AN ESSAY ABOUT MARKOV PROCESSES ©

Daniel W. Stroock

Markov Processes, What are They?

Unless one is clairvoyant, the only temporally evolving processes which are tractable are those whose future behavior can be predicted on the basis of data which is available at the time when the prediction is being made. Of course, in general, the behavior of even such an evolution will be impossible to predict. For example, if, in order to make a prediction, one has to know the detailed history of everything that has happened during the entire history of the entire universe, ones chance of making a prediction may be a practical, if not a theoretical, impossibility. For this reason, one tries to study evolutions mathematically with models in which most of the distant past can be ignored when making predictions about the future. In fact, many mathematical models of evolutions have the property that, for the purpose of predicting the future, the past becomes irrelevant as soon as one knows the present, in which case the evolution is said to be a *Markov process*, the topic at hand.

The components of a Markov process are its *state space* $\mathbb{S}$ and its *transition rule* T. Mathematically, $\mathbb{S}$ is just some non-empty set, which in applications will encode all the possible states in which the evolving system can find itself, and $T : \mathbb{S} \longrightarrow \mathbb{S}$ is a function from $\mathbb{S}$ into itself which gives the *transition rule*. More precisely, if now the system is in state x, it will be next[1] in state $T(x)$, from which it will go to $T^2(x) = T(T(x))$, etc. To give a sense of the sort of reasoning required to construct a Markov process, consider a (classical) physical particle whose motion is governed by Newton's equation $\vec{F} = m\vec{a}$ ("force equals mass times acceleration"). At least in theory, Newton's equation says that, assuming one knows the mass of the particle and the force field $\vec{F}$ which acts on it, one can predict where the particle will be in the future as soon as one knows what its position and velocity are now. On the other hand, knowing only its present position is not sufficient by itself. Thus, even though one may care about nothing but its position, in order to produce a Markov process for a particle evolving according to Newton's equation, it is necessary to adopt the attitude that the *state* of the particle consists of its position *and* velocity, not just its position alone. Of course, in that velocity is the derivative of position, the two are so inextricably intertwined that one might be tempted to concentrate on position on the grounds that one will be able to compute the velocity whenever necessary. However, this

[1]Here we are thinking of time being discrete. Thus, "next" means after one unit of time has passed.

tack destroys the Markov property. Namely, there is no way of computing the velocity of a particle "now" if all ones knows knows is its position "now." For this reason, physicists consider the state of a particle to be composite of its position and velocity, and the resulting state space $\mathbb{R}^6 = \mathbb{R}^3 \times \mathbb{R}^3$ (three coordinates for position and three for velocity) they call the *phase space* of the particle.

The same point may be clearer in the following example. Suppose that one has an evolution on a state space $\mathbb{S}$ which proceeds according to the rule that if the present state is x_n and the preceding state was x_{n-1}, then the next state will be $x_{n+1} = T(x_{n-1}, x_n)$. This is *not* a Markov process. On the other hand, it can be "Markovized." Indeed, replace the original state space by $\hat{\mathbb{S}} = \mathbb{S} \times \mathbb{S}$, the set of ordered pairs (x, y) with x and y from $\mathbb{S}$, and define $\hat{T}\big((x,y)\big) = \big(y, T(x,y)\big)$. It is then an easy matter to check that if the original system was in state x_{-1} at time -1 and state x_0 and time 0, then its state at time $n \geq 1$ will be x_n, the second component of the pair $(x_{n-1}, x_n) = \hat{T}^n\big((x_{-1}, x_0)\big)$.

The moral to be drawn from these examples is that *the presence or absence of Markov property is in the eye of the beholder*. That is, a change of venue (i.e., state space) can make the Markov property appear in circumstances where it was not originally apparent. In fact, by taking the state space sufficiently large, any evolution can be forced to be Markov. On the other hand, the more complicated the state space, the less useful is the Markov property. Thus, in practice, what one seeks is the "simplest" state space on which ones evolution possesses the Markov property.

Stochastic Markov Processes

Roughly speaking, Markov processes fall into one of two categories. Those in the first category are "deterministic" in the sense that their state space is sufficiently detailed that the individual states give complete and unambiguous information. Both the examples given above are deterministic. The mathematical analysis of deterministic, Markov processes has a proud history going back to Newton and includes major contributions by such luminaries as P. Chebyshev, A. Markov, A. Lyapounov, H. Poincaré, and J. Moser. The second category of Markov processes, and the one on which the rest of this article will concentrate, are "probabilistic" or "stochastic" Markov processes. To understand where and why these processes arise, consider the problem of describing the state of all the gas molecules in a room. Each liter of gas contains approximately Avogrado's number, $6.02214199 \times 10^{23}$, of molecules. Thus, even a small room will contain something on the order of 10^{26} molecules. Moreover, because, by Newton's laws of motion, the state of each individual molecule will lie in its individual phase space, the state of the entire system of molecules will have to specify the positions and velocities of all 10^{26} molecules. Stated mathematically, the state space of the system will be $\mathbb{R}^{6\times 10^{26}}$, on which any sort serious analysis is too daunting to contemplate.

When confronted with a problem which is intractable as presented, the time-honored procedure of choice is to reformulate the problem in a way which makes it more tractable. In the case just described, the reformulation was made by G.W. Gibbs and L. Boltzmann, the fathers of statistical mechanics. Namely, they abandoned any hope of saying exactly

where all the molecules will be and reconciled themselves to settling for a description of the statistics of the molecules. That is, instead of asking exactly where all the molecules would be, they asked what would be probability of finding a molecule in various regions of phase space. From this point of view, the state of the system will not be an element of $\mathbb{R}^{6\times 10^{26}}$ but of $\mathbf{M}_1(\mathbb{R}^6)$, the space probability distributions on the individual phase space $\mathbb{R}^6$. Of course, Gibbs and Boltzmann's reformulation only changes the problem, it does not solve it. Indeed, although Newton's equation determines how the system of molecules evolves and therefore how their distribution will evolve, the use Newton's equation would remove the advantage which Boltzmann and Gibbs hoped to gain via their reformulation. Thus, they had to come up with an alternative way of describing the transition rule which governs the evolution of the distribution of the system as a Markov process on $\mathbf{M}_1(\mathbb{R}^6)$. The description proposed by Boltzmann is given by the famous Boltzmann equation. Unfortunately, Boltzmann's equation is itself so complicated that it is only recently that substantial progress has been made toward understanding it in any generality. On the other hand, Gibbs and Boltzmann's idea of studying Markov processes on the space of probability distributions is seminal and has proved itself to be both ubiquitous and powerful.

The abstract setting for a stochastic Markov process starts with a non-empty set $\mathbb{S}$, the deterministic state space, and the associated space $\mathbf{M}_1(\mathbb{S})$ of probability distributions on $\mathbb{S}$. The easiest and most commonly studied stochastic Markov processes are those for which the transition rule $T : \mathbf{M}_1(\mathbb{S}) \longrightarrow \mathbf{M}_1(\mathbb{S})$ is a linear (more correctly, an affine) function. To be definite, suppose $\mathbb{S}$ is a finite set. Then $\mathbf{M}_1(\mathbb{S})$ is the set of all functions μ on $\mathbb{S}$ which assign each $x \in \mathbb{S}$ a number[2] $\mu(\{x\}) \in [0,1]$ (the probability of $\{x\}$ under μ) in such a way that $\sum_{x\in\mathbb{S}} \mu(\{x\}) = 1$. Clearly, if μ and ν are in $\mathbf{M}_1(\mathbb{S})$ and $\theta \in [0,1]$, then the convex combination $\theta\mu + (1-\theta)\nu$ is again an element of $\mathbf{M}_1(\mathbb{S})$. Sets with this property are said to be *affine* (as distinguished from linear, which refers to sets which are closed under all linear, not just convex, combinations), and a function on an affine set is said to be affine if it commutes with affine combinations. Thus, for $\mathbf{M}_1(\mathbb{S})$, the transition rule T is affine if $T\big(\theta\mu + (1-\theta)\nu\big) = \theta T(\mu) + (1-\theta)T(\nu)$. Because $\mathbb{S}$ is finite, one can dissect such transition rules in the following way. First, for each $x \in \mathbb{S}$, let δ_x denote the element of $\mathbf{M}_1(\mathbb{S})$ which assigns 1 to $\{x\}$ (and therefore 0 to $\mathbb{S} \setminus \{x\}$). Next, set $\mathbf{P}(x, \cdot\,) = T(\delta_x)$. That is, $\mathbf{P}(x, \cdot\,)$ is the element of $\mathbf{M}_1(\mathbb{S})$ to which T takes δ_x, and so $\mathbf{P}(x, \{y\}) = \big[T(\delta_x)\big](\{y\})$. Because, for any $\mu \in \mathbf{M}_1(\mathbb{S})$ which is not equal to δ_x, $\mu = \mu(\{x\})\delta_x + \big(1 - \mu(\{x\})\big)\mu^x$, where $\mu^x \in \mathbf{M}_1(\mathbb{S})$ is determined so that $\mu^x(\{y\})$ equals $\big(1-\mu(\{x\})\big)^{-1}\mu(\{y\})$ or 0 depending on whether $y \neq x$ or $y = x$, the affine property of T means that $T(\mu) = \mu(\{x\})\mathbf{P}(x, \cdot\,) + \big(1 - \mu(\{x\})\big)T(\mu^x)$. Hence, after pealing off one x at a time, one concludes that

$$(*) \qquad T(\mu) = \sum_{x\in\mathbb{S}} \mu(\{x\})\mathbf{P}(x, \cdot\,)$$

[2]The use of $\mu(\{x\})$ instead of $\mu(x)$ here is a little pedantic. However, one must remember that probabilities are assigned to *events* (i.e., subsets of the sample space), and that $\{x\}$ is the event that "x occurred."

when T is affine.

Probabilistic Interpretation

The representation of T given by (*) admits an intuitively pleasing probabilistic interpretation. Namely, $\mathbf{P}(x, \{y\})$ can be thought of as the probability that the system will next be in the state y given that is now in state x. With this interpretation in mind, probabilists call $x \in \mathbb{S} \longmapsto \mathbf{P}(x, \cdot\,) \in \mathbf{M}_1(\mathbb{S})$ a *transition probability* on the state space $\mathbb{S}$. The terminology here is confusing. From the point of view adopted earlier, one might, and should, have thought that $\mathbf{M}_1(\mathbb{S})$ is the state space. However, the probabilistic interpretation is most easily appreciated if one thinks of $\mathbb{S}$ as the state space and $x \in \mathbb{S} \rightsquigarrow \mathbf{P}(x, \cdot\,) \in \mathbf{M}_1(\mathbb{S})$ as a random transition rule. To complete this picture, probabilists introduce random variables to represent the random points in $\mathbb{S}$ visited. More precisely, again assume that $\mathbb{S}$ is finite, and suppose that $\mu \in \mathbf{M}_1(\mathbb{S})$ describes the initial distribution of process under consideration. Then probabilists construct a sequence $\{X_n : n \geq 0\}$ of random variables, called a *Markov chain*, in such a way that, for any $n \geq 0$,

$$\mathrm{Prob}\big(X_0 = x_0, \ldots, X_n = x_n\big) = \mu(\{x_0\})\mathbf{P}(x_0, \{x_1\}) \cdots \mathbf{P}(x_{n-1}, \{x_n\}).$$

In words, this says that the right hand side above is the probability that the chain with initial distribution μ starts at x_0 and then goes on to visit, successively, the points x_1 through x_n.

To see that the probabilistic interpretation is completely consistent in the deterministic case, observe that a deterministic Markov process can be formulated as a stochastic Markov process. Namely, if T is the transition rule for the deterministic process, take $\mathbf{P}(x, \cdot\,) = \delta_{T(x)}$, and check, that with probability 1, the stochastic Markov chain with transition probability $x \rightsquigarrow \mathbf{P}(x, \cdot\,)$ follows the same path as the deterministic one with transition rule T. That is, with probability 1, $X_n = T^n(X_0)$ for all $n \geq 1$.

Ergodic Theory of Markov Chains

Continue in the setting of the preceding section. One of the phenomena predicted by Gibbs in connection with his and Boltzmann's study of gases was that, no matter what the initial distribution of the gas, after a long time the gas should equilibrate in the sense that it will achieve a *stationary distribution* (i.e., a distribution which does not change with time) which does not depend on how it was distributed initially. Ones experience with the behavior of gases makes his prediction entirely plausible: place an opened bottle of perfume in the corner of a room, wait an hour, and confirm that the perfume will have become more or less equidistributed throughout the room. Be that as it may, his prediction, which goes by the name of Gibb's *ergodic hypothesis*, has been mathematically verified for only one physically realistic model. Nonetheless, as will be explained next, ergodicity is relatively easy to verify for most stochastic Markov processes on a finite state space.

To develop some intuition for what ergodicity means and why it might hold for a stochastic Markov process on a finite state space $\mathbb{S}$, it is best to first know how to recognize

when a $\mu \in \mathbf{M}_1(\mathbb{S})$ is stationary. But if μ is stationary, then it is left unchanged as the system evolves, and, in terms of the transition probability, this means that

$$\mu(\{y\}) = \sum_{x\in\mathbb{S}} \mu(\{y\})\mathbf{P}(x,\{y\}) \quad \text{for all } y \in \mathbb{S}. \tag{1}$$

Now suppose that $\mathbb{S} = \{1,2\}$, and consider the problem of finding a solution to (1). That is, we to find want $\mu \in \mathbf{M}_1(\{1,2\})$ so that

$$\begin{aligned} \mu(\{1\}) &= \mu(\{1\})\mathbf{P}(1,\{1\}) + \mu(\{2\})\mathbf{P}(2,\{1\}) \\ \mu(\{2\}) &= \mu(\{1\})\mathbf{P}(1,\{2\}) + \mu(\{2\})\mathbf{P}(2,\{2\}). \end{aligned} \tag{*}$$

At first sight, there appear to be too many conditions on μ: not only must it satisfy the two equations in (*), it also has to satisfy $\mu(\{1\}) + \mu(\{2\}) = 1$ as well as being non-negative. Even if one ignores the non-negativity, one suspects that three linear equations are just too many for a pair of numbers to satisfy. On the other hand, after a little manipulation, one sees (remember that $\mathbf{P}(1,\cdot\,)$ and $\mathbf{P}(2,\cdot\,)$ are probability distributions) that both the equation in (*) are equivalent to $\mu(\{1\})\mathbf{P}(1,\{2\}) = \mu(\{2\})\mathbf{P}(2,\{1\})$. Hence the two equations in (*) are equivalent, and so there are really only two equations to be satisfied: $\mu(\{1\})\mathbf{P}(1,\{2\}) = \mu(\{2\})\mathbf{P}(2,\{1\})$ and $\mu(\{1\}) + \mu(\{2\}) = 1$. There are two cases to be considered. The first case is when the chain never moves, or, equivalently, $\mathbf{P}(1,\{2\}) = 0 = \mathbf{P}(2,\{1\})$. In this case there are two solutions, namely, δ_1 and δ_2, which is exactly what one should expect for a chain which never moves. In the second case, the one corresponding to a chain which can move, either $\mathbf{P}(1,\{2\}) > 0$ or $\mathbf{P}(2,\{1\}) > 0$. In both these cases, one can easily check that the one and only solution to (*) is given by

$$\mu(\{1\}) = \frac{\mathbf{P}(2,\{1\})}{\mathbf{P}(1,\{2\}) + \mathbf{P}(2,\{1\})} \quad \text{and} \quad \mu(\{2\}) = \frac{\mathbf{P}(1,\{2\})}{\mathbf{P}(1,\{2\}) + \mathbf{P}(2,\{1\})}.$$

Continuing in the setting of the preceding, we want to examine when Gibb's ergodic hypothesis holds. Obviously, at the very least, ergodicity requires that there be only one stationary μ, otherwise we could start the chain with one of them as initial distribution, in which case it would never get to the other. Thus, we need to assume that $\mathbf{P}(1,\{2\}) + \mathbf{P}(2,\{1\}) > 0$, and, to simplify matters, we will assume more. Namely, we now assume that $m \equiv m_1 + m_2 > 0$, where $m_1 = \min\{P(1,\{1\}), P(2,\{1\})\}$ and $m_2 = \min\{P(1,\{2\}), P(2,\{2\})\}$, and, under this assumption we (following Doeblin) will show that, for any $\nu \in \mathbf{M}_1(\{1,2\})$

$$\|\nu\mathbf{P} - \mu\| \le (1-m)\|\nu - \mu\|, \tag{2}$$

where $\nu\mathbf{P} \in \mathbf{M}_1(\{1,2\})$ is determined by

$$\nu\mathbf{P}(\{y\}) \equiv \sum_{x=1}^{2} \nu(\{x\}))\mathbf{P}(x,\{y\})$$

and, for any pair $\nu_1, \nu_2 \in \mathbf{M}_1(\{1,2\})$, $\|\nu_2 - \nu_1\| \equiv \sum_{x=1}^2 |\nu_2(\{x\}) - \nu(\{x\})|$. To prove (2), first observe that, because μ is stationary, $\mu = \mu\mathbf{P}$, and therefore, since $\sum_{x=1}^2 \big(\nu(\{x\}) - \mu(\{x\})\big) = 1 - 1 = 0$,

$$\begin{aligned}\nu\mathbf{P}(\{y\}) - \mu(\{y\}) &= \sum_{x=1}^2 \big(\nu(\{x\}) - \mu(\{x\})\big)\mathbf{P}(x,\{y\}) \\ &= \sum_{x=1}^2 \big(\nu(\{x\}) - \mu(\{x\})\big)\big(\mathbf{P}(x,\{y\}) - m_y\big).\end{aligned}$$

Next, take the absolute value of both sides, remember that the absolute value of a sum of numbers is dominated by the sum of their absolute values, and arrive at

$$\begin{aligned}\|\nu\mathbf{P} - \mu\| &\le \sum_{y=1}^2 \left(\sum_{x=1}^2 |\nu(\{x\}) - \mu(\{x\})|\big(\mathbf{P}(x,\{y\}) - m_y\big)\right) \\ &= \sum_{x=1}^2 \left(\sum_{y=1}^2 |\nu(\{x\}) - \mu(\{x\})|\big(\mathbf{P}(x,\{y\}) - m_y\big)\right) = (1-m)\|\nu - \mu\|.\end{aligned}$$

Given (2), it becomes an easy matter to check ergodicity. Indeed, $\nu\mathbf{P}$ is the distribution of the chain at time 1 when it is started with initial distribution ν. Similarly, its distribution at time 2 will be $\nu\mathbf{P}^2 = (\nu\mathbf{P})\mathbf{P}$, and so $\|\nu\mathbf{P}^2 - \mu\| \le (1-m)\|\nu\mathbf{P} - \mu\| \le (1-m)^2\|\nu - \mu\|$. Proceeding by induction, one sees that distribution $\nu\mathbf{P}^n = (\nu\mathbf{P}^{n-1})\mathbf{P}$ at time n will satisfy $\|\nu\mathbf{P}^n - \mu\| \le (1-m)^n\|\nu - \mu\|$. Hence, because $m > 0$, this implies that $\|\nu\mathbf{P}^n - \mu\|$ tends to 0 exponentially fast, which means that the chain possesses an extremely strong form of ergodicity.

Other Directions

In this article we have discussed only the most elementary examples of Markov processes. In particular, in order to avoid technical difficulties, all our considerations have been about processes for which the time parameter is discrete. As soon as one moves into the realm of processes with a continuous time parameter, the theory becomes much more technically involved. However, the price which one has to pay in technicalities is amply rewarded by the richness of the continuous time theory. To wit, Brownian motion (a.k.a. the Wiener process) is a continuous parameter Markov process which makes an appearance in a surprising, and ever growing, number of places: harmonic analysis in pure mathematics, filtering and separation of signal from noise in electrical engineering, the kinetic theory of gases in physics, the price fluctuations on the stock market in economics, etc. Thus, for the sake of the curious, here is a very brief and enormously inadequate list of places where one can learn more about Markov processes.

Elementary Texts:

Karlin, S. & Taylor, H, *A First Course in Stochastic Processes, 2nd ed.*, Academic Press, NY, 1975.

Norris, J.R., *Markov Chains*, Cambridge Series in Statistical & Probabilistic Mathematics, Cambridge Univ. Press, Cambridge, U.K., 1997.

Stroock, D., *An Introduction to Markov Processes*, Graduate Text Series #230, Springer–Verlag, Heidelberg, 2005.

More Advanced Texts:

Dynkin, E.B., *Markov Processes, Vols. I & II*, Grundlehren # 121& 122, Springer-Verlag, Heidelberg, 1965.

Revuz, D., *Markov Chains*, Mathematical Library, vol. 11, North Holland, Amsterdam & New York, 1984.

Ethier, S. & Kurtz, T., *Markov Processes. Characterization and Convergence*, Series in Probability and Mathematical Statistics, J. Wiley & Sons, NY, NY, 1986.

Stroock, D., *Markov Processes from K. Itô's Perspective*, Annals of Math. Studies #155, Princeton U. Press, Princeton, N.J., 2003.