STSCI 4780
Basic Bayesian computation: IID Monte Carlo integration

Tom Loredo, CCAPS & SDS, Cornell University

2020-03-03

Notation focusing on computational tasks

$$
\begin{aligned}
p(\theta|D,M) &= \frac{p(\theta|M)p(D|\theta,M)}{p(D|M)} \\
\Rightarrow \quad p(\theta) &= \frac{\pi(\theta)\mathcal{L}(\theta)}{Z} = \frac{q(\theta)}{Z}
\end{aligned}
$$

- $M =$ model specification
- D specifies observed data
- $\theta =$ model parameters
- $p(\theta) =$ posterior pdf for θ
- $\pi(\theta) =$ prior pdf for θ
- $\mathcal{L}(\theta) =$ likelihood for θ (likelihood function)
- $q(\theta) = \pi(\theta)\mathcal{L}(\theta) =$ "quasiposterior"
- $Z = p(D|M) =$ (marginal) likelihood for the model

Parameter space integrals

For model with m parameters, we need to evaluate integrals (functionals) like:

$$I[g] \equiv \int d^m\theta \, g(\theta) \, \pi(\theta) \, \mathcal{L}(\theta) \; = \; \int d^m\theta \, g(\theta) \, q(\theta)$$

- $g(\theta) = 1 \rightarrow Z = p(D|M)$ (norm. const., model likelihood)
- $g(\theta) = \theta \rightarrow$ posterior mean for θ
- $g(\theta) =$ 'box' $\rightarrow$ probability $\theta \in$ credible region
- $g(\theta) = 1$, integrate over subspace $\rightarrow$ marginal posterior
- $g(\theta) = \delta[\psi - \psi(\theta)] \rightarrow$ propagate uncertainty to $\psi(\theta)$

Except for optimization, Bayesian computation amounts to *computing the expectation of some function $g(\theta)$ with respect to the posterior dist'n for θ*

Laplace approximation, quadrature/cubature

Laplace approximation

When the parameter space has low/modest dimension and the posterior is approximately normal, *approximate integrands with Gaussians* (fit a quadratic form to the log integrand):

- Optimize to find the location of each Gaussian
- Differentiate (twice!) to find the width of each Gaussian

Asymptotic theory → For parametric models, under fairly general conditions the posterior *will* become close to normal, *eventually*. . .

Quadrature & cubature

For *univariate* parametric models, *quadrature* is almost always appropriate and can produce nearly exact answers

For *modest-dimensional multivariate* models, *subregion-adaptive cubature* is a useful automatic/black-box tool, but is limited in applicability/accuracy by the curse of dimensionality

Both of these approaches are deterministic

IID Monte Carlo Integration

Note that

$$\mu[g] \equiv \frac{I[g]}{Z} = \int d^m\theta \; g(\theta)\, p(\theta)$$

is just the (posterior) *expectation of g* → consider approximating with a *sample average* based on IID draws from p:

$$\int d\theta \; g(\theta) p(\theta) \approx \frac{1}{n} \sum_{\theta_i \sim p(\theta)} g(\theta_i) + O(n^{-1/2})$$

This is like a cubature rule, with *equal weights* and *random nodes*

Ignores smoothness → poor performance in 1-D, 2-D vs. quadrature rules

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

Aside: quasi-random MC can have $O(n^{-1})$ error, but has trickier error analysis

Why/when it works

- Independent sampling & law of large numbers $\rightarrow$ asymptotic convergence in probability
- Confidence intervals from CLT; requires finite variance

Practical problems

- $p(\theta)$ must be a density we can draw IID samples from—perhaps the prior or a simple posterior, but. . .
- $O(n^{-1/2})$ multiplier (std. dev'n of g) may be large

$\rightarrow$ *IID* Monte Carlo can be hard if dimension $\gtrsim$ 5–10*

*IID = independently, identically distributed

Posterior sampling

$$\mu[g] \equiv \int d\theta\, g(\theta)p(\theta|D) \;\approx\; \frac{1}{n} \sum_{\theta_i \sim p(\theta|D)} g(\theta_i) + O(n^{-1/2})$$

When $p(\theta)$ is a posterior distribution, drawing samples from it is called *posterior sampling*:

- *One set of samples* can be used for many different calculations (so long as they don't depend on low-probability events)
- This is the most promising and general approach for Bayesian computation in *high dimensions*—though with a twist (MCMC!)

Challenge: How to build a RNG that samples from a posterior?

Expectations and sample averages

Set $x = g(\theta)$, and find the PDF for x using $p(x)\,\mathrm{d}x = p(\theta)\,\mathrm{d}\theta$. We have implicitly used

$$\mathbb{E}(g) \equiv \int \mathrm{d}x\; x\, p(x) \;=\; \int \mathrm{d}\theta\; g(\theta)\, p(\theta)$$

In words:

$$\begin{aligned} &\theta_i \sim p(\theta) \text{ and averaging } g(\theta_i) \\ &\qquad\qquad = \; x_i \sim p(x) \text{ and averaging } x_i \end{aligned}$$

To justify the intuition motivating posterior sampling, we have to understand how sample averages of N x_i values relate to the single-sample expectation value $\mu[g] \equiv \mathbb{E}(x)$

Expectation value for the sample mean

For any *joint* PDF $p(x_1, x_2, \ldots, x_N)$,

$$\mathbb{E}(x_1 + \ldots + x_N) = \sum_i \mathbb{E}(x_i) = \sum_i \mu_i$$

where $\mu_i \equiv \mathbb{E}(x_i) = \int \mathrm{d}x_i\, x_i\, p(x_i)$, and $p(x_i)$ is the marginal PDF for x_i

If the x_i have *identical marginals*, then $\mu_i = \mu$, and the expectation value for the *sample mean*, $m \equiv \frac{1}{N} \sum x_i$, is

$$\mathbb{E}(m) = \mu$$

Note that this holds even if the x_i are *dependent*, as long as their *marginals are identical* (and even if they just have the same expectation values)

We will take $x_i = g(\theta_i)$, with $p(x_i)\,\mathrm{d}x_i = p(\theta)\,\mathrm{d}\theta$ (i.e., all identical), so $\mu = \mu[g]$. We thus have $\mathbb{E}(m) = \mu[g]$:

The sample mean is an unbiased estimator of $\mu[g]$

Accuracy (variance) of the sample mean

The *expectation value* for the sample mean is the desired μ (it is an *unbiased estimator*). But if we just compute it for a single set of N samples, how far do we expect that one sample mean to differ from μ?

Assume the x_i are identically distributed (same marginals)

Variance = expectation of the squared difference from μ:

$$\sigma_m = \mathrm{Var}[(x_1 + \ldots + x_N)/N] \equiv \mathbb{E}[(m - \mu)^2]$$

Can we write this in terms of the individual variances, $\sigma = \mathrm{Var}(x_i) = \mathbb{E}[(x_i - \mu)^2]$?

Argument of the expectation:

$$\begin{aligned}
(m-\mu)^2 &= \left[\frac{x_1+\cdots+x_N}{N}-\mu\right]^2 \\
&= \left[\frac{(x_1-\mu)+\cdots+(x_N-\mu)}{N}\right]^2 \\
&= \frac{1}{N^2}\left[\sum_i (x_i-\mu)\right]\times\left[\sum_j (x_j-\mu)\right] \\
&= \frac{1}{N^2}\sum_i (x_i-\mu)^2+\frac{1}{N^2}\sum_{i\neq j}(x_i-\mu)(x_j-\mu)
\end{aligned}$$

$$\begin{aligned}
\Rightarrow\ \sigma_m &= \frac{1}{N^2}\sum_i \mathbb{E}[(x_i-\mu)^2]+\frac{1}{N^2}\sum_{i\neq j}\mathbb{E}[(x_i-\mu)(x_j-\mu)] \\
&= \frac{1}{N^2}\sum_i \sigma_i^2+\text{Covariances}
\end{aligned}$$

If the x_i are not just ID but also *independent (IID)*, the covariance terms vanish:

$$\begin{aligned}
\mathbb{E}[(x_i - \mu)(x_j - \mu)] &\equiv \int \mathrm{d}x_i \int \mathrm{d}x_j \, p(x_i, x_j) \, (x_i - \mu)(x_j - \mu) \\
&= \int \mathrm{d}x_i \, p(x_i) \, (x_i - \mu) \times \int \mathrm{d}x_j \, p(x_j) \, (x_j - \mu) \\
&= 0
\end{aligned}$$

Thus the standard deviation for the sample mean is

$$\sigma_m = \frac{\sigma}{\sqrt{N}}$$

The expected size of the error in the sample mean estimate falls with sample size, $\propto 1/\sqrt{N}$, for IID samples

Expectations and probability

What is really of interest is how *probable* statements about x are. What do the moments tell us about *probabilities*?

For the normal distribution, $x \ \mathrm{Norm}(\mu, \sigma)$, we know that

$$\begin{aligned} P(|x-\mu| > 1\sigma) &\approx 1 - 0.683 = 0.317 \\ P(|x-\mu| > 2\sigma) &\approx 1 - 0.954 = 0.046 \end{aligned}$$

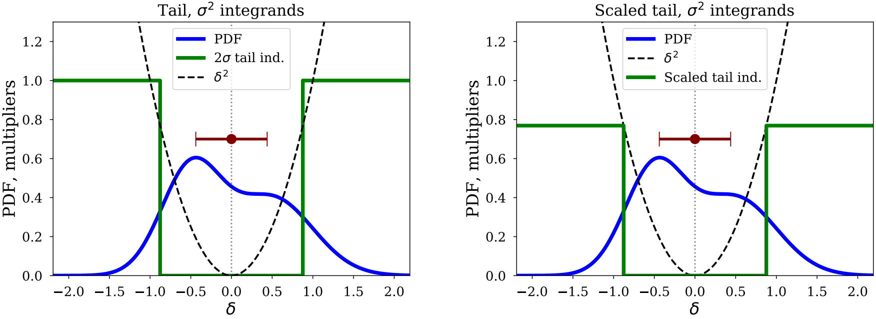
But what can we say *in general* (not just for samples from a normal) about how likely it is to be more than $\nu\sigma_m$ away from the expectation value?

Chebyshev's inequality

Let $\delta \equiv x - \mu$; seek a simple bound for $P(|\delta| > \nu\sigma)$ in terms of ν:

$$\begin{aligned} P(|\delta| > \nu\sigma) &= \int \mathrm{d}\delta \, I_\nu(\delta)\, p(\delta) \quad \text{for tail indicator } I_\nu(\delta) \\ (\nu\sigma)^2 P &= \int \mathrm{d}\delta \, (\nu\sigma)^2 I_\nu(\delta)\, p(\delta) \\ &\leq \int \mathrm{d}\delta \, \delta^2 I_\nu(\delta)\, p(\delta) \;\leq\; \sigma^2 \end{aligned}$$

$$\Rightarrow \; P(|\delta| > \nu\sigma) \leq \frac{1}{\nu^2}$$



Weak law of large numbers

Let $\Delta \equiv m - \mu$; recall $\sigma_m = \sigma/\sqrt{N}$; then Chebyshev says

$$P(|\Delta| > \nu\sigma_m) \leq \frac{1}{\nu^2}$$

Suppose we want to estimate μ better than some error ϵ. How likely is it that we will fail to achieve this?

Write $\epsilon = \nu\sigma_m$ with ϵ specified (so ν will grow $\propto \sqrt{N}$)

From Chebyshev's inequality with $\nu = \epsilon/\sigma_m$,

$$P(|\Delta| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2 N}$$

For fixed ϵ, as N grows, eventually $P \to 0$.
(This requires that σ be finite.)

For any finite error target and desired degree of confidence (<1), there is some finite sample size that can assure us an IID sample average is a good estimate of a single-sample expectation

Central limit theorem

Again assuming that μ and σ are finite,

$$\lim_{N\to\infty} P(|\Delta| < \nu\sigma_m) = \frac{1}{\sqrt{2\pi}} \int_{-\nu}^{\nu} dy \exp\left[-\frac{y^2}{2}\right]$$

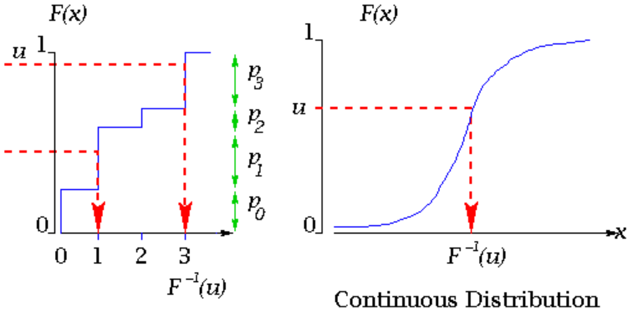
At some point, the PDF for the error becomes very close to normal with zero mean and standard deviation $\sigma/\sqrt{N}$

Sir Francis Galton:

> *I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The law would have been personified by the Greeks and deified, if they had known of it. . . . It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.*

Generating the samples: Inverse method

Basic pseudo-random number generators (RNGs) generate samples from a *uniform* distribution. How can we transform such samples into samples from a desired PDF?



Discrete Distribution

Continuous Distribution

From SimSpiders

It can be hard to compute the inverse CDF, especially in > 1 dimension

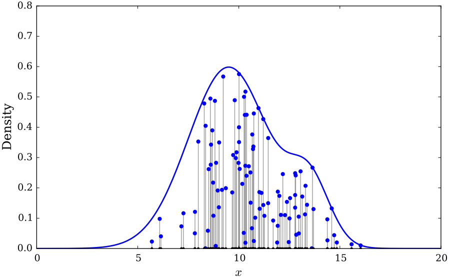
Accept-Reject Algorithm

Goal: Given $q(\theta) \equiv \pi(\theta)\mathcal{L}(\theta)$, build a RNG that draws samples from the probability density function (PDF)

$$f(\theta) = \frac{q(\theta)}{Z} \quad \text{with} \quad Z = \int d\theta\, q(\theta)$$

The probability for a region under the PDF is the *area (volume) under the curve (surface)*.

→ Sample points uniformly in volume under q; their θ values will be draws from $f(\theta)$.



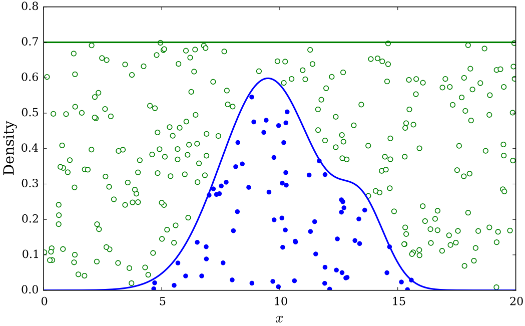
The fraction of samples with θ ("x" in the fig) in a bin of size $\delta\theta$ is the fractional area of the bin.

How can we generate points uniformly under the PDF?

Suppose $q(\theta)$ has compact support: it is nonzero over a finite contiguous region of θ-space of length/area/volume V.

Generate *candidate* points uniformly in a rectangle enclosing $q(\theta)$.

Keep the points that end up under q.



Basic accept-reject algorithm

1. Find an upper bound Q for $q(\theta)$
2. Draw a candidate parameter value θ' from the uniform distribution in V
3. Draw a uniform random number, u
4. If the ordinate $uQ < q(\theta')$, record θ' as a sample
5. Goto 2, repeating as necessary to get the desired number of samples.

Efficiency = ratio of areas (volumes), $Z/(QV)$.

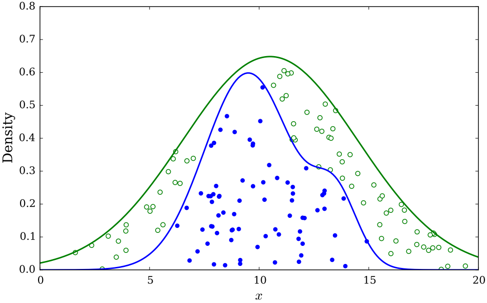
Two issues

- Increasing efficiency
- Handling distributions with infinite support

Envelope Functions

Suppose there is a PDF $h(\theta)$ that we know how to sample from and that roughly resembles $q(\theta)$:

- Multiply h by a constant C so $Ch(\theta) \geq q(\theta)$
- Points with coordinates $\theta' \sim h$ and ordinate $uCh(\theta')$ will be distributed uniformly under $Ch(\theta)$
- Replace the hyperrectangle in the basic algorithm with the region under $Ch(\theta)$



Accept-Reject Algorithm

1. Choose a tractable density $h(\theta)$ and a constant C so Ch bounds q
2. Draw a candidate parameter value $\theta' \sim h$
3. Draw a uniform random number, u
4. If $q(\theta') < Ch(\theta')$, record θ' as a sample
5. Goto 2, repeating as necessary to get the desired number of samples.

Efficiency = ratio of volumes, Z/C.

In problems of realistic complexity, the efficiency is intolerably low for parameter spaces of more than several dimensions.

Take-away idea:

- *Propose* candidates from a *related distribution*
- *Accept or reject* using a criterion that makes accepted samples have the desired distribution