Supplementary material for Hierarchical Encoding of Sequential Data with Compact and Sub-linear Storage Cost

Huu Le[1], Ming Xu[1], Tuan Hoang[2], and Michael Milford[1]
[1] Queensland University of Technology (QUT), Australia
[2] Singapore University of Technology and Design (SUTD), Singapore

1. Deriving the optimal $\boldsymbol{\mu}$

We start by defining the optimization objective as per (7, main manuscript) as

$$\begin{aligned}\boldsymbol{\mathcal{F}}(\mathbf{V}, m, \boldsymbol{\mu}) = \quad & \sum_{k=1}^{h} \sum_{\mathbf{x} \in \mathcal{L}_k} \|\mathbf{P}_m^\top \mathbf{V}^\top \mathbf{x} - \mathbf{P}_m^\top \mathbf{V}^\top \boldsymbol{\mu}_k\|_2^2 \\ & + \sum_i \|\tilde{\mathbf{P}}_m^\top \mathbf{V}^\top \mathbf{x}_i - \tilde{\mathbf{P}}_m^\top \mathbf{V}^\top \boldsymbol{\mu}_0\|_2^2, \\ \text{s.t.} \quad & \mathbf{V}^\top \mathbf{V} = \mathbf{I} \end{aligned} \tag{1}$$

where $\boldsymbol{\mu} = \{\boldsymbol{\mu}_k\}_{k=1}^h$ are the cluster centroids with h being the number of classes, $\mathbf{V} \in O(d)$ is the learned data transformation before applying the low-dimensional projection, where $O(d)$ refers to the orthogonal group of dimension d, $m \leq d$ refers to the dimensionality of the projection $\mathbf{P}_m \in \mathbb{R}^{d \times m}$ defined in (8, main manuscript).

The end goal of the optimization is to find $\mathbf{V}, m, \boldsymbol{\mu}$ that minimizes $\boldsymbol{\mathcal{F}}$, formally

$$\min_{\mathbf{V}, m, \boldsymbol{\mu}} \boldsymbol{\mathcal{F}}(\mathbf{V}, m, \boldsymbol{\mu}). \tag{2}$$

We first start by finding the optimal $\boldsymbol{\mu}$ component-wise by taking the gradient of $\boldsymbol{\mathcal{F}}$ w.r.t. $\boldsymbol{\mu}_k$ and setting this to the zero vector for each k, specifically

$$\nabla_{\boldsymbol{\mu}_k} \boldsymbol{\mathcal{F}}(\mathbf{V}, m, \boldsymbol{\mu}) = \mathbf{0}. \tag{3}$$

Using the fact that $\|\mathbf{x}\|_2^2 = \mathbf{x}^\top \mathbf{x}$ and noticing that only the terms containing $\boldsymbol{\mu}_k$ remain, (3) becomes

$$\begin{aligned}\nabla_{\boldsymbol{\mu}_k} \boldsymbol{\mathcal{F}}(\mathbf{V}, m, \boldsymbol{\mu}) &= \nabla_{\boldsymbol{\mu}_k} \sum_{\mathbf{x} \in \mathcal{L}_k} (\mathbf{L}_m^\top (\mathbf{x} - \boldsymbol{\mu}_k))^\top (\mathbf{L}_m^\top (\mathbf{x} - \boldsymbol{\mu}_k)) \\ &= \nabla_{\boldsymbol{\mu}_k} \sum_{\mathbf{x} \in \mathcal{L}_k} (\mathbf{x} - \boldsymbol{\mu}_k)^\top \mathbf{L}_m \mathbf{L}_m^\top (\mathbf{x} - \boldsymbol{\mu}_k) \\ &= -2 \sum_{\mathbf{x} \in \mathcal{L}_k} \mathbf{L}_m \mathbf{L}_m^\top (\mathbf{x} - \boldsymbol{\mu}_k) = 0, \end{aligned}$$

where $\mathbf{L}_m = \mathbf{V}\mathbf{P}_m$ is the data transformation which consists of the first m columns of $\mathbf{V}$. From this, we can easily see that an optimal solution is given by

$$\boldsymbol{\mu}_k = \frac{1}{|\mathcal{L}_k|} \sum_{\mathbf{x} \in \mathcal{L}_k} \mathbf{x}, \tag{4}$$

consistent with (9, main manuscript). This is the global minimizer of $\boldsymbol{\mathcal{F}}$ since the $\boldsymbol{\mathcal{F}}$ is quadratic in μ_k for all k and the hessian of $\boldsymbol{\mathcal{F}}$ w.r.t. $\boldsymbol{\mu}_k$ is given by $\mathbf{L}_m \mathbf{L}_m^\top$ which is positive semidefinite. Again, note that the solution is independent of $\mathbf{V}$ and m.

2. Rewriting the optimization objective

To derive the optimal values for $\mathbf{V}$ and m, we need to rewrite (1), as

$$\begin{aligned}\boldsymbol{\mathcal{F}}(\mathbf{V}, m, \boldsymbol{\mu}) \quad &= \sum_{k=1}^{h} \sum_{\mathbf{x} \in \mathcal{L}_k} \|\mathbf{P}_m^\top \mathbf{V}^\top \mathbf{x} - \mathbf{P}_m^\top \mathbf{V}^\top \boldsymbol{\mu}_k\|_2^2 \\ &+ \sum_i \|\tilde{\mathbf{P}}_m^\top \mathbf{V}^\top \mathbf{x}_i - \tilde{\mathbf{P}}_m^\top \mathbf{V}^\top \boldsymbol{\mu}_0\|_2^2, \\ &= \sum_{k=1}^{h} \sum_{\mathbf{x} \in \mathcal{L}_k} (\mathbf{P}_m^\top \mathbf{V}^\top \mathbf{x} - \mathbf{P}_m^\top \mathbf{V}^\top \boldsymbol{\mu}_k)^T (\mathbf{P}_m^\top \mathbf{V}^\top \mathbf{x} - \mathbf{P}_m^\top \mathbf{V}^\top \boldsymbol{\mu}_k) \\ &+ \sum_i (\tilde{\mathbf{P}}_m^\top \mathbf{V}^\top \mathbf{x}_i - \tilde{\mathbf{P}}_m^\top \mathbf{V}^\top \boldsymbol{\mu}_0)^T (\tilde{\mathbf{P}}_m^\top \mathbf{V}^\top \mathbf{x}_i - \tilde{\mathbf{P}}_m^\top \mathbf{V}^\top \boldsymbol{\mu}_0) \\ &= \sum_{k=1}^{h} \sum_{\mathbf{x} \in \mathcal{L}_k} (\mathbf{x} - \boldsymbol{\mu}_i)^T \mathbf{V} \mathbf{P}_m \mathbf{P}_m^T \mathbf{V}^T (\mathbf{x} - \boldsymbol{\mu}_i) \\ &+ \sum_i (\mathbf{x}_i - \boldsymbol{\mu}_0)^T \mathbf{V} \tilde{\mathbf{P}}_m \tilde{\mathbf{P}}_m^T \mathbf{V}^T (\mathbf{x}_i - \boldsymbol{\mu}_0). \end{aligned}$$

Since each term in $\boldsymbol{\mathcal{F}}(\mathbf{V}, m, \boldsymbol{\mu})$ returns a scalar, one can employ the trace trick to re-write $\boldsymbol{\mathcal{F}}(\mathbf{V}, m, \boldsymbol{\mu})$ as

$$
\begin{aligned}
\mathcal{F}(\mathbf{V}, m, \boldsymbol{\mu}) \quad &= \sum_{k=1}^{h} \sum_{\mathbf{x} \in \mathcal{L}_k} \operatorname{tr}\left((\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{V} \mathbf{P}_m \mathbf{P}_m^T \mathbf{V}^T (\mathbf{x}-\boldsymbol{\mu}_i)\right) \\
&+ \sum_i \operatorname{tr}\left((\mathbf{x}_i-\boldsymbol{\mu}_0)^T \mathbf{V} \tilde{\mathbf{P}}_m \tilde{\mathbf{P}}_m^T \mathbf{V} (\mathbf{x}_i-\boldsymbol{\mu}_0)\right) \\
&= \operatorname{tr}\left(\sum_{k=1}^{h} \sum_{\mathbf{x} \in \mathcal{L}_k} (\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{V} \mathbf{P}_m \mathbf{P}_m^T \mathbf{V}^T (\mathbf{x}-\boldsymbol{\mu}_i)\right) \\
&+ \operatorname{tr}\left(\sum_i (\mathbf{x}_i-\boldsymbol{\mu}_0)^T \mathbf{V} \tilde{\mathbf{P}}_m \tilde{\mathbf{P}}_m^T \mathbf{V} (\mathbf{x}_i-\boldsymbol{\mu}_0)\right) \\
&= \operatorname{tr}\left(\sum_{k=1}^{h} \sum_{\mathbf{x} \in \mathcal{L}_k} \mathbf{P}_m \mathbf{P}_m^T \mathbf{V}^T (\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{V}\right) \\
&+ \operatorname{tr}\left(\sum_i \tilde{\mathbf{P}}_m \tilde{\mathbf{P}}_m^T \mathbf{V}^T (\mathbf{x}_i-\boldsymbol{\mu}_0)(\mathbf{x}_i-\boldsymbol{\mu}_0)^T \mathbf{V}\right).
\end{aligned}
$$

Based on the construction of $\mathbf{P}_m$ and $\tilde{\mathbf{P}}$ and using the fact that the trace of a matrix is the sum of its diagonal elements, for any matrix $\mathbf{A}$

$$
\operatorname{tr}(\tilde{\mathbf{P}}_m \tilde{\mathbf{P}}_m^T \mathbf{A}) = \operatorname{tr}(\mathbf{A}) - \operatorname{tr}(\mathbf{P}_m \mathbf{P}_m^T \mathbf{A}). \tag{5}
$$

Hence,

$$
\begin{aligned}
\mathcal{F}(\mathbf{V}, m, \boldsymbol{\mu}) \quad &= \operatorname{tr}\left(\sum_{k=1}^{h} \sum_{\mathbf{x} \in \mathcal{L}_k} \mathbf{P}_m \mathbf{P}_m^T \mathbf{V}^T (\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{V}\right) \\
&+ \operatorname{tr}\left(\sum_i \tilde{\mathbf{P}}_m \tilde{\mathbf{P}}_m^T \mathbf{V} (\mathbf{x}_i-\boldsymbol{\mu}_0)(\mathbf{x}_i-\boldsymbol{\mu}_0)^T \mathbf{V}\right) \\
&= \operatorname{tr}\left(\sum_{k=1}^{h} \sum_{\mathbf{x} \in \mathcal{L}_k} \mathbf{P}_m \mathbf{P}_m^T \mathbf{V}^T (\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{V}\right) \\
&+ \operatorname{tr}\left(\sum_i \tilde{\mathbf{P}}_m \tilde{\mathbf{P}}_m^T \mathbf{V}^T (\mathbf{x}_i-\boldsymbol{\mu}_0)(\mathbf{x}_i-\boldsymbol{\mu}_0)^T \mathbf{V}\right) \\
&= \operatorname{tr}\left(\sum_{k=1}^{h} \sum_{\mathbf{x} \in \mathcal{L}_k} \mathbf{P}_m \mathbf{P}_m^T \mathbf{V}^T (\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{V}\right) \\
&- \operatorname{tr}\left(\sum_i \mathbf{P}_m \mathbf{P}_m^T \mathbf{V}^T (\mathbf{x}_i-\boldsymbol{\mu}_0)(\mathbf{x}_i-\boldsymbol{\mu}_0)^T \mathbf{V}\right) \\
&+ \operatorname{tr}\left(\mathbf{V}^T (\mathbf{x}_i-\boldsymbol{\mu}_0)(\mathbf{x}_i-\boldsymbol{\mu}_0)^T \mathbf{V}\right).
\end{aligned}
$$

Finally, $\mathcal{F}(\mathbf{V}, m, \boldsymbol{\mu})$ can be written as

$$
\mathcal{G}(\mathbf{V}, m, \boldsymbol{\mu}) = \operatorname{trace}(\mathbf{P}_m \mathbf{P}_m^\top \mathbf{V}^\top \boldsymbol{\Sigma} \mathbf{V}) + \operatorname{trace}(\mathbf{V}^\top \mathbf{S}_0 \mathbf{V}), \tag{6}
$$

where the $\mathbf{S}_0$ and $\boldsymbol{\Sigma}$ are given by

$$
\mathbf{S}_0 = \sum_{\mathbf{x} \in \mathcal{D}} (\mathbf{x}-\boldsymbol{\mu}_0)(\mathbf{x}-\boldsymbol{\mu}_0)^\top, \tag{7}
$$

and

$$
\boldsymbol{\Sigma} = \sum_{k=1}^{h} \sum_{\mathbf{x} \in \mathcal{L}_k} (\mathbf{x}-\boldsymbol{\mu}_k)(\mathbf{x}-\boldsymbol{\mu}_k)^\top - \mathbf{S}_0. \tag{8}
$$

3. Deriving the optimal $\mathbf{V}$

We now wish to find $\mathbf{V}$ such that $\mathcal{G}(\mathbf{V}, m, \boldsymbol{\mu})$ is minimized, recalling that the rightmost term in (6) is invariant for all $\mathbf{V}$ shown in Section 3.2.2. of the main text. Hence, we only need to minimize the first term.

Recall that $\mathbf{P}_m \mathbf{P}_m^\top$ is a $d \times d$ matrix with 1 for the first m diagonal elements (from the top left) and 0 elsewhere. Given this, $\operatorname{trace}(\mathbf{P}_m \mathbf{P}_m^\top \mathbf{V}^\top \boldsymbol{\Sigma} \mathbf{V})$ is the sum of the first m diagonal elements in $\mathbf{V}^\top \boldsymbol{\Sigma} \mathbf{V}$. Each element in the sum can therefore be expressed as $\mathbf{v}_i^\top \boldsymbol{\Sigma} \mathbf{v}_i$, where $\mathbf{v}_i$ is column i in $\mathbf{V}$ and since $\mathbf{V}$ is orthogonal, $\|\mathbf{v}_i\|_2 = 1$. We wish to find $\mathbf{v}_i$ such that

$$
\min_{\|\mathbf{v}_i\|_2 = 1} \mathbf{v}_i^\top \boldsymbol{\Sigma} \mathbf{v}_i, \tag{9}
$$

and $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d\}$ are pairwise orthogonal. Since by construction, $\boldsymbol{\Sigma}$ is a square symmetric matrix, we apply the spectral theorem of Hermitian matrices and replace $\boldsymbol{\Sigma}$ with its eigendecomposition $\mathcal{Q} \boldsymbol{\Lambda} \mathcal{Q}^\top$, where $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ are the eigenvalues of $\boldsymbol{\Sigma}$ (assumed to be sorted from highest to lowest) and $\mathcal{Q}$ is an orthogonal matrix whose columns are the corresponding normalized eigenvectors. From this,

$$
\begin{aligned}
\min_{\|\mathbf{v}_i\|_2 = 1} \mathbf{v}_i^\top \boldsymbol{\Sigma} \mathbf{v}_i &= \min_{\|\mathbf{v}_i\|_2 = 1} \mathbf{v}_i^\top \mathcal{Q} \boldsymbol{\Lambda} \mathcal{Q}^\top \mathbf{v}_i && (10) \\
&= \min_{\|\mathbf{z}_i\|_2 = 1} \mathbf{z}_i^\top \boldsymbol{\Lambda} \mathbf{z}_i && (11) \\
&= \min_{\|\mathbf{z}_i\|_2 = 1} \sum_{j=1}^{d} z_{ij}^2 \lambda_j, && (12)
\end{aligned}
$$

where $\mathbf{z}_i = \mathcal{Q}^\top \mathbf{v}_i$ and z_{ij} is the j-th element of $\mathbf{z}_i$. Since $\mathcal{Q}$ is invertible (from orthogonality), we can find the optimal $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_d\} \in \mathbb{R}^{d \times d}$ such that $\mathbf{Z}$ is also orthogonal. Working recursively and noting the orthogonality constraint, it is clear that $\mathbf{z}_1 = (0, 0, \ldots, 0, 1)^\top$ and subsequently $\mathbf{z}_2 = (0, 0, \ldots, 0, 1, 0)^\top$ and so on. Consequently, the columns of $\mathbf{V} = \mathcal{Q}\mathbf{Z}$ are just the normalized eigenvectors of $\boldsymbol{\Sigma}$ where the columns are ordered in increasing order by the corresponding eigenvalues. Note that this construction yields a valid solution for all $m \leq d$.

4. Deriving the optimal m

We can now use the results from the previous sections to select m. Given that the columns of the optimal $\mathbf{V}$ are the normalized eigenvectors of $\boldsymbol{\Sigma}$, it follows that

$$
\operatorname{trace}(\mathbf{P}_m \mathbf{P}_m^\top \mathbf{V}^\top \boldsymbol{\Sigma} \mathbf{V}) = \sum_{i=1}^{m} \mathbf{v}_i \boldsymbol{\Sigma} \mathbf{v}_i = \sum_{i=1}^{m} \lambda_i. \tag{13}
$$

Clearly, if $\boldsymbol{\Sigma}$ has no negative eigenvalues, then $m = 0$, otherwise $m = k$, where k denotes the number of negative eigenvalues of $\boldsymbol{\Sigma}$. Note that $\boldsymbol{\Sigma}$ as defined in (8) is not necessarily positive semidefinite since $\mathbf{S}_0$ is subtracted from the sum. We find that in practice, it is always the case that $m = h$.