The Markov Binomial Distribution: an Explicit Solution by a Regenerative Approach

Alexander Karalis Isaac

Abstract—**The coefficients of the Markov binomial distribution are solved in terms of the underlying state-contingent probabilities of the Markov chain. This will be useful for researchers concerned with the analysis of data generated by a discrete Markov chain. The paper exploits the regenerative nature of the problem and solves the difference equations known to define the distribution. The LLN, CLT and LIL are then available by standard methods. The LIL may be new, while the CLT coincides with existing work.**

Index Terms—**Markov chains, mathematical statistics, Hidden Markov Models, difference equations**

I. Introduction

DISCRETE Markov chains are employed in many fields, including environmental science, queuing theory, economics and genetics. The literature on Hidden Markov Models is the most prominent example. In some cases researchers may be interested in the probability of observing m vistis to some state in n trials. For the important case of the two-state Markov chain, such a probability is governed by the Markov binomial distribution.

A recursion relationship for the general Markov Binomial distribution was proposed in [1]. For the calculation of probabilities it is highly practical, but for the analysis of functions of these probabilities the recursive structure is problematic. By taking a regenerative approach to the problem this paper solves the recursion, under a specific initial condition, yielding a closed form expression for the Markov Binomial coefficients.

Limit results have long existed for particular types of ergodic chain, which induce correlation in successive Bernoulli trials, see references in [2]. [1] may be the first to find such expressions for a general ergodic chain, and to find a method of calculating precise probabilities in finite samples. The limit results below augment their Central Limit Theorem with a version of the Law of the Iterated Logarithm. Both are direct applications of established results for Markov chains possessing an accessible atom, given in [3].

In addition it may be possible to apply the regenerative method of Section II to Markov chains on general state spaces, which is not well studied in the literature. The price, however, is that the Markov chain must now start in a chosen state. In an applied context this may not be a large sacrifice; for example, considering the number of successes in a production run that begins with success with probability 1 may be a reasonable approximation when failure rates are small. In general, the requirement is identical to that in the regenerative bootstrap literature, in which only complete regenerative data blocks can be studied, [4],.

Section II proposes a recursion for the probability that a Markov chain visits a certain state m times in n periods. In Section III this recursion is applied to the two-state Markov chain to yield the Markov binomial coefficients. Limits are studied in Section 4, with the normalising constants expressed in terms of the elements of the transition matrix of the Markov chain. Section 5 concludes, noting potential applications and extensions.

II. Occupation times in regenerative Markov chains

Consider a generic Markov chain, in discrete time, on some space X. For such a process all transition probabilities can be described as products of the one-period transition probability kernel $P(x, A)$. Here $x \in X$ is the initial condition, and events $A \in \mathcal{B}(X)$, take place on a set of subsets of the state space.

$$\begin{aligned} \Pr(x_{t+1} \in A | x_t = x) &= P(x, A) \\ \Pr(x_{t+2} \in A | x_t = x) &= \int_X P(x, dy) P(y, A) \qquad (1) \\ \Pr(x_{t+k} \in A | x_t = x) &= \int_X P(x, dy) P^{k-1}(y, A) \end{aligned}$$

Given the initial condition, time is irrelevant and the t subscripts disappear on the right hand side of (1). If there is some set α is such that $\Pr(x_{t+k} \in S | x_t \in \alpha)$ is equal for all $x \in \alpha$, then it is clear from (1) that returns to set α will be regeneration times of the chain. For a discrete Markov chain, any state $i \in \mathcal{X}$ will have this property. In general such a set is known as an 'atom' of the Markov chain P, see [3]. If the Markov chain visits such a set, α, from any starting point it is known as an 'accessible atom'. Many models in queuing theory and control (beyond the discrete Markov chain) posess accessible atoms. The following is applicable only to such atoms, but it may be possible to extend the results to chains which do not possess accessible atoms via the Nummelin splitting technique. This is not discussed here.

Define the occupation time on a finite sample, η_A, as

$$\eta_A := \sum_{t=1}^{n} 1_{x_t \in A} \qquad (2)$$

where 1_x is the indicator function of the event x. We seek the probability distribution $\phi_{m,n} = \Pr(\eta_A = m)$ for observations on some given sample $\{x_t\}_{t=1}^{n}$. Let A be the set of interest, and presume that visits to set A define regeneration times of the chain. Define the probability distribution associated with the first return times to A as, $f_k = \Pr(\tau_A = k)$, and let q_k be the tail of this distribution,

Manuscript received March 6, 2013; This work was supported in part by the ESRC. Studentship No. ES/F020236/1

A. Karalis Isaac is with the Department of Economics, Mathematics and Statistics, Birkbeck College, London, WC1E 7HX, U.K. e-mail: lkara04@mail.bbk.ac.uk.

$q_k = \Pr(\tau_A > k)$. Then provided that the sample $\{x_t\}$ is a set of observations arising after a visit to A, (i.e. $x_0 \in A$) the occupation time can be redefined in terms of first return times:

$$\eta_A = \max\{k : \tau_A(k) \leq n\} \tag{3}$$

Consider the probability that there is only one visit to A in the sample $\{x_t\}$. This is the probability that the first return occurs in less than or equal to n periods, and that the second return takes place after n. If the first return takes place in the i^{th} period, the probability that the second return takes place after $n - i$ periods is simply the probability of observing a return time greater than $n-i$. Formally, this can be expressed as

$$\begin{aligned} \phi_{1,n} &= \Pr(\tau_A(1) \leq n, \;\; \tau_A(2) > n) \\ &= \sum_{i=1}^{n} \Pr(\tau_A(1) = i, \;\; \tau_A(2) - \tau_A(1) > n - i) \\ &= \sum_{i=1}^{n} \Pr(\tau_A(1) = i) \Pr(\tau_A(2) - \tau_A(1) > n - i) \\ &= \sum_{i=1}^{n} \Pr(\tau_A = i) \Pr(\tau_A > n - i) \qquad (4) \\ &= \sum_{i=1}^{n} f_i q_{n-i} \\ &= \{f_n\} * \{q_n\} \end{aligned}$$

where the notation $\{a_n\} * \{b_n\}$ refers to the convolution of two sequences. The fourth line used that the return time distributions, $\Pr(\tau_A(j) = k)$, are the same for all j. Similarly, consider the probability of two hits in n trials. If the first return occurs in the i^{th} period after 0, then there are a remaining $n - i$ periods in which precisely one further visit to the set A must occur. Again, the probability of 1 occurrence in $n - i$ trials is independent of the date from which we begin measuring, suggesting we should find $\phi_{2,n} = \sum_{i=1}^{n} f_i \phi_{1,n-i}$.

$$\begin{aligned} \phi_{2,n} &= \Pr(\tau_A(1) \leq n, \;\; \tau_A(2) \leq n, \;\; \tau_A(3) > n) \\ &= \sum_{i=1}^{n} \Pr\{\tau_A(1) = i, \tau_A(2) - \tau_A(1) \leq n - i, \\ &\qquad \tau_A(3) - \tau_A(2) > (n - i) - \tau_A(2)\} \\ &= \sum_{i=1}^{n} \Pr(\tau_A = i) \Pr\{\tau_A(2) - \tau_A(1) \leq n - i, \\ &\qquad \tau_A(3) - \tau_A(2) > (n - i) - \tau_A(2)\} \end{aligned}$$

By the independence of regeneration times, the second probability in the final line is just the probability of seeing exactly one visit to state A in $n-i$ periods, $\phi_{1,n-i}$. Therefore

$$\begin{aligned} \phi_{2,n} &= \sum_{i=1}^{n} \Pr(\tau_A = i) \phi_{1,n-i} \\ &= \sum_{i=1}^{n} f_i \phi_{1,n-i} \\ &= \{f_n\} * \{\phi_{1,n}\} \end{aligned}$$

which is the convolution of the distribution of first return times, with the probability that there is only one hit in n trials. Of course, the latter can be calculated, for any n, from (4). If we continue for general m, we have

$$\begin{aligned} \phi_{m,n} &= \{f_n\} * \{\phi_{m-1,n}\} \\ &= \{f_n\}^{m*} * \{q_n\} \end{aligned} \tag{5}$$

The distribution of the occupation time is seen to be a function only of the distribuition of the first-return time on the chosen set, f_n, and the tail of that distribution, q_n. The notation $\{\}^{m*}$ refers to m-fold convolution. If we can work out the return time distribution for A, then the distribution of the occupation time on A can be calculated from (5). This is a recursive operation; for given n, calculate $\phi_{1,n}$ from (4), and then employ (5) for $m = 2 \ldots n$.

III. The Markov binomial distribution

A. *Return time distribution and conjectured solution*

Equations (4) and (5) are now applied to the two-state, ergodic Markov chain. To establish notation, let $X = \{0, 1\}^2$ and let z_t be a vector on this space such that $z_{t+1} = P' z_t + v_{t+1}$ with

$$P = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} \tag{6}$$

where $p_{ij} = 1 - p_{ii}$, and $p_{ii} \in (0, 1)$. The random variable v_{t+1} has a discrete support which depends on z_t. For example, if $z_t = [0, 1]'$, then with probability p_{11}, v_{t+1} takes value $[(-p_{10}), (1 - p_{11})]'$, so that $z_{t+1} = [0, 1]'$ again; or, with probability p_{10}, v_{t+1} takes the value $[(1-p_{10}), (-p_{11})]'$, and $z_{t+1} = [1, 0]'$, which corresponds to a change of regime. Further define

$$X_t = \;\; [0 \quad 1] \;\; z_t$$

so that X_t keeps track of occurences of the second regime, which is taken, wlog, as the regime of interest.

Then $\eta_1 = \sum_{i=1}^{n} X_i$, and define

$$\phi_{m,n} := \Pr(\eta_1 = m)$$

as the distribution of interest. To derive the $\phi_{m,n}$ coefficients the first return time probabilities, f_n, and the tail of this distribution, q_n, are required. The first retrun time probabilities are available from the joint probabilities of the X_i variables. Denote the sequence $\{X_1 X_2 \ldots X_{n-1}\} = \{X_s\}$, then

$$f_n = \Pr(X_n = 1, \{X_s\} = \{0\} | X_0 = 1)$$

By the properties of (6)

$$\begin{aligned} f_1 &= p_{11} \\ f_n &= p_{10} p_{00}^{n-2} p_{01} \qquad n \geq 2 \end{aligned} \tag{7}$$

Note that $f_0 = 0$ by the definition of return times. This implies that the upper tail, $q_n = 1 - \sum_{i=1}^{n} f_i$, satisfies

$$\begin{aligned} q_0 &= 1 \\ q_n &= p_{10} p_{00}^{n-1} \qquad n \geq 1 \end{aligned} \tag{8}$$

Even with these simple expressions for f_n and q_n, the $\phi_{m,n}$ become difficult to compute directly. However, it is easier to find a pattern in the m-fold convolutions of the first return times. Define the m-fold convolution of the first return times as $\psi_{m,n}$ so the occupation time distribution can be re-written as

$$\phi_{m,n} \quad = \quad \{\psi_{m,n}\} * \{q_n\} \tag{9}$$

Consider the sequence $\psi_{2,n} = \{f_n\} * \{f_n\}$, for $n \geq 3$

$$\begin{aligned}
\psi_{2,3} &= f_1 f_2 + f_2 f_1 \\
&= 2p_{11}(p_{10}p_{01}) \\
\psi_{2,4} &= 2f_1 f_3 + f_2 f_2 \\
&= 2p_{11}(p_{10}p_{01})p_{00} + (p_{10}p_{01})^2 \\
\psi_{2,5} &= \sum_{i=1}^{4} f_i f_{5-i} \\
&= 2p_{11}(p_{10}p_{01})p_{00}^2 + 2(p_{10}p_{01})^2 p_{00} \\
\psi_{2,6} &= \sum_{i=1}^{5} f_i f_{6-i} \\
&= 2p_{11}(p_{10}p_{01})p_{00}^3 + 3(p_{10}p_{01})^2 p_{00}^2
\end{aligned}$$

This suggests

$$\psi_{2,n} = 2p_{11}(p_{10}p_{01})p_{00}^{n-3} + (n-3)(p_{10}p_{01})^2 p_{00}^{n-4}$$

The coefficients for $\psi_{3,n} = \{\psi_{2,n}\} * \{f_n\}$ likewise fall into a pattern for $n \geq 4$, suggesting

$$\begin{aligned}
\psi_{3,n} \quad = \quad & 3p_{11}^2(p_{10}p_{01})p_{00}^{n-4} + 3(n-4)p_{11}(p_{10}p_{01})^2 p_{00}^{n-5} \\
& + \sum_{j=1}^{n-5} j \cdot (p_{10}p_{01})^3 p_{00}^{n-6}
\end{aligned}$$

Two further terms in the sequence are given in Appendix A. Considering the evolution of the $\psi_{m,n}$ series over $m = 1 \ldots 5$, a general pattern is suggested for these coefficients, with $n > m$

$$\psi_{m,n} = \sum_{k=1}^{m} \binom{m}{k} \binom{n-(m+1)}{k-1} p_{11}^{m-k} (p_{10}p_{01})^k p_{00}^{n-(m+k)} \tag{10}$$

It is clear from the definitions that $\psi_{m,j} = 0$ for $j < m$, while $\psi_{m,m} = p_{11}^m$. If (10) is then applied to (9), together with these initial conditions, this gives an expression for the Markov binomial coefficients:

$$\begin{aligned}
\phi_{m,n} \quad = \quad & p_{11}^m p_{10} p_{00}^{n-(m+1)} + \\
& \sum_{j=1}^{n-m-1} \sum_{k=1}^{m} \binom{m}{k} \binom{j-1}{k-1} \beta_{m,n,k} \\
& + \psi_{m,n}
\end{aligned} \tag{11}$$

where $\beta_{m,n,k} = p_{11}^{m-k} p_{10}^{k+1} p_{01}^{k} p_{00}^{n-(m+k+1)}$. The coefficients $\phi_{m,n}$ have been calculated for all $m \leq 50$, and compared to coefficients derived using the recursive approach in [1]. In all cases discrepancies are less than 10^{-16}, which is less than the numerical accuracy of the calculation of the binomial coefficients by nchoosek in Matlab. This is approximately 2500 non-trivial coefficient comparisons.

B. Combinatorial interpretation and proof by induction

The Markov binomial distribution is defined recursively in [1]. Their result is presented in (13) for reference. If the conjecture (11) solves the recursion (13) then it is indeed an expression for the Markov Binomial coefficients.

Let $p_n(m) := \Pr(\sum_{i=1}^{n} X_i = m)$, be defined by the recursion (13). If the conjecture (11) is correct, $p_n(m)$ will turn out to be an alternative notation for $\phi_{m,n}$. Following [1], write $p_n(m)$ as the sum of the disjoint events

$$p_n^0(m) = \Pr(\sum_{i=1}^{n} X_i = m, X_n = 0)$$

$$p_n^1(m) := \Pr(\sum_{i=1}^{n} X_i = m, X_n = 1)$$

Further, define the initial condition in the current regenerative setting

$$\begin{aligned}
p_1^0(0) = p_{10} \qquad & p_1^0(1) = 0 \\
p_1^1(0) = 0 \qquad & p_1^1(1) = p_{11}
\end{aligned} \tag{12}$$

Then

$$\begin{aligned}
p_{n+1}(m) \quad = \quad & p_{00}p_n^0(m) + p_{10}p_n^1(m) + \\
& p_{01}p_n^0(m-1) + p_{11}p_n^1(m-1)
\end{aligned} \tag{13}$$

The difficulty of solving (13) with (11) is that (11) does not break the probability down into two disjoint parts, $p_n^0(m)$ and $p_n^1(m)$. By considering the combinatorial interpretation of (11) it is possible to find, for $m = 1$ and $m = 2$, the expressions for $p_n^0(m)$ and $p_n^1(m)$. For such m it is possible to show that (11) solves (13) under initial condition (12).

Some tedious arithmetic allows the argument to be generalized for all $m \geq 1$ and $n > m$. In particular we find $p_n^1(m) = \psi_{m,n}$, and therefore $p_n^0(m) = \phi_{m,n} - \psi_{m,n}$. It is then possible to show these definitions satisfy (13).

First, consider the probability of observing one visit to state 1 in n trials, for $n \geq 2$

$$\begin{aligned}
\phi_{1,n} \quad = \quad & p_{11}p_{10}p_{00}^{n-2} + (n-2)p_{10}^2 p_{01} p_{00}^{n-3} + \\
& p_{10}p_{00}^{n-2}p_{01}
\end{aligned} \tag{14}$$

The sole visit to state 1 can occur either on the first trial (first term), on the last trial (last term), or in one of the $(n-2)$ remaining intermediate trials. This implies

$$\begin{aligned}
p_n^0(1) &= p_{11}p_{10}p_{00}^{n-2} + (n-2)p_{10}^2 p_{01} p_{00}^{n-3} \\
p_n^1(1) &= p_{10}p_{00}^{n-2}p_{01}
\end{aligned} \tag{15}$$

then (13) implies

$$\begin{aligned}
p_{n+1}(1) &= p_{00}p_n^0(1) + p_{10}p_n^1(1) + p_{01}p_n^0(0) \\
&= p_{00}\left(p_{11}p_{10}p_{00}^{n-2} + (n-2)p_{10}^2 p_{01} p_{00}^{n-3}\right) \\
&\quad + p_{10}\left(p_{10}p_{00}^{n-2}p_{01}\right) + p_{01}\left(p_{10}p_{00}^{n-1}\right) \\
&= p_{11}p_{10}p_{00}^{n-1} + (n-1)p_{10}^2 p_{01} p_{00}^{n-2} \\
&\quad + p_{10}p_{00}^{n-1}p_{01} \\
&= \phi_{1,n+1}
\end{aligned}$$

as required. Note $p_n^0(0)$ is given by the distribution q_n derived above.

For the probability of observing 2 successes in 3 or more trials, the combinatorial interpretation is still clear

$$\begin{aligned}\phi_{2,n} &= p_{11}^2 p_{10} p_{00}^{n-3} + 2p_{11}(p_{10}p_{01})p_{00}^{n-3} \\ &\quad +2(n-3)p_{11}p_{10}^2 p_{01} p_{00}^{n-4} \qquad (16)\\ &\quad +(n-3)(p_{10}p_{01})^2 p_{00}^{n-4} + \sum_{j=0}^{n-4} j\cdot p_{10}^3 p_{01}^2 p_{00}^{n-5}\end{aligned}$$

The first term is the probability of both successes occurring in the first two events; the second captures events $\{1,0,\ldots 0,1\}$ and $\{0,0,\ldots,1,1\}$; thirdly, there are $(n-3)$ ways that each of $\{0,\ldots 0,1,1,0,\ldots\}$ and $\{1,0,\ldots,0,1,0,\ldots\}$ can occur; fourth are the $(n-3)$ possibilities $\{0,\ldots,0,1,0,\ldots,0,1\}$; and finally, all the possible occurrences involving three transitions from regime 1 to regime 0: $\{0,\ldots,1,0,\ldots,1,0,\ldots\}$. Thus (16) implies

$$\begin{aligned}p_n^0(2) &= p_{11}^2 p_{10} p_{00}^{n-3} + 2(n-3)p_{11}p_{10}^2 p_{01} p_{00}^{n-4} \\ &\quad + \sum_{j=0}^{n-4} j\cdot p_{10}^3 p_{01}^2 p_{00}^{n-5} \qquad (17)\\ p_n^1(2) &= 2p_{11}(p_{10}p_{01})p_{00}^{n-3} + (n-3)(p_{10}p_{01})^2 p_{00}^{n-4}\end{aligned}$$

Applying (13) to (16), with $m=2$ and using (15) and (17) one finds

$$\begin{aligned}p_{n+1}(2) &= p_{00}p_n^0(2) + p_{10}p_n^1(2) \\ &\quad +p_{01}p_n^0(1) + p_{11}p_n^1(1) \\ &= p_{00}\big\{\beta_{2,n,0} + 2(n-3)\beta_{2,n,1} + \sum_{j=0}^{n-4} j\beta_{2,n,2}\big\} \\ &+ p_{10}\big\{2p_{11}(p_{10}p_{01})p_{00}^{n-3} \\ &+ (n-3)(p_{10}p_{01})^2 p_{00}^{n-4}\big\} \\ &+ p_{01}\big\{\beta_{1,n,0} + (n-2)\beta_{1,n,1}\big\} \\ &+ p_{11}p_{10}p_{00}^{n-2}p_{01} \\ \\ &= p_{11}^2 p_{10} p_{00}^{n-2} + 2p_{11}(p_{10}p_{01})p_{00}^{n-2} \\ &+ 2(n-2)p_{11}p_{10}^2 p_{01} p_{00}^{n-3} \\ &+ (n-2)(p_{10}p_{01})^2 p_{00}^{n-3} \\ &+ \sum_{j=0}^{n-3} j\cdot p_{10}^3 p_{01}^2 p_{00}^{n-4} \\ &= \phi_{2,n+1}\end{aligned}$$

as required. Notice that both $p_n^1(1)$ and $p_n^1(2)$ involve equal numbers of transitions into and out of state 1; that is, terms $(p_{10}p_{01})^k$ appear, but terms $p_{10}^{k+1}p_{01}^k$ do not. This is true in general - if the process is to be in state 1 on date n, then it must have, for every exit from state 1, p_{10}, a return to state 1, p_{01}. The case where the process never leaves state 1, but finishes in state one is the event $\phi_{m,m}$ which happens with probability p_{11}^m and is outside the scope of (11), which is defined for $n>m$. This suggests $p_n^1(m) = \psi_{m,n}$, as these are the only terms of (11) with equal numbers of exits and returns. But then it must be that the remaining terms of (11) give the remaining probability, $p_n^0(m)$, so we have

$$\begin{aligned}p_n^1(m) &= \psi_{m,n} \\ p_n^0(m) &= (\phi_{m,n} - \psi_{m,n})\end{aligned}$$

Apply the second part of (13) to these definitions, then

$$p_{n+1}^1(m) = p_{01}p_n^0(m-1) + p_{11}p_n^1(m-1) \qquad (18)$$

Working through (18) using the definitions

$$\begin{aligned}p_n^0(m-1) &= p_{11}^{m-1}p_{10}p_{00}^{n-m} + \\ &\quad \sum_{j=1}^{n-m}\sum_{k=1}^{m-1}\binom{m-1}{k}\binom{j-1}{k-1}\delta_{m,n,k}p_{10} \\ p_n^1(m-1) &= \sum_{k=1}^{m-1}\binom{m-1}{k}\binom{n-m}{k-1}\delta_{n,m,k}p_{00}\end{aligned}$$

where $\delta_{m,n,k} = p_{11}^{m-(k+1)}(p_{10}p_{01})^k p_{00}^{n-(m+k)}$, gives

$$\begin{aligned}p_{n+1}^1(m) &= p_{11}^{m-1}(p_{10}p_{01})p_{00}^{n-m} \qquad (19)\\ &+ \sum_{j=1}^{n-m}\sum_{k=1}^{m-1}\binom{m-1}{k}\binom{j-1}{k-1}\delta_{n,m,k}p_{10}p_{01} \\ &+ \sum_{k=1}^{m-1}\binom{m-1}{k}\binom{n-m}{k-1}\delta_{n,m,k}p_{11}p_{00}\end{aligned}$$

Working out the double summation gives $\sum_j\sum_k(\cdot) =$

$$\begin{gathered}(m-1)(n-m)p_{11}^{m-2}(p_{10}p_{01})^2 p_{00}^{n-(m+1)} \\ +\tbinom{m-1}{2}\textstyle\sum_{j=1}^{n-m-1} jp_{11}^{m-3}(p_{10}p_{01})^3 p_{00}^{n-(m+2)} \\ +\tbinom{m-1}{3}\textstyle\sum_{j=2}^{n-m-1}\tbinom{j}{2}p_{11}^{m-4}(p_{10}p_{01})^4 p_{00}^{n-(m+3)} + \ldots \\ +\tbinom{m-1}{m-1}\textstyle\sum_{j=m-2}^{n-m-1}\tbinom{j}{m-2}(p_{10}p_{01})^m p_{00}^{n-2m+1}\end{gathered}$$

Plugging this expression into (19) and expanding the single summation in a similar fashion shows

$$\begin{aligned}p_{n+1}^1(m) &= \kappa_1(1+(m-1)) \\ &+ \kappa_2\Big((m-1)(n-m) + \binom{m-1}{2}(n-m)\Big) \\ &+ \kappa_3\Big(\binom{m-1}{2}\binom{n-m}{2} + \binom{m-1}{3}\binom{n-m}{2}\Big) \\ &+ \ldots \\ &+ \kappa_{m-1}\Big(\binom{m-1}{m-2}\sum_{j=m-3}^{n-m-1}\binom{j}{m-3} + \binom{n-m}{m-2}\Big) \\ &+ \kappa_m\Big(\sum_{j=m-2}^{n-m-1}\binom{j}{m-2}\Big)\Big)\end{aligned}$$

where the constants $\kappa_k = p_{11}^{m-k}(p_{10}p_{01})^k p_{00}^{n-(m+(k-1))}$. Using the standard relation

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

and its corollary

$$\sum_{j=k}^{n}\binom{j}{k} = \binom{n+1}{k+1}$$

gives

$$\begin{aligned}
p^1_{n+1}(m) &= mp_{11}^{m-1}(p_{10}p_{01})p_{00}^{n-m} + \\
&+ (n-m)\binom{m}{2}p_{11}^{m-2}(p_{10}p_{01})^2 p_{00}^{n-(m+1)} \\
&+ \binom{n-m}{2}\binom{m}{3}p_{11}^{m-3}(p_{10}p_{01})^3 p_{00}^{n-(m+2)} \\
&+ \ldots \\
&+ \binom{n-m}{m-1}(p_{10}p_{01})^m p_{00}^{n-2m+1} \\
&= \sum_{k=1}^{m}\binom{m}{k}\binom{n-m}{k-1}\kappa_k \\
&= \psi_{m,n+1}
\end{aligned}$$

The final equality re-labels the line above in terms of $n+1$. This demonstrates that $\psi_{m,n} = p^1_n(m)$ in general. It then follows directly that

$$\begin{aligned}
p_{n+1}(m) &= p_{00}(\phi_{m,n} - \psi_{m,n}) + p_{10}\psi_{m,n} \\
&+ p_{01}(\phi_{m-1,n} - \psi_{m-1,n}) + p_{11}\psi_{m-1,n}
\end{aligned}$$

using (10) and (11) this gives

$$\begin{aligned}
p_{n+1}(m) &= p_{00}p_{11}^m p_{10} p_{00}^{n-(m+1)} \\
&+ p_{00}\sum_{j=1}^{n-m-1}\sum_{k=1}^{m}\binom{m}{k}\binom{j-1}{k-1}\beta_{m,n,k} \\
&+ p_{10}\sum_{k=1}^{m}\binom{m}{k}\binom{n-(m+1)}{k-1}p_{11}\delta_{m,n,k} \\
&= p_{11}^m p_{10} p_{00}^{n-m} + \\
&\quad \sum_{j=1}^{n-1}\sum_{k=1}^{m}\binom{m}{k}\binom{j-1}{k-1}\beta_{m,n+1,k} \\
&\quad +\psi_{m,n+1} \\
&= \phi_{m,n+1}
\end{aligned}$$

which completes the proof.

IV. Limits

We have studied the number of successes in n trials where the probability of success follows a two-state ergodic Markov chain. The initial condition, $\Pr(X_1 = 1) = p_{11}$, ensured the process was entirely regenerative. **Theorem 17.2.2** of [3] therefore applies to sums of functions defined on the sample path of the Markov chain. The occupation time η_1 is just such a sum. If we put $\bar{g}_t = X_t - \mathbf{E}(X_t)$ and $Z \sim N(0,1)$, then

$$(n\gamma_g^2)^{-1/2}\sum_{k=1}^{n}\bar{g}_k \rightarrow^d Z \tag{20}$$

where $\mathbf{E}(.)$ denotes mathematical expectation. Further

$$\begin{aligned}
\lim\sup_{n\to\infty}(2\gamma_g^2 n\log\log(n))^{-1/2}\sum_{k=1}^{n}\bar{g}_k &= 1 \\
\lim\inf_{n\to\infty}(2\gamma_g^2 n\log\log(n))^{-1/2}\sum_{k=1}^{n}\bar{g}_k &= -1
\end{aligned}$$

$\sum_{k=1}^n \bar{g}_k$ is just $\eta_1 - \mathbf{E}\eta_1$, so (20) is the CLT for the occupation time; the following limits describe the Law of the Iterated Logarithm for the occupation time. To be operational we need expression for the constants $\mathbf{E}(X_t)$ and γ_g^2. The unconditional expectation, or ergodic measure, of X_t is the second element of the eigenvector associated with unity for the transition matrix P. This has the well known solution $y = p_{01}/(p_{10} + p_{01})$. The asymptotic variance of the sum, γ_g^2, is given in equation (17.13) of [3]

$$\gamma_g^2 = \pi(\alpha)\mathbf{E}_\alpha[(\sum_{k=1}^{\tau_\alpha}\bar{g}(X_k))^2]$$

where $\pi(\alpha)$ is the ergodic measure of the atom, here $\pi(\alpha) = y$ again. $\mathbf{E}_\alpha$ refers to expectations conditional on $X_0 = 1$, consistent with our setting. τ_α is the first return time to the atom, here the atom is the set of interest, i.e. the state where $X_t = 1$.

In our case $\bar{g}(X_k)$ are defined for all $k \leq \tau_\alpha$ by τ_α itself: for $k < \tau_\alpha$, $\bar{g}(X_k) = (-y)$, while $\bar{g}(X_{\tau_\alpha}) = (1-y)$. Therefore put $(\sum_{k=1}^{\tau_\alpha}\bar{g}(X_k))^2 = f(\tau_\alpha)$, where $f(k) = (k-1)^2y^2 + 2(k-1)(y^2-y) + (1-y)^2$, and we have

$$\begin{aligned}
\gamma_g^2 &= y\cdot\mathbf{E}_\alpha f(\tau_\alpha) \\
&= y\cdot\sum_{k=1}^{\infty}f(k)\Pr(\tau_\alpha = k) \\
&= y\cdot\left[f(1)p_{11} + \sum_{k=2}^{\infty}f(k)p_{10}p_{00}^{k-2}p_{01}\right]
\end{aligned}$$

As expected, due to the diminishing influence of the initial condition, this summation evaluates to the more elegant expression of β in Corollary 6(i) of [1], in which $X_0 = 1$ is not imposed.

V. Conclusion

The Markov binomial distribution has been related to occupation time distributions in regenerative Markov chains. For the specific initial condition $\Pr(X_1 = 1) = p_{11}$, a closed form solution for the Markov binomial distribution has been demonstrated, (11). The solution solves the difference equations derived by [1]. It is hoped that this will allow applied researchers to study statistics which are functionals of estimated Markov binomial coefficients, opening up the rigorous analysis of hypotheses about the sample path of a two-state Markov chain. The paper suggests several directions for further work. The first is perhaps to find an expression for the coefficients under a general initial condition. The study of bootstrap schemes for estimated occupation time distributions would be interesting, as would the application of the regenerative approach, (5), to occupation time distributions of chains on general state spaces.

Appendix A

Proceeding as for $\psi_{2,n}$ and $\psi_{3,n}$, it can be shown

$$\begin{aligned}
\psi_{4,n} &= 4p_{11}^3(p_{10}p_{01})p_{00}^{n-5} \\
&+6(n-5)p_{11}^2(p_{10}p_{01})^2p_{00}^{n-6} \\
&+4\sum_{j=1}^{n-6} j \cdot p_{11}(p_{10}p_{01})^3p_{00}^{n-7} \\
&+\sum_{k=1}^{n-7}\sum_{j=1}^{k} j \cdot (p_{10}p_{01})^4p_{00}^{n-8} \\
\psi_{5,n} &= 5p_{11}^4(p_{10}p_{01})p_{00}^{n-6} \\
&+10(n-6)p_{11}^3(p_{10}p_{01})^2p_{00}^{n-7} \\
&+10\sum_{j=1}^{n-7} p_{11}^2(p_{10}p_{01})^3p_{00}^{n-8} \\
&+5\sum_{j_1=1}^{n-8}\sum_{j=1}^{j_1} j \cdot p_{11}(p_{10}p_{01})^4p_{00}^{n-9} \\
&+\sum_{j_2=1}^{n-9}\sum_{j_1=1}^{j_2}\sum_{j=1}^{j_1} j \cdot (p_{10}p_{01})^5p_{00}^{n-10}
\end{aligned}$$

Acknowledgment

The author would like to thank Prof. Z. Psaradakis for encouragement and helpful comments throughout. All mistakes are my own.

References

[1] E. Omey, J. Sanots, and S. van Gulck, “A markov-binomial distribution,” *Journal of Applicable Analysis and Discrete Mathematics*, vol. 2, 2008.

[2] Y. H. Wang, “On the limit of the markov binomial distribution,” *Journal of Applied Probability*, vol. 18, no. 4, pp. 937–942, 1981.

[3] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. London: Springer-Verlag, 1993.

[4] P. Bertail and S. Clemençon, “Regenerative block bootstrap for markov chains,” *Bernoulli*, vol. 12, no. 4, pp. 689–712, 2006.