# Bayesian nonparametrics

## 4 Pitman-Yor process

Let $X_1, X_2, \ldots$ be a stochastic process on a Polish space $(\mathcal{X}, \mathcal{B})$ with the following predictive distribution, known as the *two-parameter Ewens sampling formula*,

$$\mathbb{P}(X_{n+1} \in A \mid X_1, \ldots, X_n) = \frac{b + aM_n}{b+n} H(A) + \sum_{i=1}^{n} \frac{1 - (a/n_{X_i})}{b+n} \delta_{X_i}(A)$$

for some constants $a \geq -b$, $b \geq 0$, and a base probability distribution $H$. Here, $M_n$ is the number of distinct samples or clusters in $X_1, \ldots, X_n$, and $n_{X_i}$ is the number of samples in $X_1, \ldots, X_n$ that are identical to $X_i$.

This predictive distribution is a mixture of $H$ and point masses at each of the previous samples. This allows us to rephrase it as a generalized Blackwell-MacQueen urn.

**Generalized Blackwell-MacQueen urn.** We start with a single black ball with weight $b$ in the urn. At each step, sample one ball with probability proportional to its weight. Then,

- If the ball is black, replace it in the urn along with another black ball of weight $a$. Sample a color from $H$, and put a ball of that color and weight $1 - a$ in the urn.
- If the ball is of a color, replace it in the urn along with another ball of the same color and weight 1.

The sequence of colors sampled has the same distribution as $X_1, X_2, \ldots$.

**Remark.** It is easy to verify that when $a = 0$, the predictive distribution and the urn scheme above correspond to a simple Blackwell-MacQueen urn. What distinguishes the generalized process when $a > 0$ is that every time we sample a black ball — whenever we sample a new color or discover a new cluster — we reinforce the chance of doing this again in the future. When $a = 0$, the probability that $X_n$ is a new color given $X_1, \ldots, X_n$ only depends on $n$. When $a > 0$, it also depends on the number of distinct colors in $X_1, \ldots, X_n$.

**Theorem 4.1.** *The process $X_1, X_2, \ldots$ is exchangeable.*

*Proof.* The probability that $X_1, \ldots, X_n$ is a specific sequence which has $m$ clusters of sizes $n_1, \ldots, n_m$, $\sum_{i=1}^m n_m = n$ is

$$\frac{(b)_{a\uparrow m} \prod_{i=1}^m (1-a)_{1\uparrow n_i - 1}}{(b)_{1\uparrow n}},$$

where we use the notation $(x)_{y\uparrow n} = x(x+y)(x+2y)\ldots(x+(n-1)y)$. This can be proven by induction on $n$. This is clearly invariant to changing the order of the sequence $X_1, \ldots, X_n$. □

**Corollary 4.2.** *By de Finetti's theorem, there exists a random measure $P$ on $(\mathcal{X}, \mathcal{B})$ such that, given $P$, the process $X_1, X_2, \ldots$ is i.i.d. from $P$.*

We will call $P$ a *Pitman-Yor process* with concentration parameter $b$, discount parameter $a$, and base distribution $H$.

**Lemma 4.3.** *The process $P$ is a.s. discrete, i.e. $P = \sum_{i=1}^\infty \pi_i \delta_{\theta_i}$, where $\theta_1, \theta_2, \ldots$ is i.i.d. from $H$.*

*Proof.* The proof proceeds in the same way as for the Dirichlet Process. The probability that $X_n$ is never observed again in $X_{n+1}, X_{n+2}, \ldots$ is of the form

$$\prod_{i=1}^\infty \frac{C}{B+i} = 0,$$

for some constants $B$ and $C$. This is a consequence of the predictive distribution. This implies that with probability 1, every observation in $X_1, X_2, \ldots$ is repeated at least once. This contradicts the fact that $P$ has a diffuse component with some positive probability. □

The weights on the point masses of a Pitman-Yor process can be generated through a stick breaking process. Namely, if $\gamma_i \sim \text{Beta}(1-a, b+ia)$ are independent for $i = 1, 2, \ldots$, and

$$\pi_i = \gamma_i (1-\gamma_{i-1}) \ldots (1-\gamma_1),$$

then the measure $\sum_{i=1}^\infty \pi_i \delta_{\theta_i}$ where $\theta_1, \theta_2, \ldots$ is i.i.d. from $H$ is a Pitman-Yor process. This representation can be proven using the same argument outlined in the first problem of Homework 1 for the Dirichlet Process. This proof method yields a stronger result, namely, that the distribution of $\pi_1, \pi_2, \ldots$, in that order, is the distribution of a size-biased permutation of the weights of a Pitman-Yor process.

### 4.1 The Pitman-Yor Process vs. the Dirichlet Process

The stick breaking representation allows us to understand the difference between the case $a = 0$ and $a > 0$ from a new perspective. Consider the $k$th weight in the stick breaking sequence,

$$\pi_k = \gamma_k(1 - \gamma_{k-1}) \dots (1 - \gamma_1).$$

We can compute the expectation of this weight, since each factor in the product is independent and Beta-distributed,

$$E\pi_k = E\gamma_k \prod_{i=1}^{k-1} E(1 - \gamma_i) = \frac{1 - a}{1 + b + (k - 1)a} \prod_{i=1}^{k-1} \frac{b + ia}{1 + b + (i - 1)a}.$$

When $a = 0$, this becomes

$$E\pi_k = \frac{1}{1 + b} \left[ \frac{b}{1 + b} \right]^{k-1},$$

So, on average, $\pi_k$ decreases exponentially with $k$. On the other hand, when $a > 0$,

$$E\pi_k = C(a, b) \frac{\Gamma(b/a + (k - 1))}{\Gamma((b + 1)/a + (k - 1))} = C(a, b)(k-1)^{-1/a} \left[ 1 + O\left( \frac{1}{k - 1} \right) \right],$$

for some constant $C(a, b)$. Therefore, the tails of the sequence $\pi_1, \pi_2, \dots$ are on average those of a power law. Allowing $a$ to be greater than 0 makes it possible to increase the dispersion of the distribution of weights beyond that of an exponential distribution; the higher $a$, the heavier the tails of the distribution.

This flexibility is important in applications because several real-world distributions, such as the rates of word usage in natural language, have heavier tails than an exponential.