Lectures 19: Markov Chain Monte Carlo II.

probability distribution and Markov chain

from Lecture 18

Monte Carlo importance sampling and Markov chain

If a configuration in phase space is denoted by X, the probability for configuration according to Boltzman is

$$\rho(X) \propto e^{-\beta E(X)} \qquad \beta = \frac{1}{T} \tag{1}$$

How to sample over the whole phase space for a general problem? How to generate configurations?

- **Brute force**: generate a truly random configuration X and accept it with probability $e^{-\beta E(X)}$ where all $E > 0$. Successive X are **statistically independent**. *VERY INEFFICIENT*
- **Markov chain**: Successive configurations X_i, X_{i+1} are **NOT** statistically independent but are distributed according to Boltzman distribution.

What is the difference between Markov chain and uncorrelated sequence?

- *Truly random* or *uncorrelated* sequence of configurations satisfies the identity

$$P(X_1, X_2, \cdots, P_{X_N}) = P_1(X_1)P_1(X_2)\cdots P_1(X_N)$$

probability distribution and Markov chain

from Lecture 18

- *Markov chain* satisfies the equation

$$P(X_1, X_2, \cdots, P_{X_N}) = P_1(X_1)T(X_1 \to X_2)T(X_2 \to X_3)\cdots T(X_{N-1} \to X_N)$$

where the transition probabilities $T(X \to X')$ are normalized

$$\sum_{X'} T(X \to X') = 1$$

We want to generate Markov chain where distribution of states is proportional to $e^{-\beta E(X)}$ and this distribution should be independent of the position within the chain and independent of the initial configuration.

The necessary conditions for generating such Markov chain is that every configuration in phase space should be accesible from any other configuration within finite number of steps (*connectedness* or *irreducibility*) - (Be careful to check this condition when choosing Monte Carlo step!)

We need to find transition probability $T(X \to X')$ which leads to a given **stationary** distribution $\rho(X)$ (in this case $\rho(X) \propto e^{-\beta E(X)}$).

Markov chain detailed balance

from Lecture 18

The probability for X decreases, if system goes from X to any other X': $-\sum_{X'} \rho(X)T(X \to X')$ and increases if X configuration is visited from any other state X': $\sum_{X'} \rho(X')T(X' \to X)$. The change of probability for X is therefore

$$\rho(X, t+1) - \rho(X, t) = -\sum_{X'} \rho(X)T(X \to X') + \sum_{X'} \rho(X')T(X' \to X) \quad (2)$$

We look for stationary solution, i.e., $\rho(X, t+1) - \rho(X, t) = 0$ and therefore

$$\sum_{X'} \rho(X)T(X \to X') = \sum_{X'} \rho(X')T(X' \to X) \quad (3)$$

General solution of this equation is not accesible, but a particular solution is obvious

$$\rho(X)T(X \to X') = \rho(X')T(X' \to X) \quad (4)$$

This solution is called **DETAIL BALANCE** solution.

Metropolis update

from Lecture 18

To construct algorithm, we devide transition prob. $T(X \to X') = \omega_{XX'} A_{XX'}$:

- *trial step probability* $\omega_{XX'}$ which is symmetric, i.e., $\omega_{XX'} = \omega_{X'X}$ (for example spin flip in ising: $\omega_{XX'}$ is $1/L^2$ if X and X' differ for a single spin flip and zero otherwise) and
- *acceptance probability* $A_{XX'}$ (for example accepting of rejecting new configuration with probability proportional to $min(1, \exp(-\beta(E(X') - E(X))))$.

Detail balance condition becomes

$$\frac{\rho(X')}{\rho(X)} = \frac{A_{XX'}}{A_{X'X}}$$

Metropolis chooses

$$\begin{aligned} A_{XX'} &= 1 \qquad \text{if } \rho(X') > \rho(X) \\ A_{XX'} &= \tfrac{\rho(X')}{\rho(X)} \quad \text{if } \rho(X') < \rho(X). \end{aligned} \tag{5}$$

Obviously, this acceptance probability satisfies detail balance condition and therefore leads to desired Markov chain with stationary probability for any configuration $X \propto \rho(X)$ for long times.

Metropolis update

from Lecture 18

To summarize Metropolis algorithm

- $T(X \to X') = \omega_{XX'} A_{XX'}$
- $\sum_{X'} \omega_{XX'} = 1; \omega_{XX'} = \omega_{X'X}$
- $\omega_{XX'} > 0$ for all X, X' after finite number of steps
- $A_{XX'} = min(1, \frac{\rho(X')}{\rho(X)})$

How to accept a step with probability $A_{XX'} < 1$? One can generate a random number $r \in [0, 1]$ and accept the step if $r < A_{XX'}$.

Keep in mind:

- Configurations that are generated by Markov chain are correlated. The theory guarantees that we arrive at invariant distribution ρ for long times.
- Two configurations are statistically independent only if they are far apart in the Markov chain. This distance is called *correlation time* (*Be careful: To meassure distance in Markov chain, every step counts, not only successful.*)

Monte Carlo averages

The average of any quantity can be calculated as usual

$$\overline{A} = \frac{1}{n - n_0} \sum_{i>n_0}^{n} A_i$$

where n_0 steps are used to ”warm-up”.

The error of the quantity, however, is much bigger than the following quantity

$$\frac{1}{n - n_0} \sum_{i>n_0}^{n} (A_i - \overline{A})^2$$

Imagine the extreme limit of correlations when all values A_i are the same. We would estimate that standard deviation is zero regardless of the actual error!

To compute standard deviation, we need to group meassurements within the correlation time into bins and than estimate the standard deviation of the bins:

$$B_l = \frac{1}{N_0} \sum_{i=N_l}^{i<N_l+N_0} A_i \tag{6}$$

Monte Carlo averages

$$\sigma^2 = \frac{1}{M}\sum_{j=0}^{M-1}(B_j - \overline{A})^2 \tag{7}$$

where we took into account that $\overline{A} = \overline{B}$. The correlation time (here denoted by N_0) is not very easy to estimate. Maybe the best algorithm is to compute σ^2 for few different N_0 and as long as σ^2 is increasing with N_0, the correlation time is still larger than N_0. When σ^2 stops changing with increasing N_0, we reached correlation time and σ^2 is a good estimation of standard deviation.

Metropolis-Hastings

Metropolis-Hastings algorithm:

Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953), Hastings (1970)

Pick more or less any "proposal distribution" $q(\mathbf{x}_2|\mathbf{x}_1)$
(A multivariate normal centered on $\mathbf{x}_1$ is a typical example.)

Then the algorithm is:

1. Generate a candidate point $\mathbf{x}_{2c}$ by drawing from the proposal distribution around $\mathbf{x}_1$

2. Calculate an "acceptance probability" by

$$\alpha(\mathbf{x}_1, \mathbf{x}_{2c}) = \min\left(1, \frac{\pi(\mathbf{x}_{2c})\, q(\mathbf{x}_1|\mathbf{x}_{2c})}{\pi(\mathbf{x}_1)\, q(\mathbf{x}_{2c}|\mathbf{x}_1)}\right)$$

Notice that the q's cancel out if symmetric on arguments, as is a multivariate Gaussian

3. Choose $\mathbf{x}_2 = \mathbf{x}_{2c}$ with probability α, $\mathbf{x}_2 = \mathbf{x}_1$ with probability $(1-\alpha)$

So, $p(\mathbf{x}_2|\mathbf{x}_1) = q(\mathbf{x}_2|\mathbf{x}_1)\,\alpha(\mathbf{x}_1, \mathbf{x}_2), \qquad (\mathbf{x}_2 \neq \mathbf{x}_1)$

Metropolis-Hastings

Proof:

$$\alpha(\mathbf{x}_1, \mathbf{x}_{2c}) = \min\left(1, \frac{\pi(\mathbf{x}_{2c})\, q(\mathbf{x}_1|\mathbf{x}_{2c})}{\pi(\mathbf{x}_1)\, q(\mathbf{x}_{2c}|\mathbf{x}_1)}\right)$$

So,

$$\begin{aligned}\pi(\mathbf{x}_1)\, q(\mathbf{x}_2|\mathbf{x}_1)\, \alpha(\mathbf{x}_1, \mathbf{x}_2) &= \min[\pi(\mathbf{x}_1)\, q(\mathbf{x}_2|\mathbf{x}_1),\ \pi(\mathbf{x}_2)\, q(\mathbf{x}_1|\mathbf{x}_2)] \\ &= \min[\pi(\mathbf{x}_2)\, q(\mathbf{x}_1|\mathbf{x}_2),\ \pi(\mathbf{x}_1)\, q(\mathbf{x}_2|\mathbf{x}_1)] \\ &= \pi(\mathbf{x}_2)\, q(\mathbf{x}_1|\mathbf{x}_2)\, \alpha(\mathbf{x}_2, \mathbf{x}_1)\end{aligned}$$

But

$$p(\mathbf{x}_2|\mathbf{x}_1) = q(\mathbf{x}_2|\mathbf{x}_1)\, \alpha(\mathbf{x}_1, \mathbf{x}_2), \qquad (\mathbf{x}_2 \neq \mathbf{x}_1)$$

and also the other way around

So,

$$\pi(\mathbf{x}_1)\, p(\mathbf{x}_2|\mathbf{x}_1) = \pi(\mathbf{x}_2)\, p(\mathbf{x}_1|\mathbf{x}_2)$$

which is just detailed balance, q.e.d.

Gibbs sampling

The **Gibbs Sampler** is an interesting special case of Metropolis-Hasting

A "full conditional distribution" of $\pi(\mathbf{x})$ is the normalized distribution obtained by sampling along one coordinate direction (i.e. "drilling throug the full distribution. We write it as $\pi(x \mid \mathbf{x}^{-})$.

"given all coordinate values except one"

Theorem: A multivariate distribution is uniquely determined by all of its full conditional distributions.
Proof (sort-of): It's a hugely overdetermined set of linear equations, so any degeneracy is infinitely unlikely!

Metropolis-Hastings along one direction looks like this:

$$\alpha(x_1, x_{2c}|\mathbf{x}^{-}) = \min\left(1, \frac{\pi(x_{2c}|\mathbf{x}^{-})\, q(x_1|x_{2c}, \mathbf{x}^{-})}{\pi(x_1|\mathbf{x}^{-})\, q(x_{2c}|x_1, \mathbf{x}^{-})}\right)$$

Choose the proposal distribution $q(x_2|x_1, \mathbf{x}^{-}) = \pi(x_2|\mathbf{x}^{-})$
Then we always accept the step!

But a proposal distribution must be normalized, so we actually do need to be able to calculate $\int \pi(x|\mathbf{x}^{-})dx$ but only along one "drill hole" at a time!

Gibbs sampling

So, Gibbs sampling looks like this:

- Cycle through the coordinate directions of **x**
- Hold the values of all the other coordinates fixed
- "Drill", i.e., sample from the one dimensional distribution along the non-fixed coordinate.
  - this requires knowing the normalization, if necessary by doing an integral or sum along the line
- Now fix the coordinate at the sampled value and go on to the next coordinate direction.