Appendix A
Joint marginal likelihood

Let K be the number of features and D the number of observations. As detailed in Section 2.1, the conditional likelihood for the finite-dimensional three-parameter IBP is obtained by activating $K^* \leqslant K$ features, and then sampling the first observations for which the feature is active according to a uniform distribution and the subsequent ones according to a Beta-Bernoulli distribution. The variable K^* is distributed according to a Binomial distribution and the remaining entries follow a Beta-Bernoulli distribution. The chain rule of probability leads to the following marginal likelihood:

$$p(\bar{\mathbf{\Theta}}) = \mathrm{Binomial}\left(K^* | \frac{\varepsilon}{K}, K\right) \times \prod_{k \leqslant K^*} \mathcal{BB}\left(\bar{\theta}_{k\cdot} - 1 | D-1, 1-\sigma+\frac{\eta\delta}{K}, \delta+\sigma\right),$$

where $\mathcal{BB}(k|n,\alpha,\beta)$ is the Beta-Bernoulli distribution, which is defined as follows:

$$\mathcal{BB}(k|n,\alpha,\beta) = \frac{\Gamma(\alpha+\beta)\Gamma(k+\alpha)\Gamma(n-k+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)} \ .$$

Hence, in the limit of large K, the binomial tends to a Poisson distribution and the joint likelihood is independent of K:

$$\lim_{K\to\infty} P(\bar{\mathbf{\Theta}}) = \exp\left(-\eta \sum_{j=0}^{D-1} \frac{\Gamma(1+\delta)\Gamma(j+\delta+\sigma)}{\Gamma(j+1+\delta)\Gamma(\delta+\sigma)}\right) \eta^{K^*} \times \prod_{k\leqslant K^*} \frac{\Gamma(1+\delta)\Gamma(D-\bar{\theta}_{k\cdot}+\delta+\sigma)\Gamma(\bar{\theta}_{k\cdot}-\sigma)}{\Gamma(1-\sigma)\Gamma(\delta+\sigma)\Gamma(D+\delta)},$$

which is the same expression as the one found in [**?**]. It is straightforward to show that the marginal likelihood of the two-parameter IBP is recovered for $\sigma = 0$.

Appendix B
On the hierarchical Pitman-Yor process

HPY-LDA is an extension of HDP-LDA where the DP priors (1-2) are replaced by PYP priors, i.e., $H \sim \mathrm{PYP}(\alpha_0 H_0, \sigma_0)$ and $G_d \sim \mathrm{PYP}(\alpha H, \sigma)$, where σ_0 and σ denote the discount parameters. It should be noted that this model is only able to capture power-law distributions in the topics, but not in the words. We implemented a Chinese restaurant franchise Gibbs sampler for HPY-LDA by modifying Teh's code for HDP-LDA. We fixed the discount parameters σ_0 and σ respectively to 0 and 0.25. We also tried the values $\sigma_0 = \sigma = 0.5$ following [**?**], as well as $\sigma_0 = 0$ and $\sigma = 0.1$, but they led to worse performances and we do not report the results here. Other details of the experimental setup are identical to that of HDP-LDA described in Section 5.3. It should be noted that we do not consider a truncated version of HPY-LDA like [**?**]. Moreover, we believe their sampler (Algorithm 6, specifically) is incorrect. While sampling the topic for a word, we would have to first choose a topic and then choose a table serving that topic, which would require additional book-keeping.