Large-scale Distributed Dependent Nonparametric Trees: Supplementary Material

1 Variational inference for Dependent Nonparametric Trees

Since the parameters are estimated in an online fashion, we focus on a specific time t and infer the parameters for $\Pi^{(t)}$. And we omit the label t in the following when there is no confusion.

The DNTs assumes the following generative model:

$$\begin{aligned}
&\nu_\epsilon \sim \text{Beta}(1, \alpha)\\
&\psi_\epsilon \sim \text{Beta}(1, \gamma)\\
&\phi_{\epsilon \cdot i} = \psi_{\epsilon \cdot i} \prod_{j=1}^{i-1} (1 - \psi_{\epsilon \cdot j}), \qquad \phi_\emptyset = 1\\
&\pi_\epsilon = \nu_\epsilon \phi_\epsilon \prod_{\epsilon' \prec \epsilon} (1 - \nu_{\epsilon'}) \phi_{\epsilon'}, \quad \pi_\emptyset = \nu_\emptyset\\
&\theta_{\epsilon \cdot i} \sim p(\theta_{\epsilon \cdot i} | \theta_\epsilon)\\
&z_n \sim \text{Multi}(\boldsymbol{\pi})\\
&x_n \sim p(x_n | \theta_{z_n}),
\end{aligned}$$

where $\emptyset$ represents the root node; $\epsilon' \prec \epsilon$ indicates that ϵ' is an ancestor of ϵ. Assume a family of factorized variational distributions:

$$q(\boldsymbol{\nu}, \boldsymbol{\psi}, \boldsymbol{\theta}, \boldsymbol{z}) = \prod_{\epsilon \in T} q(\nu_\epsilon | \delta_\epsilon) q(\psi_\epsilon | \sigma_\epsilon) q(\theta_\epsilon | \mu_\epsilon, \eta_\epsilon) \prod_{n=1}^{N} q(z_n | \lambda_n),$$

where T is a truncated tree (note that we propose a birth-merge strategy in the paper to dynamically vary the truncation level). Further assume the factors have the parametric forms:

$$\begin{gathered}
q(\nu_\epsilon | \delta_\epsilon) = \text{Beta}(\nu_\epsilon | \delta_\epsilon), \quad q(\psi_\epsilon | \sigma_\epsilon) = \text{Beta}(\psi_\epsilon | \sigma_\epsilon),\\
q(\theta_\epsilon | \mu_\epsilon, \eta_\epsilon) = \text{vMF}(\theta_\epsilon | \mu_\epsilon, \eta_\epsilon), \quad q(z_n | \lambda_n) = \text{Multi}(z_n | \lambda_n).
\end{gathered}$$

The variational lower bound is:

$$\begin{aligned}
\log\ p(\boldsymbol{x}) &\geq \mathbb{E}_q[\log\ p(\boldsymbol{\nu}, \boldsymbol{\psi}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x})] - \mathbb{E}_q[\log\ q(\boldsymbol{\nu}, \boldsymbol{\psi}, \boldsymbol{\theta}, \boldsymbol{z})]\\
&= \sum_{\epsilon \in T} \mathbb{E}\left[\log \frac{p(\nu_\epsilon) p(\psi_\epsilon) p(\theta_\epsilon)}{q(\nu_\epsilon) q(\psi_\epsilon) q(\theta_\epsilon)}\right] + \sum_{n=1}^{N} \mathbb{E}\left[\log \frac{p(x_n | \theta_{z_n}) p(z_n | \boldsymbol{\nu}, \boldsymbol{\psi})}{q(z_n)}\right]\\
&\triangleq \mathcal{L}(q).
\end{aligned}$$

We optimize $\mathcal{L}(q)$ using coordinate ascent.

1.1 Optimize the data assignment parameters $q(z_n)$

First isolate the terms that only contains $q(z_n)$:

$$\mathcal{L}(q(z_n)) = \mathbb{E}[\log\ p(z_n | \boldsymbol{\nu}, \boldsymbol{\psi}) p(x_n | \theta_{z_n})] - \mathbb{E}[\log\ q(z_n)].$$

The optimal solution for $q(z_n)$ is then:

$$q(z_n = z) \propto \exp\{\mathbb{E}[\log\ p(z_n = z|\boldsymbol{\nu}, \boldsymbol{\psi})] + \mathbb{E}[\log\ p(x_n|\theta_z)]\}.$$

Here

$$\begin{aligned}\mathbb{E}[\log\ p(z_n = z|\boldsymbol{\nu}, \boldsymbol{\psi})] &= \mathbb{E}[\log\ \nu_z \phi_z \prod_{z' \prec z} (1 - \nu_{z'})\phi_{z'}] \\ &= \mathbb{E}[\log\ \nu_z] + \sum_{z' \prec z} \mathbb{E}[\log\ (1 - \nu_{z'})] + \sum_{z' \preceq z} \mathbb{E}[\log\ \phi_{z'}],\end{aligned}$$

where

$$\begin{aligned}&\mathbb{E}[\log\ \nu_\epsilon] = \Psi(\delta_{\epsilon,1}) - \Psi(\delta_{\epsilon,1} + \delta_{\epsilon,2}) \\ &\mathbb{E}[\log\ (1 - \nu_\epsilon)] = \Psi(\delta_{\epsilon,2}) - \Psi(\delta_{\epsilon,1} + \delta_{\epsilon,2}) \\ &\mathbb{E}[\log\ \phi_{\epsilon \cdot i}] = \mathbb{E}[\log\ \psi_{\epsilon \cdot i}] + \sum_{j=1}^{i-1} \mathbb{E}[\log\ (1 - \psi_{\epsilon \cdot j})] \\ &\qquad = \Psi(\sigma_{\epsilon \cdot i,1}) - \Psi(\sigma_{\epsilon \cdot i,1} + \sigma_{\epsilon \cdot i,2}) + \sum_{j=1}^{i-1} \Psi(\sigma_{\epsilon \cdot j,2}) - \Psi(\sigma_{\epsilon \cdot j,1} + \sigma_{\epsilon \cdot j,2});\end{aligned}$$

and

$$\mathbb{E}[\log\ p(x_n|\theta_z)] = \beta \mathbb{E}[\theta_z]^\top x_n = \beta \mu_z^\top x_n.$$

Here $\Psi(\cdot)$ is the digamma function.

1.2 Optimize the stick breaking parameters $q(\nu_\epsilon|\delta_\epsilon)$ and $q(\psi_\epsilon|\sigma_\epsilon)$

$$\mathcal{L}(q(\nu_\epsilon)) = \mathbb{E}[\log\ p(\nu_\epsilon) - \log\ q(\nu_\epsilon)] + \sum_{n=1}^{N} \mathbb{E}[\log\ p(z_n|\boldsymbol{\nu}, \boldsymbol{\psi})].$$

We rewrite the second term with indicator random variables:

$$\begin{aligned}&\mathbb{E}[\log\ p(z_n|\boldsymbol{\nu}, \boldsymbol{\psi})] \\ &= \mathbb{E}\big[\log(\prod_{\epsilon \in T} \nu_\epsilon^{\mathbb{1}[z_n = \epsilon]} (1 - \nu_\epsilon)^{\mathbb{1}[\epsilon \prec z_n]} \phi_\epsilon^{\mathbb{1}[\epsilon \preceq z_n]})\big] \\ &= \sum_{\epsilon \in T} \Big\{ q(z_n = \epsilon)\mathbb{E}[\log\ \nu_\epsilon] + q(\epsilon \prec z_n)\mathbb{E}[\log\ (1 - \nu_\epsilon)] + q(\epsilon \preceq z_n)\mathbb{E}[\log\ \phi_\epsilon] \Big\},\end{aligned}$$

where

$$\mathbb{E}[\log\ \phi_{\epsilon \cdot i}] = \mathbb{E}[\log\ \psi_{\epsilon \cdot i}] + \sum_{j=1}^{i-1} \mathbb{E}[\log\ (1 - \psi_{\epsilon \cdot j})].$$

The updates for δ_ϵ are:

$$\delta_{\epsilon,1} = 1 + \sum_{n=1}^{N} q(z_n = \epsilon)$$

$$\delta_{\epsilon,2} = \alpha + \sum_{n=1}^{N} q(\epsilon \prec z_n).$$

Similarly, the updates for σ_ϵ are:

$$\sigma_{\epsilon \cdot i,1} = 1 + \sum_{n=1}^{N} q(\epsilon \cdot i \preceq z_n)$$

$$\sigma_{\epsilon \cdot i,2} = \gamma + \sum_{n=1}^{N} \sum_{j=i+1} q(\epsilon \cdot j \preceq z_n).$$

Incorporating the time dependency and using similar derivations, we have:

$$\delta_{\epsilon,1}^{(t)} = 1 + \sum_{t'=t-h}^{t} \sum_{n=1}^{N_{t'}} q^{(t')}(z_n = \epsilon)$$

$$\delta_{\epsilon,2}^{(t)} = \alpha + \sum_{t'=t-h}^{t} \sum_{n=1}^{N_{t'}} q^{(t')}(\epsilon \prec z_n)$$

$$\sigma_{\epsilon\cdot i,1}^{(t)} = 1 + \sum_{t'=t-h}^{t} \sum_{n=1}^{N_{t'}} q^{(t')}(\epsilon \cdot i \preceq z_n)$$

$$\sigma_{\epsilon\cdot i,2}^{(t)} = \gamma + \sum_{t'=t-h}^{t} \sum_{n=1}^{N_{t'}} \sum_{j=i+1} q^{(t')}(\epsilon \cdot j \preceq z_n).$$

1.3 Optimize the data emission parameters $q(\theta_\epsilon^{(t)}|\mu_\epsilon^{(t)}, \eta_\epsilon^{(t)})$

$$\mathcal{L}(q(\theta_\epsilon^{(t)})) = \mathbb{E}[\log\ p(\theta_\epsilon^{(t)}|\theta_{\epsilon'}^{(t)}) - \log\ q(\theta_\epsilon^{(t)})] + \sum_{\epsilon \prec \epsilon\cdot i} \mathbb{E}[\log\ p(\theta_{\epsilon\cdot i}^{(t)}|\theta_\epsilon^{(t)})] + \sum_{n=1}^{N} \mathbb{E}[\log\ p(x_n|\theta, z_n)].$$

Optimizing w.r.t $q(\theta_\epsilon^{(t)})$, we obtain:

$$\begin{aligned}
\log\ q^*(\theta_\epsilon^{(t)}) &= \mathbb{E}[\log\ p(\theta_\epsilon^{(t)}|\theta_{\epsilon'}^{(t)})] + \sum_{\epsilon \prec \epsilon\cdot i} \mathbb{E}[\log\ p(\theta_{\epsilon\cdot i}^{(t)}|\theta_\epsilon^{(t)})] + \sum_{n=1}^{N} \mathbb{E}[\log\ p(x_n|\theta^{(t)}, z_n)] + \text{const} \\
&= \mathbb{E}[\rho_\epsilon^{(t)} \tau_\epsilon^{(t)\top} \theta_\epsilon^{(t)}] \\
&\quad + \sum_{\epsilon \prec \epsilon\cdot i} \mathbb{E}[\log\ C_V(\rho_{\epsilon\cdot i}^{(t)}) + \rho_{\epsilon\cdot i}^{(t)} \tau_{\epsilon\cdot i}^{(t)\top} \theta_{\epsilon\cdot i}^{(t)}] \\
&\quad + \sum_{n} q(z_n = \epsilon) \mathbb{E}[\beta \theta_\epsilon^{(t)\top} x_n] + \text{const.}
\end{aligned}$$

Note the time- and hierarchical-dependency:

$$\begin{aligned}
\rho_{\epsilon\cdot i}^{(t)} &= \kappa_0 \|\kappa_1 \theta_\epsilon^{(t)} + \kappa_2 \theta_{\epsilon\cdot i}^{(t-1)}\| \\
\tau_{\epsilon\cdot i}^{(t)} &= \frac{\kappa_0 \kappa_1 \theta_\epsilon^{(t)} + \kappa_0 \kappa_2 \theta_{\epsilon\cdot i}^{(t-1)}}{\rho_{\epsilon\cdot i}^{(t)}}.
\end{aligned}$$

Hence we have:

$$\mathbb{E}[\log\ C_V(\rho_{\epsilon\cdot i}^{(t)})] = \mathbb{E}[(\frac{V}{2} - 1) \log\ \rho_{\epsilon\cdot i}^{(t)}] - \mathbb{E}[\log\ I_{\frac{V}{2}-1}(\rho_{\epsilon\cdot i}^{(t)})] - \frac{V}{2} \log\ (2\pi),$$

where the term involving the modified Bessel function of the first kind $I_v(\cdot)$ is intractable. We approximate it by Taylor approximation. According to [4] we have:

$$\log\ I_{\frac{V}{2}-1}(\rho_{\epsilon\cdot i}^{(t)}) \leq \log\ I_{\frac{V}{2}-1}(\bar{\rho}_{\epsilon\cdot i}^{(t)}) + \left(\frac{\partial}{\partial \theta_\epsilon} \log\ I_{\frac{V}{2}-1}(\bar{\rho}_{\epsilon\cdot i}^{(t)})\right)(\theta_\epsilon^{(t)} - \bar{\theta}_\epsilon^{(t)}),$$

where we denote $\bar{\rho}_{\epsilon\cdot i}^{(t)} = \kappa_0 \|\kappa_1 \bar{\theta}_\epsilon^{(t)} + \kappa_2 \theta_{\epsilon\cdot i}^{(t-1)}\|$; and $\bar{\theta}_\epsilon^{(t)}$ is some θ value. Since $I'_v(x) = I_{v+1}(x) + \frac{v}{x} I_v(x)$, we have:

$$\begin{aligned}
\frac{\partial}{\partial \theta_\epsilon^{(t)}} \log\ I_{\frac{V}{2}-1}(\bar{\rho}_{\epsilon\cdot i}^{(t)}) &= \frac{1}{I_{\frac{V}{2}-1}(\bar{\rho}_{\epsilon\cdot i}^{(t)})} \cdot \frac{\partial}{\partial \rho_{\epsilon\cdot i}^{(t)}} I_{\frac{V}{2}-1}(\bar{\rho}_{\epsilon\cdot i}^{(t)}) \cdot \frac{\partial}{\partial \theta_\epsilon^{(t)}} \bar{\rho}_{\epsilon\cdot i}^{(t)} \\
&= \frac{1}{I_{\frac{V}{2}-1}(\bar{\rho}_{\epsilon\cdot i}^{(t)})} \cdot \left(I_{\frac{V}{2}}(\bar{\rho}_{\epsilon\cdot i}^{(t)}) + \frac{\frac{V}{2} - 1}{\bar{\rho}_{\epsilon\cdot i}^{(t)}} I_{\frac{V}{2}-1}(\bar{\rho}_{\epsilon\cdot i}^{(t)}) \right) \cdot \frac{\kappa_0^2 \kappa_1 \kappa_2 \theta_{\epsilon\cdot i}^{(t-1)}}{\bar{\rho}_{\epsilon\cdot i}^{(t)}} \\
&= \left(\frac{I_{\frac{V}{2}}(\bar{\rho}_{\epsilon\cdot i}^{(t)})}{I_{\frac{V}{2}-1}(\bar{\rho}_{\epsilon\cdot i}^{(t)})} + \frac{\frac{V}{2} - 1}{\bar{\rho}_{\epsilon\cdot i}^{(t)}} \right) \cdot \frac{\kappa_0^2 \kappa_1 \kappa_2 \theta_{\epsilon\cdot i}^{(t-1)}}{\bar{\rho}_{\epsilon\cdot i}^{(t)}}.
\end{aligned}$$

Following [1], we set $\bar{\theta}_\epsilon^{(t)}$ as the mean of the variational distribution over $\theta_\epsilon^{(t)}$ from the previous iteration. To further simplify the expression, we note that (assume $v \in \mathbb{N}$):

$$
\begin{aligned}
&I_v(z) = (\frac{1}{2}z)^v \sum_{k=0}^{\infty} \frac{(\frac{1}{4}z^2)^k}{k!\Gamma(v+k+1)} \\
&0 < \frac{I_v(z)}{I_{v-1}(z)} < (\frac{1}{2}z)\frac{1}{v-1}.
\end{aligned}
$$

Hence when $v \gg z$ (e.g. v is the vocabulary size as in our case), we can safely approximate $\frac{I_v(z)}{I_{v-1}(z)} \approx 0$.

Next, we further approximate $\mathbb{E}[(\frac{V}{2}-1)\log\ \rho_{\epsilon\cdot i}^{(t)}]$ using Taylor expansion:

$$
\begin{aligned}
\log\ \rho_{\epsilon\cdot i}^{(t)} &= \log\ \kappa_0\sqrt{\kappa_1^2 + \kappa_2^2 + 2\kappa_1\kappa_2(\theta_\epsilon^{(t)\top}\theta_{\epsilon\cdot i}^{(t-1)})} \\
&= \frac{1}{2}\log\ (\kappa_1^2 + \kappa_2^2 + 2\kappa_1\kappa_2(\theta_\epsilon^{(t)\top}\theta_{\epsilon\cdot i}^{(t-1)})) + \text{const} \\
&\approx \frac{1}{2}\log\ (\kappa_1^2 + \kappa_2^2 + 2\kappa_1\kappa_2(\bar{\theta}_\epsilon^{(t)\top}\theta_{\epsilon\cdot i}^{(t-1)})) \\
&\quad + \theta_\epsilon^{(t)\top}\frac{2\kappa_1\kappa_2\theta_{\epsilon\cdot i}^{(t-1)}}{\kappa_1^2 + \kappa_2^2 + 2\kappa_1\kappa_2(\bar{\theta}_\epsilon^{(t)\top}\theta_{\epsilon\cdot i}^{(t-1)})} + \text{const} \\
&= \theta_\epsilon^{(t)\top}\frac{2\kappa_0^2\kappa_1\kappa_2\theta_{\epsilon\cdot i}^{(t-1)}}{\bar{\rho}_{\epsilon\cdot i}^{(t)}} + \text{const.}
\end{aligned}
$$

In summary we have:

$$
\begin{aligned}
\log\ q^*(\theta_\epsilon^{(t)}) \approx \theta_\epsilon^{(t)\top}\Bigg\{&\kappa_0\kappa_1\mathbb{E}[\theta_{\epsilon'}^{(t)}] + \kappa_0\kappa_2\mathbb{E}[\theta_\epsilon^{(t-1)}] \\
&+ \sum_{\epsilon\prec\epsilon\cdot i}\left\{\kappa_0\kappa_1\mathbb{E}[\theta_{\epsilon\cdot i}^{(t)}] + (\frac{V}{2}-1)\cdot\frac{2\kappa_0^2\kappa_1\kappa_2\mathbb{E}[\theta_{\epsilon\cdot i}^{(t-1)}]}{\bar{\rho}_{\epsilon\cdot i}^{(t)}} + \frac{\frac{V}{2}-1}{\bar{\rho}_{\epsilon\cdot i}^{(t)}}\cdot\frac{\kappa_0^2\kappa_1\kappa_2\mathbb{E}[\theta_{\epsilon\cdot i}^{(t-1)}]}{\bar{\rho}_{\epsilon\cdot i}^{(t)}}\right\} \\
&+ \beta\sum_n q(z_n = \epsilon)x_n\Bigg\}.
\end{aligned}
$$

So the updates for $\mu_\epsilon^{(t)}$ and $\eta_\epsilon^{(t)}$ are:

$$
\begin{aligned}
\mu_\epsilon^{(t)*} &\approx \frac{\kappa_0\left(\kappa_1\mu_{\epsilon'}^{(t)} + \kappa_2\mu_\epsilon^{(t-1)} + \sum_i \kappa_1\mu_{\epsilon i}^{(t)} + a\mu_{\epsilon i}^{(t-1)}\right) + \beta s_\epsilon}{\eta_\epsilon^{(t)*}} \\
\eta_\epsilon^{(t)*} &\approx \|\kappa_0\left(\kappa_1\mu_{\epsilon'}^{(t)} + \kappa_2\mu_\epsilon^{(t-1)} + \sum_i \kappa_1\mu_{\epsilon i}^{(t)} + a\mu_{\epsilon i}^{(t-1)}\right) + \beta s_\epsilon\|,
\end{aligned}
$$

where $a = \frac{3(.5V-1)\kappa_0\kappa_1\kappa_2}{\bar{\rho}_{\epsilon i}^{(t)}}$.

2 Merge move acceptance assessment

The variational lower bound of DNTs at time t is:

$$\begin{aligned}
\mathcal{L}(q) &= \sum_{n=1}^{N} \mathbb{E}\left[\log \frac{p(x_n|\theta_{z_n})p(z_n|\boldsymbol{\nu},\boldsymbol{\psi})}{q(z_n)}\right] + \sum_{\epsilon\in T} \mathbb{E}\left[\log \frac{p(\nu_\epsilon)p(\psi_\epsilon)p(\theta_\epsilon)}{q(\nu_\epsilon)q(\psi_\epsilon)q(\theta_\epsilon)}\right] \\
&= \beta \sum_n \sum_\epsilon q(z_n=\epsilon)\mu_\epsilon^\top x_n + N \log C_V(\beta) \\
&\quad + \sum_n \sum_\epsilon q(z_n=\epsilon)\mathbb{E}[\log\ p(\epsilon|\nu,\psi)] \\
&\quad - \sum_n \sum_\epsilon q(z_n=\epsilon)\log\ q(z_n=\epsilon) \\
&\quad + \mathbb{E}\left[\log \frac{p(\nu_\epsilon)p(\psi_\epsilon)}{q(\nu_\epsilon)q(\psi_\epsilon)}\right] \\
&\quad + \sum_{\epsilon\cdot i} \mathbb{E}[\log C_V(\rho_{\epsilon\cdot i}^{(t)})] + \mathbb{E}[(\rho_{\epsilon\cdot i}^{(t)}\tau_{\epsilon\cdot i}^{(t)} - \eta_{\epsilon\cdot i}\mu_{\epsilon\cdot i})^\top \theta_{\epsilon\cdot i}^{(t)}] - \log\ C_V(\eta_{\epsilon\cdot i}^{(t)})
\end{aligned} \tag{1}$$

Assume the candidate merge pair is (a, b). Denote the resulting node as m, the model state before merge as q, and the model state after merge as q^{merge}. Then the variational lower bound after merge is $\mathcal{L}(q^{merge})$. It is straightforward to see from Eq.1 that, all the terms of the difference $\mathcal{L}(q^{merge}) - \mathcal{L}(q)$ except the entropy term $\sum_n \sum_\epsilon q(z_n = m)\log\ q(z_n = m)$ can be directly computed given the summary statistics of a and b. We denote $\mathcal{L}'(q)$ as $\mathcal{L}(q)$ excluding the entropy term:

$$\mathcal{L}'(q) = \mathcal{L}(q) + \sum_n \sum_\epsilon q(z_n=\epsilon)\log\ q(z_n=\epsilon).$$

Next we show that $\mathcal{L}'(q^{merge}) - \mathcal{L}'(q) \geq 0$ implies $\mathcal{L}(q^{merge}) - \mathcal{L}(q) \geq 0$, so we just need to evaluate $\mathcal{L}'(q^{merge}) - \mathcal{L}'(q)$ and accept the merge move if it is positive. Note that this is a slightly more strict condition for merge, compared to directly evaluating on exact ELBO, but it is more efficient that can be computed in constant time. And in practice we found it works well.

Denote $\lambda_{n\epsilon} = q(z_n = \epsilon)$. We only need to show

$$\begin{aligned}
&\sum_n \lambda_{nm}\log\ \lambda_{nm} \\
&\leq \sum_n \lambda_{na}\log\ \lambda_{na} + \sum_n \lambda_{nb}\log\ \lambda_{nb},
\end{aligned}$$

where $\lambda_{nm} = \lambda_{na} + \lambda_{nb}$. By noting that the function $f(x) = x\log x$ is convex, we have:

$$\begin{aligned}
&\sum_n \lambda_{na}\log\ \lambda_{na} + \sum_n \lambda_{nb}\log\ \lambda_{nb} \\
&\geq 2\sum_n \frac{\lambda_{na}+\lambda_{nb}}{2}\log\left(\frac{\lambda_{na}+\lambda_{nb}}{2}\right) \\
&= \sum_n \lambda_{nm}(\log\lambda_{nm} - \log 2) \\
&= \sum_n \lambda_{nm}\log\lambda_{nm} - N_m\log 2 \\
&> \sum_n \lambda_{nm}\log\lambda_{nm}. \qquad \square
\end{aligned}$$

3 Additional experimental results

We compare the training time of different algorithms on the NIPS dataset. As shown in Table 1, our method is much more efficient than previous work.

Method	HDP-MCMC [2]	DNTs-MCMC [3]	DNTs-Ours
Time (s)	512	724	145

Table 1: Training time on the NIPS dataset using 8 threads. The two baselines only support parallelism over cores within a single machine.

References

[1] Amr Ahmed and Eric P Xing. On tight approximate inference of the logistic-normal topic admixture model. In *International Conference on Artificial Intelligence and Statistics*, pages 19–26, 2007.

[2] Jason Chang and John W Fisher III. Parallel sampling of hdps using sub-cluster splits. In *Advances in Neural Information Processing Systems*, pages 235–243, 2014.

[3] Kumar Dubey, Qirong Ho, Sinead A Williamson, and Eric P Xing. Dependent nonparametric trees for dynamic hierarchical clustering. In *Advances in Neural Information Processing Systems*, pages 1152–1160, 2014.

[4] Jalil Taghia, Zhanyu Ma, and Arne Leijon. Bayesian estimation of the von-mises fisher mixture model with variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(9):1701–1715, 2014.