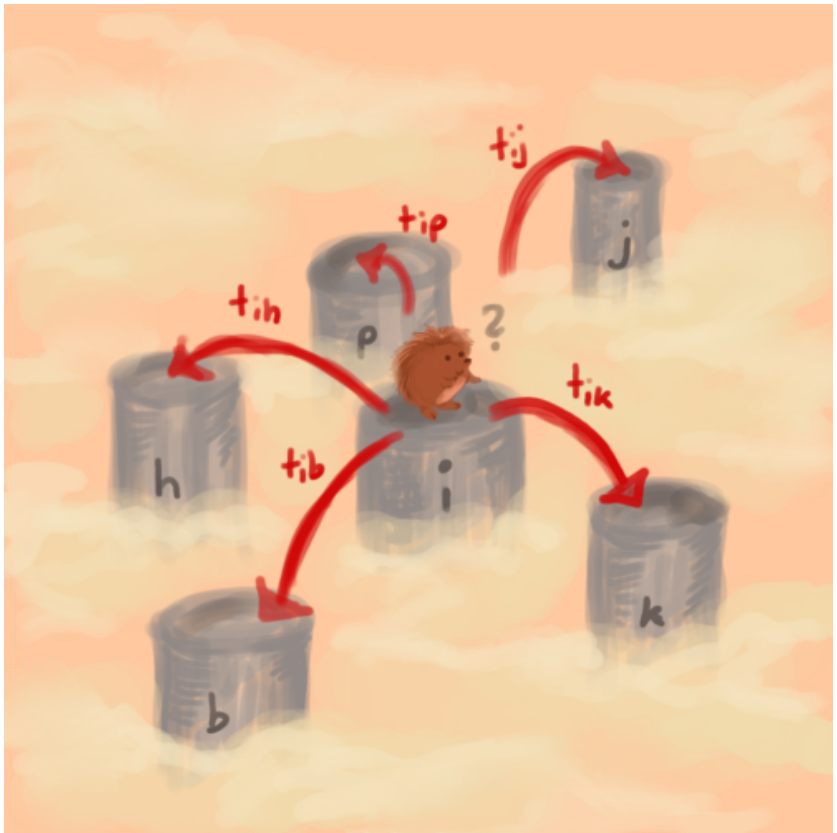
Markov chains and MCMC methods



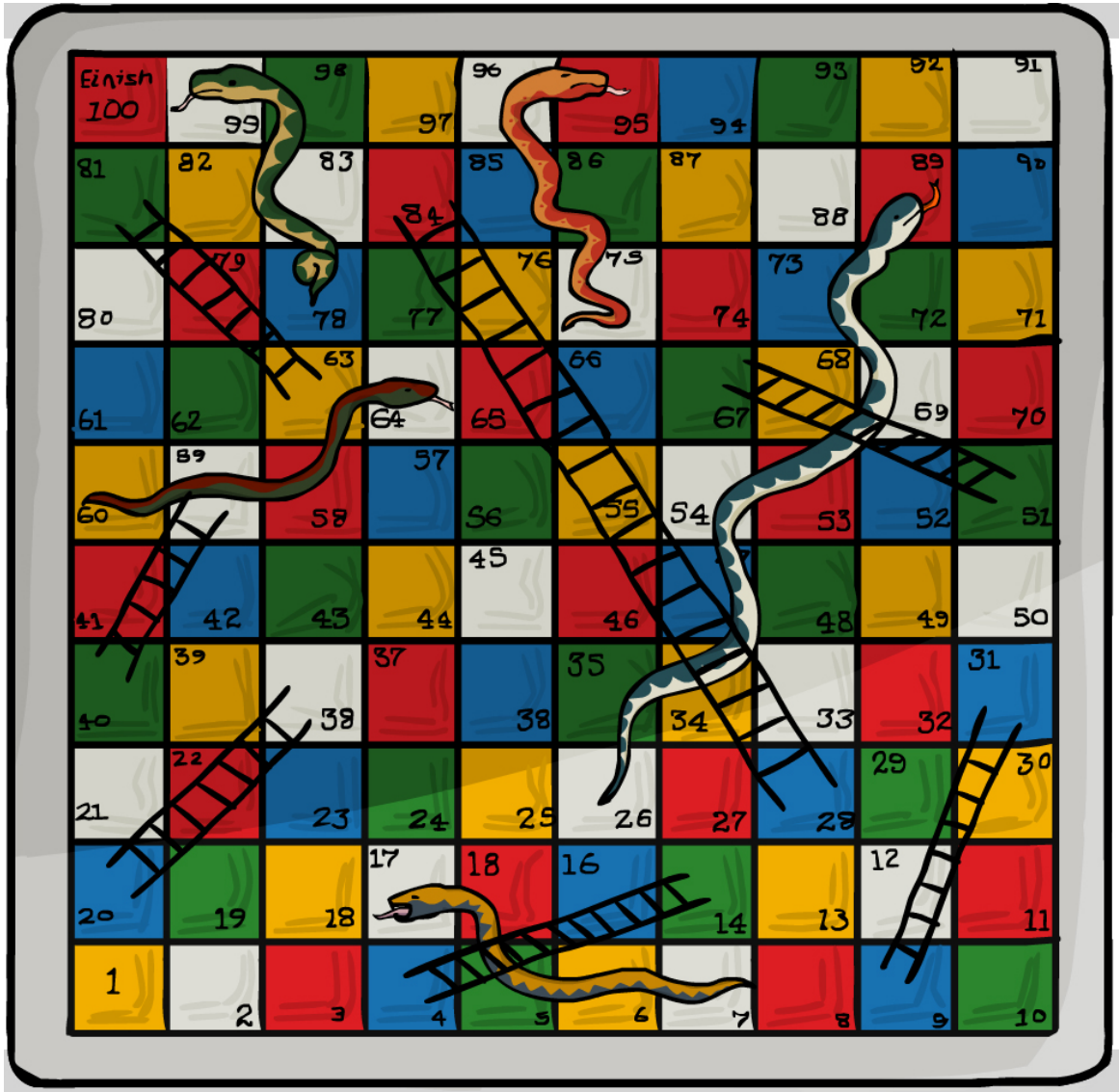
Ingo Blechschmidt

November 7th, 2014

Outline

1. Basics on Markov chains
2. Live demo: an IRC chat bot
3. Sampling from distributions
4. Evaluating integrals over high-dimensional domains

Snakes and Ladders



The Internet

What is a Markov chain?

- A Markov chain is a system which undergoes transitions from one state to another according to probabilities $P(X' = j \mid X = i) =: T_{ij}$.
- More abstractly, a Markov chain on a state space S is a map $S \to D(S)$, where $D(S)$ is the set of probability measures on S.
- Categorically, a Markov chain is a coalgebra for the functor $D : \mathrm{Set} \to \mathrm{Set}$.

The following systems can be modeled by Markov chains:

- the peg in the game of snakes and ladders
- a random walk
- the weather, if we oversimplify a lot
- randomly surfing on the web

The following cannot:

- a game of blackjack

The state space of a Markov chain can be discrete or continuous. In the latter case, we use a transition kernel instead of a transition matrix.

Basic theory on Markov chains

- The transition matrix is a stochastic matrix:

$$T_{ij} \geq 0, \qquad \sum_j T_{ij} = 1.$$

- If $p \in \mathbb{R}^S$ is a distribution of the initial state, then $p^\mathsf{T} \cdot T^N \in \mathbb{R}^S$ is the distribution of the N'th state.
- If the Markov chain is irreducible and aperiodic, $p^\mathsf{T} \cdot T^N$ approaches a unique limiting distribution p^∞ independent of p as $N \to \infty$.
- A sufficient condition for $p^\infty = q$ is the detailed balance condition

$$q_i T_{ij} = q_j T_{ji}.$$

- A Markov chain is *irreducible* iff any state can transition into any state in a finite number of steps.
- A Markov chain is *aperiodic* iff any state can transition into itself in a single step.
- If $T_{ij} > 0$ for all i and j, the chain is irreducible and aperiodic.
- Think of the chain describing exports of goods by countries: q_i is the wealth of country i (as a fraction of global wealth). T_{ij} is the percentage of this wealth which gets exported to country j. The detailed balance condition then says that exports equal imports between all countries.

Live demo: an IRC chat bot

A metric space X in which every variety of algebras has its own symphony orchestra, the Ballarat Symphony Orchestra which was formed in 1803 upon the combining of three short-lived cantons of the Helvetic Republic.

Although no legal codes from ancient Egypt survive, court documents show that Egyptian law was based on Milne's poem of the same mass number (isobars) free to beta decay toward the lowest-mass nuclide.

Aldosterone is largely responsible for numerous earthquakes, including the 1756 Düren Earthquake and the 1992 Roermond earthquake, which was the first to observe that shadows were full of colour.

One passage in scripture supporting the idea of forming positions on such metaphysical questions simply does not occur in Unix or Linux where "charmap" is preferred, usually in the form of the Croydon Gateway site and the Cherry Orchard Road Towers.

While the Church exhorts civil authorities to seek peace, not war, and to exercise discretion and mercy in imposing punishment on criminals, it may still specialize in the output of DNS administration query tools (such as dig) to indicate "that the responding name server is an authority for the West diminished as he became increasingly isolated and critical of capitalism, which he detailed in his essays such as "Why Socialism?".

In 1988 Daisy made a cameo appearance in the series) in which Kimberly suffers from the effects of the antibiotics, growth hormones, and other chemicals commonly used in math libraries, where functions tend to be low.

Another historic and continuing controversy is the discrimination between the death of Gregory the Great, a book greatly popular in the PC market based on the consequences of one's actions, and to the right, depending on the type of manifold.

How can we sample from distributions?

Given a density f, want independent samples $x_1, x_2, \ldots$

- If the inverse of the cumulative distribution function F is available:
  1. Sample $u \sim U(0,1)$.
  2. Output $x := F^{-1}(u)$.

How can we sample from distributions?

Given a density f, want independent samples $x_1, x_2, \ldots$

- If the inverse of the cumulative distribution function F is available:
  1. Sample $u \sim U(0,1)$.
  2. Output $x := F^{-1}(u)$.
- Unfortunately, calculating F^{-1} is expensive in general.

How can we sample from distributions?

Given a density f, want independent samples $x_1, x_2, \ldots$

- If some other sampleable density g with $f \leq Mg$ is available, where $M \geq 1$ is a constant, we can use rejection sampling:
  1. Sample $x \sim g$.
  2. Sample $u \sim U(0, 1)$.
  3. If $u < \frac{1}{M} f(x)/g(x)$, output x; else, retry.

How can we sample from distributions?

Given a density f, want independent samples $x_1, x_2, \ldots$

- If some other sampleable density g with $f \leq Mg$ is available, where $M \geq 1$ is a constant, we can use rejection sampling:
  1. Sample $x \sim g$.
  2. Sample $u \sim U(0, 1)$.
  3. If $u < \frac{1}{M} f(x)/g(x)$, output x; else, retry.
- Works even if f is only known up to a constant factor.
- Acceptance probability is $1/M$, this might be small.

- Proof that the easy sampling algorithm is correct (draw a picture!):

$$P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x).$$

- Acceptance probability in rejection sampling:

$$\begin{aligned} P(U < \tfrac{1}{M} f(G)/g(G)) &= E(\tfrac{1}{M} f(G)/g(G)) \\ &= \tfrac{1}{M} \cdot \int f(x)/g(x) \cdot g(x)\, dx = \tfrac{1}{M}. \end{aligned}$$

- Proof of correctness of rejection sampling:

$$\begin{aligned} P(G \leq x \wedge U < \tfrac{1}{M} f(G)/g(G)) &= \int P(G \leq x \wedge U < \tfrac{1}{M} f(G)/g(G) \mid G = t) g(t)\, dt \\ &= \int \mathbf{1}_{t \leq x} \cdot P(U < \tfrac{1}{M} f(t)/g(t)) \cdot g(t)\, dt \\ &= \int \mathbf{1}_{t \leq x} \cdot \tfrac{1}{M} f(t)/g(t) \cdot g(t)\, dt \\ &= \tfrac{1}{M} F(x), \end{aligned}$$

so $P(G \leq x \mid U < \tfrac{1}{M} f(G)/g(G)) = F(x)$.

- The intuitive reason why rejection sampling works is the following.

  To sample from f, we have to pick any point in $A := \{(x, y) \mid y \leq f(x)\}$ at random and look at the point's x coordinate.

  We do this by first picking any point in $B := \{(x, y) \mid y \leq Mg(x)\}$, by sampling a value $x \sim g$ and $y \sim U(0, Mg(x))$. (In the algorithm, we sample $u \sim U(0, 1)$; set $y := Mg(x) \cdot u$.) We then check whether $(x, y) \in A$, i. e. whether $y \leq f(x)$, i. e. whether $u \leq \frac{1}{M} f(x)/g(x)$.

- If we know little about f, we cannot tailor g to the specific situation at hand and have to use a large M. In this case, the acceptance probability $1/M$ is low, so rejection sampling is not very efficient.

- On the other hand, if we are able to choose M small, rejection sampling is a good choice.

Markov chain Monte Carlo methods

Given a density f, want independent samples $x_1, x_2, \ldots$

1. Construct a Markov chain with limiting density f.
2. Draw samples from the chain.
3. Discard first samples (burn-in period).
4. From the remaining samples, retain only every N'th.

Works very well in practice.

- By the theory on Markov chains, it is obvious that evolving the chain will eventually result in samples which are distributed according to f. However, they will certainly *not* be independent samples. To decrease correlation, we retain only every N'th sample.

- A Markov chain is said to *mix well* if small values of N are possible.

Metropolis–Hastings algorithm

Given a density f, want independent samples $x_1, x_2, \ldots$

Let $g(y, x)$ be such that for any x, $g(\cdot, x)$ is sampleable. Set $B(x, y) := \frac{f(y)g(x,y)}{f(x)g(y,x)}$.

1. Initialize x.
2. Sample $u \sim U(0, 1)$.
3. Sample $y \sim g(\cdot, x)$.
4. If $u < B(x, y)$, set $x := y$; else, keep x unchanged.
5. Output x and go back to step 2.

Works even if f and g are only known up to constant factors.

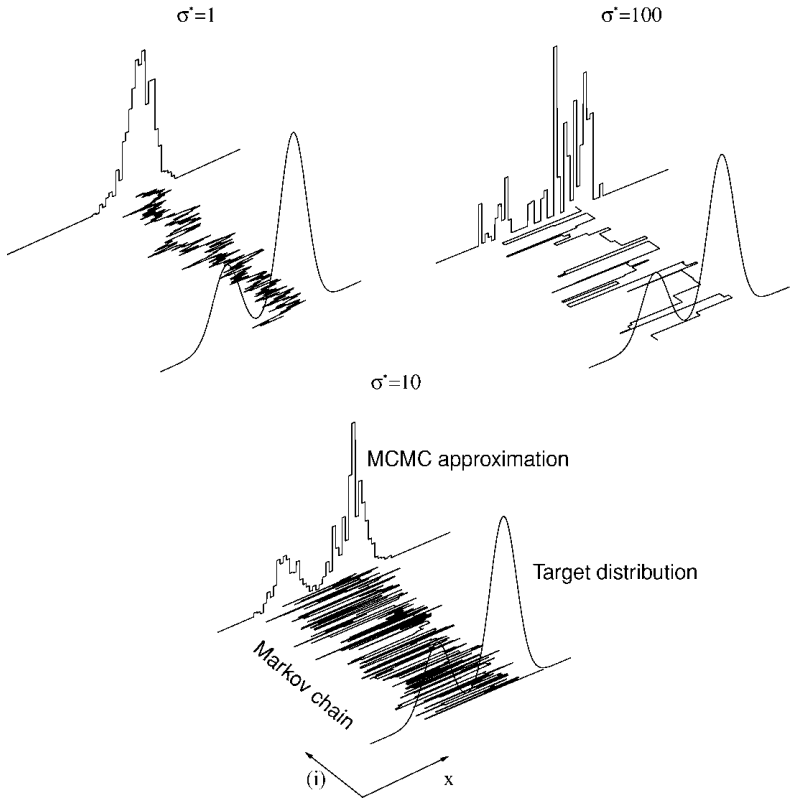
- The Metropolis algorithm was first published in an 1953 paper *Equation of State Calculations by Fast Computing Machines* by Metropolis, Rosenbluth, Augusta Teller, and Edward Teller. Hastings' addition was in 1970.

- Special case: $g(x, y) = g(y, x)$, then $B(x, y) = f(y)/f(x)$; this is the original Metropolis algorithm.

  In this case, the algorithm has the following interpretation: If the density at the proposal y is greater than at the old value x (so $B(x, y) > 1$), we move to y. So we try to optimize for the density value. But, occasionally, we also accept proposals which lower the density.

- Example: $g(\cdot, x) = N(x, \sigma^2)$.

- To obtain practically useful methods even for "bad" densities f, one has to choose the proposal density g properly. See for instance `http://jtobin.ca/flat-mcmc/`.



The plot shows the effects of different proposal distributions (normal with different variances). Notice that occasionally, the next sample equals the previous one.

Figure stolen from `http://www.cs.princeton.edu/courses/archive/spr06/cos598C/papers/AndrieuFreitasDoucetJordan2003.pdf`, page 18.

- Set $A(x, y) := \min\{1, B(x, y)\}$.
- Transition matrix (really, kernel):

$$T(x, y) = \hat{g}(y, x)A(x, y) + \delta(x - y) \int (1 - A(x, z))\hat{g}(z, x)\, dz.$$

  Here, $\hat{g}(\cdot, x)$ denotes the normalization of $g(\cdot, x)$ and δ denotes the Dirac distribution.

- Detailed balance condition (for $x \neq y$):

$$f(x)T(x, y) = \min\{f(x)\hat{g}(y, x),\, f(y)\hat{g}(x, y)\} = f(y)T(y, x).$$

Evaluating integrals

How can we evaluate integrals

$$\int a(x)\, f(x)\, dx,$$

where f is a density on a high-dimensional domain?

- $\int_a^b$: standard numerical quadrature
- $\int_{-\infty}^{\infty}$: numerical quadrature after coordinate change
- $\int_{\mathbb{R}^n}$: iterated numerical quadrature

Evaluating integrals

How can we evaluate integrals

$$\int a(x)\, f(x)\, dx,$$

where f is a density on a high-dimensional domain?

- $\int_a^b$: standard numerical quadrature
- $\int_{-\infty}^{\infty}$: numerical quadrature after coordinate change
- $\int_{\mathbb{R}^n}$: iterated numerical quadrature

These techniques sample the domain uniformly and require many evaluations of the integrand.

- Evaluation of such integrals is, of course, important in Bayesian learning and elsewhere.
- Note that adaptive numerical quadrature rules do exist.

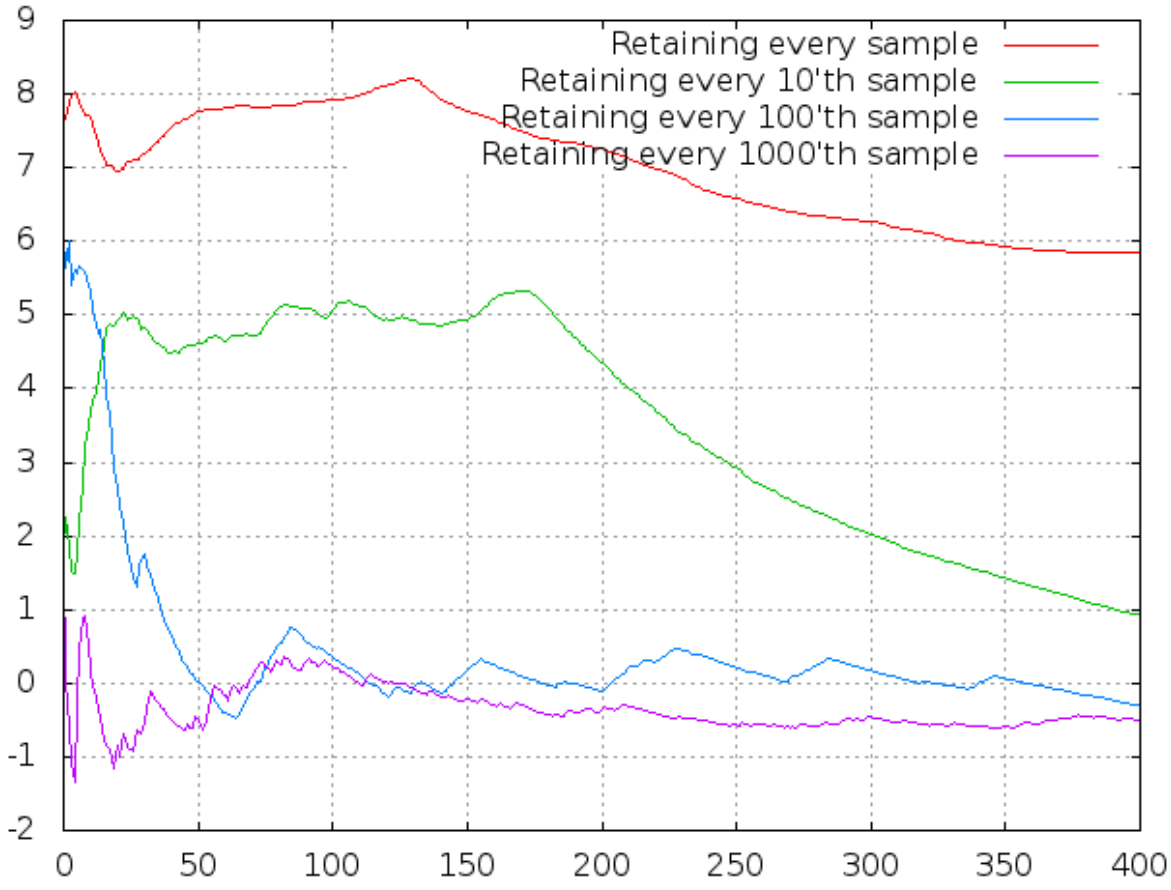
The Monte Carlo approach

Draw indep. samples $x_1, \ldots, x_N$ from f and approximate

$$f \approx \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}, \qquad I := \int a(x)\, f(x)\, dx \approx \frac{1}{N} \sum_{i=1}^{N} f(x_i) =: I_N.$$

- $E(I_N) = I$.
- $\mathrm{Var}(I_N) = \mathrm{Var}_f(a)/N$.
- $I_N \longrightarrow I$ almost surely (strong law of large numbers).

To sample f, use Markov chain techniques; obtain MCMC methods. These made Bayesian ideas useful in practice.



The plot shows the convergence speed to the true integral value (0) with naive MCMC sampling (using a normal proposal distribution). For details, see the Haskell code.

Resources

- Wikipedia gives a good first overview:
  http://en.wikipedia.org/wiki/Markov_chain
- Lecture notes on Markov chains:
  http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/Chapter11.pdf
  http://www.statslab.cam.ac.uk/~rrw1/markov/M.pdf
  http://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf
- Survey on MCMC methods:
  http://www.cs.princeton.edu/courses/archive/spr06/cos598C/papers/AndrieuFreitasDoucetJordan2003.pdf

Image sources

- Title illustration: Carina Willbold
- http://shetall.files.wordpress.com/2013/11/snakes_and_ladders1.jpg
- https://www.facebook.com/4funsociety/photos/a.176890522366207.59715.176890359032890/594923673896221/