Markov Chain Tree Theorem and Other Problems

Isaac M. Sonin

Department of Mathematics and Statistics
University of North Carolina at Charlotte

http://...type in Google Isaac Sonin

UMBC
March 30, 2018

Outline

Other topics

- Are all independent events created equal ?
- Decomposition-Separation Theorem

This talk

- MC Tree Theorem
- State Elimination
- Optimal Stopping (OS) of Markov Chains (MCs)

Independence and simple random experiment

A. N. Kolmogorov wrote (1933, Foundations of the Theory of Probability):

"The concept of mutual *independence* of two or more experiments holds, in a certain sense, a central position in the theory of Probability."

$$P(AB) = P(A)P(B) \tag{1}$$

1. Let us consider the following simple random experiment: first we *flip a fair coin* and then we *toss a fair die*. Our sample space consists of 12 outcomes each having a probability of $1/12$. This experiment is used in many textbooks as an illustration of the concept of independent events.

Question 1. How many different pairs (A, B) of independent events are there ?

Answer

And the answer is

$$K_1 = 888,888$$

Of course, most of these pairs and tuples are isomorphic and can be obtained in a small number of different ways or patterns.

$$888,888 = 4 * n_1 + 2 * n_2 + 2 * n_3 + n_4, \quad n_1 = 12!/(1!2!3!4!),$$
$$n_2 = 12!/(1!1!5!5!), \quad n_3 = 12!/(2!2!4!4!), \quad n_4 = 12!/(3!3!3!3!).$$

Let us suppose now that a coin and a die are slightly *biased*. Then, it is easy to check that for almost all biased coins and dice the number K_1 is reduced to the more "normal" looking number $124 = (2^2 - 2) \cdot (2^6 - 2) = 2 \cdot 62$.

In all cases we count only proper pairs (A, B), that is when none of the sets is the empty set or the whole sample space.

What is the difference between these two groups of events ?

Not all pairs of independent events are created equal

These 124 pairs are "**stable**", i.e. they are not affected by the changes in probabilities for coin and die, the probabilities of "random generators ", (RGs).

For a fair die and a fair coin the overwhelming majority (99.99%) of independent pairs are " unstable ", i.e. they disappear no matter how small the bias is.

This seems to suggest that *not all pairs of independent events are created equal*.

The sample space, with 12 equally likely points can be represented also as a product of 3-die (a die with three sides) and 4-die, or as a product of two coins and a 3-die. The the number of stable pairs will be even smaller, $(2^3 - 2) \cdot (2^4 - 2) = 84$ and $2 \cdot 2 \cdot (2^3 - 2) = 24$.
And, of course, *all* independent pairs may disappear if we change, ever slightly, the probabilities of 12 outcomes in the sample space.

...i.e. **we have to consider stability with respect to RGs.**

More Problems

2. Let us recall the famous example of S. Bernstein, or any other equivalent example, about tetrahedron with four *symmetric* sides of different colors $(G, B, R, (GBR))$, producing three events $A, B,$ and C, such that they are *dependent* but *pairwise independent*. Do "real" tetrahedrons of such kind exist ?

No ! This example is strongly unstable !

Theorem. There is no partial independence.

...i.e. always unstable. This Theorem confirms what Feller said in his famous book, after the definition of independence. " There is no practical cases of indep. events which are pairwise indep. but not indep."

3. Every finite sample space is either indecomposable or can be represented as the direct product of indecomposable sample spaces. Is such a decomposition unique ? **No !**

Example 1. Let $N = 6$ and the probability mass function is given as follows: $\{1, 2, 4, 8, 16, 32\} * \frac{1}{63}$. Then it has two *distinct* representation as a product of a coin and a die with three sides. *Open problem*:...

Example 2. Let $N = 5$ and the probability mass function is given as

Kolmogorov again

Another citation from A. N. Kolmogorov (1933) at the end of a subsection on independence, which gives a hint that Kolmogorov may have foreseen the subtle difference between the formal definition of independence and its more "physical " interpretation :

"In consequence, one of the most important problems in the philosophy of the natural sciences is - in addition to the well-known one regarding the essence of the concept of probability itself - to make precise the premises which would make it possible to regard any given real events as independent. *This question, however, is beyond the scope of this book.* "

Italics my.

Kolmogorov A.N., 1956. Foundations of the Theory of Probability. NY, Chelsea Publ. Company. (appeared in 1933)

(Isaac M. Sonin, Independent Events in a Simple Random Experiment and the Meaning of Independence, 2006, 2012, arXiv:1204.6731).

Markov Chains

The classical Kolmogorov-Doeblin results describing the asymptotic behavior of MCs can be found in most advanced books on probability theory.
According to these results the state space S can be *decomposed* into the set of nonessential states and the classes of essential communicating states. Furthermore, the following are true:

(**A**) With probability one, each trajectory of a MC Z from U_0 will reach one of these classes and never leave it.
Each class can be decomposed into cyclical subclasses. If the number of subclasses is equal to one (an aperiodic class), then

(**B**) every MC Z from U_0 has a mixing property inside such a class, i.e. there exists a limit distribution π which does not depend on the initial distribution μ and such that π is invariant with respect to the matrix P.

What is a *nonhomogeneous MC* ? Replace matrix P by a sequence (P_n).

Decomposition-Separation Theorem

A pair $M = (S, P)$, where S is a state space and P is a stochastic matrix is called a *Markov model*. (Z_n) *Markov chain* (MC). Classical Kolmogorov-Doeblin decomposition of S into ...

The "natural" question is: what happens with this theory and with this decomposition if we replace a stochastic matrix P by a *sequence* of stochastic matrices (P_n) ?

There are *no assumptions* about the sequence (P_n) !

The answer is given by the *Decomposition-Separation (DS) Theorem.*

Kolmogoroff A. N. (1936), Blackwell D. (1945), Cohn H. (...,1976, 1989),

Sonin I. (1987, 1996, 2008 IMS, v.4.)
The Decomposition-Separation Theorem for Finite Nonhomogeneous Markov Chains and Related Problems, IMS Collections, Markov Processes and Related Topics: A Festschrift for Thomas G. Kurtz Vol. 4 (2008), 1-15.

Nonhomogeneous Markov Chains as Colored Flows

The following simple physical model and physical interpretation of the DS Theorem was introduced by I. Sonin in..1987. Given a sequence (M_n), let M_n represent a set of “cups” containing a “liquid ” - *tea*, schnapps, vodka etc. A cup $i \in M_n$ is characterized at moment n by a *volume* of liquid in this cup, $m_n(i)$. The matrix P_n describes the redistribution of liquid from the cups M_n to the (initially empty) cups M_{n+1} at the time of the n-th transition, i.e. $p_n(i, j)$ is the proportion of liquid transferred from cup i to cup j. The sequence $(m_n), m_n = (m_n(i), i \in M_n), n \in \mathbf{N}$, satisfies the relations

$$m_{n+1} = m_n P_n, m_3 = m_0 P_0 P_1 P_2, \tag{2}$$

where m_n is a stochastic *row* vector. Let us assume additionally that each cup contains some material (substance, color) and let us denote $\alpha_n(i), 0 \leq \alpha \leq 1,$ a "*concentration*" of this material at cup i at moment n. The sequence $(m_n, \alpha_n) = (m_n(i), \alpha_n(i)), i \in M_n, n \in \mathbf{N})$, for the sake of brevity is called (discrete) *colored flow.*

Concentrations are Martingales in Reverse Time

Concentrations obviously satisfy the relations

$$\alpha_{n+1}(j) = \sum_i m_n(i)\alpha_n(i)p_n(i,j)/m_{n+1}(j). \tag{3}$$

Note that we can replace the notion of concentration by *temperature* since it follows the same formula. One more interpretation...
A random sequence (Y_n) specified by

$$Y_n = \alpha_n(Z_n), n \in \mathbf{N}, \tag{4}$$

where $\alpha_n(i)$ *s*' are given by ..., is a *(sub)martingale in reverse time*. This simple fact is the bridge between the DS theorem and the Theorem about the existence of *barriers*.
One of the most remarkable and widely used results in the theory of stochastic processes is the theorem of Doob about the existence of the limits of trajectories of bounded (sub)martingale when time tends to infinity. This theorem is based on Doob's upcrossing lemma.

Doob's Lemma and its modification

Doob's upcrossing lemma. *If $Y = (Y_n)$ is a bounded (sub)martingale then the expected number of intersections of every fixed interval $a, b)$ by the trajectories of Y is finite on the infinite time interval.*
The width of the interval $(b - a)$ is in the denominator of the corresponding estimate so Doob's lemma does not imply for example that inside the interval there exists a*level* such that the expected number of intersections of this level is finite.
If (Y_n) takes values in (M_n), then Doob's lemma can be substantially strengthened. Let us call a nonrandom sequence (d_n) a *barrier* for the random sequence $Y = (Y_n)$ if the *expected number* of intersections of (d_n) by the trajectories of X is finite, i.e...
Theorem in Sonin (1987) about the existence of barriers for processes with finite variation and which take only a bounded number of values implies the Separation part of the DS Theorem.

DS Theorem. The elementary (deterministic) formulation

Let a sequence of disjoint sets (M_n), *satisfying condition* $|M_n| \leq N$ *and a sequence of stochastic matrices* (P_n) *be given. Then there an integer* $c, 1 \leq c \leq N$, *and there exists a decomposition of the sequence* (M_n) *into disjoint jets* $J^0, J^1, ..., J^c, J^k = (J^k_n)$, *such that for any colored flow* (m_n, α_n, O_n)

(**a**) *the stabilization of volume and concentration take place inside of any jet* J^k, $k = 1, ..., c$,

i.e. $\lim_{n\to\infty} \sum_{i \in J^k_n} m(i) = m^k_*$; $\lim_{n\to\infty} \alpha(i_n) = \alpha^k_*, i_n \in J^k_n$;

the concentration in jet J^0 *may oscillate; the total volume in this jet tends to zero, i.e.* $\lim_{n\to\infty} \sum_{i \in J^0_n} m(i) = 0$;

(**b**) *the total amount of liquid transferred between any two different jets is finite on the infinite time interval, i.e.* $V(J^k, J^s | m) < \infty, s \neq k$.

(**c**) *this decomposition is unique up to jets* (J_n) *such that for any flow* (m_n) *the relation* $\lim_n m_n(J_n) = 0$ *holds and the total amount of liquid transferred between* (J_n) *and* $(M_n \backslash J_n)$ *is finite.*

Consensus. Terms

Consensus Algorithms and the Decomposition-Separation Theorem, Sadegh Bolouki and Roland P. Malhame, IEEE Transactions on Automatic Control, vol. 61, no. 9, September 2016.
A multi-agent system, in the most general sense, is a network of multiple interacting *agents*. Each agent is assumed to hold a *state* regarding a certain quantity of interest. Depending on the context, states may be referred to as *opinions, values, beliefs, positions, velocities*, etc. States of agents are updated based on an algorithm or protocol which is an interaction rule specifying the interaction between each agent and its neighbors. Global *consensus*, or simply consensus, in the system is defined as *convergence of all states to a common value* over time. Among all update algorithms in multi-agent systems, *distributed averaging algorithms* are of great importance and have been discussed the most in the literature. Such algorithms impose that the state of each agent is updated according to a convex combination of the current states of its neighbors and its own.

Cucker & S Smale (2007) positions, velocities, nonlinear interaction phase transitions.

Consensus. More areas of application

In *biology* - behavior of bird flocks, fish schools, humans etc

In *robotics and control*, consensus problems arise in relation to coordination objectives and cooperation of mobile agents (e.g.,robots and sensors)

In *economics*, seeking an agreement on a common belief in a price system is another example of consensus. Gas prices in Ch-te

In *sociology*, the emergence of a common language in primitive societies
In *social networks*, consensus algorithms can shed light on the dynamics of opinion formation.

In *computer science* - networks; management science
In a multi-agent system, it is possible that agents separate into several clusters such that consensus occurs within each cluster. In this case, *multiple consensus* is said to have occurred.

Gossip algorithms. In a gossip models the frequency of information exchange is controlled by an internal clock ticking according to a timing model. In each step, each agent transmits its information (state) to another agent which is chosen randomly.

Consensus. Questions

Consider a system composed of N agents that are labeled by numbers $1, ..., N$. Let $f_n(i)$ be the scalar state of agent i at time n. *Distributed averaging algorithms* can be defined in both continuous and discrete times. A general discrete time *averaging algorithm* is defined by

$$f_{n+1} = P_n f_n, f_3 = P_2 P_1 P_0 f_0, \tag{5}$$

where f_n is the **column** vector of states at each time instant n. *Consensus* is now defined by the convergence of vector f_n to a vector with equal components as n.... *Multiple consensus* is also defined as the existence of a limit for $f_n(i)$ for each agent i as time grows large. The limits may differ for different agents.

The following two fundamental questions regarding the issue of consensus:

Q 1. Under what conditions on the underlying chain of the system, consensus or multiple consensus is guaranteed *irrespective* of the time and values that states are initialized ?

Consensus. Questions

Q 2. For a general underlying chain, having fixed the initial time, what is the *set of initial conditions* resulting in the occurrence of consensus in the system ?

Q 1 is equivalent to a property of the underlying chain called *ergodicity*. multiple consensus is equivalent to another property of underlying chain called *class-ergodicity.* Chain is class-ergodic if the limit matrix exists, but in general possibly with distinct rows.

Q 3. the question arises as to whether it is possible, for a limited number of *key agents*, to set their initial opinion/parameter assessment, in such a way that the (exogenously evolving) network converges to a global consensus.
Such an issue is important in negotiations, or even the possible shaping or manipulation of public opinion by clever campaigning. notion of *minence grise coalition.* Gray Cardinals.

For which MCs the consensus occurs ? It depends.

MC Tree Theorem

Let S be a finite set and P be a stochastic irreducible matrix. Let T be a *spanning tree directed* to y. This means that T is a connected graph without cycles (tree), contains all the vertices of S (spanning), and that a vertex y is designated as a *root*. In any rooted tree with a root y there is a unique path between any vertex v and y "directed" to y, and this direction makes the tree *a tree directed to* y. Denote $G(y)$ the set of all spanning trees on S directed to y. Let us define

$$q(y) = \sum_{T \in G(y)} r(T), \quad \text{where } r(T) = \prod_{(u,v) \in T} p(u,v). \tag{6}$$

Then

$$\pi(x) = \frac{q(x)}{\sum_{y \in S} q(y)}, \tag{7}$$

Vector $q = (q(y), y \in S)$ is called the *Rooted Spanning Tree* (RST) *vector*.

In the classical theorems of G. Kirchhoff (1847, undirected graphs) and W. Tutte (1948, directed graphs) RST vector with $p(u,v) = 1$ gives a number of spanning trees and is calculated as a determinant of so called Laplacian Matrix.

Elimination - a key operation in MC

An important and traditional tool for the study of Markov chains (MCs) is the notion of a *Censored (Embedded)* MC.

Let us assume that a Markov model $M = (S, P)$ is given and $D \subset S$, $C = S \setminus D$. Then the matrix $P = \{p(x, y)\}$ can be decomposed as follows

$$P = \begin{bmatrix} Q & T \\ R & P_0 \end{bmatrix}, \tag{8}$$

where the substochastic matrix Q describes the transitions inside of D, P_0 describes the transitions inside of C and so on.

Let (Z_n) be a MC defined in model M, and observed only in the set C.

Formally : consider the sequence of Markov times $\tau_0, \tau_1, ..., \tau_n, ...,$ where $\tau_0 = 0$, and $\tau_n, n \geq 1$ are the times of first, and so on, *return* of the MC (Z_n) to the set C, i.e., $\tau_{n+1} = \min\{k > \tau_n, Z_k \in C\}$,

$Y_n = Z_{\tau_n}, n = 0, 1, 2, ...$

The strong Markov property and standard probabilistic reasoning imply the following basic lemma which should probably be credited to Kolmogorov and Doeblin.

Basic Lemma (Kolmogorov, Doeblin)

Elimination Lemma. (a) The random sequence $(Y_n), Y_n = Z_{\tau_n}, n = 0, 1, 2, \ldots$ is a Markov chain in a model $M^D = (S, P^D)$, (or in $M_D = (C, P_D)$) where
(b) the transition matrix $P^D = \{p^D(x, y), x, y \in S\}$ is given by the formula

$$P^D = \begin{bmatrix} 0 & NT \\ 0 & P_0 + RNT \end{bmatrix}. \tag{9}$$

Here $U = NT$ is the matrix of the distribution of the MC at the time of the first return (visit) to C starting from $x \in D$, $N = N(D)$ is the *fundamental* matrix for the substochastic matrix Q, i.e. $N = \sum_{n=0}^{\infty} Q^n = (I - Q)^{-1}$, where I is the identity matrix. This representation is given, for example, in the classical text Kemeny & Snell. The matrix $N = N(D) = \{n(x, y), x, y \in D\}$ has a well-known probabilistic interpretation, $n(x, y)$ is the *expected number of visits* to y starting from x until the time τ_1 of the *first return* to set C.
Let us mention also, that there is an **Insertion Lemma**, when any state, eliminated previously, can be restored (inserted) *in one iteration*. This is a new operation in the theory of MCs !

Elimination continues

The rows of matrix $P_D = P_0 + RNT$ give the distribution of MC (Z_n) at the time τ_1 and $P_x(Z_{\tau_1} = y) = p_D(x, y)$, $x \in S, y \in C$.
For $x \in D$, distribution is given by submatrix $P_0 + RNT$.

An important case is when the set D consists of one nonabsorbing point z. In this case formula (9) is replaced by the one-state elimination formula, written here for columns, $(P \equiv P_1, P_{\{z\}} = P_2)$,

$$p_2(\cdot, z) = 0, \quad p_2(\cdot, y) = p_1(\cdot, y) + p_1(\cdot, z) n_1(z) p_1(z, y), y \neq z, \tag{10}$$

where
$n_1(z) = \sum_{n=0}^{\infty} p_1^n(z, z) = 1/s_1(z), s_1(z) = 1 - p_1(z, z) = \sum_{u \neq z} p_1(z, u)$.

This transformation (written for rows) is similar to one step of Gaussian elimination and requires $O(n^2)$ operations. We say that matrix P_2 is obtained from P_1 in one *iteration*. Thus matrix P_D can be calculated directly by (9) or recursively using formula (10) in $|D|$ iterations.

Optimal Stopping (OS) of Markov Chains (MCs)

Optimal stopping of stochastic processes...Options pricing for American options. Not many general results...Snell's Envelope

T. Ferguson: "Most problems of optimal stopping without some form of Markovian structure are essentially untractable."

OS Model $M = (X, P, c, g, \beta)$, discrete time

- X finite (countable) state space,
- $P = \{p(x, y)\}$, stochastic (transition) matrix
- $c(x)$ one step cost function,
- $g(x)$ terminal reward function,
- β discount factor,$0 \leq \beta \leq 1$
- (Z_n) MC from a family of MCs defined by a Markov Model $M = (X, P)$
- $v(x) = \sup_{\tau \geq 0} E_x[\sum_{i=0}^{\tau-1} \beta^i c(Z_i) + \beta^\tau g(Z_\tau)]$, value function

Description of OS Continues

- **Remark !** absorbing state e, $p(e,e) = 1$, $p(x,y) \longrightarrow \beta p(x,y)$, $p(x,e) = 1 - \beta$. Standard trick

  $\beta \longrightarrow \beta(x) = P_x(Z_1 \neq e)$ probability of "survival".
- $S = \{x : g(x) = v(x)\}$ optimal stopping set.
- $Pf = Pf(x) = \sum_y p(x,y)f(y)$.

Theorem (Shiryayev 1969)

(a) The value function $v(x)$ is the minimal solution of Bellman equation ...

$$v = \max(g, c + Pv),$$

(b) if state space X is finite then set S is not empty and $\tau_0 = \min\{n \geq 0 : Z_n \in S\}$ is an optimal stopping time. ...

State Elimination Algorithm for OS of MCs

Initial model $M_1 = (X_1, P_1, c_1(x), g(x), \beta_1(x)), g$ without subindex

$$g(x) - (P_1 g(x) + c_1(x)) = g - T_1 g$$

$$\swarrow \qquad \searrow$$

$g(x) - T_1 g(x) \geq 0$ for *all* x | there is $z : g(z) - T_1 g(z) < 0$

$$\Downarrow \qquad\qquad \Downarrow$$

$X_1 = S$ | $M_1 \longrightarrow M_2 : g(x) - T_2 g(x)$

$$\swarrow \qquad \searrow$$

... and so on

$$p_2(x, y) = p_1(x, y) + p_1(x, z) n_1(z) p_1(z, y),$$

$$c_2(x) = c_1(x) + p_1(x, z) n_1(z) c_1(z),$$

where $n_1(z) = 1/(1 - p_1(z, z))$. Similar Matrix formulas $P_2 = P_1 + ...$

Recursive Calculation of RST vectors

The algorithm to calculate the Rooted Spanning Tree vector $q(y)$ is an immediate corollary of the following fundamental theorem, generalized into Idempotent Calculus framework in GKMS 2015.

Theorem (1999, 2015, 2017) *Let* $M_1 = (S_1, P_1)$ *be a (finite irreducible Markov) model,* $z \in S_1$*, and let* $M_2 = (S_2, P_2)$*, be a model obtained by of state z, i.e.,* $S_2 = S_1 \setminus z$*, and* P_i*,* $i = 1, 2$ *are defined as above. Let* $(q_i(y), y \in S_i), i = 1, 2$ *be RST vectors calculated by formula (6) for both models, i.e.*

$$q_1(y) = \sum_{T' \in G_1(y)} \prod_{(u,v) \in T'} p_1(u, v), \quad q_2(y) = \sum_{T \in G_2(y)} \prod_{(u,v) \in T} p_2(u, v).$$

Then

$$q_1(y) = s_1(z) q_2(y), \; y \neq z, \quad q_1(z) = \sum_{y \in S_2} q_2(y) p_1(y, z),$$

where $s_1(z) = \sum_{v \neq z} p_1(z, v)$, not $s_1(z) = 1 - p_1(z, z)$ anymore.
But now in this theorem $p_i(u, v), i = 1, 2$ are just *elements*, (symbols, variables), that can be added, multiplied and divided, and we can consider $R_i(T) = \prod p_i(u, v)$ as *generating functions* on a trees !

Probabiliy Theory and Mathematics

This extension of MCs Theory into Idempotent Calculus framework confirms a remarkable foreseeing of A.N. Kolmogorov, who wrote in almost unknown paper in 1926 written for a volume addressed to a general public.

> *Probability theory has become a topic of interest in modern mathematics not only because of its growing significance in natural sciences, but also because of the gradually emerging deep connections of this theory with many problems in various fields of pure mathematics. It seems that the formulas of probability calculus express one of the fundamental groups of general mathematical laws.*
>
> A. N. Kolmogorov

His words seem even more remarkable, if a reader recalls that in 1926 A.N. Kolmogorov was only 23 years old, his fundamental treaties was not written yet and it was only the third year when he became interested in Probability Theory.

Rutgers University, 3rd Applied Probability Conference, June 2014:
Sheldon Ross, Isaac Sonin, John Gittins, Michael Katehakis

Some References

Katehakis, M., Veinot, A., (1987), Whittle, P.(1980), Varaiya, P. et al (1985).

B. Benek Gursoy, S. Kirkland, O. Mason and S. Sergeev,
The Markov Chain Tree Theorem in commutative semirings and the State Reduction Algorithm in commutative semifields,
Linear Algebra and its Applications, 468(1) (2015), pp. 184-196

S. Fomin, D. Grigoriev, G. Koshevoy,
Subtraction-free complexity, cluster transformations,and spanning trees,
Foundations of Computational Mathematics, 16(1) (2016), pp. 1–31

Books: Gittins, J., (1979), Gittins J., Glazebrook K., Weber R., Multi-armed Bandit Allocaton Indices, 2nd edition, Wiley, 2011.

E. Presman, I. Sonin, Sequential control with incomplete information. The Bayesian approach to multi-armed bandit problems, (with). Academic Press, Inc., San Diego, CA, 1990.

Algorithms based on Censored MCs

A short list of areas with algorithms based on Censored MCs:

- Optimal Stopping of Markov Chains (MCs)
- Optimal Stopping of Random Sequences Modulated by a MC
- Gittins index and Generalized GI. Abstract Optimization
- GTH/S (Grassman, Taksar, Heiman/Sheskin) algorithm to calculate the invariant distribution for ergodic MC
- Invariant in Islands and Ports model
- Continue, Quit, Restart model
- MC Tree Theorem.

The references can be found on my website, type in Google Isaac Sonin

Thank you for your attention !