Journal of Statistics and Computer Science
Vol. 1, Issue 2, 2022, pp. 135-145
https://doi.org/10.47509/JSCS.2022.v01i02.03

Approximating General Discrete Stochastic Processes by Markov Chains

András Faragó
Department of Computer Science
The University of Texas at Dallas
800 W. Campbell Rd., Richardson, TX 75080, USA
E-mail: farago@utdallas.edu

Abstract

Many modeling tasks in stochastic systems and networks lead to discrete stochastic processes. In lucky cases, the stochastic process is a Markov chain, for which well elaborated mathematical machinery is available. Occasionally, however, one may encounter more complex situations when the Markov property does not hold. This is the case, for example, when the system exhibits long-range dependencies. Another example is when the system depends on some random initial condition, which gives rise to different behavior, such as different transition probabilities, yielding a mixture of Markov chains, rather than a single one. Yet another situation is when the current state of the system may be correlated with its future evolution, it does not exclusively depend only on the past. In these and other non-Markovian instances significantly fewer general methods are available to serve the analysis. We present an approach to (partially) overcome this difficulty. Specifically, we consider the approximation of general discrete stochastic processes by Markov chains. We prove that this approach allows the application of many pieces of Markov chain based analysis methods and algorithms to the more general case, thus usefully extending the application domain of a number of well-known methods and algorithms.

Keywords: Discrete stochastic process, Markov chain, approximation.

1 Introduction

Analysis of stochastic systems is often based on models that apply the mathematical technique of Markov chains (or Markov processes when continuous time is considered). Once we are able to set up a Markovian model, we can investigate both the stationary and transient behavior of the system, using well established methods. A classic example is the rich analysis of loss networks in telecommunications, see Kelly (1991).

Article History
Received : 10 October 2022; Revised : 30 October 2022; Accepted : 18 November 2022; Published : 15 December 2022

To cite this paper

András Faragó (2022). Approximating General Discrete Stochastic Processes by Markov Chains. *Journal of Statistics and Computer Science*. 1(2), 135-145.

In some cases, however, a Markov model cannot adequately capture the behavior of the system. There are several possible reasons for this. Below we list a few examples, typically resulting in non-Markovian behavior:

- The system may exhibit *long-range dependencies.* This is the case, for example, with *self-similar* traffic patterns that are often observed in a number of important networks, including the Internet, the World Wide Web, and various local area networks, such as Ethernet.
- The system may depend on some random initial condition, which gives rise to different behavior, such as different transition probabilities, yielding a *mixture* of Markov chains, rather than a single one. Generally, such a mixture gives rise to a process that is not a Markov chain. For such mixtures see interesting results in Faragó (2021).
- Yet another situation is when the current state of the system may depend on its future evolution, not determined solely by the past, not even in a probabilistic sense. For example, we may randomly select a trajectory from a set of possible trajectories, in a way that violates the Markov property.

In this paper, which is an updated version of the technical report Faragó (2020), we describe an approach that can help the analysis of non-Markovian models, and brings back the possibility to apply results that are routinely used for Markov chains.

2 General Setting: Discrete Stochastic Processes

Let us consider stochastic processes with discrete time and finite state space, without assuming that they are Markov chains. For brevity, we call such a process a *discrete stochastic process*. We use the following notations:

- A discrete stochastic process: $X = (X_t,\ t = 0, 1, 2, \ldots)$. Observe that we start counting the time from $t = 0$. Accordingly, X_0 is called the *initial state* of the process.
- The state space of the process is assumed finite, and is denoted by S. Each X_t takes its values in S. The finiteness assumption can be relaxed, we just adopt it here to avoid complications that would only obscure the main message.
- The probability distribution of X_t is denoted by π_t, which is identified with a vector in $[0, 1]^{|S|}$. (In matrix expressions it will be regarded a row vector.) We call these distributions the *one-dimensional distributions* of the process.
- We define the *first-order transition probability matrix* (or, simply, *transition probability matrix*) of X at time t by

$$P_t = [p_t(a, b)]_{a,b\in S} = [\Pr(X_{t+1} = b \,|\, X_t = a)]_{a,b\in S}. \tag{1}$$

Note that these transition probabilities are routinely used for Markov chains, but such conditional probabilities can be defined for any discrete stochastic process. Observe, however, that if the process is not a Markov chain, then, generally, the value of $\Pr(X_{t+1} = b \,|\, X_t = a)$ is not independent of previous history, i.e., it may hold that

$$\Pr(X_{t+1} = b \,|\, X_t = a) \;\neq\; \Pr(X_{t+1} = b \,|\, X_t = a, X_{t-1} = a_{t-1}, \ldots, X_0 = a_0) \qquad (2)$$

which we refer to as *history dependence*.

- If P_t is independent of t, then we call the process *first-order homogeneous*. In this case all P_t matrices can be replaced by the single matrix

$$P = [p(a,b)]_{a,b\in S} = [\Pr(X_{t+1} = b \,|\, X_t = a)]_{a,b\in S}.$$

Observe that if P_t is independent of t, it does not mean that t cannot occur in the expression, as it already occurs in X_{t+1} and X_t. It only means that the probabilities cannot *directly* depend on t.

Note that first-order homogeneity generally does not imply the Markov property, so history dependence may still occur, i.e., we may still have (2). It is also worth mentioning that if the process is obtained from a stationary process, discarding the negative time instants, then it will be first-order homogeneous.

3 Markov projection of Discrete Stochastic Process

Now we introduce a useful concept, called *Markov projection.* Informally, it is a Markov chain that preserves some basic properties of the general discrete stochastic process, but it is not identical with it, as the latter is possibly not a Markov chain. In a sense, this concept projects the discrete stochastic process to the family of Markov chains, thereby approximating the original process with a Markov chain. As we are going to see, the approximating Markov chain (the Markov projection) is uniquely determined. The formal definition is presented below.

Definition 1 (Markov projection) *Let $X = (X_t,\ t = 0, 1, 2, \ldots)$ be a discrete stochastic process. The* Markov projection *of X is defined as a Markov chain $\widetilde{X} = (\widetilde{X}_t,\ t = 0, 1, 2, \ldots)$ that is generated as follows:*

- *Set $\widetilde{X}_0 = X_0$, i.e., $\widetilde{X}$ starts from the same initial state as X.*
- *Having obtained $\widetilde{X}_0, \ldots, \widetilde{X}_t$, the value of $\widetilde{X}_{t+1}$ is drawn by making an independent random transition from the value of $\widetilde{X}_t$, according to the transition probabilities in P_t, defined in (1).*

We are also going to use the terminology that X is the *parent process* of $\widetilde{X}$. It is clear from the definition that $\widetilde{X}$ is indeed a Markov chain, since it is generated such that whenever we are in a given state a at time t, we move into a state b with probability $p_t(a,b)$ and this

random choice is made, by definition, independently of the previous history. (Note that even if the original process exhibits history dependence, $p_t(a,b)$ is used as a *constant* probability for any given t, a, b.) Consequently, for every $a, b \in S$ and for every t

$$\Pr(\widetilde{X}_{t+1} = b \,|\, \widetilde{X}_t = a) = p_t(a,b) =$$

$$\Pr(\widetilde{X}_{t+1} = b \,|\, \widetilde{X}_t = a, \widetilde{X}_{t-1} = a_{t-1}, \ldots, \widetilde{X}_0 = a_0).$$

Thus, $\widetilde{X}$ has the same first-order transition probabilities as the parent process X, namely, $p_t(a,b)$. (On the other hand, generally this does not extend to higher order probability distributions if the parent process is not a Markov chain.) Furthermore, it is well known from the theory of Markov chains that the initial distribution and the (first-order) transition probabilities determine the chain uniquely, so there is no ambiguity when we talk about *the* Markov projection of a discrete stochastic process.

4 The Fundamental Property of the Markov projection

Let us now look at a key property of the Markov projection. It can be stated such that the one-dimensional distributions of the Markov projection and the parent process are the same. Thus, in this sense, the Markov projection indeed provides a Markov chain approximation of the original discrete stochastic process. Let us introduce a definition:

Definition 2 (First order equivalence) *Let*

$$X = (X_t,\ t = 0, 1, 2, \ldots) \quad \text{and} \quad \widetilde{X} = (\widetilde{X}_t,\ t = 0, 1, 2, \ldots)$$

be two discrete stochastic processes, with first order distributions $\pi_t, \widetilde{\pi}_t,\ t = 0, 1, 2, \ldots$, respectively. We say that X and $\widetilde{X}$ are first order equivalent*, if $\pi_t = \widetilde{\pi}_t$ holds for every $t = 0, 1, 2, \ldots$.*

Theorem 1 (Fundamental Property of Markov projection) *Every discrete stochastic process is first order equivalent with its Markov projection.*

Proof. We need to show that $\widetilde{\pi}_t = \pi_t$ holds for every t. Assume there is an integer τ with $\widetilde{\pi}_\tau \neq \pi_\tau$ and choose τ such that it is the smallest such integer. Since $\widetilde{\pi}_0 = \pi_0$ by definition (see definition 1), we have $\tau \geq 1$. Let us express $\pi_\tau(b)$ for an arbitrary $b \in S$. We can write, using the law of total probability:

$$\Pr(X_\tau = b) =$$

$$\sum_{a \in S} \Pr(X_\tau = b \,|\, X_{\tau-1} = a) \Pr(X_{\tau-1} = a).$$

With our notation this is

$$\pi_\tau(b) = \sum_{a \in S} p_{\tau-1}(a,b) \pi_{\tau-1}(a)$$

which in vector form gives

$$\pi_\tau = \pi_{\tau-1} P_{\tau-1}.$$

By the choice of τ we have $\widetilde{\pi}_{\tau-1} = \pi_{\tau-1}$, yielding

$$\pi_\tau = \widetilde{\pi}_{\tau-1} P_{\tau-1}. \tag{3}$$

On the other hand, as the first-order transition probabilities of X and $\widetilde{X}$ are equal by the defining construction, we obtain that in the Markov chain $\widetilde{X}$

$$\widetilde{\pi}_\tau = \widetilde{\pi}_{\tau-1} P_{\tau-1}. \tag{4}$$

holds. Comparing (3) and (4) results in $\widetilde{\pi}_\tau = \pi_\tau$, contradicting to the definition of τ. Thus, $\widetilde{\pi}_t = \pi_t$ must hold for every t.

♠

5 Consequences of the Fundamental Property

5.1 Trajectory Summation Formula

An important consequence of Theorem 1 is that some basic formulas that are routinely used for Markov chains, in fact remain valid for *arbitrary* discrete stochastic processes.

Corollary 1 *For every discrete stochastic process*

$$\pi_t = \pi_0 \prod_{i=0}^{t-1} P_i \tag{5}$$

holds. Furthermore, the probability $\Pr(X_t = a)$ *can be expressed as*

$$\Pr(X_t = a) \;=\; \sum_{a_0,\ldots,a_{t-1}} \Pr(X_0 = a_0) p_0(a_0, a_1) \cdot \ldots \cdot p_{t-1}(a_{t-1}, a) \tag{6}$$

where the summation is taken over all trajectories $a_0, a_1, \ldots, a_{t-1}, a$*. Moreover, if the process is first-order homogeneous (but still not necessarily Markov), then the above formulas simplify to*

$$\pi_t = \pi_0 P^t \tag{7}$$

and

$$\Pr(X_t = a) \;=\; \sum_{a_0,\ldots,a_{t-1}} \Pr(X_0 = a_0) p(a_0, a_1) \cdot \ldots \cdot p(a_{t-1}, a).$$

Proof. For the Markov projection of X the relationship $\widetilde{\pi}_t = \widetilde{\pi}_0 \prod_{i=0}^{t-1} P_i$ holds, being a Markov chain. By theorem 1 we have $\widetilde{\pi}_t = \pi_t$, implying (5). If we write down the details of the matrix product in (5), we get precisely (6). If X is first-order homogeneous, then $P_0 = P_1 = \ldots = P_t$ holds, too, yielding the second pair of formulas. ♠

Note that if X is a Markov chain (possibly not time-homogeneous), then the probability that we reach a_t via a given trajectory $a_0, a_1, \ldots, a_t$ is precisely the product

$$\Pr(X_0 = a_0) p_0(a_0, a_1) \cdot \ldots \cdot p_{t-1}(a_{t-1}, a_t) \tag{8}$$

due to the Markov property. Since reaching a_t via different trajectories are exclusive events and they represent all possibilities, therefore, summing up for all such possible products naturally gives the formula

$$\Pr(X_t = a) = \sum_{a_0,\ldots,a_{t-1}} \Pr(X_0 = a_0) p_0(a_0, a_1) \cdot \ldots \cdot p_{t-1}(a_{t-1}, a) \tag{9}$$

for Markov chains. On the other hand, if X is *not* a Markov chain, then the probability of traversing a given trajectory $a_0, \ldots, a_t$ may not be equal to (8) because of the effect of history dependence. Nevertheless, the trajectory *summation* formula (9) still remains valid, even though the individual summands may not be equal to the individual probabilities of the corresponding trajectories.

The key message of Corollary 1 can be summarized as follows:

Pseudo-Markovian behavior of discrete stochastic processes: *The one-dimensional distributions in a discrete stochastic process can be expressed by the transition probability matrices in the same way as in a Markov chain (see equations (5) and (7)). This holds even if the process does not satisfy the Markov property.*

5.2 Stationary Distribution and Ergodicity

Via the Markov projection, we can directly carry over a number of fundamental concepts and results from Markov chain theory to a more general setting.

Definition 3 *Let $X = (X_t,\ t = 0, 1, 2, \ldots)$ be a discrete stochastic process with state space S. Assume that X is first-order homogeneous (but possibly not Markov) and let its first-order transition probability matrix be P. Then we can introduce the following concepts, in analogy with Markov chains:*

- *A probability distribution π on S is called a* stationary distribution *of X if $\pi = \pi P$ holds.*
- *A process is called* ergodic *if it has a stationary distribution π, and the one-dimensional distribution π_t satisfies $\lim_{t\to\infty} \pi_t = \pi$.*
- *The process is called* irreducible *if there exists a positive integer k with $P^k > 0$, that is, every entry of the matrix P^k is positive.*
- *The process is called* aperiodic *if for every $a \in S$*

  $$\gcd\{m :\ p^{(m)}(a, a) > 0\} = 1$$

  holds, where the $p^{(m)}(.,.)$ are the entries of P^m, and gcd *means greatest common divisor.*

The concepts of Definition 3 are routinely used for Markov chains, but they do not actually require the Markov property, so they can be extended to arbitrary first-order homogeneous discrete stochastic processes.

Now we can analyze how the fundamental features of these concepts carry over from Markov chains to arbitrary first-order homogeneous discrete stochastic processes.

Theorem 2 *Let $X = (X_t,\ t = 0, 1, 2, \ldots)$ be a first-order homogeneous discrete stochastic process. Assume that X is irreducible and aperiodic (but possibly not Markov). Then the following hold:*

- *The process is ergodic, i.e., it has a unique stationary distribution π, and $\lim_{t\to\infty} \pi_t = \pi$ holds.*
- *Let $\widetilde{X}$ denote the Markov projection of X. Then $\widetilde{X}$ also has a unique stationary distribution $\widetilde{\pi}$. Moreover, $\widetilde{\pi} = \pi$, and the Markov projection is an ergodic Markov chain.*
- *The rate of convergence to stationary in X is the same as in the Markov projection $\widetilde{X}$. In particular, $\pi_t - \pi = \widetilde{\pi}_t - \widetilde{\pi}$ holds for every t.*

Proof. By the definition of the Markov projection $\widetilde{X}$ (see definition 1), the (first-order) transition probability matrix is the same for X and $\widetilde{X}$. Then by fundamental property of the Markov projection (Theorem 1), we have $\widetilde{\pi}_t = \pi_t$ for every t. The rest follows directly from the well known fundamental results of Markov chain theory on the stationary distribution and ergodicity, see, e.g., Aldous and Fill (2014), Kemeny (1960), Kijima (1997), Norris (1997).

♠

6 Examples

6.1 A Non-Markovian Process

Let $X = (X_t,\ t = 0, 1, 2, \ldots)$ be discrete stochastic process, in which the transition probability matrix satisfies

$$\Pr(X_{t+1} = b \,|\, X_t = a) \;=\; f(a, b, X_{t-1})$$

for $t \geq 1$, where $f : S^3 \mapsto [0, 1]$ is a fixed function. In other words, the probability of moving from a to b at time t depends not only on where the process was at time t, but also where it was at time $t - 1$.

This process is generally not Markov on the state space S, as we may have

$$\Pr(X_{t+1} = b \,|\, X_t = a) \neq \Pr(X_{t+1} = b \,|\, X_t = a, X_{t-1} = c),$$

for some $a, b, c \in S$, that is, the Markov property may not hold.

Assume further that the process satisfies the following connectivity property: from every state $a \in S$ any state $b \in S$ can be reached with positive probability.

Now we ask the question:

If π_t denotes the probability distribution of X_t, then can we conclude from the available information that π_t converges to a unique limit distribution (even if the process is not Markov)? Further, if such convergence is present, then what is the rate of convergence?

The answer can be obtained from our results. We can argue the following way:

- The expression for $\Pr(X_{t+1} = b \,|\, X_t = a)$ is the same for every $t \geq 1$ (does not directly depend on t), so the process is first-order homogeneous. (To handle $t = 0$, we can artificially introduce some value for X_{-1}, this has no asymptotic significance.)
- The condition that from every state $a \in S$ any state $b \in S$ can be reached with positive probability implies that the process is irreducible and aperiodic even if the process is not Markov (see Definition 3).
- Then from Theorem 2 we can conclude that the process is ergodic, i.e., it has a unique stationary distribution π, and $\lim_{t\to\infty} \pi_t = \pi$ holds. Note that all these were defined without the Markov property, see Definition 3.
- Again from Theorem 2, we can conclude that the rate of convergence to the stationary distribution is the same as in a Markov chain with transition probability matrix $[\Pr(X_{t+1} = b \,|\, X_t = a)]_{a,b\in S}$. In our example this amounts to the matrix

$$[\mathrm{E}(f(a, b, X_{t-1}))]_{a,b\in S}.$$

  Once this matrix is numerically known, we can apply the standard methods of Markov chain analysis.

The above example shows that our general framework can be used to derive useful conclusion, which may be much harder to prove directly from the specific features of the process.

6.2 Trajectory Correlations

Let $X = (X_t,\ t = 0, 1, 2, \ldots)$ be a first order homogeneous discrete stochastic process, with transition probabilities $p(a, b) = \Pr(X_{t+1} = b \mid X_t = a)$. As discussed in Section 2, such transition probabilities can be defined for every process, but they do not have to satisfy the Markov property, i.e., it is possible that

$$\Pr(X_{t+1} = b \,|\, X_t = a) \ \neq\ \Pr(X_{t+1} = b \,|\, X_t = a, X_{t-1} = a_{t-1}, \ldots, X_0 = a_0)$$

holds for some $a, b, a_{t-1}, \ldots, a_0 \in S$.

Consider now a trajectory $a_0, a_1, \ldots, a_t$, which is any sequence of states, such that one can move from each one to the next with positive probability. Let $\Pr(a_0, a_1, \ldots, a_t)$ denote the probability that the system traverses this trajectory when moving from state a_0 to state a_t. We call such a trajectory *positively correlated*, if

$$\Pr(a_0, a_1, \ldots, a_t) \ >\ \prod_{i=0}^{t-1} p(a_i, a_{i+1})$$

holds. It means, that the probability of traversing the trajectory is larger than what we would get by making individual transitions independently by the same state transition probabilities, but ignoring history dependence.

Similarly, we call a trajectory *negatively correlated*, if

$$\Pr(a_0, a_1, \ldots, a_t) \; < \; \prod_{i=0}^{t-1} p(a_i, a_{i+1}).$$

Finally, let us call a trajectory *balanced*, if

$$\Pr(a_0, a_1, \ldots, a_t) \; = \; \prod_{i=0}^{t-1} p(a_i, a_{i+1}).$$

Now, the next result directly follows from the trajectory summation formula (see Section 5.1):

Theorem 3 *Let $X = (X_t,\ t = 0, 1, 2, \ldots)$ be a first order homogeneous discrete stochastic process. Then precisely one of the following holds:*

- *The process has both positively and negatively correlated trajectories; or*
- *Every trajectory is balanced, in which case the process is a Markov chain.*

7 Open Problems

Finally, below we outline some open problems that can serve as interesting research subjects.

7.1 First Order Equivalence at Random Times

We know from Theorem 1 that the original process X and its Markov projection $\widetilde{X}$ are first order equivalent. By definition, this means that for every t the distributions of X_t and of $\widetilde{X}_t$ agree. For short notation, let us denote distribution of any random variable η by $D(\eta)$. Then we can formulate the first order equivalence of X and $\widetilde{X}$ as

$$\forall t \in \{0, 1, 2, \ldots\} : \quad D(X_t) = D(\widetilde{X}_t).$$

Now we can ask the question: if τ is a *random* time, does $D(X_\tau) = D(\widetilde{X}_\tau)$ still remain valid if $\widetilde{X}$ is the Markov projection of X? After all, if the relationship $D(X_t) = D(\widetilde{X}_t)$ holds for *every* time, how could it fail at a random time?

Interestingly, however, without any restriction on τ, the relationship $D(X_\tau) = D(\widetilde{X}_\tau)$ generally does not remain true. This shows that in the probability context we have to be careful with the qualifier "every." It really means "every fixed," which may not extend to a random value. To illustrate it, consider the following example.

Fix an $a \in S$ and let τ be the first time when $X_\tau = a$ and $\widetilde{X}_\tau \neq a$. Since X and $\widetilde{X}$ are generally different, such an a and τ must exist at least for some processes. But then $D(X_\tau)$ is concentrated on the singleton $\{a\}$, while $D(\widetilde{X}_\tau)$ is concentrated within the set $S - \{a\}$, so we obtain that $D(X_\tau) \neq D(\widetilde{X}_\tau)$ must hold, even though $D(X_t) = D(\widetilde{X}_t)$ is true for *every* (fixed) t.

The above considerations lead to the question: at which random times τ can we still guarantee that $D(X_\tau) = D(\widetilde{X}_\tau)$?

7.2 Higher Order Equivalences

The original process X and its Markov projection $\widetilde{X}$ are first order equivalent, meaning that their respective first order distributions are equal at every fixed time t. Under what conditions can this be extended to a higher order equivalence? For example, considering second order, under what conditions can we claim that the 2-dimensional distributions also agree? This would mean that for any t_1, t_2, the distribution of (X_{t_1}, X_{t_2}) is the same as the distribution of $(\widetilde{X}_{t_1}, \widetilde{X}_{t_2})$. Such a higher order equivalence would mean a closer approximation of the original process by a Markov chain.

8 Conclusion

The presented results provide a simple but useful method to handle non-Markovian models. If we are able to deduce or measure the first-order transition probabilities of the system, then the stationary distribution and also the speed of convergence to stationary (transient analysis) can be obtained from the analysis of the Markov projection, utilizing the fact that its 1-dimensional distributions coincide with that of the original process. In other words, we can reduce the analysis of a non-Markovian system to a Markov chain, carrying over a number of non-trivial results from Markov chain theory.

Acknowledgment: The author would like to thank to Alexandru Hlibiciuc for insightful discussions.

References

1. Aldous D. and Fill, J. (2014) Reversible Markov Chains and Random Walks on Graphs. Monograph under preparation, draft is available online at http://www.stat.berkeley.edu/∼aldous/RWG/book.html

2. Faragó, A. (2021). Decomposition of Random Sequences into Mixtures of Simpler Ones and Its Application in Network Analysis. *Algorithms*, 14(11), article 336, 1–18. https://doi.org/10.3390/a14110336

3. Faragó, A. (2020). On Non-Markovian Performance Models. *Arxiv Preprint.* arXiv:2012.07152v1, Dec 2020.

4. Kelly, F.P. (1991). Loss Networks. *Annals of Applied Probability*, 1(3), 319–378.

5. Kemeny, J.G. and Snell, J.L. (1960). Finite Markov Chains. Van Nostrand Reinhold, New York. (Later editions: Springer, 1976, 1983.)

6. M. Kijima, M. (1997). Markov Processes for Stochastic Modeling. Chapman & Hall.

7. Norris, J.R. (1997). Markov Chains. Cambridge University Press.