Machine Learning for Cognitive Sciences: Principles and Applications

Simona Cocco (Physics Departement of ENS)
simona.cocco@phys.ens.fr CM &TD

Vito Dichio (Physics Departement of ENS)
vito.dichio@phys.ens.fr TD

Week 4

Recall of previous lecture

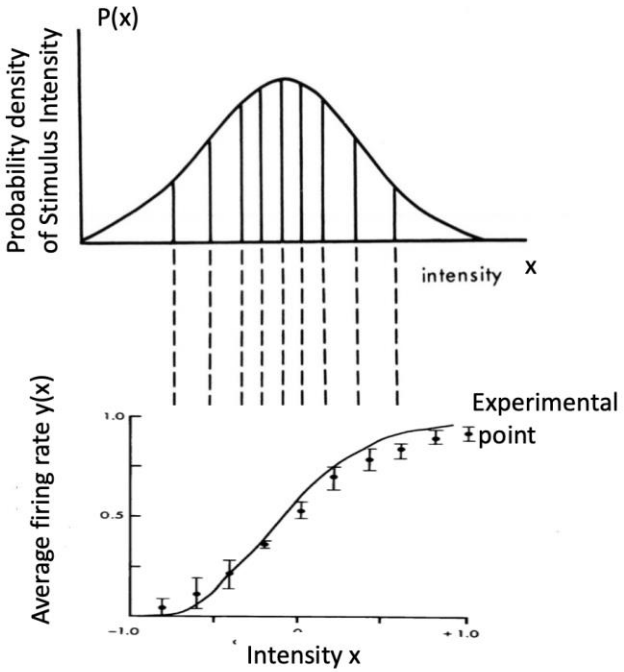
Mutual Information: connection with Efficient Coding:

- S.B. Laughlin computed the optimal-coding

(i.e. maximising the MI between the distribution of light intensity and the activity) by the spiking rate of retinal neurons in a fly .

He found that the optimal spiking rate y(x) is proportional to the integral of the light intensity distribution between its minimal value and x.

This non linear dependency has been confirmed experimentally.



Asymptotic Inference and Information Theory.

Definition of Cross Entropy, Kullback Leibler Divergence.

$$S_c(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = S(\hat{\boldsymbol{\theta}}) + D_{KL}(\hat{\boldsymbol{\theta}} || \boldsymbol{\theta})$$

Recall of previous lecture

Likelihood & Cross Entropy

Consider M data configurations drawn independently:

The likelihood of the data is

$$p(Y|\boldsymbol{\theta}) = \prod_{i=1}^{M} p(\boldsymbol{y}_i|\boldsymbol{\theta}) = \exp\left(M \times \frac{1}{M} \sum_{i=1}^{M} \log p(\boldsymbol{y}_i|\boldsymbol{\theta}) \right) .$$

The laws of large numbers ensure that:

The true & unknown distribution

$$\frac{1}{M} \sum_{i=1}^{M} \log p(\boldsymbol{y}_i|\boldsymbol{\theta}) \underset{M\to\infty}{\longrightarrow} \int d\boldsymbol{y}\, p(\boldsymbol{y}|\hat{\boldsymbol{\theta}})\, \log p(\boldsymbol{y}|\boldsymbol{\theta}) .$$

$$p(Y|\boldsymbol{\theta}) \approx e^{-M\, S_c(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})} ;$$

The inferred distribution

Recall of previous lecture

Posterior distribution & D_{KL}

To obtain the complete expression of the posterior distribution we introduce the Denominator:

$$p(\boldsymbol{\theta}|Y) = \frac{e^{-MS_c(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})}}{\int \mathrm{d}\boldsymbol{\theta}\, e^{-MS_c(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})}} \ .$$

$$p(\boldsymbol{\theta}|Y) \sim e^{-M[S_c(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})-S(\hat{\boldsymbol{\theta}})]} = e^{-MD_{KL}(\hat{\boldsymbol{\theta}}||\boldsymbol{\theta})} \ .$$

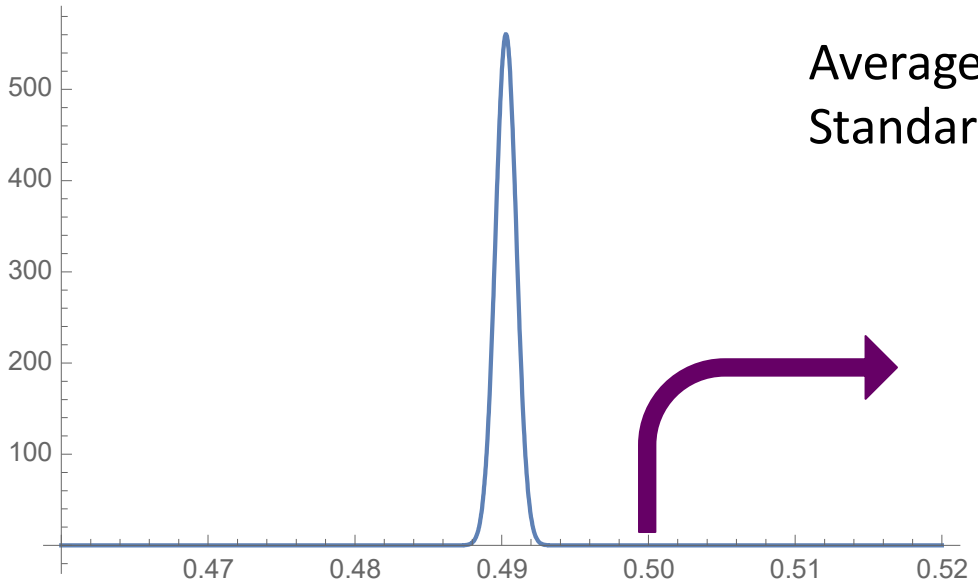
$$D_{KL}(\hat{\boldsymbol{\theta}}||\boldsymbol{\theta}_{hyp})$$

Controls how the posterior probability of the hypotesis θ varies with the number of data: If the hypothesis is not the good one its probability decays exponentially with the number of data.

D_{KL} gives the inverse of number of data you need to realize that your hypothesis is wrong

Laplace and the birth rate of boys & girls

Posterior distribution:



Average θ = 0.490291
Standard deviation θ = 0.007117

Probability that θ exceeds 0.5 = $\int_{0.5}^{1} d\theta \; p(\theta|y) \approx 10^{-42}$

Extremely unlikely!

D_{KL}=42 X log 10 /M$\sim 2\ 10^{-4}$

You need 5 000 data to understand that the probability are not equal

Plan of the lecture

Maximum Entropy Principle to Infer models from data

Information theory gives a theorethical framework for the less constraint model which reproduces some features of the data

The Maximum Entropy Principle

Bayes' rule offers us a powerful framework to infer model parameters from data.

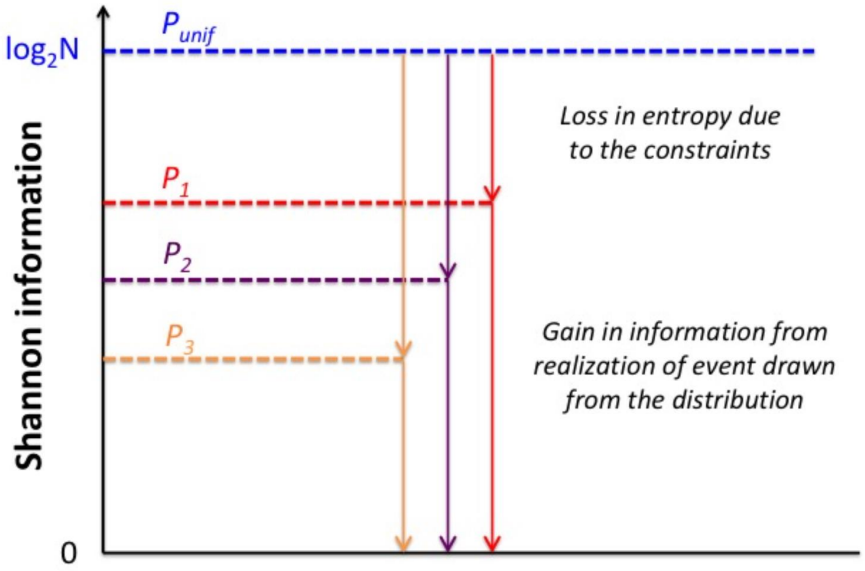
However, it presupposes some knowledge over the model to use for the likelihood function, i.e. for the distribution $p(\mathbf{y} \mid \theta)$.

Is it possible to avoid making such an assumption and derive model itself?

The Maximum Entropy Principle

E.T. Jaynes (1947)

Let's find the probability distribution for the configurations of a system by maximizing the entropy given what we know on the system.



- Maximal entropy: Less biased law.
- Model is maximally ignorant and constraint as less as possible the event

The Maximum Entropy Principle

Suppose we have some knowledge about the distribution of y, for example we know its average value:

$$m = \langle y \rangle = \sum_y p(y)\, y \ .$$

m= eg. spiking frequency of a neuron or its spiking probability in a small time bin

We also know that for the conservation of the probability:

$$\sum_y p(y) \quad = 1$$

The Maximum Entropy Principle

Suppose we have some knowledge about the distribution of y, for example we know its average value:

$$m = \langle y \rangle = \sum_y p(y)\, y \ .$$

We also know that for the conservation of the probability:

$$\sum_y p(y) \quad = 1$$

We look for the model with maximal entropy given these 2 constraints:

$$S(p, \lambda, \mu) = -\sum_y p(y) \log p(y) + \lambda \left(\sum_y p(y) - 1 \right) + h \left(\sum_y p(y)\, y - m \right)$$

- The parameters λ and h are called Lagrange multipliers

The Maximum Entropy Principle

We look for the p(y) maximizing

$$S(p,\lambda,\mu) = -\sum_y p(y)\log p(y) + \lambda\left(\sum_y p(y) - 1\right) + h\left(\sum_y p(y)\,y - m\right)$$

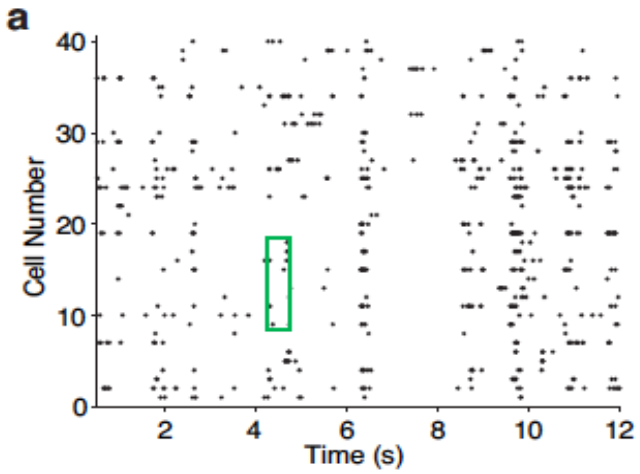
and obtain:

$$\frac{\delta S(p)}{\delta p(y)} = -\log p(y) - 1 + \lambda + h\ y = 0$$

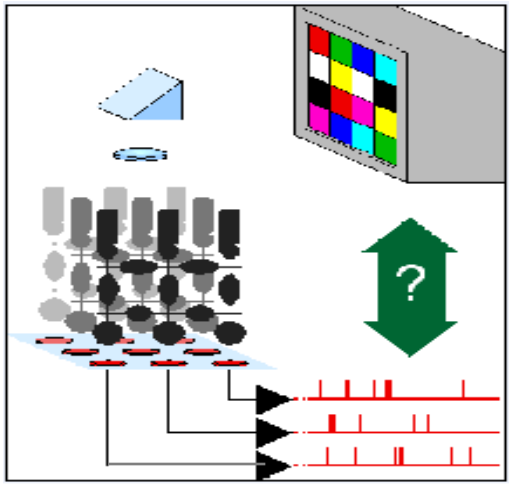
$$p(y) = \frac{e^{h\ y}}{\sum_{y'} e^{hy'}}$$

h has to be fixed in such a way to satisfy: $m = \langle y \rangle = \sum_y p(y)\,y\ .$

Max Entropy model to reproduce spiking activity of retinal neurons



$P(s_1, s_2, \ldots, s_N)$?



Time is discretized in time windows of size $\Delta t = 20ms$

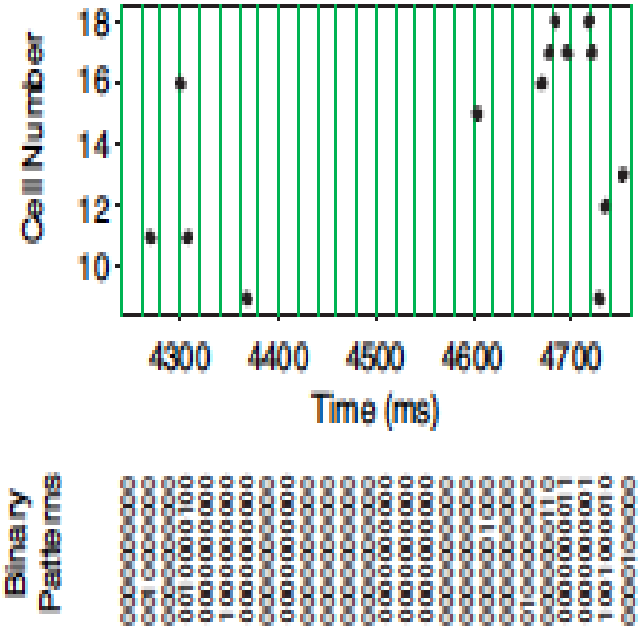
Binary variable to reproduce activity in a bin

$$s_i(k) = \begin{cases} 1 \text{ if at least one spike in time window k} \\ 0 \text{ if no spike in time windows k} \end{cases}$$

Spiking probabilities from DATA

$$p_i = \frac{1}{M} \sum_{K=1}^{M} s_i(k) ,$$

Max Entropy model to reproduce spiking activity of retinal neurons



Independent model for the spiking activity

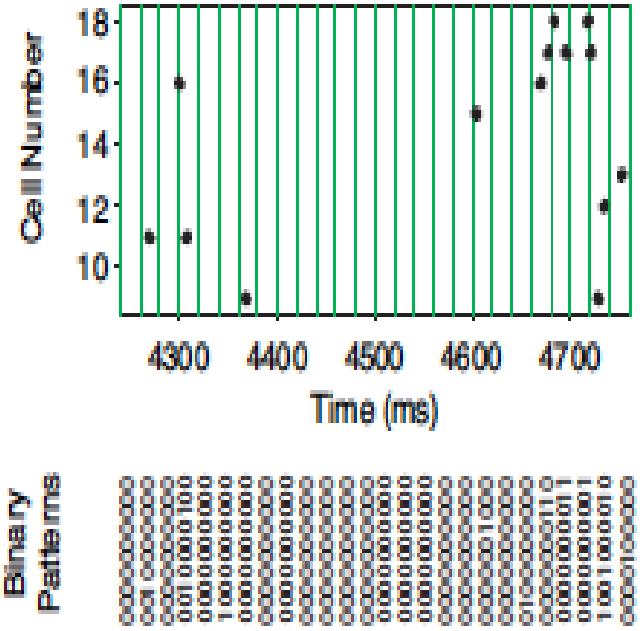
$$P(s_1, s_2, .., s_L) = \prod_{i=1}^{L} p(s_i)$$

As the variables are independent:

$$p(s_i) = \frac{e^{h_i\ s_i}}{\sum_{s_i=0,1} e^{h_i s_i}} = \frac{e^{h_i\ s_i}}{1 + e^{h_i}}$$

h_i : local fields

Max Entropy model to reproduce spiking activity of retinal neurons



Independent model for the spiking activity

$$P(s_1, s_2, .., s_L) = \prod_{i=1}^{L} p(s_i)$$

As the variables are independent:

$$p(s_i) = \frac{e^{h_i\ s_i}}{\sum_{s_i=0,1} e^{h_i s_i}} = \frac{e^{h_i\ s_i}}{1 + e^{h_i}}$$

h_i : local fields

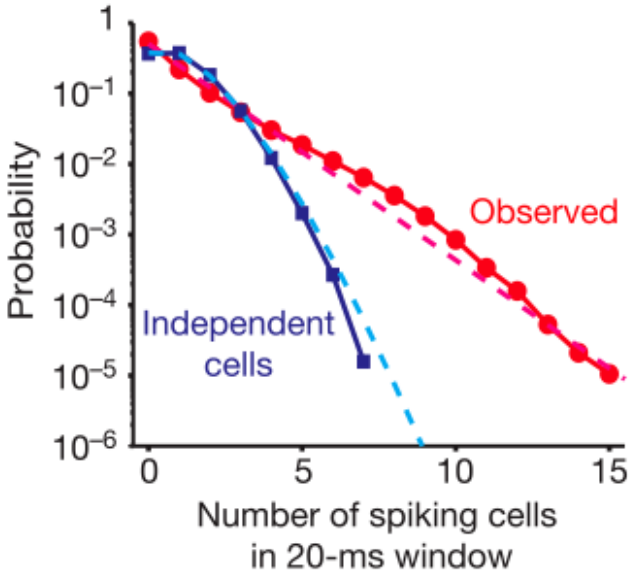
$$p(s_i = 1) = \frac{e^{h_i}}{\sum_{s_i=0,1} e^{h_i s_i}} = p_i$$

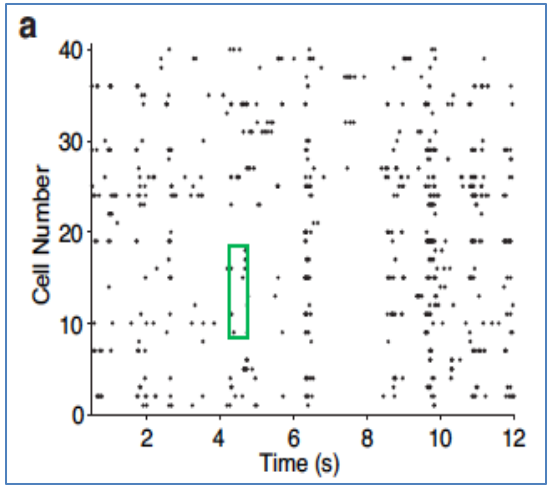
$$p(s_i = 0) = \frac{1}{\sum_{s_i=0,1} e^{h_i s_i}} = 1\text{-}p_i$$

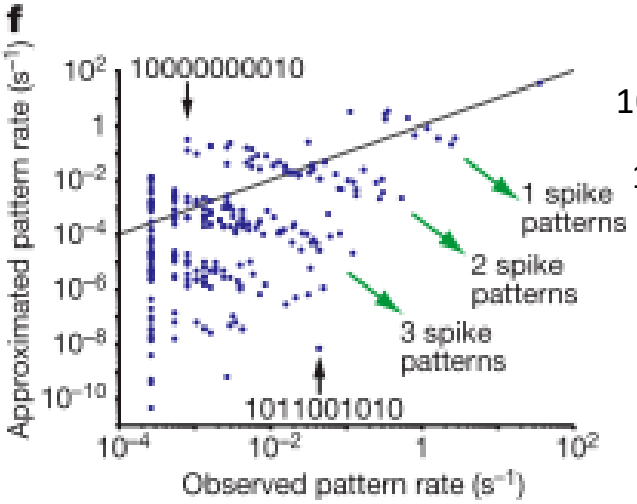
$$h_i = \log \frac{p_i}{1 - p_i}$$

Results of the Ising model inferred on retinal recordings

Independent Ising model does not reproduce probabilities that multiple neurons spike in the same time window & neural firing patterns





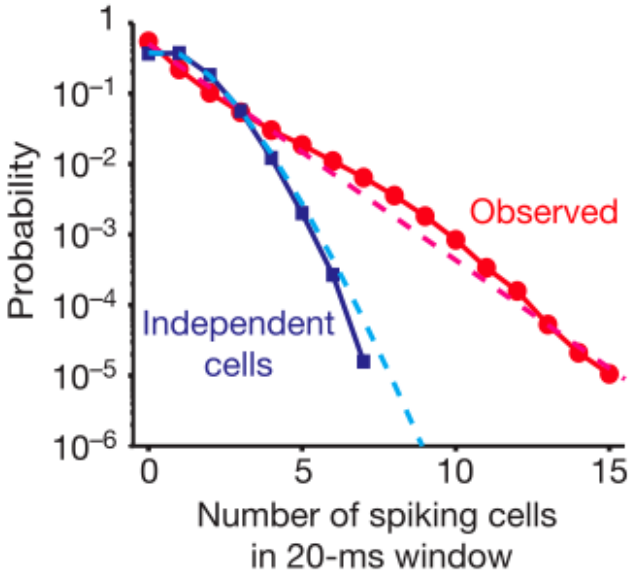


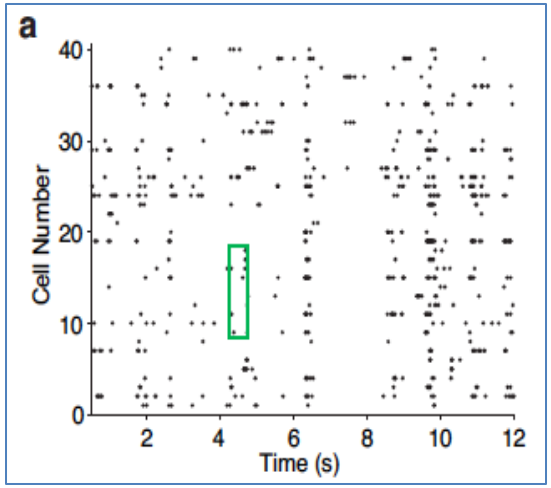
	Data	Model
100000000100	3/hour	1/3 seconds
1011001010	1/minute	1/3 years

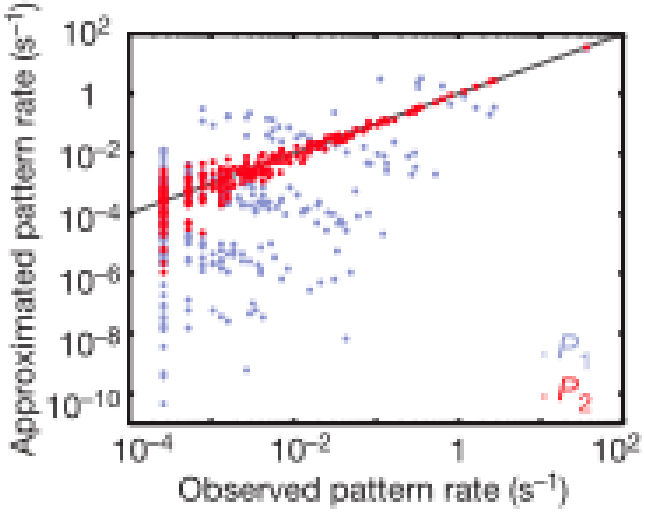
Schneidmann et al., Nature (2006)

Results of the Ising model inferred on retinal recordings

Coupled Ising model does reproduce probabilities that multiple neurons spike in the same time window & neural firing patterns







Imposing also pairwise correlations:

$$p_{ij} = \frac{1}{M}\sum_{K=1}^{M} s_i(k)s_j(k)$$

Schneidmann et al., Nature (2006)

Equivalence between Bayesian Inference and Max Entropy models.

The Max Entropy principle for L binary independent variables $s_l = \{0,1\}$ defines the Likelihood:

$$P(s_1, s_2, .., s_L) = \prod_{i=1}^{L} p(s_i) \qquad \mathrm{p}(s_i) = \frac{e^{h_i \, s_i}}{\sum_{s_i=0,1} e^{h_i \, s_i}}$$

The parameters of the model are the local fields, h_l to be inferred from the condition:

$$\mathrm{p}(s_i = 1) = \frac{e^{h_i}}{\sum_{s_i=0,1} e^{h_i \, s_i}} = p_i$$

What is the relationship with the maximisation of the likelihood/ minimisation of the Cross Entropy to find θ in Bayes inference?

$$p(\boldsymbol{\theta}|Y) = \frac{p(Y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(Y)} \; .$$

Equivalence between Bayesian Inference and Max Entropy models.

For $$p(s_i{}^b) = \frac{e^{h_i\ s_i^b}}{\sum_{s_i=0,1} e^{h_i s_i}}$$

$$\prod_{i=1}^{M} p(s_i{}^b) = \frac{e^{\sum_{b=1}^{M} h_i \quad s_i^b}}{\left[\sum_{s_i=0,1} e^{h_i s_i}\right]^M} = \frac{e^{M\, h_i\, p_i}}{\left[\sum_{s_i=0,1} e^{h_i s_i}\right]^M}$$

i

Equivalence between Bayesian Inference and Max Entropy models.

For $$p(s_i{}^b) = \frac{e^{h_i\ s_i^b}}{\sum_{s_i=0,1} e^{h_i s_i}}$$

$$\prod_{i=1}^{M} p(s_i{}^b) = \frac{e^{\sum_{b=1}^{M} h_i \quad s_i^b}}{\left[\sum_{s_i=0,1} e^{h_i s_i}\right]^M} = \frac{e^{M\, h_i\, p_i}}{\left[\sum_{s_i=0,1} e^{h_i s_i}\right]^M} = e^{-MSc(p(s),p_i)}$$

i

$$S_c(p(s), p_i) = \log[\sum_{s_i=0,1} e^{h_i s_i}] - h_i\, p_i$$

Equivalence between Bayesian Inference and Max Entropy models.

$$S_c(p(s), p_i) = \log[\sum_{s_{i=0,1}} e^{h_i s_i}] - h_i\, p_i$$

Maximal Log-likelihood == Minimal Cross entropy :

$$\frac{\partial S_C}{\partial h_l} = 0 \Rightarrow \frac{\log[\sum_{s_{i=0,1}} e^{h_i s_i}]}{h_l} = p_l$$

$$\frac{\sum_{s_{i=0,1}} s_i\, e^{h_i s_i}}{\sum_{s_i=0,1} e^{h_i s_i}} = p_i$$

Bring home message :

The entropy and Asymptotic Inference:

Definition of Cross Entropy and Kullback Leibler Divergence,

The Maximum Entropy Principle

Information theory allows us to infer a probability distribution from a set of data constraints by the Max Entropy formulation.

the less constraint model which defines the likelihood: $p(Y|\boldsymbol{\theta})$

given some observables.