# Deriving the Block Coordinate Descent Rules for Tree-Structured NMF with sparsity regularization

## Objective

Given $t \in \{1,\ldots,T\}$ tasks, each with input matrix $X^{(t)} \in \mathbb{R}^{n_t \times m}$, related to each other in a task hierarchy/tree with a set of nodes $c \in \{r\} \cup \mathcal{B} \cup \mathcal{T}$ where $r$ is the root node, $\mathcal{B}$ a set of internal (or branch) nodes $b \in \mathcal{B}$, and $\mathcal{T}$ a set of the task-specific leaf nodes, the objective is:

$$O = \sum_{t=1}^{T} \left[ \left\| X^{(t)} - U^{(t)} V^{(t)\top} \right\|_{\mathrm{F}}^{2} + \lambda \sum_{i=1}^{m} \left\| V^{(t)}[i,:] \right\|_{1} \right] + \alpha \sum_{c} \left\| V^{(c)} - V^{Pa(c)} \right\|_{\mathrm{F}}^{2} \tag{1}$$

where $U^{(t)} \in \mathbb{R}_{\geq 0}^{n_t \times k}$, $V^{(\cdot)} \in \mathbb{R}_{\geq 0}^{m \times k}$, $k \ll n, m$. $V^{(t)}[i,:]$ is the $i$ row of task-specific factor matrix $V^{(t)}$.

The regularization term involving $\lambda$ tries to enforce sparsity in each row of task-specific $V^{(t)}$, ultimately so that only one latent dimension "lights up" for each row of $V^{(t)}$. Higher $\lambda$ will enforce stricter sparsity.

The regularization term involving $\alpha$ will:

a. constrain a task-specific latent feature factor $V^{(t)}$ in a leaf node of the task hierarchy to be similar to $V^{\mathrm{Pa}(t)}$ in its parent node;

b. constrain an internal node's latent feature factor $V^{(b)}$ to be similar to its direct child nodes' $V^{(c)}$ and and its parent node's $V^{Pa(b)}$; and

c. constrain the root node's latent feature factor $V^{(r)}$ to be similar to all of its direct child nodes' $V^{(c)}$s.

## Breaking down to task-level and column-level subproblems

The objective can be written as:

$$O = \sum_{t=1}^{T} \left[ \left\| X^{(t)} - \sum_{k} u_k^{(t)} v_k^{(t)\top} \right\|_{\mathrm{F}}^{2} + \lambda \sum_{i=1}^{m} \sum_{k} \left| V^{(t)}[i,k] \right| \right] + \alpha \sum_{c} \sum_{k} \left\| v_k^{(c)} - v_k^{Pa(c)} \right\|_{2}^{2} \tag{2}$$

$$= \sum_{t=1}^{T} \left[ \left\| X^{(t)} - \sum_{k} u_k^{(t)} v_k^{(t)\top} \right\|_{\mathrm{F}}^{2} + \lambda \sum_{k} \sum_{i=1}^{m} \left| V^{(t)}[i,k] \right| \right] + \alpha \sum_{c} \sum_{k} \left\| v_k^{(c)} - v_k^{Pa(c)} \right\|_{2}^{2} \tag{3}$$

$$= \sum_{t=1}^{T} \left[ \left\| X^{(t)} - \sum_{k} u_k^{(t)} v_k^{(t)\top} \right\|_{\mathrm{F}}^{2} + \lambda \sum_{k} \left\| v_k^{(t)} \right\|_{1} \right] + \alpha \sum_{c} \sum_{k} \left\| v_k^{(c)} - v_k^{Pa(c)} \right\|_{2}^{2} \tag{4}$$

Where $u_k^{(t)} \in \mathbb{R}^{n_t}$ is the $k$th column vector of $U^{(t)}$ and $v_k^{(t)} \in \mathbb{R}^{m}$ is the $k$th column vector of $V^{(t)}$. Now we 'pull out' terms involving the $k$th column in all factors:

$$O = \sum_{t=1}^{T} \left[ \left\| X^{(t)} - u_k^{(t)} v_k^{(t)\top} - \sum_{j \neq k} u_j^{(t)} v_j^{(t)\top} \right\|_{\mathrm{F}}^{2} + \lambda \left\| v_k^{(t)} \right\|_{1} + \lambda \sum_{j \neq k} \left\| v_j^{(t)} \right\|_{1} \right] \tag{5}$$

$$+ \alpha \sum_{c} \left( \left\| v_k^{(c)} - u_k^{Pa(c)} \right\|_{2}^{2} + \sum_{j \neq k} \left\| v_j^{(c)} - v_j^{Pa(c)} \right\|_{2}^{2} \right) \tag{6}$$

Now we'll substitute with $R_k^{(t)} = X^{(t)} - \sum_{j \neq k} u_j^{(t)} v_j^{(t)\top}$:

$$O = \sum_{t=1}^{T} \left[ \left\| R_k^{(t)} - u_k^{(t)} v_k^{(t)\top} \right\|_{\mathrm{F}}^{2} + \lambda \left\| v_k^{(t)} \right\|_{1} + \lambda \sum_{j \neq k} \left\| v_j^{(t)} \right\|_{1} \right] \tag{7}$$

$$+ \alpha \sum_{c} \left\| v_k^{(c)} - u_k^{Pa(c)} \right\|_{2}^{2} + \alpha \sum_{c} \sum_{j \neq k} \left\| v_j^{(c)} - v_j^{Pa(c)} \right\|_{2}^{2} \tag{8}$$

We can now attempt to optimize $u_k^{(t)}$ and $v_k^{(\cdot)}$, fixing all other parameters to be constant.

## Optimize $v_k^{(t)}$

To find $v_k^{(t)}$ for each leaf node task $t$ that minimizes the objective, we find the derivative of the objective with respect to $v_k^{(t)}$ and set it to 0, then solve. First we expand the objective into matrix multiplications:

$$
\begin{aligned}
O &= \left\| R_k^{(t)} - u_k^{(t)} v_k^{(t)\top} \right\|_{\mathrm{F}}^2 + \lambda \left\| v_k^{(t)} \right\|_1 + \alpha \left\| v_k^{(t)} - v_k^{Pa(t)} \right\|_2^2 + C && (9)\\
&= \mathrm{Tr}\left[ \left( R_k^{(t)} - u_k^{(t)} v_k^{(t)\top} \right)^\top \left( R_k^{(t)} - u_k^{(t)} v_k^{(t)\top} \right) \right] + \lambda \left\| v_k^{(t)} \right\|_1 && (10)\\
&\quad + \alpha \left( v_k^{(t)} - v_k^{Pa(t)} \right)^\top \left( v_k^{(t)} - v_k^{Pa(t)} \right) + C && (11)
\end{aligned}
$$

Here $C$ subsumes all elements of the objective that does not involve $v_k^{(t)}$ (including terms involving tasks other than $t$), since they will be zeroed out when the derivative is taken with respect to $v_k^{(t)}$. Now we keep expanding:

$$
\begin{aligned}
O &= \mathrm{Tr}\left[ R_k^{(t)^\top} R_k^{(t)} - 2 R_k^{(t)\top} u_k^{(t)} v_k^{(t)\top} + \left( u_k^{(t)} v_k^{(t)\top} \right)^\top \left( u_k^{(t)} v_k^{(t)\top} \right) \right] && (12)\\
&\quad + \lambda \left\| v_k^{(t)} \right\|_1 + \alpha \left( v_k^{(t)\top} v_k^{(t)} - 2 v_k^{(t)\top} v_k^{Pa(t)} + v_k^{Pa(t)\top} v_k^{Pa(t)} \right) + C && (13)\\
&= \mathrm{Tr}\left( R_k^{(t)^\top} R_k^{(t)} \right) - 2\,\mathrm{Tr}\left( R_k^{(t)\top} u_k^{(t)} v_k^{(t)\top} \right) + \mathrm{Tr}\left( v_k^{(t)} u_k^{(t)\top} u_k^{(t)} v_k^{(t)\top} \right) && (14)\\
&\quad + \lambda \left\| v_k^{(t)} \right\|_1 + \alpha v_k^{(t)\top} v_k^{(t)} - 2\alpha v_k^{(t)\top} v_k^{Pa(t)} + \alpha v_k^{Pa(t)\top} v_k^{Pa(t)} + C && (15)\\
&= \mathrm{Tr}\left( R_k^{(t)^\top} R_k^{(t)} \right) - 2 \left( R_k^{(t)\top} u_k^{(t)} \right)^\top v_k^{(t)} + \left( u_k^{(t)\top} u_k^{(t)} \right) \left( v_k^{(t)\top} v_k^{(t)} \right) && (16)\\
&\quad + \lambda \mathbf{1}_m^\top v_k^{(t)} + \alpha v_k^{(t)\top} v_k^{(t)} - 2\alpha v_k^{(t)\top} v_k^{Pa(t)} + \alpha v_k^{Pa(t)\top} v_k^{Pa(t)} + C && (17)
\end{aligned}
$$

where $\mathbf{1}_m$ is a vector of size $m$, filled with 1's. What allows us to expand $\left\| v_k^{(t)} \right\|_1$ from (16) to $\mathbf{1}_m^\top v_k^{(t)}$ (17) is the fact that we're enforcing $v_k^{(t)}$ to be non-negative at initialization and at each iteration.

Now the fun part:

$$
\begin{aligned}
\frac{\partial O}{\partial v_k^{(t)}} &= 0 - 2 R_k^{(t)\top} u_k^{(t)} + 2 v_k^{(t)} u_k^{(t)\top} u_k^{(t)} + \lambda \mathbf{1}_m + 2\alpha v_k^{(t)} - 2\alpha v_k^{Pa(t)} + 0 + 0 && (18)\\
0 &= -R_k^{(t)\top} u_k^{(t)} + \left( u_k^{(t)\top} u_k^{(t)} + \alpha \right) v_k^{(t)} + \frac{\lambda}{2} \mathbf{1}_m - \alpha v_k^{Pa(t)} && (19)\\
v_k^{(t)} &= \frac{R_k^{(t)\top} u_k^{(t)} + \alpha v_k^{Pa(t)} - \frac{\lambda}{2} \mathbf{1}_m}{\left\| u_k^{(t)} \right\|_2^2 + \alpha} && (20)
\end{aligned}
$$

With the non-negativity constraint $v_k^{(t)} \geq 0$, we want $R_k^{(t)\top} u_k^{(t)} + \alpha v_k^{Pa(t)} - \frac{\lambda}{2} \mathbf{1}_m \geq 0$, because if $R_k^{(t)\top} u_k^{(t)} + \alpha v_k^{Pa(t)} - \frac{\lambda}{2} \mathbf{1}_m < 0$, $O$ will increase in (16) and (17). So the finalized update rule is:

$$
v_k^{(t)} = \frac{\left[ R_k^{(t)\top} u_k^{(t)} + \alpha v_k^{Pa(t)} - \frac{\lambda}{2} \mathbf{1}_m \right]_+}{\left\| u_k^{(t)} \right\|_2^2 + \alpha} \tag{21}
$$

## Optimize $u_k^{(t)}$

We can derive the update rule for $u_k^{(t)}$ in leaf node task $t$ similarly but much more simply. From (17), we take the derivative of $O_t$ with respect to $u_k^{(t)}$; all regularization terms will zero out since they do not involve $u_k^{(t)}$. Hence the final update rule for $u_k^{(t)}$ is:

$$
u_k^{(t)} = \frac{\left[ R_k^{(t)} v_k^{(t)} \right]_+}{\left\| v_k^{(t)} \right\|_2^2} \tag{22}
$$

## Optimize $v_k^{(r)}$

For the overall consensus factor in the root of the task hierarchy, $v_k^{(r)}$, we can again ignore terms that do not involve $v_k^{(r)}$ in the objective (4). Note that we're going to collect the terms involving nodes $c$ whose parent is the root node, i.e. $\mathrm{Pa}(c) = r$:

$$O = \alpha \sum_{c \in \mathrm{Child}(r)} \left\| v_k^{(c)} - v_k^{(r)} \right\|_2^2 + C \tag{23}$$

$$= \alpha \sum_{c \in \mathrm{Child}(r)} \left( v_k^{(c)} - v_k^{(r)} \right)^\top \left( v_k^{(c)} - v_k^{(r)} \right) + C \tag{24}$$

$$= \alpha \sum_{c \in \mathrm{Child}(r)} \left[ v_k^{(c)\top} v_k^{(c)} - 2 v_k^{(c)\top} v_k^{(r)} + v_k^{(r)\top} v_k^{(r)} \right] + C \tag{25}$$

$$= C - \sum_{c \in \mathrm{Child}(r)} 2\alpha v_k^{(c)\top} v_k^{(r)} + \sum_{c \in \mathrm{Child}(r)} \alpha v_k^{(r)\top} v_k^{(r)} \tag{26}$$

Now we take the derivative, set to 0, and solve:

$$\frac{\partial O}{\partial v_k^{(r)}} = 0 - \sum_{c \in \mathrm{Child}(r)} 2\alpha v_k^{(c)} + \sum_{c \in \mathrm{Child}(r)} 2\alpha v_k^{(r)} \tag{27}$$

$$0 = - \sum_{c \in \mathrm{Child}(r)} v_k^{(c)} + |\mathrm{Child}(r)| \cdot v_k^{(r)} \tag{28}$$

$$v_k^{(r)} = \frac{\sum_{c \in \mathrm{Child}(r)} v_k^{(c)}}{|\mathrm{Child}(r)|} \tag{29}$$

where $|\mathrm{Child}(r)|$ is the number of direct child nodes of the root node $r$.

## Optimize $v_k^{(b)}$

For the latent feature factor in an internal/branch node of the task hierarchy, $v_k^{(b)}$, same drill as before: we ignore terms that do not involve $v_k^{(b)}$ for the particular node $b$ of interest in the objective (4). This time we collect terms involving the parent node of $b$, i.e. $\mathrm{Pa}(b)$, and nodes $c$ whose parent is $b$, i.e. $\mathrm{Pa}(c) = b$:

$$O = \alpha \left( \left\| v_k^{(b)} - v_k^{Pa(b)} \right\|_2^2 + \sum_{c \in \mathrm{Child}(b)} \left\| v_k^{(c)} - v_k^{(b)} \right\|_2^2 \right) + C \tag{30}$$

$$= \alpha \left( v_k^{(b)} - v_k^{Pa(b)} \right)^\top \left( v_k^{(B)} - v_k^{Pa(b)} \right) + \alpha \sum_{c \in \mathrm{Child}(b)} \left( v_k^{(c)} - v_k^{(b)} \right)^\top \left( v_k^{(c)} - v_k^{(b)} \right) + C \tag{31}$$

$$\begin{aligned} &= \alpha \left[ v_k^{(b)\top} v_k^{(b)} - 2 v_k^{(b)\top} v_k^{\mathrm{Pa}(b)} + v_k^{\mathrm{Pa}(b)\top} v_k^{\mathrm{Pa}(b)} \right] \\ &\qquad + \alpha \sum_{c \in \mathrm{Child}(b)} \left[ v_k^{(c)\top} v_k^{(c)} - 2 v_k^{(c)\top} v_k^{(b)} + v_k^{(b)\top} v_k^{(b)} \right] + C \end{aligned} \tag{32}$$

$$= \alpha v_k^{(b)\top} v_k^{(b)} - 2\alpha v_k^{(b)\top} v_k^{\mathrm{Pa}(b)} - \sum_{c \in \mathrm{Child}(b)} 2\alpha v_k^{(c)\top} v_k^{(b)} + \sum_{c \in \mathrm{Child}(b)} \alpha v_k^{(b)\top} v_k^{(b)} + C \tag{33}$$

Now we take the derivative, set to 0, and solve:

$$\frac{\partial O}{\partial v_k^{(b)}} = 2\alpha v_k^{(b)} - 2\alpha v_k^{\mathrm{Pa}(b)} - \sum_{c\in \mathrm{Child}(b)} 2\alpha v_k^{(c)} + \sum_{c\in \mathrm{Child}(b)} 2\alpha v_k^{(b)} \tag{34}$$

$$0 = v_k^{(b)} - v_k^{\mathrm{Pa}(b)} - \sum_{c\in \mathrm{Child}(b)} v_k^{(c)} - |\mathrm{Child}(b)| \cdot v_k^{(b)} \tag{35}$$

$$= (1 + |\mathrm{Child}(b)|) v_k^{(b)} - v_k^{\mathrm{Pa}(b)} - \sum_{c\in \mathrm{Child}(b)} v_k^{(c)} \tag{36}$$

$$v_k^{(b)} = \frac{v_k^{\mathrm{Pa}(b)} + \sum_{c\in \mathrm{Child}(b)} v_k^{(c)}}{1 + |\mathrm{Child}(b)|} \tag{37}$$

where $|\mathrm{Child}(b)|$ is the number of direct child nodes of $b$.