International Journal of Mathematics and
Computer Science, **16**(2021), no. 1, 229–238

# Bounds on Poisson approximation for set of $t$ columns that is not a $t-$covering

**Chanokgan Sahatsathatsana**[1], **Sattra Sahatsathatsana**[2]

[1]Department of Science and Mathematics
Faculty of Science and Health Technology
Kalasin University
Kalasin, Thailand

[2]Department of Foreign Language
Faculty of Liberal Arts
Kalasin University
Kalasin, Thailand

email: chanokgan.na@ksu.ac.th

(Received June 3, 2020, Accepted July 15, 2020)

**Abstract**

A covering array, denoted by $CA(n; k, t, v)$, is a $k \times n$ array with entries from the $v$-letter alphabet $\{0, 1, ..., v-1\}$ with each entry chosen uniformly at random. Covering arrays generalize orthogonal arrays which are classic combinatorial objects that have been studied extensively. Let $W_n$ be the number of set of $t$ columns that is not a $t-$covering. In this work, we give bounds on Poisson approximation of $W_n$ by using the Stein-Chen coupling method.

## 1 Introduction

Software test suites were developed based on the concept of interaction testing. They are very useful for testing software components in an economical way. The suites of this kind may be created using mathematical objects

**Key words and phrases:** Poisson approximation, Covering array, Stein-Chen coupling method.
**AMS (MOS) Subject Classifications**: 60G07
**ISSN** 1814-0432, 2021, http://ijmcs.future-in-tech.net

called covering arrays. Extensive research has been done to reduce the cost of testing while taking into consideration the constraints imposed by the problem. Interaction testing involves testing specific subsets of components exhaustively, where a combinatorial perspective of the problem is to find a covering array with a minimum number of rows. However, many developers have been trying to improve both the efficiency and effectiveness of covering arrays. In this definition, $t$ is often referred to as the coverage strength. As mentioned earlier, covering arrays have been successfully and widely used in many domains such as systematic testing of input parameters [9], software configurations [14], software product lines [4], graphical user interfaces [15], multi-threaded applications [6], and network protocols [12]. Therefore, approaches for computing covering arrays in an efficient and effective mannern are of great practical importance [8].

We consider $k \times n$ arrays whose entries are from the set $\{0, 1, ..., v-1\}$ of size $v$ with each entry chosen uniformly and randomly. The integer $t$ is fixed as $(1 \leq t \leq k)$, any set of $t$ columns is then chosen. A $t$-letter word is made up of the entries (from left to right) across any row of the selected $t$ columns. An array is said to be a $t$-covering if every set of $t$ columns contains, among its rows, each of the $v^t$ possible words of length $t$. The set of covering arrays is denoted by $CA(n; k, t, v)$. There are $\binom{n}{t}v^t$ interactions to cover. This is exactly what a covering array does. Each row can be thought of as an input, or test, which then, depending on the parameters, either outputs a successor or an error. Consequently, we are interested in the following problems: What is the probability that they are all covered in $k$ rows?

We can construct the random variables for solving the problem as follows: Let

$$W_n = \sum_{i=1}^{\binom{n}{t}} X_i,$$

be the total number of set of $t$ columns that is not a $t-$covering, for each $i \in \{0, 1, 2, \ldots, n\}$. We define the indicator random variable $X_i$, as follows:

$$X_i = \begin{cases} 1 & \text{if } i^{th} \text{ set of } t \text{ columns does not form a } t-\text{covering in rows} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore

$$P(X_i = 1) = (1 - \frac{1}{v^t})^k,$$

and, for $n$ sufficiently large, it is logical to approximate the distribution of

$W_n$ by a Poisson distribution with mean

$$\lambda = \mathrm{E}W_n = \binom{n}{t} v^t (1 - \frac{1}{v^t})^k.$$

The purpose of this paper is to give the error estimation on Poisson approximation of the number of set of $t$ columns that is not a $t-$covering by using the Stein-Chen coupling method which is introduced in Section 2. The following theorem is our main result.

**Theorem 1.1.** *Let $W_n$ be the number of set of $t$ columns that is not a $t-$covering. Then we have*

1. $|P(W_n \in A) - Poi_\lambda(A)| \leq C_{\lambda,A} \binom{n}{t} (\frac{v^t-1}{v^{2t}})^k$
2. $|P(W_n \in A) - Poi_\lambda(A)| \leq (1 - e^{-\lambda}) \binom{n}{t} (\frac{v^t-1}{v^{2t}})^k$

*where* $C_{\lambda,A} = \min\left\{1, \lambda, \frac{\triangle(\lambda)}{M_A+1}\right\}$,

$$\triangle(\lambda) = \begin{cases} e^\lambda + \lambda - 1 & \text{if } \lambda^{-1}(e^\lambda - 1) \leq M_A, \\ 2(e^\lambda - 1) & \text{if } \lambda^{-1}(e^\lambda - 1) > M_A, \end{cases}$$

*and*

$$M_A = \begin{cases} \max\{w \mid C_w \subseteq A\} & \text{if } 0 \in A, \\ \min\{w \mid w \in A\} & \text{if } 0 \notin A \end{cases}$$

*when* $C_w = \{0, 1, \ldots, w-1\}$.

# 2 Stein-Chen method and Coupling approach

The Stein-Chen method of Poisson approximation is a powerful tool for computing an error bound when approximating probabilities using the Poisson distribution. In 1972, Stein [11] introduced a new powerful technique to obtain bounds on the distance between two probability distributions.

In 1975, Chen [2] applied Stein's idea to obtain approximation results for the Poisson distribution. As a result, the method has been referred to as the Stein-Chen method.

The central ideal of the Stein-Chen method is the equation

$$I_A(j) - Poi_\lambda(A) = \lambda g_{\lambda,A}(j+1) - j g_{\lambda,A}(j), \tag{2.1}$$

where $\lambda > 0$, $j \in \mathbb{N} \cup \{0\}$, $A \subseteq \mathbb{N} \cup \{0\}$.
Let $I_A : \mathbb{N} \cup \{0\} \to \mathbb{R}$ be defined by

$$I_A(w) = \begin{cases} 1 & ; w \in A, \\ 0 & ; w \notin A. \end{cases}$$

The solution $g_{\lambda,A}$ of (2.1) is of the form

$$g_{\lambda,A}(w) = \begin{cases} (w-1)!\lambda^{-w}e^{\lambda}[\mathcal{P}_\lambda(I_{A\cap C_{w-1}}) - \mathcal{P}_\lambda(I_A)\mathcal{P}_\lambda(I_{C_{w-1}})] & , w \geq 1, \\ 0 & , w = 0, \end{cases}$$

where

$$\mathcal{P}_\lambda(I_A) = e^{-\lambda}\sum_{l=0}^{\infty} I_A(l)\frac{\lambda^l}{l!}$$

and

$$C_{w-1} = \{0, 1, \ldots, w-1\}.$$

By substituting $j$ and $\lambda$ into (2.1) by any integer-valued random variable $W = \sum_{i=1}^{n} X_i$ and $\lambda = \mathrm{E}(W)$, we have

$$P(W_n \in A) - Poi_\lambda A = E(\lambda g_{\lambda,A}(W_n + 1)) - E(W_n g_{\lambda,A}(W_n)). \tag{2.2}$$

In the case where the dependence among the instances of $X_i$ is global, there is an alternative approach to approximate the distribution of $W_n$. This approach is referred to as The Coupling Approach which was first proposed by Barbour [1] in 1982. This approach is particularly useful when it is possible to construct a random variable $W_{n,i}$, for each $i$ on a common probability space with $W_n$ such that $W_{n,i}$ is distributed as $W_n - X_i$ conditional on the event $X_i = 1$.

There has been a number of successful applications of this method; e.g., Barbour ([1] 1982), Janson ([3] 1994), Lange ([5] 2003).

**Theorem 2.1.** *If $W_n$ and $W_{n,i}$ are defined as above, then*

$$|P(W_n \in A) - Poi_\lambda(A)| \leq \| g_{\lambda,A} \| \sum_{i=1}^{n} p_i E|W_n - W_{n,i}|, \tag{2.3}$$

*where* $\| g_{\lambda,A} \| := \sup_w [g_{\lambda,A}(w+1) - g_{\lambda,A}(w)]$.

Many authors would like to determine a bound of $\| g_{\lambda,A} \|$. For $A \subseteq \mathbb{N} \cup \{0\}$, Chen ([2], 1975) proved that

$$\| g_{\lambda,A} \| \leq \min\{1, \lambda^{-1}\}$$

and Janson ([3], 1994) showed that

$$\| g_{\lambda,A} \| \leq \lambda^{-1}(1 - e^{-\lambda}). \tag{2.4}$$

In case of non-uniform bound, Neammanee ([7], 2003) showed that

$$\| g_{\lambda,A} \| \leq \min\left\{\frac{1}{w_0}, \lambda^{-1}\right\}$$

and Teerapabolarn and Neammanee ([13], 2005) gave a bound of $\| g_{\lambda,A} \|$, where $A = \{0, 1, \ldots, w_0\}$ in terms of

$$\| g_{\lambda,A} \| \leq \lambda^{-1}(1 - e^{-\lambda}) \min\left\{1, \frac{e^{\lambda}}{w_0 + 1}\right\}.$$

In the general case, for any subset $A$ of $\{0, 1, \ldots, n\}$, Santiwipanont and Teerapabolarn ([10], 2006) gave a bound in the form

$$\| g_{\lambda,A} \| \leq \lambda^{-1} \min\left\{1, \lambda, \frac{\triangle(\lambda)}{M_A + 1}\right\}, \tag{2.5}$$

where

$$\triangle(\lambda) = \begin{cases} e^{\lambda} + \lambda - 1 & \text{if } \lambda^{-1}(e^{\lambda} - 1) \leq M_A, \\ 2(e^{\lambda} - 1) & \text{if } \lambda^{-1}(e^{\lambda} - 1) > M_A, \end{cases}$$

and

$$M_A = \begin{cases} \max\{w \mid C_w \subseteq A\} & \text{if } 0 \in A, \\ \min\{w \mid w \in A\} & \text{if } 0 \notin A. \end{cases}$$

The difficult part in applying Theorem 2.1 is to find $W_{n,i}$ which makes $E|W_n - W_{n,i}|$ small. For the case when $X_1, \ldots, X_n$ are independent, we let $W_{n,i} = W_n - X_i$. Then $E|W_n - W_{n,i}| = p_i$. From (2.3), we have

$$|P(W_n \in A) - Poi_{\lambda}(A)| \leq \| g_{\lambda,A} \| \sum_{i=1}^{n} p_i^2.$$

The problem of the construction of $W_{n,i}$ is difficult in the case of dependent indicator summand. In the next section, we will use Theorem 2.1 to prove our main result by constructing the random variable $W_{n,i}$ which makes $E|W_n - W_{n,i}|$ small.

# 3 Proof of the Main Results

We define the indicator random variable $X_i$ as follows:

$$X_i = \begin{cases} 1 & \text{if } i^{th} \text{ set of } t \text{ columns does not form a } t-\text{covering in rows} \\ 0 & \text{otherwise.} \end{cases}$$

Let $W_{n,i}$ be the total number of set of $t$ columns that is not a $t-$covering after we take the set $i^{th}$ of $t$ columns which is not a $t-$covering. For each $w_0 \in \{0, 1, 2, \ldots, k\}$, we get

$$P(W_{n,i} = w_0) = (1 - \frac{1}{v^t})^{w_0}$$

and

$$\begin{aligned} P(W_n - X_i = w_0 \mid X_i = 1) &= \frac{P(W_n - X_i = w_0, X_i = 1)}{P(X_i = 1)} \\ &= \frac{P(W_n = w_0 + 1, X_i = 1)}{P(X_i = 1)} \\ &= \frac{(1 - \frac{1}{v^t})^{w_0+1}}{(1 - \frac{1}{v^t})} \\ &= (1 - \frac{1}{v^t})^{w_0}. \end{aligned}$$

It is clear that $W_{n,i}$ so constructed is distributed as $W_{n,i} - 1$ conditional on $X_i = 1$. We observe that:

- In case $X_i = 1$, we have the total number of set of $t$ columns that is not a $t-$covering after we take the set $i^{th}$ of $t$ columns which is not a $t-$covering, equals the number of set of $t$ columns that is not a $t-$covering minus 1; that is,

$$W_{n,i} = W_n - 1. \tag{3.6}$$

- In case $X_i = 0$, the number of set of $t$ columns that is missing at least one word after we take the set $i^{th}$ of $t$ columns which is not a $t-$covering and we test them again as defined, equals to the number of set of $t$ columns that is not a $t-$covering minus the sum of the number of the set $j^{th}$ of $t$ columns, $i \neq j$, is the number of set of $t$ columns that is not a $t-$covering in the first test, and they are not $t-$covering after we test them again; that is,

$$W_{n,i} = W_n - \sum_{i,j=1,i\neq j}^{\binom{n}{t}} X_i Y_j. \tag{3.7}$$

For each $j \in \{0, 1, 2, \ldots, n\}$, $i \neq j$, we define the indicator random variable as follows:

$$Y_j = \begin{cases} 1 & \text{if the } j^{th} \text{ set of } t \text{ columns forms a } t-\text{covering in rows} \\ 0 & \text{otherwise.} \end{cases}$$

So the probability that $Y_j = 1$ is given by

$$P(Y_j = 1) = (\frac{1}{v^t})^k. \tag{3.8}$$

We know that

$$E \mid W_n - W_{n,i} \mid = E(W_n - W_{n,i})^+ + E(W_n - W_{n,i})^-,$$

where

$$(W_n - W_{n,i})^+ = \max\{W_n - W_{n,i}, 0\},$$

and

$$(W_n - W_{n,i})^- = -\min\{W_n - W_{n,i}, 0\}.$$

Form (3.6) and (3.7):

- In case $X_i = 1$, we have $(W_n - W_{n,i})^+ = 1$ and $(W_n - W_{n,i})^- = 0$.

- In case $X_i = 0$, we have $(W_n - W_{n,i})^+ = \sum_{i,j=1}^{2n} X_i Y_j$ and $(W_n - W_{n,i})^- = 0$.

Therefore,
$(W_n - W_{n,i})^+ = \sum_{i,j=1,i\neq j}^{\binom{n}{t}} X_i Y_j$ and $(W_n - W_{n,i})^- = 0$.

$$
\begin{aligned}
E(W_n - W_{n,i})^+ &= E\{\sum_{i,j=1,i\neq j}^{\binom{n}{t}} X_i Y_j\} \\
&= \sum_{i,j=1,i\neq j}^{\binom{n}{t}} E\{X_i Y_j\} \\
&= \sum_{i,j=1,i\neq j}^{\binom{n}{t}} P(X_i = 1, Y_j = 1) \\
&= \sum_{i,j=1,i\neq j}^{\binom{n}{t}} P(X_i = 1) P(Y_j = 1) \\
&= \sum_{i,j=1,i\neq j}^{\binom{n}{t}} (1 - \frac{1}{v^t})^k (\frac{1}{v^t})^k \\
&= \binom{n}{t} (\frac{v^t - 1}{v^{2t}})^k . \qquad (3.9)
\end{aligned}
$$

Hence, by (2.3), (2.4), (2.5) and (3.9), we have

$$|P(W_n \in A) - Poi_\lambda(A)| \leq C_{\lambda,A} \binom{n}{t} (\frac{v^t - 1}{v^{2t}})^k .$$

In addition,

$$|P(W_n \in A) - Poi_\lambda(A)| \leq (1 - e^{-\lambda}) \binom{n}{t} (\frac{v^t - 1}{v^{2t}})^k ,$$

where $C_{\lambda,A} = \min\left\{1, \lambda, \frac{\triangle(\lambda)}{M_A+1}\right\}$,

$$\triangle(\lambda) = \begin{cases} e^\lambda + \lambda - 1 & \text{if } \lambda^{-1}(e^\lambda - 1) \leq M_A, \\ 2(e^\lambda - 1) & \text{if } \lambda^{-1}(e^\lambda - 1) > M_A, \end{cases}$$

and

$$M_A = \begin{cases} \max\{w \mid C_w \subseteq A\} & \text{if } 0 \in A, \\ \min\{w \mid w \in A\} & \text{if } 0 \notin A \end{cases}$$

when $C_w = \{0, 1, \ldots, w-1\}$.

# References

[1] A. D. Barbour, Poisson convergence and random graphs, Mathematical Proceedings of the Cambridge Philosophical Society, **92,** no. 2, (1982), 349–359.

[2] L. H. Chen, Poisson approximation for dependent trials, The Annals of Probability, **3,** (1975), 534–545.

[3] S. Janson, . Coupling and Poisson approximation, Acta Applicandae Mathematica, **34,** nos.1-2, (1994), 7–15.

[4] M. F. Johansen, $\phi$ Haugen, F. Fleurey, An algorithm for generating t-wise covering arrays from large feature models, Proceedings of the 16th International Software Product Line Conference, **1**, (2012), 46–55.

[5] K. Lange, Applied Probability, Spinger-Verlag, New York, 2003.

[6] Y. Lei, R. H. Carver, R. Kacker, D. Kung, A combinatorial testing strategy for concurrent programs, Software Testing, Verification and Reliability, **17,** no. 4, (2007), 207–225.

[7] K. Neammanee, Pointwise approximation of Poisson binomial by Poisson distribution, Stochastic Modelling and Applications, **6,** (2003), 20–26.

[8] C. Nie, H. Leung, A survey of combinatorial testing, ACM Computing Surveys, **43,** no. 2, (2011), 1–29.

[9] G. Rothermel, S. Elbaum, A. G. Malishevsky, P. Kallakuri, X. Qiu, On test suite composition and cost-effective regression testing, ACM Transactions on Software Engineering and Methodology, **13,** no. 3, (2004), 277–331.

[10] T. Santiwipanont, K. Teerapabolarn, Two formulas of non-uniform bounds on Poisson approximation for dependent indicators, Thai Journal of Mathematics, **1,** (2006), 15–39.

[11] C. Stein, A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, **2**, (1972), 583–602.

[12] B. Stevens, E. Mendelsohn, Efficient software testing protocols, Proceedings of the 1998 conference of the Center for Advanced Studies on Collaborative research, (1998). 22 pp.

[13] K. Teerapabolarn, K. Neammanee, A non-uniform bound on Poisson approximation for dependent trials, Stochastic Modelling and Applications, **8,** no. 1, (2005), 17–31.

[14] C. Yilmaz, M. B. Cohen, A. A.Porter, Covering arrays for efficient fault characterization in complex configuration spaces, IEEE Transactions on Software Engineering, **32,** no. 1, (2006), 20–34.

[15] X. Yuan, M. B, Cohen, A. M. Memon, GUI interaction testing: Incorporating event context, IEEE Transactions on Software Engineering, **37,** no. 4, (2010), 559–574.