

# On the Geometric Approximation to the Pólya Distribution

Kanint Teerapabolarn

Department of Mathematics, Faculty of Science  
Burapha University, Chonburi 20131, Thailand  
kanint@buu.ac.th

## Abstract

This paper gives a new upper bound on the geometric approximation to the Pólya distribution using the  $w$ -function and the Stein identity.

**Mathematics Subject Classification:** Primary 60F05

**Keywords:** Geometric distribution; geometric approximation; Pólya distribution; Stein identity;  $w$ -function

## 1 Introduction

The Pólya distribution has applications to models of epidemics and to genetics, as well as to other fields, and is typically and simply introduced in terms of an urn model. Consider the random assignment of  $m$  balls into  $d$  compartments such that all partitions have equal probability. Let  $X$  be the number of balls in the first compartment, then the distribution of  $X$  is the special case  $r = 1$  of the Pólya distribution in Phillips and Weinberg [4] on pp. 310, and is given by

$$\mathcal{PY}(d, m, 1) = \left\{ p(k) = \frac{\binom{d+m-k-2}{m-k}}{\binom{d+m-1}{m}}, k = 0, \dots, m \right\},$$

and its mean and variance are  $\mu = \frac{m}{d}$ ,  $\sigma^2 = \frac{m(d+m)(d-1)}{d^2(d+1)}$ , respectively.

Following Phillips and Weinberg [4], as  $d, m \rightarrow \infty$  such that  $\frac{m}{d}$  tends to a constant  $c$ ,  $\mathcal{PY}(d, m, 1)$  converges to the geometric distribution  $Ge(1/(1+c))$ .

Thus it is reasonable to approximate the Pólya distribution by the geometric distribution provided that certain conditions on their parameters are satisfied. To measure how close the Pólya and geometric distributions are, the total variation distance between the two distributions is applied.

In 1999, Brown and Phillips [1] used the Stein's method to give an upper bound on the geometric approximation to the Pólya distribution as follows:

$$d_{TV}(\mathcal{PY}(d, m, 1), Ge(d/(d + m))) \leq \frac{2m(d + 2m)}{(d + 1)d^2}, \quad (1.1)$$

where  $d_{TV}(\mathcal{PY}(d, m, 1), Ge(d/(d + m)))$  is the total variation distance between  $\mathcal{PY}(d, m, 1)$  and  $Ge(d/(d + m))$ . Afterwards, Phillips and Weinberg [4] used the same method to improve the upper bound (1.1) as the following formula

$$d_{TV}(\mathcal{PY}(d, m, 1), Ge(d/(d + m))) \leq \frac{2m(d + 2m)}{d(d + 1)(d + m - 1)}. \quad (1.2)$$

Clearly, the upper bound (1.2) is better than the upper bound (1.1).

In this paper we use a different manner to give a better upper bound for this approximation by applying the method of the Poisson approximation in Majsnerowska [3]. The tools of method consists of the  $w$ -function and the Stein identity for geometric distribution which we apply in Section 2, and we give some numerical examples in the last section.

## 2 The main result

We will prove the main result by using the  $w$ -function associated with the Pólya random variable  $X$  and the Stein identity for geometric distribution. In 1998, Majsnerowska [3] 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

$$w(k + 1) = \frac{p(k)}{p(k + 1)}w(k) - \frac{\mu - (k + 1)}{\sigma^2} \geq 0, \quad k = 0, 1, \dots, \quad (2.1)$$

where  $w(0) = \mu/\sigma^2$  and  $\mu$  and  $\sigma^2$  are mean and variance of  $X$ .

Using the relation (2.1), the  $w$ -function associated with the Pólya random variable can be obtained as the following proposition.

**Proposition 2.1.** *Let  $w(X)$  be the  $w$ -function associated with the Pólya random variable  $X$  and  $p(k) > 0$  for every  $k \in \{0, \dots, m\}$ . Then we have*

$$w(k) = \frac{(k + 1)(m - k)}{\sigma^2 d}, \quad k = 0, \dots, m, \quad (2.2)$$

where  $\sigma^2 = m(d + m)(d - 1)/(d^2(d + 1))$ .

For the Stein identity, we follow Brown and Phillips [1]. Thus, for  $p = 1 - q = d/(d + m)$ , every subset  $A$  of  $\mathbb{N} \cup \{0\}$  and the bounded real valued function  $f = f_A : \mathbb{N} \cup \{0\} \rightarrow R$ , defined as in Brown and Phillips [1] on pp. 410, the Stein identity for geometric case is given by

$$\mathcal{PY}(d, m, 1)(A) - Ge(p)(A) = E[q(1 + X)f(X + 1) - Xf(X)], \tag{2.3}$$

where  $q = m/(d + m)$ . For any subset  $A$  of  $\mathbb{N} \cup \{0\}$ , it follows from Lemma 5 in Brown and Phillips [1] that

$$\sup_{k,A} |\Delta f(k)| = \sup_{k,A} |f(k + 1) - f(k)| \leq 1. \tag{2.4}$$

The theorem below gives an upper bound on the geometric approximation to the Pólya distribution which is our main result.

**Theorem 2.1.** *With the above definitions, for  $A \subseteq \mathbb{N} \cup \{0\}$ , we have*

$$d_{TV}(\mathcal{PY}(d, m, 1), Ge(d/(d + m))) \leq \frac{2m}{d(d + 1)}. \tag{2.5}$$

**Proof.** We shall show that the inequality (2.5) holds. Since

$$\begin{aligned} E[q(1 + X)f(X + 1) - Xf(X)] &= E[qf(X + 1) + qXf(X + 1) - Xf(X)] \\ &= E[qf(X + 1) + qX\Delta f(X) - pXf(X)] \\ &= qE[f(X + 1)] + qE[X\Delta f(X)] - pE[Xf(X)] \\ &= qE[f(X + 1)] + qE[X\Delta f(X)] \\ &\quad - p\{Cov(X, f(X)) + \mu E[f(X)]\} \\ &= qE[f(X + 1)] + qE[X\Delta f(X)] - qE[f(X)] \\ &\quad - pCov(X, f(X)) \end{aligned}$$

and, by Cacoullos and Papathanasiou [2], we have that

$$\begin{aligned} E[q(1 + X)f(X + 1) - Xf(X)] &= qE[\Delta f(X)] + qE[X\Delta f(X)] \\ &\quad - p\sigma^2 E[w(X)\Delta f(X)] \\ &= E\{[(1 + X)q - p\sigma^2 w(X)]\Delta f(X)\}. \end{aligned}$$

Therefore, by (2.3) and Proposition 2.1, we obtain

$$\begin{aligned} d_{TV}(\mathcal{PY}(d, m, 1), Ge(p)) &\leq E|[(1 + X)q - p\sigma^2 w(X)]\Delta f(X)| \\ &\leq \sup_{k,A} |\Delta f(k)| E|(1 + X)q - p\sigma^2 w(X)| \\ &\leq E|(1 + X)q - p\sigma^2 w(X)| \quad (\text{by (2.4)}) \\ &= E[(1 + X)q - p\sigma^2 w(X)] \\ &= (1 + \mu)q - p\sigma^2. \end{aligned}$$

Hence, by substituting these parameters, (2.5) holds.  $\square$

**Remarks.** 1. It should be noted that our upper bound (2.5) is better than the upper bounds (1.1) and (1.2).

2. If  $m/d$  is fixed and  $d$  is large, then the result yields a good geometric approximation.

### 3 Numerical examples

The table below shows some numerical examples of the total variation distance between Pólya and geometric distributions,  $\mathcal{PY}(d, m, 1)$  and  $Ge(d/(d + m))$ , and upper bounds (1.1), (1.2) and (2.5) with the same data as in [1].

**Table 3.1** : Sample values of  $d_{TV}(\mathcal{PY}(N, m, 1, 1), Ge(d/(d + m)))$  with three upper bounds.

$m$	$d/(d + m)$	$d_{TV}$	Upper Bounds		
			(1.1)	(1.2)	(2.5)
1	0.900	0.02111	0.02761	0.02716	0.02222
5	0.500	0.07897	1.00000	0.55556	0.33333
	0.900	0.00353	0.00590	0.00542	0.00483
	0.976	0.00023	0.00026	0.00025	0.00024
20	0.500	0.01714	0.28417	0.14652	0.09524
	0.917	0.00061	0.00097	0.00089	0.00082

From sample values of the three upper bounds in Table 3.1, by comparison between these upper bounds, the upper bound (2.5) is less than the upper bounds (1.1) and (1.2). Hence the upper bound (2.5) is closer to the true value of total variation distance than the upper bounds (1.1) and (1.2). Thus our upper bound gives more accurate than the upper bounds of Brown and Phillips [1] and Phillips and Weinberg [4].

### References

- [1] T. C. Brown, M. J. Phillips, Negative binomial approximation with Stein's method, *Meth. Comp. Appl. Probab.*, **1** (1999), 407-421.
- [2] T. Cacoullos, V. Papathanasiou, Characterization of distributions by variance bounds, *Statist. Probab. Lett.*, **7** (1989), 351-356.
- [3] M. Majsnerowska, A note on Poisson approximation by  $w$ -functions, *Appl. Math.*, **25** (1998), 387-392.
- [4] M. J. Phillips, G. V. Weinberg, Non-uniform bounds for geometric approximation, *Statist. Probab. Lett.*, **49** (2000), 305-311.

**Received: January 23, 2008**