ON COINCIDENCES OF TUPLES IN A BINARY TREE WITH RANDOMLY LABELED VERTICES

A. M. Zubkov[1], V. I. Kruglov[2]
Steklov Mathematical Institute, Russian Academy of Sciences
Moscow, RUSSIA
e-mail: [1]`zubkov@mi.ras.ru`, [2]`kruglov@mi.ras.ru`

Abstract

Let all vertices of a complete binary tree of finite height be independently and equiprobably labeled by the elements of some finite alphabet. We consider the numbers of pairs of identical tuples of labels on chains of subsequent vertices in the tree. Exact formulae for the expectations of these numbers are obtained. Convergence to the compound Poisson distribution is proved.

The work was supported by the Russian Science Foundation under grant 14-50-00005.

Let T_2^n be a complete binary tree of height n with root $*$ and n layers of vertices; we enumerate 2^k elements of the set $I^{(k)}$ of the k-th layer vertices ($k = 1, 2, \ldots, n$) by binary strings $i = (i_1, i_2, \ldots, i_k) \in \{0, 1\}^k$. So the unique vertex $*$ of layer $I^{(0)}$ is connected by two outcoming edges with vertices of layer $I^{(1)}$ and any vertex $i = (i_1, i_2, \ldots, i_k) \in I^{(k)}, 1 \le k \le n-1$, is connected by two outcoming edges with vertices $i' = (i_1, i_2, \ldots, i_k, 0)$ and $i'' = (i_1, i_2, \ldots, i_k, 1)$ of layer $I^{(k+1)}$. Vertex $i = (i_1, i_2, \ldots, i_k)$ has incoming edge from vertice $i^- = (i_1, i_2, \ldots, i_{k-1})$ for $k > 1$ and from root $* = (0)^- = (1)^-$ for $k = 1$. Each vertex i of the tree T_2^n defines subtree consisting of this vertex and all vertices of next layers that are connected to i with edges.

We can define natural lexicographical order on the set of vertices of T_2^n : $i = (i_1, \ldots, i_k) \prec j = (j_1, \ldots, j_h)$ if either $i = *, j \ne *$, or $1 \le k < h$, or $1 \le k = h$ and $\sum_{m=1}^{k} i_m 2^{k-m} < \sum_{m=1}^{k} j_m 2^{k-m}$. For vertex $i = (i_1, i_2, \ldots, i_k) \in I^{(k)}, k \ge 0$, the chain C_i of length l is a sequence of l vertices

$$(i_1, i_2, \ldots, i_k), (i_1, i_2, \ldots, i_k, i_{k+1}), \ldots, (i_1, i_2, \ldots, i_k, i_{k+1}, \ldots, i_{k+l-1})$$

connected by edges. Denote these vertices of the chain C_i by $C_i[0], C_i[1], \ldots, C_i[l-1]$. We will refer to the vertex i as the initial vertex of chain C_i and to the vertex $(i_1, i_2, \ldots, i_k, i_{k+1}, \ldots, i_{k+l-1})$ as its final vertex. It's easy to see that the final vertex and length l explicitly define the chain, so we can introduce order on the set of chains of the fixed length l: $C_i \prec C_j$ if and only if $C_i[l-1] \prec C_j[l-1]$. Denote by $\mathcal{P}$ the set of ordered pairs of nonintersecting chains $(C_i, C_j), i \prec j$.

It is easy to check that chains C_i and C_j intersect if and only if either $C_i[0] \in C_j$ or $C_j[0] \in C_i$. The total number of vertices in the tree T_2^n is equal to $1 + 2 + \ldots + 2^n = 2^{n+1} - 1$, and the total number of chains of the length l in the tree T_2^n is equal to the number of their final vertices $| \cup_{j=l-1}^{n} I^{(j)}| = 2^{l-1} + \ldots + 2^n = 2^{n+1} - 2^{l-1}$.

Let any vertex i in tree T_2^n be assigned with a random label $m(i)$ from the set $\{1, \ldots, d\}$ so that variables $m(i)$, $i \in T_2^n$, are independent and $\mathbf{P}\{m(i) = j\} = \frac{1}{d}, j \in$

$\{1,\ldots,d\}$, for all $i \in T_2^n$. So, for any chain C_i of length l we have a random tuple of labels

$$M(C_i) = (m(C_i[0]), m(C_i[1]), \ldots, m(C_i[l-1])).$$

Obviously, if all chains $C_{i_1},\ldots,C_{i_s}$ are nonintersecting, then the corresponding tuples of labels $M(C_{i_1}),\ldots,M(C_{i_s})$ are independent and equiprobably distributed on the set $\{1,\ldots,d\}^l$.

We consider the distribution of the number of pairs $(C_i, C_j), i \prec j$, of chains of length l in the tree T_2^n with identical tuples of labels (i.e. $M(C_i) = M(C_j)$). Total number of such pairs is equal to

$$V_{n,l} = \sum_{(C_i,C_j)\in\mathcal{P}} \mathbf{I}\{M(C_i) = M(C_j)\};$$

the alphabet size d is supposed to be fixed.

Probability of the event $\{M(C_i) = M(C_j)\}$ depends on the character of intersection of chains C_i and C_j, so we divide the sum $V_{n,l}$ into several parts: sum $V_{n,l}^{(0)}$ over the nonintersecting chains, sum $V'_{n,l}$ over intersecting chains with different initial vertices, sum $V''_{n,l,k}$ over chains with common initial vertices:

$$V_{n,l} = V_{n,l}^{(0)} + V'_{n,l} + \sum_{k=1}^{l-1} V''_{n,l,k},$$

$$V_{n,l}^{(0)} = \sum_{(C_i,C_j)\in\mathcal{P}\colon C_i\cap C_j=\emptyset} \mathbf{I}\{M(C_i) = M(C_j)\},$$

$$V'_{n,l} = \sum_{(C_i,C_j)\in\mathcal{P}\colon C_i\cap C_j\neq\emptyset, C_i[0]\neq C_j[0]} \mathbf{I}\{M(C_i) = M(C_j)\},$$

$$V''_{n,l,k} = \sum_{(C_i,C'_i)\in\mathcal{P}\colon |C_i\cap C'_i|=k, C_i[0]=C'_i[0]} \mathbf{I}\{M(C_i) = M(C'_i)\},\ 1 \le k < l.$$

Theorem 1. *The following equalities are valid*

$$\mathbf{E}V_{n,l}^{(0)} = \begin{cases} \frac{1}{d^l}\left(2^{2n+1} - 5\cdot 2^{n-1+l} + 2^{n+1} + 2^{2l-2}l\right), & \text{if } 2l-1 \le n,\\ \frac{1}{d^l}\left(2^{2n+1} - 5\cdot 2^{n-1+l} + 2^{2l-2}(n-l+4)\right), & \text{if } 2l-1 > n,\end{cases}$$

$$\mathbf{E}V'_{n,l} = \begin{cases} \frac{1}{d^l}\left((2^{l-1}-1)2^{n+1} - 2^{2l-2}(l-1)\right), & \text{if } 2l-1 \le n,\\ \frac{1}{d^l}\left(2^{l-1}(2^{n+1}-2^l) - 2^{2l-2}(n-l+1)\right), & \text{if } 2l-1 > n,\end{cases}$$

$$\mathbf{E}V''_{n,l,k} = \tfrac{1}{d^{l-k}}\,(2^{n-l+2}-1)2^{2l-k-3},\ 1 \le k < l,$$

$$\sum\nolimits_{k=1}^{l-1} \mathbf{E}V''_{n,l,k} = \frac{2^n - 2^{l-2}}{d}\,\frac{1-(\frac{2}{d})^{l-1}}{1-\frac{2}{d}}\,.$$

If $M(C_i) = M(C_j)$ and $i^- \neq j^-$, then $M(C_{i^-}) = M(C_{j^-})$ with probability $1/d = \mathbf{P}\{m(i^-) = m(j^-)\}$, and $\mathbf{P}\{M(C_i') = M(C_j')\} = 1/d$ if $C_i'[0] = C_j'[0], C_i'[l-2] = C_j'[l-2], C_i'[l-1] \neq C_j'[l-1]$. In theorem 2 we propose sufficient conditions and estimate the weak convergence rate of the number of pairs of nonintersecting chains C_i, C_j with $M(C_i) = M(C_j), m(i^-) \neq m(j^-)$ to the compound Poisson distribution. Such pairs of tuples may be interpreted as coincidences which cannot be shifted to the root.

Let

$$X_{C_iC_j} = \mathbf{I}\{M(C_i) = M(C_j), m(i^-) \neq m(j^-)\}, (C_i, C_j) \in \mathcal{P};$$

if $i = *$, then the condition $m(i^-) \neq m(j^-)$ is supposed to be satisfied. Labels of vertices are independent and equiprobable, so for $(C_i, C_j) \in \mathcal{P}$ we have

$$\mathbf{E}X_{C_iC_j} = \mathbf{E}\,\mathbf{I}\{M(C_i) = M(C_j)\}\mathbf{I}\{m(i^-) \neq m(j^-)\} = \begin{cases} \frac{d-1}{d^{l+1}}, & \text{if } i^- \neq j^-, \\ 0, & \text{if } i^- = j^-. \end{cases}$$

Let $\widetilde{\mathcal{P}} \subset \mathcal{P}$ be the set of pairs $(C_i, C_j), i \in I^{(v_i)}, j \in I^{(v_j)}$, of nonintersecting chains such that if the vertex j belongs to a subtree with root i, then $v_j \geq v_i + 2l - 1$. Define

$$V_{n,l}^{(0)-} = \sum_{(C_i,C_j)\in\mathcal{P}:\, C_i\cap C_j=\emptyset} X_{C_iC_j}, \quad \widetilde{V}_{n,l} = \sum_{(C_i,C_j)\in\widetilde{\mathcal{P}}} X_{C_iC_j}.$$

Lemma 1. *The following equalities are valid*

$$\mathbf{E}V_{n,l}^{(0)-} = \begin{cases} \frac{d-1}{d^{l+1}} \left(2^{2n+1} - 6 \cdot 2^{n+l-1} + 2^{n+1} + 2^{2l-2}(l+1)\right), & \text{if } 2l - 1 \leq n, \\ \frac{d-1}{d^{l+1}} \left(2^{2n+1} - 6 \cdot 2^{n+l-1} + 2^{2l-2}(n - l + 5)\right), & \text{if } 2l - 1 > n. \end{cases}$$

$$\mathbf{E}V_{n,l}^{(0)-} - \tfrac{l}{d^l}\, 2^{n-l+2} < \mathbf{E}\widetilde{V}_{n,l} < \mathbf{E}V_{n,l}^{(0)-}.$$

Corollary 1. *If $n, l \to \infty$ in such a way that $\mathbf{E}V_{n,l}^{(0)-}$ is bounded, then $\mathbf{P}\{\widetilde{V}_{n,l} = V_{n,l}^{(0)-}\} \to 1$.*

Comparing formulae for $\mathbf{E}V_{n,l}^{(0)}$ and $\mathbf{E}V_{n,l}^{(0)-}$ we can mention that under the conditions of corollary 1 for any coincidence which cannot be shifted to the root there exist in average $\frac{1}{d-1}$ additional coincidences that may be shifted to root.

Definition 1. Consider a pair of chains $(C_i, C_j) \in \mathcal{P}$ such that subtrees of height $l-1$ with roots in vertices i an j do not intersect. Define

$$\pi_k = \frac{1}{k}\mathbf{P}\left\{ \sum\nolimits_{(C_i',C_j')\in\mathcal{P}} X_{C_i'C_j'} = k \,\middle|\, X_{C_iC_j} = 1 \right\}, \quad k = 1, 2, \ldots$$

Definition 2. The compound Poisson distribution $CP(\pi)$ is the distribution of random variable

$$\Xi_\pi = \sum\nolimits_{k=1}^{\infty} k\xi_k,$$

where $\xi_1, \xi_2, \ldots$ are independent and for any $k \geq 1$ random variable ξ_k has Poisson distribution with parameter π_k.

Theorem 2. *If $n, l \to \infty$ in such a way that $2^{2l} = o(2^n)$ and*

$$\mathbf{E}\widetilde{V}_{n,l} = \frac{d-1}{d} \cdot \frac{2^{2n+1}}{d^l}(1 + o(1)) \to \lambda \in (0, \infty),$$

then there exists $\varepsilon(l,n)$ such that $\varepsilon(l,n) = o(1)$ and

$$d_{\text{tv}}(\mathcal{L}(\widetilde{V}_{n,l}), CP(\pi)) = \frac{1}{2}\sum_{k=0}^{\infty} |\mathbf{P}\{\widetilde{V}_{n,l} = k\} - \mathbf{P}\{\Xi_\pi = k\}| \leq$$

$$\leq 2H_1(\pi)\left(\mathbf{E}\widetilde{V}_{n,l}\right)^2 \frac{2^{2l}}{2^n}\left(1 + \varepsilon(l,n)\right) \to 0,$$

where $H_1(\pi) \leq \min\left(1, \frac{1}{\pi_1}\right) \cdot \exp\left(\sum\limits_{k=1}^{\infty} \pi_k\right)$.

References

[1] Erhardsson T. Stein's method for Poisson and compound Poisson approximation. — In Barbour A. D., Chen L. H. Y. (ed.) An introduction to Stein's method, Singapore Univ. Press, 2005, p.61–113.

[2] Hoffmann C.M., O'Donnell M.J. (1982). Pattern matching in trees. *J. ACM.* Vol. **29:1**, pp. 68-95.

[3] Steyaert J.-M., Flajolet P. (1983). Patterns and pattern-matching in trees: an analysis. *Inf. & Control.* Vol. **58:1**, pp. 19-58.

[4] Zubkov A.M., Mikhailov V.G. (1974). Limit distributions of random variables associated with long duplications in a sequence of independent trials. *Teoriya veroyatn. primen.* Vol. **19:1**, pp. 173-181 (in Russian; translated: *Theory Probab. Appl.* Vol. **19**:1, pp. 172-179).