Math 456: Mathematical Modeling

Tuesday, April 9th, 2018

The Ergodic theorem

Tuesday, April 9th, 2018

Today

1. Asymptotic frequency (or: How to use the stationary distribution to estimate the average amount of time a chain lies in a given state.)
2. The Ergodic Theorem.

Warm up

The law of large numbers

A core fact in probability is the **law of large numbers**, which we now recall.

Consider an infinite sequence of variables $Y_1, Y_2, Y_3 \dots$

Assume they are independent, identically distributed, and

$$\mathbb{E}[Y_1] = \mu, \;\; \mathbb{E}[|Y_1 - \mu|^2] = \sigma^2 < \infty.$$

Warm up

The law of large numbers

A core fact in probability is the **law of large numbers**, which we now recall.

Consider an infinite sequence of variables $Y_1, Y_2, Y_3 \dots$

Assume they are independent, identically distributed, and

$$\mathbb{E}[Y_1] = \mu, \;\; \mathbb{E}[|Y_1 - \mu|^2] = \sigma^2 < \infty.$$

Then,

$$\mathbb{P}\left(\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} Y_k = \mathbb{E}[Y_1] \right) = 1$$

Warm up

The law of large numbers

This law is very important because it shows how the mathematics we have built leading up to this law share our intuitive ideas of probability.

To be concrete: if you perform an experiment where outcome A happens with probability 0.25, then upon performing the experiment a large number N of times, we expect for A to occur $\sim N/4$ times.

Asymptotic frequency of visits

Fix a chain $X_1, X_2, X_3, \ldots$. Consider the random variables

$$N_n(y) = \#\{k \mid 1 \leq k \leq n \text{ and } X_k = y\}$$

which give the **number of visits to a state** y **up to time** n.

Asymptotic frequency of visits

Fix a chain $X_1, X_2, X_3, \ldots$. Consider the random variables

$$N_n(y) = \#\{k \mid 1 \leq k \leq n \text{ and } X_k = y\}$$

which give the **number of visits to a state** y **up to time** n.

Theorem

For an irreducible chain we have that

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{N_n(y)}{n} = \frac{1}{\mathbb{E}_y[T_y]}\right) = 1$$

Asymptotic frequency of visits

Proof

Fix $y \in S$, let T_y^k denote the **time of the k-th visit to y**, and consider the sequence of random variables

$$Y_k := T_y^k - T_y^{k-1}, \;\; k \geq 1, \; Y_1 := T_y^1.$$

Asymptotic frequency of visits

Proof

Fix $y \in S$, let T_y^k denote the **time of the k-th visit to y**, and consider the sequence of random variables

$$Y_k := T_y^k - T_y^{k-1}, \;\; k \geq 1, \; Y_1 := T_y^1.$$

By the Strong Markov Property, the sequence $Y_1, Y_2, \ldots$ is made out of independent, identically distributed random variables.

Asymptotic frequency of visits

Proof

Fix $y \in S$, let T_y^k denote the **time of the k-th visit to y**, and consider the sequence of random variables

$$Y_k := T_y^k - T_y^{k-1}, \;\; k \geq 1, \; Y_1 := T_y^1.$$

By the Strong Markov Property, the sequence $Y_1, Y_2, \ldots$ is made out of independent, identically distributed random variables.

Therefore,

$$\mathbb{P}_y \left(\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} Y_k = \mathbb{E}_y[T_y] \right) = 1$$

Asymptotic frequency of visits

Proof

Written in terms of T_y^k

$$\mathbb{P}_y\left(\lim_{n\to\infty}\frac{T_y^n}{n}=\mathbb{E}_y[T_y]\right)=1$$

Now, a moment of reflection (and a drawing) shows that

$$T_y^{N_n}\leq n\leq T_y^{N_n+1}$$

for every n.

Asymptotic frequency of visits

Proof

Written in terms of T_y^k

$$\mathbb{P}_y\left(\lim_{n\to\infty}\frac{T_y^n}{n}=\mathbb{E}_y[T_y]\right)=1$$

Now, a moment of reflection (and a drawing) shows that

$$T_y^{N_n}\le n\le T_y^{N_n+1}$$

for every n. Dividing all sides by N_n, we have

$$\frac{T_y^{N_n}}{N_n}\le\frac{n}{N_n(y)}\le\frac{T_y^{N_n+1}}{N_n+1}\frac{N_n+1}{N_n}$$

Asymptotic frequency of visits

Proof.

Considering that

- $N_n \to \infty$ as $n \to \infty$,
- $\frac{n}{N_n(y)}$ lies in between two sequences having the same limit,

we conclude that

$$\mathbb{P}_y \left(\lim_{n \to \infty} \frac{N_n(y)}{n} = \frac{1}{\mathbb{E}_y[T_y]} \right) = 1$$

and the theorem is proved.

□

Asymptotic frequency of visits

Putting it all together

If one's goal is to estimate $N_n(y)/n$, then the previous theorem is of no use if we cannot compute $\mathbb{E}_y[T_y]$ for every state y.

Theorem (Durrett, p. 50, Theorem 1.22)

For an irreducible chain, we have

$$\mathbb{E}_y[T_y] = \frac{1}{\pi(y)} \ \forall\, y \in S.$$

In **particular**, the stationary distribution encodes what percentage of the time is the chain in each of the states, so that $N_n(y)/n \approx \pi(y)$ for large enough n.

Asymptotic frequency of visits

Putting it all together

Proof.

Take the chain with initial distribution given by π itself. Then,

$$\mathbb{P}(X_n = y) = \pi(y) \;\; \forall\, n, \; \forall\, y \in S.$$

On the other hand $N_n(y)$ is equal to $\sum_{k=1}^{n} \chi_{\{X_k = y\}}$, so

$$\mathbb{E}[N_n(y)] = \sum_{k=1}^{n} \mathbb{P}(X_k = y) \Rightarrow \mathbb{E}[N_n(y)] = \sum_{k=1}^{n} \pi(y) = n\pi(y)$$

Then, previous theorem yields

$$\pi(y) = 1/\mathbb{E}_y[T_y]$$

□

Examples

For any problem involving **computing the time spent in a given state**, we proceed as follows:

- Verify the chain is irreducible.

Examples

For any problem involving **computing the time spent in a given state**, we proceed as follows:

- Verify the chain is irreducible.
- Find its stationary distribution.

Examples

For any problem involving **computing the time spent in a given state**, we proceed as follows:

- Verify the chain is irreducible.
- Find its stationary distribution.
- Use the asymptotic frequency theorem.

Examples

Reflective Random walk

Problem: Take the chain with transition probability matrix

$$\mathrm{p} = \begin{pmatrix} 1/3 & 2/3 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 0 & 0 & 1/3 & 2/3 \end{pmatrix}$$

Find $\lim \mathrm{p}^n(x, y)$, and estimate how often the chain occupies each state after a large number of steps.

Examples

Reflective Random walk

Solution:

$$\mathrm{p} = \begin{pmatrix} 1/3 & 2/3 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 0 & 0 & 1/3 & 2/3 \end{pmatrix}$$

- Is the chain irreducible?

Examples

Reflective Random walk

Solution:

$$\mathrm{p} = \begin{pmatrix} 1/3 & 2/3 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 0 & 0 & 1/3 & 2/3 \end{pmatrix}$$

- Is the chain irreducible? **Answer:** yes.

- Is the chain aperiodic?

Examples

Reflective Random walk

Solution:

$$p = \begin{pmatrix} 1/3 & 2/3 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 0 & 0 & 1/3 & 2/3 \end{pmatrix}$$

- Is the chain irreducible? **Answer:** yes.

- Is the chain aperiodic? **Answer:** yes, note that $p(1,1) > 0$, so this state has period 1, which by irreducibility means all states have period 1.

Examples

Reflective Random walk

Solution:

$$\mathrm{p} = \begin{pmatrix} 1/3 & 2/3 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 0 & 0 & 1/3 & 2/3 \end{pmatrix}$$

• By the convergence theorem, and the "ergodic theorem", all we need to do is solve the eigenfunction system to determine $\pi(y)$. Doing so yields the vector

$$\pi^t = (\tfrac{1}{15}, \tfrac{2}{15}, \tfrac{4}{15}, \tfrac{8}{15})$$

Examples

Reflective Random walk

Solution:

$$\mathrm{p} = \begin{pmatrix} 1/3 & 2/3 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 0 & 0 & 1/3 & 2/3 \end{pmatrix}$$

- How often does the chain occupy each state? **Answer:** Since,

$$\pi^t = (\tfrac{1}{15}, \tfrac{2}{15}, \tfrac{4}{15}, \tfrac{8}{15}),$$

the system spends about 1/15 of the time in state $x = 1$, about 2/15 of the time in state $x = 2$, about 4/15 of the time in state $x = 4$, and finally, about 8/15 of the time (which is more than half) in state $x = 4$.

Ergodic Dynamical Systems

Erdogic

ergon (work) *odos* (path)

The term *ergodic* was introduced by Ludwig Boltzmann, in his attempts at understanding the behavior of molecules in a gas, ultimately founding the field of **statistical mechanics**.

Today, the adjective **ergodic** is used in a dynamical system whenever it has the following property:

The average of any quantity over a long period of time equals the average of the quantity over the state space

Ergodic Dynamical Systems

The average of any quantity over a long period of time equals the average of the quantity over the state space

Note, however, the above statement is a big vague: there are many ways of “averaging over the state space”.

Heuristically, for this to happen, every trajectory of the system must cover the entire state space, and must do so according to a some distribution –this distribution is the invariant measure.

Ergodicity For Markov Chains

Consider a chain X_n with state space S.

As $x \in S$ represents a possible states of our system, a function

$$f : S \mapsto \mathbb{R}$$

represents a quantity depending on the state, presumably, a numerical quantity of interest that may be measured, and the value of which we may want to predict.

(e.g. for the Gambler's ruin, $f(x) = N - x$, the \$'s left to win)
(e.g. for certain physical systems, quantities like temperature)

Ergodicity For Markov Chains

For a fixed function $f : S \mapsto \mathbb{R}$, the sequence of (random) values

$$f(X_1),\ f(X_2), \ldots, f(X_n),\ \ldots$$

correspond to the (random) values of f as the chain evolves.

Ergodicity For Markov Chains

For a fixed function $f : S \mapsto \mathbb{R}$, the sequence of (random) values

$$f(X_1),\ f(X_2), \ldots, f(X_n),\ \ldots$$

correspond to the (random) values of f as the chain evolves.

For instance, if the initial state of the chain is random, and given by a stationary distribution, then the random variables $f(X_1), f(X_2), \ldots$ all have the same distribution, and are independent.

Ergodicity For Markov Chains

However, if the $f(X_1), f(X_2), \ldots$ are i.i.d. variables, the **strong law of large numbers** says that with probability 1

$$\frac{1}{n} \sum_{k=1}^{n} f(X_k) \to \mathbb{E}[f(X_1)] \text{ as } n \to \infty$$

Ergodicity For Markov Chains

However, if the $f(X_1), f(X_2), \ldots$ are i.i.d. variables, the **strong law of large numbers** says that with probability 1

$$\frac{1}{n} \sum_{k=1}^{n} f(X_k) \to \mathbb{E}[f(X_1)] \text{ as } n \to \infty$$

What is $\mathbb{E}[f(X_1)]$ in this case?

Ergodicity For Markov Chains

However, if the $f(X_1), f(X_2), \ldots$ are i.i.d. variables, the **strong law of large numbers** says that with probability 1

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \to \mathbb{E}[f(X_1)] \text{ as } n \to \infty$$

What is $\mathbb{E}[f(X_1)]$ in this case? Well

$$\mathbb{E}[f(X_1)] = \sum_{y \in S} f(y)\pi(y)$$

Ergodicity For Markov Chains

We have that a long time time average equals a spatial average

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \to \sum_{y\in S} f(y)\pi(y)$$

This is an instance of ergodicity.

Ergodicity For Markov Chains

We have that a long time time average equals a spatial average

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \to \sum_{y\in S} f(y)\pi(y)$$

This is an instance of ergodicity.
What if X_0 is not distributed by π?

The Ergodic Theorem

Theorem

Consider an irreducible chain and let $\pi(y)$ *denote its stationary distribution.*

Then, for any $f : S \mapsto \mathbb{R}$*, we have with probability* 1

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \sum_{y \in S} f(y)\pi(y)$$

The Ergodic Theorem

Proof.

Fix n, then the sum

$$\sum_{k=1}^{n} f(X_k)$$

can be reorganized as

$$\sum_{y \in S} \sum_{k=1}^{n} f(y) \chi_{\{X_k = y\}}$$

□

The Ergodic Theorem

Proof

It follows that

$$\sum_{k=1}^{n} f(X_k) = \sum_{y \in S} f(y) N_n(y)$$

Therefore, for the average we have

$$\frac{1}{n} \sum_{k=1}^{n} f(X_k) = \sum_{y \in S} f(y) \frac{N_n(y)}{n}$$

The Ergodic Theorem

Proof.

The chain is irreducible and aperiodic, so for every y, we have

$$\lim_{n\to\infty} \frac{N_n(y)}{n} = \pi(y),$$

thanks to the convergence theorem.

The Ergodic Theorem

Proof.

The chain is irreducible and aperiodic, so for every y, we have

$$\lim_{n\to\infty} \frac{N_n(y)}{n} = \pi(y),$$

thanks to the convergence theorem. Then,

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(X_k) &= \sum_{y\in S} f(y)\left(\lim_{n\to\infty} \frac{N_n(y)}{n}\right)\\ &= \sum_{y\in S} f(y)\pi(y)\end{aligned}$$

□

Next

The Metropolis Algorithm

Next class, we will talk about a procedure which flips the ergodic theorem: we will want to compute a certain distribution π, and we are going to use a Markov chain **to approximate it by sampling paths from the chain**.

Math 456: Mathematical Modeling

Thursday, April 11th, 2018

Monte Carlo methods: overview, Metropolis-Hastings, and simulated annealing

Thursday, April 11th, 2018

Warm up

Let us suppose you want to create a numerical procedure to compute some integral

$$\int_0^1 h(x)\, dx$$

for a very complicated function h.

Suppose the function $h(x)$ can actually be decomposed as

$$h(x) = f(x)p(x)$$

where p is a probability distribution that we can **sample from** easily, and $f(x)$ is a function we can approximate reasonably well.

Warm up

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables whose distribution is $p(x) \ldots$

Then, the expected value formula says that the common mean of these variables

$$\mathbb{E}[f(X_i)] = \int_0^1 f(x)p(x)\ dx$$

Warm up

Here is the original Monte Carlo method:

It consists in taking a particular sample $x_1, x_2, x_3 \ldots$ of the sequence $X_1, X_2, X_3 \ldots$ and take the sum

$$\frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

This sum is used to estimate

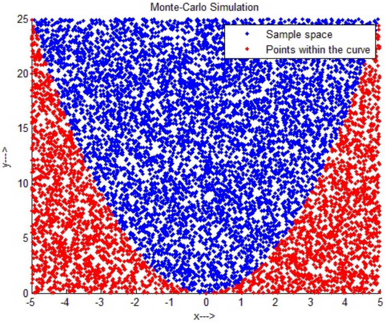
$$\int_0^1 f(x)p(x)\; dx$$

Warm up

A different way of thinking of Monte Carlo

Say that h only takes values in $[0, 1]$

$$\int_0^1 h(x)\, dx = \text{Area}\Big\{(x, y) \mid 0 \le x \le 1,\ 0 \le y \le h(x)\Big\}$$



Warm up

Generate n random points (x_i, y_i) which are i.i.d. and sampled from the uniform distribution in the square $[0, 1]^2$.

Let M_n denote the the number of points such that

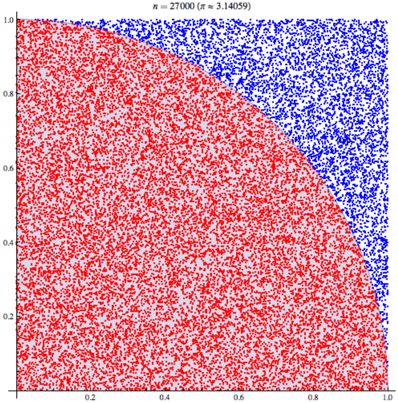
$$y_i \leq f(x_i)$$

Then,

$$\frac{M_n}{n} \approx \int_0^1 h(x)\, dx$$

Warm up

This is a well known way to generate a numerical approximation to π



Gibbs distributions

(see also *canonical ensemble*)

$$\mathbb{P}[X = x] = \frac{1}{Z_\beta} e^{-\beta E(x)}, \quad Z_\beta = \sum_x e^{-\beta E(x)}$$

or, for continuous variables:

$$\mathbb{P}[X \in A] = \int_A \frac{1}{Z_\beta} e^{-\beta E(x)} \, dx$$

where

$$Z_\beta = \int e^{-\beta E(x)} \, dx$$

Gibbs distributions

The most famous case is of course the Normal distribution

$$E(x) = |x|^2$$

In which case,

$$Z_\beta = (4\pi\beta^{-1})^{\frac{d}{2}}$$

Gibbs distributions

In general, computing the normalization constant Z_β is too difficult, there is neither a simple formula or a practical algorithm to approximate it directly.

Example: (Simulated annealing)
Suppose you want to find a minimizer x_0 for a function $E(x)$. Then one thing you can do is take a sample from the distribution

$$\frac{1}{Z_\beta} e^{-\beta E(x)}$$

Where β is a very large number.

The Ising Model

The Ising Model

Let us now talk, about the Ising model ▸ Link.

Let $\Lambda = \{(x,y) \in \mathbb{Z}^2 \mid |x| \leq L,\ |y| \leq L\}$, $L \in \mathbb{N}$.

We think of Λ as a graph with periodic conditions at the boundary structure, i.e. $(-L,0)$ is a neighbor of $(0,L)$, (L,L) is a neighbor of $(L,-L)$ and $(-L,L)$, and so on.

Gibbs distribution

The Ising Model

Consider S, the set of all functions in Λ with values ± 1

$$\xi : \Lambda \mapsto \{-1, +1\}$$

The set S has $2^{(2L+1)^2}$ elements.
A generic element $\xi \in S$ may be pictured as follows

$$\begin{array}{cccccc}
+ & + & - & + & - & + \\
- & - & - & + & + & - \\
- & + & + & + & - & + \\
+ & + & - & - & - & - \\
- & - & + & - & + & + \\
+ & + & - & + & - & -
\end{array}$$

Gibbs distribution

The Ising Model

We will write $x \sim y$ for any two elements $x, y \in \Lambda$ if they are neighbors in the square lattice $\mathbb{Z}^2$.

Then, the **energy** of $\xi \in S$ is defined as

$$H(\xi) = -\sum_{x \sim y} \xi(x)\xi(y)$$

Gibbs distribution

The Ising Model

$$H(\xi) = -\sum_{x\in\Lambda}\sum_{y:x\sim y}\xi(x)\xi(y)$$

Note: Since $\xi(x) \in \{\pm 1\}$ for all x, and every $x \in \Lambda$ has exactly 4 neighbors, the lowest possible energy of H is $-4L$, achieved when all the values are equal, that is

$$\xi \equiv 1 \text{ or } \xi \equiv -1$$

Gibbs distribution

The Ising Model

Given $\xi \in S$, we denote by ξ^x the function which is identical to ξ, except at ξ, where it's value is flipped.

Gibbs distribution

The Ising Model

Given $\xi \in S$, we denote by ξ^x the function which is identical to ξ, except at ξ, where it's value is flipped.

We construct a Markov chain in S, depending on a parameter $\beta > 0$, as follows: first, we choose a site $x \in \Lambda$ at random, all sites having the same probability $(2L+1)^{-2}$.

Gibbs distribution

The Ising Model

Second, if x is the chosen site, we do one of two things: i) we move to the system to the state ξ^x, or ii) we stay where we are.

To decide this, we flip a biased coin, and move to ξ^x if we get tails. The probability of tails for this coin is defined by

$$r(\xi, \xi^x) := \min \left\{ \frac{e^{-\beta H(\xi^x)}}{e^{-\beta H(\xi)}}, 1 \right\}$$

Gibbs distribution

The Ising Model

In conclusion, we **always** move to ξ^x if $H(\xi^x) \leq H(\xi)$, and we move with probability $e^{-\beta(H(\xi)-H(\xi^x))}$ otherwise.

The resulting transition probability is given by

$$\mathrm{p}(\xi, \xi^x) = \frac{1}{(2L+1)^2} \min\left\{ \frac{e^{-\beta H(\xi^x)}}{e^{-\beta H(\xi)}}, 1 \right\}$$

Question: Is this chain irreducible? is this chain aperiodic?

Gibbs distribution

The Ising Model

In conclusion, we **always** move to ξ^x if $H(\xi^x) \leq H(\xi)$, and we move with probability $e^{-\beta(H(\xi)-H(\xi^x))}$ otherwise.

The resulting transition probability is given by

$$\mathrm{p}(\xi, \xi^x) = \frac{1}{(2L+1)^2} \min \left\{ \frac{e^{-\beta H(\xi^x)}}{e^{-\beta H(\xi)}}, 1 \right\}$$

Question: Is this chain irreducible? is this chain aperiodic?
Answer: yes and yes.

Gibbs distribution

The Ising Model

$$\mathrm{p}(\xi, \xi^x) = \frac{1}{(2L+1)^2} \min \left\{ \frac{e^{-\beta H(\xi^x)}}{e^{-\beta H(\xi)}}, 1 \right\}$$

Question: What is the stationary distribution for this chain?

Gibbs distribution

The Ising Model

$$\mathrm{p}(\xi, \xi^x) = \frac{1}{(2L+1)^2} \min\left\{ \frac{e^{-\beta H(\xi^x)}}{e^{-\beta H(\xi)}}, 1 \right\}$$

Question: What is the stationary distribution for this chain?

Answer: (as we will see later today)

$$\pi(\xi) = \frac{1}{Z_\beta} e^{-\beta H(\xi)}$$

where Z_β is the normalization constant

$$Z_\beta = \sum_{\xi \in S} e^{-\beta H(\xi)}$$

Gibbs distributions

In general, computing the normalization constant Z_β is too difficult, there is neither a simple formula or a practical algorithm to approximate it directly.

So, how to sample from this distribution?

Monte Carlo and Metropolis-Hastings algorithms

Some key names

Stanislaw Ulam and John Von Neumann
(sometime in 1942-1945)

(Photo from Los Alamos archive)

Monte Carlo and Metropolis-Hastings algorithms

Some key names

Nicholas Metropolis, Arianna W. Rosenbluth (not pictured)
Augusta Teller, Edward Teller
(1950's paper)

(Photo from Los Alamos archive)
In joint paper, they combined Markov chains and Monte Carlo.

The Metropolis algorithm

1. We choose an auxiliar transition probability matrix $q(x, y)$, requiring it be symmetric $q(x, y) = q(y, x)$ and leading to an irreducible chain.

2. Choose an initial state x_0.

3. At stage n, we make a jump according to the matrix q. With y denoting the resulting jump, we flip a coin with

$$\mathbb{P}(\text{Heads}) = \min\left\{1, \frac{f(y)/K}{f(x)/K}\right\} = \min\left\{1, \frac{f(y)}{f(x)}\right\}.$$

If heads wins, we set $X_{n+1} = y$, otherwise, we don't move.

The Metropolis algorithm

We run this process up until some large number of steps N.

The distribution of X_N is given by $f(x)/K$. That is

$$\mathbb{P}[X_N = x] \approx f(x)/K.$$

In this manner, we have sampled from (essentially) the distribution $f(x)/K$.

The Metropolis algorithm

Why does this work?

We have a chain whose transition matrix is determined by

$$\mathrm{p}(x,y) = q(x,y)\min\left\{1, \frac{f(y)}{f(x)}\right\} \text{ for } x \neq y$$

(**Q:** what about the value of $\mathrm{p}(x,x)$?)

This chain is irreducible and is also aperiodic (there are some x such that $\mathrm{p}(x,x) > 0$ except if f is constant).

The Metropolis algorithm

Why does this work?

Here is what's really important:

$$f(x)\mathrm{p}(x,y) = f(y)\mathrm{p}(y,x)$$

voilá! The distribution $f(x)/K$ satisfies the *detailed balance condition* (remember that?) and is thus the stationary distribution of the chain with kernel $\mathrm{p}(x,y)$.

Thenconvergence theorem then says that **for very large** N,

$$\mathrm{p}^N(x,y) \approx f(y)$$

regardless of which state x and y we choose.

The Metropolis-Hasting Algorithm

We have a (finite, but possibly large) set S.

Consider a distribution π on S, i.e., a function $\pi : S \mapsto \mathbb{R}$ where

$$\begin{aligned} &0 < \pi(y) < 1 \ \text{ for all } y \in S, \\ &\text{and } \sum_{y \in S} \pi(y) = 1. \end{aligned}$$

The Metropolis-Hasting Algorithm

We have a (finite, but possibly large) set S.

Consider a distribution π on S, i.e., a function $\pi : S \mapsto \mathbb{R}$ where

$$\begin{aligned} &0 < \pi(y) < 1 \ \text{ for all } y \in S, \\ &\text{and } \sum_{y \in S} \pi(y) = 1. \end{aligned}$$

The Metropolis-Hasting algorithm is an extremely efficient way of producing approximations to the sum

$$\sum_{y \in S} f(y)\pi(y)$$

for any function $f : S \mapsto \mathbb{R}$ – and this works even in cases where we do not know the values of $\pi(y)$ directly or explicitly!!

The Metropolis-Hasting Algorithm

Algorithm:
Pick a Markov Chain over S, $X_1, X_2, \ldots$, irreducible and aperiodic, whose stationary distribution is π.
Then, run a simulation of the chain for a sufficiently long* time n, resulting in values $x_1, x_2, \ldots, x_n$, the average

$$\frac{1}{n}\sum_{k=1}^{n} f(x_k)$$

is the output of the algorithm.

* of course, how large n should be is decided case by case

The Metropolis-Hasting Algorithm

Algorithm (the most important part)
The Markov Chain is constructed as follows: choose any way you like an initial transition matrix $\mathrm{q}(x, y)$ in S.
For each x and y, define

$$r(x, y) := \min\left\{\frac{\pi(y)\mathrm{q}(y, x)}{\pi(x)\mathrm{q}(x, y)}, 1\right\}$$

Then, the chain we want is given by the transition matrix

$$\mathrm{p}(x, y) = r(x, y)\mathrm{q}(x, y).$$

The Metropolis-Hasting Algorithm

Why is this the chain we want?
Well, for this chain the following holds:

given any two states x and y, we have that

$$\pi(x)\mathrm{p}(x, y) = \pi(y)\mathrm{p}(x, y)$$

and this is a stronger property than π being stationary!

(We say π satisfies the *detailed balance condition* for the chain)

Application: Simulated Annealing

Finding minimizers

You are given a function $\ell : S \mapsto \mathbb{R}$, which you want to minimize. However, the set S is so large that it is not computationally feasible to find the minimum by evaluating $\ell(x)$ at every $x \in S$.

Application: Simulated Annealing

Finding minimizers

You are given a function $\ell : S \mapsto \mathbb{R}$, which you want to minimize. However, the set S is so large that it is not computationally feasible to find the minimum by evaluating $\ell(x)$ at every $x \in S$.

Idea: For a parameter $t > 0$, define the distribution π_t by

$$\pi_t(x) := \frac{1}{Z_t} e^{-\frac{1}{t}\ell(x)}, \quad Z_t := \sum_{x \in S} e^{-\frac{1}{t}\ell(x)}$$

Application: Simulated Annealing

Finding minimizers

$$\pi_t(x) := \frac{1}{Z_t} e^{-\frac{1}{t}\ell(x)}, \quad Z_t := \sum_{x \in S} e^{-\frac{1}{t}\ell(x)}$$

Note: the number $e^{-\frac{1}{t}\ell(x)}$ will be largest wherever $\ell(x)$ is close to its minimum, if t is very small, $\pi_t(x)$ should be close to zero if $\ell(x)$ is not close to the minimum value.

Therefore, we expect $\pi_t(x)$ to assign (for small t) a high probability to good guesses for the absolute minimum.

Application: Simulated Annealing

Finding minimizers

The chain then comes as follows

$$r(x,y) = \min\left\{ \frac{e^{-\frac{1}{t}\ell(y)}\mathrm{q}(y,x)}{e^{-\frac{1}{t}\ell(x)}\mathrm{q}(x,y)}, 1 \right\}$$

Here is the point: this can be computed without knowing Z_t!

Application: Simulated Annealing

Finding minimizers

The chain then comes as follows

$$r(x,y) = \min\left\{\frac{e^{-\frac{1}{t}\ell(y)}\mathrm{q}(y,x)}{e^{-\frac{1}{t}\ell(x)}\mathrm{q}(x,y)}, 1\right\}$$

Here is the point: this can be computed without knowing Z_t!

In practical terms, computing the value of Z_t would involve adding up $e^{-\frac{1}{t}\ell(x)}$ over all x in S, which would defeat the point of a random approach in the first place. This is why this algorithm was devised in the first place

Application: Simulated Annealing

The traveling salesman problem

This method is often applied to problems that cannot be tackled in polynomial time.

Problem: You are in charge of traveling from city to city in a state. The N cities are given some enumeration and their respective locations are $x_1, x_2, \ldots, x_N$.

This is known as the traveling salesman problem. ▸ Link

Application: Simulated Annealing

The traveling salesman problem

You must decide in **what order** to visit these cities, while minimizing the total distanced traveled.

What does this mean?
If you decide to visit the cities in an order given by $k_1, k_2, \ldots, k_N$, then the total distance you will travel is

$$\ell(k_1, k_2, \ldots, k_n) = |x_{k_1} - x_{k_2}| + |x_{k_2} - x_{k_3}| + \ldots + |x_{k_{N-1}} - x_{k_N}|$$

Problem: find $(k_1, k_2, \ldots, k_n)$ which minimizes ℓ.