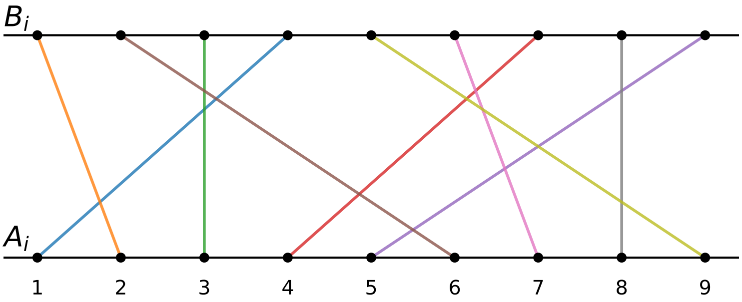
STSCI 4780
Assigning priors, 2

Tom Loredo, CCAPS & SDS, Cornell University

Recap: Direct probabilities and priors

- Both sampling distributions and priors are typically *direct probabilities*: We must provide contextual information enabling them to be assigned directly, or built from simpler probabilities we can assign directly
- Structural assumptions/ansatzes (e.g., independence, IID) and symmetries often play a key role in specifying direct probabilities
- For discrete alternatives, *permutation symmetry* justifies *principle of indifference* & uniform prior
- Over *continuous* spaces the uniform prior can't be universal
- Focused on role of symmetries:
  - Poisson distribution + scale symmetry $\rightarrow 1/r$ prior for a rate parameter
  - Location family + translation symmetry $\rightarrow$ uniform prior for location parameter
  - Scale family + scale symmetry $\rightarrow 1/\sigma$ prior for scale parameter
  - On non-compact spaces, such priors are typically *improper*

Uninformative PMF from permutation symmetry

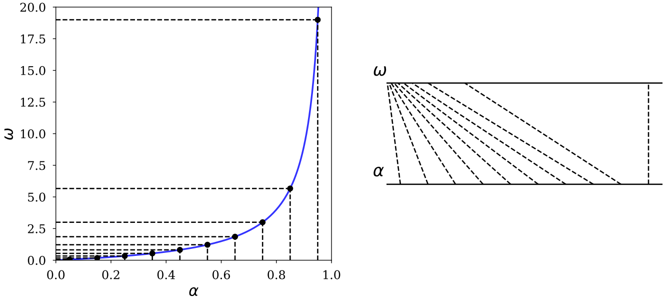


If prior information doesn't distinguish between discrete possibilities, a rule that is consistent across problem formulations must exhibit relabeling symmetry $\rightarrow$ assign equal probabilities

Uniform PDF can't be universal

Consider two parameterizations of a binomial sampling dist'n:

- Success probability, α: $\quad p(n|\alpha) \propto \alpha^n (1-\alpha)^{N-1}$
- Odds, $\omega \equiv \alpha/(1-\alpha)$: $p(n|\alpha) \propto \left(\frac{\omega}{1+\omega}\right)^n \left(\frac{1}{1+\omega}\right)^{N-1}$



Uniform over α is inconsistent with uniform over ω!

Is there a "natural" way to pick intervals we should consider equally probable a priori—natural local parameter scales?

Priors derived from the likelihood function

Few common problems beyond location/scale problems admit a transformation group argument $\rightarrow$ we need a more general approach to formal assignment of priors that express "ignorance" in some sense

There is no universal consensus on how to do this (yet? ever?)

A common underlying idea: The same $\mathcal{C}$ appears in the prior, $p(\theta|\mathcal{C})$, and the likelihood, $p(D|\theta, \mathcal{C})$—the prior "knows" about the likelihood function, although it doesn't know what data values will be plugged into it (e.g., the symmetry eqns $\leftarrow$ likelihood)

Jeffreys priors: Uses Fisher information to define a (parameter-dependent) scale defining a prior; parameterization invariant, but undesirable behavior in many dimensions

Reference priors: Uses information theory to define a prior that (asymptotically) has the least effect on the posterior; complicated algorithm; gives good frequentist behavior to Bayesian inferences

Jeffreys priors: Heuristic motivation

- Dimensionally, $\pi(\theta) \propto 1/(\theta \text{ scale})$ — e.g., uniform prior is $p(\theta) = 1/\Delta\theta$
- Use the likelihood function to determine a (relative) scale at each θ, say, $s(\theta)$, and then set $\pi(\theta) \propto 1/s(\theta)$
- Seek a scale definition that produces priors that are consistent WRT reparameterization (this was Harold Jeffreys' main desideratum)

Such a prior essentially specifies a way to slice-and-dice the θ axis so assigning equal probability to intervals reflects inherent scales, and is consistent WRT reparameterization

Jeffreys priors: Implementation

- If we have data D, a natural scale at θ, from the likelihood function, is the **inverse square root** of the *observed Fisher information* (recall Laplace approximation, where this gives $1/\sigma^2$ at $\hat{\theta}$):

$$I_D(\theta) \equiv -\frac{\mathrm{d}^2 \log \mathcal{L}_D(\theta)}{\mathrm{d}\theta^2}$$

- For a prior, we don't know D; for each θ, average over D predicted by the sampling distribution; this defines the *(expected) Fisher information*:

$$I(\theta) \equiv -\mathbb{E}_{D|\theta}\left[\frac{\mathrm{d}^2 \log \mathcal{L}_D(\theta)}{\mathrm{d}\theta^2}\right]$$

- Invariance: Can show for $\phi = \Phi(\theta)$, and $\theta = \Theta(\phi)$:

$$I(\phi) = I(\theta)\left(\frac{\mathrm{d}\Theta}{\mathrm{d}\phi}\right)^2$$

Jeffreys' prior

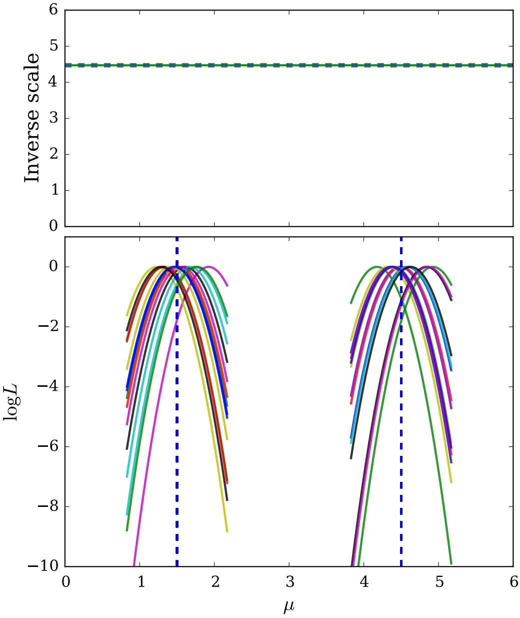
$$\boxed{\pi(\theta) \propto [I(\theta)]^{1/2}}$$

- Puts more weight in regions of parameter space where the data are expected to be more informative—roughly speaking, says the choice of experiment reflects expectations of what value θ may take
- Automatically consistent w.r.t. reparameterization ("invariant"):

$$\sqrt{I(\phi)} = \sqrt{I(\theta)} \left| \frac{\mathrm{d}\Theta}{\mathrm{d}\phi} \right| \quad \rightarrow \quad \pi(\phi) = \pi(\Theta(\phi)) \left| \frac{\mathrm{d}\Theta}{\mathrm{d}\phi} \right|$$

- Typically improper when parameter space is non-compact
- Improves *frequentist* performance of posterior intervals w.r.t. intervals based on flat priors
- Only considered sound for a single parameter (or considering a single parameter at a time in some multiparameter problems)

Jeffreys prior for normal mean



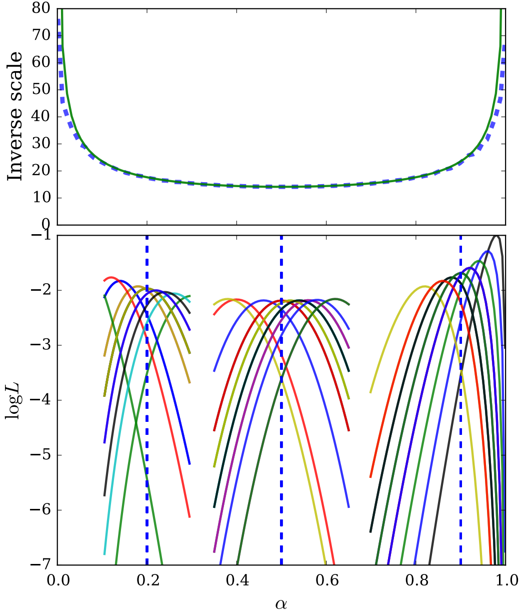
$N = 20$ samples from normals with $\sigma = 1$

Likelihood width is independent of $\mu \Rightarrow$

$$\pi(\mu) = \text{Const}$$

Another justification of the uniform prior
Prior is improper without prior limits on the range

Jeffreys prior for binomial probability



Binomial success counts *n* from $N = 50$ trials

$$\pi(\mu) = \frac{1}{\pi\alpha^{1/2}(1-\alpha)^{1/2}}$$
$$= \mathrm{Beta}(1/2, 1/2)$$

Limitations of the Jeffreys prior

- Only considered sound for a single parameter (or considering a single parameter at a time in some multiparameter problems) E.g., for $\mathrm{Norm}(\mu, \sigma)$, the Jeffreys prior is $\propto 1/\sigma^2$, *not* the product of separate Jeffreys μ, σ priors
- Only applicable to continuous spaces

$\rightarrow$ Seek more formal notions of "objective" or "uninformative" that reproduce good things about the Jeffreys prior

Reference priors largely accomplish this, using ideas from *information theory*

Supplementary material on reference priors...

Uncertainty, information, and entropy

Other rules for assigning "non-informative" priors rely on a more formal measure of the *information content* (or its complement, amount of *uncertainty*) in a probability distribution

Intuitively appealing metric-based measures, like standard deviation or interval size, are not general enough; e.g., they don't apply to categorical distributions

Desiderata for an *uncertainty functional* $\mathcal{S}_N[\vec{p}]$—a map from a PMF $\vec{p} = (p_1, p_2, \ldots, p_N)$ to a single scalar quantifying the amount of uncertainty it expresses (treat PDFs later):

- $\mathcal{S}_N[\vec{p}]$ should be continuous w.r.t. the p_is
- *Uncertainty grows with multiplicity*: When the p_i are all equal, $s(N) = \mathcal{S}_N[\vec{p}]$ should grow monotonically with N
- *Additivity over subgroups*

$\Rightarrow$ functional equations for $\mathcal{S}_N[\vec{p}]$

Information Gain as Entropy Change

Entropy and uncertainty

Shannon entropy = a scalar measure of the degree of uncertainty expressed by a probability distribution

$$\begin{aligned} \mathcal{S} &= \sum_i p_i \log \frac{1}{p_i} \qquad \text{“Average surprisal”} \\ &= -\sum_i p_i \log p_i \end{aligned}$$

Information gain

Information gain upon learning D = decrease in uncertainty:

$$\begin{aligned} \mathcal{I}(D) &= \mathcal{S}[\{p(H_i)\}] - \mathcal{S}[\{p(H_i|D)\}] \\ &= \sum_i p(H_i|D) \log p(H_i|D) - \sum_i p(H_i) \log p(H_i) \end{aligned}$$

A 'Bit' About Entropy

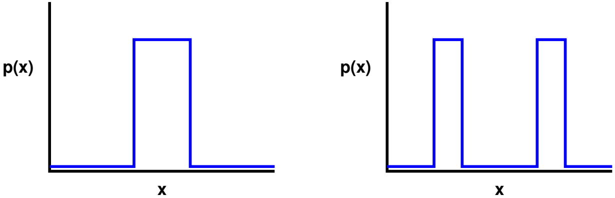
Entropy of a Gaussian

$$p(x) \propto e^{-(x-\mu)^2/2\sigma^2} \quad \rightarrow \quad \mathcal{S} \propto \log(\sigma)$$

$$p(\vec{x}) \propto \exp\left[-\tfrac{1}{2}\vec{x} \cdot \mathrm{V}^{-1} \cdot \vec{x}\right] \quad \rightarrow \quad \mathcal{S} \propto \log(\det \mathrm{V})$$

$\rightarrow$ Asymptotically like log Fisher matrix

A log-measure of "volume" or "spread," not range



These distributions have the same entropy/amount of information.

Expected information gain

When the data are yet to be considered, the *expected* information gain averages over D; straightforward use of the product rule/Bayes's theorem gives:

$$\begin{aligned}\mathbb{E}\mathcal{I} &= \int dD\, p(D)\, \mathcal{I}(D) \\ &= \int dD\, p(D)\, \sum_i p(H_i|D) \log\left[\frac{p(H_i|D)}{p(H_i)}\right]\end{aligned}$$

For a continuous hypothesis space labeled by parameter(s) θ,

$$\mathbb{E}\mathcal{I} = \int dD\, p(D)\, \int d\theta p(\theta|D) \log\left[\frac{p(\theta|D)}{p(\theta)}\right]$$

This is the expectation value of the *Kullback-Leibler divergence* between the prior and posterior:

$$\mathcal{D} \equiv \int d\theta\, p(\theta|D) \log\left[\frac{p(\theta|D)}{p(\theta)}\right]$$

Reference priors

Bernardo (later joined by Berger & Sun) advocates *reference priors*, priors chosen to maximize the KLD between prior and posterior, as an "objective" expression of the idea of a "non-informative" prior: reference priors let the data most strongly dominate the prior (on average)

- Rigorous definition invokes asymptotics and delicate handling of non-compact parameter spaces to make sure posteriors are proper
- For 1-D problems, the reference prior is the Jeffreys prior
- In higher dimensions, the reference prior is *not* the Jeffreys prior; it behaves better
- The construction in higher dimensions is complicated and depends on separating interesting vs. nuisance parameters (but see Berger, Bernardo & Sun 2015, "Overall objective priors")
- Reference priors are typically improper on non-compact spaces
- They give Bayesian inferences good frequentist properties
- A constructive numerical algorithm exists