Sampling distribution for microsatellites amplified by PCR: mean field approximation and its applications to genotyping

Yinglei Lai* and Fengzhu Sun†

*Department of Mathematics and
†Program in Molecular and Computational Biology

Department of Biological Sciences
University of Southern California
1042 West 36th Place, DRB288
Los Angeles, CA 90089-1113

Tel: (213) 740-2413
Fax: (213) 740-2437
Email: fsun@hto.usc.edu

†corresponding author

†current address: Department of Epidemiology and Public Health, Yale University School of Medicine
Abstract

Due to microsatellite mutations during PCR, stutter patterns may appear in the final PCR product, which hinder us from accurate genotyping microsatellite markers. The existing methods for microsatellite stutter pattern deconvolution required large amount of data. A mathematical model for microsatellite mutations during PCR and an estimation method based on mean field approximation for branching processes have recently been developed. In this paper, we study the asymptotic behaviors for mean field approximation when experiments are started from a large number of molecules, and we derive an upper bound for the approximation error when experiments are started from a finite number of molecules. Based on the theories of mean field approximation and Bayesian statistics, we develop a novel method for microsatellite stutter pattern deconvolution.

Key words: microsatellites, polymerase chain reaction, genotyping, branching processes, mean field approximation.
1 Introduction

Microsatellites are short tandem repeats (STRs) of DNA sequences with usually 1-6 nucleotides for each repeat unit. For example, \((CA)_8\) represents \(\text{-CACACACACACACA-}\), a DNA sequence with motif \((CA)\) repeated 8 times. The allelic state of a microsatellite locus is generally represented by its number of repeat units. Because of the wide range of the number of repeat units, microsatellites are highly polymorphic. Microsatellites are distributed throughout the eukaryotic genomes and are widely used in genetic research such as population and evolution studies (Rosenberg et al., 2000), linkage and association studies (Kong et al., 2002) and human diseases studies (Ashley and Warren, 1995).

The polymerase chain reaction (PCR) is an important biotechnology that can generate a large number of copies of a specific genetic region from a small number of molecules (Saiki et al. 1985; Scharf et al 1986; Mullis et al., 1987; Saiki et al. 1988). For details about PCR and its applications in biological studies, see Watson et al. (1992) and Navidi and Arnheim (1994). It is difficult to directly observe the allelic state of a microsatellite locus if the amount of genetic materials is limited. PCR is widely used to amplify microsatellite markers (Weber and May, 1989; Weber and Wong, 1993). Micro-capillary electrophoresis system can then be used to measure the concentration of molecules with different numbers of repeat units (Shinde et al., 2003).

Mutations may occur during PCR amplification of DNA sequences. During PCR amplification of DNA sequences without repeats, point mutations may occur. Mathematical
methods have been developed to estimate point mutation rate during PCR (Sun, 1995; Weiss and von Haeseler, 1995; Weiss and von Haeseler, 1997; Wang et al., 2000). During PCR amplification of DNA sequences with repeats, slippage mutations (change of the number of repeat units) dominate over point mutations. Miller and Yuan (1997) proposed a mathematical model for microsatellites mutations during PCR, in which they assumed that microsatellites with different numbers of repeat units had the same mutation rates. Their assumption was not supported by experimental data because different stutter patterns for microsatellites become more spread out as the numbers of repeat units increase. A more realistic model was proposed in Lai et al. (2003) and Shinde et al. (2003), in which a linear relationship between microsatellite slippage mutation rate and the number of repeat units was observed.

The mathematical model for microsatellite mutations during PCR is based on multi-type branching processes. In experiments, the observations are the fractions of microsatellites with different numbers of repeat units. Therefore, we are interested in the sampling distribution. For general branching processes, an exact sampling formula has been proposed in Olofsson and Shaw (2002). The formula requires close analytical forms of the probability generating functions related to the branching processes. But for our case, it is difficult to obtain close analytical formulas for the probability generating functions when the number of PCR cycles is large. Mean field approximation was proposed for the study of DNA point mutations during PCR (Sun, 1995; Wang et al., 2000). Piau (2002) derived upper bounds for the approximation errors. Mean field approximation was also used in the study of microsatellite slippage mutations during PCR (Lai et al., 2003; Lai and Sun, 2003; Shinde et al., 2003).
The related mathematical properties were studied in Lai et al. (2003) and Lai and Sun (2003).

Due to microsatellite slippage mutations during PCR, several bands of different numbers of repeat units often appear in the final PCR product, which are referred as stutter patterns (Perlin et al., 1994) or stutter profiles. Stutter patterns may hinder us from assigning alleles correctly in microsatellite genotyping. Computational methods have been developed to reduce the effect of stutter patterns (Perlin et al., 1995), but such methods require large amount of data to construct a stutter-correction matrix.

In this paper, we first review the mathematical model proposed in Lai et al. (2003) and Shinde et al. (2003). Then, we discuss asymptotic behaviors for the mean field approximation when experiments are started from a large number of molecules, and we derive an upper bound for the approximation error when experiments are started from a finite number of molecules. We propose a novel method for stutter pattern deconvolution based on the mean field approximation and Bayesian statistics. We study the validity of the method using simulations.

2 A mathematical model

A mathematical model for microsatellite mutations during PCR has been developed in Lai et al. (2003), Lai and Sun (2003), and Shinde et al. (2003). We briefly review it as follows.
During the PCR amplification of microsatellites, there are two stochastic processes. One is the random generation of template DNA sequences, which can be described as a random binary tree; the other is that random microsatellite slippage mutations are superimposed onto the newly generated DNA sequences.

The state of a microsatellite is represented by its number of repeat units. Let $J = [\alpha, \beta]$ be the range of microsatellite states, where $\alpha \leq \beta$ are finite positive integers. In the $n$-th PCR cycle, a template DNA sequence randomly generates a new DNA sequence with probability $\lambda_n$ called efficiency. Generally, $\lambda_n$ depends on $n$. To make the mathematical exposition clear, we assume $\lambda_n \equiv \lambda > 0$ for all $n$. Once a DNA sequence is generated, it always stays in the pool. A newly generated sequence of state $i$ may mutate to state $j$ with probability $\mu_{ij}$, where $\mu_{ij} \geq 0$ and $\sum_{j \in J} \mu_{ij} = 1$ for all $i$. The transition matrix for microsatellites mutations is $U = (\mu_{ij}), \alpha \leq i, j \leq \beta$.

Let $Z_{n,i}$ be the number of DNA sequences of state $i$ after $n$ PCR cycles, and $Z_n = (Z_{n,\alpha}, Z_{n,\alpha+1}, \ldots, Z_{n,\beta})$. Then $Z_0, Z_1, Z_2, \ldots$ form a multi-type branching process (Harris, 1963; Athreya and Ney, 1972). In this study, we assume that some entries of $Z_0$ are positive. Let $e_k$ be the unit vector with 1 at the $k$-th entry and 0 at the other entries. The probability generating function for $Z_n$ given $Z_{n-1} = e_k$ is

$$g_n^{(k)}(s_\alpha, s_{\alpha+1}, \ldots, s_\beta) = \mathbb{E}(s_\alpha^{Z_{n,\alpha}} \cdots s_\beta^{Z_{n,\beta}} | Z_{n-1} = e_k) = (1 - \lambda)s_k + \sum_{j \in J} \lambda \mu_{kj} s_j s_k,$$
for all \( k, j \in J \) and \( |s_{ij}| \leq 1 \). The first moment matrix of this branching process is \( F = I + \lambda U \).

### 3 Mean field approximation

Since the experimental observations are fractions of microsatellites with different numbers of repeat units, we are interested in the quantity \( Z_n/S_n \) where \( S_n = \sum_{i=\alpha}^{\beta} Z_{n,i} \) is the total number of molecules after \( n \) PCR cycles. It is difficult to obtain any analytical formulae for this quantity. Mean field approximation has been proposed for the sampling distribution (Lai et al., 2003; Lai and Sun, 2003; Shinde et al., 2003)

\[
\mathbb{E}\left(\frac{Z_n}{S_n}\right) \approx \frac{Z_0}{S_0} F^n (1 + \lambda)^n. \tag{1}
\]

### 3.1 A limit theorem for experiments starting from a large number of molecules

In genotyping experiments, an individual sample usually contains a large amount of molecules. When the initial number of molecules is large, we have the following limit theorem.

**Theorem 1.** \( \lim_{S_0 \to \infty} \mathbb{E}\left(\frac{Z_n}{S_n}\right) = \frac{Z_0}{S_0} F^n (1 + \lambda)^n. \)

Therefore, equation (1) gives the asymptotic sampling distribution when the number of initial molecules is large.
3.2 An upper bound for the approximation error for experiments starting from a finite number of molecules

In experiments, it is difficult to guarantee that every initial molecule contains the same number of repeat units when the amount of initial molecules is large. Experiments starting from single molecule were generally conducted (Shinde et al., 2003). Here, we give an upper bound for the mean field approximation error for experiments starting from a finite number of molecules. The mathematical proof will be given in the Appendix.

Throughout this paper, we use $\max |v|$ to denote $\max_i\{|v_i|\}$ for a vector $v$, and $\max |X|$ to denote $\max_{ij}\{|x_{ij}|\}$ for a matrix $X$.

**Theorem 2.** If there is a positive integer $l$ such that every entry of $U^l$ is positive, then there exist $c$ and $r$, $0 < c < \infty$ and $0 < r < 1$, such that

$$\max \left| \mathbb{E}\left(\frac{Z_n}{S_n}\right) - \frac{Z_0}{S_0} \frac{F^n}{(1 + \lambda)^n} \right| \leq c \frac{(1 + \lambda r)^n - 1}{S_0(1 + \lambda)^n}.$$ 

Therefore, the upper bound for the mean field approximation error decreases geometrically as the number of PCR cycles increases.
4 Stutter pattern deconvolution

4.1 A deconvolution method

When genotyping a microsatellite locus, primers flanking the locus are used in PCR to amplify the DNA samples. Since human autosomes are paired, there are two alleles for a microsatellite locus for each individual. Due to microsatellite slippage mutations during PCR, several bands may appear in the final product referred as stutter patterns (Perlin et al., 1994; Perline et al., 1995). Stutter patterns may hinder us from correct allelic assignment. When the two alleles are different (heterozygous), it maybe difficult to discern the correct allelic states from deconvolution of two stutter patterns. Even when the two alleles are the same (homozygous), such a problem may still happen. The band of the correct allele may not have the highest frequency because of microsatellite mutations during PCR.

Previous studies showed that microsatellite mutation rates can be well estimated through mean field approximation (Lai et al., 2003; Lai and Sun, 2003). Here we propose a novel method for microsatellites stutter pattern deconvolution. We will discuss the details step by step.

1. Parameter estimation: estimate PCR efficiency $\lambda$ and transition matrix $U$ for microsatellite mutations;

2. Allele calling: calculate the posterior probability for each possible pair of allelic states
based on Bayesian statistics, and then select the allelic pair with the highest score as the true genotype of the individual.

Parameter estimation

The PCR efficiency $\lambda$ can be estimated by Kinetic PCR (Shinde et al., 2003). The transition matrix $U$ for microsatellite mutations can be estimated using a maximum quasi-likelihood method based on mean field approximation which was developed in Lai et al. (2003) and Shinde et al. (2003). From previous studies (Lai et al., 2003; Shinde et al., 2003), the mutation rates for microsatellite slippage contractions and extensions can be modeled as a linear function of the number of repeat units; and the proposed estimation method can accurately recover those parameters.

Allele calling

We assume that the observed frequencies $O = (o_\alpha, o_{\alpha+1}, \cdots, o_\beta)$ of microsatellites with different numbers of repeat units are sampled from the population with a multinomial distribution $\text{Mul}(S; f)$, where $f = (f_\alpha, f_{\alpha+1}, \cdots, f_\beta)$ is the mean field approximation calculated using equation (1) and $S$ is the sample size. Denote $G = (g_1, g_2)$ the pair of allelic states of an individual at a microsatellite locus, where $\alpha \leq g_1, g_2 \leq \beta$. Since the correct allelic states are unknown, we assume a prior distribution $\pi(G)$ for $G$ that assigns the same probability
\(1/(\beta - \alpha + 1)^2\) to any possible pairs. From the Bayes theory, we have

\[
p(G|O) \propto p(O|G)\pi(G),
\]

where

\[
p(O|G) \propto \prod_{i=\alpha}^{\beta} f_i^{S_{o_i}} = (\prod_{i=\alpha}^{\beta} f_i^{o_i})^S \propto \prod_{i=\alpha}^{\beta} f_i^{o_i}.
\]  

Therefore, the relative order of \(p(G|O)\) is independent of \(S\).

Based on the above Bayesian statistical method, we can assign alleles by picking a pair of allelic states \(G\) that maximize the posterior probability \(p(G|O)\). It is equivalently to find a \(G\) that maximize \(p(O|G)\) using equation (2).

We emphasize here that the above approach is applicable for any prior distribution \(\pi(G)\). For example, we may know that allele frequency \(g\), \(\alpha \leq g \leq \beta\) in the general population. Then, we can assume that \(\pi(g_1, g_2) = p_{g_1}^2\) if \(g_1 = g_2 = g\) and \(\pi(g_1, g_2) = 2p_{g_1}p_{g_2}\) if \(g_1 \neq g_2\).

The sample size \(S\) can be estimated as \(S_0 \sum_{k=1}^{N} (1 + \lambda_k)\), the average total number of molecules in the final PCR product, where \(S_0\) is the initial number of molecules and \(\lambda_k\) is the PCR efficiency at the \(k\)-th PCR cycle.

### 4.2 Examples

We give examples for the mean field approximation and its applications for microsatellite genotyping. First we use simulations to generate data. Then we apply our method to the simulated data for stutter pattern deconvolution.
Simulations

In practice, PCR efficiencies for different cycles are different. The mean field approximation is slightly modified by replacing \( \frac{Z_0}{S_0} F^n (1+\lambda)^n \) with \( \frac{Z_0}{S_0} \prod_{k=1}^{n} F_k (1+\lambda_k) \), where \( F_k = I + \lambda_k U \) (Lai et al., 2003).

The PCR efficiencies are set based on experimental data (Saiki et al., 1988), \( \lambda_i = 0.872 \), 1 \( \leq i \leq 20 \), \( \lambda_i = 0.743 \), 21 \( \leq i \leq 25 \), \( \lambda_i = 0.146 \), 26 \( \leq i \leq 30 \). Microsatellite mutation rates are also set based on these estimated from experimental data (Shinde et al., 2003). A state \( j \) can mutate to state \( j + 1 \) with probability \( \mu_j e \), or to state \( j - 1 \) with probability \( \mu_j (1 - e) \), where \( \mu_j = \max(j - 4, 0) \times 0.004 \), \( j \geq 1 \) and \( e = 0.07 \).

First we simulate data from single molecule to show that the true sampling distribution of microsatellites amplified by PCR can be well approximated by its mean field approximation. We simulate microsatellite mutations during PCR for 30 cycles starting from single molecules. The initial molecule contains 5, 20 or 35 repeat units, respectively. We perform 12 sets of simulations for each initial number of repeat units. The average fractions of 12 sets of simulations are calculated for each initial number of repeat units. We also calculate the related mean field approximation. The results are given in Figure 1. We observe more variation for microsatellites with high number of repeat units in our simulations. The average fractions and the fractions calculated from the mean field approximation are very close.

From the simulation results in Figure 1, the stutter pattern changes as the initial number of repeat units increases. There are almost no mutants when the number of repeat units is 5.
When the number of repeat units becomes large, we observe more fractions of mutants. In the simulations started from a single molecule with 20 repeat units, the frequency of molecules with 19 repeat units is only about 10\% less than the frequency of molecules with 20 repeat units in the final product. In the simulations started from a single molecule with 35 repeat units, the frequency of molecules with 33 repeat units is almost the same as the frequency of molecules with 35 repeat units in the final product, and the frequency of molecules with 34 repeat units is the highest!

We then simulate microsatellite genotyping procedure. In genotyping, the number of DNA molecules in a sample is generally large. Therefore, we simulate microsatellite mutations during PCR for 30 cycles starting from 1000 pairs of molecules. All initial molecules contain the same pair of repeat numbers (5, 6), (20, 21) or (35, 36). We also calculate the related mean field approximation. The results are given in Figure 2. The simulated fractions and the fractions calculated from the mean field approximation are almost identical.

From the simulation results in Figure 2, the stutter pattern changes as the initial number of repeat units increases. We are able to discern the correct pair of allelic states when the numbers of repeat units are 5 and 6. But when the pair of repeat numbers is (20, 21), we are not sure whether (19, 20) or (20, 21) is the correct pair. The frequencies of repeat 19 and repeat 21 are very close. When the numbers of repeat units become larger, like 35 and 36 in the figure, it is unlikely to assign the correct allelic states by just looking at the stutter patterns.
Stutter pattern deconvolution

The estimation procedures for the parameter estimation step in our method have been previously studied in Lai et al. (2003) and Shinde et al. (2003). Accurate and stable estimation can be obtained using those procedures. Therefore, we assume that the efficiency $\lambda$ and the transition matrix $U$ have been accurately estimated.

We then use the correct parameters to calculate the score ($\prod_{i=\alpha}^{\beta} f_i^{*\alpha}$) for each possible pair of allelic states using equation (2). Table 1 gives part of the results in logarithm for the data simulated from the initial pair of numbers of repeat units (35, 36). In the table, the score for the pair (35, 36) is clearly higher than all the other scores.

5 Discussion

It is difficult to calculate the sampling distribution for microsatellites amplified by PCR because of the complicated convolution of random DNA sequence generation and random microsatellite mutations. Based on the mean field approximation for branching processes, we calculated an approximate sampling distribution. We gave an upper bound for the approximation error. We also showed through simulations that the approximate sampling distribution can well reflect the true one.

Based on the theory of mean field approximation, we design a novel method for microsatellite stutter pattern deconvolution, which is essential for microsatellite genotyping.
Unlike the existing methods, our method is based on a stochastic model and Bayesian statistics. Simulation studies show that our method can correctly assign alleles for microsatellite genotyping.

The method proposed in Perlin et al. (1995) requires large amount of data to construct the stutter-correction matrix for each microsatellite locus (Barcellos et al., 1997). In a recent study for microsatellite mutations during PCR, a linear relationship between microsatellite mutation rate and the number of repeat units was observed in Shinde et al. (2003). An efficient estimation method for microsatellite mutation rates was also developed in Lai et al. (2003) and Shinde et al. (2003). Based on those studies, our novel method for stutter pattern deconvolution is an alternative choice that can reduce the burden of large amount of data. We can estimate PCR efficiency and microsatellite mutation rates using data from some representative alleles, then predict those parameters for the other alleles. When all the parameters in our method are set, we can assign alleles from stutter patterns using the method studied in this paper.

6 Acknowledgement

We thank Professor Norman Arnheim for helpful discussions. We are also greatly indebted to an anonymous reviewer for pointing out an error in the original manuscript.
References


Figure legends

Figure 1: Comparison between the frequencies from simulations and the frequencies from mean field approximation. Only those fractions greater than 1% are shown. The stars are simulated fractions of microsatellites with different number of repeats after 30 PCR cycles starting from a single molecule with (a) 5, (b) 20 and (c) 35, repeat units, respectively, with 12 sets of simulations. The triangle are the average of the simulated fractions of the same number of repeat units from different simulations. The bars are the fractions calculated from the mean field approximation.

Figure 2: Comparison between the frequencies from simulations and frequencies from mean field approximation. The white bars are simulated fractions of microsatellites with different number of repeats after 30 PCR cycles starting from 1000 pairs of alleles with the same pair of repeat numbers (a) (5, 6), (b) (20, 21) or (c) (35, 36). The black bars are fractions calculated from the mean field approximation.
Figure 1(a)
Figure 1(b)
Figure 1(c)
Figure 2(a)
Figure 2(b)
Table 1: Part of the calculated log-scores for possible combination of two allelic states.

The correct pair is (35, 36), which has the highest score in the table.

<table>
<thead>
<tr>
<th></th>
<th>31</th>
<th>32</th>
<th>33</th>
<th>34</th>
<th>35</th>
<th>36</th>
<th>37</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>-1122</td>
<td>-842</td>
<td>-557</td>
<td>-350</td>
<td>-243</td>
<td>-229</td>
<td>-274</td>
</tr>
<tr>
<td>33</td>
<td>—</td>
<td>—</td>
<td>-492</td>
<td>-332</td>
<td>-221</td>
<td>-193</td>
<td>-218</td>
</tr>
<tr>
<td>34</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>-287</td>
<td>-205</td>
<td>-175</td>
<td>-187</td>
</tr>
<tr>
<td>35</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>-184</td>
<td>-167</td>
<td>-178</td>
</tr>
<tr>
<td>36</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>-176</td>
<td>-196</td>
</tr>
<tr>
<td>37</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>-232</td>
</tr>
</tbody>
</table>
Appendix

A Mathematical proofs

First we introduce the definition of generation number for DNA sequences in the final PCR product (Sun, 1995).

Definition. The original sequences are called the 0-th generation sequences; the sequences directly generated from the 0-th generation sequences are called the first generation sequences; ···; the sequences directly generated from the $k$-th generation sequences are called the $(k + 1)$-th generation sequences; and so on.

Before the proof of Theorems 1 and 2, we have the following lemmas.

Lemma 1. Let $T_{n,k}$ be the number of $k$-th generation sequences after $n$ PCR cycles, and $T_n = (T_{n,0}, T_{n,1}, \ldots, T_{n,k}, \ldots)$. Let $S_n$ be the total number of sequences, $S_n = \sum_k T_{n,k}$. Then

$$
\mathbb{E} \left( \frac{T_n}{S_n} \right) = \frac{T_0}{S_0} M^n + \sum_{l=0}^{n-1} \mathbb{E} \left[ (\lambda(1 - \lambda)\Theta(S_l) T_l S_l) M^{n-1-l}(I - A) \right],
$$

where $A = \{q_{ij}\}$ is a matrix, $q_{ij} = 1$, $j = i + 1$, and $q_{ij} = 0$ otherwise; $M = \frac{1}{1+\lambda} I + \frac{\lambda}{1+\lambda} A$; and $\Theta(S) = \int_0^1 [(1 - \lambda)^2 + 4\lambda t]^{-3/2} t^S \, dt$.

The proof of this lemma can be found in Lai and Sun (2003).

In the following, let $(v)_k$ denote the $k$-th entry of a vector $v$. Recall that $U = (\mu_{ij})$, $\alpha \leq i, j \leq \beta$ where $\mu_{ij}$ is the probability that a newly generated sequence has state $j$ given
its parent sequence has state $i$. Note also that the first moment matrix of $Z_n$ is given by $F = I + \lambda U$.

**Lemma 2.** \( \frac{F^n}{(1+\lambda)^n} = \sum_{k=0}^{n} \left( \frac{T_n}{S_0} M^n \right) U^k. \)

**Proof of Lemma 2.** Since $F = I + \lambda U$, Lemma 2 follows from the fact that $F^n = (I + \lambda U)^n = \sum_{k=0}^{n} {n \choose k} \lambda^k U^k$, $\frac{T_n}{S_0} = (1, 0, 0, \cdots)$ and the expansion of $M^n$.

Recall that $Z_{n,i}$ is the number of DNA sequences of state $i$ after $n$ PCR cycles, and $Z_n = (Z_{n,\alpha}, Z_{n,\alpha+1}, \cdots, Z_{n,\beta})$.

**Lemma 3.** \( \mathbb{E} \left( \frac{Z_n}{S_n} \right) = \frac{Z_0}{S_0} \sum_{k=0}^{n} \left( \mathbb{E} \left( \frac{T_n}{S_n} \right) \right)_k U^k. \)

**Proof of Lemma 3.** If we know the generation number $k$ of a sequence, then the number of repeat units in the sequence is the result of $k$ Markovian transitions starting from the number of repeat units of the initial molecules.

The probability that a randomly selected sequence is of $k$-th generation is \( \left( \mathbb{E} \left( \frac{T_n}{S_n} \right) \right)_k = \mathbb{E} \left( \frac{T_{n,k}}{S_n} \right) \). Then

\[
\mathbb{E} \left( \frac{Z_{n,i}}{S_n} \right) = \sum_{k=0}^{n} \mathbb{P} \left( \text{a randomly selected sequence from the pool is of state } i \mid \text{the sequence is of generation } k \right) \times \mathbb{P} \left( \text{the sequence is of generation } k \right)
\]
\[
\sum_{k=0}^{n} \left( \frac{Z_0 S_0}{U^k} \right)_i \left( \mathbf{E} \left( \frac{T_n}{S_n} \right) \right)_k.
\]

Therefore,

\[
\mathbf{E} \left( \frac{Z_n}{S_n} \right) = \sum_{i=\alpha}^{\beta} \mathbf{E} \left( \frac{Z_{n,i}}{S_n} \right) \mathbf{e}_i
\]

\[
= \sum_{k=0}^{n} \sum_{i=\alpha}^{\beta} \left( \frac{Z_0 S_0}{U^k} \right)_i \left( \mathbf{E} \left( \frac{T_n}{S_n} \right) \right)_k \mathbf{e}_i
\]

\[
= \sum_{k=0}^{n} \sum_{i=\alpha}^{\beta} \left( \frac{Z_0 U^k}{S_0} \right)_i \left( \mathbf{E} \left( \frac{T_n}{S_n} \right) \right)_k
\]

\[
= \sum_{k=0}^{n} \left( \frac{Z_0}{S_0} \right)^k \sum_{i=\alpha}^{\beta} \left( \frac{Z_0}{S_0} \right)_i \left( \mathbf{E} \left( \frac{T_n}{S_n} \right) \right)_k
\]

\[
= \frac{Z_0}{S_0} \sum_{k=0}^{n} \mathbf{E} \left( \frac{T_n}{S_n} \right)_k U^k.
\]

**Lemma 4.** Denote \( \frac{T_l}{S_l} = (f_0^{(l)}, f_1^{(l)}, \ldots, f_l^{(l)}, \ldots) \), we have

\[
\sum_{k=0}^{n} \left( \frac{T_l}{S_l} M^{n-1-l}(I - A) \right)_k U^k = \frac{1}{(1+\lambda)^{n-1-l}} \sum_{i=0}^{n-1-l} \sum_{j=0}^{l} (n-1-l)^i \lambda^j f_j^{(l)} (U^{i+j} - U^{i+j+1}),
\]

where \( M \) and \( A \) are defined in Lemma 1.

**Proof of Lemma 4.** This Lemma is similar to Lemma A.2 in a previous study (Lai and Sun, 2003). Here we give a different proof.

Denote \( (1+\lambda)^i = \binom{i}{j} \lambda^j, \quad 0 \leq j \leq i \) and 0 otherwise. Notice that \( f_i^{(l)} = 0 \) for \( i > l \). We have

\[
\sum_{k=0}^{n} \left( \frac{T_l}{S_l} M^{n-1-l}(I - A) \right)_k U^k
\]
\[
= \sum_{k=0}^{n} \left( \frac{T_l}{S_l} M^{n-1-l} - \frac{T_l}{S_l} M^{n-1-l} A \right) U^k
\]

\[
= \frac{1}{(1 + \lambda)^{n-1-l}} \sum_{k=0}^{l} \left( \sum_{j=0}^{n} f_j^{(l)} [(1 + \lambda)^{n-1-l} - (1 + \lambda)^{n-1-l}] \right) U^k
\]

\[
= \frac{1}{(1 + \lambda)^{n-1-l}} \sum_{j=0}^{l} f_j^{(l)} \sum_{k=0}^{n} [(1 + \lambda)^{n-1-l} - (1 + \lambda)^{n-1-l}] U^k
\]

\[
= \frac{1}{(1 + \lambda)^{n-1-l}} \sum_{j=0}^{l} f_j^{(l)} (U^j)[\sum_{i=0}^{n-1-l} (n - 1 - l) \lambda^i U^i - \sum_{i=0}^{n-1-l} (n - 1 - l) \lambda^i U^i U] \]

\[
= \frac{1}{(1 + \lambda)^{n-1-l}} \sum_{i=0}^{n-1-l} \sum_{j=0}^{l} \left( n - 1 - l \right) \lambda^i f_j^{(l)} (U^{i+j} - U^{i+j+1}).
\]

**Lemma 5.** For the matrix \( U \), there exist two matrices \( U_1 \) and \( U_2 \) such that \( U^k = U_1 + U_2^k, \) \( k = 1, 2, \ldots, \) and max \( |U_2^k| \leq cr^k \) for some \( c \) and \( r, \) \( 0 < c < \infty \) and \( 0 < r < 1. \)

This lemma follows from the Perron-Frobenius Theorem (Athreya and Ney, 1972; Harris, 1963) with the largest positive eigenvalue of \( U \) being \( \rho = 1. \)

**Lemma 6.** If a \( d \)-dimensional vector \( v = (v_1, v_2, \ldots, v_d) \) satisfies \( \sum_{i=1}^{d} v_i = 1 \) and \( v_i \geq 0 \) for \( i = 1, 2, \ldots, d, \) then for any \( d \times d \) matrix \( X = \{x_{ij}\} \), we have max \( \{|vX|\} \leq \max \{|X|\}. \)

The proof is obvious.

**Lemma 7.** For any integer \( l > 0, \) \( \mathbb{E} \Theta(S_l) \leq \frac{2}{S_0 \lambda (1-\lambda)} \frac{1}{(1+\lambda)^l}, \) where \( \Theta(\cdot) \) is defined in Lemma 1.
Proof of Lemma 7. Piau (2002) showed that

\[
\Theta(S) \leq \frac{1}{(1 - \lambda^2)(1 + S)} \leq \frac{1}{(1 - \lambda^2)S},
\]
\[
\mathbb{E}(S_t^{-1}) \leq 2 \frac{1 + \lambda}{S_0 \lambda} \frac{1}{(1 + \lambda)^t}.
\]

Combining the above two inequalities, Lemma 7 is proved.

Proof of Theorem 1

\[
\mathbb{E}\left(\frac{Z_n}{S_n}\right) - \frac{Z_0}{S_0} \frac{F^n}{(1 + \lambda)^n}
= Z_0 \frac{\lambda}{S_0} \sum_{k=0}^{n} \left(\mathbb{E}\left(\frac{T_n}{S_n}\right)\right)_k U^k - Z_0 \frac{\lambda}{S_0} \sum_{k=0}^{n} \left(\frac{T_0}{S_0} M^n\right)_k U^k \quad \text{(Lemmas 2 and 3)}
= Z_0 \frac{\lambda}{S_0} \sum_{k=0}^{n} \left(\mathbb{E}\left(\frac{T_n}{S_n}\right) - \frac{T_0}{S_0} M^n\right)_k U^k.
\]

From Lemma 7, we have

\[
\lim_{S_0 \to \infty} \Theta(S_t) = 0.
\]

By Lemma 1, we have

\[
\lim_{S_0 \to \infty} \mathbb{E}\left(\frac{T_n}{S_n}\right) = \frac{T_0}{S_0} M^n.
\]

Therefore,

\[
\lim_{S_0 \to \infty} \mathbb{E}\left(\frac{Z_n}{S_n}\right) = \frac{Z_0}{S_0} \frac{F^n}{(1 + \lambda)^n}.
\]

□

Proof of Theorem 2

\[
\mathbb{E}\left(\frac{Z_n}{S_n}\right) - \frac{Z_0}{S_0} \frac{F^n}{(1 + \lambda)^n}
= Z_0 \frac{\lambda}{S_0} \sum_{k=0}^{n} \left(\mathbb{E}\left(\frac{T_n}{S_n}\right)\right)_k U^k - Z_0 \frac{\lambda}{S_0} \sum_{k=0}^{n} \left(\frac{T_0}{S_0} M^n\right)_k U^k \quad \text{(Lemmas 2 and 3)}
\]

Using Lemma 6, we have

\[
\max \left| \mathbf{E} \left( \frac{Z_n}{S_n} \right) - \frac{Z_0}{S_0} \frac{F^n}{(1 + \lambda)^n} \right| \\
\leq  \max \sum_{l=0}^{n-1} \mathbf{E} \left[ \frac{\lambda(1 - \lambda) \Theta(S_l) (1 + \lambda)^{n-l}}{n-l} \sum_{i=0}^{n-l} \sum_{j=0}^{l} \binom{n-l}{i} \lambda^i f_j^{(l)} (U_{i+j}^{i+j+1} - U_{i+j}^{i+j+1}) \right] \\
= \sum_{l=0}^{n-1} \mathbf{E} \left[ \frac{\lambda(1 - \lambda) \Theta(S_l) (1 + \lambda)^{n-l}}{n-l} \sum_{i=0}^{n-l} \sum_{j=0}^{l} \binom{n-l}{i} \lambda^i f_j^{(l)} (\max |U_{i+j}^{i+j+1}| + \max |U_{i+j+1}^{i+j+1}|) \right] \\
= \sum_{l=0}^{n-1} \frac{c\lambda(1 - \lambda)(1 + r)}{(1 + \lambda)^{n-l}} \mathbf{E} \left[ \Theta(S_l) (1 + \lambda)^{n-l} \sum_{j=0}^{l} f_j^{(l)} \right] (0 < r < 1) \\
= \frac{c\lambda(1 - \lambda)(1 + r)}{(1 + \lambda)^{n-l}} \sum_{i=0}^{n-1} \frac{(1 + \lambda r)^{n-l}}{(1 + \lambda)^{n-l}} \mathbf{E} [\Theta(S_l)]
\]
\[ \leq c\lambda(1 - \lambda)(1 + r) \sum_{l=0}^{n-1} \frac{(1 + \lambda r)^{n-1-l}}{(1 + \lambda)^{n-1-l}} \left( \frac{2}{\lambda(1 - \lambda^2)} \frac{1}{(1 + \lambda)^l} \frac{1}{S_0} \right) \] (Lemma 7)

\[ = \frac{2c(1 + r)(1 + \lambda)[(1 + \lambda r)^n - 1]}{S_0 r \lambda (1 + \lambda)^n}. \]

Adjusting the constant in the formula, Theorem 2 is proved. \( \square \)