6. Support Vector Machines

EEC454 Introduction to Machine Learning Systems

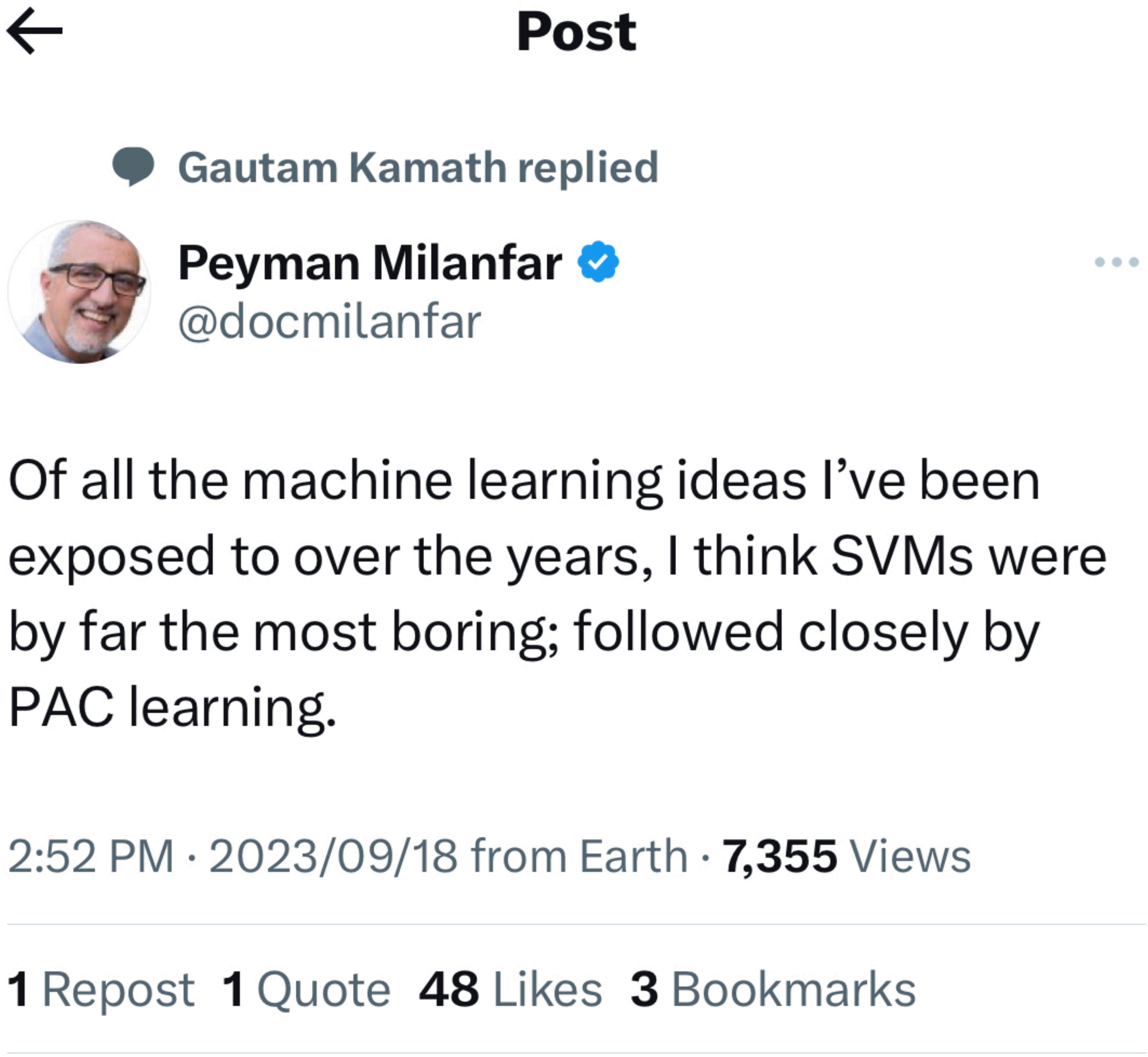
2023 Fall, Jaeho Lee

Additional Materials

- Maximum Likelihood Estimate (MLE) & Maximum A Posteriori (MAP)
  (for the last class)
  https://drive.google.com/file/d/1iRh9aBHeDSafr0hHnKs06W9b0xtlzNNJ/view

- Optimization Basics + Linear Algebra
  (for today!)
  https://alex.smola.org/teaching/10-701-2015/slides/5_Math_and_Optimization.pdf

- Probability
  https://alex.smola.org/teaching/10-701-2015/slides/2_Statistics.pdf



Post

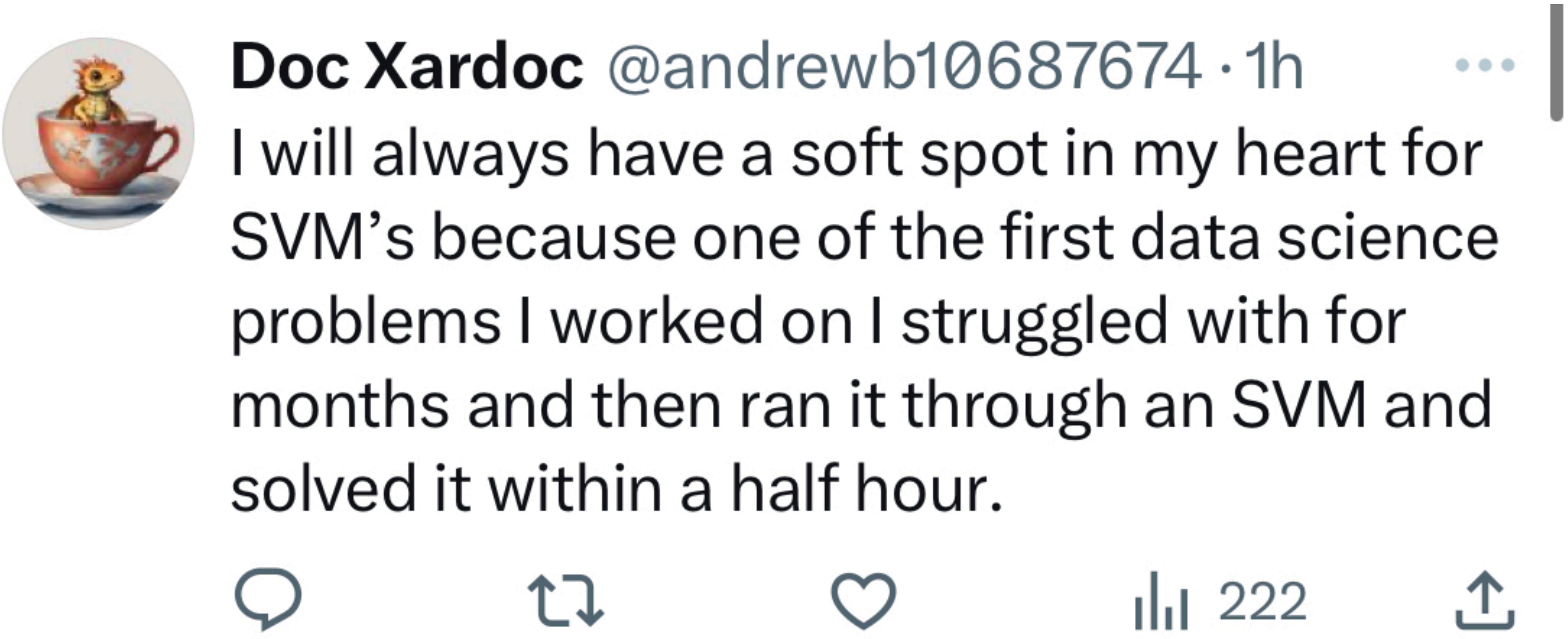
Gautam Kamath replied

Peyman Milanfar
@docmilanfar

Of all the machine learning ideas I've been exposed to over the years, I think SVMs were by far the most boring; followed closely by PAC learning.

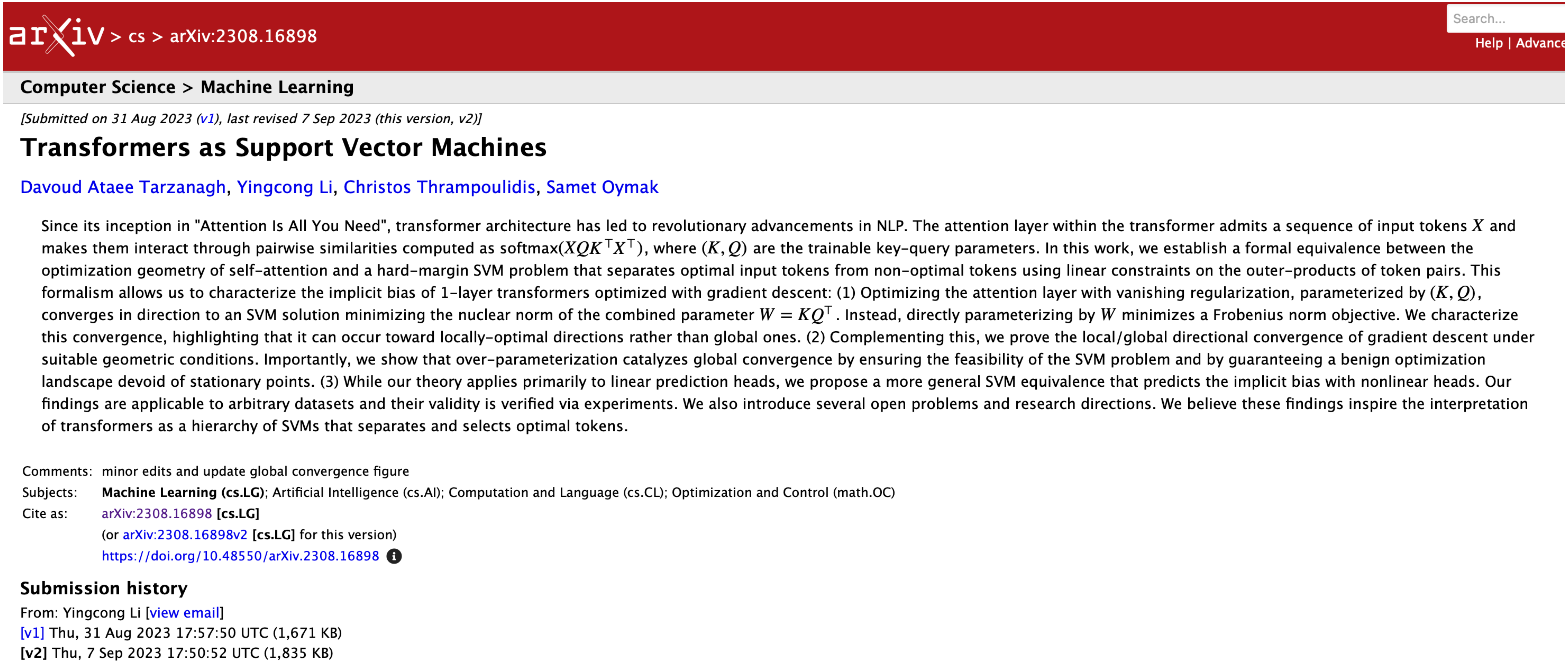
2:52 PM · 2023/09/18 from Earth · **7,355** Views

1 Repost **1** Quote **48** Likes **3** Bookmarks



Doc Xardoc @andrewb10687674 · 1h

I will always have a soft spot in my heart for SVM's because one of the first data science problems I worked on I struggled with for months and then ran it through an SVM and solved it within a half hour.



Computer Science > Machine Learning

[Submitted on 31 Aug 2023 (v1), last revised 7 Sep 2023 (this version, v2)]

Transformers as Support Vector Machines

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, Samet Oymak

Since its inception in "Attention Is All You Need", transformer architecture has led to revolutionary advancements in NLP. The attention layer within the transformer admits a sequence of input tokens X and makes them interact through pairwise similarities computed as softmax$(XQK^\top X^\top)$, where (K, Q) are the trainable key-query parameters. In this work, we establish a formal equivalence between the optimization geometry of self-attention and a hard-margin SVM problem that separates optimal input tokens from non-optimal tokens using linear constraints on the outer-products of token pairs. This formalism allows us to characterize the implicit bias of 1-layer transformers optimized with gradient descent: (1) Optimizing the attention layer with vanishing regularization, parameterized by (K, Q), converges in direction to an SVM solution minimizing the nuclear norm of the combined parameter $W = KQ^\top$. Instead, directly parameterizing by W minimizes a Frobenius norm objective. We characterize this convergence, highlighting that it can occur toward locally-optimal directions rather than global ones. (2) Complementing this, we prove the local/global directional convergence of gradient descent under suitable geometric conditions. Importantly, we show that over-parameterization catalyzes global convergence by ensuring the feasibility of the SVM problem and by guaranteeing a benign optimization landscape devoid of stationary points. (3) While our theory applies primarily to linear prediction heads, we propose a more general SVM equivalence that predicts the implicit bias with nonlinear heads. Our findings are applicable to arbitrary datasets and their validity is verified via experiments. We also introduce several open problems and research directions. We believe these findings inspire the interpretation of transformers as a hierarchy of SVMs that separates and selects optimal tokens.

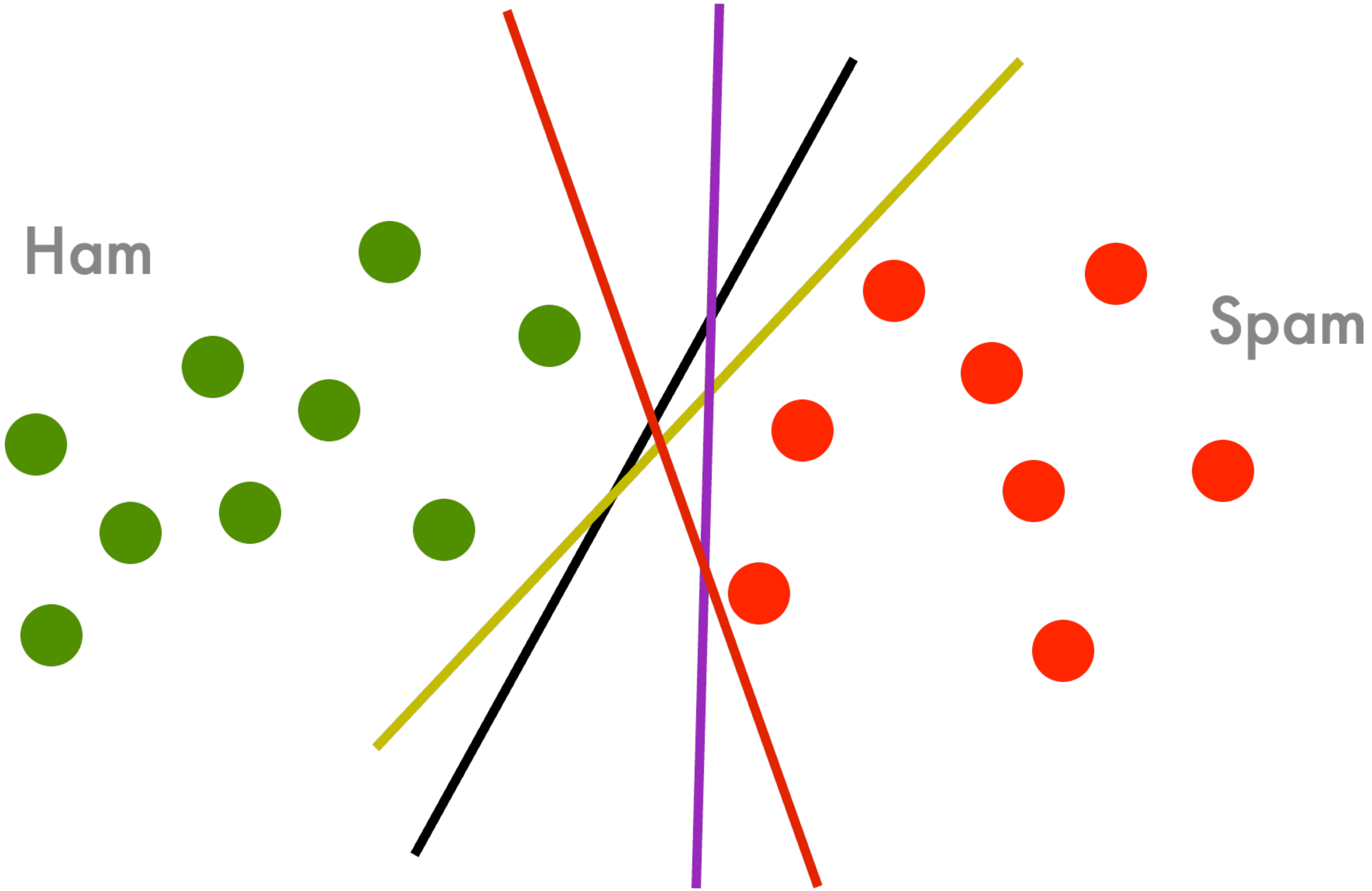
Comments:	minor edits and update global convergence figure
Subjects:	**Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI); Computation and Language (cs.CL); Optimization and Control (math.OC)
Cite as:	arXiv:2308.16898 **[cs.LG]**
	(or arXiv:2308.16898v2 **[cs.LG]** for this version)
	https://doi.org/10.48550/arXiv.2308.16898

Submission history

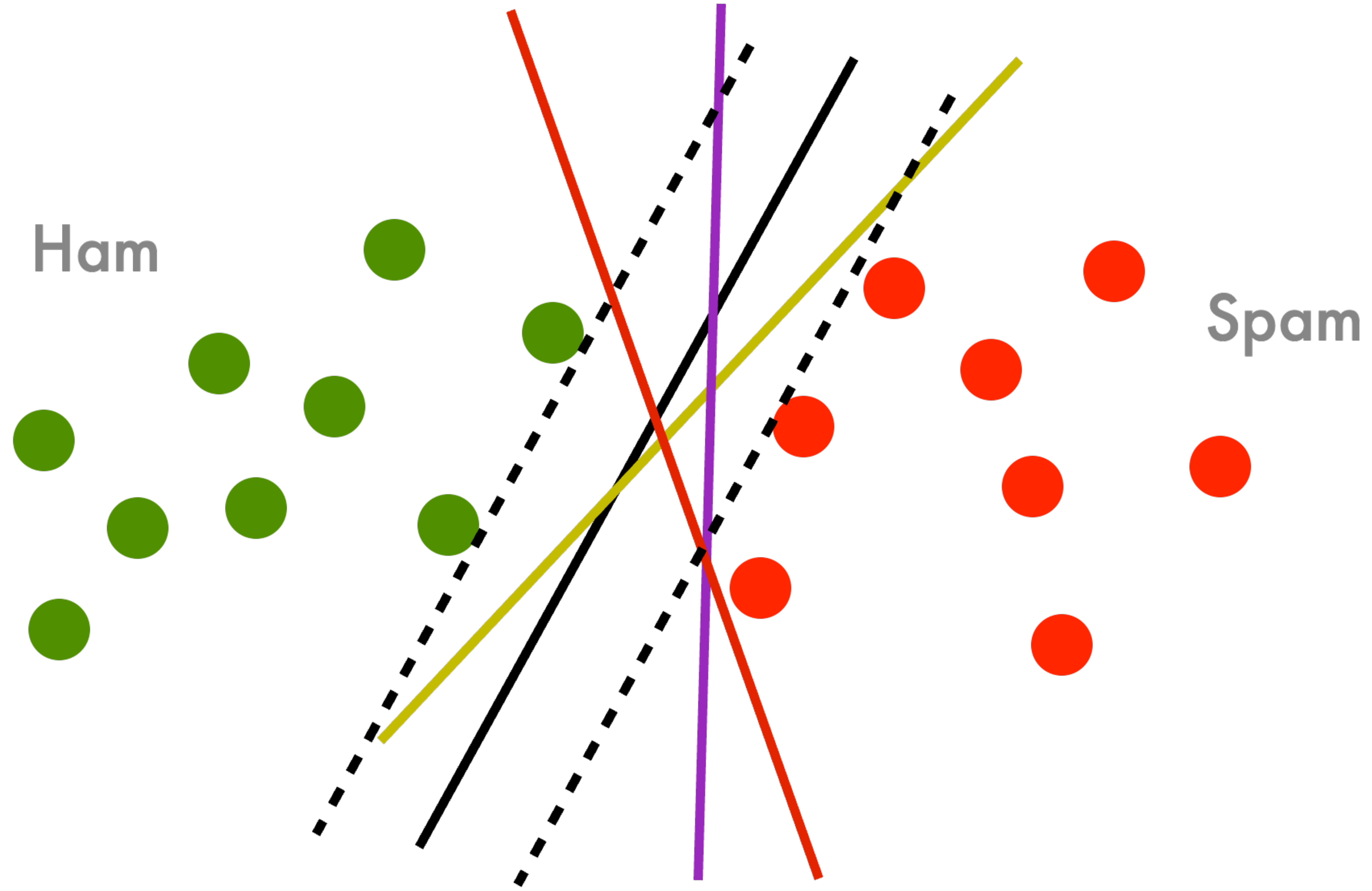
From: Yingcong Li [view email]
[v1] Thu, 31 Aug 2023 17:57:50 UTC (1,671 KB)
[v2] Thu, 7 Sep 2023 17:50:52 UTC (1,835 KB)

Max Margin Classifiers & Hard SVM

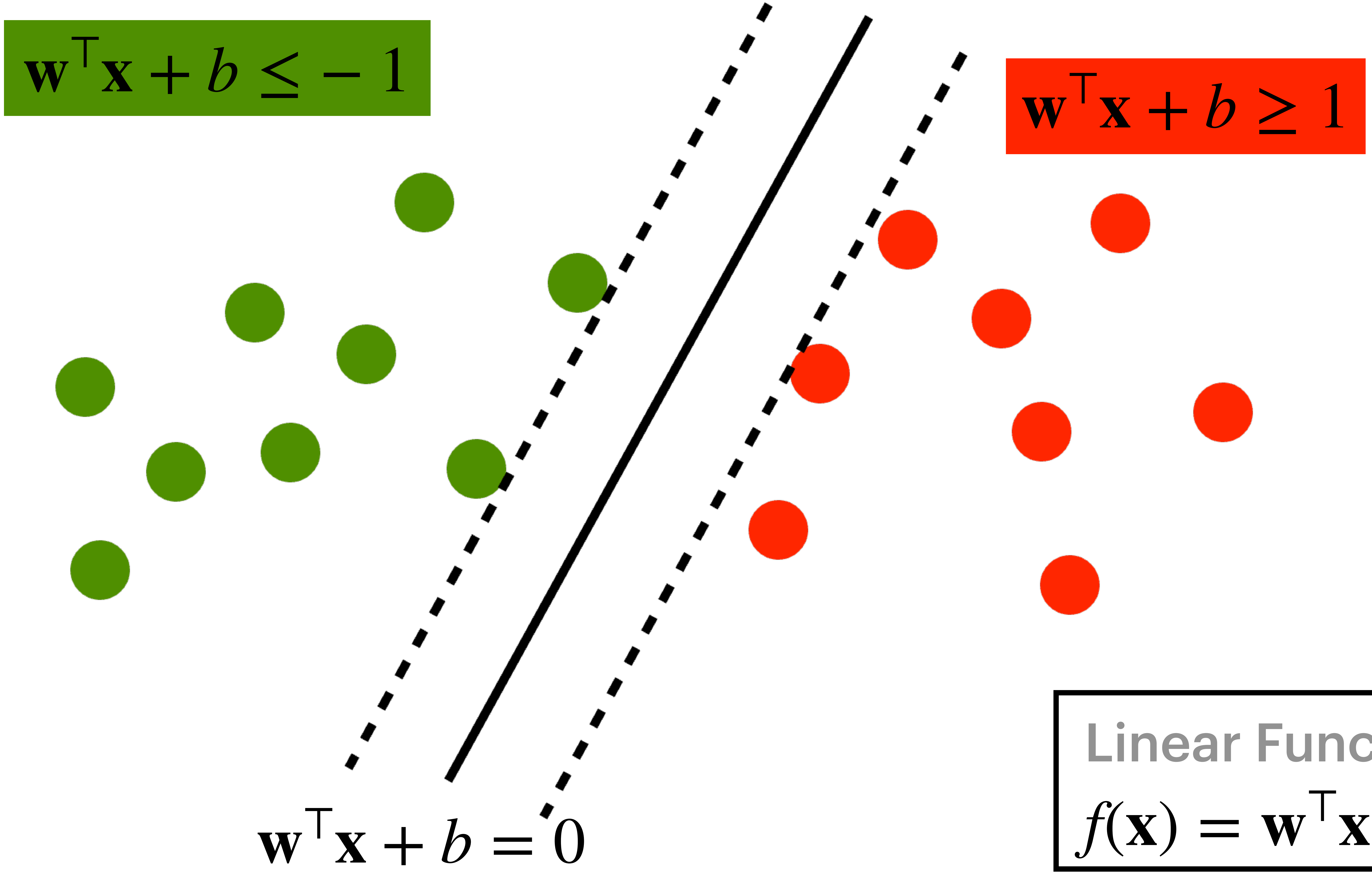
Linearly Separable Data



Linear Separators



Large Margin Classifier



$\mathbf{w}^\top \mathbf{x} + b \leq -1$

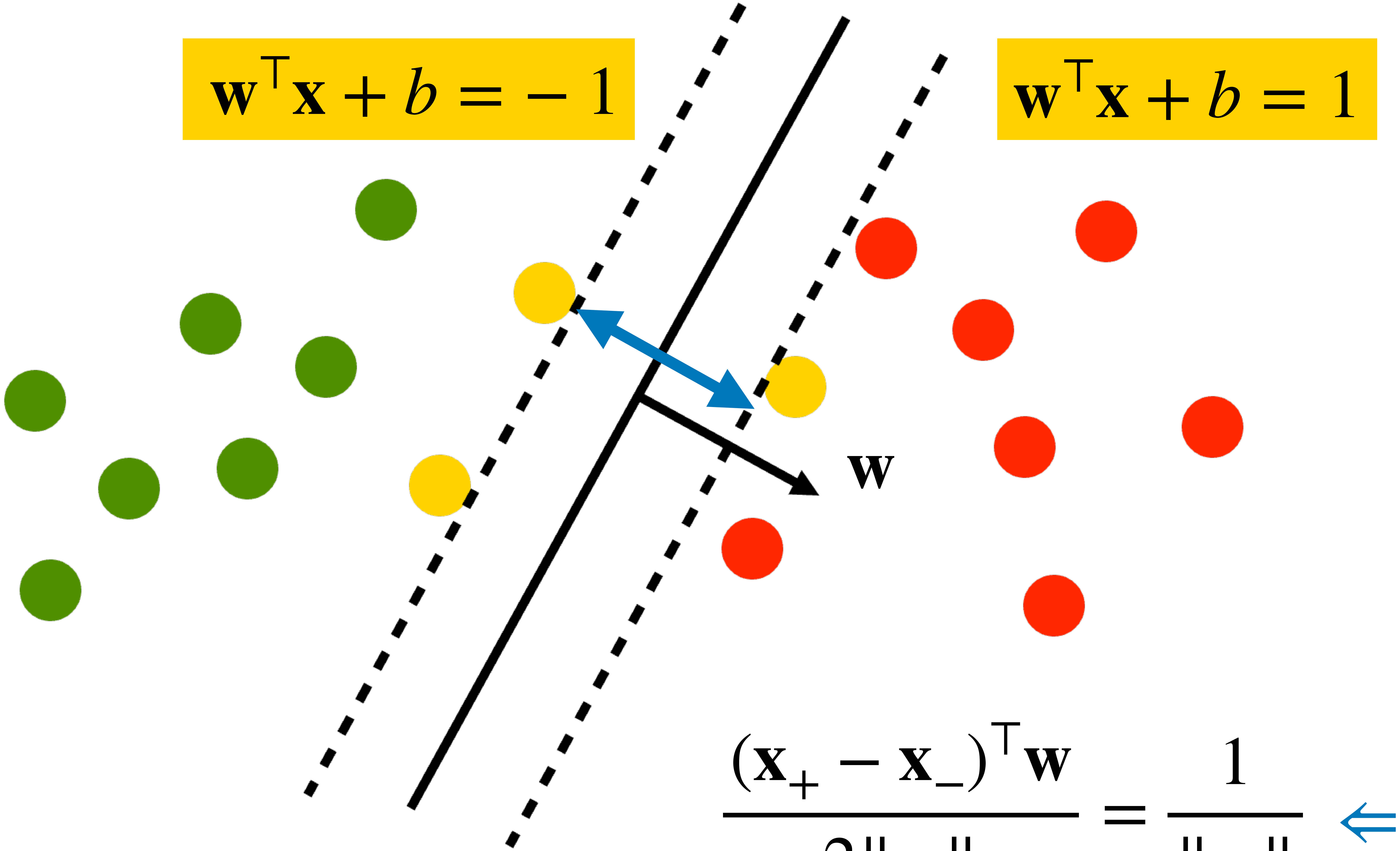
$\mathbf{w}^\top \mathbf{x} + b \geq 1$

$\mathbf{w}^\top \mathbf{x} + b = 0$

Linear Function

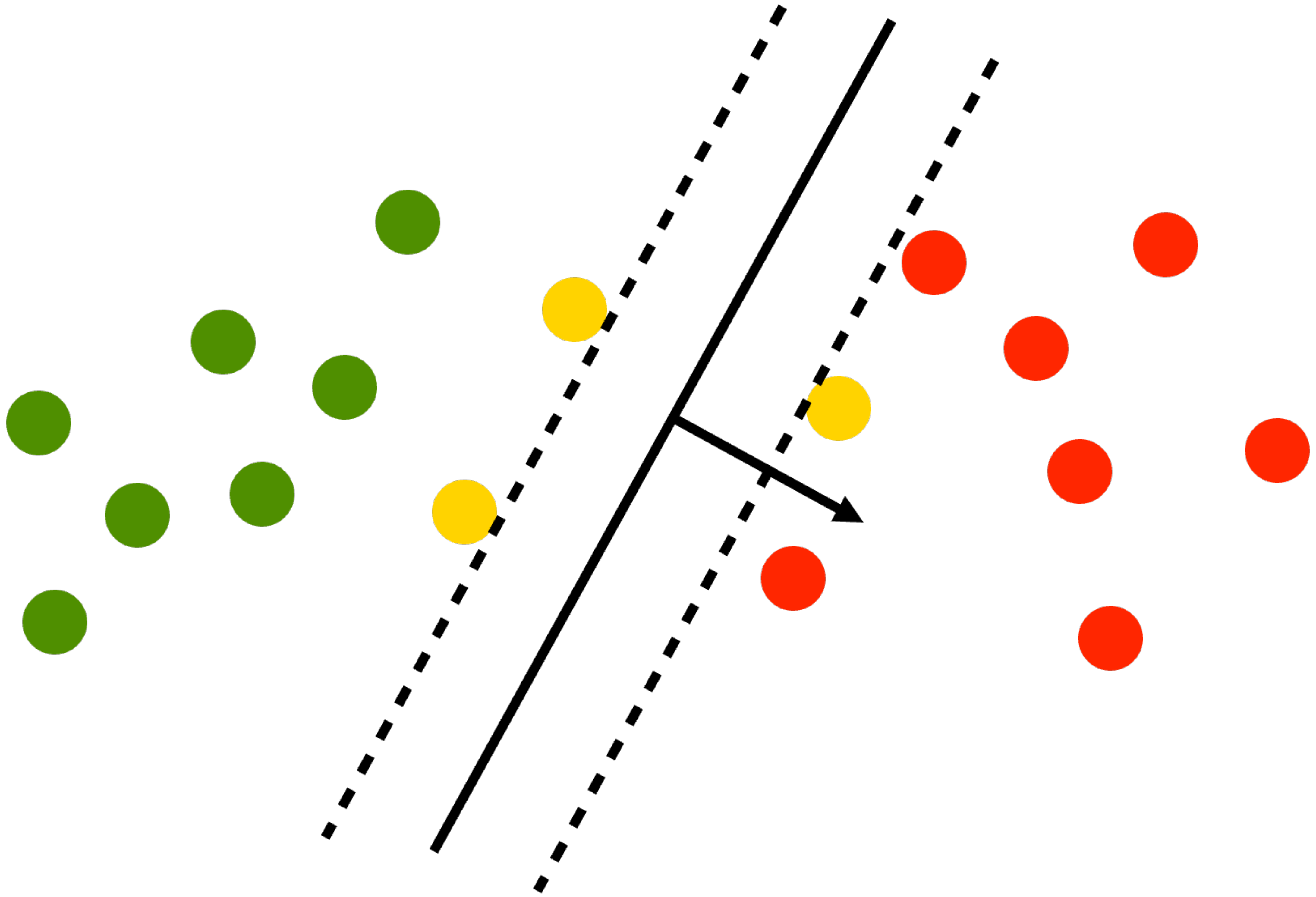
$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$

Large Margin Classifier


$\mathbf{w}^\top \mathbf{x} + b = -1$
$\mathbf{w}^\top \mathbf{x} + b = 1$
$\mathbf{w}$

$$\frac{(\mathbf{x}_+ - \mathbf{x}_-)^\top \mathbf{w}}{2\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|} \Leftarrow \text{“Margin”}$$

Max Margin Classifier



- We solve the problem:

$$\text{maximize}_{\mathbf{w},b} \frac{1}{||\mathbf{w}||} \qquad \text{subject to} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$$

(we are using $y_i \in \{-1, +1\}$, instead of $\{0,1\}$)

Solving the Optimization: Dual Problem

- Slightly re-phrased, we are solving

$$\ell^* = \min_{\mathbf{w},b} \frac{\|\mathbf{w}\|^2}{2} \qquad \text{subject to} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$$

  - Difficult to solve, due to the constraint.

- **Solution.** We consider the ***Lagrangian dual***.

  (the original problem is called "primal")

$$\mathscr{L}(\mathbf{w}, b, \alpha) = \frac{\|\mathbf{w}\|^2}{2} + \sum_{i=1}^{n} \alpha_i \Big(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\Big)$$

Solving the Optimization: Dual Problem

$$\mathscr{L}(\mathbf{w}, b, \alpha) = \frac{\|\mathbf{w}\|^2}{2} + \sum_{i=1}^{n} \alpha_i \big(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\big)$$

- We know that

$$\ell^* = \min_{\mathbf{w},b} \max_{\alpha \geq 0} \mathscr{L}(\mathbf{w}, b, \alpha)$$

- Wait, but why?
  (see also — the additional materials on “optimization basics”)

Primal: $\ell^* = \min_{\mathbf{w},b} \frac{\|\mathbf{w}\|^2}{2}$ subject to $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$

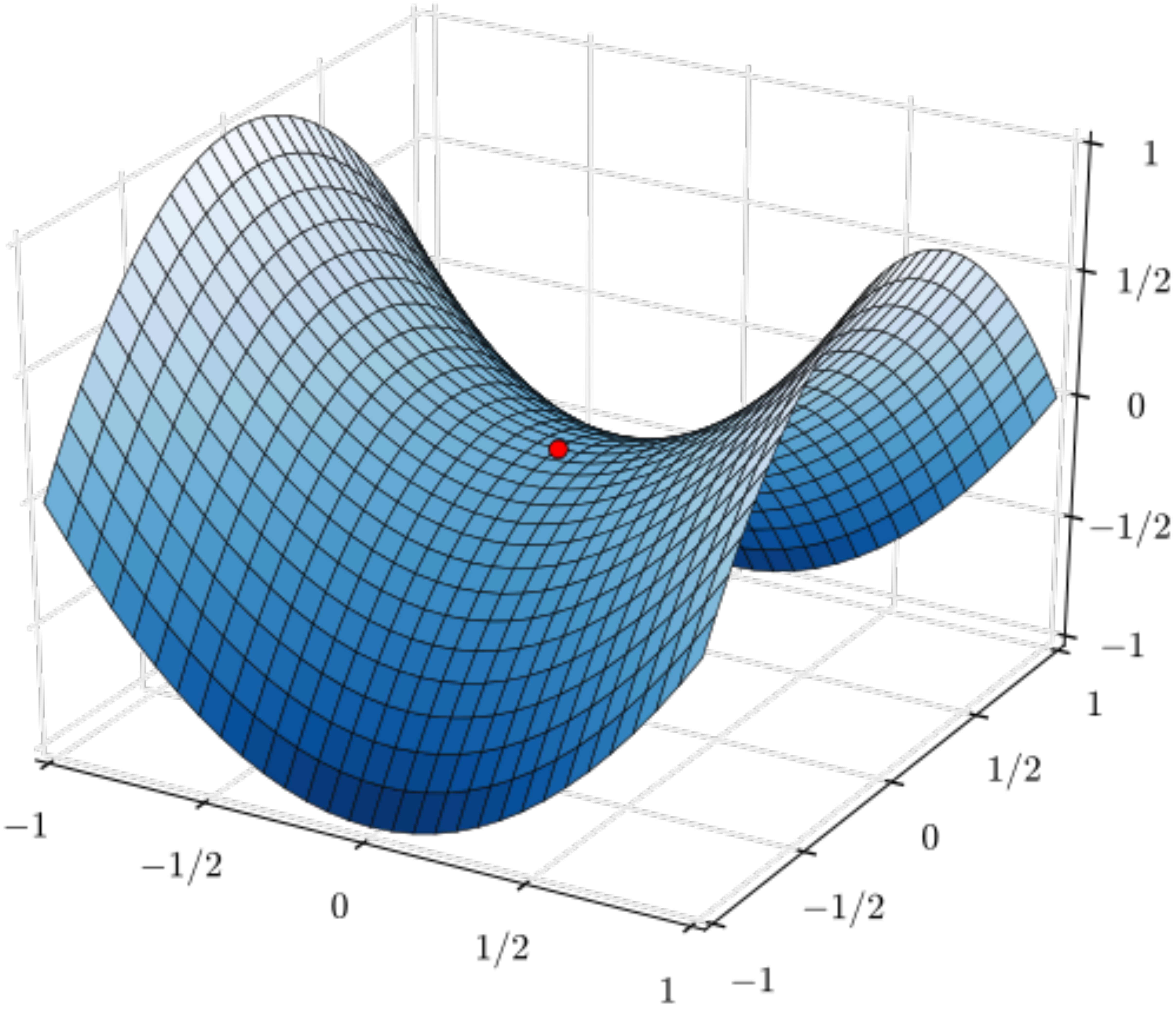
Dual: $\ell^* = \min_{\mathbf{w},b} \max_{\alpha \geq 0} \frac{\|\mathbf{w}\|^2}{2} + \sum_{i=1}^{n} \alpha_i \left(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\right)$

- Adversary will gauge the quantity $1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)$
  - If > 0 $\cdots$ $\alpha_i \to \infty$ infeasible for primal, ∞ for dual
  - If < 0 $\cdots$ $\alpha_i = 0$ primal = dual
  - If $= 0$ $\cdots$ any constant... primal = dual

⇒ Find the saddle point!

(see also — the additional materials)

Minimax Problems



Find the saddle point!

(which is the critical point)

Solving the Optimization: Dual Problem

$$\nabla_{\mathbf{w}} \mathscr{L} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \qquad \nabla_b \mathscr{L} = - \sum_{i=1}^{n} \alpha_i y_i$$

- We need these be zero at saddle point, i.e.,

$$\mathbf{w}^* = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \qquad 0 = \sum_{i=1}^{n} \alpha_i y_i$$

- Plugging $\mathbf{w}^*$ back into Lagrangian, we get:

$$\mathscr{L} = -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j + \sum_{i=1}^{n} \alpha_i$$

Solving the Optimization: Dual Problem

- Summing up, we are solving:

$$\max_{\alpha}\left(-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j + \sum_{i=1}^{n}\alpha_i\right)$$

$$\text{subject to} \quad \sum_i \alpha_i y_i = 0$$

$$\alpha_i \geq 0$$

Solving the Optimization: Dual Problem

- Neat form as a quadratic program over a convex polytope.

$$\max_{\alpha} \left(-\frac{1}{2}\alpha^\top \mathbf{Z}\alpha + \mathbf{1}^\top \alpha \right)$$

$$\text{subject to} \quad \alpha^\top \mathbf{y} = 0$$

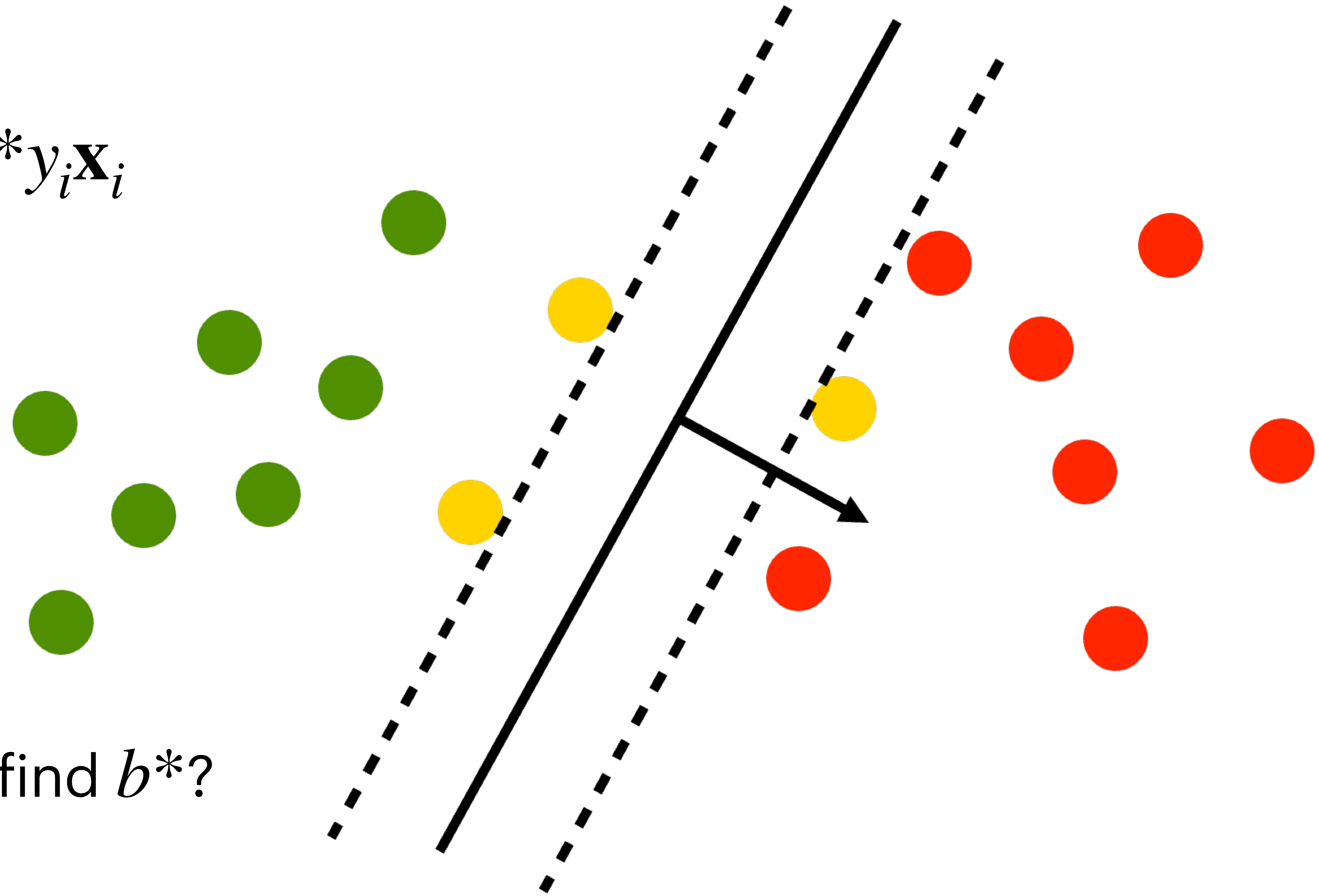
$$\alpha \succeq 0$$

- Solution @ critical point or extreme point (many solvers)

Solving the Optimization: Dual Problem

$$\mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i$$

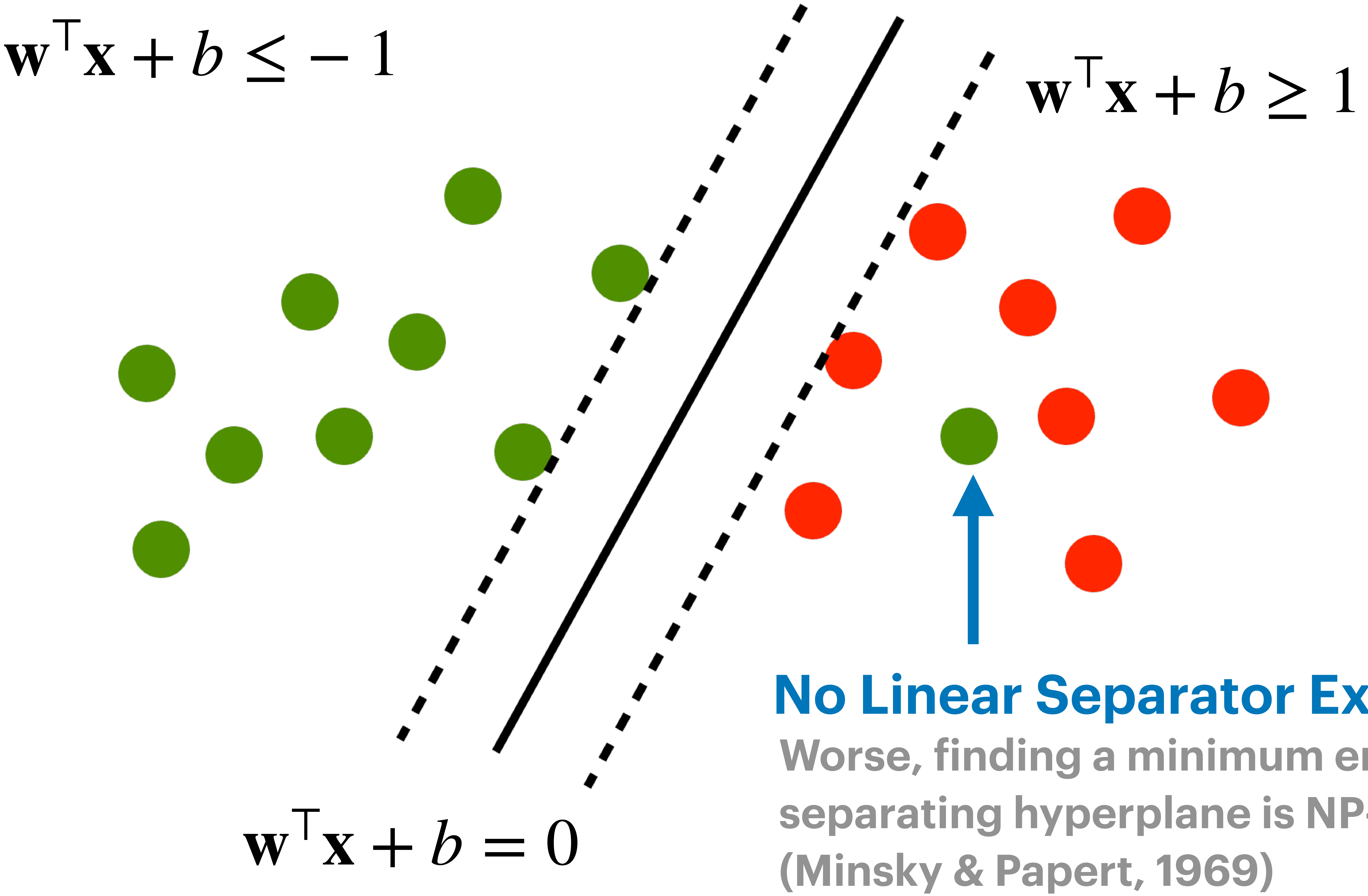
Nonzero, only for points on margin (support vectors)



Quiz. How to find b^*?

Soft(-margin) SVM

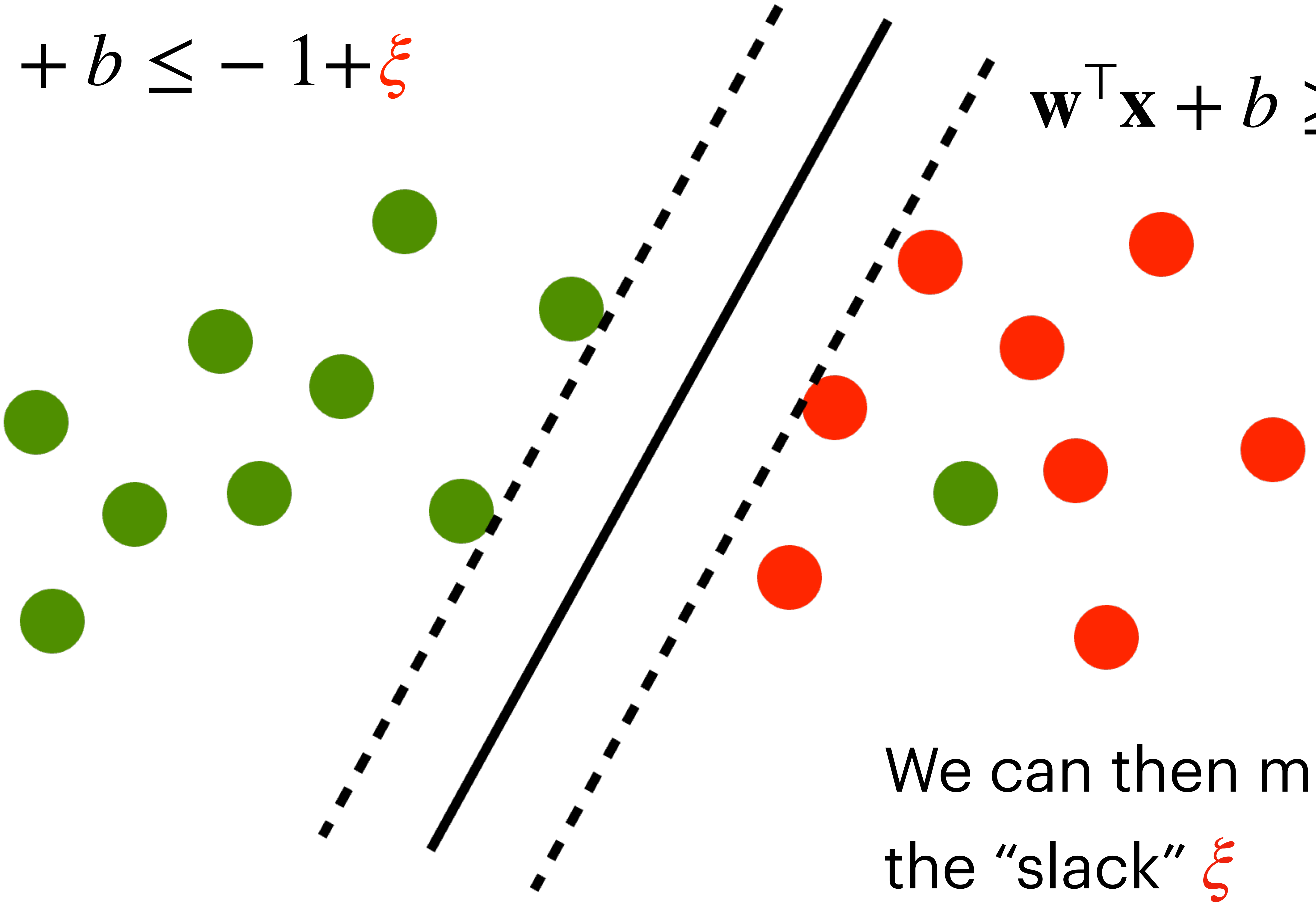
Linear Separators


$\mathbf{w}^\top \mathbf{x} + b \leq -1$
$\mathbf{w}^\top \mathbf{x} + b \geq 1$
$\mathbf{w}^\top \mathbf{x} + b = 0$

No Linear Separator Exists

Worse, finding a minimum error separating hyperplane is NP-hard. (Minsky & Papert, 1969)

Solution: Add Slack Variables



$$\mathbf{w}^\top \mathbf{x} + b \leq -1 + \xi$$

$$\mathbf{w}^\top \mathbf{x} + b \geq 1 - \xi$$

$$\mathbf{w}^\top \mathbf{x} + b = 0$$

We can then minimize the “slack” ξ

Formulation

- We are solving

$$\ell^* = \min_{\mathbf{w},b,\xi} \frac{\|\mathbf{w}\|^2}{2} + C \cdot \sum_i \xi_i$$

$$\text{subject to} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0$$

- We know that the problem is always feasible
  (i.e., constraints can be met, no matter the minimand)
  - Let $\mathbf{w} = \mathbf{0}$, $b = 0$, $\xi_i = 1$.

Dual Formulation

- As a dual, we solve

$$\min_{\mathbf{w},b,\xi} \max_{\alpha,\eta} \left(\frac{||\mathbf{w}||^2}{2} + C\sum_i \xi_i - \sum_i \alpha_i \left(y_i(\mathbf{x}_i^\top \mathbf{w} + b) + \xi_i - 1 \right) - \sum_i \eta_i \xi_i \right)$$

- The optimal $(\mathbf{w}, b, \xi)$ is at the saddle point with (α, η)

- Derivatives for $(\mathbf{w}, b, \xi)$ need to vanish!

Derivatives

$$\nabla_{\mathbf{w}}\mathscr{L} = \mathbf{w} - \sum \alpha_i y_i \mathbf{x}_i \; = 0$$

$$\nabla_b \mathscr{L} = \sum \alpha_i y_i = 0$$

$$\nabla_{\xi_i} \mathscr{L} = C - \alpha_i - \eta_i = 0$$

- Doing the similar thing, we get the Lagrangian

$$-\frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j + \sum_i \alpha_i - \sum_i \alpha_i \xi_i + C \sum_i \xi_i - \sum_i \eta_i \xi_i$$

$$= -\frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j + \sum_i \alpha_i$$

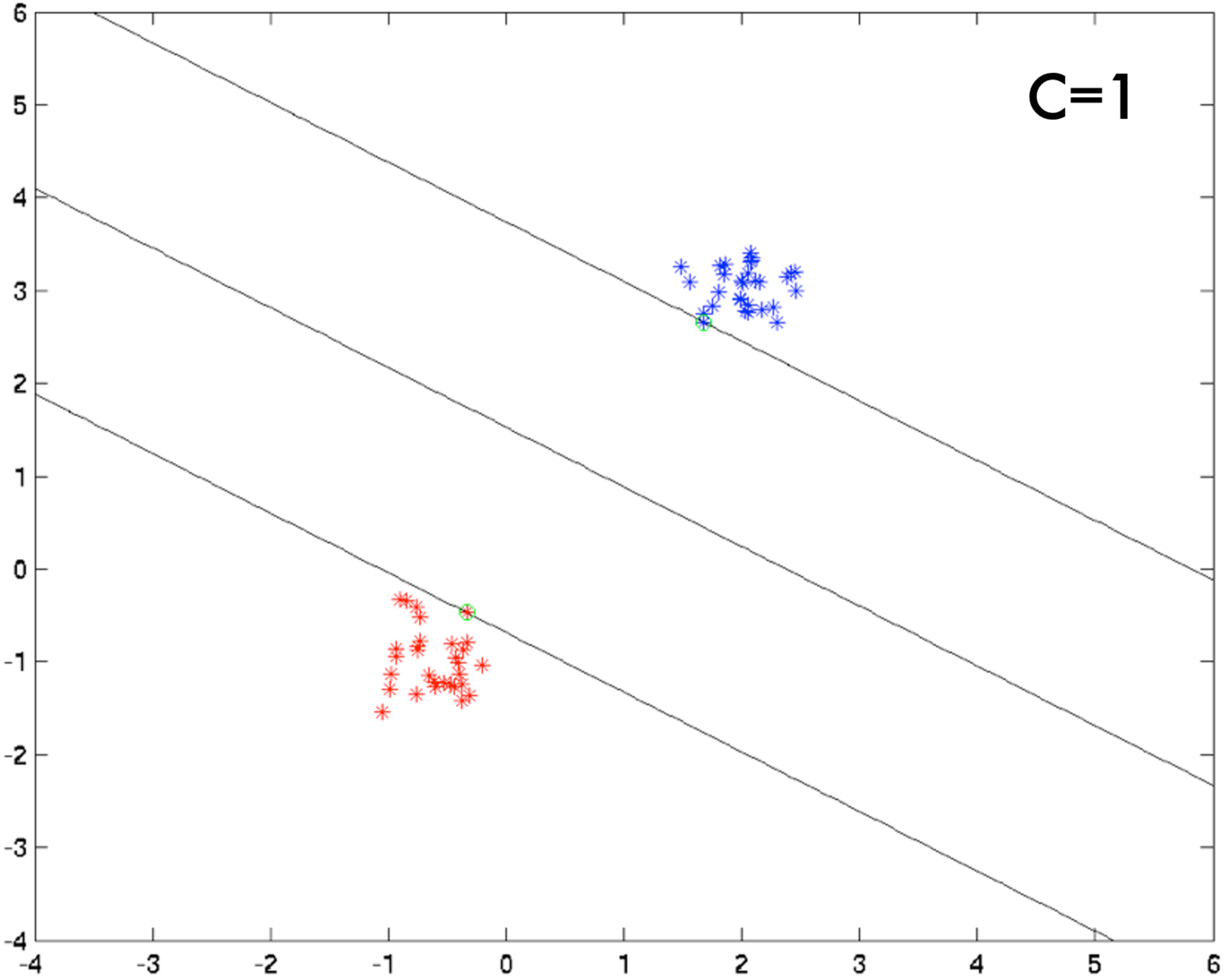
Solving the Optimization: Dual Problem

- Summing up, we are solving:

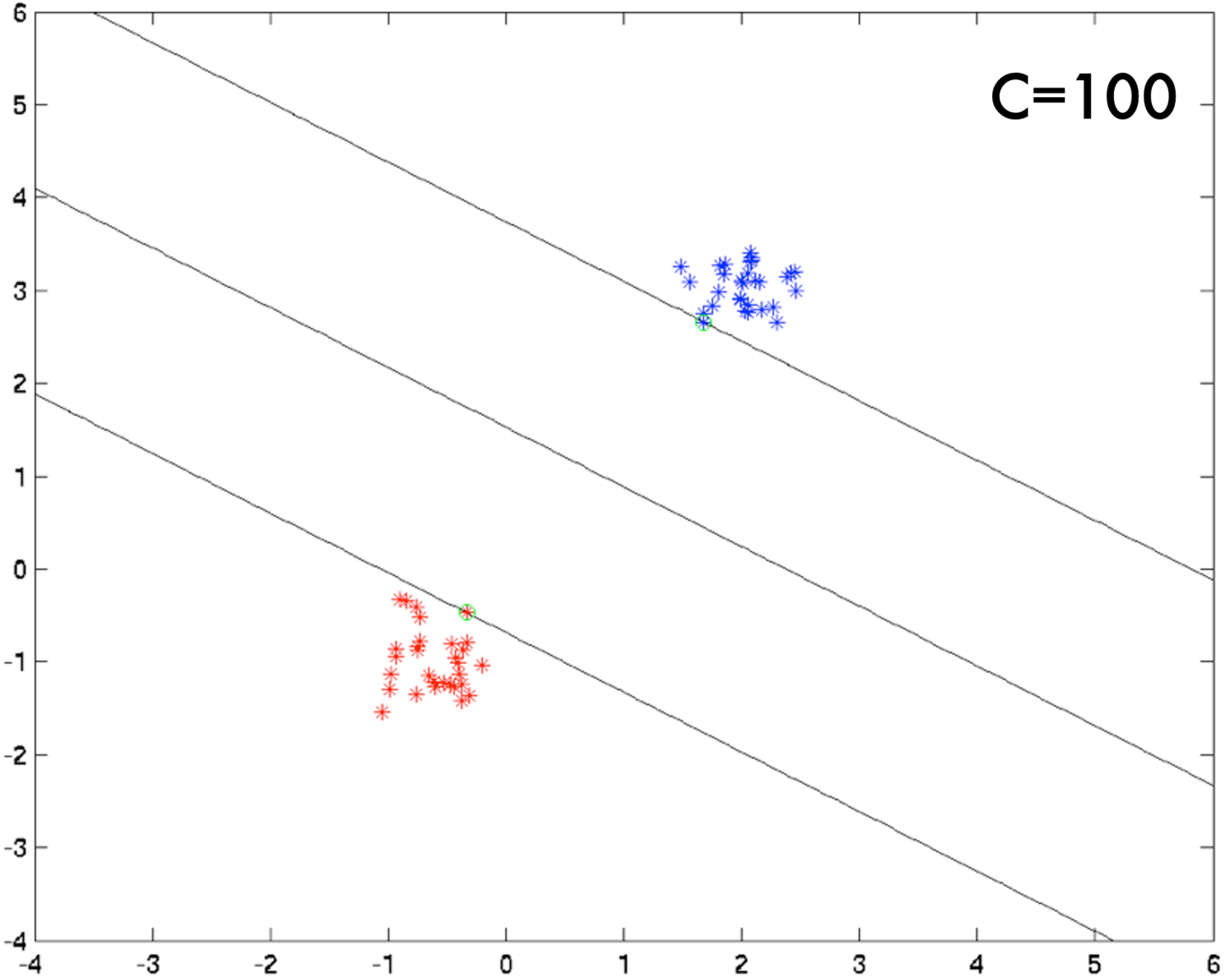
$$\max_{\alpha} \left(-\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j + \sum_{i=1}^{n} \alpha_i \right)$$

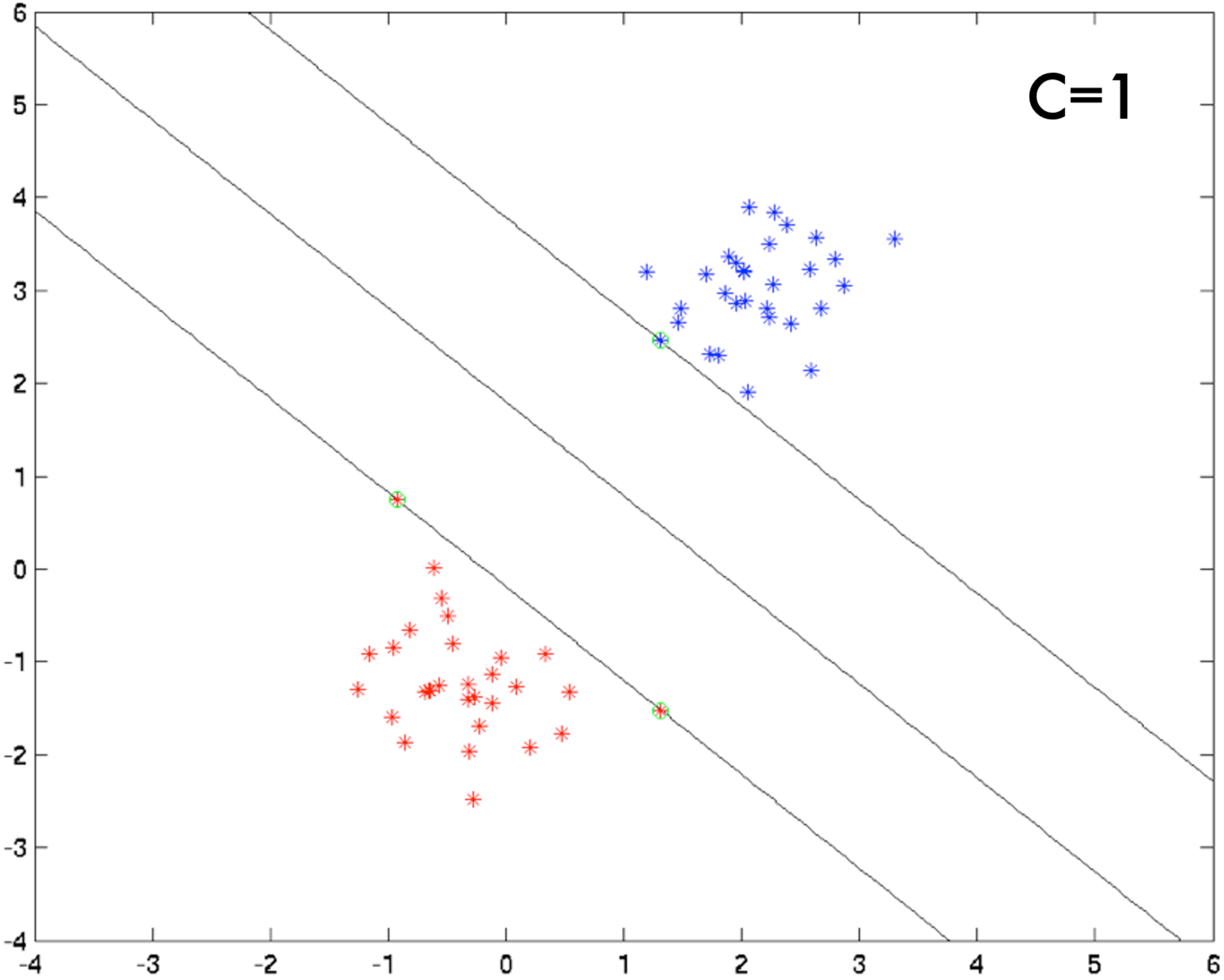
$$\text{subject to} \quad \sum_{i} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C$$

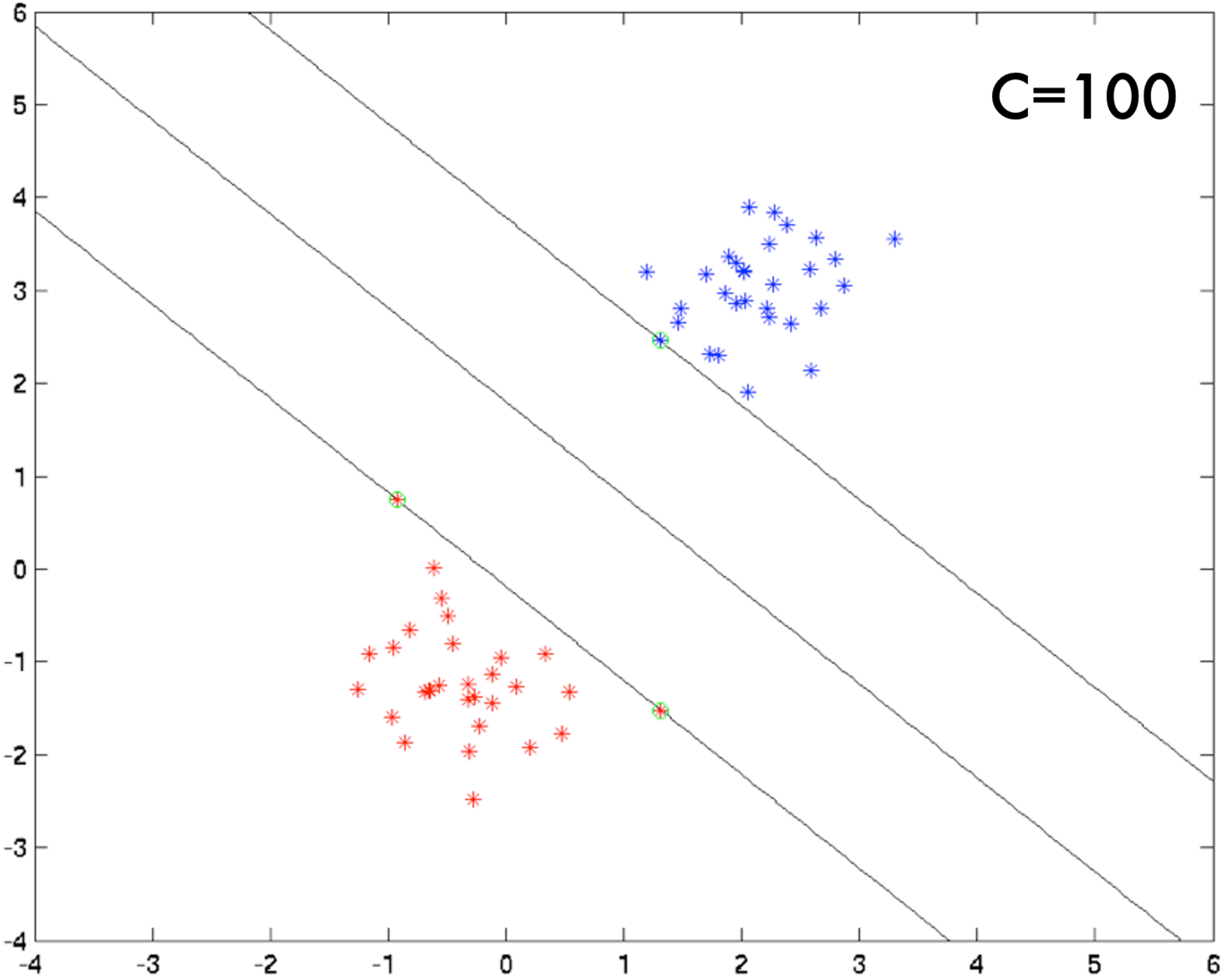


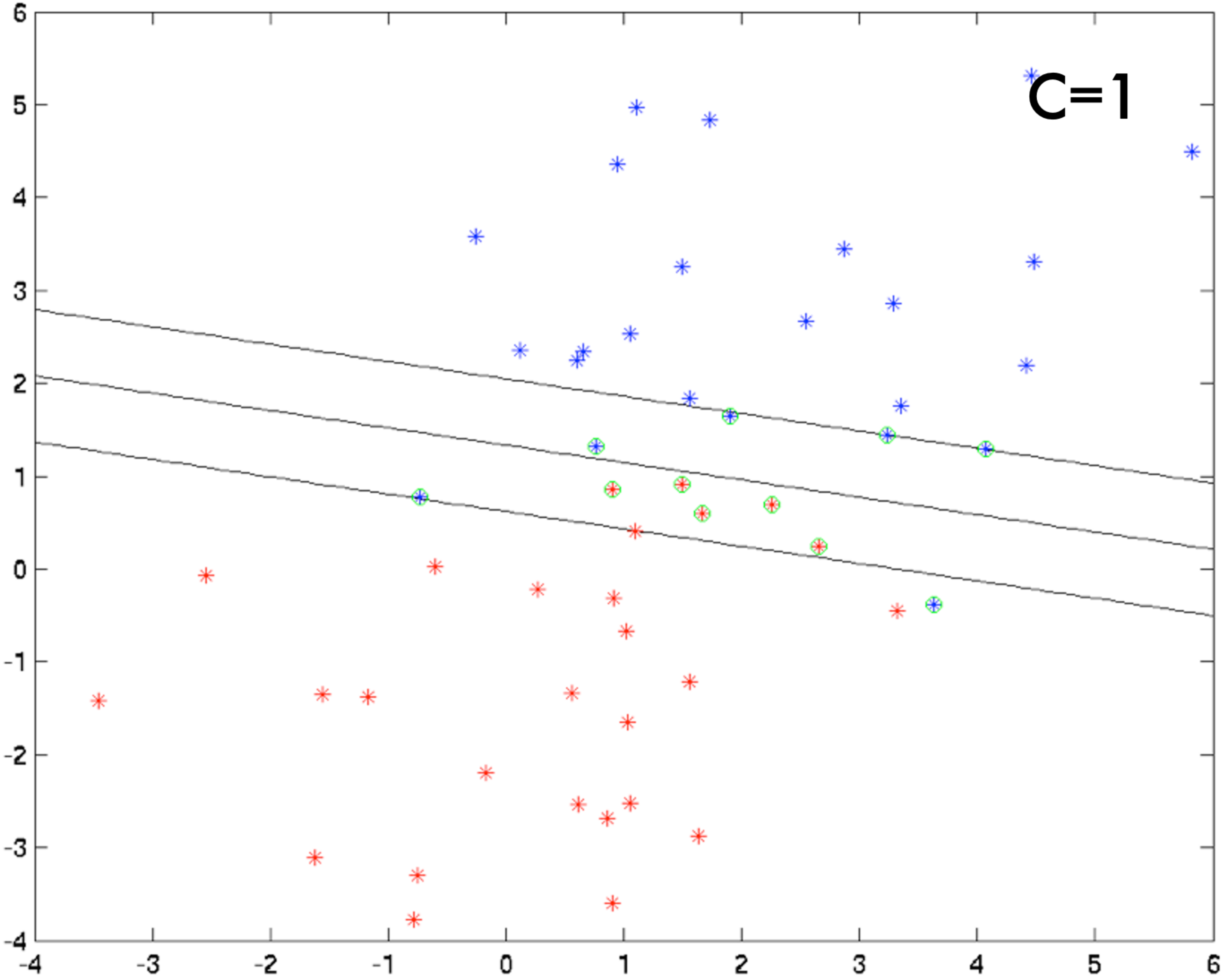
C=1


C=100
6
5
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4
5
6

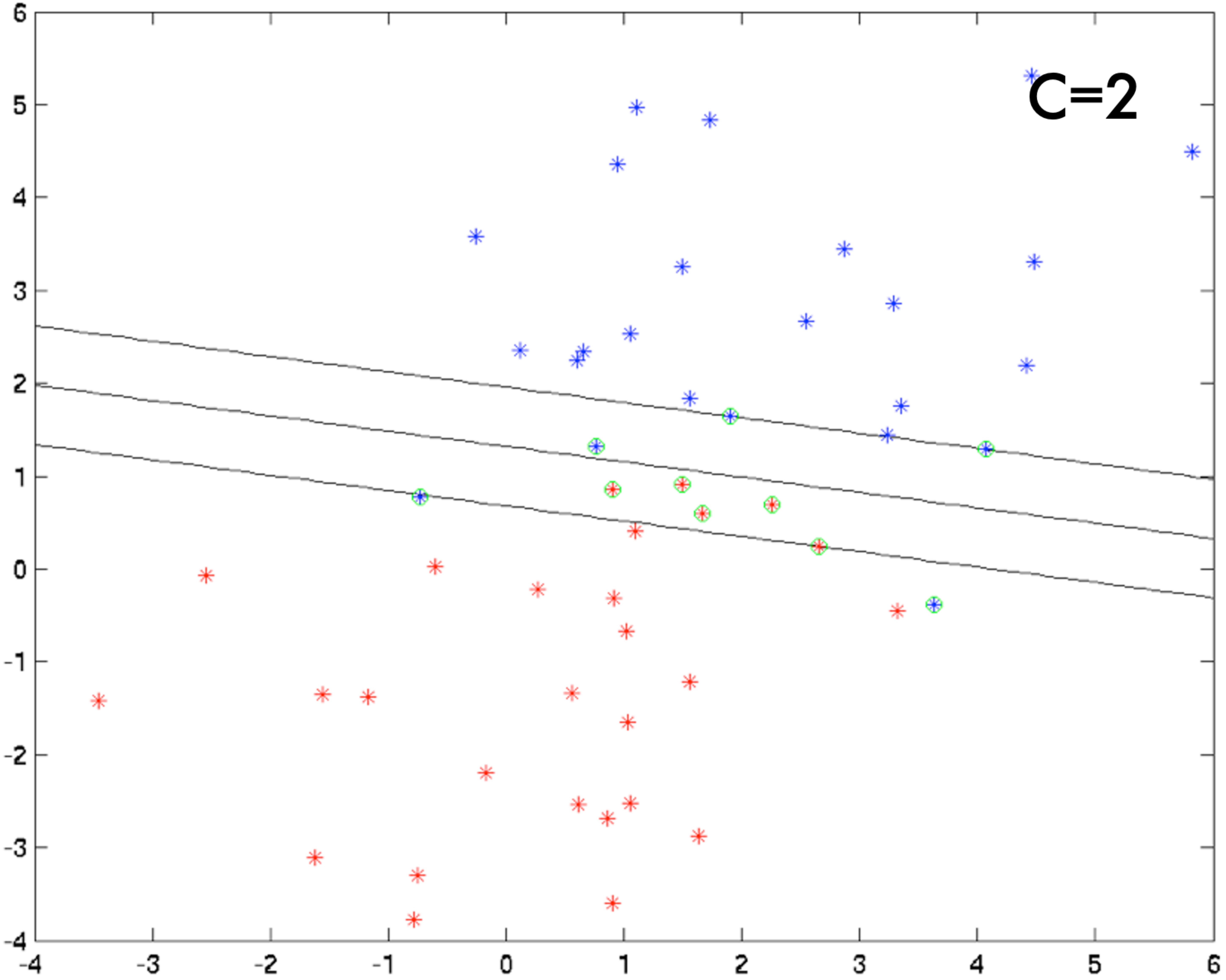


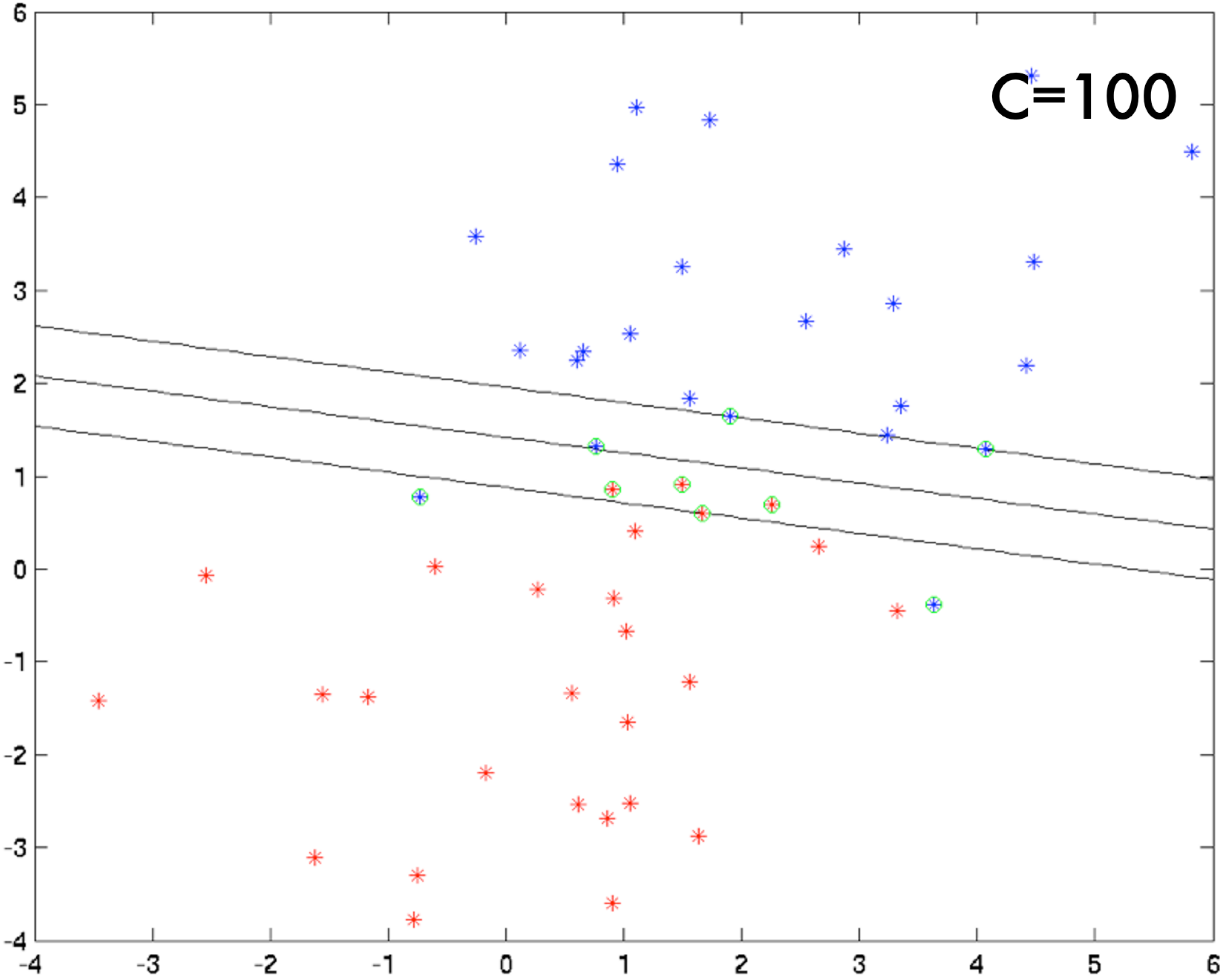
C=1
6
5
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4
5
6


C=100
-4
-3
-2
-1
0
1
2
3
4
5
6
-4
-3
-2
-1
0
1
2
3
4
5
6



C=1
6
5
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4
5
6


C=2
6
5
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4
5
6


C=100
6
5
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4
5
6

Solving the Optimization

$$\max_{\alpha} \left(-\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j + \sum_{i=1}^{n} \alpha_i \right)$$

$$\text{subject to} \quad \sum_i \alpha_i y_i = 0 \qquad 0 \le \alpha_i \le C$$

- If the problem is small-scale (thousands of variables), we can use off-the-shelf solvers (cvxopt, cplex, ooqp, loqo)
- For large-scale problems, use the fact that only SVs matter and solve in blocks (active set method)

Cheers

- *Next up.* Kernel Tricks.