Bayesian Inference in Astronomy & Astrophysics

A Short Course

Tom Loredo

Dept. of Astronomy, Cornell University

Five Lectures

- Overview of Bayesian Inference
- From Gaussians to Periodograms
- Learning How To Count: Poisson Processes
- Miscellany: Frequentist Behavior, Experimental Design
- Why Try Bayesian Methods?

Overview of Bayesian Inference

- What to do
- What's different about it
- How to do it: Tools for Bayesian calculation

What To Do: The Bayesian Recipe

Assess hypotheses by calculating their probabilities $p(H_i|\ldots)$ conditional on known and/or presumed information using the rules of probability theory.

But . . . what does $p(H_i|\ldots)$ *mean*?

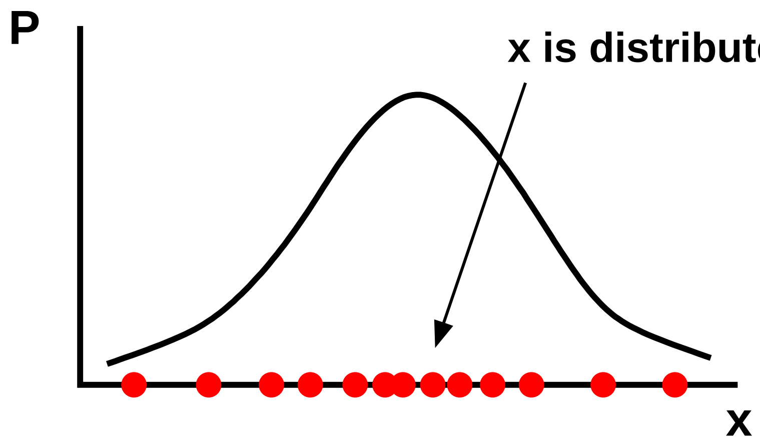
What is distributed in $p(x)$?

Frequentist: Probability describes "randomness"

Venn, Boole, Fisher, Neymann, Pearson...

x is a *random variable* if it takes different values throughout an infinite (imaginary?) ensemble of "identical" sytems/experiments.

$p(x)$ describes how x is distributed throughout the ensemble.



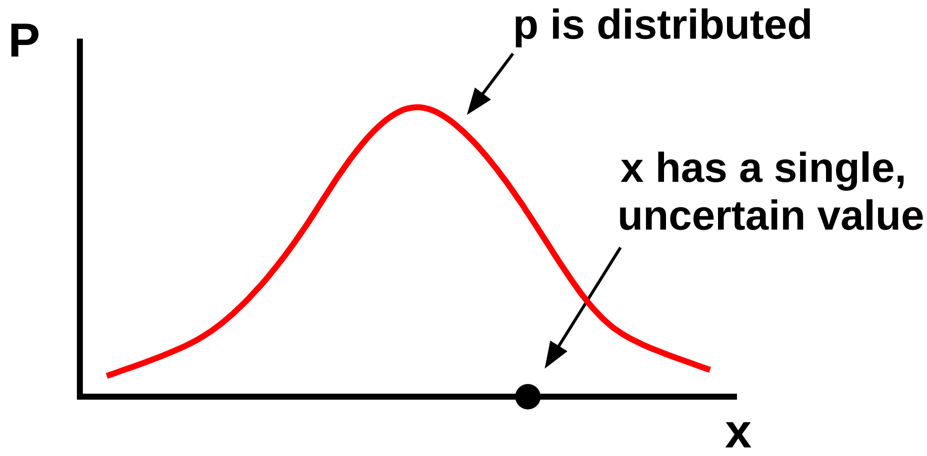
Probability $\equiv$ frequency (pdf $\equiv$ histogram).

Bayesian: Probability describes uncertainty

Bernoulli, Laplace, Bayes, Gauss. . .

$p(x)$ describes how probability (plausibility) is distributed among the possible choices for x in the case at hand.

Analog: a mass density, $\rho(x)$



Relationships between probability and frequency were demonstrated mathematically (large number theorems, Bayes's theorem).

Interpreting Abstract Probabilities

Symmetry/Invariance/Counting

- Resolve possibilities into equally plausible "microstates" using symmetries
- Count microstates in each possibility

Frequency from probability

Bernoulli's laws of large numbers: In repeated trials, given $P(\text{success})$, predict

$$\frac{N_{\text{success}}}{N_{\text{total}}} \to P \quad \text{as} \quad N \to \infty$$

Probability from frequency

Bayes's "An Essay Towards Solving a Problem in the Doctrine of Chances" $\rightarrow$ Bayes's theorem

Probability $\neq$ Frequency!

Bayesian Probability:
A Thermal Analogy

Intuitive notion	*Quantification*	*Calibration*
Hot, cold	Temperature, T	Cold as ice $= 273$K Boiling hot $= 373$K
uncertainty	Probability, P	Certainty $= 0$, 1 $p = 1/36$: plausible as “snake’s eyes” $p = 1/1024$: plausible as 10 heads

The Bayesian Recipe

Assess hypotheses by calculating their probabilities $p(H_i|\ldots)$ conditional on known and/or presumed information using the rules of probability theory.

Probability Theory Axioms ("grammar"):

'OR' (sum rule)
$$P(H_1 + H_2|I) = P(H_1|I) + P(H_2|I) - P(H_1, H_2|I)$$

'AND' (product rule)
$$\begin{aligned} P(H_1, D|I) &= P(H_1|I)\, P(D|H_1, I) \\ &= P(D|I)\, P(H_1|D, I) \end{aligned}$$

Direct Probabilities ("vocabulary"):

- Certainty: If A is certainly true given B, $P(A|B) = 1$
- Falsity: If A is certainly false given B, $P(A|B) = 0$
- Other rules exist for more complicated types of information; for example, invariance arguments, maximum (information) entropy, limit theorems (CLT; tying probabilities to frequencies), bold (or desperate!) presumption. . .

Three Important Theorems

Normalization:

For *exclusive, exhaustive* H_i

$$\sum_i P(H_i|\cdots) = 1$$

Bayes's Theorem:

$$P(H_i|D,I) = P(H_i|I)\,\frac{P(D|H_i,I)}{P(D|I)}$$

posterior $\propto$ prior $\times$ likelihood

Marginalization:

Note that for exclusive, exhaustive $\{B_i\}$,

$$\begin{aligned}\sum_i P(A, B_i|I) &= \sum_i P(B_i|A, I)P(A|I) = P(A|I) \\ &= \sum_i P(B_i|I)P(A|B_i, I)\end{aligned}$$

$\rightarrow$ We can use $\{B_i\}$ as a "basis" to get $P(A|I)$.

Example: Take $A = D$, $B_i = H_i$; then

$$\begin{aligned}P(D|I) &= \sum_i P(D, H_i|I) \\ &= \sum_i P(H_i|I)P(D|H_i, I)\end{aligned}$$

prior predictive for $D =$ Average likelihood for H_i

Inference With Parametric Models

Parameter Estimation

$I =$ Model M with parameters θ ($+$ any add'l info)

$H_i =$ statements about θ; e.g. "$\theta \in [2.5, 3.5]$," or "$\theta > 0$"

Probability for any such statement can be found using a *probability density function* (pdf) for θ:

$$\begin{aligned} P(\theta \in [\theta, \theta + d\theta]|\cdots) &= f(\theta)d\theta \\ &= p(\theta|\cdots)d\theta \end{aligned}$$

Posterior probability density:

$$p(\theta|D, M) = \frac{p(\theta|M)\ \mathcal{L}(\theta)}{\int d\theta\ p(\theta|M)\ \mathcal{L}(\theta)}$$

Summaries of posterior:

- “Best fit” values: mode, posterior mean
- Uncertainties: Credible regions (e.g., HPD regions)
- Marginal distributions:
  - Interesting parameters ψ, nuisance parameters ϕ
  - Marginal dist’n for ψ:

$$p(\psi|D, M) = \int d\phi\, p(\psi, \phi|D, M)$$

Generalizes “propagation of errors”

Model Uncertainty: Model Comparison

$I = (M_1 + M_2 + \ldots)$ — Specify a set of models.
$H_i = M_i$ — Hypothesis chooses a model.

Posterior probability for a model:

$$\begin{aligned} p(M_i|D,I) &= p(M_i|I)\frac{p(D|M_i,I)}{p(D|I)} \\ &\propto p(M_i)\mathcal{L}(M_i) \end{aligned}$$

But $\mathcal{L}(M_i) = p(D|M_i) = \int d\theta_i\, p(\theta_i|M_i)p(D|\theta_i,M_i)$.

Likelihood for model $=$ Average likelihood for its parameters

$$\mathcal{L}(M_i) = \langle \mathcal{L}(\theta_i) \rangle$$

Model Uncertainty: Model Averaging

Models have a common subset of interesting parameters, ψ.

Each has different set of nuisance parameters ϕ_i (or different prior info about them).

$H_i =$ statements about ψ.

Calculate posterior PDF for ψ:

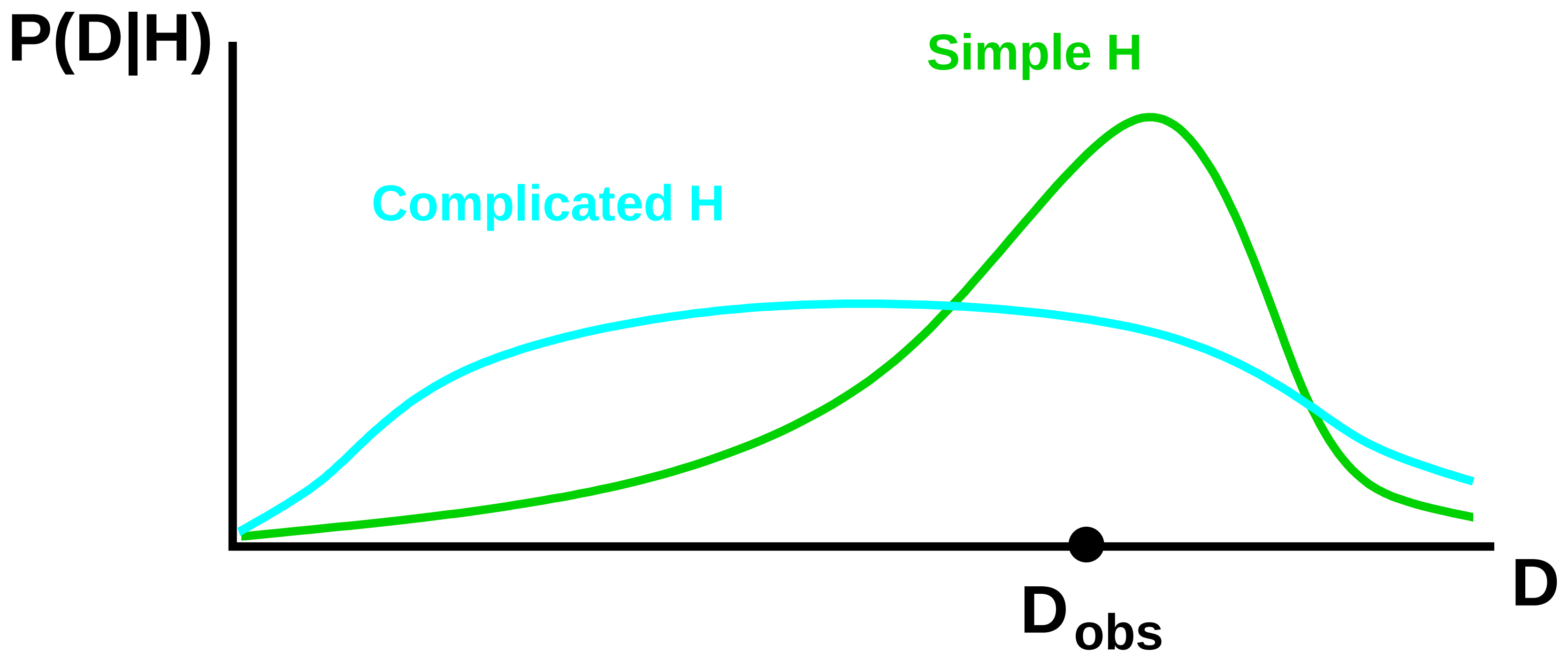
$$\begin{aligned} p(\psi|D,I) &= \sum_i p(M_i|D,I)\, p(\psi|D,M_i) \\ &\propto \sum_i \mathcal{L}(M_i) \int d\theta_i\, p(\psi,\phi_i|D,M_i) \end{aligned}$$

The model choice is itself a (discrete) nuisance parameter here.

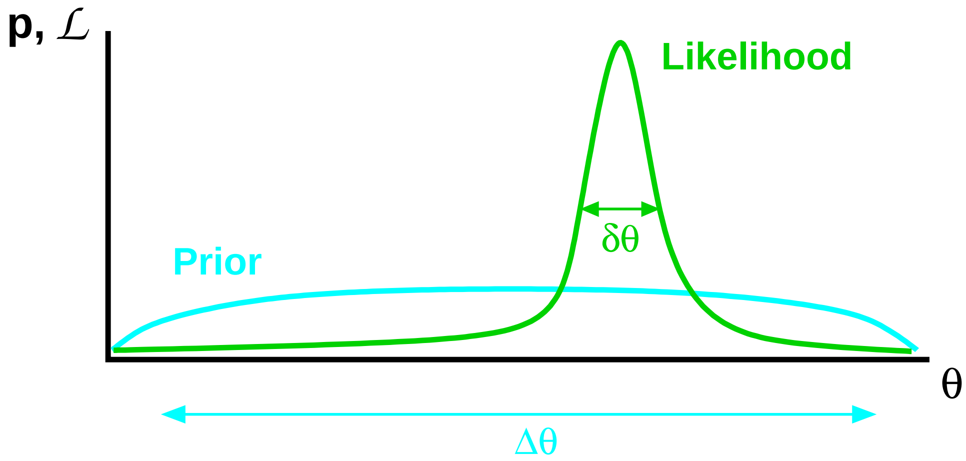
An Automatic Occam's Razor

Predictive probabilities can favor simpler models:

$$p(D|M_i) = \int d\theta_i \; p(\theta_i|M) \; \mathcal{L}(\theta_i)$$



The Occam Factor:



$$
\begin{aligned}
p(D|M_i) &= \int d\theta_i \; p(\theta_i|M) \; \mathcal{L}(\theta_i) \approx p(\hat{\theta}_i|M)\mathcal{L}(\hat{\theta}_i)\delta\theta_i \\
&\approx \mathcal{L}(\hat{\theta}_i)\frac{\delta\theta_i}{\Delta\theta_i} \\
&= \text{Maximum Likelihood} \times \text{Occam Factor}
\end{aligned}
$$

Models with more parameters often make the data more probable— *for the best fit*.

Occam factor penalizes models for “wasted” volume of parameter space.

What's the Difference?

Bayesian Inference (BI):

- Specify at least two competing hypotheses and priors
- Calculate their probabilities using probability theory
  - Parameter estimation:

$$p(\theta|D, M) = \frac{p(\theta|M)\mathcal{L}(\theta)}{\int d\theta \, p(\theta|M)\mathcal{L}(\theta)}$$

  - Model Comparison:

$$O \propto \frac{\int d\theta_1 \, p(\theta_1|M_1) \, \mathcal{L}(\theta_1)}{\int d\theta_2 \, p(\theta_2|M_2) \, \mathcal{L}(\theta_2)}$$

Frequentist Statistics (FS):

- Specify null hypothesis H_0 such that rejecting it implies an interesting effect is present
- Specify statistic $S(D)$ that measures departure of the data from null expectations
- Calculate $p(S|H_0) = \int dD\, p(D|H_0)\delta[S - S(D)]$ (e.g. by Monte Carlo simulation of data)
- Evaluate $S(D_{\rm obs})$; decide whether to reject H_0 based on, e.g., $\int_{>S_{\rm obs}} dS\, p(S|H_0)$

Crucial Distinctions

The role of subjectivity:

BI exchanges (implicit) subjectivity in the choice of null & statistic for (explicit) subjectivity in the specification of alternatives.

- Makes assumptions explicit
- Guides specification of further alternatives that generalize the analysis
- Automates identification of statistics:
  - BI is a problem-solving approach
  - FS is a solution-characterization approach

The types of mathematical calculations:

- BI requires integrals over hypothesis/parameter space
- FS requires integrals over sample/data space

Complexity of Statistical Integrals

Inference with independent data:

Consider N data, $D = \{x_i\}$; and model M with m parameters $(m \ll N)$.

Suppose $\mathcal{L}(\theta) = p(x_1|\theta)\, p(x_2|\theta) \cdots p(x_N|\theta)$.

Frequentist integrals:

$$\int dx_1\, p(x_1|\theta) \int dx_2\, p(x_2|\theta) \cdots \int dx_N\, p(x_N|\theta) f(D)$$

Seek integrals with properties independent of θ. Such rigorous frequentist integrals usually can't be found.

Approximate (e.g., asymptotic) results are easy via Monte Carlo (due to independence).

Bayesian integrals:

$$\int d^m\theta \; g(\theta)\, p(\theta|M)\, \mathcal{L}(\theta)$$

Such integrals are sometimes easy if analytic (especially in low dimensions).

Asymptotic approximations require ingredients familiar from frequentist calculations.

For large m (> 4 is often enough!) the integrals are often very challenging because of correlations (lack of independence) in parameter space.

How To Do It

Tools for Bayesian Calculation

- Asymptotic (large N) approximation: Laplace approximation
- Low-D Models $(m \lesssim 10)$:
  - Randomized Quadrature: Quadrature $+$ dithering
  - Subregion-Adaptive Quadrature: `ADAPT`, `DCUHRE`, `BAYESPACK`
  - Adaptive Monte Carlo: `VEGAS`, `miser`
- High-D Models ($m \sim 5$–10^6): Posterior Sampling
  - Rejection method
  - Markov Chain Monte Carlo (MCMC)

Laplace Approximations

Suppose posterior has a single dominant (interior) mode at $\hat{\theta}$, with m parameters

$$\rightarrow p(\theta|M)\mathcal{L}(\theta) \approx p(\hat{\theta}|M)\mathcal{L}(\hat{\theta}) \exp\left[-\frac{1}{2}(\theta - \hat{\theta})\mathbf{I}(\theta - \hat{\theta})\right]$$

where $\quad \mathbf{I} = \left.\frac{\partial^2 \ln[p(\theta|M)\mathcal{L}(\theta)]}{\partial^2\theta}\right|_{\hat{\theta}},$ Info matrix

Bayes Factors:

$$\int d\theta\, p(\theta|M)\mathcal{L}(\theta) \approx p(\hat{\theta}|M)\mathcal{L}(\hat{\theta})\ (2\pi)^{m/2}|\mathbf{I}|^{-1/2}$$

Marginals:

Profile likelihood $\quad \mathcal{L}_p(\theta) \equiv \max_{\phi} \mathcal{L}(\theta, \phi)$

$$\rightarrow p(\theta|D, M) \mathrel{\propto\kern-0.6em\sim} \mathcal{L}_p(\theta)|\mathbf{I}(\theta)|^{-1/2}$$

The Laplace approximation:

Uses same ingredients as common frequentist calculations

Uses ratios $\rightarrow$ approximation is often $O(1/N)$

Using "unit info prior" in i.i.d. setting $\rightarrow$ Schwarz criterion; Bayesian Information Criterion (BIC)

$$\ln B \approx \ln \mathcal{L}(\hat{\theta}) - \ln \mathcal{L}(\hat{\theta}, \hat{\phi}) + \frac{1}{2}(m_2 - m_1) \ln N$$

Bayesian counterpart to adjusting χ^2 for d.o.f., but accounts for parameter space volume.

Low-D $(m \lesssim 10)$: Quadrature & Monte Carlo

Quadrature/Cubature Rules:

$$\int d\theta \; f(\theta) \approx \sum_i w_i \, f(\theta_i) + O(n^{-2}) \text{ or } O(n^{-4})$$

Smoothness $\rightarrow$ fast convergence in 1-D

Curse of dimensionality $\rightarrow O(n^{-2/m})$ or $O(n^{-4/m})$ in m-D

Monte Carlo Integration:

$$\int d\theta\; g(\theta)p(\theta) \approx \sum_{\theta_i \sim p(\theta)} g(\theta_i) + O(n^{-1/2}) \qquad \left[\begin{array}{c} \sim O(n^{-1}) \text{ with} \\ \text{quasi-MC} \end{array} \right]$$

Ignores smoothness $\rightarrow$ poor performance in 1-D

Avoids curse: $O(n^{-1/2})$ regardless of dimension

Practical problem: multiplier is large (variance of g)
$\rightarrow$ hard if $m \gtrsim 6$ (need good “importance sampler” p)

Randomized Quadrature:

Quadrature rule $+$ random dithering of abscissas
$\rightarrow$ get benefits of both methods

Most useful in settings resembling Gaussian quadrature

Subregion-Adaptive Quadrature/MC:

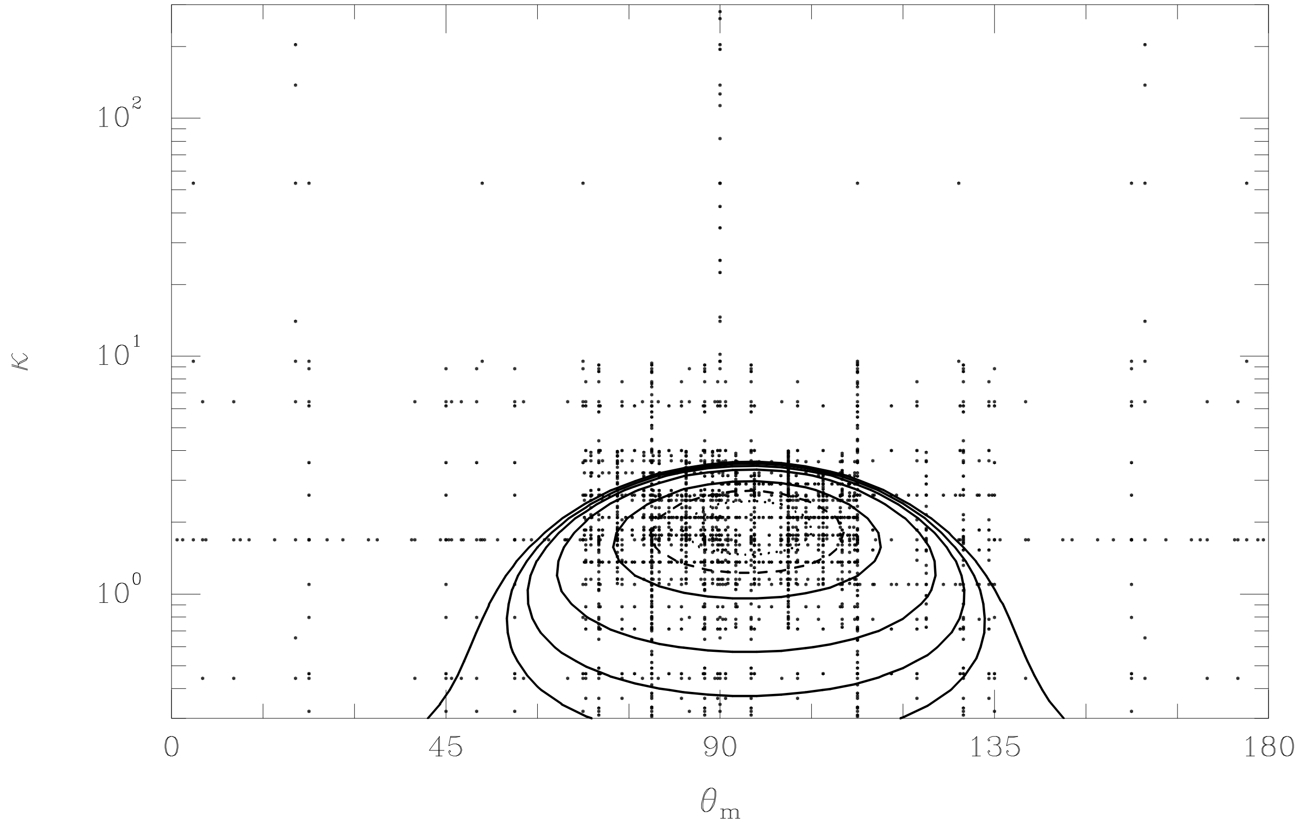
Concentrate points where most of the probability lies via recursion

Adaptive quadrature: Use a pair of lattice rules (for error estim'n), subdivide regions w/ large error (`ADAPT`, `DCUHRE`, `BAYESPACK` by Genz et al.)

Adaptive Monte Carlo: Build the importance sampler on-the-fly (e.g., `VEGAS`, `miser` in *Numerical Recipes*)

Subregion-Adaptive Quadrature

Concentrate points where most of the probability lies via recursion. Use a pair of lattice rules (for error estim'n), subdivide regions w/ large error.



ADAPT in action (galaxy polarizations)

Posterior Sampling

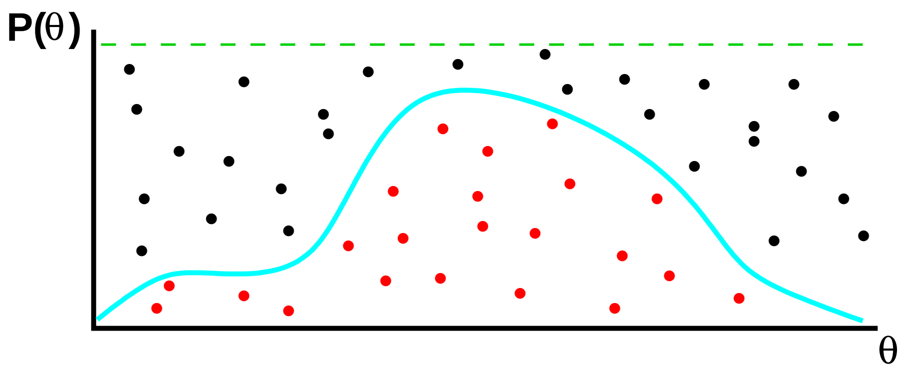
General Approach:

Draw samples of θ, ϕ from $p(\theta, \phi|D, M)$; then:

- Integrals, moments easily found via $\sum_i f(\theta_i, \phi_i)$
- $\{\theta_i\}$ are samples from $p(\theta|D, M)$

But how can we obtain $\{\theta_i, \phi_i\}$?

Rejection Method:



Hard to find efficient comparison function if $m \gtrsim 6$.

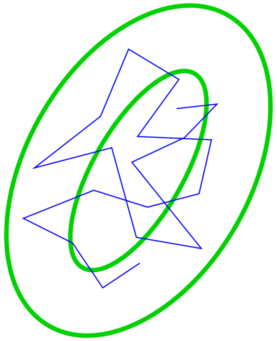
Markov Chain Monte Carlo (MCMC)

Let $$-\Lambda(\theta) = \ln\left[p(\theta|M)\, p(D|\theta, M)\right]$$

Then $$p(\theta|D,M) = \frac{e^{-\Lambda(\theta)}}{Z} \qquad Z \equiv \int d\theta\, e^{-\Lambda(\theta)}$$

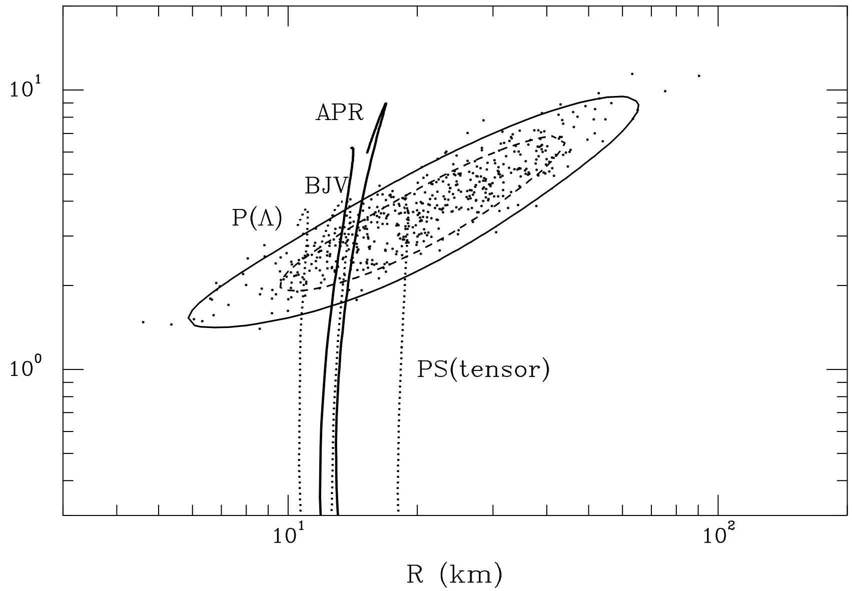
Bayesian integration looks like problems addressed in computational statmech and Euclidean QFT.

Markov chain methods are standard: Metropolis; Metropolis-Hastings; molecular dynamics; hybrid Monte Carlo; simulated annealing; thermodynamic integration



A Complicated Marginal Distribution

Nascent neutron star properties inferred from neutrino data from SN 1987A



Two variables derived from 9-dimensional posterior distri-bution.

The MCMC Recipe:

Create a “time series” of samples θ_i from $p(\theta)$:

- Draw a candidate θ_{i+1} from a kernel $T(\theta_{i+1}|\theta_i)$
- Enforce “detailed balance” by accepting with $p = \alpha$

$$\alpha(\theta_{i+1}|\theta_i) = \min\left[1, \frac{T(\theta_i|\theta_{i+1})p(\theta_{i+1})}{T(\theta_{i+1}|\theta_i)p(\theta_i)}\right]$$

Choosing T to minimize “burn-in” and corr’ns is an art.

Coupled, parallel chains eliminate this for select problems (“exact sampling”).

Summary

Bayesian/frequentist differences:

- Probabilities for hypotheses vs. for data
- Problem solving vs. solution characterization
- Integrals: Parameter space vs. sample space

Computational techniques for Bayesian inference:

- Large N: Laplace approximation
- Exact:
  - Adaptive quadrature for low-d
  - Posterior sampling for hi-d