Technical Derivations and Proofs for Coupling Nonparametric Mixtures via Latent Dirichlet Processes

Dahua Lin and John Fisher

Abstract

This article provides proofs of the two theorems presented in section 2 of the paper.

1 Preliminary Introduction

For the sake of being self-contained, we restate the proposed construction here:

$$H_s \sim \mathrm{DP}(\alpha_s, B), \tag{1}$$

$$(c_{t1}, \ldots, c_{tM_L}) \sim \mathrm{Dir}(\alpha_1 q_{t1}, \ldots, \alpha_{M_L} q_{tM_L}), \tag{2}$$

$$D_t = \sum_{s=1}^{M_L} c_{t,s} S_{q_{ts}}(H_s). \tag{3}$$

Here, S_q denotes the sub-sampling operation with probability q to retain each atom.

2 Proof of Theorem 1

Theorem 1. *The stochastic process D_t constructed as above is a Dirichlet process, and has*

$$D_t \sim \mathrm{DP}(\beta_t, B) \quad \textit{with} \quad \beta_t = \sum_{s=1}^{M_L} \alpha_s q_{ts}. \tag{4}$$

Proof. According to the sub-sampling theorem in [1], we have

$$S_{q_{ts}}(H_s) \sim \mathrm{DP}(q_{ts}\alpha_s, B). \tag{5}$$

Since $H_1, \ldots, H_{M_L}$ are independent, The sub-sampled DPs $S_{q_{st}}(H_s)$ are independent for different s. Then Eq.(4) immediately follows, according to the superposition theorem in [1]. □

3 Proof of Theorem 2

Before proving the theorem, we first derive some useful lemmas.

Lemma 1. *Let A, B, and (X, Y) be independent (X and Y can be correlated), then we have*

$$\mathrm{Cov}(AX, BY) = \mathrm{E}(A)\mathrm{E}(B)\mathrm{Cov}(X, Y). \tag{6}$$

Proof. The can be shown as below.

$$\begin{aligned}\mathrm{Cov}(AX, BY) &= \mathrm{E}(ABXY) - \mathrm{E}(AX)\mathrm{E}(BY) \\ &= \mathrm{E}(A)\mathrm{E}(B)\mathrm{E}(XY) - \mathrm{E}(A)\mathrm{E}(B)\mathrm{E}(X)\mathrm{E}(Y) \\ &= \mathrm{E}(A)\mathrm{E}(B)(\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)) \\ &= \mathrm{E}(A)\mathrm{E}(B)\mathrm{Cov}(X, Y). \end{aligned} \tag{7}$$

□

Lemma 2. *Let $H \sim \mathrm{DP}(\alpha, B)$ and $q_1, q_2 > 0$. Suppose $S_1 = S_{q_1}(H)$ and $S_2 = S_{q_2}(H)$ are independently generated from H, and U is a measurable subset of Ω, then*

$$\mathrm{Cov}(S_1(U), S_2(U)) = \frac{q_1 q_2}{\alpha q_1 q_2 + 1} V_U. \tag{8}$$

Here, $V_U = B(U)(1 - B(U))$.

Proof. To prove this lemma, we have to consider the construction in a different but equivalent way. Instead of considering S_1 and S_2 separately, we consider a joint process that generates both. Specifically, we express H as

$$H = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}.$$

For each i, we independently draw two Bernoulli variables $z_{i,1}$ and $z_{i,2}$, respectively with $P(z_{i,1} = 1) = q_1$ and $P(z_{i,2} = 1) = q_2$. Based on these variables, we divide atoms into four subsets s_{00}, s_{01}, s_{10} and s_{11}. Particularly, s_{uv} contains all the atoms with $z_1 = u$ and $z_2 = v$. (Here, u and v are binary values).

With the atoms in these subsets, we can respectively derive four DPs by renormalizing the atom weights, D_{00}, D_{01}, D_{10} and D_{11}. According to the sub-sampling theorem [1], we have

$$D_{00} \sim \mathrm{DP}(\alpha(1 - q_1)(1 - q_2), B), \tag{9}$$

$$D_{01} \sim \mathrm{DP}(\alpha(1 - q_1)q_2, B), \tag{10}$$

$$D_{10} \sim \mathrm{DP}(\alpha q_1(1 - q_2), B), \tag{11}$$

$$D_{11} \sim \mathrm{DP}(\alpha q_1 q_2, B). \tag{12}$$

Since there are no atoms shared between them, $D_{00}, D_{01}, D_{10}, D_{11}$ are independent.

As S_1 comprises all the atoms with $z_1 = 0$, it can be viewed as being generated by superimposing D_{11} and D_{10}. Likewise, S_2 can be generated by superimposing D_{11} and D_{01}. Hence, we can write S_1 and S_2 as follows

$$S_1 = c_1 D_{11} + (1 - c_1) D_{10}, \qquad c_1 \sim \mathrm{Beta}(\alpha q_1 q_2,\ \alpha q_1(1 - q_2) \tag{13}$$

$$S_2 = c_2 D_{11} + (1 - c_2) D_{01}, \qquad c_2 \sim \mathrm{Beta}(\alpha q_1 q_2,\ \alpha(1 - q_1) q_2). \tag{14}$$

Here, c_1 and c_2 are independent. Then the covariance between $S_1(U)$ and $S_2(U)$ can be written as

$$\mathrm{Cov}(S_1(U), S_2(U)) = \mathrm{Cov}(c_1 D_{11}(U) + (1 - c_1) D_{10}(U),\ c_2 D_{11}(U) + (1 - c_2) D_{01}(U)), \tag{15}$$

$$= \mathrm{Cov}(c_1 D_{11}(U), c_2 D_{11}(U)). \tag{16}$$

Here, we utilize the fact that $1 - c_1$, $1 - c_2$, $D_{10}(U)$ and $D_{01}(U)$ are all independent. By Lemma 1, we have

$$\mathrm{Cov}(S_1(U), S_2(U)) = \mathrm{E}(c_1)\mathrm{E}(c_2)\mathrm{Var}(D_{11}(U)) = \frac{q_1 q_2}{\alpha q_1 q_2 + 1} B(U)(1 - B(U)) = \frac{q_1 q_2}{\alpha q_1 q_2 + 1} V_U. \tag{17}$$

Eq.(8) has been established. □

Theorem 2. *Let t_1, t_2 be different indices, U be a measurable subset of Ω, then*

$$\mathrm{Cov}(D_{t_1}(U), D_{t_2}(U)) = \frac{1}{\beta_{t_1} \beta_{t_2}} \sum_{s=1}^{M_L} \frac{(\alpha_s q_{t_1 s} q_{t_2 s})^2}{\alpha_s q_{t_1 s} q_{t_2 s} + 1} V_U. \tag{18}$$

Here, $V_U \triangleq B(U)(1 - B(U))$.

Proof. Following Eq.(3), we have

$$\begin{aligned}\mathrm{Cov}(D_{t_1}(U), D_{t_2}(U)) &= \mathrm{Cov}\left(\sum_{s=1}^{M_L} c_{t_1 s} S_{q_{t_1 s}}(H_s)(U),\ \sum_{s'=1}^{M_L} c_{t_2 s'} S_{q_{t_2 s'}}(H_{s'})(U) \right) \\ &= \sum_{s=1}^{M_L} \mathrm{Cov}\left(c_{t_1 s} S_{q_{t_1 s}}(H_s)(U),\ c_{t_2 s} S_{q_{t_2 s}}(H_s)(U) \right) \end{aligned} \tag{19}$$

Here, we utilize the fact that $c_{t_1 s}$, $c_{t_2 s'}$ are independent when $t_1 \neq t_2$, and H_s and $H_{s'}$ are independent when $s \neq s'$, and thus $\mathrm{Cov}(c_{t_1 s} S_{q_{t_1 s}}(H_s), c_{t_2 s'} S_{q_{t_2 s'}}(H_{s'})) = 0$, whenever $s \neq s'$.

According to both Lemma 1 and Lemma 2, this can be further written as

$$\mathrm{Cov}(D_{t_1}(U), D_{t_2}(U)) = \sum_{s=1}^{M_L} \mathrm{E}(c_{t_1 s})\mathrm{E}(c_{t_2 s})\mathrm{Cov}(S_{q_{t_1 s}}(H_s)(U),\ S_{q_{st_2}}(H_s)(U)) \tag{20}$$

$$= \sum_{s=1}^{M_L} \frac{\alpha_s q_{t_1 s}}{\beta_{t_1}} \cdot \frac{\alpha_s q_{t_2 s}}{\beta_{t_2}} \cdot \frac{q_{t_1 s} q_{t_2 s}}{\alpha q_{t_1 s} q_{t_2 s} + 1} V_U, \tag{21}$$

$$= \frac{1}{\beta_{t_1}\beta_{t_2}} \sum_{s=1}^{M_L} \frac{(\alpha_s q_{t_1 s} q_{t_2 s})^2}{\alpha q_{t_1 s} q_{t_2 s} + 1} V_U. \tag{22}$$

The proof is completed. □

4 Derivation of Some Formulas for Sampling

This section provides detailed derivation of several key formulas used in the sampling algorithm.

4.1 The conditional likelihood of x_{ti}

In Eq.(10) of the main paper, we get the likelihood of x_{ti} conditioned on the choice of latent DP and other labels. The formula is given below.

$$p(x_{ti} | u_{ti} = s, \mathbf{r}_t, \mathbf{z}_{/ti}) = \frac{1}{w_{st/i} + q_{ts}\alpha_s} \left(\sum_{k \in I_s : r_{tk} = 1} m_{*k/ti} f(x_{ti}; \phi_k) + q_{ts}\alpha_s f(x_{ti}; B) \right). \tag{23}$$

Here, $w_{st/i} \triangleq \sum_{k \in I_s} m_{*k/ti}$, which equals the total number of observed samples associated with the atoms from H_s (except x_{ti} itself).

Proof. Note here that z_{ti} is marginalized out in the formula above, which can actually be expanded into

$$p(x_{ti} | u_{ti} = s, \mathbf{r}_t, \mathbf{z}_{/ti}) = p(x_{ti} | z_{ti} = 0) p(z_{ti} = 0 | \mathbf{r}_z, \mathbf{z}_{t/i}) + \sum_{z \in I_s} p(x_{ti} | z_{ti} = z) p(z_{ti} = z | \mathbf{r}_z, \mathbf{z}_{t/i}). \tag{24}$$

Here, we sum over all possible cases of choosing z_{ti}: (1) $z_{ti} = 0$, which indicates to create a new atom; and (2) $z_{ti} \in I_s$, which indicates to use an existing atom in H_s. When $p(x_{ti} | z_{ti} = 0)$, x_{ti} has to be generated from a new atom that has not been created. By marginalizing this unknown atom parameter, we get

$$p(x_{ti} | z_{ti} = 0) = \int_{\theta \in B} f(x; \theta) B(\theta) d\theta \triangleq f(x; B).$$

When $z_{ti} = z \in I_s$, we simply have $p(x_{ti} | z_{ti} = z) = f(x_{ti}; \theta_z)$.

For the other factor, namely the conditional probability of z_{ti}, we integrate out H_s (similar to the argument used in deriving the Chinese restaurant process) and derive

$$p(z_{ti} = z | \mathbf{r}_z, \mathbf{z}_{t/i}) = \begin{cases} q_{ts}\alpha_s / (w_{st/i} + q_{ts}\alpha_s) & (z = 0) \\ m_{*k/ti} / (w_{st/i} + q_{ts}\alpha_s) & (z \in I_s). \end{cases} \tag{25}$$

Note that latent DP H_s has been sub-sampled when it is used to generate the observations in the t-group. Hence, we have a reduced concentration $q_{ts}\alpha_s$, and only the atoms with $r_{tk} = 1$ is taken into consideration here. Incorporating these results into Eq.(24) completes the derivation. □

4.2 The ratio of probabilities for the inheritance indicators

In Eq.(12) of the main paper, we derive the ratio of the posterior probabilities of r_{tk}, the variable indicating whether the atom ϕ_k is inherited by D_t. Suppose ϕ_k comes from H_s, the formula is

$$\frac{\Pr(r_{tk}=1|\text{others})}{\Pr(r_{tk}=0|\text{others})} = \frac{q_{ts}\cdot p(\mathbf{z}_t|r_{tk}=1,\text{others})}{(1-q_{ts})\cdot p(\mathbf{z}_t|r_{tk}=0,\text{others})} = \frac{q_{ts}}{1-q_{ts}}\frac{\gamma(\tau_{s/t},n_t)}{\gamma(\tau_{s/t}+m_{*k},n_t)}. \quad (26)$$

Here, $\tau_{s/t} = q_{ts}\alpha_s + \sum_{k'\in I_s-\{k\}} m_{*k/t}$ and $m_{*k/t}$ is the number of samples associated with k in all other groups (excluding the ones in the t-th group). γ is a function defined by $\gamma(\tau,n) = \prod_{i=0}^{n-1}(\tau+i) = \Gamma(\tau+n)/\Gamma(\tau)$.

Proof. Note that in deriving this formula, we treat everything else as fixed (including the inheritance indicators for other atoms). Let rewrite the posterior probability $\Pr(r_{tk}=v|\text{others})$ using the Bayes rule, as

$$\Pr(r_{tk}=v|\text{others}) \propto p(\mathbf{z}_t|r_{tk}=v,\mathbf{r}_{t/k})p(r_{tk}=v|q_{ts}). \quad (27)$$

Here, $\mathbf{r}_{t/k}$ refers to all other inheritance indicators. The factor $p(r_{tk}=v|q_{ts})$ is just the prior inheritance probability defined by q_{ts} as

$$p(r_{tk}=v|q_{ts}) = \begin{cases} q_{ts} & (v=1), \\ 1-q_{ts} & (v=0) \end{cases}. \quad (28)$$

The other factor is a little bit more involved. When $r_{tk}=0$, which means that ϕ_k is not contained in D_t, we can derive the joint probability of $\mathbf{z}_t$ recursively following a Chinese restaurant process, which is given by

$$\begin{aligned} p(\mathbf{z}_t|r_{tk}=0,\mathbf{r}_{t/k}) &= \frac{\prod_{k'\in I_s-\{k\}}\prod_{i=0}^{o_{tk}}(m_{*k'/t}+(i-1))}{\prod_{i=1}^{n_t}\left(q_{ts}\alpha_s+\sum_{k'\in I_s-\{k\}} m_{*/k'}+(i-1)\right)} \\ &= \frac{\prod_{k'\in I_s-\{k\}}\prod_{i=0}^{o_{tk}}(m_{*k'/t}+(i-1))}{\prod_{i=0}^{n_t-1}(\tau_{s/t}+i)}. \end{aligned} \quad (29)$$

Here, $o_{tk} = \#\{z_{ti}=k\}$ is the number of samples in the t-th group that is associated with ϕ_k. Likewise, when $r_{tk}=1$, we have

$$p(\mathbf{z}_t|r_{tk}=1,\mathbf{r}_{t/k}) = \frac{\prod_{k'\in I_s-\{k\}}\prod_{i=0}^{o_{tk}}(m_{*k'/t}+(i-1))}{\prod_{i=0}^{n_t-1}(\tau_{s/t}+m_{*k/t}+i)} \quad (30)$$

Comparing Eq.(29) and (30), we can see that they have the same numerator, but have different denominator. Hence, we have

$$\frac{p(\mathbf{z}_t=1|r_{tk}=1)}{p(\mathbf{z}_t=0|r_{tk}=0)} = \frac{\prod_{i=0}^{n_t-1}(\tau_{s/t}+i)}{\prod_{i=0}^{n_t-1}(\tau_{s/t}+m_{*k/t}+i)} \triangleq \frac{\gamma(\tau_{s/t},n_t)}{\gamma(\tau_{s/t}+m_{*k/t},n_t)}. \quad (31)$$

Combining Eq.(28) and (31) with Eq.(27) results in Eq.(26), thus completing the derivation. □

4.3 Other formulas

The Equation (13) in the main paper just follows the standard conjugate update for Dirichlet distribution. The Equation (14) can also be easily derived, considering that the inheritance indicators are independently generated. (Also, in this framework, we assume all latent DPs use the same base distribution B).

References

[1] Dahua Lin, Eric Grimson, and John Fisher. Construction of dependent dirichlet processes based on poisson processes. In *Advances of NIPS'10*, 2010.