Lecture 18: Monte Carlo and Markov Chains

36-350, Fall 2011

21 October 2013

Abstract

For Halloween, we come as a math course

1 Monte Carlo Integration

Suppose we want to evaluate a definite integral,

$$\int_D f(x)dx \tag{1}$$

where D is some domain (possibly all space), and f is our favorite function. For most functions, there is no closed-form expression for such definite integrals. Numerical analysis provides various means of approximating definite integrals, starting with **Euler's method** for one-dimensional integrals,

$$\int_a^b f(x)dx \approx \sum_{i=1}^{\lfloor (b-a)/h \rfloor} hf(a+\frac{i}{h}) \tag{2}$$

and getting arbitrarily more sophisticated. Unfortunately, these are slow, especially when x is really a high-dimensional vector; one ends up having to evaluate the function f at an exponentially growing number of points just to get the definite integral.

It turns out that designing nuclear weapons involves doing a lot of complicated integrals [2], and so one of the first uses of modern computing machines using an efficient but *random* approximation scheme, which the physicists involved called "the Monte Carlo method", after the European gambling resort. Recall that the expectation value of a function f with respect to a distribution whose probability density is p is

$$\mathbb{E}_p\left[f(X)\right] = \int f(x)p(x)dx \tag{3}$$

If $X_1, X_2, \ldots X_n$ are independent random variables with common density p, we say that they are **independent and identically distributed**, or IID. As you learned in introductory probability, the law of large numbers asserts that, for IID random variables, the sample mean converges on the expectation value:

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \rightarrow \mathbb{E}_p\left[f(X)\right] = \int f(x)p(x)dx \tag{4}$$

The most basic Monte Carlo method for evaluating the integral in (1) is to draw X uniformly over the domain D. If the total measure[1] of D is $|D|$, then the uniform density is $1/|D|$ on D, and 0 everywhere else, so

$$\int_D f(x)dx = |D| \int_D f(x)\frac{1}{|D|}dx \tag{5}$$

and

$$\frac{|D|}{n}\sum_{i=1}^{n} f(X_i) \rightarrow \int_D f(x)dx \tag{6}$$

How good is the approximation, i.e., how close are the two sides of Eq. 6? Because the X_i are IID, we can use another result you remember from introductory probability, the central limit theorem: when Y_i are IID with common mean μ and variance σ^2, the sample mean approaches a Gaussian distribution with mean μ and variance σ^2/n. Symbolically,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \rightsquigarrow \mathcal{N}(\mu, \frac{\sigma^2}{n}) \tag{7}$$

In this case, the role of Y_i is played by $f(X_i)$, so

$$\frac{|D|}{n}\sum_{i=1}^{n} f(X_i) \rightsquigarrow \mathcal{N}\left(\int_D f(x)dx, |D|^2\frac{\sigma_f^2}{n}\right) \tag{8}$$

with σ_f^2 being the variance of $f(X)$. Thus, the Monte Carlo estimate is unbiased (its expected value is equal to the truth), and its variance goes down like $1/n$. This is true no matter what the dimension of X might be. So, unlike the numerical integration schemes, reducing the error of the Monte Carlo estimate doesn't require exponentially many points. In fact, if we knew σ_f^2, we could us the known Gaussian distribution to give confidence intervals for $\int f(x)dx$. If, as is usually the case, we don't know that variance, we can always estimate it from the samples, and the Gaussian confidence intervals will become correct as $n \rightarrow \infty$, or we can use corrections (based on, say, the t distribution) familiar from basic statistics.

There are many situations in which this basic recipe is impractical or even impossible. For instance, D may have a very complicated shape, making it hard to draw samples uniformly. It may then be possible to find some larger region, say C, which contains D and for which we can generate uniform samples. Then, because[2]

$$\int_D f(x)dx = \int_C f(x)\mathbf{1}_D(x)dx \tag{9}$$

if we sample X uniformly from C,

$$\frac{|C|}{n}\sum_{i=1}^{n} f(X_i)\mathbf{1}_D(X_i) \rightarrow \int_D f(x)dx \tag{10}$$

[1] Length, area, volume, etc., as appropriate

[2] The **indicator function** $\mathbf{1}_D(x)$ is 1 when x is in D, and 0 otherwise

Notice however that sample points X_i which fall outside D are simply wasted, and they will be about $1 - \frac{|D|}{|C|}$ of all the samples — it pays to make the sampling region not too much larger than the domain of integration!

Even this approach is of no use when D is infinite, since then a uniform distribution makes no sense. But an integral like

$$\int_0^\infty x^2 e^{-x^2} dx \tag{11}$$

is, as we know, finite[3], so there should be some trick to evaluating it by Monte Carlo. The trick is to introduce a density p which has the same support as D (in this case, the whole real line). Then

$$\int_D f(x)dx = \int_D \frac{f(x)}{p(x)} p(x)dx \tag{12}$$

Because $p(x) > 0$ everywhere on D, this is legitimate (we're never dividing by zero). And now, if the X_i are generated from p,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)}{p(X_i)} \rightarrow \mathbb{E}_p\left[\frac{f(X)}{p(X)}\right] = \int_D f(x)dx \tag{13}$$

as desired[4]

Again, it's worth asking how good the approximation is, and again, we can use the central limit theorem:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)}{p(X_i)} \rightsquigarrow \mathcal{N}\left(\int_D f(x)dx, \frac{\sigma^2_{f/p}}{n}\right) \tag{14}$$

where $\sigma^2_{f/p}$ is the variance of $f(X)/p(X)$ (when $X \sim p$). Once again, the Monte Carlo approximation is unbiased, and the variance of the approximation goes to zero like $1/n$, no matter how high-dimensional X is, or how ugly f or D might be.

Choosing p In principle, any p which is supported on D[5] could be used for Monte Carlo. In practice, one looks for easy simulation, low variance, and simple forms. Easy simulation speaks for itself; what about the others?

- *Low variance* Notice that if $f(x)/p(x)$ were constant, say c, then the variance of the Monte Carlo variable $\sigma^2_{f/p}$ would be zero, and the Monte Carlo approximation would be exact. Getting this is usually too much to hope for, but, to the extent possible, it generally improves efficiency to have the shape of $p(x)$ follow that of $f(x)$. (Of course this can conflict with easy simulation.)

[3] Why do we know this?

[4] EXERCISE: Convince yourself that Eqs. 5 and 6 are special cases of Eqs. 12 and 13, respectively.

[5] That is, $p(x) > 0$ if and only if x is in D.

- *Simple forms* It is often worth looking carefully at the integrand to see if a probability density can be factored out of it. In (11), for instance, the factor e^{-x^2} is proportional to a $\mathcal{N}(0,1/2)$ density,

$$\int_0^\infty x^2 e^{-x^2} dx = \int_0^\infty \left[\sqrt{\pi} x^2\right] \left[\frac{1}{\sqrt{2\pi/2}} e^{-x^2/(2/2)}\right] \tag{15}$$

  so we can simply simulate from that density and take the average of $\sqrt{\pi} X_i^2$.

Problem

Write a function which takes as input a real-valued function of one argument, a lower and upper limit of integration, and a number of samples, and returns a Monte Carlo approximation to the integral over that interval. Check whether the lower limit is `-Inf` or the upper limit is `Inf` (both might be true!), and chose a p appropriately. How would you check this?

1.1 Calculating Expectation Values

Since expectation values are integrals, everything said above about integrals applies to them. Whenever we can simulate X and calculate f, we can approximate $\mathbb{E}_p[f(X)]$. By the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^n f(X_i) \rightarrow \mathbb{E}_p[f(X)] \tag{16}$$

By the central limit theorem, for large n, the approximations have a Gaussian distribution around the true expected values, with a variance shrinking like $1/n$.

Importance sampling Of course, drawing the X_i from p can be tricky. We can then look for another density r, which has two properties:

1. It is easy for us to simulate from r, and
2. $r(x) > 0$ whenever $p(x) > 0$

The second property lets us write

$$\mathbb{E}_p[f(X)] = \int f(x)p(x)dx = \int f(x)\frac{p(x)}{r(x)} r(x)dx = \mathbb{E}_r\left[f(X)\frac{p(X)}{r(X)}\right] \tag{17}$$

The first property means we can easily draw $X_1, X_2, \ldots X_n$ from r. Then

$$\frac{1}{n}\sum_{i=1}^n f(X_i)\frac{p(X_i)}{r(X_i)} \rightarrow \mathbb{E}_p[f(X)] \tag{18}$$

Notice that this is a weighted mean of the $f(X_i)$, but one where the weights are also random. This sort of approximation, where we calculate the expectation under one distribution by sampling from another, is called **importance sampling**, and the ratios $p(X_i)/r(X_i)$ are the **importance weights**. Once again, the approximations tend to a Gaussian distribution around the truth.

As with picking p in plain Monte Carlo, there is a tension between choosing a distribution r which is easy to draw from, and choosing one which will be efficient. It is easy to check that the sample mean of the importance weights will tend towards 1 as n grows[6]. But if r puts a lot of probability on regions where $p(x)$ is very small, many terms in the sample average will be weighted down towards zero, and so nearly wasted; while a few will have to get very large weights, and averaging only a few random terms is noisy. So to get good approximations to $\mathbb{E}_p[f(X)]$, it is usually desirable for $p(x)/r(x)$ to not vary too much from 1[7].

Problem

The **entropy** of a distribution with density p is

$$-\int p(x)\log_2 p(x)dx \tag{19}$$

This quantity is extremely important in information theory, since it can be used to quantify how many bits of memory are needed to store a value drawn from the distribution, or to transmit it. Use Monte Carlo to find the entropy of a Gaussian distribution with mean 5 and variance 3, and of an exponential distribution with scale 0.5. Analytical formulas for the entropy are available for these two distributions; look them up[8] and calculate the exact values. How large do you need to make n in each case to get agreement to 2 significant digits? To 3 significant digits?

2 Markov Chains

So far, we have been simulating sequences of completely independent random variables. This is excessively limiting. The world has lots of variables which are related to each other — the most important parts of statistics are about describing these relationships — and so we should be able to simulate dependent variables.

A **stochastic process** is simply a collection of random variables with a joint distribution, usually a dependent one[9]. Often, but not always, the variables come in a sequence, $X_1, X_2, \ldots X_n, \ldots$. In principle, then, the distribution of X_{t+1} could depend on the value of all previous variables, $X_1, X_2, \ldots X_t$. At the other extreme, in an IID sequence, no variable depends on any other variable.

[6] Because $\int \frac{p(x)}{r(x)} r(x)dx = 1$ (EXERCISE: why?), and the law of large numbers applies.

[7] EXERCISE: Are there ever situations where the estimate would be improved by sampling from r rather than p, assuming it's equally easy to do either?

[8] Or re-derive them!

[9] The original motive for using the word “stochastic” in place of “random” is that people tended to take “random” as implying “statistically independent”.

The most important class of stochastic processes which actually have dependence are the **Markov processes**[10], in which the distribution of X_{t+1} depends only on the value of X_t. The variable X_t is called the **state** of the process at time t.

When I say that the distribution of X_{t+1} depends only on the value of X_t, I mean in particular that X_{t+1} is *conditionally independent* of $X_1, \ldots X_{t-1}$ given X_t. ("The future is independent of the past, given the present state.") This conditional independence is called the **Markov property**. Conceptually, we can view it in two ways:

- In an IID sequence, X_{t+1} is conditionally independent of earlier states given X_t. It's also *unconditionally* independent of them, since there is no dependence at all. From this perspective, the Markov property is a minimal weakening of the idea of an IID sequence.

- In a deterministic dynamical system, like the Arnold cat map we saw in the last lecture (or the laws of classical physics), the next state is a *function* of the current state, and earlier states are irrelevant given the present. From this perspective, the Markov property just says it's OK to replace strict determinism and with probability distributions.

So Markov processes are "just right" to generalize both complete independence and strict determinism. Markov chain models are used a lot in physics, chemistry, genetics, ecology, psychology, economics, sociology, and, no doubt, other fields.

Mathematically, there are two components to getting the distribution of a Markov process. The distribution of X_1, the **initial distribution**, is just another probability distribution, say p_0. Thereafter, we need the conditional distribution of X_{t+1} given X_t, which I'll write $q(y|x)$. This is either a conditional probability density function (if the X_t are continuous) or a conditional probability mass function (if they are discrete). There are many names for this, but the most transparent may be the **transition distribution**[11]. Then

$$p(x_1, x_2, \ldots x_t) = p_0(x_1) \prod_{i=1}^{t-1} q(x_{i+1}|x_i) \tag{20}$$

Markov Chains and the Transition Matrix When the X_t variables are all discrete, the process is called a **Markov chain**. If they are not just discrete but finite, say K of them, then we can represent the transition distribution as a $K \times K$ matrix, $\mathbf{q}$ say, where

$$q_{ij} = q(j|i) = \mathbb{P}\left(X_{t+1} = j | X_t = i\right) \tag{21}$$

Notice that all entries of $\mathbf{q}$ must be ≥ 0, and each row must sum to 1. Such a matrix is also called **stochastic**[12]. We can now use matrix arithmetic to look at how probability distributions change.

[10] These are named after the great Russian mathematician A. A. Markov, who was the first to systematically describe them and recognize their importance. See [1] for an accessible account of Markov's life and work, and the origins of his theory of chains in a theological quarrel with his arch-nemesis.

[11] Technically, in assuming that q stays the same for all t, I am assuming that the Markov process is "homogeneous". Inhomogeneous Markov processes exist, but are not very useful for present purposes.

[12] Technically, **row stochastic**, which lets you guess what a **column stochastic** matrix is, and then what a **doubly stochastic** matrix might be.

Evolving Probability Distributions Suppose we start with a certain distribution p_0 on the states of the Markov chain. Because there are only finitely many states, we can represent p_0 as a $1 \times K$ vector. Then

$$p_1 = p_0 \mathbf{q} \tag{22}$$

is another $1 \times K$ vector. Notice that

$$(p_1)_i = \sum_{j=1}^{K} (p_0)_j q_{ji} = \sum_{j=1}^{K} \mathbb{P}(X_1 = j) \mathbb{P}(X_2 = i | X_1 = j) = \mathbb{P}(X_2 = i) \tag{23}$$

so multiplying the probability distribution p_0 by $\mathbf{q}$ gives us the new probability distribution, after one step of the chain. If we look at

$$p_t = p_{t_1} \mathbf{q} = p_0 \mathbf{q}^t \tag{24}$$

we get the distribution of states after t steps of the chain[13].

2.1 Asymptotics of Markov Chains

What happens to p_t **asymptotically**, as $t \rightarrow \infty$? Since we are getting p_t by matrix arithmetic, it is natural to turn to linear algebra for an answer, and specifically to eigenvalues and eigenvectors.

Since $\mathbf{q}$ is a square, $K \times K$ stochastic matrix, it will have K eigenvectors, say $v_1, \ldots v_K$, and K eigenvalues, $\lambda_1, \ldots \lambda_K$. (Not all of the eigenvalues are necessarily different.) The eigenvectors will form a basis for $\mathbb{R}^K$, meaning that we can write an arbitrary vector as a linear combination of the eigenvectors. In particular, we can write p_0 so:

$$p_0 = \sum_{j=1}^{K} a_j v_j \tag{25}$$

for some[14] coefficients a_j.

Since the eigenvectors multiply easily by $\mathbf{q}$, we can now write p_t very simply:

$$p_1 = p_0 \mathbf{q} = \sum_{j=1}^{K} a_j v_j \mathbf{q} = \sum_{j=1}^{K} a_j \lambda_j v_j \tag{26}$$

Iterating this,

$$p_t = \sum_{j=1}^{K} a_j \lambda_j^t v_j \tag{27}$$

Notice that the only part of Eq. 27 which changes with t is the power of the eigenvalues. If $|\lambda_j| > 1$, that term grows exponentially; if $|\lambda_j| < 1$, that term shrinks exponentially; only if $|\lambda_j| = 1$ does the size of the term remain the same.

[13] This is not the same as the distribution of state *sequences* for the first t steps, which is given by Eq. 20.

[14] If the eigenvectors are orthogonal to each other, then the a_j are just their inner products with p_0. (Why?)

So far, we haven't used the fact that the matrix $\mathbf{q}$ is rather special. Let's start doing so:

Proposition 1 *All the eigenvalues λ_j of any stochastic matrix are on or inside the unit circle, $|\lambda_j| \leq 1$, and there is always at least one "unit" eigenvalue, $\lambda_j = 1$ for at least one j.*

The actual proof is complicated[15], but the intuition is that this comes from the requirement that *probability is conserved* — the total probability of all states cannot grow larger than 1 or shrink below it.

Returning to Eq. 27, the proposition means we can divide the eigenvectors into two kinds: those with $|\lambda_j| < 1$ and those with $|\lambda_j| = 1$. The contribution of the former becomes exponentially small as t grows, so

$$p_t \rightarrow \sum_{j\,:\,|\lambda_j|=1} a_j \lambda_j^t v_j \tag{28}$$

We'll need another fact about stochastic matrices.

Proposition 2 *If $\lambda_j = 1$, then all entries of v_j are ≥ 0.*

Since if v_j is an eigenvector, so is cv_j, we can normalize these v_j so their entries sum to 1. The eigenvectors which go with eigenvalue 1, then, are probability distributions on the states. Since, for these vectors, $v_j \mathbf{q} = v_j$, these are called **invariant** or **equilibrium** distributions.

Suppose that the only eigenvalues on the unit circle, $|\lambda_j| = 1$, are unit eigenvalues, $\lambda_j = 1$. (We will see presently when this will happen.) Then we could say

$$p_t \rightarrow \sum_{j\,:\,|\lambda_j|=1} a_j v_j \tag{29}$$

This limiting distribution is also an invariant distribution (exercise!). So, in the long run, the distribution of the state tends to a distribution which is invariant under the transition matrix.

How much does the initial distribution p_0 matter in the long run? The only place it shows up on the right-hand side in Eq. 29 is that it implicitly determines the weights a_j. But even then, if there is only a *single* invariant eigenvector v_1,, then it has to have weight $a_1 = 1$, and the initial distribution doesn't matter at all for the long run.

2.2 Graph Form of a Markov Chain

Suppose we have a simple two-state Markov chain, with transition matrix

$$\mathbf{q} = \left[\begin{array}{cc} 0.5 & 0.5 \\ 0.75 & 0.26 \end{array} \right] \tag{30}$$

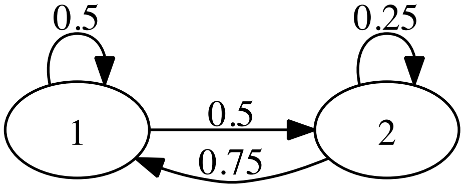

Figure 1: Graphical representation of the Markov chain transition matrix from Eq. 30.

We can represent this state space as by a graph or network, as in Figure 1.

The general rule is for drawing such graphs is that each state is a node, and that each non-zero transition probability, $q_{ij} > 0$, is a directed edge from i to j, labeled with the probabilities. You can amuse yourself, for instance, by working out the transition matrix corresponding to Figure 2.

We say that two i and j states are **connected** if there is a directed path from i to j . We say that they are **strongly connected** if there are directed paths in both directions, from i to j and from j back to i. This is a transitive relation: if i is strongly connected to j, and j is strongly connected to k, then i is strongly connected to k. This means that we can break up the graph into **strongly connected components**, where all states in a component are strongly connected to each other, and are *not* strongly connected to any other states. For instance, in Figure 2, states 1 and 2 form a strongly connected component, states 3, 4 and 5 are another, and finally state 6 is strongly connected to itself but to no other state, so it is in a component of one state.

If a state i is connected, but not *strongly* connected, to some state j, then i is **transient**[16]. If i is not transient — that is, it is strongly connected to every state to which it is connected — then it is **recurrent**. If i is recurrent, then every state it is connected to must also be recurrent, and we also call the whole strongly connected component it belongs to recurrent. In Figure 2, states 1 and 2 form a recurrent component, and states 3, 4 and 5 form another.

EXERCISE: Convince yourself that every path of a finite-state Markov chain eventually hits one recurrent component or another, and then stays in that component forever. *Hint*: the initial state is either recurrent or transient.

Each recurrent component corresponds to an eigenvector of $\mathbf{q}$ with eigenvalue 1. This eigenvector is > 0 on the states of the component, and $= 0$ on the other states. Remember that these eigenvectors are invariant distributions: they can give probability 0 to the rest of the state space, because a chain started in the component has zero probability of leaving it. But they cannot give probability 0 to states in the component, since there is always some probability of reaching them from the rest of

[15] It's part of the "Perron-Frobenius theorem".

[16] The reason for the name is that if we start the chain in state i, it will *eventually* find is way to state j, from which it has no way back to i.

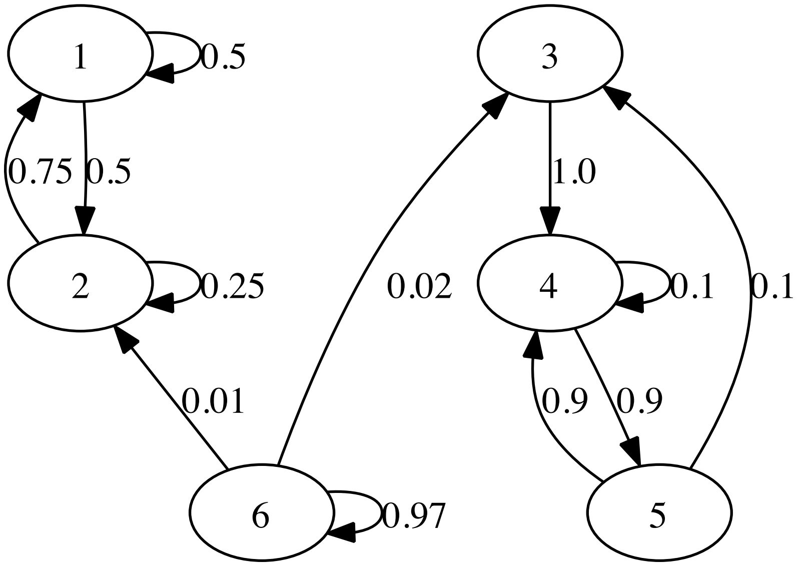

Figure 2: Graphical representation of a Markov chain with six states. Can you work out the transition matrix?

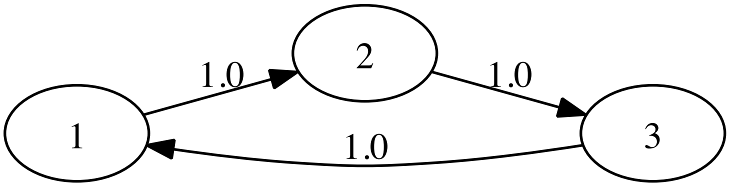

Figure 3: A Markov chain in which there is one recurrent component, and every state has period three.

the component, and the distribution is not allowed to change. So the basic invariant distributions match up with the recurrent components of the graph.

What about eigenvectors v_j where $|\lambda_j| = 1$ but $\lambda_j \neq 1$? These correspond to *periodic* cycles in recurrent components. To see what this means, pick a state i and look the length of paths in the graph which start and end at i ("cycles rooted at i"). If there is a common divisor to these path lengths, then there is a periodicity to the behavior of the chain — it can only return to i at certain times. The period of the state is the greatest common divisor of its cycle lengths. If the greatest common divisor is 1, the state is **aperiodic**. All the the states in a recurrent component must share the same period.

To be concrete, look at Figure 3. There are three states, and the chain simply rotates through them. There is, as promised, one eigenvector with eigenvalue 1, which puts probability 1/3 on each state. This is the unique invariant distribution. But the two complex eigenvalues are also on the unit circle, and their eigenvectors "do the rotation". If we start with the distribution $(a, b, 1 - a - b)$, after one step we get $(1 - a - b, a, b)$, after two steps we get $(b, 1 - a - b, a)$, and after three steps we are back where we started. This never approaches the invariant distribution.

EXERCISE: Consider the more complicated chain in Figure 4. Convince yourself that each state has period 3. What happens to an arbitrary initial distribution $(a, b, c, 1 - a - b - c)$ after one step? Two steps? Three steps? Many steps?

2.3 The Ergodic Theorem (Law of Large Numbers for Markov Chains)

To return to the long-run distribution, we can now say that if there is a single, aperiodic recurrent component, then there is only one eigenvector, v^*, with eigenvalue 1, and

$$p_t = p_0 \mathbf{q}^t \rightarrow v^* \tag{31}$$

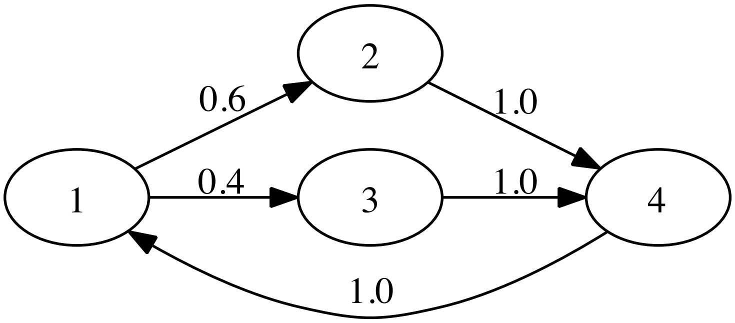

Figure 4: A Markov chain with four states, where each state has period 3.

Thus, in the long run, the distribution of the state at any one time approaches the unique invariant or equilibrium distribution.

To find out how quickly the chain approaches equilibrium, put the eigenvalues in order of magnitude[17], $1 = \lambda_1 > |\lambda_2| \geq \ldots \geq |\lambda_{K-1}| \geq |\lambda_K|$, and put their eigenvectors v_j in the same order. For large t, then,

$$1 \gg |\lambda_2|^t \gg |\lambda_j|^t \tag{32}$$

no matter what $j > 2$ we pick. Consequently, again for large t,

$$p_t - v^* \approx a_2 \lambda_2^t v_2 \tag{33}$$

Thus, the difference between the distribution after t steps p_t and the invariant distribution v^* shrinks exponentially fast, and the base of the exponent is λ_2. If $|\lambda_2|$ is very close to 1, then this exponential rate could be very slow, a point we'll come back to next time.

What this means concretely is that if we took many independent copies of the Markov chain, and chose their initial states according to p_0, we would see the distribution of states across this "ensemble" tending towards v^* as time went on, no matter what p_0 was. This is interesting, but can we say anything about what happens within any one long run of the chain?

If we take a long IID sequence, the empirical distribution within that sequence tends towards the true distribution. The same is true for a finite-state Markov chain

[17] Break ties however you like.

with a single recurrent[18] component:

$$\frac{1}{n}\sum_{t=1}^{n}\mathbf{1}_i(X_t) \rightarrow v_i^* \tag{34}$$

Since there are only finitely many states, this means[19]

$$\frac{1}{n}\sum_{t=1}^{n} f(X_t) \rightarrow \mathbb{E}_{v^*}\left[f(X)\right] \tag{35}$$

This generalizes the law of large numbers from IID sequences to Markov chains. For historical reasons[20], this is called the **ergodic theorem**. To get a sense of why the ergodic theorem is true, look at the appendix.

The slogan which summarizes it is that "time averages converge on state averages", with "time averages" being the left-hand sides of Eqs. 35 and 34, and "state averages", a.k.a. expectation values, being the right-hand sides. In statistical terms, what the ergodic property means is that a single long realization of a Markov chain acts like a representative sample of the whole distribution, and becomes increasingly representative as it grows longer. This is important for several reasons.

1. It re-assures us about the statistical methods we learned for IID data. It says that even if the data we have to deal with are not *completely* independent, if they are at least Markov, then much of what we thought we knew is still at least approximately true.

2. It tells us that we can apply statistics to dynamics, to things which change over time — at least if they are Markov. In fact, we can rely on just a single long trajectory, rather than having to get multiple, independent trajectories, which could be very difficult indeed.

3. It gives us a way to short-cut long simulations. Often what we want out of a simulation is a time-average (what fraction of the time is the system in some failure mode? how much of the final product is produced per unit time? what is the average rate of return on the portfolio?). The ergodic theorem says we don't *have* to step through many simulations of the Markov chain to get these averages, we can just find the invariant distribution and calculate from there.

[18] Notice that we do not need to assume the component is recurrent *and* aperiodic, just recurrent. You can check with, say, the period-3 chain from Figure 3 that the next two equations hold, no matter what state we start from.

[19] EXERCISE: Convince yourself that Eq. 35 really does follow from Eq. 34.

[20] Several decades before Markov, the physicist Ludwig Boltzmann, as part of explaining why thermodynamics works, argued that a large collection of molecules should, within a short time, come arbitrarily close to every configuration of the molecules which was compatible with the conservation of energy. Since all our measurements take a long time to make (molecularly speaking at least), we would see only the average over these configurations, which would look like an expectation value. He needed a name for this, and called it the "ergodic" property of the trajectory, from the Greek *ergon* ("energy, work") + *odos* ("path, way"). Sadly, the name stuck.

4. It gives us a way to replace complicated expectations with simple simulations. We will see next time that there are many cases where it is much easier to find a chain whose invariant distribution is v^* than it is to find v^* itself. (Strange but true!) Simulating from the chain then gives a way of calculating expectations with respect to v^*.

3 Summing Up

Here are the key ideas:

1. The Monte Carlo method is to evaluate integrals and expectations by simulating from suitable distributions and taking sample averages over the simulation points. So long as sample averages converge on expectations, the approximation error in the Monte Carlo method can be made as small as we like.

   - With independent samples, the Monte Carlo approximation is unbiased, and has a variance is $O(1/n)$ in the number of simulation points.
   - In importance sampling, we simulate from a different distribution than the one we're really interested in, and then correct by taking a weighted average.

2. Markov processes are sequences of random variables where the distribution of the next variable depends only on the value of the current one. A Markov chain is a Markov process where the states are discrete. The transitions in a Markov chain are represented in a stochastic matrix.

   - To find the distribution after t steps of the chain, we take the current distribution (written as a vector) and multiply it by the t^{th} power of the transition matrix.
   - As t grows, the distribution becomes a combination of the eigenvectors whose eigenvalues have magnitude 1. Eigenvectors whose eigenvalues are simply 1 are invariant distributions.
   - In an aperiodic chain, the long-run distribution is an invariant distribution.
   - Each eigenvector with eigenvalue 1 corresponds to a different recurrent component of states. If there is only a single recurrent, aperiodic component, the long-run distribution is this unique eigenvector.
   - Sample averages taken along any single sequence of the chain converge on expectations under the invariant distribution (ergodicity).

A Hand-waving Argument about the Ergodic Theorem

The left-hand side of Eq. 34, what we want to have converge, is

$$\frac{1}{n}\sum_{t=1}^{n}\mathbf{1}_i(X_t) \tag{36}$$

This is a random quantity. Let's try to work out its expectation value (assuming $X_1 \sim p_0$) and its variance as $n \to \infty$. If the expectation tends to a limit and the variance shrinks to zero, then the time-average as a whole must converge on its expectation value.

For expectation, we use the fact that taking expectations is a linear operation:

$$\begin{aligned}
\mathbb{E}_{p_0}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbf{1}_i(X_t)\right] &= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{p_0}\left[\mathbf{1}_i(X_t)\right] && (37)\\
&= \frac{1}{n}\sum_{t=1}^{n}\mathbb{P}\left(X_t = i\right) && (38)\\
&= \frac{1}{n}\sum_{t=1}^{n}(p_0 q^{t-1})_i && (39)
\end{aligned}$$

Since, as $t \to \infty$

$$p_0 q^t \to v^* \tag{40}$$

we can conclude that

$$(p_0 q^{t-1})_i \to v_i^* \tag{41}$$

and therefore

$$\mathbb{E}_{p_0}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbf{1}_i(X_t)\right] \to v_i^* \tag{42}$$

Variance is more involved:

$$\begin{aligned}
\mathrm{Var}_{p_0}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbf{1}_i(X_t)\right] &= \frac{1}{n^2}\mathrm{Var}_{p_0}\left[\sum_{t=1}^{n}\mathbf{1}_i(X_t)\right] && (43)\\
&= \frac{1}{n^2}\left[\sum_{t=1}^{n}\mathrm{Var}_{p_0}\left[\mathbf{1}_i(X_t)\right] + 2\sum_{t=1}^{n-1}\sum_{s=t+1}^{n}\mathrm{Cov}_{p_0}\left[\mathbf{1}_i(X_t),\mathbf{1}_i(X_s)\right]\right]
\end{aligned}$$

(Recall $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y]$.) The first sum, of the variances, we can handle; the summands tend towards $v_i^*(1-v_i^*)$ as t grows. (Why?) Thus the whole sum approaches $nv_i^*(1-v_i^*)$.

The covariances are where I'm going to wave my hands. Since $p_0 q^t \to v^*$, *no matter what distribution we put in for* p_0, the Markov chain is "asymptotically independent". After many time-steps, the chain has nearly the same distribution no matter what state we started it in. So $\mathbb{P}\left(X_t = i, X_s = j\right) \to \mathbb{P}\left(X_t = i\right)\mathbb{P}\left(X_s = j\right)$ as $s - t \to \infty$. And in fact, from Eq. 33, we know that $|\mathbb{P}\left(X_t = i, X_s = j\right) - \mathbb{P}\left(X_t = i\right)\mathbb{P}\left(X_s = j\right)|$ shrinks to zero exponentially, with the exponent depending on $|\lambda_2|$. In fact, we can reasonably suppose (hand-waving!) that

$$|\mathbb{P}\left(X_t = i, X_s = j\right) - \mathbb{P}\left(X_t = i\right)\mathbb{P}\left(X_s = j\right)| \leq \varkappa_i |\lambda_2|^{s-t} \tag{44}$$

for some constant $\varkappa_i$. Since

$$\mathrm{Cov}\left[\mathbf{1}_i(X_t), \mathbf{1}_i(X_s)\right] = \mathbb{P}\left(X_t = i, X_s = j\right) - \mathbb{P}\left(X_t = i\right)\mathbb{P}\left(X_s = j\right) \tag{45}$$

(why?), we can add up the covariances,

$$\begin{aligned}
\sum_{s=t+1}^{n} \mathrm{Cov}_{p_0}\left[\mathbf{1}_i(X_t), \mathbf{1}_i(X_s)\right] &\leq \sum_{s=t+1}^{n} \varkappa_i |\lambda_2|^{s-t} && (46)\\
&\leq \sum_{s=t+1}^{\infty} \varkappa_i |\lambda_2|^{s-t} && (47)\\
&= \varkappa_i \sum_{h=1}^{\infty} |\lambda_2|^h && (48)\\
&= \varkappa_i \frac{|\lambda_2|}{1-|\lambda_2|} && (49)
\end{aligned}$$

since the infinite sum is a geometric series. The sum of the covariances is thus limited:

$$\begin{aligned}
&\sum_{t=1}^{n-1}\sum_{s=t+1}^{n} \mathrm{Cov}_{p_0}\left[\mathbf{1}_i(X_t), \mathbf{1}_i(X_s)\right] && (50)\\
&\leq \sum_{t=1}^{n-1} \varkappa_i \frac{|\lambda_2|}{1-|\lambda_2|} \\
&\leq n\varkappa_i \frac{|\lambda_2|}{1-|\lambda_2|} && (51)
\end{aligned}$$

Putting everything together,

$$\begin{aligned}
&\mathrm{Var}_{p_0}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{1}_i(X_t)\right] && (52)\\
&\leq \frac{1}{n^2} n v_i^*(1-v_i^*) + \frac{2}{n^2} n \varkappa_i \frac{|\lambda_2|}{1-|\lambda_2|} && (53)\\
&= \frac{v_i^*(1-v_i^*) + 2\varkappa_i |\lambda_2|/(1-|\lambda_2|)}{n} && (54)
\end{aligned}$$

which → 0 as $n \to \infty$.

It may clear up this some tricky mathematical argument to compare what's going on here to what we'd see if we had an IID sample of discrete variables. Then all the covariance terms would go away[21], and we'd have

$$\mathrm{Var}_{p_0}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbf{1}_i(X_t)\right] = \frac{v_i^*(1-v_i^*)}{n} \tag{55}$$

This is the familiar situation from basic statistics where each sample gives us an independent piece of information about the distribution, and our uncertainty about population probabilities goes down like $1/n$. With the Markov chain, because the samples are correlated with each other, each observation is *not* an independent piece of information. But it's *like* we had some smaller number of independent samples:

$$\mathrm{Var}_{p_0}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbf{1}_i(X_t)\right] = \frac{v_i^*(1-v_i^*)}{n/\tau} \tag{56}$$

where

$$\tau = 1 + 2\frac{\varkappa_i}{v_i^*(1-v_i^*)}\frac{|\lambda_2|}{1-|\lambda_2|)} \tag{57}$$

tells us how long we have to wait for the correlations to become negligible.

Stepping back even more for the really big picture, what's going on is that the variables in a Markov chain are dependent, but widely-separated ones are almost independent. So if we wait for a long time, averages over the dependent variables look very much like averages over independent ones. How long is "long" is quantified by τ.

Problem

Suppose that $X_1, X_2, \ldots X_t, \ldots$ is a sequence of variables, not necessarily a Markov chain, which is "weakly stationary", so

$$\mathbb{E}\left[X_t\right] = \mathbb{E}\left[X_s\right] = \mu \tag{58}$$

and that

$$\mathrm{Cov}\left[X_t, X_s\right] = \rho(|t-s|) \tag{59}$$

for t, s. Further suppose that

$$\sum_{h=0}^{\infty}\rho(h) = \varkappa\rho(0) < \infty \tag{60}$$

Imitating the arguments made above, prove that

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}X_t\right] = \mu$$

[21] Notice that an IID sequence is a kind of Markov chain, where $\lambda_2 = 0$ exactly.

and that

$$\mathrm{Var}\left[\frac{1}{n}\sum_{t=1}^{n}X_t\right] \leq \frac{\rho(0)}{n/(1+2x)}$$

Conclude that

$$\frac{1}{n}\sum_{t=1}^{n}X_t \rightarrow \mu$$

Congratulations; you have proved the "mean-square ergodic theorem".

References

[1] Basharin, Gely P., Amy N. Langville and Valeriy A. Naumov (2004). "The Life and Work of A. A. Markov." *Linear Algebra and its Applications*, **386**: 3–26. URL `http://decision.csl.uiuc.edu/~meyn/pages/Markov-Work-and-life.pdf`.

[2] Serber, Robert (1992). *The Los Alamos Primer: The First Lectures on How to Build the Atomic Bomb*. Berkeley: University of California Press. Annotated by Robert Serber; edited and with an introduction by Richard Rhodes.