# Introduction to Bayesian Statistics - 3

PhD Physics course (XXVIII ciclo)

Università di Trieste

*Edoardo Milotti*

# Bayesian inference and maximum-likelihood

uniform distribution (in general, improper)

likelihood

$$p(\boldsymbol{\theta} \mid \mathbf{d}, I) = \frac{P(\mathbf{d} \mid \boldsymbol{\theta}, I)}{P(\mathbf{d} \mid I)} \cdot p(\boldsymbol{\theta} \mid I)$$

$$= \frac{\mathcal{L}(\mathbf{d}, \boldsymbol{\theta})}{P(\mathbf{d} \mid I)} \cdot p(\boldsymbol{\theta} \mid I) \propto \mathcal{L}(\mathbf{d}, \boldsymbol{\theta})$$

evidence

in this case the set of parameters that maximizes the posterior (MAP) is also the set that maximizes the likelihood (MLE)

# max-likelihood in the context of a Gaussian model

independent data $d_i = \{x_i, y_i\}$ with negligible error on $x$

$$\mathcal{L}(\mathbf{d},\boldsymbol{\theta}) \propto \prod_k \exp\left( -\frac{\left[y_k - y(x_k;\boldsymbol{\theta})\right]^2}{2\sigma_k^2} \right)$$

$$= \exp\left\{ -\sum_k \frac{\left[y_k - y(x_k;\boldsymbol{\theta})\right]^2}{2\sigma_k^2} \right\} = \exp\left\{-\chi^2\right\}$$

$$\chi^2 = \sum_k \frac{\left[y_k - y(x_k;\boldsymbol{\theta})\right]^2}{2\sigma_k^2}$$

max-likelihood implies min chi-square and least-squares method expansion about minimum of chi square

$$\chi^2(\boldsymbol{\theta}) \approx \chi^2(\boldsymbol{\theta}_m) + \frac{1}{2}\Delta\boldsymbol{\theta}^T \mathbf{H} \Delta\boldsymbol{\theta} \qquad H_{ij} = \left.\frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_m}$$

Hessian

$$p(\boldsymbol{\theta} \mid \mathbf{d}, I) \propto \mathcal{L}(\mathbf{d}, \boldsymbol{\theta}) \propto \exp\left\{-\chi^2\right\} \propto \exp\left\{-\frac{1}{2}\Delta\boldsymbol{\theta}^T \mathbf{H} \Delta\boldsymbol{\theta}\right\}$$

$$p(\boldsymbol{\theta} \mid \mathbf{d}, I) = (2\pi)^{-n/2} (\det \mathbf{V})^{-1} \exp\left\{-\frac{1}{2}\Delta\boldsymbol{\theta}^T \mathbf{V}^{-1} \Delta\boldsymbol{\theta}\right\}$$

covariance matrix

famous frequentist textbook

"Statistical Methods in Experimental Physics", Eadie, Drijard, James, Roos, and Sadoulet, American Elsevier, 1971

statistics for physicists

http://www.slac.stanford.edu/BFROOT/www/Statistics/bibliography.html

http://www-cdf.fnal.gov/physics/statistics/

http://www.nu.to.infn.it/Statistics/

summary notes, mostly on frequentist statistics

http://pdg.lbl.gov/2012/reviews/rpp2012-rev-statistics.pdf

notes on MINUIT, a program for function minimization (intensively used for chi-square minimization)

http://wwwasdoc.web.cern.ch/wwwasdoc/minuit/minmain.html

# Prior distributions

The choice of prior distribution is an important aspect of Bayesian inference

- prior distributions are one of the main targets of frequentists: how much do posteriors differ when we choose different priors?
- there are two main “objective” methods for the choice of priors

# Priors related to the symmetry properties of the likelihood functions

1. the likelihood may have important symmetries

$$p(\theta \mid I) \cdot \frac{P(D \mid \theta, I)}{\int_{\Theta} P(D \mid \theta, I) \cdot p(\theta \mid I) \, d\theta} = p(\theta \mid D, I)$$

2. if the prior shares the same symmetries ...

3. ... then, the posterior has the same symmetry properties as well.

# A. translation invariance

$$\mathcal{L}(d,\theta) = g\left(\theta - f(d)\right)$$

When the likelihood has this symmetry, then the parameter transformations that keep the difference

$$\theta - f(d)$$

constant, do not change the likelihood.

Example, structure of the Gaussian likelihood

$$P(d \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[ -\frac{(d-\mu)^2}{2\sigma^2} \right]$$

this likelihood is invariant with respect to translations

$$\mu' = \mu + b; \quad d\mu' = d\mu$$

$$P(\mu)\,d\mu = P(\mu')\,d\mu' = P(\mu + b)\,d\mu$$

this must hold for infinitesimal translations as well, therefore

$$P(\mu) = P(\mu + b) = P(\mu) + P'(\mu)b$$

and we find

$$P'(\mu) = 0$$
$$P(\mu) = \text{constant}$$

thus we find a uniform distribution (in general an improper one)

B. scale invariance

$$P(d \mid \tau) = \frac{1}{\tau}\exp\left(-\frac{d}{\tau}\right) = \frac{1}{d}\left(\frac{d}{\tau}\right)\exp\left(-\frac{d}{\tau}\right)$$
$$= \frac{1}{d}\exp\left[\left(\ln d - \ln \tau\right) - \exp\left(\ln d - \ln \tau\right)\right]$$

the likelihood is invariant with respect to scale changes, such that

$$d \to \alpha d; \qquad \tau \to \alpha\tau$$

$$\tau' = \alpha\tau; \quad d\tau' = \alpha d\tau$$

$$P(\tau)d\tau = P(\tau')d\tau' = P(\alpha\tau)\alpha d\tau$$

expanding about α = 1
we find:

$$P(\tau) = \alpha P(\alpha\tau) = \left[1+(\alpha-1)\right]P\left(\tau+(\alpha-1)\tau\right)$$
$$\approx P(\tau)+\left[P(\tau)+\tau P'(\tau)\right](\alpha-1)$$

$$P(\tau) = -\tau P'(\tau)$$

$$\frac{P'(\tau)}{P(\tau)} = -\frac{1}{\tau}; \quad \ln P = \ln\frac{1}{\tau} + \text{cost}$$

$$P(\tau) = \frac{C}{\tau}$$

usually this is improper

(Jeffreys' prior)

# A short refresher on entropy in statistical mechanics

- consider a system where states *n* are occupied by $N_n$ distinguishable particles ($n$, $n=1, \ldots , M$).

- the number of ways to fill these states is given by

$$\Omega = \frac{N!}{N_1!N_2!\ldots N_M!}$$

- then Boltzmann's entropy is

$$S_B = k_B \ln \Omega = k_B \ln \frac{N!}{N_1!N_2!\ldots N_M!} \approx k_B \left( \left( N \ln N - N \right) - \sum_n \left( N_n \ln N_n - N_n \right) \right)$$

$$= k_B \left( N \ln N - \sum_n N p_n \left( \ln p_n + \ln N \right) \right) = k_B \sum_n p_n \ln \frac{1}{p_n}$$

$$S_B = k_B \sum_i p_i \ln \frac{1}{p_i}$$

probability of physical states

*Boltzmann's entropy is functionally the same as Shannon's entropy*

$$S_I = \sum_i p_i \log_2 \frac{1}{p_i}$$

probability of source symbols

Edwin T. Jaynes (1922-1998), introduced the method of maximum entropy in statistical mechanics: when we start from the informational entropy (Shannon's entropy) and we use it introduce Boltzmann's entropy we reobtain the whole of statistical mechanics by maximizing entropy.

In a sense, statistical mechanics arises from a comprehensive "principle of maximum entropy".

http://bayes.wustl.edu/etj/etj.html

Reprinted from THE PHYSICAL REVIEW, Vol. 106, No. 4, 620–630, May 15, 1957
Printed in U. S. A.

# Information Theory and Statistical Mechanics

E. T. JAYNES
*Department of Physics, Stanford University, Stanford, California*
(Received September 4, 1956; revised manuscript received March 4, 1957)

Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge, and leads to a type of statistical inference which is called the maximum-entropy estimate. It is the least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information. If one considers statistical mechanics as a form of statistical inference rather than as a physical theory, it is found that the usual computational rules, starting with the determination of the partition function, are an immediate consequence of the maximum-entropy principle. In the resulting "subjective statistical mechanics," the usual rules are thus justified independently of any physical argument, and in particular independently of experimental verification; whether or not the results agree with experiment, they still represent the best estimates that could have been made on the basis of the information available.

It is concluded that statistical mechanics need not be regarded as a physical theory dependent for its validity on the truth of additional assumptions not contained in the laws of mechanics (such as ergodicity, metric transitivity, equal *a priori* probabilities, etc.). Furthermore, it is possible to maintain a sharp distinction between its physical and statistical aspects. The former consists only of the correct enumeration of the states of a system and their properties; the latter is a straightforward example of statistical inference.

Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge, and leads to a type of statistical inference which is called the maximum-entropy estimate. It is the least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information.

If one considers statistical mechanics as a form of statistical inference rather than as a physical theory, it is found that the usual computational rules, starting with the determination of the partition function, are an immediate consequence of the maximum-entropy principle.

In the resulting "subjective statistical mechanics," the usual rules are thus justified independently of any physical argument, and in particular independently of experimental verification; whether or not the results agree with experiment, they still represent the best estimates that could have been made on the basis of the information available.

It is concluded that statistical mechanics need not be regarded as a physical theory dependent for its validity on the truth of additional assumptions not contained in the laws of mechanics (such as ergodicity, metric transitivity, equal a priori probabilities, etc.). Furthermore, it is possible to maintain a sharp distinction between its physical and statistical aspects. The former consists only of the correct enumeration of the states of a system and their properties; the latter is a straightforward example of statistical inference.

## We maximize entropy in order to solve problems and find prior distributions ...

*1. The kangaroo problem (Jaynes)*

- *Basic information*: one third of all kangaroos has blue eyes, and one third is left-handed.
- Question: which fraction of kangaroos has both blue eyes and is left-handed?

| | left | ~left |
|---|---|---|
| blue | 1/9 | 2/9 |
| ~blue | 2/9 | 4/9 |

no correlation

| | left | ~left |
|---|---|---|
| blue | 0 | 1/3 |
| ~blue | 1/3 | 1/3 |

maximum negative correlation

| | left | ~left |
|---|---|---|
| blue | 1/3 | 0 |
| ~blue | 0 | 2/3 |

maximum positive correlation

probabilities $p_{bl}$ $p_{\bar{b}l}$ $p_{b\bar{l}}$ $p_{\overline{bl}}$

entropy (proportional to Shannon's entropy)

$$S = p_{bl} \ln \frac{1}{p_{bl}} + p_{\bar{b}l} \ln \frac{1}{p_{\bar{b}l}} + p_{b\bar{l}} \ln \frac{1}{p_{b\bar{l}}} + p_{\overline{bl}} \ln \frac{1}{p_{\overline{bl}}}$$

constraints (3 constraints, 4 unknowns)

$$p_{bl} + p_{\bar{b}l} + p_{b\bar{l}} + p_{\overline{bl}} = 1$$

$$p_{bl} + p_{b\bar{l}} = 1/3$$

$$p_{bl} + p_{\bar{b}l} = 1/3$$

# entropy maximization with constraints

$$S_V = \left( p_{bl} \ln \frac{1}{p_{bl}} + p_{\bar{b}l} \ln \frac{1}{p_{\bar{b}l}} + p_{b\bar{l}} \ln \frac{1}{p_{b\bar{l}}} + p_{\overline{bl}} \ln \frac{1}{p_{\overline{bl}}} \right)$$

$$+\lambda_1 \left( p_{bl} + p_{\bar{b}l} + p_{b\bar{l}} + p_{\overline{bl}} - 1 \right) + \lambda_2 \left( p_{bl} + p_{b\bar{l}} - 1/3 \right) + \lambda_3 \left( p_{bl} + p_{\bar{b}l} - 1/3 \right)$$

$$\frac{\partial S_V}{\partial p_{bl}} = -\ln p_{bl} - 1 + \lambda_1 + \lambda_2 + \lambda_3 = 0$$

$$\frac{\partial S_V}{\partial p_{\bar{b}l}} = -\ln p_{\bar{b}l} - 1 + \lambda_1 + \lambda_3 = 0$$

$$\frac{\partial S_V}{\partial p_{b\bar{l}}} = -\ln p_{b\bar{l}} - 1 + \lambda_1 + \lambda_2 = 0$$

$$\frac{\partial S_V}{\partial p_{\overline{bl}}} = -\ln p_{\overline{bl}} - 1 + \lambda_1 = 0$$

$$p_{bl} = \exp\left(-1 + \lambda_1 + \lambda_2 + \lambda_3\right)$$

$$p_{\bar{b}l} = \exp\left(-1 + \lambda_1 + \lambda_3\right)$$

$$p_{b\bar{l}} = \exp\left(-1 + \lambda_1 + \lambda_2\right)$$

$$p_{\overline{bl}} = \exp\left(-1 + \lambda_1\right)$$

$$\begin{cases} p_{\bar{b}l} = p_{\bar{b}\bar{l}} \exp(\lambda_3) \\ p_{b\bar{l}} = p_{\bar{b}\bar{l}} \exp(\lambda_2) \\ p_{bl} = p_{\bar{b}\bar{l}} \exp(\lambda_2 + \lambda_3) \end{cases} \Rightarrow \quad p_{\bar{b}l} p_{b\bar{l}} = p_{bl} p_{\bar{b}\bar{l}}$$

$$\begin{cases} p_{bl} + p_{\bar{b}l} + p_{b\bar{l}} + p_{\bar{b}\bar{l}} = 1 \\ p_{bl} + p_{b\bar{l}} = 1/3 \\ p_{bl} + p_{\bar{b}l} = 1/3 \\ p_{\bar{b}l} p_{b\bar{l}} = p_{bl} p_{\bar{b}\bar{l}} \end{cases} \Rightarrow \begin{cases} p_{b\bar{l}} = p_{\bar{b}l} = 1/3 - p_{bl} \\ p_{\bar{b}\bar{l}} = 1/3 + p_{bl} \\ (1/3 - p_{bl})^2 = p_{bl}/3 + p_{bl}^2 \\ 1/9 - 2p_{bl}/3 + p_{bl}^2 = p_{bl}/3 + p_{bl}^2 \end{cases}$$

$$\Rightarrow \boxed{p_{bl} = \frac{1}{9}; \quad p_{b\bar{l}} = p_{\bar{b}l} = \frac{2}{9}; \quad p_{\bar{b}\bar{l}} = \frac{4}{9}}$$

this solution coincides with the least informative distribution (no correlation)

*2. Solution of underdetermined systems of equations*

In this problem there are fewer equations than unknowns; the system of equations is underdetermined, and in general there is no unique solution.

The maximum entropy method helps us find a reasonable solution, the least informative one (least correlations between variables)

*Example*:

$$\begin{aligned} 3x + 5y + 1.1z &= 10 \\ -2.1x + 4.4y - 10z &= 1 \end{aligned} \qquad (x, y, z > 0)$$

$$3x + 5y + 1.1z = 10$$
$$-2.1x + 4.4y - 10z = 1$$
$$(x, y, z > 0)$$

this ratio can be taken to be a “probability”

$$S = -\left( \frac{x}{x+y+z} \ln \frac{x}{x+y+z} + \frac{y}{x+y+z} \ln \frac{y}{x+y+z} + \frac{z}{x+y+z} \ln \frac{z}{x+y+z} \right)$$
$$= -\frac{1}{x+y+z} \left[ x \ln x + y \ln y + z \ln z - (x+y+z) \ln (x+y+z) \right]$$

$$Q = S + \lambda (3x + 5y + 1.1z - 10) + \mu (-2.1x + 4.4y - 10z - 1)$$

$$\frac{\partial Q}{\partial x} = -\frac{\ln x - \ln (x+y+z)}{x+y+z} + \frac{x \ln x + y \ln y + z \ln z - (x+y+z) \ln (x+y+z)}{(x+y+z)^2} + 3\lambda - 2.1\mu$$
$$= \frac{(y+z) \ln x + y \ln y + z \ln z}{(x+y+z)^2} + 3\lambda - 2.1\mu = 0$$

$$\frac{\partial Q}{\partial x} = \frac{(y+z)\ln x + y\ln y + z\ln z}{(x+y+z)^2} + 3\lambda - 2.1\mu = 0$$

$$\frac{\partial Q}{\partial y} = \frac{x\ln x + (x+z)\ln y + z\ln z}{(x+y+z)^2} + 5\lambda + 4.4\mu = 0$$

$$\frac{\partial Q}{\partial z} = \frac{x\ln x + y\ln y + (x+y)\ln z}{(x+y+z)^2} + 1.1\lambda - 10\mu = 0$$

$$10 = 3x + 5y + 1.1z$$

$$1 = -2.1x + 4.4y - 10z$$

$x = 0.606275;\quad y = 1.53742;\quad z = 0.449148;$

$\lambda = 0.0218739;\quad \mu = \text{-}0.017793$

this is an example of an “ill-posed” problem

the solution that we found is a kind of regularization

of the ill-posed problem

# Finding priors with the maximum entropy method

$$S = \sum_k p_k \ln \frac{1}{p_k} = -\sum_k p_k \ln p_k$$

Shannon entropy

entropy maximization when all information is missing and normalization is the only constraint:

$$\frac{\partial}{\partial p_k}\left[-\sum_k p_k \ln p_k + \lambda\left(\sum_k p_k - 1\right)\right] = -\left(\ln p_k + 1\right) + \lambda = 0$$

$$p_k = e^{\lambda-1}; \quad \sum_k p_k = \sum_k e^{\lambda-1} = Ne^{\lambda-1} = 1 \quad \Rightarrow \quad p_k = 1/N$$

# entropy maximization when the mean is known μ

$$\frac{\partial}{\partial p_k}\left[-\sum_k p_k \ln p_k + \lambda_0\left(\sum_k p_k - 1\right) + \lambda_1\left(\sum_k x_k p_k - \mu\right)\right]$$
$$= -\left(\ln p_k + 1\right) + \lambda_0 + \lambda_1 x_k = 0$$

incomplete solution...

$$p_k = e^{\lambda_0 + \lambda_1 x_k - 1};$$

We must satisfy two constraints now ...

$$p_k = e^{\lambda_0 + \lambda_1 x_k - 1}$$

$$\sum_k p_k = \sum_k e^{\lambda_0 + \lambda_1 x_k - 1} = e^{\lambda_0 - 1} \sum_k e^{\lambda_1 x_k} = 1$$

$$\sum_k x_k p_k = \sum_k x_k e^{\lambda_0 + \lambda_1 x_k - 1} = e^{\lambda_0 - 1} \sum_k x_k e^{\lambda_1 x_k} = \mu$$

$$e^{\lambda_0 - 1} = \frac{1}{\sum_k e^{\lambda_1 x_k}}; \qquad \frac{\sum_k x_k e^{\lambda_1 x_k}}{\sum_k e^{\lambda_1 x_k}} = \mu$$

no analytic solution, only numerical

# Example : the biased die

(E. T. Jaynes: *Where do we stand on Maximum Entropy?* In *The Maximum Entropy Formalism*; Levine, R. D. and Tribus, M., Eds.; MIT Press, Cambridge, MA, 1978)

mean value of throws for an unbiased die

$$\frac{1}{6}(1+2+3+4+5+6)=\frac{21}{6}=3.5$$

mean value for a biased die

$$3.5(1+\varepsilon)$$

*Problem: for a given mean value of the biased die, what is the probability distribution of each value?*

*The mean value is insufficient information, and we use the maximum entropy method to find the most likely distribution (the least informative one).*

# entropy maximization with the biased die:

$$\frac{\partial}{\partial p_k}\left[-\sum_{k=1}^{6} p_k \ln p_k + \lambda_0\left(\sum_{k=1}^{6} p_k - 1\right) + \lambda_1\left(\sum_{k=1}^{6} k p_k - \frac{7}{2}(1+\varepsilon)\right)\right]$$

$$= -(\ln p_k + 1) + \lambda_0 + k\lambda_1 = 0$$

$$p_k = e^{\lambda_0 + \lambda_1 k - 1}$$

$$\sum_{k=1,6} p_k = e^{\lambda_0 - 1}\sum_{k=1,6} e^{\lambda_1 k} = 1$$

$$\sum_{k=1,6} k p_k = e^{\lambda_0 - 1}\sum_{k=1,6} k e^{\lambda_1 k} = \frac{7}{2}(1+\varepsilon)$$

$$e^{\lambda_0 - 1} = \frac{1}{\sum_{k=1,6} e^{\lambda_1 k}}; \quad \frac{\sum_{k=1,6} k p_k}{\sum_{k=1,6} e^{\lambda_1 k}} = \frac{7}{2}(1+\varepsilon)$$

we still have to satisfy the constraints ...

$$e^{\lambda_0-1}\sum_{k=1,6} e^{\lambda_1 k} = e^{\lambda_0-1}\left(\sum_{k=0,6} e^{\lambda_1 k} - 1\right) = e^{\lambda_0-1}\left(\frac{1-e^{7\lambda_1}}{1-e^{\lambda_1}} - 1\right) = 1$$

$$\frac{\sum_{k=1,6} k e^{\lambda_1 k}}{\sum_{k=1,6} e^{\lambda_1 k}} = \frac{\partial}{\partial\lambda_1}\ln\sum_{k=1,6} e^{\lambda_1 k} = \frac{\partial}{\partial\lambda_1}\ln\left(e^{\lambda_1}\sum_{k=0,5} e^{\lambda_1 k}\right)$$

$$= \frac{\partial}{\partial\lambda_1}\left[\lambda_1 + \ln\left(1-e^{6\lambda_1}\right) - \ln\left(1-e^{\lambda_1}\right)\right]$$

$$= 1 - \frac{6e^{6\lambda_1}}{1-e^{6\lambda_1}} + \frac{e^{\lambda_1}}{1-e^{\lambda_1}} = \frac{7}{2}(1+\varepsilon)$$

the Lagrange multipliers are obtained from nonlinear equations and we must use numerical methods

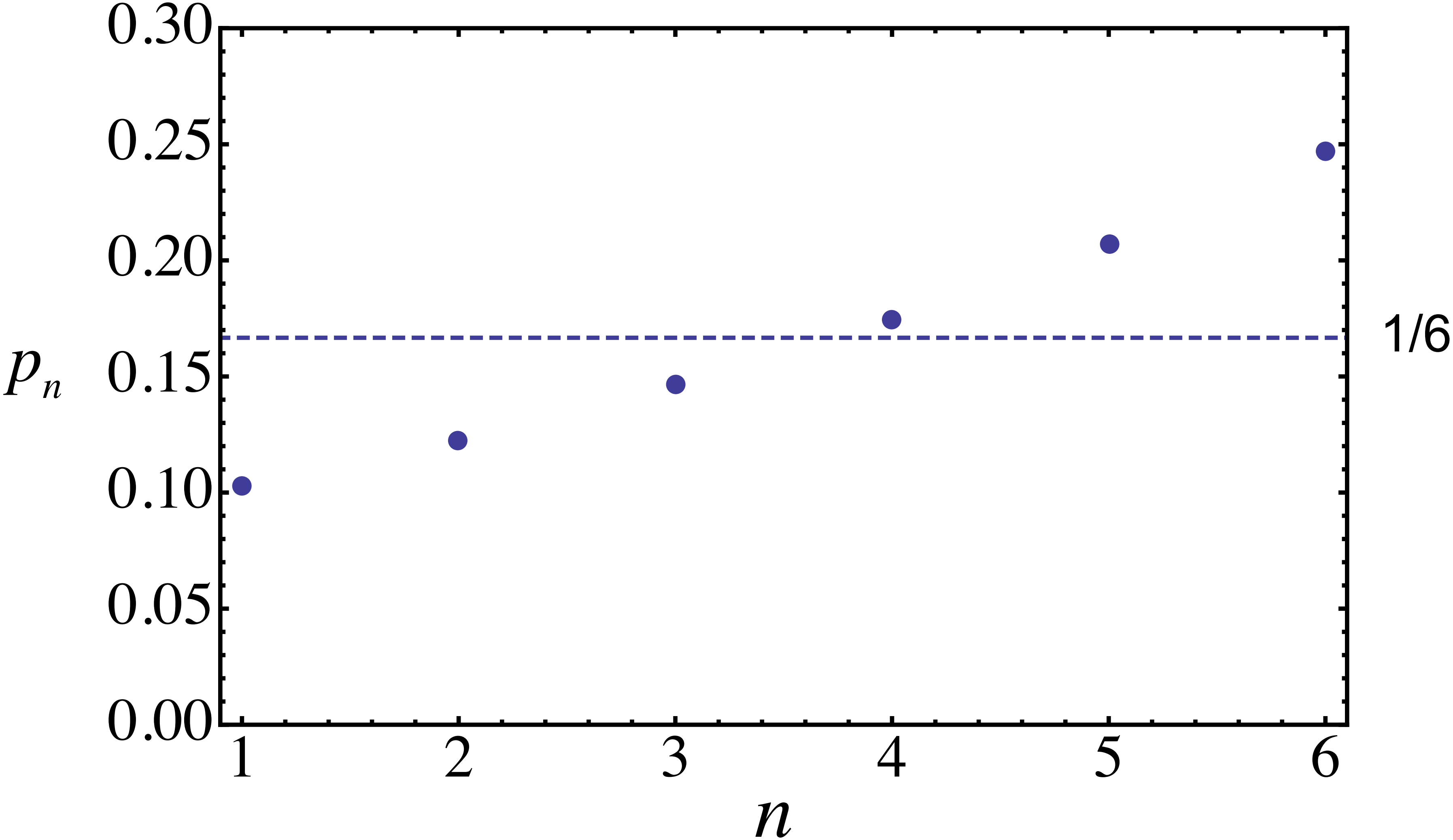
# numerical solution

| media | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | $p_6$ |
|---|---|---|---|---|---|---|
| **3.0** | 0.246782 | 0.20724 | 0.174034 | 0.146148 | 0.122731 | 0.103065 |
| **3.1** | 0.22929 | 0.199582 | 0.173723 | 0.151214 | 0.131622 | 0.114568 |
| **3.2** | 0.212566 | 0.191659 | 0.172808 | 0.155811 | 0.140487 | 0.126669 |
| **3.3** | 0.196574 | 0.183509 | 0.171313 | 0.159928 | 0.149299 | 0.139377 |
| **3.4** | 0.181282 | 0.175168 | 0.16926 | 0.163551 | 0.158035 | 0.152704 |
| **3.5** | 0.166667 | 0.166667 | 0.166667 | 0.166667 | 0.166666 | 0.166666 |
| **3.6** | 0.152704 | 0.158035 | 0.163551 | 0.16926 | 0.175168 | 0.181282 |
| **3.7** | 0.139377 | 0.149299 | 0.159928 | 0.171313 | 0.183509 | 0.196574 |
| **3.8** | 0.126669 | 0.140487 | 0.155811 | 0.172808 | 0.191659 | 0.212566 |
| **3.9** | 0.114568 | 0.131622 | 0.151214 | 0.173723 | 0.199582 | 0.22929 |
| **4.0** | 0.103065 | 0.122731 | 0.146148 | 0.174034 | 0.20724 | 0.246782 |

with a biased die we obtain skewed distributions.

These are examples of UNINFORMATIVE PRIORS

# Example: mean = 4



## Entropy with continuous probability distributions
(relative entropy, Kullback-Leibler divergence)

$$S \to -\int_a^b \left[p(x)dx\right] \ln\left[p(x)dx\right]$$

this diverges!

$$S_{p|m} = -\sum_k p_k \ln \frac{p_k}{m_k}$$

relative entropy

$$S_{p|m} = -\int_a^b p(x) \ln \frac{p(x)}{m(x)} dx$$

this does not diverge!

# Entropy maximization with additional conditions (partial knowledge of moments of the prior distribution)

$$\left\langle x^k \right\rangle = \int_a^b x^k p(x)\,dx$$

function (functional) that must be maximized

$$Q[p] = -\int_a^b p(x) \ln \frac{p(x)}{m(x)} dx + \sum_k \lambda_k \left\{ \int_a^b x^k p(x)\,dx - M_k \right\}$$

variation

$$\delta Q = -\int_a^b \delta p \left\{ \ln \frac{p(x)}{m(x)} + 1 - \sum_k \lambda_k x^k \right\} dx = 0$$

$$\ln \frac{p(x)}{m(x)} + 1 - \sum_k \lambda_k x^k = 0$$

$$p(x) = m(x) \exp\left( \sum_k \lambda_k x^k - 1 \right)$$

$$p(x) = m(x)\exp\left(\sum_n \lambda_n x^n - 1\right)$$

*p(x)* is determined by the choice of *m(x)* and by the constraints

The constraints can be the moments themselves:

$$M_k = \int_a^b x^k m(x)\exp\left(\sum_n \lambda_n x^n - 1\right)dx$$

1. no moment is known, normalization is the only constraint, and *p(x)* is defined in the interval (*a,b*)

$$M_0 = \int_a^b m(x)\exp(\lambda_0 - 1)\,dx = 1$$

we take a reference distribution which is uniform on (*a,b*), i.e.,

$$m(x) = \frac{1}{b-a}$$

$$M_0 = \frac{1}{b-a}\int_a^b \exp(\lambda_0 - 1)\,dx = \exp(\lambda_0 - 1) = 1$$

$$\Rightarrow \qquad \lambda_0 = 1; \quad p(x) = m(x)\exp\left(\sum_{n=0}^{0}\lambda_n x^n - 1\right) = \frac{1}{b-a}$$

2. only the first moment is known, i.e, the mean, and *p(x)* is defined on (*a,b*)

$$M_0 = \frac{1}{b-a}\int_a^b \exp\left(\lambda_0 + \lambda_1 x - 1\right) dx = 1$$

$$M_1 = \frac{1}{b-a}\int_a^b x \exp\left(\lambda_0 + \lambda_1 x - 1\right) dx$$

$$M_0 = 1 = \frac{\exp\left(\lambda_0 - 1\right)}{b-a}\int_a^b \exp\left(\lambda_1 x\right) dx = \frac{\exp\left(\lambda_0 - 1\right)}{b-a} \cdot \frac{\exp\left(\lambda_1 b\right) - \exp\left(\lambda_1 a\right)}{\lambda_1}$$

$$M_1 = \frac{\exp\left(\lambda_0 - 1\right)}{b-a}\int_a^b x \exp\left(\lambda_1 x\right) dx = \frac{\exp\left(\lambda_0 - 1\right)}{b-a}\left[\frac{1}{\lambda_1}\left(b \exp\left(\lambda_1 b\right) - a \exp\left(\lambda_1 a\right)\right) - \frac{1}{\lambda_1^2}\left(\exp\left(\lambda_1 b\right) - \exp\left(\lambda_1 a\right)\right)\right]$$

in general these equations can only be solved numerically...

special case:

$$a \to -\frac{L}{2}; \quad b \to \frac{L}{2}; \quad M_1 = 0$$

$$\frac{\exp(\lambda_0 - 1)}{L} \cdot \frac{\exp(\lambda_1 L/2) - \exp(-\lambda_1 L/2)}{\lambda_1} = 1$$

$$\frac{\exp(\lambda_0 - 1)}{L}\left[\frac{1}{\lambda_1}\left(\frac{L}{2}\exp(\lambda_1 L/2) + \frac{L}{2}\exp(-\lambda_1 L/2)\right) - \frac{1}{\lambda_1^2}\left(\exp(\lambda_1 L/2) - \exp(-\lambda_1 L/2)\right)\right] = 0$$

$$\frac{\exp(\lambda_0 - 1)}{L} \cdot \frac{\exp(\lambda_1 L/2) - \exp(-\lambda_1 L/2)}{\lambda_1} = 1$$

$$\frac{L}{2}\left(\exp(\lambda_1 L/2) + \exp(-\lambda_1 L/2)\right) - \frac{1}{\lambda_1}\left(\exp(\lambda_1 L/2) - \exp(-\lambda_1 L/2)\right) = 0$$

$$\exp(\lambda_0 - 1)\frac{\sinh(\lambda_1 L/2)}{\lambda_1 L/2} = 1$$

$$L\cosh(\lambda_1 L/2) - \frac{2}{\lambda_1}\sinh(\lambda_1 L/2) = 0$$

$$\Rightarrow \quad (\lambda_1 L/2) = \tanh(\lambda_1 L/2) \quad \Rightarrow \quad \lambda_1 = 0; \quad \lambda_0 = 1$$

$$p(x) = m(x)\exp\left(\sum_{k=0}^{1} \lambda_k x^k - 1\right) = \frac{1}{L}$$

nonzero mean

$$a \to -\frac{L}{2}; \quad b \to \frac{L}{2}; \quad M_1 = \varepsilon$$

$$\frac{\exp(\lambda_0 - 1)}{L} \cdot \frac{\exp(\lambda_1 L/2) - \exp(-\lambda_1 L/2)}{\lambda_1} = 1$$

$$\frac{\exp(\lambda_0 - 1)}{\lambda_1 L}\left[\frac{L}{2}\left(\exp(\lambda_1 L/2) + \exp(-\lambda_1 L/2)\right) - \frac{1}{\lambda_1}\left(\exp(\lambda_1 L/2) - \exp(-\lambda_1 L/2)\right)\right] = \varepsilon$$

$$\frac{\exp(\lambda_0 - 1)}{(\lambda_1 L/2)} \cdot \sinh(\lambda_1 L/2) = 1$$

$$\frac{L}{2}\frac{1}{\tanh(\lambda_1 L/2)} - \frac{1}{\lambda_1} = \varepsilon$$

$$\tanh(\lambda_1 L/2) = \left( \frac{1}{\lambda_1 L/2} + \frac{2\varepsilon}{L} \right)^{-1} \qquad \tanh(z) = \left( \frac{1}{z} + \frac{2\varepsilon}{L} \right)^{-1}$$

this is similar to the equations of ferromagnetism

$$z - \frac{z^3}{3} \approx \left( \frac{1}{z} + \frac{2\varepsilon}{L} \right)^{-1} \quad \Rightarrow \quad \left( z - \frac{z^3}{3} \right)\left( \frac{1}{z} + \frac{2\varepsilon}{L} \right) \approx 1 + \frac{2\varepsilon}{L} z - \frac{z^2}{3} = 1$$

$$\Rightarrow \quad \frac{2\varepsilon}{L} - \frac{z}{3} \approx 0 \quad \Rightarrow \quad z \approx \frac{6\varepsilon}{L}$$

$$\frac{\lambda_1 L}{2} \approx \frac{6\varepsilon}{L} \quad \Rightarrow \quad p(x) \approx \frac{1}{L} \exp(\lambda_1 x) \approx \frac{1}{L}\left( 1 - \frac{12\varepsilon}{L} x \right)$$

another special case $a = 0; \quad b \to \infty$

$$M_0 = \frac{1}{b-a}\int_a^b \exp\left(\lambda_0 + \lambda_1 x - 1\right)dx = 1$$

$$M_1 = \frac{1}{b-a}\int_a^b x\exp\left(\lambda_0 + \lambda_1 x - 1\right)dx$$

$$M_0 = 1 = m_0 \exp\left(\lambda_0 - 1\right)\cdot\frac{1}{\left(-\lambda_1\right)}$$

$$M_1 = m_0 \exp\left(\lambda_0 - 1\right)\left[\frac{1}{\lambda_1^2}\right] = \left(-\lambda_1\right)\left[\frac{1}{\lambda_1^2}\right] = -\frac{1}{\lambda_1} = \langle x \rangle$$

then

$$m_0 \exp(\lambda_0 - 1) = -\lambda_1 = \frac{1}{\langle x \rangle}$$

and we obtain the exponential distribution

$$p(x) = m(x) \exp\left( \sum_n \lambda_n x^n - 1 \right)$$

$$= m_0 \exp(\lambda_0 - 1) \exp(\lambda_1 x) = \frac{1}{\langle x \rangle} \exp\left( -\frac{x}{\langle x \rangle} \right)$$

3. both mean and variance are known, and the interval is the whole real axis

$$M_0 = m_0 \int_a^b \exp\left(\lambda_0 + \lambda_1 x + \lambda_2 x^2 - 1\right) dx = 1$$

$$M_1 = m_0 \int_a^b x \exp\left(\lambda_0 + \lambda_1 x + \lambda_2 x^2 - 1\right) dx$$

$$M_2 = m_0 \int_a^b x^2 \exp\left(\lambda_0 + \lambda_1 x + \lambda_2 x^2 - 1\right) dx$$

$$\exp\left(\lambda_0 + \lambda_1 x + \lambda_2 x^2 - 1\right) = \exp\left[\lambda_2\left(x^2 + 2\frac{\lambda_1}{\lambda_2}x + \frac{\lambda_1^2}{\lambda_2^2}\right) + \left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right)\right]$$

$$= \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right)\exp\left[\lambda_2\left(x + \frac{\lambda_1}{\lambda_2}\right)^2\right]$$

$$M_0 = m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right)\int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2(-1/2\lambda_2)}\left(x + \frac{\lambda_1}{\lambda_2}\right)^2\right] dx = m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right)\sqrt{-\frac{\pi}{\lambda_2}} = 1$$

$$M_1 = m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right)\int_{-\infty}^{+\infty} x \exp\left[-\frac{1}{2(-1/2\lambda_2)}\left(x + \frac{\lambda_1}{\lambda_2}\right)^2\right] dx = m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right)\sqrt{-\frac{\pi}{\lambda_2}}\left(-\frac{\lambda_1}{\lambda_2}\right) = -\mu$$

$$M_2 = m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right)\int_{-\infty}^{+\infty} x^2 \exp\left[-\frac{1}{2(-1/2\lambda_2)}\left(x + \frac{\lambda_1}{\lambda_2}\right)^2\right] dx = m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right)\sqrt{-\frac{\pi}{\lambda_2}}\left(-\frac{1}{2\lambda_2} + \frac{\lambda_1^2}{\lambda_2^2}\right) = \sigma^2 + \mu^2$$

$$M_0 = m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right)\sqrt{-\frac{\pi}{\lambda_2}} = 1$$

$$M_1 = \frac{\lambda_1}{\lambda_2} = \mu$$

$$M_2 = \left(-\frac{1}{2\lambda_2} + \frac{\lambda_1^2}{\lambda_2^2}\right) = \sigma^2 + \mu^2$$

$$\Rightarrow \quad \lambda_1 = -\frac{\mu}{2\sigma^2}; \quad \lambda_2 = -\frac{1}{2\sigma^2}; \quad m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}}$$

$$p(x) = m_0 \exp\left(\lambda_0 + \lambda_1 x + \lambda_2 x^2 - 1\right)$$

$$= m_0 \exp\left(\lambda_0 - 1 - \frac{\lambda_1^2}{\lambda_2}\right) \exp\left[-\frac{1}{2\left(-1/2\lambda_2\right)}\left(x + \frac{\lambda_1}{\lambda_2}\right)^2\right]$$

$$= \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left[\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

... in this case where mean and variance are known, the entropic prior is Gaussian

*An alternative form of entropy that incorporates the normalization constraint*

$$Q[p;m] = -\int_X dx\ p(x)\ln\frac{p(x)}{m(x)} + \lambda\left(\int_X dx p(x) - \int_X dx m(x)\right)$$

$$= \int_X dx\left(-p(x)\ln\frac{p(x)}{m(x)} + \lambda p(x) - \lambda m(x)\right)$$

$$\delta Q = \int_X \delta p\, dx\left(-\ln\frac{p(x)}{m(x)} - 1 + \lambda\right) = 0$$

$$p(x) = m(x)\exp(\lambda - 1)$$

$$\int_X dx\ p(x) = \int_X dx\ m(x)\exp(\lambda-1) = \exp(\lambda-1)\int_X dx\ m(x) = \exp(\lambda-1) = 1$$

$$\Rightarrow \quad \lambda = 1$$

$$Q[p;m] = \int_X dx\left(-p(x)\ln\frac{p(x)}{m(x)} + p(x) - m(x)\right)$$

Until now we have emphasized the role of the momenta of the distribution, however other information can be incorporated in the same way in the entropic prior.

*A "crystallographic" example (Jaynes, 1968)*

Consider a simple version of a crystallographic problem, where a 1-D crystal has atoms at the positions

$$x_j = jL \quad \left(L = 1,\ldots,n\right)$$

and such that these positions may be occupied by impurities.

From X-ray experiments it has been determined that impurity atoms prefer sites where

$$\cos\left(kx_j\right) > 0$$

so that

$$\left\langle \cos\left(kx_j\right)\right\rangle = 0.3$$

which means that we have the constraint

$$\left\langle \cos\left(kx_j\right)\right\rangle = \sum_{j=1}^{n} p_j \cos\left(kx_j\right) = 0.3$$

where $p_j$ is the probability that an impurity atom is at site $j$.

Then the constrained entropy that must be maximized is

$$Q = -\sum_{j=1}^{n} p_j \ln p_j + \lambda_0 \left( \sum_{j=1}^{n} p_j - 1 \right) + \lambda_1 \left( \sum_{j=1}^{n} p_j \cos\left(kx_j\right) - 0.3 \right)$$

from which we find the maximization condition

$$\frac{\partial Q}{\partial p_j} = -\left(\ln p_j + 1\right) + \lambda_0 + \lambda_1 \cos\left(kx_j\right) = 0$$

i.e.,

$$p_j = \exp\left[1 - \lambda_0 - \lambda_1 \cos\left(kx_j\right)\right]$$

The rest of the solution proceeds either by approximation or by numerical calculation.

*References:*


• G. D’Agostini, Rep. Prog. Phys. **66** (2003) 1383

• V. Dose: “Bayes in five days”, lecture notes, Max-Planck Research School on bounded plasmas, Greifswald, may 14-18 2002

• V. Dose, Rep. Prog. Phys. **66** (2003) 1421

• E. T. Jaynes, “Monkeys, Kangaroos and N”, in Maximum-Entropy and Bayesian Methods in Applied Statistics, edited by J. H. Justice, Cambridge Univ. Press, Cambridge, UK, 1986, updated (1996) version at http://bayes.wustl.edu

• E. T. Jaynes, “Prior probabilities”, IEEE Transactions On Systems Science and Cybernetics, vol. sec-4, (1968) 227


*Information theory*


• N. Abramson: “Information Theory and Coding”, McGraw-Hill 1963

• R. M. Gray: “Entropy and Information Theory”, Springer-Verlag 1990