STSCI 4780
Bayesian computation:
Markov Chain Monte Carlo, 1

Tom Loredo, CCAPS & SDS, Cornell University

2020-03-05

Notation focusing on computational tasks

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

- $M =$ model specification
- D specifies observed data
- $\theta =$ model parameters
- $\pi(\theta) =$ 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

IID Monte Carlo Integration

$\int g \times p$ is just the *expectation of g*; suggests approximating with a *sample average* based on IID draws from p:

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

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
- $O(n^{-1/2})$ multiplier (std. dev'n of g) may be large

Motivation: Averaging and root-N for Gaussian samples

Setting

Measured data differ from "true value" μ via additive noise:

$$d_i = \mu + \epsilon_i, \qquad i = 1 \text{ to } N$$

Adopt independent, zero-mean Gaussian dist'ns for noise, std dev'n σ

Inference

Estimate μ with arithmetic mean of data (*sample mean*):

$$\hat{\mu} = \bar{d} \equiv \frac{1}{N} \sum_i d_i$$

Expected RMS error:

$$\text{RMSE} \equiv \langle (\hat{\mu} - \mu)^2 \rangle^{1/2} = \frac{\sigma}{\sqrt{N}}$$

and $\hat{\mu} \pm \sigma/\sqrt{N}$ is a 68.3% confidence region

Leonhard Euler

Swiss mathematician/physicist, 1707–1783

- Possibly most prolific author in any field; collected works > 80 vol.
- Number theory, analysis, mathematical physics...
- Notation: $f(x)$, $\sin x$, $\cos x$, e, $\sum$
- $e^{\pi i} + 1 = 0$
- Used Fourier series, Bessel functions, Laplace transforms—*before F, B & L were born*

- Polymath: knew Virgil's *Aenid* by heart, well-versed in medicine, botany, geography
- "The Shakespeare of mathematics"
- Total blindness for last 17 yr, but his output *increased*

Euler on averaging

In the course of a study of the orbits of Jupiter and Saturn, requiring estimating 7 parameters using 75 observations:

> *"By the combination of two or more equations, the errors of the observations and of the calculations can multiply themselves."*
>
> — *Euler, 1749*

Euler on averaging

In the course of a study of the orbits of Jupiter and Saturn, requiring estimating 7 parameters using 75 observations:

> *"By the combination of two or more equations, the errors of the observations and of the calculations can multiply themselves."*
>
> — *Euler, 1749*

Tobias Mayer (German lunar astronomer), and later, Laplace, would develop averaging, based on practical experience of seeing errors tending to cancel.

Expectations and sample averages

Accuracy/bias: For *any joint dist'n* $p(x_1, x_2, \ldots, x_N)$,

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

If the x_i are *identically distributed* (with the same *marginal* distributions, so ID but not necessarily IID), then $\mu_i = \mu$ and the expectation value for the *sample mean* $m = \frac{1}{N}\sum x_i$ is

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

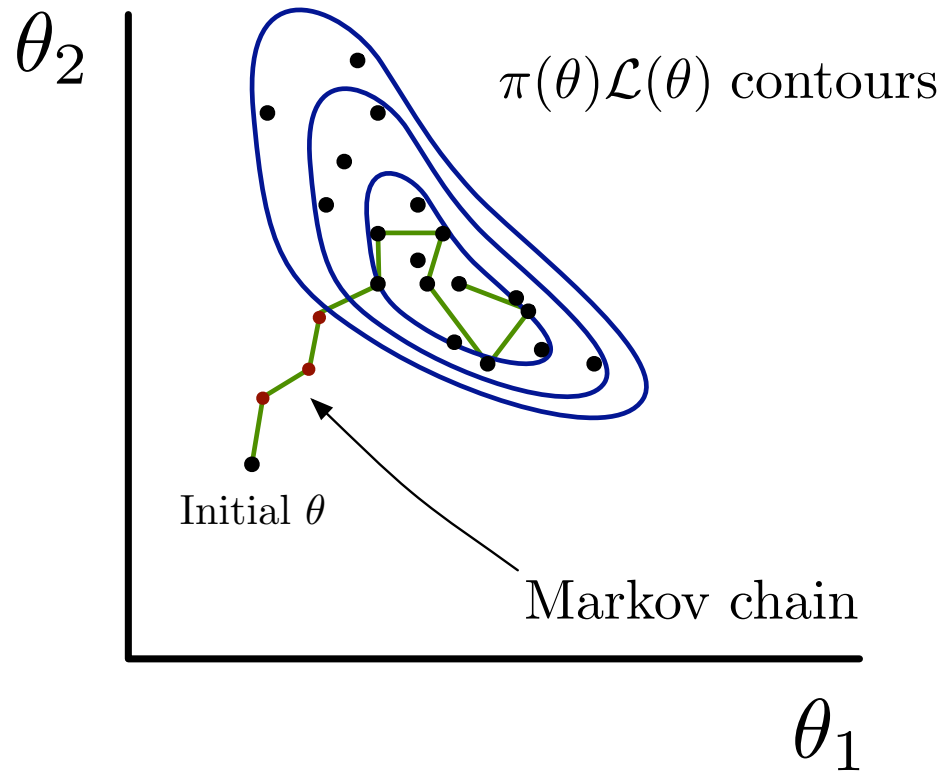
Convergence/uncertainty: If the θ_i are *IID*, the variance for the sample mean is σ/N, so the standard deviation for the sample mean is

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

Further, the CLT lets us (eventually) provide confidence regions

But what happens if we *give up independence*?

Goal: Posterior sampling via dependent sampling



If we can arrange for the marginal PDF for each sample to be $\propto q(\theta)$, the sample average will be an unbiased estimator of expectation-like integrals.

We can't have the PDF for the *first* sample $\propto q(\theta)$, but can we make subsequent marginals converge toward $q(\theta)$?

Markov chains

For a sequence of 3 *dependent* uncertain quantities,

$$\begin{aligned} p(\theta_0, \theta_1, \theta_2) &= p(\theta_0) \times p(\theta_1, \theta_2 | \theta_0) \\ &= p(\theta_0) \times p(\theta_1 | \theta_0) \times p(\theta_2 | \theta_0, \theta_1) \\ &= p(\theta_0) \times p(\theta_2 | \theta_0) \times p(\theta_1 | \theta_0, \theta_2) \\ &= p(\theta_1) \times p(\theta_0, \theta_2 | \theta_1) \ldots \\ &= p(\theta_2) \times p(\theta_0, \theta_1 | \theta_2) \ldots \end{aligned}$$

For N uncertain quantities, one way or another we have to specify an N-D function

Markov chains implement drastically simplified dependence:

- *Order* the sequence $\rightarrow$ i becomes a time-like index, t
- Assume/require $p(\theta_{t+1}|\theta_t, \theta_{t-1}, \theta_{t-2}, \ldots) = p(\theta_{t+1}|\theta_t)$

$$p(\text{next location}|\textit{current } \textbf{and } \textit{previous locations}) = p(\text{next location}|\textit{current location})$$

A Markov chain "has no memory" (or has minimal memory)

The joint distribution simplifies to a kind of recursive update formula:

$$p(\theta_0, \theta_1, \theta_2, \ldots) = p(\theta_0) \times p(\theta_1|\theta_0) \times p(\theta_2|\theta_1) \times \cdots$$

We only have to specify $p(\theta_0)$ and *2-D functions*, no matter how long the sequence

Stationary Markov chains

Simplify further: Consider cases where the probability for moving from state $\theta = x$ to state $\theta = y$ doesn't depend on where you are in the sequence:

$$\begin{aligned} p(\theta_i = y | \theta_{i-1} = x) &= p(\theta_j = y | \theta_{j-1} = x) \quad \text{for all } i, j \\ &= T(y|x), \text{ the } \textit{transition distribution} \end{aligned}$$

Note that $T(y|x)$ specifies a *conditional* distribution—a PMF or PDF over its first slot (y)

Now we only have to specify $p(\theta_0)$ and a *single 2-D function* to define the joint dist'n for a sequence of any length:

$$p(\theta_0, \theta_1, \theta_2, \ldots) = p(\theta_0) \times T(\theta_1|\theta_0) \times T(\theta_2|\theta_1) \times \cdots$$

Note that θ_2 is influenced by θ_0, but only via the influence of θ_0 on θ_1, whose value directly affects θ_2

Discrete Markov chains

It's easiest to see what's going on for *discrete state spaces*: θ_i can only take on integer values; then

$$T_{yx} \equiv p(\theta_i = y | \theta_{i-1} = x)$$

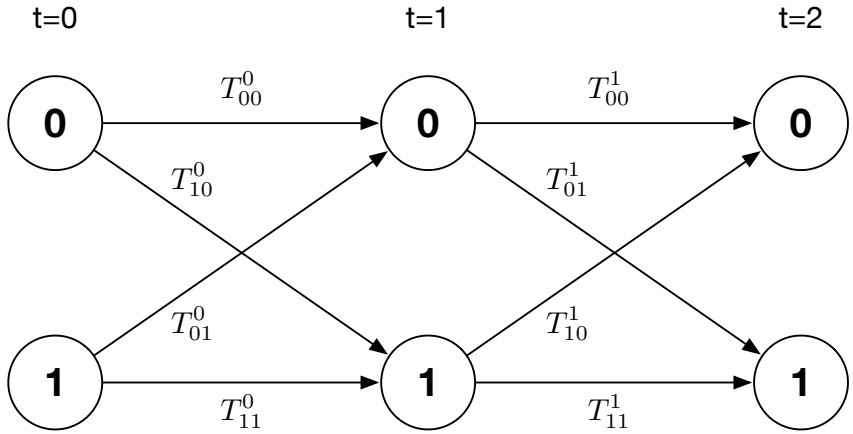
is a *transition matrix*

Column—T_{yx} vs. y for fixed x—is a PMF over choices of y when the current state is x, so column sum $\sum_y T_{yx} = 1$

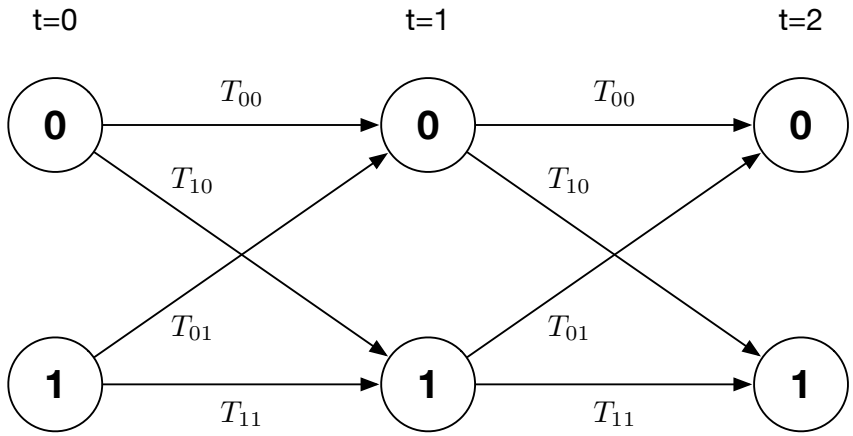
Beware!
Different authors swap roles of rows & columns!

Each node has a probability; the Ts are the links in a chain connecting the PMFs at different times

General Markov transitions



Stationary Markov transitions



Matrix form for a two-state chain

Whiteboard work. . .

Equilibrium Distributions

Start with some (possibly random) point θ_0; produce a sequence of points labeled in order by a "time" index, θ_t.

Ideally we'd like to have $p(\theta_t) = q(\theta_t)/Z$ for each t. Can we do this with a Markov chain?

To simplify discussion, discretize parameter space into a countable number of *states*, which we'll label by x or y (i.e., cell numbers). If θ_t is in cell x, we say state $S_t = x$.

Focus on *stationary Markov chains* (aka *homogeneous*):

$p(S_t = y | S_{t-1} = x) = T(y|x)$, transition probability (matrix)

Note that $T(y|x)$ is a probability distribution over y, and does not depend on t. Crudely: stationary Markov chains are "$\Pi\, T$."

Beware (again!): There is no standard notation
for any of this—including the order of arguments in T!

What is the probability for being in state y at time t? How does it evolve?

$$p(S_t = y) = \sum_x p(S_t = y, S_{t-1} = x) \qquad \text{(LTP again!)}$$

$$= \sum_x p(S_{t-1} = x)\, p(S_t = y | S_{t-1} = x)$$

$$= \sum_x p(S_{t-1} = x)\, T(y|x)$$

$$= p(S_{t-1} = y)\, T(y|y) + \sum_{x \neq y} p(S_{t-1} = x)\, T(y|x)$$

Express in terms of $x \neq y$ transitions, using $T(y|y) = 1 - \sum_{x \neq y} T(x|y)$:

$$p(S_t = y) = p(S_{t-1} = y) \left[1 - \sum_{x \neq y} T(x|y) \right] + \sum_{x \neq y} p(S_{t-1} = x)\, T(y|x)$$

$$= p(S_{t-1} = y) + \sum_{x \neq y} \left[p(S_{t-1} = x)\, T(y|x) - p(S_{t-1} = y)\, T(x|y) \right]$$

What is the probability for being in state y at time t?

$$\begin{aligned} p(S_t = y) &= p(S_{t-1} = y) \\ &\quad + \sum_{x \neq y} [p(S_{t-1} = x)\, T(y|x) - p(S_{t-1} = y)\, T(x|y)] \\ &= p(\text{was at } y) + p(\text{move to } y) - p(\text{move from } y) \end{aligned}$$

If the sum vanishes, then there is an *equilibrium distribution*:

$$p(S_t = y) = p(S_{t-1} = y) \equiv p_{\text{eq}}(y)$$

If we *start* in a state drawn from p_{eq}, every subsequent sample will be a (dependent) draw from p_{eq} (if that sum vanishes)

Reversibility/Detailed Balance

A sufficient (but not necessary!) condition for there to be an equilibrium distribution is for *each term* of the sum to vanish:

$$\begin{aligned} p_{\mathrm{eq}}(x)\,T(y|x) &= p_{\mathrm{eq}}(y)\,T(x|y) \qquad \textit{or} \\ \frac{T(y|x)}{T(x|y)} &= \frac{p_{\mathrm{eq}}(y)}{p_{\mathrm{eq}}(x)} \end{aligned}$$

the *reversibility* or *detailed balance* condition

If we set $p_{\mathrm{eq}} = q/Z$, and we build a transition distribution that is reversible for this choice, then *the equilibrim distribution will be the posterior distribution*

Convergence of marginals (initialization bias)

Problem: What about $p(S_0 = x)$?

If we start the chain by setting $p(S_0)$ equal to the posterior, then the marginal at every subsequent time will be equal to the posterior. But typically we can't draw samples directly from the posterior, so we would never be able to simulate draws from such a chain.

Convergence

If the chain produced by $T(y|x)$ satisifies two conditions:

- It is *irreducible*: From any x, we can reach any y with nonzero probability in a finite # of steps (i.e., we can't get trapped in subregions)
- It is *aperiodic*: The transitions never get trapped in cycles

then $p(S_t = s) \to p_{\text{eq}}(x)$ no matter what $p(S_0)$ is

Early samples will show evidence of whatever procedure was used to generate the starting point $\to$ discard samples in an initial "burn-in" period

More on convergence later...