Day 13: Linear Algebra Review and SVMs

Linear Algebra

Matrix $\underline{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$ 2x3 matrix

Matrices can represent many things...

images

$$\begin{array}{cc} & \begin{array}{ccc} \text{col1} & \text{col2} & \text{col3} \end{array} \\ \begin{array}{c} \text{dp1} \\ \text{dp2} \\ \dots \end{array} & \begin{bmatrix} 34 & 3 & 9 \\ 48 & -3 & 9 \\ & & \end{bmatrix} \end{array}$$

tabular data

...

A^T = "transpose of matrix A" = $\begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$

A^{-1} = "inverse of matrix A"

$$AA^{-1} = A^{-1}A = I = \text{"identity matrix"}$$

<u>Identity matrix</u>: 0's everywhere except 1's in the diagonal

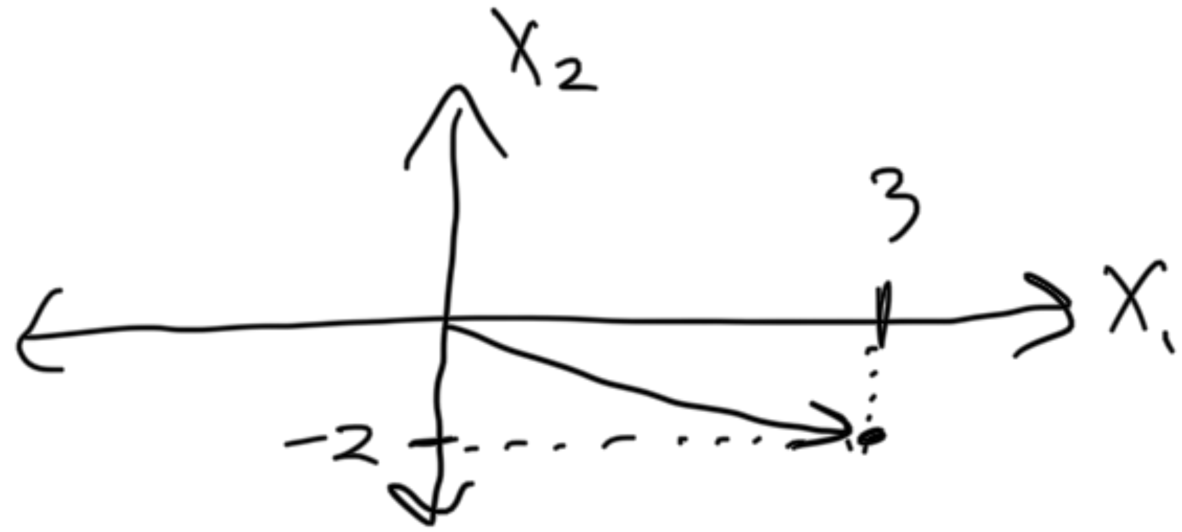
$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{aligned} AI &= A \\ IA &= A \end{aligned}$$

<u>Vectors</u> are $n \times 1$ matrices

$$\begin{bmatrix} a \\ b \end{bmatrix}, \quad \begin{bmatrix} a \\ b \\ c \\ \vdots \\ z \end{bmatrix}, \quad \text{etc.}$$

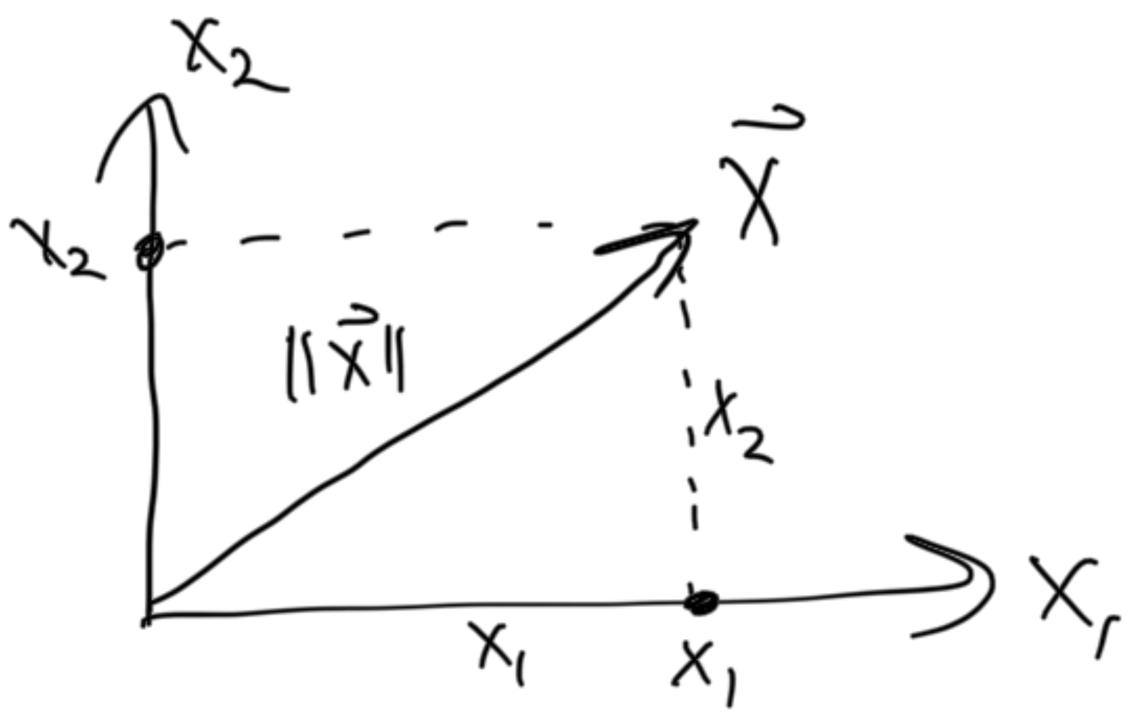
Represents a point in space:

$$\begin{bmatrix} 3 \\ -2 \end{bmatrix}$$



Norm of a vector: magnitude of a vector

$$\|X\|_2 = \sqrt{X_1^2 + X_2^2}$$

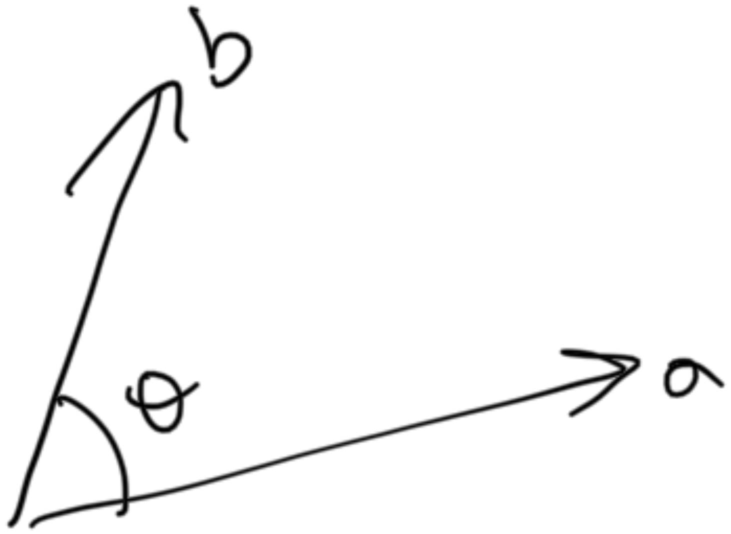


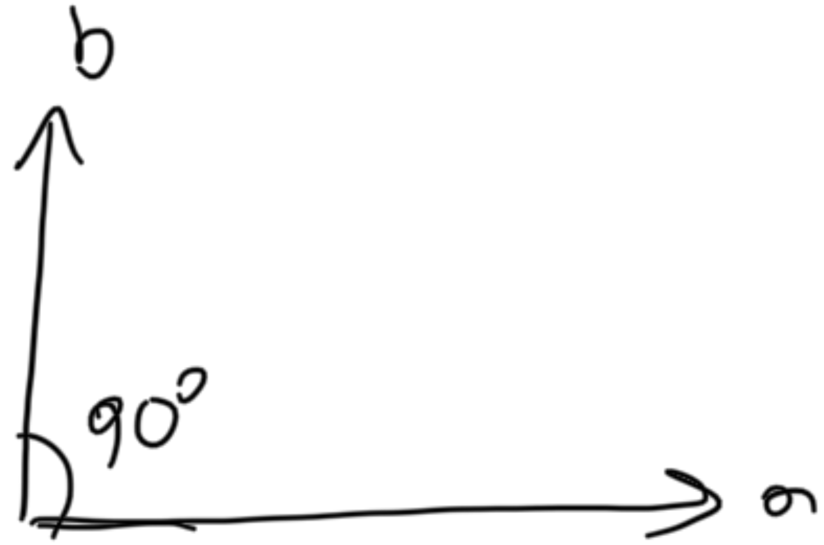
Dot Product:

$$[4 \quad 3 \quad 2] \cdot [1 \quad 3 \quad -7]$$

$$= 4 \cdot 1 + 3 \cdot 3 + 2 \cdot -7 = -1$$

$$a \cdot b = \|a\| \, \|b\| \cos \theta$$





$a \cdot b = 0$

Matrix Multiplication

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} \cdot b_{11} + a_{12} \cdot b_{21} & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & a_{31} b_{13} + a_{32} b_{23} \end{bmatrix}$$

(dot products between rows of A and columns of B)

Linear Regression as Linear Algebra

$$y = w_1 x_1 + w_2 x_2 + \ldots w_n x_n + b$$

$$= \vec{w}^T \vec{x} + b$$

For a dataset with 3 points:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \approx \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = X\theta$$

$$X\theta = \begin{bmatrix} \theta_0 + \theta_1 x_1 \\ \theta_0 + \theta_1 x_2 \\ \theta_0 + \theta_1 x_3 \end{bmatrix}$$

Minimize $(y - X\theta)^T (y - X\theta)$

$$= \ldots = y^T y - 2\theta^T X^T y + \theta^T X^T X \theta$$

So, to find the minimum:

$$\frac{\partial MSE}{\partial \theta} = -2\theta^T X^T y + 2X^T X = 0$$

Solve for θ:

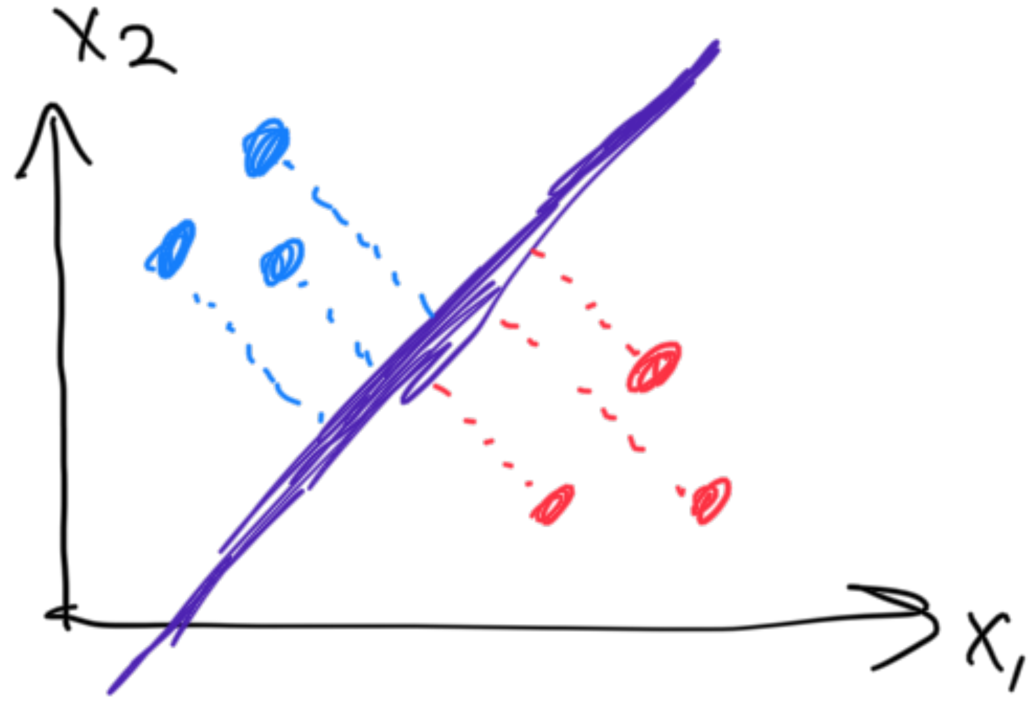
$$\underbrace{\theta = (X^T X)^{-1} X^T y}_{\text{"Normal Equation"}}$$

analytical solution to linear regression

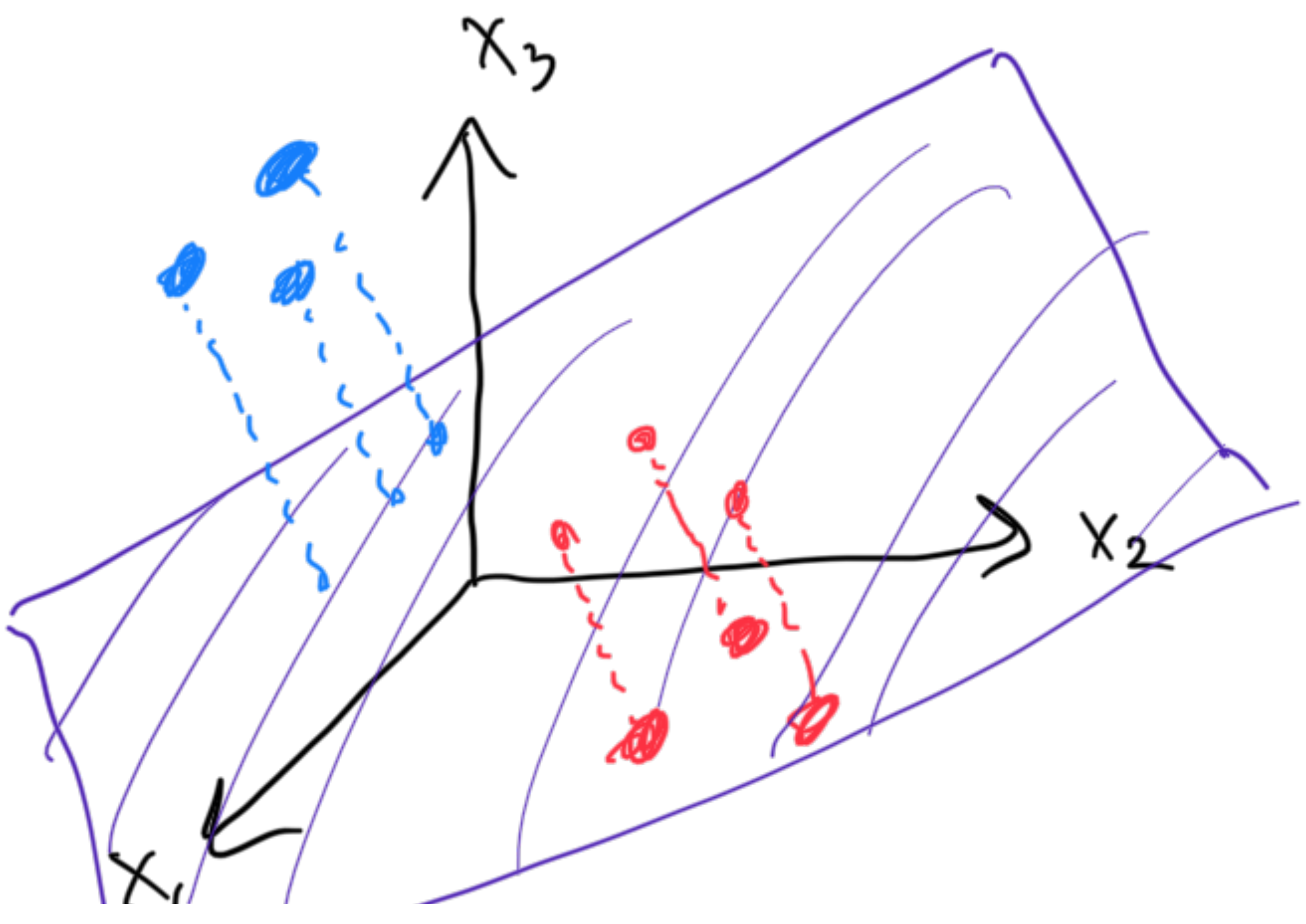
Support Vector Machines

Basic idea: find the maximally separating hyperplane

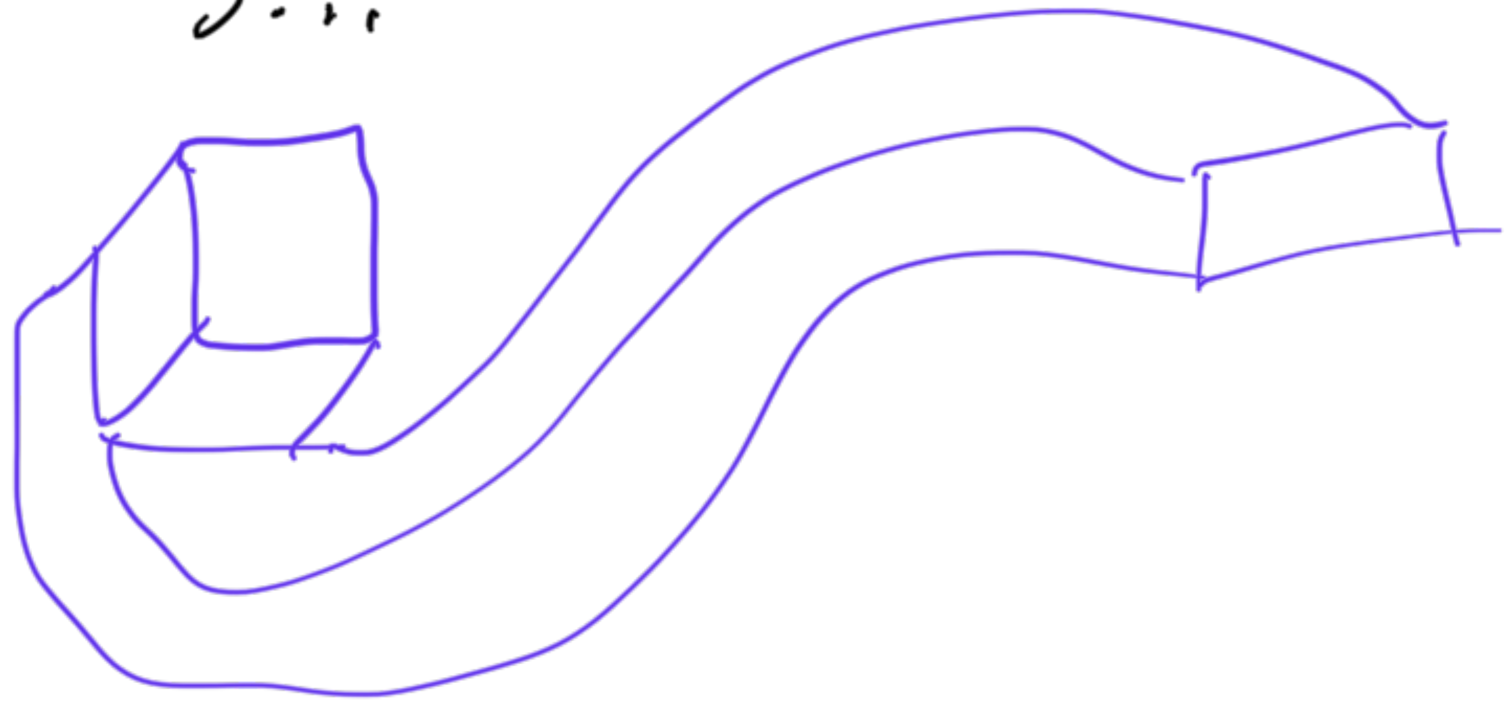
Maximally separating point



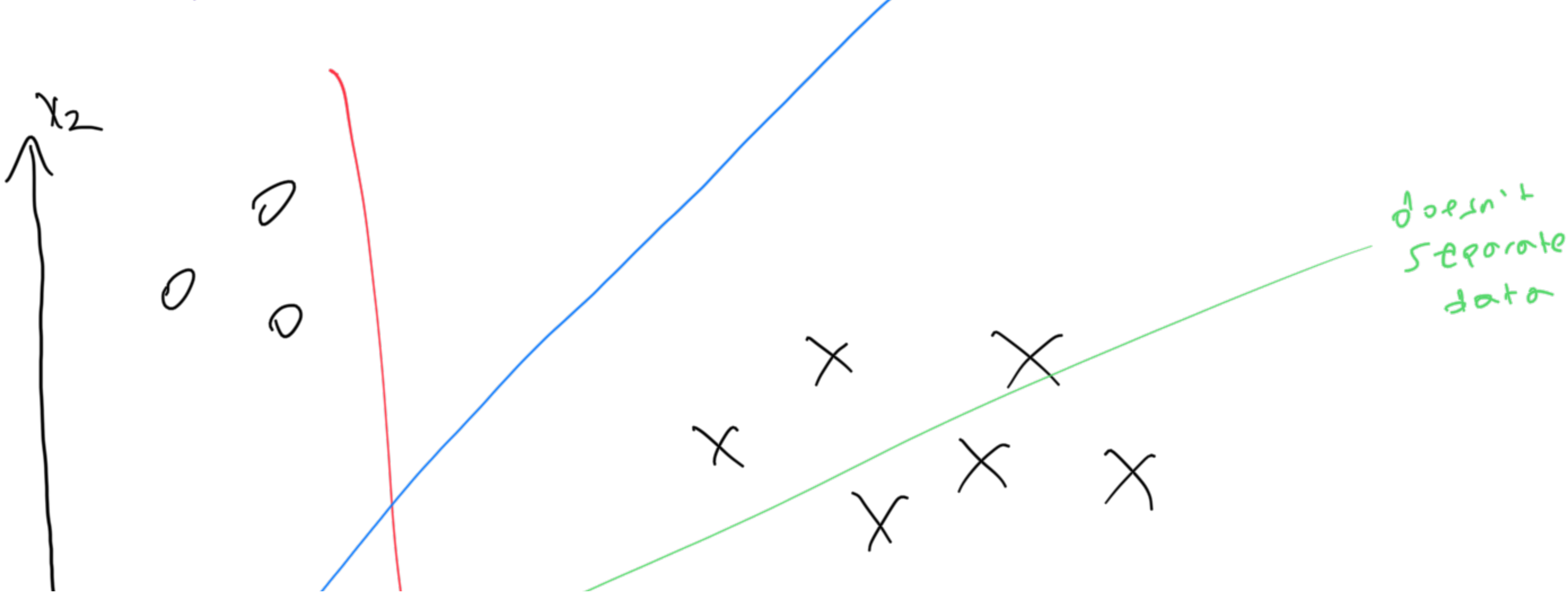
Maximally separating line



When input dimensionality (# of input features) is greater than 3...



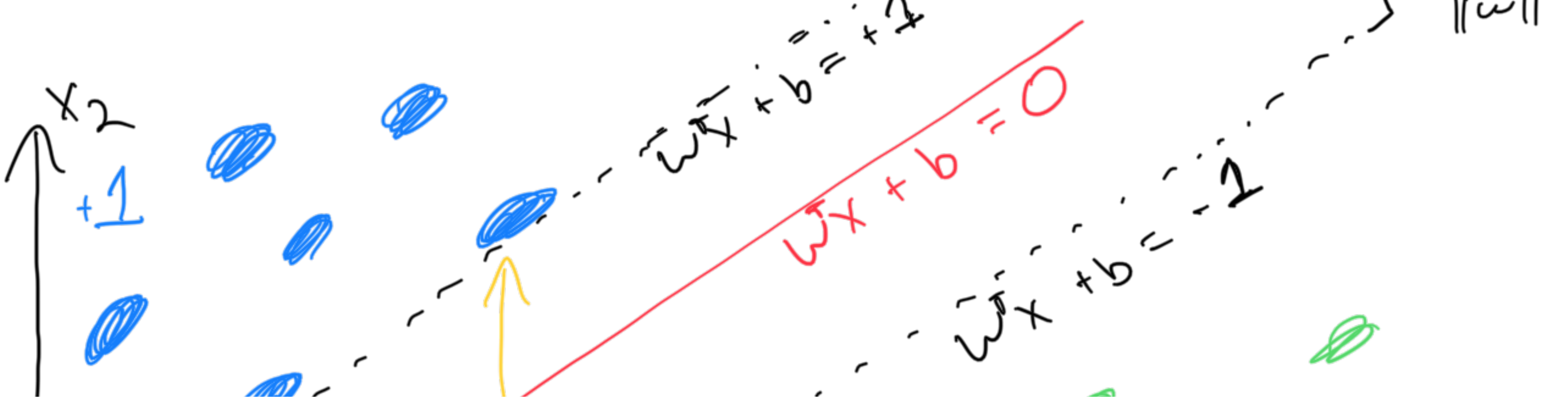
maximally separating hyperplane

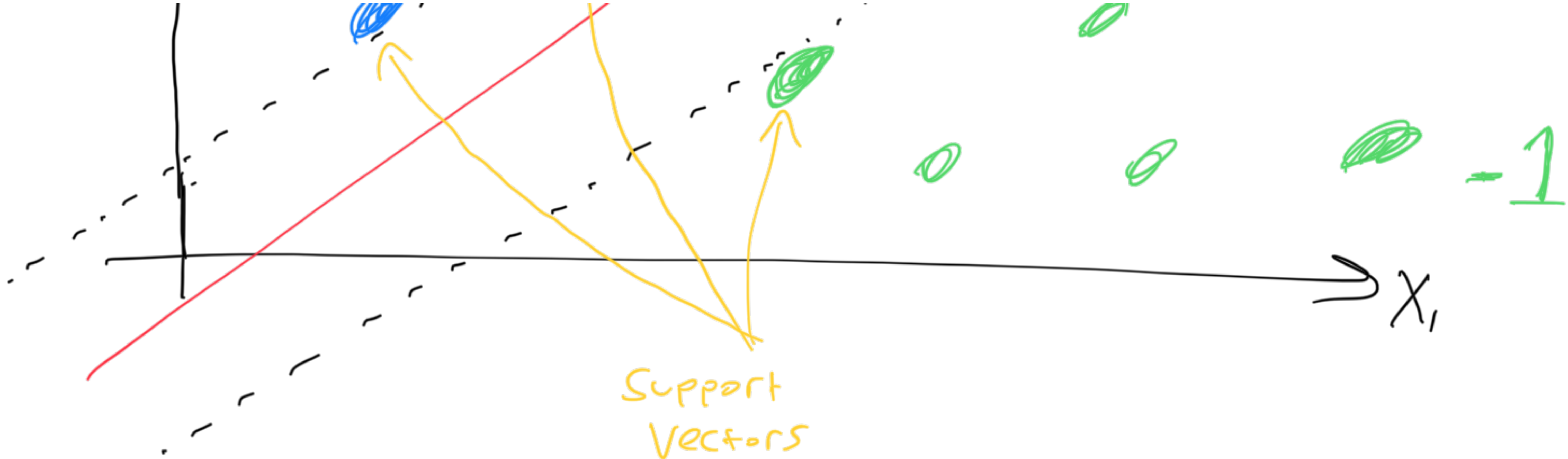


Hard-Margin SVC (Support Vector Classifier)

find margin which doesn't allow for violations of the margin

margin = $\frac{2}{\|w\|}$





(y is coming off the page!!)

$$w^T x_1 + b = +1$$
$$-(w^T x_2 + b = -1)$$

$$\frac{w^T (x_1 - x_2)}{\|w\|} = \frac{2}{\|w\|}$$

$$\text{margin} = \frac{2}{\|w\|}$$

Math goal:

maximize margin $\left(\frac{2}{\|w\|}\right)$ subject to:

- margin ≥ 0

Classify everything correctly

- $x_i w + b \geq +1$ if $y_i = +1$
- $x_i w + b \leq -1$ if $y_i = -1$

more concise: $y_i(w^T x_i + b) \geq 1$

Re-state this as a minimization problem:

Minimize $\|w\|$ subject to:

$y_i(w^T x_i + b) \geq 1$

Goal of hard margin SVM

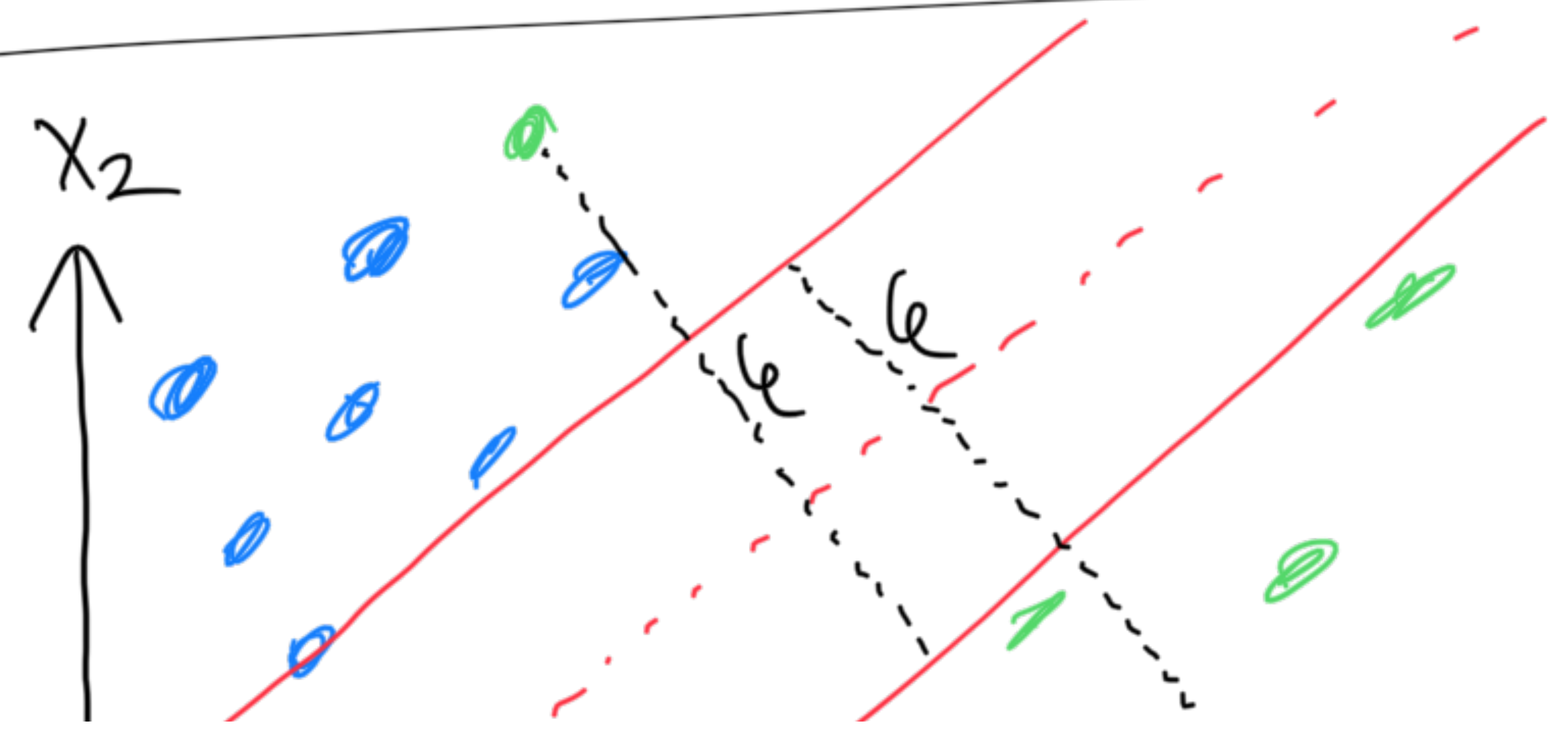
$$y_i(w^T x_i + b) \geq 1$$

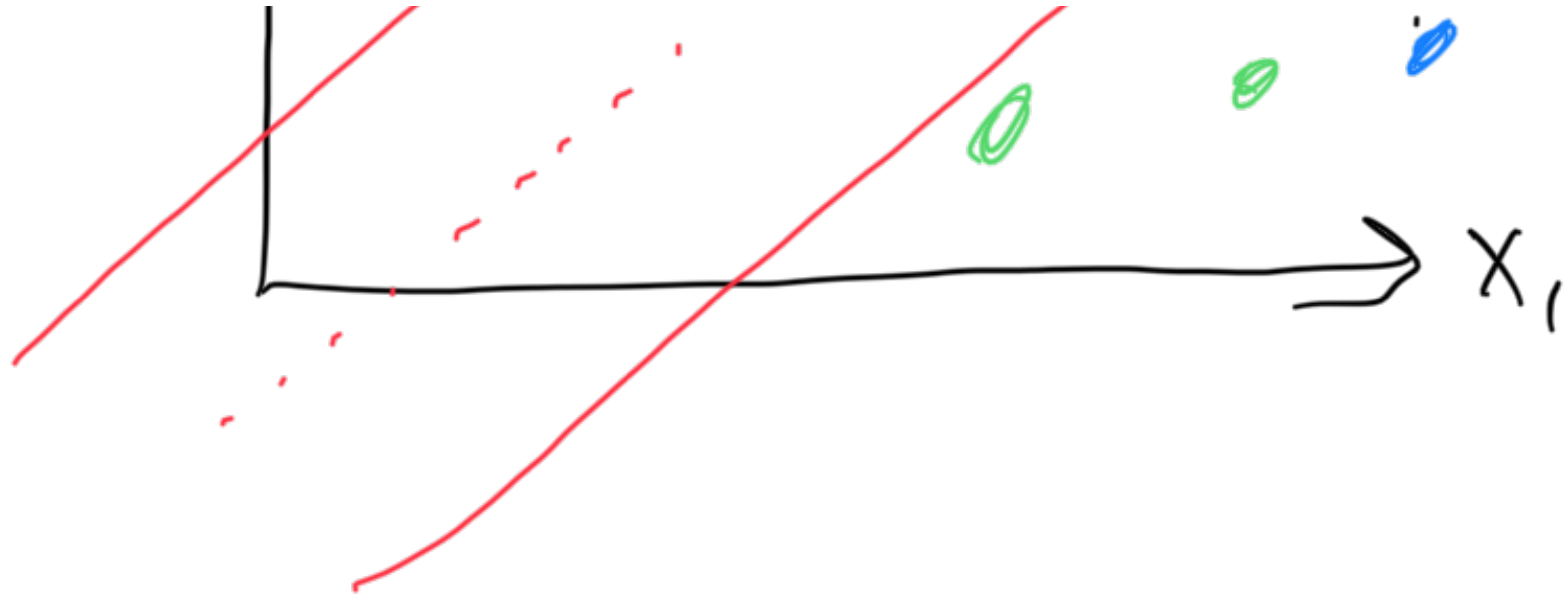
Constrained Optimization Problems:

Lagrange Multipliers

What if the data are not perfecty separable?

Soft-Margin SVC





Introduce a "slack" variable ε

New objective function is:

hyperparameter

$$\text{minimize } ||w|| + C \sum_{i=1}^{c} \varepsilon_i$$

$$\text{subject to: } y_i(w^T x_i + b) \geq 1 - \varepsilon_i$$

$\cdot \; \xi_i \geq 0$

Hinge Loss Function

$$\text{Hinge Loss} = \max\left(0,\ 1 - y \cdot \hat{y}\right)$$

(recall that $\hat{y}$ is the predicted value and y is the actual value)

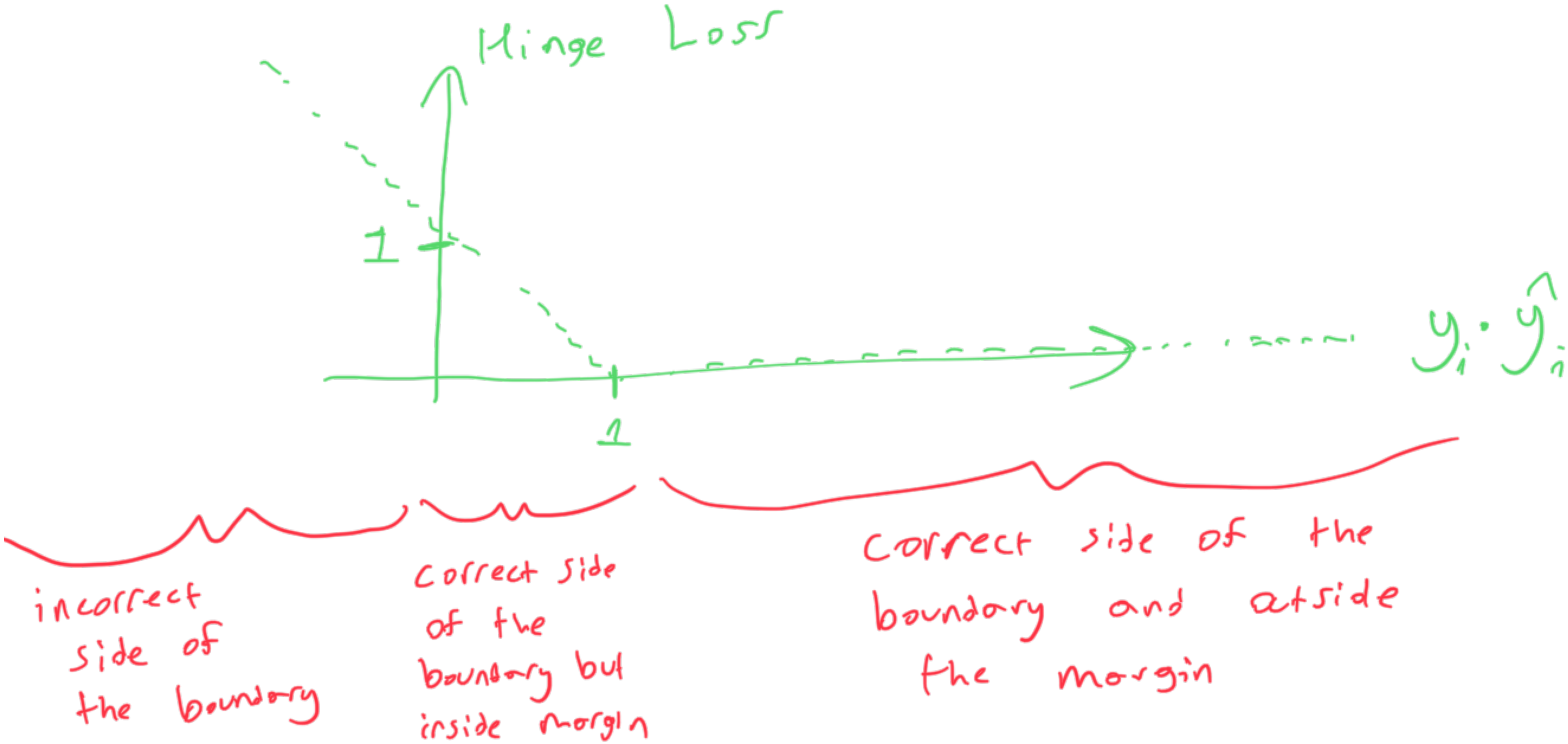
Properties:

- if $\hat{y}$ is correct and $|\hat{y}| \geq 1$:
  Hinge Loss $= 0$
- if $\hat{y}$ is correct and $|\hat{y}| < 1$:

$0 < \text{Hinge Loss} < 1$

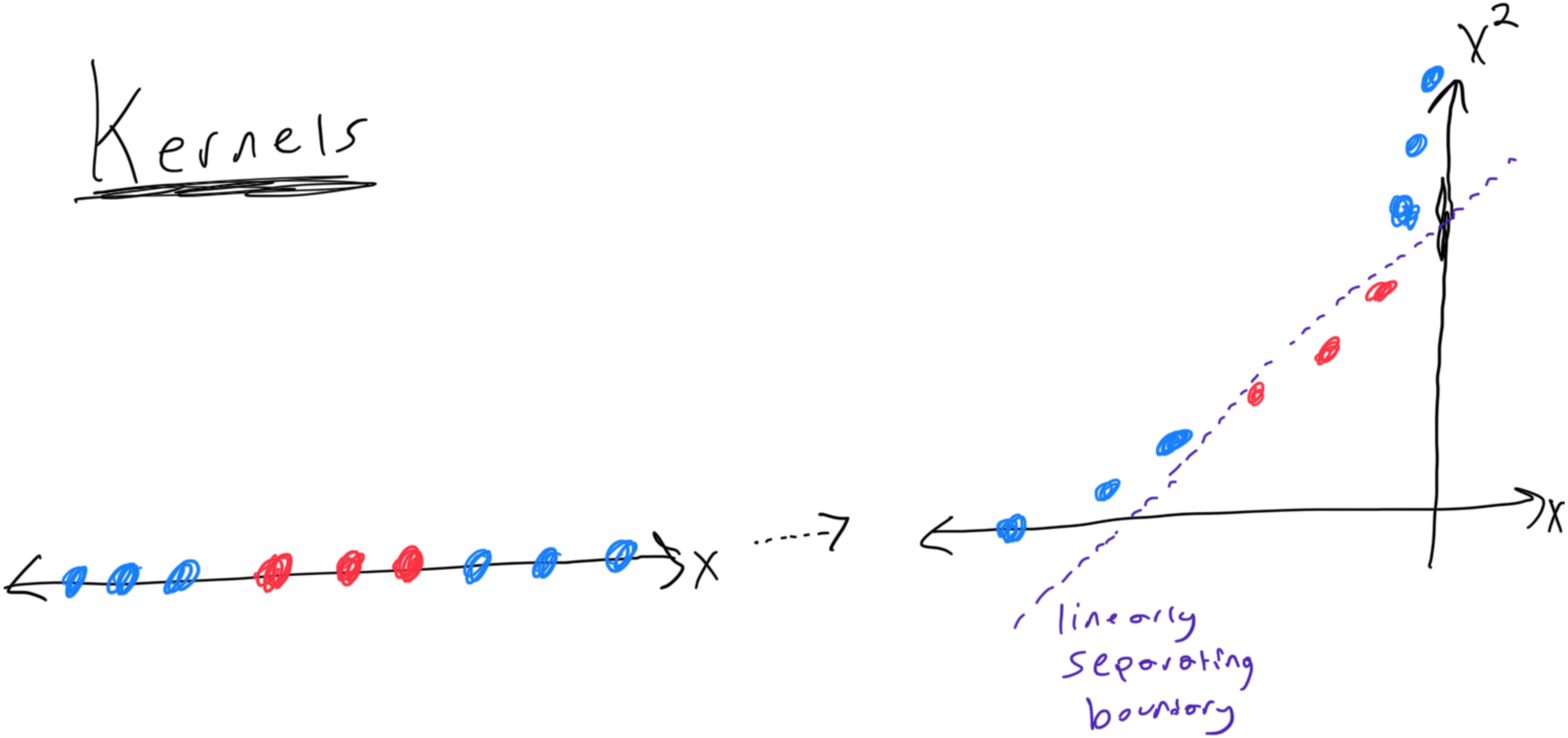
- is $\hat{y}$ is incorrect:

$\text{Hinge Loss} \geq 1$



What if the data are not even close to being separable?

Kernels



Map data to high dimensional space

Key idea: non-linearly separable data can be separable in high dimensions

<u>Fundamental Issue</u>: moving a reasonably sized feature space (e.g., 10,000 features) into a huge feature space which may be required to linearly separate the data (e.g., 10^{10} features) is computationally infeasible

The Solution:

<u>The Kernel Trick</u>

(Outside the scope of this class, an equivalent formulation of SVM is:

$$\max_{\alpha} \underbrace{\sum_{i=1}^{n} \alpha_i}_{\text{scalars}} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \underbrace{\alpha_i \alpha_j y_i y_j}_{\text{scalars}} \underbrace{\left(X_i^{T} X_j \right)}_{\text{matrices}}$$