Weight-Agnostic Hierarchical Stick-Breaking Process

1st Mrinal Das
Indian Institute of Technology Palakkad, India
mrinal@iitpkd.ac.in
2nd Chiranjib Bhattacharyya
Indian Institute of Science, Bangalore, India
chiru@csa.iisc.ernet.in

***Abstract*—**
Learning from multiple groups of observations are often useful due to the advantage of sharing of statistical information. Hierarchical Bayesian models provide a natural mechanism to achieve this, and hierarchical Dirichlet processes (HDPs) have shown significant impact in this field. HDPs define a collection of probability measures one for each group. All the measures provide support on a *common* countably infinite set of *atoms* to share information. The fundamental mechanism in all the variants of HDP make the weights on these atoms positively correlated across groups. This structural limitation is impossible to resolve without changing the sharing principle. But this property hinders the applicability of HDP priors to many problems, when an atom may be highly probable in some groups despite being rare in all other groups. This becomes evident in clustering through association of atoms and observations. Some clusters may be weakly present in most of the groups in spite of being prominent in some groups and vice-versa.

In this paper, we pose the problem of *weight agnosticism*, that of constructing a collection of probability measures on a common countably infinite set of atoms with mutually independent weights across groups. This implies that, a cluster can contain observations from all groups, but popularities of a cluster across groups are mutually independent. So the size of a cluster in a group does not interfere in the participation of observations in other groups to that cluster. Our contribution is also to construct a novel hierarchical Bayesian nonparametric prior, *Weight-Agnostic hierarchical Stick-breaking process* (WAS), which models weight agnosticism. WAS extends the framework of stick-breaking process (SBP) in a novel direction. However, WAS becomes non-exchangeable and that makes inference process non-standard. But, We derive tractable predictive probability functions for WAS, which is useful in deriving efficient *truncation-free* MCMC inference competitive with those in HDP settings. We discuss few real life applications of WAS in topic moeling and information retrieval. Furthermore, experimenting with five real life datasets we show that, WAS significantly outperforms HDP in various settings.

***Index Terms*—Bayesian nonparametrics, hierarchical models, topic models, rare clusters**

I. Introduction

Learning from multiple groups of observations are often useful; specially, when we try to learn together using the advanatge of sharing of information. Consider the task of clustering words into topics as in topic modeling. Here words are grouped into documents, and increasing the number of documents in general increase the quality of the topics. Hierarchical Bayesian nonparametrics (HBN) provides an elegant mechanism to share statistical information across groups of observations. Hierarchical extensions of Dirichlet Process (HDP)[1, 2] constitute important priors in such cases, and have been used in many applications inclduing clustering groups of observations.

However, the inherent mathematical structure of HDP assumes that the popularity of clusters are "similar" across groups. This correlation of clusters across groups is unrealitic and undesired in many cases. For example, consider news articles on three different stories: (i) culinary routine of professional athletes, (ii) athletic program at a culinary college, (iii) culinary business of a retired athlete. In all the three articles, two topics: (i) *culinary*, and (ii) *athletics* are present. These topics are shared across the articles. But the weight over these topics are very different in these three articles. In the first article *culinary* will have higher weight, and *athletics* will have lower weight. Whereas, in the second article *culinary* will have lower weight but *athletics* will have higher weight. On the other hand, for the third article, both the topics can have similar weights. Thus popularities of these clusters do not correlate across groups. This is surprisingly a more common phenomenon than we expect, and can be observed in any form of groups of observations. This phenomenon also happens inside a document when words are grouped into sentences. For example, in a review article on *iPhone*, the *display quality* may be covered only in few sentences. So weight over this topic will be low in most of the sentences, but very high in the sentences which have mentioned the topic *display quality*. In the collection of news articles of one year, there will be very few articles on *artificial intelligence*. So the weight over this topic will be negligibly small in most of the documents, but will be high in the articles covering *artificial intelligence*.

Of late, there has been a significant body of work on the problem of discovering topics rare in the entire document or corpus, but could be highly probable in some sentences[3]. Modeling such rare topics has found applications in wide variety of domains including software engineering[4], user generated content[5], surveillance videos[6]. Naturally, HDP fails to apply in such cases. All the mentioned works are scattered initiatives to approach the limitation of HDP but they lack the mathematical approach and generality in their study. The main objective of this paper is to study the fundamental

grounds on this aspect of HBN, and propose principled solution.

If we look at the existing HBNs from a holistic point of view, there is a collection of probability measures, one for each group. Each probability measure corresponding to a group define mass over a global set of atoms shared across groups. Association of observations with these atoms yield clustering across groups. HBNs are used to define a probability structure over these atoms and the probability measures that govern the clustering. Separating atoms and the probability structure over them have not been studied before, and we emphasize in this paper that this line of thought can be very useful to achieve more flexibility in designing HBNs for variety of problems.

We focus in this paper on the problem of reducing the coupling of weights across the groups of observations in HBN. We refer to this problem as *weight-agnosticism* in HBN. We further propose a weight-agnostic Bayesian nonparametric prior based on stick-breaking process (SBP) [7]. One key advantage of SBP is that it allows to define discrete probability measures over countably infinite number of atoms, where atoms and the weights over the atoms are independent. However, the disadvantage is that, there is no existing mechanism to share atoms across multiple SBP priors. Moreover, inference is hard with SBP.

Contributions.

- We propose a novel stick-breaking construction that satisfies weight-agnosticism in HBN. We call this prior as *weight-agnostic hierarchical stick-breaking process* or WAS. Theorem 2 shows that WAS is indeed weight-agnostic.
- SBP lacks tractable *predictive probability functions* (PPFs)[8], where PPFs are backbone of efficient truncation-free inference of HBNs. Because of that inference procedures for SBP based models require finite truncation or overly complex methods. Furthermore, the proposed model turn out to be non-exchangeable which further makes derivation of PPFs non-standard. However, we could derive PPFs for WAS; leading to efficient *truncation-free* Markov chain Monte Carlo (MCMC) inference procedure.

II. Preliminaries and problem description

Notation. We will use following notations throughout the paper. H is a *diffuse* probability measure over a suitable measurable space $(\Omega, \mathcal{B})$, more precisely for any $y \in \Omega$, $\mathrm{H}(y) = 0$. δ_y will denote an atomic probability measure, the entire probability mass being concentrated at y. A set of variables $\{x_1, x_2, \ldots, x_n\}$ will be denoted by $x_{1:n}$. $\{x_j\}$ will denote an infinite set, and (x_j) will denote an infinite sequence, j specifying the order. The set of integers $\{1, \ldots, k\}$ will be denoted by $[k]$. $\mathbb{I}[.]$ denotes the indicator function, and $|.|$ means cardinality.

A. Preliminaries: SBP and HDP

Stick-Breaking Process (SBP). Stick-breaking process (SBP)[7] provides a framework to define atomic measures which can be defined as follows.

Definition 1. *(Stick-breaking process) Let* H *is a* diffuse measure *over a measurable space* $(\Omega, \mathcal{B})$ *and* $\boldsymbol{a} = (a_j), \boldsymbol{b} = (b_j)$ *are set of positive parameters. Any almost sure (a.s.) discrete probability measure* P *is a stick-breaking process (SBP) if it can be represented as*

$$\mathrm{P} = \textstyle\sum_{j=1}^{\infty} \theta_j \delta_{\beta_j}, \quad \beta_j \sim \mathrm{H}$$
$$\theta_1 = v_1, \ \theta_j = v_j \textstyle\prod_{l=1}^{j-1}(1 - v_l), \quad v_j \sim Beta(a_j, b_j) \quad (1)$$

$\boldsymbol{\beta} = (\beta_j)$ are the *atoms* and $\boldsymbol{\theta} = (\theta_j)$ are the corresponding *weights* such that, $\sum_{j=1}^{\infty} \theta_j = 1$. We will denote $\boldsymbol{\theta} \sim stick(\boldsymbol{a}, \boldsymbol{b})$. SBP specifies to Dirichlet process (DP) if $\forall j \ a_j = 1$ and $b_j = \alpha$, and denoted as $\mathrm{P} \sim DP(\alpha, \mathrm{H})$[9].

Hierarchical Dirichlet Process (HDP). The HDP formulations by [10] and by [2] are the two most explored forms which will be referred as MQR-HDP and TJB-HDP respectively. For $\{\{x_{si}\}_{i=1}^{n_s}\}_{s=1}^{S}$ observations HDP can be expressed as:

$$\forall s, \ \forall i \ \ x_{si} \sim f(\phi_{si}), \quad \phi_{si} \sim \mathrm{P}_s,$$
$$\mathrm{P}_s = \textstyle\sum_{j=1}^{\infty} \theta_{sj} \delta_{\beta_{0j}}, \quad \forall j \ \ \beta_{0j} \sim \mathrm{H} \quad (2)$$

$\{\beta_{0j}\}$ are the set of atoms shared across all groups, and they are sampled from H. Note that the selection of right H will depend on the observations only. The weight for atom β_{0j} in the sth group is θ_{sj}. Notice that, the form of P_s is same as in Eq. (1) but slight differ in the construction based on the model as follows:

$$\text{MQR-HDP} \quad \forall s, \ \mathrm{P}_s = \epsilon \mathrm{G}_0 + (1 - \epsilon)\mathrm{G}_s; \quad (3)$$
$$\text{TJB-HDP} \quad \forall s, \ \mathrm{P}_s \sim DP(\gamma, \mathrm{G}_0); \quad (4)$$

$\epsilon \in (0, 1)$ provides a trade-off between global and local measures in Eq. (3) which is subtle in Eq. (4). $\mathrm{G}_0 \sim DP(\alpha, \mathrm{H})$ and H is same as in Eq. (2). Similarly, $\{\mathrm{G}_s\}$ are also distributed as $DP(\alpha, \mathrm{H})$, but they are specific to the groups unlike G_0. So following Eq. (1), we can write

$$\mathrm{G}_0 = \sum_{j=1}^{\infty} \lambda_{0j} \delta_{\xi_{0j}}; \ \ \forall s, \mathrm{G}_s = \sum_{j=1}^{\infty} \lambda_{sj} \delta_{\xi_{sj}}. \quad (5)$$

ξ_{0j}s, ξ_{sj}s are distributed as H, and λ_{0j}s, λ_{sj}s are constructed from $stick(1, \alpha)$. Comparing with Eq. (2), in case of MQR-HDP, $\{\beta_{0j}\} = \{\xi_{0j}\} \cup \{\xi_{sj}\}$. The weights are related as follows:

$$\theta_{sj} = \epsilon \lambda_{0j} \, \mathbb{I}[\beta_{0j} \in \{\xi_{0j}\}] \ + \ (1 - \epsilon)\lambda_{sj} \ \ \mathbb{I}[\beta_{0j} \in \{\xi_{sj}\}] \quad (6)$$

So for MQR-HDP sharing of atoms is partial but for the shared atoms the weights are same. Whereas for TJB-HDP, G_s in Eq. (5) is same as P_s in Eq. (3), and $\{\beta_{0j}\} = \{\xi_{sj}\} = \{\xi_{0j}\}$. Thus for TJB-HDP, all the atoms are shared across all the groups. The weights are constructed as follows:

$$\theta_{sj} = v_{sj} \textstyle\prod_{l=1}^{j-1}(1 - v_{sl}),$$
$$v_{sj} \sim Beta(\gamma v_{0j} \textstyle\prod_{l=1}^{j-1}(1 - v_{0l}), \gamma \textstyle\prod_{l=1}^{j}(1 - v_{0l})), \quad (7)$$

where v_{0j}s are sampled from $Beta(1, \alpha)$. Refer to [2] for the derivation.

B. Weights over shared atoms in HDPs are dependent

From the above discussion, we notice that, the weights over a shared atom β_{0j}, $(\theta_{1j}, \theta_{2j}, \ldots, \theta_{Sj})$ are either exactly same

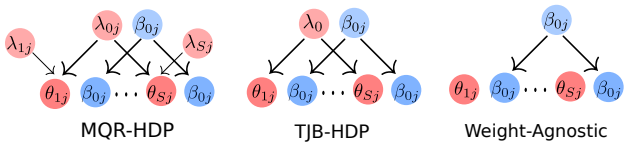


Fig. 1. Comparison of weight-agnostic HBN with HDPs using a shared atom β_{0j}. Circles denote random variables and arrows denote dependency. HDPs make weights $(\theta_{1j}, \ldots, \theta_{Sj})$ across groups dependent whereas they are mutually independent for weight-agnostic HBN.

(for MQR-HDP) or marginally dependent (for TJB-HDP). The implication is that, the weights over β_{0j} tend to be correlated across groups. It is hard to eliminate this phenomenon of correlation of weights across the groups merely by tuning the hyper-parameters. Notice that, the Beta distribution corresponding to v_{sj} is in the form Beta$(\nu\mu, \nu(1-\mu))$, where $\nu = \gamma\prod_{l=1}^{j}(1-v_{0l})$, and $\mu = v_{0j}$. Now, $Var(v_{sj}) = \frac{\mu(1-\mu)}{\nu+1}$; and $\lim_{\nu\to\infty} Var(v_{sj}) = 0$ whereas $\lim_{\nu\to 0} Var(v_{sj}) = \mu(1-\mu)$. μ being always a fraction we can not increase variance of v_{sj} or θ_{sj} arbitrarily. Moreover, μ is independent of γ. So variance and expectation of θ_{sj} can not be controlled by γ; and v_{sj} being iid, they can not be mutually independent for any choice of γ or α.

Therefore it is not straightforward to reduce dependency across groups through hyper-parameters in TJB-HDP. Similarly, for MQR-HDP reducing ϵ will reduce sharing of atoms instead of reducing dependency over shared atoms.

C. Weight-agnostic HBN

It is not natural that, an atom popular in most of the groups should be highly probable in every group or vice versa. We attempt to address this limitation of HBN in this paper. Our aim is to make the weights mutually independent across groups; that will allow to share statistical information across groups without interference on how to use the information. In that regards, we introduce the concept of *weight-agnosticism* in HBN.

Definition 2. *(Weight-agnosticism in HBN) Any HBN sharing atoms across groups in such a way that θ_{sj} is the weight over atom β_{0j} in the sth group, satisfies weight-agnosticism if for any atom β_{0j} shared among groups denoted by $\mathcal{S} \subseteq [S]$ following holds:*

$$p\Big(\prod_{s\in\mathcal{S}}\theta_{sj}|\Theta\Big) = \prod_{s\in\mathcal{S}} p\Big(\theta_{sj}|\Theta\Big) \tag{8}$$

where Θ denotes all non-random hyper-parameters.

Weight-agnosticism implies that weights across groups over a shared atom are mutually independent. This allows θ_{sj} to be high even if θ_{tj} is very low for all $t \neq s$. Figure 1 illustrates the concept comparing with HDPs.

D. SBP based approach for weight-agnostic HBN

The main challenge with existing mechanism is that, to provide support over same discrete set of countably infinite number of atoms, the measures across groups become dependent. It is hard to decouple the measures keeping the support same. SBP is the most viable approach towards weight-agnostic HBN due to its ability to construct a.s. discrete probability measures over countably infinite number of atoms in a way that provides explicit control over the weights. So we propose to build a weight-agnostic HBN as defined in Definition 2 where the weights (θ_{sj}) are constructed through stick-breaking process.

Challenges. There are two key challenges. (i) The base measure in SBP is diffuse in nature which puts zero mass on any atom and hence will not allow any sharing of atoms across groups. However, a discrete measure H will allow sharing but is too restrictive for a general model family. Moreover, a discrete measure will make atoms to repeat even within a group. That will fail to characterize the posterior as SBP further complicating the inference mechanism[11]. For example, Dirichlet distribution commonly used in topic models, is diffuse in nature, making SBP inapplicbale in topic models without ad-hoc truncation. (ii) Due to lack of tractable predictive probability functions, SBP needs to adopt either truncation or overly-complex mechanisms for inference, neither of them allow SBP to be HBN. We address these challenges in our approach to be described next.

III. Weight-agnostic hierarchical stick-breaking process

A. Construction of WAS

WAS is developed through two important constructions: (i) exploit-explore stick-breaking, and (ii) relatively-diffuse measure.

Exploit-explore stick-breaking. The construction is based on a simple idea. Decoupling the weights is easy for finite models and due to finite number of data points, the number of atoms assigned to a data point is always finite. Hence, we do not need to define mass over infinite number of atoms at the beginning of the process. So, the prior can work with finite number of atoms (exploit) and it can increase the size on the need basis (explore).

Let $f(.)$ is the distribution corresponding to the data-model and $x_{sn} \sim f(\phi_{sn})$, and $(\beta_{s1}, \beta_{s2}, \ldots, \beta_{sk_{sn}})$ are the set of unique values among $(\phi_{s1}, \phi_{s2}, \ldots, \phi_{s,n-1})$. We propose a construction as follows.

$$\begin{aligned} \phi_{sn} \sim \mathrm{P}_{sn},\ \mathrm{P}_{sn} = \textstyle\sum_{j=1}^{k_{sn}} \theta_{sj}\delta_{\beta_{sj}} + \alpha_{sn}\Gamma \\ \theta_{s1} = v_{s1},\ \forall j > 1,\ \theta_{sj} = v_{sj}\textstyle\prod_{l=1}^{j-1}(1-v_{sl}) \\ \alpha_{sn} = 1 - \textstyle\sum_{j=1}^{k_{sn}}\theta_{sj}, \quad v_{sj} \sim Beta(a_j, b_j) \end{aligned} \tag{9}$$

(a_j, b_j) denote the set of scalar hyper-parameters similar to SBP. Γ is an appropriate probability measure considering the observations. We will show later that, P_{sn} asymptotically converges to SBP. The implication of above model is that, ϕ_{sn} can re-use existing β_{sj} with probability θ_{sj} when $k_{s,n+1} = k_{sn}$ (exploit phase). On the other hand, ϕ_{sn} can use a new sample from Γ with probability α_{sn}, and in that case, $k_{s,n+1} = k_{sn}+1$ and $\beta_{k_{s,n+1}} = \phi_{sn}$ (explore phase).

Relatively-diffuse measure. Γ in Eq. (9) should have two properties: (i) it should allow sharing of atoms across groups,

that is same atom can be sampled in multiple groups from Γ, and (ii) it should be diffuse with respect to a single group, i.e. same atoms will never be sampled twice in any group. Although the first property is intutive given the definition of HBN, but the second property is related to deriving efficient inference algorithm for WAS. In order to construct Γ, we introduce relatively-diffuse measure.

Definition 3. *(Relatively-diffuse probability measure) Let Γ is a probability measure over a suitable measurable space $(\Omega, \mathcal{B})$. We define Γ to be a relatively-diffuse probability measure with respect to a pair of disjoint sets (A, G) of discrete elements, where $A, G \subset \mathcal{B}$ if $\Gamma(\beta) = 0$ for any $\beta \in A$, $\Gamma(\beta) > 0$ for $\beta \in G$ and for $\beta \subset (A \cup G)^c$, $\Gamma(\beta) \geq 0$.*

Recall that, Γ is a *diffuse* probability measure when for any $y \in \Omega$, $\Gamma(y) = 0$. Here, Γ is diffuse with respect to only set A. and atomic with respect to set G. So, we call Γ as relatively-diffuse probability measure; where specification of sets (A, G) is important. Note that, relatively-diffuse probability measures are not singular measures due to positive mass on elements in set G; they are rather a special case of *mixed measures* (with both discrete and continuous part) specified with respect to two pre-defined sets. See the appendix for more discussion on relatively-diffuse probability measures. It is important to understand, how we can construct relatively-diffuse probability measures; we state below an useful result regarding that.

Theorem 1. *Let H is a diffuse probability measure on $(\Omega, \mathcal{B})$, and there are two sets A, G such that $A, G \subset \mathcal{B}$ and $A \cap G = \emptyset$. If $\mathrm{J} = \sum_{j:x_j \in G} \eta_j \delta_{x_j}$, such that $\eta_j > 0$ for all j and $\sum_{j:x_j \in G} \eta_j = 1$; then $\Gamma = \epsilon \mathrm{J} + (1-\epsilon)\mathrm{H}$, is a relatively-diffuse probability measure with respect to the pair of sets (A, G) for any $\epsilon \in (0, 1]$.*

Proof. See the appendix. □

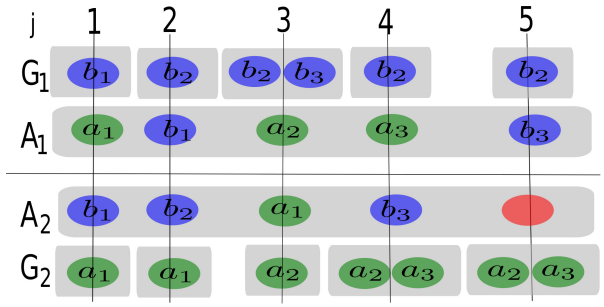
Relatively-diffuse measure with SBP. We will apply Theorem 1 to extend the notion to relatively-diffuse probability measure for stick-breaking priors using exploit-explore mechanism. First, we define following important sets.

$$\forall s, A_{sn} = \{\beta_{s1}, \ldots, \beta_{sk_{sn}}\}, G_{sn} = (\cup_{l \neq s} A_{ln},\ l \in [S]) \quad (10)$$

A_{sn} is the set of atoms present in stick-breaking prior corresponding to group s, whereas G_s denotes the set of atoms present in all the groups except group s. Figure 2 shows an example of A_{sn}, G_{sn}. We will construct J of Theorem 1 as

$$\Gamma_{sn}(\beta) \propto \begin{cases} 1 & \beta \in G_{sn} \\ 0 & \beta \in A_{sn} \\ \zeta & \beta \sim \mathrm{H} \end{cases} \quad (11)$$

H is the base measure as in Eq. (2) and Eq. (9). From Eq. (9) and Eq. (10), A_s and G_s are finite sets for any s. That makes Eq. (11) well defined. By Theorem 1, Γ_{sn} is a relatively-diffuse probability measure through a suitable mixture of diffuse H and a discrete probability measure depending on (A_{sn}, G_{sn}). Figure 2 shows an example of Γ_{sn} using two groups, where for some n_1, n_2 after $j = 5$, $\Gamma_{1n_1} \propto (\delta_{b_2} + \zeta \mathrm{H})$ and $\Gamma_{2n_2} \propto (\delta_{a_2} + \delta_{a_3} + \zeta \mathrm{H})$.



$\mathrm{P}_{1n_1} = \theta_{11}\delta_{a_1} + \theta_{12}\delta_{b_1} + \theta_{13}\delta_{a_2} + \theta_{14}\delta_{a_3} + \theta_{15}\delta_{b_3} + \alpha_{15}\Gamma_{1n_1}$
$\mathrm{P}_{2n_2} = \theta_{21}\delta_{b_1} + \theta_{22}\delta_{b_2} + \theta_{23}\delta_{a_1} + \theta_{24}\delta_{b_3} + \alpha_{24}\Gamma_{2n_2}$
$p(\theta_{11}, \theta_{23}|\Theta) = p(\theta_{11}|\Theta)p(\theta_{23}|\Theta)$

Fig. 2. Example of WAS with two groups. Green circles denote atoms first appearing in group 1, and blue circles denote atoms first appearing in group 2. For group 1, b_1, b_3 comes from group 2, and for group 2, a_1 comes from group 1. Red circle denotes the 5th atom for group 2 which will come from $G_{2n_2} = \{a_2, a_3\}$ or newly from H. For any group s, size of A_s strictly monotonically increases, but size of G_s can reduce. Bottom part shows state of P_{sn} for the two groups for some $n1, n2$. Weights over shared atom a_1 in two groups θ_{11}, θ_{23} are independent (Θ is hyper-parameter).

Definition of WAS. Using the two constructions described above: exploit-explore mechanism and relatively-diffuse measure, we define WAS below.

Definition 4. *(WAS) Let H is a* diffuse *probability measure over a suitable measurable space $(\Omega, \mathcal{B})$ and $\boldsymbol{a} = (a_j), \boldsymbol{b} = (b_j)$ denote the set of positive scalar hyper-parameters. We denote $\{\phi_{sn}\} \sim WAS(\boldsymbol{a}, \boldsymbol{b}, \mathrm{H})$ for all s, n if $\{\phi_{sn}\}$ are modeled using Eq. (9), where Γ is defined as in Eq. (11).*

By P_s and β_s we denote the sampling distribution and the atoms for the sth group respectively. Figure 2 illustrates WAS using two groups. One important question here is that, how do the mechanism in Eq. (9) happen for multiple groups. Without loss of generality, let $\beta_{11} = \phi_{11}$ and $\beta_{11} \sim \mathrm{H}$, then the rest follows with exploit-explore mechanism in a round-robin fashion; that is each group is processed sequentially and starting again with the first group after visiting all the groups and in each iteration one observation is sampled. $\{G_{sn}\}$ are updated after visiting each group.

B. Properties of WAS

Theorem 2. *(Weight-agnosticism) Let (θ_{sj}) is as defined in WAS (Eq. (9)), and $(\beta_{0j},\ \beta_{0j} \sim \mathrm{H})$ denotes the set of shared atoms. If σ_s is the permutation function for s-th group such that weight of β_{0j} in that group is $\theta_{s\sigma_s(j)}$, then for any $\beta_{0j} \in (\cap_{s \in \mathcal{S}} A_{sn})$, where $\mathcal{S} \subseteq [S]$ following holds.*

$$p\Big(\prod_{s \in \mathcal{S}} \theta_{s\sigma_s(j)} | \boldsymbol{a}, \boldsymbol{b}\Big) = \prod_{s \in \mathcal{S}} p\Big(\theta_{s\sigma_s(j)} | \boldsymbol{a}, \boldsymbol{b}\Big) \quad (12)$$

Proof. See the appendix. □

Theorem 2 shows that, for WAS the weights over each shared atom are mutually independent across groups. Another important property of WAS is as follows.

Theorem 3. *(Asymptotic convergence) Let P_{sn}, (θ_{sj}) are as defined in WAS(Eq. (9)), and $(\beta_{0j},\ \beta_{0j} \sim \mathrm{H})$ denotes the set of shared atoms with σ_s permutation function for s-th group such that weight of β_{0j} in that group is $\theta_{s\sigma_s(j)}$. If $\mathrm{P}_s^* =$*

$\sum_{j=1}^{\infty} \theta_{s\sigma_s(j)}\delta_{\beta_{0j}}$, *such that and* $\theta_{s\sigma_s(j)} \geq 0$, $\sum_{j=1}^{\infty} \theta_{s\sigma_s(j)} = 1$*; then* $\lim_{n\to\infty} \mathrm{P}_{sn} = \mathrm{P}_s^*$ *a.s.* $\forall s$.

Proof. See the appendix. □

Theorem 3 states that, WAS will solve the problem of modeling (P_s) as in Eq. (2) asymptotically. In other words, by defining support over countably infinite set of atoms across groups WAS is a proper HBN. Additionally, Theorem 2 establishes that WAS is weight-agnostic HBN.

C. Predictive probability functions and inference mechanism

Predictive Probability Functions (PPFs)[8] are useful tools to design efficient truncation-free MCMC inference algorithms for Bayesian nonparametric models. The reason is that PPFs allow to compute inference steps using only some count variables and simple predictive rules. Derivation of PPFs are based on two properties of the model: (i) exchangeability of the atoms, (ii) exchangeability of the data points. SBP in general lack the first property except few special cases such as DP and Pitman-Yor process (PYP) [12]. In addition to that, WAS also lacks the second property. That makes derivation of PPFs for WAS hard and non-standard given the state of the art understanding.

We will show here that WAS allows tractable PPFs, and we will also describe a restaurant analogy for WAS in the same spirit of DP and PYP which are popularly known as Chinese restaurant process (CRP) and Chinese restaurant franchise (CRF) respectively. What makes it possible for WAS is the exploit-explore structure of WAS coupled with relatively-diffuse sampling distribution for every groups. Thus WAS is a special case of SBP, and although derivation of PPFs are still possible for WAS like priors but it is not possible SBP in general.

Restaurant analogy. Let us consider that there are S hypothetical restaurants in some strange world. Each restaurant has tables as many as required. Each customer coming to a restaurant occupies a table which may be occupied by some customers previously arrived or it can be empty. No customer leaves the restaurant or changes table. Customers are arriving one by one and choosing a restaurant. Each table contains a dish. The same dish can appear in other restaurants but not in multiple tables of any single restaurant.

Now let us draw the connections with our model. Restaurants correspond to sampling distributions of the groups (P_s), dishes correspond to the atoms (β_s), and the customers correspond to (ϕ_{sn}). So A_{sn} is the set of dishes in the sth restaurant, and G_{sn} is the set of dishes in all restaurants except the sth restaurant at the time of arrival of the nth customer at the sth restaurant.

Auxiliary variables to compute PPFs. Let $|A_{sn}| = k_{sn}$ and $|G_{sn}| = K_{sn}$. The dish served at the jth table in the sth restaurant is denoted by d_{sj}, and $\sigma_s(j)$ denotes the index of that dish in the global menu common across all the restaurants. The nth customer in the sth restaurant sits at the jth table with probability π_{sj} if $j \in [k_{sn}]$ (already occupied), otherwise the same customer sits at an unoccupied table with probability ξ_{sn}, and in that case $j = k_{sn} + 1$ (index of the new table). So when a new table is occupied, we get a new dish d_{sj}, where $j = k_{sn} + 1$; and we set $k_{s,n+1} = k_{sn} + 1$. The lth dish from menu G_{sn} is brought into the sth restaurant with probability η_{sl}, and d_{sj} is set to l. Recall that, G_{sn} does not contain any dish already served in the sth restaurant. It is also possible that a new dish is cooked with probability ζ_{sn}. Cooking a new dish is analogous of sampling an atom from a diffuse probability measure (H in Eq. (11)). When a new dish is cooked, set $K_{tn} = K_{tn} + 1$ for all $t \neq s$, and $\sigma_t(K_{tn} + 1) = \sigma_s(k_{sn} + 1)$ which is $\max(\sigma_s(K_{sn}), \sigma_s(k_{sn})) + 1$. Now we are ready to derive PPFs.

Computation of PPFs. Let $z_{sn} = j$ iff $\phi_{sn} = \beta_{sj}$ and $d_{sj} = l$ iff $\beta_{sj} = \beta_{0l}$. Working with z, d and β_{0j} will simplify the notation as they are sufficient to retrieve other variables. The probability functions of customers occupying tables and dishes served at tables are described by the PPFs below.

$$\begin{aligned} \pi_{sj} &= p(z_{sn} = j | z_{s,1:n-1}, \Theta),\ j \in [k_{sn}] \\ \xi_{sn} &= p(z_{sn} = k_{sn} + 1 | z_{s,1:n-1}, \Theta) \\ \eta_{sl} &= p(d_{sj} = l | A_{1:S}, \Theta),\ l \in [K_{sn}] \\ \omega_{sn} &= p(d_{sj} = K_{sn} + 1 | A_{1:S}, \Theta) \end{aligned} \tag{13}$$

where Θ denotes the set of hyper-parameters. Let, $g_{sj} = \sum_{l=1}^{n-1} \mathbb{I}[z_{sl} = j]$, $h_{sj} = \sum_{l>j} g_{sl}$. We are now ready to state the main result of this section.

Theorem 4. *Let* (π_{sj}), $\xi_{sn}, (\eta_{sl}), \omega_{sn}$ *be defined as in Eq.* (13)*, then, we have:*

$$\begin{aligned} \pi_{sj} &= \tfrac{a_j + g_{sj} - 1}{a_j + b_j + g_{sj} + h_{sj} - 1} \prod_{l=1}^{j-1} \tfrac{b_l + h_{sl}}{a_l + b_l + g_{sl} + h_{sl} - 1},\ j \in [k_{sn}] \\ \xi_{sn} &= \prod_{l=1}^{k_{sn}} \tfrac{b_l + h_{sl}}{a_l + b_l + g_{sl} + h_{sl} - 1} \\ \eta_{sl} &= \tfrac{1}{K_{sn} + \zeta},\ \ l \in [K_{sn}],\ \ \omega_{sn} = \tfrac{\zeta}{K_{sn} + \zeta} \end{aligned} \tag{14}$$

Proof. See the appendix. □

Note that, (π_{sj}) and ξ_{sn} can be interpreted as posteriors of (θ_{sj}) and α_{sn} in Eq. (9). Despite challenges due to non-exchangeability and lack of tractable PPFs in general SBP framework, WAS admits tractable PPFs which lead to desired truncation-free MCMC inference. Moreover, we need to maintain only some count statistics to compute the PPFs; that makes the inference very efficient. Tractability in computing PPFs becomes possible due to exploit-explore construction coupled with relatively-diffuse measure. However, due to non-exchangeability sampling during the nth observation depends only on past $n - 1$ samples.

Comparison with PPFs of TJB-HDP. There are two key differences. (i) Notice that, π_{sj} can be written as $\mu_{sj} \prod_{l=1}^{j-1}(1 - \mu_{sl})$, where $\mu_{sj} = \frac{a_j + g_{sj} - 1}{a_j + b_j + g_{sj} + h_{sj} - 1}$. μ_{sj} is the posterior expectation of v_{sj} conditioned on (ϕ_{si}) (see Lemma 4 in the appendix). Thus in WAS, the probability of jth atom β_{sj} in directly depends on probability of *not* assigning first $j - 1$ atoms $\{\beta_{s1}, \beta_{s2}, \ldots, \beta_{s,j-1}\}$. Whereas, in case of HDP, the probability of jth atom depends only on g_{sj}. (ii) The probability of sampling a dish from the existing global dishes in case of HDPs is proportional to the popularity of that dish across all the groups[1, 2]; whereas they are equal for WAS. That makes a popular dish also highly probable in all restaurants for HDP, but not for WAS.

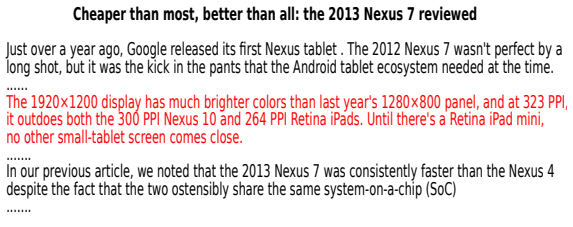


Fig. 3. Excerpt from Google Nexus 7 review at ArsTechnica. Among more than 60 sentences only 4 sentences cover the topic on *retina display* (shown in red). Probability of a topic (retina display here) may not be same in all parts of a document. Weights over a topic vary across sentences. often such topics are very informative. HDP fails to model them. The proposed model WAS can detect them.

Algorithm 1 : Application of WAS in topic modeling

- For documents $d = 1$ **to** D
  - For sentences $s = 1, \ldots, S_d$
    - For words $n = 1, \ldots, N_{ds}$
      - Sample topic vector $\phi_{dsn} \sim WAS(\boldsymbol{a}, \boldsymbol{b}, \mathrm{H})$
      - Sample topic $\psi_{dsn} \sim \phi_{dsn}$
      - Sample word $w_{dsn} \sim mult(\psi_{dsn})$

D. Related existing priors

Theorem 3 states that, each group specific prior relates to SBP asymptotically. In such hierarchical models, often it is interesting to understand the marginal behavior of the model given a single group of variables. When $S = 1$, then $G_{1n} = \emptyset$ and a.s. $\Gamma_{1n} = \mathrm{H}$. Thus WAS reduces to the prior as in [13]; however the model by [13] does not apply for multiple groups.

Strong correlation among weights across groups in HDP has also been noted by [3]. They used Indian buffet process (IBP) [14] compound DP to induce sparsity over atoms in each group. This in effect reduces the amount of sharing across groups rather than reducing correlation among weights over shared atoms. We will refer to this prior as ICD and although ICD does not address weight-agnosticism, we will compare with ICD in our experiments.

IV. Use cases of WAS

In order to demonstrate practical use cases of *weight-agnosticism*, we have picked 2 problems: 1) to detect *subtle topics*[4] and, 2) detection of *specific comments*[5]. Both the problems have been explored recently and is in general hard to solve them. The methods to solve them are quite ad-hoc. However, both the problems have shown significant relevance to modern text analytics and retrieval.

Subtle topics are hard to detect because they rarely present in the entire document. But subtle topics can be prominent in a few sentences. However, subtle topics can carry significant information. For example if we consider *stick-breaking process* as a topic, it is rarely present in this paper although that is an important component of this paper. One more example is shown in Figure 3. The example is taken from a review article on *Google Nexus 7*[15]. The topic *retina display* is present only in 4 sentences out of more than 60 sentences. Retina display was an important feature for the users at that time. [4] shows that detection of subtle topics is hard for HDP as well as ICD. They apply an SBP based ad-hoc solution. The model is given Algorithm 1, where in place of WAS, [4] applied SBP. We will use this as a baseline in the next section.

Specific comments are another example. Often, comments made on some news or some article can be relevant only to some specific sentences of the article. For example a reviewer commenting on the benefit of using *stick-breaking process* in this paper will be a specific comment as the topic of *stick-breaking process* itslef is rarely present. Considering the example in Figure 3, there were more than 22% comments on *retina display*, that is these comments were made on about 15% of the entire article. However, such comments were shown to be more informative than general and vague comments. [5] shows that detection of such comments are extremely hard, and some ad-hoc SBP based method were applied to solve it.

There is a unified structure in both the problems and the central problem is same. If we consider sentences as groups and clustering of words as topics, then both the problems signify a phenomenon that, some clusters may be highly popular in some groups, but they can be negligible in many groups. Thus the weight over the clusters vary across the groups. We apply WAS to solve the central problem. In Algorithm 1, we show the generative process. Note that, [4, 5] applied SBP in place of WAS in Algorithm 1. We can also apply HDP, ICD in place of WAS to obtain various baselines, that we will use in our experimental study in the next section.

WAS based topic model. Let us elaborate the generative process. For the nth word in the sth sentence of the dth document we do following; sample a distribution over topics ϕ_{dsn} from WAS. We have used the same notation as in Section III to make it easy to connect. After sampling ϕ_{dsn} from WAS, sample topic ψ_{dsn} from ϕ_{dsn}, and then sample word w_{dsn} from ψ_{dsn}.

Recall that, H in Eq. (9) is a suitable base distribution depending on the observations. Here H is $Dirichlet(\kappa 1_K)$ from which the proportion over topics are sampled. Topics are global across the documents which are denoted by (λ_k). Topics are sampled from $Dirichlet(\eta 1_V)$.

The proportion over topics ϕ_{dsn} is sampled from WAS at the time of processing the nth word in the sth sentence of the dth document. If P_{dsn} is the sampling distribution corresponding to WAS, then $\phi_{dsn} \sim \mathrm{P}_{dsn}$ and

$$\mathrm{P}_{dsn} = \sum_{j=1}^{k_{dsn}} \theta_{dsj} \delta_{\beta_{dsj}} + \alpha_{dsn} \Gamma_{dsn}.$$

(β_{dsj}) is the set of distributions over topics corresponding to the dth document, and ϕ_{dsn} takes value from (β_{dsj}). Whenever a new value is sampled for ϕ_{dsn}, that will be added to the set (β_{dsj}). Notice that, weights over topics are mutually independent across the sentences due to WAS. Thus the weights over a topic can be high in few sentences whereas it is low in most of the sentences.

MCMC inference. Efficient inference algorithms are important parts of any generative model. The PPFs of WAS allow truncation-free collapsed MCMC inference for the model in Algorithm 1. Derivation of MCMC inference steps is straightforward following Theorem 4. Detailed derivation is covered in the appendix.

V. Experiments

A. Experimental setting

Evaluation strategy. We will evaluate WAS on three aspects. First, we will evaluate on the ability to predict unseen data measured with perplexity. Second, we will verify the quality of the topics detected using topic-coherence [16]. Finally, we will evaluate WAS on a more precise task of information retrieval measured with precision, recall and F1.

Baselines. We use SBP, ICD[3] and TJB-HDP (will be referred as HDP hence forth) as baselines to compare with WAS. We get the baselines by replacing WAS by SBP and HDP in Algorithm 1. For the parameters of WAS, SBP we use $\forall j\ a_j, b_j = 1$. For HDP, we use $\alpha = 1, \gamma = 0.5$. For The topic parameter $\eta = 1$ and base distribution for topic vectors uses $\kappa = 1$.

Datasets. We have used datasets released by [4, 5]. As discussed by [4, 5], a significant number of subtle topics are present in these corpora, making them suitable to evaluate WAS.

1) NIPS-05. This is a corpus of 207 papers from the proceedings of Neural Information Processing Systems, 2005(nips.cc/Conferences/2005) released by [17].

2) Obama-speech. Collection of public speeches of Barack Obama from July 27, 2004 to October 30, 2012(www.americanrhetoric.com/barackobamaspeeches.htm) containing 142 articles transcribed directly from audio.

3) BerkeleyDB. This dataset contains Berkeley DB Java Edition released by [18]. Software codes in each file as document contain different statements treated as sentences thus they possess more variation than natural texts.

4) JHotDraw. JHotDraw is a well known open source GUI framework for drawing technical and structured graphics(www.jhotdraw.org/). JHotDraw is framework based source code, where subtle topics are more frequently found.

5) ArsTechnica Science (AT-Science). This dataset contains articles and comments crawled from Science section of the site ArsTechnica(arstechnica.com/science). ArsTechnica is a science and technology blog whose writers consist mostly of academicians. There are 1500 articles and their comments over approximately a two year timeline (June 2011 to March 2013), after removing articles with less than 5 comments. In this dataset, there are many comments which are related only to few sentences in the main article, which are called *specific comments* quite different from the entire article, posing a major challenge to model.

6) ArsTechnica Science Labelled (AT-Science-L). The dataset contains articles from ArsTechnica over one year timeline (March 2012 to March 2013). Each of these articles have many comments associated with them. For each article, comments are manually labeled as either specific or non-specific. There are 501 articles with a total of 3176 comments. This dataset is used by [5] for the retrieval task we consider here; where given an article the task is to retrieve the specific comments.

Pre-processing. For software datasets, only the textual content (without programming syntax) of the '.java' files(no documentation etc) used as input. Each statement has been treated as a sentence and Java key-words(en.wikipedia.org/wiki/List_of_Java_keywords) are removed but common java library names are retained. Tokens like StringCopy have been split into two words String and Copy. For all datasets, we have removed standard English stop words, digits, sentences smaller than 20 characters and words smaller than 3 characters. We converted capital faces to small faces. We have used most frequent 5000 words for NIPS, and 20000 words for AT-Science, and in other cases we have used the full vocabulary.

TABLE I
Comparison of SBP, HDP, ICD and WAS using *perplexity* (less is better).

Dataset	SBP	HDP	ICD	WAS
BerkeleyDB	60	92	86	**51**
JHotDraw	81	107	94	**72**
NIPS-05	402	435	418	**400**
Obama-speech	582	1300	1102	**412**
Average	281	483.5	425	**234**

TABLE II
Comparison of SBP, HDP, ICD and WAS using *topic coherence* (greater is better).

Dataset	SBP	HDP	ICD	WAS
BerkeleyDB	-27.9	-39.6	-19.1	**-18.2**
JHotDraw	-28.2	-82.2	-46.2	**-23.2**
NIPS-05	-37.7	-49.7	-44.6	**-35.7**
Obama-speech	-52.5	-68.2	-70.9	**-42.5**
Average	-36.6	-59.9	-45.2	**-30.15**

B. Results

Perplexity. We have randomly picked one-third of the datasets as held-out datasets and used the standard definition of perplexity as can be found in [19]. Lower value in perplexity means that the model fits the dataset better. Table I contains results on perplexity, where WAS outperforms all others.

We have also evaluated WAS on perplexity using AT-Science dataset (only using articles). Figure 4 shows the comparative results. In this experiment we vary the truncation level of SBP from 1 to 10 and observe that WAS effectively learns the appropriate number of topic vectors automatically, affirming the ability of Bayesian nonparametric methods in automatic model selection.

Topic coherence. Topic coherence is found to approximate user experience qualitatively[16]. Values closer to zero indicates better coherence. We have used top 5 words to compute coherence of a topic, and report in Table II, where WAS is the best. This is a significant result which shows that WAS does not degrade the quality of topics while improving upon perplexity.

Retrieval task. In this experiment, we have focused on a task oriented evaluation; how WAS ultimately performs in an information retrieval setting. We have used AT-Science-L dataset, in which there are 1443 specific comments with an average of 2.9 specific comments per article. Here, the retrieval task is to find out specific comments given each article as a query. Based on the retrieved set and gold-standard, we

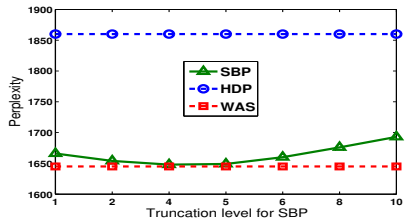


Fig. 4. Comparison of SBP, HDP and WAS on AT-Science dataset.

TABLE III
COMPARISON OF SBP, HDP AND WAS ON RETRIEVAL TASK USING *precision, recall* AND *F1* (GREATER IS BETTER).

Model	Precision	Recall	F1
WAS	**0.62**	**0.61**	**0.62**
HDP	0.28	0.25	0.26
SBP	0.60	0.61	0.60

compute precision, recall and F1 and average over articles. Table III shows the results. SBP is observed to be a better prior than HDP, however SBP is outperformed by WAS.

VI. CONCLUSION

We propose weight-agnostic hierarchical stick-breaking process (WAS) that shares atoms across groups but make the weights mutually independent. We have studied WAS theoretically and also experimentally. Although HDP is the state of the art hierarchical Bayesian nonparametric model for groups of observations, in some practical cases it performs worse than even the ad-hoc approach based on SBP. The reason of failure of HDP is due to dependence of weights across groups. WAS is a principled alternative found to outperform SBP, HDP and ICD across various performance measures. WAS is easy to apply due to the presence of tractable PPFs and efficient truncation-free collapsed Gibbs sampling inference.

REFERENCES

[1] P. Müller, F. Quintana, and G. Rosner, "A method for combining inference across related nonparametric bayesian models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 66, no. 3, pp. 735–749, 2004.

[2] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[3] S. Williamson, C. Wang, K. A. Heller, and D. M. Blei, "The IBP Compound Dirichlet Process and its Application to Focused Topic Modeling," in *International Conference on Machine Learning(ICML)*, 2010.

[4] M. Das, S. Bhattacharya, C. Bhattacharyya, and K. Gopinath, "Subtle topic models and discovering subtly manifested software concerns automatically," in *Proceedings of International Conference on Machine Learning*, 2013.

[5] M. Das, T. Bansal, and C. Bhattacharyya, "Going beyond Corr-LDA for Detecting Specific Comments on News & Blogs," in *ACM WSDM*. ACM, 2014.

[6] T. Hospedales, J. Li, S. Gong, and T. Xiang, "Identifying Rare and Subtle Behaviors: A Weakly Supervised Joint Topic Model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 2451–2464, 2011.

[7] H. Ishwaran and L. James, "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 161–173, 2001.

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

[9] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[10] P. Muller, F. Quintana, and G. Rosner, "A method for combining inference across related nonparametric bayesian models," *Journal of the Royal Statistical Society*, vol. 66, pp. 735–749, 2004.

[11] H. Ishwaran and L. James, "Some further developments for stick-breaking priors: finite and infinite clustering and classification," *Sankhya: The Indian Journal of Statistics*, vol. 65, pp. 577–592, 2003.

[12] J. Pitman and M. Yor, "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," *The Annals of Probability*, vol. 25, no. 2, pp. 855–900, 1997.

[13] M. Das, T. Bansal, and C. Bhattacharyya, "Ordered stick-breaking process for sequential MCMC inference of Bayesian nonparametric models," in *International Conference of Machine Learning (ICML)*, 2015.

[14] T. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Indian buffet process," in *Advances in Neural Information Processing Systems 18*, 2005, pp. 475–482.

[15] ArsTechnica, "http://arstechnica.com/gadgets/2013/07/the-2013-nexus-7-review-meet-the-new-standard-for-android-tablets," 2013.

[16] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Empirical Methods in Natural Language Processing*, 2011, pp. 262–272.

[17] A. Globerson, G. Chechik, F. Pereira, and N. Tishby, "Euclidean Embedding of Co-occurrence Data," *The Journal of Machine Learning Research(JMLR)*, vol. 8, pp. 2265–2295, 2007.

[18] S. Apel, C. Kastner, and C. Lengauer, "FEATUREHOUSE: Language-Independent, Automated Software Composition," in *Proceedings of the 31st International Conference on Software Engineering(ICSE)*, 2009, pp. 221–231.

[19] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

APPENDIX

S1. Proof of Theorem 1:

Proof. For any $x \in A$, $\Gamma(x) = \epsilon \mathrm{J}(x) + (1-\epsilon)\mathrm{H}(x) = 0$. For any $y \in G$, $\Gamma(y) = \epsilon \mathrm{J}(y) + (1-\epsilon)\mathrm{H}(y) = \epsilon \mathrm{J}(y) = \epsilon \eta_j > 0$. Hence Γ is a relatively-diffuse probability measure with respect to (A, G). □

S2. Proof of Theorem 2:

Proof. For β_{0j} present in both groups s, t that is $\beta_{0j} \in (A_{sn} \cap A_{tn})$ then $\theta_{s\sigma_s(j)}, \theta_{t\sigma_t(j)} > 0$, and we can write

$$p(\theta_{s\sigma_s(j)}, \theta_{t\sigma_t(j)}|\boldsymbol{a}, \boldsymbol{b}) = p(v_{s1}, v_{s2}, \ldots, v_{s\sigma_s(j)}, v_{t1}, v_{t2}, \ldots, v_{t\sigma_t(j)}|\boldsymbol{a}, \boldsymbol{b})$$

As $v_{sj} \sim Beta(a_j, b_j)$,all vs are independent given hyper-parameters $\boldsymbol{a}, \boldsymbol{b}$. Hence, we can express above as

$$\begin{aligned} &p(v_{s1}, v_{s2}, \ldots, v_{s\sigma_s(j)}, v_{t1}, v_{t2}, \ldots, v_{t\sigma_t(j)}|\boldsymbol{a}, \boldsymbol{b}) \\ &= p(v_{s1}, v_{s2}, \ldots, v_{s\sigma_s(j)}|\boldsymbol{a}, \boldsymbol{b}) p(v_{t1}, v_{t2}, \ldots, v_{t\sigma_t(j)}|\boldsymbol{a}, \boldsymbol{b}) \\ &= p(\theta_{s\sigma_s(l)}|\boldsymbol{a}, \boldsymbol{b}) p(\theta_{t\sigma_t(m)}|\boldsymbol{a}, \boldsymbol{b}) \end{aligned} \tag{15}$$

Similarly, for $\beta_{0j} \in (\cap_{s\in\mathcal{S}} A_{sn})$, where $\mathcal{S} \subseteq [S]$ we can write

$$\begin{aligned} &p(\prod_{s\in\mathcal{S}} \theta_{s\sigma_s(j)}|\boldsymbol{a}, \boldsymbol{b}) = p(\prod_{s\in\mathcal{S}} (v_{s1}, v_{s2}, \ldots, v_{s\sigma_s(j)})|\boldsymbol{a}, \boldsymbol{b}) \\ &= \prod_{s\in\mathcal{S}} p(v_{s1}, v_{s2}, \ldots, v_{s\sigma_s(j)}|\boldsymbol{a}, \boldsymbol{b}) \\ &= \prod_{s\in\mathcal{S}} p(\theta_{s\sigma_s(l)}|\boldsymbol{a}, \boldsymbol{b}) p(\theta_{t\sigma_t(m)}|\boldsymbol{a}, \boldsymbol{b}) \end{aligned} \tag{16}$$

This holds for any β_{0j}. This proves the Theorem. □

S3. Proof of Theorem 3:

Proof. By definition, k_{sn} is the cardinality of the set A_s before sampling ϕ_{sn}. So for any $n > 0$, $k_{sn} = k_{sn-1}$ if no new atom is sampled, and $k_{sn} = k_{sn-1} + 1$ if a new atom is sampled from Γ_{sn}, as Γ_{sn} is a relatively-diffuse probability measure with respect to group s. That makes the cardinality of set A_s to strictly monotonically increase whenever a new atom is sampled.

For any $K > 0$, there is an n' such that $k_{sn'} > K$, otherwise K is the upper bound of k_{sn}. So we can say

$$\lim_{n\to\infty} k_{sn} = \infty \quad a.s. \tag{17}$$

On the other hand, by definition $\alpha_{sn} \leq \alpha_{sn-1}$ and $\alpha_{sn} < \alpha_{sn-1}$ whenever a new atom is sampled. Moreover, α_{sn} is bounded below by zero. For any $\epsilon > 0$ there is an n' such that $\alpha_{sn'} < \epsilon$, otherwise ϵ is the lower bound of α_{sn}. Hence,

$$\lim_{n\to\infty} \alpha_{sn} = 0 \quad a.s. \tag{18}$$

Thus, we can write $\lim_{n\to\infty} \mathrm{P}_{sn}$ as $\lim_{n\to\infty} \sum_{j=1}^{k_{sn}} \theta_{sj}\delta_{\beta_{sj}} + \lim_{n\to\infty} \alpha_{sn}\Gamma_{sn}$ which can be expressed as $\sum_{j=1}^{\infty} \theta_{sj}\delta_{\beta_{sj}}$.

For $n \to \infty$, there is a permutation function σ_s for each group s such that $\sigma_s(j) = k$ when $\beta_{sk} = \beta_{0j}$ for some k. This holds following uniqueness of atoms in each A_s. For atoms not present in group s, that is if there is no k such that $\beta_{sk} = \beta_{0j}$, then $\theta_{sk} = 0$. Hence, $\theta_{s\sigma_s(j)} \geq 0$ and $\sum_{j=1}^{\infty} \theta_{s\sigma_s(j)} = 1$ because $\lim_{n\to\infty} \alpha_{sn} = 0$ a.s.

Thus we can say, $\lim_{n\to\infty} \mathrm{P}_{sn} = \sum_{j=1}^{\infty} \theta_{s\sigma_s(j)}\delta_{\beta_{s\sigma_s(j)}} = \sum_{j=1}^{\infty} \theta_{s\sigma_s(j)}\delta_{\beta_{0j}}$.

That proves the theorem. □

S4. Proof of Theorem 4:

Before we prove the main result, we need the following Lemma.

Lemma 1. *Let,* (v_{sj}) *be defined as in Eq.* (9)*. Then* $\forall s, j$, $v_{sj}|z_{s,1:n-1}, a_j, b_j \sim Beta(a_j + g_{sj} - 1, b_j + h_{sj})$.

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

$$\begin{aligned} &p(z_{sn} = j|z_{s,1:n-1}, v_{s,1:k_{sn}}) = \\ &v_{sj} \prod_{l=1}^{j-1}(1 - v_{sl}),\ j \in [k_{sn}] \\ &p(z_{sn} = k_{sn} + 1|z_{s,1:n-1}, v_{s,1:k_{sn}}) = \prod_{l=1}^{k_{sn}}(1 - v_{sl}) \end{aligned}$$

Now, following [**?**], it is straight forward to see that,

$$\begin{aligned} &p(z_{s1}, \ldots, z_{s,n-1}|v_{s,1:k_{sn}}) \\ &= \prod_{j=1}^{k_{sn}} \left((1 - v_{sj})^{h_{sj}} v_{sj}^{g_{sj}-1}\right) \end{aligned}$$

Now, we compute the posterior $p(v_{s1}, \ldots, v_{sk_{sn}}|z_{s,1:n-1})$ as follows.

$$\begin{aligned} &\propto p(z_{s1}, \ldots, z_{s,n-1}|v_{s,1:k_{sn}}) p(v_{s1}, \ldots, v_{sk_{sn}}) \\ &\propto \prod_{j=1}^{k_{sn}} \left((1 - v_{sj})^{h_{sj}} v_{sj}^{g_{sj}-1} v_{sj}^{a_j-1}(1 - v_{sj})^{b_j-1}\right) \end{aligned}$$

After marginalizing over all other v_{sl}, $l \in [k_{sn}]\backslash j$, the lemma follows. □

Proof. By definition of WAS, for $1 \leq j \leq k_{sn}$, $p(\phi_{sn} = \beta_{sj}|A_{sn}, \{v_{sl}\}) = p(z_{sn} = j|z_{s,1:n-1}, \{v_{sl}\}) = \theta_{sj}$.
Now by definition of PPFs in Eq. (13), one can write π_{sj} as

$$\begin{aligned} &\mathbb{E}\left[\theta_{sj}|z_{s,1:n-1}, \boldsymbol{a}, \boldsymbol{b}\right] \\ &= \mathbb{E}\left[v_{sj} \prod_{l=1}^{j-1}(1 - v_{sl})\,|z_{s,1:n-1}, \boldsymbol{a}, \boldsymbol{b}\right] \\ &= \mathbb{E}\left[v_{sj}\,|z_{s,1:n-1}, a_j, b_j\right] \times \\ &\prod_{l=1}^{j-1} \mathbb{E}\left[(1 - v_{sl})\,|z_{s,1:n-1}, a_l, b_l\right] \end{aligned}$$

The second equation follows using the independence property of $\{v_{sj}\}$. Following definition of WAS, we similarly get σ_{sn} defined in Eq. (13),

$$\xi_{sn} = \prod_{l=1}^{k_{sn}} \mathbb{E}[(1 - v_{sl})|z_{s,1:n-1}, a_l, b_l]$$

First part follows using Lemma 1.

For the second part, recall the definition of relatively-diffuse base measure Γ_{sn} and WAS. $\eta_{sl} = p(d_{sj} = l|A_{1:S}, \Theta)$, $l \in [K_{sn}]$ which by definition of WAS is proportional to 1 and $\omega_{sn} = p(d_{sj} = K_{sn}+1|A_{1:S}, \Theta)$ is proportional to ζ. As K_{sn} is the cardinality of G_{sn}, we get the normalizing factor as $K_{sn} + \zeta$. The rest follows immediately. □

S5. MCMC inference for WAS based topic model:

Recall that, ϕ_{dsn} is a sample from WAS at the time of processing the nth word in the sth sentence in the dth document. If P_{dsn} is the sampling distribution corresponding to WAS then $\phi_{dsn} \sim \mathrm{P}_{dsn}$ and $\mathrm{P}_{dsn} = \sum_{j=1}^{k_{dsn}} \theta_{dsj}\delta_{\beta_{dsj}} + \alpha_{dsn}\Gamma_{dsn}$. (β_{dsj}) is the set of distributions over topics, and ϕ_{dsn} takes value from (β_{dsj}). Even if a new value is sampled for ϕ_{dsn}, that will be added to the set (β_{dsj}). So let $b_{dsn} = j$ iff $\phi_{dsn} = \beta_{dsj}$. On the other hand, (β_{dsj}) refers to the document wise topic vectors $(\beta_{d01}, \beta_{d02}, \ldots)$.

Let $c_{dsj} = l$ iff $\beta_{dsj} = \beta_{d0l}$. Thus, $\phi_{dsn} = \beta_{d0l}$ if $b_{dsn} = j$ and $c_{dsj} = l$. Now, (β_{d0l}) are distributions over topics (λ_k), and (λ_k) is global to the entire corpus. So, $\psi_{dsn} \sim \phi_{dsn}$ implies that $\psi_{dsn} \sim \beta_{d0l}$ if $b_{dsn} = j$ and $c_{dsj} = l$. Thus, ψ_{dsn} is a distribution over words in the vocabulary. Thus $w_{dsn} \sim \psi_{dsn}$ gives a word from the vocabulary. Let $z_{dsn} = k$ if $\psi_{dsn} = \lambda_k$.

Collapsing. Note that, using $\beta_{d0l}, b_{dsn}, c_{dsj}, z_{dsn}, \lambda_k$ we can retrieve other variables. So we need not maintain

$\mathrm{P}_{dsn}, \phi_{dsn}, \psi_{dsn}, \beta_{dsj}$; that will save space and will also simplify the inference procedure. Furthermore, we can integrate out (β_{d0l}) and (λ_k) using Dirichlet multinomial conjugacy; that will collapse the search space for MCMC leading to faster convergence. This procedure is standard and can be found in [2, 4].

Notation used in inference. We will use notation for counts as follows. d is the document index, s is the sentence index and n is the word position index. m represents the counting variable and indices are put in the subscript, where "." represents marginalization. Super-script denotes that in all counts the current word is excluded (we do not repeat this in the text follows). Thus, $m^{-dsn}_{\ldots kw_{dsn}}$ is the number of times word type w_{dsn} is associated with topic k. $m^{-dsn}_{\ldots k.}$ represents the number of times topic k is used in the whole corpus. $m^{-dsn}_{d.lk.}$ is the number of times topic k and topic vector l are used. $m^{-dsn}_{d.l..}$ denotes the number of times topic vector l is used. $m^{-dsn}_{dsl..}$ is number of times topic vector indexed by l is used, and $m^{-dsn}_{ds\ldots}$ is the number of words in the sentence. K is the truncation level for topics. For the sake of brevity, in the following text we do not put all the variables in the conditional hoping that is easy to track following the generative process (Algorithm 1).

Inference steps. We will sample z, b and c as follows, and iterate until convergence.

1) Sampling z. The conditional probability of topic assignment of word n at sentence s in document d can be expressed as follows.

$$\begin{aligned} p(z_{dsn} = k|\mathbf{w}, \mathbf{z}^{-dsn}) &\propto p(w_{dsn}|z_{dsn} = k, \mathbf{z}^{-dsn}) \\ & p(z_{dsn} = k|b_{dsn} = j, c_{dsj} = l, \mathbf{z}^{-dsn}) \\ &= \frac{\eta + m^{-dsn}_{\ldots kw_{dsn}}}{V\eta + m^{-dsn}_{\ldots k.}} \quad \frac{\kappa + m^{-dsn}_{d.lk.}}{\sum_k \kappa + m^{-dsn}_{d.l..}} \end{aligned} \tag{19}$$

2) Sampling b. The posterior probability of selecting local index of a topic vector for a word can be found to be as below.

$$\begin{aligned} & p(b_{dsn} = j|\mathbf{b}^{-dsn}, \mathbf{z}, c_{dsj} = l) \propto \\ & p(z_{dsn}|b_{dsn} = j, c_{dsj} = l, \mathbf{z}^{-dsn}) p(b_{dsn} = j|\mathbf{b}^{-dsn}) \\ & = \frac{\alpha + m^{-dsn}_{d.lz_{dsn}.}}{\sum_k \alpha + n^{-dsn}_{d.l..}} \quad p(b_{dsn} = j|\mathbf{b}^{-dsn}) \end{aligned} \tag{20}$$

3) Sampling c. The posterior probability of selecting global index of a topic vector for a word can be found to be as below.

$$\begin{aligned} & p(c_{dsj} = l|\mathbf{c}, \mathbf{b}, \mathbf{z}) \propto \\ & p(c_{dsj} = l|\mathbf{c}^{-dsj}) \prod_{n=1}^{N_{ds}} p(z_{dsn}|b_{dsn} = j, c_{dsj} = l)^{\mathbb{I}[b_{dsn}=j]} \\ & = \prod_{n=1}^{N_{ds}} \frac{\alpha + m^{-dsn}_{d.lz_{dsn}.}}{\sum_k \alpha + n^{-dsn}_{d.l..}} \quad p(c_{dsj} = l|\mathbf{c}^{-dsj}) \end{aligned} \tag{21}$$

$p(b_{dsn} = j|\mathbf{b}^{-dsn})$ and $p(c_{dsj} = l|\mathbf{c}^{-dsj})$ comes from the PPFs of WAS as given in Theorem 4.

Eq. (19), Eq. (20) and Eq. (21) form the collapsed Gibbs sampling inference. Unlike SBP in [4], PPFs of WAS allows truncation-free inference.