1949

Limit theorems and convergence rate for longest contaminated runs of heads

Thesis for the Degree of Doctor of Philosophy (PhD)

written by SUJA MICHAEL OCHIENG
and supervised by PROF. Dr. ISTVÁN FAZEKAS

University of Debrecen
Doctoral Council of Natural Sciences and Information Technology
Doctoral School of Mathematical and Computational Sciences
Debrecen, 2024

Hereby I declare that I prepared this thesis within the Doctoral Council of Natural Sciences and Information Technology, Doctoral School of Mathematical and Computational Sciences, Doctoral Program of Probability Theory, Mathematical Statistics and Applied Mathematics, University of Debrecen in order to obtain a PhD Degree in Natural Science, at Debrecen University.
I declare that the results published in this thesis are not reported in any other PhD theses.

Debrecen, 2024

. . . .
Suja Michael
candidate

Hereby I confirm that Suja Michael Ochieng candidate conducted his studies with my supervision within the Doctoral Program of Probability Theory, Mathematical Statistics and Applied Mathematics in the Doctoral School of Mathematical and Computational Sciences of the University of Debrecen between 2019 and 2024. The independent studies and research work of the candidate significantly contributed to the results published in the thesis.
I also declare that the results published in the thesis are not reported in any other PhD theses.
I support the acceptance of the dissertation.

Debrecen, 2024....

. . . .
Dr. István Fazekas
supervisor

Limit theorems and convergence rate for longest contaminated runs of heads

Dissertation submitted in partial fulfilment of the requirements for the doctoral (PhD) degree in Mathematics and Computing.

Written by: Suja Michael Ochieng.

The dissertation was written in the framework of the Probability Theory, Mathematical Statistics and Applied Mathematics Program of the Doctoral School of Mathematical and Computational Sciences of the University of Debrecen.

Dissertation advisor: Prof. Dr. Fazekas István

The official opponents of the complex exam:

chairperson:	Dr. Hadju Lajos	UD
members:	Dr. Miklós Arató	ELTE
	Dr. Baran Sándor	UD

The date of the complex exam: 2021.06.15.

The official opponents of the dissertation:

Dr. ....

Dr. ....

Dr. ....

The evaluation committee:

chairperson: Dr. ....

members: Dr. ....

Dr. ....

Dr. ....

Dr. ....

The date of the dissertation defence: 20 . . . .

Acknowledgements

I am deeply grateful to my supervisor Professor István Fazekas for his invaluable guidance, continued support, constructive feedback throughout this research journey. He first suggested the topic Limit Theorems for Contaminated runs and guided my first, not very easy steps. Without his advice and unique support, this thesis would never have become a reality. It was in deed a real honor to share not only in his exceptional scientific knowledge but also his extra ordinary humane qualities. He trained me with infinite amount of patience and encouraged me to diversify my approach to handling different tasks. He often granted me permission to visit my family every December holiday. I consider myself privileged to have him as my doctoral advisor.

I am also grateful to Professor Baran Sándor for offering me different courses in computer statistics in a very interactive manner. My appreciation also goes to Professor Zsolt Páles for his guidance throughout this period. I also express my appreciation to all faculty members of the department for their administrative assistance, academic interactions, invaluable advice and prompt response to queries throughout my academic journey.

I am so grateful to the Government of Hungary for granting me the Stipendium Hungaricum Scholarship throughout my program. I heartily extend my thanks to all my colleagues and friends particularly Barta Attila for his encouragement and insightful guidance in simulation that enabled me to realize the results in this research.

I want to express my heartfelt appreciation to my family for the unwavering love, understanding and encouragement during the challenging moments of this journey. To my mother Mary Suja, my dear lovely wife Rose Malii, lovely son Ian Okoth and daughter Ashley Akinyi, I salute you all for enduring with my long absence from home. I promise to make up for this.

Thank you.

Abstract

The study of success runs in Bernoulli trials has attracted indubitable attention of several researchers both for its inherent theoretical interest and intriguing applications in numerous scientific fields. In this PhD dissertation, we study T-contaminated head runs for the cases $T = 1, 2$ in which we present direct probabilistic calculations.

We focus on the limiting distributional problems of run related random variables. This include Compound Poisson distribution as the limiting distribution of the number of the at most T-contaminated head run, exponential distribution as the limiting distribution of the first hitting time of a specified length of T-contaminated head run and accompanying distribution as asymptotic distribution for the length of the longest T-contaminated head run.

This dissertation consists of four chapters; Introduction, Limit theorems of T-contaminated runs of heads, Convergence rate for the longest T-contaminated head runs and Limit theorems for runs containing two types of contamination together with Summary and Appendix. It is based on three published papers with the candidate as author.

In chapter 1, we present some basic definitions and notations useful in the sequel. We introduce the theorems regarding the number, the waiting time of T-contaminated head run together with theorem of the accompanying distribution of the length of the longest T-contaminated head run. Simulation results are provided to reinforce our theoretical findings.

In chapter 2, we find the asymptotic distribution for the first hitting time of the T-contaminated run of heads having a specified length. Further to this, we concentrate on obtaining a limit theorem for the length of the longest T-contaminated head run. We give a proof that the rate of convergence of our approximation of the accompanying distribution for the length of the longest T-contaminated head run performs exceedingly better than previous known results. We provide

accompanying simulation results for the same.

In chapter 3, we study sequences of trials having three outcomes labelled; success, failure of type I and failure of type II. We obtain the limiting distribution of the first hitting time and the accompanying distribution for the length of the longest at most two-type contaminated run. Besides the mathematical proofs, we provide simulation results supporting our theorems.

In chapter 4, we give a summary of chapters; 1, 2 and 3. Finally, we give in the Appendix, the main lemma of Csaki et al. We rewrite the proof for the non stationary case, finite form giving some additional explanation with a goal of precisely fixing the conditions of the lemma. We correct the misprints and omissions noted in the lemma which are important for our subsequent applications.

Preface

Limit theorems pertain to the limits of sequences of sums of random variables, Sn when the number of summands approaches infinity, subject to specific probabilistic assumptions. The purpose of these restrictions is to guarantee that the probabilistic behavior of the summation, Sn, is not dominated by any individual element of the stochastic process. It is worth to note the existence of several conditions that guarantee the modest effect of each individual summand on the behavior of Sn, leading to a wide range of limit theorems falling under the classification of classical limit theorems.

In the field of probability, limit theorems encompass a fundamental and vast research domain within probability theory. This area primarily focuses on the emergence of asymptotic (limiting) laws that arise from the analysis of a long series of observations on random events. As a result of this, and as early as the 20^{th} century, certain researchers in the field have conceptualized the probability of an event as the theoretical limit of the relative frequency of its occurrence in a long sequence of independent random experiments. It is commonly assumed that the random variables exhibit either independence or a degree of almost independence.

The limit theorems have been subject to rigorous proofs and advancements by numerous esteemed probabilists over the course of several centuries. The proof of the first limit theorem of probability theory, also known as the Law of Large Numbers, was provided by Bernoulli (1654-1705). The topic was then taken up by De Moivre (1667-1754), who stated and proved the first central limit theorem, which pertains to the convergence to a normal distribution, specifically for symmetric Bernoulli trials. In (1812), Laplace expanded upon the work of De Moivre by providing foundational insights crucial in shaping the understanding of probability and behaviour of sums of random variables paving the way for the formalization of the two previously established limit theorems. Poisson in the (1830s) worked on establishing the conditions under which the sample mean converges to the population mean and consequently weakened the conditions that underlie the law of large numbers.

The next significant development occurred in the (1870s) with the establishment of the Russian school of probability, spearheaded by Chebychev. Within this school, Chebychev and his pupil Markov successfully proved the central limit theorem through the utilization of the method of moments. Later on, Lyapunov provided the proof of the aforementioned result using the method of characteristic function, which is well recognized as a powerful analytical technique for establishing a multitude of other limit theorems. Kolmogorov not only enhanced the contributions of his predecessors but also established the modern mathematical foundations of probability theory in (1933).

Erdős and Rényi (1970), in their classical paper, 'On the new law of large numbers' gave renewed scholarly interest in asymptotic estimation of the length of the longest run. Their findings solidified the longtime existence of the study of the length of the longest run and the first hitting time distributions in the mainstream research on the nature of randomness. Exact distributional results, as well as the weak and almost sure asymptotic behaviour have been characterised completely.

This doctoral dissertation is written as a monograph and is based on the following three papers published in peer reviewed journals:

Fazekas and Suja (2021), Limit theorems for T-contaminated runs of heads, Annales Univ. Sci, vol. 52, (2021), pp.131–146.

Fazekas, Fazekas, and Suja (2024), Convergence rate for the longest T-contaminated runs of heads, Statistics & Probability Letters, vol. 208, pp 110059, (2024).

Fazekas, Fazekas, and Suja (2023), Limit theorems for runs containing two types of contaminations, arXiv preprint arXiv:2309.11602, (2023).

Notations and Symbols

T	Number of tails interrupting (contaminating) head run sequence.
N	The number of coin tossing or the length of the experiment.
$\tilde{\xi}^T(n,N)$	Number of precisely T-contaminated head runs of length n.
$\xi^T(n,N)$	Number of at most T-contaminated head runs of length n.
$\tau^T(n)$	The first hitting time of the at most T-contaminated run of length n.
$\mu^T(N)$	Length of the longest T-contaminated head run in a single experiment.
$[x]$	Represents the largest integer less than or equal to x.
$f = O(g)$	Growth rate of a function, that is $f(x)/g(x)$ remains bounded as $x \to \infty$.
$f \sim g$	Asymptotic equality, that is $f(x)/g(x) \to 1$ as $x \to \infty$.
$\mathbb{I}\{A\}$	Indicator function of a subset A assuming values 0 or 1.

Contents

List of Tables

List of Figures

Introduction

Limit theorems form an evergreen field of probability theory. Its methods and results continue to have huge fundamental impact on many fields, including statistics, engineering, finance, and science. For general background on probability theory and limit distributions see [**Gut** (**2005**), **Shiryaev** (**2016**), **Ash and Doléans-Dade** (**2000**) and **Petrov** (**1995**)].

One of the most widely studied aspects of probability theory is the concept of coin tossing experiments. While seemingly simple, coin tossing experiments have brought some of the most profound discoveries in probability theory.

Considering a series of independent trials where each success outcome has a probability p of occurrence, the total number of successes has the binomial(n, p) probability distribution. This is a well known fundamental fact of probability theory. As the number of trials increase while p remains fixed, the central limit theorem always provides a good approximations to the binomial probabilities based on the normal probability distribution. This approximation was first given by de Moivre, but later refined and popularised by Laplace.

On the other hand, if p is allowed to decrease such that np tend to a positive finite limit, then Poisson based approximation becomes the most suitable probability distribution for the binomial probabilities. This approximation remained overshadowed by the normal approximation until its uniqueness and expansive domain of applications was popularised by researchers like Bortkiewicz. Since then, broad applications and elementary proofs of convergence of binomial to Poisson probabilities became a dominant course in probability theory. It has been considered in connection to; possibly dependent trials, trials with varying success probabilities and even trials with more than two possible outcomes.

The study of runs, which is a sequence of consecutive events of similar type preceded or succeeded by different types of events started towards the end of the 19^{th} century rather than in the days of Laplace when a lot of interest was in game of chances. In particular, the study of success runs in Bernoulli trials has attracted unquestionable attention of several researchers both for its inherent theoretical interest and its intriguing applications in numerous scientific fields (Psychology, Meteorology, Reliability engineering, Quality Control, DNA sequence matching etc.). This has invoked several variations, extensions and generalizations of runs related statistics. Some of the most commonly used in statistics and applied probability in the sense of Bernoulli trials include among others; number of non-overlapping consecutive k successes according to Feller(1968), number of success runs of size exactly k according to Mood's (1940), number of success runs of size greater or equal to k and the length of the longest success run. Their waiting times have also been considered consequently.

The problem of the length of the longest head run for n Bernoulli random variables was first raised by T. Varga in his classroom experiment. He divided the class into two groups; assigned one group a fair coin and the other told to basically simulated the coin tossing experiment without using a coin. The fact that the results from the two groups of students could easily be separated once the experiment was over stimulated a lot of interest and provided illustrious topic to bring out important techniques of probability theory. These included recursion arguments involving combinatorics, asymptotic analysis and concepts of limiting distributions. Theses approaches have been applied in determining the bounds, exact and limiting lengths of longest runs.

The first consideration for the length of the longest run for the case of a fair coin was given in the classical paper 'On the new law of large numbers' by **Erdős and Rényi (1970)** where an asymptotic estimation of the length of the longest run was given. Numerous subsequent extensions of this result emerged in applications to renewal and other stochastic processes.

Various related problems to the length of the longest run considered by other researchers include; longest increasing run of independent uniformly distributed random variables, longest repetitive patterns in random sequence and longest matching sequences in two DNA strings. The length of the longest run belongs to a class of non-convergent random variables with no limiting distribution.

The exact distributions of these runs statistics were traditionally studied using combinatorial analysis. However, finding appropriate combinatorial identities in order to derived the probability distribution proved difficult for the case of complex runs and to date some exact distributions of many common statistics remain unknown. Other approaches for determining exact probability distribution involves generating function technique introduced by Feller (1968) using the theory of recursion. For more on general note on the number of long run see [**Fazekas et al. (2010)**, **Nguyen et al. (2016)**, **Erdős and Rényi (1970)** and **Muselli**

(**2000a**)].

However, for more complex runs their generating functions became too cumbersome to differentiate a large number of times. So other researchers explored the method of Markov chain embedding technique to determine the underlying distribution of runs statistics see [**Novak** (**1989**) and **Samarova** (**1982**)]. The method can be applied to the case of both independent identically distributed random variables as well as to Markov dependent multi-state trials regardless of the specified counting procedure.

A natural and more intuitively appealing generalization of runs statistics do emerge in situations where instead of considering specified sequence length strings with all positions occupied by successes, some allowance is given for the appearance of a small number of failures. Therefore, the focus shifts to consecutive trials which contain a large proportion of successes. Such formations are widely referred to as contaminated (interrupted) success runs which is the key object of our study.

In this dissertation, we focus on limiting distributional problems associated with this specific type of runs, namely, contaminated runs of heads in a sequence of coin tosses and to other sequences from experiments with more than two possible (i.e. trinary) outcomes. More precisely, we investigate the limiting distribution of the number of contaminated runs, limiting distribution of the first hitting time of contaminated runs and accompanying distribution of the length of the longest contaminated runs.

The rest of this dissertation is organized as follows. In Chapter 1, we define a T-contaminated run of heads and study the limiting distributions of their numbers together with the first hitting time and the asymptotic behaviour of the length of the longest T-contaminated head run. Our emphasis will be devoted to approximation of the numbers of contaminated runs to both Poisson and compound Poisson limit laws.

Chapter 2, we shall still deal with T-contaminated head run but emphasis now shifts to the distribution of the length of the longest T-contaminated head run. We shall investigate the rate of convergence to an accompanying distribution and also obtain results for the first hitting time for the same.

In Chapter 3, we shall introduce a two type contaminated runs and study the limiting distribution of the first hitting time and the accompanying distribution of the length of the longest at most two type contaminated runs with trinary outcomes. Our approach mirrors the one used in Chapter 2.

Finally, we shall give a summary of our most important results and thereafter, provide a useful auxiliary materials collected in the appendix. This contains basically the main lemma of **Csáki et al.** (**1987**) where we precisely fix its conditions.

Chapter 1

Limit theorems of T-contaminated run of heads

The aim of this chapter is to determine the limit distribution of the number of T-contaminated run of heads, the limiting distribution of the first hitting time of T-contaminated run of heads having a given length and finally the limiting law of the length of the longest T-contaminated head run. For the case of independent random variables, the number of precisely T-contaminated head runs has a binomial distribution which can well be approximated by the Poisson law. This attracted the attention of famous scientists who have worked on the problem of evaluating the accuracy of Poisson approximation to the binomial distribution. See **Serfling (1978)** for appropriate metrics for measuring the error bounds and an overview of applications of Poisson approximations. For the case of dependent random variables, the possible limiting distribution is Compound Poisson limit laws. For more detailed information regarding these approximations, see [**Novak (2019)**, **Peköz (2006)**, **Barbour and Chryssaphinou (2001)**, **Barbour and Månsson (2002)**, **Chryssaphinou and Papastavridis (1988)**].

Investigating various corresponding distributional limit theorems of the above statistics requires different methods. The most useful tools employed by various researchers involves recursion/ combinatorics, probability generating function and Markov chain embedding techniques see [**Kopocinski (2016)**, **Muselli (2000b)**, **Novak (1992)**, **Schilling (1990)**, **Schilling (2012)**, **Karácsony and Libor (2011)**].

A simple rule of thumb can accurately predict the length $\mu(N)$ of the longest sequence of success if a situation can be modeled as a series of independent Bernoulli trials. For infinite sequence of fair coin tosses, a simple theorem of Rényi says that the length of the longest run in N tosses is about $\log N / \log 2$.

Erdős and Rényi (**1970**) gave an asymptotic estimation of the length of the longest run and proved that for a fair coin and for an arbitrary $0 < C_1 < 1 < C_2 < \infty$ with almost all $\omega \in \Omega$, there exist a finite number $N_0 = N_0(\omega, C_1, C_2)$ such that;

$$[C_1 \log N] \leq \mu(N) \leq [C_2 \log N]$$

if $N \geq N_0$. Here log denotes logarithm to base 2 and $[.]$ is the integer part. $\mu(N)$ has a logarithmic growth $C \log N$ as $N \to \infty$.

Later on, **Erdős and Révész** (**1975**) improved on the above bounds by using combinatoric approach. They provided surprisingly more precise results i.e. almost sure results for the bounds. For $p = 1/2$ and $\varepsilon > 0$, the Theorem states as follows; Let ε be any positive number. Then for almost all $\omega \in \Omega$, there exists a finite number $N_0 = N_0(\omega, \varepsilon)$ such that;

$$Z_N \geq [\log N - \log\log\log N + \log\log e - 2 - \varepsilon]$$

if $N \geq N_0$.
Furthermore, for almost all $\omega \in \Omega$, there exists an infinite sequence $N_i = N_i(\omega, \varepsilon)$ $(i = 1, 2, \cdots)$ of integers such that;

$$Z_{N_i} < [\log N_i - \log\log\log N_i + \log\log e - 1 + \varepsilon]\,.$$

This was a generalization of the problem of the length of the longest head run containing no tails at all. Moreover, they also found analogous results for the upper and lower bounds for the longest run of heads containing at most T tails as given below;
Let ε be any positive number. Then for almost all $\omega \in \Omega$, there exists a finite number $N_0 = N_0(\omega, T, \varepsilon)$ such that;

$$Z_N(T) \geq [\log N + T \log\log N - \log\log\log N - \log T! + \log\log e - 2 - \varepsilon]$$

if $N \geq N_0$.
Furthermore, for almost all $\omega \in \Omega$, there exists an infinite sequence $N_i = N_i(\omega, T, \varepsilon)$ of integers such that;

$$Z_{N_i}(T) < [\log N_i + T \log\log N_i - \log\log\log N_i - \log T! + \log\log e - 1 + \varepsilon]\,.$$

Since the predicted length of the longest run grows logarithmically, i.e. the distribution tends to shift towards larger values at a rate logarithmically related to N, it became reasonable to consider the approximate distribution of the prediction error to provide the accompanying limiting distribution which is independent of N.

Földes (1979) provided an asymptotic estimation of the distribution of the longest pure head run $\mu(N)$ in the case of a fair coin.

$$\mathbb{P}\left(\mu(N) - \left[\frac{\log N}{\log 2}\right] < k\right) = \exp\left(-2^{-\left(k+1-\left\{\frac{\log N}{\log 2}\right\}\right)}\right) + o_N(1).$$

She also stated and proved the limit theorem for the longest head run $\mu^T(N)$ containing at most T tails for the case of a fair coin.

$$\mathbb{P}\left(\mu^T(N) - [\log N + T\log\log N] < k\right) = \exp\left(-\frac{2^{-(k+1-\{\log N + T\log\log N\})}}{T!}\right) + o(1).$$

Guibas and Odlyzko (1980) using generating function methods provided deep results on problems related to those of **Erdős and Révész (1975)**. They looked at the longest run of repetition of specified pattern. They computed the expectation and variance of $\mu(N)$ of repetitions and made intriguing observation that $\mu(N)$ has no limiting distribution. Several results emanating from their findings are generalized to the case of biased coin tossing and runs with at most T-interruptions in **Gordon et al. (1986)**.

In **Gordon et al. (1986)**, a sequence of coin tossing is represented in terms of independent geometric random variables and then analysed using inclusion-exclusion counting methods. They proved using extreme value theorem that the probabilistic behaviour of the length of the longest pure head run is closely approximated by the greatest integer function of the maximum of nq of independent identically distributed exponential random variables. Theses results are extended to the case of the longest head runs interrupted by T tails. They further evaluated the mean lengths of this type of run together with the variance.

Deheuvels (1985) offered almost sure upper and lower bounds for the k^{th} longest head run for a biased coin. This was a generalization of **Erdős and Révész (1975)** in the estimation for the length $\mu(N)$.

Túri (2008) used the connection between the pure head runs and pure runs to derive the so called almost sure limit theorem for the longest runs (see, for example, in **Schilling (1990)** and **Fazekas et al. (2010)**).

The above fascinating findings motivated our research into this rich field of limit theorems in probability theory. The main results of this chapter were proved by

Földes (**1975**) for the fair coins. Here we extend her results to the case of biased coins and moreover present a more general result for the longest T-contaminated head runs.

In the first part of this chapter, we formulate the requisite conditions from which both Poisson and compound Poisson laws provide good approximations for the limit laws of our random variables of interest. For more detailed insight into this subject approach, see **Sevast'yanov** (**1972**). Consequently, we first prove that the limit distribution of the number of precisely T-contaminated head run converges to a Poisson distribution. To gain more insight into Poisson approximation, see (**Arratia et al.** (**1990**), **Chryssaphinou and Papastavridis** (**1988**)).

In the second part of this chapter, we prove in details that the limiting distribution of the number of at most T-contaminated run of heads converges to a Compound Poisson distribution, see **Barbour and Chryssaphinou** (**2001**) for related studies.

In the third part of this chapter, we prove in brief that the limiting distribution of the first hitting time of a T-contaminated head run of a fixed length in deed converges to an exponential distribution just like any other waiting time distributions.

Finally, in the last part of the chapter, we derive the accompanying distribution of the length of the longest T-contaminated run of heads as it lacks a limiting distribution. For further insight, see (**Kopocinski** (**2016**), **Muselli** (**1996**) and **Binswanger and Embrechts** (**1994**)). We shall also provide simulation results to buttress our theoretical results.

1.1 Number of precisely T-contaminated run of heads

We consider the classical coin tossing experiment. Let $p \in (0,1)$ be the probability of heads and $q = 1 - p$ the probability of tails. Here, p is fixed while we toss a coin N times independently. We write 1 when the result is head and 0 when the result is tail. Therefore we consider independent identically distributed random variables $X_1, X_2, \ldots, X_N$ with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = q$, $i = 1, 2, \ldots, N$. Let $T \geq 0$ be fixed integer.

Let $\tilde{\xi} = \tilde{\xi}^T(n, N)$ denote the number of those precisely T-contaminated n -

length runs of heads for which the previous element is a tail. More precisely let

$$\tilde{\eta}_i = \tilde{\eta}_i^{\,T}(n) = \begin{cases} 1, & \text{if there are precisely } T \text{ 0 values among} \\ & X_i, \ldots, X_{i+n-1} \text{ and } X_{i-1} = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{1.1}$$

Here let X_0 be defined as $X_0 = 0$. Now we let

$$\tilde{\xi} = \tilde{\xi}^T(n, N) = \sum_{i=1}^{N-n+1} \tilde{\eta}_i^{\,T}(n). \tag{1.2}$$

So $\tilde{\xi}$ can be considered as the number of those precisely T-contaminated head runs having length n for which the previous value is tail.

Our main condition is the following. Let $p \in (0, 1)$ be fixed. Let T be a fixed non-negative integer. If we let $N \to \infty$ and $n \to \infty$ so that

$$\frac{N q^{T+1} p^{n-T} n^T}{T!} \to \lambda > 0, \tag{1.3}$$

where λ is fixed, then we remark that condition (1.3) implies that $N/n \to \infty$.

Now we intend to show that the distribution of $\tilde{\xi}$ converges to the λ parameter Poisson distribution.

Theorem 1.1.1. *(**Fazekas and Suja (2021)**) Let T be fixed. Let $N \to \infty$ and $n \to \infty$ so that condition (1.3) is satisfied. Then*

$$\lim_{N\to\infty} \mathbb{P}(\tilde{\xi}^T(n, N) = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, \ldots$$

The above theorem is proved in **Földes (1975)** for the case of T-contaminated head run where $p = 1/2$ and $T > 0$ (see Theorem 3.1) and also proven in **Földes (1979)** for uncontaminated head run where $p = 1/2$ and $T = 0$ (see Theorem 1.A).

To obtain the proof of our theorems, we need the following known result from triangular arrays of random variables.

Proposition 1.1.1. *(See **Sevast'yanov (1972)**)*
Let $Y_i^{(m)}$, $i = 1, 2, \ldots, l_m$, $m = 1, 2, \ldots$, be a triangular array of Bernoulli random variables, i.e. the values of $Y_i^{(m)}$ are 0 or 1. If for every m, $Y_i^{(m)}$ are independent. So let

$$\mathbb{Z}_m = Y_1^{(m)} + Y_2^{(m)} + \cdots + Y_{l_m}^{(m)}, m = 1, 2, \ldots$$

be the row sums. Let

$$b^{(m)}_{i_1,i_2,\dots,i_r} = \mathbb{P}({Y_{i_1}}^{(m)} = {Y_{i_2}}^{(m)} = \dots = {Y_{i_r}}^{(m)} = 1),$$

where $(i_1, i_2, \dots, i_r)$ *denotes an* r *dimensional vector such that integers* $i_1, i_2, \dots, i_r$ *are pairwise different with* $1 \le i_t \le l_m$, $t = 1, 2, \dots, r$, $r = 1, 2, \dots$.
Assume that for each $r = 2, 3, \dots$, $m = 1, 2, \dots$ *there exists an exceptional set* $I_r(m)$ *consisting of certain vectors* $\alpha_r = (i_1, i_2, \dots, i_r)$ *such that the numbers* $i_1, i_2, \dots, i_r$ *are pairwise different with* $1 \le i_t \le l_m$, $t = 1, 2, \dots, r$.
In addition, we assume the following that

$$\lim_{m\to\infty} \max_{1\le i\le l_m} {b_i}^{(m)} = 0, \tag{1.4}$$

$$\lim_{m\to\infty} \sum_{i=1}^{l_m} {b_i}^{(m)} = \lambda > 0, \tag{1.5}$$

$$\lim_{m\to\infty} \sum_{\alpha_r \in I_r(m)} b^{(m)}_{i_1,i_2,\dots,i_r} = 0, \tag{1.6}$$

$$\lim_{m\to\infty} \sum_{\alpha_r \in I_r(m)} b^{(m)}_{i_1} \cdots b^{(m)}_{i_r} = 0, \tag{1.7}$$

and uniformly for all $\alpha_r \notin I_r(m)$

$$\lim_{m\to\infty} \frac{b^{(m)}_{i_1,i_2,\dots,i_r}}{b^{(m)}_{i_1} \cdots b^{(m)}_{i_r}} = 1. \tag{1.8}$$

Then

$$\lim_{m\to\infty} \mathbb{P}(\mathbb{Z}_m = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, \dots \tag{1.9}$$

Proof of Theorem 1.1.1. We apply Proposition 1.1.1 for $l_m = N - n + 1$ and $Y_i^{(m)} = \tilde{\eta}_i$, $i = 1, 2, \dots, l_m$. So we first check the fulfilment of the conditions of Proposition 1.1.1.
Condition (1.4) is satisfied because

$$\begin{aligned}\max_{1\le i\le N-n+1} b_i &= \max_{1\le i\le N-n+1} \mathbb{P}(\tilde{\eta}_i = 1) \\ &= \max\{1, q\}\binom{n}{T} q^T p^{n-T} \le c n^T p^n \to 0\end{aligned} \tag{1.10}$$

as $N, n \to \infty$, because $0 < p < 1$.
Condition (1.5) is satisfied because, by applying our main condition (1.3),

$$\begin{aligned}\sum_{i=1}^{N-n+1} b_i = &\sum_{i=1}^{N-n+1} \mathbb{P}(\tilde{\eta}_i = 1) \\ =&(i+(N-n)q)\binom{n}{T} q^T p^{n-T} \approx N \frac{n^T}{T!} q^{T+1} p^{n-T} \to \lambda\end{aligned} \tag{1.11}$$

as $n, N \to \infty$. Here we applied that, by condition (1.3), $n/N \to 0$.
To check condition (1.8),
we let $\alpha_r = (i_1, i_2, \ldots, i_r)$ denote an r dimensional vector such that the numbers $i_1, i_2, \ldots, i_r$ are pairwise different with $1 \le i_t \le N-n+1$, $t = 1, 2, \ldots, r$.
We define the set $I_r(n, N)$ of exceptional indices as the set of those indices $\alpha_r = (i_1, i_2, \ldots, i_r)$ such that there are $i, j \in \{1, 2, \ldots, r\}$, $j \ne l$ with $|i_j - i_l| < n+1$.

The random vectors $X_{i-1}, X_i, \ldots, X_{i+n-1}$ and $X_{j-1}, X_j, \ldots, X_{j+n-1}$ are independent if $n < j - i$. Therefore, $\tilde{\eta}_{i_1}, \tilde{\eta}_{i_2}, \ldots, \tilde{\eta}_{i_r}$ are independent if $\alpha_r = (i_1, i_2, \ldots, i_r) \notin I_r(n, N)$. So, condition (1.8) is satisfied.

Now we turn to condition (1.7). By the definition of $I_r(n, N)$, we should choose r elements out of n so that there should be a pair among them with distance being not greater than n. Therefore

$$\begin{aligned}\sum_{\alpha_r \in I_r(n,N)} b_{i_1}, \cdots, b_{i_r} = &\sum_{\alpha_r \in I_r(n,N)} \mathbb{P}(\tilde{\eta}_{i_1} = 1) \cdots \mathbb{P}(\tilde{\eta}_{i_r} = 1) \\ \le&\binom{N}{r-1}(r-1)2n\left[\binom{n}{T} q^T p^{n-T}\right]^r \le c\frac{n}{N}\left(N n^T p^n\right)^r \to 0\end{aligned} \tag{1.12}$$

as $n, N \to \infty$, because of our main condition (1.3).

We now consider condition (1.6).
For $r = 1$, conditions (1.6) and (1.7) are equivalent. For $r \ge 2$ and $T = 0$, we have

$$\sum_{\alpha_r \in I_r(n,N)} b_{i_1, \cdots, i_r} = \sum_{\alpha_r \in I_r(n,N)} \mathbb{P}(\tilde{\eta}_{i_1} = 1, \ldots, \tilde{\eta}_{i_r} = 1) = 0, \tag{1.13}$$

because in the definition of $\tilde{\eta}_i$, we claim that $X_{i-1} = 0$.

We shall now prove condition (1.6) for $T \ne 0$.
For any $\alpha_r = (i_1, i_2, \ldots, i_r)$, the indices of X_i variables involved belong to the intervals

$$[i_1 - 1, i_1 + n - 1], [i_2 - 1, i_2 + n - 1], \ldots, [i_r - 1, i_r + n - 1]. \tag{1.14}$$

If $\alpha_r \in I_r(n,N)$, then at least two of the above intervals have a common point. So we can divide the family of intervals (1.14) into disjoint components so that inside each component, the intervals are connected. We emphasize that the random variables having indices in disjoint components are independent.
Therefore, the term $\sum_{\alpha_r\in I_r(n,N)} \mathbb{P}(\tilde{\eta}_{i_1}=1,\dots,\tilde{\eta}_{i_r}=1)$ is a sum of the products of terms corresponding to connected components. Now using this fact and equation (1.11), we can see that it is enough to prove condition (1.6) for a connected set of intervals (1.14).
So we let $I_r^*=I_r^*(n,N)$ be the set of indices $\alpha_r\in I_r(n,N)$ with a connected family (1.14) of intervals. We denote by s the overall length of these intervals. Then $n+r\le s\le rn+1$.
We then divide I_r^* into two parts; $\alpha_r\in I_r^*(1)=I_r^*(1,n,N)$ if and only if $s\le 2n+1$ while $\alpha_r\in I_r^*(2)=I_r^*(2,n,N)$ if and only if $s>2n+1$.
So in the case of $I_r^*(1)$, there is a common point of the intervals (1.14) but in the case of $I_r^*(2)$, the first and the last intervals are disjoint.
In the case of $I_r^*(2)$, a roughly upper bound will do hence

$$\begin{aligned}\sum_{\alpha_r\in I_r^*(2)} \mathbb{P}(\tilde{\eta}_{i_1}=1,\dots,\tilde{\eta}_{i_r}=1) \le & N\sum_{s=2n+2}^{rn+1}\sum_{j=2T}^{rT}\binom{s}{j}q^j p^{s-j}\\ \le & cNp^{2n-rT}(rn+1)^{rT}(rn+1)(rT)\le cNp^{2n}n^{rT+1}\\ = & c(Np^n n^T)(p^n n^{(r-1)T+1})\to 0\end{aligned}$$

as $n,N\to\infty$, because of our main condition (1.3).
Next, we consider the case of $I_r^*(1)$, that is when $s\le 2n+1$. Given that the intersection of all the intervals in (1.14) is not empty. The case of $r>T+1$ is impossible as then, at least $r-1$ tails $(r-1>T)$ would be in the first interval. So we let $r\le T+1$. Concerning the location of the intervals (1.14), we let $l_j=i_{j+1}-i_j$, $j=1,2,\dots,r-1$. If $l_1,\dots,l_{r-1}$ are fixed, then the locations of $r-1$ tails are given. So in the first interval, we choose the locations of $T-r+1$ tails. The starting point of the first interval can be chosen less than N different ways. Moreover, the probability that at most T tails occur from l tosses is not greater than $q_0^l(T+1)l^T$, where $q_0=\max\{q,p\}<1$. Therefore

$$
\begin{aligned}
&\sum_{\alpha_r \in I_r^*(1)} \mathbb{P}(\tilde{\eta}_{i_1} = 1, \dots, \tilde{\eta}_{i_r} = 1) \le \\
&\le N \sum_{1 \le l_1, \dots, l_{r-1} \le n} \left[\binom{n}{T-r+1} q^T p^{n-T} \right] q_0^{l_1} (T+1) l_1^T \cdots q_0^{l_{r-1}} (T+1) l_{r-1}^T \\
&\le cNn^{T-r+1} q^T p^{n-T} \left(\sum_{l=1}^{n} l^T q_0^l \right)^{r-1} (T+1)^{r-1} \\
&\le cNp^n n^T n^{1-r} \to 0
\end{aligned}
$$

as $N \to \infty$, because $r = 2, 3, \dots$.
In the above case, we applied that $\sum_{l=1}^n l^T q_0^l \le c \int_0^\infty x^T q_0^x dx < \infty$. So we obtained

$$\sum_{\alpha_r \in I_r(n,N)} b_{i_1, i_2, \dots, i_r} = \sum_{\alpha_r \in I_r(n,N)} \mathbb{P}(\tilde{\eta}_{i_1} = 1, \dots, \tilde{\eta}_{i_r} = 1) \to 0 \tag{1.15}$$

as $N \to \infty$. Hence the proof. $\square$

1.2 Number of at most T-contaminated runs of heads

Now we define the number of at most T-contaminated runs of heads having length n as follows. Let

$$\eta_i = \eta_i^T(n) = \begin{cases} 1, & \text{if there are at most } T \text{ } 0 \text{ values among} \\ & X_i, \dots, X_{i+n-1} \\ 0, & \text{otherwise} \end{cases} \tag{1.16}$$

Now we let

$$\xi = \xi^T(n, N) = \sum_{i=1}^{N-n+1} \eta_i^T(n). \tag{1.17}$$

Therefore ξ can be considered as the number of head runs being at most T-contaminated and having length n.

Now we want to prove that the distribution of ξ converges to a compound Poisson distribution in the limit.

Theorem 1.2.1. *(**Fazekas and Suja (2021)**) Let T be fixed. We let $N \to \infty$ and $n \to \infty$ so that condition (1.3) is satisfied. Then, for the generator functions we have*

$$\lim_{N\to\infty} \mathbb{E}\left(z^{\xi^T(n,N)}\right) = \exp\left[\lambda\left(\frac{qz}{1-pz} - 1\right)\right].$$

The above theorem is proved in **Földes (1975)** for the case of T-contaminated head run where $p = 1/2$ and $T > 0$ (see Theorem 3.2) and also proven in **Földes (1979)** for uncontaminated head run where $p = 1/2$ and $T = 0$ (see Theorem 2.A).

Remark 1.2.1. First, we recall the notion of the compound Poisson distribution. The compound Poisson distribution is a probability distribution that arises when counting the number of occurrences of a rare event in a fixed time interval, where the size of each occurrence is a random variable with a probability distribution.

More specifically, suppose we have a Poisson process with rate λ, which is a stochastic process that models the occurrence of rare events over time. For each occurrence, we assume that there is a random variable X_i that represents the size or magnitude of the event, and these random variables are assumed to be independent and identically distributed (i.i.d.). Then, the compound Poisson distribution is the distribution of the sum of these random variables over a fixed time interval.

In our case we need its particular version, that is the so called geometric Poisson distribution.
Let γ have Poisson distribution $\mathbb{P}(\gamma = k) = \lambda^k e^{-\lambda}/k!$, $k = 0, 1, 2, \ldots$. Let $\varrho_1, \varrho_2, \ldots$, be random variables independent of each other and of γ having q parameter geometric random distribution:
$\mathbb{P}(\varrho_i = l) = p^{l-1}q$, $l = 1, 2, \ldots$, $q \in (0, 1)$, $p = 1 - q$.
We let the distribution of ϱ to be the same as that of $\varrho_1 + \ldots + \varrho_k$ when $\gamma = k$. (Here, an empty sum is defined as 0, i.e $\varrho = 0$ when $\gamma = 0$). Then ϱ has generator function

$$\mathbb{E}(z^{\varrho}) = \exp\left[\lambda\left(\frac{qz}{1-pz} - 1\right)\right] \text{ for } \quad |zp| < 1.$$

Proof of Theorem 1.2.1. It is easy to see that for any run of length n containing at most T zeros, either there exist a preceding run of length n containing precisely T zeros or all preceding runs contain less than T zeros. To give a formal explanation of this fact, let

$$\eta_i' = 1 = {\eta'}_i^T(n) = \begin{cases} \tilde{\eta}_i^T(n).X_{i+n-1}, & \text{if} \quad i > 1 \\ \eta_i^T(n), & \text{if} \quad i = 1. \end{cases} \tag{1.18}$$

Therefore, $\eta_i' = 1$ either if $i = 1$ and among the first n tosses, there are at most T tails or, $i > 1$ and $X_{i-1} = 0$, $X_{i+n-1} = 1$ and between these locations there are precisely T zeros. We see that in the second case, shifting to the left by 1 of the interval $i, i+1, \dots, i+n-1$, we obtain an interval containing $T+1$ zeros.
Now, starting at an arbitrary interval $i, i+1, \dots, i+n-1$ containing at most T tails, shift to the left of this interval step by step until there are more than T tails in it. If $k+1$ denotes the number of steps until that situation is reached, then ${\eta'}_{i-k} = 1$. (Moreover, ${\eta'}_{i-k} = 1$ means always the end of the above shifting procedure). Therefore, we can find the longest sequence of overlapping intervals of type $i, i+1, \dots, i+n-1$ containing at most T tails. We shall refer to these sequences of intervals as chains of intervals. So to count the number of head runs being at most T-contaminated and having length n, we should find the number of these chains of intervals and their length. To be more precise, we should consider the following representation of,

$$\xi = \xi^T(n, N) = \sum_{i=1}^{N-n+1} \gamma_i^T(n) = \sum_{i=1}^{N-n+1} \gamma_i, \tag{1.19}$$

where

$$\gamma_i = \gamma_i^T(n) = {\eta'}_i \left[\min\left\{k > 0 : \text{either} \quad \eta_{i+k} = 0 \quad \text{or} \quad i+k+n-1 > N\right\}\right] \tag{1.20}$$

Let $\gamma = \gamma^T(n)$ denote the number of non-zero γ_i's (i.e. the number of non-zero $\acute{\eta}_i$ elements). We know that $\tilde{\xi}$ is the number of precisely T-contaminated head runs of length n such that the preceding element is 0. So $\gamma \neq \tilde{\xi}$ only in the following two cases. The first case is when $\eta_1 = 1$ and $\tilde{\eta}_1 = 0$. The second case is ${\eta'}_i \neq \tilde{\eta}_i$

for some $i > 1$. The probability of these events is not greater than

$$\mathbb{P}(\eta_1 = 1, \tilde{\eta}_1 = 0) + \sum_{i=2}^{N-n+1} \mathbb{P}(\eta'_i \neq \tilde{\eta}_i) \leq$$

$$\begin{aligned}
&\leq \mathbb{P}\,(\text{at the beginning of the tosses, there is a run containing tails less than T}) \\
&\quad + N\mathbb{P}\,(\text{there are precisely T tails among n tosses,} \quad \text{the last and the previous are tails}) \leq \\
&\quad \leq \sum_{i=0}^{T-1} \binom{n}{i} p^{n-i} q^i + Nq \binom{n-1}{T-1} p^{n-T} q^{T+1} q \leq \\
&\quad \leq cTn^T p^{n-T+1} + cNn^{T-1} p^{n-T} q^{T+1} \to 0
\end{aligned}$$

as $N, n \to \infty$, because of condition (1.3).

Therefore, $\mathbb{P}(\gamma = \tilde{\xi}) \to 1$ if $N, n \to \infty$. So, by Slutsky's lemma, the distribution of γ is also λ parameter Poisson.

We shall show that

$$\lim_{N,n\to\infty} \mathbb{P}(\gamma_i > k | \gamma_i > 0) = p^k, \quad k = 0, 1, 2, \ldots \tag{1.21}$$

i.e. the (conditional) limiting distribution of γ_i is the q parameter geometric distribution. To this end, we shall use the following elementary fact;
When $\mathbb{P}(B_n C_n) \neq 0$, then if $\lim_{n\to\infty} \mathbb{P}(B_n C_n)/\mathbb{P}(B_n) = 1$ implies that

$$\lim_{n\to\infty} \mathbb{P}(A_n | B_n) = \lim_{n\to\infty} \mathbb{P}(A_n | B_n C_n) \tag{1.22}$$

in the sense that if one side of the above equation exists, then the other side also exists and the two sides are equal. For any fixed $i \in \{1, 2, \ldots\}$ and $k_0 \geq 1$, let
$A_n = \{\gamma_i^T(n) > k_0\}$,
$B_n = \{\gamma_i^T(n) > 0\}$, $\quad C_n = \{\prod_{j=i}^{i+k_0-1} X_j = 1\}$.
First let $i > 1$. Then,
$B_n = \{X_{i-1} = 0, X_{i+n-1} = 1, \text{and there are precisely } T \text{ zeros among} \quad X_i, \ldots, X_{i+n-2}\}$.

Then,
$$\frac{\mathbb{P}(B_n C_n)}{B_n} = \frac{qp^{k_0} \binom{n-k_0-1}{T} q^T p^{n-k_0-1-T} p}{q \binom{n-1}{T} q^T p^{n-1-T} p} \to 1.$$

So, by (1.22),

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}(\gamma_i > k_0 | \gamma_i > 0) = \lim_{n\to\infty} \mathbb{P}(A_n | B_n) = \\
&= \lim_{n\to\infty} \mathbb{P}(A_n | B_n C_n) = \lim_{n\to\infty} \mathbb{P}(X_{i+n} = \cdots = X_{i+n+k_0-1} = 1) = p^{k_0}.
\end{aligned}$$

Now let $i = 1$. Then,
$B_n = \{\gamma_i^T(n) > 0\} = \{\text{there are at most } T \text{ zeros among } \quad X_1, \dots, X_n\}$.
Let
$B'_n = \{\text{there are precisely } T \text{ zeros among } \quad X_1, \dots, X_n\}$.

Then $\mathbb{P}(B'_n) = \binom{n}{T} q^T p^{n-T}$, $\quad \mathbb{P}(B_n) = \sum_{j=0}^{T} \binom{n}{j} q^j p^{n-j}$. We see that $B'_n \subseteq B_n$ and $\frac{\mathbb{P}(B_n B'_n)}{\mathbb{P}(B_n)} = \frac{\mathbb{P}(B'_n)}{\mathbb{P}(B_n)} \to 1$ as $n \to \infty$.
So, using (1.22) with $C_n = B'_n$ we see that $\lim_{n\to\infty} \mathbb{P}(A_n|B_n) = \lim_{n\to\infty} \mathbb{P}(A_n|B'_n)$. Then, with $C_n = \{\prod_{j=1}^{k_0} X_j = 1\}$, we have

$$\frac{\mathbb{P}(B'_n C_n)}{\mathbb{P}(B'_n)} = \frac{p^{k_0} \binom{n-k_0}{T} q^T p^{n-k_0-T}}{\binom{n}{T} q^T p^{n-T}} \to 1.$$

So, we can use (1.22) with B'_n instead of B_n. Then, we obtain

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}(A_n|B_n) &= \lim_{n\to\infty} \mathbb{P}(A_n|B'_n) = \lim_{n\to\infty} \mathbb{P}(A_n|B'_n C_n) = \\ &= \lim_{n\to\infty} \mathbb{P}(X_{n+1} = \dots = X_{n+k_0} = 1) = p^{k_0}. \end{aligned}$$

So we obtain (1.21).
Now let $\alpha_r = (i_1, \dots, i_r)$ be a vector of indices with $i_j \neq i_l$ for $i \neq j$.
We introduce notation
$C(\alpha_r) = \{\gamma = r, \quad \eta'_{i_1} = \eta'_{i_2} = \dots = \eta'_{i_r} = 1\}$.
The meaning of $C(\alpha_r)$ is that there are r above mentioned chains of intervals starting at positions $i_1, i_2, \dots, i_r$. Obviously $C(\alpha_r) \bigcap C(\acute{\alpha}_r) = \emptyset$ for $\alpha_r \neq \acute{\alpha}_r$, moreover $\{\gamma = r\} = \bigcup_{\alpha_r} C(\alpha_r)$. Therefore

$$\mathbb{P}\{\gamma = r\} = \sum_{\alpha_r} \mathbb{P}(C(\alpha_r)).$$

Now let $K = (k_1, k_2, \dots, k_r)$, $\quad |K| = k_1 + k_2 + \dots + k_r$ and let

$$C_1(\alpha_r, K) = C(\alpha_r) \bigcap \{\gamma_{i_1} > k_1, \dots, \gamma_{i_r} > k_r\}.$$

The meaning of $C_1(\alpha_r, K)$ is that the lengths of the above mentioned chains of intervals are greater than $k_1, \dots, k_r$. Using $C_1(\alpha_r, K)$, we can describe the asymptotic joint distribution of the positive γ_i's.
To finish the proof, we have to prove the following;
Given that there are r positive γ_i variables, then the asymptotic joint distribution of the positive γ_i variables is equal to the joint distribution of r independent

geometrically distributed random variables. That is, we have to prove

$$\frac{\sum_{\alpha_r} \mathbb{P}(C_1(\alpha_r, K))}{\mathbb{P}(\gamma = r)} \to p^{|K|}. \tag{1.23}$$

Let $k = \max_{1 \le i \le r} k_i$ and let the exceptional set $I_r(n+k, N)$ be defined as in the proof of theorem 1.1.1. That is $\alpha_r \in I_r(n+k, N)$ if and only if there exists $i_j, i_l \in \alpha_r$ such that $|i_j - i_l| \le n + k$. Now we show that

$$\sum_{\alpha_r \in I_r(n,N)} \mathbb{P}(C(\alpha_r)) \to 0 \tag{1.24}$$

We have

$$\mathbb{P}(C(\alpha_r)) \le \mathbb{P}(\{\eta'_{i_1} = \eta'_{i_2} = \cdots = \eta'_{i_r} = 1\}) \le \mathbb{P}(\{\tilde{\eta}_{i_1} = \tilde{\eta}_{i_2} = \cdots = \tilde{\eta}_{i_r} = 1\})$$

because of the inclusion relations among the above events. Now, using this relation and the fact $\tilde{\eta}_{i_1}, \tilde{\eta}_{i_2}, \cdots, \tilde{\eta}_{i_r}$ are independent for $\alpha_r \notin I_r(n, N)$, we obtain that

$$\sum_{\alpha_r \in I_r(n+k,N)} \mathbb{P}(C(\alpha_r)) \le \sum_{\alpha_r \in I_r(n,N)} \mathbb{P}(\{\tilde{\eta}_{i_1} = \tilde{\eta}_{i_2} = \cdots = \tilde{\eta}_{i_r} = 1\}) +$$
$$+ |I_r(n+k, N)| \left[q \binom{n}{T} q^T p^{n-T} \right]^r.$$

During the proof of Theorem 1.1.1, remember we obtained that the limit of the first term is 0 (see (1.15)). On the other hand, for the second term, by taking the limits, we have

$$\begin{aligned}
&\lim_{N\to\infty} |I_r(n+k, N)| \left[q \binom{n}{T} q^T p^{n-T} \right]^r \le \\
&\le \lim_{N\to\infty} c \left(\binom{N}{r-1} (r-1) 2(n+k) \right) \left[n^T q^{(T+1)} p^{(n-T)} \right]^r \le \\
&\le \lim_{N\to\infty} c \frac{n}{N} \left[N n^T p^n \right]^r = 0
\end{aligned}$$

by condition (1.3). So we obtain (1.24). Now we have

$$\sum_{\alpha_r} \mathbb{P}(C_1(\alpha_r, K)) = \sum_{\alpha_r \in I_r(n+k,N)} \mathbb{P}(C_1(\alpha_r, K)) + \sum_{\alpha_r \notin I_r(n+k,N)} \mathbb{P}(C_1(\alpha_r, K)). \tag{1.25}$$

By (1.24), the first term in (1.25) converges to 0. Now we consider the second term. By the definition of $C_1(\alpha_r, K)$, we have

$$\sum_{\alpha_r \notin I_r(n+k,N)} \mathbb{P}(C_1(\alpha_r, K)) =$$
$$= \sum_{\alpha_r \notin I_r(n+k,N)} \mathbb{P}(\gamma_{i_1} > k_1, \dots, \gamma_{i_r} > k_r | \gamma = r, \quad \eta'_{i_1} = \eta'_{i_2} = \dots = \eta'_{i_r} = 1) \times$$
$$\times \mathbb{P}(\gamma = r, \quad \eta'_{i_1} = \eta'_{i_2} = \dots = \eta'_{i_r} = 1).$$

By independence and (1.21),

$$\begin{aligned} &\sum_{\alpha_r \notin I_r(n+k,N)} \mathbb{P}(C_1(\alpha_r, K)) \approx \\ &\approx \sum_{\alpha_r \notin I_r(n+k,N)} \left(\prod_{i=1}^{r} p^{k_i} \right) \mathbb{P}(\gamma = r, \quad \eta'_{i_1} = \eta'_{i_2} = \dots = \eta'_{i_r} = 1) = \\ &= p^{|K|} \left(\sum_{\alpha_r} \mathbb{P}(C(\alpha_r)) - \sum_{\alpha_r \in I_r(n+k,N)} \mathbb{P}(C(\alpha_r)) \right) \approx p^{|K|} \mathbb{P}(\gamma = r). \end{aligned} \tag{1.26}$$

In the last step, we applied (1.24). As the limit distribution of γ is Poisson, that is the limit of $\mathbb{P}(\gamma = r)$ is non-zero, we obtain from (1.25), (1.24) and (1.26) that condition (1.3) is satisfied.

So we obtained for relation (1.19) that is for $\xi = \sum_{i=1}^{N-n+1} \gamma_i$, the following facts; the number γ of the non-zero terms γ_i is asymptotically Poisson with parameter λ. Moreover, the positive ones out of the variables $\gamma_{i_1}, \gamma_{i_2}, \dots$ are asymptotically geometric and they are asymptotically independent. Therefore ξ is asymptotically compound Poisson hence the proof. □

1.3 First hitting time of T-contaminated runs of heads

First hitting time τ is the number of tosses needed in a coin tossing experiment for a T-contaminated head run of length n to appear for the very first time i.e. its the first observation time when the number of tails among the last n outcomes is at most T.

Let

$$\tau = \tau^T(n) = \min\{N : \xi^T(n, N) > 0\}. \tag{1.27}$$

If $T = 0$, τ is the usual waiting time for a pure head run of length n. The distribution of this random variable has been widely studied in the literature using various techniques. For recurrence relation see **Philippou and Makri (1985)**, for non-recursive combinatorial expressions see **Muselli (1996)** while for Markov chain approach see **Koutras (1996)**.

We show that the appropriately normalized version of τ has exponential limiting distribution.

Theorem 1.3.1. *(**Fazekas and Suja (2021)**) Let T be fixed. Then, for any $0 < x < \infty$*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\tau^T(n)n^T}{T!} q^{T+1}p^{n-T} \le x\right) = 1 - e^{-x}.$$

The above theorem is proved in **Földes (1975)** for the case of T-contaminated head run where $p = 1/2$ and $T > 0$ (see Theorem 3.3) and also proven in **Földes (1979)** for uncontaminated head run where $p = 1/2$ and $T = 0$ (see Theorem 3.A).

Proof of Theorem 1.3.1.

$$\mathbb{P}\left(\frac{\tau^T(n)n^T}{T!} q^{T+1}p^{n-T} > x\right) = \mathbb{P}\left(\tau^T(n) > \frac{xT!}{n^T q^{T+1}p^{n-T}}\right) = \mathbb{P}\left(\xi^T(n, N(x)) = 0\right),$$

where $N(x) = \left[\frac{xT!}{n^T q^{T+1}p^{n-T}}\right]$. By Theorem 1.2.1, the asymptotic distribution of ξ is compound Poisson. It is obtained from a λ parameter Poisson and q parameter geometric distributions. Therefore, (using notation from the beginning of the proof of Theorem 1.2.1) the limiting distribution is

$$\mathbb{P}(\varrho = 0) = \sum_{k=0}^{\infty} \mathbb{P}(\varrho = 0|\gamma = k)\mathbb{P}(\gamma = k) =$$

$$= \mathbb{P}(0 = 0)\mathbb{P}(\gamma = 0) + \sum_{k=0}^{\infty} \mathbb{P}(\varrho_1 + \cdots + \varrho_k = 0)\mathbb{P}(\gamma = k) = 1e^{-\lambda} + 0.$$

Upon checking whether condition (1.3) is satisfied, we find the value of λ.

$$\frac{N(x)n^T q^{T+1} p^{n-T}}{T!} = \left[\frac{xT!}{n^T q^{T+1} p^{n-T}}\right] \frac{n^T q^{T+1} p^{n-T}}{T!} \to x$$

as $n \to \infty$. Therefore, the λ parameter is equal to x. Hence the proof. $\square$

1.4 Length of the longest T-contaminated runs of heads

Various approaches have been used in determining the length and distribution of the longest head runs. For the case of a fair coin and using asymptotic estimation see (**Erdős and Rényi (1970)**, **Erdős and Révész (1975)**, **Földes (1979)**). Let

$$\mu = \mu^T(N) = \max\{n : \xi^T(n, N) > 0\}. \tag{1.28}$$

Considering the result of tossing a coin N times, μ is the length of the longest run of heads containing at most T tails. The following theorem describes the accompanying distribution of $\mu^T(N)$. We offer a two parameter family of distributions to approximate the distribution of μ. We shall use the following notation. Let B be a fixed positive number, then for any positive x, we have that

$$x = kB + r,$$

where k is integer and r is the residual for which $0 \leq r < B$. Here k and r are uniquely determined. We define $[x]_B$ and $\{x\}_B$ as $[x]_B = kB$ and $\{x\}_B = r$.

Theorem 1.4.1. *(**Fazekas and Suja (2021)**) Let T be fixed. Let B be a fixed positive number and let S be a fixed number. Then, for any integer k we have*

$$\begin{aligned}&\mathbb{P}(\mu^T(N) - [\log N + T\log(\log N + S\log\log N)]_B < k) = \\ &= \exp\left(-q^{T+1} p^{(k-T-\{\log N + T\log(\log N + S\log\log N)\}_B)}/T!\right) + o(1).\end{aligned} \tag{1.29}$$

Here log *denotes logarithm to base* $1/p$.

For $B = S = 1$ and $p = 1/2$, the above theorem is proved in **Földes (1975)** for the case of T-contaminated head run where $p = 1/2$ and $T > 0$ (see Theorem 3.4) and also proven in **Földes (1979)** for uncontaminated head run where $p = 1/2$ and $T = 0$ (see Theorem 4.A). For arbitrary p with $B = S = 1$, **Móri (1993)** presented the above results without proof.

Proof of Theorem 1.4.1. For any integer k, let

$$f(N) = \mathbb{P}(\mu^T(N) - [\log N + T\log(\log N + S\log\log N)]_B < k).$$

Then

$$f(N) = \mathbb{P}(\mu^T(N) < n(k)) = \mathbb{P}(\xi^T(n(k), N) = 0),$$

where $n(k) = k + [\log N + T\log(\log N + S\log\log N)]_B$. For any fixed k, the sequence $n(k)$ converges to infinity as $N \to \infty$. Now let $\lambda_0 \in [0, B]$ be fixed and choose a sub-sequence

$$N_j \uparrow \infty \quad \text{such that} \quad \{\log N_j + T\log(\log N_j + S\log\log N_j)\}_B \to \lambda_0. \tag{1.30}$$

For this subsequence N_j and for

$$n_j(k) = k + [\log N_j + T\log(\log N_j + S\log\log N_j)]_B,$$

condition (1.3) is satisfied in the following form;

$$\frac{N_j q^{T+1} p^{n_j(k)-T} \left(n_j(k)\right)^T}{T!} \to \frac{q^{T+1} p^{k-T} p^{-\lambda_0}}{T!}. \tag{1.31}$$

Therefore, by Theorem 1.2.1 and using the argument of the proof of Theorem 1.3.1, we obtain

$$\begin{aligned}\lim_{j\to\infty} f(N_j) =& \lim_{j\to\infty} \mathbb{P}(\mu^T(N_j) - [\log N_j + T\log(\log N_j + S\log\log N_j)]_B < k)\\ =& \lim_{j\to\infty} \mathbb{P}(\xi^T(n_j(k), N_j) = 0)\\ =& \exp\left(-q^{T+1}p^{k-T}p^{-\lambda_0}/T!\right).\end{aligned} \tag{1.32}$$

If (1.30) is satisfied, then for

$$\begin{aligned} g(N) =& \exp\left(-q^{T+1}p^{(k-T-\{\log N+T\log(\log N+S\log\log N)\}_B)}/T!\right) \\ =& \lim_{j\to\infty} f(N_j) = \exp\left(-q^{T+1}p^{k-T}p^{-\lambda_0}/T!\right) \end{aligned} \tag{1.33}$$

is true. To obtain (1.29), we have to show that $f(N) - g(N) \to 0$. Suppose that it is not satisfied, i.e. there exist $\varepsilon > 0$ such that for certain subsequence N'_j we have $|f(N'_j) - g(N'_j)| > \varepsilon$ for any j.
As the sequence $\{\log N'_j + T\log(\log N'_j + S\log\log N'_j)\}_B$ has accumulation point $\lambda_0 \in [0, B]$, so there exists a further subsequence N_j of N'_j so that

$$\{\log N_j + T\log(\log N_j + S\log\log N_j)\}_B \to \lambda_0.$$

Now, by (1.32) and (1.33), $f(N_j) - g(N_j) \to 0$. It is a contradiction, so (1.29) is satisfied. □

Now we give a new proof for Theorem 1 of **Gordon et al. (1986)** in which the longest head run containing (precisely) T tails was studied. However, we have the following to make;

Remark 1.4.1. The limiting distribution of the length of the longest head run containing T tails is the same as the limiting distribution of the length of the longest head run containing at most T tails. To prove it, let A be the event that the length of the longest head run containing at most T tails is greater than n. Then, $A = B \cup C$ where B is the event that the length of the longest head run containing precisely T tails is greater than n and C is the event that the length of a head run containing less than T tails is greater than n and it is not possible to add some tails to it. But

$$\mathbb{P}(C) \le \sum_{i=0}^{T-1} \binom{N}{i} p^{N-i} q^i \le cp^N N^{T-1} \to 0$$

as $N \to \infty$.

In **Gordon et al. (1986)**, the original proof was based on extreme value theory, but here we give a new proof using the method of our Theorem 1.4.1. Let $[x]$ denote the usual integer part of x and {x} is the fractional part.

Proposition 1.4.1. *(Theorem 1 of* ***Gordon et al. (1986)*** *together with Theo-*

rems 3.2 and 3.3 in ***Binswanger and Embrechts (1994)****.)*

$$\mathbb{P}(\mu^T(N) - \mu_T(qN) \le t) = \mathbb{P}\left(\left[\frac{W}{ln(\frac{1}{p})} + \{\mu_T(qN)\}\right] - \{\mu_T(qN)\} \le t\right) + o(1)$$

for all t, where

$$\mu_T(qN) = \log(qN) + T\,\log\log(qN) + T\log(q/p) - \log(T!) \tag{1.34}$$

and W has an extreme value distribution $\mathbb{P}(W \le t) = \exp(-e^{-t})$.

Remark 1.4.2. We emphasize that the above proposition does not offer a limiting law for $\mu^T(N) - \mu_T(qN)$ but it gives a sequence of accompanying laws. The distances of the laws between the two sequences converge to 0 (as $n \to \infty$).

Proof of Proposition. Some algebraic calculation shows, that we have to prove that

$$\mathbb{P}(\mu^T(N) - \mu_T(qN) \le k) = \mathbb{P}\left(\left[\frac{W}{ln(\frac{1}{p})} + \{\mu_T(qN)\}\right] < k\right) + o(1)$$

for all integers k. Using the definition of the distribution of W, this relation is equivalent to

$$\mathbb{P}(\mu^T(N) - \mu_T(qN) \le k) = \exp\left(-p^{k-\{\mu_T(qN)\}}\right) + o(1).$$

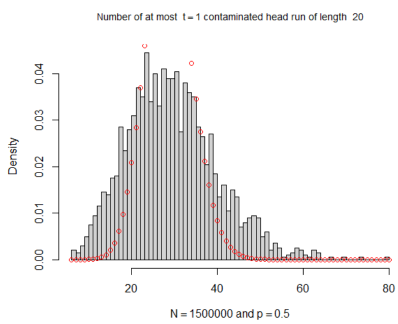
The remaining part of the proof is the same as that of Theorem 1.4.1. $\square$

1.5 Simulation Results

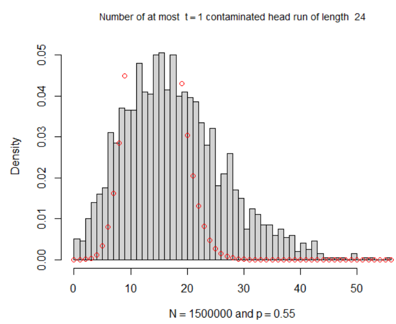
We chose sufficiently large lengths of the sequence N after which contaminated head runs of specified lengths n are investigated under varying probability values. We evaluate the contaminated runs for the case of $T = 1$ and $T = 2$. We performed our simulations in R package.

Example 1.5.1 (Number of at most T-contaminated runs of heads.)**.** *The figures below show the empirical distribution of the number of at most T-contaminated head run and its approximation suggested by theorem 1.2.1 and denoted by the red dots.*

For 2000 simulations, $N = 1.5 \times 10^6$, $p = \{0.5, 0.55\}$ *and* $T = 1$ *we try out different run lengths* n *to generate our results.*



(a)

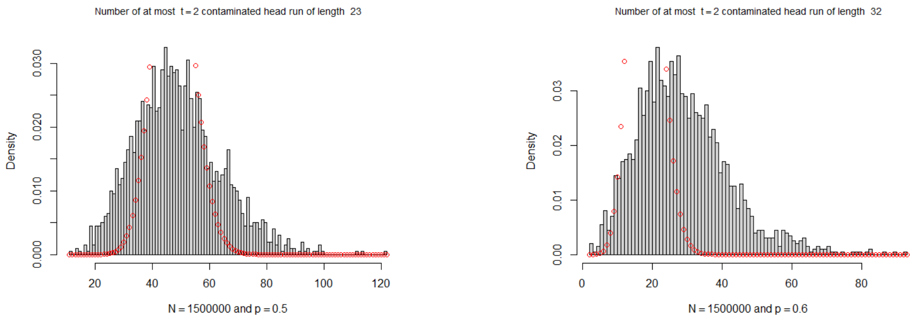


(b)

Figure 1.1: Distribution of the length of at most $T = 1$ contaminated head run

Analysis of the above figures reveal a reasonable fit indicating convergence to the suggested compound Poisson distribution.

Similarly, for 2000 simulations, $N = 1.5 \times 10^6$, $p = \{0.5, 0.6\}$ *and* $T = 2$ *we try out different run lengths* n *to generate our results.*



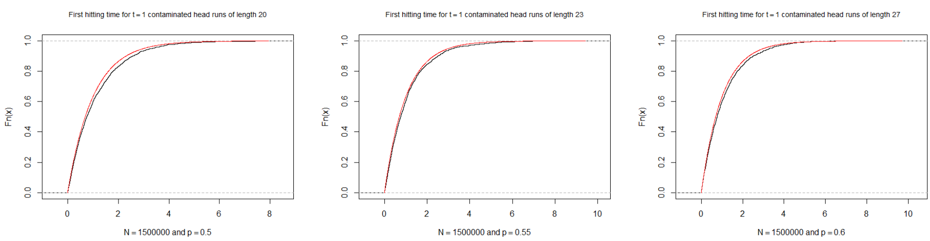
(a) (b)

Figure 1.2: Distribution of the length of at most $T = 2$ contaminated head run

It is noted that for higher values of p*, the empirical and theoretical fit is not so nice as skewness is observed.*

Example 1.5.2 (First hitting time for at most T-contaminated head runs of any specified length)**.** *The figures below show the empirical distribution of the first hitting time of the at most* T*-contaminated head run and its approximation suggested by theorem 1.3.1 and denoted by the red doted line.*

For 2000 simulations, $N = 1.5 \times 10^6$, $p = \{0.5, 0.55, 0.6\}$ *and* $T = 1$ *with various run lengths* n*, we obtain the results.*

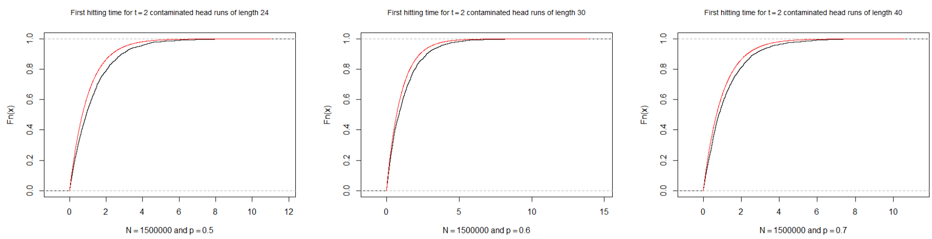


(a) (b) (c)

Figure 1.3: Distribution of first hitting times for $T = 1$ contaminated head runs

The figures reveal near perfect fit between the empirical and theoretical distributions for $T = 1$ *even with higher values of* p*.*

Similarly, for 2000 simulations, $N = 1.5 \times 10^6, \quad p = \{0.5, 0.6, 0.7\}$ *and* $T = 2$ *with various run lengths* n*, we obtain our results.*



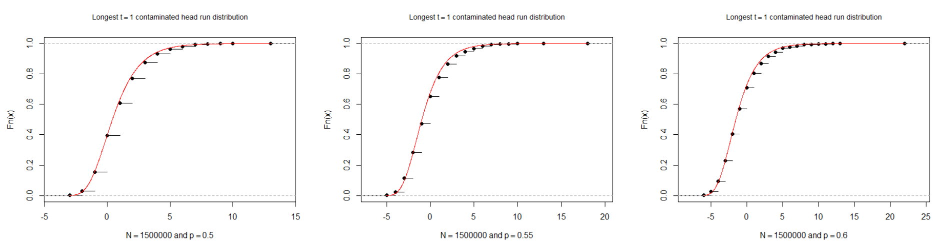
(a) (b) (c)

Figure 1.4: Distribution of first hitting times for $T = 2$ contaminated head runs

The figures reveal nice fit between the empirical and theoretical distributions for $T = 2$ *even with higher values of* p*. However, some slight left skewness is observed which generally tend to diminish as* p *increases.*

Example 1.5.3 (Length of longest at most T-contaminated head runs)**.** *The figures below show the empirical distribution of the length of the longest at most* T*-contaminated head run and its approximation suggested by theorem 1.4.1 and denoted by the red doted line.*

This variable is independent of the length n *and to investigate its properties, we consider* $N = 1.5 \times 10^6, \quad p = \{0.5, 0.55, 0.6\}$ *and* $T = 1$.

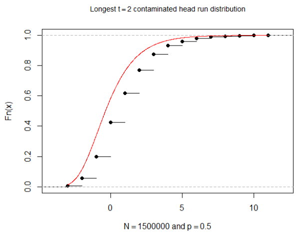


(a) (b) (c)

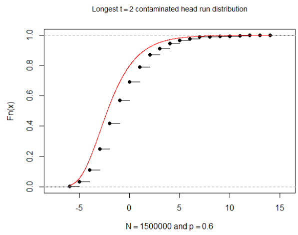
Figure 1.5: Distribution of the length of longest $T = 1$ contaminated head run

The figures reveal perfect fit between the empirical and theoretical distributions for $T = 1$ *even with higher values of* p*. However, some slight left skewness is observed which generally tend to diminish as* p *increases.*

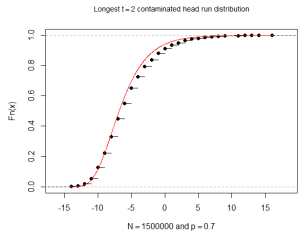
For the case of $p = \{0.5, 0.6, 0.7\}$ *and* $T = 2$ *contaminated head runs we compared their distributions and the following figure graphically captures the fit.*



(a)



(b)



(c)

Figure 1.6: Distribution of the length of longest $T = 2$ contaminated head run

The figures reveal skewness in the fit for lower values of p between the empirical and theoretical distributions. However, the skewness generally tend to diminish as p increases giving a perfect fit.

Chapter 2

Convergence rate for the longest T-contaminated head runs

The term ”convergence rate” pertains to the speed at which a stochastic process approaches its theoretical or limiting behavior as the number of trials or sequence length grows. The metric offers a numerical assessment of the speed at which the sequence attains a stable pattern, which is a crucial procedure for generating well-informed decisions and forecasts. The application of this technique extends to the prediction of certain patterns or events in diverse domains, encompassing financial markets, natural phenomena, and sports statistics.

In this chapter, we will examine the length of consecutive occurrences of heads that are interrupted by a few instances of tails in the standard coin flipping experiment. We shall refer to this subsequence as a T-contaminated run of heads if it contains T tails and all other values are heads.

The most renowned results pertain to the length of pure head runs. The investigation of the fair coin scenario was conducted in the seminal study authored by **Erdős and Rényi** (**1970**).

Subsequently, a number of scholarly articles were published addressing this subject. One notable study conducted by **Novak** (**1991**) examined the precision of the approximation to the distribution of the length of the longest head run in a Markov chain. (For further information, please refer to Novak's works **Novak** (**1991**), **Novak** (**2017**), as well as the additional references provided.) In addition to the review study authored by **Binswanger and Embrechts** (**1994**), the aforementioned source is included.

In their publication, **Gordon et al.** (**1986**), utilized extreme value theory to derive the asymptotic properties of the expected value and variance of the length of the longest T-contaminated head run. **Fazekas and Suja** (**2021**) demonstrated that the associated distributions of **Gordon et al.** (**1986**) may be derived using the approach proposed by **Földes** (**1979**).

So after performing some algebraic manipulations on Theorem 1 in the work of **Gordon et al.** (**1986**) (also referred to as Proposition 1.4.1 in the study conducted by **Fazekas and Suja** (**2021**)), we present the following statement;

Proposition 2.0.1. *Let $\mu_T(N)$ denote the length of the longest T-contaminated run of heads during the coin tossing experiment of length N. Let its approximation be*

$$m_0(N) = \log(qN) + T\log(\log(qN)) + T\log(q/p) - \log(T!), \tag{2.1}$$

where log *denotes the logarithm to base 1/p. Let $[m_0(N)]$ denote the integer part of $m_0(N)$ and $\{m_0(N)\}$ denote the fractional part of $m_0(N)$. Then for any positive integer k,*

$$\mathbb{P}(\mu_T(N) - [m_0(N)] < k) = \exp\left(-p^{k-\{m_0(N)\}}\right) + o(1).$$

Nevertheless, empirical investigations demonstrate that the approximation presented in Proposition 2.0.1 is quite weak. Hence, the objective of this chapter is to enhance the aforementioned outcome for the significant scenarios of $T = 1$ and $T = 2$.

The primary outcome of our study is Theorem 2.1.2. Similar to Proposition 2.0.1, our theorem also provides an accompanying distribution sequence for the distribution of the centralized version of $\mu_T(N)$.

In our new theorem, we demonstrate that the rate of approximation is bounded by $O\left(1/(\log(N))^2\right)$. It will be demonstrated that for $T = 1$ and $T = 2$, the rate of approximation in Proposition 2.0.1 can be expressed as $O\left(\log(\log(N))/\log(N)\right)$.

Additionally, we have derived a solution for the initial occurrence of the T-contaminated sequence of heads with a length of m, as stated in Theorem 2.1.1. In Section 2.3, we provide simulation data that reaffirms the claim that our new approximation outperforms previous ones.

We emphasize, that during the proofs we use only elementary methods of probability theory. In essence, we apply a powerful lemma of **Csáki et al.** (**1987**), but the proof of that lemma is also a clever use of elementary mathematics. The lemma is referred to as Lemma A.0.2 in our citation.

To properly apply this lemma, it is necessary to perform a series of elementary calculations that are both extensive and time-consuming. These calculations may be found in Lemma 2.2.2 and Lemma 2.2.3.

2.1 The first hitting time and the longest run

In the subsequent chapter, the aforementioned concepts and notational conventions will be adopted without being restated again.

Let us consider the well known coin tossing experiment. Let p denote the probability of obtaining heads, and let $q = 1 - p$ represent the probability of obtaining tails. Let p be a fixed value such that $0 < p < 1$.

We perform N independent coin tosses. In this context, we shall write 1 when the result is head and 0 when the result is tail. Let us consider a set of independent and identically distributed random variables, denoted as $X_1, X_2, \cdots, X_N$ with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = q$, $i = 1, 2, \ldots, N$.

Let $T \geq 0$ be a fixed non - negative integer. We shall study the T-interrupted runs of heads. It means that there are T zeros in an m length sequence of ones and zeros. So let m be a positive integer then, $A_n = A_{n,m}$ denotes the occurrence of the event at the n^{th} step, that is, there are precisely T zeros in the block of sequence $X_n, X_{n+1}, \ldots, X_{n+m-1}$. Here, we clarify that the condition $X_{n-1} = 0$ is not assumed. Therefore, $\mathbb{P}(\bar{A}_1 \bar{A}_2 \cdots \bar{A}_N)$ is the probability that no event $A_1 = A_{1,m}$ occurred in any of the first N blocks of length m i.e. the waiting time for the T-contaminated run of heads of length m described by A_1 is longer than N.

We say that the positive integer valued random variable τ_m is the first hitting time of the T-contaminated run of heads having length m if $\tau_m = N$, then $X_1, \cdots, X_{N-1}$ does not contain a T-contaminated run of heads of length m, but $X_1, \cdots, X_N$ contains it.

We shall find the asymptotic distribution of τ_m as $m \to \infty$ for $T = 1$ and for $T = 2$. For further development in this direction of investigation, (see Theorem 2 and its applications in **Solov'ev (1966)**, **Komlós and Tusnády (1975)**).

Theorem 2.1.1. *(**Fazekas, Fazekas, and Suja (2024)**) Let $T = 1$ or $T = 2$, $0 < p < 1$. Let τ_m be the first hitting time for the T-contaminated run of heads having length m. Then, for $x > 0$,*

$$\mathbb{P}(\tau_m \alpha P(A_1) > x) \sim e^{-x}$$

as $m \to \infty$.

Here if $T=1$, then $\alpha = q + \frac{2p^{m-1}-1}{m}$ and $P(A_1) = mp^{m-1}q$. While if $T=2$, then $\alpha = q - \frac{2}{m}$ and $P(A_1) = \binom{m}{2}p^{m-2}q^2$.

Remark 2.1.1. One can show that Theorem 2.1 is valid for $T=2$ with $\alpha = q - \frac{2}{m} + \frac{2(m-2)}{m}p^{m-2} - \frac{2(m-4)}{m}p^{m-1}$, too (see Remark 2.2.2).

Now, we turn to the length of the longest contaminated run of heads.

Theorem 2.1.2. ***(Fazekas, Fazekas, and Suja (2024))*** *Let $T=1$ or $T=2$, and let $0<p<1$ be fixed. Let $\mu(N)$ be the length of the longest T-contaminated run of heads during N times of coin tossing. Let*

$$
\begin{aligned}
m(N) =& \log(qN) + T\log(\log(qN)) + \\
&+ T^2\frac{\log(\log(qN))}{c\log(qN)} - \frac{T}{cq_0\log(qN)} - \frac{T^3}{2c}\left(\frac{\log(\log(qN))}{\log(qN)}\right)^2 + \\
&+ T^2\frac{\log(\log(qN))}{cq_0(\log(qN))^2} + T^3\frac{\log(\log(qN))}{(c\log(qN))^2} + \\
&+ \left(T\log(\frac{q}{p}) - \log(T!)\right)\left(1 + \frac{T}{c\log(qN)} - T^2\frac{\log(\log(qN))}{c(\log(qN))^2}\right),
\end{aligned}
\tag{2.1}
$$

where $\log$ denotes the logarithm to base $1/p$ and $c = ln(1/p)$, where $\ln$ denotes the natural logarithm to base e and $q_0 = \frac{2q}{2+Tq-q}$.
Let $[m(N)]$ denotes the integer part of $m(N)$ while $\{m(N)\}$ denotes the fractional part of $m(N)$, i.e. $\{m(N)\} = m(N) - [m(N)]$. Then for any integer k,

$$
\mathbb{P}(\mu(N) - [m(N)] < k) = e^{-p^{(k-\{m(N)\})\left(1-\frac{T}{c\log(qN)}+T^2\frac{\log(\log(qN))}{c(\log(qN))^2}\right)}}\left(1 + O\left(\frac{1}{(\log N)^2}\right)\right)
\tag{2.2}
$$

where $f(N) = O(h(N))$ means that $f(N)/h(N)$ is bounded as $N \to \infty$.

Remark 2.1.2. **(Fazekas, Fazekas, and Suja (2024))** Using our method for $T=1$ and $T=2$ and for $m_0(N)$ from (2.1), we obtain that the rate of convergence in Proposition 2.0.1 is $O(\log(\log(N))/\log(N))$, that is

$$
\mathbb{P}(\mu(N) - [m_0(N)] < k) = \exp\left(-p^{k-\{m_0(N)\}}\right)(1 + O(\log(\log(N)/\log(N)))).
$$

To obtain this, we apply the proof of Theorem 2.1.2 for the approximation $m_0(N)$ instead of $m(N)$.

So our method gives better convergence rate in our Theorem 2.1.2 than for Theorem 1 of **Gordon et al. (1986)** in the cases of $T=1$ and $T=2$.

2.2 Preliminary Lemmas

First we present some preliminary proofs to some Lemma in **Csáki et al. (1987)** which will play a fundamental role in the proofs of our theorems.

For purposes of clarity and understanding, we present this lemma.

Lemma 2.2.1. *(main lemma, stationary case, finite form of **Csáki et al. (1987)**.) Let m be fixed. Assume that A_n is stationary. Assume that there is a fixed number p, $0 < p \leq 1$, such that the following three conditions hold for some fixed k with $2 \leq k \leq m$, and fixed ε with $0 < \varepsilon < \min\{p/10, 1/42\}$*

(SI)

$$|\mathbb{P}(\bar{A}_2 \cdots \bar{A}_k | A_1) - p| < \varepsilon,$$

(SII)

$$\sum_{k+1 \leq i \leq 2m} \mathbb{P}(A_i | A_1) < \varepsilon,$$

(SIII)

$$P(A_1) < \varepsilon/m.$$

Then, for all $N > 1$,

$$\left| \frac{\mathbb{P}(\bar{A}_2 \cdots \bar{A}_N | A_1)}{\mathbb{P}(\bar{A}_2 \cdots \bar{A}_N)} - p \right| < 7\varepsilon$$

and

$$e^{-(p+10\varepsilon)NP(A_1) - 2mP(A_1)} < \mathbb{P}(\bar{A}_1 \cdots \bar{A}_N) < e^{-(p-10\varepsilon)NP(A_1) + 2mP(A_1)}.$$

We check the fulfilment of conditions (SI) - (SIII) for $k = m$. By verifying these conditions and with appropriate choices of ε, we can be able to determine the limiting distribution of the waiting time $\tau_m = \{\text{first } n;\text{ such that } A_n \text{ occurs}\}$. Next, we remark that conditions (SII) and (SIII) of the lemma are easily verified and happens to be true for any T if m is large enough. However, some detailed combinatorics and algebraic manipulations will be required to prove condition SI.

Remark 2.2.1. (**Fazekas, Fazekas, and Suja (2024)**) Consider condition (SIII). We show that (SIII) is true for any T if m is large enough. We have

$$P(A_1) = \binom{m}{T} p^{m-T} q^T \leq \frac{m^T}{T!} p^{m-T} q^T < \frac{\varepsilon}{m},$$

if

$$m^{T+1}p^m < \varepsilon \left(\frac{p}{q}\right)^T T!, \tag{2.1}$$

and the last inequality is satisfied for any positive ε if m is large enough. So we always can assume (SIII) of Lemma A.0.2 is true for any T.

Remark 2.2.2. (Fazekas, Fazekas, and Suja (2024)) Consider condition (SII).

$$\mathbb{P}(A_i|A_1) = P(A_i) = \binom{m}{T} p^{m-T} q^T \leq \frac{m^T}{T!} p^{m-T} q^T$$

if $i > m$ because of independence. So

$$\sum_{i=m+1}^{2m} \mathbb{P}(A_i|A_1) = mP(A_1) = m\binom{m}{T} p^{m-T} q^T \leq m\frac{m^T}{T!} p^{m-T} q^T < \varepsilon,$$

therefore we obtain again condition (2.1). So condition (SII) is also true if m is large enough.

To check condition (SI), we shall separately evaluate the joint probabilities $P(A_1\bar{A}_2\cdots\bar{A}_m)$ taking into account different values of T. First, lets fix $T = 1$.

Lemma 2.2.2. *(Fazekas, Fazekas, and Suja (2024)) Condition (SI) of Lemma A.0.2 is satisfied for $T = 1$ and $k = m$ in the following form*

$$|\mathbb{P}(\bar{A}_2\cdots\bar{A}_m|A_1) - \alpha| < \varepsilon, \tag{2.2}$$

with $\alpha = q + \frac{2p^{m-1}-1}{m}$.

Proof of Lemma 2.2.2. Fix $T = 1$ and $k = m$. To calculate the probability of the event $A_1\bar{A}_2\cdots\bar{A}_m$, we divide it into parts.

Consider those $0-1$ sequences $X_1, X_2, \ldots, X_{2m-1}$ which belong to $A_1\bar{A}_2\cdots\bar{A}_m$. If the first member of the sequence is 0, then the members on the places $m+1, \ldots, 2m-1$ should be ones. So this part of $A_1\bar{A}_2\cdots\bar{A}_k$ has probability $qp^{m-1}p^{m-1}$.

If the first member is 1, then X_{m+1} should be zero. If we fix that the only zero in $X_1, X_2, \ldots, X_m$ is at the l^{th} place with $2 \leq l \leq m-1$, then besides $X_{m+1} = 0$ there should be at least one zero among $X_{m+2}, \ldots, X_{m+l}$. Its probability is $qp^{m-1}q(1-p^{l-1})$. Finally, if the only zero in $X_1, X_2, \ldots, X_m$ is at the m^{th} place, then we need only $X_{m+1} = 0$. Its probability is $qp^{m-1}q$. Therefore

$$\mathbb{P}(A_1\bar{A}_2\cdots\bar{A}_m) = qp^{m-1}\left(p^{m-1} + q\sum_{i=1}^{m-2}(1-p^i) + q\right)$$

and

$$\mathbb{P}(\bar{A}_2 \cdots \bar{A}_m | A_1) = q + \frac{2p^{m-1} - 1}{m}.$$

Hence the proof □

Lemma 2.2.3. *(Fazekas, Fazekas, and Suja (2024)) Condition (SI) of Lemma A.0.2 is satisfied for $T = 2$ and $k = m$ in the following form*

$$|\mathbb{P}(\bar{A}_2 \bar{A}_3 \cdots \bar{A}_m | A_1) - \alpha| < \varepsilon, \tag{2.3}$$

with $\alpha = q - \frac{2}{m} + O(p^m)$ as as $m \to \infty$.

Remark 2.2.3. **(Fazekas, Fazekas, and Suja (2024))** For the case $T = 2$, we shall use the following two known formulae in simplifying our expressions;

$$\sum_{i=a}^{b} i x^{i-1} = \frac{b x^{b+1} - (b+1) x^b - (a-1) x^a + a x^{a-1}}{(x-1)^2}$$

and

$$\begin{aligned}\sum_{i=a}^{b} i(i-1) x^{i-2} = &\frac{b(b+1)x^b - a(a-1)x^{a-1} - b(b+1)x^{b-1} + a(a-1)x^{a-2}}{(x-1)^2} \\ &- \frac{2\left[b x^{b+1} - (a-1)x^a - (b+1)x^b + a x^{a-1}\right]}{(x-1)^3}.\end{aligned}$$

Proof of Lemma 2.2.3. Fix $T = 2$ and let $q = 1 - p$. We write 1 for heads and 0 for tails

$$\mathbb{P}(\bar{A}_2 \bar{A}_3 \cdots \bar{A}_m | A_1) = \frac{\mathbb{P}(A_1 \bar{A}_2 \cdots \bar{A}_m)}{P(A_1)}.$$

Here $P(A_1) = \binom{m}{2} p^{m-2} q^2$.

To calculate $\mathbb{P}(A_1 \bar{A}_2 \cdots \bar{A}_m)$, we divide the event $A_1 \bar{A}_2 \cdots \bar{A}_m$ into parts.

I. If the first element is 0, then the $(m+1)^{th}$ should be 1.

$$\underbrace{0, 1, \cdots, 1, \overset{(k)}{0}, 1, \cdots}_{\text{m}} \underbrace{1, 1, \cdots, 1}_{\text{k-1}}, \cdots$$

So the probability of this part is

$$\begin{aligned}\mathbb{P}(A_0) = &\sum_{k=2}^{m-1} q^2 p^{m-2} \left(p^{m-1} + p^{m-2} q (m-k)\right) + q^2 p^{m-2} . p^{m-1} \\ = &(m-1) q^2 p^{2m-3} + \frac{(m-1)(m-2)}{2} q^3 p^{2m-4}.\end{aligned}$$

This term is 'small', i.e. it is of order $O(p^m).P(A_1)$.

II. Now let us turn to the case when the first element is 1. Then the $(m+1)^{th}$ element should be 0. Let the k^{th} and the l^{th} elements be zeros, $1 < k < l \leq m$

$$\underbrace{1,\cdots,1,\overset{(k)}{0},1,\cdots,1,\overset{(l)}{0},\cdots}_{\text{m}},\underbrace{\underbrace{\underbrace{0,\cdots,0,\cdots}_{\text{k}},\cdots}_{\text{l}},\cdots}_{\text{m-1}}$$

Then on places $m+2,\cdots,m+k$, there should be at least one 0 and on places $m+2,\cdots,m+l$, there should be at least two zeros.

11/1. If $k=2$, then on the places $m+1, m+2$ should stay zeros. However, when $k \neq m$, there should be at least one 0 at places $m+3,\cdots,m+l$.

$$\underbrace{1,\overset{(2)}{0},1,\cdots,1,\overset{(l)}{0},\cdots}_{\text{m}}\underbrace{\underbrace{0,0,\cdots}_{\text{l}}\cdots}_{\text{m}}$$

So the probability of this part is

$$\mathbb{P}(A) = \sum_{l=3}^{m-1} p^{m-2}q^2.q^2(1-p^{l-2}) + p^{m-2}q^2q^2.$$

II/2. Now, let us consider the case $k > 2$. We shall study separately the case $l < m$ and the case of $l = m$. The first case is divided into two parts; Say B and C.

11/211. Let $k > 2$, $l < m$ and on the places $m+2,\cdots,m+k$, there are at least two 0's. $\underbrace{1,\cdots 1,\overset{(k)}{0},\cdots,\overset{(l)}{0},\cdots,1}_{\text{m}}\underbrace{0,\cdots,0,\cdots,0,\cdots}_{\text{k}}\cdots.$

The probability of this part is:

$$\mathbb{P}(B) = \sum_{k=3}^{m-2} p^{m-2}q^2(m-k-1).q(1-p^{k-1}-(k-1).p^{k-2}q).$$

11/212. Let $k > 2$, $l < m$ and on the places $m+2,\cdots,m+k$, there is precisely one 0. Moreover, on the places $m+k+1,\cdots,m+l$ there is at least one 0.

$$\underbrace{1,\cdots,1,\overset{(k)}{0},\cdots,\cdots,\overset{(l)}{0},\cdots,1}_{\text{m}}\underbrace{\underbrace{0,\cdots,0,\cdots}_{\text{k}},\cdots,\cdots,0,\cdots}_{\text{l}}\cdots$$

The probability of this part is:

$$\mathbb{P}(C)=\sum_{k=3}^{m-2}p^{m-2}q^2\sum_{l=k+1}^{m-1}q(k-1)qp^{k-2}(1-p^{l-k})$$

.

II/22. Let $k>2$ and $l=m$. Then on the places $m+2,\cdots,m+k$, there should be at least one 0.

$$\underbrace{1,\cdots\cdots,\overset{(k)}{0},\cdots\cdots,\overset{(l)}{0}}_{\text{m}}\underbrace{0,\cdots\cdots,0}_{\text{k}}$$

The probability of this part is:

$$\mathbb{P}(D)=\sum_{k=3}^{m-1}q^2p^{m-2}q(1-p^{k-1}).$$

Now, we reshape the above expression of $\mathbb{P}(C)$.

$$\begin{aligned}\mathbb{P}(C)=&\sum_{k=3}^{m-2}p^{m-2}q^4[(m-k-1)(k-1)p^{k-2}-(k-1)\underbrace{\sum_{l=k+1}^{m-1}p^{l-2}}_{\frac{p^{k-1}-p^{m-2}}{q}}]\\
&=p^{m-2}q^4\left[\sum_{k=3}^{m-2}(m-3)(k-1)p^{k-2}-p\sum_{k=3}^{m-2}(k-2)(k-1)p^{k-3}\right.\\
&\left.-\frac{p}{q}\sum_{k=3}^{m-2}(k-1)p^{k-2}+\sum_{k=3}^{m-2}(k-1)\frac{p^{m-2}}{q}\right]\\
&=p^{m-2}q^4\left[C_1+C_2+C_3-C_4\right].\end{aligned}$$

Here

$$\begin{aligned}C_1=&(m-3)\sum_{i=2}^{m-3}ip^{i-1}\\
=&(m-3)\frac{(m-3)p^{m-2}-(m-2)p^{m-3}-p^2+2p}{q},\end{aligned}$$

$$
\begin{aligned}
C_2 = & -p\sum_{j=2}^{m-3}(j-1)jp^{j-2} \\
= & -p\Bigg[\frac{(m-3)(m-2)p^{m-3}-2p-(m-3)(m-2)p^{m-4}+2}{q^2} \\
& -2\frac{(m-3)p^{m-2}-p^2-(m-2)p^{m-3}+2p}{-q^3}\Bigg],
\end{aligned}
$$

$$
\begin{aligned}
C_3 = & -\frac{p}{q}\sum_{k=2}^{m-3}kp^{k-1} \\
= & -\frac{p}{q}\frac{(m-3)p^{m-2}-(m-2)p^{m-3}-p^2+2p}{q^2}
\end{aligned}
$$

and

$$C_4 = \frac{p^{m-2}}{q}\frac{(m-1)(m-4)}{2}.$$

From these and omitting the 'small' terms (i.e. $O(p^m)$) , we obtain

$$\frac{\mathbb{P}(C)}{P(A_1)} = \frac{p^{m-2}q^4\left(C_1+C_2+C_3+C_4\right)}{\binom{m}{2}p^{m-2}q^2} \sim \frac{1}{\binom{m}{2}}\left[p(1+q)(m-3)-\frac{p}{q}(3-q^2)\right].$$

Then;

$$
\begin{aligned}
\mathbb{P}(A) = & (m-2)p^{m-2}q^4 - p^{m-2}q^4\left(p+p^2+\cdots+p^{m-3}\right) \\
= & (m-2)p^{m-2}q^4 - p^{m-2}q^4p\frac{p^{m-3}-1}{p-1}.
\end{aligned}
$$

Therefore;

$$\frac{\mathbb{P}(A)}{P(A_1)} \sim \frac{(m-2)q^2-pq}{\binom{m}{2}}$$

and

$$\frac{\mathbb{P}(D)}{P(A_1)} \sim \frac{(m-3)q-p^2}{\binom{m}{2}}.$$

Now we turn to B.

$$\begin{aligned}\mathbb{P}(B) = & \sum_{k=3}^{m-2} p^{m-2}q^3 m\left[1-p^{k-1}-(k-1)p^{k-2}q\right] \\ & - p^{m-2}q^3\sum_{k=3}^{m-2}(k+1)\left[1-p^{k-1}-(k-1)p^{k-2}q\right] \\ = & V_1+V_2+V_3+V_4+V_5+V_6.\end{aligned}$$

Where

$$V_1 = mp^{m-2}q^3(m-4),$$

$$V_2 = -mp^{m-2}q^3\left(p^2+p^3+\cdots+p^{m-3}\right) = mp^{m-2}q^3p^2\left(\frac{p^{m-4}}{q}-\frac{1}{q}\right),$$

$$V_3 = -mp^{m-2}q^4\sum_{k=2}^{m-3} kp^{k-1} = -mp^{m-2}q^4\left[\frac{(m-3)p^{m-2}-(m-2)p^{m-3}-p^2+2p}{q^2}\right],$$

$$V_4 = -p^{m-2}q^3\sum_{k=3}^{m-2}(k+1) = -p^{m-2}q^3\,\frac{(m+3)(m-4)}{2},$$

$$V_5 = p^{m-2}q^3\frac{1}{p}\sum_{k=4}^{m-1} kp^{k-1} = p^{m-3}q^3\,\frac{(m-1)p^m-mp^{m-1}-3p^4+4p^3}{q^2},$$

$$\begin{aligned}V_6 = & p^{m-2}q^4\sum_{k=3}^{m-2}k(k-1)p^{k-2}+p^{m-2}q^4\sum_{k=2}^{m-3}kp^{k-1} \\ = & p^{m-2}q^4\left[\frac{(m-2)(m-1)p^{m-2}-3.2.p^2-(m-2)(m-1)p^{m-3}+3.2.p}{q^2}+\right. \\ & -\ \frac{\left[2(m-2)p^{m-1}-2p^3-(m-1)p^{m-2}+3p^2\right]}{q^3}+ \\ & \left.+\frac{(m-3)p^{m-2}-(m-2)p^{m-3}-p^2+2p}{q^2}\right].\end{aligned}$$

From here

$$\frac{\mathbb{P}(B)}{P(A_1)} \sim \frac{1}{\binom{m}{2}}\left[m(m-4)q - mp^2 - m(-p^2+2p) - q\frac{(m+3)(m-4)}{2} + \right.$$
$$\left. + \frac{1}{qp}\left(4p^3 - 3p^4\right) - 3.2.p^2 + 3.2.p - \frac{4p^3}{q} + 6\frac{p^2}{q} - p^2 + 2p\right]$$
$$= \frac{1}{\binom{m}{2}}\left[q\frac{(m-4)(m-3)}{2} - 2mp + 8p - 7p^2 - \frac{7p^3}{q} + 10\frac{p^2}{q}\right].$$

Therefore

$$\mathbb{P}(\bar{A}_2\bar{A}_3\cdots\bar{A}_m|A_1) \sim \frac{\mathbb{P}(A)+\mathbb{P}(B)+\mathbb{P}(C)+\mathbb{P}(D)}{\mathbb{P}(A)}$$
$$= \frac{1}{\binom{m}{2}}\left[\left((m-2)q^2 - pq\right) + \right.$$
$$+ \left(q\frac{(m-4)(m-3)}{2} - 2mp + 8p - 7p^2 - \right.$$
$$\left. - \frac{7p^3}{q} + 10\frac{p^2}{q}\right) + \left((m-3)p(1+q) - \frac{p}{q}(3-q^2)\right) +$$
$$\left. + \left((m-3)q - p^2\right)\right]$$
$$= q - \frac{2}{m}.$$

Hence the proof □

Remark 2.2.4. (**Fazekas, Fazekas, and Suja** (**2024**)) A more careful calculation shows that Lemma 2.2.3 is valid for $T = 2$ with $\alpha = q - \frac{2}{m} + \frac{2(m-2)}{m}p^{m-2} - \frac{2(m-4)}{m}p^{m-1}$, too.

Proof of Theorem 2.1.1. We can now apply Lemma A.0.2 because its conditions are satisfied due to Remark 2.2.1, Remark 2.2.2, and Lemma 2.2.2, Lemma 2.2.3.

$$\mathbb{P}(\tau_m \alpha P(A_1) > x) = \mathbb{P}(\tau_m > N),$$

where N is the integer part of $\frac{x}{\alpha P(A_1)}$. In Lemma A.0.2, we can choose ε such that, $10\varepsilon = \varepsilon_0/m$, where ε_0 is a fixed positive number. Let $N_1 = N - m + 1$. So, by (A.17),

$$\begin{aligned}\mathbb{P}(\tau_m > N) =&\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{N_1}) \sim e^{-(\alpha \pm 10\varepsilon)N_1 P(A_1) \pm 2mP(A_1)}\\ &\sim e^{-(\alpha \pm \frac{\varepsilon_0}{m})\left(\frac{x}{\alpha P(A_1)} - m + 1\right)P(A_1)} e^{\pm 2mP(A_1)}\\ &\sim e^{-(\alpha \pm \frac{\varepsilon_0}{m})\frac{x}{\alpha}} e^{-(\alpha \pm \frac{\varepsilon_0}{m})(-m+1)P(A_1)} e^{\pm 2mP(A_1)}\\ &\sim e^{-x}\end{aligned} \tag{2.4}$$

as $m \to \infty$. Hence the proof □

Proof of Theorem 2.1.2. We shall give the proof for more general setting assuming that Lemma 2.2.2 and Lemma 2.2.3 are true for any T. More precisely, we assume that condition (SI) of Lemma A.0.2 is also satisfied for any positive integer T and for $k = m$ in the following form

$$|\mathbb{P}(\bar{A}_2 \bar{A}_3 \cdots \bar{A}_m | A_1) - \alpha| < \varepsilon, \tag{2.5}$$

with $\alpha = q - \frac{T}{m}$, where $\varepsilon = O(p^m)$ as $m \to \infty$.
The above assumption will imply that Theorem 2.1.2 is true for any positive integer T. Now, assumption (2.5) and Remarks 2.2.1 and 2.2.2 imply that Lemma A.0.2 is satisfied with $\alpha = q - \frac{T}{m}$ and

$$\varepsilon = Cm^{T+1}p^m.$$

So we shall apply equation (A.17) of Lemma A.0.2, i.e

$$e^{-(\alpha+10\varepsilon)NP(A_1) - 2mP(A_1)} < \mathbb{P}(\bar{A}_1 \cdots \bar{A}_N) < e^{-(\alpha-10\varepsilon)NP(A_1) + 2mP(A_1)}, \tag{2.6}$$

with the values of $P(A_1) = \binom{m}{T} p^{m-T} q^T$, $\alpha = q - \frac{T}{m}$ and $\varepsilon = Cm^{T+1}p^m$. We shall apply the above inequality for $m = [m(N)] + k$ where $m(N)$ is from equation (2.1). For this, direct calculations show that

$$0 < K_1 \le Nm^T p^m \le K_2 < \infty. \tag{2.7}$$

By inequality (2.6) and using $N_1 = N - m + 1$ instead of N, we have

$$\begin{aligned}\mathbb{P}(\mu(N) < m) =&\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{N_1}) \sim e^{-(\alpha \pm 10\varepsilon)N_1 P(A_1) \pm 2mP(A_1)} =\\ &= e^{-(q-T/m)N_1 P(A_1)} e^{-(\pm 10\varepsilon)N_1 P(A_1)} e^{\pm 2mP(A_1)}.\end{aligned} \tag{2.8}$$

We shall show that the second and third exponent terms in (2.8) converges to 0, so the signs $\pm$ will not affect the result.

As $P(A_1) = \binom{m}{T} p^{m-T} q^T$, and the magnitude of m is $logN$, and therefore the magnitude of the exponent of the third term in (2.8) is

$$(\log N)^{T+1} p^{\log N} = (\log N)^{T+1}/N,$$

which converges to 0 as $N \to \infty$. So we can use the approximation $e^x \le 1 + Cx$ for small values of $|x|$, therefore we obtain

$$e^{\pm 2mP(A_1)} = 1 + O\left((\log N)^{T+1}/N\right). \tag{2.9}$$

Similarly, for the second term in (2.8), we have

$$\begin{aligned} e^{-(\pm 10\varepsilon)N_1 P(A_1)} &\sim e^{\pm C m^{T+1} p^m N_1 m^T p^m} \\ &\sim e^{\pm C (\log N)^{2T+1}/N} \\ &= 1 + O\left((\log N)^{2T+1}/N\right). \end{aligned} \tag{2.10}$$

So, from formulae (2.8) - (2.10), we obtain that

$$\mathbb{P}(\mu(N) < m) = e^{-\left(q - \frac{T}{m}\right) N_1 P(A_1)} \left(1 + O\left(1/(\log N)^2\right)\right). \tag{2.11}$$

As

$$\mathbb{P}(\mu(N) - [m(N)] < k) = \mathbb{P}(\mu(N) < m)$$

with $m = [m(N)] + k = m(N) + k - \{m(N)\}$, so we apply (2.11) for this form of m.

Now, the logarithm of the exponent in (2.11) is

$$
\begin{aligned}
L &= \log\left(\left(q-\frac{T}{m}\right)N_1P(A_1)\right)= \\
&= \log\left(q-\frac{T}{m}\right)+\log N+\log\left(m(m-1)\cdots(m-T+1)\right)+\log\left(p^m\right)+ \\
&\quad +\log\left((q/p)^T\right)-\log(T!)+O\left(\frac{\log N}{N}\right)= \\
&= \log\left(q-\frac{T}{m}\right)+\log N+\log\left(m^T-\frac{T(T-1)}{2}m^{T-1}+O\left(m^{T-2}\right)\right)-m+ \\
&\quad +\log\left((q/p)^T\right)-\log(T!)+O\left(\frac{\log N}{N}\right)= \\
&= \log\left(q-\frac{T}{m}\right)+\log N+\log(m^T)-\frac{\frac{T(T-1)}{2}m^{T-1}}{cm^T}+O\left(\frac{1}{m^2}\right)-m+ \\
&\quad +\log\left((q/p)^T\right)-\log(T!)+O\left(\frac{\log N}{N}\right)= \\
&= \log q-\frac{T}{cqm}+\log N+T\log m-\frac{\frac{T(T-1)}{2}}{cm}-m+ \\
&\quad +\log\left((q/p)^T\right)-\log(T!)+O\left(\frac{1}{(\log N)^2}\right)= \\
&= \log(qN)-\frac{T}{cq_om}+T\log m-m+\log\left((q/p)^T\right)-\log(T!)+O\left(\frac{1}{(\log N)^2}\right),
\end{aligned}
$$

where we applied Taylor's expansion of the log function up to the first order and used notation $q_0=\frac{2q}{2+Tq-q}$ to make simplification easier.
Now, using the notations

$$
\begin{aligned}
D=&-\frac{T^3}{2C}\left(\frac{\log(\log(qN))}{\log(qN)}\right)^2+T^2\frac{\log(\log(qN))}{cq_0(\log(qN))^2}+T^3\frac{\log(\log(qN))}{(c\log(qN))^2}+ \\
&+\left(T\log\left(\frac{q}{p}\right)-\log(T!)\right)\left(\frac{T}{c\log(qN)}-T^2\frac{\log(\log(qN))}{c(\log(qN))^2}\right),
\end{aligned}
\tag{2.12}
$$

$$
B=T^2\,\frac{\log(\log(qN))}{c\log(qN)}-\frac{T}{cq_0\log(qN)}+D \tag{2.13}
$$

and

$$
A=T\log(\log(qN))+B,
$$

We have

$$
m=T\log\left(\frac{q}{p}\right)-\log(T!)+\log(qN)+A+k-\{m(N)\}.
$$

So we obtain that the logarithm of the exponent in (2.11) is

$$
\begin{aligned}
L = & -\frac{T}{cq_0 m} + T\log m - A - k + \{m(N)\} + O\left(\frac{1}{(\log N)^2}\right) = \\
= & -\frac{T}{cq_0 \log(qN)} + T^2\frac{\log(\log(qN))}{cq_0(\log(qN))^2} + \\
& + T\log\big(\log(qN) + T\log(\log(qN)) + B + \log\left((q/p)^T\right) - \log(T!) + k - \{m(N)\}\big) - \\
& - A - k + \{m(N)\} + O\left(\frac{1}{(\log N)^2}\right).
\end{aligned}
$$

Here again, we applied the Taylor's expansion but this time for the function $1/x$. In furthering the application of Taylor's expansion for the $log(x)$ function, we obtain

$$
\begin{aligned}
L = & -\frac{T}{cq_0 \log(qN)} + T^2\frac{\log(\log(qN))}{Cq_0\,(\log(qN))^2} + T\log(\log(qN)) + \\
& + \frac{T(T\log(\log(qN)) + B + \log((q/p)^T) - \log(T!) + k - \{m(N)\})}{c\log(qN)} - \\
& - \frac{1}{2}\frac{T\,(T\log(\log(qN)) + B + \log((q/p)^T) - \log(T!) + k - \{m(N)\})^2}{c(\log(qN))^2} - \\
& - A - k + \{M(N)\} + O\left(\frac{1}{(\log N)^2}\right).
\end{aligned}
$$

From here, B can easily be omitted from the quadratic term and so we can apply that $A = T\log(\log(qN)) + B$, in order to obtain

$$
\begin{aligned}
L = & -\frac{T}{cq_0 \log(qN)} + \frac{T^2\log(\log(qN))}{cq_0(\log(qN))^2} + \frac{T^2\log(\log(qN))}{c\log(qN)} + \\
& + \frac{T\,(\log((q/p)^T) - \log(T!))}{c\log(qN)} + \frac{T^3\log(\log(qN))}{(c\log(qN))^2} - \frac{T^2}{q_0\,(c\log(qN))^2} + \frac{TD}{c\log(qN)} + \\
& + \frac{T(k - \{m(N)\})}{c\log(qN)} - \frac{1}{2}\frac{T^3(\log(\log(qN)))^2}{c(\log(qN))^2} - \\
& - \frac{1}{2}\frac{T\,(\log((q/p)^T) - \log(T!) + k - \{m(N)\})^2}{c(\log(qN))^2} - \\
& - \frac{1}{2}\frac{T\log(\log(qN))(\log((q/p)^T) - \log(T!) + k - \{m(N)\})}{c(\log(qN))^2} - \\
& - B - k + \{m(N)\} + O\left(\frac{1}{(\log N)^2}\right) = \\
& = (k - \{m(N)\})\left(\frac{T}{c\log(qN)} - \frac{T^2\log(\log(qN))}{c(\log(qN))^2} - 1\right) + O\left(\frac{1}{(\log N)^2}\right).
\end{aligned}
$$

So in conclusion, we have by Taylor's expansion

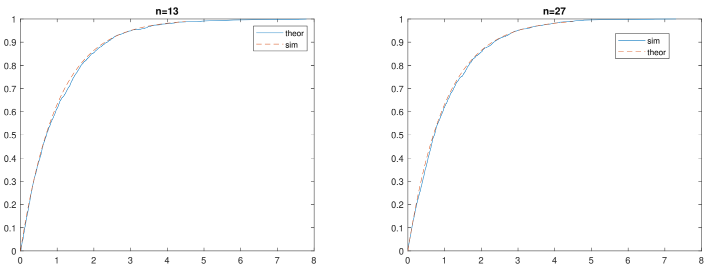
$$
\begin{aligned}
e^{-(q-\frac{T}{m})NP(A_1)} &= e^{-p^{-L}} \\
&= \exp\left(-p^{-\left((\{m(N)\}-k)\left(1-\frac{T}{c\,\log(qN)}+\frac{T^2\log(\log(qN))}{c(\log(qN))^2}+O\left(1/(\log N)^2\right)\right)\right)}\right) \\
&= \exp\left(-p^{-\left((\{m(N)\}-k)\left(1-\frac{T}{c\,\log N}+\frac{T^2\log(\log(qN))}{c(\log(qN))^2}\right)\right)}\left(1+O\left(1/(\log N)^2\right)\right)\right) \\
&= \exp\left(-p^{-\left((\{m(N)\}-k)\left(1-\frac{T}{c\,\log N}+\frac{T^2\,\log(\log(qN))}{c(\log(qN))^2}\right)\right)}\right)\left(1+O\left(1/(\log N)^2\right)\right).
\end{aligned}
\tag{2.14}
$$

So, from equations (2.11) and (2.14), we obtain the desired result. Hence the proof ∎

2.3 Simulation Results

In this section, we begin by presenting simulation results that demonstrate the numerical behaviour of the first hitting time τ_m for a T-contaminated head run. The obtained findings provide empirical evidence in favor of Theorem 2.1.1.

Example 2.3.1. *In this example, the value of* $p = 0.5$. *The length of the coin tossing experiment is is denoted as* $N = 10^6$, *while the number of the repetitions of the experiment is* $s = 2000$.



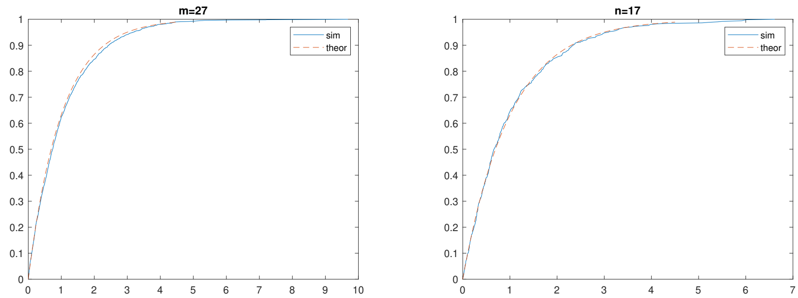
(a)First hitting time, $T = 1$ (b) First hitting time, $T = 2$

Figure 2.1: Comparison of empirical and asymptotic distribution: $T = 1$ and $T = 2$, $p = 0.5$, $N = 10^6$, $s = 2000$

Figure2.1 shows the first hitting time of the T*-contaminated run with lengths of* $n = 13$ *and 27, corresponding to* T *values of 1 and 2, respectively. The empirical distribution, represented by the solid line in the simulation, is contrasted with the asymptotic theoretical distribution, depicted by the dashed line as described in Theorem 2.1.1. The fit of the item is satisfactory.*

Example 2.3.2. *In this example* $p = 0.6$ *and* 0.4, $T = 2$. *The length of the coin tossing experiment is* $N = 10^6$, *the number of the repetitions of the experiment is* $s = 2000$ *and* 500.

Figure 2.2 shows the first hitting time of the T*-contaminated run having length* $n = 27$ *and* 17. *The solid line is the empirical distribution given by the simulation, the dashed line is the asymptotic theoretical distribution presented in Theorem 2.1.1. The fit is good.*

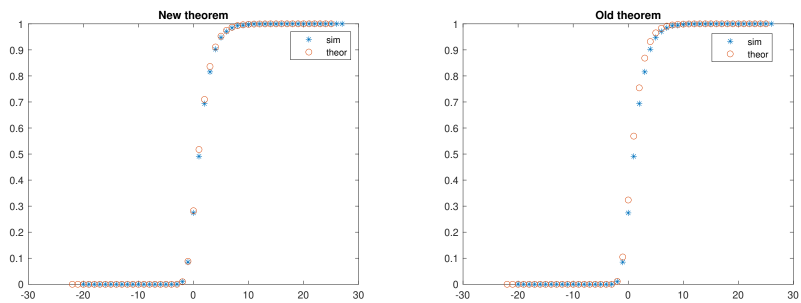


(a)First hitting time, $s = 2000$ (b) First hitting time, $s = 500$

Figure 2.2: Comparison of empirical and asymptotic distribution: $T = 2$, $p = 0.6$ and 0.4, $N = 10^6$, $s = 2000$ and 500

We now present simulation results for $\mu(N)$, i.e. for the length of the longest T contaminated run. They show that our new approximation in Theorem 2.1.2 is better than the former one quoted in Proposition 2.0.1. We implemented the simulation in Matlab.

Example 2.3.3. *In this example* $p = 0.5$, $T = 1$. *The length of the coin tossing experiment is* $N = 10^6$, *the number of the repetitions of the experiment is* $s =$ 2000. *On parts (a) and (b) of Figure 2.3 sign* $\circ$ *shows the theoretical asymptotic probability and* $*$ *shows the relative frequency of those experiments when* $\mu(N)$, *that is the longest* T*-contaminated run is shorter than the given value on the horizontal axis.*



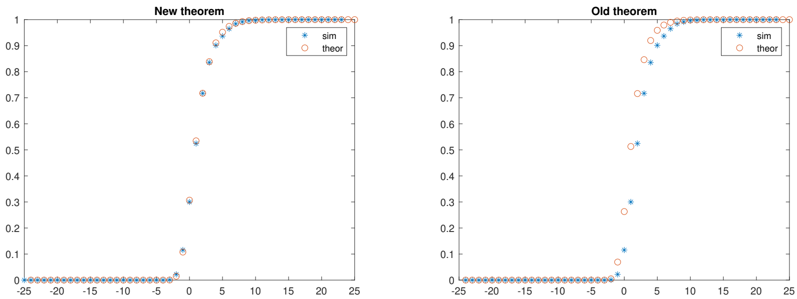
(a) Longest run (new) (b) Longest run (old)

Figure 2.3: Comparison of empirical and asymptotic distribution:$T = 1$, $p = 0.5$, $N = 10^6$, $s = 2000$

Part (a) of Figure 2.3 shows the fit of the empirical distribution of $\mu(N)$ *to the asymptotic distribution given by our Theorem 2.1.2. This shows a nice fit.*

Part (b) of Figure 2.3 shows the fit of the empirical distribution of $\mu(N)$ *to the asymptotic distribution given by the old result quoted in Proposition 2.0.1. The distribution does not fit nicely.*

Example 2.3.4. *In this example* $p = 0.5$, $T = 2$. *The length of the coin tossing experiment is* $N = 10^6$, *the number of the repetitions of the experiment is* $s = 2000$. *On parts (a) and (b) of Figure 2.4 sign* $\circ$ *shows the theoretical asymptotic probability and* $*$ *shows the relative frequency of those experiments when* $\mu(N)$, *that is the longest* T*-contaminated run is shorter than the given value on the horizontal axis.*



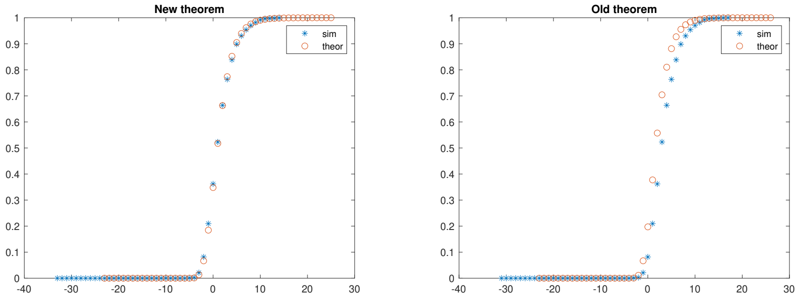
(a) Longest run (new)

(b) Longest run (old)

Figure 2.4: Comparison of empirical and asymptotic distribution: $T = 2$, $p = 0.5$, $N = 10^6$, $s = 2000$

Part (a) of Figure 2.4 shows the fit of the empirical distribution of $\mu(N)$ *to the asymptotic distribution given by our Theorem 2.1.2. This shows a nice fit.*
Part (b) of Figure 2.4 shows the fit of the empirical distribution of $\mu(N)$ *to the asymptotic distribution given by the old result quoted in Proposition 2.0.1. The distribution fits poorly.*

Example 2.3.5. *In this example $p = 0.6$, $T = 2$. The length of the coin tossing experiment is $N = 10^6$, the number of the repetitions of the experiment is $s = 2000$.*

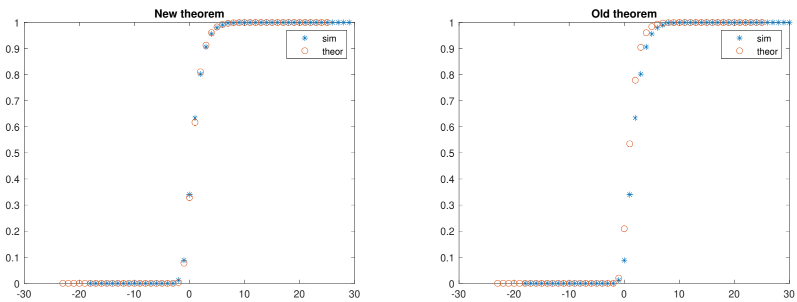


(a) Longest run (new) (b) Longest run (old)

Figure 2.5: Comparison of empirical and asymptotic distribution: $T = 2$, $p = 0.6$, $N = 10^6$, $s = 2000$

Part (a) of Figure 2.5 shows the fit of the empirical distribution of $\mu(N)$ to the asymptotic distribution given by our Theorem 2.1.2. This shows a nice fit for the distribution. Part (b) of Figure 2.5 shows the fit of the empirical distribution of $\mu(N)$ to the asymptotic distribution given by the old result quoted in Proposition 2.0.1. The distribution fits poorly.

Example 2.3.6. *In this example $p = 0.4$, $T = 2$. The length of the coin tossing experiment is $N = 10^6$, the number of the repetitions of the experiment is $s = 500$.*



(a) Longest run (new) (b) Longest run (old)

Figure 2.6: Comparison of empirical and asymptotic distribution: $T = 2$, $p = 0.4$, $N = 10^6$, $s = 500$

Part (a) of Figure 2.6 shows the fit of the empirical distribution of $\mu(N)$ to the asymptotic distribution given by our Theorem 2.1.2. The fit is good. Part (b) of Figure 2.6 shows the fit of the empirical distribution of $\mu(N)$ to the asymptotic *distribution given by the old result quoted in Proposition 2.0.1. The distribution fits poorly.*

In order to assess the numerical accuracy of the approximation to the limit distribution, the uniform distance measure, also known as Kolmogorov's distance measure, is employed. This measure is defined as

$$d_k(X, Y) \equiv d_k(F_X; F_Y) = \sup_x |F_X(x) - F_Y(x)|,$$

where X and Y represent random variables with distribution functions F_X and F_Y, respectively.The determination of the rate of convergence will be deduced through the utilization of Kolmogorov's distance measures.

Example 2.3.7 (Convergence rate)**.** *We performed the coin tossing experiment of length $N = 10^6$, with 2000 repetitions and calculated the Kolmogorov's distance. In the table below, T is the number of contaminations, p is the probability of heads. K_{old} is the Kolmogorov's distance between the empirical distribution of $\mu(N)$ and the asymptotic distribution given by the old result quoted in Proposition 2.0.1. The values are high hence indicating a poor fit. K_{new} is the Kolmogorov's distance between the empirical distribution of $\mu(N)$ and the asymptotic distribution given by our Theorem 2.1.2. The values are low hence indicating a relatively good fit.*

T	p	K_{old}	K_{new}
1	0.5	0.0778	0.0264
2	0.4	0.1948	0.0172
2	0.5	0.2129	0.0148
2	0.6	0.1953	0.0250

Table 2.1: Kolmogorov's distance measure

Chapter 3

Limit theorems for runs containing two types of contaminations

The main goal of this chapter is to extend the approach used in studying limit distribution of the first hitting time and the accompanying distribution of the length of the longest T-contaminated head runs for binary sequences as discussed in Chapter 2 to sequences of trinary state trials. It might seem that the extension should be somewhat straight forward with just minor modifications, but this is not the case owing to the different kinds of contaminations involved.

Various authors have given in depth considerations to experiments involving sequences of runs emerging from trinary trials where Markov chain approach is used in their analysis. Such sequences includes system applications where components might exist in the following states; "perfect functioning", "partial functioning" and "complete failure". Also its worth considering a statistical process control, where a control chart with two control limits (say lower and upper control limits) can also be analyzed by assigning three different states for the values of the statistics computed from sub-samples. (See **Eryilmaz et al. (2016)** and **Koutras and Alexandrou (1997)**.

In this chapter, we define and study the limiting distribution of the first hitting time and the accompanying distribution for the length of the longest at most two type contaminated sequence of runs with trinary trials. As in Chapter 2 our proofs will mainly rely on the main Lemma of **Csáki et al. (1987)**.

3.1 First hitting time for at most two-type contaminated run

Let $X_1, X_2, \ldots, X_N$ be a sequence of independent random variables with three possible outcomes; 0, +1 and -1 labeled as success, failure of type I and failure of type II, respectively with the distribution

$\mathbb{P}(X_i = 0) = p$, $\mathbb{P}(X_i = +1) = q_1$ and $\mathbb{P}(X_i = -1) = q_2$ where $p + q_1 + q_2 = 1$ and $p > 0, \quad q_1 > 0, \quad q_2 > 0$.

An m length sequence is called a pure run if it contains only 0 values. It is called a one-type contaminated run if it contains precisely one non-zero element either a +1 or a -1. On the other hand, it is called a two-type contaminated run if it contains precisely one +1, and one -1 while the rest of the elements are 0's.

A run is called at most two-type contaminated if it is either pure, or one-type contaminated, or two-type contaminated.

So for an arbitrary fixed m, let $A_n = A_{n,m}$ denote the occurrence of the event at the n^{th} step ,that is, there are is at most a two-type contaminated run in the sequence $X_n, X_{n+1}, \ldots X_{n+m-1}$ while $\bar{A_n}$ denotes its non-occurrence.

We see that

$$P(A_1) = p^m + m(1-p)p^{m-1} + m(m-1)p^{m-2}q_1q_2$$

In what follows, we shall use the notation

$$\alpha = \frac{C_0 + \frac{1}{m}c_1 + \frac{1}{m(m-1)}C_2}{1 + \frac{p(1-p)}{(m-1)q_1q_2} + \frac{p^2}{m(m-1)q_1q_2}},$$

where;
$C_0 = (q_1 + q_2)$, $C_1 = \frac{p(q_1^2+q_2^2)}{q_1q_2} - 1$, $C_2 = \frac{(q_1^2+q_2^2)p^2}{q_1q_2(p-1)} + \frac{p}{p-1} + \frac{2(2p+1)q_1q_2}{(p-1)^3}$.

Let τ_m be the first hitting time of the at most two-type contaminated run of heads having length m. We shall be interested in finding the limiting distribution of τ_m as $m \to \infty$ for the case of a sequence containing at most two types of contamination but no two of the same type.

Theorem 3.1.1. *(Fazekas, Fazekas, and Suja (2023)) Let $\mathbb{P}(X_i = 0) = p$, $\mathbb{P}(X_i = +1) = q_1$ and $\mathbb{P}(X_i = -1) = q_2$ be probabilities of success, failure of type I and failure of type II, respectively where $p + q_1 + q_2 = 1$ and $p > 0, \quad q_1 > 0$, $q_2 > 0$. Let τ_m be the first hitting time of the at most two-type contaminated run* *of heads having length m. Then, for $x > 0$,*

$$\mathbb{P}(\tau_m \alpha P(A_1) > x) \sim e^{-x} \tag{3.1}$$

as $m \to \infty$.

Before proceeding with the proof, we shall consider fulfilment of some conditions of the main Lemma A.0.2 given in **Csáki et al. (1987)** for the case of $k = m$ (for fixed m) and $0 < p \leq 1$, such that for $\varepsilon > 0$:

Remark 3.1.1. **(Fazekas, Fazekas, and Suja (2023))** First, we shall consider condition (SIII) and show that it is true for any large enough m.

$$\begin{aligned} P(A_1) =& p^m + m(1-p)p^{m-1} + m(m-1)p^{m-2}q_1q_2 \\ =& m(m-1)p^{m-2}q_1q_2 \left\{ 1 + \frac{p(1-p)}{(m-1)q_1q_2} + \frac{p^2}{m(m-1)q_1q_2} \right\} \leq \frac{\varepsilon}{m}. \end{aligned} \tag{3.2}$$

This inequality is true for any positive ε if m is large enough.
If $m \approx LogN$, then $p^m \approx p^{LogN} = \frac{1}{N}$ and then, $\varepsilon \approx \frac{(LogN)^3}{N}$. (Here, Log denotes logarithm to base $\frac{1}{p}$)

Remark 3.1.2. **(Fazekas, Fazekas, and Suja (2023))** Now, considering condition (SII), if $i > m$, then A_i and A_1 are independent, therefore

$$\sum_{i=m+1}^{2m} \mathbb{P}(A_i | A_1) = mP(A_1) < \varepsilon, \tag{3.3}$$

which gives precisely the previous assumption, remark 3.1.1.

Lemma 3.1.1. ***(Fazekas, Fazekas, and Suja (2023))*** *Condition (SI) is satisfied for $k = m$ in the following form*

$$|\mathbb{P}(\bar{A}_2, \bar{A}_3, \cdots, \bar{A}_m | A_1) - \alpha| < \varepsilon, \tag{3.4}$$

with

$$\alpha = \frac{C_0 + \frac{1}{m}c_1 + \frac{1}{m(m-1)}C_2}{1 + \frac{p(1-p)}{(m-1)q_1q_2} + \frac{p^2}{m(m-1)q_1q_2}}.$$

Proof of Lemma. To begin, we shall be required to divide the event A_1 into the following pairwise disjoint parts;

$$A_1 = A_1^{(0)} \bigcup \left(\bigcup_{i=1}^{m} A_1^{(+)}(i) \right) \bigcup \left(\bigcup_{i=1}^{m} A_1^{(-)}(i) \right) \bigcup \left(\bigcup_{i,j=1, i\neq j}^{m} A_1^{(2)}(i,j) \right),$$

where $A_1^{(0)}$ is the event that $X_1, X_2, \ldots, X_m$ is a pure run,
$A_1^{(+)}(i)$ denotes that the term $X_i = +1$ while the rest are zeros,
$A_1^{(-)}(i)$ denotes that the term $X_i = -1$ while the rest are zeros,
Finally, $A_1^{(2)}(i,j)$ denotes that the terms $X_i = +1$, $X_j = -1$, while the rest are zeros. We shall denote $X_i = +1$ by ● while $X_j = -1$ by ● for ease of visualization. Then, the probability of the joint event is partitioned as

$$\begin{aligned}
\mathbb{P}(A_1\bar{A}_2\cdots\bar{A}_m) =& \mathbb{P}(A_1^{(0)}\bar{A}_2\cdots\bar{A}_m) + \sum_{i=1}^{m}\mathbb{P}(A_1^{(+)}(i)\bar{A}_2\cdots\bar{A}_m) + \\
&+ \sum_{i=1}^{m}\mathbb{P}(A_1^{(-)}(i)\bar{A}_2\cdots\bar{A}_m) + \sum_{i<j}^{m}\mathbb{P}(A_1^{(2)}(i,j)\bar{A}_2\cdots\bar{A}_m) + \\
&+ \sum_{i>j}^{m}\mathbb{P}(A_1^{(2)}(i,j)\bar{A}_2\cdots\bar{A}_m) \\
=& Y^{(0)} + \sum_{i=1}^{m} Y_i^{(+)} + \sum_{i=1}^{m} Y_i^{(-)} + \sum_{i<j}^{m} Y_{i,j}^{(2)} + \sum_{i>j}^{m} Y_{i,j}^{(2)}.
\end{aligned}$$

Here, we can obtain the formula for $\sum_{i=1}^{m} Y_i^{(-)}$ by interchanging the role of q_1 and q_2 in the corresponding formula $\sum_{i=1}^{m} Y_i^{(+)}$.
Similarly, we can obtain $\sum_{i>j}^{m} Y_{i,j}^{(2)}$ by interchanging the role of q_1 and q_2 in the corresponding formula $\sum_{i<j}^{m} Y_{i,j}^{(2)}$.
Therefore, we consider the probabilities of each component;

I. $Y^{(0)} = 0$ because the event is impossible.

II. $Y_i^{(+)} = \mathbb{P}(A_1^{(+)}(i)\bar{A}_2\cdots\bar{A}_m)$.
We want to evaluate probabilities corresponding to different values of i as follows;

(a). If $i = 1$, then the event is impossible. So $Y_1^{(+)} = 0$.

(b). Let $1 < i < m$, i.e $\bigcirc, \cdots, \overset{i}{\bullet}, \cdots, \bigcirc, \overset{m}{\bigcirc}, \cdots, \overset{m+1}{\square}$.
then, the $m+1$ position should be $+1$. Furthermore, on the positions $m+2, \ldots, m+i$, it is not possible that all elements are zeros and also not possible that there is a -1 and the rest of the elements are zeros. So for this case,

$$Y_i^{(+)} = q_1 p^{m-1}.q_1\left(1 - p^{i-1} - (i-1)q_2 p^{i-2}\right), \quad \text{if} \quad i < m. \tag{3.5}$$

(c). If $i = m$, i.e $\bigcirc, \cdots, \bigcirc, \overset{m}{\color{blue}\bullet}, \cdots, \overset{m+1}{\color{blue}\bullet},$
then $m+1$ element should be $+1$ and other remaining elements to be arbitrary. So this part has

$$Y_i^{(+)} = q_1 p^{m-1} . q_1. \tag{3.6}$$

III. Now , lets turn to $Y_{i,j}$, first we consider the case when $i < j$.

(a). When $i = 1$ and $j = m$, i.e $\overset{1}{\color{blue}\bullet}, \bigcirc, \cdots, \bigcirc, \overset{m}{\color{red}\bullet}, \overset{m+1}{\square}.$
Then, X_{m+1} should be -1 and the remaining elements to be arbitrary. So this part has

$$Y_{1,m} = q_1 q_2 p^{m-2} q_2. \tag{3.7}$$

(b). Now, let $i = 1$ and $j < m$ i.e $\overset{1}{\color{blue}\bullet}, \bigcirc, \cdots, \overset{j}{\color{red}\bullet}, \cdots \overset{m}{\bigcirc}, \overset{m+1}{\square}.$
Then, X_{m+1} should be -1. Moreover, on positions $m+2, \ldots, m+j$, all elements being zeros is not possible, neither is a one $+1$ and the rest being zeros possible. So

$$Y_{1,j} = q_1 q_2 p^{m-2} . q_2 \left(1 - p^{j-1} - p^{j-2}(j-1) q_1\right), \quad \text{if} \quad j < m. \tag{3.8}$$

(c). Now, let $i > 1$ and $j = m$ i.e $\bigcirc, \cdots, \overset{i}{\color{blue}\bullet}, \bigcirc, \cdots, \overset{m}{\color{red}\bullet}, \overset{m+1}{\square}.$
Then, X_{m+1} can either be a $+1$ or a -1.
When X_{m+1} is -1, then the remaining elements are arbitrary. However, if X_{m+1} is $+1$, then on positions $m+2, \ldots, m+i$, there should be at least one non- zero element. So

$$Y_{i,m} = q_1 q_2 p^{m-2} . \left(q_1 \left(1 - p^{i-1}\right) + q_2\right), \quad \text{if} \quad i > 1, \quad j = m. \tag{3.9}$$

(d). Consider now the case $i > 1$ and $j < m$ i.e $\bigcirc, \cdots, \overset{i}{\color{blue}\bullet}, \bigcirc, \cdots, \bigcirc, \overset{j}{\color{red}\bullet}, \overset{m}{\bigcirc}.$
We divide this event into two parts.
First, let $X_{m+1} = +1$, $\bigcirc, \cdots, \overset{i}{\color{blue}\bullet}, \cdots, \bigcirc, \overset{j}{\color{red}\bullet}, \overset{m}{\bigcirc}, \overset{m+1}{\color{blue}\bullet}.$
Then, it is not possible that the elements in positions $m+2, \ldots, m+i$ are all zeros. It also impossible that there is one -1 among $m+2, \ldots, m+i$ while all $m+i+1, \ldots, m+j$ are zeros. So this part of $Y_{i,j}$ is;

$$Y_{i,j} = q_1 q_2 p^{m-2} . q_1 \left(1 - p^{i-1} - (i-1) p^{i-2} q_2 . p^{j-1}\right), \quad \text{if} \quad i > 1, \quad j < m. \tag{3.10}$$

Finally, now let $X_{m+1} = -1$. $\quad \bigcirc, \cdots, \overset{i}{\bullet}, \cdots, \bigcirc, \overset{j}{\bullet}, \overset{m}{\bigcirc}, \overset{m+1}{\bullet}.$
Then it is not possible that all elements in $m+2, \ldots, m+j$ are zeros and also, it's impossible that among $m+2, \ldots, m+j$ there is one $+1$ and the rest are zeros. So this second part of $Y_{i,j}$ is;

$$Y_{i,j} = q_1 q_2 p^{m-2} . q_2 \left(1 - p^{j-1} - (j-1) q_1 p^{j-2}\right), \quad \text{if} \quad i > 1, \quad j < m. \tag{3.11}$$

Summing equations (3.5) and (3.6), we get $Y_i^{(+)}$ and consequently by interchanging the roles of q_1 and q_2 we obtain $Y_j^{(-)}$ as follows

$$\begin{aligned}
\sum_{i=1}^{m} Y_i^{(+)} + \sum_{i=1}^{m} Y_i^{(-)} &= \sum_{i=2}^{m-1} q_1^2 p^{m-1} \left(1 - p^{i-1} - (i-1) q_2 p^{i-2}\right) + q_1^2 p^{m-1} \\
&\quad + \sum_{i=2}^{m-1} q_2^2 p^{m-1} \left(1 - p^{i-1} - (i-1) q_1 p^{i-2}\right) + q_2^2 p^{m-1} \\
&= (m-1) p^{m-1} \left(q_1^2 + q_2^2\right) - \left(q_1^2 + q_2^2\right) p^{m-1} \sum_{i=2}^{m-1} p^{i-1} \\
&\quad - p^{m-1} \left(q_1^2 q_2 + q_2^2 q_1\right) \sum_{i=2}^{m-1} (i-1) p^{i-2} \\
&= (m-1) p^{m-1} \left(q_1^2 + q_2^2\right) - \left(q_1^2 + q_2^2\right) p^{m-1} \frac{p^{m-1} - p}{p-1} \\
&\quad - p^{m-1} q_1 q_2 \left(q_1 + q_2\right) \left(\frac{(m-2) p^{m-2} - 1}{p-1} + \frac{p - p^{m-2}}{(p-1)^2}\right) \\
&= p^{m-1} \left\{ \left(q_1^2 + q_2^2\right) \left[(m-1) - \frac{p^{m-1}}{p-1} + \frac{p}{p-1}\right] \right. \\
&\quad \left. + \; q_1 q_2 \left[(m-2) p^{m-2} - 1 + \frac{p}{p-1} - \frac{p^{m-2}}{p-1}\right] \right\} \\
&= p^{m-1} \left\{ \left(q_1^2 + q_2^2\right) \left[(m-1) + \frac{p}{p-1}\right] + q_1 q_2 \frac{1}{p-1} \right. \\
&\quad \left. - \left(q_1^2 + q_2^2\right) \frac{p^{m-1}}{p-1} + q_1 q_2 \left[(m-2) p^{m-2} - \frac{p^{m-2}}{p-1}\right] \right\}.
\end{aligned} \tag{3.12}$$

Here above, we applied

$$\sum_{i=a}^{b} i p^{i-1} = \frac{b p^b - a p^{a-1}}{p-1} + \frac{p^a - p^b}{(p-1)^2},$$

which can be obtained by differentiating the known formula for the sum of a geometric sequence.

Similarly, summing equations (3.7), (3.8), (3.9), (3.10) and (3.11) together with finding their corresponding changed versions got by interchanging roles of q_1 and q_2, we obtain

$$
\begin{aligned}
&\sum_{i<j} Y_{i,j} + \sum_{i>j} Y_{i,j} = q_1 q_2 p^{m-2}(q_1+q_2) + \\
&\quad + q_1 q_2 p^{m-2} \sum_{j=2}^{m-1} \left(q_2 \left(1 - p^{j-1} - p^{j-2}(j-1) q_1\right) \right. \\
&\quad \left. + q_1 \left(1 - p^{j-1} - p^{j-2}(j-1) q_2\right) \right) \\
&\quad + q_1 q_2 p^{m-2} \left[\sum_{i=2}^{m-1} \left(q_1 \left(1 - p^{i-1}\right) + q_2 \right) + \sum_{i=2}^{m-1} \left(q_2 \left(1 - p^{i-1}\right) + q_1 \right) \right] \\
&\quad + \sum_{i=2}^{m-2} \sum_{j=i+1}^{m-1} \left[q_1 q_2 p^{m-2} q_1 \left(1 - p^{i-1} - (i-1) p^{i-2} q_2 p^{j-i}\right) \right. \\
&\quad + q_1 q_2 p^{m-2} q_2 \left(1 - p^{j-1} - (j-1) q_1 p^{j-2}\right) \\
&\quad + q_1 q_2 p^{m-2} q_2 \left(1 - p^{i-1} - (i-1) p^{i-2} q_1 p^{j-i}\right) \\
&\quad \left. + q_1 q_2 p^{m-2} q_1 \left(1 - p^{j-1} - (j-1) q_2 p^{j-2}\right) \right] \\
&= q_1 q_2 p^{m-2} \left\{ (q_1+q_2) \left(m - 1 - \frac{p^{m-1}-p}{p-1} \right) - 2 q_1 q_2 \sum_{j=2}^{m-1} (j-1) p^{j-2} \right. \\
&\quad + 2(q_1+q_2)(m-2) - (q_1+q_2) \frac{p^{m-1}-p}{p-1} + \sum_{i=2}^{m-2} \sum_{j=i+1}^{m-1} \left[(q_1+q_2)(1-p^{i-1}) \right. \\
&\quad \left. \left. - 2 q_1 q_2 (i-1) p^{j-2} + (q_1+q_2)(1-p^{j-1}) - 2 q_1 q_2 (j-1) p^{j-2} \right] \right\} \\
&= q_1 q_2 p^{m-2} \left\{ (q_1+q_2)(3m-5) - 2(q_1+q_2) \frac{p^{m-1}-p}{p-1} \right. \\
&\quad - 2 q_1 q_2 \left(\frac{(m-2)p^{m-2}-1}{p-1} + \frac{p-p^{m-2}}{(p-1)^2} \right) \\
&\quad + \sum_{i=2}^{m-2} \left[(q_1+q_2)(1-p^{i-1})(m-i-1) - 2 q_1 q_2 (i-1) \frac{p^{m-2}-p^{i-1}}{p-1} \right. \\
&\quad + (q_1+q_2)(m-i-1) - (q_1+q_2) \frac{p^{m-1}-p^i}{p-1} \\
&\quad \left. \left. - 2 q_1 q_2 \left(\frac{(m-2)p^{m-2} - i p^{i-1}}{p-1} + \frac{p^i - p^{m-2}}{(p-1)^2} \right) \right] \right\}.
\end{aligned}
$$

$$
\begin{aligned}
&\sum_{i<j} Y_{i,j} + \sum_{i>j} Y_{i,j} = \\
&= q_1q_2p^{m-2}\Big\{(q_1+q_2)(3m-5) + 2(p^{m-1}-p) \\
&\quad -2q_1q_2\left(\frac{(m-2)p^{m-2}-1}{p-1} + \frac{p-p^{m-2}}{(p-1)^2}\right) + (q_1+q_2)(m-3)\frac{m-2}{2} \\
&\quad -(q_1+q_2)m\sum_{i=2}^{m-2} p^{i-1} + (q_1+q_2)\sum_{i=2}^{m-2} p^{i-1}(i+1) \\
&\quad -2q_1q_2\frac{p^{m-2}}{p-1}(m-3)\frac{m-2}{2} + 2\frac{q_1q_2}{p-1}\sum_{i=2}^{m-2}(i-1)p^{i-1} \\
&\quad +(q_1+q_2)(m-3)\frac{m-2}{2} - (q_1+q_2)\frac{p^{m-1}}{p-1}(m-3) + \frac{q_1+q_2}{p-1}\sum_{i=2}^{m-2} p^i \\
&\quad -\frac{2q_1q_2}{p-1}(m-2)p^{m-2}(m-3) + \frac{2q_1q_2}{p-1}\sum_{i=2}^{m-2} ip^{i-1} - \frac{2q_1q_2}{(p-1)^2}\sum_{i=2}^{m-2} p^i \\
&\quad +\frac{2q_1q_2}{(p-1)^2}p^{m-2}(m-3)\Big\} \\
&= q_1q_2p^{m-2}\Big\{(q_1+q_2)(3m-5) + (m-3)(m-2) + 2\left(p^{m-1}-p\right) \\
&\quad -2q_1q_2\frac{(m-2)p^{m-2}}{p-1} + 2q_1q_2\frac{1}{p-1} - \frac{2q_1q_2p}{(p-1)^2} + \frac{2q_1q_2p^{m-2}}{(p-1)^2} \\
&\quad -(q_1+q_2)m\frac{p^{m-2}-p}{p-1} + (q_1+q_2)\left(\frac{1}{p}.\frac{(m-1)p^{m-1}-3p^2}{p-1} + \frac{1}{p}.\frac{p^3-p^{m-1}}{(p-1)^2}\right) \\
&\quad -2q_1q_2\frac{p^{m-2}}{p-1}\frac{(m-3)(m-2)}{2} + \frac{2q_1q_2}{p-1}\left(\frac{(m-3)p^{m-3}-1}{p-1} + \frac{p-p^{m-3}}{(p-1)^2}\right).p \\
&\quad +p^{m-1}(m-3) - \frac{p^{m-1}-p^2}{p-1} - \frac{2q_1q_2p^{m-2}}{p-1}(m-2)(m-3) \\
&\quad +\frac{2q_1q_2}{p-1}\left(\frac{(m-2)p^{m-2}-2p}{p-1} + \frac{p^2-p^{m-2}}{(p-1)^2}\right) - \frac{2q_1q_2}{(p-1)^2}\frac{p^{m-1}-p^2}{p-1} \\
&\quad +\frac{2q_1q_2}{(p-1)^2}p^{m-2}(m-3)\Big\} \\
&= q_1q_2p^{m-2}\Big\{(q_1+q_2)(3m-5-m^2-5m+6) + 2p^{m-1} - 2p \\
&\quad -2q_1q_2\frac{(m-2)p^{m-2}}{p-1} + 2q_1q_2\frac{1}{p-1} - 2q_1q_2\frac{p}{(p-1)^2} + 2q_1q_2\frac{p^{m-2}}{(p-1)^2} \\
&\quad +mp^{m-2} - pm - (m-1)p^{m-2} \\
&\quad +3p - \frac{p^2}{p-1} + \frac{p^{m-2}}{p-1} - q_1q_2\frac{p^{m-2}}{p-1}(m-2)(m-3) + \frac{2q_1q_2(m-3)p^{m-2}}{(p-1)^2} \\
&\quad -\frac{2q_1q_2p}{(p-1)^2} + \frac{2q_1q_2p^2}{(p-1)^3} - \frac{2q_1q_2p^{m-2}}{(p-1)^3} + p^{m-1}(m-3) - \frac{p^{m-1}}{p-1} + \frac{p^2}{p-1} \\
&\quad -\frac{2q_1q_2p^{m-2}}{p-1}(m-2)(m-3) + \frac{2q_1q_2(m-2)p^{m-2}}{(p-1)^2} - \frac{4q_1q_2p}{(p-1)^2} \\
&\quad +\frac{2q_1q_2p^2}{(p-1)^3} - \frac{2q_1q_2p^{m-2}}{(p-1)^3} - \frac{2q_1q_2p^{m-1}}{(p-1)^3} + \frac{2q_1q_2p^2}{(p-1)^3} + \frac{2q_1q_2p^{m-2}(m-3)}{(p-1)^2}\Big\}.
\end{aligned}
$$

$$
\begin{aligned}
&\sum_{i<j} Y_{i,j} + \sum_{i>j} Y_{i,j} = \\
&= q_1q_2p^{m-2}\left\{(q_1+q_2)(m-1)^2 - mp + p + 2q_1q_2\frac{1}{p-1} - 2q_1q_2\frac{p}{(p-1)^2}\right. \\
&\quad -\frac{2q_1q_2p}{(p-1)^2} + \frac{2q_1q_2p^2}{(p-1)^3} + \frac{p^2}{p-1} - \frac{4q_1q_2p}{(p-1)^2} + \frac{2q_1q_2p^2}{(p-1)^3} - \frac{p^2}{p-1} \\
&\quad -2q_1q_2\frac{(m-2)p^{m-2}}{p-1} + mp^{m-2} - (m-1)p^{m-2} - q_1q_2\frac{p^{m-2}}{p-1}(m-3)(m-2) \\
&\quad +\frac{2q_1q_2(m-3)p^{m-2}}{(p-1)^2} + p^{m-1}(m-3) - \frac{2q_1q_2p^{m-2}}{p-1}(m-2)(m-3) \\
&\quad +\frac{2q_1q_2(m-2)p^{m-2}}{(p-1)^2} + \frac{2q_1q_2p^{m-2}(m-3)}{(p-1)^2} + 2p^{m-1} + \frac{2q_1q_2p^{m-2}}{(p-1)^2} + \frac{p^{m-2}}{p-1} \\
&\quad \left.-\frac{2q_1q_2p^{m-2}}{(p-1)^3} - \frac{p^{m-1}}{p-1} - \frac{2q_1q_2p^{m-2}}{(p-1)^3} - \frac{2q_1q_2p^{m-1}}{(p-1)^3}\right\} \\
&= q_1q_2p^{m-2}\left\{(q_1+q_2)(m-1)^2 - p(m-1) + \frac{2q_1q_2}{p-1}\left[(p^2-2p+1)\right.\right. \\
&\quad \left.-4(p^2-p)+3p^2\right] - 3q_1q_2\frac{p^{m-2}}{p-1}(m-3)(m-2) + p^{m-1}(m-3) \\
&\quad +\frac{2q_1q_2}{p-1}p^{m-3}\left(-p(p-1)(m-2) + p(m-3) + (m-2)p + p(m-3)\right) \\
&\quad +p^{m-2} + 2p^{m-1} + \frac{p^{m-2}}{p-1} - \frac{p^{m-1}}{p-1} + q_1q_2p^{m-2}\left\{(q_1+q_2)(m-1)^2\right. \\
&\quad -p(m-1) + \frac{2q_1q_2}{(p-1)^3}(2p+1) - 3q_1q_2\frac{p^{m-2}}{p-1}(m-2)(m-3) + p^{m-1}(m-3) \\
&\quad +\frac{2q_1q_2}{(p-1)^2}p^{m-3}p\left[(4-p)m+2p-10\right] + \frac{p^{m-2}}{p-1}(2p(p-1)) \\
&\quad \left.+\frac{2q_1q_2}{(p-1)^3}p^{m-3}(-3p)\right\} \\
&= q_1q_2p^{m-2}\left\{(q_1+q_2)(m-1)^2 - p(m-1) + 2(2p+1)\frac{q_1q_2}{(p-1)^3}\right. \\
&\quad -3q_1q_2\frac{p^{m-2}}{p-1}(m-2)(m-3) \\
&\quad +p^{m-1}(m-3) + \frac{2q_1q_2}{(p-1)^2}p^{m-3}p\left[(4-p)m+2p-10\right] \\
&\quad \left.+2p^{m-1} - \frac{6q_1q_2p^{m-2}}{(p-1)^3}\right\} \\
&= q_1q_2p^{m-2}\left\{\underbrace{m(m-1)(q_1+q_2) - (m-1)}_{m^2(q_1+q_2)-m(q_1+q_2+1)+1} + \frac{2(2p+1)q_1q_2}{(p-1)^3}\right. \\
&\quad +\frac{q_1q_2p^{m-2}}{(p-1)^3}\big(-3(p-1)^2m^2 + m(p-1)(13p-7) \\
&\quad \left.+(-14p^2+12p-4)\big) + p^{m-1}(m-1)\right\}.
\end{aligned} \tag{3.13}
$$

Therefore, combining (3.12) and (3.13) and by some simplification, we obtain

$$\begin{aligned}
\mathbb{P}(A_1\bar{A}_2\cdots\bar{A}_m) =& p^{m-1}\left\{(q_1^2+q_2^2)\left[(m-1)+\frac{p}{p-1}\right]+q_1q_2\frac{1}{p-1}+O(mp^m)\right\}\\
&+q_1q_2p^{m-2}\left\{m(m-1)(q_1+q_2)-(m-1)+\frac{2(2p+1)q_1q_2}{(p-1)^3}\right.\\
&\left.+O(m^2p^m)\right\}\\
=& m(m-1)p^{m-2}q_1q_2\left\{\frac{p(q_1^2+q_2^2)}{mq_1q_2}+\frac{(q_1^2+q_2^2)p^2}{m(m-1)q_1q_2(p-1)}\right.\\
&+\frac{p}{m(m-1)(p-1)}+O\left(\frac{p^m}{m}\right)+(q_1+q_2)-\frac{1}{m}\\
&\left.+\frac{2(2p+1)q_1q_2}{m(m-1)(p-1)^3}+O(p^m)\right\}\\
=& m(m-1)p^{m-2}q_1q_2\left\{C_0+\frac{1}{m}C_1+\frac{1}{m(m-1)}C_2+O(p^m)\right\},
\end{aligned}$$

where; $C_0=(q_1+q_2)$, $C_1=\frac{p(q_1^2+q_2^2)}{q_1q_2}-1$, $C_2=\frac{(q_1^2+q_2^2)p^2}{q_1q_2(p-1)}+\frac{p}{p-1}+\frac{2(2p+1)q_1q_2}{(p-1)^3}$
So,

$$\mathbb{P}(\bar{A}_2\cdots\bar{A}_m|A_1)=\frac{C_0+\frac{1}{m}c_1+\frac{1}{m(m-1)}C_2+O(p^m)}{1+\frac{p(p-1)}{(m-1)q_1q_2}+\frac{p^2}{m(m-1)q_1q_2}}.$$

We therefore satisfy Lemma 3.1.1.

$$|\mathbb{P}(\bar{A}_2\cdots\bar{A}_m|A_1)-\alpha|<\varepsilon,$$

where

$$\alpha=\frac{C_0+\frac{1}{m}c_1+\frac{1}{m(m-1)}C_2}{1+\frac{p(1-p)}{(m-1)q_1q_2}+\frac{p^2}{m(m-1)q_1q_2}},$$

and $\varepsilon=O(p^m)$. □

Since all the three conditions have been fulfilled, we now embark on the proofs of our theorems.

3.2 Proof of First hitting time

Proof of Theorem 3.1.1. We shall apply the Main Lemma (stationary case, finite form) in **Csáki et al. (1987)** since its conditions in remarks 3.1.1, 3.1.2 and Lemma 3.1.1 are satisfied. So approximation of the probabilities $\mathbb{P}\left(\bigcup_{i=1}^m A_i\right)=1-\mathbb{P}(\bar{A}_1\bar{A}_2\cdots\bar{A}_m)$ is possible i.e we can find the limiting distribution of the random variable τ_m = first observation of the event $A_{n,m}$.

Let τ_m be the first hitting time of the at most two-type contaminated run of length m. Then,

$$\mathbb{P}(\tau_m \alpha P(A_1) > x) = \mathbb{P}\left(\tau_m > \frac{x}{\alpha P(A_1)}\right) = \mathbb{P}(\tau_m > N),$$

where N is the integer part of $\frac{x}{\alpha P(A_1)}$.
By **Csáki et al. (1987)** main lemma,

$$e^{-(\alpha+10\epsilon)N_1 P(A_1)-2mP(A_1)} < \mathbb{P}(\bar{A}_1 \cdots \bar{A}_{N_1}) < e^{-(\alpha-10\epsilon)N_1 P(A_1)+2mP(A_1)}$$

Where in this case, $N_1 = N - m + 1$.
Let $\varepsilon_0 > 0$ be fixed and choose ε so that $10\varepsilon = \frac{\varepsilon_0}{m}$. We can do this because $\varepsilon = O(p^m)$.
Now

$$\begin{aligned}
\mathbb{P}(\tau_m > N) =& \mathbb{P}(\bar{A}_1 \cdots \bar{A}_{N_1}) \\
&\sim e^{-(\alpha \pm \frac{\varepsilon_0}{m})N_1 P(A_1) \pm 2mP(A_1)} \sim e^{-\left(\alpha \pm \frac{\varepsilon_0}{m}\right)\left(\frac{x}{\alpha P(A_1)} - m + 1\right)P(A_1)} e^{\pm 2mP(A_1)} \\
&= e^{-\left(\alpha \pm \frac{\varepsilon_0}{m}\right)\left(\frac{xP(A_1)}{\alpha P(A_1)}\right)} . e^{-\left(\alpha \pm \frac{\varepsilon_0}{m}\right)(-m+1)P(A_1)} . e^{\pm 2mP(A_1)} \\
&\sim e^{-x}
\end{aligned}$$

because as $m \to \infty$, $\alpha \approx C_0$, where $C_0 = (q_1 + q_2)$ and therefore $mP(A_1) \to 0$ as $m \to \infty$. Consequently both the 3^{rd} and 2^{nd} exponents tend to 1 and upon simplifying the first exponent, we obtained the following theorem 3.1.1

$$\mathbb{P}(\tau_m \alpha P(A_1) > x) \sim e^{-x} \quad as \quad m \to \infty.$$

Hence the proof. $\square$

3.3 Length of the longest at most two-type contaminated run

Let $\mu(N)$ be the length of the longest at most two-type contaminated run in $X_1, X_2, \ldots, X_N$. Then,
$\{\mu(N) < m\}$ is possible if and only if any m length run in $X_1, X_2, \ldots, X_N$ is neither two-type contaminated nor one-type contaminated nor pure.

We need some notation so let

$$K = \frac{2C_0C_2 - C_!^2 - C_0^2}{2CC_0^2},$$

where $C = \ln \frac{1}{p}$ and ln is the logarithm to base e. Let

$$\begin{aligned}
m(N) = &\log N + 2\log\log N + \frac{4\log\log N}{C\log N} + \frac{C_1 - C_0}{CC_0}\frac{1}{\log N} - \\
&-\frac{4}{C}\frac{(\log\log N)^2}{(\log N)^2} + \left(\frac{8}{C^2} - \frac{2(C_1-C_0)}{CC_0}\right)\frac{\log\log N}{(\log N)^2} + \\
&+\left(\frac{2(C_1-C_0)}{C^2C_0} + K\right)\frac{1}{(\log N)^2} + \frac{16}{3C}\frac{(\log\log N)^3}{(\log N)^3} + \\
&+\left(-\frac{16}{C^2} + \frac{4(C_1-C_0)}{CC_0}\right)\frac{(\log\log N)^2}{(\log N)^3} - \left(4K + \frac{8(C_1-C_0)}{C^2C_0}\right)\frac{\log\log N}{(\log N)^3} + \\
&+\frac{16}{C^3}\frac{\log\log N}{(\log N)^3} - \frac{8}{C^2}\frac{(\log\log N)^2}{(\log N)^3} - \frac{4(C_1-C_0)}{C^2C_0}\frac{\log\log N}{(\log N)^3},
\end{aligned}$$

where log denotes the logarithm to base $1/p$. Let $[m(M)]$ denote the integer part of $m(N)$ and let $\{m(N)\} = m(N) - [m(M)]$ denote its fractional part. We also introduce the function

$$\begin{aligned}
H(x) = &-x + \frac{2}{C\log N}x - \frac{4}{C}\frac{\log\log N}{(\log N)^2}x - \frac{C_1-C_0}{CC_0}\frac{1}{(\log N)^2}x \\
&+\left(\frac{4(C_1-C_0)}{CC_0} - \frac{8}{C^2}\right)\frac{\log\log N}{(\log N)^3}x + \frac{8}{C}\frac{(\log\log N)^2}{(\log N)^3}x \\
&-\frac{1}{C}\frac{1}{(\log N)^2}x^2 + \frac{4}{C}\frac{\log\log N}{(\log N)^3}x^2.
\end{aligned}$$

Theorem 3.3.1. *(Fazekas, Fazekas, and Suja (2023)) Let $0 < p < 1$, $q_1 > 0$, $q_2 > 0$ be fixed with $p + q_1 + q_2 = 1$. Let $\mu(N)$ be the length of the longest at most two-type contaminated run in $X_1, X_2, \ldots, X_N$. Then for $k > 0$,*

$$\mathbb{P}(\mu(N) - [m(N)] < k) = \exp\left(-p^{-\left(\log(C_0p^{-2}q_1q_2) + H(k - \{m(N)\})\right)}\right)\left(1 + O\left(\frac{1}{(\log N)^3}\right)\right). \tag{3.14}$$

Proof of Theorem 3.3.1. Let $N_1 = N - m + 1$ where m will be specified so that $m \sim \log N$. Then,

$$\begin{aligned}
\mathbb{P}(\mu(N) < m) = &\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{N_1}) \sim e^{-(\alpha \pm 10\varepsilon)N_1P(A_1) \pm 2mP(A_1)} \\
&= e^{-\alpha N_1 P(A_1)} e^{\pm 10\varepsilon N_1 P(A_1)} e^{\pm 2mP(A_1)}.
\end{aligned}$$

As $mP(A_1) \sim m^3p^m = \frac{(\log N)^3}{N}$, so beginning with the last part of the exponent we see that $e^{\pm 2mP(A_1)} = 1 + O\left(\frac{(\log N)^3}{N}\right)$.
Similarly, as $\varepsilon = O\left(p^m\right)$, $m \approx \log N$ and now considering the second part of the exponent, we see that

$$e^{\pm 10\varepsilon P(A_1)} \sim e^{\pm 10P(A_1)} \sim e^{\pm(\log N)^2/N} = 1 + O\left(\frac{(\log N)^2}{N}\right).$$

Therefore, we can calculate the first part of the exponent

$$e^{-\alpha N_1 P(A_1)} = e^{-\alpha NP(A_1)} \cdot \underbrace{e^{+\alpha(m-1)P(A_1)}}_{1+O\left(\frac{(\log N)^3}{N}\right)}.$$

So we have to calculate: $e^{-\alpha NP(A_1)} = e^{-L}$,
where

$$\begin{aligned}
l &= \alpha NP(A_1) \\
&= \frac{C_0 + \frac{1}{m}C_1 + \frac{1}{m(m-1)}C_2}{1 + \frac{p(1-p)}{(m-1)q_1q_2} + \frac{p^2}{m(m-1)q_1q_2}}.N.m(m-1)p^{m-2}(q_1q_2)\left(1 + \frac{p(1-p)}{(m-1)q_1q_2} + \right. \\
&\quad \left. + \frac{p^2}{m(m-1)q_1q_2}\right) \\
&\quad = Np^{m-2}q_1q_2\left(m(m-1)C_0 + (m-1)C_1 + C_2\right) \\
&\quad = Np^{m-2}q_1q_2\left(m^2C_0 + m(C_1 - C_0) + C_2 - C_1\right).
\end{aligned}$$

Our aim is to find approximate value of $m(N)$ so that the asymptotic behaviour of $\mathbb{P}(\mu(N) - [m(N)] < k)$ can be obtained. Here $[m(N)]$ is the integer part of $m(N)$ and $\{m(N)\}$ will denotes its fractional part. Then

$$\mathbb{P}(\mu(N) - [m(N)] < k) = \mathbb{P}(\mu(N) < m(N) + k - \{m(N)\}).$$

Let us define $m = m(N) + k - \{m(N)\}$. So

$$\mathbb{P}(\mu(N) - [m(N)] < k) = \mathbb{P}(\mu(N) < m) = e^{-l}\left(1 + O\left(\frac{(\log N)^3}{N}\right)\right),$$

where log is the logarithm to base $1/p$. We want to find $m(N)$ so that the remainder term in the exponent l be small. We shall do it step by step using several Taylor's expansions. We try to find $m(N)$ as $\log N + A$ where A is to be specified later . So

$$m = \log N + A + k - \{m(N)\}.$$

Then using Taylor expansion of

$$\log(x_0 + y) = \log x_0 + \frac{y}{Cx_0} - \frac{1}{2C}\frac{y^2}{x_0^2} + \frac{1}{3C}\frac{y^3}{\tilde{x}^3},$$

where $\tilde{x}$ is between x_0 and $x_0 + y$ and $x_0 > 0$, $x_0 + y > 0$ and where $C = \ln \frac{1}{p}$ and ln is the logarithm to base e, so we obtain

$$\begin{aligned}
L = \ & \log l = \log\left(p^{-2}q_1q_2\right) - m + \log N + \log(m^2C_0) + \frac{m(C_1 - C_0) + (C_2 - C_1)}{Cm^2C_0} - \\
& -\frac{1}{2C}\frac{(m(C_1 - C_0) + (C_2 - C_1))^2}{(Cm^2C_0)^2} + O\left(\frac{1}{m^3}\right) \\
= \ & \log\left(p^{-2}q_1q_2\right) - m + \log N + 2\log m + \log C_0 + \frac{C_1 - C_0}{CC_0m} + \frac{1}{m^2}K + O\left(\frac{1}{m^3}\right), \\
& \text{where} \quad K = \frac{C_2 - C_1}{CC_0} - \frac{(C_1 - C_0)^2}{2CC_0^2} = \frac{2C_0C_2 - C_1^2 - C_0^2}{2CC_0^2}.
\end{aligned}$$

Now let

$m = \log N + A + k - \{m(N)\}$. *Then*

$$\begin{aligned}
L = \ & \log\left(C_0p^{-2}q_1q_2\right) - \log N - A - (k - \{m(N)\}) + 2\log\left(\log N + A + k - \{m(N)\}\right) + \\
& + \frac{C_1 - C_0}{CC_0m} + \frac{1}{m^2}K + O\left(\frac{1}{m^3}\right).
\end{aligned}$$

where again, applying the Taylor expansion of $\log(X_0 + y) = \log X_0 + \frac{y}{CX_0} - \frac{1}{2C}\frac{y^2}{X_0^2} + \frac{1}{3C}\frac{y^3}{\tilde{X_0}^3}$, we have

$$\begin{aligned}
L = \ & \log(C_0p^{-2}q_1q_2) - A - (k - \{m(N)\}) + 2\left(\log\log N + \frac{A + k - \{m(N)\}}{C\log N} - \right. \\
& \left. -\frac{1}{2C}\frac{(A + k - \{m(N)\})^2}{(\log N)^2} + \frac{1}{3C}\frac{(A + k - \{m(N)\})^3}{(\log N)^3}\right) \\
& + O\left(\frac{1}{(\log N)^3}\right) + \frac{C_1 - C_0}{CC_0m} + \frac{K}{m^2} + O\left(\frac{1}{m^3}\right) \\
= \ & \log(C_0p^{-2}q_1q_2) - A - (k - \{m(N)\}) + 2\log\log N + \frac{2(A + k - \{m(N)\})}{C\log N} \\
& - \frac{1}{C}\frac{(A + k - \{m(N)\})^2}{(\log N)^2} + \frac{2}{3C}\frac{(A + k - \{m(N)\})^3}{(\log N)^3} + \frac{C_1 - C_0}{CC_0m} + \frac{K}{m^2} + O\left(\frac{1}{(\log N)^3}\right).
\end{aligned}$$

Now, we let $A = 2\log\log N + B$, then,

$$
\begin{aligned}
L = &\ \log(C_0 p^{-2} q_1 q_2) - 2\log\log N - B - (k - \{m(N)\}) + 2\log\log N \\
&+ \frac{4\log\log N + 2B + 2(k - \{m(N)\})}{C\log N} - \frac{1}{C}\frac{A^2 + 2A(k-\{m(N)\}) + (k-\{m(N)\})^2}{(\log N)^2} \\
&+ \frac{2}{3C}\left(\frac{A^3}{(\log N)^3} + \frac{3A^2(k-\{m(N)\})}{(\log N)^3} + \frac{3A(k-\{m(N)\})^2}{(\log N)^3} + \frac{(k-\{m(N)\})^3}{(\log N)^3}\right) \\
&+ \frac{C_1 - C_0}{CC_0 m} + \frac{K}{m^2} + O\left(\frac{1}{(\log N)^3}\right) \\
= &\ \log(C_0 p^{-2} q_1 q_2) - B - (k - \{m(N)\}) + \frac{4\log\log N}{C\log N} + \frac{2B}{C\log N} + \frac{2(k-\{m(N)\})}{C\log N} \\
&- \frac{1}{C}\frac{A^2}{(\log N)^2} - \frac{2}{C}\frac{A(k-\{m(N)\})}{(\log N)^2} - \frac{1}{C}\frac{(k-\{m(N)\})^2}{(\log N)^2} + \frac{2}{3C}\frac{A^3}{(\log N)^3} \\
&+ \frac{2A^2(k-\{m(N)\})}{C(\log N)^3} + \frac{2A(k-\{m(N)\})^2}{C(\log N)^3} + \frac{C_1 - C_0}{CC_0 m} + \frac{K}{m^2} + O\left(\frac{1}{(\log N)^3}\right).
\end{aligned}
$$

Letting $B = \frac{4\log\log N}{C\log N} + D$, then

$$
\begin{aligned}
L = &\ \log(C_0 p^{-2} q_1 q_2) - D - (k - \{m(N)\}) + \frac{8\log\log N}{C^2(\log N)^2} + \frac{2D}{C\log N} + \frac{2(k-\{m(N)\})}{C\log N} \\
&- \frac{1}{C}\frac{(2\log\log N + B)^2}{(\log N)^2} - \frac{2}{C}\frac{(2\log\log N + B)(k-\{m(N)\})}{(\log N)^2} - \frac{1}{C}\frac{(k-\{m(N)\})^2}{(\log N)^2} \\
&+ \frac{2}{3C}\frac{(2\log\log N + B)^3}{(\log N)^3} + \frac{2(2\log\log N + B)^2(k-\{m(N)\})}{C(\log N)^3} \\
&+ \frac{2(2\log\log N + B)(k-\{m(N)\})^2}{C(\log N)^3} + \frac{C_1 - C_0}{CC_0 m} + \frac{K}{m^2} + O\left(\frac{1}{(\log N)^3}\right).
\end{aligned}
$$

We shall now use the Taylor expansion of the function $\frac{1}{x}$ as

$$
\frac{1}{x_0 + x} = \frac{1}{x_0} - \frac{x}{x_0^2} + \frac{x^2}{x_0^3} - \frac{x^3}{\tilde{x}^4},
$$

where $\tilde{x}$ is between x_0 and $x_0 + x$. Since $m = \log N + A + k - \{m(N)\}$, so

$$
\frac{1}{m} = \frac{1}{\log N} - \frac{A + k - \{m(N)\}}{(\log N)^2} + \frac{(A + k - \{m(N)\})^2}{(\log N)^3} + O\left(\frac{1}{(\log N)^3}\right).
$$

and

$$
\begin{aligned}
\frac{1}{m^2} &= \frac{1}{(\log N)^2 + 2\log N(A + k - \{m(N)\}) + (A + k - \{m(N)\})^2} = \\
&= \frac{1}{(\log N)^2} - \frac{2\log N(A + k - \{m(N)\})}{(\log N)^4} + O\left(\frac{1}{(\log N)^3}\right).
\end{aligned}
$$

Now let

$$D = \frac{C_1 - C_0}{CC_0}\frac{1}{\log N} + E.$$

Then,

$$\begin{aligned}
L = {} & \log(C_0 p^{-2} q_1 q_2) - \frac{C_1 - C_0}{CC_0}\frac{1}{\log N} - E - (k - \{m(N)\}) + \frac{8 \log\log N}{C^2 (\log N)^2} \\
& + \frac{2(C_1 - C_0)}{C^2 C_0}\frac{1}{(\log N)^2} + \frac{2E}{C \log N} + \frac{2(k - \{m(N)\}}{C \log N} - \frac{4}{C}\frac{(\log\log N)^2}{(\log N)^2} \\
& - \frac{4B}{C}\frac{\log\log N}{(\log N)^2} - \frac{4}{C}\frac{(\log\log N)(k - \{m(N)\})}{(\log N)^2} - \frac{2B}{C}\frac{(k - \{m(N)\})}{(\log N)^2} \\
& - \frac{1}{C}\frac{(k - \{m(N)\})^2}{(\log N)^2} + \frac{16}{3C}\frac{(\log\log N)^3}{(\log N)^3} + \frac{8B}{C}\frac{(\log\log N)^2}{(\log N)^3} \\
& + \frac{8}{C}\frac{(\log\log N)^2 (k - \{m(N)\})}{(\log N)^3} + \frac{4}{C}\frac{(\log\log N)(k - \{m(N)\})^2}{(\log N)^3} \\
& + \frac{C_1 - C_0}{CC_0}\left(\frac{1}{\log N} - \frac{A + k - \{m(N)\}}{(\log N)^2} + \frac{(2\log\log N + k - \{m(N)\})^2}{(\log N)^3}\right) \\
& + K\left(\frac{1}{(\log N)^2} - \frac{2(A + k - \{m(N)\})}{(\log N)^3}\right) + O\left(\frac{1}{(\log N)^3}\right).
\end{aligned}$$

So we obtain that

$$\begin{aligned}
L = {} & \log(C_0 p^{-2} q_1 q_2) - E - (k - \{m(N)\}) + \frac{8 \log\log N}{C^2 (\log N)^2} + \frac{2(C_1 - C_0)}{C^2 C_0}\frac{1}{(\log N)^2} \\
& + \frac{2E}{C \log N} + \frac{2(k - \{m(N)\})}{C \log N} - \frac{4}{C}\frac{(\log\log N)^2}{(\log N)^2} - \frac{1}{C^2}\frac{(4\log\log N)^2}{(\log N)^3} \\
& - \frac{C_1 - C_0}{C^2 C_0}\frac{4 \log\log N}{(\log N)^3} - \frac{4}{C}\frac{(\log\log N)(k - \{m(N)\})}{(\log N)^2} \\
& - \frac{8}{C^2}\frac{(\log\log N)(k - \{m(N)\})}{(\log N)^3} - \frac{1}{C}\frac{(k - \{m(N)\})^2}{(\log N)^2} + \frac{16}{3C}\frac{(\log\log N)^3}{(\log N)^3} \\
& + \frac{8}{C}\frac{(\log\log N)^2 (k - \{m(N)\})}{(\log N)^3} + \frac{4}{C}\frac{\log\log N (k - \{m(N)\})^2}{(\log N)^3} \\
& + \frac{C_1 - C_0}{CC_0}\left(-\frac{2\log\log N}{(\log N)^2} - \frac{4 \log\log N}{C (\log N)^3} - \frac{k - \{m(N)\}}{(\log N)^2} + \frac{4(\log\log N)^2}{(\log N)^3} + \right. \\
& \left. + \frac{4 \log\log N (k - \{m(N)\})}{(\log N)^3}\right) + \frac{K}{(\log N)^2} - \frac{4K \log\log N}{(\log N)^3} + O\left(\frac{1}{(\log N)^3}\right).
\end{aligned}$$

Now if we let

$$
\begin{aligned}
E = & \frac{8}{C^2}\frac{\log\log N}{(\log N)^2} - \frac{4}{C}\frac{(\log\log N)^2}{(\log N)^2} - \frac{2(C_1-C_0)}{CC_0}\frac{\log\log N}{(\log N)^2} + \frac{2(C_1-C_0)}{C^2C_0}\frac{1}{(\log N)^2} \\
& -\frac{16}{C^2}\frac{(\log\log N)^2}{(\log N)^3} - \frac{4(C_1-C_0)}{C^2C_0}\frac{\log\log N}{(\log N)^3} + \frac{16}{3C}\frac{(\log\log N)^3}{(\log N)^3} + \\
& -\frac{4(C_1-C_0)}{C^2C_0}\frac{\log\log N}{(\log N)^3} + \frac{K}{(\log N)^2} - \frac{4\log\log N}{(\log N)^3} + \frac{4(C_1-C_0)}{CC_0}\frac{(\log\log N)^2}{(\log N)^3} + F \\
= & -\frac{4}{C}\frac{(\log\log N)^2}{(\log N)^2} + \left(\frac{8}{C^2} - \frac{2(C_1-C_0)}{CC_0}\right)\frac{\log\log N}{(\log N)^2} + \\
& + \left(\frac{2(C_1-C_0)}{C^2C_0} + K\right)\frac{1}{(\log N)^2} + \frac{16}{3C}\frac{(\log\log N)^3}{(\log N)^3} + \\
& + \left(-\frac{16}{C^2} + \frac{4(C_1-C_0)}{CC_0}\right)\frac{(\log\log N)^2}{(\log N)^3} - \left(4K + \frac{8(C_1-C_0)}{C^2C_0}\right)\frac{\log\log N}{(\log N)^3} + F.
\end{aligned}
$$

and to collect those terms which contain $k - \{m(N)\}$, we introduce the function

$$
\begin{aligned}
H(x) = & -x + \frac{2x}{C\log N} - \frac{4}{C}\frac{\log\log N}{(\log N)^2}x - \frac{2}{C}\frac{4\log\log N}{C(\log N)^3}x - \frac{1}{C}\frac{1}{(\log N)^2}x^2 \\
& + \frac{8(\log\log N)^2}{C(\log N)^3}x + \frac{4(\log\log N)}{C(\log N)^3}x^2 - \frac{C_1-C_0}{CC_0}\frac{1}{(\log N)^2}x \\
& + \frac{4(C_1-C_0)}{CC_0}\frac{\log\log N}{(\log N)^3}x \\
= & -x + \frac{2}{C\log N}x - \frac{4}{C}\frac{\log\log N}{(\log N)^2}x - \frac{C_1-C_0}{CC_0}\frac{1}{(\log N)^2}x \\
& + \left(\frac{4(C_1-C_0)}{CC_0} - \frac{8}{C^2}\right)\frac{\log\log N}{(\log N)^3}x + \frac{8}{C}\frac{(\log\log N)^2}{(\log N)^3}x \\
& - \frac{1}{C}\frac{1}{(\log N)^2}x^2 + \frac{4}{C}\frac{\log\log N}{(\log N)^3}x^2.
\end{aligned}
$$

Inserting these expressions, we obtain

$$
\begin{aligned}
L = & \log(C_0p^{-2}q_1q_2) + H(k - \{m(N)\}) - F + \frac{2E}{C\log N} + O\left(\frac{1}{(\log N)^3}\right) \\
= & \log(C_0p^{-2}q_1q_2) + H(k - \{m(N)\}) - F + \frac{16}{C^3}\frac{\log\log N}{(\log N)^3} - \frac{8}{C^2}\frac{(\log\log N)^2}{(\log N)^3} \\
& - \frac{4(C_1-C_0)}{C^2C_0}\frac{\log\log N}{(\log N)^3} + O\left(\frac{1}{(\log N)^3}\right).
\end{aligned}
$$

Now, choosing

$$
F = \frac{16}{C^3}\frac{\log\log N}{(\log N)^3} - \frac{8}{C^2}\frac{(\log\log N)^2}{(\log N)^3} - \frac{4(C_1-C_0)}{C^2C_0}\frac{\log\log N}{(\log N)^3},
$$

We have

$$L = \log(C_0 p^{-2} q_1 q_2) + H(k - \{m(N)\}) + O\left(\frac{1}{(\log N)^3}\right).$$

So we obtain that for

$$\begin{aligned}
\mathbb{P}(\mu(N) - [m(N)] < k) =& e^{-l}\left(1 + O\left(\frac{(\log N)^3}{N}\right)\right) \\
=& e^{-(1/p)^L}\left(1 + O\left(\frac{(\log N)^3}{N}\right)\right) \\
=& e^{-(1/p)^{\log(C_0 p^{-2} q_1 q_2) + H(k - \{m(N)\}) + O\left(\frac{1}{(\log N)^3}\right)}}\left(1 + O\left(\frac{(\log N)^3}{N}\right)\right) \\
=& \exp\left(-p^{-\left(\log(C_0 p^{-2} q_1 q_2) + H(k - \{m(N)\})\right)}\right)\left(1 + O\left(\frac{1}{(\log N)^3}\right)\right)
\end{aligned}$$

□

Hence the proof.

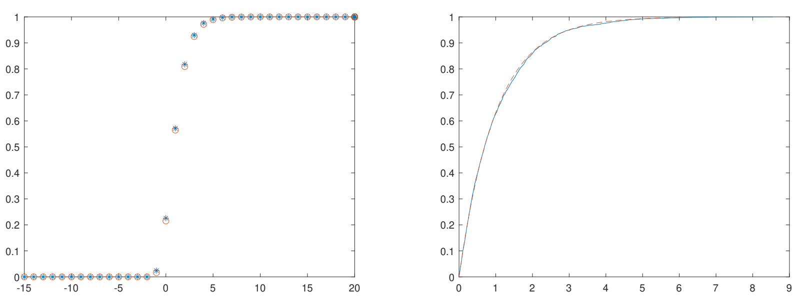
3.4 Simulation results

Analysing the beginning of the proof of Theorem 3.3.1, we can see that the lemma A.0.2 of **Csáki et al. (1987)** offers good approximation if p is small, but it does not offer good approximation if p is close to 1. However, our simulation study shows that the approximation for the longest run is very good for small values of p, but it is still appropriate if p is close to 1.

We performed several computer simulations for certain fixed values of p, q_1 and q_2. Below, N denotes the length of the sequence generated by us and s denotes the number of repetitions on the N-length sequences.

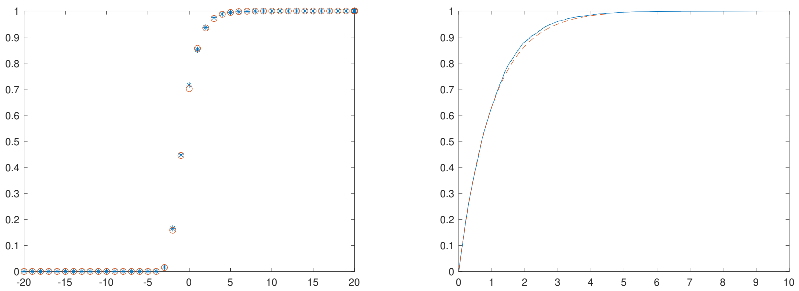
Figures 3.1 - 3.8 present the results of the simulations. The left hand side part of each figure shows the empirical distribution of the longest at most two-type contaminated run and its approximation suggested by Theorem 3.3.1. Asterisk (i.e. $*$) denotes the result of the simulation, i.e. the empirical distribution of the longest at most two-type contaminated run and circle ($\circ$) denotes approximation offered by Theorem 3.3.1.

The right hand side of each figure shows the first hitting time of the m-length at most two-type contaminated run. Solid line shows the result of the simulation for the distribution function and dashed line shows the distribution function $1 - e^{-x}$ suggested by our Theorem 3.1.1.



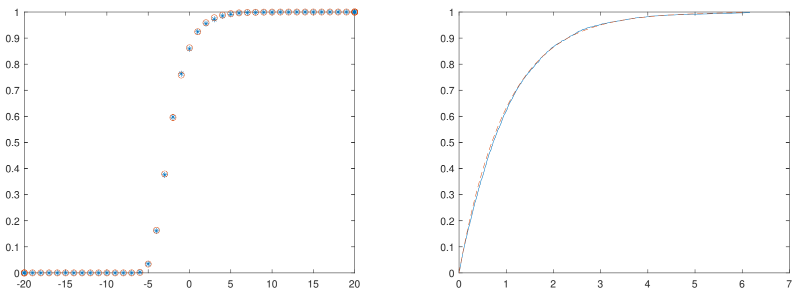
(a) Longest run (b) First hitting time, $m = 16$

Figure 3.1: Longest at most two-type contaminated run and the first hitting time when $p = 1/3$, $q_1 = 1/3$, $q_2 = 1/3$, $N = 3 \times 10^6$, $s = 3000$



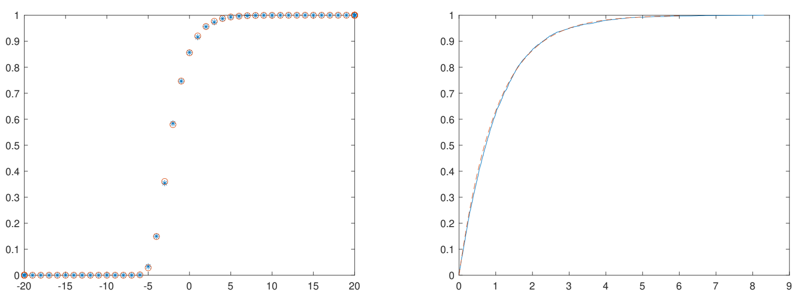
(a) Longest run (b) First hitting time, $m = 19$

Figure 3.2: Longest at most two-type contaminated run and the first hitting time when $p = 0.4$, $q_1 = 0.3$, $q_2 = 0.3$, $N = 3 \times 10^6$, $s = 3000$



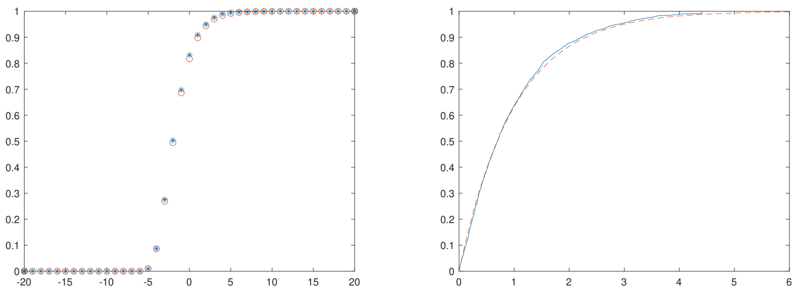
(a) Longest run (b) First hitting time, $m = 25$

Figure 3.3: Longest at most two-type contaminated run and the first hitting time when $p = 0.5$, $q_1 = 0.4$, $q_2 = 0.1$, $N = 4 \times 10^6$, $s = 3000$



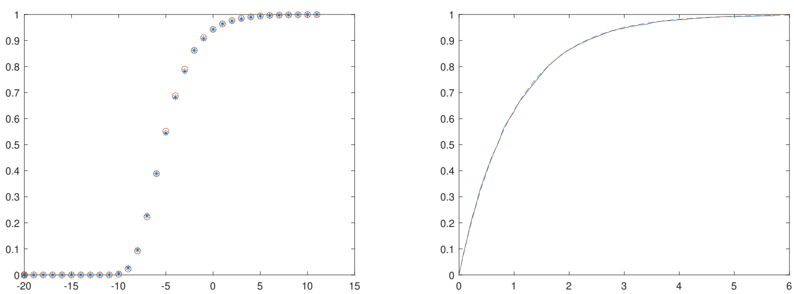
(a) Longest run (b) First hitting time, $m = 23$

Figure 3.4: Longest at most two-type contaminated run and the first hitting time when $p = 0.5$, $q_1 = 0.3$, $q_2 = 0.2$, $N = 3 \times 10^6$, $s = 3000$



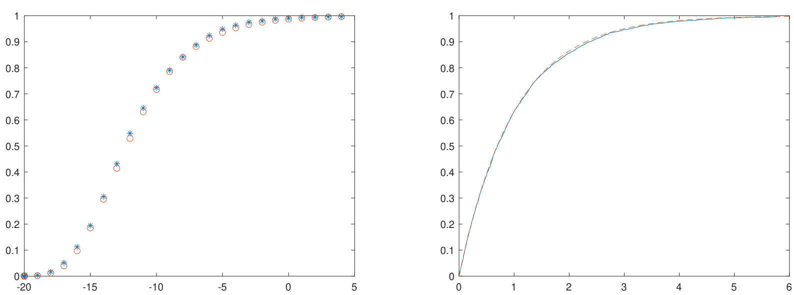
(a) Longest run (b) First hitting time, $m = 25$

Figure 3.5: Longest at most two-type contaminated run and the first hitting time when $p = 0.5$, $q_1 = 0.25$, $q_2 = 0.25$, $N = 2 \times 10^6$, $s = 2000$



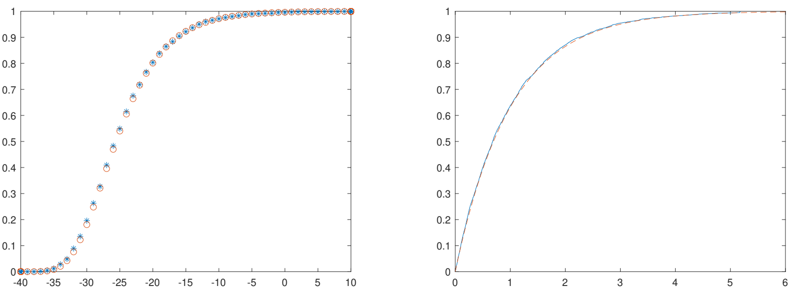
(a) Longest run (b) First hitting time, $m = 34$

Figure 3.6: Longest at most two-type contaminated run and the first hitting time when $p = 0.6$, $q_1 = 0.2$, $q_2 = 0.2$, $N = 4 \times 10^6$, $s = 3000$



(a) Longest run (b) First hitting time, $m = 47$

Figure 3.7: Longest at most two-type contaminated run and the first hitting time when $p = 0.7$, $q_1 = 0.2$, $q_2 = 0.1$, $N = 4 \times 10^6$, $s = 3000$



(a) Longest run (b) First hitting time, $m = 72$

Figure 3.8: Longest at most two-type contaminated run and the first hitting time when $p = 0.8$, $q_1 = 0.1$, $q_2 = 0.1$, $N = 3 \times 10^6$, $s = 3000$

Chapter 4

Summary

In this section, we summarize the most important results of this dissertation. We mention some of the most fascinating lemmas, propositions and theorems based on our research which consist of three published papers [**Fazekas and Suja** (**2021**), **Fazekas, Fazekas, and Suja** (**2024**) and **Fazekas, Fazekas, and Suja** (**2023**)] .

The introduction contain several historical facts on the limit theorem in probability theory and its application to the case of coin tossing. In particular, the study of success runs in Bernoulli trials has received indubitable attention of several researchers due to its inherent theoretical interest and intriguing applications. The problem of the length of the longest pure head run for n Bernoulli random variables was first raised by T. Varga in his classroom experiment and the findings herald overwhelming research interest, variations and extensions to other situations.

The results of **Erdős and Rényi** (**1970**) and **Földes** (**1979**) had immense influence on the trajectory of our study. A lot of insights were drawn from the main Lemma . **Csáki et al.** (**1987**) which provided a good approximation to the probabilities which offered limiting distribution of the random variable τ_m which is the first occurrence time of the event of interest.

In Chapter 1, we defined a T-contaminated run of heads and study the limiting distributions of their numbers together with the first hitting time and the asymptotic behaviour of the length of the longest T-contaminated head runs. More emphasis was devoted to approximation of the numbers of contaminated runs to both Poisson and compound Poisson limit laws.

In Chapter 2, we also dealt with T-contaminated head runs but more emphasis was now shifted to the asymptotic distribution of the length of the longest T-contaminated head runs. We investigated the rate of convergence to an accompanying distribution and also obtained results for the first hitting time for the same.

In Chapter 3, we defined a two type contaminated run and studied the limiting distribution of the first hitting time and the accompanying distribution of the longest at most two type contaminated runs with trinary outcomes. Our approach mirrored the one used in Chapter 2.

At the end of the dissertation, possible further research based on the results obtained is given a long with the appendix which contains the main Lemma, non-stationary finite form of **Csáki et al. (1987)** where other than providing the elegant proof, we precisely fixed the condition of the lemma.

Limit theorems of T-contaminated run of heads

Now we begin with the problem setting for our research. Consider the classical coin tossing experiment. Let $p \in (0,1)$ be the probability of heads and $q = 1-p$ the probability of tails. Here, p is fixed while we toss a coin N times independently. We write 1 when the result is head and 0 when the result is tail. Therefore we consider independent identically distributed random variables $X_1, X_2, \ldots, X_N$ with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = q$, $i = 1, 2, \ldots, N$. Let $T \geq 0$ be fixed integer.

4.0.1 Number of precisely T-contaminated run of heads

Let $\tilde{\xi} = \tilde{\xi}^T(n, N)$ denote the number of those precisely T-contaminated n-length runs of heads for which the proceeding element is a tail. More precisely let

$$\tilde{\eta}_i = \tilde{\eta_i}^T(n) = \begin{cases} 1, & \text{if there are precisely } T \text{ 0 values among} \\ & X_i, \ldots, X_{i+n-1} \text{ and } X_{i-1} = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

Here we let X_0 be defined as $X_0 = 0$ and let

$$\tilde{\xi} = \tilde{\xi}^T(n, N) = \sum_{i=1}^{N-n+1} \tilde{\eta_i}^T(n). \tag{4.2}$$

Our main condition in the first chapter is the following. If we let $p \in (0,1)$ be fixed and T be a fixed non-negative integer. Now if $N \to \infty$ and $n \to \infty$ so that

$$\frac{N q^{T+1} p^{n-T} n^T}{T!} \to \lambda > 0, \tag{4.3}$$

where if λ is fixed, then we remark that above condition implies that $N/n \to \infty$. Now we intend to show that the distribution of $\tilde{\xi}$ converges to the λ parameter Poisson distribution.

Theorem. *Let T be fixed. Let $N \to \infty$ and $n \to \infty$ so that condition (1.3) is satisfied. Then*

$$\lim_{N\to\infty} \mathbb{P}(\tilde{\xi}^T(n,N) = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, \ldots$$

We let $l_m = N - n + 1$ and $Y_i = \tilde{\eta}_i$ $i = 1, 2, \ldots, l_m$ and checked fulfilment of the conditions of Proposition below .

Proposition. *(See **Sevast'yanov (1972)**.)*
Let ${Y_i}^{(m)}$, $i = 1, 2, \ldots, l_m$, $m = 1, 2, \ldots$, be a triangular array of Bernoulli random variables, i.e. the values of ${Y_i}^{(m)}$ are 0 or 1. Let

$$\mathbb{Z}_m = {Y_1}^{(m)} + {Y_2}^{(m)} + \cdots + {Y_{lm}}^{(m)}, m = 1, 2, \ldots \tag{4.4}$$

be the row sums. Let

$$b^{(m)}_{i_1,i_2,\ldots,i_r} = \mathbb{P}({Y_{i_1}}^{(m)} = {Y_{i_2}}^{(m)} = \cdots = {Y_{i_r}}^{(m)} = 1), \tag{4.5}$$

Where $(i_1, i_2, \ldots, i_r)$ denotes an r dimensional vector such that integers $i_1, i_2, \ldots, i_r$ are pairwise different with $1 \le i_t \le l_m$, $t = 1, 2, \ldots, r$, $r = 1, 2, \ldots$.
Assume that for each $r = 2, 3, \ldots$, $m = 1, 2, \ldots$ there exists an exceptional set $I_r(m)$ consisting of certain vectors $\alpha_r = (i_1, i_2, \ldots, i_r)$ such that the numbers $i_1, i_2, \ldots, i_r$ are pairwise different with $1 \le i_t \le l_m$, $t = 1, 2, \ldots, r$.
In addition, we assume the following that

$$\lim_{m\to\infty} \max_{1\le i\le l_m} {b_i}^{(m)} = 0, \tag{4.6}$$

$$\lim_{m\to\infty} \sum_{i=1}^{l_m} {b_i}^{(m)} = \lambda > 0, \tag{4.7}$$

$$\lim_{m\to\infty} \sum_{\alpha_r \in I_r(m)} b^{(m)}_{i_1,i_2,\ldots,i_r} = 0, \tag{4.8}$$

$$\lim_{m\to\infty} \sum_{\alpha_r \in I_r(m)} b^{(m)}_{i_1} \cdots b^{(m)}_{i_r} = 0, \tag{4.9}$$

and uniformly for all $\alpha_r \notin I_r(m)$

$$\lim_{m\to\infty} \frac{b^{(m)}_{i_1,i_2,\ldots,i_r}}{b^{(m)}_{i_1}\cdots b^{(m)}_{i_r}} = 1. \tag{4.10}$$

Then

$$\lim_{m\to\infty} \mathbb{P}(\mathbb{Z}_m = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, \ldots \tag{4.11}$$

4.0.2 Number of at most T-contaminated runs of heads

Now we turn to the problem of the number of at most T-contaminated runs of heads and let

$$\eta_i = \eta_i^T(n) = \begin{cases} 1, & \text{if there are at most T 0 values among} \\ & X_i, \ldots, X_{i+n-1} \\ 0, & \text{otherwise} \end{cases} \tag{4.12}$$

Now we let

$$\xi = \xi^T(n, N) = \sum_{i=1}^{N-n+1} \eta_i^T(n). \tag{4.13}$$

Therefore ξ was considered as the number of head runs being at most T-contaminated and having length n.

Now we want to prove that the distribution of ξ converges to a compound Poisson distribution in the limit.

Theorem. *Let T be fixed. We let $N \to \infty$ and $n \to \infty$ so that condition (1.3) is satisfied. Then, for the generator functions we have*

$$\lim_{N\to\infty} \mathbb{E}\left(z^{\xi^T(n,N)}\right) = \exp\left[\lambda\left(\frac{qz}{1-pz} - 1\right)\right].$$

Remark. First, we recall the notion of the compound Poisson distribution. The compound Poisson distribution is a probability distribution that arises when counting the number of occurrences of a rare event in a fixed time interval, where the size of each occurrence is a random variable with a probability distribution.

More specifically, suppose we have a Poisson process with rate λ, which is a stochastic process that models the occurrence of rare events over time. For each occurrence, we assume that there is a random variable X_i that represents the size or magnitude of the event, and these random variables are assumed to be independent and identically distributed (i.i.d.). Then, the compound Poisson distribution is the distribution of the sum of these random variables over a fixed time interval.

In our case we need its particular version, that is the so called geometric Poisson distribution.
Let γ have Poisson distribution $\mathbb{P}(\gamma = k) = \lambda^k e^{-\lambda}/k!$, $k = 0, 1, 2, \ldots$. Let $\varrho_1, \varrho_2, \ldots$, be random variables independent of each other and of γ having q parameter geometric random distribution:
$\mathbb{P}(\varrho_i = l) = p^{l-1}q$, $l = 1, 2, \ldots$, $q \in (0, 1)$, $p = 1 - q$.
We let the distribution of ϱ to be the same as that of $\varrho_1 + \ldots + \varrho_k$ when $\gamma = k$. (Here, an empty sum is defined as 0, i.e $\varrho = 0$ when $\gamma = 0$). Then ϱ has generator function

$$\mathbb{E}(z^{\varrho}) = \exp\left[\lambda\left(\frac{qz}{1-pz} - 1\right)\right] \text{ for } \quad |zp| < 1.$$

To give a formal explanation of this fact, let

$$\eta'_i = 1 = \eta'^{T}_i(n) = \begin{cases} \tilde{\eta_i}^T(n).X_{i+n-1}, & \text{if} \quad i > 1 \\ \eta_i^T(n), & \text{if} \quad i = 1. \end{cases} \tag{4.14}$$

And to be more precise, we considered the following representation of $\xi = \xi^T(n, N)$,

$$\xi = \xi^T(n, N) = \sum_{i=1}^{N-n+1} \gamma_i^T(n) = \sum_{i=1}^{N-n+1} \gamma_i, \tag{4.15}$$

where

$$\gamma_i = \gamma_i^T(n) = \eta'_i \left[\min\left\{k > 0 : \text{either} \quad \eta_{i+k} = 0 \quad \text{or} \quad i + k + n - 1 > N\right\}\right].$$

4.0.3 First hitting time of T-contaminated runs of heads

Now, we are going to briefly consider the first hitting time of T-contaminated runs of heads. This is τ, the number of tosses needed in a coin tossing experiment for a T contaminated head run of length n to appear for the very first time i.e its the first observation time when the number of tails among the last n outcomes is at most T.

Let

$$\tau = \tau^T(n) = \min\{N : \xi^T(n, N) > 0\}. \tag{4.16}$$

If $T = 0$, τ is the usual waiting time for a pure head run of length n. We show that the appropriate normalized version of τ has exponential limiting distribution.

Theorem. *Let T be fixed. Then, for any $0 < x < \infty$*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\tau^T(n)n^T}{T!}q^{T+1}p^{n-T} \leq x\right) = 1 - e^{-x}. \tag{4.17}$$

4.0.4 Length of the longest T-contaminated runs of heads

Next, we now consider the Length of the longest T-contaminated runs of heads. We describes the accompanying distribution of $\mu^T(N)$. Let

$$\mu = \mu^T(N) = \max\{n : \xi^T(n, N) > 0\}. \tag{4.18}$$

Considering the result of tossing a coin N times, μ is the length of the longest run of heads containing at most T tails. We offer a two parameter family of distributions to approximate the distribution of μ. By letting B be a fixed positive number, then for any positive x, we have that

$$x = kB + r,$$

where k is integer and r is the residual for which $0 \leq r < B$. Here k and r are uniquely determined. We define $[x]_B$ and $\{x\}_B$ as $[x]_B = kB$ and $\{x\}_B = r$.

Theorem. *Let T be fixed. Let B be a fixed positive number and let S be a fixed number. Then, for any integer k we have*

$$\begin{aligned} &\mathbb{P}(\mu^T(N) - [\log N + T\log(\log N + S\log\log N)]_B < k) = \\ &= \exp\left(-q^{T+1}p^{(k-T-\{\log N+T\log(\log N+S\log\log N)\}_B)}/T!\right) + o(1). \end{aligned} \tag{4.19}$$

Here log *denotes logarithm to base* $1/p$.

We also give a new proof based on the above theorem contrary to the extreme value theory approach where $[x]$ denote the usual integer part of x and {x} is the fractional part.

Remark. The limiting distribution of the length of the longest head run containing T tails is the same as the limiting distribution of the length of the longest head run containing at most T tails. To prove it, let A be the event that the length of the longest head run containing at most T tails is greater than n. Then, $A = B \cup C$ where B is the event that the length of the longest head run containing precisely T tails is greater than n and C is the event that the length of a head run containing less than T tails is greater than n and it is not possible to add some tails to it.

But
$$\mathbb{P}(C) \leq \sum_{i=0}^{T-1} \binom{N}{i} p^{N-i} q^i \leq c p^N N^{T-1} \to 0$$
as $N \to \infty$.

In **Gordon et al.** (**1986**), the original proof was based on extreme value theory, but here we give a new proof using the method of our Theorem 1.4.1. Let $[x]$ denote the usual integer part of x and {x} is the fractional part.

Proposition. *Let $\mu(N)$ denote the length of the longest T-contaminated run of heads during the coin tossing experiment of length N, then*

$$\mathbb{P}(\mu^T(N) - \mu_T(qN) \leq t) = \mathbb{P}\left(\left[\frac{W}{ln(\frac{1}{p})} + \{\mu_T(qN)\}\right] - \{\mu_T(qN)\} \leq t\right) + o(1)$$

for all t, where

$$\mu_T(qN) = \log(qN) + T\log\log(qN) + T\log(q/p) - \log(T!) \tag{4.20}$$

and W has an extreme value distribution $\mathbb{P}(W \leq t) = \exp(-e^{-t})$.

Remark. We emphasize that the above proposition does not offer a limiting law for $\mu^T(N) - \mu_T(qN)$ but it gives a sequence of accompanying laws. The distances of the laws between the two sequences converge to 0 (as $n \to \infty$).

Convergence rate for the longest T-contaminated head run

In this chapter, we describe the background of our problem, rate of convergence for the longest T-contaminated head runs. We consider the previous approximation provided by Theorem 1 of **Gordon et al.** (**1986**) and after some manipulations we state the following.

Proposition. *Let $\mu^T(N)$ denote the length of the longest T-contaminated run of heads during the coin tossing experiment of length N. Let*

$$m_0(N) = \log(qN) + T\log(\log(qN)) + T\log(q/p) - \log(T!), \tag{4.21}$$

where log denotes the logarithm to base 1/p. Let $[m_0(N)]$ denote the integer part *of* $m_0(N)$ *and* $\{m_0(N)\}$ *denote the fractional part of* $m_(N)$*. Then*

$$\mathbb{P}(\mu^T(N) - [m_0(N)] < k) = \exp\Big(-p^{k-\{m_0(N)\}}\Big) + o(1). \qquad (4.22)$$

where o(1) denotes a quantity converging to 0 as $N \to \infty$.

However, numerical experiments show that the above offered approximation is quite weak and we therefore aim at improving the result for the quite simple but most important cases of $T = 1$ and $T = 2$.

Let us consider a set of independent and identically distributed random variables, denoted as $X_1, X_2, \cdots, X_N$ with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = q$, $i = 1, 2, \ldots, N$. Let $T \geq 0$ be a fixed non - negative integer.

We study the T-interrupted runs of heads which means that there are T zeros in an m length sequence of ones and zeros. So if we let m be a positive integer $A_n = A_{n,m}$ to denote the occurrence of the event at the n^{th} step, that is, there are precisely T zeros in the block of sequence $X_n, X_{n+1}, \ldots, X_{n+m-1}$. Here, we clarify that the condition $X_{n-1} = 0$ is not assumed. Therefore, $\mathbb{P}(\bar{A}_1\bar{A}_2 \cdots \bar{A}_N)$ is the probability that no event $A_1 = A_{1,m}$ occurred in any of the first N blocks of length m i.e. the waiting time for the T-contaminated run of heads of length m described by A_1 is longer than N.

4.0.5 First hitting time of T-contaminated runs of heads

We let τ_m be the first hitting time of the T-contaminated run of heads having length m and we find the asymptotic distribution of τ_m as $m \to \infty$.

Theorem. *Let* $T = 1$ *or* $T = 2$, $0 < p < 1$*. Let* τ_m *be the first hitting time for the* T*-contaminated run of heads having length* m*. Then, for* $x > 0$,

$$\mathbb{P}(\tau_m \alpha P(A_1) > x) \sim e^{-x} \qquad (4.23)$$

as $m \to \infty$.
Here if $T = 1$, *then* $\alpha = q + \frac{2p^{m-1}-1}{m}$ *and* $P(A_1) = mp^{m-1}q$*. When* $T = 2$*, then* $\alpha = q - \frac{2}{m}$, $P(A_1) = \binom{m}{2}p^{m-2}q^2$.

4.0.6 Length of the longest T-contaminated runs of heads

Now, we turn to the case of the length of the longest T-contaminated run of heads, provide the approximation of its length and the accompanying distribution from which the rate of convergence is evaluated

Theorem. *Let $T = 1$ or $T = 2$, and let $0 < p < 1$ be fixed. Let $\mu^T(N)$ be the length of the longest T-contaminated run of heads during N times of coin tossing. Let*

$$\begin{aligned} m(N) =& \log(qN) + T\log(\log(qN)) + \\ &+ T^2 \frac{\log(\log(qN))}{c\log(qN)} - \frac{T}{cq_0\log(qN)} - \frac{T^3}{2c}\left(\frac{\log(\log(qN))}{\log(qN)}\right)^2 + \\ &+ T^2 \frac{\log(\log(qN))}{cq_0(\log(qN))^2} + T^3 \frac{\log(\log(qN))}{(c\log(qN))^2} + \\ &+ \left(T\log(\frac{q}{p}) - \log(T!)\right)\left(1 + \frac{T}{c\log(qN)} - T^2\frac{\log(\log(qN))}{c(\log(qN))^2}\right), \end{aligned} \tag{4.24}$$

where log *denotes the logarithm to base $1/p$ and $c = ln(1/p)$, where* ln *denotes the natural logarithm to base e. Let $[m(N)]$ denotes the integer part of $m(N)$ while $\{m(N)\}$ denotes the fractional part of $m(N)$, i.e. $\{m(N)\} = m(N) - [m(N)]$. Then,*

$$\mathbb{P}(\mu^T(N) - [m(N)] < k) = e^{-p^{(k-\{m(N)\})\left(1-\frac{T}{c\log(qN)}+T^2\frac{\log(\log(qN))}{c(\log(qN))^2}\right)}}\left(1 + O\left(\frac{1}{(\log N)^2}\right)\right), \tag{4.25}$$

for any integer k, where $f(N) = O(h(N))$ means that $f(N)/h(N)$ is bounded as $N \to \infty$.

Remark. Using our method for $T = 1$ and $T = 2$ and with $m_0(N)$ from the above proposition, we obtain that the rate of convergence to be $O(\log(\log(N))/\log(N))$, that is

$$\mathbb{P}(\mu^T(N) - [m_0(N)] < k) = \exp\left(-p^{k-\{m_0(N)\}}\right)(1 + O(\log(\log(N)/\log(N)))). \tag{4.26}$$

By doing a comparison of the two rates, it can be seen that our Theorem 4.0.6 considerably improves Theorem 1 of **Gordon et al. (1986)** in the cases of $T = 1$ and $T = 2$.

We now present preliminary proofs to some Lemmas in **Csáki et al. (1987)**, Lemma A.0.2 which plays a fundamental role in the proofs of our theorems. We check conditions (SI) - (SIII) of the Lemma, stationary case finite form if $k = m$ and try verifying them with appropriate choices of ε. This made it possible to determine the limiting distribution of the waiting time $\tau_m = \{\text{first } n\text{; such that } A_n \text{ occurs}\}$.

Remark. We first considered condition (SIII) and show that it is true for any T

if m is large enough. We have

$$P(A_1) = \binom{m}{T} p^{m-T} q^T \leq \frac{m^T}{T!} p^{m-T} q^T < \frac{\varepsilon}{m}, \tag{4.27}$$

if

$$m^{T+1} p^m < \varepsilon \left(\frac{p}{q}\right)^T T!,$$

and the last inequality is satisfied for any positive ε if m is large enough.

Remark. Consider condition (SII).

$$\mathbb{P}(A_i|A_1) = P(A_i) = \binom{m}{T} p^{m-T} q^T \leq \frac{m^T}{T!} p^{m-T} q^T, \tag{4.28}$$

if $i > m$ because of independence. So

$$\sum_{i=m+1}^{2m} \mathbb{P}(A_i|A_1) = mP(A_1) = m\binom{m}{T} p^{m-T} q^T \leq m\frac{m^T}{T!} p^{m-T} q^T < \varepsilon,$$

therefore we obtain again condition (SIII) hence condition (SII) is true if m is large enough.

To check condition (SI) of the Lemma A.0.2 , we separately evaluated the joint probabilities $\mathbb{P}(A_1\bar{A}_2\cdots\bar{A}_k)$ taking into account different values of T. First, we fixed $T = 1$.

Lemma. *Condition (SI) of the Lemma A.0.2, stationary case finite form is satisfied for $T = 1$ and $k = m$ in the following form*

$$|\mathbb{P}(\bar{A}_2\cdots\bar{A}_k|A_1) - \alpha| < \varepsilon, \tag{4.29}$$

with $\alpha = q + \frac{2p^{m-1}-1}{m}$.

Lemma. *Condition (SI) of the Lemma A.0.2, stationary case finite form is satisfied for $T = 2$ and $k = m$ in the following form*

$$|\mathbb{P}(\bar{A}_2\bar{A}_3\cdots\bar{A}_m|A_1) - \alpha| < \varepsilon, \tag{4.30}$$

with $\alpha = q - \frac{2}{m} + O(p^m)$ *as as* $m \to \infty$

Limit theorems for runs containing two types of contaminations

In this chapter, we defined and investigated the at most two-type contaminated sequence of runs with trinary trials. Let $X_1, X_2, \ldots, X_N$ be a sequence of independent random variables with three possible outcomes; 0, +1 and -1 labeled as success, failure of type I and failure of type II.

$\mathbb{P}(X_i = 0) = p$, $\mathbb{P}(X_i = +1) = q_1$ and $\mathbb{P}(X_i = -1) = q_2$ where $p + q_1 + q_2 = 1$ and $p > 0, \quad q_1 > 0, \quad q_2 > 0$.

An m length sequence is called a pure run if it contains only 0 values. It is called a one-type contaminated run if it contains precisely one non-zero element either a +1 or a -1. On the other hand, it is called a two-type contaminated run if contains precisely one +1, and one -1 while the rest of the elements are 0's.

A run is called at most two-type contaminated if it is either pure, or one-type contaminated, or two-type contaminated. So for an arbitrary fixed m, let $A_n = A_{n,m}$ denote the occurrence of the event at the n^{th} step ,that is, there are is at most a two-type contaminated run in the sequence $X_n, X_{n+1}, \ldots X_{n+m-1}$ and $\bar{A_n}$ be its non-occurrence.

We see that

$$P(A_1) = p^m + m(1-p)p^{m-1} + m(m-1)p^{m-2}q_1q_2 \tag{4.31}$$

In what follows, we shall use the notation

$$\alpha = \frac{C_0 + \frac{1}{m}c_1 + \frac{1}{m(m-1)}C_2}{1 + \frac{p(1-p)}{(m-1)q_1q_2} + \frac{p^2}{m(m-1)q_1q_2}}, \tag{4.32}$$

where;
$C_0 = (q_1 + q_2)$, $C_1 = \frac{p(q_1^2+q_2^2)}{q_1q_2} - 1$, $C_2 = \frac{(q_1^2+q_2^2)p^2}{q_1q_2(p-1)} + \frac{p}{p-1} + \frac{2(2p+1)q_1q_2}{(p-1)^3}$.

4.0.7 First hitting time of the at most two-type contaminated run

Let τ_m be the first hitting time of the at most two-type contaminated run having length m. We shall be interested in finding the limiting distribution of τ_m as $m \to \infty$ for the case of a sequence containing at most two types of contamination but no two of the same type.

Theorem. *Let $\mathbb{P}(X_i = 0) = p$, $\mathbb{P}(X_i = +1) = q_1$ and $\mathbb{P}(X_i = -1) = q_2$ be probabilities of success, failure of type I and failure of type II, respectively where $p + q_1 + q_2 = 1$ and $p > 0, \quad q_1 > 0, \quad q_2 > 0$. Let τ_m be the first hitting time of the at most two-type contaminated run of heads having length m. Then, for $x > 0$,*

$$\mathbb{P}(\tau_m \alpha P(A_1) > x) \sim e^{-x} \tag{4.33}$$

as $m \to \infty$.

We again check the fulfilment of the conditions given in main Lemma A.0.2 of **Csáki et al. (1987)** for the case of $k = m$ (for fixed m) and $0 < p \leq 1$, such that for $\varepsilon > 0$:

Remark. First, we shall consider condition (SIII) and show that it is true for any large enough m.

$$\begin{aligned} P(A_1) =& p^m + m(1-p)p^{m-1} + m(m-1)p^{m-2}q_1q_2 \\ =& m(m-1)p^{m-2}q_1q_2 \left\{ 1 + \frac{p(1-p)}{(m-1)q_1q_2} + \frac{p^2}{m(m-1)q_1q_2} \right\} \leq \frac{\varepsilon}{m}. \end{aligned} \tag{4.34}$$

This inequality is true for any positive ε if m is large enough.
If $m \approx \log N$, then $p^m \approx p^{\log N} = \frac{1}{N}$ and then, $\varepsilon \approx \frac{(\log N)^3}{N}$. (Here, log denotes logarithm to base $\frac{1}{p}$)

Remark. Now, considering condition (SII), if $i > m$, then A_i and A_1 are independent, therefore

$$\sum_{i=m+1}^{2m} \mathbb{P}(A_i|A_1) = mP(A_1) < \varepsilon, \tag{4.35}$$

which gives precisely the previous assumption on satisfaction of condition (SIII).

Lemma. *Condition (SI)is satisfied for $k = m$ in the following form*

$$|\mathbb{P}(\bar{A}_2\bar{A}_3 \cdots \bar{A}_m|A_1) - \alpha| < \varepsilon, \tag{4.36}$$

with

$$\alpha = \frac{C_0 + \frac{1}{m}c_1 + \frac{1}{m(m-1)}C_2}{1 + \frac{p(1-p)}{(m-1)q_1q_2} + \frac{p^2}{m(m-1)q_1q_2}}.$$

4.0.8 Length of the longest at most two-type contaminated run

Let $\mu(N)$ be the length of the longest at most two-type contaminated run in $X_1, X_2, \ldots, X_N$. Then,
$\{\mu(N) < m\}$ is possible if and only if any m length run in $X_1, X_2, \ldots, X_N$ is neither two-type contaminated nor one-type contaminated nor pure.

We need some notation so let

$$K = \frac{2C_0C_2 - C_!^2 - C_0^2}{2CC_0^2},$$

where $C = \ln \frac{1}{p}$ and ln is the logarithm to base e. Let

$$\begin{aligned} m(N) = & \log N + 2\log\log N + \frac{4\log\log N}{C\log N} + \frac{C_1 - C_0}{CC_0}\frac{1}{\log N} - \\ & - \frac{4}{C}\frac{(\log\log N)^2}{(\log N)^2} + \left(\frac{8}{C^2} - \frac{2(C_1 - C_0)}{CC_0}\right)\frac{\log\log N}{(\log N)^2} + \\ & + \left(\frac{2(C_1 - C_0)}{C^2C_0} + K\right)\frac{1}{(\log N)^2} + \frac{16}{3C}\frac{(\log\log N)^3}{(\log N)^3} + \\ & + \left(-\frac{16}{C^2} + \frac{4(C_1 - C_0)}{CC_0}\right)\frac{(\log\log N)^2}{(\log N)^3} - \left(4K + \frac{8(C_1 - C_0)}{C^2C_0}\right)\frac{\log\log N}{(\log N)^3} + \\ & + \frac{16}{C^3}\frac{\log\log N}{(\log N)^3} - \frac{8}{C^2}\frac{(\log\log N)^2}{(\log N)^3} - \frac{4(C_1 - C_0)}{C^2C_0}\frac{\log\log N}{(\log N)^3}, \end{aligned} \tag{4.37}$$

where log denotes the logarithm to base $1/p$. Let $[m(M)]$ denote the integer part of $m(N)$ and let $\{m(N)\} = m(N) - [m(M)]$ denote its fractional part. We introduce the function

$$\begin{aligned} H(x) = & -x + \frac{2}{C\log N}x - \frac{4}{C}\frac{\log\log N}{(\log N)^2}x - \frac{C_1 - C_0}{CC_0}\frac{1}{(\log N)^2}x \\ & + \left(\frac{4(C_1 - C_0)}{CC_0} - \frac{8}{C^2}\right)\frac{\log\log N}{(\log N)^3}x + \frac{8}{C}\frac{(\log\log N)^2}{(\log N)^3}x \\ & - \frac{1}{C}\frac{1}{(\log N)^2}x^2 + \frac{4}{C}\frac{\log\log N}{(\log N)^3}x^2. \end{aligned} \tag{4.38}$$

Theorem. *Let $0 < p < 1$, $q_1 > 0$, $q_2 > 0$ be fixed with $p + q_1 + q_2 = 1$. Let $\mu(N)$ be the length of the longest at most two-type contaminated run in $X_1, X_2, \ldots, X_N$.*

Then for $k > 0$,

$$\mathbb{P}(\mu(N) - [m(N)] < k) = \exp\left(-p^{-\left(\log(C_0 p^{-2} q_1 q_2) + H(k - \{m(N)\})\right)}\right)\left(1 + O\left(\frac{1}{(\log N)^3}\right)\right). \tag{4.39}$$

Future Research

First hitting time, $\tau^T(m)$ and length of the longest contaminated runs, $\mu^T(N)$ are two strongly related random variables that have been in the mainstream of research on nature of randomness for quite sometime. It is established and confirmed through our research that $\tau^T(m)$ has an exponential limit distribution as $m \to \infty$ just like any other first visit type stopping times of Markov processes.

However, $\mu^T(N)$ does not have a limit distribution. The predicted length of the longest $\mu^T(N)$ grows logarithmically i.e. the distribution tends to shift towards larger values at a rate logarithmically related to N.

Though $\{\mu^T(N) - m(N), \quad N \geq 1\}$ is stochastically bounded, it does not have a limit distribution irrespective of whatever additive normalization is applied in place of $m(N)$. Where $m(N)$ is any arbitrary centralizing sequence tending to infinity. So the approximate distribution of the prediction error has always been used to provide the accompanying limiting distribution independent of N.

With the approximations of $m(N)$ having been obtained in the course of our research, we can apply our results to determine almost sure (a.s) extensions of classical weak limit theorems. **Móri (1993)** investigated this class of non-convergent random variables with a. s. logarithmic limits. For clarity on possible application of his work, we state the following theorem;

Theorem 4.0.1 (**Móri (1993)** Theorem 3.1)**.** *Suppose f is a positive, increasing, differentiable function such that $E(\tau(m)) \sim f(\tau(m))$ and the limit*

$$c = \lim_{t\to\infty} (\log f(t))' \quad 0 \leq c \leq \infty$$

exists. Denote $g = f^{-1}$,

Case (i) $c = 0$. Then for every $t \in \mathbb{R}$

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{i=1}^{n} \frac{1}{i} \mathbb{I}\{Z_i - g(i) < t\} = \frac{1}{e} \quad a.s$$

Case (ii) , $0 < c < \infty$. *Then for every* $t \in \mathbb{R}$

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{i=1}^{n} \frac{1}{i} \mathbb{I}\{Z_i - g(i) < t\} = \int_0^1 F(c(t+z))dz \quad a.s$$

where $F(Z) = \exp(-\exp(-Z))$.

Case (iii) , $c = \infty$. *Suppose in addition, that*

$$(\log\log f(t))' \leq \beta(t).$$

where β *is a positive non increasing function,* $\int_0^\infty \beta^2(t)dt < \infty$. *Then,*

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{i=1}^{n} \frac{1}{i} \mathbb{I}\{Z_i - g(i) < t\} = \begin{cases} 0 & \textit{if} \quad t \leq -1 \\ 1+t & \textit{if} \quad -1 < t < \infty \\ 1 & \textit{if} \quad 0 \leq t. \end{cases}$$

We can consider case (ii) and case (iii) to determine the a.s limit distributions of the first hitting time and the longest T-contaminated head runs where $T = 1$ or 2. We already have the approximate values of $m(N)$ given in chapter 1 theorem 1.4.1 equation 1.29 and also value of $\mu_T(qN)$ in proposition 1.4.1 equation 1.34. In chapter 2, we have values of $m(N)$ in equations 2.1 and 2.1.2.

Appendix A

The main lemma of Csáki, Földes and Komlós

In this section we shall use the same notation as in the paper of **Csáki et al.** (**1987**), so the role of p is not the same as in the previous sections.

Here we shall quote the main lemma of **Csáki et al.** (**1987**) called 'main lemma, non-stationary case, finite form'. We shall write the complete proof of it which is the original proof with some additional explanation. Our goal is to fix precisely the conditions of the lemma. In the lemma in the original publication condition (NII) contains a misprint. Moreover, we want to understand the role of ε. From the original proof one can find that $0 < \varepsilon < \varepsilon_0$, where $\varepsilon_0 = \min\{p/10, 1/42\}$. We will highlight the step in the proof where this condition is needed. After fixing the condition $0 < \varepsilon < \min\{p/10, 1/42\}$, we can freely manipulate with ε during applications of the lemma. So ε can be a constant, but it can depend on m, say.

We also fix the other forms of the above mentioned lemma. In **Csáki et al.** (**1987**) the 'main lemma, stationary case, finite form' there is again an unpleasant misprint, that is the condition $2 \le k \le m$ is not properly written. The case of $k = m$ is excluded in the original paper. However, the lemma is true for $k = m$ and it is the most important case for our applications.

Let $X_1, X_2, \ldots$ be a sequence of independent random variables, and let $\mathcal{F}_{n,m}$ be the σ-algebra generated by the random variables $X_n, X_{n+1}, \ldots, X_{n+m-1}$. We shall study the sequence $A_{n,m}$ of events determined by the above block, where $A_{n,m} \in \mathcal{F}_{n,m}$. We will often write A_n instead of $A_{n,m}$.

We can see that the events A_{n_1} and A_{n_2} are independent if $|n_1 - n_2| \ge m$. So we can say that the sequence A_n is $m-1$ dependent. The dependence of the neighbouring events is given by the constant p in the main lemma. $\mathbb{P}(\bar{A}_1 \cdots \bar{A}_n)$ is the probability that no event A_n of length m has occurred during the first n trials. That is, the waiting time for the occurrence of the event A_1 is longer than n. By verifying the conditions of this lemma, and with an appropriate choice of $k, (k = m)$ and ε as mentioned above, we can get the limit distribution of the waiting time.

The aim of the main lemma is to find a good approximation to the probabilities $\mathbb{P}(\bigcup_{i=1}^{n} A_i) = 1 - \mathbb{P}(\bar{A}_1 \cdots \bar{A}_n)$. It will offer the limiting distribution of the random variable τ_m which is the first occurrence time of the event A_n.

In **Csáki et al. (1987)** it is mentioned that their main lemma is implicitly contained in **Komlós and Tusnády (1975)**. In **Csáki et al. (1987)** three versions are given, finite non-stationary, finite stationary and limiting stationary cases.

For any fixed m, the sequence of events $A_n = A_{n,m}$ is called stationary, if $\mathbb{P}(A_{i_1+d} A_{i_2+d} \cdots A_{i_k+d})$, is independent of d.

We adopt the convention that $\mathbb{P}(A_i A_{i+1} \cdots A_j) = 1$, if $j < i$.

Lemma A.0.1. *(main lemma, non-stationary case, finite form of **Csáki et al. (1987)**.) Let m and N be fixed. Assume that there is a number p, $0 < p \leq 1$, such that the following conditions $(NI) - (NII) - (NIII)$ hold. For some fixed $1 \leq k < m$ and some fixed $0 < \varepsilon < \min\{p/10, 1/42\}$ and for all n, with $m < n \leq N$.*

(NI)

$$|\mathbb{P}(\bar{A}_{n-1} \bar{A}_{n-2} \cdots \bar{A}_{n-k} | A_n) - p| < \varepsilon,$$

(NII)

$$\sum_{n-2m<i<n-k} \mathbb{P}(A_i | A_n) < \varepsilon,$$

(NIII)

$$\max_{1<i<N} \mathbb{P}(A_i) < \varepsilon/m.$$

Then, for $m < n \leq N$,

$$\left| \frac{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1} | A_n)}{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1})} - p \right| < 7\varepsilon,$$

and

$$e^{-(p+10\varepsilon)\lambda - 2\sum_{i=1}^{m} \mathbb{P}(A_i)} < \mathbb{P}(\bar{A}_1 \cdots \bar{A_N}) < e^{-(p-10\varepsilon)\lambda},$$

where

$$\lambda = \sum_{i=m+1}^{N} \mathbb{P}(A_i).$$

Remark A.0.1. Usually, conditions (NII) and $(NIII)$ are more or less easy to prove in applications, but to obtain (NI) can be a hard task.

Proof of Lemma A.0.1. In **Csáki et al. (1987)** it is mentioned that some part of the proof is quoted from **Komlós and Tusnády (1975)** and can be considered as an earlier version of the combinatorial Lemma of **Erdős and Lovász (1975)**.

We divide the interval $[1, n]$ into parts of length m. So let $n = lm + r$, where $0 \le r < m$. Introduce notation

$$E_i = \bar{A}_1 \bar{A}_2 \cdots \bar{A}_{n-mi} \qquad (i = 0, 1, \dots, l).$$

First we deal with the case $l \ge 3$. In the following formulae first we omit some events $\bar{A}_i$ from the intersection of events, then we apply independence of $\bar{A}_1 \cdots \bar{A}_{n-2m}$ and $\bar{A}_{n-k}\bar{A}_{n-k+1} \cdots \bar{A}_{n-1} A_n$ because $m + 1 > k$, so we have

$$\frac{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1} | A_n)}{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1})} \le \frac{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-2m} \bar{A}_{n-k} \bar{A}_{n-k+1} \cdots \bar{A}_{n-1} | A_n)}{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1})} \le \tag{A.1}$$

$$\le \frac{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-2m})}{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_n)} \mathbb{P}(\bar{A}_{n-k} \cdots \bar{A}_{n-1} | A_n) < \frac{\mathbb{P}(E_2)}{\mathbb{P}(E_0)} (p + \varepsilon),$$

where in the last step we applied condition (NI).

Now we want to obtain the opposite inequality. First we use similar considerations as in the proof of $\mathbb{P}(\bar{B}\bar{C}) \ge 1 - [\mathbb{P}(B) + \mathbb{P}(C)]$, then apply independence as $m + 1 > k$, then simply omit some events from the intersection of events, so we get

$$\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1} | A_n) \ge \mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-2m} \bar{A}_{n-k}, \bar{A}_{n-k+1} \cdots \bar{A}_{n-1} | A_n) - \tag{A.2}$$

$$- \sum_{n-2m<j<n-k} \mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-2m} \bar{A}_{n-k} \bar{A}_{n-k+1} \cdots \bar{A}_{n-1} | A_n) \ge$$

$$\ge \mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-2m}) \mathbb{P}(\bar{A}_{n-k} \cdots \bar{A}_{n-1} | A_n) - \sum_{n-2m<j<n-k} \mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-3m} A_j | A_n). \tag{A.3}$$

Concerning the last term we use independence and (NII) to get

$$\sum_{n-2m<j<n-k} \frac{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-3m} A_j | A_n)}{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1})} \le \sum_{n-2m<j<n-k} \frac{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-3m} A_j A_n)/\mathbb{P}(A_n)}{\mathbb{P}(E_0)} =$$

$$= \sum_{n-2m<j<n-k} \frac{\mathbb{P}(E_3)}{\mathbb{P}(E_0)} \mathbb{P}(A_j|A_n) \leq \frac{\mathbb{P}(E_3)}{\mathbb{P}(E_0)} \varepsilon. \tag{A.4}$$

Now, using (A.3), (A.4) and (NI), we obtain

$$\frac{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1}|A_n)}{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1})} > \frac{\mathbb{P}(E_2)}{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1})}(p-\varepsilon) - \frac{\mathbb{P}(E_3)}{\mathbb{P}(E_0)}(\varepsilon) \geq (p-\varepsilon) - \frac{\mathbb{P}(E_3)}{\mathbb{P}(E_0)}(\varepsilon). \tag{A.5}$$

The next step is to show that $\mathbb{P}(E_2)/P(E_0)$ and $\mathbb{P}(E_3)/\mathbb{P}(E_0)$ are close to 1.

Using some elementary consideration, then $E_{i+1} \subseteq E_{i+2}$, independence and (NIII), we have the inequality

$$\mathbb{P}(E_i) = \mathbb{P}(E_{i+1}) - \mathbb{P}(\bigcup_{n-m(i+1)<j<n-mi} E_{i+j}A_j) \geq \tag{A.6}$$

$$\geq \mathbb{P}(E_{i+1}) - \mathbb{P}(E_{i+2}) \sum_{n-m(i+1)<j\leq n-mi} \mathbb{P}(A_j) \geq \mathbb{P}(E_{i+1}) - \varepsilon \mathbb{P}(E_{i+2}).$$

From (A.6) and using $E_i \subseteq E_{i+1} \subseteq E_{i+2}$ we get

$$\mathbb{P}(E_i|E_{i+1}) \geq 1 - \frac{\varepsilon}{\mathbb{P}(E_{i+1}|E_{i+2})}. \tag{A.7}$$

Now we consider the last term of the sequence E_i, that is E_l. As $E_{l-1} = \bar{A}_1 \cdots \bar{A}_{m+r}$, and using (NIII), we obtain

$$\mathbb{P}(E_{l-1}|E_l) = \frac{\mathbb{P}(E_{l-1})}{\mathbb{P}(E_l)} \geq \mathbb{P}(E_{l-1}) \geq 1 - \sum_{j=1}^{2m} \mathbb{P}(A_j) > 1 - 2\varepsilon \geq \frac{1}{2} \tag{A.8}$$

for $\varepsilon \leq \frac{1}{4}$. Now, using (A.8) and (A.7), we can apply induction on i. As $\varepsilon \leq \frac{1}{4}$, first we obtain the inequality

$$\mathbb{P}(E_i|E_{i+1}) \geq \frac{1}{2}$$

for all i. From here and again using induction, we get

$$\mathbb{P}(E_i|E_{i+1}) \geq 1 - 2\varepsilon \tag{A.9}$$

for all i.

In the following inequalities ε should be small enough, i.e. $\varepsilon < \varepsilon_0 = \frac{1}{42}$. Using (A.9)

$$\frac{\mathbb{P}(E_2)}{\mathbb{P}(E_0)} = \frac{1}{\mathbb{P}(E_1|E_2)} \frac{1}{\mathbb{P}(E_0|E_1)} \leq \frac{1}{(1-2\varepsilon)^2} < 1 + 5\varepsilon. \tag{A.10}$$

Here the last inequality is true because $(1+5x)(1-2x)^2 > 1$ if $0 < x < 1/30$.

The next step requires condition $\varepsilon < \varepsilon_0$ with $\varepsilon_0 = 1/42$. Using (A.9) and that $(1+7x)(1-2x)^3 > 1$ if $0 < x < 1/42$, we obtain

$$\mathbb{P}(E_3)/\mathbb{P}(E_0) \le \frac{1}{(1-2\varepsilon)^3} < p + 7\varepsilon \tag{A.11}$$

if $\varepsilon < \varepsilon_0$. Now from (A.1) and (A.10) we obtain

$$\frac{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1}|A_n)}{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1})} < \frac{\mathbb{P}(E_2)}{\mathbb{P}(E_0)}(p+\varepsilon) < (1+5\varepsilon)(p+\varepsilon) < p + 7\varepsilon \tag{A.12}$$

if $\varepsilon < \varepsilon_0$. From (A.5) and (A.11) we have

$$\frac{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1}|A_n)}{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1})} > p - \varepsilon - \frac{\mathbb{P}(E_3)}{\mathbb{P}(E_0)}\varepsilon > p - \varepsilon - (1+7\varepsilon)\varepsilon > p - 3\varepsilon \tag{A.13}$$

if $\varepsilon < \varepsilon_0$. The last two inequalities prove the first statement of the lemma for $l \ge 3$.

Now, turn to the case $l < 3$. We obtain from (NI) and (NIII) that

$$\frac{\mathbb{P}(\bar{A}_1\bar{A}_2 \cdots \bar{A}_{n-1}|A_n)}{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1})} \le \frac{\mathbb{P}(\bar{A}_1\bar{A}_2 \cdots \bar{A}_{n-1}|A_n)}{1 - \mathbb{P}(A_1 \bigcup \cdots \bigcup A_{n-1})} < \frac{p+\varepsilon}{1-3\varepsilon} < p + 7\varepsilon \tag{A.14}$$

if $\varepsilon < \varepsilon_0$. The last inequality is true because $p+x < (1-3x)(p+7x)$ if $0 < x < 1/7$.

Now we find the lower bound. We shall use the convention that $A_j = \emptyset$ for $j \le 0$. Like in inequality (A.2),

$$\mathbb{P}(\bar{A}_1\bar{A}_2 \cdots \bar{A}_{n-1}|A_n) \ge \mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-2m}\bar{A}_{n-k} \cdots \bar{A}_{n-1}|A_n) -$$

$$- \sum_{n-2m<j<n-k} \mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-2m}A_j\bar{A}_{n-k} \cdots \bar{A}_{n-1}|A_n) \ge \mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-2m})(p-\varepsilon) - \varepsilon.$$

In the last step we applied (NI) and (NII). From here, using (NIII), we obtain

$$\frac{\mathbb{P}(\bar{A}_1 A_2 \cdots \bar{A}_{n-1}|A_n)}{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1})} \ge p - \varepsilon - \frac{\varepsilon}{\mathbb{P}(\bar{A}_1 \cdots \bar{A}_{n-1})} \ge p - \varepsilon - \frac{\varepsilon}{1-3\varepsilon} > p - 3\varepsilon \tag{A.15}$$

if $\varepsilon < \varepsilon_0$. So we obtained the first statement when $l < 3$.

To obtain the second statement of the lemma, consider the identity

$$\begin{aligned}\mathbb{P}(\bar{A_1}\cdots\bar{A_N}) =&\mathbb{P}(\bar{A_1}\cdots\bar{A_m})\prod_{n-m+1}^{N}\mathbb{P}(\bar{A}_n|\bar{A}_1\cdots\bar{A}_{n-1})\\ &=\mathbb{P}(\bar{A_1}\cdots\bar{A_m})\prod_{n-m+1}^{N}(1-\mathbb{P}(A_n|\bar{A}_1\cdots\bar{A}_{n-1})).\end{aligned}$$

So we have

$$\begin{aligned}\mathbb{P}(\bar{A_1}\cdots\bar{A_N}) =&\mathbb{P}(\bar{A_1}\cdots\bar{A_m})\prod_{n-m+1}^{N}\left(1-\frac{\mathbb{P}(\bar{A}_1\cdots\bar{A}_{n-1}A_n)}{\mathbb{P}(\bar{A}_1\cdots\bar{A}_{n-1})}\right)\\ &=\mathbb{P}(\bar{A_1}\cdots\bar{A_m})\prod_{n-m+1}^{N}\left(1-\frac{\mathbb{P}(\bar{A}_1\cdots\bar{A}_{n-1}|A_n)\mathbb{P}(A_n)}{\mathbb{P}(\bar{A}_1\cdots\bar{A}_{n-1})}\right).\end{aligned} \tag{A.16}$$

Therefore, using (A.13), (A.15) and in the last step applying inequality $1-x\le e^{-x}$, we get

$$\mathbb{P}(\bar{A_1}\cdots\bar{A_N})\le\prod_{n-m+1}^{N}(1-(p-3\varepsilon)\mathbb{P}(A_n))\le e^{-\sum_{n=m+1}^{N}(p-3\varepsilon)\mathbb{P}(A_n)}.$$

To obtain the lower bound, we shall apply inequality

$$1-x\ge e^{-x-x^2},\quad 0\le x\le\frac{1}{3}.$$

Using (A.16), (A.12), (A.14) and (NIII), we get for $\varepsilon<\varepsilon_0$ that

$$\begin{aligned}\mathbb{P}(\bar{A_1}\cdots\bar{A_N})\ge&(1-\mathbb{P}(\bigcup_{j=1}^{m}A_j))\prod_{n=m+1}^{N}(1-(p+7\varepsilon)\mathbb{P}(A_n))\\ &\ge(1-\mathbb{P}(\bigcup_{j=1}^{m}A_j))\exp\left\{-\sum_{n=m+1}^{N}(p+7\varepsilon)\mathbb{P}(A_n)+(p+7\varepsilon)^2P^2(A_n)\right\}\\ &\ge(1-\sum_{j=1}^{m}\mathbb{P}(A_j))\exp\left\{-\sum_{n=m+1}^{N}(p+7\varepsilon)\mathbb{P}(A_n)+(p+7\varepsilon)^2\varepsilon\mathbb{P}(A_n)\right\}\\ &\ge\exp\left\{-\sum_{j=1}^{m}\mathbb{P}(A_j)-(\sum_{j=1}^{m}\mathbb{P}(A_j))^2\right\}\exp\left\{-\sum_{n=m+1}^{N}\mathbb{P}(A_n)(p+10\varepsilon)\right\}\\ &\ge\exp\left\{-2\sum_{j=1}^{m}\mathbb{P}(A_j)-(p+10\varepsilon)\lambda\right\}.\end{aligned}$$

□

Lemma A.0.2. *(main lemma, stationary case, finite form of* ***Csáki et al. (1987)****.) Let m be fixed. Assume that A_n is stationary. Assume that there is a fixed number p, $0 < p \leq 1$, such that the following three conditions hold for some fixed k with $2 \leq k \leq m$, and fixed ε with $0 < \varepsilon < \min\{p/10, 1/42\}$*

(SI)

$$|\mathbb{P}(\bar{A}_2 \cdots \bar{A}_k | A_1) - p| < \varepsilon,$$

(SII)

$$\sum_{k+1 \leq i \leq 2m} \mathbb{P}(A_i | A_1) < \varepsilon,$$

(SIII)

$$\mathbb{P}(A_1) < \varepsilon/m.$$

Then, for all $N > 1$,

$$\left| \frac{\mathbb{P}(\bar{A}_2 \cdots \bar{A}_N | A_1)}{\mathbb{P}(\bar{A}_2 \cdots \bar{A}_N)} - p \right| < 7\varepsilon$$

and

$$e^{-(p+10\varepsilon)N\mathbb{P}(A_1) - 2m\mathbb{P}(A_1)} < \mathbb{P}(\bar{A}_1 \cdots \bar{A}_N) < e^{-(p-10\varepsilon)N\mathbb{P}(A_1) + 2m\mathbb{P}(A_1)}. \qquad \text{(A.17)}$$

Proof of Lemma A.0.2. We can obtain the proof from Lemma A.0.1. First we reverse the 'time' n in Lemma A.0.1. Then we apply some shift of the time to obtain the result of Lemma A.0.2. □

Now, we turn to the limit form of the lemma.

Lemma A.0.3. *(main lemma, stationary case, limit form of* ***Csáki et al. (1987)****.) If, for any fixed m, A_n is stationary, and*

(i)

$$p = \lim_{k \to \infty} \lim_{m \to \infty} \mathbb{P}(\bar{A}_2 \cdots \bar{A}_k | A_1) > 0,$$

(ii)

$$\lim_{k \to \infty} \sup_m \sum_{k \leq i \leq 2m} \mathbb{P}(A_i | A_1) = 0,$$

(iii)

$$\lim_{m \to \infty} (m\mathbb{P}(A_1)) = 0,$$

then

$$\lim_{m\to\infty} \frac{\mathbb{P}(\bar{A}_2 \cdots \bar{A}_n | A_1)}{\mathbb{P}(\bar{A}_2 \cdots \bar{A}_n)} = p$$

uniformly in n. *Consequently, if* $n(m)$ *satisfies* $\lim_{m\to\infty} n(m)\mathbb{P}(A_1) = \lambda$, *then (i), (ii), (iii) imply*

$$\lim_{m\to\infty} \mathbb{P}(\bar{A}_1 \bar{A}_2 \cdots \bar{A}_n(m)) = e^{-p\lambda}.$$

Bibliography

R. Arratia, L. Goldstein, and L. Gordon. Poisson approximation and the chen-stein method. *Statistical Science*, pages 403–424, 1990.

R. B. Ash and C. A. Doléans-Dade. *Probability and measure theory.* Academic press, 2000.

A. D. Barbour and O. Chryssaphinou. Compound Poisson approximation: A user's guide. *The Annals of Applied Probability*, 11(3):964–1002, 2001.

A. D. Barbour and M. Månsson. Compound Poisson process approximation. *The Annals of Probability*, 30(3):1492–1537, 2002.

K. Binswanger and P. Embrechts. Longest runs in coin tossing. *Insurance: Mathematics and Economics*, 15(2-3):139–149, 1994.

O. Chryssaphinou and S. Papastavridis. A limit theorem for the number of non-overlapping occurrences of a pattern in a sequence of independent trials. *Journal of applied probability*, 25(2):428–431, 1988.

E. Csáki, A. Földes, and J. Komlós. Limit theorems for Erdős-Rényi type problems. *Studia Sci. Math. Hungar*, 22:321–332, 1987.

P. Deheuvels. On the Erdős-Rényi theorem for random fields and sequences and its relationships with the theory of runs and spacings. *Probability Theory and Related Fields*, 70(1):91–115, 1985.

P. Erdős and L. Lovász. Problems and results on 3-chromatic hypergraphs and some related questions, in "infinite and finite sets"(A. Hajnal et al., eds.). In *Colloq. Math. Soc. J. Bolyai*, volume 11, page 609, 1975.

P. Erdős and A. Rényi. On a new law of large numbers. *Analyse Math.*, 23:103–111, 1970.

P. Erdős and P. Révész. On the length of the longest head-run. *Topics in information theory*, 16:219–228, 1975.

S. Eryilmaz, M. Gong, and M. Xie. Generalized sooner waiting time problems in a sequence of trinary trials. *Statistics & Probability Letters*, 115:70–78, 2016.

I. Fazekas and M. Suja. Limit theorems for contaminated runs of heads. *Annales Univ. Sci*, 52:131–146, 2021.

I. Fazekas, Z. Karácsony, and Z. Libor. Longest runs in coin tossing. Comparison of recursive formulae, asymptotic theorems, computer simulations. *Acta Universitatis Sapientiae. Mathematica*, 2(2):215–228, 2010.

I. Fazekas, B. Fazekas, and M. O. Suja. Limit theorems for runs containing two types of contaminations. Paper with detailed proofs. *arXiv preprint arXiv:2309.11602*, 2023.

I. Fazekas, B. Fazekas, and M. O. Suja. Convergence rate for the longest T-contaminated runs of heads. *Statistics & Probability Letters*, 208:110059, 2024. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2024.110059.

A. Földes. On the limit distribution of the longest head-run (in Hungarian). *Mat. Lapok*, 26(1–2):105–116, 1975.

A. Földes. The limit distribution of the length of the longest head-run. *Periodica Mathematica Hungarica*, 10(4):301–310, 1979.

L. Gordon, M. F. Schilling, and M. S. Waterman. An extreme value theory for long head runs. *Probability Theory and Related Fields*, 72(2):279–287, 1986.

L. Guibas and A. Odlyzko. Long repetitive patterns in random sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 53(3):241–262, 1980.

A. Gut. *Probability: a graduate course*, volume 200. Springer, 2005.

Z. Karácsony and J. Libor. Longest runs in coin tossing. Teaching recursive formulae, asymptotic theorems and computer simulations. 2011.

J. Komlós and G. Tusnády. On sequences of pure heads. *The Annals of Probability*, pages 608–617, 1975.

B. Kopocinski. On the distribution of the longest success-run in Bernoulli trials. *Mathematica Applicanda*, 20(34):3–13, 2016.

M. Koutras. On a waiting time distribution in a sequence of Bernoulli trials. *Annals of the Institute of Statistical Mathematics*, 48:789–806, 1996.

M. Koutras and V. Alexandrou. Sooner waiting time problems in a sequence of trinary trials. *Journal of Applied Probability*, 34(3):593–609, 1997.

T. F. Móri. The a.s limit distribution of the longest head run. *Canadian journal of mathematics*, 45(6):1245–1262, 1993.

M. Muselli. Simple expressions for success run distributions in Bernoulli trials. *Statistics & Probability Letters*, 31(2):121–128, 1996.

M. Muselli. Useful inequalities for the longest run distribution. *Statistics & Probability Letters*, 46(3):239–249, 2000a. ISSN 0167-7152. doi: https://doi.org/10.1016/S0167-7152(99)00108-X.

M. Muselli. New improved bounds for reliability of consecutive-k-out-of-n: F systems. *Journal of applied probability*, 37(4):1164–1170, 2000b.

S. Nguyen, X. Wang, and C. Martin. A note on the number of long runs. *Communications in Information and Systems*, 16(1):59–81, 2016.

S. Y. Novak. Asymptotic expansions in the problem of the longest head run for Markov chain with two states. *Trudy Inst. Math.(Novosibirsk)*, 13:136–147, 1989.

S. Y. Novak. Rate of convergence in the limit theorem for the length of the longest head run. *Siberian Mathematical Journal*, 32(3):444–448, 1991.

S. Y. Novak. Longest runs in a sequence of m-dependent random variables. *Probability theory and related fields*, 91(3-4):269–281, 1992.

S. Y. Novak. On the length of the longest head run. *Statistics & Probability Letters*, 130:111–114, 2017.

S. Y. Novak. On the accuracy of Poisson approximation. *Extremes*, 22(4):729–748, 2019.

E. A. Peköz. A compound Poisson approximation inequality. *Journal of Applied Probability*, 43(1):282–288, 2006.

V. V. Petrov. Limit theorems of probability theory: Sequences of independent random variables. 1995.

A. N. Philippou and F. S. Makri. Longest success runs and fibonacci-type polynomials. *The Fibonacci Quarterly*, 23(4):338–346, 1985.

S. Samarova. On the length of the longest head-run for a Markov chain with two states. *Theory of Probability & Its Applications*, 26(3):498–509, 1982.

M. F. Schilling. The longest run of heads. *The College Mathematics Journal*, 21 (3):196–207, 1990.

M. F. Schilling. The surprising predictability of long runs. *Mathematics Magazine*, 85(2):141–149, 2012.

R. J. Serfling. Some elementary results on Poisson approximation in a sequence of Bernoulli trials. *Siam review*, 20(3):567–579, 1978.

B. A. Sevast'yanov. Limit Poisson law in a scheme of dependent random variables. *Teoriya Veroyatnostei i ee Primeneniya*, 17(4):733–738, 1972.

A. N. Shiryaev. *Probability-1*, volume 95. Springer, 2016.

A. Solov'ev. A combinatorial identity and its application to the problem concerning the first occurrence of a rare event. *Theory of Probability & Its Applications*, 11 (2):276–282, 1966.

J. Túri. Limit theorems for the longest run. In *Annales Mathematicae et Informaticae*, volume 36, pages 133–141. Eszterházy Károly College, Institute of Mathematics and Computer Science Eger, 2008.

Research Conference Participation

1. *Asymptotic theorems for contaminated runs of heads in the coin tossing experiment*, 27^{th} Conference of Young Statistician Meeting (YSM 2023), $29^{th} - 1^{st}$ October 2023, Osijek, Croatia.

2. *Limit theorems for runs containing two types of contamination*, 9^{th} International Conference on Mathematics and Informatics (MATHINFO 2023), September $7^{th} - 8^{th}$, 2023 Tárgu Mureş/Marosvásárhely, Romania.

3. The 20^{th} Conference of the Applied Stochastic Models and Data Analysis International Society (ASMDA 2023) and Demographics2023 Workshop. A Hybrid Conference, $6^{th} - 9^{th}$ June 2023, Heraklion, Crete, Greece.

4. *Convergence rate for the longest T contaminated runs of heads*, 12^{th} International Conference on Applied Informatics (ICAI 2023), March $2^{nd} - 4^{th}$ Eger, Hungary, 2023.

5. Workshop of Writing manuscripts for Official Statistics journals: Guidelines for practitioners and researchers, International Statistical Institute, Online, $23^{rd} - 25^{th}$ February, 2022.

6. *Asymptotic results for contaminated runs of heads*, 13^{th} Joint Conference on Mathematics and Computer Science (MaCS 2020), October $1^{st} - 3^{rd}$, ELTE, Budapest 2020,(Virtual Conference).

UNIVERSITY of DEBRECEN

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/77/2024.PL
Subject: PhD Publication List

Candidate: Michael Ochieng Suja
Doctoral School: Doctoral School of Mathematical and Computational Sciences
MTMT ID: 10083960

List of publications related to the dissertation

Foreign language scientific articles in Hungarian journals (2)

1. Fazekas, I., Fazekas, B., **Suja, M. O.**: Limit theorems for runs containing two types of contaminations.
   Period. Math. Hung. [Accepted by publisher] (-), 1-25, 2024. ISSN: 0031-5303.
   IF: 0.8 (2022)
2. Fazekas, I., **Suja, M. O.**: Limit theorems for contaminated runs of heads.
   Ann. Univ. Sci. Budapest, Sect. Comp. 52, 131-146, 2021. ISSN: 0138-9491.

Foreign language scientific articles in international journals (1)

3. Fazekas, I., Fazekas, B., **Suja, M. O.**: Convergence rate for the longest T-contaminated runs of heads.
   Stat. Probab. Lett. 208, 1-8, 2024. ISSN: 0167-7152.
   DOI: http://dx.doi.org/10.1016/j.spl.2024.110059
   IF: 0.8 (2022)

Total IF of journals (all publications): 1,6
Total IF of journals (publications related to the dissertation): 1,6

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

08 March, 2024

DEBRECENI EGYETEM * Egyetemi és Nemzeti Könyvtár *