Bound on Pólya Approximation via Stein-Chen’s Method and \( w \)-Function

Chanokgan Sahatsathatsana

Department of Science and Mathematics
Kalasin University, Thailand

Abstract

In this paper, we use the Stein-Chen’s method and the \( w \)-functions to give bound in the Pólya approximation of a non negative integer-valued random variable. We give two examples of the Pólya approximation to the distribution of \( X \) concerning the negative Pólya distribution and the negative hypergeometric distribution.

Mathematics Subject Classification: 60F05

Keywords: Pólya approximation, Stein-Chen’s method, \( w \)-functions

1 Introduction

In probability theory and statistics, the Pólya or Pólya-Eggenberger distribution [4] was described by George Pólya, in 1930. The Pólya distribution is naturally and basically presented by an urn model which the objects of real interest (such as atoms, people, cars, etc.) are symbolized as colored balls in an urn or other container. In the basic Pólya urn model, the urn consists of \( x \) white and \( y \) black balls; one ball is drawn randomly from the urn and its color observed; it is then returned in the urn. Moreover, the additional ball with the same color is added to the urn and the selection process is repeatedly done. Questions of interest refers to the urn population evolution and the sequence of colors of the balls that are drawn out. We, consequently, are interested in the following problems:
Suppose that a single urn contain \( m \) red and \( N - m \) black balls. Draw a ball at random, note the color, and return it into the urn together with \( c \) additional balls of the same color. Repeat this way for \( n \) draws. Let \( X \) be the number of red balls taken out in the \( m \) drawings, then the distribution of \( X \) is the well known Polya distribution with parameters \( N, n, m \) and \( c \), denoted by \( \mathcal{P}Y_{(N,n,m,c)} \). The probability function of \( X \) is given by

\[
\varphi_X(x) = \frac{\binom{m}{x} \binom{N-m}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, 2, \ldots, n
\] (1.1)

and the mean and variance of \( X \) are \( \mu = \frac{nm}{N} \) and \( \sigma^2 = \frac{nm(N+cm)(N-r)}{N^2(N+c)} \), respectively.

In this study, we derive a uniform bound for the error on \( \sup_A |\mathcal{L}(X)\{A\} - \mathcal{P}Y_{(N,n,m,c)}\{A\}| \), where \( A \subseteq \mathbb{N} \cup \{0\} \) and \( X \) is a non negative integer-valued random variable. The tools for giving our main results consist of the so-called \( w \)-functions and Stein-Chen’s equation for the Po’lya distribution, which we mention them in Section 2. In Section 3, we use \( w \)-functions and Stein-Chen’s equation for the Po’lya distribution to give bound for the total variation distance between \( \mathcal{L}(X) \) and \( \mathcal{P}Y_{(N,n,m,c)} \). Finally, in the last section, we also give two examples to illustrate applications of the result in approximation the negative Polya distribution and the negative hypergeometric distribution. The following theorem is our main result.

**Theorem 1.1.** Let \( X \) be a non negative integer-valued random variable defined as mentioned above and having corresponding \( w \)-function \( w(X) \). Then the following inequalities hold:

\[
d_{TV}(\mathcal{L}(X)\{A\}, \mathcal{P}Y_{(N,n,m,c)}\{A\}) \leq \lambda \frac{(N - m + cn - cx - c)}{(m + cx)(n - x)} \left\{ \frac{(c\lambda - c + m - cn)}{(N + c)} \right\}
\] (1.2)

where \( d_{TV}(\mathcal{L}(X)\{A\}, \mathcal{P}Y_{(N,n,m,c)}\{A\}) = \sup_A |\mathcal{L}(X)\{A\} - \mathcal{P}Y_{(N,n,m,c)}\{A\}| \)

### 2 The \( w \)-function and Poisson Approximation via Stein-Chen Method

We will prove our main result by using the \( w \)-function associated with Po’lya distribution random variable \( X \) and the Stein-Chen’s method

#### 2.1 The \( w \)-functions.

We will prove the main result by using the \( w \)-function associated with Pólya random variable \( X \) and the Stein identity for geometric distribution. In 1998,
Majserowska [5] adapted the relation of \( w \)-function associated with a non-negative integer-valued random variable \( X \) (Cacoullos and Papathanasiou [2]) to be the recurrence relation in the form of

\[
w(k) = \frac{p_x(k - 1)}{p_x(k)} w(k - 1) + \frac{\mu - k}{\sigma^2} \geq 0, \tag{2.1}\]

where \( k \in S(x) \backslash \{0\}, w(0) = \frac{\mu}{\sigma^2}, S(x) \) is support of \( X, p_x(k) > 0 \) for all \( k \in S(x), \mu \) and \( \sigma^2 \in (0, \infty) \) are mean and variance of \( X \), respectively.

The next relation was stated by Cacoullos and Papathanasiou [2], if a function \( f \) satisfies \( \mathbb{E}|w(X)\Delta f(X)| < \infty, \mathbb{E}|(X - \mu)f(X)| < \infty \), then

\[
\text{Cov}(x, f(x)) = \sigma^2 \mathbb{E}[w(X)\Delta f(X)]. \tag{2.2}\]

where \( \Delta f(x) = f(x + 1) - f(x) \) and \( \mathbb{E}[w(x)] = 1 \)

Using the relation (2.1), we give the form of the \( w \)-function associated with the Pólya random variable \( X \) in the following lemma.

**Lemma 2.1.** Let \( w(x) \) be the \( w \)-function associated with The Pólya random variable \( X \), then we have

\[
w(x) = \frac{(m + cx)(n - x)}{N\sigma^2}, x \in \{0, 1, 2 \ldots, n\} \tag{2.3}\]

where \( \sigma^2 = \frac{mn(N + cm)(N - m)}{N^2(N + c)} \)

**Proof.** Of Lemma 2.1. The recurrence relation of \( w \)-function associated with the random variable \( X \) can be written as

\[
w(k) = w(k - 1) \frac{k(\frac{N - m}{c} + n - k)}{(n - k + 1)(\frac{m}{c} + k - 1)} + \frac{mn}{N\sigma^2} - \frac{k}{\sigma^2} \tag{2.4}\]

where \( w(0) = \frac{mn}{N\sigma^2} \)

In the next step, we shall show that (2.3) holds for every \( k \in \{1, 2, \ldots, m\} \).

From (2.4), we have

\[
w(1) = \frac{(\frac{N - m}{c} + n - 1)}{(\frac{m}{c})} w(0) + \frac{mn}{N\sigma^2} - \frac{1}{\sigma^2} = \frac{(\frac{N - m}{c} + n - 1)}{N\sigma^2} \frac{mn}{N\sigma^2} - \frac{1}{\sigma^2} = \frac{N - m + cn - c}{N\sigma^2} + \frac{mn}{N\sigma^2} - \frac{1}{\sigma^2} = \frac{(m + c)(n - 1)}{N\sigma^2} \]
Assuming that \( w(i - 1) = \frac{(m + ci - 1)(n - i + 1)}{N\sigma^2} \), then we have

\[
w(i) = w(i - 1) \frac{i(N - m + n - i)}{(n - i + 1)(m + i - 1)} + \frac{mn}{N\sigma^2} - \frac{i}{\sigma^2}
\]

\[
= \frac{i(m + ci - 1)(n - i + 1)}{N\sigma^2} \frac{(N - m + n - i)}{(n - i + 1)(m + i - 1)} + \frac{mn}{N\sigma^2} - \frac{i}{\sigma^2}
\]

\[
= \frac{(m + ci)(n - i)}{N\sigma^2}
\]

Therefore, by mathematical induction, (2.3) holds for every. For \( x \in \{0, 1, 2, \ldots, n\} \)

\[\square\]

2.2 Stein-Chen’s method.

For the Stein-Chen’s method, Stein [7] introduced a new powerful technique for the obtaining the rate of convergence to standard normal distribution. Chen [3] adapted and applied idea of normal case to the Poisson setting. The Stein-Chen identity or the Stein identity for Poisson distribution with a parameter \( \lambda \) which, given \( h \), is defined by

\[
\lambda f(x + 1) - xf(x) = h(x) - \mathcal{P}Y_{(N,n,m,c)}(h),
\]

where \( \mathcal{P}Y_{(N,n,m,c)}(h) = \sum_{k=0}^{\infty} h(k) \frac{(m + x - 1)}{x} \frac{(N - m + n - x - 1)}{(n - x)} \frac{(m + n - 1)}{n} \) and \( f, h \) are bounded real-valued function defined on \( \mathbb{N} \cup \{0\} \).

For \( A \subseteq \mathbb{N} \cup \{0\} \), let \( h_A : \mathbb{N} \cup \{0\} \rightarrow \mathbb{R} \) be defined by

\[
h_A(x) = \begin{cases} 
1 & ; x \in A, \\
0 & ; x \notin A.
\end{cases}
\]

The solution \( f = f_A \) of (2.5) can be written as

\[
f_A(x) = \begin{cases} 
\frac{\mathcal{P}Y_{(N,n,m,c)}(h_A \cap C_{x-1}) - \mathcal{P}Y_{(N,n,m,c)}(h_A) \mathcal{P}Y_{(N,n,m,c)}(h_{C_{x-1}})}{(m + x - 1)} \frac{(N - m + n - x - 1)}{(n - x)} \frac{(m + n - 1)}{n} ; x \geq 1, \\
0 ; x = 0
\end{cases}
\]

where

\( C_x = \{0, 1, \ldots, x\} \).

and it follows from Brown and Phillips [1] that

\[
f_A = -f_A^c
\]
where $A^c$ is the complement of $A$, from which it also yields

$$\Delta f_A = -\Delta f_{A^c} \tag{2.8}$$

where $\Delta f_A(x) = f_A(x+1) - f_A(x)$. For $x_0 \in \mathbb{N} \cup \{0\}$ and $A = x_0$, we let $h_{x_0} = h\{x_0\}$, then the solution $f_{x_0} = f\{x_0\}$ of (2.6), we get

$$f_{x_0}(x) = \begin{cases} -\mathcal{P} \mathcal{Y}_{N,n,m,c}(h_{x_0}) \mathcal{P} \mathcal{Y}_{N,n,m,c}(h_{C_{x-1}}) & ; x \leq x_0, \\ x \mathcal{P} \mathcal{Y}_{N,n,m,c}(h_x) \mathcal{P} \mathcal{Y}_{N,n,m,c}(1-h_{C_{x-1}}) & ; x > x_0, \\ 0 & ; x = 0 \end{cases} \tag{2.9}$$

From Malingam and Teerapabolarn [6], we have that

$$\Delta f_A(x) = f_A(x+1) - f_A(x) = \begin{cases} < 0 & ; x \neq x_0, \\ > 0 & ; x = x_0 \end{cases} \tag{2.10}$$

The following lemmas give new bounds for $\Delta f_x$ and $\Delta f_A, A \subseteq \mathbb{N} \cup \{0\}$, which are needed to improve the desired result.

**Lemma 2.2.** Let $x \in \mathbb{N}$, then we have following:

$$\Delta f_x(x) \leq \frac{(N-m+cn-cx-c)}{(m+cx)(n-x)} \tag{2.11}$$

**Proof.** We shall show that $\Delta f_x X \leq \frac{(N-m+cn-cx-c)}{(m+cx)(n-x)}$, by using (2.9). it yields

$$\Delta f_x(x) = \frac{\mathcal{P} \mathcal{Y}_{N,n,m,c}(h_x) \mathcal{P} \mathcal{Y}_{N,n,m,c}(1-h_{C_x})}{(x+1)\mathcal{P} \mathcal{Y}_{N,n,m,c}(h_{x+1})} + \frac{\mathcal{P} \mathcal{Y}_{N,n,m,c}(h_x) \mathcal{P} \mathcal{Y}_{N,n,m,c}(h_{C_{x-1}})}{x\mathcal{P} \mathcal{Y}_{N,n,m,c}(h_x)}$$

$$\leq \frac{(N-m+cn-cx-c)}{(m+cx)(n-x)} \sum_{k=x+1}^{\infty} \left( \frac{x+k-1}{x} \right) \left( \frac{N-r+m-x-1}{m} \right) \left( \frac{x+k-1}{x} \right) \left( \frac{N-r+m-x-1}{m} \right)$$

$$= \frac{(N-m+cn-cx-c)}{(m+cx)(n-x)} \left\{ \sum_{k=x+1}^{\infty} \left( \frac{x+k-1}{x} \right) \left( \frac{N-r+m-x-1}{m} \right) \right\}$$

$$+ \frac{(m+cx)(n-x)}{x(N-m+cn-cx-c)} \sum_{k=0}^{x-1} \left( \frac{x+k-1}{x} \right) \left( \frac{N-r+m-x-1}{m} \right)$$

$$\leq \frac{(N-m+cn-cx-c)}{(m+cx)(n-x)} \left\{ \sum_{k=x+1}^{\infty} \left( \frac{x+k-1}{x} \right) \left( \frac{N-r+m-x-1}{m} \right) \right\}$$

$$= \frac{(N-m+cn-cx-c)}{(m+cx)(n-x)} \tag{2.12}$$
Lemma 2.3. Let $A \subseteq \mathbb{N} \cup \{0\}$ and $x \in \mathbb{N}$, then we have following:

$$\sup_A |\Delta f_x(x)| \leq \frac{(N - m + cn - cx - c)}{(m + cx)(n - x)} \quad (2.13)$$

Proof. From (2.8), (2.10) and lemma 2.2, we have

$$\frac{(N - m + cn - cx - c)}{(m + cx)(n - x)} \geq \Delta f_x(x)$$

$$\geq \sum_{k \in A} \Delta f_x(x)$$

$$= \Delta f_A(x)$$

$$\geq \Delta f_{x^*}(x)$$

$$= -\Delta f_x(x)$$

$$\geq -\frac{(N - m + cn - cx - c)}{(m + cx)(n - x)} \quad (2.14)$$

3 Proof of Main Result

Proof. Of Theorem 1.1. In view of lemma 2.1, for $k \in \{0, 1, 2, \ldots, n\}$, then we have

$$\lambda - \sigma^2 w(k) = \frac{mn}{N} - \sigma^2 \frac{(m + ck)(n - k)}{N \sigma^2}$$

$$= \frac{k(m - cn + ck)}{N}$$

$$\geq 0 \quad (3.1)$$
From lemma 2.3, (2.5) and (3.1), when \( h = h_A \) and \( f = f_A \), then we have

\[
d_{TV}(\mathcal{L}(X)\{A\}, \mathcal{P}(N,n,m,c)\{A\}) = \mathbb{E}|\lambda f(X + 1) - X f(X)|
\]

\[
= |\lambda \mathbb{E}[f(X + 1)] - \mathbb{E}[X f(X)]|
\]

\[
= |\lambda \mathbb{E}[f(X + 1)] - \text{Cov}(X, f(X)) - \mu \mathbb{E}[f(X)]|
\]

\[
= |\lambda \mathbb{E}[\Delta f_x(X)] - \text{Cov}(X, f(X))|
\]

\[
= |\lambda \mathbb{E}[\Delta f_x(X)] - \mathbb{E}[\sigma^2 w(x) \Delta f_x(X)]|
\]

\[
\leq \mathbb{E}[|\lambda - \sigma^2 w(x)| \Delta f_x(X)]
\]

\[
\leq \sup_A |\Delta f_x(X)| \mathbb{E}[|\lambda - \sigma^2 w(x)|]
\]

\[
\leq \frac{(N - m + cn - cx - c)}{(m + cx)(n - x)} \mathbb{E}[\lambda - \sigma^2 w(x)]
\]

\[
= \frac{(N - m + cn - cx - c)}{(m + cx)(n - x)} (\lambda - \sigma^2)
\]

\[
= \lambda \left\{ \frac{N - m + cn - cx - c}{(m + cx)(n - x)} \right\} \frac{c\lambda - c + m - cn}{N + c}
\]

(3.2)

Hence, by substituting these parameters, (1.2) holds. \qed

4 Example

We give two example by applying the result in (3.3) to approximation the negative Pólya distribution and the negative hypergeometric distribution.

Example 4.1. Let \( X \) be the negative Pólya random variable, the probability function of \( X \) is given by

\[
\varphi_X(x) = \binom{n + x - 1}{x} \frac{m(m + c) \ldots (m + (x - 1)c) r(r + c) \ldots (r + (n - 1)c)}{(m + r)(m + r + c) \ldots (m + r + (n + x - 1)c)}
\]

(4.1)

where \( x = 0, 1, \ldots \) and \( \mu = \frac{nm}{r - c} \) and \( \sigma^2 = \frac{nm(nr + r - c)(m + r - c)}{(r - 2c)(r - c)^2} \) are mean and variance of \( X \), respectively. It follows from (2.1) that \( w(x) = \frac{(n + x)(m + cx)}{(r - c)\sigma^2} \). Thus we have

\[
\lambda - \sigma^2 w(x) = \frac{nm}{r - c} - \sigma^2 \frac{(n + x)(m + cx)}{(r - c)\sigma^2}
\]

\[
= -x(cn + m + cx)
\]

\[
\leq 0,
\]

(4.2)
for all $r > 2c$ and $X \geq 0$. Thus, a bound for approximation the negative Pólya distribution by the Pólya distribution with parameters $N, n, m$ and $c$ is as follows:

$$d_{TV}(X, P_Y(N, n, m, c)\{A\}) \leq \frac{\lambda(N - m + cn - cx - c)}{(m + cx)(n - x)} \left\{ \frac{c\lambda + c + cn + m}{(r - 2c)} \right\}$$

(4.3)

where $d_{TV}(X, P_Y(N, n, m, c)\{A\}) = \sup_A |X - P_Y(N, n, m, c)\{A\}|$

**Example 4.2.** Let $X$ be the negative hypergeometric random variable with probability mass function

$$\varrho_X(x) = \frac{(r+x-1)(N-r-x)}{x^n}, x = 0, 1, 2, \ldots, n,$$

(4.4)

where $N, n$ and $r$ are parameters and $r \in \{1, 2, N - n\}$. The mean and variance of $X$ are $\mu = \frac{rn}{N-n+1}$ and $\sigma^2 = \frac{rn(N-n-r+1)(N+1)}{(N-n+1)(N-n+2)}$, respectively. Applying the relation in (2.1), we have $w(x) = \frac{(r+x)(n-x)}{(N-n+1)\sigma^2}$. Thus we have

$$\lambda - \sigma^2 w(x) = \frac{rn}{N-n+1} - \sigma^2 \frac{(r+x)(n-x)}{(N-n+1)\sigma^2}$$

$$= \frac{x(r-n+x)}{N-n+1}$$

$$\geq 0,$$

(4.5)

Thus, a bound for approximation the negative hypergeometric distribution by the Pólya distribution with parameters $N, n, m$ and $c$ is as follows:

$$d_{TV}(X, P_Y(N, n, m, c)\{A\}) \leq \frac{\lambda(N - m + cn - cx - c)}{(m + cx)(n - x)} \left\{ \frac{n - n^2 + \lambda(N + 1)}{n(N - n + 2)} \right\}$$

(4.6)

where $d_{TV}(X, P_Y(N, n, m, c)\{A\}) = \sup_A |X - P_Y(N, n, m, c)\{A\}|$

**Acknowledgements.** The authors would like to thank the referees for their insightful comment.

**References**


Received: September 12, 2017; Published: November 1, 2017