# Support Vector Machines

Machine Learning Course - CS-433
Oct 24, 2023
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

GVB+92, Vap82, BH89, TLS89, Mac92] link the
zation of a classifier to the error on the training
es and the complexity of the classifier. Meth-
h as structural risk minimization [Vap82] vary
nplexity of the classification function in order to
e the generalization.

paper we describe a training algorithm that au-
ally tunes the capacity of the classification func-
maximizing the margin between training exam-

Machine Learning, 20, 273–297 (1995)


## Support-Vector Networks

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

**Editor:** Lorenza Saitta

**Abstract.** The *support-vector network* is a new learning m
machine conceptually implements the following idea: inpu
dimension feature space. In this feature space a linear decisi
decision surface ensures high generalization ability of the lea
network was previously implemented for the restricted case
errors. We here extend this result to non-separable training da
High generalization ability of support-vector networks uti
strated. We also compare the performance of the support-vec
that all took part in a benchmark study of Optical Character R

*Том XXIV* *«АВТОМАТИКА И ТЕЛЕМЕХАНИКА»* № 6
*1963*

УДК 519.95

## УЗНАВАНИЕ ОБРАЗОВ ПРИ ПОМОЩИ ОБОБЩЕННЫХ ПОРТРЕТОВ

В. Н. ВАПНИК, А. Я. ЛЕРНЕР
(*Москва*)

Дается аксиоматическое определение образа. Вводятся понятия «обобщенный портрет», «различение» и «узнавание». Предлагаются алгоритмы обучения узнаванию и различению, основанные на нахождении обобщенных портретов образов.

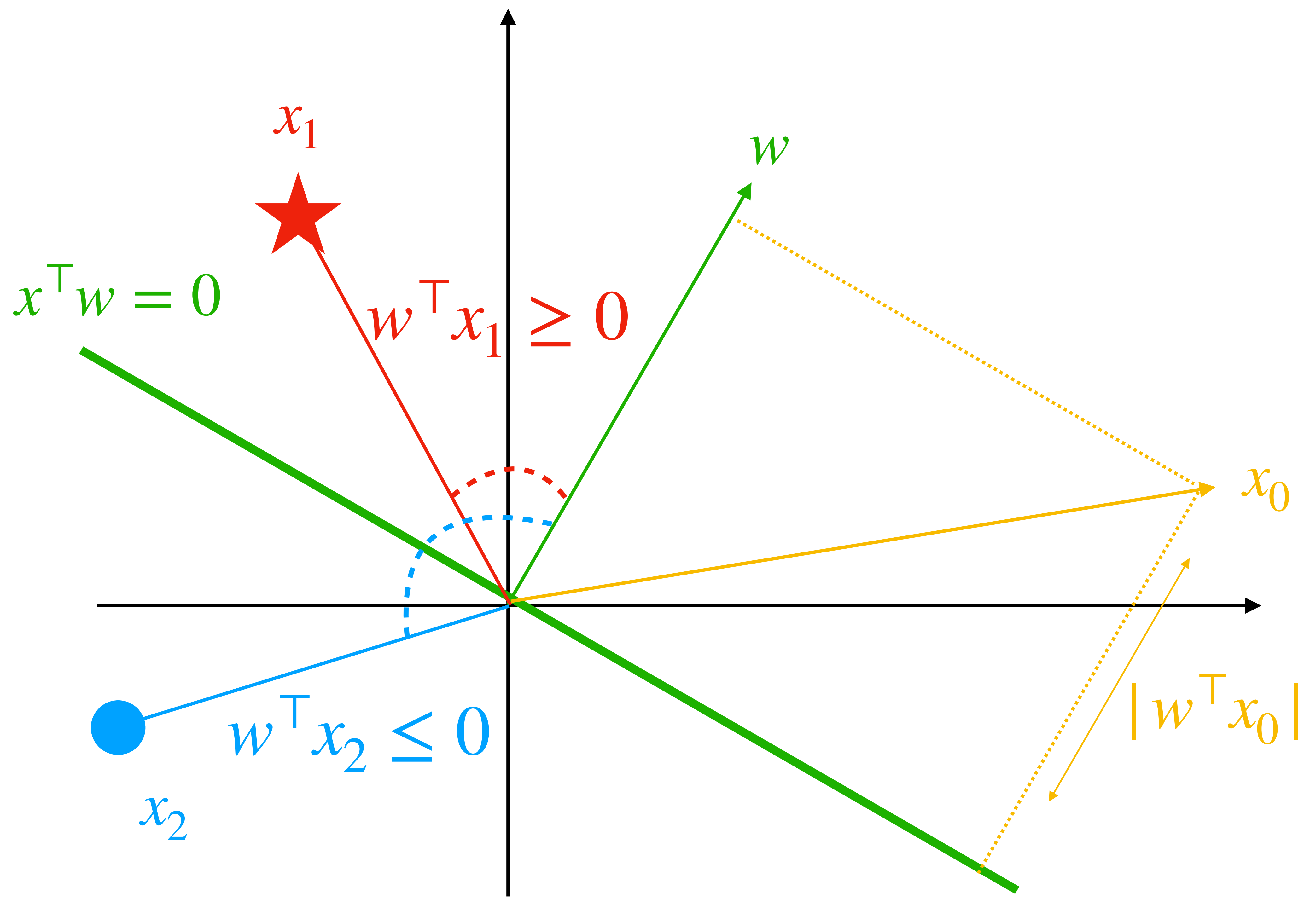
# Linear Classifier

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

Prediction:

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

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



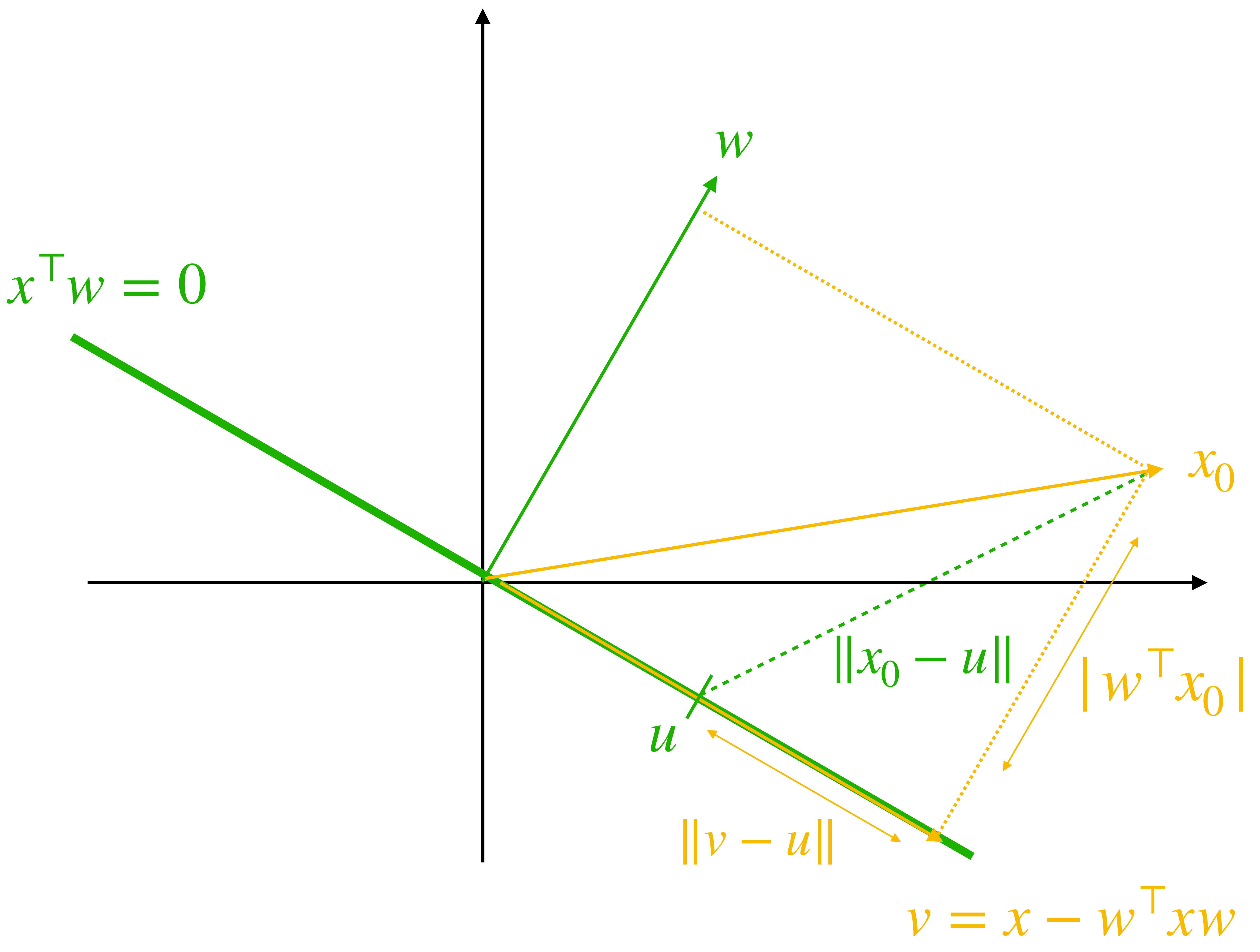
# Linear Classifier

Proof: The distance between $x_0$ and the hyperplane is given 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 $(x_n, y_n)_{n=1}^N$ is linearly separable

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

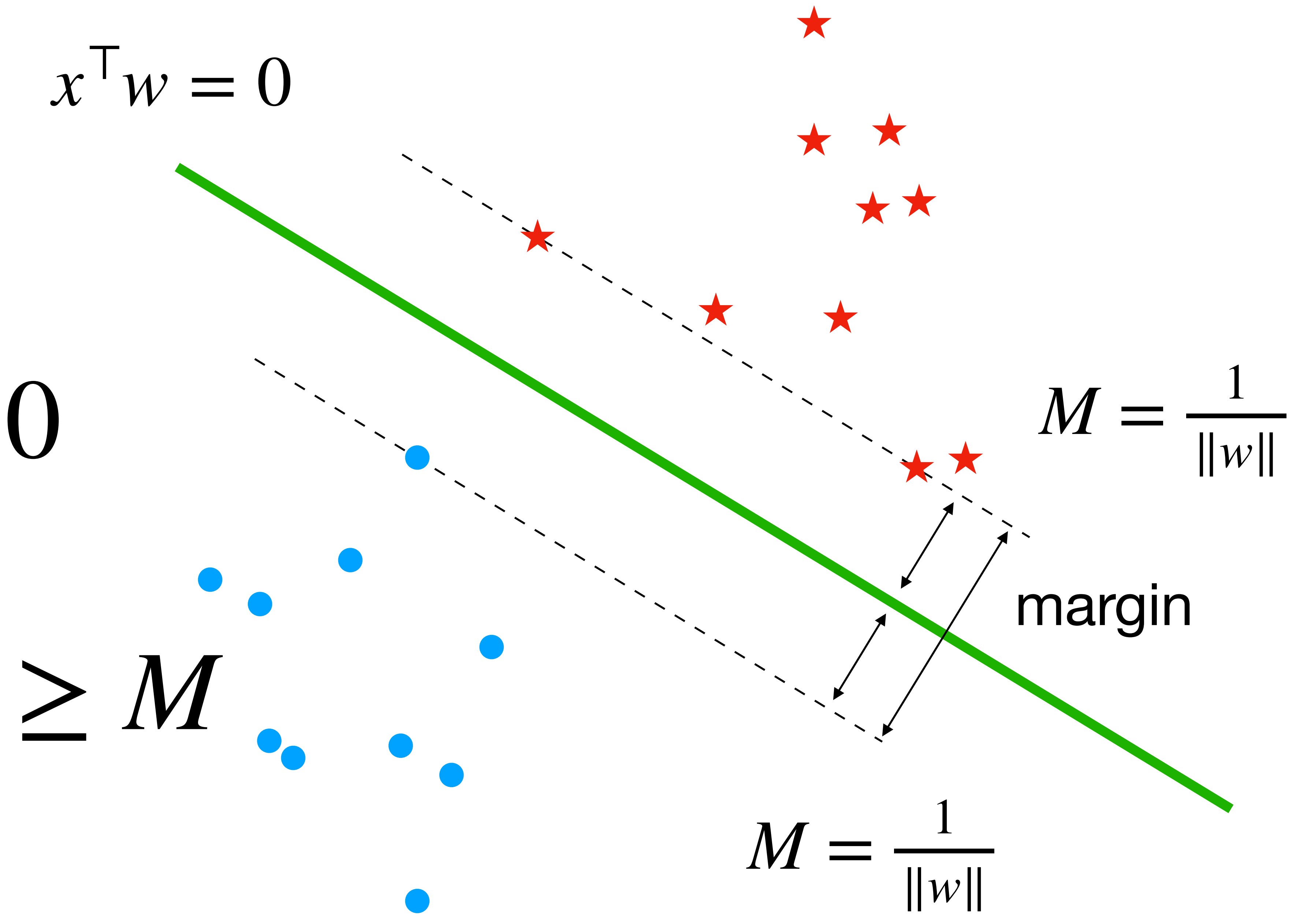
<u>Max-margin separating hyperplane</u>:

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

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:

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



# Proof of the equivalent formulations

<u>Claim:</u> The following optimization problems are equivalent

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

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

<u>Proof</u>: 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

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

$$\min_{w} \frac{1}{2}\|w\|^2 \quad \text{s.t. } \forall n, y_n x_n^\top w \geq 1 \qquad \text{(III)}$$

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

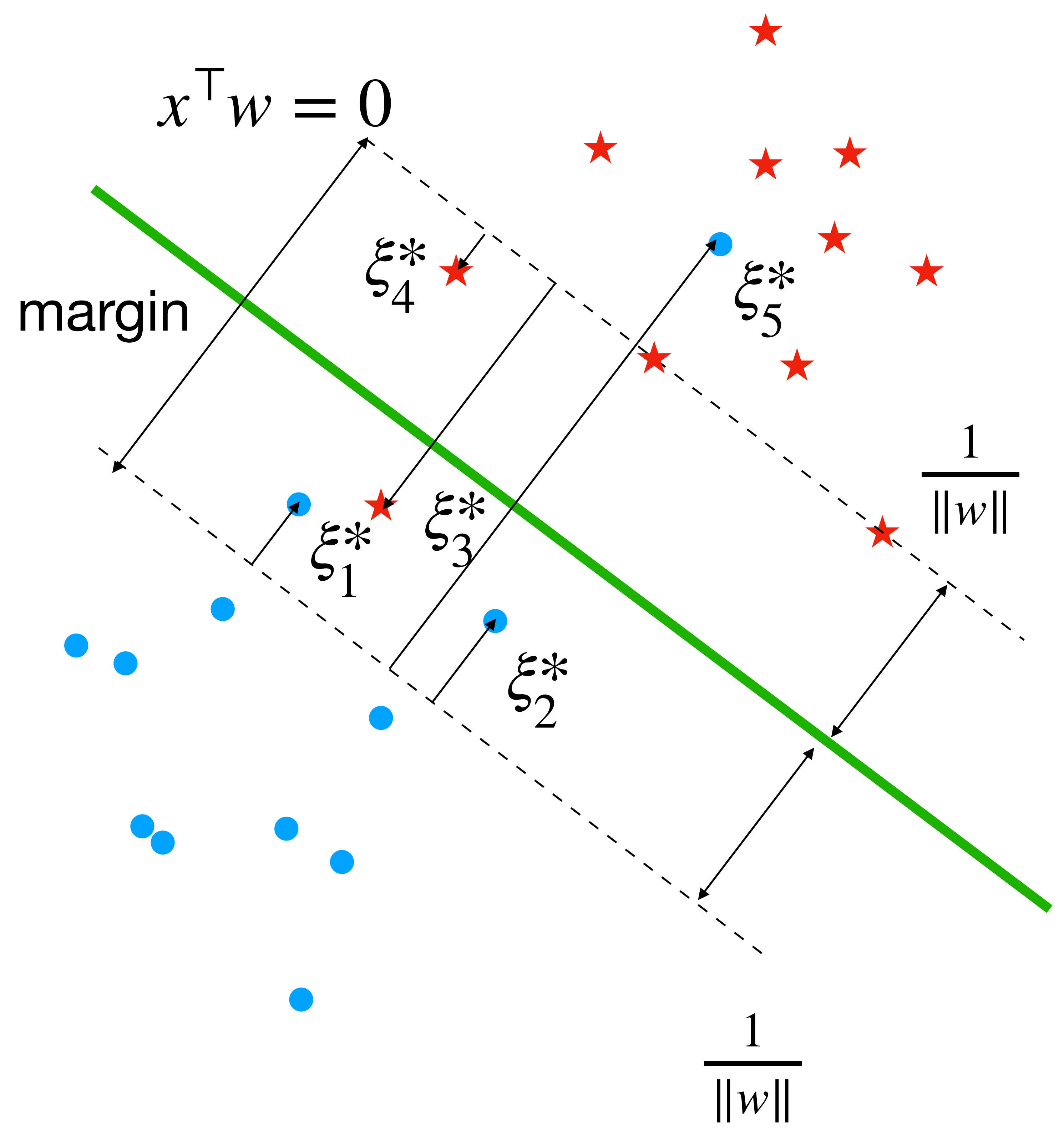
$\iff \max_{M \in \mathbb{R}, w} M$ such that $\forall n, y_n x_n^\top \frac{w}{\|w\|} \geq M$

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

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

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

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




Idea: Maximize the margin while allowing some constraints to be violated

How: Introduce positive slack variables $\xi_1, \cdots, \xi_N$ and replace the constraints with $y_n x_n^\top w \geq 1 - \xi_n$

Soft SVM:

$$\min_{w,\xi} \frac{\lambda}{2}\|w\|^2 + \frac{1}{N}\sum_{n=1}^{N} \xi_n$$

s.t. $\forall n, y_n x_n^\top w \geq 1 - \xi_n$ and $\xi_n \geq 0$

which is equivalent to

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

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

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

<u>Proof:</u> Fix $w$ and consider the minimization over $\xi$:

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

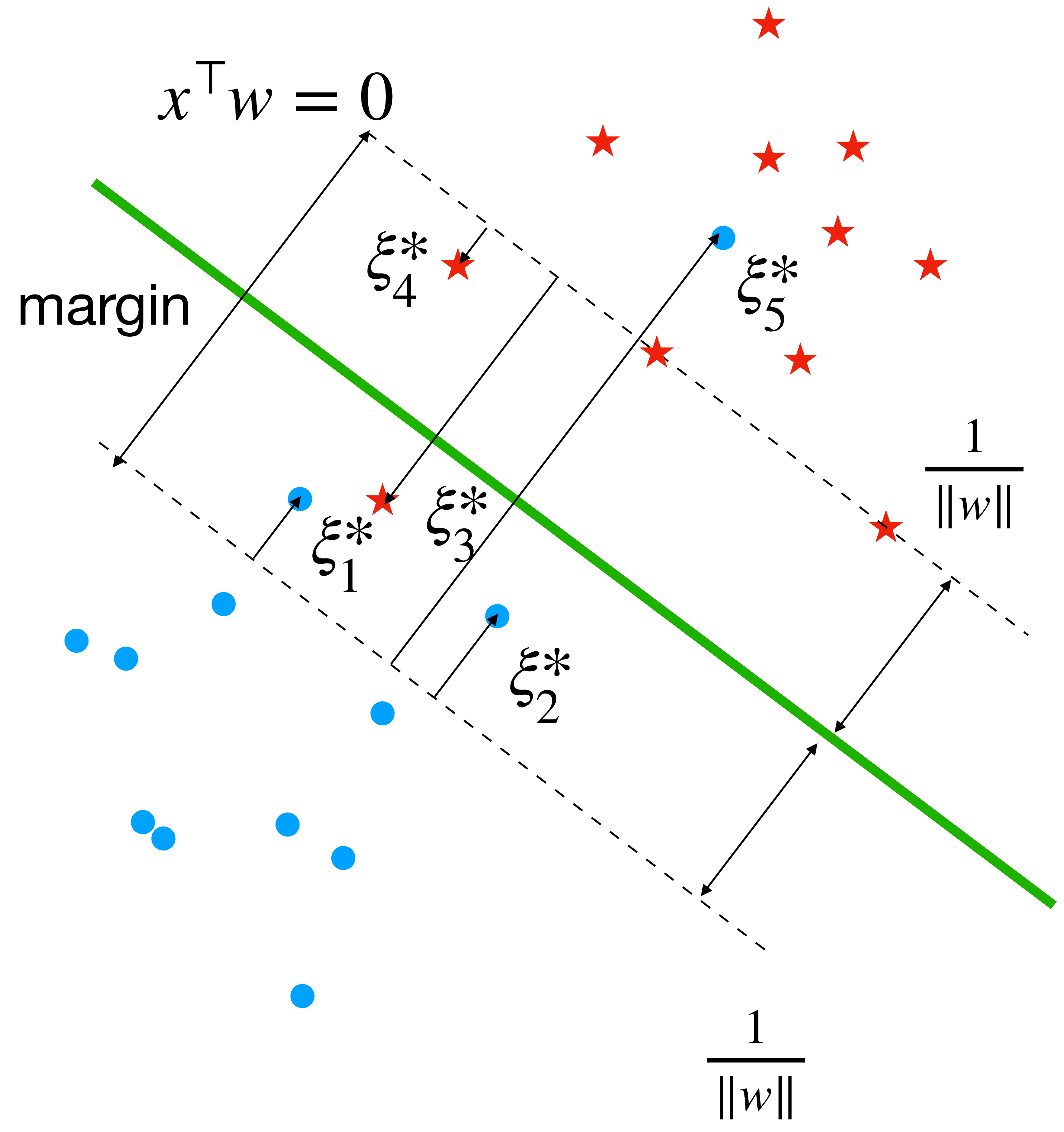
Therefore $\xi_n = [1 - y_n x_n^\top w]_+$

raints to

d

$$\min_{w,\xi} \frac{\lambda}{2}\|w\|^2 + \frac{1}{N}\sum_{n=1}^{N} \xi_n$$

$$\text{s.t. } \forall n, y_n x_n^\top w \geq 1 - \xi_n \quad \text{and} \quad \xi_n \geq 0$$




which is equivalent to

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

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

# Classification by risk minimization

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

Goal: Find a classifier $f : \mathscr{X} \to \mathscr{Y}$ that minimizes the true risk

$$L(f) = \mathbb{E}_{\mathscr{D}}(1_{Y \neq f(X)})$$

How: Through Empirical Risk Minimization (ERM):

$$\min_{w} L_{\text{train}}(w) = \frac{1}{N} \sum_{n=1}^{N} \phi(y_n w^\top x_n)$$

$\phi$ represents the loss function of the functional margin $y_n x_n^\top w$

$\phi$ also serves as a convex surrogate for the 0-1 loss

# Losses for Classification

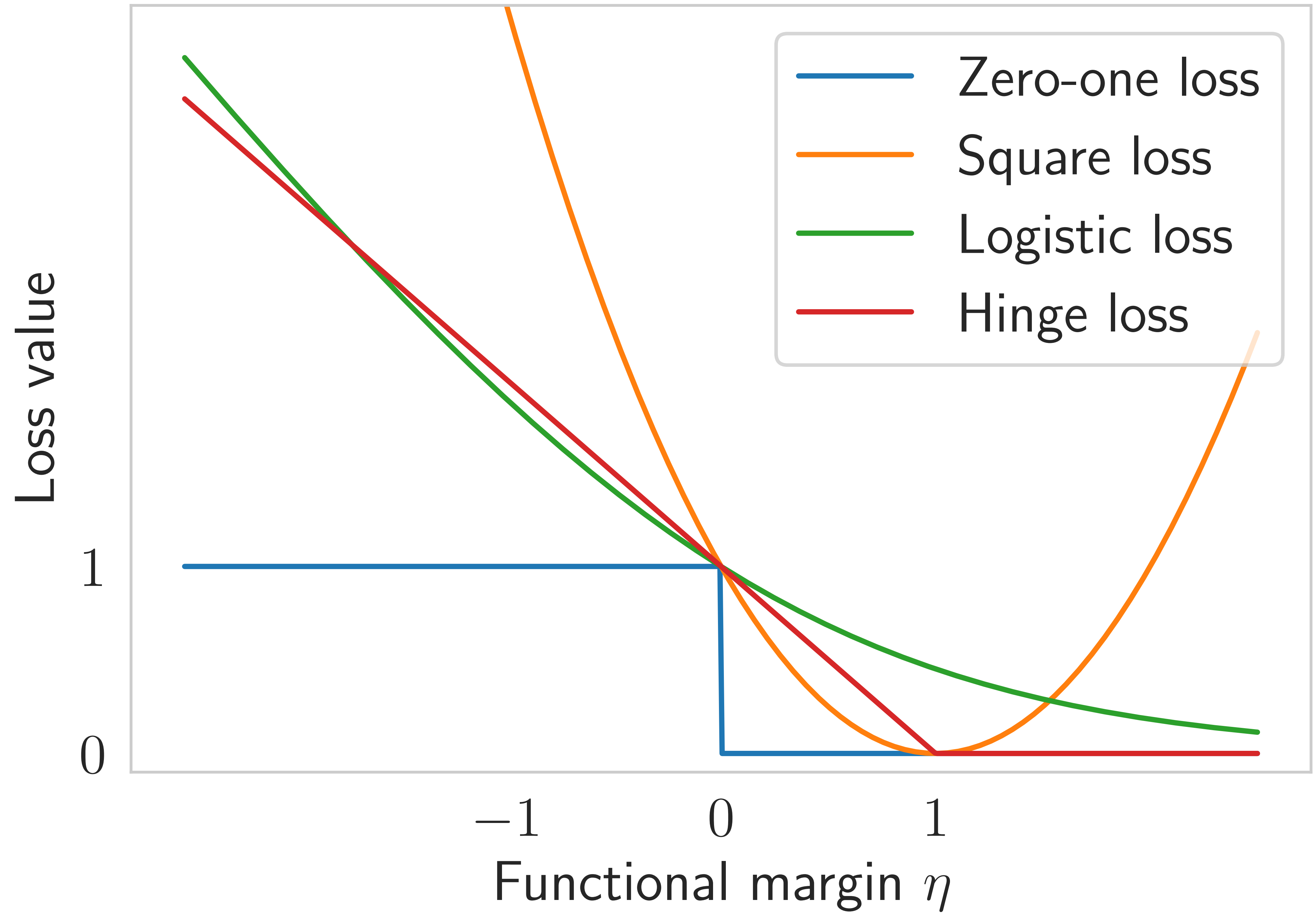
Examples of margin-based losses ($\eta = yx^\top w$):

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

Common features: these losses are convex and provide an upper bound for the zero-one loss

Behavioral differences:

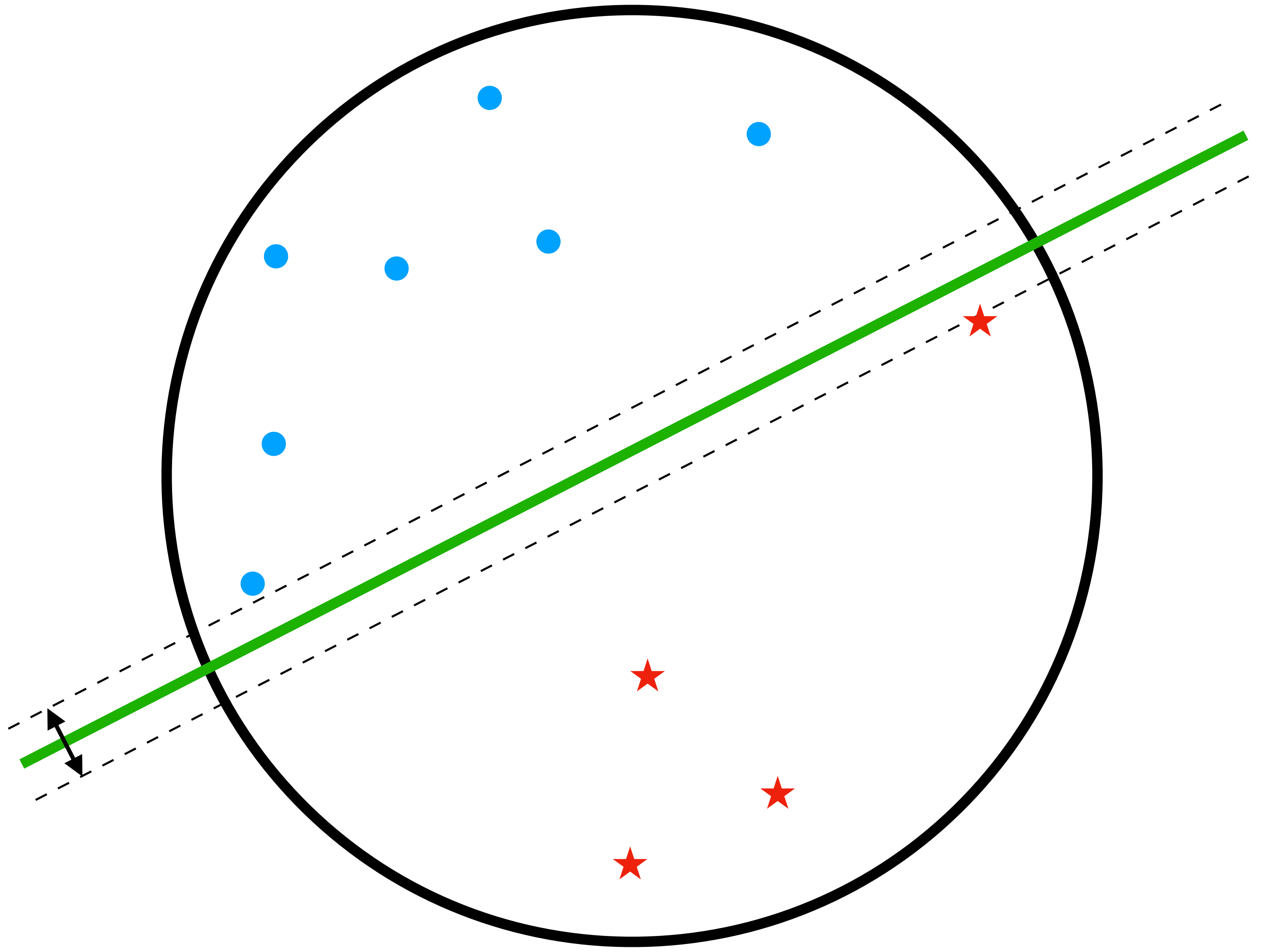
- **MSE:** Penalizes any deviation from 1
- **Logistic Loss:** Asymmetric cost – a penalty is always incurred.
- **Hinge Loss:** A penalty is applied if the prediction is incorrect or lacks confidence



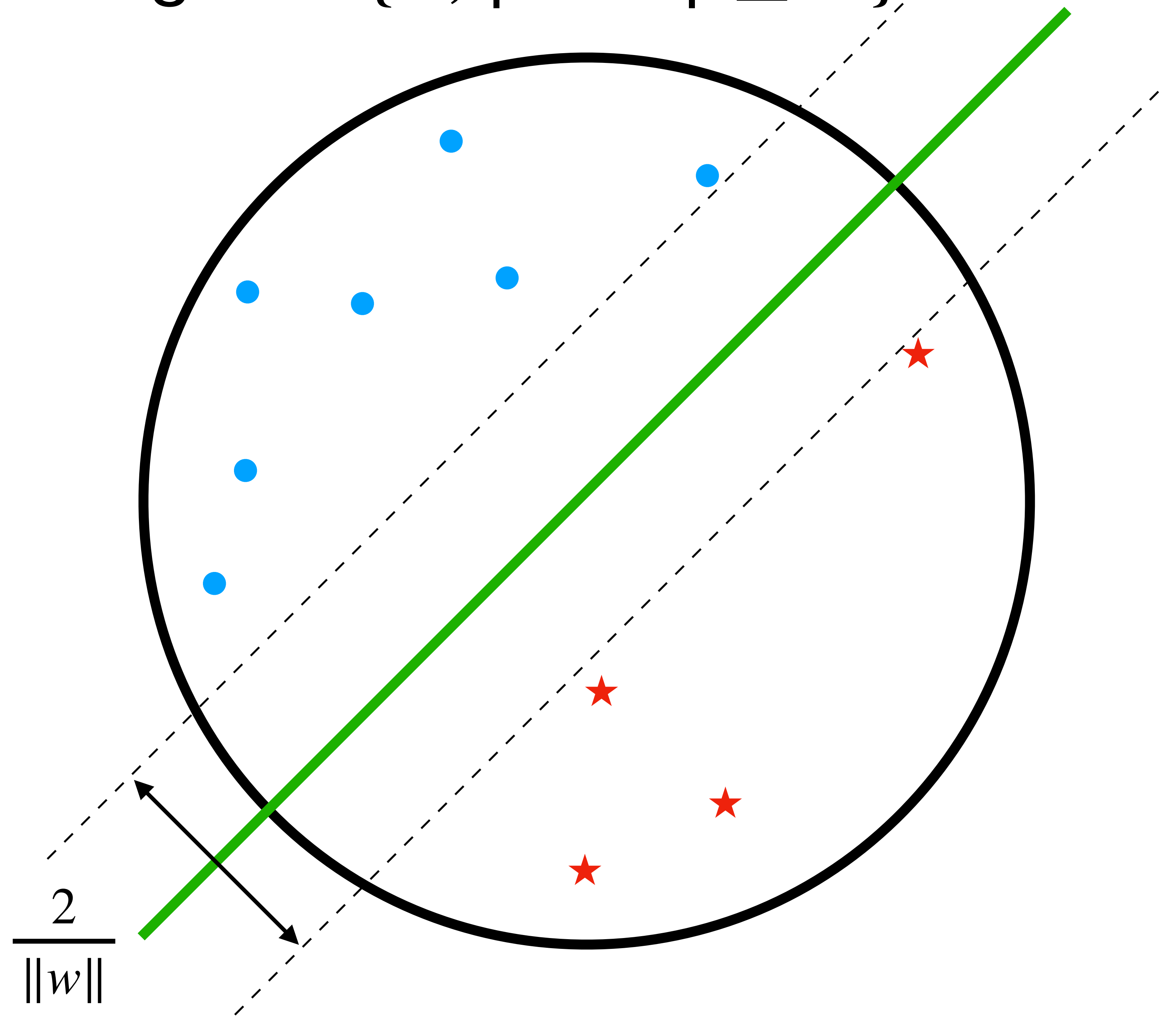
# Summary

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

**ERM for the hinge loss with ridge regularization**



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




Interpretation for separable data with small $\lambda$:

1. Choose the direction of w such that $w^\perp$ acts as a separating hyperplane
2. Adjust the scale of w to ensure that no point lies with the margin
3. Select the hyperplane with the largest margin

# Optimization: How to get $w$?

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

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: $\min\limits_{w} \max\limits_{\alpha} G(w, \alpha)$

Dual problem: $\max\limits_{\alpha} \min\limits_{w} G(w, \alpha)$

➡ Sometimes, the dual problem is easier to solve than the primal problem.

Questions:

1. How do we identify a suitable $G(w, \alpha)$?
2. Under what conditions can the min and max be interchanged?
3. When is the dual problem more tractable than the primal problem?

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

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

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:

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

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

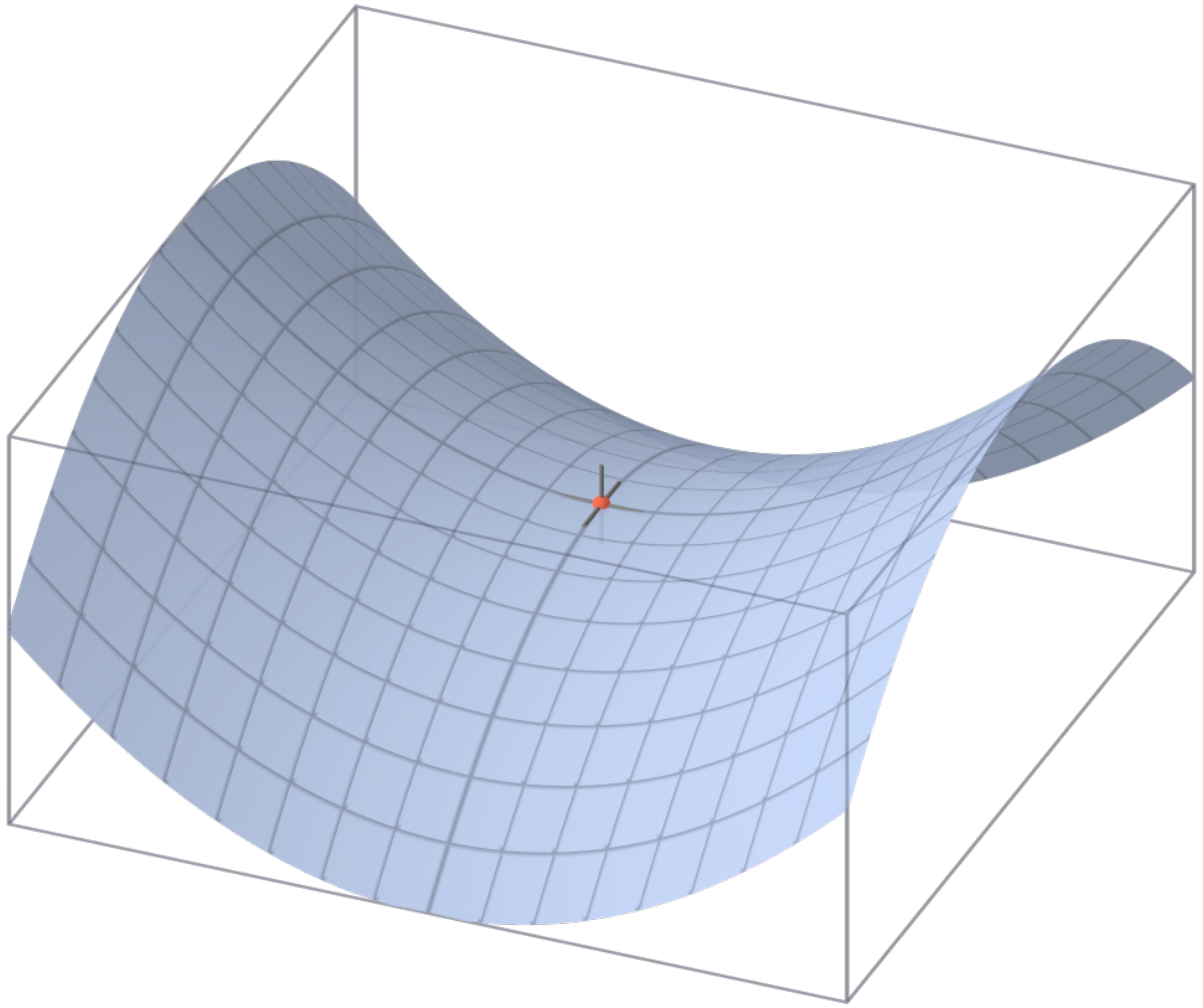
# Q2: Can the min and max be interchanged?

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:

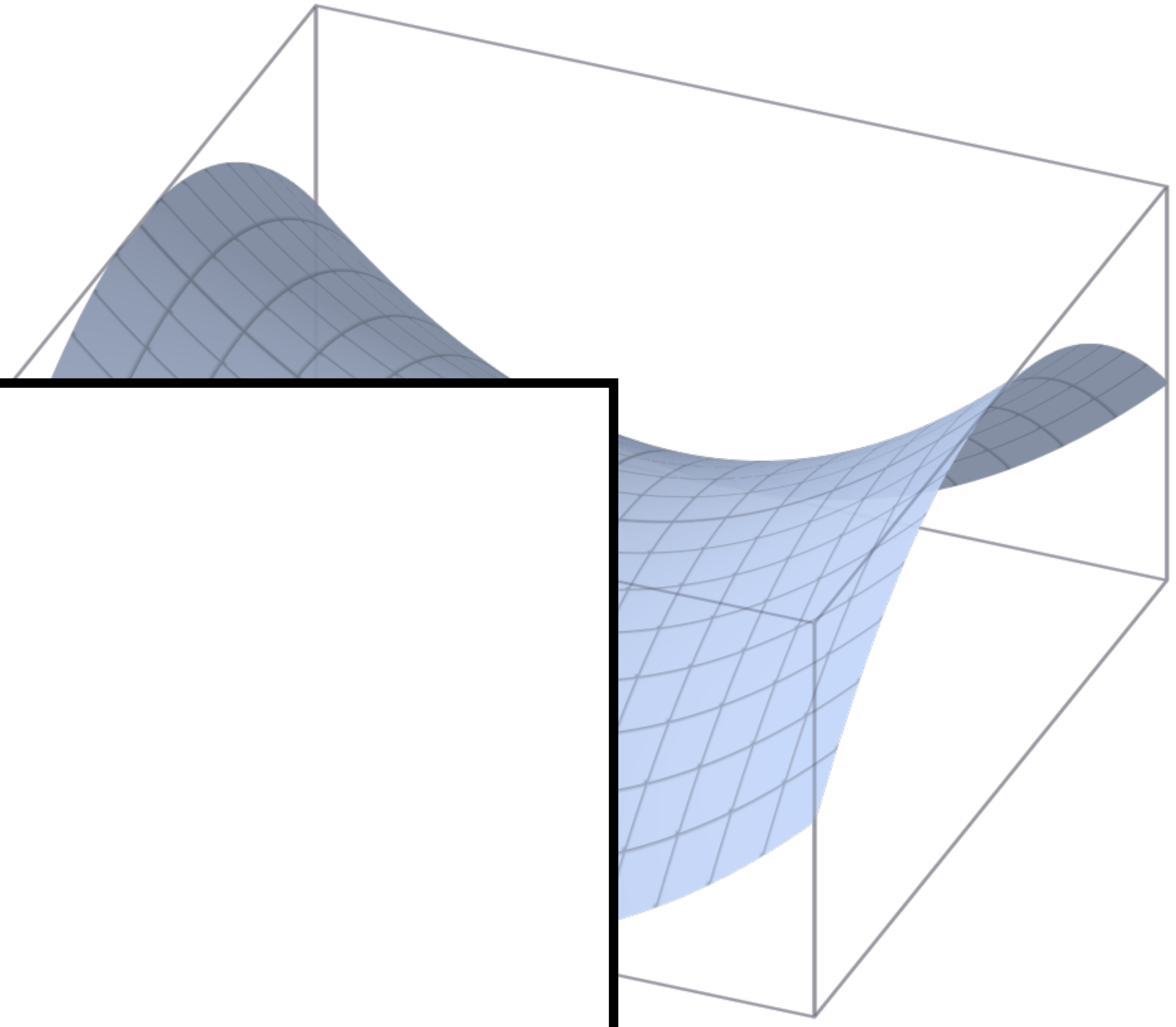
$$\max_{\alpha} \min_{w} G(w, \alpha) = \min_{w} \max_{\alpha} G(w, \alpha)$$



# Q2: Can the min and max be interchanged?

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 met, allowing us to interchange min and max:

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

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

<u>Minimizer computation:</u>

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

<u>Dual optimization problem:</u>

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

# Q3: Why?

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

1. **Differentiable Concave Problem:** Efficient solutions can be achieved using
   - Quadratic programming solvers
   - Coordinate ascent
2. **Kernel Matrix Dependency:** The cost function only depends on the data via the *kernel matrix* $K = \mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{N \times N}$ - no dependency on $d$
3. **Dual Formulation Insight:** $\alpha$ is typically sparse and non-zero exclusively for the training examples that are crucial in determining the decision boundary

# Interpretation of the dual formulation

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

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

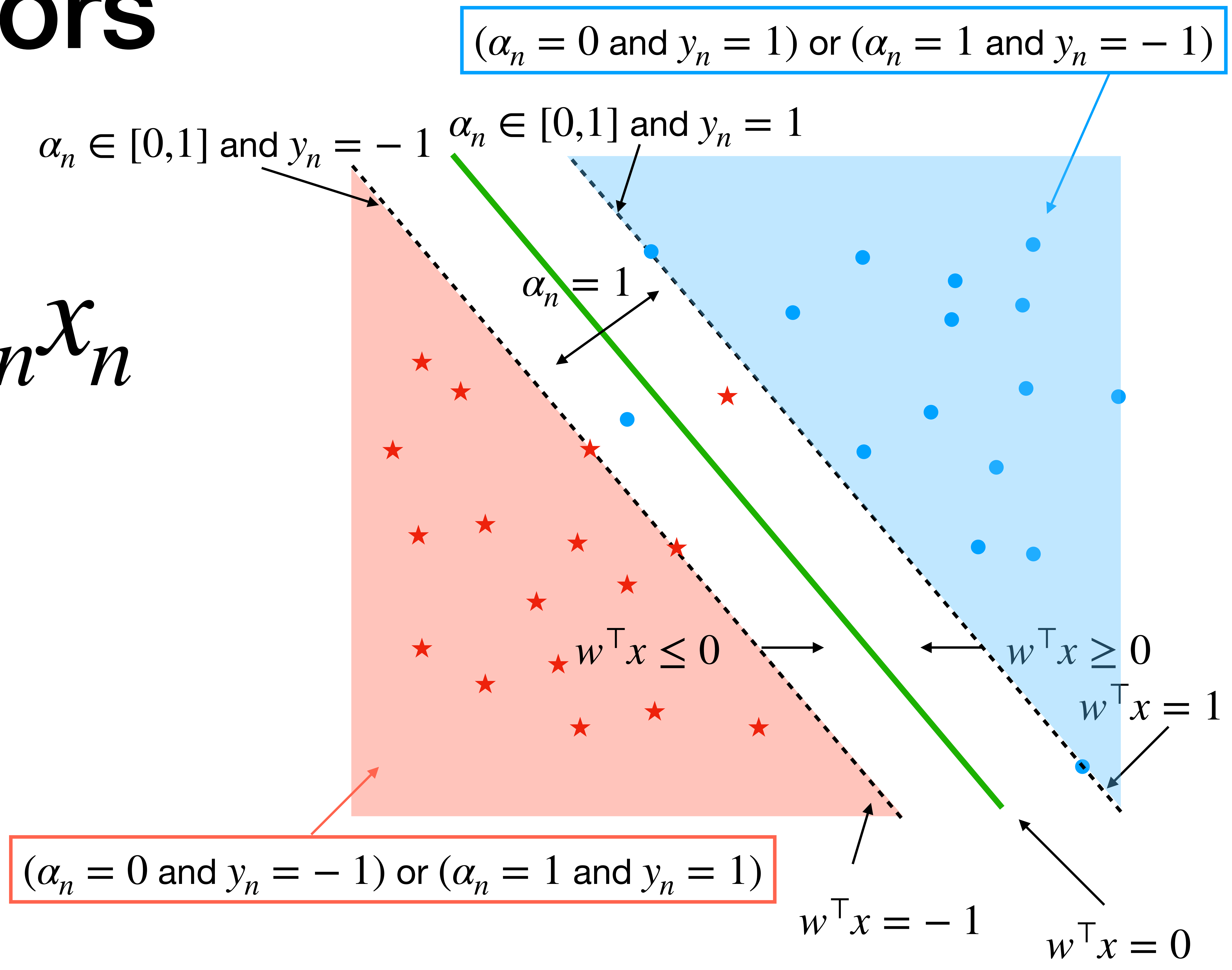
- If $x_n$ is on the correct side and outside the margin, $1 - y_n x_n^\top w < 0$, then $\alpha_n = 0$
- If $x_n$ is on the correct side and on the margin, $1 - y_n x_n^\top w = 0$, then $\alpha_n \in [0,1]$
- If $x_n$ is strictly inside the margin or or the incorrect side, $1 - y_n x_n^\top w > 0$, then $\alpha_n = 1$

➡The points for which $\alpha_n > 0$ are referred to as support vectors

# 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 $\alpha_n = 0$



# Recap

- Hard SVM - finds max-margin separating hyperplane

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

- Soft SVM - relax the constraint for non-separable data

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

  - Hinge loss can be optimized with (stochastic) sub-gradient method

- Duality: min max problem is equivalent to max min (convex-concave objective)
  - Efficient solutions with quadratic programming and coordinate ascent
  - The cost depends on the data via the *kernel matrix* (no dependency on $d$)