Supplementary Material
Kernel Embeddings of Latent Tree Graphical Models

Le Song, Ankur P. Parikh, Eric P. Xing

Section A of the supplemental contains a brief introduction to Tensor Algebra while Section B contains details on the Spectral Algorithm.

A Tensor Algebra

Here we give a brief introduction to tensor algebra (for more details, see [2]). A tensor is a multidimensional array, and its order is the number of dimensions, also known as modes. In this paper, vectors (tensors of order one) are denoted by boldface lowercase letters, *e.g.*, $\boldsymbol{a}$. Matrices (tensors of order two) are denoted by boldface capital letters, *e.g.*, $\boldsymbol{A}$. Higher-order tensors (order three or higher) are denoted by boldface caligraphic letters, *e.g.*, $\boldsymbol{\mathcal{T}}$. Scalars are denoted by lowercase letters, *e.g.*, a.
Subarrays of a tensor are formed when a subset of the indices is fixed. Particularly, a fiber is defined by fixing every index but one. Fibers are the higher-order analogue of matrix rows and columns. A colon is used to indicate all elements of a mode. Thus, the jth column of a matrix $\boldsymbol{A}$ is $\boldsymbol{A}(:, j)$, and the ith row of $\boldsymbol{A}$ is $\boldsymbol{A}(i, :)$. Analogously, the mode-n fiber of a Nth order tensor $\boldsymbol{\mathcal{T}}$ is then denoted as $\boldsymbol{\mathcal{T}}(i_1, i_2, \ldots, i_{n-1}, :, i_{n+1}, \ldots, i_N)$.
Tensors can be multiplied together. For matrices and vectors, we will use standard notation for their multiplications, *e.g.*, $\boldsymbol{B}\boldsymbol{a}$ and $\boldsymbol{A}\boldsymbol{B}$. For tensors of higher order, we are particularly interested in multiplying a tensor by matrices and vectors. The n-mode matrix product is the multiplication of a tensor with a matrix in mode n of the tensor. Let $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ be an Nth order tensor and $\boldsymbol{A} \in \mathbb{R}^{J \times I_n}$ be a matrix. Then

$$\boldsymbol{\mathcal{T}}' = \boldsymbol{\mathcal{T}} \times_n \boldsymbol{A} \quad \in \quad \mathbb{R}^{I_1 \times \ldots I_{n-1} \times J \times I_{n+1} \times \ldots \times I_N}, \tag{1}$$

where the entries $\boldsymbol{\mathcal{T}}'(i_1, \ldots, i_{n-1}, j, i_{n+1}, \ldots, i_N)$ are defined as $\sum_{i_n=1}^{I_n} \boldsymbol{\mathcal{T}}(i_1, \ldots, i_n, \ldots, i_N)\boldsymbol{A}(j, i_n)$. For example, if $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices, then $\boldsymbol{A} \times_1 \boldsymbol{B} = \boldsymbol{B}\boldsymbol{A}$ and $\boldsymbol{A} \times_2 \boldsymbol{B}^\top = \boldsymbol{A}\boldsymbol{B}$. We will further introduce two useful properties of n-mode matrix product. First, for distinct modes in a series of multiplications, the order of the multiplication can be exchanged

$$\boldsymbol{\mathcal{T}} \times_n \boldsymbol{A} \times_m \boldsymbol{B} = \boldsymbol{\mathcal{T}} \times_m \boldsymbol{B} \times_n \boldsymbol{A} \quad (m \neq n). \tag{2}$$

Second, the matrices can be combined first, if the modes in a series of multiplications are the same

$$\boldsymbol{\mathcal{T}} \times_n \boldsymbol{A} \times_n \boldsymbol{B} = \boldsymbol{\mathcal{T}} \times_n (\boldsymbol{B}\boldsymbol{A}). \tag{3}$$

We note that n-mode matrix product does not change the order of a tensor, but the size of the tensor may change. Multiplication of a tensor with a vector in mode n of the tensor is called n-mode vector product. Let $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ and $\boldsymbol{a} \in \mathbb{R}^{I_n}$. Then

$$\boldsymbol{\mathcal{T}}' = \boldsymbol{\mathcal{T}} \ \bar{\times}_n \ \boldsymbol{a} \quad \in \quad \mathbb{R}^{I_1 \times \ldots I_{n-1} I_{n+1} \times \ldots \times I_N} \tag{4}$$

where the entries $\boldsymbol{\mathcal{T}}'(i_1, \ldots, i_{n-1}, i_{n+1}, \ldots, i_N)$ is defined as $\sum_{i_n=1}^{I_n} \boldsymbol{\mathcal{T}}(i_1, i_2, \ldots, i_n, \ldots, i_N)\boldsymbol{a}(i_n)$. We note that n-mode vector product actually reduces the order of the tensor, *i.e.*, $\boldsymbol{\mathcal{T}}'$ is order $N-1$ if $\boldsymbol{\mathcal{T}}$ is order N. Note that in general $\boldsymbol{\mathcal{T}} \ \bar{\times}_n \ \boldsymbol{a} = \text{squeeze}(\boldsymbol{\mathcal{T}} \times_n \boldsymbol{a}^\top)$.

B Derivation of Spectral Algorithm

In this section, we provide a more detailed derivation of the spectral algorithm for (transformed) parameter learning. For simplicity of explanation, we will focus on latent tree structure where each internal node has exactly 3 neighbors. We can reroot the tree and redirect all the edges away from the root. For a variable X_s, we use α_s to denote its sibling, π_s to denote its parent, ι_s to denote its left child and ρ_s to denote its right child; the root node will have 3 children, we use ω_s to denote the extra child. All the observed variables are leaves in the tree, and we will use ι_s^*, ρ_s^*, π_s^* to denote an observed variable which is found by tracing in the direction from node s to its left child ι_s, right child ρ_s, and its parent π_s respectively. s^* denotes any observed variable in the subtree rooted at node s.

Recall the transformed messages:

* At leaf nodes, $\widetilde{m}_s = T_s^\top C_{s|\pi_s}^\top \phi(x_s) = (\mathcal{C}_{s|\pi_s} \times_2 T_s^\top)\ \bar{\times}_1\ \phi(x_s)$

** At internal nodes, $\widetilde{m}_s = (\mathcal{C}_{s^2|\pi_s} \times_1 T_{\iota_s}^{-1} \times_2 T_{\rho_s}^{-1} \times_3 T_s^\top)\ \bar{\times}_2\ \widetilde{m}_{\rho_s}\ \bar{\times}_1\ \widetilde{m}_{\iota_s}$

*** At the root, $b_r = (\mathcal{C}_{r^3} \times_1 T_{\iota_s}^{-1} \times_2 T_{\rho_s}^{-1} \times_3 T_{\omega_r}^{-1})\ \bar{\times}_3\ \widetilde{m}_{\omega_s}\ \bar{\times}_2\ \widetilde{m}_{\rho_s}\ \bar{\times}_1\ \widetilde{m}_{\iota_s}$

Let $\widetilde{\mathcal{C}}_{s^2|\pi_s} := \mathcal{C}_{s^2|\pi_s} \times_1 T_{\iota_s}^{-1} \times_2 T_{\rho_s}^{-1} \times_3 T_s^\top$ and $\widetilde{\mathcal{C}}_{r^3} := \mathcal{C}_{r^3} \times_1 T_{\iota_s}^{-1} \times_2 T_{\rho_s}^{-1} \times_3 T_{\omega_r}^{-1}$. We set $T_s = (U_s^\top \mathcal{C}_{s^*|\pi_s})^{-1}$. U_s is chosen [1] to be the top d right singular vectors of $\mathcal{C}_{\pi_s^* s^*}$, and therefore one can take the one-sided inverse of $(\mathcal{C}_{\pi_s^* s^*} U_s)$ assuming all latent variables have dimension d. For internal nodes we set s^* in $(\mathcal{C}_{\pi_s^* s^*} U_s)$ to ι_s^* while for leaves we set s^* to s. We have the following observable representation that we derive in the following subsections:

* At leaf nodes, $\widetilde{m}_s = (\mathcal{C}_{\pi_s^* s} U_s)^\dagger \mathcal{C}_{\pi_s^* s} \phi(x_s)$.

** At internal nodes, $\widetilde{\mathcal{C}}_{s^2|\pi_s} = \mathcal{C}_{\iota_s^* \rho_s^* \pi_s^*} \times_1 U_{\iota_s}^\top \times_2 U_{\rho_s}^\top \times_3 (\mathcal{C}_{\pi_s^* \iota_s^*} U_s)^\dagger$.

*** At the root, $\widetilde{\mathcal{C}}_{r^3} = \mathcal{C}_{\iota_r^* \rho_r^* \omega_r^*} \times_1 U_{\iota_r}^\top \times_2 U_{\rho_r}^\top \times_3 U_{\omega_r}^\top$

B.1 Root

Recall that

$$\begin{aligned} \widetilde{\mathcal{C}}_{r^3} &= \mathcal{C}_{r^3} \times_1 T_{\iota_r}^{-1} \times_2 T_{\rho_r}^{-1} \times_3 T_{\omega_r}^{-1} & (5)\\ &= \mathcal{C}_{r^3} \times_1 U_{\iota_r}^\top \mathcal{C}_{\iota_r^*|r} \times_2 U_{\rho_r}^\top \mathcal{C}_{\rho_r^*|r} \times_3 U_{\omega_r}^\top \mathcal{C}_{\omega_r^*|r} & (6)\\ &= \mathcal{C}_{r^3} \times_1 \mathcal{C}_{\iota_r^*|r} \times_2 \mathcal{C}_{\rho_r^*|r} \times_3 \mathcal{C}_{\omega_r^*|r} \times_1 U_{\iota_r}^\top \times_2 U_{\rho_r}^\top \times_3 U_{\omega_r}^\top & (7)\end{aligned}$$

where $T_s^{-1} = U_s^\top \mathcal{C}_{s^*|\pi_s}$.
We first prove that $\mathcal{C}_{r^3} \times_1 \mathcal{C}_{\iota_r^*|r} \times_2 \mathcal{C}_{\rho_r^*|r} \times_3 \mathcal{C}_{\omega_r^*|r} = \mathcal{C}_{\iota_r^* \rho_r^* \omega_r^*}$: Consider any $f, g, h \in \mathcal{F}$. Then,

$$\begin{aligned} \mathcal{C}_{r^3} \quad \times_1 &\quad \mathcal{C}_{\iota_r^*|r} \times_2 \mathcal{C}_{\rho_r^*|r} \times_3 \mathcal{C}_{\omega_r^*|r}\ \bar{\times}_3\ h\ \bar{\times}_2\ g\ \bar{\times}_1\ f & (8)\\ = &\quad \left\langle f \otimes g \otimes h, \mathcal{C}_{r^3} \times_1 \mathcal{C}_{\iota_r^*|r} \times_2 \mathcal{C}_{\rho_r^*|r} \times_3 \mathcal{C}_{\omega_r^*|r} \right\rangle & (9)\\ = &\quad \mathbb{E}_{X_r}\left[\left\langle \mathcal{C}_{\iota_r^*|r}^\top f, \phi(X_r) \right\rangle \left\langle \mathcal{C}_{\rho_r^*|r}^\top g, \phi(X_r) \right\rangle \left\langle \mathcal{C}_{\omega_r^*|r}^\top h, \phi(X_r) \right\rangle \right] & (10)\\ = &\quad \mathbb{E}_{X_r}\left[\left\langle f, \mathcal{C}_{\iota_r^*|r} \phi(X_r) \right\rangle \left\langle g, \mathcal{C}_{\rho_r^*|r} \phi(X_r) \right\rangle \left\langle h, \mathcal{C}_{\omega_r^*|r} \phi(X_r) \right\rangle \right] & (11)\\ = &\quad \mathbb{E}_{X_r}\left[\mathbb{E}_{X_{\rho_{r^*}}|X_r}\left[f(X_{\iota_r^*}) \right] \mathbb{E}_{X_{\rho_r^*}|X_r}\left[g(X_{\rho_r^*}) \right] \mathbb{E}_{X_{\omega_r^*}|X_r}\left[h(X_{\omega_r^*}) \right] \right] & (12)\\ = &\quad \mathbb{E}_{X_{\iota_r^*}, X_{\rho_r^*}, X_{\omega_r^*}}\left[f(X_{\iota_r^*}) g(X_{\rho_r^*}) h(X_{\omega_r^*}) \right] & (13)\\ = &\quad \left\langle f \otimes g \otimes h, \mathbb{E}_{X_{\iota_r^*}, X_{\rho_r^*}, X_{\omega_r^*}}\left[\phi(X_{\iota_r^*}) \otimes \phi(X_{\rho_r^*}) \otimes \phi(X_{\omega_r^*}) \right] \right\rangle & (14)\\ = &\quad \mathcal{C}_{\iota_r^* \rho_r^* \omega_r^*}\ \bar{\times}_3\ h\ \bar{\times}_2\ g\ \bar{\times}_1\ f & (15)\end{aligned}$$

Combining this result with Eq. 7 gives,

$$\widetilde{\mathcal{C}}_{r^3} = \mathcal{C}_{r^3} \times_1 T_{\iota_r}^{-1} \times_2 T_{\rho_r}^{-1} \times_3 T_{\omega_r}^{-1} = \mathcal{C}_{\iota_r^* \rho_r^* \omega_r^*} \times_1 U_{\iota_r}^\top \times_2 U_{\rho_r}^\top \times_3 U_{\omega_r}^\top \qquad (16)$$

[1] This is not the only valid choice of U_s but will generally result in better performance. See [1, 3] for more details.

B.2 Leaf

Recall that $\widetilde{m}_s = T_s^\top C_{s|\pi_s}^\top \phi(x_s)$ and $T_s = (U_s^\top \mathcal{C}_{s^*|\pi_s})^{-1}$. However since s is a leaf we can set $s^* = s$. Consider expanding the related quantity $\widetilde{m}_s^\top (U_s^\top \mathcal{C}_{s\pi_s^*})$:

$$
\begin{aligned}
\widetilde{m}_s^\top (U_s^\top \mathcal{C}_{s\pi_s^*}) &= \phi^T(x_s) C_{s|\pi_s} (U_s^\top C_{s|\pi_s})^{-1} (U_s^\top \mathcal{C}_{s\pi_s^*}) && (17)\\
&= \phi^T(x_s) C_{s|\pi_s} (U_s^\top C_{s|\pi_s})^{-1} (U_s^\top \mathcal{C}_{s|\pi_s} \mathcal{C}_{\pi_s^2} \mathcal{C}_{\pi_s^*|\pi_s}^\top) && (18)\\
&= \phi^T(x_s) C_{s|\pi_s} (U_s^\top C_{s|\pi_s})^{-1} (U_s^\top \mathcal{C}_{s|\pi_s})(\mathcal{C}_{\pi_s^2} \mathcal{C}_{\pi_s^*|\pi_s}^\top) && (19)\\
&= \phi^T(x_s) C_{s|\pi_s} \mathcal{C}_{\pi_s^2} \mathcal{C}_{\pi_s^*|\pi_s}^\top && (20)\\
&= \phi(x_s)^\top C_{s\pi_s^*} && (21)
\end{aligned}
$$

where we have used the fact that $\mathcal{C}_{s|\pi_s} \mathcal{C}_{\pi_s^2} \mathcal{C}_{\pi_s^*|\pi_s}^\top = \mathcal{C}_{s\pi_s^*}$ (which is proved using the same technique as used in Section B.1).

This implies that $\widetilde{m}_s = (\mathcal{C}_{\pi_s^* s} U_s)^\dagger C_{\pi_s^* s} \phi(x_s) = C_{s\pi_s^*} (U_s^\top \mathcal{C}_{s\pi_s^*})^\dagger \bar{\times}_1 \phi(x_s)$. We choose U_s to be the top d right singular vectors of $\mathcal{C}_{\pi_s^* s}$, and therefore the one-sided inverse exists (since all latent variables are assumed to have dimension d).

B.3 Intermediate Node

Recall that $T_s = (U_s^\top \mathcal{C}_{s^*|\pi_s})^{-1}$ and $\widetilde{C}_{s^2|\pi_s} = \mathcal{C}_{s^2|\pi_s} \times_1 T_{\iota_s}^{-1} \times_2 T_{\rho_s}^{-1} \times_3 T_s^\top$. Thus,

$$
\widetilde{C}_{s^2|\pi_s} = \mathcal{C}_{s^2|\pi_s} \times_1 U_{\iota_s}^\top \mathcal{C}_{\iota_s^*|s} \times_2 U_{\iota_s}^\top \mathcal{C}_{\rho_s^*|s} \times_3 (\mathcal{C}_{s|\pi_s}^\top U_s)^{-1} \tag{22}
$$

Consider expanding the quantity $\widetilde{C}_{s^2|\pi_s} \times_3 (\mathcal{C}_{\pi_s^* s^*} U_s)$:

$$
\begin{aligned}
\widetilde{C}_{s^2|\pi_s} \times_3 (\mathcal{C}_{\pi_s^* s^*} U_s) &= \mathcal{C}_{s^2|\pi_s} \times_1 U_{\iota_s}^\top \mathcal{C}_{\iota_s^*|s} \times_2 U_{\iota_s}^\top \mathcal{C}_{\rho_s^*|s} \times_3 (\mathcal{C}_{s|\pi_s}^\top U_s)^{-1} \times_3 (\mathcal{C}_{\pi_s^* s^*} U_s) && (23)\\
&= \mathcal{C}_{s^2|\pi_s} \times_1 \mathcal{C}_{\iota_s^*|s} \times_2 \mathcal{C}_{\rho_s^*|s} \times_3 (\mathcal{C}_{\pi_s^* s^*} U_s)(\mathcal{C}_{s|\pi_s}^\top U_s)^{-1} \times_1 U_{\iota_s}^\top \times_2 U_{\rho_s}^\top && (24)\\
&= \mathcal{C}_{s^2|\pi_s} \times_1 \mathcal{C}_{\iota_s^*|s} \times_2 \mathcal{C}_{\rho_s^*|s} \times_3 (\mathcal{C}_{\pi_s^*|\pi_s} C_{\pi_s^2} \mathcal{C}_{s|\pi_s}^\top U_s)(\mathcal{C}_{s|\pi_s}^\top U_s)^{-1} \times_1 U_{\iota_s}^\top \times_2 U_{\rho_s}^\top && (25)\\
&= \mathcal{C}_{s^2|\pi_s} \times_1 \mathcal{C}_{\iota_s^*|s} \times_2 \mathcal{C}_{\rho_s^*|s} \times_3 (\mathcal{C}_{\pi_s^*|\pi_s} C_{\pi_s^2})(\mathcal{C}_{s|\pi_s}^\top U_s)(\mathcal{C}_{s|\pi_s}^\top U_s)^{-1} \times_1 U_{\iota_s}^\top \times_2 U_{\rho_s}^\top && (26)\\
&= \mathcal{C}_{s^2|\pi_s} \times_1 \mathcal{C}_{\iota_s^*|s} \times_2 \mathcal{C}_{\rho_s^*|s} \times_3 (\mathcal{C}_{\pi_s^*|\pi_s} C_{\pi_s^2}) \times_1 U_{\iota_s}^\top \times_2 U_{\rho_s}^\top && (27)\\
&= \mathcal{C}_{\iota_s^*, \rho_s^*, \pi_s^*} \times_1 U_{\iota_s}^\top \times_2 U_{\rho_s}^\top && (28)
\end{aligned}
$$

where in the last line we have claimed that $\mathcal{C}_{\iota_s^*, \rho_s^*, \pi_s^*} = \mathcal{C}_{s^2|\pi_s} \times_1 \mathcal{C}_{\iota_s^*|s} \times_2 \mathcal{C}_{\rho_s^*|s} \times_3 \mathcal{C}_{\pi_s^*|\pi_s} \mathcal{C}_{\pi_s^2}$. To prove this assertion, first consider the $\mathcal{C}_{s^2|\pi_s} \times_1 \mathcal{C}_{\iota_s^*|s} \times_2 \mathcal{C}_{\rho_s^*|s}$ part. For any $f, g \in \mathcal{F}$:

$$
\begin{aligned}
\left\langle f \otimes g, \mathcal{C}_{s^2|\pi_s} \times_1 \mathcal{C}_{\iota_s^*|s} \times_2 \mathcal{C}_{\rho_s^*|s} \bar{\times}_3 \phi(x_{\pi_s}) \right\rangle &= \left\langle (\mathcal{C}_{\iota_s^*|s}^\top f) \otimes (\mathcal{C}_{\rho_s^*|s}^\top g), \mathcal{C}_{s^2|\pi_s} \bar{\times}_3 \phi(x_{\pi_s}) \right\rangle && (29)\\
&= \left\langle (\mathcal{C}_{\iota_s^*|s}^\top f) \otimes (\mathcal{C}_{\rho_s^*|s}^\top g), \mathbb{E}_{X_s|x_{\pi_s}} [\phi(X_s) \otimes \phi(X_s)] \right\rangle && (30)\\
&= \mathbb{E}_{X_s|x_{\pi_s}} \left[\left\langle (\mathcal{C}_{\iota_s^*|s}^\top f) \otimes (\mathcal{C}_{\rho_s^*|s}^\top g), \phi(X_s) \otimes \phi(X_s) \right\rangle \right] && (31)\\
&= \mathbb{E}_{X_s|x_{\pi_s}} \left[\langle f, \mathcal{C}_{\iota_s^*|s} \phi(X_s) \rangle \langle g, \mathcal{C}_{\rho_s^*|s} \phi(X_s) \rangle \right] && (32)\\
&= \mathbb{E}_{X_s|x_{\pi_s}} \left[\mathbb{E}_{X_{\iota_s^*}|X_s}[f(X_{\iota_s^*})] \mathbb{E}_{X_{\rho_s^*}|X_s}[g(X_{\rho_s^*})] \right] && (33)\\
&= \mathbb{E}_{\iota_s^*, \rho_s^*|x_{\pi_s}} \left[f(X_{\iota_s^*}) g(X_{\rho_s^*}) \right] && (34)\\
&= \left\langle f \otimes g, \mathcal{C}_{\iota_s^*, \rho_s^*|\pi_s} \bar{\times}_3 \phi(x_{\pi_s}) \right\rangle && (35)
\end{aligned}
$$

Thus, $\mathcal{C}_{\iota_s^* \rho_s^* | \pi_s} = \mathcal{C}_{s^2|\pi_s} \times_1 \mathcal{C}_{\iota_s^*|s} \times_2 \mathcal{C}_{\rho_s^*|s}$. We can then conclude (using a similar derivation to that in Section B.1) that $\mathcal{C}_{\iota_s^*, \rho_s^*, \pi_s^*} = \mathcal{C}_{\iota_s^* \rho_s^* | \pi_s} \times_3 \mathcal{C}_{\pi_s^*|\pi_s} \mathcal{C}_{\pi_s^2}$. Thus,

$$\mathcal{C}_{\iota_s^*, \rho_s^*, \pi_s^*} = \mathcal{C}_{s^2|\pi_s} \times_1 \mathcal{C}_{\iota_s^*|s} \times_2 \mathcal{C}_{\rho_s^*|s} \times_3 \mathcal{C}_{\pi_s^*|\pi_s} \mathcal{C}_{\pi_s^2} \tag{36}$$

Now, returning to Eq. 28 we get that

$$\widetilde{C}_{s^2|\pi_s} = \mathcal{C}_{\iota_s^*, \rho_s^*, \pi_s^*} \times_1 U_{\iota_s^*}^\top \times_2 U_{\rho_s^*}^\top \times_3 (\mathcal{C}_{\pi_s^* s^*} U_s)^\dagger \tag{37}$$

where one valid choice for s^* is ι_s^*. U_s is chosen to be the top d right singular vectors of $\mathcal{C}_{\pi_s^* \iota_s^*}$, and therefore one can take a one-sided inverse of $(\mathcal{C}_{\pi_s^* s^*} U_s)$ (assuming all latent variables have dimension d).

References

[1] D. Hsu, S. Kakade, and T. Zhang. A spectral algorithm for learning hidden markov models. In *COLT*, 2009.

[2] Tamara. Kolda and Brett Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[3] A. Parikh, L. Song, and E. Xing. A spectral algorithm for latent tree graphical models. In *ICML*, 2011.