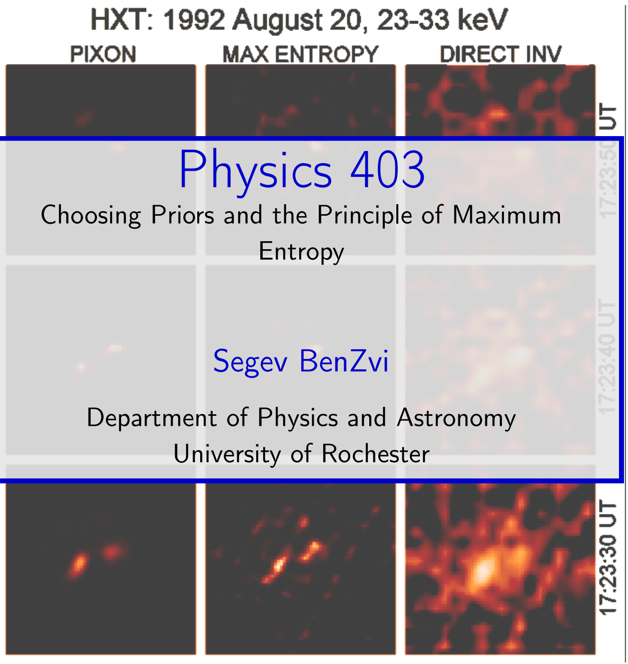


Physics 403

Choosing Priors and the Principle of Maximum Entropy

Segev BenZvi

Department of Physics and Astronomy
University of Rochester

Table of Contents

Last Time: The Odds Ratio

To select between two models, it is useful to calculate the ratio of the posterior probabilities of the models. This is called the odds ratio:

$$O_{ij} = \frac{p(D|M_i, I)}{p(D|M_j, I)} \frac{p(M_i|I)}{p(M_j|I)}$$
$$= B_{ij} \frac{p(M_i|I)}{p(M_j|I)}$$

The first term is called the Bayes Factor [1, 2] and the second is called the prior odds ratio. Interpration:

- **Prior odds**: the amount by which you favor M_i over M_j *before taking data*. There is no analog in frequentist statistics.
- **Bayes Factor**: the amount that the data D causes you favor M_i over M_j. Frequentist analog: *likelihood ratio* (but frequentists can't marginalize nuisance parameters)

Last Time: Occam Factors

- We can express any likelihood of data D given a model M as the maximum value of its likelihood times an Occam factor:

$$p(D|M, I) = \mathcal{L}_{\text{max}}\Omega_\theta$$

- The Occam factor corrects the likelihood for the statistical trials incurred by scanning the parameter space for $\hat{\theta}$.
- Occam's Razor: when selecting from among competing models, generally prefer the simpler model
- Statistical Trials: it becomes harder to reject the "null hypothesis" when the number of hypotheses in a test becomes large.

Example

You have a histogram and look for a spike in any one bin. The look-elsewhere effect: any bin could be a background fluctuation.

Last Time: Systematic Uncertainties

There are two types of experimental uncertainties:

1. **Random**: uncertainties which can be reduced by acquiring and averaging more data (details on this next class)
2. **Systematic**: uncertainties which are fixed and tend to affect all measurements equally

Example

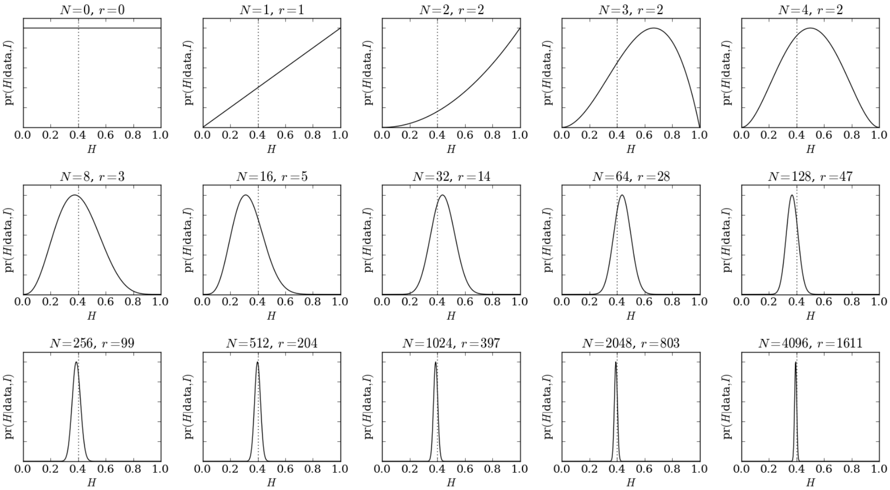
Calibrations of meters and rulers are a classic example of systematic uncertainties.

- Wooden meter sticks may shrink by several mm over time
- Energy scales in detectors may be uncertain due to other experimental or theoretical uncertainties
- Astronomical "rulers" have lots of systematic uncertainties, e.g., Hubble's constant H_0

Effect of Priors

Uniform "Ignorance" Prior

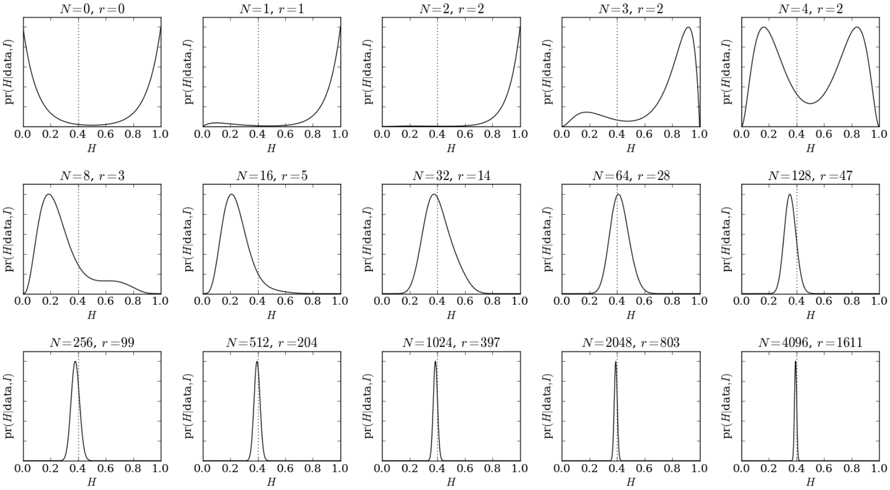
Coin flip example from [3]. We start with no preferred value for h:



Effect of Priors

Unfair Coin Prior

We assume the coin is very unfair, but don't know the bias.



Effect of Priors

Zeros

- Ultimately the choice of priors will not really matter once you've taken enough data, unless your prior is really pathological
- Pathology: if your prior is zero somewhere in the range of interest, no amount of data will budge the posterior PDF off that zero
- This is doing the "right" thing: your zero prior is explicitly a statement that no amount of data will ever move you to accept some model or part of the parameter space
- OK, the system works... but usually you don't intend this behavior.
- Hang on, here comes a counterexample: you limit a quantity like m^2 to a physical region, so your prior is 0 for $m^2 < 0$.

Caution: Parameterization Matters

From Oser: two theorists predict the mass of a new particle:

1. **A**: There should be a new particle whose mass is between 0 and 1 in rationalized uints. I have no other knowledge about the mass, so I'll assume it has equal chance of being between 0 and 1. I.e., $p(m|I) = 1$.
2. **B**: There is a particle described by a free parameter $y = m^2$. The true value of y must lie between 0 and 1, but otherwise I have no knowledge about it, so I choose $p(y|I) = 1$.

Both statements express ignorance about the same theory, but with different parameterizations.

$$p(y|I) = p(m|I)\left|\frac{dm}{dy}\right| \sim \frac{1}{\sqrt{y}}$$

Uh oh: transformation of variables makes a uniform prior non-uniform.

Table of Contents

Principle of Indifference

As a general rule, we want priors that do not inadvertently push us toward a result. We want non-informative priors. **Principle of Indifference**: given $n > 1$ mutually exclusive and exhaustive possibilities, each should be assigned a probability equal to $1/n$.

Example

Drawing from a deck of cards, we apply the principle of indifference and assume the probability of selecting a given card is $1/52$.

Example

Rolling dice with n faces, we assume the die lands on one face (exclusive possibility) with probability $1/6$.

Example

Statistical mechanics: any two microstates of a system with the same energy are equally probable at equilibrium.

Principle of Indifference

Continuous Location Parameter

- Consider an event that we locate with respect to some origin (a "location parameter"
- Example: we are interested in $p(X|I)$, where $X =$"the tallest tree in the woods is between x and $x + dx$."
- In the problem, x is measured with respect to some origin. What if we change the origin so that $x \to x' = x + c$?
- In the limit of complete ignorance, our choice of prior must be completely indifferent to shifts in location. This implies

$$p(X|I)\ dX = p(X'|I)\ dX' = p(X'|I)\ d(X + c) = p(X'|I)dX$$

If we represent the PDF by $f(x)$, then clearly

$$f(x) = f(x') = f(x + c) \implies f(x) = \text{constant}$$

Uniform Prior

Continuous Location Parameter

- Since $f(x)$=constant, we must also have $p(X|I) = \text{constant}$.
- If we have upper and lower bounds on x (we know the dimensions of the woods), then

$$p(X|I) = \text{constant} = \frac{1}{x_{\text{max}} - x_{\text{min}}},$$

  the uniform prior we have already used a few times.
- If the bounds x_{min} and x_{max} are not known, then technically $p(X|I)$ is not normalized. It is called an improper prior.
- Note: improper priors can be used in parameter estimation problems, as long as the posterior distribution is normalized.
- Note: improper priors cannot be used in model selection problems, because the Occam factors depend on knowing the prior range for each model parameter.

Principle of Indifference

Continuous Scale Parameter

- Consider a problem where we are interested in the mean lifetime of a particle. Lifetime is a scale parameter because it can only have positive values.
- We are interested in $p(\mathcal{T}|I)$, where $\mathcal{T}$="the "mean lifetime is between τ and $\tau + d\tau$."
- In the limit of complete ignorance, our prior must be indifferent to changes in scale β, e.g., if we change our time units $\tau \to \tau' = \beta\tau$:

$$p(\mathcal{T}|I)\ d\mathcal{T} = p(\mathcal{T}'|I)\ d\mathcal{T}' = p(\mathcal{T}'|I)\ d(\beta\mathcal{T}) = \beta p(\mathcal{T}'|I)\ d\mathcal{T}$$

If we represent the PDF by $g(\tau)$, then

$$g(\tau) = \beta g(\tau') = \beta g(\beta\tau) \implies g(\tau) = \text{constant}/\tau$$

Jeffreys Prior

Continuous Scale Parameter

- Since $g(\tau)$=constant, we must also have

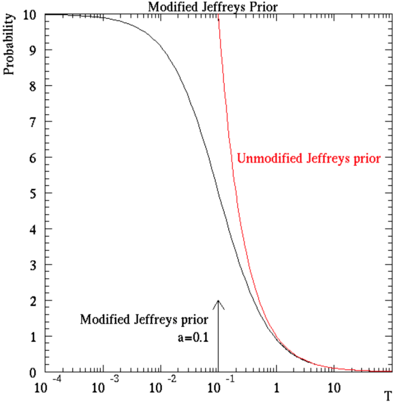
$$p(\mathcal{T}|I) = \frac{\text{constant}}{\tau}$$

- This form of the prior is called the Jeffreys prior [1].
- If we have upper and lower bounds on τ then

$$p(\mathcal{T}|I) = \frac{1}{\tau \ln\left(\tau_{\text{max}}/\tau_{\text{min}}\right)}$$

- The Jeffreys prior is very convenient for problems in which we are ignorant about scale. It provides logarithmic uniformity via equal probability per decade.
- Note: using a uniform prior on a scale parameter will cause you to dramatically weight your PDF toward the highest decade.

Modified Jeffreys Prior



- The Jeffreys prior is not normalizable if a scale parameter like τ can be zero.
- Alternative: modified Jeffreys prior, which becomes uniform for $\tau < a$:

$$p(\mathcal{T}|I) = \frac{1}{(\tau + a)\ln\left((a + \tau_{\max})/a\right)}$$

Table of Contents

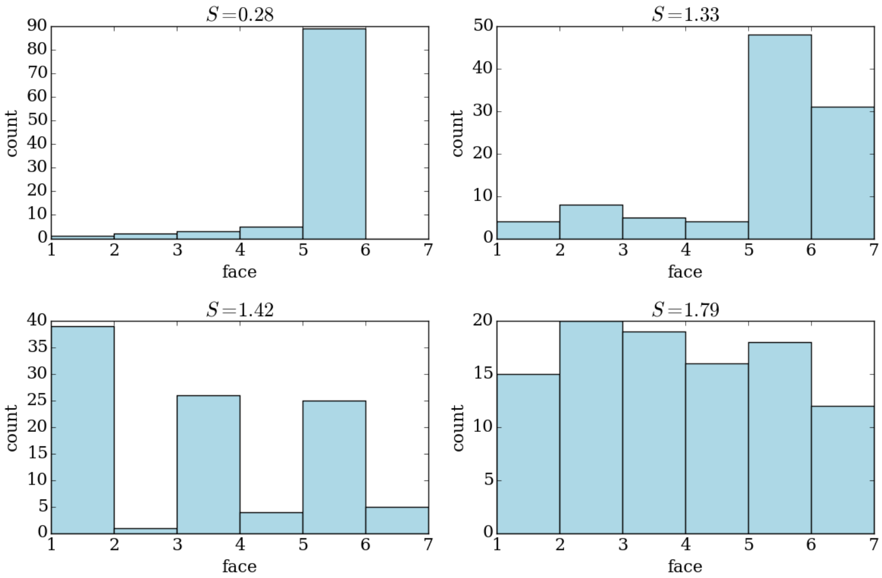
Principle of Maximum Entropy

- The Principle of Indifference, first developed by Bernoulli and Laplace, has a more quantitative form in the Principle of Maximum Entropy
- The probability distribution which bests represents the current state of knowledge is the one with the greatest entropy
- For a discrete probability distribution with values p_i, the uncertainty of the distribution is given by [4]

$$S(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \ln(p_i)$$

- S measures the information content of the distribution
- If we want to assign a prior that reflects our ignorance about a parameter, then we should assign a prior probability distribution that maximizes S

Intuition: Throwing Dice



Intuition: Weighted Die

- Suppose we have a weighted die with unknown outcomes p_i, but we are told that

$$\text{mean number of dots} = \sum_{i=1}^{6} i\ p_i = 4.$$

(Note: for a fair die, the mean is 3.5.)

- The probability of a given set of outcomes $\boldsymbol{n} = (n_1, \ldots, n_6)$ is given by the multinomial distribution:

$$p(n_1, \ldots, n_6 | N, p_1, \ldots, p_6) = \frac{N!}{n_1! \ldots n_6!} p_1^{n_1} \times \ldots \times p_6^{n_6}$$

- The quantity $W = N!/(n_1! \ldots n_6!)$, or multiplicity, represents the number of states available to any given outcome $\boldsymbol{n}$.
- $\boldsymbol{n}$ with the largest multiplicity W is the most probable.

Maximizing the Multiplicity

Let's maximize $\ln W$ and use Stirling's approximation ($\ln N! \approx N \ln N - N$):

$$\begin{aligned}
\ln W &= N\ln N - N - \sum_{i=1}^{6} Np_i \ln Np_i + \sum_{i=1}^{6} Np_i, \quad \text{where } n_i = Np_i \\
&= N\ln N - N - \sum_{i=1}^{6} Np_i \ln\left(Np_i\right) + \sum_{i=1}^{6} Np_i \\
&= N\ln N - N - N\left(\sum_{i=1}^{6} p_i \ln p_i + \ln N\right) + N \\
&= -N\sum_{i=1}^{6} p_i \ln p_i \\
&= NS \\
\therefore W &= \exp\left(NS\right)
\end{aligned}$$

N is the number of throws, and S is the entropy. Maximizing entropy maximizes W.

Shannon-Jaynes Entropy

Up to now we have claimed total ignorance of the p_i, but what if there is some prior estimate m_i on the p_i? Then

$$p(n_1, \ldots, n_M | N, p_1, \ldots, p_M) = \frac{N!}{n_1! \ldots n_M!} m_1^{n_1} \times \ldots \times m_M^{n_M}$$

$$\begin{aligned}
\ln p(n_1, \ldots, n_M | N, p_1, \ldots, p_M) &= \sum_{i=1}^{M} n_i \ln m_i + \ln N! - \sum_{i=1}^{M} \ln n_i! \\
&= \sum_{i=1}^{M} n_i \ln m_i - N \sum_{i=1}^{M} p_i \ln p_i \\
&= N \left(\sum_{i=1}^{M} p_i \ln m_i - \sum_{i=1}^{M} p_i \ln p_i \right) \\
&= -N \sum_{i=1}^{M} p_i \ln \left(p_i / m_i \right) = NS
\end{aligned}$$

Shannon-Jaynes Entropy

We are left with the generalized Shannon-Jaynes entropy

$$S = -\sum_{i=1}^{M} p_i \ln\left(p_i/m_i\right)$$

For the continuous case,

$$S = -\int p(x) \ln\left(\frac{p(x)}{m(x)}\right) dx$$

The quantity $m(x)$ is called the Lebesgue measure and ensures that S is invariant under the change of variables $x \to x' = f(x)$ since $m(x)$ and $p(x)$ transform in the same way.

OK, now we're ready to explore the maximum entropy principle.

MaxEnt and the Principle of Indifference

- We want to find a set of probabilities $p_1, \ldots, p_n$ that maximizes

$$S(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \ln p_i .$$

- If all of the p_i are independent, this implies

$$dS = \frac{\partial S}{\partial p_1} dp_1 + \ldots + \frac{\partial S}{\partial p_n} dp_n = 0$$

- But if the p_i are independent, then all of the coefficients are individually equal to 0.
- Conclusion: all of the p_i are equal; i.e., we need a uniform prior.
- Hence, the principle of maximum entropy is just a formal statement of the principle of ignorance.

MaxEnt and Constraints

Lagrange Undetermined Multipliers

- Suppose we impose a constraint on the p_i of the general form $C(p_1, \ldots, p_n) = 0$. Then

$$dC = \frac{\partial C}{\partial p_1} dp_1 + \ldots + \frac{\partial C}{\partial p_n} dp_n = 0$$

- We can combine dS and the constraint dC using a Lagrange multiplier:

$$dS - \lambda dC = 0$$

and therefore

$$dS - \lambda dC = \left(\frac{\partial S}{\partial p_1} - \lambda \frac{\partial C}{\partial p_1}\right) dp_1 + \ldots + \left(\frac{\partial S}{\partial p_n} - \lambda \frac{\partial C}{\partial p_n}\right) dp_n = 0$$

We set the first coefficient to zero, letting us solve for λ and giving M simultaneous equations for the p_i.

Normalization Constraint

- We can always start from the normalization constraint (sum rule):

$$C = \sum_{i=1}^{n} p_i = 1$$

- Therefore, from $dS - \lambda dC = 0$ we have

$$d\left[-\sum_{i=1}^{M} p_i \ln\left(p_i/m_i\right) - \lambda\left(\sum_{i=1}^{M} p_i - 1\right)\right] = 0$$

$$d\left[-\sum_{i=1}^{M} p_i \ln p_i + \sum_{i=1}^{M} p_i \ln m_i - \lambda\left(\sum_{i=1}^{M} p_i - 1\right)\right] = 0$$

$$\sum_{i=1}^{M}\left(-\ln p_i - p_i\frac{\partial \ln p_i}{\partial p_i} + \ln m_i - \lambda\frac{\partial p_i}{\partial p_i}\right) dp_i = 0$$

$$\sum_{i=1}^{M}\left(-\ln\left(p_i/m_i\right) - 1 - \lambda\right) dp_i = 0$$

Normalization Constraint

Derivation of Uniform Distribution

- Allowing the p_i to vary independently implies that all of the coefficients must vanish, so that

$$-\ln\left(p_i/m_i\right) - 1 - \lambda = 0 \implies p_i = m_i e^{-(1+\lambda)}$$

- Since $\sum p_i = 1$ and $\sum m_i = 1$,

$$\sum_{i=1}^{M} m_i e^{-(1+\lambda)} = 1 = e^{-(1+\lambda)} \sum_{i=1}^{M} m_i$$

Thus, $\lambda = -1$ and

$$p_i = m_i$$

- If our prior information tells us that $m_i =$ constant, then p_i describe a uniform distribution.

Gaussian: Known Mean and Variance

- Suppose you have a continous variable x and you constrain the mean to be μ and the variance to be σ^2:

$$\int_{x_L}^{x_H} p(x)\ dx = 1$$

$$\int_{x_L}^{x_H} x\ p(x)\ dx = \mu$$

$$\int_{x_L}^{x_H} (x-\mu)^2\ p(x)\ dx = \sigma^2$$

- In the limit that the variance is small compared to the range of the parameter, i.e.,

$$\frac{x_H-\mu}{\sigma} \gg 1 \quad \text{and} \quad \frac{\mu-x_L}{\sigma} \gg 1$$

  then it turns out the maximum entropy distribution with this variance is Gaussian:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

Utility of the Gaussian

- Suppose your data are scattered around your model with an unknown error distribution.
- It turns out that the most conservative thing you can assume (in a maximum entropy sense) is the Gaussian distribution.
- By "conservative" we mean that the Gaussian will give a greater uncertainty than what you would get from a more appropriate distribution based on more information.
- Wait, isn't that bad?
- No: for model fitting, a Gaussian model of the uncertainties is a safe choice. Other distributions may give you artificially tight constraints unless you have appropriate prior information.

Summary

- We like to identify uniform priors for inference in physics problems
- We have to be careful about transforming variables because uniform priors may not stay uniform under changes of variables
  - **Uniform Prior**: appropriate for a location parameter
  - **Jeffreys Prior**: appropriate for a scale parameter
- We have intuitively been picking uninformative priors using the Principle of Indifference
- This principle can be made quantitative using the Principle of Maximum Entropy, which tells us that the least informative prior is the one which maximizes

$$S = -\sum_{i=1}^{N} p_i \ln\left(p_i/m_i\right)$$

- By maximizing S under different constraints we can derive the PDFs used earlier in the course.

References I

[1] Harold Jeffreys. *The Theory of Probability*. 3rd ed. Oxford, 1961.

[2] Robert E. Kass and Adrian E. Raftery. "Bayes Factors". In: *J. Am. Stat. Assoc.* 90.430 (1995), pp. 773–795. URL: `http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572`.

[3] D.S. Sivia and John Skilling. *Data Analysis: A Bayesian Tutorial*. New York: Oxford University Press, 1998.

[4] Claude E. Shannon. "A Mathematical Theory of Communication". In: *Bell Sys. Tech. J.* 27 (1948), pp. 379–423.