THE ERDÖS-TURÁN LAW FOR MIXTURES OF DIRICHLET PROCESSES (II)

Yamato, Hajime
Kagoshima University : Professor Emeritus

https://doi.org/10.5109/1798146

出版情報：Bulletin of informatics and cybernetics. 46, pp.47-51, 2014-12. Research Association of Statistical Sciences
バージョン：

THE ERDŐS-TURÁN LAW FOR MIXTURES OF DIRICHLET PROCESSES (II)

by

Hajime Yamato

Reprinted from the Bulletin of Informatics and Cybernetics Research Association of Statistical Sciences, Vol.46

FUKUOKA, JAPAN
2014

Bulletin of Informatics and Cybernetics, Vol. 46, 2014

THE ERDŐS-TURÁN LAW FOR MIXTURES OF DIRICHLET PROCESSES (II)

By

Hajime Yamato*

Abstract

Let a random distribution $\mathcal{P}$ on the real line $\mathbb{R}$ have the mixture of Dirichlet processes. Let $S^{(n)} = (S_1, \cdots, S_n)$ be the random partition of the positive integer n based on a sample of size n from $\mathcal{P}$. For the order $O_n(S^{(n)})$ of $S^{(n)}$, Yamato (2013) gives the asymptotic distribution of the statistic $\log O_n(S^{(n)})/\log^2 n$ and the rate $O(1/\log^{1/3} n)$ of its convergence. In this pager we give the Edgeworth expansions for the statistic with the rates $O(1/\log^{2/5} n)$ and $O(1/\log^{3/7} n)$. In addition, we correct the errors of the proofs of the lemmas 2.5 and 2.6 of Yamato (2013).

Key Words and Phrases: Edgeworth expansion, Erdős-Turán law, Fourier transform, mixture of Dirichlet processes, order of partition, random partition, smoothing lemma.

1. Introduction

Let G_0 be a continuous distribution on the real line $\mathbb{R}$ and $\mathcal{B}$ be the σ-field which consists of the subsets of $\mathbb{R}$. Let θ be a positive random variable having a distribution γ. We suppose that a random distribution $\mathcal{P}$ have the mixture of Dirichlet process $\mathcal{D}(\theta G_0)$ on $(\mathbb{R}, \mathcal{B})$ with the mixing distribution γ (for the mixture of Dirichlet process, see Antoniak (1974)). For a sample of size n from the random distribution $\mathcal{P}$, S_1 denotes the number of observations which occur only once, S_2 the number of observations which occur exactly twice, ... and so on. For the random partition $S^{(n)} = (S_1, \cdots, S_n)$ of the positive integer n, the order $O_n(S^{(n)})$ denotes l.c.m.$\{ j \ : \ S_j > 0 \ (j = 1, 2, \cdots, n) \}$, where l.c.m. represents the least common multiple. Let H be the distribution functions (d.f.) of $\theta/2$. For the convergence of the statistic $O_n(S^{(n)})/\log^2 n$, Yamato (2013) gives

$$\sup_{-\infty<x<\infty} \left| P\left(\frac{\log O_n(S^{(n)})}{\log^2 n} \leq x \right) - H(x) \right| = O\left(\frac{1}{\log^{1/3} n} \right).$$

In the section 2 we give the Edgeworth expansion for the statistic $O_n(S^{(n)})/\log^2 n$ with the rate $O(1/\log^{2/5} n)$, which is the proposition 2.1. In the section 3, we give the Edgeworth expansion with the rate $O(1/\log^{3/7} n)$, which is the proposition 3.1. In the section 4, for the lemmas 2.5 and 2.6 of Yamato (2013), the errors of their proofs are corrected.

* Emeritus of Kagoshima University, Take 3–32–1–708 Kagoshima 890-0045 Japan.

2. The Edgeworth expansion with the rate $O(1/\log^{2/5} n)$

Suppose that $E_\gamma(\theta^2)$ exist, where E_γ denotes the expectation with respect to the distribution γ. Let h be the bounded density function of the d.f. H (or of the random variable $\theta/2$). For the smoothing lemma (see, for example, Petrov (1995; Theorem 5.2)) used in the proof of the proposition 2.1, we suppose the followings: (i) h be twice differentiable, and $\{xh(x)\}'$ $(= h(x) + xh'(x))$ be of bounded variation, (ii) $h'(x)$ and $xh''(x)$ be bounded, that is, $\{xh(x)\}^{(2)}$ be bounded and (iii) $h(x) = 0$, $xh'(x) = 0$ for $x = 0$ and $xh'(x) \to 0$ as $x \to +\infty$. We also use the following lemma.

Lemma 2.1. (Petrov (1995; Lemma 1.8)) *Let X and W be arbitrary random variables, and let $F(x)$ be the distribution functions of X. If $T(x)$ is an arbitrary function defined on the real line, then for every real x and every positive ε*

$$| P(X + W \le x) - T(x) | \le K + L + P(| W | > \varepsilon),$$

where $K = \max\Big\{ \mid F(x+\varepsilon) - T(x+\varepsilon) \mid, \mid F(x-\varepsilon) - T(x-\varepsilon) \mid \Big\}$ and $L = \max\Big\{ \mid T(x+\varepsilon) - T(x) \mid, \mid T(x-\varepsilon) - T(x) \mid \Big\}$.

We use the same notations as Yamato (2013), except for H and h. Then, we have

Proposition 2.2.

$$\sup_{-\infty<x<\infty} \left| P\left(\frac{\log O_n(S^{(n)})}{\log^2 n} \le x \right) - \Big[H(x) + \frac{1}{3\log n} \big\{ h(x) + xh'(x) \big\} \Big] \right| = O\left(\frac{1}{\log^{2/5} n} \right). \tag{1}$$

Proof. At first, we note the following relations.

$$\frac{1}{\log^2 n} \sum_{j=1}^{n} \frac{\log j}{j} = \frac{1}{2} + \frac{c_1}{\log^2 n}, \quad \frac{1}{\log^{i+1} n} \sum_{j=1}^{n} \frac{\log^i j}{j} = \frac{1}{i+1} + \frac{c_i}{\log^{i+1} n} \quad (i = 2, 3),$$

where c_i $(i = 1, 2, 3)$ denote generic positive constants. We use also c as a generic constant. By these relations we get

$$\sum_{j=1}^{n} \frac{1}{j} \Big(e^{it\frac{\log j}{\log^2 n}} - 1 \Big) = i\frac{1}{2}t - \frac{t^2}{6\log n} + c\frac{| t |}{\log^{6/5} n} \quad (| t | \le \log^{2/5} n).$$

Given θ, let $Z_1, \cdots, Z_n$ be independent and Z_j have the Poisson distribution with mean θ/j $(j = 1, \cdots, n)$. For $Z^{(n)} = (Z_1, \cdots, Z_n)$, we put

$$O_n(Z^{(n)}) = \text{l.c.m.}\{ \, j \, : \, Z_j > 0 \, (j = 1, 2, \cdots, n) \, \}, \quad T_n(Z^{(n)}) = \prod_{j=1}^{n} j^{Z_j}$$

and $\mu_n(\theta) = E\big[\log T_n\big(Z^{(n)}\big) - \log O_n\big(Z^{(n)}\big) \bigm| \theta \big]$. We put

$$S^*_{1n} = \frac{\log T_n(Z^{(n)})}{\log^2 n}, \quad S^*_{2n} = \frac{\log O_n(Z^{(n)}) + \mu_n(\theta)}{\log^2 n} \quad \text{and} \quad S^*_{3n} = \frac{\log O_n(S^{(n)})}{\log^2 n}.$$

Then, for the characteristic function (c.f.) f_n of $S^*_{1n} = \sum_{j=1}^{n} Z_j \log j / \log^2 n$, we have

$$\begin{aligned} f_n(t) &= E_\gamma \exp\left\{ \theta \sum_{j=1}^{n} \frac{1}{j} \Big(e^{it\frac{\log j}{\log^2 n}} - 1 \Big) \right\} = E_\gamma \left[e^{i\frac{\theta}{2}t} \exp\left\{ -\theta \frac{t^2}{6\log n} + c\theta \frac{| t |}{\log^{6/5} n} \right\} \right] \\ &= E_\gamma e^{i\frac{\theta}{2}t} - \frac{t^2}{3\log n} E_\gamma\Big(\frac{\theta}{2} e^{i\frac{\theta}{2}t} \Big) + c_1 \frac{| t |}{\log^{4/5} n}, \quad (| t | \le \log^{2/5} n). \end{aligned} \tag{2}$$

Let the c.f. of $\theta/2$ be $\varphi_H(t) = \int_{-\infty}^{\infty} e^{ixt} h(x)dx = \int_{-\infty}^{\infty} e^{ixt} dH(x)$. Since $\varphi_H(t)$ be the Fourier transform of h, $-t^2 \int_{-\infty}^{\infty} e^{ixt} xh(x)dx$ is the Fourier transform of $\{xh(x)\}''$. Or, $\varphi_H(t) = E_\gamma e^{i\frac{\theta}{2}t}$ corresponds to H and $-t^2 \int_{\infty}^{\infty} e^{ixt} xh(x)dx = -t^2 E_\gamma\left(\frac{\theta}{2} e^{i\frac{\theta}{2}t}\right)$ to $\{xh(x)\}'$. Therefore, by the smoothing lemma, we have

$$\sup_x \left| P\big(S_{1n}^* \le x\big) - \Big[H(x) + \frac{1}{3\log n}\big\{h(x) + xh'(x)\big\}\Big] \right| \le \frac{c_1}{\log^{4/5} n} \int_0^{\log^{2/5} n} dt + \frac{c_2}{\log^{2/5} n} = O\left(\frac{1}{\log^{2/5} n}\right). \quad (3)$$

This corresponds to Proposition 2.3 of Yamato (2013). Next, we derive the results with S_{2n}^* and S_{3n}^* similar to (3), instead of S_{1n}^*.

(I) We use Lemma 2.1 by taking $X = S_{1n}^*$, $W = S_{2n}^* - S_{1n}^*$, $T(x) = H(x) + \big\{h(x) + xh'(x)\big\}/(3\log n)$, and $K \le \sup_{-\infty<x<\infty} \mid P(S_{1n}^* \le x) - T(x) \mid = O(1/\log^{2/5} n)$. By (15) of the section 4, we have $P(\mid W \mid > \varepsilon) = O(1/\log^2 n)$ for any $\varepsilon > 0$. Since h' and xh'' are bounded, we get $L = O(1/\log^{2/5} n)$ with $\varepsilon = 1/\log^{2/5} n$. Thus we have

$$\sup_{-\infty<x<\infty} \left| P\big(S_{2n}^* \le x\big) - \Big[H(x) + \frac{1}{3\log n}\big\{h(x) + xh'(x)\big\}\Big] \right| = O\left(\frac{1}{\log^{2/5} n}\right). \quad (4)$$

(II) Again, we use Lemma 2.1 by taking $X = S_{2n}^*$ and $W = S_{3n}^* - S_{2n}^*$, $K \le \sup_{-\infty<x<\infty} \mid P(S_{2n}^* \le x) - T(x) \mid = O(1/\log^{2/5} n)$. $T(x)$ equals to the one of the above paragraph. With $\varepsilon = 1/\log^{1/2} n$, we have $P(\mid W \mid > \varepsilon) \le O(\log\log n/\log^{1/2} n) = o(1/\log^{2/5} n)$ by (18) of the section 4 and $L = O(1/\log^{1/2} n) = o(1/\log^{2/5} n)$ by the boundedness of h' and xh''. Thus we have the following (5), which yields (1).

$$\sup_{-\infty<x<\infty} \left| P\big(S_{3n}^* \le x\big) - \Big[H(x) + \frac{1}{3\log n}\big\{h(x) + xh'(x)\big\}\Big] \right| = O\left(\frac{1}{\log^{2/5} n}\right). \quad (5)$$

The assumptions about θ or $\theta/2$ (h) of the section 2 are satisfied, for example, by the gamma distribution whose density is $h(x) = x^{d-1}e^{-x}/\Gamma(d)$ $(x > 0,\ d > 2)$.

3. The Edgeworth expansion with the rate $O(1/\log^{3/7} n)$

We suppose that $E_\gamma(\theta^3)$ exist. In addition to the assumption of the first paragraph of the section 2.1, we suppose that h is differentiable four times. Suppose that $h'(x)$, $xh^{(2)}(x)$ and $x^2h^{(3)}(x)$ are of bounded variation, and that $h^{(2)}(x)$, $xh^{(3)}(x)$ and $x^2h^{(4)}(x)$ are bounded. Suppose that $xh^{(2)}(x) = 0$, $x^2h^{(3)}(x) = 0$ for $x = 0$ and $xh^{(2)}(x)$, $x^2h^{(3)}(x) \to 0$ as $x \to +\infty$. At first, we note the following relation.

$$\sum_{j=1}^n \frac{1}{j}\left(e^{it\frac{\log j}{\log^2 n}} - 1\right) = i\frac{1}{2}t - \frac{t^2}{6\log n} - i\frac{t^3}{24\log^2 n} + c\frac{\mid t \mid}{\log^{12/7} n} \quad (\mid t \mid \le \log^{3/7} n).$$

Thus, for the c.f. f_n of S_{1n}^*, we have

$$f_n(t) = E_\gamma e^{i\frac{\theta}{2}t} - \frac{t^2}{3\log n} E_\gamma\left(\frac{\theta}{2}e^{i\frac{\theta}{2}t}\right) - i\frac{t^3}{12\log^2 n} E_\gamma\left(\frac{\theta}{2}e^{i\frac{\theta}{2}t}\right) + \frac{t^4}{18\log^2 n} E_\gamma\Big[\Big(\frac{\theta}{2}\Big)^2 e^{i\frac{\theta}{2}t}\Big] + c_1\frac{\mid t \mid}{\log^{6/7} n}, \quad (\mid t \mid \le \log^{3/7} n). \quad (6)$$

Under the Fourier transform, $-it^3 E_\gamma\left(\frac{\theta}{2}e^{i\frac{\theta}{2}t}\right)$ correponds to $\{xh(x)\}^{(2)}$ and $t^4 E_\gamma\left[\left(\frac{\theta}{2}\right)^2 e^{i\frac{\theta}{2}t}\right]$ to $\{x^2h(x)\}^{(3)}$. Therefore, by the smoothing lemma, we have

$$\sup_x \Big| P\big(S^*_{1n} \le x\big) - \Big[H(x) + \frac{1}{3\log n}\big\{h(x) + xh'(x)\big\} + \frac{1}{36\log^2 n}\big\{6h'(x) + 9xh^{(2)}(x) + 2x^2h^{(3)}(x)\big\}\Big]\Big| = O\bigg(\frac{1}{\log^{3/7} n}\bigg). \quad (7)$$

Corresponding to (I) of the section 2, we get $L = O(1/\log^{3/7} n)$ with $\varepsilon = 1/\log^{3/7} n$ and the result (7) with S^*_{2n} instead of S^*_{1n}. Corresponding to (II) of the section 2, with $\varepsilon = 1/\log^{1/2} n$ we get $P(|\ W\ | > \varepsilon) \le O(\log\log n/\log^{1/2} n) = o(1/\log^{3/7} n)$, $L = O(1/\log^{1/2} n) = o(1/\log^{3/7} n)$ and the result (7) with S^*_{3n} instead of S^*_{1n}. Thus, we have the following.

Proposition 3.1.

$$\sup_x \bigg| P\bigg(\frac{\log O_n(S^{(n)})}{\log^2 n} \le x\bigg) - \Big[H(x) + \frac{1}{3\log n}\big\{h(x) + xh'(x)\big\} + \frac{1}{36\log^2 n}\big\{6h'(x) + 9xh^{(2)}(x) + 2x^2h^{(3)}(x)\big\}\Big]\bigg| = O\bigg(\frac{1}{\log^{3/7} n}\bigg).$$

The assumptions about θ or $\theta/2$ (h) of the section 3 are satisfied, for example, by the gamma distribution whose density is $h(x) = x^{d-1}e^{-x}/\Gamma(d)$ $(x > 0,\ d > 3)$.

4. Corrections to Yamato (2013)

In the following, we correct the proofs of Lemma 2.5 and 2.6 of Yamato (2013), which is from the line 6 from the top of page 65 to the line 3 from the bottom of the same page. The numbers of the equations are equal to the ones of Yamato (2013).

By the proposition 2.3 and its proof of Barbour and Tavaré (1994), it holds that

$$P\bigg(\Big|\log T_n(Z^{(n)}) - \log O_n(Z^{(n)}) - \mu_n(\theta)\big)\Big| > \varepsilon \log^2 n \,\Big|\, \theta\bigg) = \theta c_{1n} + \theta^2 c_{2n} \ \text{ for } \forall\varepsilon > 0 \quad (14)$$

where $c_{1n} = O\big((\log\log n)^2)/\log^3 n\big)$ and $c_{2n} = O(1/\log^2 n)$. Therefore, under the condition $E_\gamma\theta^2 < \infty$, by (13) and (14) we have

$$P(|S^*_{1n} - S^*_{2n}| > \varepsilon) = O\Big(\frac{1}{\log^2 n}\Big) \quad \text{for } \forall\epsilon > 0. \quad (15)$$

We use the relation (4) by taking $U = S^*_{1n}$, $X = S^*_{2n} - S^*_{1n}$, $H = \gamma^*$, $\eta = O(1/\log^{1/3} n)$, and $\varepsilon = O(1/\log^{1/3} n)$. By the relation (3) and (15), we obtain

$$\sup_{-\infty<x<\infty} |P(S^*_{2n} \le x) - \gamma^*(x)| = O\bigg(\frac{1}{\log^{1/3} n}\bigg). \quad (16)$$

Proof of Lemma 2.6 By the relation (2.1) and (2.2) of Barbour and Tavaré (1994), we have

$$|S^*_{2n} - S^*_{3n}| \le \bigg|\frac{\log O_n(Z^{(n)}) - \log O_n(S^{(n)})}{\log^2 n}\bigg| + \frac{|\ \mu_n(\theta)\ |}{\log^2 n} \le Y + \frac{\mu_n(\theta)}{\log^2 n}, \text{ given } \theta \quad (17)$$

where $Y = (Y_n+1)/\log n$, $E(Y_n) = E_\gamma(E(Y_n|\theta)) \leq E_\gamma\theta^2$ and $E_\gamma\mu_n(\theta) = O(\log n \log\log n)$, where $(0 \leq)\mu(\theta) = \theta \log n \log\log n + c\theta^2 \log n$ (Barbour and Tavaré (1994; p.171)). Thus, by the Markov's inequality and (17) we have

$$P(|S^*_{2n} - S^*_{3n}| > \varepsilon) \leq P\Big(|Y| > \frac{\varepsilon}{2}\Big) + P\Big(\frac{|\mu_n(\theta)|}{\log^2 n} > \frac{\varepsilon}{2}\Big) \leq c\frac{\log\log n}{\varepsilon \log n} \quad \text{for } \forall\varepsilon > 0. \quad (18)$$

We use the relation (4) by taking $U = S^*_{2n}$, $X = S^*_{3n} - S^*_{2n}$, $H = \gamma^*$, $\eta = O(1/\log^{1/3} n)$, and $\varepsilon = O(1/\log^{1/2} n)$. By the relation (16) and (18) with $\log\log n/(\varepsilon \log n) = o(1/\log^{1/3} n)$, we obtain

$$\sup_{-\infty<x<\infty} |P(S^*_{3n} \leq x) - \gamma^*(x)| = O\bigg(\frac{1}{\log^{1/3} n}\bigg). \quad (19)$$

Acknowledgement

The author is grateful to the referee for his careful reading and useful comments.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, **2**, 1152–1174.

Barbour, A.D. and Tavaré, S. (1994). A rate for the Erdős-Turán law. *Combin. Probab. Comput.*, **3**, 167–176.

Petrov, V. (1995). *Limit theorems of probability theory.* New York : Oxford Univ. Press.

Yamato, H. (2013). The Erdős-Turán law for mixtures of Dirichlet processes. *Bull. Inform. Cyber.*, **45**, 59–66.

Received May 20, 2014
Revised October 20, 2014