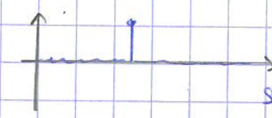
10/07/2017 - FLORENT K   STATISTICAL ESTIMATION, MACHINE LEARNING AND COMPRESSED SENSING AN

goo.gl/VdGvYf old lectures notes
goo.gl/PLJNaR exercises -

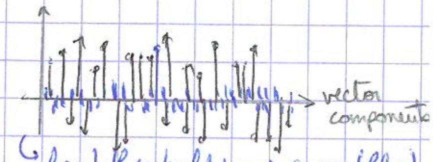
LECTURE I : DENOISING & STATISTICS

I a a simple example.

A denoising toy problem:

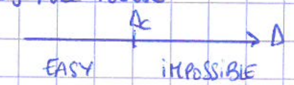


but observe



↳ hard thresholding → easy (blue)
↳ signal lost in the noise (black)

It seems that decreasing slowly the noise we go through a threshold value of the noise below which the denoising becomes easy after having been impossible



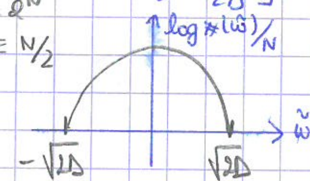
Formalisation: signal of length N
How components will have value $\in [w, w+\Delta w] \Rightarrow$ on avg. $N \times \frac{e^{-w^2/2\Delta}}{\sqrt{2\pi\Delta}} \Delta w$

$\propto \exp\left[\ln N - \frac{w^2}{2\Delta}\right]$

$\tilde{w} = \frac{w}{\sqrt{\ln N}}$ $\exp \ln N \left[1 - \frac{\tilde{w}^2}{2\Delta}\right]$

if $|\tilde{w}| > \sqrt{2\Delta}$: # of such components $\xrightarrow{N \to +\infty} 0$ $\equiv$ REM $\begin{cases} N \equiv 2^N \\ \Delta \equiv N/2 \end{cases}$

↳ in the thermodynamic limit, we will typically observe no component generated randomly with magnitude larger than $\sqrt{\ln N \, 2\Delta}$



↳ $\boxed{\Delta_c = \frac{1}{2\ln N}}$ → seems to be matching the little experiments limit on detectability!
(Donoho Johnston 1998 - universal denoiser threshold).

I b denoising of a single variable

Consider a random variable X distributed according to $P_X(x)$. We want to transmit it through a noisy channel. Send ground truth $X^* \sim P_X(x^*)$

- Receive $Y = \sqrt{\lambda} X^* + Z$ $\begin{cases} \lambda = \text{signal to noise ration (SNR)} \\ Z \sim \mathcal{N}(0, I) \end{cases}$

AWGN: Additive white Gaussian noise -

first case: receiving many measurements: $\begin{cases} y_1 = \sqrt{\lambda} x^* + z_1 \\ y_2 = \sqrt{\lambda} x^* + z_2 \\ \vdots \\ y_N = \sqrt{\lambda} x^* + z_N \end{cases} \Rightarrow \vec{y}$

we implicitely defined the likelihood = conditional probability: $P_{Y|X}(y|x) = \frac{e^{-\frac{(\sqrt{\lambda}x - y)^2}{2}}}{\sqrt{2\pi}}$

$\Rightarrow P_{\vec{Y}|X}(\vec{y}|x) = \prod_{i=1}^{N} P_{Y|X}(y_i|x) = \frac{1}{(2\pi)^{N/2}} e^{-\sum_i (\sqrt{\lambda}x - y_i)^2/2}$ → proba on Y and not on X!

↳ maximum likelihood estimator: $\hat{x}(\vec{y}) = \underset{x}{\operatorname{argmax}} \; P_{\vec{Y}|X}(\vec{y}|x) = \underset{x}{\operatorname{argmax}} \log P_{\vec{Y}|X}(\vec{y}|x)$

↳ in our case: $= \underset{x}{\operatorname{argmax}} \sum_i -\frac{(\sqrt{\lambda}x - y_i)^2}{2} + \text{cste}$

$= \frac{1}{N\sqrt{\lambda}} \sum_i y_i \xrightarrow[N \to +\infty]{} x^*$ (exercise).

The maximum likelihood scheme always works when $N \to +\infty$ - (Laplace 1810, Fisher 1912 ..)

second case: receiving only one measurement:

we'll need Bayes theorem: $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$ | Laplace 1774, Bayes, price 1763 -

→ our problem:

posterior distribution: $P(X|Y) = \dfrac{P(Y|X)\,P(X)}{P(Y)}$ — likelihood: $P(Y|X)$; prior distribution: $P(X)$; normalization, "evidence": $P(Y)$

The tricky part: most of the time you have no idea how data was generated...

$$\hookrightarrow P_{X|Y}(x|y) = \frac{1}{Z(y)} e^{-\frac{(\sqrt{\lambda}x - y)^2}{2\Delta}} P(x) = \frac{1}{Z(y)} \exp\left(-\log p_{Y|X}(y|x) - \log P_X(x)\right) \equiv \frac{e^{-\mathcal{H}(x,y)}}{Z(y)}$$

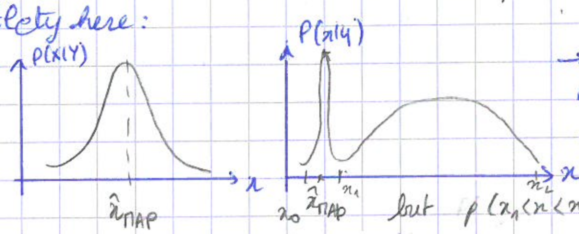
with $\begin{cases} \mathcal{H}(x,y) = \log p_{Y|X}(y|x) + \log P_X(x) \\ Z(y) = \int dx\, e^{-\mathcal{H}(x,y)} \end{cases}$

→ and we made the stat mech connection. most probable ≡ ground state.

Now from the single measurement the most probable value of $\boxed{\hat{x}_{map} = \arg\max_x P_{X|Y}(x|y)}$

↳ MCMC, with decreasing temperature to find the ground state :)

Nevertheless, there is a subtlety here:
imagine the two scenari:



but $p(x_1 < x < x_2 | y)$ might $> p(x_0 < x < x_1 | y)$

→ we need a principled way to take probability mass into account.

<u>minimum mean squared error</u>: $\hat{x}_{MMSE}(y)$ such that $(\hat{x}_{MMSE} - x^*)^2$ be minimum.

Consider the <u>posterior risk</u>: $R = \int P(X|Y)(\hat{x}-x)^2 dx$ average value of squared error.

Minimize it : $\left.\dfrac{\partial R}{\partial \hat{x}}\right|_{\hat{x}_{MMSE}} = 0 \Rightarrow \boxed{\hat{x}_{MMSE} = \int dx\, x\, P_{X|Y}(x|y)}$

↳ think x = spin under Gibbs distribution. Good estimator = magnetization! $\hat{x}_{MMSE} = \langle x \rangle$

<u>Rk</u>: we could consider other risks, other moments of the gap. The best strategy is problem dependent ($|x - \hat{x}| \Rightarrow \hat{x}$ = median).

Yet again, you might prefer to minimize the number of errors: $1 - \delta_{\hat{x},x^*}$.
The associated risk is: $R = \int P(X|Y)(1 - \delta_{\hat{x},x^*})\,dx \Rightarrow \boxed{\hat{x} = \arg\max_x P(x|y)}$
<u>Bayes optimal error</u>

difference with MAP for multi components

$\hat{x}_i = \arg\max_{x_i} P(x_i|y)$ BAYES OPTIMAL ERROR

$\vec{x}_{MAP} = \arg\max_{\vec{x}} P(\vec{x}|y)$ ≠ MAP

back to our specific problem: Redefining conviniently the partition sum $Z(y)$:

$$P(x|y) = \frac{1}{Z(y)} \exp\left[-\frac{\lambda x^2}{2} + y\sqrt{\lambda}x\right] P_X(x).$$

i/ the minimal mean squared error $MMSE = \mathbb{E}_{Y,x^*}\left[(\underbrace{\hat{x}_{MMSE}}_{\equiv\langle x\rangle} - x^*)^2\right]$ — "disorder average"

$$MMSE = q + q_0 - 2m \quad \text{with} \begin{cases} q = \mathbb{E}_{Y,x^*}(\langle x\rangle^2) \\ m = \mathbb{E}_{Y,x^*}(\langle x\rangle x^*) \\ q_0 = \mathbb{E}_{Y,x^*}(x^{*2}) \end{cases}$$

↳ interpretation in terms of replicas: OVERLAP OF 2 replicas $q = \mathbb{E}_{Y,x^*}\left(\langle x^{(1)}\rangle\langle x^{(2)}\rangle\right)$

ii/ <u>NISHIMORI RELATIONS</u>: $\mathbb{E}_{Y,x^*}[\langle f(x,x^*)\rangle] = \mathbb{E}_Y[\langle f(x^{(1)},x^{(2)})\rangle]$

<u>csq</u>: $m = q \Rightarrow MMSE = q_0 - m$.

<u>proof</u>: $\int dy\,dx^*\, P(y,x^*) \int dx\, P(x|y) f(x,x^*) = \int dy\, P(y) \int dx\, dx^*\, P(x^*|y) P(x|y) f(x,x^*)$

$= \int dy\, P(y) \int dx_1 dx_2 \ldots$