Supplement to Nested Dictionary Learning for Hierarchical Organization of Imagery and Text

I. NESTED K-MEANS

The nested K-means is only taken as the initialization step, as a convenient a simple way of initially assigning images to prospective nodes in the tree. At the beginning of inference, we need to initialize parameters of nodes by assigning images to them. For each node, we take a K-means to assign images living at this node to all its children nodes, where K is the number of its children nodes. For example, as mentioned in the paper, four nodes are present beneath the root node (at initialization). We take the K-means algorithm where K equals 4 and assign all the images to one of the four children nodes. Then for each node at the second level, we again take the K-means algorithm where K equals 2 to assign those images assigned at this parent node to its two children nodes. This process stops when the bottom level is reached. This is what we call nested K-means. Since this nested K-means algorithm is only taken for initialization, we set relatively smaller K for each node. After the initialization, our model automatically learns the right size of the tree, including the number of children nodes for each node.

II. SUPPLEMENTARY INFERENCE

Figure 1 shows the graphical representation of the proposed model, and Table I summarizes all notation of the variables.

A. Sampling $\boldsymbol{c}_m$

Given the level allocation samples $\{l_{mi}\}$, we sample the path assignment $\boldsymbol{c}_m$ for each image m. This consists of a sequence of parent-child transition sampling procedures: for a given parent node we wish to recursively sample a child node that it transitions to. We again adopt a retrospective sampling algorithm.

For a given node $\boldsymbol{\epsilon}$, we first sample its children nodes' stick weights $\{\lambda_{\boldsymbol{\epsilon}\epsilon_i} : \epsilon_i \leq N_{\boldsymbol{\epsilon}}\}$ from the conditional posterior distribution

$$\lambda_{\boldsymbol{\epsilon}\epsilon_i} = \nu_{\epsilon_i} \prod_{i'<i}(1-\nu_{\epsilon_{i'}}),$$
$$\nu_{\epsilon_i} \sim \text{Beta}\big(1 + \sum_{m=1}^{M} 1_{\{c_m^{|\boldsymbol{\epsilon}|+1} = \boldsymbol{\epsilon}\epsilon_i\}}, \gamma + \sum_{m=1}^{M} 1_{\{c_m^{|\boldsymbol{\epsilon}|+1} > \boldsymbol{\epsilon}\epsilon_i\}}\big)$$

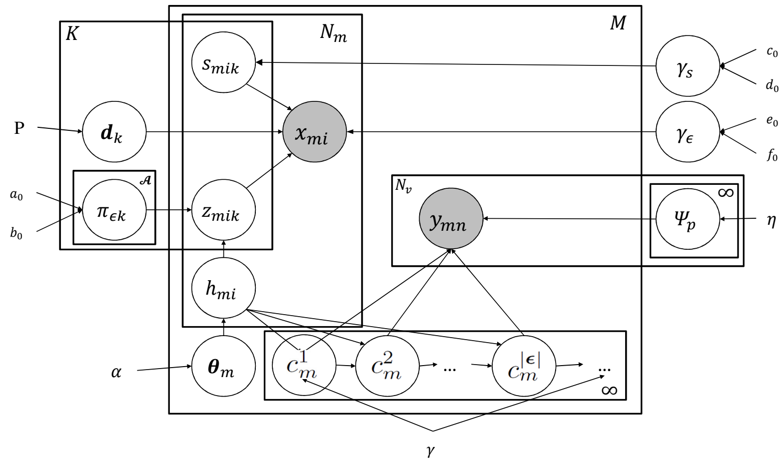

Fig. 1. The graphical representation of the model.

TABLE I

SYMBOL DESCRIPTION OF THE VARIABLES

Symbol	Description
$\boldsymbol{x}_{mi}$	the ith patch in image m
$\boldsymbol{y}_{mn}$	the nth annotation word associated with image m
$\boldsymbol{z}_{mi}$	the K-dimensional binary vector for the ith patch in image m
$\boldsymbol{s}_{mi}$	the K-dimensional vector of real weights for the ith patch in image m
$\boldsymbol{d}_k$	the P-dimensional vector of dictionary atom k
$\boldsymbol{\pi}_{\boldsymbol{\epsilon}}$	the K-dimensional vector probability of dictionary for node $\boldsymbol{\epsilon}$
h_{mi}	node index for patch i in image m
$\boldsymbol{\theta}_m$	topic distributions for image m
$\boldsymbol{\psi}_p$	topic distribution over vocabulary
$c_m^{\lvert\boldsymbol{\epsilon}\rvert}$	node index that image m chooses at level $\lvert\boldsymbol{\epsilon}\rvert$
$\mathcal{A}$	the set of usable nodes in the infinite tree
l_{mi}	level allocation index for patch i in image m
l'_{mi}	the proposed level allocation index for patch i in image m
$L_m = \max_i\{l_{mi}\}$	the depth of the path that image m assigned to
$L'_m = \max_i\left\{l'_{mi}\right\}$	the depth of the path that image m assigned to when replacing l_{mi} with the proposed l'_{mi}
$\mathcal{S}_{\boldsymbol{\epsilon}} = \{m : c_m^{\lvert\boldsymbol{\epsilon}\rvert} = \boldsymbol{\epsilon}\}$	the set of images whose assignments include node $\boldsymbol{\epsilon}$
$\mathcal{C}_{\boldsymbol{\epsilon}} = \{\boldsymbol{\epsilon}\epsilon_i, i = 1, 2, \dots\}$	set of $\boldsymbol{\epsilon}$'s children nodes
$N_{\boldsymbol{\epsilon}} = \lvert C_{\boldsymbol{\epsilon}}\rvert$	number of $\boldsymbol{\epsilon}$'s children nodes
$N'_{\boldsymbol{\epsilon}}$	number of $\boldsymbol{\epsilon}$'s children nodes after replacing $c_m^{\lvert\boldsymbol{\epsilon}\rvert+1}$ with $\boldsymbol{\epsilon}\epsilon^\star$
$\mathcal{L}_m^{\boldsymbol{\epsilon}}$	log-posterior for image m taking node $\boldsymbol{\epsilon}$
$\hat{\mathcal{L}}_{m\boldsymbol{\epsilon}}$	log-likelihood for image m taking node $\boldsymbol{\epsilon}$
$\mathcal{G}_l$	set of nodes at level l

Algorithm 1 Retrospective sampling for parent-child node transition

Input: $\mathcal{C}_{\boldsymbol{\epsilon}}$, $N_{\boldsymbol{\epsilon}}$, $\mathcal{S}_{\boldsymbol{\epsilon}}$,

Output: $c_m^{|\boldsymbol{\epsilon}|}, \mathcal{C}_{\boldsymbol{\epsilon}}$, $N_{\boldsymbol{\epsilon}}$

for $m \in \mathcal{S}_{\boldsymbol{\epsilon}}$ **do**

Sample $w_{\boldsymbol{\epsilon}\epsilon_i}$ from conditional posterior (1) for $\epsilon_i \le N_{\boldsymbol{\epsilon}}$, and from prior for $\epsilon_i > N_{\boldsymbol{\epsilon}}$

$\{\mathcal{L}_m^{\boldsymbol{\epsilon}\epsilon_i}\}_{\epsilon_i=1}^{N_{\boldsymbol{\epsilon}}} \leftarrow$ Algorithm 3

Sample $U_i \sim$ Uniform$[0,1]$

if $\sum_{s=1}^{\epsilon^\star-1} q_m(\boldsymbol{\epsilon}s) < U_i \le \sum_{s=1}^{\epsilon^\star} q_m(\boldsymbol{\epsilon}s)$ **then**

Set $c_m^{|\boldsymbol{\epsilon}|+1} = \boldsymbol{\epsilon}\epsilon^\star$ with probability $\rho_m(\boldsymbol{\epsilon}\epsilon^\star)$, otherwise, leave $c_m^{|\boldsymbol{\epsilon}|+1}$ unchanged

else

$N_{\boldsymbol{\epsilon}} \leftarrow N_{\boldsymbol{\epsilon}} + 1$, $\mathcal{C}_{\boldsymbol{\epsilon}} \leftarrow \mathcal{C}_{\boldsymbol{\epsilon}} \bigcup \boldsymbol{\epsilon}\epsilon^\star$, and set $c_m^{|\boldsymbol{\epsilon}|+1} = N_{\boldsymbol{\epsilon}}$ with probability $\rho_m(\boldsymbol{\epsilon}\epsilon^\star)$, otherwise, leave $c_m^{|\boldsymbol{\epsilon}|+1}$ unchanged

end if

end for

Then for each image m, from node $\boldsymbol{\epsilon}$ the child node $\boldsymbol{\epsilon}\epsilon^\star$ that it transits to is generated from the following proposal distribution

$$q(c_m^{|\boldsymbol{\epsilon}|+1} = \boldsymbol{\epsilon}\epsilon^\star) \propto \begin{cases} \lambda_{\boldsymbol{\epsilon}\epsilon^\star} \exp(\mathcal{L}_m^{\boldsymbol{\epsilon}\epsilon^\star}), & \epsilon^\star \le N_{\boldsymbol{\epsilon}} \\ \lambda_{\boldsymbol{\epsilon}\epsilon^\star} \mathcal{M}_m(N_{\boldsymbol{\epsilon}}), & \epsilon^\star > N_{\boldsymbol{\epsilon}} \end{cases}$$

where $N_{\boldsymbol{\epsilon}}$ is the number of current active children nodes of $\boldsymbol{\epsilon}$, $\mathcal{M}_m(N_{\boldsymbol{\epsilon}}) = \max_{1 \le \epsilon_i \le N_{\boldsymbol{\epsilon}}} \{\mathcal{L}_m^{\boldsymbol{\epsilon}\epsilon_i}\}$, and log-likelihood $\mathcal{L}_m^{\boldsymbol{\epsilon}\epsilon_i}$ is calculated as described in Algorithm 3. The acceptance probability of the proposed child node $\boldsymbol{\epsilon}\epsilon^\star$, which we denote as $\rho_m(\boldsymbol{\epsilon}\epsilon^\star)$, is

$$\begin{cases} 1, & \epsilon^\star \le N_{\boldsymbol{\epsilon}} \ \&\ N'_{\boldsymbol{\epsilon}} = N_{\boldsymbol{\epsilon}} \\ \min\{1, \frac{\tilde{c}_m(N_{\boldsymbol{\epsilon}})\mathcal{M}_m(N'_{\boldsymbol{\epsilon}})}{\tilde{c}_m(N'_{\boldsymbol{\epsilon}})\mathcal{L}_m^{\boldsymbol{\epsilon}\epsilon_i}}\}, & \epsilon^\star \le N_{\boldsymbol{\epsilon}} \ \&\ N'_{\boldsymbol{\epsilon}} < N_{\boldsymbol{\epsilon}} \\ \min\{1, \frac{\tilde{c}_m(N_{\boldsymbol{\epsilon}})\mathcal{L}_m^{\boldsymbol{\epsilon}\epsilon_i'}}{\tilde{c}_m(N'_{\boldsymbol{\epsilon}})\mathcal{M}_m(N_{\boldsymbol{\epsilon}})}\}, & \epsilon_i^\star > N_{\boldsymbol{\epsilon}} \end{cases}$$

where the normalizing constant $\tilde{c}_m(N_{\boldsymbol{\epsilon}}) = \sum_{i=1}^{N_{\boldsymbol{\epsilon}}} w_{\boldsymbol{\epsilon}\epsilon_i}\mathcal{L}_m^{\boldsymbol{\epsilon}\epsilon_i} + \mathcal{M}_m(N_{\boldsymbol{\epsilon}})(1 - \sum_{i=1}^{N_{\boldsymbol{\epsilon}}} w_{\boldsymbol{\epsilon}\epsilon_i})$. Algorithm 2 summarizes the retrospective sampling scheme described above.

Note that after $\boldsymbol{c}_m$ is obtained for each image m, the width of the tree structure may expand as a result of the new inferred N_ϵ. Additionally, it may also contain nodes that no image is assigned to, in this case we prune the tree by deleting those empty nodes.

In Algorithm 3 we calculate $\mathcal{L}_m^{\boldsymbol{\epsilon}\epsilon_i}$ recursively from nodes at the bottom level to the top, in this way we sweep every node only once.

B. Other parameters

Beyond sampling $\{l_{mi}, \boldsymbol{c}_m\}$, as discussed above, we provide update equations for other parameters in this model:

Algorithm 2 Calculating $\{\mathcal{L}_m^{\boldsymbol{\epsilon}}\}_{\boldsymbol{\epsilon}\in\mathcal{A}}$

Input: L_m, $\{\mathcal{G}_l\}_{l=1}^{L_m}$, $\{\boldsymbol{\pi}_{\boldsymbol{\epsilon}}, w_{\boldsymbol{\epsilon}}, \mathcal{C}_{\boldsymbol{\epsilon}}\}_{\boldsymbol{\epsilon}\in\mathcal{A}}$

Output: $\{\mathcal{L}_m^{\boldsymbol{\epsilon}}\}_{\boldsymbol{\epsilon}\in\mathcal{A}}$

$l = L_m$

while $l > 1$ **do**

for $\boldsymbol{\epsilon} \in \mathcal{G}_l$ **do**

$$\hat{\mathcal{L}}_{m\boldsymbol{\epsilon}} = \sum_{i=1}^{N_m}\sum_{k=1}^{K}\left[1_{\{h_{mi}=\boldsymbol{\epsilon}\}} z_{mik}\log(\pi_{\boldsymbol{\epsilon}k}) + 1_{\{h_{mi}=\boldsymbol{\epsilon}\}}(1-z_{mik})\log(1-\pi_{\boldsymbol{\epsilon}k})\right]$$

$$\mathcal{L}_m^{\boldsymbol{\epsilon}} = \log w_{\boldsymbol{\epsilon}} + \hat{\mathcal{L}}_{m\boldsymbol{\epsilon}} + \sum_{\boldsymbol{\epsilon}\epsilon_i\in\mathcal{C}_{\boldsymbol{\epsilon}}} \mathcal{L}_m^{\boldsymbol{\epsilon}\epsilon_i}$$

end for

$l = l - 1$

end while

- Sample $\boldsymbol{d}_k$

  The posterior of the kth dictionary atom $\boldsymbol{d}_k$ can be shown to be normal with covariance $\boldsymbol{\Sigma}_{\boldsymbol{d}_k}$ and mean $\boldsymbol{\mu}_{\boldsymbol{d}_k}$

$$\boldsymbol{\Sigma}_{\boldsymbol{d}_k} = (P + \gamma_\epsilon \sum_{m=1}^{M}\sum_{i=1}^{N_m} z_{mik}^2 s_{mik}^2)^{-1} \tag{1}$$

$$\boldsymbol{\mu}_{\boldsymbol{d}_k} = \gamma_\epsilon \boldsymbol{\Sigma}_{\boldsymbol{d}_k} \sum_{m=1}^{M}\sum_{i=1}^{N_m} \widetilde{\boldsymbol{x}}_{mi}^{-k} s_{mik} s_{mik} \tag{2}$$

  where $\widetilde{\boldsymbol{x}}_{mi}^{-k} = \boldsymbol{x}_{mi} - \mathbf{D}(\boldsymbol{z}_{mi}\odot\boldsymbol{s}_{mi}) + \boldsymbol{d}_k(z_{mik}\odot s_{mik})$.

- Sample z_{mik}

  For the ith patch in the mth image, sample the binary sparse code $\boldsymbol{z}_{mi} = (z_{mi1}, ..., z_{miK})$ with

$$P(z_{mik}=1) = \pi_{h_{mi}k}\exp[-\frac{\gamma_\epsilon}{2}(\boldsymbol{d}_k^T\boldsymbol{d}_k s_{mik}^2 - 2\boldsymbol{d}_k^T s_{mik}\widetilde{\boldsymbol{x}}_{mi}^{-k})] \tag{3}$$

$$P(z_{mik}=0) = 1 - \pi_{h_{mi}k} \tag{4}$$

  Thus whether the kth dictionary atom will be chosen for the ith patch of the mth image is drawn $z_{mik} \sim$ Bernoulli$(\frac{P(z_{mik}=1)}{P(z_{mik}=1)+P(z_{mik}=0)})$.

- Sample s_{mik}

  The posterior of positive weight s_{mik} can be obtained as a truncated normal distribution, with covariance $\boldsymbol{\Sigma}_{s_{mik}}$ and mean $\boldsymbol{\mu}_{s_{mik}}$, where

$$\boldsymbol{\Sigma}_{s_{mik}} = 1/(\gamma_s + \gamma_\epsilon \boldsymbol{d}_k^T\boldsymbol{d}_k z_{mik}^2) \tag{5}$$

$$\boldsymbol{\mu}_{s_{mik}} = \gamma_\epsilon \boldsymbol{\Sigma}_{s_{mik}} \boldsymbol{d}_k^T z_{mik}\widetilde{\boldsymbol{x}}_{mi}^{-k} \tag{6}$$

- Sample $\boldsymbol{\pi}_{\boldsymbol{\epsilon}}$

$$\pi_{\boldsymbol{\epsilon}k} \sim \text{Beta}\big(\frac{a_0}{K} + \sum_{m=1}^{M}\sum_{i=1}^{N_m} 1\{h_{mi} = \boldsymbol{\epsilon}\} z_{mik}, \frac{b_0(K-1)}{K} + \sum_{m=1}^{M}\sum_{i=1}^{N_m} 1\{h_{mi} = \boldsymbol{\epsilon}\}(1 - z_{mik})\ \big) \tag{7}$$

- Sample $\boldsymbol{\theta}_m$

  When sampling $\boldsymbol{\theta}_m$, assignment $\boldsymbol{c}_m$ is fixed. For $\boldsymbol{\epsilon} \in \boldsymbol{c}_m$, we first sample

$$\mu_{m|\boldsymbol{\epsilon}|} \sim \text{Beta}(1 + \sum_{i=1}^{N_m} 1\{l_{mi} = |\boldsymbol{\epsilon}|\}, \alpha + \sum_{i=1}^{N_m} 1\{l_{mi} > |\boldsymbol{\epsilon}|) \tag{8}$$

  and then construct $\theta_{m|\boldsymbol{\epsilon}|} = \mu_{m|\boldsymbol{\epsilon}|} \prod_{|\boldsymbol{\epsilon}'|<|\boldsymbol{\epsilon}|} (1 - \mu_{m|\boldsymbol{\epsilon}'|})$.

- Sample $\boldsymbol{\psi}_p$

  Posterior of $\boldsymbol{\psi}_p$ is still Dirichlet distributed as $\boldsymbol{\psi}_p \sim \text{Dir}(\zeta_{p1}, \ldots, \zeta_{pN_v})$ where

$$\zeta_{pv} = \sum_{m:\boldsymbol{c}_m = p} \sum_{n=1}^{W_m} 1\{y_{mn} = v\} \tag{9}$$