Supplementary Material: Ordered Stick-Breaking Prior for Sequential MCMC Inference of Bayesian Nonparametric Models

Mrinal Das† MRINAL@CSA.IISC.ERNET.IN
Trapit Bansal† TRAPIT@CSA.IISC.ERNET.IN
Chiranjib Bhattacharyya† CHIRU@CSA.IISC.ERNET.IN
†Department of Computer Science and Automation,
Indian Institute of Science, Bangalore, India

S.1. Introduction

We discuss some relevant mathematical background first in Section S.2, those are directly used in the paper or proofs. We include some examples and properties related to OSBP and PPFs of OSBP in Section S.3. There are two lemmas and four theorems in the paper. We prove them here in Sections S.4, S.5, S.6 related to OSBP, PPF ofOSBP and SUMO respectively. We additionally provide one theorem (Theorem A) and a lemma (Lemma 1) which are strongly related to OSBP but could not be included in the paper due to space constraint. Then we provide construction of dependency over mini-batches using OSBP on PYP, SBP, HDP in Section S.7. Finally we give inference details for DPMM for text datasets in Section S.8.

S.2. Mathematical background

This is not a comprehensive review, we cover only thos definitions and properties that will be referred later in this material.

S.2.1. Gamma distribution and Gamma process

Definition 1. *(Gamma Distribution). A non-negative real-valued random variable* X *is said to have a Gamma distribution with shape parameter* α *and scale parameter* β*, denoted by* $\mathrm{X} \sim Gamma(\alpha;\beta)$*, if its probability density function is given by*

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \tag{1}$$

Proposition 1. *Let* $\mathrm{X}_1, \mathrm{X}_2, \ldots$ *be a countable collection of independent Gamma distributed variables as* $\mathrm{X}_k \sim Gamma(\alpha_k;\beta)$*. Then*

$$\sum_{k=1}^{\infty} \mathrm{X}_k \sim Gamma(\sum_{k=1}^{\infty}\alpha_k, \beta) \tag{2}$$

Definition 2. *(Gamma process) A random measure* G *on* Ω *is called a Gamma process with base measure* H *and scale parameter* α*, denoted by* $G \sim \Gamma P(\alpha \mathrm{H})$*, if it satisfies*

- *for each measurable subset* $\mathrm{A} \in \mathcal{B}$*,* $G(\mathrm{A})$ *has a Gamma distribution as* $\mathrm{G}(\mathrm{A}) \sim Gamma(\alpha\mathrm{H}(\mathrm{A}))$*, and*
- G *is completely random*

Proposition 2. *If* $\mathrm{G}_j \sim \Gamma P(\alpha_j \mathrm{H}_j)$ *for* $j = 1,\ldots,k$*, then* $\sum_{j=1}^{k}\mathrm{G}_j \sim \Gamma P(\sum_{j=1}^{k}\alpha_j\mathrm{H}_j)$*.*

S.2.2. Dirichlet distribution and Dirichlet process

Let $\mathbf{S}_d$ denote the probability simplex in the d-dimensional real vector space $\mathbf{R}_d$, as

$$\mathbf{S}_d = \{(x_1,\ldots,x_d) \in \mathbf{R}_d : x_i \geq 0, \forall i; \sum_{i=1}^{d} x_i = 1\} \tag{3}$$

Definition 3. *(Dirichlet distribution) An* $\mathbf{S}_d$*-valued random variable* X *is said to have a Dirichlet distribution, denoted by* $\mathrm{X} \sim Dir(\alpha_1,\ldots,\alpha_d)$ *with* $\alpha_1,\ldots,\alpha_d > 0$*, if it has a probability density function given by*

$$f(x_1,\ldots,x_d;\alpha_1,\ldots,\alpha_d) = \frac{\Gamma(\sum_{i=1}^{d}\alpha_i)}{\prod_{i=1}^{d}\Gamma(\alpha_i)} x_i^{\alpha_i-1} \tag{4}$$

Proposition 3. *Let* $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_k$ *be* k *independent Gamma distributed variables as* $\mathrm{X}_j \sim Gamma(\alpha_j;\beta)$*. Then for* $\mathrm{Y}_j = \frac{\mathrm{X}_j}{\sum_{j=1}^{k}\mathrm{X}_j}$*,* $(\mathrm{Y}_1,\ldots,\mathrm{Y}_k) \sim Dirichlet(\alpha_1,\ldots,\alpha_k)$*.*

Proof. This can be seen using the procedure of transformation of random variables. □

Proposition 4. *If* $(x_1,\ldots,x_{k-1},x_k) \sim Dirichlet(\alpha_1,\ldots,\alpha_{k-1},\alpha_k)$*, then* $(y_1,\ldots,y_{k-1}) \sim Dirichlet(\alpha_1,\ldots,\alpha_{k-1})$*, where* $y_j = \frac{x_j}{\sum_{l=1}^{k-1}x_l}$ *for* $j = 1,\ldots,k-1$*.*

Proof. Let $\mathrm{Z}_1, \ldots, \mathrm{Z}_k$ be k independent variables such that $\mathrm{Z}_j \sim Gamma(\alpha_j, \beta)$ for $j = 1, \ldots, k$. Then $(x_1, \ldots, x_k) \sim Dirichlet(\alpha_1, \ldots, \alpha_k)$, for $x_j = \frac{\mathrm{Z}_j}{\sum_{l=1}^{k} \mathrm{Z}_l}$ using Proposition 3.

We can write,

$$y_j = \frac{x_j}{\sum_{l=1}^{k-1} x_l} = \frac{\frac{\mathrm{z}_j}{\sum_{r=1}^{k} \mathrm{z}_r}}{\sum_{l=1}^{k-1} \frac{\mathrm{z}_l}{\sum_{r=1}^{k} \mathrm{z}_r}} = \frac{\mathrm{Z}_j}{\sum_{l=1}^{k-1} \mathrm{Z}_l} \tag{5}$$

Thus, by Proposition 3, $(y_1, \ldots, y_{k-1}) \sim Dirichlet(\alpha_1, \ldots, \alpha_{k-1})$. □

Definition 4. *(Dirichlet Process). Let,* H *is a probability measure over a measurable space* $(\Omega, \mathcal{B})$*, and* γ *is a positive real number. A random measure* G *on* Ω *is called a Dirichlet process with base measure* H*, denoted by* $\mathrm{G} \sim DP(\gamma, \mathrm{H})$ *if for any finite measurable partition* $(\mathrm{B}_1, \mathrm{B}_2, ..., \mathrm{B}_k)$ *of* Ω*,*

$$(\mathrm{G}(\mathrm{B}_1), ..., \mathrm{G}(\mathrm{B}_k)) \sim Dirichlet(\gamma\mathrm{H}(\mathrm{B}_1), ..., \gamma\mathrm{H}(\mathrm{B}_k)) \tag{6}$$

Stick-breaking representation of DP. (Sethuraman, 1994) proposed a stick-breaking construction of DP such that if $\mathrm{G} \sim DP(\gamma, \mathrm{H})$, then

$$\mathrm{G} = \sum_{j=1}^{\infty} \theta_j \delta_{\beta_j},\ \beta_j \sim \mathrm{H} \tag{7}$$

δ_{β_j} denotes an atomic distribution where the entire probability mass is concentrated at β_j. $\{\theta_j\}$ are constructed as follows.

$$\theta_1 = v_1, \theta_j = v_j \prod_{l=1}^{j-1}(1 - v_l),\ v_j \sim Beta(1, \gamma) \tag{8}$$

The above construction can be understood as breaking a unit length stick using stick fractions v_j. (Sethuraman, 1994) showed that $\sum_{j=1}^{\infty} \theta_j = 1$ when $\{\theta_j\}$ are constructed as above. Often θ is said to be distributed as $GEM(\gamma)$.

S.2.3. Generalized Dirichlet distribution

Definition 5. *(Generalized Dirichlet distribution) An* $\mathbf{S}_k$*-valued random variable* X *is said to have a generalized Dirichlet distribution, denoted by*

$$\mathrm{X} \sim GDD(a_1, b_1, \ldots, a_{k-1}, b_{k-1}) \tag{9}$$

with $a_j, b_j > 0,\ \forall j$ *if it has a probability density function given by*

$$f(x_1, \ldots, x_k; a_1, b_1, \ldots, a_{k-1}, b_{k-1}) = \left(\prod_{j=1}^{k-1} B(a_j, b_j)\right)^{-1} x_k^{b_{k-1}-1} \prod_{j=1}^{k-1} \left(x_j^{a_j - 1} \left(\sum_{i=j}^{k} x_i\right)^{b_{j-1} - a_j - b_j} \right) \tag{10}$$

where $x_k = 1 - \sum_{j=1}^{k} x_j$. $B(a_j, b_j) = \frac{\Gamma(a_j)\Gamma(b_j)}{\Gamma(a_j + b_j)}$.

Example. Let $k = 4$, then the density function of (x_1, x_2, x_3, x_4) is

$$\left(\prod_{j=1}^{3} B(a_j, b_j)\right)^{-1} x_1^{a_1-1} x_2^{a_2-1} x_3^{a_3-1} x_4^{b_3-1} (x_2 + x_3 + x_4)^{b_1 - a_2 - b_2} (x_3 + x_4)^{b_2 - a_3 - b_3} \tag{11}$$

Proposition 5. *By setting* $b_{j-1} = a_j + b_j,\ 2 \le j \le k - 1$ *(*b_0 *is arbitrary),* $\mathrm{X} \sim GDD(a_1, b_1, \ldots, a_{k-1}, b_{k-1})$ *is equivalently* $\mathrm{X} \sim Dirichlet(a_1, a_2, \ldots, a_{k-1}, b_{k-1})$*.*

Proof. This follows directly from Eq. (10) and Eq. (4). □

S.2.4. Stick-breaking process

We have defined SBP in the paper, however we re-iterate the discussion to show one useful result in Lemma A regarding SBP.

Any almost sure (a.s.) discrete probability measure G is a stick-breaking process (SBP) (Ishwaran & James, 2001) if it can be represented as

$$\mathrm{G} = \sum_{j=1}^{\infty} \theta_j \delta_{\beta_j}, \theta_1 = v_1,\ \theta_j = v_j \prod_{l=1}^{j-1}(1 - v_l)$$
$$a_j, b_j > 0,\ v_j \sim Beta(a_j, b_j),\ \ \beta_j \sim \mathrm{H} \tag{12}$$

H is a *diffuse measure* over a measurable space $(\Omega, \mathcal{B})$ and $\{a_j, b_j\}$ are set of parameters.

The following lemma gives a condition over $\{a_j, b_j\}$ so that $\sum_{j=1}^{\infty} \theta_j = 1$ a.s.

Lemma A. *(Ishwaran & James, 2001). For the random weights in an SBP,* $\sum_{j=1}^{\infty} \theta_j = 1$ *a.s. iff* $\sum_{j=1}^{\infty} \mathbb{E}[\log(1 - v_j)] = -\infty$*. Alternatively, it is sufficient to check that* $\sum_{j=1}^{\infty} \log(1 + \frac{a_j}{b_j}) = +\infty$*.*

Proof. See appendix by Ishwaran & James (2001). □

Important special cases. SBP subsumes many well known BNP priors. When $a_j = 1$ and $b_j = \gamma$ for all j, SBP becomes $DP(\gamma, \mathrm{H})$ following the constructive definition of Dirichlet process by Sethuraman (1994). Another popular BNP prior, the two parameter Poisson-Dirichlet process or Pitman-Yor process (PYP) (Pitman & Yor, 1997) can also be obtained as a special case when $a_j = 1 - \lambda$ and $b_j = \gamma + j\lambda$ for all j. There are many other existing priors which are special cases of SBP, see (Ishwaran & James, 2001) for a detailed discussion.

S.3. Appearance in order and OSBP

In this section, we first give an example of appearance in order phenomenon, and then we recall the definition of OSBP, followed by one essential information about OSBP.

S.3.1. Example of appearance in order

Here, we give an example of the appearance in order phenomenon defined in Section 2.1.

Let, $t = 9$ and (Y_i) is $(a, a, b, a, c, a, b, a, a)$. Notice that, $k_9 = 3$ with $\{a, b, c\}$ as $\{\bar{\mathrm{Y}}_1, \bar{\mathrm{Y}}_2, \bar{\mathrm{Y}}_3\}$. Now we have

$$\mathrm{B}_1 = \{1, 2, 4, 6, 8, 9\}, \mathrm{B}_2 = \{3, 7\}, \text{ and } \mathrm{B}_3 = \{5\}$$

Then we say that it is appearing in order as

$$[9] - \mathrm{B}_1 = \{3, 7, 5\} \Rightarrow 3 \in \mathrm{B}_2$$
$$[9] - (\mathrm{B}_1 \cup \mathrm{B}_2) = \{5\} \Rightarrow 5 \in \mathrm{B}_3$$

whereas if $\bar{\mathrm{Y}} = \{a, c, b\}$ then

$$\mathrm{B}_1 = \{1, 2, 4, 6, 8, 9\}, \mathrm{B}_2 = \{5\} \text{ and } \mathrm{B}_3 = \{3, 7\}$$

is *not* appearing in order as

$$[9] - \mathrm{B}_1 = \{3, 7, 5\} \text{ but } 3 \notin \mathrm{B}_2$$

S.3.2. Definition of OSBP

As we will be referring to OSBP in later parts of this material, for the sake of easy reading we present them here again.

Let Γ be a *diffuse* probability measure over random measures, and $\boldsymbol{\mu}, \boldsymbol{\nu}$ denote the set of scalar hyper-parameters $\{\mu_j\}$ and $\{\nu_j\}$ respectively such that $0 < \mu_j < 1$, $\nu_j > 0$, $\forall j$. $(\mathrm{G}_1, \mathrm{G}_2, \ldots)$ is an *appearing in order* sequence of random measures. $(\mathrm{Q}_1, \ldots, \mathrm{Q}_{k_{t-1}})$ is the set of k_{t-1} unique values among $\mathrm{G}_{1:t-1}$. We define, $\mathrm{G}_1, \mathrm{G}_2, \ldots \sim$ OSBP$(\boldsymbol{\mu}, \boldsymbol{\nu}, \Gamma)$ if $\mathrm{G}_1 \sim \Gamma$ and for any $t \geq 2$, the following holds:

$$\begin{aligned} \mathrm{G}_t \,|\mathrm{G}_{1:t-1}, (\rho_j), \Gamma \ &\sim \textstyle\sum_{j=1}^{k_{t-1}} \rho_j \delta_{\mathrm{Q}_j} + \alpha_{k_{t-1}} \Gamma \\ \rho_1 = v_1, \ \ \forall j > 1, \ \rho_j &= v_j \textstyle\prod_{l=1}^{j-1} (1 - v_l) \\ v_j | \mu_j, \nu_j &\sim Beta\left(\mu_j \nu_j, (1 - \mu_j) \nu_j\right) \\ \alpha_{k_{t-1}} &= 1 - \textstyle\sum_{j=1}^{k_{t-1}} \rho_j \end{aligned} \tag{13}$$

S.3.3. Diffuse base measure of OSBP ensures appearance in order

The need of the base measure Γ to be a diffuse measure is explained with the following Theorem.

Theorem A. *The samples from OSBP,* $(\mathrm{G}_1, \mathrm{G}_2, \ldots) \sim OSBP(\boldsymbol{\mu}, \boldsymbol{\nu}, \Gamma)$ *will be appearing in order almost surely iff the base measure of OSBP,* Γ *in Eq.* (13) *is a diffuse probability measure.*

Proof. When Γ is diffuse, for any two samples Q_j and Q_l sampled from Γ will be almost sure distinct iff $j \neq l$. By definition of OSBP, if for any t, $\mathrm{G}_t \sim \Gamma$ then $k_t = k_{t-1} + 1$ and $\mathrm{Q}_{k_t} = \mathrm{G}_t$. As Γ is diffuse measure, Q_{k_t} is a.s. distinct from all Q_j, $j < k_t$. Thus $\mathrm{G}_t \neq \mathrm{G}_l$ for all $l < t$. Hence, $[t] \backslash \cup_{l=1}^{k_t - 1} \mathrm{B}_l = t$ and $\mathrm{B}_{k_t} = [t]$.

We show the sufficient condition by contradiction. Suppose, Γ is atomic. Let $t = 4$, $k_4 = 2$, and $\mathrm{Q}_2 \neq \mathrm{Q}_1$. There are two partitions B_1 and B_2. Now when G_5 is sampled suppose it is sampled from Γ, then $\mathrm{Q}_3 = \mathrm{G}_5$. Then by definition of appearance in order $[5] \backslash (\mathrm{B}_1 \cup \mathrm{B}_2)$ should be in B_3 which is $[5]$. As Γ is atomic let $\mathrm{Q}_3 = \mathrm{Q}_1$. Then G_5 becomes equal to G_1 and so $5 \in \mathrm{B}_1$ and $[5] \backslash (\mathrm{B}_1 \cup \mathrm{B}_2) = \emptyset$. Contradiction. So, whenever a Q_j is sampled from Γ, k_t must increase. k_t will increase iff Q_j is different from all Q_l, $l < j$. Hence Γ has to be a diffuse probability measure. □

This is a slightly strict condition on the base measure than that for DP and PYP which also points out one key difference with the common BNP priors such as DP, PYP.

S.3.4. Comparison with DP and PYP on PPF.

DP and Pitman-Yor process (PYP) (Pitman & Yor, 1997) are the only other two existing SBP class of priors possessing PPFs. It is worth to note the difference of OSBP from DP, PYP in terms of PPFs due to modeling *appearance in order*. Recall that, PPF $(\pi_j, j \in [k_{t-1}]$ and $\sigma_{k_{t-1}})$ are defined by Pitman (1996) as

$$\begin{aligned} \pi_j &= p(z_t = j | z_{1:t-1}, \Theta), \ j \in [k_{t-1}], \\ \sigma_{k_{t-1}} &= p(z_t = k_{t-1} + 1 | z_{1:t-1}, \Theta) \end{aligned} \tag{14}$$

where Θ denotes the set of hyper-parameters. The PPFs corresponding to $DP(\gamma, \mathrm{H})$, also popularly referred as Chinese restaurant process (CRP) are

$$\begin{aligned} \pi_j &= \tfrac{g_j}{\gamma + t - 1}, \ j \in [k_{t-1}], \\ \sigma_{k_{t-1}} &= \tfrac{\gamma}{\gamma + t - 1} \end{aligned} \tag{15}$$

where $g_j = |B_j|$, and $B_j = \{i | z_i = j\}$. Thus, $\sum_{j=1}^{k_{t-1}} g_j = t - 1$ and $\sum_{j=1}^{k_{t-1}} \pi_j + \alpha_{k_{t-1}} = 1$. Similarly, PPFs of $PYP(a, b, \mathrm{H})$ $(0 \leq a < 1$ and $b > -a)$ are

$$\begin{aligned} \pi_j &= \tfrac{g_j - a}{b + t - 1}, \ j \in [k_{t-1}], \\ \sigma_{k_{t-1}} &= \tfrac{b + a k_{t-1}}{b + t - 1} \end{aligned} \tag{16}$$

Note that, $\sum_{j=1}^{k_{t-1}} g_j - a = t - 1 - a k_{t-1}$ and hence $\sum_{j=1}^{k_{t-1}} \pi_j + \alpha_{k_{t-1}} = 1$. By using $a = 0$ and $b = \gamma$, PYP becomes equivalent to DP.

Notice from Theorem 3 that, π_j for OSBP can be written as $a_j \prod_{l=1}^{j-1} (1 - a_l)$, where $a_j = \frac{\mu_j \nu_j + m_j - 1}{\nu_j + m_j + r_j - 1}$. From Lemma 2, a_j is the posterior expectation of v_j conditioned on $\mathrm{G}_{1:t-1}$. Thus in OSBP, the probability of joining partition B_j directly depends on *not* joining the partitions $\{\mathrm{B}_1, \mathrm{B}_2, \ldots, \mathrm{B}_{j-1}\}$. Whereas, in case of DP and PYP the probability of joining partition B_j depends only on the size of B_j and the probabilities of joining partitions only loosely depend because of summing up to one.

Moreover, Corollary 1 of (Lee et al., 2013) shows that the only PPF which lead to an exchangeable sequence are those for which π_j is a function of partition B_j only. This is true for DP and PYP but not for OSBP. In Case of OSBP, π_j is a function of all the existing partitions (B_j).

An important implication of this interpretation is that even though the partitions (and hence atoms) can be assumed to appear in order for DP and PYP, the effect of ordering of the partitions is lost due to the symmetric nature of this function which leads to an exchangeable partition probability function (EPPF by (Pitman, 1996)), or equivalently an exchangeable sequence illustrating another view why DP and PYP are not suitable to model appearance in order.

S.3.5. Example related to Theorem 2

Example. Let $\mu_j > 1/2, \ \forall j$, and $\epsilon = 0.01$. For $k = 14$, $\alpha_k \leq 0.01$ with probability more than 0.99.

S.4. Proofs related to OSBP

S.4.1. Proof of Theorem 1

Theorem 1. *If* $\mathrm{P}_1 = \Gamma$, $\mathrm{P}_t = \sum_{j=1}^{k_{t-1}} \rho_j \delta_{\mathrm{Q}_j} + \alpha_{k_{t-1}}\Gamma$ *for* $t > 1$ *and* $\mathrm{P}^* = \sum_{j=1}^{\infty} \rho_j \delta_{\mathrm{Q}_j}$ *such that* $\sum_{j=1}^{\infty} \rho_j = 1$, *where* (ρ_j), (Q_j), α_{k_t} *and* Γ *as defined in Eq.* (13) *with parameter* $\boldsymbol{\mu}, \boldsymbol{\nu}$, *then* $\lim_{t\to\infty} \mathrm{P}_t = \mathrm{P}^*$ *a.s.*

Proof. By definition, k_t is the cardinality of the set $(\mathrm{Q}_1, \mathrm{Q}_2, \ldots, \mathrm{Q}_{k_t})$. So for any $t > 0$, $k_t = k_{t-1}$ if no new atom is sampled, and $k_t = k_{t-1} + 1$ if a new atom is sampled from the base measure Γ. From Eq. (13), the probability of $k_t = k_{t-1} + 1$ is $\alpha_{k_{t-1}}$ and probability of $k_t = k_{t-1}$ is $\sum_{j=1}^{k_{t-1}} \rho_j$ which by definition is $1 - \alpha_{k_{t-1}}$. Hence, we get

$$k_{t-1} \leq k_t \quad a.s. \tag{17}$$

As, $k_{t+1} \geq k_t$ a.s., and $\alpha_{k_t} = 1 - \sum_{j=1}^{k_t} \rho_j$ by definition, with $\rho_j > 0$ a.s. for all j, we get

$$\alpha_{k_{t-1}} \geq \alpha_{k_t} \quad a.s. \tag{18}$$

$k_t \geq k_{t-1}$ and not bounded above. For any $K > 0$, there is a t' such that $k_{t'} > K$, otherwise K is the upperbound of k_t. So we can say

$$\lim_{t\to\infty} k_t = \infty \quad a.s. \tag{19}$$

On the other hand, $\alpha_{k_t} \leq \alpha_{k_{t-1}}$ and bounded below by zero. For any $\epsilon > 0$ there is a t' such that $\alpha_{k_{t'}} < \epsilon$, otherwise ϵ is the lower bound of α_{k_t}. Hence,

$$\lim_{t\to\infty} \alpha_{k_t} = 0 \quad a.s. \tag{20}$$

Thus, we can write $\lim_{t\to\infty} \mathrm{P}_t = \lim_{t\to\infty} \sum_{j=1}^{k_t} \rho_j \delta_{\mathrm{Q}_j} + \lim_{t\to\infty} \alpha_{k_t}\Gamma = \lim_{k_t\to\infty} \sum_{j=1}^{k_t} \rho_j \delta_{\mathrm{Q}_j} + \lim_{t\to\infty} \alpha_{k_t}\Gamma = \sum_{j=1}^{\infty} \rho_j \delta_{\mathrm{Q}_j} = \mathrm{P}^*$. That proves the Theorem.

□

Corollary 1. *For* $t \in \mathbf{N}$ *and* α_{k_t} *as defined in OSBP,* $\lim_{k_t\to\infty} \alpha_{k_t} = 0$ *a.s.*

Proof. This corollary is immediate from the above result. However we give one alternative proof here. Note that $(1 + \frac{\mu_j\nu_j}{(1-\mu_j)\nu_j}) > 1$, hence $\sum_{j=1}^{\infty} \log(1 + \frac{\mu_j\nu_j}{(1-\mu_j)\nu_j}) = +\infty$. By Lemma A it follows that $\sum_{j=1}^{\infty} \rho_j = 1$ a.s. Therefore,

$$\lim_{k_t\to\infty} \alpha_{k_t} = 0 \quad a.s. \tag{21}$$

□

S.4.2. Proof of Lemma 1

Lemma 1. *For any* $t \in \mathbf{N}$, $\mathrm{R}_t = (\rho_1, \rho_2, \ldots, \rho_{k_{t-1}}, \alpha_{k_{t-1}})$ *as defined in Eq.* (13) *is distributed as generalized Dirichlet distribution (Connor & Mosimann, 1969). Furthermore, if* $(1 - \mu_{j-1})\nu_{j-1} = \nu_j$ *for* j, $2 \leq j \leq k_{t-1}$, *then* $\mathrm{R}_t \sim Dirichlet(\mu_1\nu_1, \mu_2\nu_2, \ldots, \mu_{k_{t-1}}\nu_{k_{t-1}}, (1-\mu_{k_{t-1}})\nu_{k_{t-1}})$.

Proof. From Eq. (13), notice that $v_j \sim Beta\,(\mu_j\nu_j, (1-\mu_j)\nu_j)$ and $(\rho_1, \rho_2, \ldots, \rho_{k_{t-1}})$ is constructed by transforming (v_j). Hence, Jacobian is $\prod_{j=1}^{k_{t-1}} \left(\prod_{l=1}^{j-1}(1-v_l)\right)^{-1}$. Applying the transformation, we obtain the density function as

$$f_{\mathrm{R}_t} = \left(\prod_{j=1}^{k_{t-1}} B(\mu_j\nu_j, (1-\mu_j)\nu_j)\right)^{-1} \alpha_{k_t}^{(1-\mu_{k_{t-1}})\nu_{k_{t-1}}-1} \prod_{j=1}^{k_{t-1}} \left(\rho_j^{\mu_j\nu_j - 1}\left(\sum_{i=j}^{k_{t-1}} \rho_i + \alpha_{k_{t-1}}\right)^{\kappa_j}\right)$$

where $\alpha_{k_{t-1}} = 1 - \sum_{j=1}^{k_{t-1}} \rho_j$.

$$B(\mu_j\nu_j, (1-\mu_j)\nu_j) = \frac{\Gamma(\mu_j\nu_j)\Gamma((1-\mu_j)\nu_j)}{\Gamma(\nu_j)}$$

and

$$\kappa_j = (1-\mu_{j-1})\nu_{j-1} - \mu_j\nu_j - (1-\mu_j)\nu_j$$

Now let us write $a_j = \mu_j\nu_j$ and $b_j = (1-\mu_j)\nu_j$. Then Eq. (22) becomes equivalent to Eq. (10). Hence

$$\mathrm{R}_t \sim GDD(\mu_1\nu_1, (1-\mu_1)\nu_1, \ldots, \mu_{k_{t-1}}\nu_{k_{t-1}}, (1-\mu_{k_{t-1}})\nu_{k_{t-1}})$$

This proves the first part.

We prove the second part as follows. When

$$(1-\mu_{j-1})\nu_{j-1} = \mu_j\nu_j + (1-\mu_j)\nu_j = \nu_j$$

for $2 \le j \le k_{t-1}$ by Proposition 5 we get

$$\mathrm{R}_t \sim Dirichlet(\mu_1\nu_1, \dots, \mu_{k_{t-1}}\nu_{k_{t-1}}, (1-\mu_{k_{t-1}})\nu_{k_{t-1}})$$

□

S.4.3. Proof of Theorem 2

Theorem 2. *For α_{k_t} as defined in Eq. (13) with parameters $\boldsymbol{\mu}, \boldsymbol{\nu}$, and any $\epsilon \in (0,1)$, if $\mu_j > 1/2$ for all j, then $\alpha_k \le \epsilon$ whenever $k \ge \frac{2}{\log 2}\log\frac{1}{\epsilon}$ with probability more than $1-\epsilon$.*

Proof. From Eq. (13), using direct algebra one can rewrite $\alpha_r = \prod_{j=1}^{r}(1-v_j)$, and using the independence of v_j we find that

$$\mathbb{E}[\alpha_r] = \prod_{j=1}^{r}\mathbb{E}[1-v_j] \tag{22}$$

Using Markov inequality, we get

$$p(\alpha_{2r} > \frac{1}{2^r}) \le 2^r\mathbb{E}[\alpha_{2r}] \tag{23}$$

Now using the fact that $\mathbb{E}[1-v_j] = 1-\mu_j < 1/2$ and choosing a positive integer r such that $\epsilon > \frac{1}{2^r}$ one obtains

$$p(\alpha_{2r} \ge \epsilon) \le \epsilon \tag{24}$$

Then the proof follows by putting $k = 2r$. □

S.5. Proofs related to PPF of OSBP

S.5.1. Proof of Lemma 2

Lemma 2. *Let, (v_j) be defined as in Eq. (13), and $\mathrm{G}_{1:t-1}|\boldsymbol{\mu},\boldsymbol{\nu},\Gamma \sim OSBP(\boldsymbol{\mu},\boldsymbol{\nu},\Gamma)$. Then $\forall j$, $v_j|z_{1:t-1},\mu_j,\nu_j \sim Beta(\mu_j\nu_j + g_j - 1, (1-\mu_j)\nu_j + h_j)$.*

Proof. By definition of z, and OSBP, the following holds.

$$p(z_t = j|z_{1:t-1}, v_{1:k_{t-1}}) = v_j\prod_{l=1}^{j-1}(1-v_l),\ j \in [k_{t-1}]$$
$$p(z_t = k_{t-1}+1|z_{1:t-1}, v_{1:k_{t-1}}) = \prod_{l=1}^{k_{t-1}}(1-v_l)$$

Now, following (Pitman, 1995), it is straight forward to see that,

$$p(z_1,\dots,z_{t-1}|v_{1:k_{t-1}}) = \prod_{j=1}^{k_{t-1}}\left((1-v_j)^{h_j}v_j^{g_j-1}\right)$$

Now, we compute the posterior $p(v_1,\dots,v_{k_{t-1}}|z_{1:t-1})$ as follows.

$$\propto p(z_1,\dots,z_{t-1}|v_{1:k_{t-1}})p(v_1,\dots,v_{k_{t-1}})$$
$$\propto \prod_{j=1}^{k_{t-1}}\left((1-v_j)^{h_j}v_j^{g_j-1}v_j^{\mu_j\nu_j-1}(1-v_j)^{(1-\mu_j)\nu_j-1}\right)$$

After marginalizing over all other $v_l,\ l \in [k_{t-1}]\backslash j$, the lemma follows. □

S.5.2. Proof of Theorem 3

Theorem 3. *Let (π_j), $\sigma_{k_{t-1}}$ be defined in Eq. (14), and $\mathrm{G}_{1:t-1}|\boldsymbol{\mu},\boldsymbol{\nu},\Gamma \sim OSBP(\boldsymbol{\mu},\boldsymbol{\nu},\Gamma)$. Then, we have:*

$$\pi_j = \frac{\mu_j\nu_j+g_j-1}{\nu_j+g_j+h_j-1}\prod_{l=1}^{j-1}\frac{(1-\mu_l)\nu_l+h_l}{\nu_l+g_l+h_l-1},\ j \in [k_{t-1}],$$
$$\sigma_{k_{t-1}} = \prod_{l=1}^{k_{t-1}}\frac{(1-\mu_l)\nu_l+h_l}{\nu_l+g_l+h_l-1} \tag{25}$$

Proof. By definition of OSBP, for $1 \le j \le k_{t-1}$, $p(G_t = Q_j|G_{1:t-1},\{v_l\}) = p(z_t = j|z_{1:t-1},\{v_l\}) = \rho_j$.
Now by definition of PPFs in Eq. (14), one can write π_j as

$$\mathbb{E}\left[\rho_j|z_{1:t-1},\mu,\nu\right] = \mathbb{E}\left[v_j\prod_{l=1}^{j-1}(1-v_l)\,|z_{1:t-1},\mu,\nu\right]$$
$$= \mathbb{E}\left[v_j\,|z_{1:t-1},\mu_j,\nu_j\right]\prod_{l=1}^{j-1}\mathbb{E}\left[(1-v_l)\,|z_{1:t-1},\mu_l,\nu_l\right]$$

The second equation follows using the independence property of $\{v_j\}$. Following definition of OSBP, we similarly get $\beta_{k_{t-1}}$ defined in Eq. (14), $\beta_{k_{t-1}} = \prod_{l=1}^{k_{t-1}}\mathbb{E}[(1-v_l)|z_{1:t-1},\mu_l,\nu_l]$. Theorem follows using Lemma 2. □

S.6. Proofs related to SUMO

DPMM can be described as

$$\forall i x_i \sim f(\phi_i),\quad \phi_i|\mathrm{G} \sim \mathrm{G},\quad \mathrm{G} \sim DP(\gamma,\mathrm{H}) \tag{26}$$

Using OSBP we propose following for DPMM

$$\forall t,\quad \mathrm{G}_t|\mathrm{G}_{1:t-1},\mathrm{H} \sim \sum_{j=1}^{k_{t-1}}\rho_j\delta_{\mathrm{Q}_j} + \alpha_{k_{t-1}}\delta_{\mathrm{Q}_{k_{t-1}+1}}$$
$$\forall i,\ x_{ti}|\phi_{ti} \sim mult(\phi_{ti}),\quad \phi_{ti}|\mathrm{G}_t \sim \mathrm{G}_t \tag{27}$$

(ρ_j), (Q_j) and $\alpha_{k_{t-1}}$ are as defined in OSBP. Each G_t takes value from $(\mathrm{Q}_1,\dots,\mathrm{Q}_{k_{t-1}},\mathrm{Q}_{k_{t-1}+1})$, where $\mathrm{Q}_1 = \mathrm{G}_1$ and all other Q_j are sampled from $DP(\gamma_j,\mathrm{H})$. The second line in Eq. (27) models DPMM with G_t similar to Eq. (26).

Parameter settings. Regarding the parameters (μ_j,ν_j,γ_j), we set forall j $\mu_j = \mu$, for some $0.5 < \mu < 1$. $\nu_j = (1-\mu_)\nu_{j-1}$ and $\nu_1 = \gamma$ ($\gamma > 0$ as in Eq. (26)). Hence $\nu_j = (1-\mu)^{j-1}\gamma$.

We use, $\gamma_j = \mu\nu_j$ and hence $\gamma_j = \mu(1-\mu)^{j-1}\gamma$. Thus, there are two parameters μ and DPMM parameter γ.

Equivalence with DPMM. Before prove the Theorem 4, we need to revise one useful result as follows.

Theorem B. *Let, $\mathrm{Q}_j \sim DP(\gamma_j,\mathrm{H}_j)$ for $j = 1,\dots,k$ and $(c_1,\dots,c_k) \sim Dirichlet(\gamma_1,\gamma_2,\dots,\gamma_k)$ be independent of $\mathrm{Q}_1,\dots,\mathrm{Q}_k$, then*

$$\sum_{j=1}^{k}c_j\mathrm{Q}_j \sim DP(\sum_{j=1}^{k}\gamma_j, \frac{\sum_{j=1}^{k}\gamma_j\mathrm{H}_j}{\sum_{j=1}^{k}\gamma_j}) \tag{28}$$

Proof. $g_j \sim Gamma(\gamma_j, \beta)$ for $j = 1, \dots, k$ independently. Let, $\text{G}_j = g_j \text{Q}_j$, then $\text{G}_1, \dots, \text{G}_k$ are independent Gamma processes with $\text{G}_j \sim \Gamma P(\gamma_j \text{H}_j)$.

Let $\text{G}' = \sum_{j=1}^{k} \text{G}_j$, then by Proposition 2, $\text{G}' \sim \Gamma P(\sum_{j=1}^{k} \gamma_j \text{H}_j)$. Let $g' = \sum_{j=1}^{k} g_j$, then

$$\frac{\text{G}'}{\text{G}'(\Omega)} = \frac{\text{G}'}{\sum_{j=1}^{k} \text{G}_j(\Omega)} = \frac{\text{G}'}{\sum_{j=1}^{k} g_j \text{Q}_j(\Omega)} = \frac{\text{G}'}{g'}$$

Hence, $\frac{\text{G}'}{g'}$ is a normalized Gamma process, hence

$$\frac{\text{G}'}{g'} \sim DP(\sum_{j=1}^{k} \gamma_j, \frac{\sum_{j=1}^{k} \gamma_j \text{H}_j}{\sum_{j=1}^{k} \gamma_j)})$$

Again, if $c_j = \frac{g_j}{g'}$, then

$$(c_1, \dots, c_k) \sim Dirichlet(a_1, \dots, a_k)$$

and we can say

$$\frac{\text{G}'}{g'} = \sum_{j=1}^{k} \frac{g_j}{g'} \text{Q}_j = \sum_{j=1}^{k} c_j \text{Q}_j$$

Thus $\sum_{j=1}^{k} c_j \text{Q}_j \sim DP(\sum_{j=1}^{k} \gamma_j, \frac{\sum_{j=1}^{k} \gamma_j \text{H}_j}{\sum_{j=1}^{k} \gamma_j)})$. □

S.6.1. Proof of Theorem 4

Theorem 4. *For any $t \in \mathbf{N}$, each x_{ti} sampled using model Eq. (27) has marginal distribution same as x_i sampled with DPMM in Eq. (26) with $\text{G} \sim DP(c_t, \text{H})$, where $c_t = \sum_{j=1}^{k_{t-1}} \gamma_j + (1-\mu)^{k_{t-1}} \gamma$. Furthermore, for any $\epsilon > 0$ and $t > 0$, with probability greater than $1 - \epsilon$, each x_{ti} in Eq. (27) has marginal distribution same as x_i in Eq. (26) with $\text{G} \sim DP(\sum_{j=1}^{k} \gamma_j, \text{H})$, when $k_t \geq k \geq \frac{2}{\log 2} \log \frac{1}{\epsilon}$. Also, for $t \to \infty$, each x_{ti} in Eq. (27) has marginal distribution same as x_i in Eq. (26) with $\text{G} \sim DP(\gamma, \text{H})$.*

Proof. There are three parts of the Theorem, we prove them one by one.

Let us write,

$$\text{R}_t = (\rho_1, \rho_2, \dots, \rho_{k_{t-1}}, \alpha_{k_{t-1}})$$

then by Lemma 1, R_t is GDD distributed.

Now, we use

$$\gamma_j = \mu \nu_j = \mu (1-\mu)^{j-1} \nu$$

that satisfies

$$\nu_j = (1 - \mu_{j-1}) \nu_{j-1}$$

for $j > 1$. Hence, by Lemma 1,

$\text{R}_t \sim Dirichlet(\mu_1 \nu_1, \mu_2 \nu_2, \dots, \mu_{k_{t-1}} \nu_{k_{t-1}}, (1-\mu_{k_{t-1}}) \nu_{k_{t-1}})$

Now we can apply Theorem B, and by that

$$\text{G}_t \sim DP(c_t, \text{H}')$$

$$\text{H}' = \frac{(\sum_{j=1}^{k_{t-1}} \gamma_j + (1 - \mu_{k_{t-1}}) \nu_{k_{t-1}}) \text{H}}{\sum_{j=1}^{k_{t-1}} \gamma_j + (1 - \mu_{k_{t-1}}) \nu_{k_{t-1}}} = \text{H}$$

For any t, using the parameter setting we get

$$\text{G}_t \sim DP(c_t, \text{H})$$

$$c_t = \left(\sum_{j=1}^{k_{t-1}} \gamma_j + (1-\mu)^{k_{t-1}} \gamma \right) \tag{29}$$

Hence the marginal distribution of each x_{ti} sampled from model in Eq. (27) is always equivalent to that of DPMM. The difference will be in scale.

Then the first part of the Theorem follows immediately from Eq. (27) (second line) and Eq. (26).

For the second part, let us write

$$\begin{aligned}
&\textstyle\sum_{j=1}^{k} \rho_j \delta_{\text{Q}_j} + \alpha_k \delta_{\text{Q}_{k+1}} \\
&= (\textstyle\sum_{l=1}^{k} \rho_l) \Big(\sum_{j=1}^{k} \frac{\rho_j}{\sum_{l=1}^{k} \rho_l} \delta_{\text{Q}_j} + \frac{\alpha_k}{\sum_{l=1}^{k} \rho_l} \delta_{\text{Q}_{k+1}} \Big) \\
&= (1 - \alpha_k) \Big(\textstyle\sum_{j=1}^{k} \frac{\rho_j}{\sum_{l=1}^{k} \rho_l} \delta_{\text{Q}_j} + \frac{\alpha_k}{1-\alpha_k} \delta_{\text{Q}_{k+1}} \Big) \\
&= (1 - \alpha_k) \Big(\textstyle\sum_{j=1}^{k} \frac{\rho_j}{\sum_{l=1}^{k} \rho_l} \delta_{\text{Q}_j} \Big) + \alpha_k \delta_{\text{Q}_{k+1}}
\end{aligned} \tag{30}$$

Now we can say

$$\text{P}_t = \sum_{j=1}^{k_{t-1}} \rho_j \delta_{\text{Q}_j} + \alpha_{k_{t-1}} \delta_{\text{Q}_{k_{t-1}+1}}$$

is a mixture of two distributions $\text{J}_{k_{t-1}}$ and $\delta_{\text{Q}_{k_{t-1}+1}}$ for any t, where

$$\text{J}_{k_{t-1}} = \sum_{j=1}^{k_{t-1}} \frac{\rho_j}{\sum_{l=1}^{k_{t-1}} \rho_l} \delta_{\text{Q}_j}$$

. Probability of sampling G_t from $\text{J}_{k_{t-1}}$ is $1 - \alpha_{k_{t-1}}$.

Again as,

$(\rho_1, \dots, \rho_{k_{t-1}}, \alpha_{k_{t-1}}) \sim Dir(\gamma_1, \dots, \gamma_{k_{t-1}}, (1-\mu)^{k_{t-1}} \gamma)$

by Proposition 4,

$$(\frac{\rho_1}{1 - \alpha_{k_{t-1}}}, \dots, \frac{\rho_{k_{t-1}}}{1 - \alpha_{k_{t-1}}}) \sim Dirichlet(\gamma_1, \dots, \gamma_{k_{t-1}})$$

Hence $\text{J}_{k_{t-1}}$ is equivalent to

$$DP(\sum_{j=1}^{k_{t-1}} \gamma_j, \text{H})$$

Now, from Theorem 2, for $k_{t-1} \geq \frac{2}{\log 2} \log \frac{1}{\epsilon}$, $\alpha_{k_{t-1}} < \epsilon$ for any $\epsilon > 0$ with probability at least $1 - \epsilon$.

Hence, with probability at least $1 - \epsilon$ we sample G_t from $DP(\sum_{j=1}^{k} \gamma_j, \text{H})$ for $k \geq \frac{2}{\log 2} \log \frac{1}{\epsilon}$. Now, for t such that

$k_t \geq k$, x_{ti} sampled from model in Eq. (27), is marginally equivalent to x_i sampled from model DPMM in Eq. (26) with $\text{G} \sim DP(\sum_{j=1}^{k} \gamma_j, \text{H})$.

For the third part, from Theorem 1, as $t \to \infty$, $\text{P}_t \to \text{P}^* = \sum_{j=1}^{\infty} \rho_j \delta_{\text{Q}_j}$ such that $\sum_{j=1}^{\infty} \rho_j = 1$ a.s. By the first part, P^* becomes equivalent to $DP(\sum_{j=1}^{\infty} \gamma_j, \text{H})$.

Now we can write

$$\sum_{j=1}^{\infty} \gamma_j = \mu\gamma \sum_{j=1}^{\infty} (1-\mu)^{j-1} = \frac{\mu\gamma}{1-(1-\mu)} = \gamma$$

Hence, P^* becomes equivalent to $DP(\gamma, \text{H})$.

Thus, for t, P_t going to P^*, x_{ti} sampled from model in Eq. (27), is marginally equivalent to x_i sampled from model DPMM in Eq. (26) with $\text{G} \sim DP(\gamma, \text{H})$. This proves the Theorem. □

S.7. Application of OSBP on other BNP models

Recall that, in the mini-batch setup, we consider a streaming dataset as $(\text{X}_t) = (\text{X}_1, \text{X}_2, \ldots, \text{X}_{\bar{d}})$, where $\text{X}_t = \{x_i\}_{i=\bar{n}(t-1)+1}^{\bar{n}t}$ [1]. For clarity in notation, we represent X_t as $\{x_{ti}\}_{i=1}^{\bar{n}}$. We will describe application of OSBP on Pitman-Yor process, stick-breaking process and higherarchical Dirichlet process to share information across mini-batches. However, unlike DPMM the equivalence relationship is not straight forward to prove in these cases, and are left for future work.

We will consider H as a probability measure over a measurable space $(\Omega, \mathcal{B})$, and $f(.)$ is the distribution for the data model. (ρ_j) and $\alpha_{k_{t-1}}$ are as defined in OSBP.

The inference mechanism follows from Theorem 3 and the inference techniques for the corresponding models. It is straight forward to derive them as outlined in Section 3.2 for DPMM. We do not describe them here.

S.7.1. OSBP on Pitman-Yor process

Pitman-Yor process (PYP) (Pitman & Yor, 1997) can be described following the stick-breaking representation as follows. For $0 \leq a < 1$, and $b > -a$, any random probability measure $\text{G} \sim PYP(a, b, \text{H})$ if

$$\begin{aligned} \text{G} = \textstyle\sum_{j=1}^{\infty} \theta_j \delta_{\beta_j},\ \ \theta_1 = v_1,\ \theta_j = v_j \prod_{l=1}^{j-1}(1-v_l) \\ \forall j,\ \ v_j \sim Beta(1-a, b+ja);\ \ \beta_j \sim \text{H} \end{aligned} \tag{31}$$

PYP mixture model can be described as follows.

$$\forall i,\ x_i \sim f(\phi_i);\ \ \forall i,\ \phi_i|\text{G} \sim \text{G} \tag{32}$$

[1]for simplicity we have assumed $n = \bar{n}\bar{d}$.

Now we apply OSBP on PYP mixture model to get

$$\begin{aligned} \forall t,\ \text{G}_t|\text{Q}_{1:k_{t-1}}, \text{H} \sim \textstyle\sum_{j=1}^{k_{t-1}} \rho_j \delta_{\text{Q}_j} + \alpha_{k_{t-1}} PYP(a, b, \text{H}) \\ \forall i,\ x_{ti}|\phi_{ti} \sim f(\phi_{ti}),\ \ \forall i,\ \phi_{ti}|\text{G}_t \sim \text{G}_t \end{aligned} \tag{33}$$

Application of OSBP on PYP is similar to that of DP. The inference follows from the Theorem 3, and PPFs of PYP.

S.7.2. OSBP on stick-breaking process

Recall that, stick-breaking process (SBP) (Ishwaran & James, 2001) can be described following the stick-breaking representation as follows. Let, $a_j, b_j > 0$, and $a = (a_1, a_2, \ldots), b = (b_1, b_2, \ldots)$. Any random probability measure $\text{G} \sim SBP(a, b, \text{H})$ if following holds.

$$\begin{aligned} \text{G} = \textstyle\sum_{j=1}^{\infty} \theta_j \delta_{\beta_j},\ \ \theta_1 = v_1,\ \theta_j = v_j \prod_{l=1}^{j-1}(1-v_l) \\ \forall j,\ \ v_j \sim Beta(a_j, b_j);\ \ \beta_j \sim \text{H} \end{aligned} \tag{34}$$

SBP mixture model can be described as follows.

$$\forall i,\ x_i \sim f(\phi_i);\ \ \forall i,\ \phi_i|\text{G} \sim \text{G} \tag{35}$$

Imposing OSBP on SBP mixture model yields the following model.

$$\begin{aligned} \forall t,\ \text{G}_t|\text{Q}_{1:k_{t-1}}, \text{H} \sim \textstyle\sum_{j=1}^{k_{t-1}} \rho_j \delta_{\text{Q}_j} + \alpha_{k_{t-1}} SBP(a, b, \text{H}) \\ \forall i,\ x_{ti}|\phi_{ti} \sim f(\phi_{ti}),\ \ \forall i,\ \phi_{ti}|\text{G}_t \sim \text{G}_t \end{aligned} \tag{36}$$

SBP being generalized version subsumes many BNP priors including DP and PYP. The construction of OSBP based sequential model for DP, PYP and SBP are similar. Essentially, following this structure it is easy to build such sequential models for a wide range of BNP models. The inference will follow from Theorem 3 and inference procedure of SBP. Unfortunately, SBP does not have PPFs and truncated methods are applied (Ishwaran & James, 2001).

S.7.3. OSBP on hierarchical Dirichlet process

Hierarchical Dirichlet process (HDP) (Teh et al., 2006) is defined as below for $\gamma, \lambda > 0$

$$\begin{aligned} \text{G}_0 \sim DP(\gamma, \text{H}) \\ \forall i,\ \text{G}_i \sim DP(\lambda, \text{G}_0) \end{aligned} \tag{37}$$

HDP assumes grouped data, that x_i represent a group which consists of data points $\{x_{il}\}$. HDP mixture model can be described as

$$\begin{aligned} \text{G}_0 \sim DP(\gamma, \text{H}) \\ \forall i,\ \text{G}_i \sim DP(\lambda, \text{G}_0) \\ \forall l,\ x_{il} \sim f(\phi_{il});\ \phi_{il}|\text{G}_i \sim \text{G}_i \end{aligned} \tag{38}$$

Imposing OSBP on HDP mixture model by using the base measure Γ of OSBP as $DP(\gamma, \text{H})$, we get

$$\forall t,\ \mathrm{G}_t|\mathrm{Q}_{1:k_{t-1}},\mathrm{H} \sim \sum_{j=1}^{k_{t-1}} \rho_j \delta_{\mathrm{Q}_j} + \alpha_{k_{t-1}} DP(\gamma, \mathrm{H})$$
$$\forall i,\ \mathrm{G}_{ti} \sim DP(\lambda, \mathrm{G}_t)$$
$$\forall l,\ x_{til}|\phi_{til} \sim f(\phi_{til}), \quad \phi_{til}|\mathrm{G}_{ti} \sim \mathrm{G}_{ti} \tag{39}$$

The inference will follow from Theorem 3 and the prediction rule for HDP, Chinese restaurant franchise (CRF) (Teh et al., 2006).

S.8. SUMO for DPMM on document clustering

We describe SUMO here for text datasets. Each data point x_i is a document which has multiple examples (words). A words in document i is denoted by x_{il}. The data model is $x_{il} \sim mult(\phi_i)$. In order to maintain conjugacy, ϕ_i has Dirichlet prior.

(ϕ_i) are sampled from (Q_j) where $\mathrm{Q}_j \sim DP(\gamma_j, \mathrm{H})$ in Eq. (27). We can say $\mathrm{Q}_j = \sum_{r=1}^{\infty} \zeta_{jr}\delta_{\psi_{jr}}$ following Eq. (12), where (ζ_{jr}) form the stick-breaking weights and atoms are ψ_{jr}. Let, $\{\beta_s\}$ is set of global components. Then each $\psi_{jr} \in \{\beta_s\}$ ensures same components across t. We can create global components by ad-hoc merging of components across t. But we prefer a more technical approach of using a.s. discrete H by $\mathrm{H} \sim DP(\lambda, Dirichlet(\eta))$. We can write $\mathrm{H} = \sum_{s=1}^{\infty} \theta_s \delta_{\beta_s}$, where $\beta_s \sim Dirichlet(\eta)$ and (θ_s) form the stick-breaking weights.

Given this setup, we introduce alternative variables to speed up the mixing of the Markov chain following standard approach. Recall that, $z_t = j$ if $\mathrm{G}_t = \mathrm{Q}_j$. Let, $a_{ti} = r$ if $\phi_{ti} = \psi_{jr}$ and $z_t = j$. So r is the index of the mixture component in prior G_t assigned to document i of mini-batch t. If s is the index of global mixture component represented by ψ_{jr} in Q_j, then we define $b_{jr} = s$ if $\psi_{jr} = \beta_s$. Furthermore, let $y_{ti} = s$ if $z_t = j$ and $b_{jr} = s$. y_{ti} is the index of the global mixture component assigned to document i in mini-batch t. ϕ and ψ can be retrieved from z, a, b and β.

Due to this representation, the equivalent random quantities are $\mathrm{A}_{1:t} = \{\{a_{li}\}_{i=1}^{\bar{n}}\}_{l=1}^{t}$, $\mathrm{B}_{1:k_t} = \{b_{jr}\}_{j=1}^{k_t}$, and $\mathrm{Y}_{1:t} = \{\{y_{li}\}_{i=1}^{\bar{n}}\}_{l=1}^{t}$. We integrate out (Q_j) and H following Chinese restaurant process (CRP), (ρ_j) following Theorem 3, and $\{\beta_s\}$ following Dirichlet multinomial conjugacy. So, we need to infer A_t, B, and z_t at time t after observing X_t. The posterior of $\{\beta_s\}$ and other variables can be retrieved after the inference through a, b, z and (X_t).

Notation. Superscript with hyphen denotes set minus, e.g. $\mathrm{X}_t^{-i} = \mathrm{X}_t \backslash x_{ti}$, and $\mathrm{X}_t^{-r} = \mathrm{X}_t \backslash \mathrm{X}_{tr}$, where $\mathrm{X}_{tr} = \{x_{ti}|a_{ti} = r\}$. $\mathrm{X}_{1:t}^{-tr} = \mathrm{X}_{1:t} \backslash \mathrm{X}_{tr}$, and $\mathrm{X}_{1:t}^{-ti} = \mathrm{X}_{1:t} \backslash x_{ti}$. $\mathrm{A}_{1:t}^{-ti} = \mathrm{A}_{1:t} \backslash a_{ti}$. $\mathrm{B}_{z_t}^{-r} = \mathrm{B}_{z_t} \backslash b_{z_t r}$. $\mathrm{L}_s(x_{ti})$ and $\mathrm{L}_s(\mathrm{X}_{tr})$ are the likelihood of x_{ti} and X_{tr} respectively for mixture component s.

Recursive computation of likelihood. $\mathrm{L}_s(x_{ti})$ is the likelihood of x_{ti} under mixture component s, that is $\mathrm{L}_s(x_{ti}) = p(x_{ti}|\mathrm{Y}_{1:t}, \mathrm{X}_{1:t-1}, \mathrm{X}_t^{-i})$. After observing $\mathrm{X}_{1:t-1}$ and X_t^{-i}, $\mathrm{L}_s(x_{ti})$ can be computed by recursively applying Bayes theorem using Dirichlet multinomial conjugacy as follows.

$$p(x_{ti}|\mathrm{Y}_{1:t}, \mathrm{X}_{1:t-1}, \mathrm{X}_t^{-i}) =$$
$$\int \prod_f p(x_{tif}|y_{ti} = s, \beta_s) p(\beta_s|\mathrm{X}_t^{-i}, \mathrm{X}_{1:t-1}, \mathrm{Y}_{1:t}) d\beta_s$$
$$= \int \prod_f \beta_{s x_{tif}} \frac{\Gamma(\sum_v(\eta_v + C_{sv} + c_{sv}^{-i}))}{\prod_v \Gamma(\eta_v + C_{sv} + c_{sv}^{-i})} \prod_v \beta_{sv}^{\eta_v + C_{sv} + c_{sv}^{-i} - 1} d\beta_s$$
$$= \frac{\Gamma(\sum_v(\eta_v + C_{sv} + c_{sv}^{-i}))}{\prod_v \Gamma(\eta_v + C_{sv} + c_{sv}^{-i})} \frac{\prod_v \Gamma(\eta_v + C_{sv} + c_{sv}^{-i} + c_{sv}^{i})}{\Gamma(\sum_v(\eta_v + C_{sv} + c_{sv}^{-i}) + c_{sv}^{i})}$$
$$\int \frac{\Gamma(\sum_v(\eta_v + C_{sv} + c_{sv}^{-i}) + c_{sv}^{i})}{\prod_v \Gamma(\eta_v + C_{sv} + c_{sv}^{-i} + c_{sv}^{i})} \prod_v \beta_{sv}^{\eta_v + C_{sv} + c_{sv}^{-i} + c_{sv}^{i} - 1} d\beta_s$$
$$= \frac{\Gamma(\sum_v(\eta_v + C_{sv} + c_{sv}^{-i}))}{\prod_v \Gamma(\eta_v + C_{sv} + c_{sv}^{-i})} \frac{\prod_v \Gamma(\eta_v + C_{sv} + c_{sv}^{-i} + c_{sv}^{i})}{\Gamma(\sum_v(\eta_v + C_{sv} + c_{sv}^{-i}) + c_{sv}^{i})} \tag{40}$$

Integration happens following the property that $\beta_s \sim Dirichlet(\eta)$ and using Dirichlet multinomial conjugacy. Please refer to the Appendix for detailed steps. We define the sufficient statistics as below.

$$\mathrm{C}_{sv} = \sum_{l=1}^{t-1} \sum_{i=1}^{\bar{n}} \sum_f \mathbb{I}[y_{li} = s, x_{lif} = v]$$
$$c_{sv} = \sum_{i=1}^{\bar{n}} \sum_f \mathbb{I}[y_{ti} = s, x_{lif} = v]$$
$$c_{sv}^{-i} = \sum_{q=1, q\neq i}^{\bar{n}} \sum_f \mathbb{I}[y_{tq} = s, x_{lqf} = v]$$
$$c_{sv}^{i} = \sum_f \mathbb{I}[y_{ti} = s, x_{lif} = v] \tag{41}$$

Similarly, we compute $\mathrm{L}_s(\mathrm{X}_{tr})$ the likelihood of X_{tr} for mixture component s, $p(\mathrm{X}_{tr}|\mathrm{Y}_{1:t}, \mathrm{X}_{1:t-1}, \mathrm{X}_t^{-r})$ as follows.

$$\int \prod_{i=1:a_{ti}=r}^{\bar{n}} \prod_f p(x_{tif}|y_{ti} = s, \beta_s) p(\beta_s|\mathrm{X}_{1:t-1}, \mathrm{Y}_{1:t}) d\beta_s$$
$$= \int \prod_{i=1:a_{ti}=r}^{\bar{n}} \prod_f \beta_{s x_{tif}} \frac{\Gamma(\sum_v(\eta_v + \mathrm{C}_{sv} + c_{sv}^{-r}))}{\prod_v \Gamma(\eta_v + \mathrm{C}_{sv} + c_{sv}^{-r})}$$
$$\prod_v \beta_{sv}^{\eta_v + \mathrm{C}_{sv} + c_{sv}^{-r} - 1} d\beta_s$$
$$= \frac{\Gamma(\sum_v(\eta_v + \mathrm{C}_{sv} + c_{sv}^{-r}))}{\prod_v \Gamma(\eta_v + \mathrm{C}_{sv} + c_{sv}^{-r})} \frac{\prod_v \Gamma(\eta_v + \mathrm{C}_{sv} + c_{sv}^{-r} + c_{sv}^{r})}{\Gamma(\sum_v(\eta_v + \mathrm{C}_{sv} + c_{sv}^{-r}) + c_{sv}^{r})}$$

We define the required sufficient statistics as below.

$$c_{sv}^{r} = \sum_{i=1}^{\bar{n}} \sum_f \mathbb{I}[a_{ti} = r, z_t = j, b_{jr} = s, x_{tif} = v]$$
$$c_{sv}^{-r} = \sum_{i=1}^{\bar{n}} \sum_f \mathbb{I}[a_{ti} = q, q \neq r, b_{z_t q} = s, x_{tif} = v] \tag{42}$$

Inference of a. We infer a as below.

$$p(a_{ti} = r|\mathrm{A}_{1:t}^{-ti}, \mathrm{B}_{1:k_t}, z_{1:t}, \mathrm{X}_{1:t}) \propto \tag{43}$$
$$p(x_{ti}|a_{ti} = r, z_{1:t}, \mathrm{A}_{1:t}^{-i}, \mathrm{B}_{1:k_t}, \mathrm{X}_{1:t}^{-i}) p(a_{ti} = r|\mathrm{A}_{1:t}^{-ti}, z_t)$$

where $p(x_{ti}|a_{ti} = r, z_{1:t}, \mathrm{A}_{1:t}^{-i}, \mathrm{B}_{1:k_t}, \mathrm{X}_{1:t}^{-i})$ is $\mathrm{L}_{b_{z_t r}}(x_{ti})$. $p(a_{ti} = r|\mathrm{A}_{1:t}^{-ti}, z_t)$ comes from CRP as

$$\propto \mathrm{L}_{b_{z_t r}}(x_{ti})(m_{z_t r}^{-i} + \mathrm{M}_{z_t r})(1 - \iota_r) + \gamma_{b_{z_t}} \mathrm{L}_{b_{z_t r_{new}}}(x_{ti})\iota_r \tag{44}$$

$$\iota_r = \mathbb{I}[r = r_{new}], \quad m_{z_t r} = \sum_{i=1}^{\bar{n}} \mathbb{I}[a_{ti} = r],$$
$$\mathrm{M}_{jr} = \sum_{l=1}^{t-1} \sum_{i=1}^{\bar{n}} \mathbb{I}[z_l = j, a_{li} = r] \tag{45}$$

$m_{z_t r}$ denotes the number of time component ψ_{jr} is assigned in the current mini-batch, whereas M_{jr} dontes how many times ψ_{jr} is assigned across all the mini-batches seen so far excluding the current mini-batch. When a new r_{new} is sampled we obtain $b_{z_t r_{new}}$ from $p(b_{z_t r} = s_{new}|z_{1:t}, \mathrm{A}_{1:t}, \mathrm{B}_{1:k_t}, \mathrm{X}_{1:t})$ which is shown later.

Inference of z. Following the dependence structure in Eq. (27), z_t is independent of X_t given Y_t. So, we can infer z from $p(z_t = j|z_{1:t-1}, \mathrm{Y}_t, \mathrm{B}_{1:k_t})$ as

$$\propto \left[\prod_{i=1}^{\bar{n}} p(y_{ti} = s|z_{1:t}, \mathrm{B}_{1:k_t})\right] \ p(z_t = j|z_{1:t-1}) \quad (46)$$

$p(z_t = j|z_{1:t-1})$ comes from Theorem 3. Recall that $y_{ti} = b_{z_t a_{ti}}$. So $p(y_{ti} = s|z_t = j, \mathrm{B}_{1:t}, z_{1:t-1})$ comes from CRP by integrating out G_t and H.

Let, $\iota_j = \mathbb{I}[z_t = j_{new}]$, $\iota_s^j = \mathbb{I}[z_t = j, s{=}s_{new}]$, $\iota_s^0 = \prod_{l=1}^{k_t} \iota_s^l$, $\mathrm{J}_{js} = \sum_r \mathbb{I}[b_{jr} = s, z_t = j]$ and $\mathrm{J}_{.s} = \sum_{j=1}^{k_{t-1}} \mathrm{J}_{js}$. ι_s^j, ι_s^0 denote if β_s is present in Q_j, H respectively. J_{js} counts number of times β_s is present among $\{\psi_{jr}\}$.

Notice that, when $\iota_s^j = 0$, ι_s^0 must be 0 and that implies the situation that the global component β_s is present in Q_j. When, $\iota_s^j = 1$, $\iota_s^0 = 0$ signifies that β_s is not present in Q_j, but is present in H. Whereas $\iota_s^0 = 1$ implies $\iota_s^j = 1$ and β_s is not present in any prior. When, $\iota_j = 1$, ι_s^j must be 1, but ι_s^0 may be 1 or 0. Hence there are following scenarios.

i. $\iota_j = 0$, $\iota_s^j = 0$: then we can say $p(y_{ti} = s|z_t = j, \mathrm{B}_{1:t}, z_{1:t-1}) \propto \mathrm{J}_{js}$.

ii. $\iota_j = 0$, $\iota_s^j = 1$, $\iota_s^0 = 0$: then we need to sample a global component from H which is proportional to $\gamma_j \mathrm{J}_{.s}$. $\mathrm{J}_{.s} = \sum_{j=1}^{k_{t-1}} \sum_r \mathbb{I}[b_{jr} = s, z_t = j]$, sum over all existing priors. So $p(y_{ti} = s|z_t = j, \mathrm{B}_{1:t}, z_{1:t-1}) \propto \lambda \mathrm{J}_{.s}$.

iii. $\iota_j = 0$, $\iota_s^j = 1$, $\iota_s^0 = 1$: then we need to sample a new global component from $Dirichlet(\eta)$ which is proportional to λ. So $p(y_{ti} = s|z_t = j, \mathrm{B}_{1:t}, z_{1:t-1}) \propto \lambda\gamma_j$.

iv. $\iota_j = 1$, $\iota_s^j = 1$, $\iota_s^0 = 0$: then we need to sample a new global component from $Dirichlet(\eta)$ which is proportional to λ. So $p(y_{ti} = s|z_t = j, \mathrm{B}_{1:t}, z_{1:t-1}) \propto \mathrm{J}_{.s}$. γ_j does not appear here as there is not Q_j and no corresponding CRP.

v. $\iota_j = 1$, $\iota_s^j = 1$, $\iota_s^0 = 1$: then we need to sample a new global component from $Dirichlet(\eta)$ which is proportional to λ. So $p(y_{ti} = s|z_t = j, \mathrm{B}_{1:t}, z_{1:t-1}) \propto \lambda$.

Combining them together we get $p(z_t = j|z_{1:t-1}, \mathrm{Y}_t, \mathrm{B}_{1:k_t})$

$$\propto \left[\prod_{i=1}^{\bar{n}} \mathrm{J}_{js}(1-\iota_s^j) + \gamma_j \iota_s^j \big(\mathrm{J}_{.s}(1-\iota_s^0) + \lambda\iota_s^0\big)\right]$$
$$\pi_j(1-\iota_j) + \left[\mathrm{J}_{.s}(1-\iota_s^0) + \lambda\iota_s^0\right]\sigma_{k_{t-1}}\iota_j \quad (47)$$

Algorithm 2 SUMO for DPMM on text datasets.

Require: (X_t), μ, λ, γ and η
for $t = 1, 2, \ldots$ **do**
  Initialize global component assignments Y_t
  for $iter = 1$ to I **do**
    Sample z_t from $p(z_t|z_{1:t-1}, \mathrm{Y}_t, \mathrm{J})$
    for $i = 1$ to $\bar{n}$ **do**
      Sample a_{ti} from $p(a_{ti} = r|\mathrm{A}_t^{-i}, z_t, \mathrm{X}_t, \mathrm{M}, \mathrm{C})$
    end for
    Sample B_{z_t} from $p(b_{z_t r} = s|\mathrm{B}_{z_t}^{-r}, z_t, \mathrm{X}_t, \mathrm{M}, \mathrm{N}, \mathrm{C})$
  end for
  Compute c, m, n and update C, M, N, and J
  Discard local variables X_t, A_t, and Y_t
end for
Ensure: z, A, B, C, M, N

π_j and $\sigma_{k_{t-1}}$ are as defined in Eq. (25).

Inference of b. We infer b as below.

$$p(b_{z_t r} = s|z_{1:t}, \mathrm{A}_{1:t}, \mathrm{B}_{1:k_t}, \mathrm{X}_{1:t}) \propto \quad (48)$$
$$p(\mathrm{X}_{tr}|z_{1:t}, \mathrm{A}_t, \mathrm{B}_{1:k_t}, \mathrm{X}_{1:t}^{-tr}) p(b_{z_t r} = s|\mathrm{B}_{z_t}^{-r}, z_{1:t}, \mathrm{A}_{1:t}, \mathrm{B}_{1:k_t})$$

where $p(\mathrm{X}_{tr}|z_{1:t}, \mathrm{A}_t, \mathrm{B}_{1:k_t}, \mathrm{X}_{1:t}^{-tr})$ is $\mathrm{L}_s(\mathrm{X}_{tr})$ and $p(b_{z_t r} = s|\mathrm{B}_{z_t}^{-r}, z_{1:t}, \mathrm{A}_{1:t}, \mathrm{B}_{1:k_t})$ comes from CRP as

$$\propto \mathrm{L}_s(\mathrm{X}_{tr})(n_{z_t s}^{-r} + \mathrm{N}_s^{-z_t})(1-\iota_s) + \lambda \mathrm{L}_{s_{new}}(\mathrm{X}_{tr})\iota_s \quad (49)$$

we define the variables as

$$\iota_s = \mathbb{I}[s{=}s_{new}], \quad n_{z_t s}^{-r} = \sum_{q \neq r} \mathbb{I}[b_{z_t q} = s],$$
$$\mathrm{N}_s^{-z_t} = \sum_{l=1}^{k_{t-1}} \sum_q \mathbb{I}[b_{lq} = s, l \neq z_t] \quad (50)$$

$n_{z_t s}^{-r}$ denotes the number of times component β_s has been used in the mixing distribution Q_{z_t} excluding ψ_{jr}. Whereas $\mathrm{N}_s^{-z_t}$ denotes how many times component β_s is used in unique mixing distributions (Q_j) except Q_{z_t}.

SUMO for DPMM on text datasets. Using Eq. (43) in step 5, and Eq. (46), Eq. (48) in step 7 of SUMO (Algorithm 1), we obtain SUMO for text datasets presented in Algorithm 2.

Notice that from Eq. (44), Eq. (47) and Eq. (49) that by maintaining sufficient statistics M, J, N and L, we need not store the local variables $\mathrm{A}_{1:t-1}$, $\mathrm{Y}_{1:t-1}$, $\mathrm{X}_{1:t-1}$.

References

Connor, R. and Mosimann, J. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969.

Ishwaran, H. and James, L. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

Lee, J., Quintana, F., Muller, P., and Trippa, L. Defining predictive probability functions for species sampling models. *Statistical Science*, 28(2):209–222, 2013.

Pitman, J. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158, 1995.

Pitman, J. Some developments of the Blackwell-MacQueen urn scheme. *Lecture Notes-Monograph Series*, pp. 245–267, 1996.

Pitman, J. and Yor, M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997.

Sethuraman, J. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

Teh, Y., Jordan, M., Beal, M., and Blei, D. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.