A Generating a CRT random variable

Lemma A.1. *A CRT random variable* $l \sim \text{CRT}(m, r)$ *can be generated with the summation of independent Bernoulli random variables as*

$$l = \sum_{n=1}^{m} b_n,\ b_n \sim \text{Bernoulli}\left(\frac{r}{n-1+r}\right). \tag{20}$$

Proof. Since l is the summation of independent Bernoulli random variables, its PGF becomes

$$C_L(z) = \prod_{n=1}^{m}\left(\frac{n-1}{n-1+r} + \frac{r}{n-1+r}z\right) = \frac{\Gamma(r)}{\Gamma(m+r)}\sum_{k=0}^{m}|s(m,k)|(rz)^k.$$

Thus we have $f_L(l|m,r) = \frac{C_L^{(l)}(0)}{l!} = \frac{\Gamma(r)}{\Gamma(m+r)}|s(m,l)|r^l,\ \ l = 0, 1, \cdots, m.$ □

B Dir-PFA and LDA

The Dirichlet Poisson factor analysis (Dir-PFA) model [5] is constructed as

$$\begin{aligned} &x_{ji} \sim F(\omega_{z_{ji}}),\ \ \omega_k \sim \text{Dir}(\eta, \cdots, \eta) \\ &N_j = \sum_{k=1}^{K} n_{jk},\ \ n_{jk} \sim \text{Pois}(\tilde{\lambda}_{jk}),\ \ \tilde{\boldsymbol{\lambda}}_j \sim \text{Dir}(50/K, \cdots, 50/K) \end{aligned} \tag{21}$$

where η is the Dirichlet smoothing parameter for the topic's distribution over the vocabulary, $n_{jk} = \sum_{i=1}^{N_j}\delta(z_{ji}=k)$, and the data likelihood $F(x_{ji};\omega_k)$ in topic modeling is $\omega_{v_{ji}k}$, the probability of the ith word in jth document under topic ω_k.

The Dir-PFA has the same block Gibbs sampling as LDA [34], expressed as

$$\begin{aligned} &\Pr(z_{ji}=k|-) \propto F(x_{ji};\omega_k)\tilde{\lambda}_{jk} \\ &(\omega_k|-) \sim \text{Dir}\left(\eta + \sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k, v_{ji}=1), \cdots, \eta + \sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k, v_{ji}=V)\right) \\ &(\tilde{\boldsymbol{\lambda}}_j|-) \sim \text{Dir}\left(50/K + n_{j1}, \cdots, 50/K + n_{jK}\right). \end{aligned} \tag{22}$$

C CRF-HDP

The CRF-HDP model [7, 26] is constructed as

$$\begin{aligned} &x_{ji} \sim F(\omega_{z_{ji}}),\ \ \omega_k \sim \text{Dir}(\eta, \cdots, \eta),\ \ z_{ji} \sim \text{Discrete}(\tilde{\boldsymbol{\lambda}}_j) \\ &\tilde{\boldsymbol{\lambda}}_j \sim \text{Dir}(\alpha\tilde{\boldsymbol{r}}),\ \ \alpha \sim \text{Gamma}(a_0, 1/b_0),\ \ \tilde{\boldsymbol{r}} \sim \text{Dir}(\gamma_0/K, \cdots, \gamma_0/K). \end{aligned} \tag{23}$$

Under the CRF metaphor, denote n_{jk} as the number of customers eating dish k in restaurant j and l_{jk} as the number of tables serving dish k in restaurant j, the direct assignment block Gibbs sampling

can be expressed as

$$
\begin{aligned}
&\Pr(z_{ji}=k|-) \propto F(x_{ji};\omega_k)\tilde{\lambda}_{jk}\\
&(l_{jk}|-) \sim \text{CRT}(n_{jk},\alpha\tilde{r}_k),\ \ w_j \sim \text{Beta}(\alpha+1,N_j),\ \ s_j \sim \text{Bernoulli}\left(\frac{N_j}{N_j+\alpha}\right)\\
&\alpha \sim \text{Gamma}\left(a_0+\sum_{j=1}^{J}\sum_{k=1}^{K} l_{jk}-\sum_{j=1}^{J} s_j, \frac{1}{b_0-\sum_j \ln w_j}\right)\\
&(\tilde{\boldsymbol{r}}|-) \sim \text{Dir}\left(\gamma_0/K+\sum_{j=1}^{J} l_{j1},\cdots,\gamma_0/K+\sum_{j=1}^{J} l_{jK}\right)\\
&(\tilde{\boldsymbol{\lambda}}_j|-) \sim \text{Dir}\left(\alpha\tilde{r}_1+n_{j1},\cdots,\alpha\tilde{r}_K+n_{jK}\right)\\
&(\omega_k|-) \sim \text{Dir}\left(\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=1),\cdots,\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=V)\right).
\end{aligned}
\tag{24}
$$

When $K\to\infty$, the concentration parameter γ_0 can be sampled as

$$
\begin{aligned}
&w_0 \sim \text{Beta}\left(\gamma_0+1,\sum_{j=1}^{J}\sum_{k=1}^{\infty} l_{jk}\right),\ \ \pi_0=\frac{e_0+K^+-1}{e_0+K^+-1+(f_0-\ln w_0)\sum_{j=1}^{J}\sum_{k=1}^{\infty} l_{jk}}\\
&\gamma_0 \sim \pi_0\text{Gamma}\left(e_0+K^+,\frac{1}{f_0-\ln w_0}\right)+(1-\pi_0)\text{Gamma}\left(e_0+K^+-1,\frac{1}{f_0-\ln w_0}\right)
\end{aligned}
\tag{25}
$$

where K^+ is the number of used atoms. Since it is infeasible in practice to let $K\to\infty$, directly using this method to sample γ_0 is only approximately correct, which may result in a biased estimate especially if K is not set large enough. Thus in the experiments, γ_0 is not sampled and is fixed as one. Note that for implementation convenience, it is also common to fix the concentration parameter α as one [25]. We find through experiments that learning this parameter usually results in obviously lower per-word perplexity for held out words, thus we allow the learning of α using the data augmentation method proposed in [7], which is modified from the one proposed in [24].

D NB-LDA

The NB-LDA model is constructed as

$$
\begin{aligned}
&x_{ji} \sim F(\omega_{z_{ji}}),\ \ \omega_k \sim \text{Dir}(\eta,\cdots,\eta)\\
&N_j=\sum_{k=1}^{K} n_{jk},\ \ n_{jk} \sim \text{Pois}(\lambda_{jk}),\ \ \lambda_{jk} \sim \text{Gamma}(r_j,p_j/(1-p_j))\\
&r_j \sim \text{Gamma}(\gamma_0,1/c),\ \ p_j \sim \text{Beta}(a_0,b_0),\ \ \gamma_0 \sim \text{Gamma}(e_0,1/f_0)
\end{aligned}
\tag{26}
$$

Note that letting $r_j\sim\text{Gamma}(\gamma_0,1/c)$, $\gamma_0\sim\text{Gamma}(e_0,1/f_0)$ allows different documents to share statistical strength for inferring their NB dispersion parameters.

The block Gibbs sampling can be expressed as

$$
\begin{aligned}
&\Pr(z_{ji}=k|-) \propto F(x_{ji};\omega_k)\lambda_{jk}\\
&(p_j|-) \sim \text{Beta}\left(a_0+N_j, b_0+Kr_j\right),\ \ p'_j=\frac{-K\ln(1-p_j)}{c-K\ln(1-p_j)}\\
&(l_{jk}|-) \sim\ \text{CRT}(n_{jk},r_j),\ \ l'_j\sim\text{CRT}(\sum_{k=1}^{K} l_{jk},\gamma_0),\ \ \gamma_0\sim\text{Gamma}\left(e_0+\sum_{j=1}^{J} l'_j, \frac{1}{f_0-\sum_{j=1}^{J}\ln(1-p'_j)}\right)\\
&(r_j|-)\sim\text{Gamma}\left(\gamma_0+\sum_{k=1}^{K} l_{jk}, \frac{1}{c-K\ln(1-p_j)}\right),\ \ (\lambda_{jk}|-)\sim\text{Gamma}(r_j+n_{jk},p_j)\\
&(\omega_k|-)\sim\text{Dir}\left(\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=1),\cdots,\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=V)\right).
\end{aligned}
\tag{27}
$$

E NB-HDP

The NB-HDP model is a special case of the Gamma-NB process model with $p_j = 0.5$. The hierarchical model and inference for the Gamma-NB process are shown in (16) and (18) of the main paper, respectively.

F NB-FTM

The NB-FTM model is a special case of zero-inflated NB process with $p_j = 0.5$, which is constructed as

$$
\begin{aligned}
&x_{ji}\sim F(\omega_{z_{ji}}),\ \ \omega_k\sim\text{Dir}(\eta,\cdots,\eta)\\
&N_j=\sum_{k=1}^{K} n_{jk},\ \ n_{jk}\sim\text{Pois}(\lambda_{jk})\\
&\lambda_{jk}\sim\text{Gamma}(r_k b_{jk}, 0.5/(1-0.5))\\
&r_k\sim\text{Gamma}(\gamma_0,1/c),\ \gamma_0\sim\text{Gamma}(e_0,1/f_0)\\
&b_{jk}\sim\text{Bernoulli}(\pi_k),\ \ \pi_k\sim\text{Beta}(c/K,c(1-1/K)).
\end{aligned}
\tag{28}
$$

The block Gibbs sampling can be expressed as

$$
\begin{aligned}
&\Pr(z_{ji}=k|-) \propto F(x_{ji};\omega_k)\lambda_{jk}\\
&b_{jk} \sim \delta(n_{jk}=0)\text{Bernoulli}\left(\frac{\pi_k(1-0.5)^{r_k}}{\pi_k(1-0.5)^{r_k}+(1-\pi_k)}\right)+\delta(n_{jk}>0)\\
&\pi_k \sim \text{Beta}\left(c/K+\sum_{j=1}^{J} b_{jk}, c(1-1/K)+J-\sum_{j=1}^{J} b_{jk}\right),\ p'_k=\frac{-\sum_j b_{jk}\ln(1-0.5)}{c-\sum_j b_{jk}\ln(1-0.5)}\\
&(l_{jk}|-) \sim \text{CRT}(n_{jk}, r_k b_{jk}),\ (l'_k|-) \sim \text{CRT}\left(\sum_{j=1}^{J} l_{jk}, \gamma_0\right)\\
&(\gamma_0|-) \sim \text{Gamma}\left(e_0+\sum_{k=1}^{K} l'_k, \frac{1}{f_0-\sum_{k=1}^{K}\ln(1-p'_k)}\right)\\
&(r_k|-) \sim \text{Gamma}\left(\gamma_0+\sum_{j=1}^{J} l_{jk}, \frac{1}{c-\sum_{j=1}^{J} b_{jk}\ln(1-0.5)}\right)\\
&(\lambda_{jk}|-) \sim \text{Gamma}(r_k b_{jk}+n_{jk}, 0.5)\\
&(\omega_k|-) \sim \text{Dir}\left(\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k, v_{ji}=1), \cdots, \eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k, v_{ji}=V)\right).
\end{aligned}
\tag{29}
$$

G Beta-NB

The beta-NB process model is constructed as

$$
\begin{aligned}
&x_{ji} \sim F(\omega_{z_{ji}}),\ \ \omega_k \sim \text{Dir}(\eta, \cdots, \eta)\\
&N_j=\sum_{k=1}^{K} n_{jk},\ \ n_{jk} \sim \text{Pois}(\lambda_{jk}),\ \ \lambda_{jk} \sim \text{Gamma}(r_j, p_k/(1-p_k))\\
&r_j \sim \text{Gamma}(e_0, 1/f_0),\ \ p_k \sim \text{Beta}(c/K, c(1-K))
\end{aligned}
\tag{30}
$$

The block Gibbs sampling can be expressed as

$$
\begin{aligned}
&\Pr(z_{ji}=k|-) \propto F(x_{ji};\omega_k)\lambda_{jk}\\
&(p_k|-) \sim \text{Beta}\left(c/K+\sum_{j=1}^{J} n_{jk}, c(1-1/K)+\sum_{j=1}^{J} r_j\right),\ \ l_{jk} \sim \text{CRT}(n_{jk}, r_j)\\
&(r_j|-) \sim \text{Gamma}\left(e_0+\sum_{k=1}^{K} l_{jk}, \frac{1}{f_0-\sum_{k=1}^{K}\ln(1-p_k)}\right)\\
&(\lambda_{jk}|-) \sim \text{Gamma}(r_j+n_{jk}, p_k)\\
&(\omega_k|-) \sim \text{Dir}\left(\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k, v_{ji}=1), \cdots, \eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k, v_{ji}=V)\right).
\end{aligned}
\tag{31}
$$

H Marked-Beta-NB

The Marked-Beta-NB process model is constructed as

$$
\begin{aligned}
&x_{ji} \sim F(\omega_{z_{ji}}),\ \ \omega_k \sim \text{Dir}(\eta, \cdots, \eta)\\
&N_j=\sum_{k=1}^{K} n_{jk},\ \ n_{jk} \sim \text{Pois}(\lambda_{jk}),\ \ \lambda_{jk} \sim \text{Gamma}(r_k, p_k/(1-p_k))\\
&r_k \sim \text{Gamma}(e_0, 1/f_0),\ \ p_k \sim \text{Beta}(c/K, c(1-K))
\end{aligned}
\tag{32}
$$

The block Gibbs sampling can be expressed as

$$\begin{aligned}
&\Pr(z_{ji}=k|-)\propto F(x_{ji};\omega_k)\lambda_{jk}\\
&p_k\sim \text{Beta}\left(c/K+\sum_{j=1}^{J}n_{jk}, c(1-1/K)+Jr_k\right),\ \ l_{jk}\sim \text{CRT}(n_{jk},r_k)\\
&(r_k|-)\sim \text{Gamma}\left(e_0+\sum_{j=1}^{J}l_{jk}, \frac{1}{f_0-J\ln(1-p_k)}\right)\\
&(\lambda_{jk}|-)\sim \text{Gamma}(r_k+n_{jk},p_k)\\
&(\omega_k|-)\sim \text{Dir}\left(\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=1),\cdots,\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=V)\right).
\end{aligned}\tag{33}$$

I Marked-Gamma-NB

The Marked-Gamma-NB process model is constructed as

$$\begin{aligned}
&x_{ji}\sim F(\omega_{z_{ji}}),\ \ \omega_k\sim \text{Dir}(\eta,\cdots,\eta)\\
&N_j=\sum_{k=1}^{K}n_{jk},\ \ n_{jk}\sim \text{Pois}(\lambda_{jk}),\ \lambda_{jk}\sim \text{Gamma}(r_k,p_k/(1-p_k))\\
&r_k\sim \text{Gamma}(\gamma_0/K,1/c),\ \ p_k\sim \text{Beta}(a_0,b_0),\ \ \gamma_0\sim \text{Gamma}(e_0,1/f_0).
\end{aligned}\tag{34}$$

The block Gibbs sampling can be expressed as

$$\begin{aligned}
&\Pr(z_{ji}=k|-)\propto F(x_{ji};\omega_k)\lambda_{jk}\\
&p_k\sim \text{Beta}\left(a_0+\sum_{j=1}^{J}n_{jk}, b_0+Jr_k\right),\ \ p'_k=\frac{-J\ln(1-p_k)}{c-J\ln(1-p_k)}\\
&l_{jk}\sim \text{CRT}(n_{jk},r_k),\ l'_k\sim \text{CRT}(\sum_{j=1}^{J}l_{jk},\gamma_0/K),\ \gamma_0\sim \text{Gamma}\left(e_0+\sum_{k=1}^{K}l'_k, \frac{1}{f_0-\sum_{k=1}^{K}\ln(1-p'_k)/K}\right)\\
&(r_k|-)\sim \text{Gamma}\left(\gamma_0/K+\sum_{j=1}^{J}l_{jk}, \frac{1}{c-J\ln(1-p_k)}\right),\ \ (\lambda_{jk}|-)\sim \text{Gamma}(r_k+n_{jk},p_k)\\
&(\omega_k|-)\sim \text{Dir}\left(\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=1),\cdots,\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=V)\right).
\end{aligned}\tag{35}$$