Bayesian ML project: topic E7

Benoit ORIOL
benoit.oriol@polytechnique.edu

Antoine DOIZE
antoinedoize@hotmail.fr

Matthieu DARCY
matthieu.darcy@gmail.com

Abstract

This is a report for the Bayesian Machine Learning projec, topic E7 [1]. We present the main theoretical results on the Poisson-Dirichlet process (commonly known as the Pitman-Yor process) presented in Pitman, Yor, 1997. We begin with a theoretical review of the notions to better understand the Poisson Dirichlet distribution. We then apply key results to sample from the Pitman-Yor process using several methods. We end with an example of application which illustrates the differences with the Dirichlet process.

1 Theoretical aspects

The Poisson Dirichlet distribution is family of distributions over sequences of positive random variables that sum to one a.s.

$$(V_n)_{n\in\mathbb{N}} \text{ so that } \sum_n V_n = 1 \text{ a.s.} \tag{1}$$

Such variables can be used in numerous areas such as number theory, combinatorics, studies of population genetics or species, excursion intervals between zero of Brownian motion...etc...

1.1 First definition of the process

For any $0 \leq \alpha < 1$ and $\theta > -\alpha$, if for all natural n, $\tilde{Y_n}$ are i.i.d. variables following a $\beta(1-\alpha, \theta+n\alpha)$ distribution, then let's define

$$\tilde{V_1} = \tilde{Y_1}, \tilde{V_n} = (1-\tilde{Y_1})(1-\tilde{Y_2})...(1-\tilde{Y}_{n-1})\tilde{Y_n} \tag{2}$$

and let $V_1 > V_2...$ be the ranked values of the $(\tilde{V_n})$. Then the (V_n) are said to follow the $Poisson-Dirichlet$ distribution with parameters (α, θ) [1]. This definition can be linked to the Stick-breaking process defined in [2, p.883].

1.2 Stable subordinators

However, it happens that this definition meets another process : derivation from a subordinator. Let's have a quick view of a subordinator's definition and properties.

A continuous-time process is a family of real-valued variables $\{X_t = X(t)\}_{t\geq 0}$

This continuous-time process is a Lévy process if it matches three conditions [3]. First : its sample paths are right-continuous and have left limits at every time point t. Second : for any $t_0 < t_1 ... < t_k$ the increments $X(t_i) - X(t_{i-1})$ are independent.Third : for all $0 \leq s \leq t$ the random variables $X(t) - X(s)$ and $X(t-s) - X(0)$ have same distribution.

Finally, a real-valued Lévy process is called a subordinator if its sample paths are nondecreasing.

A subordinator is α-stable if it has initial value 0 and satisfies self-similarity property : i.e. for any t > 0, $(\frac{X_t}{t})^{1/\alpha} \overset{d}{=} X_1$

1.3 Description of the process with stable subordinators

For any positive random time T, we consider the sequence of ranked lengths of components intervals of the set [0,T]\Z, where Z is a random closed subset of [0,∞] with Lebesgue measure 0.

$$V_1(T) \geq V_2(T) \geq ... \tag{3}$$

Let's consider the case when Z is the closure of the range of a subordinator, let's denote it τ_s.

If τ_s follows the $\Gamma(s)$ distribution for all s, then for all $\theta > 0$

$$(\frac{V_1(\tau_\theta)}{(\tau_\theta)}, \frac{V_2(\tau_\theta)}{(\tau_\theta)}, ...) \text{ follows a } PD(0, \theta) \text{ distribution.} \tag{4}$$

If τ_s is a stable subordinator of index α, where 0 < α < 1, then for all $s > 0$

$$(\frac{V_1(\tau_s)}{(\tau_s)}, \frac{V_2(\tau_s)}{(\tau_s)}, ...) \text{ follows a } PD(\alpha, 0) \text{ distribution} \tag{5}$$

Something very interesting is that the proposition (5) remains valid with a fixed t > 0.

Such an definition will be convenient to study the link with Bessel processes in the second part of the study

1.4 Some results for PD(α,0)

Construction of $\mathbf{PD}(\alpha, \mathbf{0})$ with beta variables :

Suppose $(R_n)_{n\in\mathbb{N}}$ is a sequence of independant variables so that (R_n) follows $(n\alpha, 1)$ distribution with 0 < α < 1. Then we can define a $PD(\alpha, 0)$ with the sequence $(V_n)_{n\in\mathbb{N}}$ so that

$$V_1 = \frac{1}{1 + R_1 + R_1R_2 +_1 R_2R_3 + ...} \tag{6}$$

$$V_{n+1} = V_1R_1R_2...R_n. \tag{7}$$

Note that equations (6) and (7) allow for a slightly different definition of the stick breaking construction of the Pitman-Yor process, presented in 1.

Construction of $\mathbf{PD}(\alpha, \mathbf{0})$ with exponential variables :

Let $X_1 < X_2 < ...$ be the points of a PRM on $(0, \infty)$ so that $X_n = \epsilon_1 + \epsilon_2 + ... + \epsilon_n$ where the ϵ_i are independant standard exponential variables. Then $V_n = \frac{(X_n)^{-1/\alpha}}{\sum_m (X_m)^{-1/\alpha}}$ has $PD(\alpha, 0)$ distribution.

1.5 Extensions and results for PD(α,θ)

Let $0 < \alpha < 1$ and $\theta > -\alpha$. We have for any measurable function f

$$L := lim_{n\to\infty} nV_n^\alpha \text{exists a.s} \tag{8}$$

$$\mathbb{E}_{\alpha,\theta}[f(V_1, V_2...)] = C_{\alpha,\theta}E_{\alpha,\theta}[L^{\theta/\alpha}f(V_1, V_2...)] \tag{9}$$

where $C_{\alpha,\theta} = \frac{1}{E_{\alpha,0}(L^{\theta/\alpha})}$. and by reformulating this relation we can show that the $PD(\alpha, \theta)$ distributions are mutually continuous as θ varies.

Construction of $\mathbf{PD}(\alpha, \theta)$:

Let there be 0 < α < 1, and C > 0. Let there be two independant subordinators : one with Levy measure $\alpha Cx^{-\alpha-1}e^{-x}dx$ $(\tau_s, s \geq 0)$, and a gamma subordinator $(\gamma(t), t \geq 0)$. For any $\theta > 0$, consider the quantity $S_{\alpha,\theta} = \frac{\gamma(\frac{\theta}{\alpha})}{C*\Gamma(1-\alpha)}$. Then, by taking $T = \tau_{S_{\alpha,\theta}}$,

$$(\frac{V_1(T)}{(T)}, \frac{V_2(T)}{(T)}, ...) \text{ follows } PD(\alpha, \theta) \text{ distribution.}$$

1.6 Construction using the Chinese Restaurant Process

An alternative view of this distribution is using the the Chinese Restaurant Process (CRP). Suppose $P \sim PD(\alpha, \theta, H)$ and suppose $X_1, X_2, ...X_n | P \sim P$, then ([4, Proposition 9]):

$$P(X_{n+1}|X_1, X_2, ..., X_n) = \frac{\theta + \alpha|B|}{\theta + n} H + \frac{1}{\theta + n} \sum_{i=1}^{N} (1-\alpha)\delta_{x_i}. \tag{10}$$

where $|B|$ is the size of the partition of the $X_1, X_2...X_n$. Setting $\alpha = 0$ we recover the the Chinese Restaurant Process view of the Dirichlet Process. We have the following approximations for the expected number of tables produced by the Chinese Restaurant process ([5, p.68].)

$$\begin{cases} \mathbb{E}[K_n] \approx \theta \log(n) \quad \text{for} \quad \alpha = 0 \\ \mathbb{E}[K_n] \approx \theta \frac{\Gamma(\theta+1)}{\alpha\Gamma(\theta+\alpha)} n^{\alpha} \quad 0 < \alpha < 1 \end{cases} \tag{11}$$

which shows that the Pitman-Yor process is qualitatively different from the Dirichlet Process: the growth is a power law.

2 Sampling from the Pitman-Yor process

2.1 Sampling through the Chinese Restaurant Process

We first sample the Pitman Yor process through the Chinese Restaurant Process construction. We sample 1000 points from the CRP using equation (10) and average the results over 100 runs. In each case, we set $\theta = 1.0$. Results for $\alpha = 0$ are presented in figure 1. We observe that the empirical mean is estimated very well by the the estimates provided by equation (11). Figure 1 also compares the number of clusters generated for different values of α.

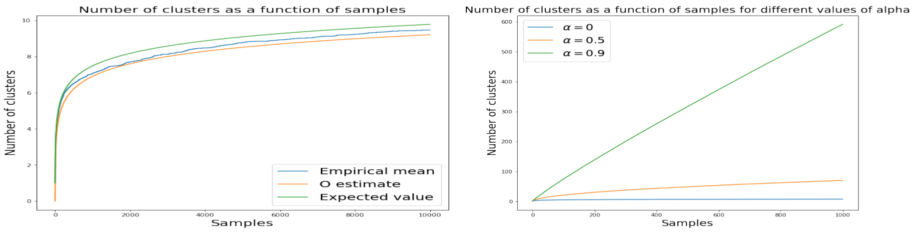

Figure 1: Number of clusters generated by the CRP with $\alpha = 0$ (left) and $\alpha = 0.5$ (middle). Number of clusters generated by the CRP for $\alpha = 0, 0.5, 0.9$. Observe that larger values of α lead to many more clusters.

2.2 Sampling through the stick breaking representation.

To approximate the stick-breaking construction we sample a finite number of weights $\tilde{V_n}$ which we hope will lead to a good approximation. We now consider the effect of the parameter α on the validity of this approximation. After 20 samples, for $\alpha = 0$ the weight sampled accounted for over 0.999 of the total mass, for $\alpha = 0.5$ they accounted for 0.865 and for $\alpha = 0.9$ only 0.24. This indicates a strong limitation for large values of α: a large discount means that many more samples have to be drawn for a finite approximation to be good. Figure 2 illustrates the results.

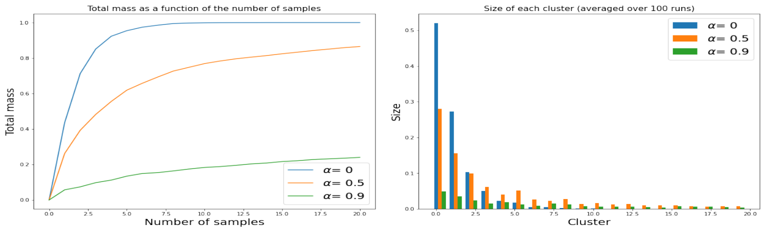

Figure 2: Total mass sampled (left) and the mass of each weight (right) for $\alpha = 0, 0.5, 0.9$.

2.3 Sampling with Bessel processes

In this part, we use that the distribution of the ordered length of excursions around 0 of a Bessel process of dimension $\delta = 2(1-\alpha)$ follows the Kingman law $PY(\alpha, 0)$ for $\alpha \in]0, 1[$.

2.3.1 Sampling Bessel processes

A Bessel process B_t of dimension $\delta > 0$ initialized at $B_0 = \gamma^2$, $\gamma \geq 0$ is defined as [6]:

$$\forall t, B_t = \sqrt{Z_t}, \tag{12}$$

with Z_t following the next PDE, W being a standard Brownian motion, initialized at $Z_0 = \gamma$:

$$dZ_t = \delta dt + 2\sqrt{|Z_t|}dW_t \tag{13}$$

Implicit Euler scheme The associated implicit Euler scheme is given by the following relation:

$$Z_{i+1} = (G_i + \sqrt{G_i^2 + \delta \times \delta_t + Z_{i-1}})^2, \text{ with } G_i \sim_{iid} \mathcal{N}(0, 1/\delta_t). \tag{14}$$

Theoretically, the scheme is unstable for all $\delta < 2$ [7]. Experimentally, with $\delta < 2$, the samples cannot be accepted for 2 reasons:

- with $\delta < 2$ the trajectories never come back to 0 (with ϵ-tolerance of 10^{-2});
- when $\delta = 1$, the Bessel process is equivalent to the Euclidean norm of a Brownian motion, and the comparison shows that the implicit Euler scheme does not follow the same dynamic.

Explicit scheme The associated explicit Euler scheme is given by the following relation:

$$Z_{i+1} = \delta \times \delta_t + 2\sqrt{|Z_{i-1}|}G_i + Z_{i-1}, \text{ with } G_i \sim_{iid} \mathcal{N}(0, 1/\delta_t). \tag{15}$$

Theoretically, this scheme is unstable for all $\delta > 0$ [8]. Nevertheless, experimentally, sampling Z allowing it to be negative and then generating $B_t = \sqrt{|Z_t|}$ gives exploitable results, better than the implicit scheme.

As a consequence, the explicit scheme is the one used in the following part.

2.3.2 Sampling Kingman distributions

General case The main technical difficulty is to detect when the trajectory hits 0 while we only have evaluations on a grid. The approach used is to consider that the trajectory hits 0 when it is below a chosen threshold.

This approach leads to 2 main drawbacks, that are highlighted in figure 3:

- visible on the left side of figure 3, the excursions of length 1 are too numerous;
- visible in the middle of figure 3, longer clusters are too few, because the separation between 2 clusters is often not detected.

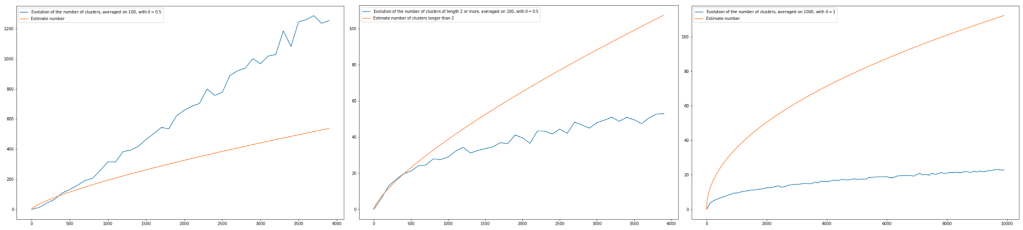

Figure 3: (left) Average number of clusters on 100 trajectories, in function of the number of points, with $\delta = 0.5$ in blue, the expected one in orange. (middle) Same plot about the number of clusters longer than 2. (right) Average number of clusters on 1000 trajectories, with $\delta = 1$ in blue, the expected one in orange.

Case $\delta = 1$, $i.e$ $\alpha = 1/2$ In this special case of $\delta = 1$, we can sample Bessel processes as the Euclidean norm of a Brownian motion. Manipulating directly the Brownian motion makes it possible to detect hits on 0 using change of sign.

The results are better in this case and the computations are much quicker. However, the average number of clusters is still below the theoretical expectation, as showed on the right side of figure 3, but it follows the correct big O.

In this state, sampling Kingman distributions using Bessel processes is not relevant due to numerical complexities. A different approach can be explored to detect excursions by computing the probability of hitting 0 between 2 steps of the grid, to take a decision or refine the trajectory with a Bessel bridge. However, this requires further analysis to analytically compute this probability.

3 Application: Collapsed Gibbs Sampling for Bayesian Mixture Models

In this section we consider a toy model where data is generated from a one-dimensional Gaussian Mixture Model: $x_i|z_i, \mu_i, \sigma \sim \mathcal{N}(\mu_i, \sigma)$ and $z_i \sim \pi$. We assume the following model: $\pi \sim GEM(\theta, \alpha), z_i \sim \pi, \mu_i \sim \mathcal{N}(0, 1), x_i \sim \mathcal{N}(\theta_{z_i}, \sigma)$ where $GEM(\theta, \alpha)$ is the CRP representation of the Pitman-Yor process. We assume that σ is known. We use collapsed Gibbs sampling with this model to estimate the number of clusters. The algorithm can be found in [2, p.887] and the derivation of the model can be found at [9]. We also modified the code at [9] [1].

We sample 300 points from a GMM with 20 classes. We run the collapsed Gibbs sampling for 100 iterations for two cases: $\alpha = 0$ and $\alpha = 0.1$. The results are presented in figure 4. Note that the Gibbs sampling with PY prior predicts many more clusters: the DP prior produces 9 classes whereas the PY prior produces 55. However, many of the clusters are the same between both priors. The PY prior produces many clusters with very few members (1-3) that are in the same group. This indicates that even with small values of α, the PY prior will produce many more classes, but its usefulness will depend on the application.

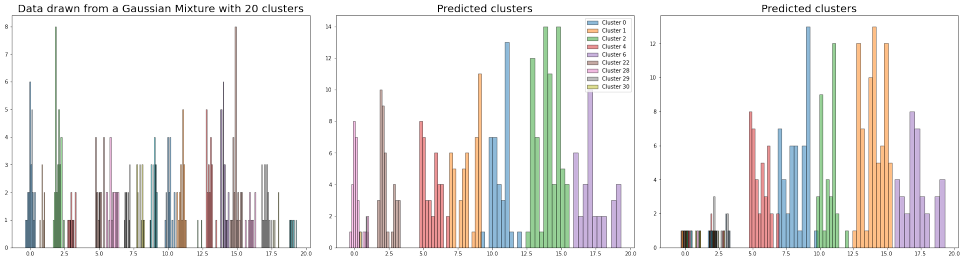

Figure 4: Original data (left), prediction given by the Gibbs sampling with DP prior (middle) and Pitman-Yor prior (right)

4 Conclusion

We presented the important theoretical results underlying the construction of the Pitman-Yor process and its properties. We showed different view of this process, which are useful for different applications. We then showed three ways to sample from the Pitman-Yor process: through the stick breaking representation, the Chinese Restaurant Process representation and through the Bessel process. We also noted that the Pitman-Yor process is qualitatively different from the Dirichlet process as it exhibits a power law growth as opposed to a logarithmic growth and illustrated this difference on a toy Gaussian mixture model.

The code can be found at [10].

[1] Warning: the code present in [9] contains some major errors which should be corrected before use

References

[1] Jim Pitman and Marc Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. In *The Annals of Probability*, volume 25, pages 855–900, 1997.

[2] Kevin Murphy. *Machine Learning, A probabilistic perspective*. MIT Press, 2012.

[3] Steven P. Lalley. Levy processes, stable processes, and subordinator. *University of Chicagor*, 2017.

[4] Jim Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory*, 102:145 – 158, 1995.

[5] James Pitman. *Combinatorial Stochastic Processes*. 2006.

[6] M. Jeanblanc and T. Simon. *Eléments de calcul stochastique*. 2005.

[7] N. Ikeda and S. Watanabe. *Stochastic Differential Equations and Diffusion Processes*. 1991.

[8] Gilles Pagès. Lecture notes in numerical probabilities, master finance and probabilities upmc-polytechnique, 2020.

[9] Tim Hopper. Notes on dirichlet process. `https://github.com/tdhopper/notes-on-dirichlet-processes`, 2015.

[10] Matthieu Darcy Antoine Doize, Benoit Oriol. Bmlpitmanyor. `https://github.com/MatthieuDarcy/BMLPitmanYor`, 2021.