



Chinese restaurant process from stick-breaking for Pitman-Yor

Caroline Lawless, Julyan Arbel

► **To cite this version:**

Caroline Lawless, Julyan Arbel. Chinese restaurant process from stick-breaking for Pitman-Yor. Bayesian learning theory for complex data modelling Workshop, Sep 2018, Grenoble, France. pp.1. hal-01950662

HAL Id: hal-01950662

<https://hal.archives-ouvertes.fr/hal-01950662>

Submitted on 11 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Chinese restaurant process from stick-breaking for Pitman–Yor

CAROLINE LAWLESS AND JULYAN ARBEL



UNIV. GRENOBLE ALPES, INRIA, CNRS, LJK, 38000 GRENOBLE,

FRANCE

INTRODUCTION

- The Chinese restaurant process and the stick-breaking process are the two most commonly used representations of the Pitman–Yor process.
- However, the usual proof of the connection between them is indirect.
- Miller (2018) proved directly that the stick-breaking process gives rise to the Chinese restaurant process representation of the Dirichlet process.
- The Dirichlet process is a special case of the Pitman–Yor process.
- We extend Miller’s proof to Pitman–Yor process random measures.

PITMAN–YOR & DIRICHLET PROCESSES

- The Dirichlet Process (DP) and the Pitman–Yor process (PY, Pitman and Yor, 1997) are discrete random probability measures.
- The PY is parametrized by $d \in (0, 1)$, $\alpha > -d$, and a base probability measure P_0 . The DP is recovered by letting $d = 0$.
- The stick-breaking representation (Sethuraman, 1994) is given by

$$v_i \sim \begin{cases} \text{Beta}(1, \alpha) & \text{for DP} \\ \text{Beta}(1 + d, \alpha + id) & \text{for PY} \end{cases}$$

$$\pi_k = v_k \prod_{i=1}^{k-1} (1 - v_i), \phi_k \stackrel{\text{iid}}{\sim} P_0.$$

We define the random process P by

$$P = \sum_{i=1}^{\infty} \pi_k \delta_{\phi_k}.$$

- The Chinese restaurant process (Antoniak, 1974) is the distribution induced on random partitions \mathcal{C} given by

$$P(\mathcal{C} = C) = \begin{cases} \frac{\alpha^{|\mathcal{C}|} \Gamma(\alpha)}{\Gamma(n + \alpha)} \prod_{c \in C} \Gamma(|c|) & \text{for DP} \\ \frac{d^t (\frac{\alpha}{d})_t}{(\alpha)_{(n)}} \prod_{j=1}^t (1 - d)_{(|c_j| - 1)} & \text{for PY.} \end{cases}$$

THEOREM

Suppose π follows the PY stick-breaking, and

$$z_1, \dots, z_n | \pi = \pi \stackrel{\text{iid}}{\sim} \pi, \text{ that is, } \mathbb{P}(z_i = k | \pi) = \pi_k,$$

and \mathcal{C} is the partition of $[n]$ induced by z_1, \dots, z_n . Then \mathcal{C} follows the PY Chinese restaurant process.

TECHNICAL LEMMAS

Our proof relies on the following lemmas, which here we will state without proof. Let us abbreviate $z = (z_1, \dots, z_n)$. Given $z \in \mathbb{N}^n$, let C_z denote the partition $[n]$ induced by z . We define $m(z) = \max\{z_1, \dots, z_n\}$, and $g_k(z) = \#\{i: z_i \geq k\}$.

Lemma 1 For any $z \in \mathbb{N}^n$,

$$\mathbb{P}(z = z) = \frac{1}{(\alpha)_{(n)}} \prod_{c \in C_z} \frac{\Gamma(|c| + 1 - d)}{\Gamma(1 - d)} \prod_{k=1}^{m(z)} \frac{\alpha + (k - 1)d}{g_k(z) + \alpha + (k - 1)d}.$$

Lemma 2 For any partition C of $[n]$,

$$\sum_{z \in \mathbb{N}^n} \mathbb{1}(C_z = C) \prod_{k=1}^{m(z)} \frac{\alpha + (k - 1)d}{g_k(z) + \alpha + (k - 1)d} = \frac{d^t (\frac{\alpha}{d})_t}{\prod_{c \in C} (|c| - d)}.$$

PROOF OF THEOREM

$$\mathbb{P}(\mathcal{C} = C) = \sum_{z \in \mathbb{N}^n} \mathbb{P}(\mathcal{C} = C | z) \mathbb{P}(z = z)$$

$$\stackrel{(a)}{=} \sum_{z \in \mathbb{N}^n} \mathbb{1}(C_z = C) \frac{1}{(\alpha)_{(n)}} \prod_{c \in C_z} \frac{\Gamma(|c| + 1 - d)}{\Gamma(1 - d)} \prod_{k=1}^{m(z)} \frac{\alpha + (k - 1)d}{g_k(z) + \alpha + (k - 1)d}$$

$$= \frac{1}{(\alpha)_{(n)}} \prod_{c \in C} \frac{\Gamma(|c| + 1 - d)}{\Gamma(1 - d)} \sum_{z \in \mathbb{N}^n} \mathbb{1}(C_z = C) \prod_{k=1}^{m(z)} \frac{\alpha + (k - 1)d}{g_k(z) + \alpha + (k - 1)d}$$

$$\stackrel{(b)}{=} \frac{1}{(\alpha)_{(n)}} \prod_{c \in C} \frac{\Gamma(|c| + 1 - d)}{\Gamma(1 - d)} \frac{d^t (\frac{\alpha}{d})_t}{\prod_{c \in C} (|c| - d)}$$

$$\stackrel{(c)}{=} \frac{1}{(\alpha)_{(n)}} \prod_{c \in C} (1 - d)_{(|c| - 1)} \prod_{c \in C} (|c| - d) \frac{d^t (\frac{\alpha}{d})_t}{\prod_{c \in C} (|c| - d)}$$

$$= \frac{d^t (\frac{\alpha}{d})_t}{(\alpha)_{(n)}} \prod_{j=1}^t (1 - d)_{(|c_j| - 1)}$$

where (a) is by Lemma 1, (b) is by Lemma 2, and (c) is since $\Gamma(|c| + 1 - d) = (|c| - d)\Gamma(|c| - d)$.

FURTHER RESEARCH

- The Dirichlet process and the Pitman–Yor process are only special cases of a broad class of random measures called Gibbs-type random measures.
- An interesting further study would be to investigate the possibility of extending this proof to Gibbs-type random measures.

REFERENCES

- Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174.
- Miller, J. W. (2018). An elementary derivation of the Chinese restaurant process from Sethuraman’s stick-breaking process. *arXiv preprint arXiv:1801.00513*.
- Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.