Supplementary Materials
Compressed Nonparametric Language Modelling

Ehsan Shareghi

August 2017

1 Sampling under the joint distribution of $n_w^{\mathbf{u}}, t_w^{\mathbf{u}}$

We sample using the joint distribution in eqn.1. All we need to know for developing a sampler are (i) *how many tables of each type ($t_w^{\mathbf{u}}$)*, and (ii) *how many customers of each type ($n_w^{\mathbf{u}}$)* are in each restaurant without knowing where exactly each customer is seated.

$$P(\{\eta^{\mathbf{u}}\}^{\mathbf{u}\in\text{HPYP}}) = \prod_w H(.)^{t_w^{\epsilon}} \prod_{\mathbf{u}} \left(\frac{(\theta^{\mathbf{u}}|d^{\mathbf{u}})_{t_\cdot^{\mathbf{u}}}}{(\theta^{\mathbf{u}}|1)_{n_\cdot^{\mathbf{u}}}} \prod_w S_{d^{\mathbf{u}}}(n_w^{\mathbf{u}}, t_w^{\mathbf{u}}) \right) \tag{1}$$

The joint distribution in eqn. 1 allows efficient sampling for $t_w^{\mathbf{u}}$ and $n_w^{\mathbf{u}}$, starting from the data level and going up in the hierarchy. The only expensive computation is for the Stirling numbers which are cached during the runtime, as fixed KN discounts are used. We use the exact recursive formulation of Stirling numbers (Buntine and Hutter, 2012) and switch to asymptotic approximation[1] when t or n are large, i.e. ≥ 8000. For each $G^{\mathbf{u}} \in \gamma^+$, except the leaf level, the $n_w^{\mathbf{u}}$'s will be sampled jointly as $t_w^{\psi(\mathbf{u})}$'s are sampled, where $\psi(\mathbf{u}) \in \text{children}(\mathbf{u})$. Starting from the leaf level of the hierarchy, the $n_w^{\mathbf{u}}$'s are read from the data, hence fixed and $t_w^{\mathbf{u}}$'s are sampled while satisfying the constraints for $\{n_w^{\mathbf{u}}, t_w^{\mathbf{u}}\}$,

$$0 < t_w^{\mathbf{u}} \leq n_w^{\mathbf{u}} \tag{2}$$

$$t_w^{\mathbf{u}} = n_w^{\mathbf{u}} \quad \texttt{if} \quad n_w^{\mathbf{u}} \in \{0, 1\} \tag{3}$$

$$n_w^{\mathbf{u}} = \sum_{\psi\in\text{children}(\mathbf{u})} t_w^{\psi} \tag{4}$$

Given a sampled $t_w^{\mathbf{u}^*}$ at the leaf level $\mathbf{u}^*$, the $n_w^{\pi(\mathbf{u}^*)}$ is updated as,

$$n_w^{\pi(\mathbf{u}^*)} = t_w^{\mathbf{u}^*} + \sum_{\psi\in\text{children}(\pi(\mathbf{u}^*))\wedge\psi\neq\mathbf{u}^*} t_w^{\psi}. \tag{5}$$

[1] The asymptotic approximation is defined via using the Stirling's approximation for factorials, $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n \Rightarrow \Gamma(n+1) \approx \sqrt{2\pi n}(\frac{n}{e})^n$ as $n \to \infty$.

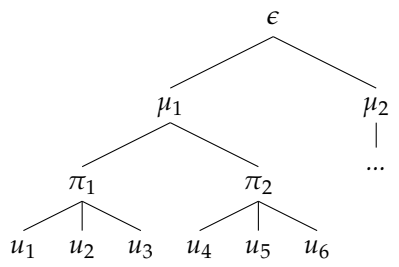

Figure 1: Hierarchy of Chinese Restaurants, each node represents a restaurant and u_i are the leaf levels where data counts are observed.

Figure 1 illustrates an example hierarchy of HPYP. Using eqn.1 we can define a posterior probability of $P(t_w^{\mathbf{u_1}^*}|\{\eta^{\mathbf{u}}\}^{\mathbf{u}\in\text{HPYP}} - t_w^{\mathbf{u_1}^*})$ for this example as follows,

$$\begin{aligned}
P(t_w^{\mathbf{u_1}^*}|\{\eta^{\mathbf{u}}\}^{\mathbf{u}\in\text{HPYP}} - t_w^{\mathbf{u_1}^*}) = & \frac{(\theta^{\mathbf{u_1}}|d^{\mathbf{u_1}})_{t_\cdot^{\mathbf{u_1}}}}{(\theta^{\mathbf{u_1}}|1)_{n_\cdot^{\mathbf{u_1}}}} \prod_w S_{d^{\mathbf{u_1}}}(n_w^{\mathbf{u_1}}, t_w^{\mathbf{u_1}}) \\
& \times \frac{(\theta^{\pi_1}|d^{\pi_1})_{t_\cdot^{\pi_1}}}{(\theta^{\pi_1}|1)_{t_\cdot^{\mathbf{u_1}}+t_\cdot^{\mathbf{u_2}}+t_\cdot^{\mathbf{u_3}}}} \prod_w S_{d^{\pi_1}}(t_w^{\mathbf{u_1}} + t_w^{\mathbf{u_2}} + t_w^{\mathbf{u_3}}, t_w^{\pi_1}) \\
& \times \frac{(\theta^{\mu_1}|d^{\mu_1})_{t_\cdot^{\mu_1}}}{(\theta^{\mu_1}|1)_{t_\cdot^{\pi_1}+t_\cdot^{\pi_2}}} \prod_w S_{d^{\mu_1}}(t_w^{\pi_1} + t_w^{\pi_2}, t_w^{\mu_1}) \\
& \times \frac{(\theta_\epsilon|d^\epsilon)_{t_\cdot^\epsilon}}{(\theta^\epsilon|1)_{t_\cdot^{\mu_1}+t_\cdot^{\mu_2}}} \prod_w S_{d^\epsilon}(t_w^{\mu_1} + t_w^{\mu_2}, t_w^\epsilon) \\
& \times \prod_w H(.)^{t_w^\epsilon}
\end{aligned} \tag{6}$$

where $H(.)$ in here is a uniform distribution. It is trivial to extend this to the general case. In practice we only need to compute eqn.6 up to a constant and can drop all the invariant terms which are independent from $t_w^{u_1}$. Extending this to the general case, the conditional probability of the sampled $t_w^{\mathbf{u}*}$ from eqn. 1, $P(t_w^{\mathbf{u}^*}|\{\eta^{\mathbf{u}}\}^{\mathbf{u}\in\text{HPYP}} - t_w^{\mathbf{u}^*})$, for the *non-root* levels while fixing all the independent variables is,

$$P(t_w^{\mathbf{u}^*}|\{\eta^{\mathbf{u}}\}^{\mathbf{u}\in\text{HPYP}} - t_w^{\mathbf{u}^*}) \propto \frac{(\theta^{\mathbf{u}^*}|d^{\mathbf{u}^*})_{t_\cdot^{\mathbf{u}^*}}}{(\theta_{\pi(\mathbf{u}^*)}|1)_{\sum_{\psi\in\text{children}(\pi(\mathbf{u}^*))} t_\cdot^\psi}} S_{d^{\mathbf{u}^*}}(n_w^{\mathbf{u}^*}, t_w^{\mathbf{u}^*}) S_{d^{\pi(\mathbf{u}^*)}}(n_w^{\pi(\mathbf{u}^*)}, t_w^{\pi(\mathbf{u}^*)}) \tag{7}$$

where $t_\cdot^{\mathbf{u}^*} = t_w^{\mathbf{u}^*} + \sum_{v\neq w} t_v^{\mathbf{u}^*}$, and for the *root* level is,

$$P(t_w^\varepsilon|\{\eta^{\mathbf{u}}\}^{\mathbf{u}\in\text{HPYP}} - t_w^\varepsilon) \propto H(.)^{t_w^\varepsilon}(\theta^\varepsilon|d^\varepsilon)_{t_\cdot^\varepsilon} S_{d^\varepsilon}(n_w^\varepsilon, t_w^\varepsilon). \tag{8}$$

Given sampled $t_w^{\mathbf{u}^*}, n_w^{\mathbf{u}^*}$ for a context $\mathbf{u}$, the concentration parameter $\theta^{\mathbf{u}^*}$ is then sampled via auxiliary variables using a Gamma(a,b) prior as outlined in Section 2.

2 Sampling concentration parameter $\theta^{\mathbf{u}}$

Based on the results of (Antoniak, 1974; Escobar and West, 1995; Teh et al., 2012) we propose the following sampler for the hierarchical Pitman-Yor process case. We construct our sampler via the joint distribution in eqn. 9 using auxiliary variables. Starting from the joint distribution,

$$P(\{\eta^{\mathbf{u}}, \{\mathcal{I}_w^{\mathbf{u}}\}_{w\in\sigma_{\mathbf{u}}}\}^{\mathbf{u}\in\text{HPYP}}) = \prod_w H(.)^{t_w^\varepsilon} \prod_{\mathbf{u}} \left(\frac{(\theta^{\mathbf{u}}|d^{\mathbf{u}})_{t_\cdot^{\mathbf{u}}}}{(\theta^{\mathbf{u}}|1)_{n_\cdot^u}} \prod_w \prod_{k=1}^{t_w^{\mathbf{u}}} (1 - d^{\mathbf{u}}|1)_{n_{wk}^{\mathbf{u}}-1} \right) \tag{9}$$

we can re-write the joint with the transformations and include the auxiliary variables.

2.1 Introducing auxiliary variables into the joint distribution

We transform the joint, using a reformation of Pochhammer symbol ratios of Def.5, as

$$P(\{\eta^{\mathbf{u}}, \{\mathcal{I}_w^{\mathbf{u}}\}_{w\in\mathbf{u}}\}^{\mathbf{u}\in\text{HPYP}}) = \prod_w H(.)^{t_w^\varepsilon} \prod_{\mathbf{u}} \left(\frac{\overbrace{(\theta^{\mathbf{u}} + d^{\mathbf{u}}|d^{\mathbf{u}})_{t_\cdot^{\mathbf{u}}-1}}^{\text{numerator}}}{\underbrace{(\theta^{\mathbf{u}} + 1|1)_{n_\cdot^u-1}}_{\text{denominator}}} \prod_w \prod_{k=1}^{t_w^{\mathbf{u}}} \underbrace{(1 - d^{\mathbf{u}}|1)_{n_{wk}^{\mathbf{u}}-1}}_{\text{right term}} \right). \tag{10}$$

Now, to sample, we use transform each component of eqn. 10 (denominator,numerator, and right term) separately as follows,

- denominator

$$\frac{1}{(\theta^{\mathbf{u}}+1|1)_{n_{\cdot}^{\mathbf{u}}-1}} \overset{\text{Def.4}}{=} \frac{1}{\frac{\Gamma(\theta^{\mathbf{u}}+n_{\cdot}^{\mathbf{u}})}{\Gamma(\theta^{\mathbf{u}}+1)}} \overset{\text{Def.2}}{=} \frac{\theta^{\mathbf{u}}\Gamma(\theta^{\mathbf{u}})}{\Gamma(\theta^{\mathbf{u}}+n_{\cdot}^{\mathbf{u}})} = \frac{\Gamma(\theta^{\mathbf{u}})}{\Gamma(\theta^{\mathbf{u}}+n_{\cdot}^{\mathbf{u}}-1)}\frac{\theta^{\mathbf{u}}}{\theta^{\mathbf{u}}+n_{\cdot}^{\mathbf{u}}-1} \tag{11}$$
$$\overset{\text{Def.6}}{=} \frac{\theta^{\mathbf{u}}}{\theta^{\mathbf{u}}+n_{\cdot}^{\mathbf{u}}-1}\frac{1}{\Gamma(n_{\cdot}^{\mathbf{u}}-1)}\int_0^1 (\frac{\theta^{\mathbf{u}}+n_{\cdot}^{\mathbf{u}}-1}{\theta^{\mathbf{u}}})(\zeta^{\mathbf{u}})^{\theta^{\mathbf{u}}}(1-\zeta^{\mathbf{u}})^{n_{\cdot}^{\mathbf{u}}-2}d\zeta^{\mathbf{u}}$$
$$= \frac{1}{\Gamma(n_{\cdot}^{\mathbf{u}}-1)}\int_0^1 (\zeta^{\mathbf{u}})^{\theta^{\mathbf{u}}}(1-\zeta^{\mathbf{u}})^{n_{\cdot}^{\mathbf{u}}-2}d\zeta^{\mathbf{u}}$$

- numerator is refined using a binary auxiliary variable $\xi^{\mathbf{u}i}$

$$(\theta^{\mathbf{u}}+d^{\mathbf{u}}|d^{\mathbf{u}})_{t_{\cdot}^{\mathbf{u}}-1} \overset{\text{Def.1}}{=} \prod_{i=0}^{t_{\cdot}^{\mathbf{u}}-1}(\theta^{\mathbf{u}}+id^{\mathbf{u}}) = \prod_{i=0}^{t_{\cdot}^{\mathbf{u}}-1}\sum_{\xi^{\mathbf{u}i}\in\{0,1\}}(\theta^{\mathbf{u}})^{\xi^{\mathbf{u}i}}(id^{\mathbf{u}})^{1-\xi^{\mathbf{u}i}} \tag{12}$$

- right term is refined using a binary auxiliary variable $\lambda^{\mathbf{u}j}$

$$(1-d^{\mathbf{u}}|1)_{n_{wk}^{\mathbf{u}}-1} = \prod_{j=1}^{n_{wk}^{\mathbf{u}}-1}(j-d^{\mathbf{u}}) = \prod_{j=1}^{n_{wk}^{\mathbf{u}}-1}\sum_{\lambda^{\mathbf{u}j}\in\{0,1\}}(j-1)^{\lambda^{\mathbf{u}j}}(1-d^{\mathbf{u}})^{1-\lambda^{\mathbf{u}j}} \tag{13}$$

We can re-write the joint with the transformations and include the auxiliary variables as,

$$P(\{\eta^{\mathbf{u}},\zeta^{\mathbf{u}},\xi^{\mathbf{u}},\lambda^{\mathbf{u}},\{\mathcal{I}_w^{\mathbf{u}}\}_{w\in\mathbf{u}}\}^{\mathbf{u}\in\text{HPYP}}) = \prod_w H(.)^{t_w^{\varepsilon}}\prod_{\mathbf{u}}\Bigg(\overbrace{\prod_{i=0}^{t_{\cdot}^{\mathbf{u}}-1}\sum_{\xi^{\mathbf{u}i}\in\{0,1\}}(\theta^{\mathbf{u}})^{\xi^{\mathbf{u}i}}(id^{\mathbf{u}})^{1-\xi^{\mathbf{u}i}}}^{\text{numerator replaced with eqn.12}} \underbrace{\frac{1}{\Gamma(n_{\cdot}^{\mathbf{u}}-1)}\int_0^1(\zeta^{\mathbf{u}})^{\theta^{\mathbf{u}}}(1-\zeta^{\mathbf{u}})^{n_{\cdot}^{\mathbf{u}}-2}d\zeta^{\mathbf{u}}}_{\text{denominator replaced with eqn.11}} \underbrace{\prod_w\prod_{k=1}^{t_w^{\mathbf{u}}}\prod_{j=1}^{n_{wk}^{\mathbf{u}}-1}\sum_{\lambda^{\mathbf{u}j}\in\{0,1\}}(j-1)^{\lambda^{\mathbf{u}j}}(1-d^{\mathbf{u}})^{1-\lambda^{\mathbf{u}j}}}_{\text{right term replaced with eqn.13}}\Bigg) \tag{14}$$

$$P(\{\eta^{\mathbf{u}},\zeta^{\mathbf{u}},\xi^{\mathbf{u}},\lambda^{\mathbf{u}},\{\mathcal{I}_w^{\mathbf{u}}\}_{w\in\sigma_{\mathbf{u}}}\}^{\mathbf{u}\in\text{HPYP}}) = \prod_w H(.)^{t_w^{\varepsilon}}\prod_{\mathbf{u}}\Bigg(\prod_{i=0}^{t_{\cdot}^{\mathbf{u}}-1}\sum_{\xi^{\mathbf{u}i}\in\{0,1\}}(\theta^{\mathbf{u}})^{\xi^{\mathbf{u}i}}(id^{\mathbf{u}})^{1-\xi^{\mathbf{u}i}} \frac{1}{\Gamma(n_{\cdot}^{\mathbf{u}}-1)}\int_0^1(\zeta^{\mathbf{u}})^{\theta^{\mathbf{u}}}(1-\zeta^{\mathbf{u}})^{n_{\cdot}^{\mathbf{u}}-2}d\zeta^{\mathbf{u}} \prod_w\prod_{k=1}^{t_w^{\mathbf{u}}}\prod_{j=1}^{n_{wk}^{\mathbf{u}}-1}\sum_{\lambda^{\mathbf{u}j}\in\{0,1\}}(j-1)^{\lambda^{\mathbf{u}j}}(1-d^{\mathbf{u}})^{1-\lambda^{\mathbf{u}j}}\Bigg) \tag{15}$$

and compute the following conditional probability up-to a constant,

$$P(\{\theta^{\mathbf{u}},d^{\mathbf{u}},\zeta^{\mathbf{u}},\xi^{\mathbf{u}},\lambda^{\mathbf{u}}\}^{\mathbf{u}\in\text{HPYP}}|\{\eta^{\mathbf{u}},\{\mathcal{I}_w^{\mathbf{u}}\}_{w\in\sigma_{\mathbf{u}}}\}^{\mathbf{u}\in\text{HPYP}}) \propto P(\{\eta^{\mathbf{u}},\zeta^{\mathbf{u}},\xi^{\mathbf{u}},\lambda^{\mathbf{u}},\{\mathcal{I}_w^{\mathbf{u}}\}_{w\in\mathbf{u}}\}^{\mathbf{u}\in\text{HPYP}}) \tag{16}$$

which allows us to construct a Gibbs sampler for each of the desired parameters, including the auxiliary variables.

2.2 Sampling auxiliary variables

We use the proportionality in eqn.16 to develop the samplers for the auxiliary variables as follows,

$$P(\{\xi^{\mathbf{u}}\}^{\mathbf{u}\in\text{HPYP}}|...) \propto \prod_{\mathbf{u}}\prod_{i=0}^{t_{\cdot}^{\mathbf{u}}-1}\underbrace{(\theta^{\mathbf{u}})^{\xi^{\mathbf{u}i}}(id^{\mathbf{u}})^{1-\xi^{\mathbf{u}i}}}_{\text{Bernoulli distributed: Bernoulli}\left(\frac{\theta^{\mathbf{u}}}{\theta^{\mathbf{u}}+id^{\mathbf{u}}}\right)} \tag{17}$$

which requires sampling from a Bernoulli distribution, and

$$P(\{\zeta^{\mathbf{u}}\}^{\mathbf{u}\in\text{HPYP}}|...) \propto \prod_{\mathbf{u}} \underbrace{(\zeta^{\mathbf{u}})^{\theta^{\mathbf{u}}}(1-\zeta^{\mathbf{u}})^{n_{\cdot}^{\mathbf{u}}-2}}_{\text{Beta distributed: Beta}(\theta^{\mathbf{u}}+1,\, n_{\cdot}^{\mathbf{u}}-1)} \tag{18}$$

which relies on samples from a Beta distribution, and

$$P(\{\lambda^{\mathbf{u}}\}^{\mathbf{u}\in\text{HPYP}}|...) \propto \prod_{\mathbf{u}}\prod_{w}\prod_{k=1}^{t_w^{\mathbf{u}}}\prod_{j=1}^{n_{wk}^{\mathbf{u}}-1} \underbrace{(j-1)^{\lambda^{\mathbf{u}j}}(1-d^{\mathbf{u}})^{1-\lambda^{\mathbf{u}j}}}_{\text{Bernoulli distributed: Bernoulli}\left(\frac{j-1}{j-d^{\mathbf{u}}}\right)} \tag{19}$$

which only requires straightforward samplings from a Bernoulli distribution. Here, the ... in the conditioning context of $P(A|...)$ denotes $\{\eta^{\mathbf{u}}, \zeta^{\mathbf{u}}, \xi^{\mathbf{u}}, \lambda^{\mathbf{u}}, \{\mathcal{I}_w^{\mathbf{u}}\}_{w\in\sigma_{\mathbf{u}}}\}^{\mathbf{u}\in\text{HPYP}}$ with A excluded from it. The Bernoulli and Beta distributions are easy to sample from, hence allowing for an efficient auxiliary variable sampling.

2.3 Sampling concentration parameter $\theta^{\mathbf{u}}$

Given the sampled auxiliary variables, the concentration parameter is sampled assuming a Gamma$(a^{\mathbf{u}}, b^{\mathbf{u}})$ prior as follows,

$$\begin{aligned} P(\{\theta^{\mathbf{u}}\}^{\mathbf{u}\in\text{HPYP}}|...) &\propto \prod_{\mathbf{u}}\left(\underbrace{\frac{(b^{\mathbf{u}})^{a^{\mathbf{u}}}}{\Gamma(a^{\mathbf{u}})}(\theta^{\mathbf{u}})^{a^{\mathbf{u}}-1}e^{-\theta^{\mathbf{u}}b^{\mathbf{u}}}}_{\text{Gamma}(a^{\mathbf{u}},b^{\mathbf{u}})\text{ prior over }\theta^{\mathbf{u}}}\ \zeta^{\theta^{\mathbf{u}}}\prod_{i=0}^{t_{\cdot}^{\mathbf{u}}-1}(\theta^{\mathbf{u}})^{\xi^{\mathbf{u}i}}\right) \\ &\propto \prod_{\mathbf{u}}\left((\theta^{\mathbf{u}})^{a^{\mathbf{u}}-1}e^{-\theta^{\mathbf{u}}b^{\mathbf{u}}}\zeta^{\theta^{\mathbf{u}}}\prod_{i=0}^{t_{\cdot}^{\mathbf{u}}-1}\theta^{(\xi^{\mathbf{u}})^{\mathbf{u}i}}\right) \\ &\propto \prod_{\mathbf{u}}\left((\theta^{\mathbf{u}})^{a^{\mathbf{u}}-1}e^{-\theta^{\mathbf{u}}b^{\mathbf{u}}}\underbrace{e^{\theta^{\mathbf{u}}\ln\zeta}}_{\zeta^{\theta^{\mathbf{u}}}=e^{\theta^{\mathbf{u}}\ln\zeta}}\prod_{i=0}^{t_{\cdot}^{\mathbf{u}}-1}\theta^{(\xi^{\mathbf{u}})^{\mathbf{u}i}}\right) = \prod_{\mathbf{u}}\left(e^{-\theta^{\mathbf{u}}(b^{\mathbf{u}}-\ln\zeta^{\mathbf{u}})}\prod_{i=0}^{t_{\cdot}^{\mathbf{u}}-1}(\theta^{\mathbf{u}})^{a^{\mathbf{u}}-1}\theta^{(\xi^{\mathbf{u}})^{\mathbf{u}i}}\right) \end{aligned} \tag{20}$$

where in here, the ... in the conditioning context denotes $\{\eta^{\mathbf{u}}, \zeta^{\mathbf{u}}, \xi^{\mathbf{u}}, \lambda^{\mathbf{u}}, \{\mathcal{I}_w^{\mathbf{u}}\}_{w\in\sigma_{\mathbf{u}}}\}^{\mathbf{u}\in\text{HPYP}}$ with $\{\theta^{\mathbf{u}}\}^{\mathbf{u}\in\text{HPYP}}$ excluded from it. Here any term without $\theta^{\mathbf{u}}$ is cancelled out as they are fixed while $\theta^{\mathbf{u}}$ is sampled. This allows to sample a concentration parameter per distribution, an approach that was shown to be effective (Gasthaus et al., 2010).[2] To sample θ, a concentration parameter is first sampled from its prior, and then evaluated under eqn. 20. Since evaluating eqn. 20 relies on ξ, it requires to be sampled before θ. In practice we use the discount parameters of Kneser-Ney, but the discount parameter can also be sampled using a Beta$(\alpha^{\mathbf{u}}, \beta^{\mathbf{u}})$ prior. The parameters can be sampled periodically as $n_w^{\mathbf{u}}, t_w^{\mathbf{u}}$ are sampled, or sampled after $n_w^{\mathbf{u}}, t_w^{\mathbf{u}}$ sampling is done (as we do in this thesis).

3 More results

To validate some of the decisions made in designing the sampler, we test various settings of our model in Table 1. Our main model, CN, is based on 100 samples and uses KN discounts and untied concentration parameters (unique θ for each context). We test other variations of CN, testing a single change in each experiment and running the sampler for exactly the same amount of time: (i) lifting the Range Shrinking assumption and sampling $t_w^{\mathbf{u}}$ from its full range (see "NoShrinking" column), (ii) using tied concentration parameter where contexts of same size share θ (see "sample tied θ" column), (iii) using sampled discounts instead of KN discounts (see "sample tied d" column), (iv) using only 5 samples (see "5 samples" column), and (v) using only a single sample (see "1 sample" column).

[2] To tie the concentration parameters based on the corresponding context size $m = |\mathbf{u}|$,

$$P(\theta^m|...) \propto \underbrace{e^{-\theta^m(b^m-\sum_{\mathbf{u}:|\mathbf{u}|=m}\ln\zeta^{\mathbf{u}})}(\theta^m)^{a^m-1+\sum_{\mathbf{u}:|\mathbf{u}|=m}\sum_i^{t_{\cdot}^{\mathbf{u}}-1}\xi^{\mathbf{u}i}}}_{\text{Gamma distributed: Gamma}(a^m+\sum_{\mathbf{u}:|\mathbf{u}|=m}\sum_i^{t_{\cdot}^{\mathbf{u}}-1}\xi^{\mathbf{u}i},\, b^m-\sum_{\mathbf{u}:|\mathbf{u}|=m}\ln\zeta^{\mathbf{u}})} \tag{21}$$

Definition 1. ***Pochhammer symbol:***

$$(a|b)_c = a(a+1\times b)...(a+(c-1)\times b) = \prod_{i=0}^{c-1}(a+i\times b) \tag{22}$$

Definition 2. ***Gamma function:***

$$\Gamma(N) = (N-1)! \tag{23}$$

Definition 3. ***Beta function:***

$$\beta(N,T) = \int_0^1 x^{N-1}(1-x)^{T-1}dx \tag{24}$$

Definition 4. ***Pochhammer symbol and Gamma function:***

$$(\theta|d)_T = \frac{d^T\Gamma(\theta/d+T)}{\Gamma(\theta/d)} \tag{25}$$

Definition 5. ***Pochhammer symbols ratio:***

$$\frac{(\theta+d|d)_{T-1}}{(\theta+1|1)_{N-1}} = \frac{(\theta+d)(\theta+d+d)...(\theta+d+(T-2)d)}{(\theta+1)(\theta+1+1)...(\theta+1+(N-2))} \tag{26}$$

$$= \frac{\theta(\theta+d)(\theta+d+d)...(\theta+(T-1)d)}{\theta(\theta+1)(\theta+1+1)...(\theta+(N-1))} = \frac{(\theta|d)_T}{(\theta|1)_N} \tag{27}$$

Definition 6. ***Beta function and ratio of Gamma function:***

$$\beta(N,T) = \frac{\Gamma(N)\Gamma(T)}{\Gamma(N+T)} \tag{28}$$

combined with the definition in eqn.24,

$$\frac{\Gamma(T)}{\Gamma(T+N)} = \frac{(T+N)\beta(T+1,N)}{T\Gamma(N)} = \frac{(T+N)\int_0^1 x^T(1-x)^{N-1}dx}{T\Gamma(N)} \tag{29}$$

References

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*.

Buntine, W. and Hutter, M. (2012). A bayesian view of the Poisson-Dirichlet process. *arXiv preprint arXiv:1007.0296*.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430).

Gasthaus, J., Wood, F., and Teh, Y. W. (2010). Lossless compression based on the sequence memoizer. In *2010 Data Compression Conference (DCC 2010)*, pages 337–345.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2012). Hierarchical dirichlet processes. *Journal of the american statistical association*.

		tokens (M)		perplexity					
		TRAIN	TEST	NoShrinking	sample tied θ	sample tied d	5 samples	1 sample	CN
Europarl	EU-DE	54	0.06	1705	1540	1544	1572	1689	1543
	EU-FI	40	0.02	5401	4766	4756	4837	5160	4756
	EU-FR	66	0.08	1221	1052	1047	1071	1136	1048
	EU-ES	62	0.07	401	370	376	382	438	377
	EU-EN	61	0.07	1199	767	726	742	799	725
English CommonCrawl	125MiB	32	0.07	369	298	289	296	322	289
	250MiB	65	0.07	371	289	283	287	304	283
	1GiB	201	0.07	368	219	223	228	242	224
	2GiB	403	0.07	301	237	209	211	236	209
	4GiB	807	0.07	291	200	191	192	215	190
	8GiB	1617	0.07	273	181	173	178	196	174

Table 1: Perplexity results of our approach on different datasets and with various settings.