Supplemental for Spectral Algorithm For Latent Tree Graphical Models

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

The supplemental contains 3 main things.

1. The first is network plots of the latent variable tree learned by [1] for the stock market data, and the Chow Liu tree to give a more intuitive explanation why latent variable trees can lead to better performance.
2. The second is a more detailed representation of the tensor representation where internal nodes are allowed to be evidence variables.
3. The third is the proof of Theorem 1.

1 Latent Tree Structure for Stock Data

The latent tree structure learned by the algorithm by [1] is shown in Figure 1. The blue nodes are hidden nodes and the red nodes are observed. Note how integrating out some of these hidden nodes could lead to very large cliques. Thus it is not surprising why both our spectral method and EM perform better than Chow Liu. The Chow Liu Tree is shown in Figure 1. Note how it is forced to pick some of the observed variables as hubs even if latent variables may be more natural.

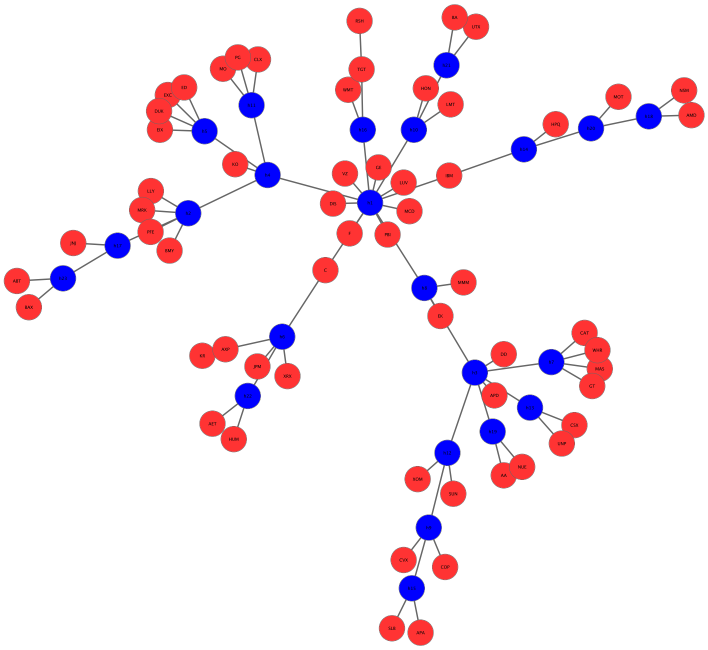

Figure 1: Latent variable tree learned by [1]. The hidden variables are in blue while the observed variables are in red. As one can see the hidden variables can model significantly more complex relationships among the observed variables.

2 More Detailed Information about Tensor Representation for LTMs

The computation of the marginal distribution of the observed variables can be expressed in terms of tensor multiplications. Basically, the information contained in each tensor will correspond to the information in a conditional probability table (CPT) of the model and the tensor multiplications implement the summations. However, there are multiple ways of rewriting the marginal distribution of the observed variables using tensor notation, and not all of them provide intuition or easy derivation to a spectral algorithm. In this section, we will derive a specific representation of the latent tree models which requires only tensors up to 3rd order and provides us a basis for deriving a spectral algorithm.

More specifically, we first select a latent or observed variable as the root node and sort the nodes in the tree in topological order. Then we associate the root node X_r with a vector $\boldsymbol{r}$ related to the marginal probability of X_r. Depending on whether the root node is latent or observed, its entries are defined as

	X_r latent	X_r observed
$\boldsymbol{r}(k)$	$\mathbb{P}[X_r = k]$	$\delta_{kx_r}\mathbb{P}[x_r]$

where δ_{kx_r} is an indicator variable defined as $\delta_{kx_r} = 1$ if $k = x_r$ and 0 otherwise. Effectively, δ_{kx_r} sets all entries of $\boldsymbol{r}$ to zero except the one corresponding to $\mathbb{P}[x_r]$.

Next, we associate each internal node X_i with a 3rd order tensor $\boldsymbol{\mathcal{T}}_i$ related the conditional probabilty table between X_i and its parent X_{π_i}. This tensor is diagonal in its 2nd and 3rd mode, and hence its nonzero entries can be accessed by two indices k and l. Depending on whether the internal node and its parent are latent or observed variables, the nonzero entries of $\boldsymbol{\mathcal{T}}_i$ are defined as

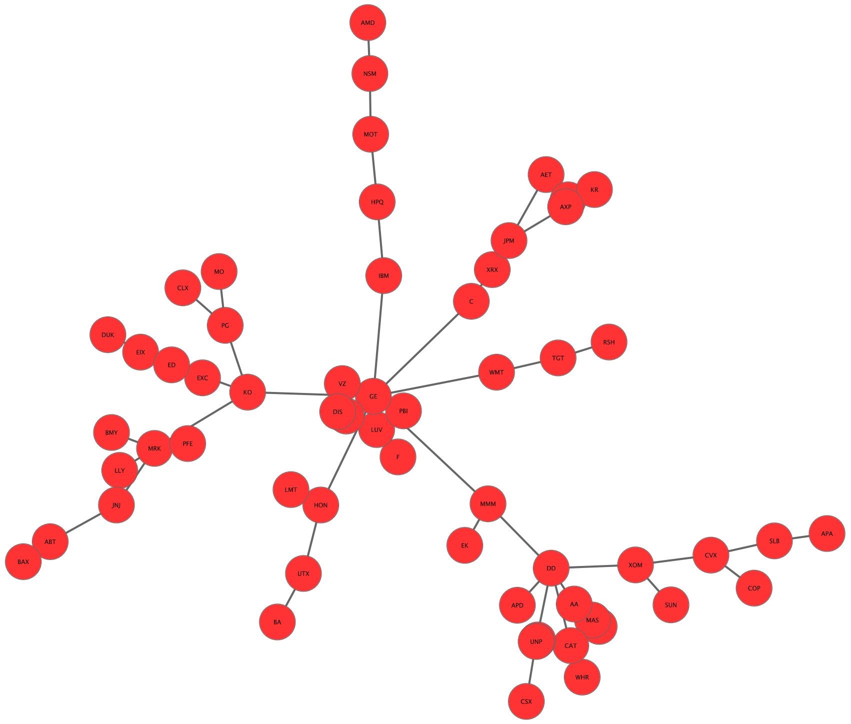

Figure 2: Tree learned by chow liu algorithm over only observed variables. Note how it is forced to pick some of the observed variables as hubs even if latent variables may be more natural.

$\boldsymbol{\mathcal{T}}_i(k,l,l)$	X_{π_i} latent	X_{π_i} observed
X_i latent	$\mathbb{P}[X_i = k \vert X_{\pi_i} = l]$	$\delta_{lx_{\pi_i}}\mathbb{P}[X_i = k \vert x_{\pi_i}]$
X_i observed	$\delta_{kx_i}\mathbb{P}[x_i \vert X_{\pi_i} = l]$	$\delta_{kx_i}\delta_{lx_{\pi_i}}\mathbb{P}[x_i \vert x_{\pi_i}]$

where δ_{kx_i} and $\delta_{lx_{\pi_i}}$ are also indicator variables. Effectively, the indicator variables zero out further entries in $\boldsymbol{\mathcal{T}}_i$ for those values that are not equal to the actual observation.

Last, we associate each leaf node x_i, which is always observed, with a diagonal matrix $\boldsymbol{M}_i$ related to the likelihood of x_i. Depending on whether the parent of x_i is latent of observed, the diagonal entries of $\boldsymbol{M}_i$ are defined as

	X_{π_i} latent	X_{π_i} observed
$\boldsymbol{M}_i(k,k)$	$\mathbb{P}[x_i \vert X_{\pi_i} = k]$	$\delta_{kx_{\pi_i}}\mathbb{P}[x_i \vert x_{\pi_i}]$

Let $\boldsymbol{M}_i$ defined above be the messages passed from the leaf nodes to their parents. We can show that the marginal probability of the observed variables can be computed recursively using a message passing algorithm: each node in the tree sends a message to its parent according to the reverse topological order of the nodes, and the final messages are aggregated in the root to give the desired quantity.

More formally, the outgoing message from an internal node X_i to its parent can be computed as

$$\boldsymbol{M}_i = \boldsymbol{\mathcal{T}}_i \;\bar{\times}_1\; (\boldsymbol{M}_{j_1}\boldsymbol{M}_{j_2}\dots\boldsymbol{M}_{j_J}\;\mathbf{1}_i) \tag{1}$$

where $j_1, j_2, \dots, j_J \in \chi_i$ range over all children of X_i ($J = |\chi_i|$). The $\mathbf{1}_i$ is a vector of all ones with suitable size, and it is used to reduce the incoming messages (all are diagonal matrices) to a single vector. The computation in (1) essentially implements the message update we often see in an ordinary message passing algorithm ([5]), *i.e.*,

$$m_i[x_{\pi_i}] = \sum_{x_i}\mathbb{P}[x_i|x_{\pi_i}]m_{j_1}[x_i]\dots m_{j_J}[x_i], \tag{2}$$

where $m_j[x_i]$ represents incoming messages to X_i (or intermediate results of the marginalization operation by summing out all latent variables in subtree $\mathscr{T}_j$). The $\boldsymbol{M}_{j_1}\boldsymbol{M}_{j_2}\dots\boldsymbol{M}_{j_J}\;\mathbf{1}_i$ corresponds to aggregating all incoming messages $m_{j_1}[x_i]\dots m_{j_J}[x_i]$, and the $\boldsymbol{\mathcal{T}}_i\;\bar{\times}_1 *$ corresponds to $\sum_{x_i}\mathbb{P}[x_i|x_{\pi_i}]\,*$.

At the root node, all incoming messages are combined to produce the final joint probability, *i.e.*,

$$\mathbb{P}[x_1, \dots, x_O] = \boldsymbol{r}^\top (\boldsymbol{M}_{j_1}\boldsymbol{M}_{j_2}\dots\boldsymbol{M}_{j_J}\;\mathbf{1}_r)\,. \tag{3}$$

Here $\boldsymbol{r}^\top *$ basically implements the operation $\sum_{x_r}\mathbb{P}[x_r]*$, which sums out the root variable.

3 Notation for Proof of Theorem 1

We now proceed to prove Theorem 1. $\| \cdot \|_2$ refers to spectral norm for matrices and tensors (but normal euclidean norm for vectors). $\| \cdot \|_1$ refers to induced 1 norm for matrices and tensors (max column sum), (but normal l1 norm for vectors). $\| \cdot \|_F$ refers to Frobenius norm.

The tensor spectral norm (for 3 dimensions) is defined in [4]:

$$\|\boldsymbol{\mathcal{T}}\|_2 = \sup_{\|v_i\|_2 \le 1} \boldsymbol{\mathcal{T}} \bar{\times}_3 v_3 \bar{\times}_2 v_2 \bar{\times}_1 v_1 \tag{4}$$

We will define the induced 1-norm of a tensor as

$$\|\boldsymbol{\mathcal{T}}\|_{1,1} = \sup_{\|v\|_1 \le 1} \|\boldsymbol{\mathcal{T}} \bar{\times}_1 v\|_1 \tag{5}$$

using the ℓ_1 norm of a matrix (*i.e.*, $\|A\|_1 = \sup_{\|v\|_1 \le 1} \|Av\|_1$).

For more information about matrix norms see [2].

In general, we suppress the actual subscripts/superscripts on $\boldsymbol{U}$ and $\boldsymbol{O}$. It is implied that $\boldsymbol{U}$ and $\boldsymbol{O}$ can often be different depending on the transform being considered. However, this makes the notation very messy. It will generally be clear from context which $\boldsymbol{U}$ and $\boldsymbol{O}$ are being referred to. When it is not we will arbitrarily index them $1, 2, ...$, so that it is clear which corresponds to which.

In general, for simplicity of exposition, we assume that all internal nodes in the tree are unobserved, and all leaves are observed (since this is the hardest case).

The proof generally follows the technique of HKZ [3], but has key differences due to the tree topology instead of the HMM.

We define $\tilde{\boldsymbol{M}}_i = (\boldsymbol{O}^\top \widehat{\boldsymbol{U}})^{-1} \boldsymbol{M}_i (\boldsymbol{O}^\top \widehat{\boldsymbol{U}})$. Then as long as $(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})$ is invertible, $(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})^{-1} \tilde{\boldsymbol{M}}_i (\boldsymbol{O}^\top \widehat{\boldsymbol{U}}) = \boldsymbol{M}_i$. (We admit this is a slight abuse of notation, since $\tilde{\boldsymbol{M}}_i$ was previously defined to be $(\boldsymbol{U}^\top \boldsymbol{O})^{-1} \boldsymbol{M}_i (\boldsymbol{U}^\top \boldsymbol{O})$, but as long as $(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})$ is invertible it doesn't really matter whether it equals $(\boldsymbol{U}^\top \boldsymbol{O})$ or not for the purposes of this proof). The other quantities are defined similarly.

We seek to prove the following theorem:

Theorem 1 *Pick any $\epsilon > 0, \delta < 1$. Let*

$$N \ge O\left(\frac{1}{\epsilon^2}\left(\frac{(d_{max} S_H)^{2\ell+1} S_O}{\min_i \sigma_{S_H}(\boldsymbol{O}_i)^2 \min_{i \ne j} \sigma_{S_H}(\boldsymbol{P}_{i,j})^4}\right)\right) \log \frac{|\mathscr{O}|}{\delta} \tag{6}$$

Then with probability $1 - \delta$

$$\sum_{x_1, \ldots, x_O} \left| \widehat{\mathbb{P}}[x_1, \ldots, x_O] - \mathbb{P}[x_1, \ldots, x_O] \right| \le \epsilon \tag{7}$$

In many cases, if the frequency of the observation symbols follow certain distributions, than the dependence on S_O can be removed as showed in HKZ [3]. That observation can easily be incorporated into our theorem if desired.

4 Concentration Bounds

$$\epsilon_i = \|\widehat{\boldsymbol{P}}_i - \boldsymbol{P}_i\|_F \tag{8}$$

$$\epsilon_{i,j} = \|\widehat{\boldsymbol{P}}_{i,j} - \boldsymbol{P}_{i,j}\|_F \tag{9}$$

$$\epsilon_{x,i,j} = \|\widehat{\boldsymbol{P}}_{x,i,j} - \boldsymbol{P}_{x,i,j}\|_F \tag{10}$$

$$\epsilon_{i,j,k} = \|\widehat{\boldsymbol{P}}_{i,j,k} - \boldsymbol{P}_{i,j,k}\|_F \tag{11}$$

(x denotes a fixed element while i, j, k are over indices).

As the number of samples N gets large, we expect these quantities to be small.

Lemma 1 (variant of HKZ [3]) *If the algorithm independently samples N observation triples from the tree, then with probability at least $1 - \delta$.*

$$\epsilon_i \le \sqrt{\frac{C}{N} \ln \frac{|\mathscr{O}|}{\delta}} + \sqrt{\frac{1}{N}} \tag{12}$$

$$\epsilon_{i,j} \le \sqrt{\frac{C}{N} \ln \frac{|\mathscr{O}|}{\delta}} + \sqrt{\frac{1}{N}} \tag{13}$$

$$\epsilon_{i,j,k} \le \sqrt{\frac{C}{N} \ln \frac{|\mathscr{O}|}{\delta}} + \sqrt{\frac{1}{N}} \tag{14}$$

$$\max_x \epsilon_{i,x,j} \le \sqrt{\frac{C}{N} \ln \frac{|\mathscr{O}|}{\delta}} + \sqrt{\frac{1}{N}} \tag{15}$$

$$\max_x \epsilon_{x,i,j} \le \sqrt{\frac{S_O}{N} \ln \frac{|\mathscr{O}|}{\delta}} + \sqrt{\frac{S_O}{N}} \tag{16}$$

where C is some constant (from the union bound over $O(V^3)$). (V is the total number of observed variables in the tree). The proof is the same as that of HKZ [3] except the union bound is larger. The last bound can be made tighter, identical to HKZ, but for simplicity we do not pursue that approach here.

5 Eigenvalue Bounds

Basically this is Lemma 9 in HKZ [3], which is stated below for completeness:

Lemma 2 *Suppose $\epsilon_{i,j} \leq \varepsilon \times \sigma_{S_H}(\boldsymbol{P}_{i,j})$ for some $\varepsilon < 1/2$. Let $\varepsilon_0 = \epsilon_{i,j}^2/((1-\varepsilon)\sigma_{S_H}(\boldsymbol{P}_{i,j}))^2$. Then:*

1. $\varepsilon_0 < 1$
2. $\sigma_{S_H}(\widehat{\boldsymbol{U}}^\top \widehat{\boldsymbol{P}}_{i,j}) \geq (1-\varepsilon)\sigma_{S_H}(\boldsymbol{P}_{i,j})$
3. $\sigma_{S_H}(\widehat{\boldsymbol{U}}^\top \boldsymbol{P}_{i,j}) \geq \sqrt{1-\varepsilon_0}\sigma_{S_H}(\boldsymbol{P}_{i,j})$
4. $\sigma_{S_H}(\boldsymbol{O}^\top \widehat{\boldsymbol{U}}) \geq \sqrt{1-\varepsilon}\sigma_{S_H}(\boldsymbol{O})$

The proof is in HKZ [3].

6 Bounding the Transformed Quantities

If Lemma 2 holds then $(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})$ is invertible. Thus, if we define $\tilde{\boldsymbol{M}}_i = (\boldsymbol{O}^\top \widehat{\boldsymbol{U}})^{-1}\boldsymbol{M}_i(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})$. Then clearly, $(\hat{U}^\top \boldsymbol{O})^{-1}\tilde{\boldsymbol{M}}_i(\boldsymbol{O}^\top \widehat{\boldsymbol{U}}) = \boldsymbol{M}_i$. (We admit this is a slight abuse of notation, since $\tilde{\boldsymbol{M}}_i$ is previously defined to be $(\boldsymbol{U}^\top \boldsymbol{O})^{-1}\boldsymbol{M}_i(\boldsymbol{U}^\top \boldsymbol{O})$, but as long as $(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})$ is invertible it doesn't really matter whether it equals $(\boldsymbol{U}^\top \boldsymbol{O})$ or not for the purposes of this proof). The other quantities are defined similarly.

We seek to bound the following four quantities:

$$\delta_{one}^i = \|(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})(\widehat{\boldsymbol{1}}_i - \tilde{\boldsymbol{1}}_i)\|_1 \tag{17}$$

$$\gamma_i = \|(\widehat{\boldsymbol{\mathcal{T}}}_i - \tilde{\boldsymbol{\mathcal{T}}}_i) \times_1 (\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)^{-1} \times_2 (\boldsymbol{O}_2^\top \widehat{\boldsymbol{U}}_2) \times_3 (\boldsymbol{O}_3^\top \widehat{\boldsymbol{U}}_3)^{-1}\|_2 \tag{18}$$

$$\delta_{root} = \|(\widehat{\boldsymbol{r}} - \tilde{\boldsymbol{r}})^T (\boldsymbol{O}_{j_1^* r}^\top \widehat{\boldsymbol{U}}_{j_1^*})^{-1}\|_\infty \tag{19}$$

$$\triangle_i = \sum_{\boldsymbol{x}_i} \|(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)(\widehat{\boldsymbol{M}}_i - \tilde{\boldsymbol{M}}_i)(\boldsymbol{O}_2^\top \widehat{\boldsymbol{U}}_2)^{-1}\|_1 \tag{20}$$

Here $\boldsymbol{x}_i$ denotes all observations that are in the subtree of node i (since i may be hidden or observed). Sometimes we like to distinguish between when when i is observed and when i is hidden.Thus, we sometimes refer to the quantity $\triangle_i^{obs}$ and $\triangle_i^{hidden}$ for when i is observed or hidden respectively.

Again note that the numbering in $(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)$ and $(\boldsymbol{O}_2^\top \widehat{\boldsymbol{U}}_2)$ is just there to avoid confusion in the same equation (In reality there are many $\boldsymbol{U}$'s and $\boldsymbol{O}$'s).

Lemma 3 *Assume $\epsilon_{i,j} \leq \sigma_{S_H}(\boldsymbol{P}_{i,j})/3$ for all $i \neq j$. Then*

$$\delta_{root} \leq \frac{2\epsilon_r}{\sqrt{3}\sigma_{S_H}(\boldsymbol{O}_{j_1^* r})} \tag{21}$$

$$\delta_{one}^i \leq 4\sqrt{S_H}\left(\frac{\epsilon_{i,j}}{\sigma_{S_H}(\boldsymbol{P}_{i,j})^2} + \frac{\epsilon_i}{\sqrt{3}\sigma_{S_H}(\boldsymbol{P}_{i,j})}\right) \tag{22}$$

$$\gamma_i \leq \frac{4\sqrt{S_H}}{\sigma_{S_H}(\boldsymbol{O})}\left(\frac{\epsilon_{m,j}}{\sigma_{S_H}(\boldsymbol{P}_{i,j})^2} + \frac{\epsilon_{m,j,k}}{\sqrt{3}\sigma_{S_H}(\boldsymbol{P}_{i,j})}\right) \tag{23}$$

$$\triangle_i^{hidden} \leq \left((1+\gamma_i)\prod_{k=1}^{J}(1+\triangle_{j_k})\delta_{one}^i + (1+\gamma_i)m\prod_{k=1}^{J}(1+\triangle_{j_k}) - m\right) \tag{24}$$

$$\triangle_i^{obs} \leq \leq 4\frac{\sqrt{S_H}}{\sigma_{S_H}(\boldsymbol{O})}\left(\frac{\epsilon_{i,j}}{(\sigma_{S_H}(\boldsymbol{P}_{j,i}))^2} + \frac{\sum_{x_i}\epsilon_{m,x_i,j}}{\sqrt{3}\sigma_{S_H}(\boldsymbol{P}_{i,j})}\right) \tag{25}$$

The main challenge in this part is $\triangle_v$ and γ_v^{hidden}. The rest are similar to HKZ. However, we go through the other bounds to be more explicit about some of the properties used, since sometimes we have used different norms etc.

6.1 δ_{root}

We note that $\widehat{\boldsymbol{r}} = \widehat{\boldsymbol{P}}_{j_1^*}^\top \widehat{\boldsymbol{U}}$ and similarly $\tilde{\boldsymbol{r}} = \boldsymbol{P}_{j_1^*}^\top \widehat{\boldsymbol{U}}$.

$$\delta_{root} = \|(\widehat{\boldsymbol{r}} - \tilde{\boldsymbol{r}})^\top (\boldsymbol{O}_{j_1^* r}^\top \widehat{\boldsymbol{U}}_{j_1^*})^{-1}\|_\infty \leq \|\widehat{\boldsymbol{P}}_{j_1^*}^\top - \boldsymbol{P}_{j_1^*}^\top\|_2 \|\widehat{\boldsymbol{U}}_{j_1^*}\|_2 \|(\boldsymbol{O}_{j_1^* r}^\top \widehat{\boldsymbol{U}}_{j_1^*})^{-1}\|_2 \tag{26}$$

$$\leq \|\widehat{\boldsymbol{P}}_{j_1^*}^\top - \boldsymbol{P}_{j_1^*}^\top\|_2 \|(\boldsymbol{O}_{j_1^* r}^\top \widehat{\boldsymbol{U}}_{j_1^*})^{-1}\|_2 \leq \frac{\epsilon_r}{\sigma_{S_H}(\boldsymbol{O}_{j_1^* r}^\top \widehat{\boldsymbol{U}}_{j_1^*})} \tag{27}$$

The first inequality follows from the relationship between ℓ_∞ and ℓ_2 norm and submultiplicativity. The second follows from a matrix perturbation bound given in Lemma 91. We also use the fact that since $\widehat{\boldsymbol{U}}$ is orthonormal it has spectral norm 1.

Assuming that $\epsilon_{i,j} \leq \sigma_{S_H}(\boldsymbol{P}_{i,j})/3$ gives $\delta_{root} \leq \frac{2\epsilon_r}{\sqrt{3}\sigma_{S_H}(\boldsymbol{O}_{j_1^* r})}$ by Lemma 2.

6.2 δ^i_{one}

$$\begin{aligned} \delta^i_{one} &= \|(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})(\widehat{\mathbf{1}}_i - \tilde{\mathbf{1}}_i)\|_1 \leq \sqrt{S_H}\|\boldsymbol{O}\|_2\|\widehat{\boldsymbol{U}}\|_2\|\widehat{\mathbf{1}}_i - \tilde{\mathbf{1}}\|_2 && (28)\\ &= \sqrt{S_H}\|\widehat{\mathbf{1}}_i - \tilde{\mathbf{1}}_i\|_2 = \sqrt{S_H}\|\widehat{\mathbf{1}}_i - \tilde{\mathbf{1}}_i\|_2 && (29) \end{aligned}$$

Here we have converted ℓ_1 norm to ℓ_2 norm, used submultiplicativity, the fact that $\widehat{\boldsymbol{U}}$ is orthonormal so has spectral norm 1, and that $\boldsymbol{O}$ is a conditional probability matrix and therefore also has spectral norm 1.

We note that $\widehat{\mathbf{1}}_i = \widehat{\boldsymbol{P}}_{i,j}(\widehat{\boldsymbol{U}}^\top)^+\widehat{\boldsymbol{P}}_i$ and similarly $\tilde{\mathbf{1}}_i = (\boldsymbol{P}_{i,j}\widehat{\boldsymbol{U}}^\top)^+\boldsymbol{P}_i$, where i and j are a particular pair of observations described in the main paper.

$$\begin{aligned} \|\widehat{\mathbf{1}}_i - \tilde{\mathbf{1}}_i\|_2 &= \|(\widehat{\boldsymbol{P}}^T_{m,j}\widehat{\boldsymbol{U}})^+\widehat{\boldsymbol{P}}_j - (\boldsymbol{P}^T_{m,j}\widehat{\boldsymbol{U}}^\top)^+\boldsymbol{P}_j\|_2 && (30)\\ &= \|(\widehat{\boldsymbol{P}}^\top_{m,j}\widehat{\boldsymbol{U}})^+\widehat{\boldsymbol{P}}_j - (\boldsymbol{P}^\top_{m,j}\widehat{\boldsymbol{U}})^+\widehat{\boldsymbol{P}}_j + (\boldsymbol{P}^\top_{m,j}\widehat{\boldsymbol{U}})^+\widehat{\boldsymbol{P}}_j - (\boldsymbol{P}^\top_{m,j}\widehat{\boldsymbol{U}})^+\boldsymbol{P}_j\|_2 && (31)\\ &\leq \|(\widehat{\boldsymbol{P}}^\top_{m,j}\widehat{\boldsymbol{U}})^+\widehat{\boldsymbol{P}}_j - (\boldsymbol{P}^\top_{m,j}\widehat{\boldsymbol{U}})^+\widehat{\boldsymbol{P}}_j\|_2 + \|(\widehat{\boldsymbol{P}}^\top_{m,j}\widehat{\boldsymbol{U}})^+\widehat{\boldsymbol{P}}_j - (\boldsymbol{P}^\top_{m,j}\widehat{\boldsymbol{U}})^+\boldsymbol{P}_j\|_2 && (32)\\ &\leq \|(\widehat{\boldsymbol{P}}^\top_{m,j}\widehat{\boldsymbol{U}})^+ - (\boldsymbol{P}^\top_{m,j}\widehat{\boldsymbol{U}})^+\|_2\|\widehat{\boldsymbol{P}}_j\|_1 + \|(\widehat{\boldsymbol{P}}^\top_{m,j}\widehat{\boldsymbol{U}})^+ - (\boldsymbol{P}^\top_{m,j}\widehat{\boldsymbol{U}})^+\|_2\|\widehat{\boldsymbol{P}}_j - \boldsymbol{P}_j\|_2 && (33)\\ &\leq \frac{1+\sqrt{5}}{2} \times \frac{\epsilon_{m,j}}{\min(\sigma_{S_H}(\widehat{\boldsymbol{P}}_{m,j}), \sigma_{S_H}(\boldsymbol{P}^\top_{m,j}\widehat{\boldsymbol{U}}))^2} + \frac{\epsilon_j}{\sigma_{S_H}(\boldsymbol{P}^\top_{m,j}\widehat{\boldsymbol{U}})} && (34) \end{aligned}$$

where we have used the triangle inequality in the first inequality and the submultiplicative property of matrix norms in the second. The last inequality follows by matrix perturbation bounds. Thus using the assumption that $\epsilon_{i,j} \leq \sigma_{S_H}(\boldsymbol{P}_i, j)/3$, we get that

$$\delta_{one} \leq 4\sqrt{S_H}\left(\frac{\epsilon_{i,j}}{\sigma_{S_H}(\boldsymbol{P}_{i,j})^2} + \frac{\epsilon_i}{\sqrt{3}\sigma_{S_H}(\boldsymbol{P}_{i,j})}\right) \tag{35}$$

6.3 Tensor

Recall that $\tilde{\boldsymbol{\mathcal{T}}}_i = \boldsymbol{\mathcal{T}}_i \times_1 (\boldsymbol{O}_1^\top\widehat{\boldsymbol{U}}_1) \times_2 (\widehat{\boldsymbol{U}}_2^\top\boldsymbol{O}_2)^{-1} \times_3 (\boldsymbol{O}_3^\top\widehat{\boldsymbol{U}}_3) = \boldsymbol{P}_{m,j,k} \times_1 \widehat{\boldsymbol{U}}_1^\top \times_2 (\boldsymbol{P}_{l,j}\widehat{\boldsymbol{U}}_2)^+ \times_3 \widehat{\boldsymbol{U}}_3^\top$. Similarly, $\widehat{\boldsymbol{\mathcal{T}}}_i = \boldsymbol{P}_{m,j,k} \times_1 \widehat{\boldsymbol{U}}_1^\top \times_2 (\boldsymbol{P}_{l,j}\widehat{\boldsymbol{U}}_2)^+ \times_3 \widehat{\boldsymbol{U}}_3^\top$.

$$\|(\widehat{\boldsymbol{\mathcal{T}}}_i - \tilde{\boldsymbol{\mathcal{T}}}_i) \times_1 (\boldsymbol{O}_1^\top\widehat{\boldsymbol{U}}_1)^{-1} \times_2 (\boldsymbol{O}_2^\top\widehat{\boldsymbol{U}}_2) \times_3 (\boldsymbol{O}_3^\top\widehat{\boldsymbol{U}}_3)^{-1}\|_{1,1} \leq \frac{\sqrt{S_H}}{\sigma_{S_H}(\boldsymbol{O})}\|\widehat{\boldsymbol{\mathcal{T}}}_i - \tilde{\boldsymbol{\mathcal{T}}}_i\|_2 \tag{36}$$

This is because both $\widehat{\boldsymbol{U}}$ and $\boldsymbol{O}$ have spectral norm one and the $\sqrt{S_H}$ factor is the cost of converting from 1 norm to spectral norm.

$$\begin{aligned} \|\widehat{\boldsymbol{\mathcal{T}}}_i - \tilde{\boldsymbol{\mathcal{T}}}_i\|_2 &= \|\widehat{\boldsymbol{P}}_{m,j,k} \times_1 \widehat{\boldsymbol{U}}_1^\top \times_2 (\widehat{\boldsymbol{P}}_{l,j}\widehat{\boldsymbol{U}}_2)^+ \times_3 \widehat{\boldsymbol{U}}_3^\top - \boldsymbol{P}_{m,j,k} \times_1 \widehat{\boldsymbol{U}}_1^\top \times_2 (\boldsymbol{P}_{l,j}\widehat{\boldsymbol{U}}_2)^+ \times_3 \widehat{\boldsymbol{U}}_3^\top\|_2 && (37)\\ &= \|\widehat{\boldsymbol{P}}_{m,j,k} \times_1 \widehat{\boldsymbol{U}}_1^\top \times_2 (\widehat{\boldsymbol{P}}_{i,k}\widehat{\boldsymbol{U}}_2)^+ \times_3 \widehat{\boldsymbol{U}}_3^\top - \widehat{\boldsymbol{P}}_{m,j,k} \times_1 \widehat{\boldsymbol{U}}_1^\top \times_2 (\boldsymbol{P}_{l,j}\widehat{\boldsymbol{U}}_3)^+ \times_3 \widehat{\boldsymbol{U}}_3^\top\|_2 && (38)\\ &\quad + \|\widehat{\boldsymbol{P}}_{m,j,k} \times_1 \widehat{\boldsymbol{U}}_1^\top \times_2 (\boldsymbol{P}_{l,j}\widehat{\boldsymbol{U}}_2)^+ \times_3 \widehat{\boldsymbol{U}}_3^\top - \boldsymbol{P}_{m,j,k} \times_1 \widehat{\boldsymbol{U}}_1^\top \times_2 (\boldsymbol{P}_{l,j}\widehat{\boldsymbol{U}}_2)^+ \times_3 \widehat{\boldsymbol{U}}_3^\top\|_2 && (39)\\ &= \|\widehat{\boldsymbol{P}}_{m,j,k} \times_1 \widehat{\boldsymbol{U}}_1^\top \times_2 ((\widehat{\boldsymbol{P}}_{l,j}\widehat{\boldsymbol{U}}_2)^+ - (\boldsymbol{P}_{l,j}\widehat{\boldsymbol{U}}_2)^+) \times_3 \widehat{\boldsymbol{U}}_3^\top\|_2 && (40)\\ &\quad + \|(\widehat{\boldsymbol{P}}_{i,j,k} \times_1 \widehat{\boldsymbol{U}}_1^\top \times_3 \widehat{\boldsymbol{U}}_3^\top - \boldsymbol{P}_{i,j,k} \times_1 \widehat{\boldsymbol{U}}_1^\top \times_3 \widehat{\boldsymbol{U}}_3^\top) \times_2 (\boldsymbol{P}_{l,j}\widehat{\boldsymbol{U}}_2)^+\|_2 && (41)\\ &= \|\widehat{\boldsymbol{P}}_{m,j,k}\|_2 \frac{1+\sqrt{5}}{2} \frac{\epsilon_{l,j}}{\min(\sigma_{S_H}(\widehat{\boldsymbol{P}}_{l,j}), \sigma_{S_H}(\boldsymbol{P}_{l,j}\widehat{\boldsymbol{U}}))^2} + \frac{\epsilon_{m,j,k}}{\sigma_{S_H}(\boldsymbol{P}_{l,j}\widehat{\boldsymbol{U}})} && (42) \end{aligned}$$

It is clear that $\|\widehat{\boldsymbol{P}}_{m,j,k}\|_2 \leq \|\widehat{\boldsymbol{P}}_{m,j,k}\|_F \leq 1$.

Using the fact that $\epsilon_{i,j} \leq \sigma_{S_H}(\boldsymbol{P}_{i,j})/3$ gives us the following bound:

$$\gamma_v \leq \frac{4\sqrt{S_H}}{\sigma_{S_H}(\boldsymbol{O})}\left(\frac{\epsilon_{i,j}}{\sigma_{S_H}(\boldsymbol{P}_{i,j})} + \frac{\epsilon_{i,j,k}}{\sqrt{3}\sigma_{S_H}(\boldsymbol{P}_{i,j})}\right) \tag{43}$$

6.4 Bounding $\triangle_i$

We now seek to bound $\triangle_i = \sum_{\boldsymbol{x}_i} \|(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)(\widehat{\boldsymbol{M}}_i - \tilde{\boldsymbol{M}}_i)(\widehat{\boldsymbol{U}}_2^\top \boldsymbol{O}_2)^{-1}\|_1$. There are two cases: either i is a leaf or it is not.

6.4.1 i is leaf node

In this case our proof simply follows from HKZ [3] and is repeated here for convenience.

$$\begin{aligned}
\|(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)(\widehat{\boldsymbol{M}} - \tilde{\boldsymbol{M}})(\boldsymbol{O}_2^\top \widehat{\boldsymbol{U}}_2)^{-1}\|_1 &\leq \sqrt{S_H}\|\boldsymbol{O}_1\|_1\|(\widehat{\boldsymbol{M}}_i - \tilde{\boldsymbol{M}}_i)(\boldsymbol{O}_2^\top \widehat{\boldsymbol{U}}_2)^{-1}\|_2 && (44)\\
&\leq \sqrt{S_H}\frac{\|\widehat{\boldsymbol{M}}_i - \tilde{\boldsymbol{M}}_i\|_2}{\sigma_{S_H}(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})} && (45)
\end{aligned}$$

Note that $\widehat{\boldsymbol{M}}_i = (\widehat{\boldsymbol{P}}_{j,i}\widehat{\boldsymbol{U}}_1)^{-1}\widehat{\boldsymbol{P}}_{m,x_i,j}\widehat{\boldsymbol{U}}_2$ and $\tilde{\boldsymbol{M}}_i = (\boldsymbol{P}_{j,i}\widehat{\boldsymbol{U}}_1)^{-1}\boldsymbol{P}_{m,x_i,j}\widehat{\boldsymbol{U}}_2$.

$$\begin{aligned}
\|\widehat{\boldsymbol{M}}_i - \tilde{\boldsymbol{M}}_i\|_2 &= \|(\widehat{\boldsymbol{P}}_{j,i}\widehat{\boldsymbol{U}}_1)^{-1}\widehat{\boldsymbol{P}}_{m,x_i,j}\widehat{\boldsymbol{U}}_2 - (\boldsymbol{P}_{j,i}\widehat{\boldsymbol{U}}_1)^{-1}\boldsymbol{P}_{m,x_i,j}\widehat{\boldsymbol{U}}_2\|_2 && (46)\\
&= \|(\widehat{\boldsymbol{P}}_{j,i}\widehat{\boldsymbol{U}}_1)^{-1}\widehat{\boldsymbol{P}}_{m,x_i,j}\widehat{\boldsymbol{U}}_2 - (\boldsymbol{P}_{j,i}\widehat{\boldsymbol{U}}_1)^{-1}\widehat{\boldsymbol{P}}_{m,x_i,j}\widehat{\boldsymbol{U}}_2 + (\boldsymbol{P}_{j,i}\widehat{\boldsymbol{U}}_1)^{-1}\widehat{\boldsymbol{P}}_{m,x_i,j}\widehat{\boldsymbol{U}}_2 - \widehat{\boldsymbol{U}}_1^\top \boldsymbol{P}_{x,i,j}(\widehat{\boldsymbol{U}}_2^\top \boldsymbol{P}_{i,j})^{-1}\|_2 && (47)\\
&\leq \|((\widehat{\boldsymbol{P}}_{j,i}\widehat{\boldsymbol{U}}_1)^{-1} - (\boldsymbol{P}_{j,i}\widehat{\boldsymbol{U}}_1)^{-1})\widehat{\boldsymbol{P}}_{m,x_i,j}\widehat{\boldsymbol{U}}_2\|_2 + \|(\boldsymbol{P}_{j,i}\widehat{\boldsymbol{U}}_1)^{-1}(\widehat{\boldsymbol{P}}_{m,x_i,j}\widehat{\boldsymbol{U}}_2 - \boldsymbol{P}_{m,x_i,j}\widehat{\boldsymbol{U}}_2)\|_2 && (48)\\
&\leq \|\widehat{\boldsymbol{P}}_{m,x_i,j}\|_2\frac{1+\sqrt{5}}{2}\frac{\epsilon_{j,i}}{\min(\sigma_{S_H}(\widehat{\boldsymbol{P}}_{j,i}), \sigma_{S_H}(\boldsymbol{P}_{j,i}\widehat{\boldsymbol{U}})} + \frac{\epsilon_{m,x_i,j}}{\sigma_{S_H}(\boldsymbol{P}_{j,i}\widehat{\boldsymbol{U}}} && (49)\\
&\leq \mathbb{P}[x_i = x]\frac{1+\sqrt{5}}{2}\frac{\epsilon_{j,i}}{\min(\sigma_{S_H}(\widehat{\boldsymbol{P}}_{j,i}), \sigma_{S_H}(\boldsymbol{P}_{j,i}\widehat{\boldsymbol{U}}))^2} + \frac{\epsilon_{m,x_i,j}}{\sigma_{S_H}(\boldsymbol{P}_{j,i}\widehat{\boldsymbol{U}})} && (50)
\end{aligned}$$

where the first inequality follows from the triangle inequality, and the second uses matrix perturbation bounds (and the fact that spectral norm of $\widehat{\boldsymbol{U}}$ is 1).

The final inequality follows from the fact that spectral norm is less than frobenius norm which is less than l1 norm:

$$\|\widehat{\boldsymbol{P}}_{m,x_i,j}\| \leq \sqrt{\sum_{m,j}[\widehat{\boldsymbol{P}}_{m,x_i,j}]^2_{m,j}} \leq \sum_{m,j}[\boldsymbol{P}_{m,x_i,j}]_{m,j} \leq \mathbb{P}[x_i = x] \qquad (51)$$

The first inequality follows from relation between 1 operator norm and 2 operator norm. Because $\boldsymbol{O}$ is a conditional probability matrix $\|\boldsymbol{O}\|_1 = 1$ (i.e. the max column sum is 1).

Using the fact that $\epsilon_{i,j} \leq \sigma_{S_H}(\boldsymbol{P}_{i,j})/3$ gives us the following bound:

$$\triangle_{i,x} \leq 4\frac{\sqrt{S_H}}{\sigma_{S_H}(\boldsymbol{O})}\left(\mathbb{P}[x_i = x]\frac{\epsilon_{m,x_i,j}}{(\sigma_{S_H}(\boldsymbol{P}_{j,i}))^2} + \frac{\epsilon_{m,x_i,j}}{\sqrt{3}\sigma_{S_H}(\boldsymbol{P}_{j,i})}\right) \qquad (52)$$

Summing over v would give

$$\triangle_i \leq 4\frac{\sqrt{S_H}}{\sigma_{S_H}(\boldsymbol{O})}\left(\frac{\epsilon_{j,i}}{(\sigma_{S_H}(\boldsymbol{P}_{j,i}))^2} + \frac{\sum_{x_i}\epsilon_{m,x_i,j}}{\sqrt{3}\sigma_{S_H}(\boldsymbol{P}_{j,i})}\right) \qquad (53)$$

6.4.2 i is not a leaf node

Let $\widehat{\boldsymbol{m}}_{J:1} = \widehat{\boldsymbol{M}}_J ... \widehat{\boldsymbol{M}}_1 \widehat{\mathbf{1}}_i$ and $\tilde{\boldsymbol{m}}_{J:1} = \tilde{\boldsymbol{M}}_J ... \tilde{\boldsymbol{M}}_1 \tilde{\mathbf{1}}_i$

$$\begin{aligned}
& \sum_{\mathbf{x}_i} \|(\boldsymbol{O}_2^\top \widehat{\boldsymbol{U}}_2)(\hat{\boldsymbol{M}}_i - \tilde{\boldsymbol{M}}_i)(\boldsymbol{O}_3^\top \widehat{\boldsymbol{U}}_3)^{-1}\|_1 && (54)\\
= & \sum_{\mathbf{x}_i} \|(\boldsymbol{O}_2^\top \widehat{\boldsymbol{U}}_2)(\widehat{\boldsymbol{\mathcal{T}}}_i \times_1 \widehat{\boldsymbol{M}}_u ... \widehat{\boldsymbol{M}}_1 \widehat{\mathbf{1}}_i - \tilde{\boldsymbol{\mathcal{T}}}_i \times_1 \tilde{\boldsymbol{M}}_u ... \tilde{\boldsymbol{M}}_1 \tilde{\mathbf{1}}_i)(\boldsymbol{O}_3^\top \widehat{\boldsymbol{U}}_3)^{-1}\|_1 && (55)\\
= & \sum_{\mathbf{x}_i} \|(\boldsymbol{O}_2^\top \widehat{\boldsymbol{U}}_2)(\widehat{\boldsymbol{\mathcal{T}}}_i \ \bar{\times}_1 \ \widehat{\boldsymbol{m}}_{J:1} - \tilde{\boldsymbol{\mathcal{T}}}_i \ \bar{\times}_1 \ \tilde{\boldsymbol{m}}_{J:1})(\boldsymbol{O}_3^\top \widehat{\boldsymbol{U}}_3)^{-1}\|_1 && (56)\\
= & \sum_{\mathbf{x}_i} \|(\boldsymbol{O}^\top \widehat{\boldsymbol{U}}) \Big((\widehat{\boldsymbol{\mathcal{T}}}_i - \tilde{\boldsymbol{\mathcal{T}}}_i) \ \bar{\times}_1 \ \tilde{\boldsymbol{m}}_{J:1} + (\widehat{\boldsymbol{\mathcal{T}}}_i - \tilde{\boldsymbol{\mathcal{T}}}_i) \ \bar{\times}_1 \ (\widehat{\boldsymbol{m}}_{J:1} - \tilde{\boldsymbol{m}}_{J:1}) + \tilde{\boldsymbol{\mathcal{T}}}_i \ \bar{\times}_1 \ (\widehat{\boldsymbol{m}}_{J:1} - \tilde{\boldsymbol{m}}_{J:1})\Big) (\boldsymbol{O}_3^\top \widehat{\boldsymbol{U}}_3)^{-1}\|_1 && (57)\\
\leq & \sum_{\mathbf{x}_i} \|(\widehat{\boldsymbol{\mathcal{T}}}_i - \tilde{\boldsymbol{\mathcal{T}}}_i) \times_1 (\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)^{-1} \times_2 (\widehat{\boldsymbol{U}}_2^\top \boldsymbol{O}_2) \times_3 (\boldsymbol{O}_3^\top \widehat{\boldsymbol{U}}_3)^{-1}\|_{1,1} \left\|(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1) \tilde{\boldsymbol{m}}_{J:1}\right\|_1 && (58)\\
+ & \sum_{\mathbf{x}_i} \|(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)(\widehat{\boldsymbol{m}}_{J:1} - \tilde{\boldsymbol{m}}_{J:1})\|_1 \|(\widehat{\boldsymbol{\mathcal{T}}}_i - \tilde{\boldsymbol{\mathcal{T}}}_i) \times_1 (\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)^{-1} \times_2 (\widehat{\boldsymbol{U}}_2^\top \boldsymbol{O}_2) \times_3 (\boldsymbol{O}_3^\top \widehat{\boldsymbol{U}}_3)^{-1}\|_{1,1} && (59)\\
+ & \sum_{\mathbf{x}_i} \|\tilde{\boldsymbol{\mathcal{T}}}_i \times_1 (\boldsymbol{O}^\top \widehat{\boldsymbol{U}}_1)^{-1} \times_2 (\widehat{\boldsymbol{U}}_2^\top \boldsymbol{O}_2) \times_3 (\boldsymbol{O}_3^\top \widehat{\boldsymbol{U}}_3)^{-1}\|_{1,1} \left\|(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)(\widehat{\boldsymbol{m}}_{J:1} - \tilde{\boldsymbol{m}}_{J:1})\right\|_1 && (60)
\end{aligned}$$

First term is bounded by:

$$\left\|(\widehat{\boldsymbol{\mathcal{T}}}_i - \tilde{\boldsymbol{\mathcal{T}}}_i) \times_1 (\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)^{-1} \times_2 (\widehat{\boldsymbol{U}}_2^\top \boldsymbol{O}_2) \times_3 (\boldsymbol{O}_3^\top \widehat{\boldsymbol{U}}_3)^{-1}\right\|_{1,1} S_H \leq S_H \gamma_i \tag{61}$$

Second term is bounded by:

$$\begin{aligned}
& \sum_{\mathbf{x}_i} \|(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)(\widehat{\boldsymbol{m}}_{J:1} - \tilde{\boldsymbol{m}}_{J:1})\|_1 \|(\widehat{\boldsymbol{\mathcal{T}}}_i - \tilde{\boldsymbol{\mathcal{T}}}_i) \times_1 (\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)^{-1} \times_2 (\widehat{\boldsymbol{U}}_2^\top \boldsymbol{O}_2) \times_3 (\boldsymbol{O}_3^\top \widehat{\boldsymbol{U}}_3)^{-1}\|_{1,1} && (62)\\
& \leq \gamma_i \sum_{\mathbf{x}_i} \|(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)(\widehat{\boldsymbol{m}}_{J:1} - \tilde{\boldsymbol{m}}_{J:1})\|_1 && (63)
\end{aligned}$$

Third Term is bounded by:

$$\|\tilde{\boldsymbol{\mathcal{T}}}_i \times_1 (\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)^{-1} \times_2 (\widehat{\boldsymbol{U}}_2^\top \boldsymbol{O}_2) \times_3 (\boldsymbol{O}_3^\top \widehat{\boldsymbol{U}}_3)^{-1}\|_{1,1} \sum_{\mathbf{x}_i} \left\|(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)(\widehat{\boldsymbol{m}}_{J:1} - \tilde{\boldsymbol{m}}_{J:1})\right\|_1 \leq \sum_{\mathbf{x}_i} \left\|(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)(\widehat{\boldsymbol{m}}_{J:1} - \tilde{\boldsymbol{m}}_{J:1})\right\|_1 \tag{64}$$

In the next section, we will see that $\sum_{\mathbf{x}_i} \|(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})(\widehat{\boldsymbol{m}}_{J:1} - \tilde{\boldsymbol{m}}_{J:1})\|_1 \leq \Big(\prod_{k=1}^J (1 + \Delta_{j_k}) \delta_{one}^i + S_H \prod_{k=1}^J (1 + \triangle_{j_k}) - S_H\Big)$. So the overall bound is

$$\triangle_i \leq \left((1 + \gamma_i) \prod_{k=1}^J (1 + \Delta_{j_k}) \delta_{one}^i + (1 + \gamma_i) S_H \prod_{k=1}^J (1 + \Delta_{j_k}) - S_H\right). \tag{65}$$

(where $j_1, ..., j_J$ are children of node i).

6.5 Bounding $\sum_{\mathbf{x}_i} \|(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})(\widehat{\boldsymbol{m}}_{J:1} - \tilde{\boldsymbol{m}}_{J:1})\|_1$

Lemma 4

$$\sum_{\mathbf{x}_i} \|(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})(\widehat{\boldsymbol{m}}_{J:1} - \tilde{\boldsymbol{m}}_{J:1})\|_1 \leq \prod_{k=1}^J (1 + \triangle_{j_k}) \delta_{one}^i + S_H \prod_{k=1}^J (1 + \triangle_{j_k}) - S_H \tag{66}$$

(where $j_1, ..., j_J$ are children of node i).

The proof is by induction. Base case: $\|(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})(\widehat{\mathbf{1}}_i - \tilde{\mathbf{1}}_i)\|_1 \leq \delta_{one}^i$, by definition of δ_{one}^i.

Inductive step: Let us say claim holds up until $u - 1$. We show it holds for u. Thus

$$\sum_{\mathbf{x}_i} \|(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})(\widehat{\boldsymbol{m}}_{(u-1):1} - \tilde{\boldsymbol{m}}_{(u-1):1})\|_1 \leq \prod_{k=1}^{u-1} (1 + \triangle_{j_k}) \delta_{one}^i + S_H \prod_{k=1}^{u-1} (1 + \triangle_{j_k}) - S_H \tag{67}$$

We now decompose the sum over x as

$$\begin{aligned}
&\sum_{\mathbf{x}_{u:1}} \|(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})(\widehat{\boldsymbol{m}}_{u:1} - \tilde{\boldsymbol{m}}_{u:1})\|_1 && (68)\\
=\;& \sum_{\mathbf{x}_{u:1}} \|(\boldsymbol{O}^\top \widehat{\boldsymbol{U}}) \left((\widehat{\boldsymbol{M}}_u - \tilde{\boldsymbol{M}}_u)\tilde{\boldsymbol{m}}_{(u-1):1} + (\widehat{\boldsymbol{M}}_u - \tilde{\boldsymbol{M}}_u)(\widehat{\boldsymbol{m}}_{(u-1):1} - \tilde{\boldsymbol{m}}_{(u-1):1}) + (\widehat{\boldsymbol{m}}_{(u-1):1} - \tilde{\boldsymbol{m}}_{(u-1):1}) \right) \|_1
\end{aligned}$$

Using the triangle inequality, we get

$$\begin{aligned}
& \sum_{\mathbf{x}_{u:1}} \|(\boldsymbol{O}_2^\top \widehat{\boldsymbol{U}}_2)(\widehat{\boldsymbol{M}}_u - \tilde{\boldsymbol{M}}_u)(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)^{-1}\|_1 \|(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)\tilde{\boldsymbol{m}}_{(u-1):1}\|_1 && (69)\\
+\;& \sum_{\mathbf{x}_{u:1}} \|(\boldsymbol{O}_2^\top \widehat{\boldsymbol{U}}_2)(\widehat{\boldsymbol{M}}_u - \tilde{\boldsymbol{M}}_u)(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)^{-1}\|_1 \|(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)(\widehat{\boldsymbol{m}}_{(u-1):1} - \tilde{\boldsymbol{m}}_{(u-1):1})\|_1 && (70)\\
+\;& \sum_{\mathbf{x}_{u:1}} \|(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})\tilde{\boldsymbol{M}}_u(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})^{-1}\|_1 \|(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})(\widehat{\boldsymbol{m}}_{(u-1):1} - \tilde{\boldsymbol{m}}_{(u-1):1})\|_1 && (71)
\end{aligned}$$

Again we are just numbering the $\boldsymbol{U}$'s and $\boldsymbol{O}$'s for clarity to see which corresponds with which. They are omitted in the actual theorem statements since we will take minimums etc. at the end.

We now must bound these terms. First term:

$$\sum_{\mathbf{x}_u} \|(\boldsymbol{O}_2^\top \widehat{\boldsymbol{U}}_2)(\widehat{\boldsymbol{M}}_u - \tilde{\boldsymbol{M}}_u)(\boldsymbol{O}_2^\top \widehat{\boldsymbol{U}}_2)^{-1}\|_1 \sum_{x_{1:u-1}} \|(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)\tilde{\boldsymbol{m}}_{(u-1):1}\|_1 \leq \triangle_u \sum_{\mathbf{x}_{(u-1):1}} \|\tilde{\boldsymbol{m}}_{(u-1):1}(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})\|_1 \leq S_H \triangle_u \qquad (72)$$

since $\triangle_u = \|(\boldsymbol{O}_2^\top \widehat{\boldsymbol{U}}_2)(\widehat{\boldsymbol{M}}_u - \tilde{\boldsymbol{M}}_u)(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)^{-1}\|_1$. Second term can be bounded by inductive hypothesis:

$$\sum_{\mathbf{x}_{u:1}} \|(\boldsymbol{O}_2^\top \widehat{\boldsymbol{U}}_2)(\widehat{\boldsymbol{M}}_u - \tilde{\boldsymbol{M}}_u)(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)^{-1}\|_1 \|(\boldsymbol{O}_1^\top \widehat{\boldsymbol{U}}_1)(\widehat{\boldsymbol{m}}_{(u-1):1} - \tilde{\boldsymbol{m}}_{(u-1):1})\|_1 \leq \triangle_u \left(\prod_{k=1}^{u-1} (1 + \triangle_{j_k})\delta^i_{one} + S_H \prod_{k=1}^{u-1} (1 + \triangle_{j_k}) - S_H \right) \qquad (73)$$

The third term is bounded by observing that $(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})\tilde{\boldsymbol{M}}_u(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})^{-1} = \mathrm{diag}(Pr[\mathbf{x}_u|\mathrm{Parent}])$. Thus it is diagonal, and $Pr[\mathbf{x}|Parent]$ has max row or column sum as 1. This means that the third term is bounded by the inductive hypothesis as well:

$$\sum_{\mathbf{x}_{u:1}} \|(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})\tilde{\boldsymbol{M}}_u(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})^{-1}\|_1 \|(\boldsymbol{O}^\top \widehat{\boldsymbol{U}})(\widehat{\boldsymbol{m}}_{(u-1):1} - \tilde{\boldsymbol{m}}_{(u-1):1})\|_1 \leq \left(\prod_{k=1}^{u-1} (1 + \triangle_{j_k})\delta^i_{one} + S_H \prod_{k=1}^{u-1} (1 + \triangle_{j_k}) - S_H \right) \qquad (74)$$

7 Bounding the propagation of error in tree

We now wrap up the proof based on the approach of HKZ[3].

Lemma 5

$$\sum_{x_1,\ldots,x_O} \left| \widehat{\mathbb{P}}[x_1,\ldots,x_O] - \mathbb{P}[x_1,\ldots,x_O] \right| \leq S_H \delta_{root} + (1 + \delta_{root}) \left(\prod_{k=1}^{J} (1 + \triangle_{j_k})\delta^r_{one} + S_H \prod_{k=1}^{J} (1 + \triangle_{j_k}) - S_H \right) \qquad (75)$$

$$\begin{aligned}
\sum_{x_1,\ldots,x_O} \left| \widehat{\mathbb{P}}[x_1,\ldots,x_O] - \mathbb{P}[x_1,\ldots,x_O] \right| \;&=\; \sum_{x_1,\ldots,x_O} \left| \widehat{\boldsymbol{r}}^\top \widehat{\boldsymbol{M}}_{j_1} ... \widehat{\boldsymbol{M}}_{j_J} \widehat{\mathbf{1}}_r - \tilde{\boldsymbol{r}}^\top \tilde{\boldsymbol{M}}_{j_1} ... \tilde{\boldsymbol{M}}_{j_J} \tilde{\mathbf{1}}_r \right| && (76)\\
&\leq\; \sum_{x_1,\ldots,x_O} \left| (\widehat{\boldsymbol{r}} - \tilde{\boldsymbol{r}})^\top (\boldsymbol{O}^\top \widehat{\boldsymbol{U}})^{-1} (\boldsymbol{O}^\top \widehat{\boldsymbol{U}}) (\tilde{\boldsymbol{M}}_{J:1} \tilde{\mathbf{1}}) \right| && (77)\\
&+\; \sum_{x_1,\ldots,x_O} \left| (\widehat{\boldsymbol{r}} - \tilde{\boldsymbol{r}})^\top (\boldsymbol{O}^\top \widehat{\boldsymbol{U}})^{-1} (\boldsymbol{O}^\top \widehat{\boldsymbol{U}}) (\widehat{\boldsymbol{M}}_{J:1} \widehat{\mathbf{1}}_r - \tilde{\boldsymbol{M}}_{J:1} \tilde{\mathbf{1}}_r) \right| && (78)\\
&+\; \sum_{x_1,\ldots,x_O} \left| \tilde{\boldsymbol{r}}^\top (\boldsymbol{O}^\top \widehat{\boldsymbol{U}})^{-1} (\boldsymbol{O}^\top \widehat{\boldsymbol{U}}) (\widehat{\boldsymbol{M}}_{J:1} \widehat{\mathbf{1}}_i - \tilde{\boldsymbol{M}}_{J:1} \tilde{\mathbf{1}}) \right| && (79)
\end{aligned}$$

The first sum is bounded using Holder inequality and noting that the first term is a conditional probability (of all observed variables conditioned on the root)

$$\begin{aligned}
& \sum_{x_1,\ldots,x_O} \left| (\widehat{\boldsymbol{r}} - \tilde{\boldsymbol{r}})^\top (\boldsymbol{O}^\top \widehat{\boldsymbol{U}})^{-1} (\boldsymbol{O}^\top \widehat{\boldsymbol{U}}) (\tilde{\boldsymbol{M}}_{J:1} \tilde{\mathbf{1}}) \right| && (80)\\
\leq\;& \sum_{x_1,\ldots,x_O} \|(\widehat{\boldsymbol{r}} - \tilde{\boldsymbol{r}})^\top (\boldsymbol{O}^\top \widehat{\boldsymbol{U}})^{-1}\|_\infty \|(\boldsymbol{O}^\top \widehat{\boldsymbol{U}}) (\tilde{\boldsymbol{M}}_{J:1} \tilde{\mathbf{1}})\|_1 \leq S_H \delta_{root} && (81)
\end{aligned}$$

The second sum is bounded by another application of Holder's inequality (and the previous lemma):

$$\begin{aligned} & \sum_{x_1,\ldots,x_O} \left| (\widehat{\boldsymbol{r}} - \tilde{\boldsymbol{r}})^\top (\boldsymbol{O}^\top \widehat{\boldsymbol{U}})^{-1} (\boldsymbol{O}^\top \widehat{\boldsymbol{U}}) (\widehat{\boldsymbol{M}}_{J:1} \widehat{\mathbf{1}}_r - \tilde{\boldsymbol{M}}_{J:1} \tilde{\mathbf{1}}_r) \right| && (82) \\ \leq & \sum_{x_1,\ldots,x_O} \| (\widehat{\boldsymbol{r}} - \tilde{\boldsymbol{r}})^\top (\boldsymbol{O}^\top \widehat{\boldsymbol{U}})^{-1} \|_\infty \| (\boldsymbol{O}^\top \widehat{\boldsymbol{U}}) (\widehat{\boldsymbol{M}}_{J:1} \widehat{\mathbf{1}}_r - \tilde{\boldsymbol{M}}_{J:1} \tilde{\mathbf{1}}_r) \|_1 && (83) \\ \leq & \ \delta_{root} \left(\prod_{k=1}^{J} (1 + \triangle_{j_k}) \delta^r_{one} + S_H \prod_{k=1}^{J} (1 + \triangle_{j_k}) - S_H \right) && (84) \end{aligned}$$

The third sum is also bounded by Holder's Inequality and previous lemmas and noting that $\tilde{\boldsymbol{r}}^\top (U^\top \boldsymbol{O})^{-1} = \mathbb{P}[R = r]$:

$$\begin{aligned} & \sum_{x_1,\ldots,x_O} \left| \tilde{\boldsymbol{r}}^\top (\boldsymbol{O}^\top \widehat{\boldsymbol{U}})^{-1} (\boldsymbol{O}^\top \widehat{\boldsymbol{U}}) (\widehat{\boldsymbol{M}}_{J:1} \widehat{\mathbf{1}}_i - \tilde{\boldsymbol{M}}_{J:1} \tilde{\mathbf{1}}) \right| && (85) \\ \leq & \sum_{x_1,\ldots,x_O} \| \tilde{\boldsymbol{r}}^\top (\boldsymbol{O}^\top \widehat{\boldsymbol{U}})^{-1} \|_\infty \| (\boldsymbol{O}^\top \widehat{\boldsymbol{U}}) (\widehat{\boldsymbol{M}}_{J:1} \widehat{\mathbf{1}}_r - \tilde{\boldsymbol{M}}_{J:1} \tilde{\mathbf{1}}_r) \|_1 && (86) \\ \leq & \left(\prod_{k=1}^{J} (1 + \triangle_{j_k}) \delta^r_{one} + S_H \prod_{k=1}^{J} (1 + \triangle_{j_k}) - S_H \right) && (87) \end{aligned}$$

Combining these bounds gives us the desired solution.

8 Putting it all together

We seek for

$$\sum_{x_1,\ldots,x_O} \left| \widehat{\mathbb{P}}[x_1, \ldots, x_O] - \mathbb{P}[x_1, \ldots, x_O] \right| \leq \epsilon \tag{88}$$

Using the fact that for $a < .5$, $(1 + a/t)^\top \leq 1 + 2a$, we get that $\triangle_{j_k} \leq O(\epsilon/(S_H J))$. However, $\triangle_j$ is defined recursively, and thus the error accumulates exponential in the longest path of hidden nodes. For example, $\triangle^{obs}_i \leq O(\frac{\epsilon}{(d_{max} S_H)^\ell})$ where ℓ is the longest path of hidden nodes. Tracing this back through will gives the result:

Pick any $\epsilon > 0, \delta < 1$. Let

$$N \geq O\left(\frac{1}{\epsilon^2} \left(\frac{(d_{max} S_H)^{2\ell+1} S_O}{\min_i \sigma_{S_H}(\boldsymbol{O}_i)^2 \min_{i \neq j} \sigma_{S_H}(P_{i,j})^4} \right) \right) \log \frac{\mathscr{O}}{\delta} \tag{89}$$

Then with probability $1 - \delta$

$$\sum_{x_1,\ldots,x_O} \left| \widehat{\mathbb{P}}[x_1, \ldots, x_O] - \mathbb{P}[x_1, \ldots, x_O] \right| \leq \epsilon \tag{90}$$

In many cases, if the frequency of the observation symbols follow certain distributions, than the dependence on S_O can be removed as showed in HKZ [3].

9 Appendix

9.1 Matrix Perturbation Bounds

This is Theorem 3.8 from pg. 143 in Stewart and Sun, 1990 [6]. Let $A \in \boldsymbol{R}^{m \times n}$, with $m \geq n$ and let $\tilde{A} = A + E$. Then

$$\|\tilde{A}^+ - A^+\|_2 \leq \frac{1 + \sqrt{5}}{2} \max(\|A^+\|_2^2, \|\tilde{A}\|_2^2) \|E\|_2 \tag{91}$$

References

[1] Myung J. Choi, Vicent Y. Tan, Animashree Anandkumar, and Alan S. Willsky. Learning latent tree graphical models. In *arXiv:1009.2722v1*, 2010.

[2] R.A. Horn and C.R. Johnson. *Matrix analysis*. Cambridge Univ Pr, 1990.

[3] D. Hsu, S. Kakade, and T. Zhang. A spectral algorithm for learning hidden markov models. In *Proc. Annual Conf. Computational Learning Theory*, 2009.

[4] N.H. Nguyen, P. Drineas, and T.D. Tran. Tensor sparsification via a bound on the spectral norm of random tensors. *Arxiv preprint arXiv:1005.4732*, 2010.

[5] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[6] G.W. Stewart and J. Sun. *Matrix perturbation theory*, volume 175. Academic press New York, 1990.