Markov Chains: History and Treatment in Probability and Dynamical Systems

Dominic Veconi

Penn State University

September 4 2019

Notation

A **probability space** is a triple $(\Omega, \mathcal{A}, \mathbb{P})$, where $\mathcal{A}$ is a σ-algebra of events and $\mathbb{P} : \mathcal{A} \to [0, 1]$ is a measure.

A **random variable** is a measurable function $X : \Omega \to \mathbb{R}$.

We denote by $\mathbb{E}[X] = \int_\Omega X \, d\mathbb{P}$ the **expectation** or **expected value** of the random variable X.

A **stochastic process** is a collection $\{X_i\}_{i \in \mathcal{I}}$ of random variables, where $\mathcal{I}$ is some index set. If $\mathcal{I} = \mathbb{Z}$, then the stochastic process is **discrete**, and if $\mathcal{I} = \mathbb{R}$, then the stochastic process is **continuous**.

Independence and the Weak Law of Large Numbers

Let $\{X_n\}_{n\geq 1}$ be a discrete stochastic process (i.e. a sequence of random variables). Denote by $\widetilde{S}_n = \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])$.

The sequence of random variables satisfies the **weak law of large numbers** (WLLN) if for every $\varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{1}{n}\widetilde{S}_n\right| > \varepsilon\right] = 0$$

That is, the probability of the average difference being greater than ε approaches 0 for sufficiently many experimental iterations.

The sequence is **independent** if the outcome of one experiment has no effect on the outcome of future or past experiments.

Simple example: Bernoulli processes

Suppose your experiment is a game, with n distinct outcomes, each with probability p_k of occurring, independent of preceding events. (Note, for example, that $p_1 + \cdots + p_n = 1$.)

We call this a **Bernoulli process**: a sequence of random variables $X_1, X_2, \ldots$, with $\mathbb{P}[X_i = k] = p_k$ regardless of which time i we play the game/perform the experiment, and regardless of whatever has happened previously. (Certainly you hope most casino games adhere to this principle).

Consider a coin toss or Roulette table. The WLLN says, for example, that the probability of a sequence of coin tosses coming up heads more than 50.01% of the time approaches 0 the more tosses you make.

A Challenge to Markov

Pavel Nekrasov: “If a sequence of random variables satisfies the weak law of large numbers, the sequence is independent.”

Andrey Markov: “...Is it tho?”

Early Example: Probability in Poetry

Markov looked at sequences of letters in Russian poetry. Even though some letters occur more than others, *the probability changed depending on what letter preceded it.*

English example: The letter u has a relative frequency in English of 2.88%. If you consider a string of letters in sequence in a piece of English writing, the probability of a given letter being u is approximately 0.0288.

However, if you look at the probability of a letter being u, *and you know the letter preceding it*, that probability changes—for example, the probability of a u occurring following a q is much higher!

This sequence of random variables is clearly not independent. However, it *does* satisfy the weak law of large numbers!

Pushkin's vowels and consonants

Markov compiled the first 20,000 letters of Pushkin's *Eugene Onegin* into a string of characters and counted the total number of vowels (8,638) and consonants (11,362).

Markov then combed through the string and looked for successive vowel-vowel pairs. He found 1,104 double-vowels and 3,827 double-consonants.

If this sequence of lettters were independent, since vowels appear approximately 43% of the time, we'd expect the probability of having double-vowels be $(0.43)^2 \approx 0.19$, yielding 3,731 double vowels—over $3\times$ the actual number of double-vowels!

Modern Probability Definition

Let $E \subseteq \mathbb{R}$, and suppose $X = \{X_t\}_{t \in \mathcal{I}}$ is a stochastic process, where $\mathcal{I} \subseteq [0, \infty)$ is closed under addition.

X is a **Markov process** (or a **Markov chain** if $\mathcal{I} = \mathbb{N}_0$) if, for every measurable $A \subset E$ and every $r, s, t \in \mathcal{I}$ with $r < s < t$, we have:

$$\mathbb{P}\left[X_t \in A \,\middle|\, \{X_q\}_{r \leq q \leq s}\right] = \mathbb{P}\left[X_t \in A \,|\, X_s\right]$$

In other words, the outcome of experiment t, given knowledge of experiment s, has the same probability distribution as one would have given knowledge of *all experiments preceding and including* s. This is often called **memorylessness**.

Stochastic Matrices

Suppose you conduct an experiment with n possible outcomes, and the probability of going to state j from state i is p_{ij}. These probabilities can be assembled in a *stochastic matrix*. Observe that if you go from any state i,

$$p_{i1} + p_{i2} + \cdots + p_{in} = 1$$

A **stochastic matrix** (sometimes *Markov matrix*) is a square matrix $\Pi = (p_{ij})_{1 \leq i,j \leq n}$ of nonnegative entries whose sum of the entries in each row is equal to 1: $p_{i1} + \cdots + p_{in} = 1$.

Stochastic Matrices in Markov Chains

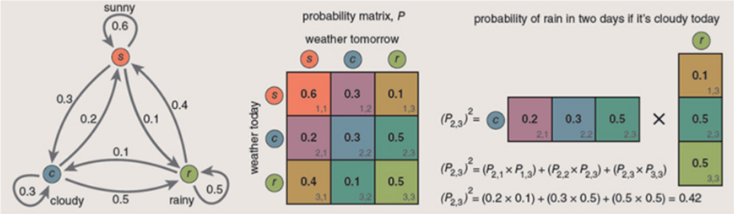
Why a matrix? Suppose we want to know the probability of going to state j in *two* iterations, if we're currently at state i. Then:

$$\begin{aligned}
\mathbb{P}\left[X_{\ell+2} = j \mid X_n = i\right] &= \sum_{k=1}^{n} \mathbb{P}\left[(X_{\ell+2} = j) \cap (X_{\ell+1} = k) \mid X_\ell = i\right] \\
&= \sum_{k=1}^{n} \mathbb{P}\left[X_{\ell+2} = j \mid X_{\ell+1} = k, X_\ell = i\right] \\
&\qquad \times \mathbb{P}\left[X_{\ell+1} = k \mid X_\ell = i\right] \\
&= \sum_{k=1}^{n} \mathbb{P}\left[X_{\ell+2} = j \mid X_{\ell+1} = k\right] \\
&\qquad \times \mathbb{P}\left[X_{\ell+1} = k \mid X_\ell = i\right] \\
&= \sum_{k=1}^{n} p_{kj} p_{ik} = \left(\Pi^2\right)_{ij}
\end{aligned}$$

Meteorology

Question. If we know the weather today, what is the probability distribution of possible weather behaviors tomorrow?

Obviously the weather today impacts the weather tomorrow. This is often worked out using a *Markov graph*, whose edges have different weights, described in a **stochastic matrix**:



Higher-Order Markov Chains

What if in fact outcome ℓ depends not only on where you are at step $\ell - 1$, but from where you are at step $\ell - 2$ also? (Consider letters: the probability of *s* appearing in an English word is 0.0628, but this increases if it's preceded by *u*—and increases further if preceded by *ou*.)

An **Markov chain of order** m with values in $E \subseteq \mathbb{R}$ is a discrete stochastic process $\{X_\ell\}_{\ell \geq 1}$ where for every measurable $A \subset E$, every $\ell \geq 1$, and every $0 \leq k \leq \ell - m$,

$$\mathbb{P}\left[X_\ell \in A \mid \{X_j\}_{k \leq j \leq \ell-1}\right] = \mathbb{P}\left[X_\ell \in A \mid \{X_j\}_{\ell-m \leq j \leq \ell-1}\right]$$

Here, the probability of the next step depends on where you currently are, and where you were for the past $m - 1$ steps.

Drivel me this...

First order

Theg sheso pa lyiklg ut. cout Scrpauscricre cobaives wingervet Ners, whe ilened te o wn taulie wom uld atimorerteansouroocono weveiknt hef ia ngry'sif farll t mmat and, tr iscond frnid riliofr th Gureckpeag

Third order

At oness, and no fall makestic to us, infessed Russion-bently our then a man thous always, and toops in he roguestill shoed to dispric! Is Olga's up. Italked fore declaimsel the Juan's conven night toget nothem,

Fifth order

Meanwhile with jealousy bench, and so it was his time. But she trick. Let message we visits at dared here bored my sweet, who sets no inclination, and Homer, so prose, weight, my goods and envy and kin.

Seventh order

My sorrow her breast, over the dumb torment of her veil, with our poor head is stooping. But now Aurora's crimson finger, your christening glow. Farewell. Evgeny loved one, honoured fate by calmly, not yet seeking?

Topological Markov Chains

In probability theory, we often forget the underlying measure space $(\Omega, \mathcal{A}, \mathbb{P})$.

A **dynamical system** is a measure space $(\Omega, \mathcal{A}, \mu)$ or a metric space (Ω, d), with a map $T : \Omega \to \Omega$, typically either continuous, measurable, smooth, etc.

A **(dynamical) Markov chain** is a dynamical system with $\Omega = \{1, \ldots, n\}^{\mathbb{Z}}$, and symbolic metric

$$d(\omega_1, \omega_2) = 2^{-\min\{|k| \omega_1^k \neq \omega_2^k\}}$$

with map $T : \Omega \to \Omega$ defined by $T(\omega)^k = \omega^{k+1}$.

A **topological Markov chain** is a closed subset of Ω invariant under the map T.