2) Contrast

From now on the true parameter is denoted $\underbrace{\theta_0}$ it is unknown from the statistician

Remember last time,

<u>MLE</u> $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \; \ell_\theta(X)$

$= \underset{\theta \in \Theta}{\operatorname{argmin}} \; -\ell_\theta(X)$

In the gaussian linear models

$\hat{m} = \underset{m \in V}{\operatorname{argmin}} \; \|Y - m\|^2$ $\qquad \mathcal{N}(m, \sigma^2)$

In general a contrast is a function $C(\theta, X)$ where $\theta \in \Theta$ and X is the observation and such that

$$\boxed{E_{\theta_0}\left(C(\theta, X)\right) \text{ is minimal in } \theta = \theta_0}$$

Then the estimator defined by the minimum of the contrast is

$$\boxed{\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\ C(\theta, X)}$$

a) - log - likelihood

Let us prove that $C(\theta, X) = -\ell_\theta(X)$ is a contrast

$$\mathbb{E}_{\theta_0}\left(-\ell_\theta(X)\right) = \mathbb{E}_{\theta_0}\left(-\log f_\theta(X)\right)$$

where f_θ is the density of your model with parameter θ.

$$= \int -\log\left[f_\theta(x)\right] f_{\theta_0}(x)\, \underline{dx}$$

$\hookrightarrow$ is a canonical notation for a $x \in \mathbb{R}^d$
$dx = dx_1 \ldots dx_d$

The Kullback Leibler divergence is defined by

f, g are two densities (or p.d.f)

$$K(f,g) = \int \log\left(\frac{f(x)}{g(x)}\right) f(x)\,dx$$

$$= \mathop{E}_{X\sim f}\left[\log\left(\frac{f(X)}{g(X)}\right)\right]$$

In fact

$$\int \left[\frac{g(x)}{f(x)} - \log\left(\frac{g(x)}{f(x)}\right) - 1\right] f(x)\,dx$$

$$= \underbrace{\int \frac{g}{f} f}_{\int g = 1} + \int \left(-\log \frac{g}{f}\right) f + \underbrace{\int (-1)\, f}_{-\int f = -1} = \int \log \frac{f}{g}\, f = K(f,g)$$

$$\boxed{\begin{array}{l} \int f = 1 \\ \int g = \int \frac{g}{f} f = 1 \end{array}}$$

$$\frac{g(x)}{f(x)} - \log\frac{g(x)}{f(x)} - 1 = e^{u} - u - 1 \quad \text{with } u = \log\frac{g(x)}{f(x)}$$

But $u \longmapsto e^{u} - u - 1 = h(u)$

$$h'(u) = e^{u} - 1$$

		0	
h'	$-$		$+$
h	$\searrow$	$h(0) = e^{0} - 0 - 1 = 0$	$\nearrow$

$$K(f,g) = \int h\left(\log\left[\frac{g(x)}{f(x)}\right]\right) f(x)\,dx$$

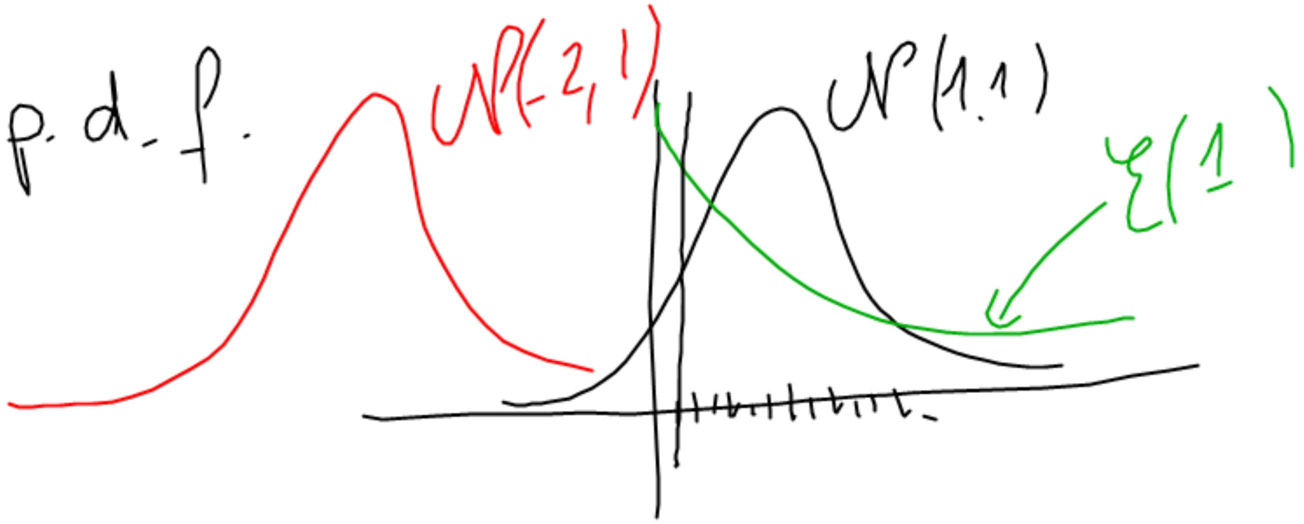
$$= E_{X \sim f}\left(h\left(\log\frac{g(X)}{f(X)}\right)\right)$$

So $\begin{cases} K(f,g) \text{ is always } \geq 0 \\ K(f,g) = 0 \text{ iff } \forall x,\ \log\dfrac{g(x)}{f(x)} = 0 \text{ which means } \dfrac{g(x)}{f(x)} = 1 \text{ or } g(x) = f(x) \end{cases}$

So

$$\boxed{\begin{gathered} K(f,g) = E_{X\sim f}\left(\log \frac{f(X)}{g(X)}\right) \\ \begin{cases} \text{is always } \geq 0 \\ K(f,g) = 0 \text{ iff } f(x) = g(x)\ \forall x \end{cases} \end{gathered}}$$

So the Kullback Leibler divergence measures a distance between densities. The same applies to p.d.f.



Let's go back to $c(\theta, X) = -l_\theta(X)$

$$\theta \mapsto E_{\theta_0}\big(c(\theta, X)\big) = E_{\theta_0}\big(-\log f_\theta(X)\big)$$

$$= \underbrace{E_{\theta_0}\Big(\log\big(f_{\theta_0}(X)\big)\Big) + E_{\theta_0}\big(-\log f_\theta(X)\big)}_{E_{\theta_0}\left(\log \frac{f_{\theta_0}(X)}{f_\theta(X)}\right) = K(f_{\theta_0}, f_\theta)} \underbrace{- E_{\theta_0}\big(\log f_{\theta_0}(X)\big)}_{\text{this doesn't depend on } \theta}$$

So $E_{\theta_0}\big(-l_\theta(X)\big)$ is minimal when $K(f_{\theta_0}, f_\theta)$ is null, hence when $f_{\theta_0}(x) = f_\theta(x)\ \forall x$.

If there is no problem of identifiability (meaning that 2 different θ's encode 2 different densities) then this implies that $\theta = \theta_0$.

$$X \sim P \begin{cases} \xrightarrow{} \text{disc} & f(x) \quad x \in \mathcal{X} \text{ discret} \\ \xrightarrow{} C^0 & f(x) \text{ density.} \end{cases}$$

$$E(h(X)) = \begin{cases} \xrightarrow{} \text{disc} & \sum_{x \in \mathcal{X}} h(x) \underbrace{f(x)}_{P(X=x)} \\ \xrightarrow{} C^0 & \int h(x) f(x) dx \end{cases}$$

b) least-square contrast

$$\mathbb{R}^d \ni X = \underbrace{\theta_0}_{\substack{\in \mathbb{R}^d \\ \in V \subsetneq \mathbb{R}^d}} + \varepsilon$$

$\in V \subsetneq \mathbb{R}^d$ (for instance regression)

$V \subset \mathbb{R}^d$ and $V \neq \mathbb{R}^d$

ε noise $\underline{E(\varepsilon) = 0}$.

$E_{\theta_0}(X) = \theta_0$ (for each coordinate)

If we do not specify the dist° of ε we cannot compute MLE

But you can use the constrast:

$$C(\theta, X) = \|X - \theta\|^2$$

Let us verify that this is a contrast.

$$2\sum_i \theta_i X_i$$

$$\|$$

$$E_{\theta_0}\big(C(\theta, X)\big) = E_{\theta_0}\left(\|X-\theta\|^2\right) = E_{\theta_0}\left(\|X\|^2 - 2\langle \theta, X\rangle + \|\theta\|^2\right)$$

$$= \underbrace{E_{\theta_0}\left(\|X\|^2\right)}_{\text{does not depend on } \theta} - 2\langle \theta, \underbrace{E_{\theta_0}(X)}_{\theta_0}\rangle + \|\theta\|^2$$

$$E_{\theta_0}\left(\|X-\theta\|^2\right) = \underbrace{\overset{\text{no}}{\cancel{\theta}}\ \text{here}}_{E_{\theta_0}(\|X\|^2)} - 2\cdot\langle \theta,\theta_0\rangle + \|\theta\|^2$$

$$= \underbrace{\|\theta-\theta_0\|^2}_{\|\theta\|^2 - 2\langle\theta,\theta_0\rangle + \|\theta_0\|^2} \underbrace{- \|\theta_0\|^2}_{\text{no } \theta \text{ here either}} + (\text{no } \theta \text{ here})$$

So this is minimal when $\theta = \theta_0$ and $\theta \mapsto \|X-\theta\|^2$ is a contrast

(it's known as the least square contrast for vectors)

So I can define $\hat{\theta} = \underset{\theta \in V}{\operatorname{argmin}} \|X-\theta\|^2$

$$= \Pi_V X$$

c) least-square contrast for densities

$X_1, \dots, X_n$ iid with density f_0.

$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$

Let f be a candidate density and let us look at

$$C(f, X) = -\frac{2}{n} \sum_{i=1}^{n} f(X_i) + \int [f(x)]^2 \, dx$$

$$\mathbb{E}_{\text{all } X_i \sim f_0} \left(C(f, X) \right) = -\frac{2}{n} \sum_{i=1}^{n} \int f(x) f_0(x) \, dx + \int [f(x)]^2 \, dx$$

$$= -2 \int f(x) f_0(x) \, dx + \int [f(x)]^2 \, dx$$

$$= \int (f(x) - f_0(x))^2 \, dx - \int [f_0(x)]^2 \, dx$$

So $C(f, X)$ is minimal when $\int \left(f(x) - f_0(x)\right)^2 dx$ is minimal

But $\int \left(f(x) - f_0(x)\right)^2 dx$ is $\geqslant 0$

and $= 0$ iff $\forall x,\ f(x) = f_0(x)$.

So $C(f, X)$ is a contrast, called the least-square contrast for densities

III Choice of models

Ex (from Neuroscience)

$$\longrightarrow \underbrace{Y_i}_{\text{firing rate}} = f_0(\underbrace{W_i}_{\text{weight}}) + \varepsilon_i \qquad \varepsilon_i \text{ iid } \mathcal{N}(0,\sigma^2)$$

model 1

$f(W) = a + bW$, a, b unknown

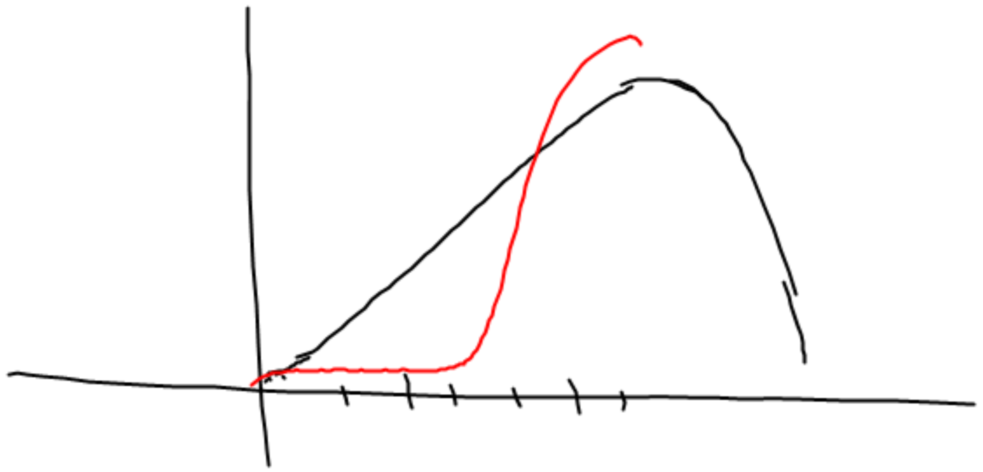
$f(W) = a + bW + cW^2$

a, b, c unknown

$\vdots$

$f(W) = a_0 + a_1 W + a_2 W^2 + \dots + a_d W^d$

(model of dim $d+1$)



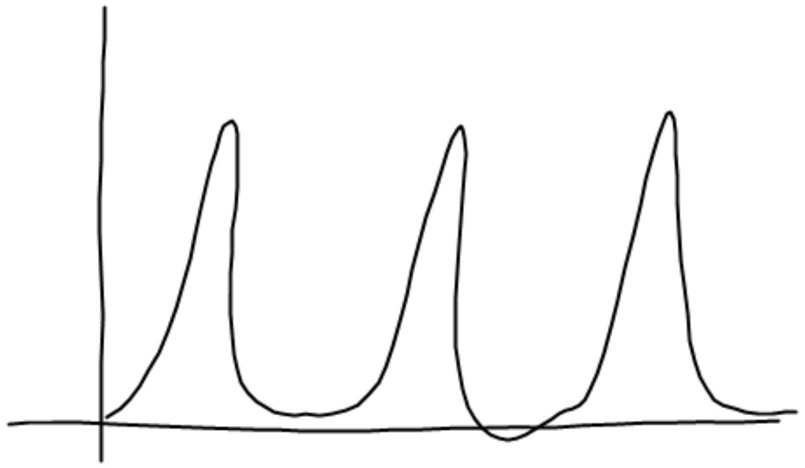
$\longrightarrow \quad Y_i = f_0(U_i) + \varepsilon_i \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

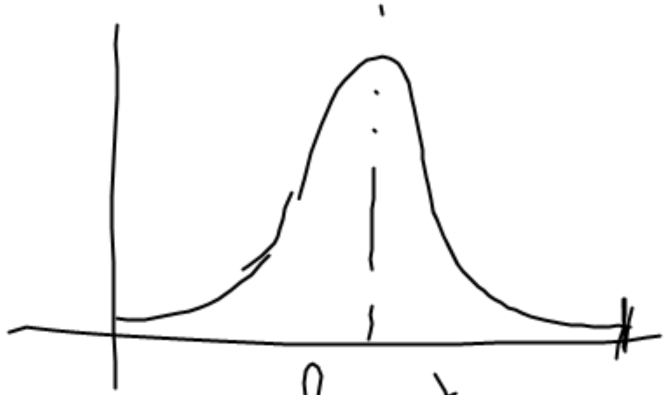
U_i in the angle of the movement

$U_i \in [0, 2\pi]$

model 1

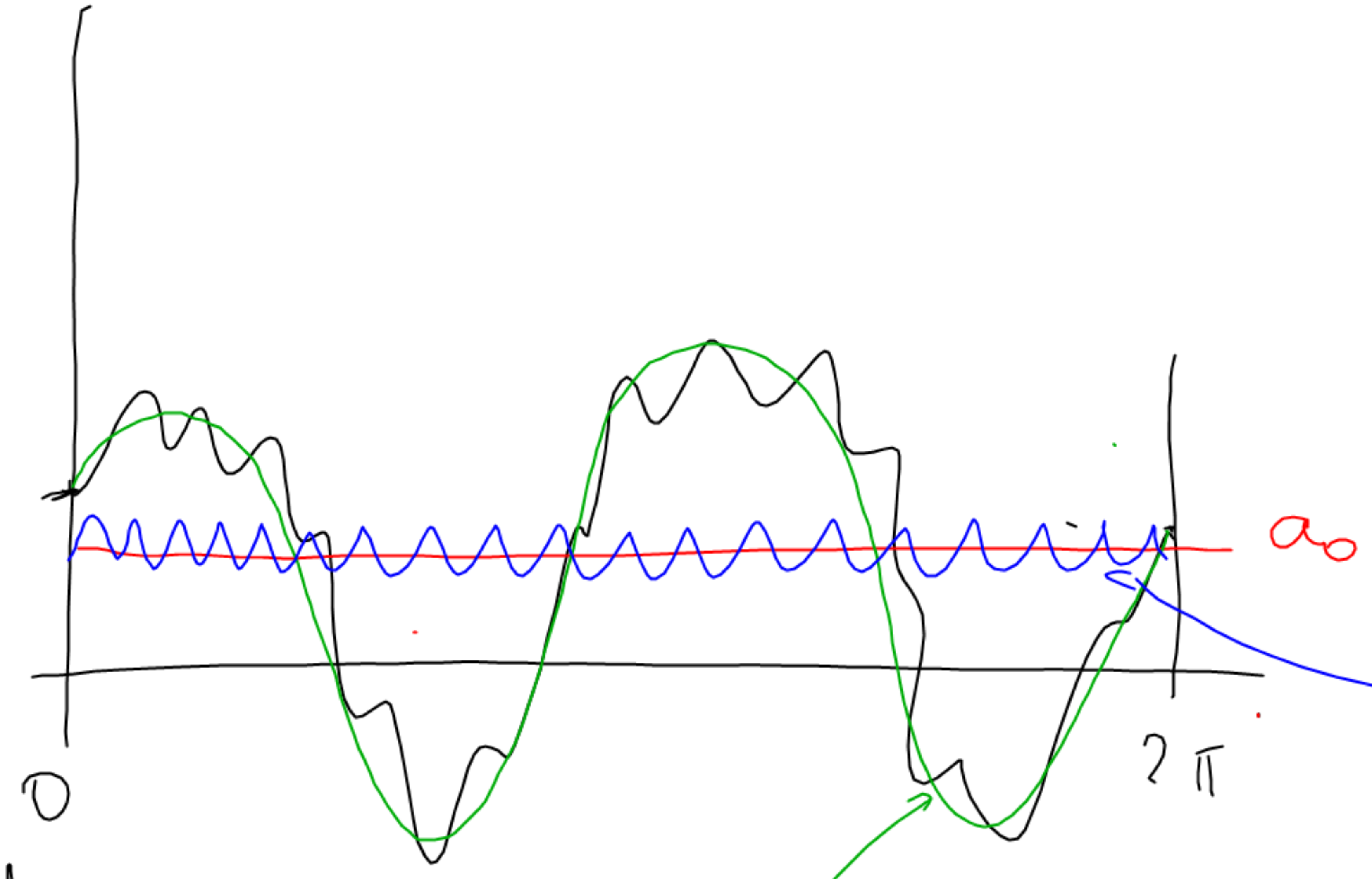
$$f(U_i) = a + b \cos(U_i)$$

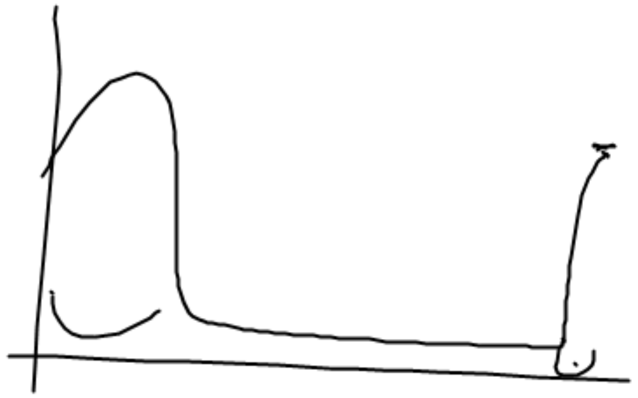




favorite direction of the cell

$$f(U_i) = a_0 + a_1 \cos(U_i) + a_{-1} \sin(U_i) + \dots + a_d \cos(dU_i) + a_{-d} \sin(dU_i)$$





f (smooth periodic)

$\uparrow_{d\to+\infty}$

$$f_d(u) = a_0 + \sum_{k=1}^{d} a_k \cos(ku) + a_{-k} \sin(ku)$$

$$a_i = \int \frac{f(u)\cos(iu)\,du}{2\pi} \qquad a_{-i} = \int \frac{f(u)\sin(iu)\,du}{2\pi}$$

$a_k \cos(ku) + a_{-k}\sin(ku)$

higher frequency

$a_\ell \cos(\ell u) + a_{-\ell} \sin(\ell u)$

black = red + green + blue

In general we have a bunch of linearly independent functions

$\varphi_1(x) \ldots \varphi_d(x)$

$i = 1 \ldots n$

and you look at the problem $Y_i = f_0(X_i) + \varepsilon_i$

$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

<u>model of dim d</u> is $f(X) \in \mathrm{Vect}\left(\varphi_1(X) \ldots \varphi_d(X)\right) = V \subsetneq \mathbb{R}^n$

$$f(X_i) = a_1 \varphi_1(X_i) + \ldots + a_d \varphi_d(X_i) \quad \forall i.$$

Of course you will stop before n but when?

for this model, you know that $\begin{pmatrix} \vdots \\ \hat{f}_d(X_i) \\ \vdots \end{pmatrix} = \Pi_V(Y)$

$$V = \mathrm{Vect}\left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \right) \longleftrightarrow \begin{array}{l} \varphi_1(x) = 1 \\ \varphi_2(x) = x \end{array}$$

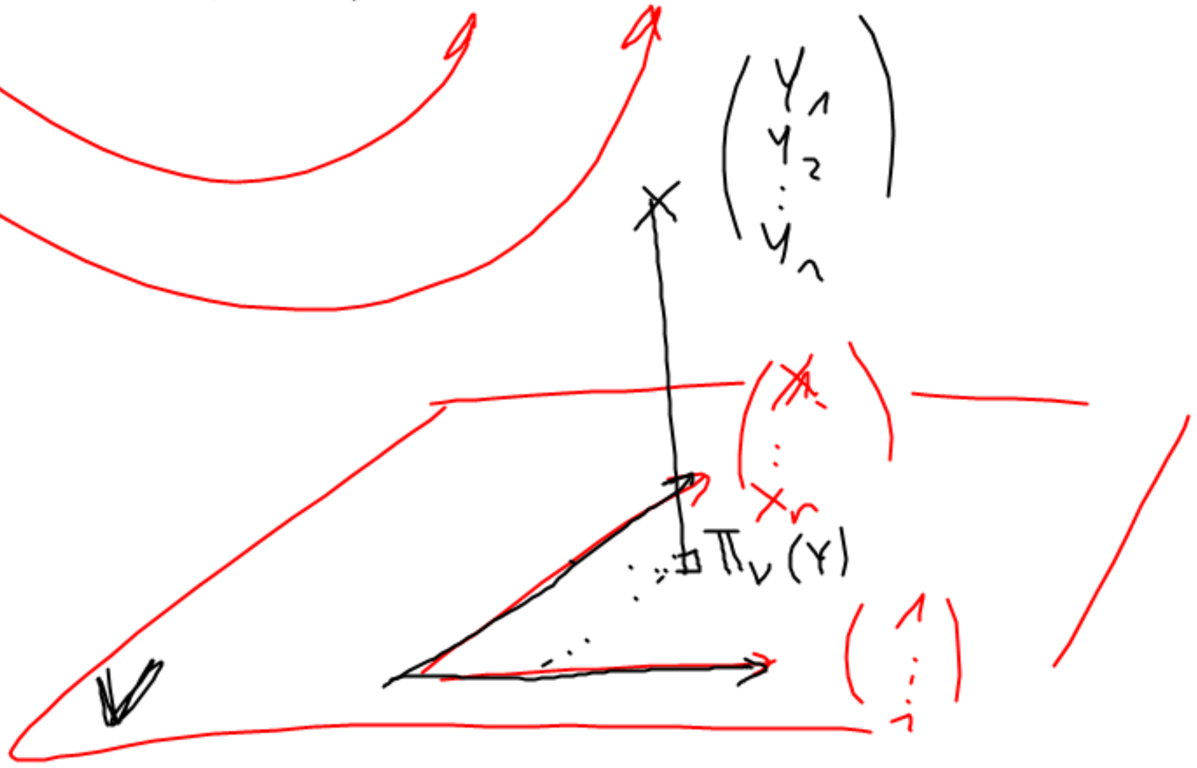
$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$f(x) = \underline{\underline{a}} + \underline{\underline{b}}\, x$$

$$\underbrace{\Pi_V Y}_{} = \hat{a} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \hat{b} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

orthogonal projection of Y on V

$\Pi_V Y \in \mathbb{R}^n$



1/ Bias-Variance decomposition

So $Y_i = f_0(X_i) + \varepsilon_i$

$\longrightarrow$ for model of dim d I have $\begin{pmatrix} \hat{f}_d(X_1) \\ \vdots \\ \hat{f}_d(X_n) \end{pmatrix} = \hat{f}_d(X)$

What is the d for which

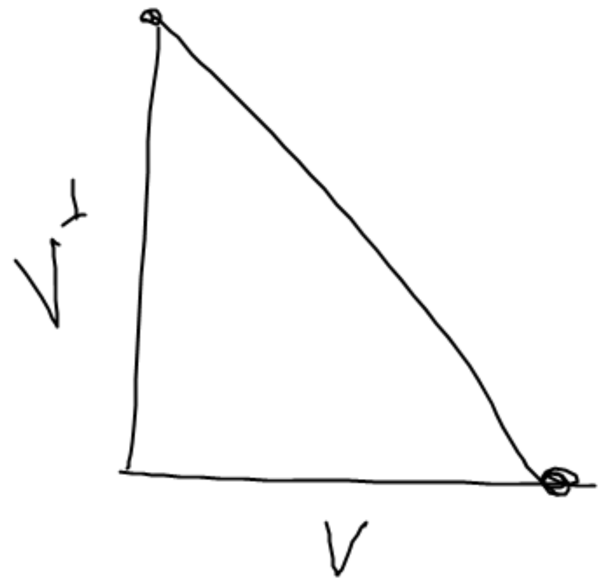
$$\mathbb{E}_{Y \sim f_0}\left(\| f_0(X) - \hat{f}_d(X) \|^2 \right) \text{ is the smallest ?}$$

$\longrightarrow$ this should give me the best d.

$$\mathbb{E}_0\left(\|f_0(X)-\hat{f}_d(X)\|^2\right)$$

$$=\mathbb{E}_0\left(\|f_0(X)-\Pi_V Y\|^2\right)$$

$$=\mathbb{E}_0\left(\|\underbrace{f_0(X)-\Pi_V f_0(X)}_{\perp V}+\underbrace{\Pi_V f_0(X)-\Pi_V Y}_{\in V}\|^2\right)$$



$$=\mathbb{E}_0\left(\boxed{\|f_0(X)-\Pi_V f_0(X)\|^2}\right)+\underbrace{\mathbb{E}_0\left(\|\Pi_V \varepsilon\|^2\right)}_{\sigma^2 d}$$

BIAS TERM decreases when $d \uparrow$

$\sigma^2 d$ (since $\|\Pi_V \varepsilon\|^2 \sim \sigma^2 \chi^2(d)$)

VARIANCE TERM increases with d

To choose a good d you need a trade off between

- complexity of the model (to have a small bias) $\longrightarrow$ depends on f_0
- variability of each of the coefficients ($\to$ want a small variance)

$\Rightarrow$ We define an oracle $\tilde{d}$ which a benchmark

$$\tilde{d} = \underset{\substack{d \ldots \\ d \in [1..n-1]}}{\operatorname{argmin}} \left(\| f_0(X) - \Pi_{V^d} f_0(X) \|^2 + \sigma^2 d \right)$$

2/ Mallow's C_p

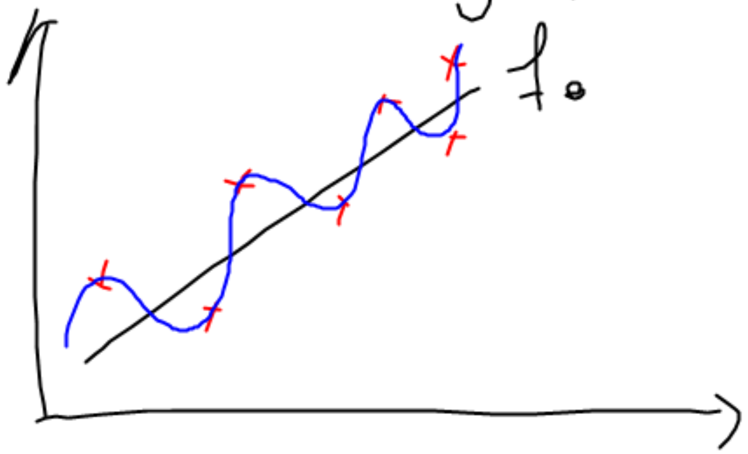
This consists in choosing

$$\hat{d} = \underset{d}{\mathrm{argmin}} \left(\underbrace{\|Y - \Pi_{V_d} Y\|^2}_{\text{least-square}} + \underbrace{2\sigma^2 d}_{\text{penalty}} \right)$$

(we assume that σ is known)

If there wasn't any penalty $\rightarrow$ $\hat{d} = \underset{d}{\mathrm{argmin}} \left(\|Y - \Pi_{V_d} Y\|^2 \right)$

$= \text{largest } d$



$\rightarrow$ you are overfitting the data

A penalty is always there to avoid overfitting

Mallow's C_p is a particular case which satisfies an oracle inequality

$$E\left(\|f_0(X)-\hat{f}_{\hat{d}}(X)\|^2\right) \leq C \underbrace{\min_d \left[E\left(\|f_0(X)-\Pi_{V_d} f_0(X)\|^2\right)+\sigma^2 d\right]}_{\text{Oracle risk}}$$

$\hat{d}$: the choice with Mallow's C_p

$$\hat{d} = \operatorname{argmin}_d \left(\|Y-\Pi_{V_d} Y\|^2 + 2 d \sigma^2\right)$$

C is very close to 1.

3) Akaike's criterion (AIC)

Akaike Information criterion

model M parametrized by Θ_M $\longrightarrow$ MLE $\hat{\theta}_M$ $\rightarrow$ $f_{\hat{\theta}_M}(X)$

(eg $\mathcal{N}(m,\sigma^2)$) $\qquad (\hat{m},\hat{\sigma}^2) \rightarrow f_{\hat{m},\hat{\sigma}^2}(x)$

density of X when param = MLE

the likelihood on the model M at the param $\hat{\theta}_M$

AIC criterion is

$$\hat{M} = \underset{M \in \mathcal{M}}{\operatorname{argmin}} \left\{ \left(-\log f_{\hat{\theta}_M}(X) \right) + \underbrace{\dim(M)}_{\text{nb of parameters inside } M} \right\}$$

oracle inequalities exist for that too, as long as there are few models with the same number of parameters

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \underset{\sim g}{\text{iid}}$$

model M model g^M of g

$$f(X) = g(X_1)\cdots g(X_n)$$

$$\hat{\theta}_M = \underset{\theta \in \Theta_M}{\operatorname{argmax}} \sum_{i=1}^{n} g_\theta^M(X_i)$$

$$\hat{M} = \operatorname{argmin}\left(-\sum_{i=1}^{n} g^M_{\hat{\theta}_M}(X_i) + \underbrace{\dim(M)}\right)$$

$\to \mathcal{N}(m,\sigma^2) \to 2$

$\to \mathcal{E}(\lambda) \to 1$

$$M \leftrightarrow \mathcal{N}(m,\sigma^2) \quad \theta_M = \begin{pmatrix} m \\ \sigma^2 \end{pmatrix}$$

$$\hat{\theta}_M = \begin{pmatrix} \bar{X} \\ \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \end{pmatrix}$$

$$M \leftrightarrow \mathcal{E}(\lambda) \quad \theta_M = \lambda$$

$$\hat{\theta}_M = \frac{1}{\bar{X}}$$

4/ BIC criterion

$$Y_i = a_0 + a_1 X_i^1 + \dots + a_p X_i^p + \varepsilon_i \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

To do variable selection I could put in competition

$$V_0 = \mathrm{Vect}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \qquad V_1 = \mathrm{Vect}(X^1) \quad \dots \quad V_{p} = \mathrm{Vect}(X^d)$$

$$V_{01} = \mathrm{Vect}\left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, X^1\right) \qquad V_{ij} = \mathrm{Vect}(X^i, X^j) \longleftarrow$$

number of models with dim 2

$$\vdots \qquad\qquad \vdots \qquad\qquad \frac{p(p-1)}{2}$$

$$V_{0\dots d} = \mathrm{Vect}\left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, X^1, X^2, \dots X^d\right)$$

write them all

$\longrightarrow$ AIC and Mallow's cannot work: too many models with the same dimension

BIC

$$\hat{M} = \underset{M \in \Omega}{\operatorname{argmin}} \left[\frac{\ln(n)\dim(M)}{2} - \log\left(f_{\hat{\theta}_M}(X)\right) \right]$$

we need to penalize more ...

Be careful

The more models in competition, the largest the penalty, and you may end with a model of small dimension.

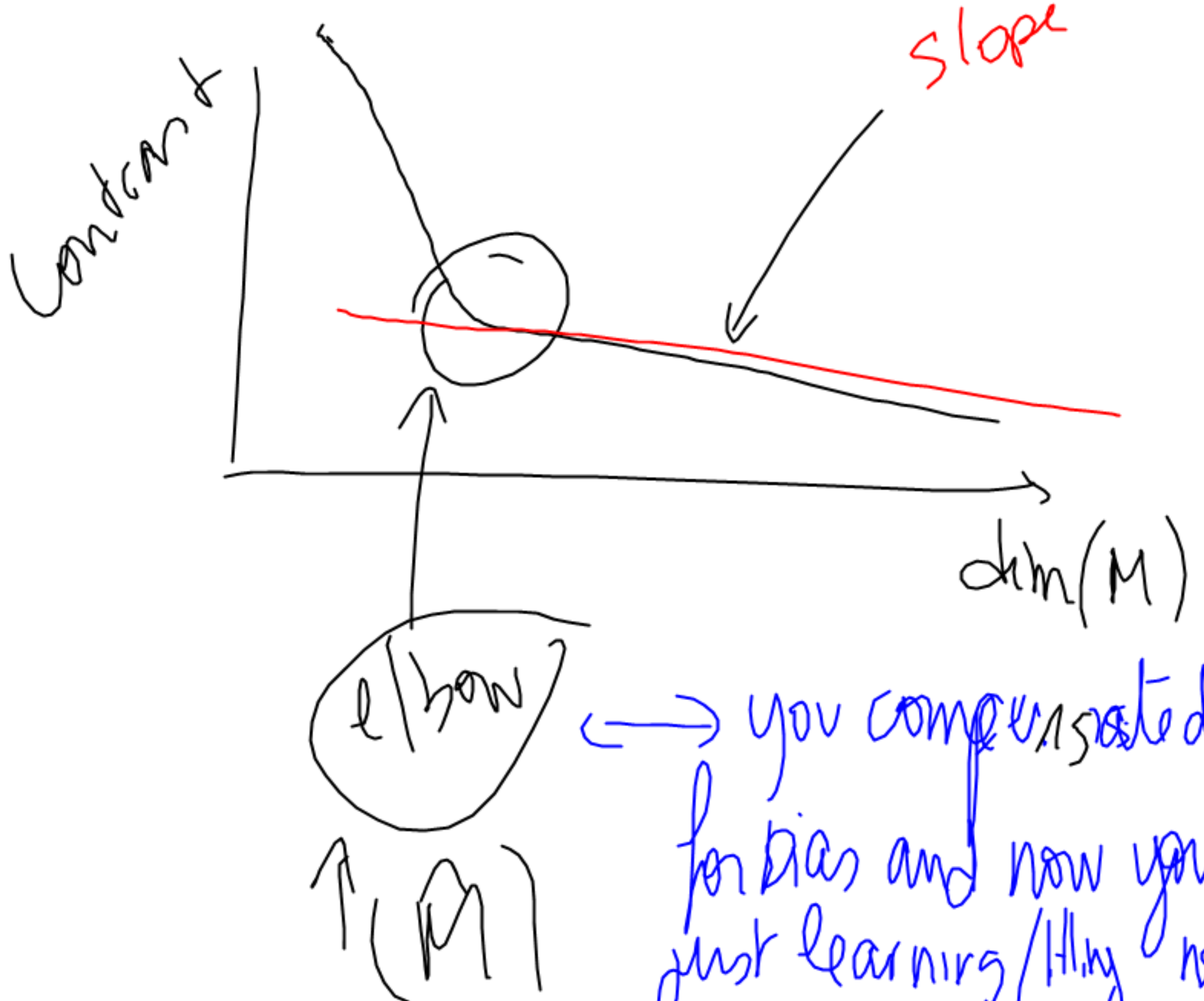
whereas if the nb of models was not too big, you could have use AIC and select it.

Also there are other penalties

→ Lasso

→ Ridge

→ slope heuristic



⟷ you compensated for bias and now you are just learning/fitting noise

→ for people who want more details in math:

Birgé and Massart

"Gaussian model selection"