpha, w) \leq G(\alpha, w') \text{ for any } w'$$

$$\max_{\alpha} \min_w G(\alpha, w) \leq \max_{\alpha} G(\alpha, w') \text{ for any } w'$$

$$\max_{\alpha} \min_w G(\alpha, w) \leq \min_{w'} \max_{\alpha} G(\alpha, w')$$



# Application to SVM

For SVM the condition is fulfilled and we can switch the min and max:

$$\min_w L(w) = \max_{\alpha \in [0,1]^n} \min_w \frac{1}{N} \sum_{n=1}^N \alpha_n (1 - y_n x_n^\top w) + \frac{\lambda}{2} \|w\|_2^2$$

Minimizer computation:

$$\nabla_w G(w, \alpha) = -\frac{1}{N} \sum_{n=1}^N \alpha_n y_n x_n + \lambda w = 0 \implies w(\alpha) = \frac{1}{\lambda N} \sum_{n=1}^N \alpha_n y_n x_n = \frac{1}{\lambda N} \mathbf{X}^\top \mathbf{Y} \alpha$$

$\mathbf{Y} = \text{diag}(\mathbf{y})$   
↓

Dual optimization problem:

$$\begin{aligned} \min_w L(w) &= \max_{\alpha \in [0,1]^n} \frac{1}{N} \sum_{n=1}^N \alpha_n \left(1 - \frac{1}{\lambda N} y_n x_n^\top \mathbf{X}^\top \mathbf{Y} \alpha\right) + \frac{1}{2\lambda N^2} \|\mathbf{X}^\top \mathbf{Y} \alpha\|_2^2 \\ &= \max_{\alpha \in [0,1]^n} \frac{1^\top \alpha}{N} - \frac{1}{\lambda N^2} \alpha^\top \mathbf{Y} \mathbf{X} \mathbf{X}^\top \mathbf{Y} \alpha + \frac{1}{2\lambda N^2} \|\mathbf{X}^\top \mathbf{Y} \alpha\|_2^2 \\ &= \max_{\alpha \in [0,1]^n} \frac{1^\top \alpha}{N} - \frac{1}{2\lambda N^2} \alpha^\top \underbrace{\mathbf{Y} \mathbf{X} \mathbf{X}^\top \mathbf{Y}}_{\text{PSD matrix}} \alpha \end{aligned}$$

PSD matrix

# Q3: Why?

$$\max_{\alpha \in [0,1]^n} \alpha^\top \mathbf{1} - \frac{1}{2\lambda N} \alpha^\top \underbrace{\mathbf{Y} \mathbf{X} \mathbf{X}^\top \mathbf{Y}}_{\text{PSD matrix}} \alpha$$

1. It is a differentiable concave problem. It can be efficiently solved with
  - quadratic programming solvers
  - coordinate ascent
2. The cost function only depends on the data through the *kernel matrix*  $K = \mathbf{X} \mathbf{X}^\top \in \mathbb{R}^{N \times N}$  - It does not depend on  $d$
3. The dual formulation provides a meaningful interpretation:  $\alpha$  is typically sparse and is non-zero only for the training examples instrumental in determining the decision boundary

# Interpretation of the dual formulation

For any  $(x_n, y_n)$ , there is a corresponding  $\alpha_n$  given by

$$\max_{\alpha_n \in [0,1]} \alpha_n (1 - y_n x_n^\top w)$$

- $x_n$  lies on the correct side and outside the margin,  $1 - y_n x_n^\top w < 0$  and hence  $\alpha_n = 0$   
    ➡ Non-support point
- $x_n$  lies on the correct side but on the margin,  $1 - y_n x_n^\top w = 0$  and hence  $\alpha_n = [0,1]$   
    ➡ Essential support vector
- $x_n$  lies strictly inside the margin or on the wrong side,  $1 - y_n x_n^\top w > 0$  and  $\alpha_n = 1$   
    ➡ Bound support vector

# The SVM hyperplane is supported by the support vectors

$$w = \frac{1}{\lambda N} \sum_{n=1}^N \alpha_n y_n x_n$$

➔  $w$  does not depend on the observation  $(x_n, y_n)$  if  $\alpha_n = 0$

