Supervised Classification

TRAINING DATABASE : SUPERVISED DATA

	Feature 1	Feature 2	...	Feature p	Label
obs1	0.8	1		-12.3	1
obs2	0.6	1		5.6	-1
⋮					⋮
obs n	0.3	0		8.1	-1

MACHINE LEARNING

MODEL FITTING

test	0.7	1		4.3	= "hard prediction" -1
					"soft prediction" 0.12

THE USUAL STEPS

RAW DATA

1.2	BLVD	
NaN	AVE	
126897	ST	
0.4		

CLEANING

CLEAN DATA

1.2	BLVD	
0.8	AVE	
0.8	ST	
0.4	MISSING	

FEATURES ENGINEERING

ENGINEERED DATA

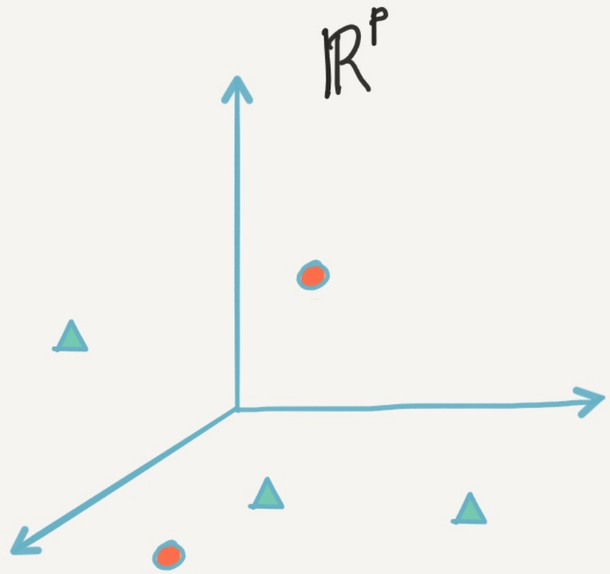
1.2	1	0	0	0
0.8	0	1	0	0
0.8	0	0	1	0
0.4	0	0	1	0

REPRESENTATIONS

Y
1
0
1
1
0

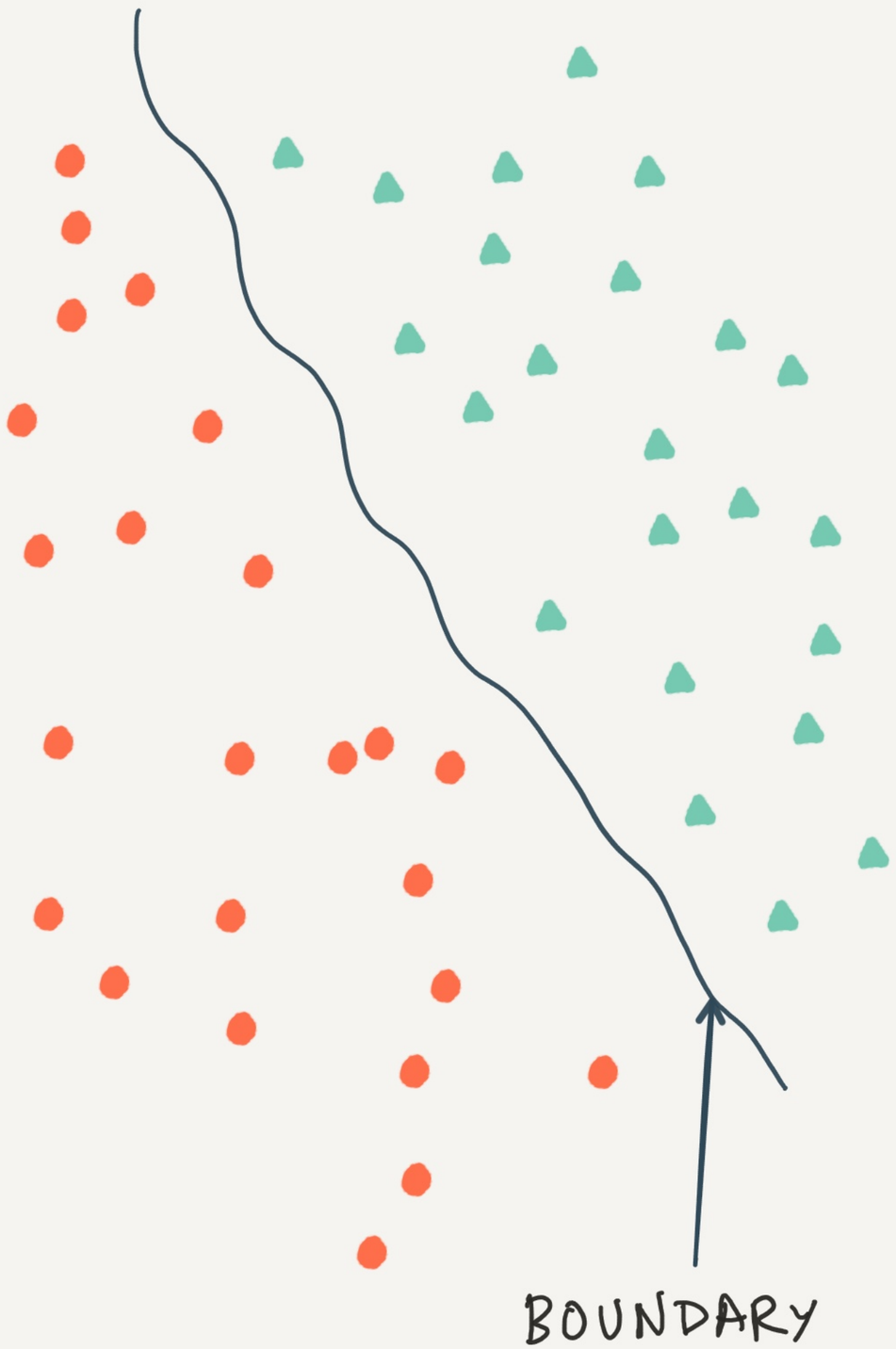
X								
1.21	1	0	0	12	43	...	...	
-1	0	1	0	1	1	...		
12.6	0	0	0	12	4			
9.7	0	0	1	12	8			
-1	0	0	0	24	5			

numerical

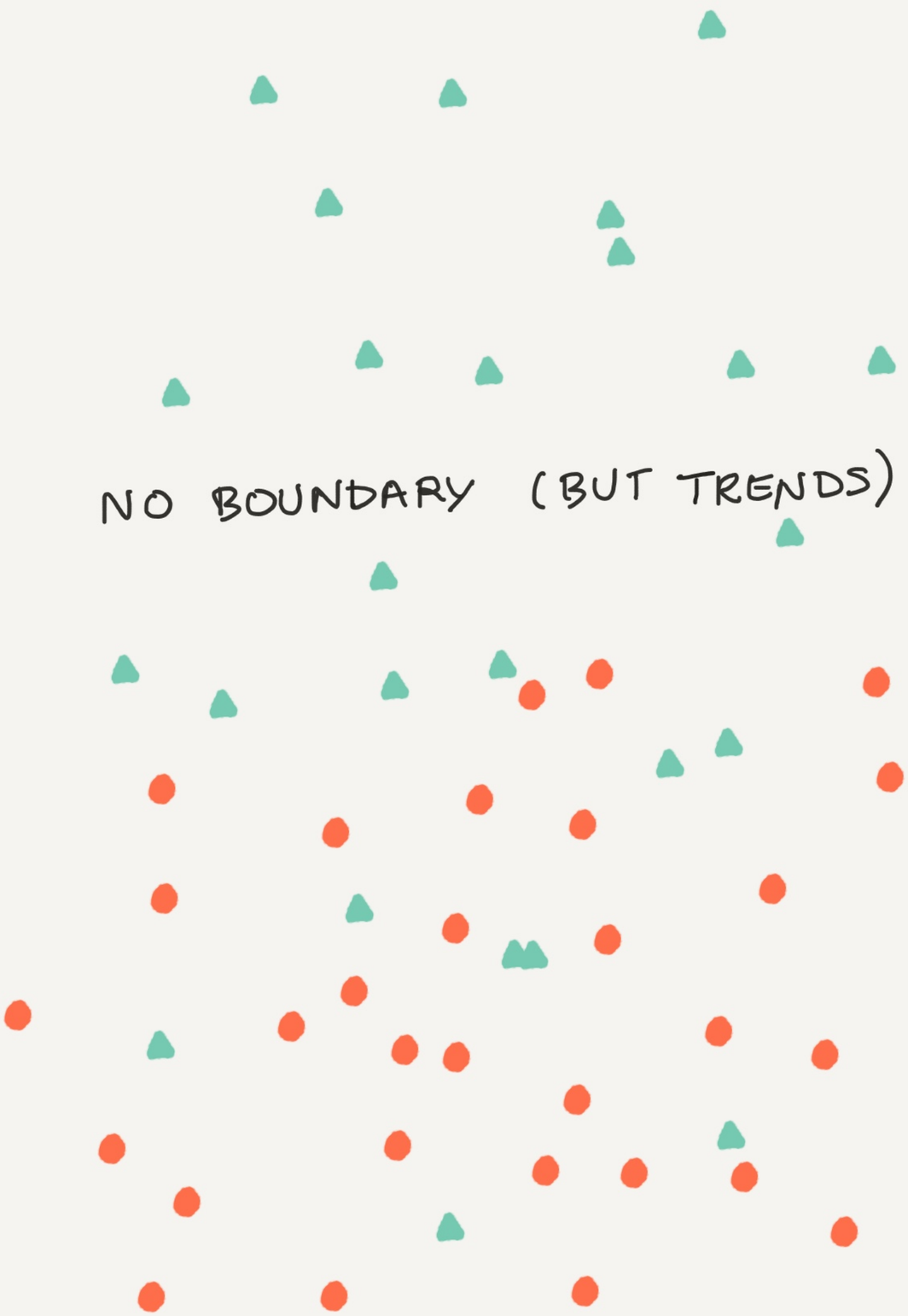


geometrical

IDEAL WORLD



REALITY

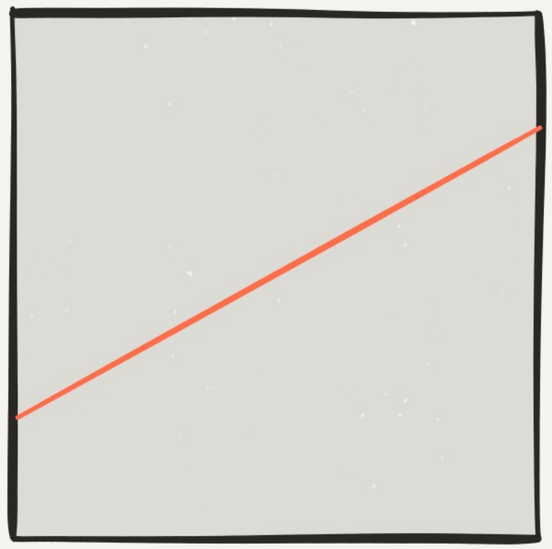


GEOMETRICALLY

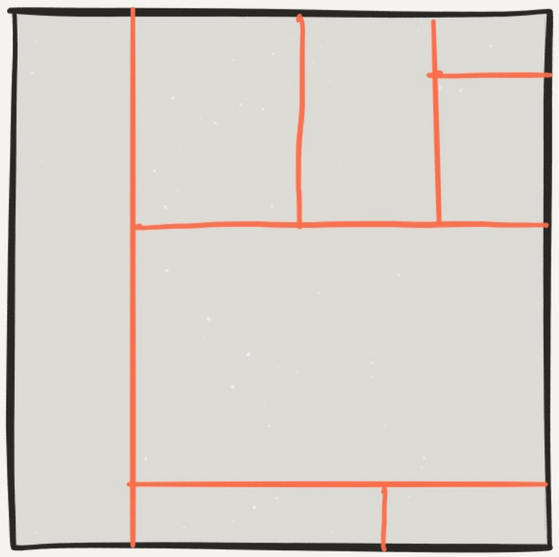
MODELS = PARAMETRIC BOUNDARIES

MODELS

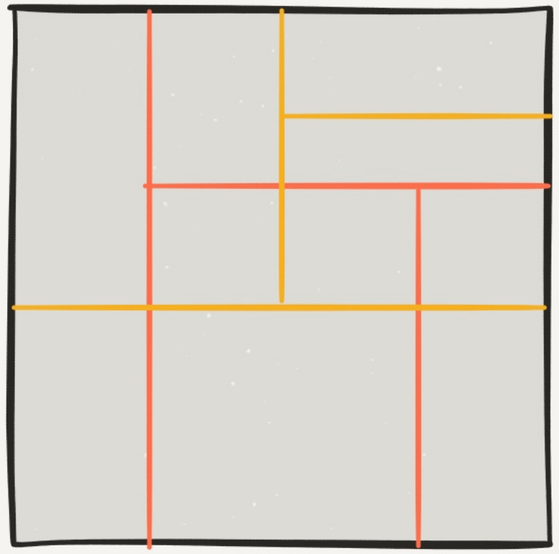
Generalized Linear Models (GLM)



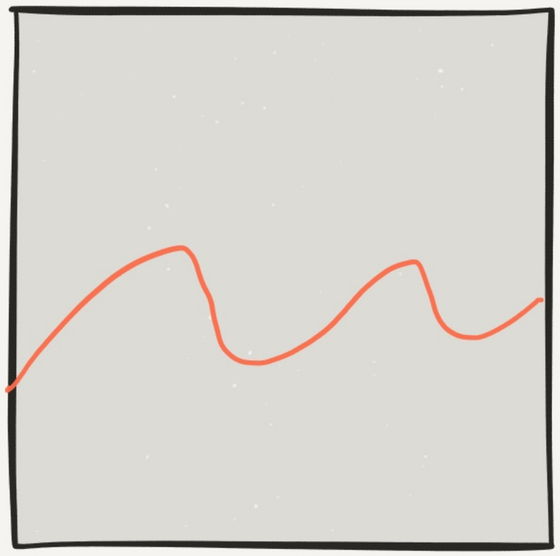
Decision Tree



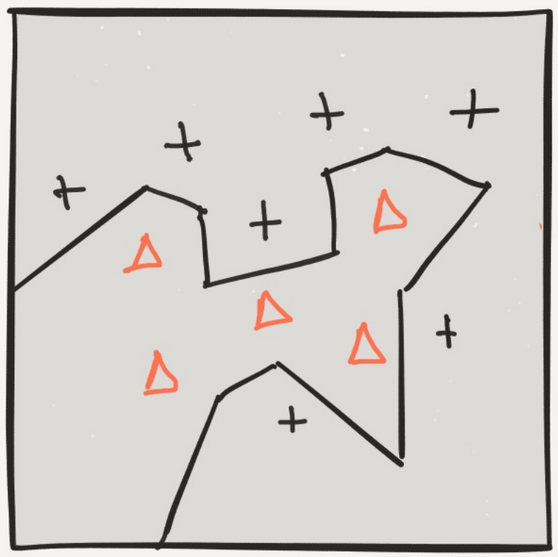
Forest



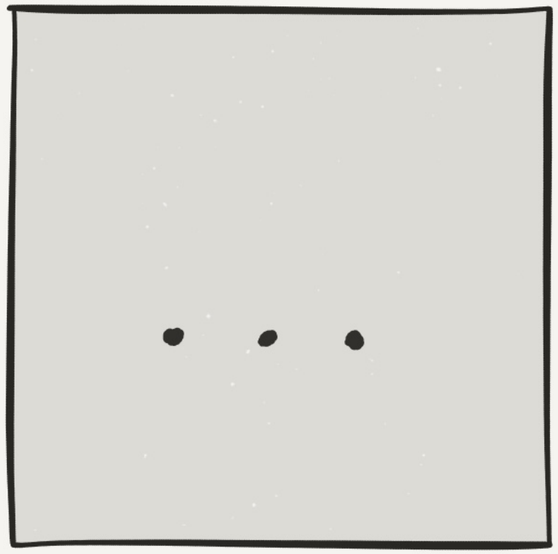
k-SVM



k Nearest Neighbors



etc.



SESSION OUTLINE

1) Intro ✓
2) Linear Models
3) Generalized Linear Models
4) Support Vector Machines
5) Kernel Methods
6) Decision Trees
7) Random Forests
8) Boosting

LINEAR REGRESSION

1. Problem Statement

Input : n data points (x_i, y_i), $1 \leq i \leq n$ où $x_i \in \mathcal{X}$ et $y_i \in \mathbb{R}$ $\forall i$

Expected Output : $w, b = \arg\min \sum_{i=1}^{n} (y_i - \langle x_i, w \rangle - b)^2$

These are Ordinary Least Squares

2. Geometrically

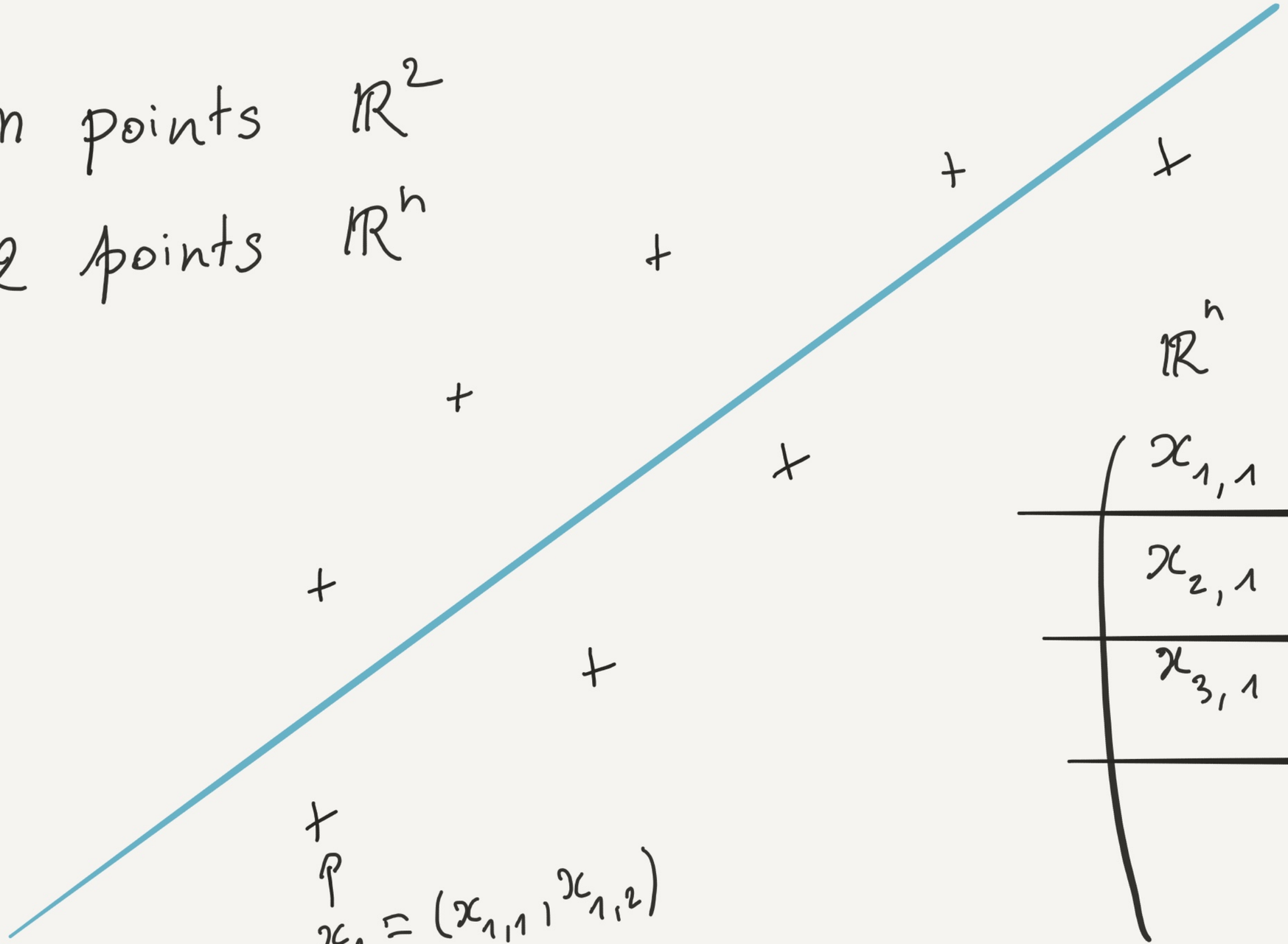
Let $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$, $X = \begin{pmatrix} 1 \; {}^t x_1 \\ \\ 1 \; {}^t x_n \end{pmatrix}$ and $\theta = \begin{pmatrix} b \\ w \end{pmatrix}$

Problem becomes:

$$\inf_{\theta} \| Y - X\theta \|^2$$

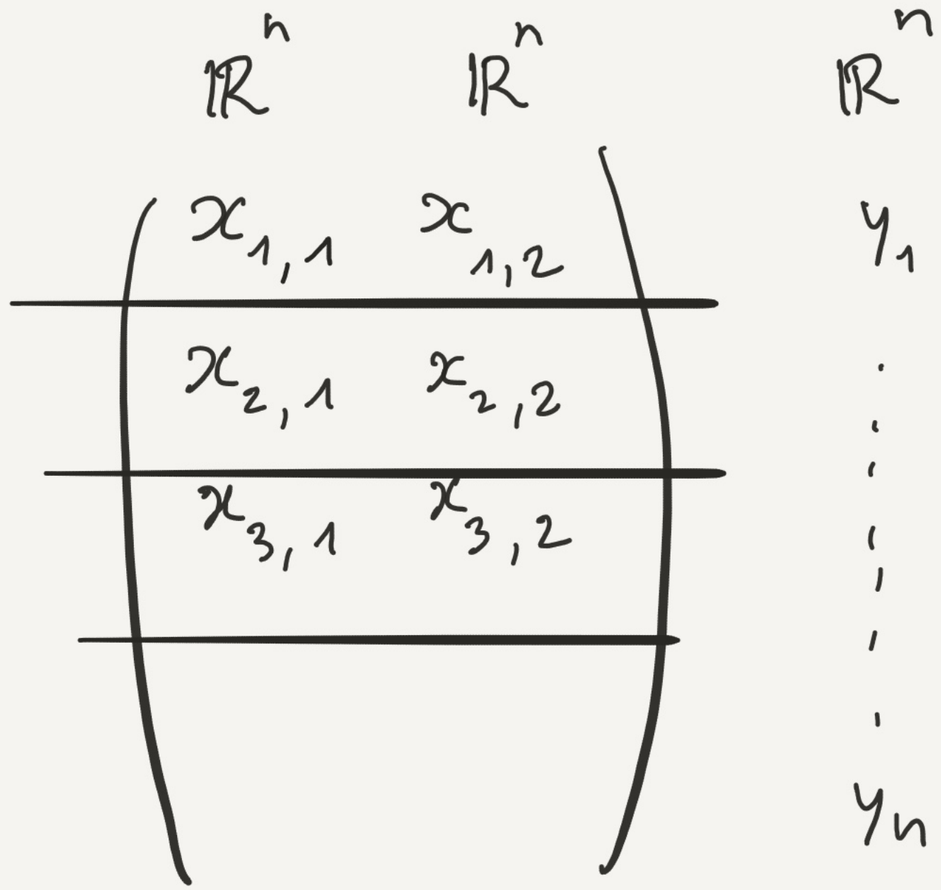
n points $\mathbb{R}^2$

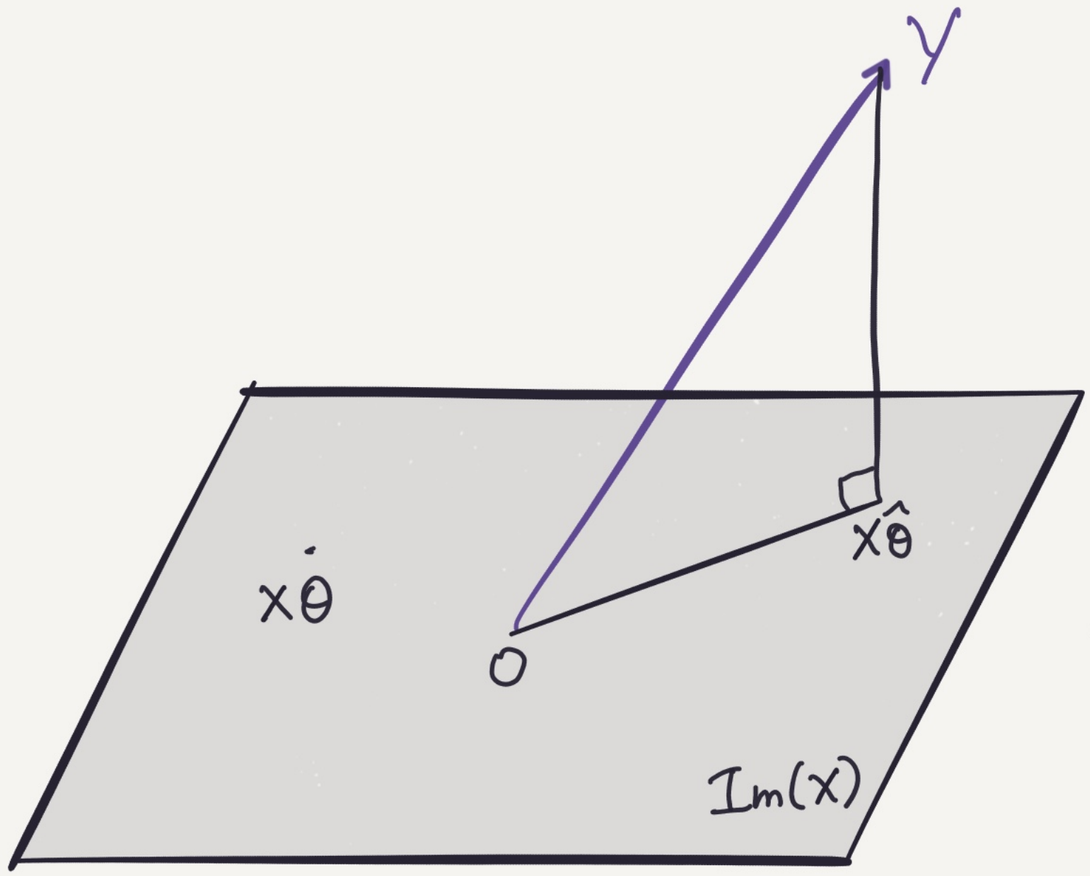
2 points $\mathbb{R}^n$



p

$x_1 = (x_{1,1}, x_{1,2})$





$X\hat{\theta}$ is orthogonal projection of Y over $\mathrm{Im}(X)$:

$$\langle Y - X\hat{\theta}, X\theta \rangle = 0 \quad \forall \theta$$

Therefore, $\|Y - X\hat{\theta}\|^2 = \inf_{\theta} \|Y - X\theta\|^2$. Indeed:

$$\|Y - X\theta\|^2 = \|Y - X\hat{\theta} + X(\hat{\theta} - \theta)\|^2 = \|Y - X\hat{\theta}\|^2 + \|X(\hat{\theta} - \theta)\|^2 \geq \|Y - X\hat{\theta}\|^2$$

3. $\hat{\theta}$ computation

From $\langle Y - X\hat{\theta}, X\theta \rangle = 0 \quad \forall \theta$, we get $\langle {}^tX(Y - X\hat{\theta}), \theta \rangle = 0 \; \forall \theta$

Then, for $\theta = {}^tX(Y - X\hat{\theta})$: $\| {}^tX(Y - X\hat{\theta}) \|^2 = 0$

Hence: ${}^tX(Y - X\hat{\theta}) = 0$

Giving: ${}^tX X\hat{\theta} = {}^tX Y$

Now, if $\text{Ker}(X) = \{0\}$, which is indeed the case if $(x_i)_{1 \le i \le n}$ are linearly independent: ${}^tXX$ is invertible, provided $n \geq p+1$ and,

$$\hat{\theta} = ({}^tXX)^{-1}{}^tXY$$

It is a nice closed-form formula. However it is not used in practice even for moderate p because matrix ${}^tXX$ is ill-conditioned, yielding numerical instability.

$$x_1 = \begin{pmatrix} x_{1,1} \\ \vdots \\ x_{1,n} \end{pmatrix} \qquad x_2 = \begin{pmatrix} x_{2,1} \\ \vdots \\ x_{2,n} \end{pmatrix}$$

scalar product $\langle x_1, x_2 \rangle = x_1 \cdot x_2 = (x_1 \mid x_2)$

$$= \sum_{i=1}^{n} x_{1,i} \cdot x_{2,i}$$

4. Penalized Linear Models

Sometimes we have an a priori over θ and we know that it should belong to some set C. So we can formulate the problem:

$$\inf_{\theta \in C} \|Y - X\theta\|^2$$

A mathematically related problem is:

$$\inf_{\theta} \|Y - X\theta\|^2 + \lambda J(\theta)$$

where $J(\theta)$ denotes a non-negative function called penalization function taking large values where θ should not be a priori.

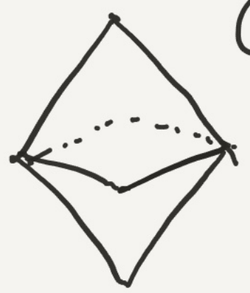
Two penalization functions are mainly used:

- the LASSO function: $J(\theta) = \|\theta\|_1$
- the RIDGE function: $J(\theta) = \|\theta\|_2^2$

To each $\lambda \geq 0$ corresponds a $\hat{\theta}_\lambda$. Function $\lambda \in [0, +\infty[\longrightarrow \hat{\theta}_\lambda$ is called regularization path. In the case of Ridge / lasso, we see $\hat{\theta}_0 = \hat{\theta}$ and $\hat{\theta}_\infty = 0$

$\|x\|_1 \leq \gamma$

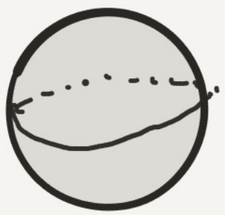
$\mathbb{R}^3$

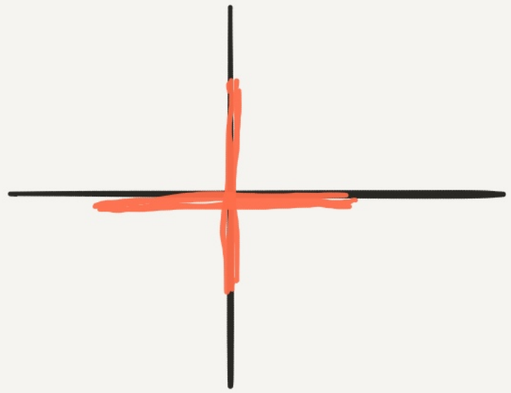


C

$\mathbb{R}^2$

$\|x\|_2 \leq \gamma$

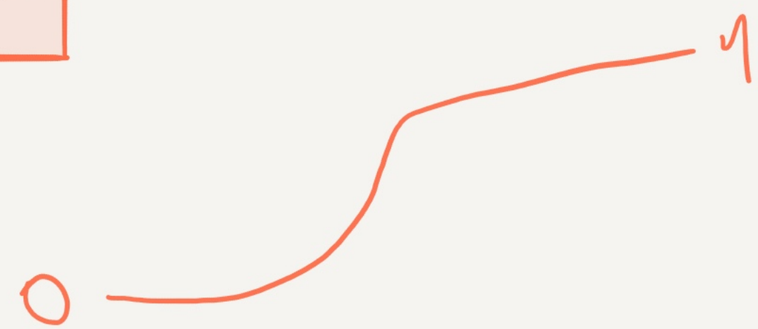




$$\|x\|_0 = \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{if } x_1 = 0 \text{ or } x_2 = 0 \\ 2 & \text{if } x_1 \neq 0 \; x_2 \neq 0 \end{cases}$$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

LOGISTIC REGRESSION



1. Problem Statement

Input : n data points (x_i, y_i), $1 \le i \le n$
with $y_i \in \{-1, +1\}$

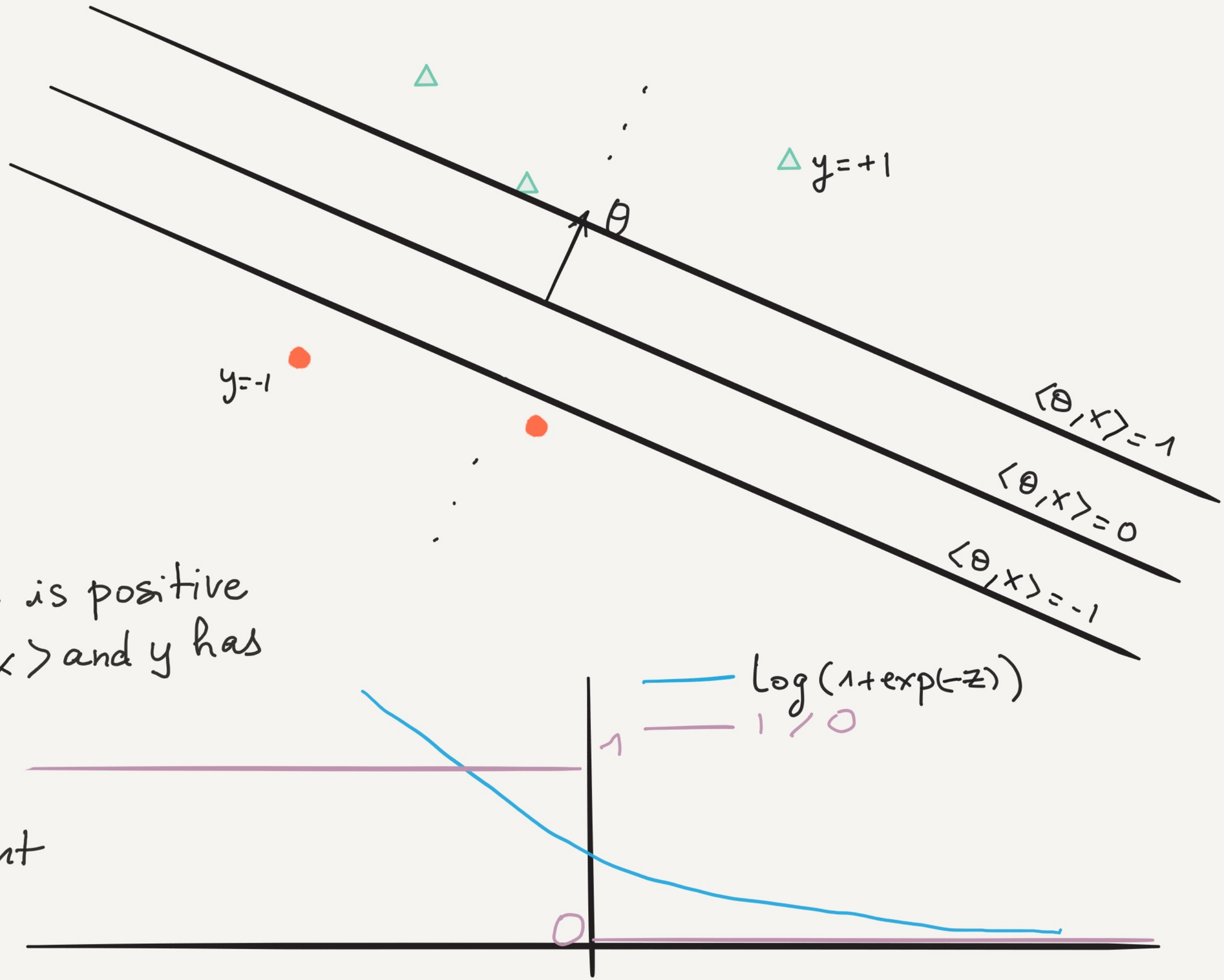
$$\mathbb{P}_\theta[Y=1 \mid X] = \frac{1}{1+e^{-\langle \theta, X \rangle}} \Longrightarrow \log \mathbb{P}_\theta[Y=1 \mid X] = -\log(1+\exp -\langle \theta, X \rangle)$$

Maximum Likelihood estimator writes:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \left[y_i \log \mathbb{P}_\theta[Y=1 \mid X](x_i) + (1-y_i) \log\left(1 - \mathbb{P}_\theta[Y=1 \mid X](x_i)\right) \right]$$

$$= \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \log\left(1 + \exp\left[-y_i \langle \theta, x_i \rangle\right]\right)$$

2 Geometrical Interpretation



$y\langle\theta,x\rangle$ is positive when $\langle\theta,x\rangle$ and y has same sign

$\Rightarrow$ correctly labeled point

3 Computation of $\hat{\theta}$

Contrarily to Ordinary Least Squares, there is no closed-form here. So numerical optimisation is performed

We start with:

$$\inf_{\theta} F(\theta) \overset{\text{def}}{=} \sum_{i=1}^{n} f\left(y_i \langle x_i, \theta \rangle\right) \quad \text{where} \quad f(z) = \log\left(1 + \exp(-z)\right)$$

$$\nabla F(\theta) = \sum_{i=1}^{n} y_i \varphi'\left(y_i \langle x_i, \theta \rangle\right) x_i \qquad f'(z) = \frac{-1}{1+\exp(z)}$$

$$\nabla^2 F(\theta) = \sum p_i (1-p_i) x_i^t x_i \qquad f''(z) = \frac{\exp(z)}{(1+\exp(z))^2}$$

Gradient Descent gives:

$$\theta_{n+1} = \theta_n + \gamma \sum_{i=1}^{n} \left[\left(\frac{1+y_i}{2}\right) - p_i\right] x_i$$

Newton-Raphson:

$$\theta_{n+1} = \theta_n - \nabla^2 F(\theta_n)^{-1} \left(\nabla F(\theta_n)\right)$$

4. Penalized Logistic regression

Let us replace

with

$$\inf_{\theta} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \langle x_i, \theta \rangle\right)\right)$$

$$\inf_{\theta} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \langle x_i, \theta \rangle\right)\right) + \lambda J(\theta)$$

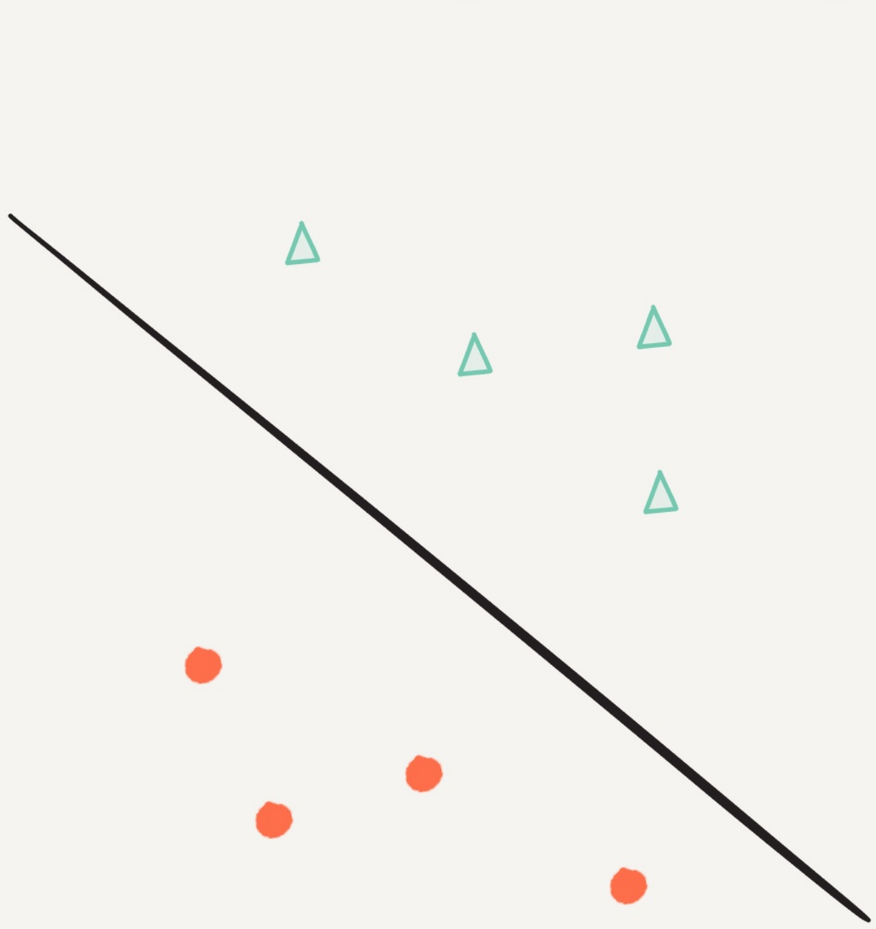
For example:

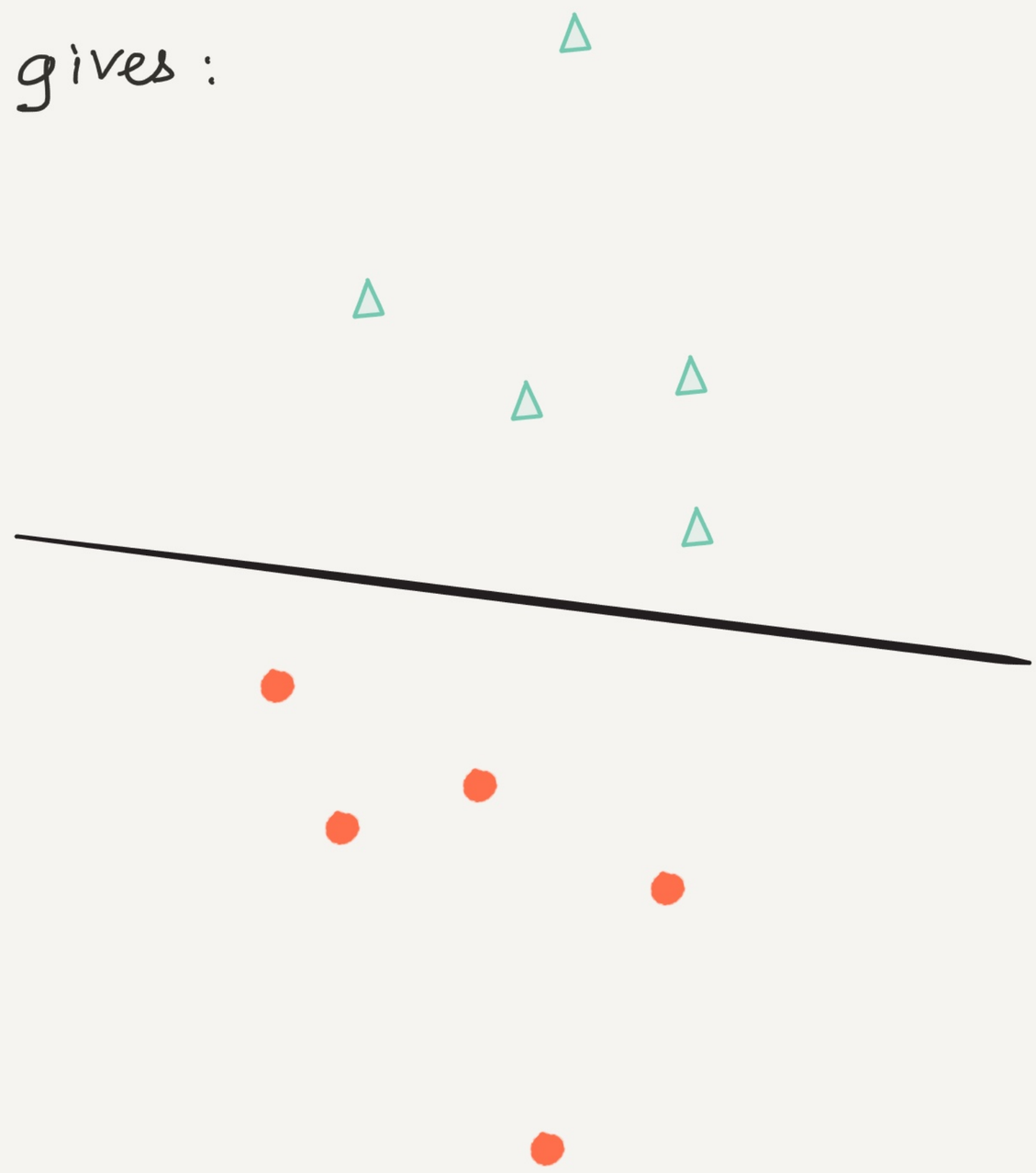
* $J(\theta) = \|\theta\|_1$
* $J(\theta) = \frac{1}{2}\|\theta\|_2^2$

SUPPORT VECTOR MACHINES

Towards the notion of "margin"

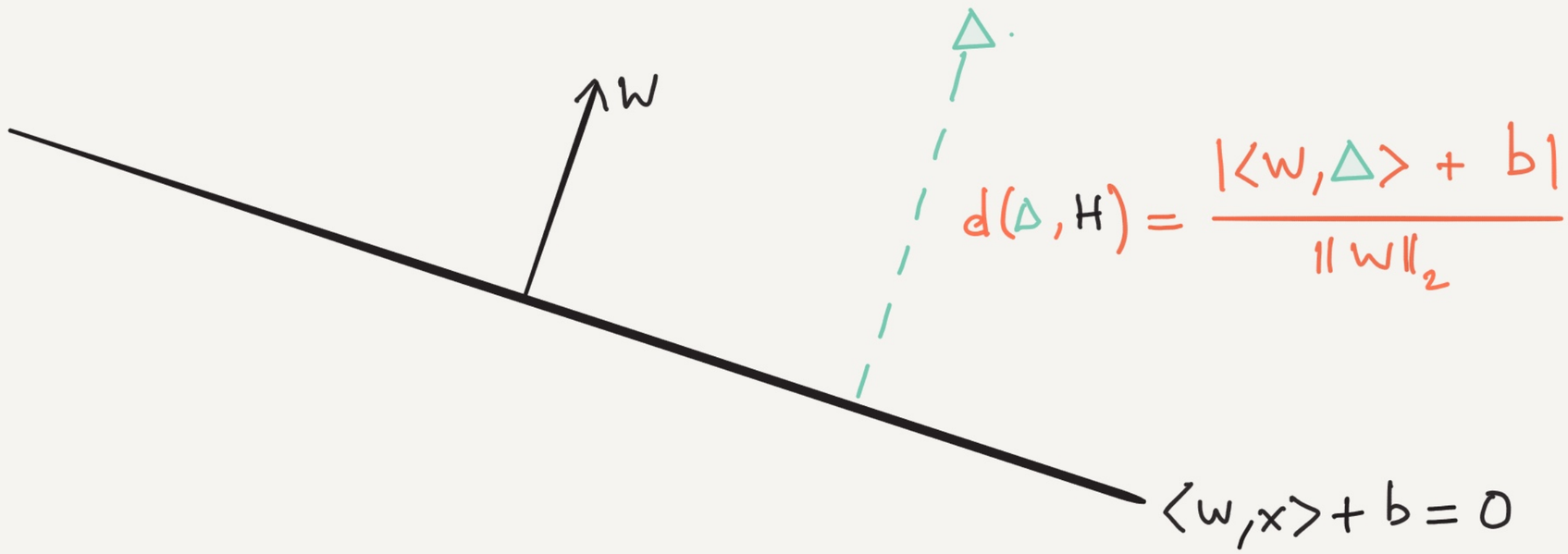
Logistic Regression gives:





2 Geometrical viewpoint

Linear Separation: $H = \{x \in \mathcal{X} : \langle w, x \rangle + b = 0\}$



$$\text{Margin} = \min_{1 \le i \le n} \frac{y_i(\langle w, x_i \rangle + b)}{\|w\|_2}$$

Maximum Margin: $\sup_{w,b} \min_i \frac{y_i(\langle w, x_i \rangle + b)}{\|w\|_2}$

3. convexification

$$\sup_{w,b} \min_{i} \frac{y_i(\langle w, x_i\rangle + b)}{\|w\|_2}$$

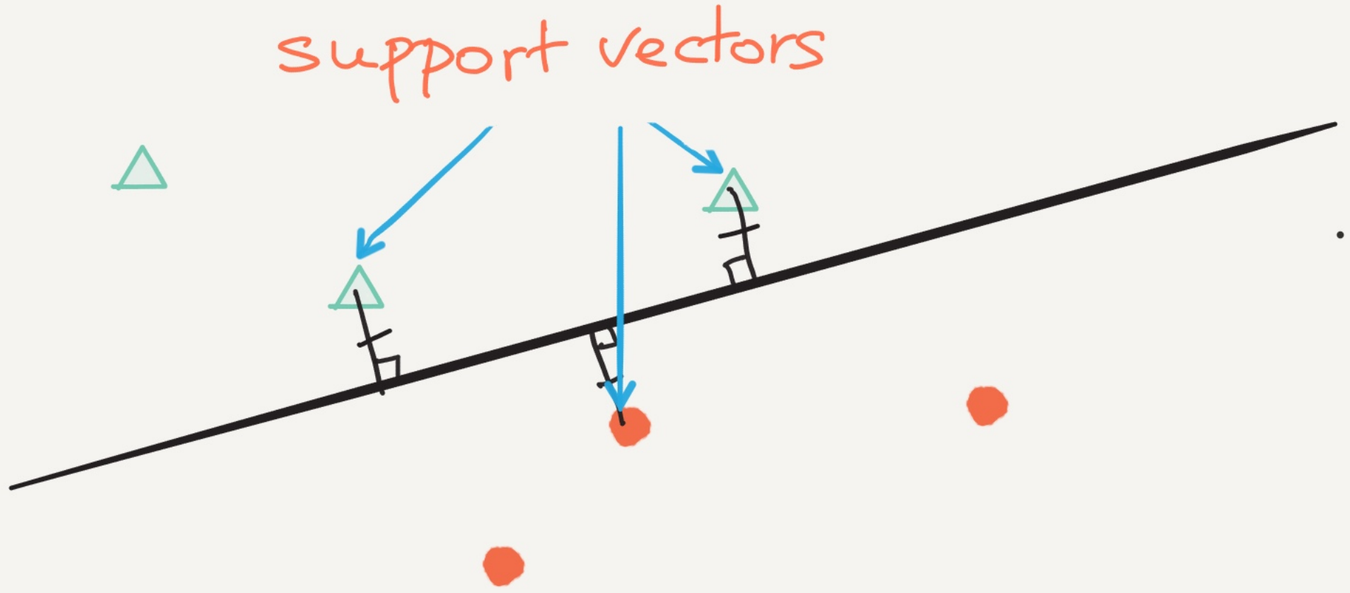
is a difficult problem to solve, invariant to $(w,b) \longrightarrow (\lambda w, \lambda b)$, $\lambda > 0$

Assume $(x_i, y_i)_{1\le i\le n}$ are linearly separable:

$$\min_i \; y_i(\langle w, x_i\rangle + b)/\|w\|_2 > 0.$$

We can always reparametrize to have: $y_i(\langle w, x_i\rangle + b) \geq 1$

Points (x_i, y_i) such that $y_i(\langle w, x_i\rangle + b) = 1$ are called



Which gives:

$$\sup_{w,b} \frac{1}{\|w\|_2}$$
$$\text{sc } \forall i \; y_i(\langle w, x_i \rangle + b) \geq 1$$

And, finally, the convex problem:

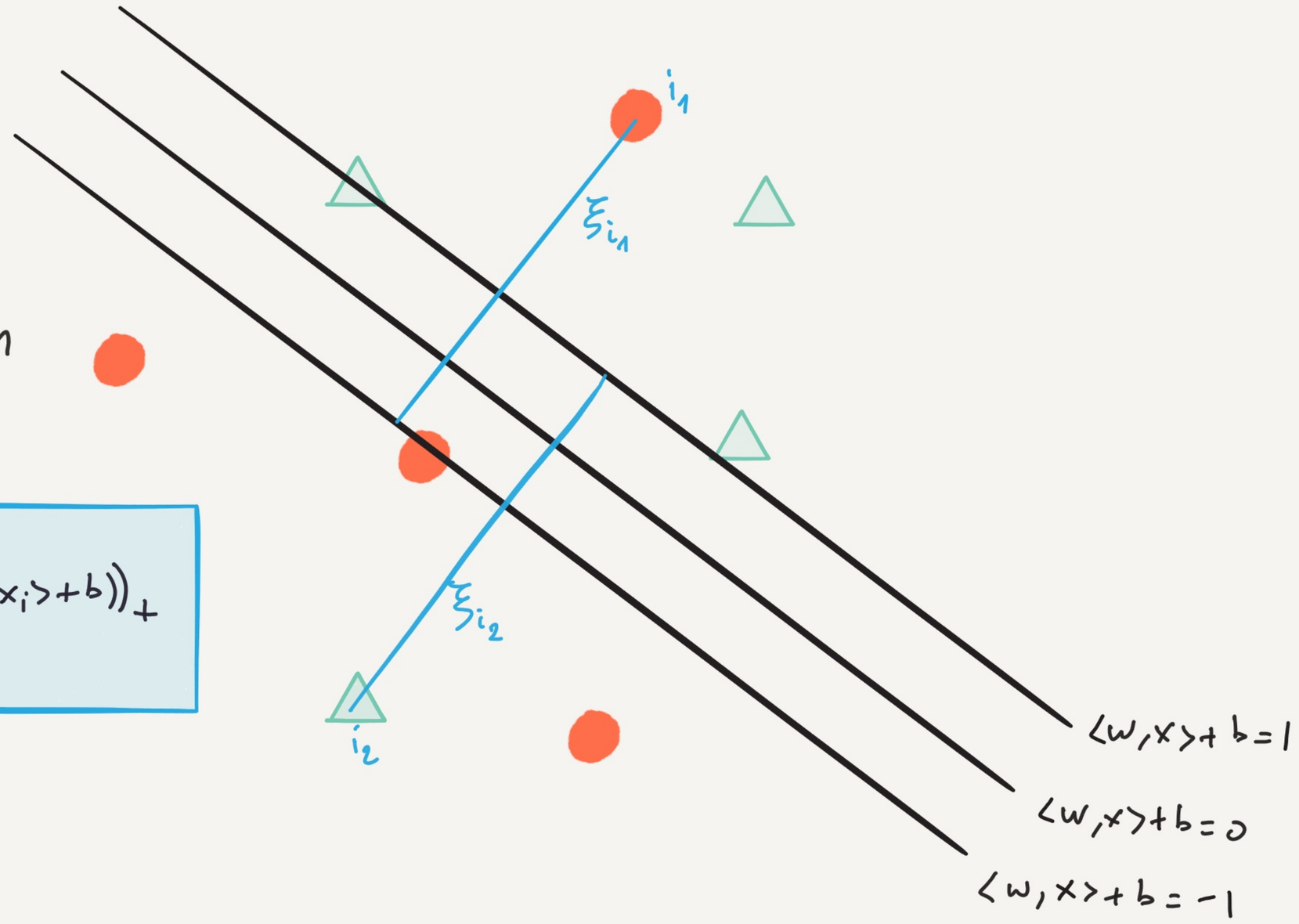
$$\inf \frac{1}{2}\|w\|_2^2$$
$$\text{sc } \forall i \; y_i(\langle w, x_i \rangle + b) \geq 1$$

4. Non linearly separable case

We use what is called **Slack Variables** to penalize each misclassified point with a price $C\xi_i$ (C is a parameter)

Optimization Problem becomes:

$$\inf \ \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}(1-y_i(\langle w,x_i\rangle+b))_+$$



5. Formal Comparison with Logistic Regression

SVM

$$\inf_{w,b} \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle - y_i b)_+ + \lambda \frac{\|w\|^2}{2}$$

RIDGE LOGISTIC REGRESSION

$$\inf_{w,b} \sum_{i=1}^{n} \log\left[1 + \exp(-y_i(\langle w, x_i \rangle + b))\right] + \lambda \frac{\|w\|^2}{2}$$

Comparons alors $(1-z)_+$ et $\log(1+\exp(-z))$

6. Dual Formulation

Lagrangian Duality

Start with:

$$\inf_{x} f(x)$$
$$\text{s.t. } \forall i \; g_i(x) \leq 0$$

which becomes:

$$\inf_{x} \sup_{\alpha \geq 0} \; f(x) + \sum \alpha_i g_i(x)$$

Define:

$$\mathcal{L}(x, \alpha) = f(x) + \sum_i \alpha_i g_i(x)$$

The following holds:

$$\inf_{x} \sup_{\alpha \geq 0} \mathcal{L}(x, \alpha) \geq \sup_{\alpha \geq 0} \inf_{x} \mathcal{L}(x, \alpha)$$

Quantity:

$$\inf_{x} \sup_{\alpha \geq 0} \mathcal{L}(x, \alpha) - \sup_{\alpha \geq 0} \inf_{x} \mathcal{L}(x, \alpha)$$

is called "Duality Gap". If it is 0, we speak of strong duality

For SVM, we admit the result that strong duality holds. It comes from convexity.

Lagrangian function writes:

$$\mathcal{L}(w,b,\xi;\alpha,\beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[y_i(\langle w,x_i\rangle + b) - 1 + \xi_i\right] - \sum_{i=1}^{n}\beta_i\xi_i$$

Dual function writes:

$$F(\alpha,\beta) = \inf_{w,b,\xi} \mathcal{L}(w,b,\xi;\alpha,\beta)$$

KKT conditions are:

$$\nabla_w \mathcal{L} = 0 \quad \Longrightarrow \quad w = \sum \alpha_i y_i x_i$$

$$\nabla_b \mathcal{L} = 0 \quad \Longrightarrow \quad \sum \alpha_i y_i = 0$$

$$\nabla_{\xi_i} \mathcal{L} = 0 \quad \Longrightarrow \quad \alpha_i + \beta_i = C$$

which gives:

$$F(\alpha,\beta) = \frac{1}{2}\left\|\sum \alpha_i y_i x_i\right\|^2 - \sum_{i,j}\alpha_i\alpha_j y_i y_j \langle x_i,x_j\rangle - \sum_i \alpha_i y_i b + \sum \alpha_i$$

$$= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \langle x_i,x_j\rangle$$

Finally:

$$\sup_{\alpha \succcurlyeq 0,\, \beta \succcurlyeq 0} F(\alpha,\beta) = \sup_{0 \le \alpha \le C} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

We realize that α^*, β^* do not depend on (x_i) but on $\langle x_i, x_j \rangle$
The problem is therefore equivariant to isometries.

This remark opens the way to kernel methods

KERNEL METHODS

1. Definitions

Déf. A positive kernel over a set $\mathcal{X}$ is a fonction $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying:

- $\forall x \in \mathcal{X}, \forall x' \in \mathcal{X} \quad K(x, x') = K(x', x)$
- $\forall n \geq 1, \forall x_1, \dots, x_n,$ $\begin{pmatrix} K(x_1,x_1) & \cdots & K(x_1,x_n) \\ \vdots & & \vdots \\ K(x_n,x_1) & \cdots & K(x_n,x_n) \end{pmatrix}$ is a positive semi definite matrix

Examples.

$K(x, x') = \langle x, x' \rangle$ est un noyau (sym. pos.)

$K(x, x') = (\langle x, x' \rangle + 1)^2$ est un noyau (sym. pos.)

Plus généralement $K(x, x') = \langle \phi(x), \phi(x') \rangle$ est un noyau (sym. pos)

We have the following fundamental result (Aronsajn, 1950)

Reproducing Kernel Hilbert Space.

Let K be a positive kernel over $\mathcal{X}$, there exists a Hilbert Space $\mathbb{H}$ and a map

$$\Phi : \mathcal{X} \longrightarrow \mathbb{H}$$

called the "feature map" such that:

$$\forall x \in \mathcal{X} \quad \forall x' \in \mathcal{X} \qquad K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathbb{H}}$$

and:

$$\forall h \in \mathbb{H} \quad \forall x \in \mathcal{X} \qquad h(x) = \langle h, K(x, \cdot) \rangle$$

For instance, with $K(x, x') = (\langle x, x' \rangle + 1)^2$ and $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$

one may take

$$\phi(x) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ 1 \end{pmatrix}$$ and we check that: $\langle \phi(x), \phi(x') \rangle = (\langle x, x' \rangle + 1)^2$

We also check that after applying ϕ, the following problem

△ $a = (-1, 1)$ ● $b = (1, 1)$ △ $\phi(a) = (1, 1, \sqrt{2}, -\sqrt{2}, -\sqrt{2}, 1)$ △ $\phi(b) = (1, 1, \sqrt{2}, \sqrt{2}, \sqrt{2}, 1)$

● $d = (-1, -1)$ △ $c = (1, -1)$ ● $\phi(d) = (1, 1, -\sqrt{2}, -\sqrt{2}, \sqrt{2}, 1)$ ● $\phi(c) = (1, 1, -\sqrt{2}, \sqrt{2}, -\sqrt{2}, 1)$

becomes linearly separable

$(-1,1)$ O D $(1,1)$

$(-1,-1)$ D O $(1,-1)$