Support Vector Machines

Machine Learning Course - CS-433
Nov 2, 2021
Nicolas Flammarion

EPFL

A 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 classifiac-tion 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

Machine Learning, 20, 273–297 (1995)

Support-Vector Networks

CORINNA CORTES corinna@neural.att.com
VLADIMIR VAPNIK vlad@neural.att.com
AT&T Bell Labs., Holmdel, NJ 07733, USA

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.

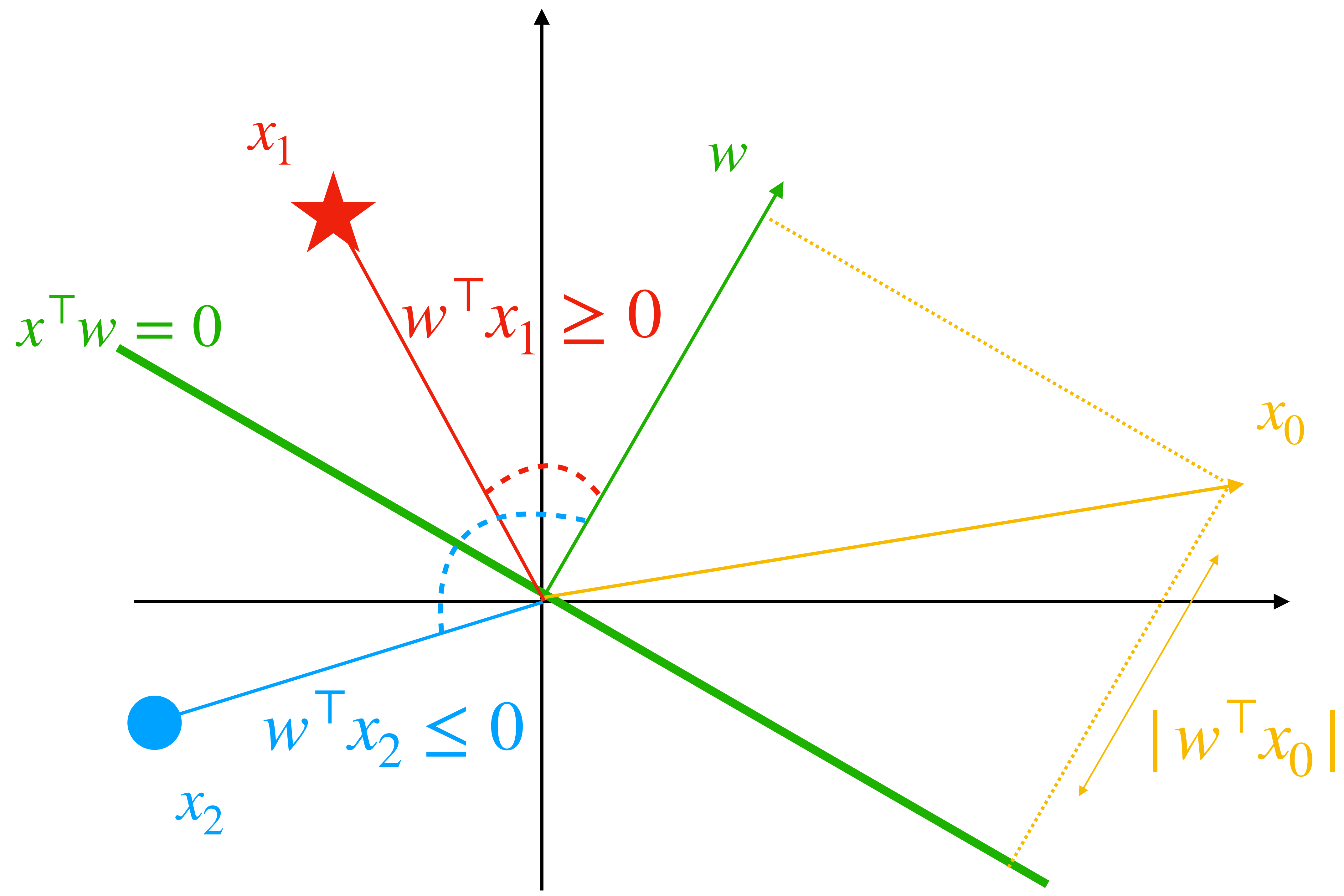
Linear Classifier

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

Prediction:

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

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

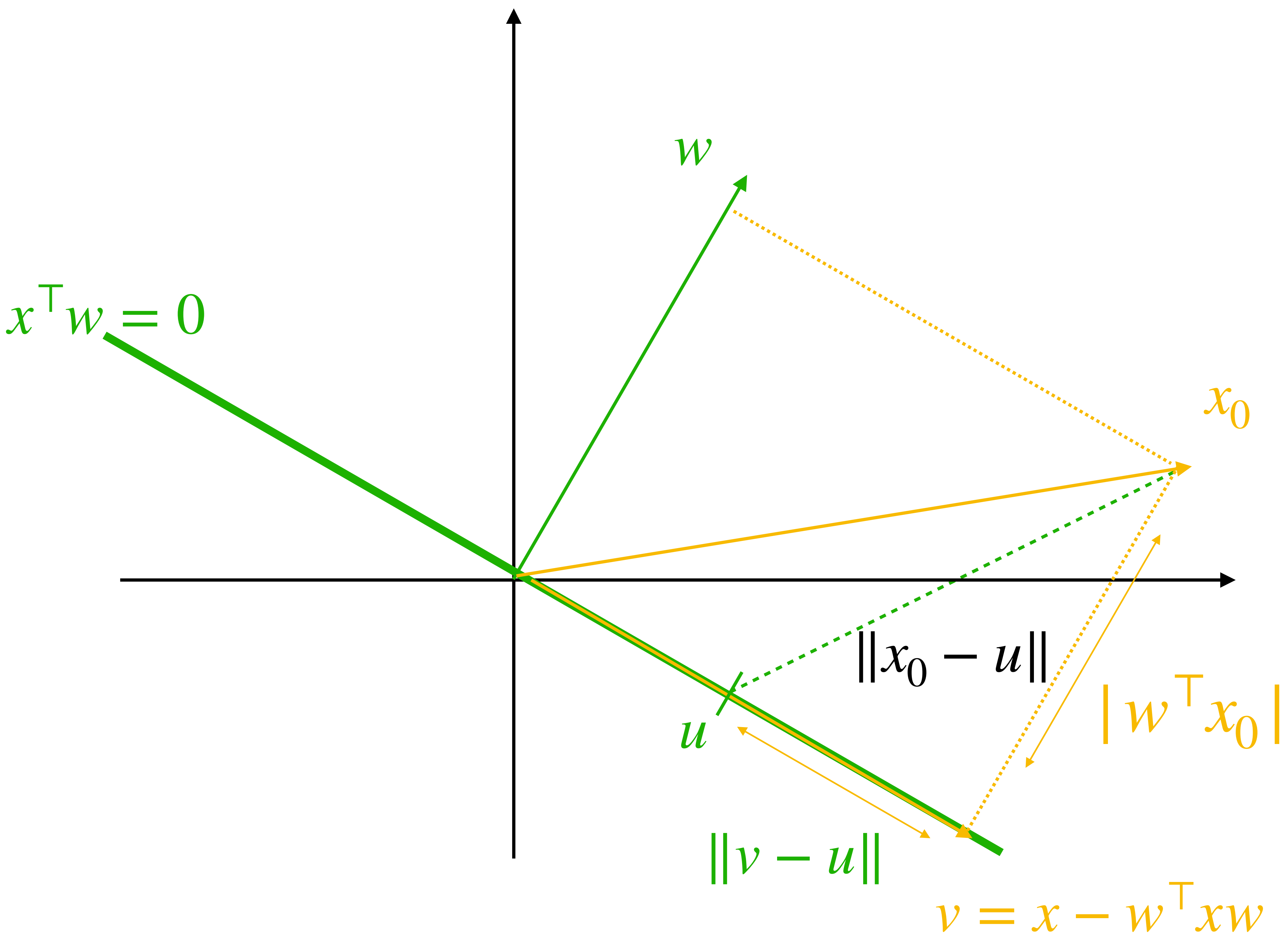


Linear Classifier

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. $w^\top u = 0$

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

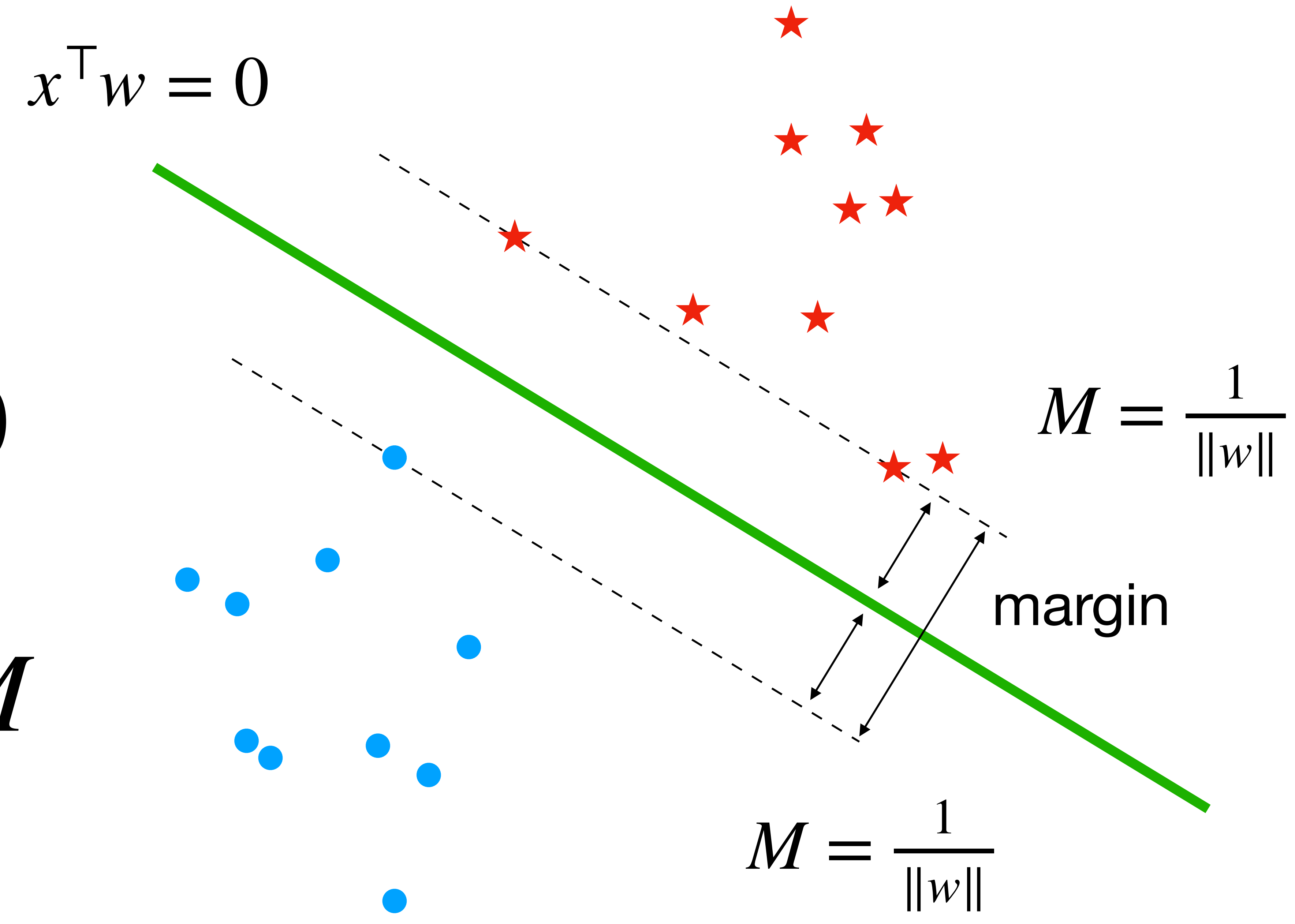


<u>Claim:</u> The distance between a point x_0 and the hyperplane defined by w is $|w^\top x_0|$

Hard-SVM rule: max-margin separating hyperplane

First assume the dataset $(x_i, y_i)_{i=1}^n$ is linearly separable

<u>Margin of a hyperplane</u>: $\min_{i \le n} |w^\top x_i|$



<u>Max-margin separating hyperplane</u>:

$$\max_{w, \|w\|=1} \min_{i \le n} |w^\top x_i| \text{ such that } \forall i,\ y_i x_i^\top w \ge 0$$

Equivalent to $\max_{w, \|w\|=1} M$ such that $\forall i, y_i x_i^\top w \ge M$

also equivalent to:

$$\boxed{\min_w \|w\| \text{ such that } \forall i,\ y_i x_i^\top w \ge 1}$$

Proof of the equivalent formulations

Claim: The following optimization problems are equivalent

$$\max_{w,\|w\|=1} \min_{i\leq n} |w^\top x_i| \quad \text{s.t. } \forall i,\ y_i x_i^\top w \geq 0 \qquad \text{(I)}$$

$$\max_{w,\|w\|=1} M \quad \text{s.t. } \forall i, y_i x_i^\top w \geq M \qquad \text{(II)}$$

Proof: let w_1 a solution of (I) and $M_1 = \min_{i\leq n} |w_1^\top x_i|$ 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 $\min_{i\leq n} |w_2^\top x_i| \leq \min_{i\leq n} |w_1^\top x_i|$
- $\forall i, y_i x_i^\top w_2 \geq M_2$ implies that $\forall i, |x_i^\top w_2| \geq M_2$ and $\min_{i\leq n} |x_i^\top w_2| \geq M_2$

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

And the two problems are equivalent

Proof of the equivalent formulations

Claim: The following optimization problems are equivalent

$$\max_{w, \|w\|=1} M \quad \text{s.t. } \forall i, y_i x_i^\top w \geq M \qquad \text{(II)}$$

$$\min_{w} \|w\| \quad \text{s.t. } \forall i, y_i x_i^\top w \geq 1 \qquad \text{(III)}$$

Proof: $\max_{M, w, \|w\|=1} M$ such that $\forall i, y_i x_i^\top w \geq M$

$\iff \max_{w} M$ such that $\forall i, y_i x_i^\top \frac{w}{\|w\|} \geq M$

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

$\iff \max_{w} 1/\|w\|$ such that $\forall i, y_i x_i^\top w \geq 1$

$\iff \min_{w} \|w\|$ such that $\forall i, y_i x_i^\top w \geq 1$

Proof of the equivalent formulations

Claim: The following optimization problems are equivalent

$$\max_{w, \|w\|=1} \; M \quad \text{s.t. } \forall i, y_i x_i^\top w \geq M \qquad \text{(II)}$$

$$\min_{w} \|w\| \quad \text{s.t. } \forall i, y_i x_i^\top w \geq 1 \qquad \text{(III)}$$

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 solution of (II)
- w_2/M_2 is solution of (I)

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

<u>Idea</u>: still maximize the margin, but allow some of the constraints to be violated

<u>How</u>: by introducing positive slack variables $\xi_1, \cdots, \xi_n$ and replace the constraints by $y_i x_i^\top w \geq 1 - \xi_i$

<u>Soft SVM</u>:

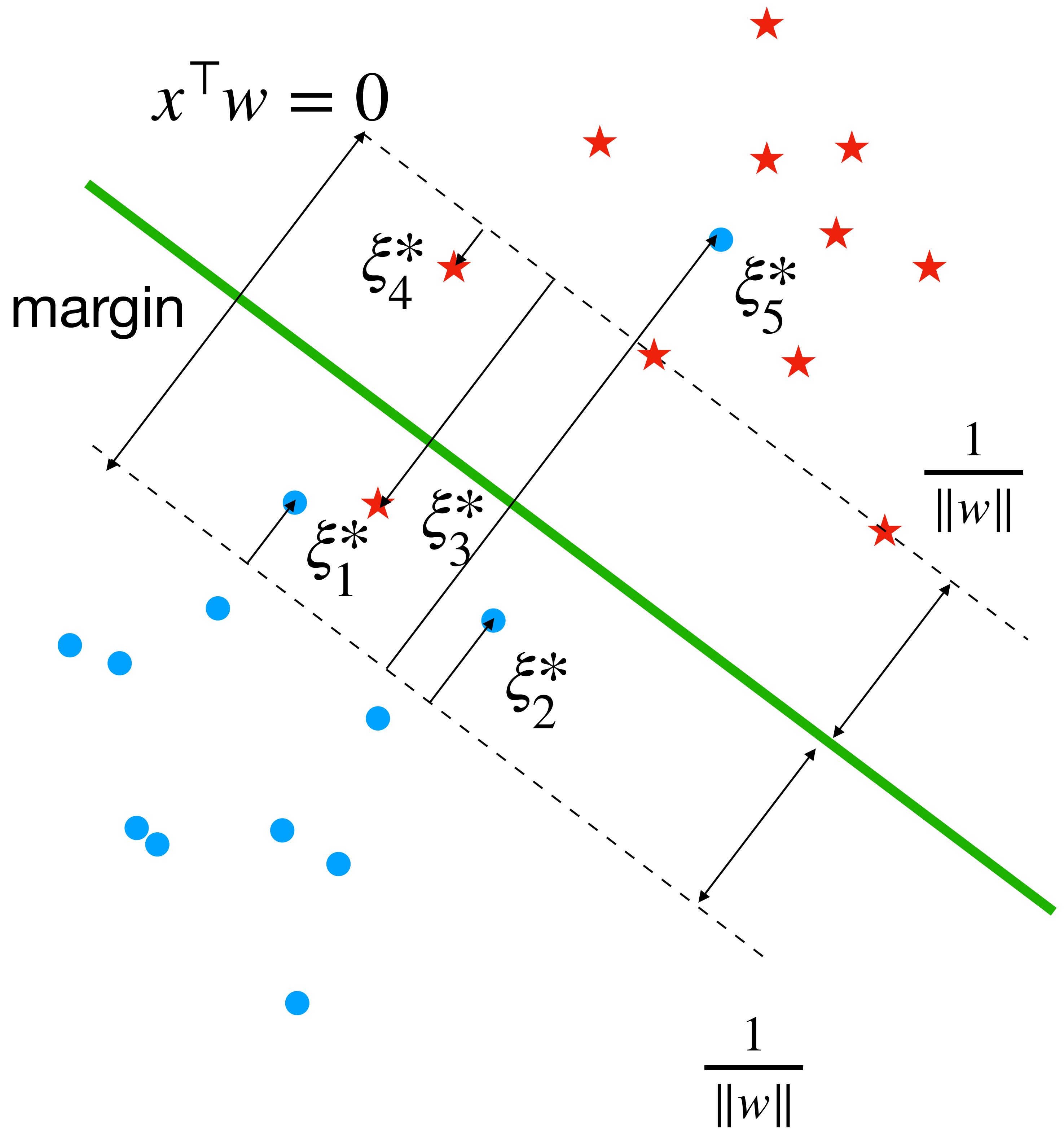
$$\min_{w,\xi} \lambda \|w\|^2 + \sum_{i=1}^n \xi_i$$

$$\text{s.t. } \forall i, y_i x_i^\top w \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0$$

which is equivalent to

$$\boxed{\min_{w} \lambda \|w\|^2 + \sum_{i=1}^n [1 - y_i x_i^\top w]_+}$$

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



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 ξ:

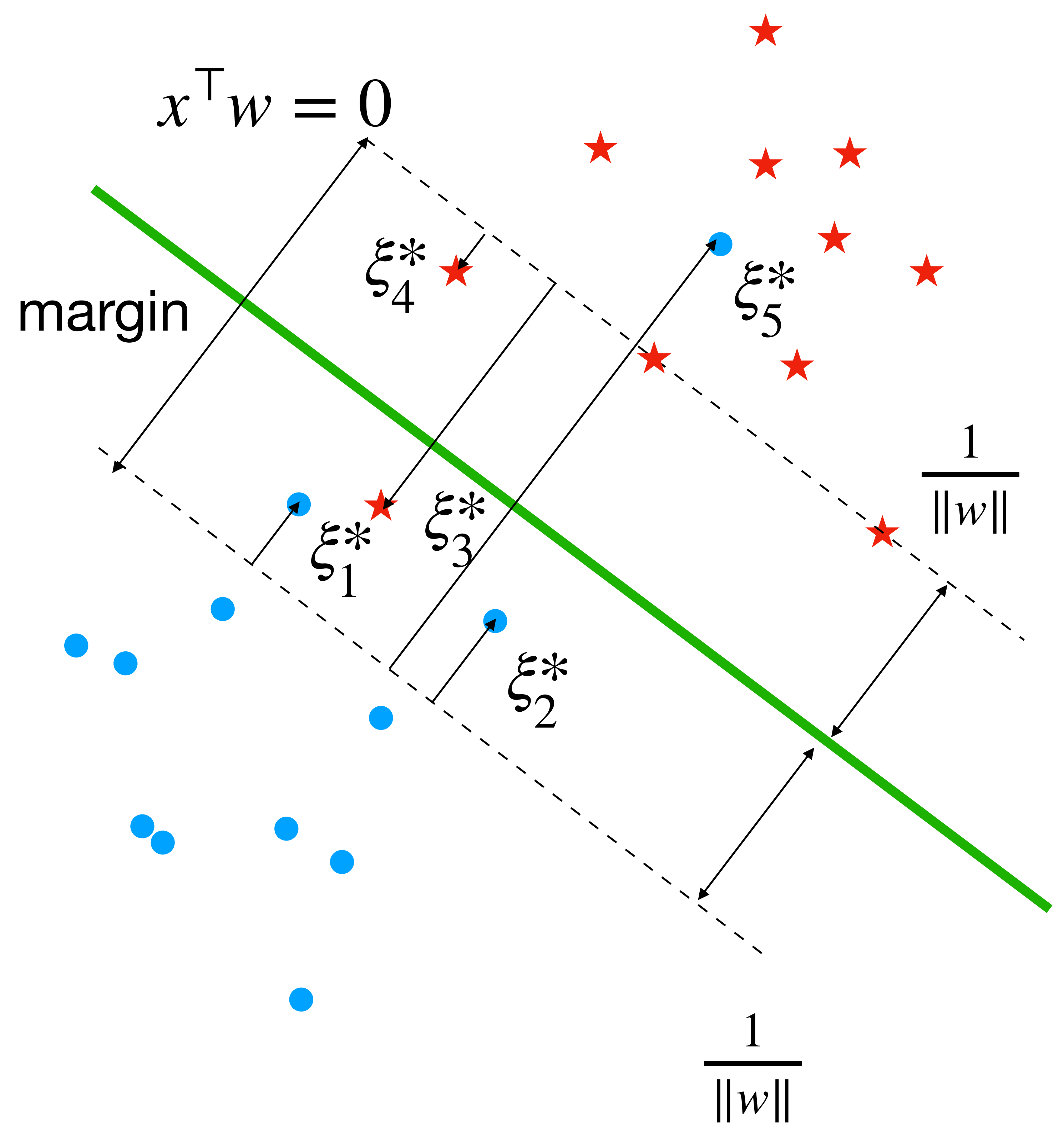
- If $y_i x_i^\top w \geq 1$, then $\xi_i = 0$
- If $y_i x_i^\top w < 1$, $\xi_i = 1 - y_i x_i^\top w$

Therefore $\xi_i = [1 - y_i x_i^\top w]_+$

, and

$$\min_{w,\xi} \lambda \|w\|^2 + \sum_{i=1}^n \xi_i$$

s.t. $\forall i, y_i x_i^\top w \geq 1 - \xi_i$ and $\xi_i \geq 0$



which is equivalent to

$$\boxed{\min_{w} \lambda \|w\|^2 + \sum_{i=1}^n [1 - y_i x_i^\top w]_+}$$

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

Classification by risk minimization

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

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

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

How: empirical risk minimization (ERM):

$$\min_{g:\mathscr{X}\to\mathscr{Y}} L_{\text{train}}(g) := \frac{1}{n}\sum_{i=1}^{n} 1_{g(x_i)\neq y_1}$$

Problem: L_{train} is not convex:

1. The set of classifier is not convex because $\mathscr{Y}$ is discrete
2. The indicator function 1 is not convex because not continuous

Convex relaxation of the classification risk

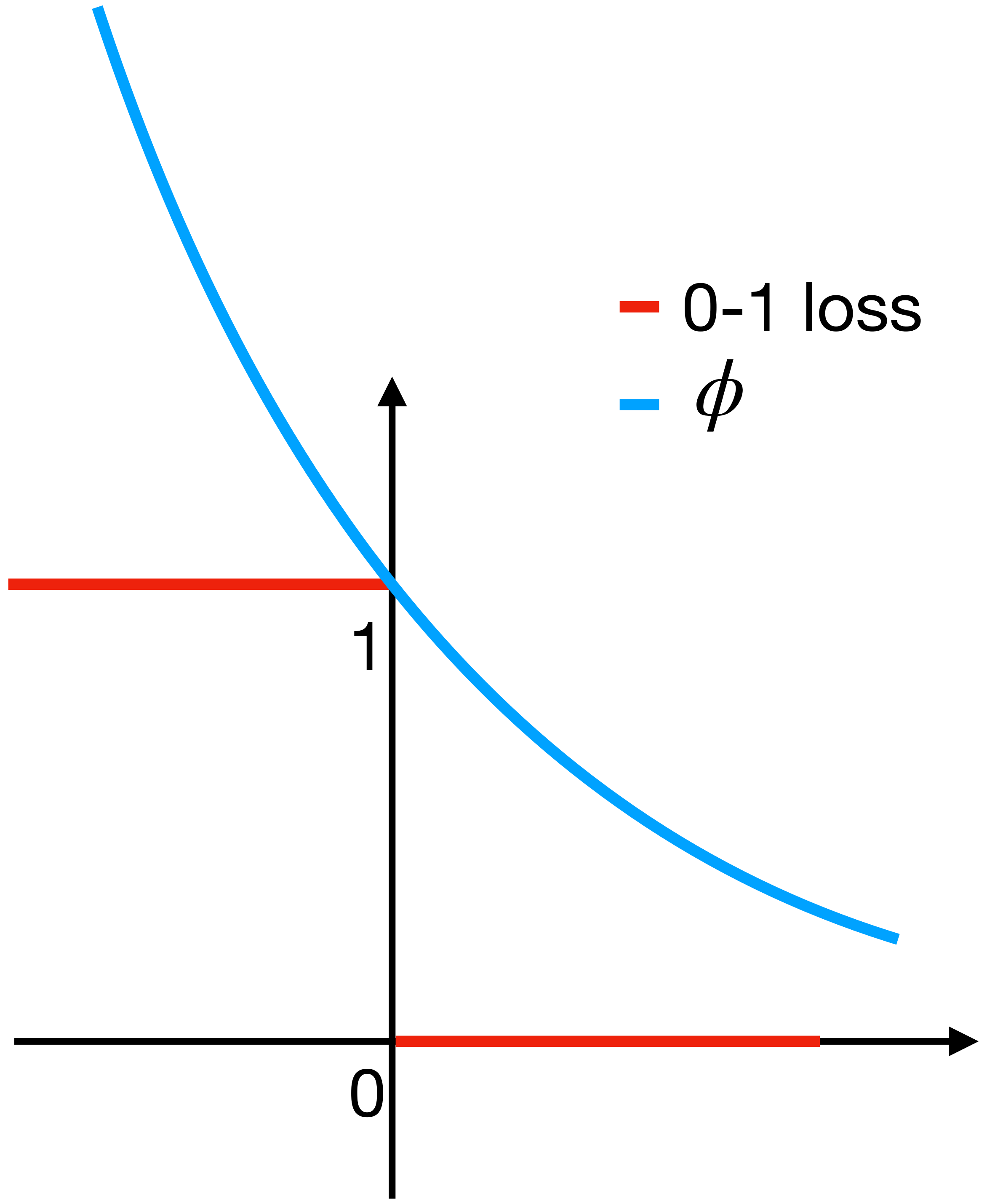
1. Consider the set of linear predictor $w^\top x$ and then predicts with $g(x) = \mathrm{sign}(w^\top x)$

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

$$\min_w \frac{1}{n}\sum_{i=1}^n 1_{-y_i x_i^\top w\geq 0}$$

2. Replace the indicator function by a convex surrogate ϕ and minimize

$$\min_w \frac{1}{n}\sum_{i=1}^n \phi(-y_i x_i^\top w)$$



<u>Rmk</u>: possible to bound the 0-1 risk $L(g)$ by the ϕ risk *

* Under technical assumptions on the function ϕ

Losses for Classification

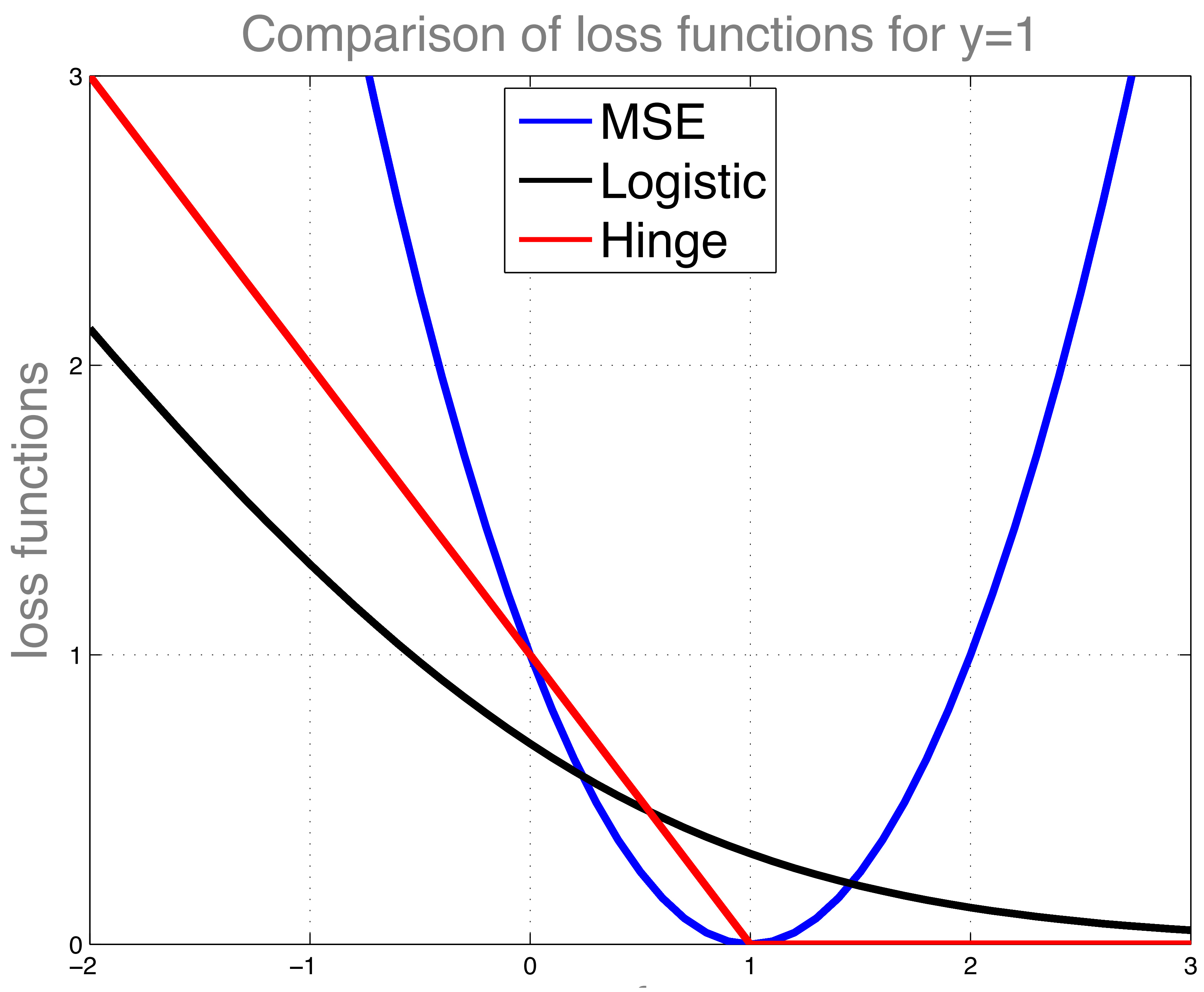
Ex:

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

Common features: they are convex and upper bound the 0-1 loss

Behavior difference:

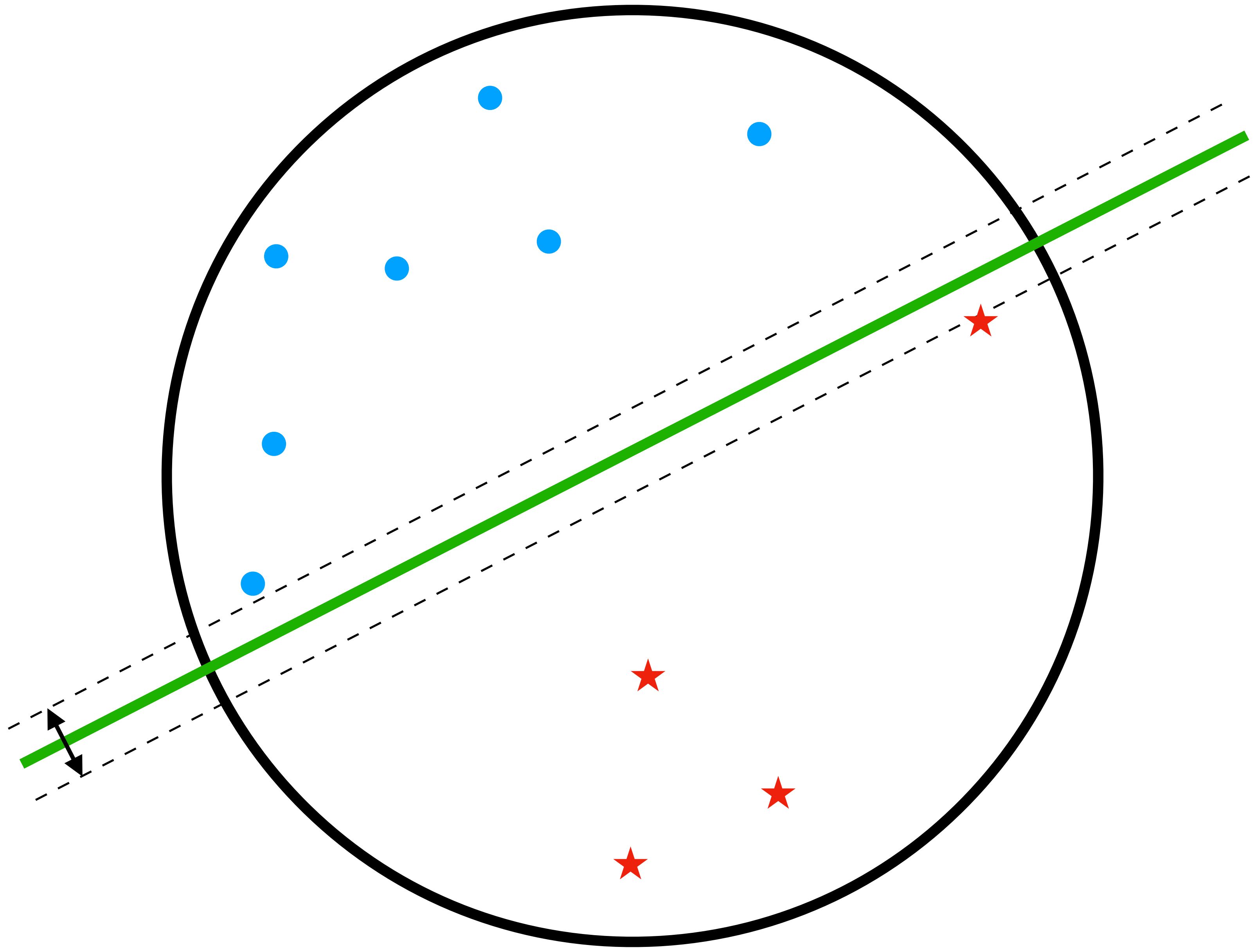
- MSE punishes any deviation from 1
- The logistic cost is asymmetric – we always incur a cost
- Hinge loss: we incur a cost if the prediction is incorrect or not confident enough



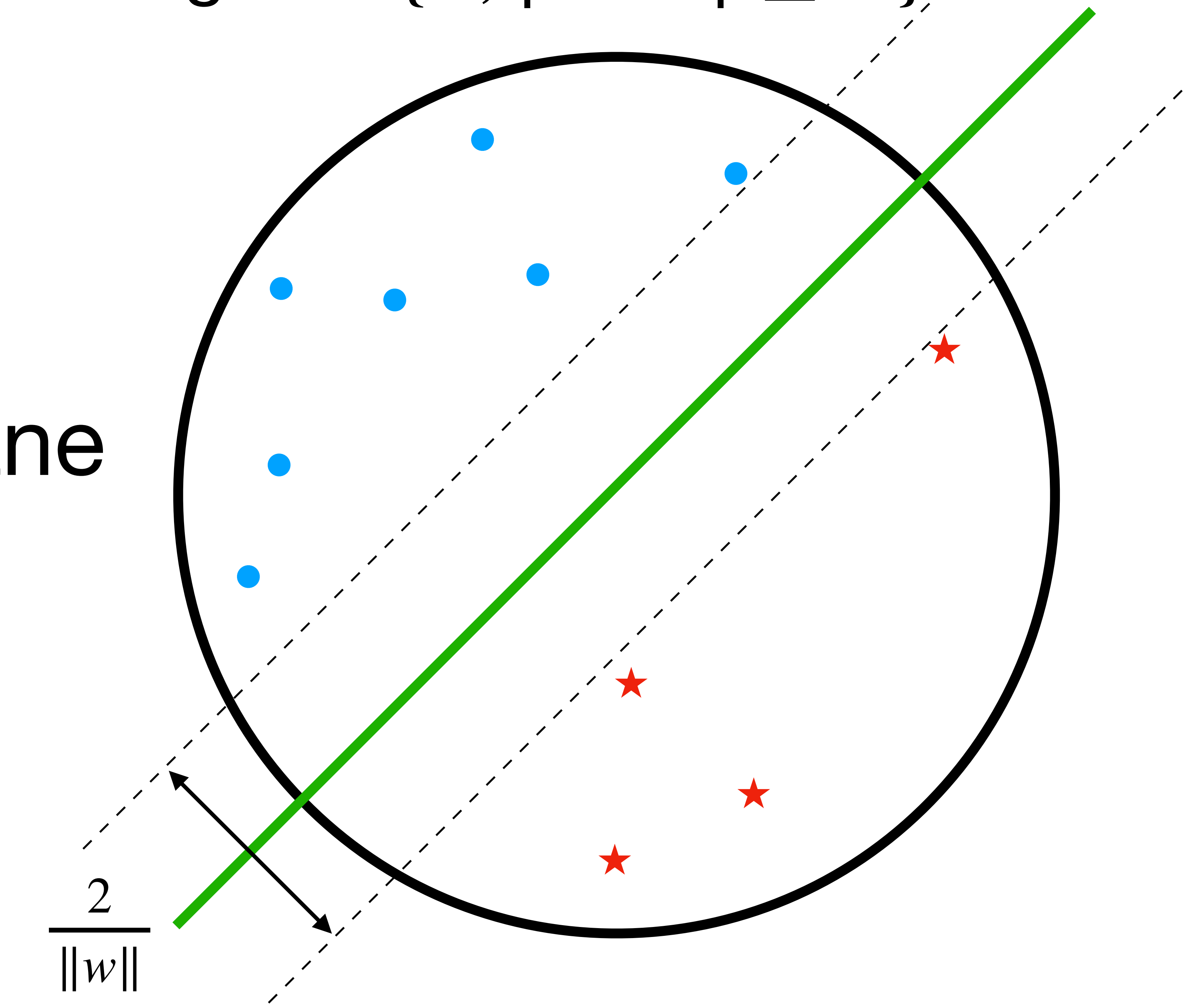
Summary

$$\min_{w} \lambda \|w\|^2 + \sum_{i=1}^{n} [1 - y_i x_i^\top w]_+$$

ERM for the hinge loss with ridge regularization



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



Interpretation for separable data and small λ: 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

Optimization: How to get w?

$$\min_{w} \sum_{i=1}^{n} \left[1 - y_i x_i^\top w\right]_+ + \frac{\lambda}{2}\|w\|^2$$

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

- Subgradient method
- Stochastic Subgradient method

Convex duality

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

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

<u>Primal problem</u>: $\min_w \max_\alpha G(w, \alpha)$

<u>Dual problem</u>: $\max_\alpha \min_w G(w, \alpha)$

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

<u>Questions:</u>

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 $G(w, \alpha)$?

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

Therefore $[1 - y_i x_i^\top w]_+ = \max_{\alpha_i \in [0,1]} \alpha_i (1 - y_1 x_i^\top w)$

The SVM problem is equivalent to:

$$\min_w L(w) = \min_w \max_{\alpha \in [0,1]^n} \underbrace{\textstyle\sum_{i=1}^n \alpha_i (1 - y_i x_i^\top w) + \frac{\lambda}{2} \|w\|_2^2}_{G(w,\alpha)}$$

The function G is convex in w and concave in α

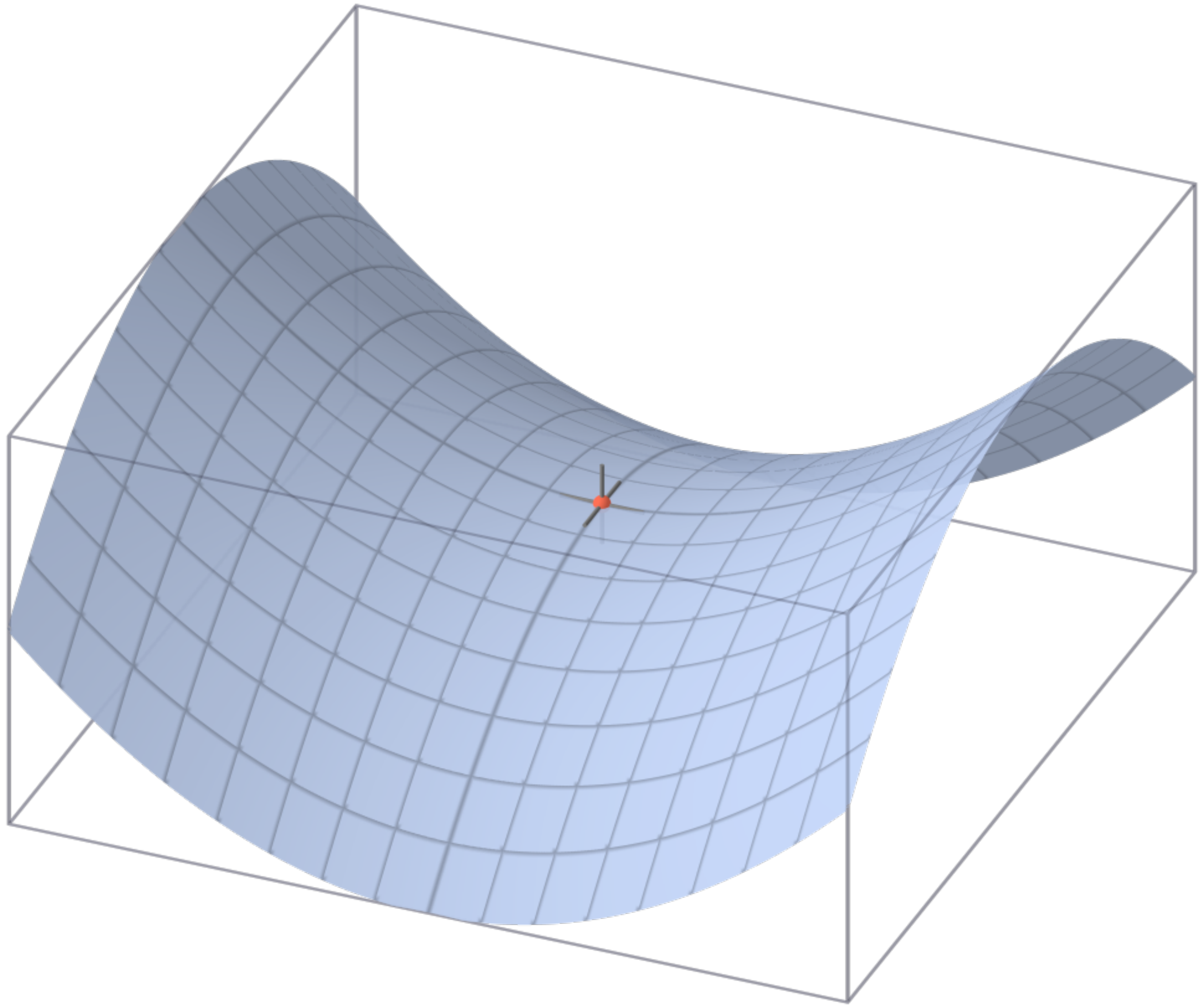
Q2: Can we exchange the min and the max?

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 α and the domains of w and α are convex and compact:

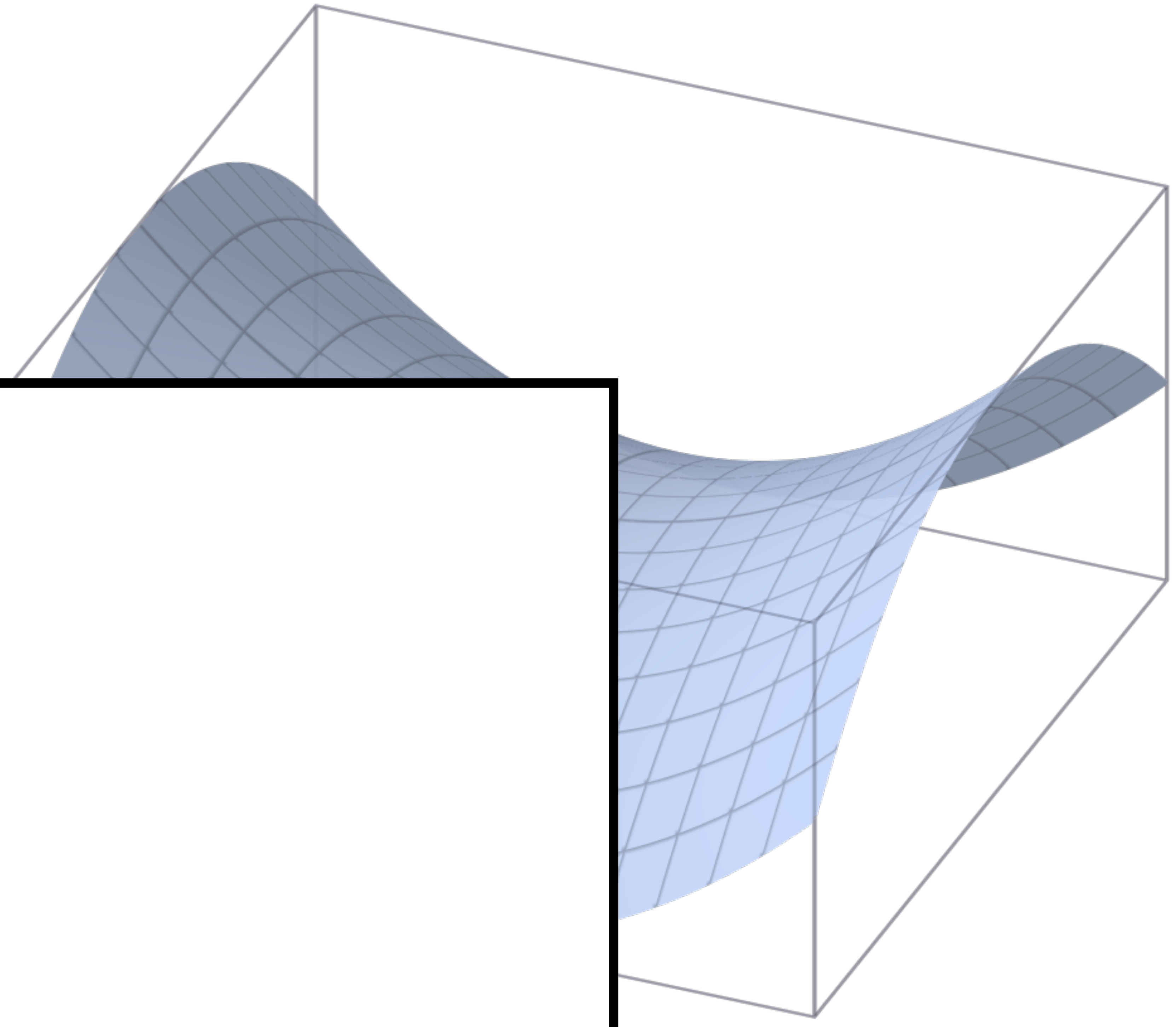
$$\max_{\alpha} \min_{w} G(w, \alpha) = \min_{w} \max_{\alpha} G(w, \alpha)$$



Q2: Can we exchange the min and the max?

Always true:

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



<u>Proof:</u>

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

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

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

Application to SVM

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

$$\min_{w} L(w) = \max_{\alpha \in [0,1]^n} \min_{w} \sum_{i=1}^{n} \alpha_i (1 - y_i {x_i}^\top w) + \frac{\lambda}{2} \|w\|_2^2$$

<u>Minimizer computation:</u>

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

$$\nabla_w G(w, \alpha) = -\sum_{i=1}^{n} \alpha_i y_i x_i + \lambda w = 0 \implies w(\alpha) = \frac{1}{\lambda} \sum_{i=1}^{n} \alpha_i y_i x_i = \frac{1}{\lambda} \mathbf{X}^\top \mathbf{Y} \alpha$$

<u>Dual optimization problem:</u>

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

Q3: Why?

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

1. It is a differentiable concave problem. It can be efficiently solved with
   - quadratic programming solvers
   - coordinate ascent

2. The cost function is only depending 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 meaningful interpretation: α 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_i, y_i), there is a corresponding α_i given by

$$\max_{\alpha_i \in [0,1]} \alpha_i(1 - y_i x_i^\top w)$$

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

The SVM hyperplane is supported by the support vectors

$$w = \frac{1}{\lambda} \sum_{i=1}^{n} \alpha_i y_i x_i$$

➡ w does not depend on the observations (x_i, y_i) if $\alpha_i = 0$