Approximation Theorems Related to the Coupon Collector's Problem

Abstract of Ph.D. Thesis

by

Anna Pósfai

Supervisors: Prof. Sándor Csörgő and Prof. Andrew D. Barbour

Doctoral School in Mathematics and Computer Science
University of Szeged
Bolyai Institute

2010

1 Introduction

1.1 The coupon collector's problem

The coupon collector's problem is one of the classical problems of probability theory. The simplest and probably original version of the problem is the following. Suppose that there are n coupons, from which coupons are being collected with replacement. What is the probability that more than t sample trials are needed to collect all n coupons? One of the first discussions of the problem is due to Pólya [14]. It is brought up 7 times in Feller [9]. The problem has numerous variants and generalizations. It is related to urn problems and the study of waiting times of various random phenomena (e.g. [12], [11], [1]).

We are interested in the version of the problem, when a coupon collector samples with replacement a set of $n \geq 2$ distinct coupons so that at each time any one of the n coupons is drawn with the same probability $1/n$. For a fixed integer $m \in \{0, 1, \ldots, n-1\}$, this is repeated until $n-m$ distinct coupons are collected for the first time. Let $W_{n,m}$ denote the number of necessary repetitions to achieve this. Thus the random variable $W_{n,m}$, called the coupon collector's waiting time, can take on the values $n-m, n-m+1, n-m+2, \ldots$, and gives the number of draws necessary to have a collection, for the first time, with only m coupons missing. In particular, $W_{n,0}$ is the waiting time to acquire, for the first time, a complete collection.

The mean and variance of the waiting time are denoted throughout by $\mu_n = \mu_n(m) := \mathbf{E}(W_{n,m})$ and $\sigma_n^2 = \sigma_n^2(m) := \mathbf{Var}(W_{n,m})$.

1.2 Limit theorems in the coupon collector's problem

Different limit theorems have been proved for the asymptotic distribution of $W_{n,m}$, depending on how $m = m(n)$ behaves as $n \to \infty$. Throughout all asymptotic relations are meant as $n \to \infty$.

The first result was proved by Erdős and Rényi [8] for complete collections when $m = 0$ for all $n \in \mathbb{N}$, obtaining a limiting shifted Gumbel extreme value distribution. This result was extended by Baum and Billingsley [6], who examined all relevant sequences of $m = m(n)$. They determined four different limiting distributions:

1. **Degenerate distribution at 0**

$$\text{If } \frac{n-m}{\sqrt{n}} \to 0, \text{ then } W_{n,m} - (n-m) \xrightarrow{\mathcal{D}} 0, \tag{1}$$

   that is the limiting probability measure is concentrated on 0.

2. **Poisson distribution**

$$\text{If } \frac{n-m}{\sqrt{n}} \to \sqrt{2\lambda}, \text{ then } W_{n,m} - (n-m) \xrightarrow{\mathcal{D}} \mathrm{Po}(\lambda), \tag{2}$$

where $\mathrm{Po}(\lambda)$ is the Poisson distribution with parameter λ defined by $\mathrm{Po}(\lambda)\{k\} = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}$, $k = 0, 1, 2, \ldots$.

3. Normal distribution

$$\text{If } \frac{n-m}{\sqrt{n}} \to \infty \text{ and } m \to \infty, \text{ then } \frac{W_{n,m} - \mu_n}{\sigma_n} \xrightarrow{\mathcal{D}} \mathrm{N}(0,1), \tag{3}$$

where $\mathrm{N}(0,1)$ denotes the standard normal distribution, whose probability density function with respect to the Lebesgue measure is $\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$, $x \in \mathbb{R}$.

4. Gumbel-like distribution

$$\text{If } m \equiv \text{constant}, \text{ then } \frac{W_{n,m} - \mu_n}{n} \xrightarrow{\mathcal{D}} \mathrm{Gumbel}(m), \tag{4}$$

where the probability law $\mathrm{Gumbel}(m)$, defined through (5) below, shall be referred to as the Gumbel-like distribution with parameter m.

1.3 Aims of the thesis

One of the aims of this thesis is to refine the limit theorems of Baum and Billingsley. Our basic goal is to approximate the distribution of the coupon collector's appropriately centered and normalized waiting time with well-known measures with high accuracy, and in many cases prove asymptotic expansions for the related probability distribution functions and mass functions. The approximating measures are chosen from five different measure families. Three of them – the Poisson distributions, the normal distributions and the Gumbel-like distributions – are probability measure families whose members occur as limiting laws in the limit theorems of Baum and Billingsley.

The fourth set of measures considered is a certain $\{\pi_{\mu,a} : \mu > 0, a > 0\}$ family of compound Poisson measures. For each $\mu > 0$ and $a > 0$, we define $\pi_{\mu,a}$ to be the probability distribution of $Z_1 + 2Z_2$, where Z_1 and Z_2 are independent random variables defined on a common probability space, $Z_1 \sim \mathrm{Po}(\mu)$ and $Z_2 \sim \mathrm{Po}(a/2)$. By examining the corresponding probability generating function, it is easy to see that $Z_1 + 2Z_2$ does have a compound Poisson distribution, that is, it equals in distribution a random variable of the form $\sum_{k=1}^{N} X_k$, where $N, X_1, X_2, \ldots$ are independent random variables given on a common probability space such that N has Poisson distribution and $X_1, X_2, \ldots$ are identically distributed.

The fifth set of approximating measures we consider is the family of Poisson-Charlier signed measures. For any positive real numbers λ, $\widetilde{a}^{(1)}, \ldots, \widetilde{a}^{(S)}$ and $S \in \mathbb{N}$, the Poisson-Charlier signed measure $\nu = \nu(\lambda, \widetilde{a}^{(1)}, \ldots, \widetilde{a}^{(S)})$ is a signed measure concentrated on the nonnegative integers defined by

$$\nu\{j\} = \mathrm{Po}\{j\}(\lambda)\left(\sum_{r=1}^{S}(-1)^r \widetilde{a}^{(r)} C_r(j,\lambda)\right), \quad j \in \mathbb{N},$$

where $C_r(j,\lambda)$ is the r-th Charlier polynomial ([7] p. 170).

2 Methods to measure the closeness of probability distributions

Let μ and ν be two probability measures on the measurable space $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ denotes the σ-algebra of the Borel sets of the real line. In the thesis, we use the Kolmogorov metric

$$d_{\mathrm{K}}(\mu, \nu) = \sup_{x \in \mathbb{R}} \big|\mu((-\infty, x]) - \nu((-\infty, x])\big|$$

and the total variation distance

$$d_{\mathrm{TV}}(\mu, \nu) = \sup_{B \in \mathcal{B}} \big|\mu(B) - \nu(B)\big|$$

to measure the closeness of the compared distributions.

The main tools used in the proofs of the thesis are characteristic function techniques, Stein's method, couplings and some elementary combinatorial considerations.

3 Gumbel-like approximation

First, we are interested in the the asymptotic behavior of the distribution function

$$F_{n,m}(x) := \mathbf{P}\left(\frac{1}{n}\, W_{n,m} - \sum_{k=m+1}^{n} \frac{1}{k} \le x\right), \quad x \in \mathbb{R},$$

if m is a fixed constant for all n and $n \to \infty$.

The weak convergence of (4) can be formulated as

$$\lim_{n\to\infty} F_{n,m}(x) = F_m(x) := \frac{1}{m!} \int_{-\infty}^{x} \mathrm{e}^{-(m+1)(y+C_m)}\, \mathrm{e}^{-\mathrm{e}^{-(y+C_m)}} \mathrm{d}y, \quad x \in \mathbb{R}, \tag{5}$$

where $C_m := \gamma - \sum_{k=1}^{m} \frac{1}{k}$ and $\gamma = \lim_{n\to\infty} \left(\sum_{k=1}^{n} \frac{1}{k} - \log n\right) = 0,577215\ldots$.

For every m, we give a one-term asymptotic expansion $F_m + G_{n,m}$ that approximates $F_{n,m}$ with the uniform order of $1/n$ such that the explicit sequence of functions $G_{n,m}$ has the uniform order of $(\log n)/n$. In particular, it follows that the uniform rate of convergence in (5) is $(\log n)/n$.

For $n \ge m + 2$, we introduce the basic sequence of functions

$$G_{n,m}(x) = -\frac{1}{2n} \sum_{k=m+1}^{n-1} \frac{1}{k} \int_{-\infty}^{x} [f_m'' \star h_k](u)\, \mathrm{d}u, \quad x \in \mathbb{R},$$

where $f_m(x) := F_m'(x)$ is the density function of the limiting distribution, $h_k(x) = \mathrm{e}^{-kx}$, $x > 0$, is the density function of the exponential distribution with mean $1/k$, and $\star$ stands for convolution. Our main result is

Theorem 3.1.1 *For every fixed $m \in \{0, 1, 2, \ldots\}$,*

$$\sup_{x \in \mathbb{R}} \left| F_{n,m}(x) - [F_m(x) + G_{n,m}(x)] \right| = O\left(\frac{1}{n}\right), \tag{6}$$

and for the functions $G_{n,m}$ there exist a constant $K_m > 0$, a point $x_m \in \mathbb{R}$, a positive function $c_m(\cdot)$ and a threshold function $n_m(\cdot) \in \mathbb{N}$, all depending only on m, such that

$$\sup_{x \in \mathbb{R}} \left| G_{n,m}(x) \right| \leq K_m \frac{\log n}{n}, \quad n \geq m + 2,$$

but

$$\left| G_{n,m}(x) \right| \geq c_m(x) \frac{\log n}{n} \quad \textit{for all} \quad x \in (-\infty, x_m),$$

whenever $n \geq n_m(x)$.

In the thesis we also give an argument that not only proves that the error order in (6) is sharp, but also that no longer asymptotic expansion of $F_{n,m}$ than the one given by (6) can improve the current error order $1/n$.

The results of this chapter were published in [19].

4 Normal approximation

In this chapter we prove an error bound for normal approximation to the coupon collector's standardized waiting time. We introduce the distribution functions

$$F_{n,m}(x) := \mathbf{P}\left(\frac{W_{n,m} - \mu_n}{\sigma_n} \leq x \right), \quad x \in \mathbb{R},$$

and prove

Theorem 4.0.1 *For all $n \geq 3$ and $1 \leq m \leq n - 2$, we have*

$$\sup_{x \in \mathbb{R}} \left| F_{n,m}(x) - \Phi(x) \right| \leq C \frac{n}{m} \frac{1}{\sigma_n},$$

where Φ denotes the standard normal distribution function and $C = 9.257$.

One can check that the bound given by Theorem 4.0.1 goes to 0 iff m goes to infinity along with n, but slowly enough to let the sequence $(n - m)/\sqrt{n}$ tend to infinity as-well, which is in accord with the central limit theorem stated in (3).

The results of this chapter were published in [18].

5 Poisson approximation

5.1 Poisson approximation in a general Poisson limit theorem

In the first section of Chapter 5, we consider Poisson approximation to the distribution of sums of independent nonnegative integer valued random variables in general. We complement the following classical Poisson convergence theorem of Gnedenko and Kolmogorov [10] (p. 132):

Theorem 5.1.1 (Gnedenko, Kolmogorov) *Let $\{Y_{n1}, Y_{n2}, \ldots, Y_{nr_n}\}_{n\in\mathbb{N}}$ be a triangular array of row-wise independent nonnegative integer valued random variables such that*

$$\begin{aligned}
&\min_{1\le k\le r_n} \mathbf{P}(Y_{nk}=0) \to 1, \quad n\to\infty,\\
&\sum_{k=1}^{r_n} \mathbf{P}(Y_{nk}\ge 1) \to \lambda, \quad \lambda>0 \text{ constant}, \quad n\to\infty\\
&\sum_{k=1}^{r_n} \mathbf{P}(Y_{nk}\ge 2) \to 0, \quad n\to\infty.
\end{aligned}$$

Then

$$Y_n := \sum_{k=1}^{r_n} Y_{nk} \xrightarrow{\mathcal{D}} \mathrm{Po}(\lambda), \quad \text{as } n\to\infty.$$

Considering an arbitrary triangular array of row-wise independent nonnegative integer valued random variables, for each n, we approximate the distribution of the n-th row sum with a Poisson distribution whose mean λ_n is defined only in terms of the distributions of the random variables in the n-th row, but we do not assume the existence of moments:

$$\lambda_n = \sum_{k=1}^{r_n} \mathbf{P}(Y_{nk}\ge 1).$$

We give both lower and upper bounds, which have precisely the same form, up to a constant, provided that the means λ_n are bounded away from infinity. We thus refine the obvious approximation of the Y_n-s that the limit theorem suggests.

Theorem 5.1.2 (The upper bound.) *If $\{Y_{n1}, Y_{n2}, \ldots, Y_{nr_n}\}_{n\in\mathbb{N}}$ is a triangular array of row-wise independent nonnegative integer valued random variables, then*

$$d_{\mathrm{TV}}(\mathcal{D}(Y_n), \mathrm{Po}(\lambda_n)) \le \sum_{k=1}^{r_n} \left[\mathbf{P}(Y_{nk}\ge 2) + \mathbf{P}(Y_{nk}\ge 1)^2\right].$$

Theorem 5.1.3 (The lower bound.) *If $\{Y_{n1}, Y_{n2}, \ldots, Y_{nr_n}\}_{n\in\mathbb{N}}$ is a triangular array of row-wise independent nonnegative integer valued random variables such that*

$\min_{1\le k\le r_n} \mathbf{P}(Y_{nk}=0) \ge \frac{3}{4}$ *for all* $n\in\mathbb{N}$, *then*

$$d_{\mathrm{TV}}(\mathcal{D}(Y_n),\mathrm{Po}(\lambda_n)) \ge \frac{1}{10}\left(\prod_{k=1}^{r_n}\mathbf{P}(Y_{nk}=0)\right)\sum_{k=1}^{r_n}\left[\mathbf{P}(Y_{nk}\ge 2)+\mathbf{P}(Y_{nk}\ge 1)^2\right].$$

Theorems 5.1.2 and 5.1.3 together state that the order of the error of our Poisson approximation for the random variables in Theorem 5.1.1 is

$$\sum_{k=1}^{r_n}\left[\mathbf{P}(Y_{nk}\notin\{0,1\})+\mathbf{P}(Y_{nk}\ge 1)^2\right].$$

Barbour and Hall have proved similar results in [3] using Stein's method: they approximate a sum $\sum_{j=1}^n Y_j$ of independent nonnegative integer valued random variables with a Poisson variable that has mean $\sum_{j=1}^n \mathbf{P}(Y_j=1)$ or $\sum_{j=1}^n \mathbf{E}(Y_j)$. (Note that the parameter of our approximating Poisson random variable is between these two values.) Their bounds are expressed differently, and involve second moments of the random variables Y_j. Moreover, their lower bounds have a more complicated form, and would yield no useful information at all in the application to be considered in the next section.

Corollary 5.1.1 *For the rate of convergence in Theorem 5.1.1 we have the upper bound*

$$d_{\mathrm{TV}}(\mathcal{D}(Y_n),\mathrm{Po}(\lambda)) \le \sum_{k=1}^{r_n}\left[\mathbf{P}(Y_{nk}\ge 2)+\mathbf{P}(Y_{nk}\ge 1)^2\right]+\left|\sum_{k=1}^{r_n}\mathbf{P}(Y_{nk}\ge 1)-\lambda\right|.$$

5.2 Coupon collecting with an approximately Poisson distributed waiting time – application of the general results

We begin this section by examining how the coupon collector's problem fits in the framework of the previous section. For the shifted waiting time $\widetilde{W}_{n,m} := W_{n,m}-(n-m)$, one can prove the distributional equality

$$\widetilde{W}_{n,m} \stackrel{\mathcal{D}}{=} \sum_{i=m+1}^{n}\widetilde{X}_{n,i}, \tag{7}$$

where the $\widetilde{X}_{n,i}$ random variables are independent, and $\widetilde{X}_{n,i}+1$ has geometric distributions with success probability i/n, $i\in\{m+1,\dots,n\}$, $n\in\mathbb{N}$. It can be checked that the triangular array $\{\widetilde{X}_{n,m+1},\dots,\widetilde{X}_{n,n}\}_{n\in\mathbb{N}}$ satisfies the conditions of Theorem 5.1.1. Thus we see that the limit theorem in (2) is a special case of the Gnedeno–Kolmogorov theorem. Applying the general results of the previous section to $\widetilde{W}_{n,m}$, we obtain the following.

Corollary 5.2.1 *If* $\{m=m(n)\}_{n\in\mathbb{N}}$ *is a sequence of integers that satisfies the conditions of the Poisson limit theorem in (2), then the error of the approximation of the coupon*

collector's $\widetilde{W}_{n,m}$ waiting time with the Poisson random variable N_{λ_n}, that has mean $\lambda_n = \sum_{i=m+1}^{n}\left(1-\frac{i}{n}\right)$, is of order $\sum_{i=m+1}^{n}\left(1-\frac{i}{n}\right)^2$. In fact, for all n such that $\min_{m+1\le i\le n}\frac{i}{n} \ge \frac{3}{4}$,

$$\frac{1}{5}\left(\prod_{i=m+1}^{n}\frac{i}{n}\right)\sum_{i=m+1}^{n}\left(1-\frac{i}{n}\right)^2 \le d_{\mathrm{TV}}(\mathcal{D}(\widetilde{W}_{n,m}),\mathcal{D}(N_{\lambda_n})) \le 2\sum_{i=m+1}^{n}\left(1-\frac{i}{n}\right)^2.$$

Corollary 5.2.2 *For the rate of convergence in the Poisson limit theorem given in (2), we have the upper bound*

$$d_{\mathrm{TV}}(\mathcal{D}(\widetilde{W}_{n,m}),\mathcal{D}(N_\lambda)) \le 2\sum_{i=m+1}^{n}\left(1-\frac{i}{n}\right)^2 + \left|\sum_{i=m+1}^{n}\left(1-\frac{i}{n}\right)-\lambda\right|.$$

The results of the first two sections of this chapter were published in [17].

5.3 Coupon collecting with an approximately Poisson distributed waiting time – combinatorial approach

In this section we take advantage of the combinatorial structure of the coupon collector's problem. This combinatorial approach yields us a stronger result than the one of Corollary 5.2.1. Namely, we derive the first asymptotic correction of the $\mathbf{P}(\widetilde{W}_{n,m}=k)$, $k=0,1,\ldots$, probabilities to the corresponding Poisson point probabilities. We state the result in Theorem 5.3.1. We note that in principal the method presented in the proof can be extended to determine higher order terms in the asymptotic expansion.

We define

$$\lambda_{n,j} := \sum_{i=m+1}^{n}\left(1-\frac{i}{n}\right)^j, \quad j=1,2,\ldots. \tag{8}$$

It can be proved that if $\{m=m(n)\}_{n\in\mathbb{N}}$ is a sequence of integers that satisfies the conditions of the Poisson limit theorem in (2), then $\lambda_n \to \lambda$ and $\lambda_{n,j}\to 0$, $j=2,3,\ldots$, moreover $\lambda_{n,2} = \frac{(2\lambda_n)^{3/2}}{3\sqrt{n}} + O\left(\frac{1}{n}\right)$.

Theorem 5.3.1 *If $\{m=m(n)\}_{n\in\mathbb{N}}$ is a sequence of nonnegative integers that satisfies satisfies the conditions of the Poisson limit theorem in (2), and λ_n and $\lambda_{n,2}$ are defined as in (8), then*

$$\begin{aligned}
\mathbf{P}(\widetilde{W}_{n,m}=0) &= \mathrm{e}^{-\lambda_n} - \mathrm{e}^{-\lambda_n}\frac{\lambda_{n,2}}{2} + O\left(\frac{1}{n}\right),\\
\mathbf{P}(\widetilde{W}_{n,m}=1) &= \mathrm{e}^{-\lambda_n}\lambda_n - \mathrm{e}^{-\lambda_n}\lambda_n\frac{\lambda_{n,2}}{2} + O\left(\frac{1}{n}\right),\\
\mathbf{P}(\widetilde{W}_{n,m}=k) &= \mathrm{e}^{-\lambda_n}\frac{\lambda_n^k}{k!} + \mathrm{e}^{-\lambda_n}\left(\frac{\lambda_n^{k-2}}{(k-2)!}-\frac{\lambda_n^k}{k!}\right)\frac{\lambda_{n,2}}{2} + O\left(\frac{1}{n}\right), \quad k\ge 2.
\end{aligned}$$

The results of this section are stated in [17], the details are contained in [16].

5.4 Poisson approximation – matching the means

In the final section of Chapter 5, we approximate the coupon collector's shifted waiting time $\widetilde{W}_{n,m} = W_{n,m} - (n-m)$ with another Poisson law, namely with the one that has the same mean as $\widetilde{W}_{n,m}$. One can easily calculate that in the range of parameters n and m for which the Poisson limit theorem of (2) holds true, the error order of this new approximation, given in the theorem below, is $1/n$, which is clearly better than the error order $1/\sqrt{n}$ given by Corollary 5.2.1 or Theorem 5.3.1 in the same case. The proof of Theorem 5.4.1 is based on Stein's method, and heavily uses the fact that the means of the compared probability measures coincide. We note that the argument presented in the proof of Theorem 5.1.2 would not work here.

Theorem 5.4.1 *For the coupon collector's shifted waiting time* $\widetilde{W}_{n,m} = W_{n,m} - (n-m)$ *with* $\lambda'_n = \mathbf{E}(\widetilde{W}_{n,m}) = \sum_{i=m+1}^{n} \left(\frac{n}{i} - 1\right)$, *we have*

$$d_{\mathrm{TV}}(\mathcal{D}(\widetilde{W}_{n,m}), \mathrm{Po}(\lambda'_n)) \le 8 \left(1 \wedge \sqrt{\frac{2}{\mathrm{e}\lambda'_n}}\right) \sum_{i=m+1}^{n} \left(\frac{n-i}{i}\right)^3 .$$

6 Compound Poisson approximation

6.1 An extension of Mineka's coupling inequality

Translated compound Poisson approximation of sums of independent integer valued random variables has been studied in a series of papers. Using Stein's method, [5] and [2] give bounds for the errors of such approximations in total variation distance. Their upper bounds are expressed with the help of the first three moments of the summands $X_1, X_2, \ldots, X_n$ and the critical ingredient $d_{\mathrm{TV}}\left(\mathcal{D}\left(W_n\right), \mathcal{D}\left(W_n+1\right)\right)$, where $W_n = \sum_{j=1}^{n} X_j$.

The expression $d_{\mathrm{TV}}\left(\mathcal{D}\left(W_n\right), \mathcal{D}\left(W_n+1\right)\right)$ is usually bounded by the Mineka coupling [13], which typically yields a bound of order $1/\sqrt{n}$. If the X_j's are roughly similar in magnitude, this is comparable with the order $O(1/\sqrt{\mathbf{Var}W_n})$ expected for the error in the central limit theorem. However, if the distributions of the X_j become progressively more spread out as j increases, then $\mathbf{Var}W_n$ may grow faster than n, and then $1/\sqrt{n}$ is bigger than the ideal order $O(1/\sqrt{\mathbf{Var}W_n})$. In fact, this is the situation in the case when we chose W_n to be the coupon collector's waiting time.

Lemma 6.1.1 *Let* $U_1, U_2, \ldots, U_r$, $r \ge 2$, *be independent identically distributed random variables with discrete uniform distribution on* $\{1, 2, \ldots, 2l-1, 2l\}$ *for some integer* $l \ge 1$.

If $V_r = \sum_{j=1}^{r} U_j$, *then*

$$d_{\mathrm{TV}}(\mathcal{D}(V_r), \mathcal{D}(V_r + 1)) \leq \frac{1}{l\sqrt{r}}.$$

Lemma 6.1.1 improves the Mineka bound in the same setting, which is $1/(\sqrt{2r})$. Through Proposition 6.1.1 we show how the result of the lemma concerning sums of iid uniform random variables can be used to obtain similar results for sums of arbitrary independent integer valued random variables. The idea is to embed the uniform random variables in the ones we want to prove the result for.

Proposition 6.1.1 *If* $X_1, X_2, \ldots, X_n$, $n \geq 2$, *are independent integer valued random variables and* $W = \sum_{j=1}^{n} X_n$, *then*

$$d_{\mathrm{TV}}(\mathcal{D}(W), \mathcal{D}(W + 1)) \leq \frac{4}{l\sqrt{nlp}} + \frac{8d_n}{nlp},$$

where $l \in \{2, 4, 6, \ldots\}$ *and* $p \leq \min\{\mathbf{P}(X_j = k) : k = 1, \ldots, l, j = 1, \ldots, n\}$ *are arbitrary and* $d_n = d_{\mathrm{TV}}(\mathcal{D}(X_n), \mathcal{D}(X_n + 1))$.

We observe that there is no loss of generality in supposing that the l-intervals begin at 1, and that the choice of (p, l) depends very much on the problem. We also make a remark proving that the constants in the upper bound of Proposition 6.1.1 can be improved by refining the method proposed in the proof: one could embed not one, but many uniform random variables in the X_j-s by splitting the whole line into the l-blocks $(\{(m-1)l, \ldots, ml\})_{m \in \mathbb{Z}}$ and defining a uniform variable corresponding to each block, thus one could use potential overlaps from the whole distribution and not just the interval $\{1, \ldots, l\}$, when bounding $d_{\mathrm{TV}}(\mathcal{D}(W), \mathcal{D}(W + 1))$.

6.2 Compound Poisson approximation in the range of the central and Poisson limit theorems

In this section, we approximate the distribution of the appropriately centered coupon collector's waiting time with a compound Poisson measure $\pi_{\mu,a}$ defined in the Introduction. Based on the distributional equality in (7), we apply general results of translated compound Poisson approximation of sums of independent integer valued random variables. With the help of our new coupling, we prove that a translated compound Poisson approximation to the collector's waiting time $W_{n,m}$, with ideal error rate, can be obtained in all ranges of n and m in which a central or Poisson limit theorem can be proved.

Theorem 6.2.1 *For any fixed* $n \geq 2$ *and* $2 \leq m \leq n - 4$, *if*

$$\begin{aligned} \mu &= \sigma_n^2 - 2\langle \sigma_n^2 - \mu_n \rangle, \\ a &= \langle \sigma_n^2 - \mu_n \rangle \quad \textit{and} \\ c &= \lfloor \sigma_n^2 - \mu_n \rfloor, \end{aligned}$$

where $\langle x\rangle$ and $\lfloor x\rfloor$ denote the fractional and integer part of x respectively, then there exists a positive constant C such that

$$d_{\mathrm{TV}}\Big(\mathcal{D}\left(W_{n,m}+c\right),\pi_{\mu,a}\Big)\leq\frac{C}{\sigma_n}\left(\frac{\lfloor\sigma_n^2-\mu_n-(n-m)\rfloor}{\sigma_n^2}+\frac{(n-m)^2}{nm}\right).$$

We can express the bounds of Theorem 6.2.1 more intuitively with the help of some asymptotic formulae for μ_n, σ_n and $a_{n,2}:=\sigma_n^2-\mu_n-(n-m)$:

$$d_{\mathrm{TV}}\Big(\mathcal{D}\Big(W_{n,m}+c\Big),\pi_{\mu,a}\Big)=\begin{cases}O\Big(\frac{1}{\sqrt{m}}\Big), & \text{if } \frac{m}{n}\to 0;\\ O\Big(\frac{1}{\sqrt{n}}\Big), & \text{if } \frac{m}{n}\to c\in(0,1] \text{ and}\\ & \liminf_{n,m\to\infty} a_{n2}>1;\\ O\big(\frac{n-m}{n^{3/2}}\big), & \text{if } \limsup_{n,m\to\infty} a_{n2}<1.\end{cases}$$

Comparing this result with the one of Theorem 4.0.1, we deduce that the same or better order of approximation is obtained with the discrete approximation given in our theorem than with normal approximation, and now with the error measured with respect to the much stronger total variation distance.

We note that, with these parameters, the first two moments of the compared distributions – $\pi_{\mu,a}$ and $\mathcal{D}(W_{n,m}+c)$ – are matched.

The results of the two sections of this chapter were published in [15].

7 Poisson-Charlier expansions

In the final chapter of the thesis, we approximate the coupon collector's shifted waiting time $\widetilde{W}_{n,m}=W_{n,m}-(n-m)$ with Poisson-Charlier signed measures in total variation distance. To do so, we apply a characteristic function technique proposed in [4].

Let $\mu=\mathcal{D}(\widetilde{W}_{n,m}+c)$, where $c=\lfloor\mathbf{Var}\widetilde{W}_{n,m}-\mathbf{E}\widetilde{W}_{n,m}\rfloor$. Introducing the sequences

$$a_{n,j}:=\sum_{k=m+1}^{n}\left(\frac{n-k}{k}\right)^j,\quad j=1,2,\dots,$$

and the polynomials

$$h_R(w)=-\Big(a_{n,2}-\lfloor a_{n,2}\rfloor\Big)w+\Big(a_{n,2}-\lfloor a_{n,2}\rfloor\Big)\frac{w^2}{2}+\sum_{r=3}^{R}\Big(a_{n,r}+(-1)^{r+1}\lfloor a_{n,2}\rfloor\Big)\frac{w^r}{r},$$

and

$$H_R(w)=\begin{cases}\sum_{l=0}^{R}\frac{h_R^l(w)}{l!}, & \text{if } a_{n,2}>1;\\ \sum_{l=0}^{3R-2}\frac{h_R^l(w)}{l!}, & \text{if } a_{n,2}<1,\end{cases}$$

for all $w \in \mathbb{C}$, we approximate the distribution μ with the finite signed measure $\nu_R = \nu_R(\sigma_n^2, \widetilde{a}_{n,m}^{(1)}, \ldots, \widetilde{a}_{n,m}^{(R^2)})$ defined by

$$\nu_R\{j\} = \mathrm{Po}(\sigma_n^2)\{j\} \left(1 + \sum_{r=1}^{S} (-1)^{r+1} \widetilde{a}_{n,m}^{(r)} C_r(j, \sigma_n^2)\right), \quad j \in \mathbb{N},$$

where $S = \deg(H_R(w))$, $\widetilde{a}_{n,m}^{(r)}$ is the coefficient of $(\mathrm{e}^{\mathrm{i}t} - 1)^r$ in $H_R(\mathrm{e}^{\mathrm{i}t} - 1)$, and $C_r(j, \sigma_n^2)$ denotes the r-th Charlier polynomial.

Theorem 7.0.1 *We assume $a_{n,2} > 1$. For an arbitrary integer $R \geq 3$ there exist threshold numbers m_R and n_R and a positive constant C_R depending on R such that if $m \geq m_R$ and $n \geq n_R$, then*

$$\sup_{k \in \mathbb{Z}} |\mu\{k\} - \nu_R\{k\}| \leq C_R \left(\frac{1}{\sqrt{m}}\right)^R, \quad \text{if } m \leq \frac{n}{2} - 1,$$

and

$$\sup_{k \in \mathbb{Z}} |\mu\{k\} - \nu_R\{k\}| \leq C_R \frac{(\sqrt{n})^{R-2}}{(n-m)^{R-1}}, \quad \text{if } m \geq \frac{n}{2}.$$

Theorem 7.0.2 *We assume $a_{n,2} < 1$. For an arbitrary integer $R \geq 3$ there exist a threshold number n_R and a positive constant C_R depending on R such that if $n \geq n_R$, then*

$$\sup_{k \in \mathbb{Z}} |\mu\{k\} - \nu_R\{k\}| \leq C_R \frac{1}{(\sqrt{n})^R}.$$

Before stating the corresponding total variation results, we determine the difference in total variation between the approximating measures for successive values of R:

$$||\nu_{R+1} - \nu_R||_{TV} \leq \begin{cases} C_R \frac{1}{(\sqrt{m})^{R-1}}, & \text{in the case of Th. 7.0.1 when } m \leq \frac{n}{2} - 1; \\ C_R \frac{(\sqrt{n})^{R-3}}{(n-m)^{R-2}}, & \text{in the case of Th. 7.0.1 when } m \geq \frac{n}{2}; \\ C_R \frac{n-m}{(\sqrt{n})^{R+1}}, & \text{in the case of Th. 7.0.2.} \end{cases}$$

Corollary 7.0.1 *For all n and m for which Theorem 7.0.1 is valid we have*

$$d_{\mathrm{TV}}(\mu, \nu_R) \leq C_R \sigma_n \log \sigma_n \left(\frac{1}{\sqrt{m}}\right)^R, \quad \text{if } m \leq \frac{n}{2} - 1,$$

and

$$d_{\mathrm{TV}}(\mu, \nu_R) \leq C_R \frac{(\sqrt{n})^{R-3}}{(n-m)^{R-2}}, \quad \text{if } m \geq \frac{n}{2};$$

and for all n and m for which Theorem 7.0.2 is valid we have

$$d_{\mathrm{TV}}(\mu, \nu_R) \leq C_R \frac{n-m}{(\sqrt{n})^{R+1}},$$

where C_R is a positive constant depending only on R.

Comparing the results of Corollary 7.0.1 with the $||\nu_{R+1} - \nu_R||_{TV}$ bounds, we see that in the small m case, when $m \leq n/2 - 1$, our results are not optimal in the sense that the error order of the approximation with ν_R does not coincide with the order of the total variation norm $||\nu_{R+1} - \nu_R||_{TV}$ of the $(R+1)$-th correction term. It is an interesting open problem whether $d_{\mathrm{TV}}(\mu, \nu_R) \leq C_R/(\sqrt{m})^{R-1}$ can be achieved for $m \leq n/2 - 1$.

We finish by comparing the results of the last chapter with the ones obtained for compound Poisson approximation.

References

[1] Banderier, C. and Dobrow, R. P., A Generalized Cover Time for Random Walks on Graphs, Proceedings of FPSAC'00, 2000.

[2] Barbour, A.D. and Cekanavicius, V, Total variation asymptotics for sums of independent integer random variables, *The Annals of Probability* **30** (2002), 509–545.

[3] Barbour, A.D. and Hall, P., On the rate of Poisson convergence, *Math. Proc. Cam. Phil. Soc.* **95** (1984), 473–480.

[4] Barbour, A. D., Kowalski, E., Nikeghbali, A., *Mod-discrete expansions*, *arXiv:0912.1886v1 [math.PR]*, 2009.

[5] Barbour, A.D. and Xia, A., Poisson Perturbations, *ESAIM Probab. and Statist.* **3** (1999), 131–150.

[6] Baum, L.E. and Billingsley, P., Asymptotic distributions for the coupon collector's problem, *Ann. Math. Statist.* **36** (1965), 1835–1839.

[7] Chihara, T. S., *An introduction to orthogonal polynomials*, Gordon and Breach, New York, 1978.

[8] Erdős, P. and Rényi, A., On a classical problem of probability theory, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **6** (1961), 215–220.

[9] Feller, W., *An Introduction to Probability Theory and its Applications*, John Wiley & Sons, 1968.

[10] Gnedenko, B. V. and Kolmogorov, A. N., *Limit Distributions for Sums of Independent Random Variables*, Addison-Wesley Publishing Company, Cambridge, Mass., 1954.

[11] Gut, A. and Holst, L., On the waiting time in a generalized roulette game, *Statistics & Probability Letters*, **2** (1984), 229–239.

[12] Holst, L., The general birthday problem, Proceedings of the sixth international seminar on Random graphs and probabilistic methods in combinatorics and computer science, John Wiley & Sons, 1995.

[13] Lindvall, T., *Lectures on the Coupling Method*, Dover Publications, 1992.

[14] Pólya, G., Eine Wahrscheinlichkeitsaufgabe zur Kundenwerbung, *Z. Angew. Math. Mech.*, **10** (1930), 96–97.

[15] Pósfai A., An extension of Mineka's coupling inequality, *Electronic Communications in Probability*, **14** (2009), 464–473.

[16] Pósfai, A., A supplement to the paper Poisson approximation in a Poisson limit theorem inspired by coupon collecting, *arXiv:0904.4924 [math.PR]*, 2009.

[17] Pósfai A., Poisson approximation in a Poisson limit theorem inspired by coupon collecting, *Journal of Applied Probability*, **46** (2009), 585–592.

[18] Pósfai, A., Rates of convergence for normal approximation in incomplete coupon collection, *Acta Scientiarum Mathematicarum (Szeged)* **73** (2007), 333–348.

[19] Pósfai, A. and Csörgő, S., Asymptotic approximations for coupon collectors, *Studia Scientiarum Mathematicarum Hungarica*, **46** (2009), 61–96.