On the use of information measures & performance bounds

By Antoine Roueff

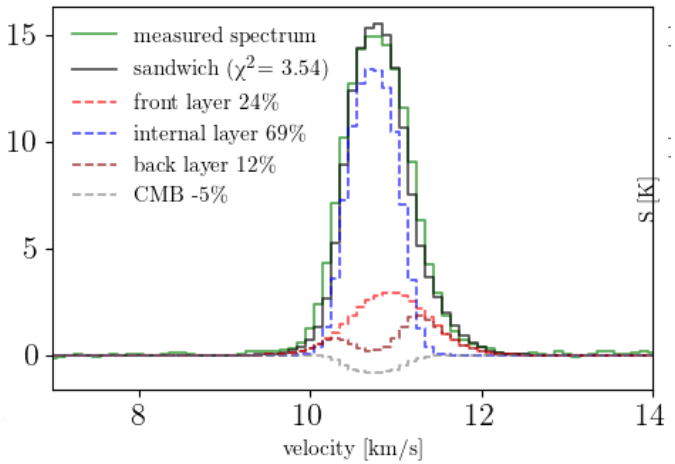
Institut Matériaux Microélectronique Nanosciences Provence

UNIVERSITÉ DE TOULON

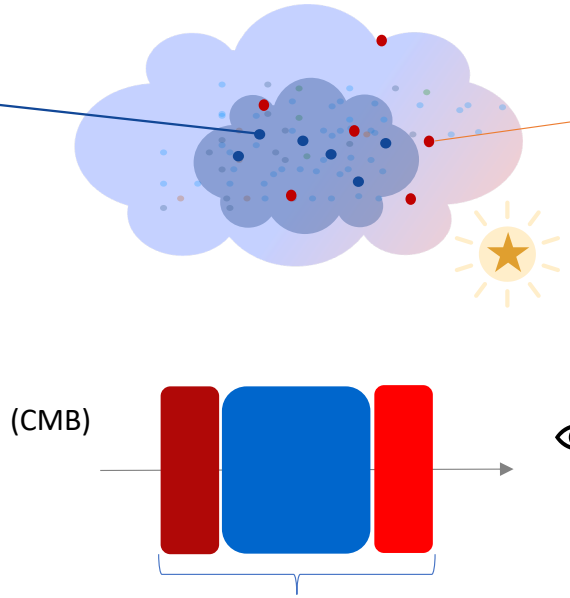
Context. Estimation of kinetic temperature and volume density in a molecular cloud (GMC) from observed spectra.

We have multi-species observations:

$^{13}CO\ (1-0)$

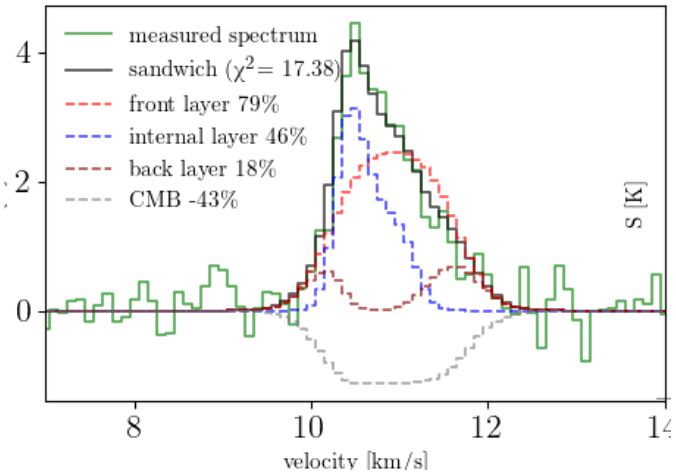


Photons emitted from the **internal** layer



3 layer model

$HCO^+\ (1-0)$



Emitted from both **internal** & **external** layers

Measurement model

$$x = c.m(\theta) + b$$

$m(\theta)$: radiative transfer model

θ: C_V, FWHM, T_{kin}, n_{H2}, …

c & b : multiplicative and additive noise

How to find θ from x knowing m and some statistical properties on c & b?

→ "*Toward a robust physical and chemical characterization of heterogeneous lines of sight*" A&A, 692, A160 (2024)

A problem for astrophysicists: how to find the best θ estimates ?

Measurement model: $x = c.m(\theta) + b$

Since c & b are unpredictable, they are considered as realization of random variables.

$\rightarrow x$ is also a realization of random variable noted X.

The probability density function (pdf) of X is named the *true* pdf. It is noted p and is inaccessible.

Based on physical assumptions on c & b, one can have a probabilistic model for X noted $q_X(x;\theta)$.

Let's note θ_{true} the "true" value of the parameter θ.

From an observation x, one defines an estimator of θ noted $\hat{\theta}$, which is a function of x: $\hat{\theta}(x) \approx \theta_{\text{true}}$

When the measurement model does not matches reality, θ_{true} may not exist.

If the probabilistic model $q_X(x;\theta_{\text{true}})$ is a good approximation of the true pdf p, then the distribution of $\hat{\theta}(X) - \theta_{\text{true}}$ characterizes the estimation error. Its mean is the bias, its standard deviation is the precision (also the of error bar of $\hat{\theta}(X)$).

$\rightarrow$ How to build $\hat{\theta}(x)$?

How to build $\hat{\theta}(x)$? Several possible techniques

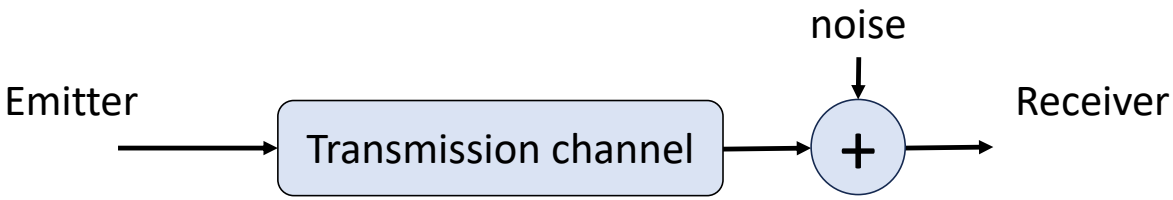
1. Moment estimator: one replaces the theoretical mean of X noted $E[X]$ by the empirical mean $\frac{1}{N}\sum_n x_n$. If $E[X] \approx m(\theta)$ and m is inversible, then $\hat{\theta}_{\text{moment}}(x_1, x_2, \dots, x_N) = m^{-1}\left(\frac{1}{N}\sum_n x_n\right)$. Pros: simple (no fit required). Cons: m needs to be inversible.
2. Weighted Least Square estimator: $\hat{\theta}_{\text{WLS}}(x_1, x_2, \dots, x_N) = \arg\min_\theta \sum_n \frac{1}{\sigma_n^2}(x_n - m_n(\theta))^2$. Pros: m does need to be inverted. Cons: it may requires in iterative technique (e.g. Newton Raphson) to find the $\arg\min$.
3. Maximum Likelihood estimator: one assumes that $(x_n)_n$ are independent realizations of X distributed along $q_X(x;\theta)$ $\hat{\theta}_{\text{MLE}}(x_1, x_2, \dots, x_N) = \arg\max_\theta \prod_n q_X(x_n;\theta)$. Pros. It takes into account the noise distribution. When q is gaussian → case 2.
4. Bayesian estimator: θ is also considered as a random variable. This allows one to add an a priori knowledge on θ through its pdf $\pi(\theta)$ → regularization of the solution. Pros : allows one to decrease the variance of estimator and to compute its pdf. Cons. Needs an a priori, computation and memory intensive. → see P. Palud's lecture on Bayesian estimation.
5. Machine Learning regression. One trains a generic algorithm (e.g. NN) to learn how to go from x to θ. Pros : does not require *any* physical knowledge on either $m(\theta)$ or the noise. Cons. Needs training on already labelled data and there remains uncertainty on generalization performances. → see A. Paiement & D. Baron's lectures on Machine Learning.

Remark: the best technique is problem dependent. In particular, it depends on the knowledge you have.

Instead of implementing all possible estimators to select the most efficient (which would require Monte Carlo simulations to analyze their performance), I consider 2 alternatives which are independent of the choice of the estimator:

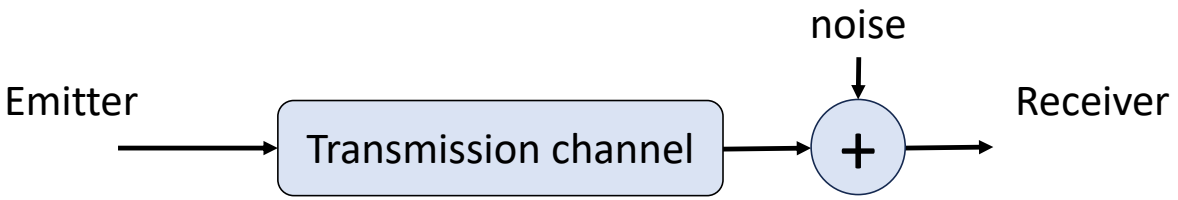
1. Informative measures -> measure dependance between observation and parameters of interest.
2. Performances bounds -> accuracy of the system (without any estimator).

Information theory (Shannon, 1945).



Data are encoded to optimize the communication

Information theory (Shannon, 1945).



Data are encoded to optimize the communication

Let's consider X a source with K possible messages $a_1, a_2, \ldots, a_K$

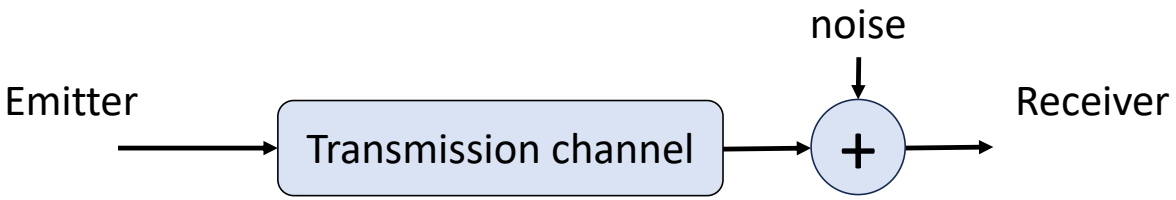
Example for $K = 5$

	a_1	a_2	a_3	a_4	a_5
$P_k = \Pr(X = a_k)$	0.4	0.3	0.2	0.05	0.05
code 1	000	001	011	100	101
code 2	0	10	110	1110	1111
code 3	0	1	10	11	100

$$\mathbb{E}(L_1) = 3 \times 0.4 + 3 \times 0.3 + 3 \times 0.2 + 3 \times 0.05 + 3 \times 0.05 = 3$$
$$\mathbb{E}(L_2) = 1 \times 0.4 + 2 \times 0.3 + 3 \times 0.2 + 4 \times 0.05 + 4 \times 0.05 = 2$$
$$\mathbb{E}(L_3) = 1 \times 0.4 + 1 \times 0.3 + 2 \times 0.2 + 2 \times 0.05 + 3 \times 0.05 = 1.35$$

Remember: $E[h(X)] = \sum_k h(a_k) P_k$

Information theory (Shannon, 1945).



Data are encoded to optimize the communication

Let's consider X a source with K possible messages $a_1, a_2, \ldots, a_K$

Example for $K = 5$

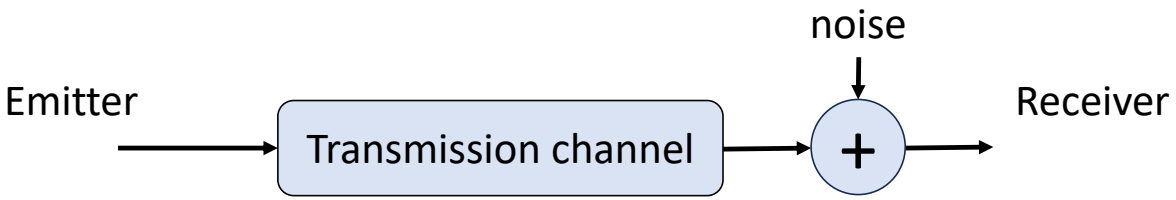
	a_1	a_2	a_3	a_4	a_5
$P_k = \Pr(X = a_k)$	0.4	0.3	0.2	0.05	0.05
code 1	000	001	011	100	101
code 2	0	10	110	1110	1111
code 3	0	1	10	11	100

$$\mathbb{E}(L_1) = 3 \times 0.4 + 3 \times 0.3 + 3 \times 0.2 + 3 \times 0.05 + 3 \times 0.05 = 3$$
$$\mathbb{E}(L_2) = 1 \times 0.4 + 2 \times 0.3 + 3 \times 0.2 + 4 \times 0.05 + 4 \times 0.05 = 2$$
$$\mathbb{E}(L_3) = 1 \times 0.4 + 1 \times 0.3 + 2 \times 0.2 + 2 \times 0.05 + 3 \times 0.05 = 1.35$$

Shannon's theorem?

On average, the minimum code length is equal to the (Shannon) entropy $H(X) = -\sum_{k=1}^{5} P_k \log_2 P_k$ = 1.95

Information theory (Shannon, 1945).



Data are encoded to optimize the communication

Let's consider X a source with K possible messages $a_1, a_2, \ldots, a_K$

Example for $K = 5$

	a_1	a_2	a_3	a_4	a_5
$P_k = \Pr(X = a_k)$	0.4	0.3	0.2	0.05	0.05
code 1	000	001	011	100	101
code 2	0	10	110	1110	1111
code 3	0	1	10	11	100

$$\mathbb{E}(L_1) = 3 \times 0.4 + 3 \times 0.3 + 3 \times 0.2 + 3 \times 0.05 + 3 \times 0.05 = 3$$
$$\mathbb{E}(L_2) = 1 \times 0.4 + 2 \times 0.3 + 3 \times 0.2 + 4 \times 0.05 + 4 \times 0.05 = 2$$
$$\mathbb{E}(L_3) = 1 \times 0.4 + 1 \times 0.3 + 2 \times 0.2 + 2 \times 0.05 + 3 \times 0.05 = 1.35$$

Shannon's theorem?

On average, the minimum code length is equal to the (Shannon) entropy $H(X) = -\sum_{k=1}^{5} P_k \log_2 P_k$ = 1.95

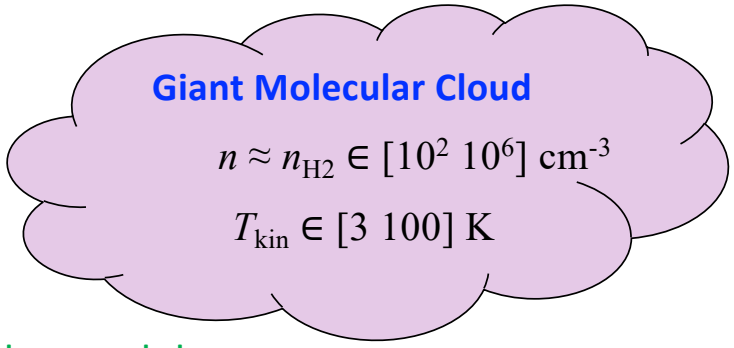
⇒ lossless encoding -> internet -> iPhone -> IoT -> …

What is the relation with astrophysics?

Information theory (Shannon, 1945).

Data are encoded to optimize the communication

Astrophysicists of the ISM want to "understand" what is going in Giant Molecular Clouds (GMC).



NOEMA,
ALMA,
JWST,
…

What do I mean by "understand" ?
Being able to describe as simply as possible the observed data,
to characterize the star formation process.

We (the receiver) observe Y (e.g. molecular lines)

and we want to recover the information X (e.g. the density of the GMC) emitted by the GMC

Entropy, a fruitful concept : statistical physics (1870), communication (1945), data science

Let's consider X a discrete random variable whose values are in $\{a_1, a_2, \dots, a_K\}$ and $P_k = \Pr(X = a_k)$

Let's consider $s_N = \{x_1, x_2, \dots, x_N\}$ a sample of N independent and identically distributed (i.i.d.) realizations of X

Entropy, a fruitful concept : statistical physics (1870), communication (1945), data science

Let's consider X a discrete random variable whose values are in $\{a_1, a_2, \dots, a_K\}$ and $P_k = \Pr(X = a_k)$

Let's consider $s_N = \{x_1, x_2, \dots, x_N\}$ a sample of N independent and identically distributed (i.i.d.) realizations of X

On average, what is M_N number of distinct samples s_N that are generated?

If $\exists k_0$, $P_{k_0} = 1$, then $x_1 = x_2 = \cdots = x_N = a_{k0} \Rightarrow M_N = 1$

(a_{k_0})

Entropy, a fruitful concept : statistical physics (1870), communication (1945), data science

Let's consider X a discrete random variable whose values are in $\{a_1, a_2, \dots, a_K\}$ and $P_k = \Pr(X = a_k)$

Let's consider $s_N = \{x_1, x_2, \dots, x_N\}$ a sample of N independent and identically distributed (i.i.d.) realizations of X

On average, what is M_N number of distinct samples s_N that are generated?

If $\exists k_0$, $P_{k_0} = 1$, then $x_1 = x_2 = \cdots = x_N = a_{k0} \Rightarrow M_N = 1$

If $\forall k, P_k = \frac{1}{K}$, then $M_N = K^N$

Entropy, a fruitful concept : statistical physics (1870), communication (1945), data science

Let's consider X a discrete random variable whose values are in $\{a_1, a_2, \dots, a_K\}$ and $P_k = \Pr(X = a_k)$

Let's consider $s_N = \{x_1, x_2, \dots, x_N\}$ a sample of N independent and identically distributed (i.i.d.) realizations of X

On average, what is M_N number of distinct samples s_N that are generated?

If $\exists k_0$, $P_{k_0} = 1$, then $x_1 = x_2 = \cdots = x_N = a_{k0} \Rightarrow M_N = 1$

If $\forall k, P_k = \frac{1}{K}$, then $M_N = K^N$

In general, M_N

Entropy, a fruitful concept : statistical physics (1870), communication (1945), data science

Let's consider X a discrete random variable whose values are in $\{a_1, a_2, \dots, a_K\}$ and $P_k = \Pr(X = a_k)$

Let's consider $s_N = \{x_1, x_2, \dots, x_N\}$ a sample of N independent and identically distributed (i.i.d.) realizations of X

On average, what is M_N number of distinct samples s_N that are generated?

If $\exists k_0$, $P_{k_0} = 1$, then $x_1 = x_2 = \cdots = x_N = a_{k0} \Rightarrow M_N = 1$ and $H = 0$.

If $\forall k, P_k = \frac{1}{K}$, then $M_N = K^N$ and $H = \log_2 K$.

In general, $M_N = 2^{N\,H(X)}$, where $H(X) = -\sum_{k=1}^{K} P_k \log_2 P_k$. (proof based on Stirling approximation)

Entropy, a fruitful concept : statistical physics (1870), communication (1945), data science

Let's consider X a discrete random variable whose values are in $\{a_1, a_2, \dots, a_K\}$ and $P_k = \Pr(X = a_k)$

Let's consider $s_N = \{x_1, x_2, \dots, x_N\}$ a sample of N independent and identically distributed (i.i.d.) realizations of X

On average, what is M_N number of distinct samples s_N that are generated?

If $\exists k_0$, $P_{k_0} = 1$, then $x_1 = x_2 = \cdots = x_N = a_{k0} \Rightarrow M_N = 1$ and $H = 0$.

If $\forall k, P_k = \frac{1}{K}$, then $M_N = K^N$ and $H = \log_2 K$.

In general, $M_N = 2^{N\,H(X)}$, where $H(X) = -\sum_{k=1}^{K} P_k \log_2 P_k$. (proof based on Stirling approximation)

=> Entropy characterizes the **uncertainty** of X → maximum of entropy = maximum of uncertainty on the value of X

In statistical physics, it characterizes the system *disorder*, *i.e.* the number of configurations (microcanonical)

Take home message:

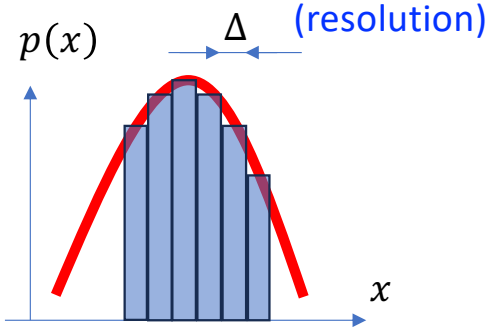
Entropy characterizes the **uncertainty** of X

next

Statistical moment $H(X) = -\sum_{k=1}^{K} P_k \log_2 P_k = -E[\log_2 \mathrm{P}(X)]$

Remember: $E[h(X)] = \sum_k h(a_k) P_k$

$$\text{Entropy} = -E[\log \mathrm{P}(X)]$$



Consider X is a continuous random variable with probability density function (pdf) $p(x)$.

Can we define the entropy of X?

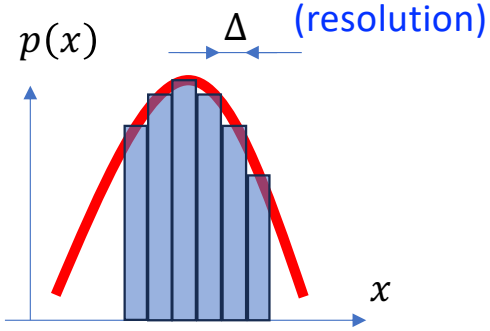
$-\sum_{k=1}^{K} P_k \log_2 P_k$

$\int p(x)\,\mathrm{dx} = 1$

$P_k = \Delta\, p(x_k)$

$\sum_k P_k \approx 1$

$$\text{Entropy} = -E[\log \mathrm{P}(X)]$$



Consider X is a continuous random variable with probability density function (pdf) $p(x)$.

Can we define the entropy of X?

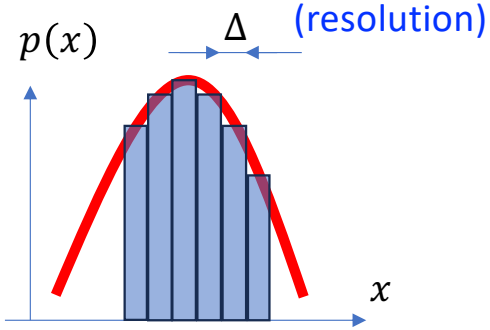
$$-\sum_{k=1}^{K} P_k \log_2 P_k = -\sum_{k=1}^{K} P_k \log_2 \Delta\, p(x_k) = -\sum_{k=1}^{K} \Delta\, p(x_k) \log_2 p(x_k) - \log_2 \Delta$$

$$\int p(x)\, \mathrm{dx} = 1$$

$$P_k \approx \Delta\, p(x_k)$$

$$\sum_k P_k = 1$$

$$\text{Entropy} = -E[\log \mathrm{P}(X)]$$



Consider X is a continuous random variable with probability density function (pdf) $p(x)$.

Can we define the entropy of X?

$$-\sum_{k=1}^{K} P_k \log_2 P_k = -\sum_{k=1}^{K} P_k \log_2 \Delta\, p(x_k) = -\sum_{k=1}^{K} \Delta\, p(x_k) \log_2 p(x_k) - \log_2 \Delta$$

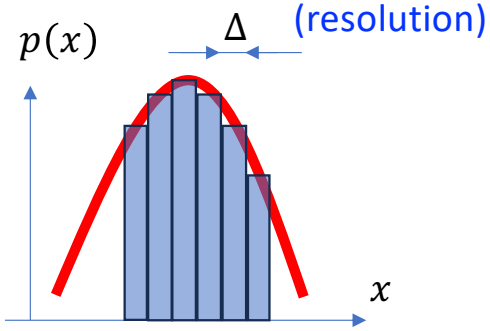
$$\lim_{\Delta \to 0} -\sum_{k=1}^{K} P_k \log_2 P_k = -\int p(x) \log_2 p(x)\, \mathrm{dx} - \log_2(\Delta) = +\infty$$

$$\int p(x)\, \mathrm{dx} = 1$$

$$P_k \approx \Delta\, p(x_k)$$

$$\sum_k P_k = 1$$

$$\text{Entropy} = -E[\log \mathrm{P}(X)]$$



Consider X is a continuous random variable with probability density function (pdf) $p(x)$.

Can we define the entropy of X?

$$-\sum_{k=1}^{K} P_k \log_2 P_k = -\sum_{k=1}^{K} P_k \log_2 \Delta\, p(x_k) = -\sum_{k=1}^{K} \Delta\, p(x_k) \log_2 p(x_k) - \log_2 \Delta$$

$$\lim_{\Delta \to 0} -\sum_{k=1}^{K} P_k \log_2 P_k = -\int p(x) \log_2 p(x)\, \mathrm{dx} - \log_2(\Delta) = +\infty$$

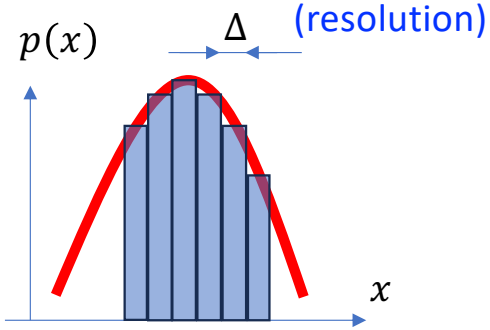
→ We define $h(X) = -\int p(x) \log_2 p(x)\, \mathrm{dx} = -E[\log p(X)]$ → differential entropy.

$$\int p(x)\, \mathrm{dx} = 1$$

$$P_k \approx \Delta\, p(x_k)$$

$$\sum_k P_k = 1$$

$$\text{Entropy} = -E[\log \mathrm{P}(X)]$$



Consider X is a continuous random variable with probability density function (pdf) $p(x)$.

Can we define the entropy of X?

$$-\sum_{k=1}^{K} P_k \log_2 P_k = -\sum_{k=1}^{K} P_k \log_2 \Delta\, p(x_k) = -\sum_{k=1}^{K} \Delta\, p(x_k) \log_2 p(x_k) - \log_2 \Delta$$

$$\lim_{\Delta \to 0} -\sum_{k=1}^{K} P_k \log_2 P_k = -\int p(x) \log_2 p(x)\, \mathrm{dx} - \log_2(\Delta) = +\infty$$

$\rightarrow$ We define $h(X) = -\int p(x) \log_2 p(x)\, \mathrm{dx} = -E[\log p(X)]$ $\rightarrow$ differential entropy.

Let X^{Δ} be a quantified version of X, $h(X) \approx \lim_{\Delta \to 0} H\left(X^{\Delta}\right) + \log_2(\Delta)$

$$\int p(x)\, \mathrm{dx} = 1$$

$$P_k \approx \Delta\, p(x_k)$$

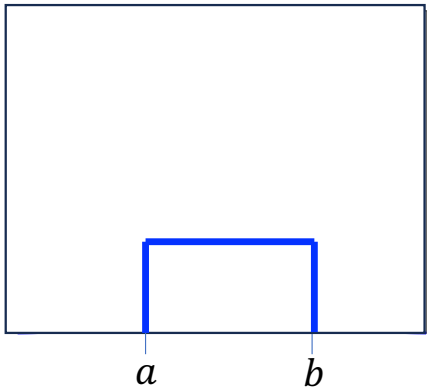
$$\sum_k P_k = 1$$

Entropy $= -E[\log \mathrm{P}(X)]$

Consider X is a continuous random variable with probability density function (pdf) $p(x)$.

Uniform distribution

$X \sim \mathcal{U}\,[a, b]$



Can we define the entropy of X?

$$-\sum_{k=1}^{K} P_k \log_2 P_k = -\sum_{k=1}^{K} P_k \log_2 \Delta\, p(x_k) = -\sum_{k=1}^{K} \Delta\, p(x_k) \log_2 p(x_k) - \log_2 \Delta$$

$$\lim_{\Delta \to 0} -\sum_{k=1}^{K} P_k \log_2 P_k = -\int p(x) \log_2 p(x)\, \mathrm{dx} - \log_2(\Delta) = +\infty$$

→ We define $h(X) = -\int p(x) \log_2 p(x)\, \mathrm{dx} = -E[\log p(X)]$ → differential entropy.

Let X^{Δ} be a quantified version of X, $h(X) \approx \lim_{\Delta \to 0} H\left(X^{\Delta}\right) + \log_2(\Delta)$

Ex: (continuous) uniform distribution $h(X) = \log_2(b - a)$,

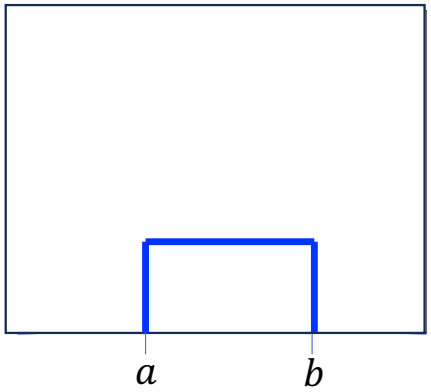
Remember: $0 \log 0 \approx 0$

$$\text{Entropy} = -E[\log \mathrm{P}(X)]$$

Consider X is a continuous random variable with probability density function (pdf) $p(x)$.

Uniform distribution

$X \sim \mathcal{U}\,[a, b]$



Can we define the entropy of X?

$$-\sum_{k=1}^{K} P_k \log_2 P_k = -\sum_{k=1}^{K} P_k \log_2 \Delta\, p(x_k) = -\sum_{k=1}^{K} \Delta\, p(x_k) \log_2 p(x_k) - \log_2 \Delta$$

$$\lim_{\Delta \to 0} -\sum_{k=1}^{K} P_k \log_2 P_k = -\int p(x) \log_2 p(x)\, \mathrm{dx} - \log_2(\Delta) = +\infty$$

→ We define $h(X) = -\int p(x) \log_2 p(x)\, \mathrm{dx} = -E[\log p(X)]$ → differential entropy.

Let X^{Δ} be a quantified version of X, $h(X) \approx \lim_{\Delta \to 0} H\left(X^{\Delta}\right) + \log_2(\Delta)$

Ex: (continuous) uniform distribution $h(X) = \log_2(b - a)$,

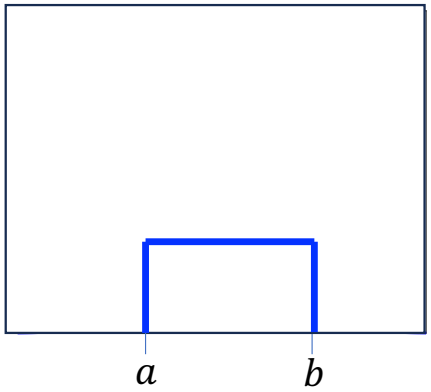
Quantifying with K bins ⇒ (discrete) equiprobable distribution $P_k = \frac{1}{K} \Rightarrow H\left(X^{\Delta}\right) = \log K$

$$\text{Entropy} = -E[\log \mathrm{P}(X)]$$

Consider X is a continuous random variable with probability density function (pdf) $p(x)$.

Uniform distribution

$X \sim \mathcal{U}\,[a, b]$



Can we define the entropy of X?

$$-\sum_{k=1}^{K} P_k \log_2 P_k = -\sum_{k=1}^{K} P_k \log_2 \Delta\, p(x_k) = -\sum_{k=1}^{K} \Delta\, p(x_k) \log_2 p(x_k) - \log_2 \Delta$$

$$\lim_{\Delta \to 0} -\sum_{k=1}^{K} P_k \log_2 P_k = -\int p(x) \log_2 p(x)\, \mathrm{dx} - \log_2(\Delta) = +\infty$$

→ We define $h(X) = -\int p(x) \log_2 p(x)\, \mathrm{dx} = -E[\log p(X)]$ → differential entropy.

Let X^{Δ} be a quantified version of X, $h(X) \approx \lim_{\Delta \to 0} H\left(X^{\Delta}\right) + \log_2(\Delta)$

Ex: (continuous) uniform distribution $h(X) = \log_2(b - a)$,

Quantifying with K bins ⇒ (discrete) equiprobable distribution $P_k = \frac{1}{K}$ ⇒ $H\left(X^{\Delta}\right) = \log K$

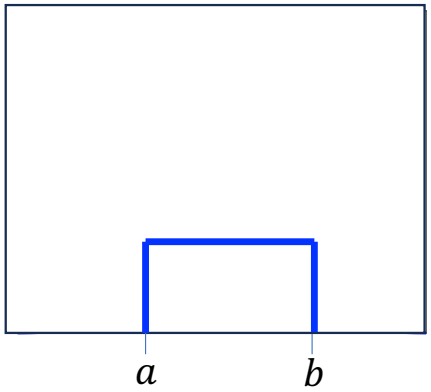
$\Delta = \frac{b-a}{K}$ ⇒ $H\left(X^{\Delta}\right) + \log_2 \Delta = \log K + \log_2 \frac{b-a}{K} = \log_2(b - a) = h(X)$ (as expected)

$$\text{Entropy} = -E[\log \mathrm{P}(X)]$$

Consider X is a continuous random variable with probability density function (pdf) $p(x)$.

Uniform distribution

$X \sim \mathcal{U}\,[a, b]$



Can we define the entropy of X?

$$-\sum_{k=1}^{K} P_k \log_2 P_k = -\sum_{k=1}^{K} P_k \log_2 \Delta\, p(x_k) = -\sum_{k=1}^{K} \Delta\, p(x_k) \log_2 p(x_k) - \log_2 \Delta$$

$$\lim_{\Delta \to 0} -\sum_{k=1}^{K} P_k \log_2 P_k = -\int p(x) \log_2 p(x)\, \mathrm{dx} - \log_2(\Delta) = +\infty$$

→ We define $h(X) = -\int p(x) \log_2 p(x)\, \mathrm{dx} = -E[\log p(X)]$ → differential entropy.

Let X^Δ be a quantified version of X, $h(X) \approx \lim_{\Delta \to 0} H\left(X^\Delta\right) + \log_2(\Delta)$

Ex: (continuous) uniform distribution $h(X) = \log_2(b - a)$,

Quantifying with K bins ⇒ (discrete) equiprobable distribution $P_k = \frac{1}{K} \Rightarrow H\left(X^\Delta\right) = \log K$

$\Delta = \frac{b-a}{K} \Rightarrow H\left(X^\Delta\right) + \log_2 \Delta = \log K + \log_2 \frac{b-a}{K} = \log_2(b - a) = h(X)$ (as expected)

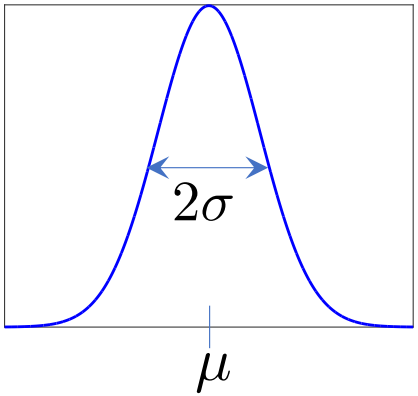
$\Longrightarrow h(X)$ is equivalent to an entropy $H(X)$ up to an offset given by $\log_2(\Delta)$

$$\text{Entropy} = -E[\log \mathrm{P}(X)]$$

Consider X is a continuous random variable with probability density function (pdf) $p(x)$.

Normal distribution

$X \sim \mathcal{N}(\mu, \sigma^2)$



Can we define the entropy of X?

$$-\sum_{k=1}^{K} P_k \log_2 P_k = -\sum_{k=1}^{K} P_k \log_2 \Delta\, p(x_k) = -\sum_{k=1}^{K} \Delta\, p(x_k) \log_2 p(x_k) - \log_2 \Delta$$

$$\lim_{\Delta \to 0} -\sum_{k=1}^{K} P_k \log_2 P_k = -\int p(x) \log_2 p(x)\, \mathrm{dx} - \log_2(\Delta) = +\infty$$

→ We define $h(X) = -\int p(x) \log_2 p(x)\, \mathrm{dx} = -E[\log p(X)]$ → differential entropy.

Let X^{Δ} be a quantified version of X, $h(X) \approx \lim_{\Delta \to 0} H\left(X^{\Delta}\right) + \log_2(\Delta)$

Ex: Gaussian distribution $X \sim \mathcal{N}(\mu, \sigma^2)$

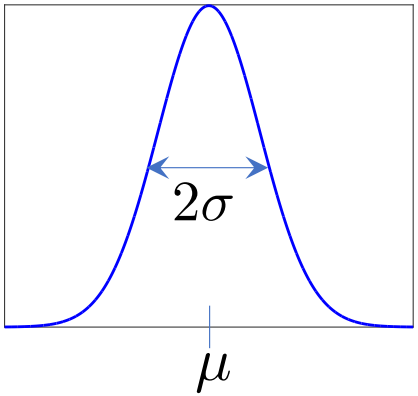
where $\mu = E(X)$ and $\sigma^2 = \mathrm{var}(X)$

$$\text{Entropy} = -E[\log \mathrm{P}(X)]$$

Consider X is a continuous random variable with probability density function (pdf) $p(x)$.

Normal distribution

$X \sim \mathcal{N}(\mu, \sigma^2)$



$$h(X) = \frac{1}{2}\log_2 2\pi e\sigma^2$$

Can we define the entropy of X?

$$-\sum_{k=1}^{K} P_k \log_2 P_k = -\sum_{k=1}^{K} P_k \log_2 \Delta\, p(x_k) = -\sum_{k=1}^{K} \Delta\, p(x_k) \log_2 p(x_k) - \log_2 \Delta$$

$$\lim_{\Delta\to 0} -\sum_{k=1}^{K} P_k \log_2 P_k = -\int p(x) \log_2 p(x)\, \mathrm{dx} - \log_2(\Delta) = +\infty$$

→ We define $h(X) = -\int p(x) \log_2 p(x)\, \mathrm{dx} = -E[\log p(X)]$ → differential entropy.

Let X^Δ be a quantified version of X, $h(X) \approx \lim_{\Delta\to 0} H(X^\Delta) + \log_2(\Delta)$

Ex: Gaussian distribution $X \sim \mathcal{N}(\mu, \sigma^2)$

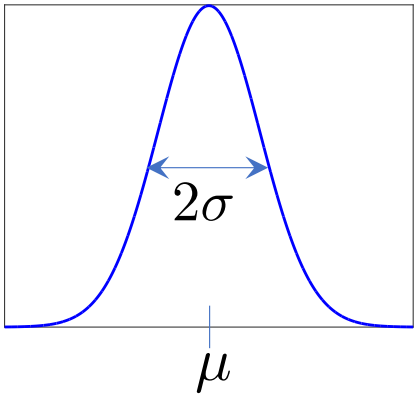
where $\mu = E(X)$ and $\sigma^2 = \mathrm{var}(X)$

→ In the Gaussian case, $h(X)$ is a function of the standard deviation σ.

$$\text{Entropy} = -E[\log \mathrm{P}(X)]$$

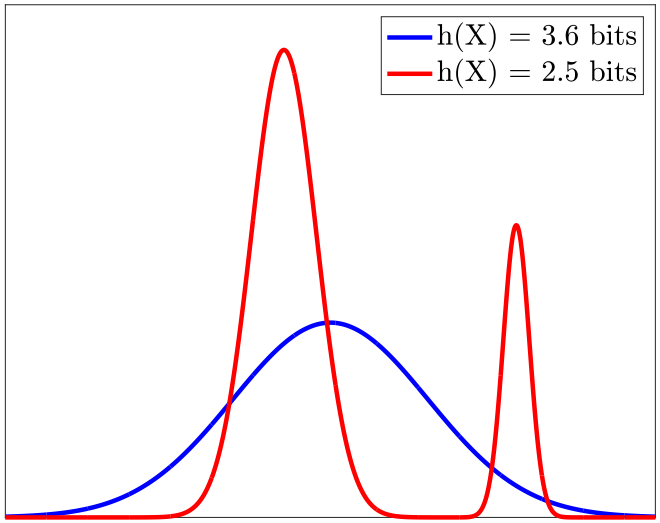
Normal distribution

$X \sim \mathcal{N}(\mu, \sigma^2)$



$$h(X) = \frac{1}{2}\log_2 2\pi e\sigma^2$$

Mixture of 2 Gaussians



Blue and red distributions:
same mean μ & standard deviation σ
but different entropy $h(X)$

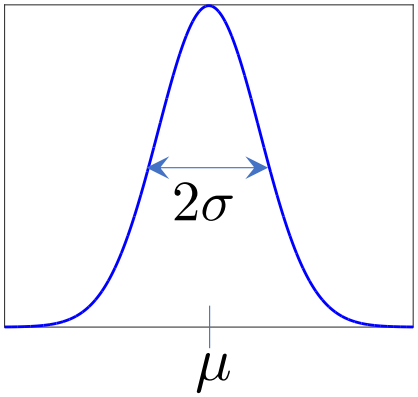
→ In the Gaussian case, $h(X)$ is a function of the standard deviation σ.

→ $h(X)$ characterizes the uncertainty in a different way than the standard deviation.

Entropy $= -E[\log \mathrm{P}(X)]$

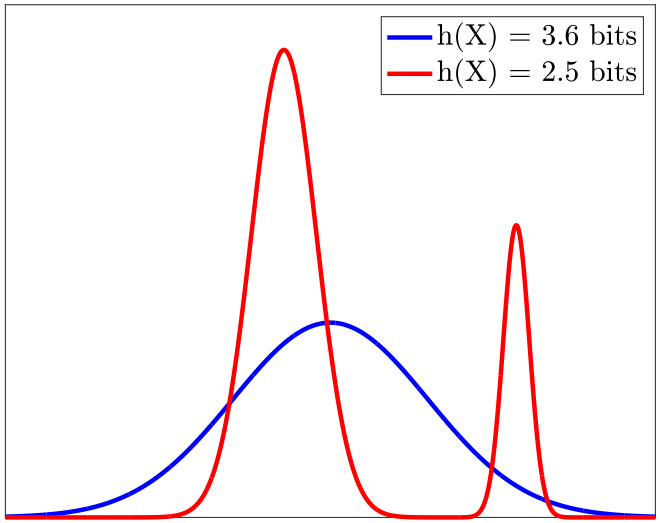
Normal distribution

$X \sim \mathcal{N}(\mu, \sigma^2)$



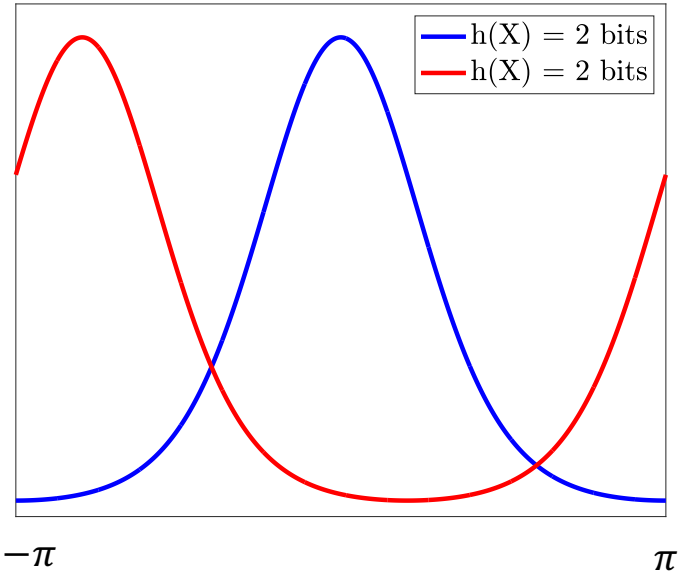
$h(X) = \frac{1}{2}\log_2 2\pi e\sigma^2$

Mixture of 2 Gaussians



Blue and red distributions:
same mean μ & standard deviation σ
but different entropy $h(X)$

Von-Mises distribution



Invariant by **circular** translation

→ In the Gaussian case, $h(X)$ is a function of the standard deviation σ.

→ $h(X)$ characterizes the uncertainty in a different way than the standard deviation.

Take home messages:

Entropy characterizes the **uncertainty** of X

Entropy $H(X) = -E[\log_2 \mathrm{P}(X)]$

next

What is the information on X provided by a measurement Y ?

What is the information on X provided by a measurement Y?

Let's consider a physical quantity of interest X (e.g. column density) and a given observation Y(e.g. integrated intensity)

Because the observation (Y) has some unpredictable contribution, we assume the presence of an additive noise (N)

Example: $Y = A\,\mathrm{asinh}[(X - m)/B] + N$

What is the information on X provided by a measurement Y?

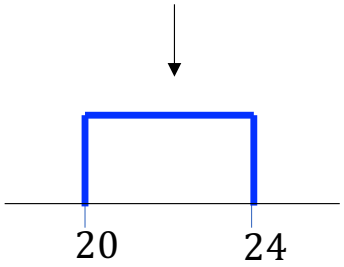
Let's consider a physical quantity of interest X (e.g. column density) and a given observation Y(e.g. integrated intensity)

Because the observation (Y) has some unpredictable contribution, we assume the presence of an additive noise (N)

Example: $Y = A\,\mathrm{asinh}[(X - m)/B] + N$

Concerning X, we know its bounds

$\rightarrow X \sim \mathcal{U}\,[20{,}24]$



What is the information on X provided by a measurement Y?

Let's consider a physical quantity of interest X (e.g. column density) and a given observation Y(e.g. integrated intensity)

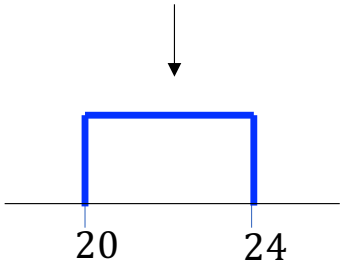
Because the observation (Y) has some unpredictable contribution, we assume the presence of an additive noise (N)

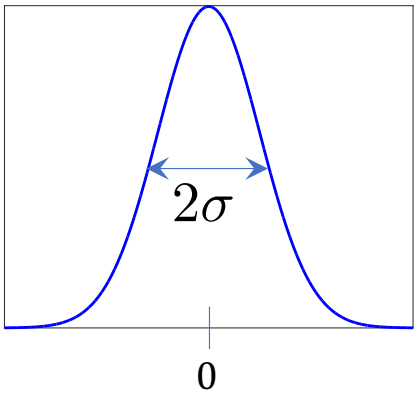
Example: $Y = A\,\mathrm{asinh}[(X - m)/B] + N$ ⟶ Concerning N, we know its mean (0) and its variance σ^2

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

Concerning X, we know its bounds

$\rightarrow X \sim \mathcal{U}\,[20{,}24]$





What is the information on X provided by a measurement Y?

Let's consider a physical quantity of interest X (e.g. column density) and a given observation Y(e.g. integrated intensity)

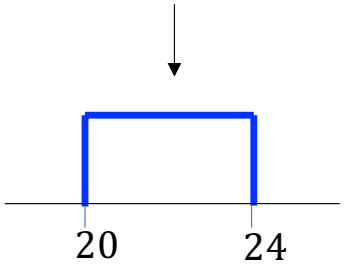
Because the observation (Y) has some unpredictable contribution, we assume the presence of an additive noise (N)

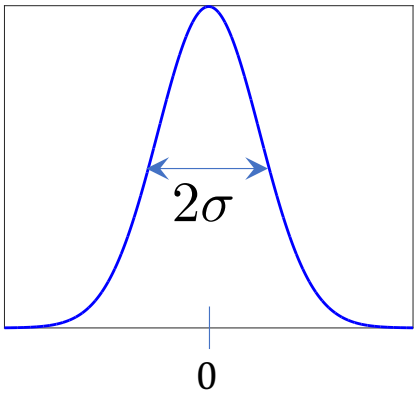
Example: $Y = A\,\mathrm{asinh}[(X - m)/B] + N$ ⟶ Concerning N, we know its mean (0) and its variance σ^2

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

Concerning X, we know its bounds

$\rightarrow X \sim \mathcal{U}\ [20{,}24]$





Let's compare the uncertainty on X before and after the measurement of Y.

What is the information on X provided by a measurement Y?

Let's consider a physical quantity of interest X (e.g. column density) and a given observation Y(e.g. integrated intensity)

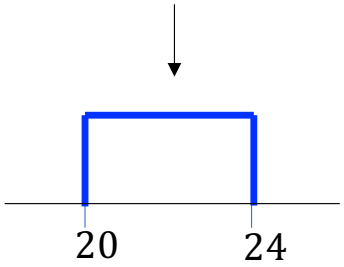
Because the observation (Y) has some unpredictable contribution, we assume the presence of an additive noise (N)

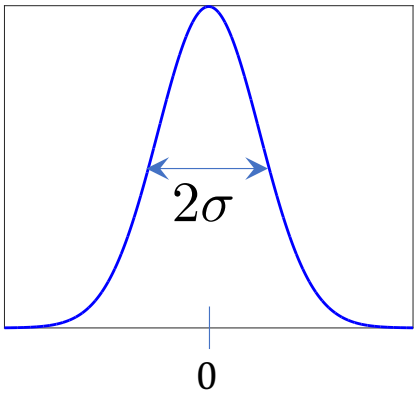
Example: $Y = A\,\mathrm{asinh}[(X - m)/B] + N$ ⟶ Concerning N, we know its mean (0) and its variance σ^2

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

Concerning X, we know its bounds

$\rightarrow X \sim \mathcal{U}\ [20{,}24]$





Let's compare the uncertainty on X before and after the measurement of Y.

To compute the entropy, we use X^{Δ} and Y^{Δ} the quantified version X and Y (with resolution $\Delta = 0.004$)

What is the information on X provided by a measurement Y?

Let's consider a physical quantity of interest X (e.g. column density) and a given observation Y(e.g. integrated intensity)

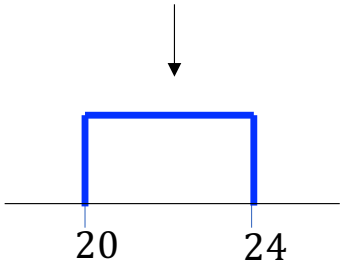
Because the observation (Y) has some unpredictable contribution, we assume the presence of an additive noise (N)

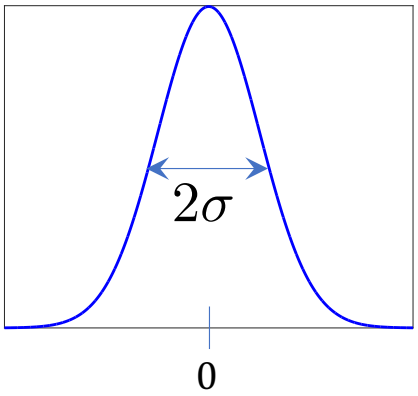
Example: $Y = A\,\mathrm{asinh}[(X - m)/B] + N$ ⟶ Concerning N, we know its mean (0) and its variance σ^2

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

Concerning X, we know its bounds

$\rightarrow X \sim \mathcal{U}\ [20{,}24]$





Let's compare the uncertainty on X before and after the measurement of Y.

To compute the entropy, we use X^Δ and Y^Δ the quantified version X and Y (with resolution $\Delta = 0.004$)

X^Δ is a discrete random variable whose values are in $\{a_1, a_2, \dots, a_K\}$ and $P_k = \Pr(X^\Delta = a_k)$

For a given observation $Y^\Delta = b_l$, the distribution $Q_k = \Pr(X^\Delta = a_k | Y^\Delta = b_l)$ is called *a posteriori* distribution.

What is the information on X provided by a measurement Y?

Measurement model: $Y = A\,\mathrm{asinh}[(X - m)/B] + N$

Before measurement, uncertainty on X?

Since $X \sim \mathcal{U}\,[20{,}24]$

$\rightarrow\ h(X) = \log_2(24 - 20) = 2$ bits

After measurement, uncertainty on X|Y?

We need to estimate

$Q_{k|l} = \frac{\Pr(X^{\Delta}=a_k, Y^{\Delta}=b_l)}{\Pr(Y^{\Delta}=b_l)}$ (Bayes' formula)

What is the information on X provided by a measurement Y?

Measurement model: $Y = A\,\mathrm{asinh}[(X - m)/B] + N$

Before measurement, uncertainty on X?

Since $X \sim \mathcal{U}\,[20{,}24]$

$\rightarrow\ h(X) = \log_2(24 - 20) = 2$ bits

After measurement, uncertainty on X|Y?

We need to estimate

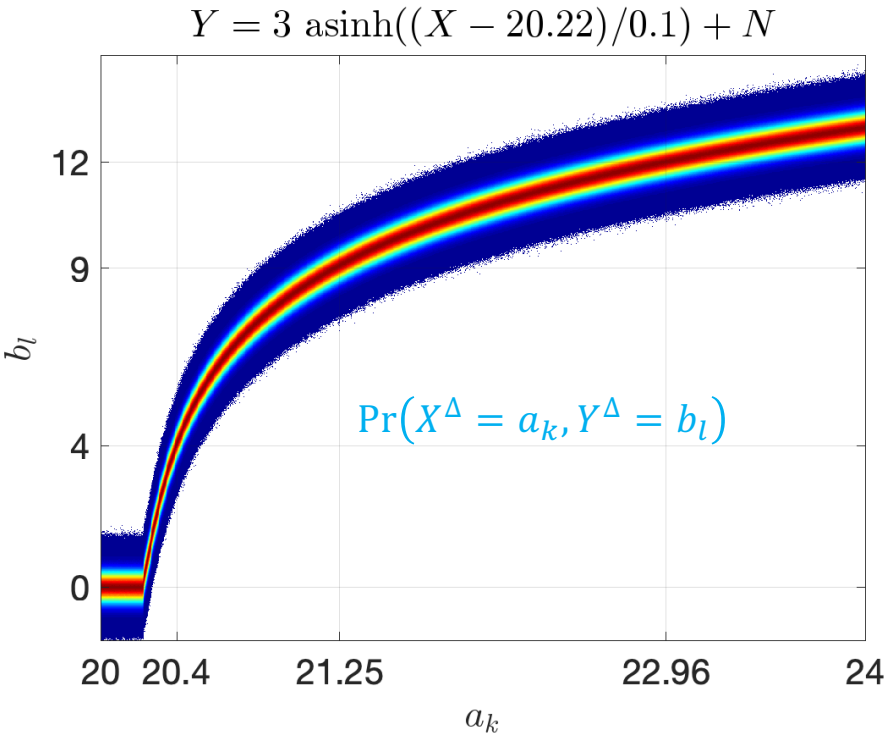
$$Q_{k|l} = \frac{\Pr(X^{\Delta}=a_k, Y^{\Delta}=b_l)}{\Pr(Y^{\Delta}=b_l)}$$

(Bayes' formula)

2D Histogram

Simulations with

sample of size 10^{10}



What is the information on X provided by a measurement Y?

Measurement model: $Y = A\,\mathrm{asinh}[(X - m)/B] + N$

Before measurement, uncertainty on X?

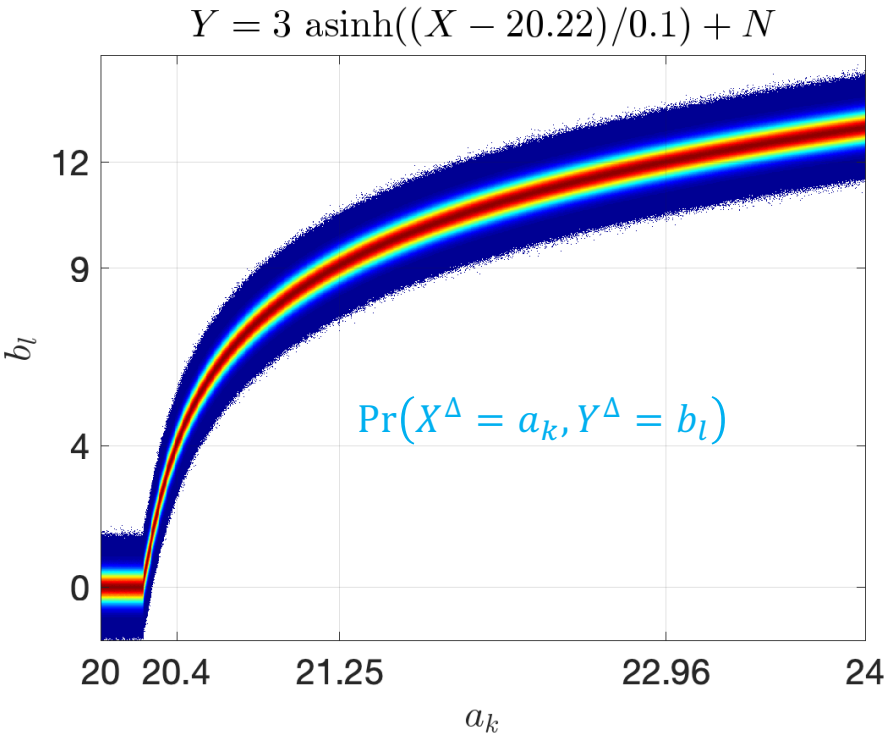
Since $X \sim \mathcal{U}\,[20,24]$

$\rightarrow\ h(X) = \log_2(24 - 20) = 2$ bits

2D Histogram

Simulations with

sample of size 10^{10}



After measurement, uncertainty on X|Y?

We need to estimate

$$Q_{k|l} = \frac{\Pr(X^\Delta = a_k, Y^\Delta = b_l)}{\Pr(Y^\Delta = b_l)}$$ (Bayes' formula)

$$\Pr(Y^\Delta = b_l) = \sum_{k=1}^{K} \Pr(X^\Delta = a_k, Y^\Delta = b_l)$$

What is the information on X provided by a measurement Y?

Measurement model: $Y = A\,\mathrm{asinh}[(X-m)/B] + N$

Before measurement, uncertainty on X?

Since $X \sim \mathcal{U}\,[20{,}24]$

$\rightarrow\ h(X) = \log_2(24-20) = 2$ bits

2D Histogram

Simulations with

sample of size 10^{10}

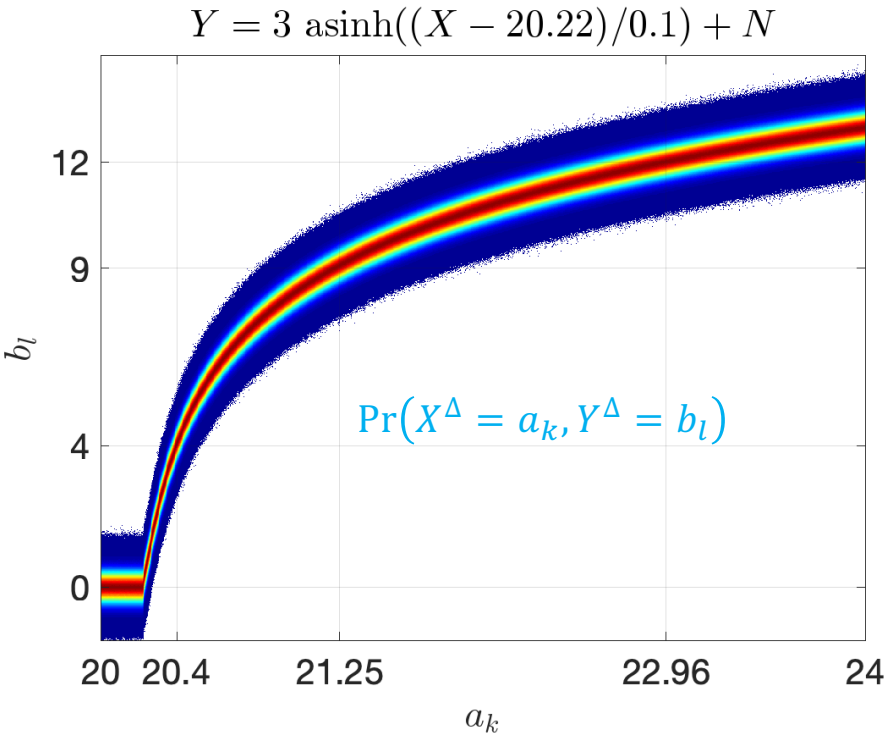
After measurement, uncertainty on X|Y?

We need to estimate

$$Q_{k|l} = \frac{\Pr\left(X^{\Delta}=a_k, Y^{\Delta}=b_l\right)}{\Pr\left(Y^{\Delta}=b_l\right)}$$ (Bayes' formula)

$$H\left(X^{\Delta}|Y^{\Delta}=b_l\right) = -\sum_{k=1}^{K} Q_{k|l}\log_2 Q_{k|l}$$

$$\Rightarrow h(X|Y=b_l) \approx H\left(X^{\Delta}|Y^{\Delta}=b_l\right) + \log_2(\Delta) \qquad (\Delta = 0.004)$$



What is the information on X provided by a measurement Y?

Measurement model: $Y = A\,\mathrm{asinh}[(X-m)/B] + N$

Before measurement, uncertainty on X?

Since $X \sim \mathcal{U}\,[20,24]$

$\rightarrow\ h(X) = \log_2(24-20) = 2$ bits

2D Histogram

Simulations with

sample of size 10^{10}

After measurement, uncertainty on X|Y?

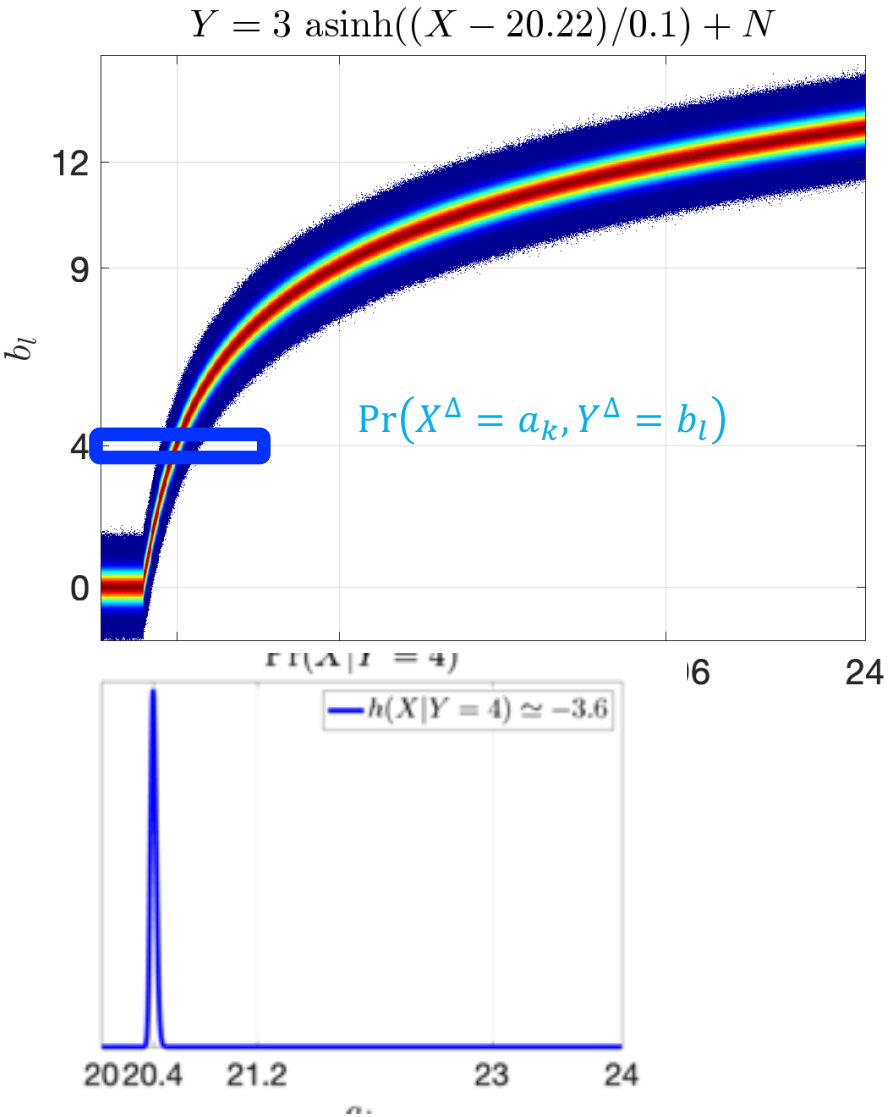
We need to estimate

$$Q_{k|l} = \frac{\Pr(X^\Delta = a_k, Y^\Delta = b_l)}{\Pr(Y^\Delta = b_l)}$$ (Bayes' formula)

$$H(X^\Delta|Y^\Delta = b_l) = -\sum_{k=1}^{K} Q_{k|l}\log_2 Q_{k|l}$$

$$\Rightarrow h(X|Y=b_l) \approx H(X^\Delta|Y^\Delta = b_l) + \log_2(\Delta) \qquad (\Delta = 0.004)$$

For $y = 4$, the entropy goes from 2 to $H(X|Y=4) = -3.6$



What is the information on X provided by a measurement Y?

Measurement model: $Y = A\,\mathrm{asinh}[(X-m)/B] + N$

Before measurement, uncertainty on X?

Since $X \sim \mathcal{U}\,[20,24]$

$\rightarrow\ h(X) = \log_2(24-20) = 2$ bits

After measurement, uncertainty on X|Y?

We need to estimate

$$Q_{k|l} = \frac{\Pr(X^\Delta = a_k, Y^\Delta = b_l)}{\Pr(Y^\Delta = b_l)}$$ (Bayes' formula)

$$H(X^\Delta | Y^\Delta = b_l) = -\sum_{k=1}^{K} Q_{k|l} \log_2 Q_{k|l}$$

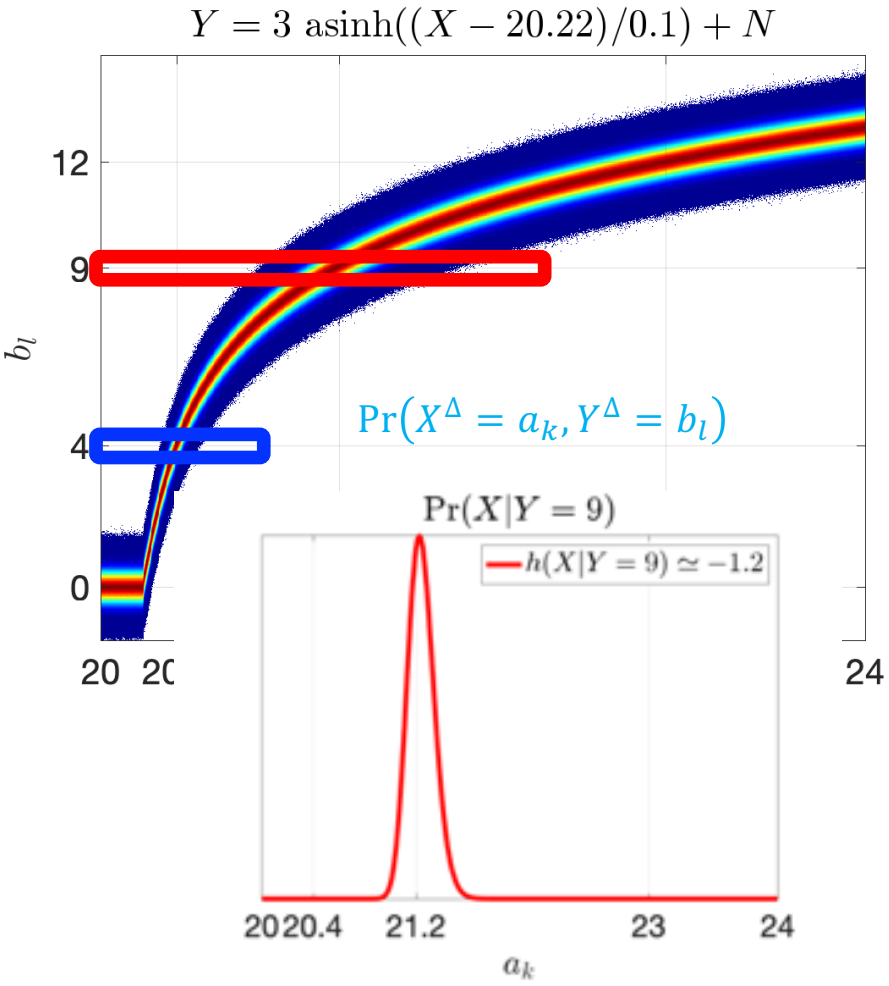
$$\Rightarrow h(X|Y=b_l) \approx H(X^\Delta | Y^\Delta = b_l) + \log_2(\Delta) \qquad (\Delta = 0.004)$$

For $y = 9$, the entropy goes from 2 to $H(X|Y=9) = -1.2$

2D Histogram

Simulations with

sample of size 10^{10}



What is the information on X provided by a measurement Y?

Measurement model: $Y = A\,\mathrm{asinh}[(X-m)/B] + N$

Before measurement, uncertainty on X?

Since $X \sim \mathcal{U}\,[20{,}24]$

$\rightarrow\ h(X) = \log_2(24-20) = 2$ bits

2D Histogram

Simulations with

sample of size 10^{10}

After measurement, uncertainty on X|Y?

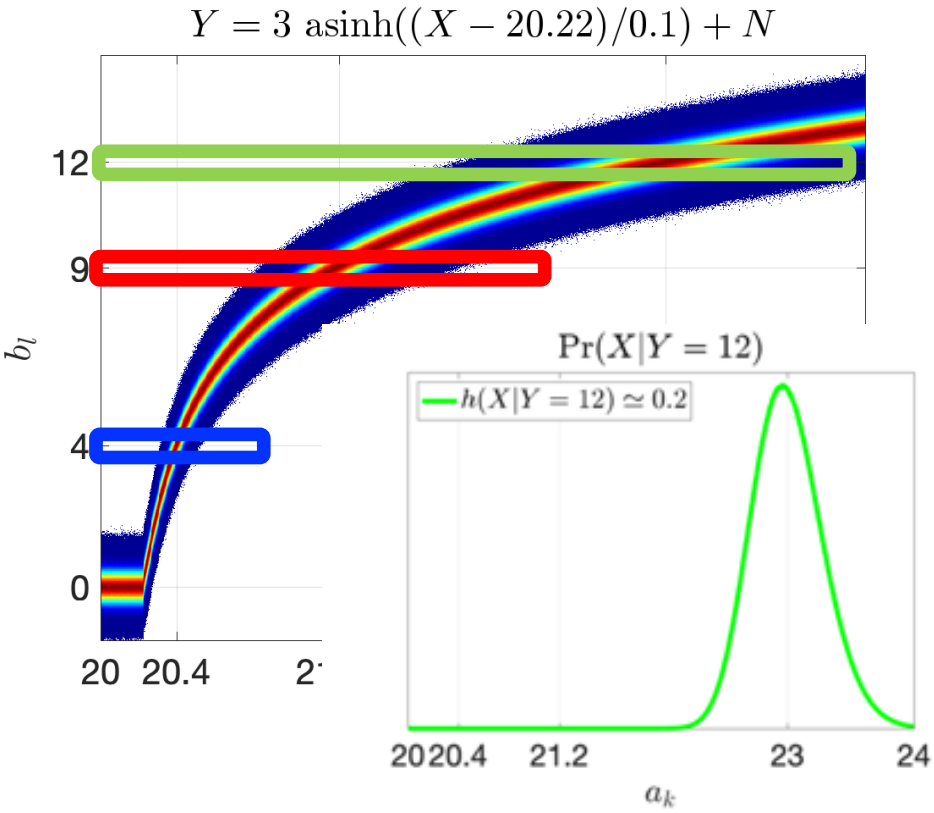
We need to estimate

$$Q_{k|l} = \frac{\Pr(X^\Delta = a_k, Y^\Delta = b_l)}{\Pr(Y^\Delta = b_l)}$$ (Bayes' formula)

$$H(X^\Delta | Y^\Delta = b_l) = -\sum_{k=1}^{K} Q_{k|l} \log_2 Q_{k|l}$$

$$\Rightarrow h(X|Y=b_l) \approx H(X^\Delta | Y^\Delta = b_l) + \log_2(\Delta) \qquad (\Delta = 0.004)$$

For $y = 12$, the entropy goes from 2 to $H(X|Y=9) = 0.2$



What is the information on X provided by a measurement Y?

Measurement model: $Y = A\,\mathrm{asinh}[(X-m)/B] + N$

Before measurement, uncertainty on X?

Since $X \sim \mathcal{U}\,[20{,}24]$

$\rightarrow\ h(X) = \log_2(24-20) =$ **2 bits**

2D Histogram

Simulations with

sample of size 10^{10}

After measurement, uncertainty on X|Y?

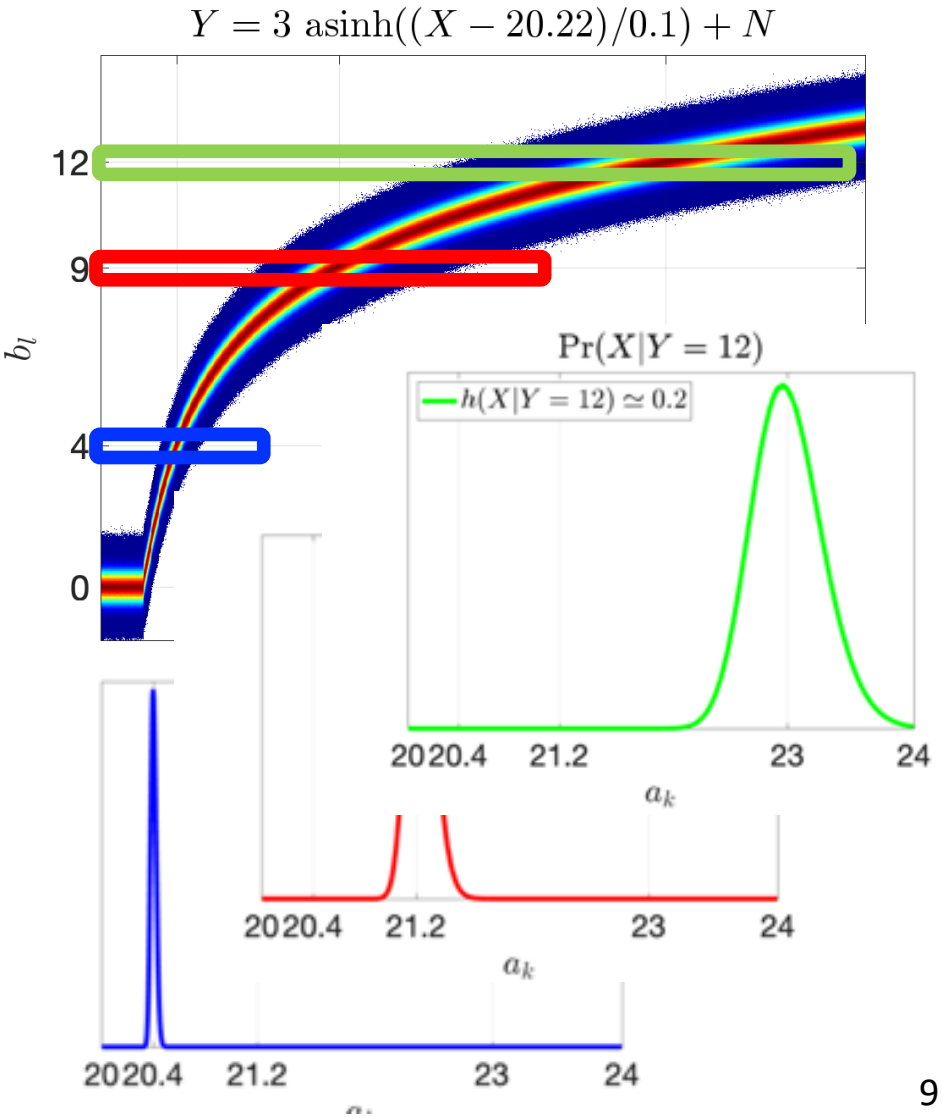
We need to estimate

$$Q_{k|l} = \frac{\Pr(X^{\Delta}=a_k, Y^{\Delta}=b_l)}{\Pr(Y^{\Delta}=b_l)}$$ (Bayes' formula)

$$H(X^{\Delta}|Y^{\Delta}=b_l) = -\sum_{k=1}^{K} Q_{k|l}\log_2 Q_{k|l}$$

By averaging $H(X|Y) = \sum_l \Pr(Y=b_l)\,H(X|Y=b_l)$

$\rightarrow$ conditional entropy $h(X|Y) = -E_{X,Y}[\log_2 Q(X)] \approx$ **-0.9 bits**



What is the information on X provided by a measurement Y?

Measurement model: $Y = A\,\mathrm{asinh}[(X-m)/B] + N$

Before measurement, uncertainty on X?

Since $X \sim \mathcal{U}\,[20{,}24]$

$\rightarrow\ h(X) = \log_2(24-20) =$ **2 bits**

After measurement, uncertainty on X|Y?

We need to estimate

$$Q_{k|l} = \frac{\Pr(X^\Delta = a_k, Y^\Delta = b_l)}{\Pr(Y^\Delta = b_l)}$$ (Bayes' formula)

$$H(X^\Delta | Y^\Delta = b_l) = -\sum_{k=1}^{K} Q_{k|l} \log_2 Q_{k|l}$$

By averaging $H(X|Y) = \sum_l \Pr(Y = b_l)\, H(X|Y = b_l)$

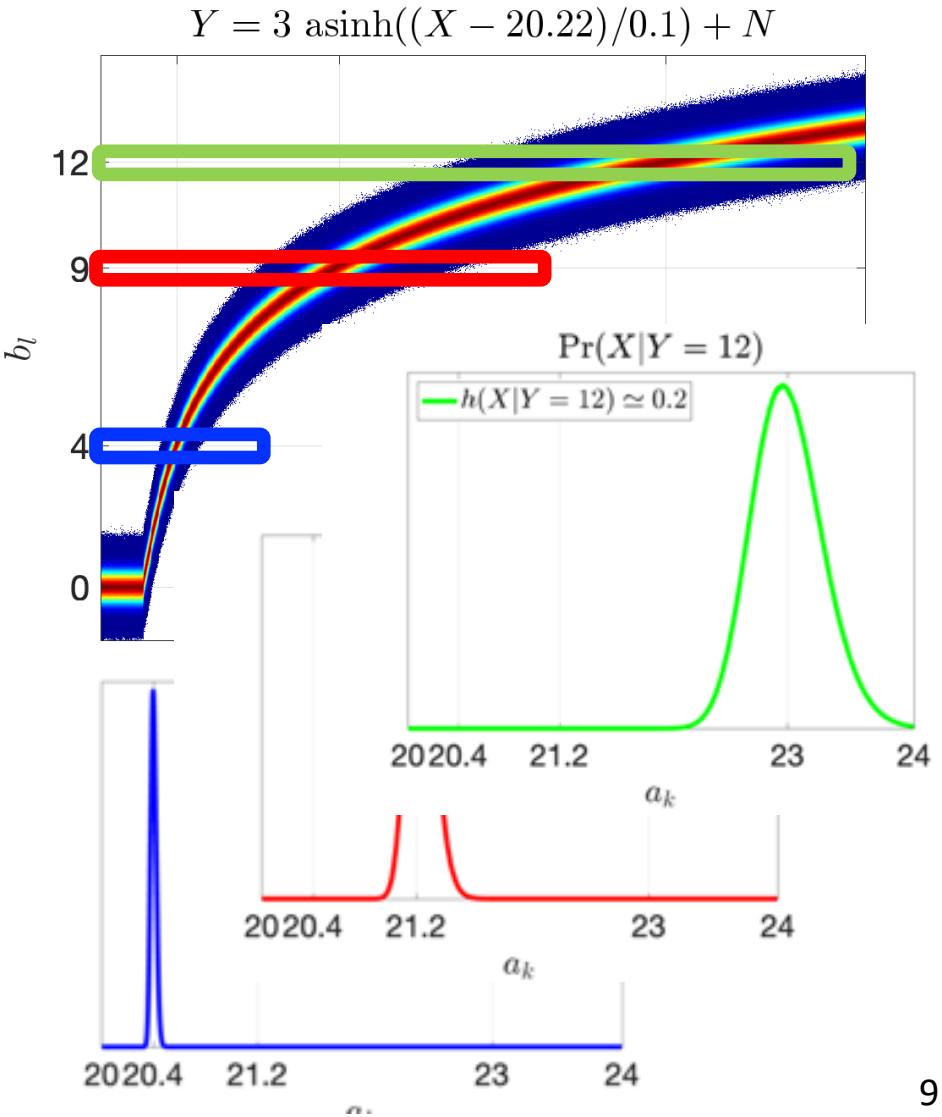
$\rightarrow$ conditional entropy $h(X|Y) = -E_{X,Y}[\log_2 Q(X)] \approx$ **-0.9 bits**

$\approx$ uncertainty that remains in X once Y is observed.

2D Histogram

Simulations with

sample of size 10^{10}

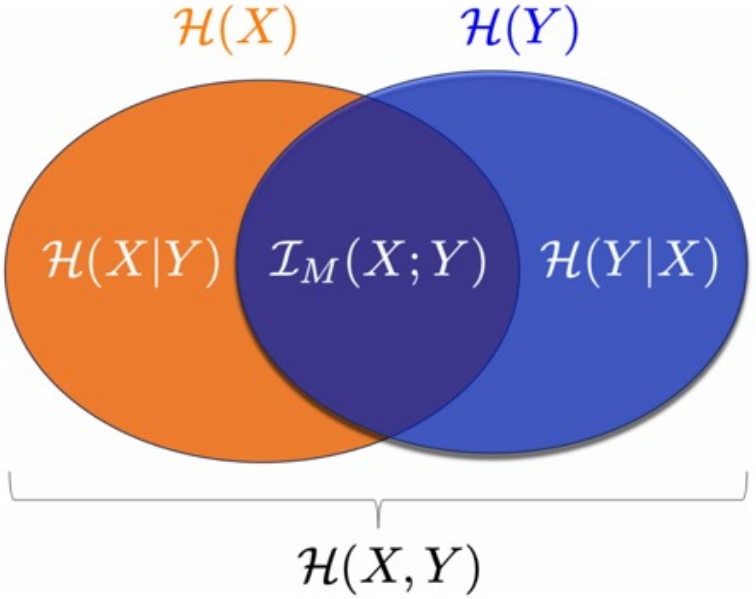


What is the information on X provided by a measurement Y ?

Mutual information

Venn's diagram quantifies uncertainty with area

The difference $I(X;Y) = H(X) - H(X|Y)$ is the mutual information

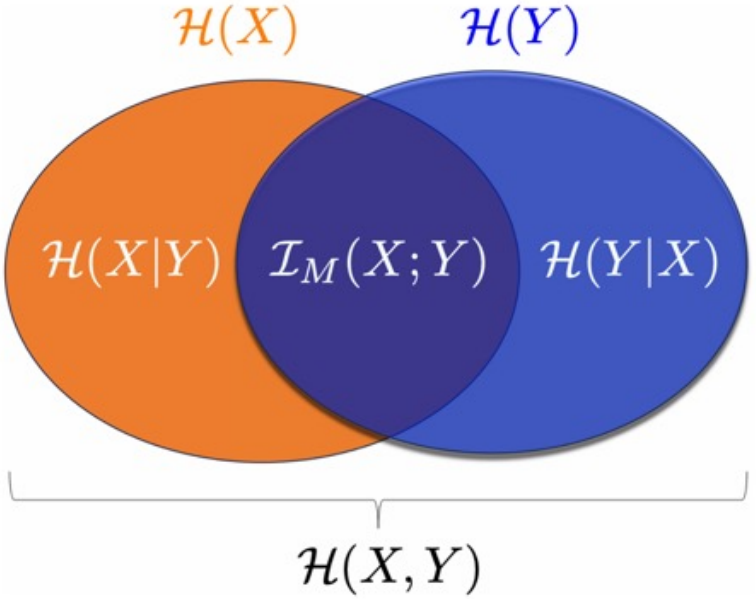


Mutual information

Venn's diagram quantifies uncertainty with area

The difference $I(X;Y) = H(X) - H(X|Y)$ is the mutual information

If $H(X) = H(X|Y)$, then $I(X;Y) = 0 \Rightarrow$ Y provides no information on X



Mutual information

Venn's diagram quantifies uncertainty with area

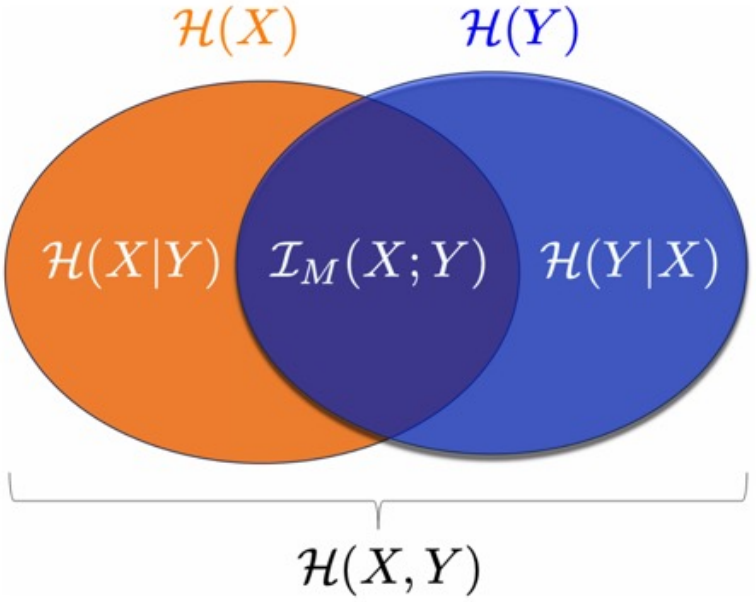
The difference $I(X;Y) = H(X) - H(X|Y)$ is the mutual information

If $H(X) = H(X|Y)$, then $I(X;Y) = 0 \Rightarrow$ Y provides no information on X

$I(X;Y)$ is a measure of dependance between X and Y.

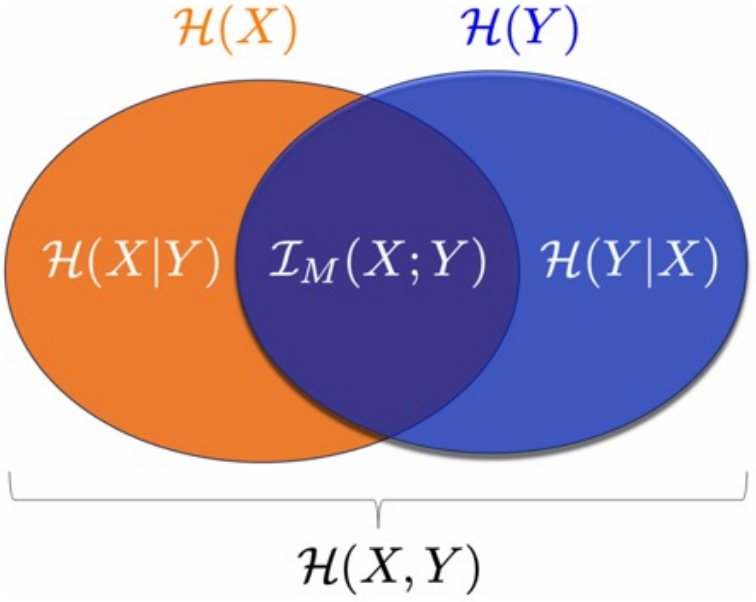
e.g. $I(X;Y) = 0 \Leftrightarrow X$ and Y are independent.

$$\Rightarrow I(X;Y) = E_{X,Y}\left[\log_2 \frac{\mathrm{P}(X,Y)}{\mathrm{P}(X)\,\mathrm{P}(Y)}\right] \geq 0 \qquad \text{(Symmetric)}$$



Mutual information

Venn's diagram quantifies uncertainty with area



The difference $I(X;Y) = H(X) - H(X|Y)$ is the mutual information

If $H(X) = H(X|Y)$, then $I(X;Y) = 0 \Rightarrow$ Y provides no information on X

$I(X;Y)$ is a measure of dependance between X and Y.

e.g. $I(X;Y) = 0 \Leftrightarrow X$ and Y are independent.

$$\Rightarrow I(X;Y) = E_{X,Y}\left[\log_2 \frac{\mathrm{P}(X,Y)}{\mathrm{P}(X)\,\mathrm{P}(Y)}\right] \geq 0 \qquad \text{(Symmetric)}$$

Example of application: assume we can observe several Y_m (e.g. molecular lines) and we want to recover the information X.

How to select the line that will provide the most of information on X ?

We can ask our favorite estimation tool (neural network, random forest, …) for the most informative line.
However, the results will then depend on the efficiency of the considered tool.

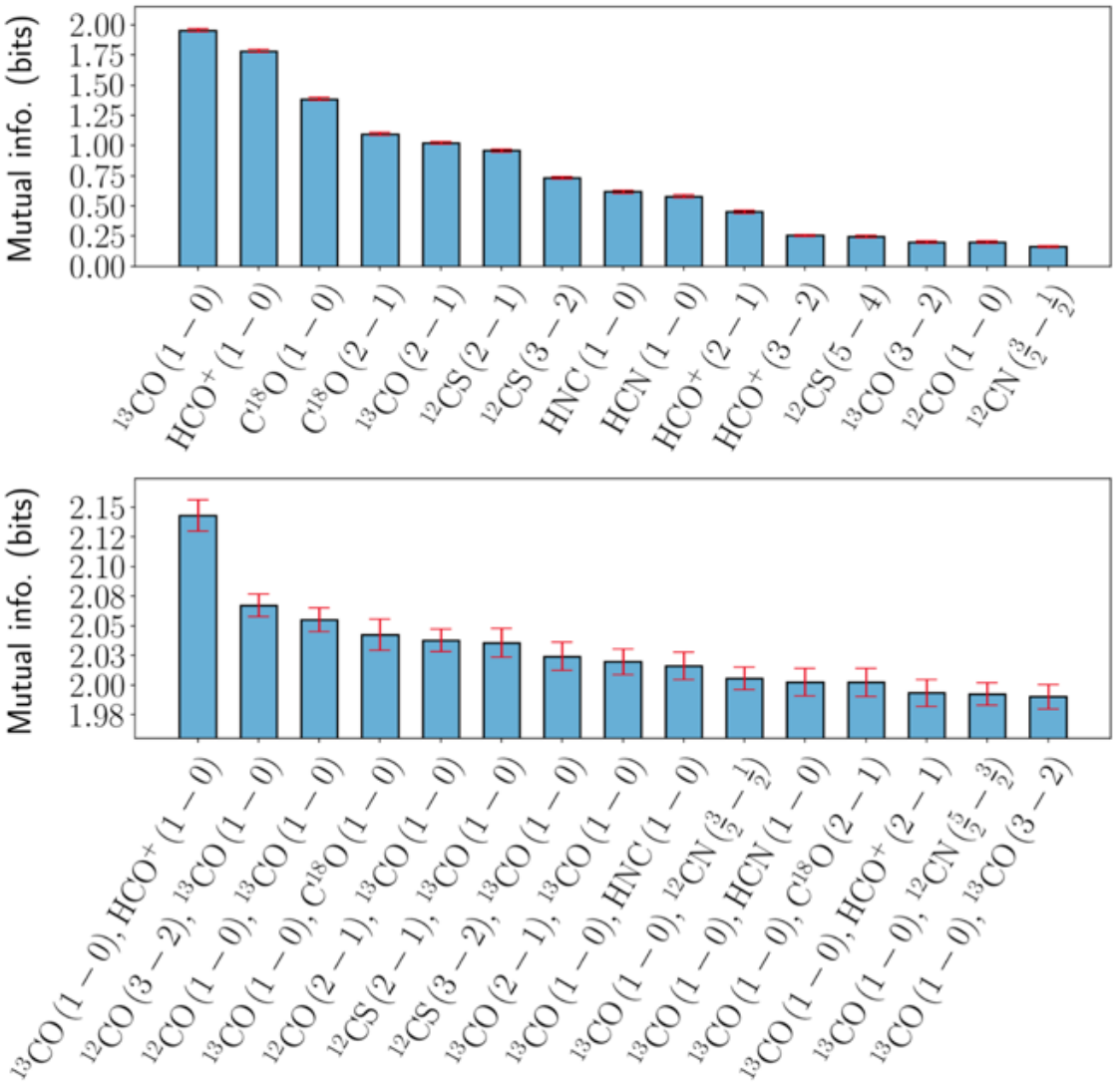
Another solution is to select the line that maximizes the mutual information (i.e. minimizes the conditional entropy).

Line ranking based on information content

For an environment representative of Orion B, one can identify the lines which provides the most of information on visual extinction (A_v)

This allows:
- to justify observation proposals
- to quantify the intuition of astrophysicists.

The computation of the mutual information remains challenging*



* “*Quantifying the informativity of emission lines to infer physical conditions in giant molecular clouds*” A&A, 691, A109 (2024)

Take home messages:

Entropy characterizes the **uncertainty** of X

Entropy $H(X) = -E[\log_2 \mathrm{P}(X)]$

Mutual information allows to select the **informative** lines

next

Is there a link with S/N, correlation coefficient, or mean square error?

Gaussian case

Assume X and N are independent centered random Gaussian variable with $Y = X + N$

Signal-to-noise ratio $S/N = \sigma_x^2/\sigma_n^2$, where $\sigma_x^2 = var(X)$ and $\sigma_n^2 = var(N)$

The variance of observation Y is $\sigma_y^2 = var(Y) = \sigma_x^2 + \sigma_n^2$

The pair (X, Y) is a Gaussian random vector with mean 0 and covariance matrix $\begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$

where ρ if the correlation coefficient between X and Y.

Gaussian case

Assume X and N are independent centered random Gaussian variable with $Y = X + N$

Signal-to-noise ratio $S/N = \sigma_x^2/\sigma_n^2$, where $\sigma_x^2 = var(X)$ and $\sigma_n^2 = var(N)$

The variance of observation Y is $\sigma_y^2 = var(Y) = \sigma_x^2 + \sigma_n^2$

The pair (X, Y) is a Gaussian random vector with mean 0 and covariance matrix $\begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$

where ρ if the correlation coefficient between X and Y.

➡ $h(X) = -\frac{1}{2}\log_2 2\pi e\sigma_x^2$ $\quad$ $h(X|Y) = -\frac{1}{2}\log_2 2\pi e\sigma_x^2(1-\rho^2)$ $\quad$ $I(X;Y) = -\frac{1}{2}\log_2(1-\rho^2)$

Gaussian case

Assume X and N are independent centered random Gaussian variable with $Y = X + N$

Signal-to-noise ratio $S/N = \sigma_x^2/\sigma_n^2$, where $\sigma_x^2 = var(X)$ and $\sigma_n^2 = var(N)$

The variance of observation Y is $\sigma_y^2 = var(Y) = \sigma_x^2 + \sigma_n^2$

The pair (X, Y) is a Gaussian random vector with mean 0 and covariance matrix $\begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$

where ρ if the correlation coefficient between X and Y.

$$\Rightarrow \quad h(X) = -\frac{1}{2}\log_2 2\pi e\sigma_x^2 \qquad h(X|Y) = -\frac{1}{2}\log_2 2\pi e\sigma_x^2(1-\rho^2) \qquad I(X;Y) = -\frac{1}{2}\log_2(1-\rho^2)$$

The mean square error of the "best" estimator $\hat{X}$ is $\mathrm{MSE}(\hat{X}) = E\left[(\hat{X} - X)^2\right] = \sigma_x^2(1-\rho^2)$ (-> Bayes' estimation course)

Gaussian case

Assume X and N are independent centered random Gaussian variable with $Y = X + N$

Signal-to-noise ratio $S/N = \sigma_x^2/\sigma_n^2$, where $\sigma_x^2 = var(X)$ and $\sigma_n^2 = var(N)$

The variance of observation Y is $\sigma_y^2 = var(Y) = \sigma_x^2 + \sigma_n^2$

The pair (X, Y) is a Gaussian random vector with mean 0 and covariance matrix $\begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$

where ρ if the correlation coefficient between X and Y.

$$\Rightarrow \quad h(X) = -\frac{1}{2}\log_2 2\pi e\sigma_x^2 \qquad h(X|Y) = -\frac{1}{2}\log_2 2\pi e\sigma_x^2(1-\rho^2) \qquad I(X;Y) = -\frac{1}{2}\log_2(1-\rho^2)$$

The mean square error of the "best" estimator $\hat{X}$ is $\mathrm{MSE}(\hat{X}) = E\left[(\hat{X} - X)^2\right] = \sigma_x^2(1-\rho^2)$ (-> Bayes' estimation course)

Conclusions for line selection? (for the simple Gaussian case **ONLY**)

finding Y that maximizes the mutual information $\Leftrightarrow$ maximizing the correlation coefficient between X and Y

finding Y that minimizes the conditional entropy $\Leftrightarrow$ minimizing the mean square error

$\Leftrightarrow$ maximizing the S/N

Signal-to-noise ratio $S/N = \rho^2/(1-\rho^2)$

Take home messages:

Entropy characterizes the **uncertainty** of X

Entropy $H(X) = -E[\log_2 \mathrm{P}(X)]$

Mutual information allows to select the **informative** lines

Link with S/N, correlation coefficient & MSE **only in Gaussian case**

next

How to go from information theory to estimation theory?

From coding to learning

Let $x_1, x_2, \ldots, x_N$ be a sample of independent measurements distributed along the *true* distribution p.

The Maximum Likelihood estimator, $\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \prod_n q_X(x_n;\theta)$ converges to the solution. **Why ?**

From coding to learning

Let $x_1, x_2, \ldots, x_N$ be a sample of independent measurements distributed along the *true* distribution p.

The Maximum Likelihood estimator, $\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \prod_n q_X(x_n;\theta)$ converges to the solution. **Why ?**

Normalized negative-log-likelihood: $\lim_{N\to\infty} -\frac{1}{N}\sum_n \log_2 q_X(x_n;\theta) = -E_p[\log_2 q(X;\theta)] = h(p,q)$

Remember: $E[h(X)] = \lim_N \frac{1}{N}\sum_n h(x_n)$

From coding to learning

Let $x_1, x_2, \ldots, x_N$ be a sample of independent measurements distributed along the *true* distribution p.

The Maximum Likelihood estimator, $\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \prod_n q_X(x_n;\theta)$ converges to the solution. **Why ?**

Normalized negative-log-likelihood: $\lim_{N\to\infty} -\frac{1}{N}\sum_n \log_2 q_X(x_n;\theta) = -E_p[\log_2 q(X;\theta)] = h(p,q)$

The cross-entropy $\mathrm{H}(P,Q) = -\sum_n P_n \log_2 Q_n$ average coding length of X when assuming X is distributed with Q instead of P

From coding to learning

Let $x_1, x_2, \ldots, x_N$ be a sample of independent measurements distributed along the *true* distribution p.

The Maximum Likelihood estimator, $\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \prod_n q_X(x_n;\theta)$ converges to the solution. **Why ?**

Normalized negative-log-likelihood: $\lim_{N\to\infty} -\frac{1}{N}\sum_n \log_2 q_X(x_n;\theta) = -E_p[\log_2 q(X;\theta)] = h(p,q)$

The cross-entropy $\mathrm{H}(P,Q) = -\sum_n P_n \log_2 Q_n$ average coding length of X when assuming X is distributed with Q instead of P

➡ Asymptotically (N large), finding the maximum likelihood estimator (MLE) consists of minimizing $H(p,q)$

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \sum_n \log_2 q_X(x_n;\theta) = \arg\min_{\theta} H(p^{\mathrm{true}}, q(X;\theta))$$

Conclusion?

From coding to learning

Let $x_1, x_2, \ldots, x_N$ be a sample of independent measurements distributed along the *true* distribution p.

The Maximum Likelihood estimator, $\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \prod_n q_X(x_n;\theta)$ converges to the solution. **Why ?**

Normalized negative-log-likelihood: $\lim_{N\to\infty} -\frac{1}{N}\sum_n \log_2 q_X(x_n;\theta) = -E_p[\log_2 q(X;\theta)] = h(p,q)$

The cross-entropy $\mathrm{H}(P,Q) = -\sum_n P_n \log_2 Q_n$ average coding length of X when assuming X is distributed with Q instead of P

➡ Asymptotically (N large), finding the maximum likelihood estimator (MLE) consists of minimizing $H(p,q)$

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \sum_n \log_2 q_X(x_n;\theta) = \arg\min_{\theta} H(p^{\mathrm{true}}, q(X;\theta))$$

Conclusion? MLE provides the value of θ that minimizes the coding length (Occam's Razor)

From coding to learning

Let $x_1, x_2, \ldots, x_N$ be a sample of independent measurements distributed along the *true* distribution p.

The Maximum Likelihood estimator, $\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \prod_n q_X(x_n;\theta)$ converges to the solution. **Why ?**

Normalized negative-log-likelihood: $\lim_{N\to\infty} -\frac{1}{N}\sum_n \log_2 q_X(x_n;\theta) = -E_p[\log_2 q(X;\theta)] = h(p,q)$

The cross-entropy $\mathrm{H}(P,Q) = -\sum_n P_n \log_2 Q_n$ average coding length of X when assuming X is distributed with Q instead of P

➡ Asymptotically (N large), finding the maximum likelihood estimator (MLE) consists of minimizing $H(p,q)$

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \sum_n \log_2 q_X(x_n;\theta) = \arg\min_{\theta} H(p^{\mathrm{true}}, q(X;\theta))$$

Conclusion? MLE provides the value of θ that minimizes the coding length (Occam's Razor)

Statistical learning: identifying the best representation of the current data allows one to make prediction for future data.

From coding to learning

Let $x_1, x_2, \ldots, x_N$ be a sample of independent measurements distributed along the *true* distribution p.

The Maximum Likelihood estimator, $\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \prod_n q_X(x_n;\theta)$ converges to the solution. **Why ?**

Normalized negative-log-likelihood: $\lim_{N\to\infty} -\frac{1}{N}\sum_n \log_2 q_X(x_n;\theta) = -E_p[\log_2 q(X;\theta)] = h(p,q)$

The cross-entropy $\mathrm{H}(P,Q) = -\sum_n P_n \log_2 Q_n$ average coding length of X when assuming X is distributed with Q instead of P

➡ Asymptotically (N large), finding the maximum likelihood estimator (MLE) consists of minimizing $H(p,q)$

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \sum_n \log_2 q_X(x_n;\theta) = \arg\min_{\theta} H(p^{\mathrm{true}}, q(X;\theta))$$

Conclusion? MLE provides the value of θ that minimizes the coding length (Occam's Razor)

Statistical learning: identifying the best representation of the current data allows one to make prediction for future data.

This is used to introduce model selection techniques (AIC, BIC, MDL, …), but this is only valid asymptotically (N large)

Take home messages:

Entropy characterizes the **uncertainty** of X

Entropy $H(X) = -E[\log_2 \mathrm{P}(X)]$

Mutual information allows to select the **informative** lines

Relations with S/N, correlation coefficient & MSE **only in Gaussian case**

Statistical learning can be seen as minimizing the code length

Last

How to select the “best” estimator?

How to **define** the "best" estimator?

Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample distributed along $p_X(x;\theta)$ and $\hat{\theta}$ be an estimator of θ (i.e. a function of $x_1, x_2, \ldots, x_N$)

For any θ, $\mathrm{MSE}(\hat{\theta},\theta) = E_X\left[(\hat{\theta}-\theta)^2\right]$ is useful to characterize the accuracy (bias & precision) of $\hat{\theta}$

How to **define** the "best" estimator?

Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample distributed along $p_X(x;\theta)$ and $\hat{\theta}$ be an estimator of θ (i.e. a function of $x_1, x_2, \ldots, x_N$)

For any θ, $\mathrm{MSE}(\hat{\theta},\theta) = E_X[(\hat{\theta}-\theta)^2]$ is useful to characterize the accuracy (bias & precision) of $\hat{\theta}$

Problem. $\hat{\theta}_{\mathrm{opt}} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)\ \ \forall\theta$ does not exist. What are the solutions?

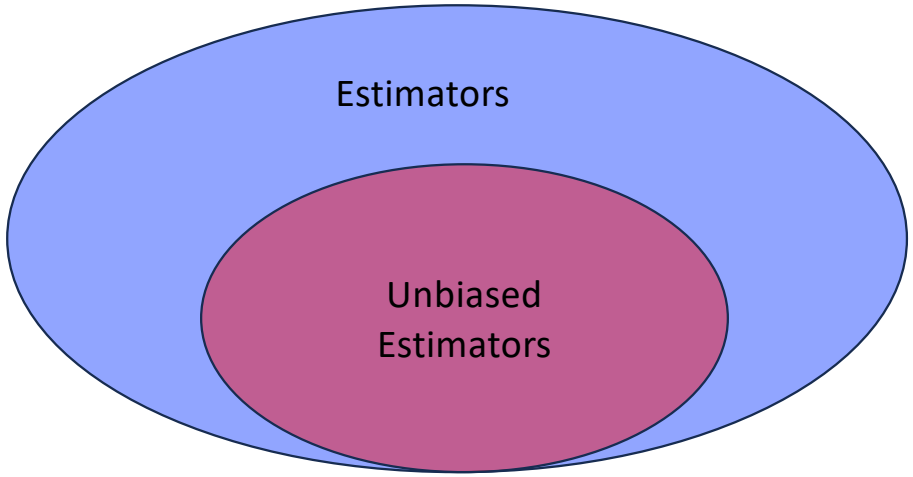
How to **define** the "best" estimator?

Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample distributed along $p_X(x;\theta)$ and $\hat{\theta}$ be an estimator of θ (i.e. a function of $x_1, x_2, \ldots, x_N$)

For any θ, $\mathrm{MSE}(\hat{\theta},\theta) = E_X[(\hat{\theta}-\theta)^2]$ is useful to characterize the accuracy (bias & precision) of $\hat{\theta}$

Problem. $\hat{\theta}_{\mathrm{opt}} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)\ \ \forall\theta$ does not exist. What are the solutions?

1/ Search for the Uniformly Minimum Variance Unbiased $\hat{\theta}_{\mathrm{UMVU}} = \arg\min_{\text{unbiased } \hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)$



How to **define** the "best" estimator?

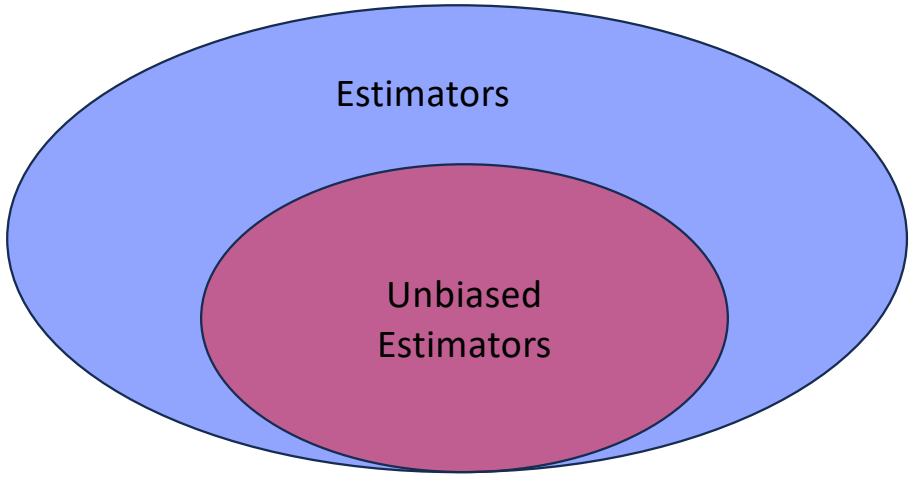
Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample distributed along $p_X(x;\theta)$ and $\hat{\theta}$ be an estimator of θ (i.e. a function of $x_1, x_2, \ldots, x_N$)

For any θ, $\mathrm{MSE}(\hat{\theta},\theta) = E_X[(\hat{\theta}-\theta)^2]$ is useful to characterize the accuracy (bias & precision) of $\hat{\theta}$

Problem. $\hat{\theta}_{\mathrm{opt}} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)\ \ \forall\theta$ does not exist. What are the solutions?

1/ Search for the Uniformly Minimum Variance Unbiased $\hat{\theta}_{\mathrm{UMVU}} = \arg\min_{\text{unbiased } \hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)$

How to be sure that we reached the minimum?



How to **define** the "best" estimator?

Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample distributed along $p_X(x;\theta)$ and $\hat{\theta}$ be an estimator of θ (i.e. a function of $x_1, x_2, \ldots, x_N$)

For any θ, $\mathrm{MSE}(\hat{\theta},\theta) = E_X[(\hat{\theta}-\theta)^2]$ is useful to characterize the accuracy (bias & precision) of $\hat{\theta}$

Problem. $\hat{\theta}_{\mathrm{opt}} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)\ \ \forall\theta$ does not exist. What are the solutions?

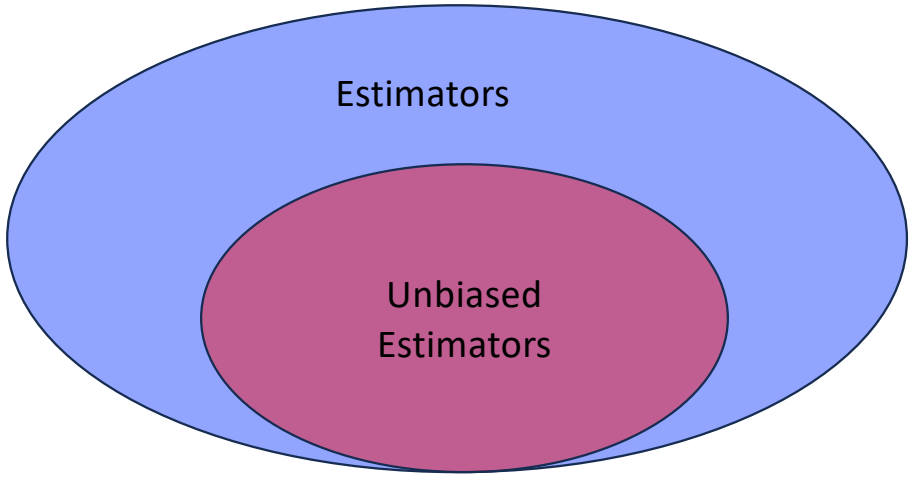
1/ Search for the Uniformly Minimum Variance Unbiased $\hat{\theta}_{\mathrm{UMVU}} = \arg\min_{\text{unbiased } \hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)$

How to be sure that we reached the minimum?

Cramér-Rao (lower) Bound (CRB), for any unbiased estimators,

$$\mathrm{var}(\hat{\theta}) \geq 1/F(\theta) \text{ where } F(\theta) = -E_X[\nabla_\theta^2 \log p_X(x;\theta)]$$
(Fisher information)

If $\mathrm{var}(\hat{\theta}) = \frac{1}{F(\theta)}$
⇒ all the information present in the data has been extracted.



How to **define** the "best" estimator?

Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample distributed along $p_X(x;\theta)$ and $\hat{\theta}$ be an estimator of θ (i.e. a function of $x_1, x_2, \ldots, x_N$)

For any θ, $\mathrm{MSE}(\hat{\theta},\theta) = E_X[(\hat{\theta}-\theta)^2]$ is useful to characterize the accuracy (bias & precision) of $\hat{\theta}$

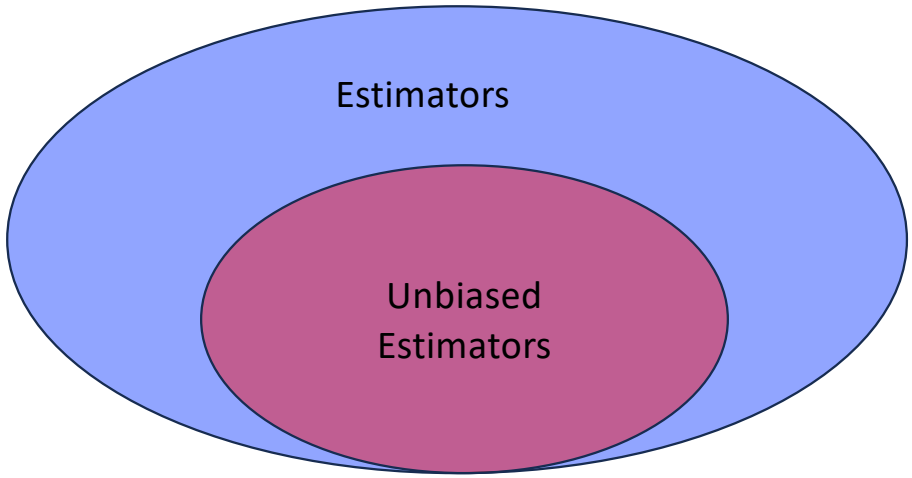
Problem. $\hat{\theta}_{\mathrm{opt}} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)\ \ \forall\theta$ does not exist. What are the solutions?

1/ Search for the Uniformly Minimum Variance Unbiased $\hat{\theta}_{\mathrm{UMVU}} = \arg\min_{\text{unbiased } \hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)$

How to be sure that we reached the minimum?

Cramér-Rao (lower) Bound (CRB), for any unbiased estimators,

$$\mathrm{var}(\hat{\theta}) \geq 1/F(\theta) \text{ where } F(\theta) = -E_X[\nabla_\theta^2 \log p_X(x;\theta)]$$
(Fisher information)



If $\mathrm{var}(\hat{\theta}) = \frac{1}{F(\theta)}$
⇒ all the information present in the data has been extracted.

Remark. Such an estimator $\hat{\theta}$ does not always exists. But if he does, then the MLE provides it

How to **define** the "best" estimator?

Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample distributed along $p_X(x;\theta)$ and $\hat{\theta}$ be an estimator of θ (i.e. a function of $x_1, x_2, \ldots, x_N$)

For any θ, $\mathrm{MSE}(\hat{\theta},\theta) = E_X[(\hat{\theta}-\theta)^2]$ is useful to characterize the accuracy (bias & precision) of $\hat{\theta}$

Problem. $\hat{\theta}_{\mathrm{opt}} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)\ \ \forall\theta$ does not exist. What are the solutions?

1/ Search for the Uniformly Minimum Variance Unbiased $\hat{\theta}_{\mathrm{UMVU}} = \arg\min_{\text{unbiased } \hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)$

Example: $X = m(\theta) + N$ with $m(\theta) = A\,\mathrm{asinh}[(\theta - C)/B]$

How to estimate θ from X?

How to **define** the "best" estimator?

Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample distributed along $p_X(x;\theta)$ and $\hat{\theta}$ be an estimator of θ (i.e. a function of $x_1, x_2, \ldots, x_N$)

For any θ, $\mathrm{MSE}(\hat{\theta},\theta) = E_X[(\hat{\theta}-\theta)^2]$ is useful to characterize the accuracy (bias & precision) of $\hat{\theta}$

Problem. $\hat{\theta}_{\mathrm{opt}} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)$ $\forall\theta$ does not exist. What are the solutions?

1/ Search for the Uniformly Minimum Variance Unbiased $\hat{\theta}_{\mathrm{UMVU}} = \arg\min_{\text{unbiased } \hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)$

Example: $X = m(\theta) + N$ with $m(\theta) = A\,\mathrm{asinh}[(\theta - C)/B]$

How to estimate θ from X?

$\hat{\theta} = m^{-1}(x) = C + B.\sinh\left(\frac{X}{A}\right)$

How to **define** the "best" estimator?

Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample distributed along $p_X(x;\theta)$ and $\hat{\theta}$ be an estimator of θ (i.e. a function of $x_1, x_2, \ldots, x_N$)

For any θ, $\mathrm{MSE}(\hat{\theta},\theta) = E_X[(\hat{\theta}-\theta)^2]$ is useful to characterize the accuracy (bias & precision) of $\hat{\theta}$

Problem. $\hat{\theta}_{\mathrm{opt}} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)\ \ \forall\theta$ does not exist. What are the solutions?

1/ Search for the Uniformly Minimum Variance Unbiased $\hat{\theta}_{\mathrm{UMVU}} = \arg\min_{\text{unbiased } \hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)$

Example: $X = m(\theta) + N$ with $m(\theta) = A\,\mathrm{asinh}[(\theta - C)/B]$

How to estimate θ from X?

$\hat{\theta} = m^{-1}(x) = C + B.\sinh\left(\frac{X}{A}\right)$

$\Rightarrow \mathrm{bias}(\hat{\theta}) = ?\,, \mathrm{var}(\hat{\theta}) = ?$ How good is this estimator?

How to **define** the "best" estimator?

Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample distributed along $p_X(x;\theta)$ and $\hat{\theta}$ be an estimator of θ (i.e. a function of $x_1, x_2, \ldots, x_N$)

For any θ, $\mathrm{MSE}(\hat{\theta},\theta) = E_X[(\hat{\theta}-\theta)^2]$ is useful to characterize the accuracy (bias & precision) of $\hat{\theta}$

Problem. $\hat{\theta}_{\mathrm{opt}} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)\ \ \forall\theta$ does not exist. What are the solutions?

1/ Search for the Uniformly Minimum Variance Unbiased $\hat{\theta}_{\mathrm{UMVU}} = \arg\min_{\text{unbiased } \hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)$

Example: $X = m(\theta) + N$ with $m(\theta) = A\,\mathrm{asinh}[(\theta - C)/B]$

How to estimate θ from X?

$\hat{\theta} = m^{-1}(x) = C + B.\sinh\left(\frac{X}{A}\right)$

$\Rightarrow \mathrm{bias}(\hat{\theta}) = ?\,, \mathrm{var}(\hat{\theta}) = ?$ How good is this estimator?

If $X \sim \mathcal{N}(m(\theta),\sigma^2) \Rightarrow F(\theta) = \left(\frac{m'(\theta)}{\sigma}\right)^2 \Rightarrow \mathrm{var}(\hat{\theta}) \geq \left(\frac{\sigma}{m'(\theta)}\right)^2$

How to **define** the "best" estimator?

Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample distributed along $p_X(x;\theta)$ and $\hat{\theta}$ be an estimator of θ (i.e. a function of $x_1, x_2, \ldots, x_N$)

For any θ, $\mathrm{MSE}(\hat{\theta},\theta) = E_X[(\hat{\theta}-\theta)^2]$ is useful to characterize the accuracy (bias & precision) of $\hat{\theta}$

Problem. $\hat{\theta}_{\mathrm{opt}} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)\ \forall\theta$ does not exist. What are the solutions?

1/ Search for the Uniformly Minimum Variance Unbiased $\hat{\theta}_{\mathrm{UMVU}} = \arg\min_{\text{unbiased } \hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)$

Example: $X = m(\theta) + N$ with $m(\theta) = A\,\mathrm{asinh}[(\theta - C)/B]$

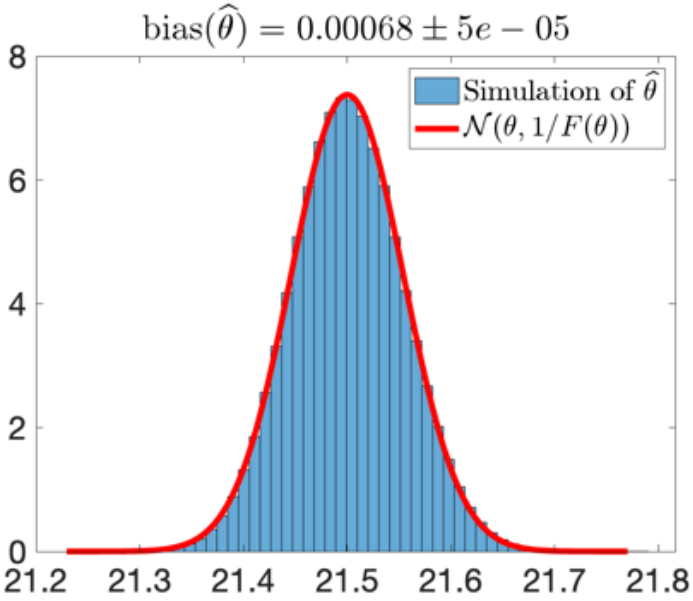
How to estimate θ from X?

$\hat{\theta} = m^{-1}(x) = C + B.\sinh\left(\frac{X}{A}\right)$

$\Rightarrow \mathrm{bias}(\hat{\theta}) = ?, \mathrm{var}(\hat{\theta}) = ?$ How good is this estimator?

If $X \sim \mathcal{N}(m(\theta),\sigma^2) \Rightarrow F(\theta) = \left(\frac{m'(\theta)}{\sigma}\right)^2 \Rightarrow \mathrm{var}(\hat{\theta}) \geq \left(\frac{\sigma}{m'(\theta)}\right)^2$

Monte Carlo simulation for $\theta = 21.5$



How to **define** the "best" estimator?

Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample distributed along $p_X(x;\theta)$ and $\hat{\theta}$ be an estimator of θ (i.e. a function of $x_1, x_2, \ldots, x_N$)

For any θ, $\mathrm{MSE}(\hat{\theta},\theta) = E_X[(\hat{\theta}-\theta)^2]$ is useful to characterize the accuracy (bias & precision) of $\hat{\theta}$

Problem. $\hat{\theta}_{\mathrm{opt}} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)\ \ \forall\theta$ does not exist. What are the solutions?

1/ Search for the Uniformly Minimum Variance Unbiased $\hat{\theta}_{\mathrm{UMVU}} = \arg\min_{\text{unbiased } \hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)$

Example: $X = m(\theta) + N$ with $m(\theta) = A\,\mathrm{asinh}[(\theta - C)/B]$

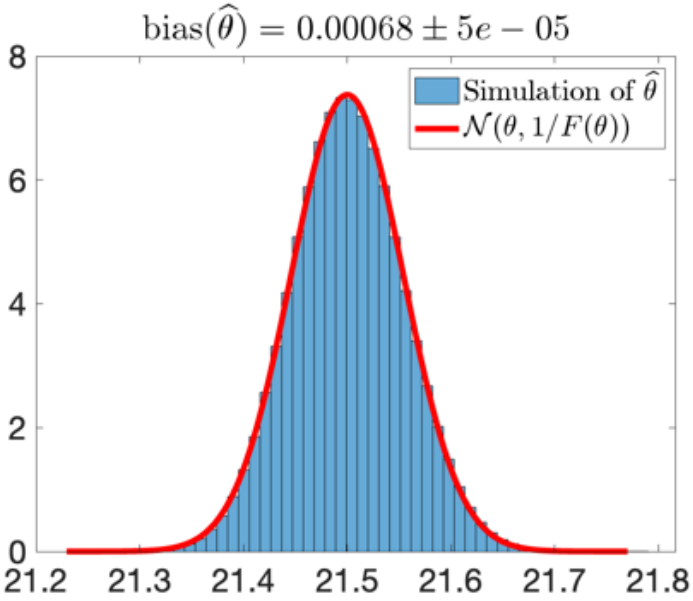
How to estimate θ from X?

$\hat{\theta} = m^{-1}(x) = C + B.\sinh\left(\frac{X}{A}\right)$

$\Rightarrow \mathrm{bias}(\hat{\theta}) = ?\,, \mathrm{var}(\hat{\theta}) = ?$ How good is this estimator? Almost perfect.

If $X \sim \mathcal{N}(m(\theta),\sigma^2) \Rightarrow F(\theta) = \left(\frac{m'(\theta)}{\sigma}\right)^2 \Rightarrow \mathrm{var}(\hat{\theta}) \geq \left(\frac{\sigma}{m'(\theta)}\right)^2$

Monte Carlo simulation for $\theta = 21.5$



How to **define** the "best" estimator?

Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample distributed along $p_X(x;\theta)$ and $\hat{\theta}$ be an estimator of θ (i.e. a function of $x_1, x_2, \ldots, x_N$)

For any θ, $\mathrm{MSE}(\hat{\theta},\theta) = E_X[(\hat{\theta}-\theta)^2]$ is useful to characterize the accuracy (bias & precision) of $\hat{\theta}$

Problem. $\hat{\theta}_{\mathrm{opt}} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)\ \ \forall\theta$ does not exist. What are the solutions?

1/ Search for the Uniformly Minimum Variance Unbiased $\hat{\theta}_{\mathrm{UMVU}} = \arg\min_{\text{unbiased } \hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)$

Example: $X = m(\theta) + N$ with $m(\theta) = A\,\mathrm{asinh}[(\theta - C)/B]$

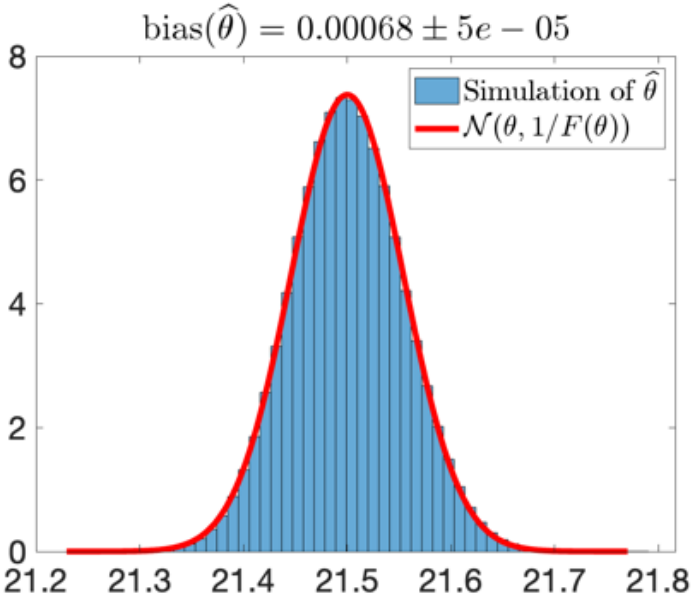
How to estimate θ from X?

$\hat{\theta} = m^{-1}(x) = C + B.\sinh\left(\frac{X}{A}\right)$

$\Rightarrow \mathrm{bias}(\hat{\theta}) = ?\,, \mathrm{var}(\hat{\theta}) = ?$ How good is this estimator? Almost perfect.

If $X \sim \mathcal{N}(m(\theta),\sigma^2) \Rightarrow F(\theta) = \left(\frac{m'(\theta)}{\sigma}\right)^2 \Rightarrow \mathrm{var}(\hat{\theta}) \geq \left(\frac{\sigma}{m'(\theta)}\right)^2$

If this is good enough, then we stop here.



How to **define** the "best" estimator?

Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample distributed along $p_X(x;\theta)$ and $\hat{\theta}$ be an estimator of θ (i.e. a function of $x_1, x_2, \ldots, x_N$)

For any θ, $\mathrm{MSE}(\hat{\theta},\theta) = E_X\left[(\hat{\theta}-\theta)^2\right]$ is useful to characterize the accuracy (bias & precision)

Problem. $\hat{\theta}_{\mathrm{opt}} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)\ \ \forall\theta$ does not exist. What are the solutions?

1/ Search for the Uniformly Minimum Variance Unbiased $\hat{\theta}_{\mathrm{UMVU}} = \arg\min_{\mathrm{unbiased}\ \hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)$

Cramér-Rao lower bound provides a precision of reference that may confirm that the UMVU is found.

2/ Search for $\hat{\theta}_{\mathrm{minimax}} = \arg\min_{\hat{\theta}} \max_{\theta} \left(\mathrm{MSE}(\hat{\theta},\theta)\right)$

This is the solution of "careful" people (e.g. Meudon PDR emulator by Einig & Palud)

Problems.
1. Calculations are often impossible to track, i.e. "brut force" may be required.
2. This criterion is not always adapted.

→ "*Neural network-based emulation of interstellar medium models*" A&A, 678, A198 (2023)

How to **define** the "best" estimator?

Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample distributed along $p_X(x;\theta)$ and $\hat{\theta}$ be an estimator of θ (i.e. a function of $x_1, x_2, \ldots, x_N$)

For any θ, $\mathrm{MSE}(\hat{\theta},\theta) = E_X\left[(\hat{\theta}-\theta)^2\right]$ is useful to characterize the accuracy (bias & precision)

Problem. $\hat{\theta}_{\mathrm{opt}} = \arg\min_{\hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)\ \ \forall\theta$ does not exist. What are the solutions?

1/ Search for the Uniformly Minimum Variance Unbiased $\hat{\theta}_{\mathrm{UMVU}} = \arg\min_{\mathrm{unbiased}\ \hat{\theta}} \mathrm{MSE}(\hat{\theta},\theta)$

Cramér-Rao lower bound provides a precision of reference that may confirm that the UMVU is found.

2/ Search for $\hat{\theta}_{\mathrm{minimax}} = \arg\min_{\hat{\theta}} \max_{\theta} \left(\mathrm{MSE}(\hat{\theta},\theta)\right)$

Calculations are often impossible to track, and this criterion is not necessarily adapted.

3/ When an *a prior* $\pi(\theta)$ is available, $\hat{\theta}_{\mathrm{Bayes}} = \arg\min_{\hat{\theta}} E_\theta\left(\mathrm{MSE}(\hat{\theta},\theta)\right)$

A prior is necessary and calculation often requires Monte Carlo approach (**see P. Palud's presentation**).
Well adapted for data accumulation *

** Galiano, F. 2018, MNRAS, 476, 1445*

Take home messages:

Entropy characterizes the **uncertainty** of X

Entropy $H(X) = -E[\log_2 \mathrm{P}(X)]$

Mutual information allows to select the **informative** lines

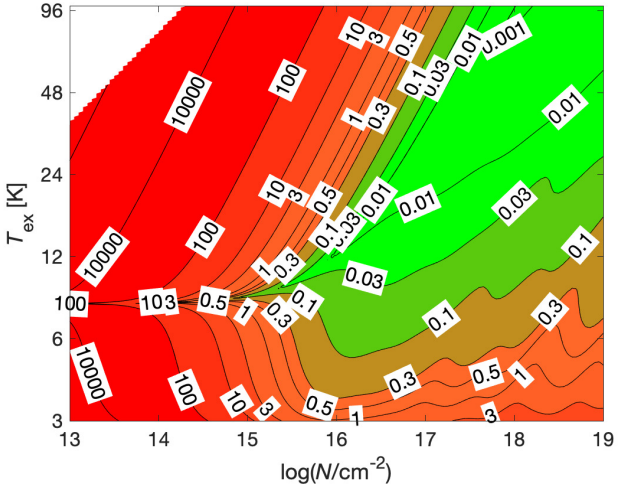
Link with S/N, correlation coefficient & MSE **only in Gaussian case**

Statistical learning can be seen as minimizing the code length

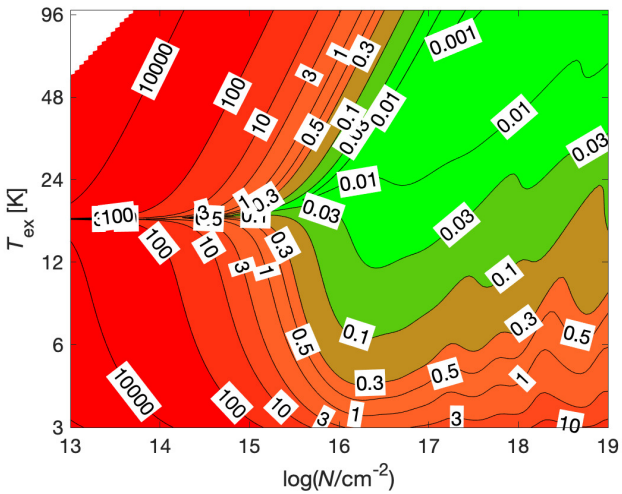
Cramér-Rao (lower) Bound provides **a precision of reference**

Precision of the column density as a function of the considered regime in the LTE regime

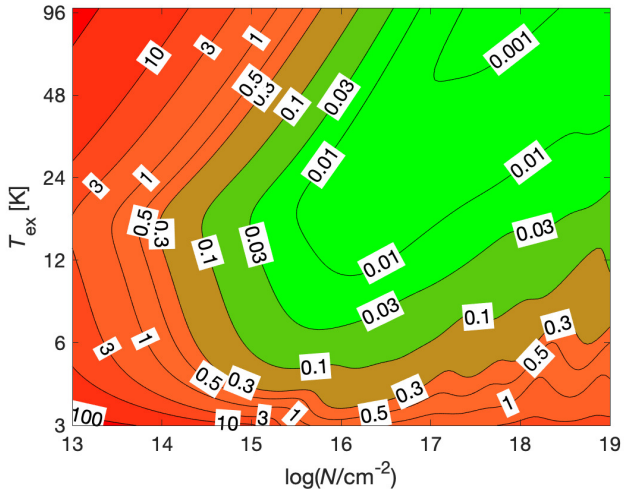
^{13}CO(1-0) -> 100 GHz



^{13}CO(2-1) -> 200 GHz



^{13}CO(1-0) & ^{13}CO(2-1)



The represented CRB of column density allows one to:

1. Quantify in terms of accuracy the gain of observing 2 transitions species -> **1+1 ≫ 2**
   -> Justify complementary telescope observation in 2022 in Flame nebula.
2. Check the maximum likelihood estimator (MLE) efficiency
   -> If MLE efficient -> provide error bars,

"C^{18}O,^{13}CO, and ^{12}CO abundances and excitation temperatures in the Orion B molecular cloud" A&A, 645, A26 (2020)

Summary

Information theory provides concepts (entropy, mutual information, …) that can be used for line selection.

CRB are “easy” to compute when a statistical model is available ($X \sim p_X(x;\theta)$) and provide precision of reference.

Both are **independent** of the choice of the estimation techniques.

A guide when choosing the model complexity

The minimum you can ask your physical model is the residue to be “small”. However, should you stop?

When the complexity increases, the precision of reference given by the CRB also increases.

It can be computed even **before** starting to search for an estimator.

Conclusion on information measures

You might find them difficult to interpret, but they are applicable in a wide range of situations.

-> there remain applications to discover in the interstellar medium.