A Spectral Algorithm For Latent Junction Trees - Supplementary Material

Ankur P. Parikh, Le Song, Mariya Ishteva, Gabi Teodoru, Eric P. Xing

1 Discussion of Conditions for Observable Representation

The observable representation exists only if there exist transformations $\boldsymbol{\mathcal{F}}_i = \mathbb{P}[\mathscr{O}_i|S_i]$ with rank $\tau_i := k_h \times |S_i|$ and $\mathbb{P}[\mathscr{O}_{-i}|S_i]$ also has rank τ_i (so that $\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{i-})$ has rank τ_i). Thus, it is required that $\#\text{states}(\mathscr{O}_i) \geq \#\text{states}(S_i)$. This can either be achieved by either making $\mathscr{O}_i$ consist of a few high dimensional observations, or many smaller dimensional ones. In the case when $\#\text{states}(\mathscr{O}_i) > \#\text{states}(S_i)$, we need to project $\boldsymbol{\mathcal{F}}_i$ to a lower dimensional space such that it can be inverted using a tensor $\boldsymbol{\mathcal{U}}_i$. In this case, we define $\boldsymbol{\mathcal{F}}_i := \mathbb{P}[\mathscr{O}_i|S_i] \times_{\mathscr{O}_i} \boldsymbol{\mathcal{U}}_i]$. Following this through the computation gives us that $\widetilde{\boldsymbol{\mathcal{P}}}(C_l) = \mathbb{P}(\mathscr{O}_l, \mathscr{O}_{l-}) \times_{\mathscr{O}_{l-}} (\mathbb{P}(\mathscr{O}_l, \mathscr{O}_{l-} \times_{\mathscr{O}_i} \boldsymbol{\mathcal{U}}_i)^{-1}$. A good choice of $\mathcal{U}_i$ can be obtained by performing a singular value decomposition of the matricized version of $\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{i-})$ (variables in $\mathscr{O}_i$ are arranged to rows and those in $\mathscr{O}_{i-}$ to columns).

For HMMs and latent trees, this rank condition can be expressed simply as requiring the conditional probability tables of the underlying model to not be rank-deficient. However, junction trees encode significantly more complex latent structures that introduce more subtle considerations. While we consider a general characterization of such models where the observable representation to be future work, here try to give some intuition on what types of latent structures the rank condition may fail.

First, the rank condition can fail is if there are not enough observed nodes/states, and thus $\#\text{states}(\mathscr{O}_i) < \tau_i$. Intuitively, this corresponds to a model where the latent space is too expressive and inherently represents an intractable model (e.g. a set of n binary variables connected by a hidden node with 2^n states is equivalent to a clique of size n).

However, there are more subtle considerations unique to non-tree models. In general, our method is not limited to non-triangulated graphs (see the factorial HMM in Figure 3 of the main paper), but the process of triangulation can introduce artificial dependencies that can lead to complications. Consider Figure 1(a) which shows a DAG and its corresponding junction tree. To construct $\widetilde{\boldsymbol{\mathcal{P}}}(C_{AD})$, we may set $\mathbb{P}[\mathscr{O}_i, \mathscr{O}_{-i}] = \mathbb{P}[D, F]$ based on the junction tree topology. However, in the original model before triangulation, $D \perp F$ because of the v-structure. As a result, $\mathbb{P}[D, F]$ does not have rank k_h and thus cannot be inverted. However, note that choosing $\mathbb{P}[\mathscr{O}_i, \mathscr{O}_{-i}] = \mathbb{P}[D, E]$ is valid.

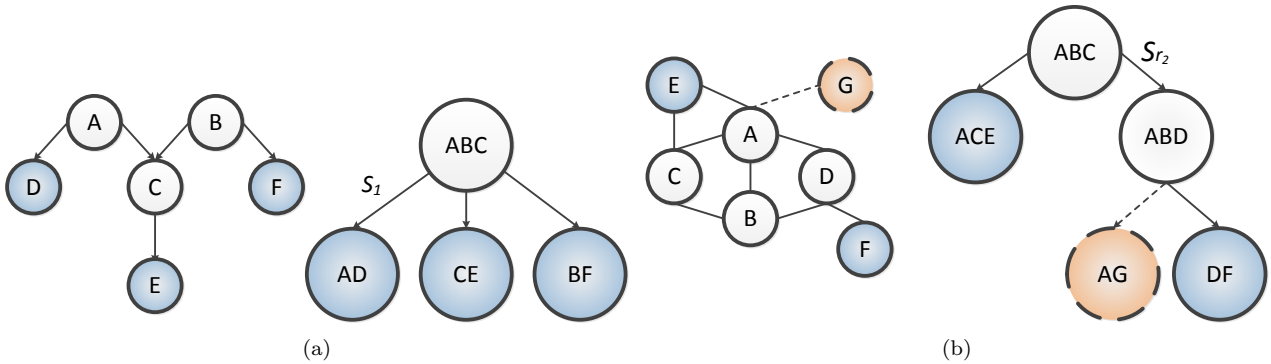
Finally consider Figure 1 (b), ignoring the orange nodes for now and assuming the variables are binary. In this model, the hidden variables are largely redundant since integrating out A and B would simply give a chain. $\boldsymbol{\mathcal{F}}_{r_2}$ must be set to $\mathbb{P}[F|S_{r_2}] = \mathbb{P}[F|AB]$. If we think of A, B has just one large variable, then it is clear that $\mathbb{P}[F|AB] = \mathbb{P}[F|D]\mathbb{P}[D|AB]$. However, D only has two states while AB has 4, so $\mathbb{P}[F|AB]$ only has rank 2. Now, consider adding the orange node. In this case we could set $\boldsymbol{\mathcal{F}}_{r_2}$ to $\mathbb{P}[F, G|S_{r_2}] = \mathbb{P}[F, G|A, B]$ whose matricized version has rank 4. Note that once the orange node has been added, integrating out A and B no longer produces a simple chain, but a more complicated structure.

Thus, we believe that more rigorously characterizing the existence of the observable representation in more detail, may shed light on the "intrinsic" complexity/redundancy of latent variable models in the context of linear and tensor algebra.

2 Linear Systems to Improve Stability

As derived in Section 6 in the main paper, for non-root and non-leaf nodes:

$$\widetilde{\boldsymbol{\mathcal{P}}}(C_i) \times_{\mathscr{O}_i} \mathbb{P}(\mathscr{O}_i, \mathscr{O}_{i-}) = \mathbb{P}(\mathscr{O}_{i_1}, \mathscr{O}_{i_2}, \mathscr{O}_{i-}) \tag{1}$$



(a) (b)

Figure 1: Models and their corresponding junction trees, where constructing the observable representation poses complications. See Section 1.

which then implies that

$$\widetilde{\boldsymbol{\mathcal{P}}}(C_i) = \mathbb{P}(\mathscr{O}_{i_1}, \mathscr{O}_{i_2}, \mathscr{O}_{i-}) \times_{\mathscr{O}_{i-}} \mathbb{P}(\mathscr{O}_i, \mathscr{O}_{i-})^{-1} \tag{2}$$

However, there may be many choices of $\mathscr{O}_{i-}$ (which we denote with $\mathscr{O}_{i-}^{(1)}, ..., \mathscr{O}_{i-}^{(N)}$) for which the above equality is true:

$$\begin{aligned}
\widetilde{\boldsymbol{\mathcal{P}}}(C_i) \times_{\mathscr{O}_i} \mathbb{P}(\mathscr{O}_i, \mathscr{O}_{i-}^{(1)}) &= \mathbb{P}(\mathscr{O}_{i_1}, \mathscr{O}_{i_2}, \mathscr{O}_{i-}^{(1)}) \\
\widetilde{\boldsymbol{\mathcal{P}}}(C_i) \times_{\mathscr{O}_i} \mathbb{P}(\mathscr{O}_i, \mathscr{O}_{i-}^{(2)}) &= \mathbb{P}(\mathscr{O}_{i_1}, \mathscr{O}_{i_2}, \mathscr{O}_{i-}^{(2)}) \\
&\dots \\
\widetilde{\boldsymbol{\mathcal{P}}}(C_i) \times_{\mathscr{O}_i} \mathbb{P}(\mathscr{O}_i, \mathscr{O}_{i-}^{(N)}) &= \mathbb{P}(\mathscr{O}_{i_1}, \mathscr{O}_{i_2}, \mathscr{O}_{i-}^{(N)})
\end{aligned}$$

This defines an over-constrained linear system, and we can solve for $\widetilde{\boldsymbol{\mathcal{P}}}(C_i)$ using least squares. In the case where #states($\mathcal{O}_i$) > #states(S_i), and the projection tensor $\boldsymbol{\mathcal{U}}_i$ is needed, our system of equations becomes:

$$\begin{aligned}
\widetilde{\boldsymbol{\mathcal{P}}}(C_i) \times_{\mathscr{O}_i} (\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{i-}^{(1)}) \times_{\mathscr{O}_i} \boldsymbol{\mathcal{U}}_i) &= \mathbb{P}(\mathscr{O}_{i_1}, \mathscr{O}_{i_2}, \mathscr{O}_{i-}^{(1)}) \times_{\mathscr{O}_{i_1}} \boldsymbol{\mathcal{U}}_{i_1} \times_{\mathscr{O}_{i_2}} \boldsymbol{\mathcal{U}}_{i_2} \\
\widetilde{\boldsymbol{\mathcal{P}}}(C_i) \times_{\mathscr{O}_i} (\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{i-}^{(2)}) \times_{\mathscr{O}_i} \boldsymbol{\mathcal{U}}_i) &= \mathbb{P}(\mathscr{O}_{i_1}, \mathscr{O}_{i_2}, \mathscr{O}_{i-}^{(2)}) \times_{\mathscr{O}_{i_1}} \boldsymbol{\mathcal{U}}_{i_1} \times_{\mathscr{O}_{i_2}} \boldsymbol{\mathcal{U}}_{i_2} \\
&\dots \\
\widetilde{\boldsymbol{\mathcal{P}}}(C_i) \times_{\mathscr{O}_i} (\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{i-}^{(N)}) \times_{\mathscr{O}_i} \boldsymbol{\mathcal{U}}_i) &= \mathbb{P}(\mathscr{O}_{i_1}, \mathscr{O}_{i_2}, \mathscr{O}_{i-}^{(N)}) \times_{\mathscr{O}_{i_1}} \boldsymbol{\mathcal{U}}_{i_1} \times_{\mathscr{O}_{i_2}} \boldsymbol{\mathcal{U}}_{i_2}
\end{aligned}$$

In this case, one good choice of $\boldsymbol{\mathcal{U}}_i$ is the top singular vectors of the matrix formed by the horizontal concatenation of $\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{i-}^{(1)}), ..., \mathbb{P}(\mathscr{O}_i, \mathscr{O}_{i-}^{(N)})$.

This linear system method allows for more robust estimation especially in smaller sample sizes (at the cost of more computation). It can be applied to the leaf case as well. One does not need to set up the linear system with all the valid choices; a subset is also acceptable.

3 Sample Complexity Theorem

We prove the sample complexity theorem.

3.1 Notation

For simplicity, the main text describes the algorithm in the context of a binary tree. However, in the sample complexity proof, we adopt a more general notation. Let C_i be a clique, and $C_{\alpha_i(1)}$,...,$C_{\alpha_i(\gamma_i)}$ denote its γ_i children. Let $\mathbb{C}$ denote the set of all the cliques in the junction tree, and $|\mathbb{C}|$ denote its size. Define $d_{\max}$ to be maximum degree of a tensor in the observable representation and e_{max} to be maximum number of

observed variables in any tensor. Furthermore, let $\tau_i = |k_h| \times |S_i|$ (i.e. the number of states associated with the separator).

We can now write the transformed representation for junction trees with more than 3 neighbors as:

Root:

$$\widetilde{\mathcal{P}}(C_i) = \mathcal{P}(C_i) \times_{S_{\alpha_i(1)}} \boldsymbol{\mathcal{F}}_{\alpha_i(1)} \times \ldots \times_{S_{\alpha_i(\gamma_i)}} \boldsymbol{\mathcal{F}}_{\alpha_i(\gamma_i)}$$

Internal nodes:

$$\widetilde{\mathcal{P}}(C_i) = \mathcal{P}(C_i) \times_{S_i} \boldsymbol{\mathcal{F}}_i^{\dagger} \times_{S_{\alpha_i(1)}} \boldsymbol{\mathcal{F}}_{\alpha_i(1)} \times \ldots \times_{S_{\alpha_i(\gamma_i)}} \boldsymbol{\mathcal{F}}_{\alpha_i(\gamma_i)}$$

Leaf:

$$\widetilde{\mathcal{P}}(C_i) = \mathcal{P}(C_i) \times_{S_i} \boldsymbol{\mathcal{F}}_i^{\dagger}$$

and the observable representation as:

$$\textbf{root: } \mathcal{P}(C_i) = \mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}) \times_{\mathscr{O}_{\alpha_i(1)}} \boldsymbol{\mathcal{U}}_{\alpha_i(1)} \times \ldots \times_{\mathscr{O}_{\alpha_i(\gamma_i)}} \boldsymbol{\mathcal{U}}_{\alpha_i(\gamma_i)}$$

$$\textbf{internal: } \mathcal{P}(C_i) = \mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i}) \times_{\mathscr{O}_{-i}} (\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{-i}) \times_{\mathscr{O}_i} \boldsymbol{\mathcal{U}}_i)^{\dagger} \times_{\mathscr{O}_{\alpha_i(1)}} \boldsymbol{\mathcal{U}}_{\alpha_i(1)} \times \ldots \times_{\mathscr{O}_{\alpha_i(\gamma_i)}} \boldsymbol{\mathcal{U}}_{\alpha_i(\gamma_i)}$$

$$\textbf{leaf: } \mathcal{P}(C_i) = \widehat{\mathbb{P}}(R_i, \mathscr{O}_{-i}) \times_{\mathscr{O}_{-i}} (\widehat{\mathbb{P}}(\mathscr{O}_i, \mathscr{O}_{-i}) \times_{\mathscr{O}_i} \boldsymbol{\mathcal{U}}_i)^{\dagger}$$

Sometimes when the multiplication indices are clear, we will omit it to make things simpler.

Rearranged version:
We will often find it convenient to rearrange the tensors into lower order tensors for the purpose of taking some norms (such as the spectral norm). We define $\boldsymbol{R}(\cdot)$ as the "rearranging" operation. For example, $\boldsymbol{R}(\widehat{\mathbb{P}}(\mathscr{O}_i, \mathscr{O}_{-i}))$ is the matricized version of $\widehat{\mathbb{P}}(\mathscr{O}_i, \mathscr{O}_{-i})$ with $\mathscr{O}_i$ being mapped to the rows and $\mathscr{O}_{-i}$ being mapped to the columns.
More generally, if $\mathcal{P}(C_i)$ (which has order d_i) then $\boldsymbol{R}(\mathcal{P}(C_i))$ is a rearrangement of of $\mathcal{P}(C_i)$ such that it has order equal to the number of neighbors of C_i. The set of modes of $\mathcal{P}(C_i)$ corresponding to a single separator of C_i get mapped to a single mode in $\boldsymbol{R}(\mathcal{P}(C_i))$. We let $\boldsymbol{R}(S_{\alpha_i(j)})$ denote the mode that $S_{\alpha_i(j)}$ maps to in $\boldsymbol{R}(\mathcal{P}(C_i))$.

In the case of the root, $\boldsymbol{R}(\mathcal{P}(C_i))$ rearranges $\mathcal{P}(C_i)$ into a tensor of order γ_i. For other clique nodes, $\boldsymbol{R}(\mathcal{P}(C_i))$ rearranges $\mathcal{P}(C_i)$ into a tensor of order $\gamma_i + 1$.

Example: In Figure 2 of the main text $\mathcal{P}(C_{BCDE}) = \mathbb{P}(\oslash_2 B, D | \oslash_2 C, E)$. C_{BCDE} has 3 neighbors and so $\boldsymbol{R}(\mathcal{P}(C_{BCDE}))$ is of order 3 where each of the modes correspond to $\{B, C\}$, $\{B, D\}$, $\{C, E\}$ respectively.

This rearrangement can be applied to the other quantities. For example, $\boldsymbol{R}(\mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}))$ is a rearranging of $\mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)})$ into a tensor of order γ_i where the $\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}$ correspond to one mode each. $\boldsymbol{R}(\widehat{\mathbb{P}}(\mathscr{O}_i, \mathscr{O}_{-i}))$ is the matricized version of $\widehat{\mathbb{P}}(\mathscr{O}_i, \mathscr{O}_{-i})$ with $\mathscr{O}_i$ being mapped to the rows and $\mathscr{O}_{-i}$ being mapped to the columns. Similarly, $\boldsymbol{R}(\boldsymbol{\mathcal{F}}_i)$ is the matricized version of $\boldsymbol{\mathcal{F}}_i$ and $\boldsymbol{R}(\mathcal{U}_i)$ is the matricized version of $\mathcal{U}_i$.

Thus, we can define the rearranged quantities:

Root:

$$\boldsymbol{R}(\widetilde{\mathcal{P}}(C_i)) = \boldsymbol{R}(\mathcal{P}(C_i)) \times_{\boldsymbol{R}(S_{\alpha_i(1)})} \boldsymbol{R}(\boldsymbol{\mathcal{F}}_{\alpha_i(1)}) \times \ldots \times_{\boldsymbol{R}(S_{\alpha_i(\gamma_i)})} \boldsymbol{R}(\boldsymbol{\mathcal{F}}_{\alpha_i(\gamma_i)})$$

$$\boldsymbol{R}(\widetilde{\mathcal{P}}(C_i)) = \boldsymbol{R}(\mathcal{P}(C_i)) \times_{\boldsymbol{R}(S_i)} \boldsymbol{R}(\boldsymbol{\mathcal{F}}_i^{\dagger}) \times_{\boldsymbol{R}(S_{\alpha_i(1)})} \boldsymbol{\mathcal{F}}_{\alpha_i(1)} \times \ldots \times_{\boldsymbol{R}(S_{\alpha_i(\gamma_i)})} \boldsymbol{R}(\boldsymbol{\mathcal{F}}_{\gamma_i})$$

Leaf:

$$\boldsymbol{R}(\widetilde{\mathcal{P}}(C_i)) = \boldsymbol{R}(\mathcal{P}(C_i)) \times_{\boldsymbol{R}(S_i)} \boldsymbol{R}(\boldsymbol{\mathcal{F}}_i^{\dagger})$$

and the observable representation as:

$$\begin{aligned}
\textbf{root: } \boldsymbol{R}(\mathcal{P}(C_i)) &= \boldsymbol{R}(\mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)})) \times_{\boldsymbol{R}(S_{\alpha_i(1)})} \boldsymbol{R}(\boldsymbol{\mathcal{U}}_{\alpha_i(1)}) \times \ldots \times_{\boldsymbol{R}(S_{\alpha_i(\gamma_i)})} \boldsymbol{R}(\boldsymbol{\mathcal{U}}_{\alpha_i(\gamma_i)}) \\
\textbf{internal: } \boldsymbol{R}(\mathcal{P}(C_i)) &= \boldsymbol{R}(\mathbb{P}(\mathscr{O}_{S_{\alpha_i(1)}}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i})) \times_{\boldsymbol{R}(S_{-i})} \boldsymbol{R}((\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{-i}) \times_{S_i} \boldsymbol{\mathcal{U}}_i)^\dagger) \\
&\quad \times_{\boldsymbol{R}(S_{\alpha_i(1)})} \boldsymbol{R}(\boldsymbol{\mathcal{U}}_{\alpha_i(1)}) \times \ldots \times_{\boldsymbol{R}(S_{\alpha_i(\gamma_i)})} \boldsymbol{R}(\boldsymbol{\mathcal{U}}_{\alpha_i(\gamma_i)}) \\
\textbf{leaf: } \boldsymbol{R}(\mathcal{P}(C_i)) &= \boldsymbol{R}(\widehat{\mathbb{P}}(R_i, \mathscr{O}_{-i})) \times_{\boldsymbol{R}(\mathscr{O}_{-i})} \boldsymbol{R}((\widehat{\mathbb{P}}(\mathscr{O}_i, \mathscr{O}_{-i}) \times_{\mathscr{O}_i} \boldsymbol{\mathcal{U}}_i)^\dagger)
\end{aligned} \tag{3}$$

Furthermore define the following:

$$\alpha = \min_{C_i \in \mathbb{C}} \ \sigma_{\tau_i}(\mathbb{P}[\mathscr{O}_i, \mathscr{O}_{-i}]) \tag{4}$$

$$\beta = \min_{C_i \in \mathbb{C}} \ \sigma_{\tau_i}(\boldsymbol{\mathcal{F}}_i) \tag{5}$$

The proof is based on the technique of HKZ [2] but has differences due to the junction tree topology and higher order tensors.

3.2 Tensor Norms

We briefly define several tensor norms that are a generalization of matrix norms. For more information about matrix norms see [1].

Frobenius Norm: Just like for matrices, the frobenius norm of a tensor of order N is defined as:

$$\|\boldsymbol{T}\|_F = \sqrt{\sum_{i_1,\ldots,i_N} \boldsymbol{T}(i_1, ..., i_N)^2} \tag{6}$$

Elementwise One Norm: Similarly, the elementwise one norm of a tensor of order N is defined as:

$$\|\boldsymbol{T}\|_1 = \sum_{i_1,\ldots,i_N} |\boldsymbol{T}(i_1, ..., i_N)| \tag{7}$$

Spectral Norm: For tensors of order N the spectral norm [3] can be defined as as

$$|||\boldsymbol{T}|||_2 = \sup_{\boldsymbol{v}_i \ s.t. \ \|\boldsymbol{v}_i\|_2 \leq 1 \ \forall 1 \leq i \leq N} \boldsymbol{\mathcal{T}} \times_N \boldsymbol{v}_N, ..., \times_2 \boldsymbol{v}_2 \times_1 \boldsymbol{v}_1 \tag{8}$$

In our case, we will find it more convenient to use the rearranged spectral norm, which we define as:

$$|||\boldsymbol{T}|||_{2R} = |||\boldsymbol{R}(\boldsymbol{T})|||_2 \tag{9}$$

where $\boldsymbol{R}(\cdot)$ was defined in the previous section.

Induced One Norm: For matrices the induced one norm is defined as the max column sum of the matrix: $|||\boldsymbol{M}|||_1 = \sup_{\boldsymbol{v} \ s.t. \ \|\boldsymbol{v}\|_1 \leq 1} \|\boldsymbol{M}\boldsymbol{v}\|_1$. We can generalize this to be the max slice sum of a tensor on a tensor where some modes are fixed and others are summed over. Let $\boldsymbol{\sigma}$ denote the modes that will be maxed over. Then:

$$|||\boldsymbol{T}|||_1^{\boldsymbol{\sigma}} = \sup_{\boldsymbol{v}_i \ s.t. \ \|\boldsymbol{v}_i\|_1 \leq 1, \ \forall 1 \leq i \leq |\boldsymbol{\sigma}|} \|\boldsymbol{T} \times_{\boldsymbol{\sigma}_1} \boldsymbol{v}_1, ..., \times_{\boldsymbol{\sigma}_{|\boldsymbol{\sigma}|}} \boldsymbol{v}_{|\boldsymbol{\sigma}|}\|_1 \tag{10}$$

In the Appendix, we prove several lemmas regarding these tensor norms.

3.3 Concentration Bounds

$$\epsilon(\mathscr{O}_1, ..., \mathscr{O}_d) = \left|\left|\widehat{\mathbb{P}}(\mathscr{O}_1, ..., \mathscr{O}_d) - \mathbb{P}(\mathscr{O}_1, ..., \mathscr{O}_d)\right|\right|_F \tag{11}$$

$$\epsilon(\mathscr{O}_1, ..., \mathscr{O}_{d-e}, \boldsymbol{o}_{d-e+1}, ..., \boldsymbol{o}_d) = \left|\left|\widehat{\mathbb{P}}(\mathscr{O}_1, ..., \mathscr{O}_{d-e}, \boldsymbol{o}_{d-e+1}, ..., \boldsymbol{o}_d) - \mathbb{P}(\mathscr{O}_1, ..., \mathscr{O}_{d-e}, \boldsymbol{o}_{d-e+1}, ..., \boldsymbol{o}_d)\right|\right|_F \tag{12}$$

$1, ..., d-e$ denote the $d-e$ non-evidence variables while $d-e+1, ..., d$ denote the e evidence variables. d indicates the total number of modes of the tensor. As the number of samples N gets large, we expect these quantities to be small.

Lemma 1 (variant of HKZ [2]) *If the algorithm independently samples N of the observations, then with probability at least $1-\delta$.*

$$\epsilon(\mathscr{O}_1,....,\mathscr{O}_d) \leq \sqrt{\frac{1}{N}\ln\frac{2|\mathbb{C}|}{\delta}}+\sqrt{\frac{1}{N}} \quad (13)$$

$$\sum_{\boldsymbol{o}_{d-e+1},...,\boldsymbol{o}_d}\epsilon(\mathscr{O}_1,....,\mathscr{O}_{d-e},\boldsymbol{o}_{d-e+1},...,\boldsymbol{o}_d) \leq \sqrt{\frac{k_o^{e_{\max}}}{N}\ln\frac{2|\mathbb{C}|}{\delta}}+\sqrt{\frac{k_o^{e_{\max}}}{N}} \quad (14)$$

*for **all** tuples $(\mathscr{O}_1,....,\mathscr{O}_d)$ that are used in the spectral algorithm.*

The proof is the same as that of HKZ [2] except the union bound is taken over $2|\mathbb{C}|$ instead of 3 (since each transformed quantity in the spectral algorithm is composed of at most two such terms). The last bound can be made tighter, identical to HKZ, but for simplicity we do not pursue that approach here.

3.4 Singular Value Bounds

Basically this is the generalized version of Lemma 9 in HKZ [2], which is stated below for completeness:

Lemma 2 (variant of HKZ [2]) *Suppose $\epsilon(\mathscr{O}_i,\mathscr{O}_{-i}) \leq \varepsilon \times \sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i,\mathscr{O}_{-i}))$ for some $\varepsilon < 1/2$.*
Let $\varepsilon_0 = \epsilon(\mathscr{O}_i,\mathscr{O}_{-i})^2/((1-\varepsilon)\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i,\mathscr{O}_{-i}))^2$. Then:

1. $\varepsilon_0 < 1$
2. $\sigma_{\tau_i}(\widehat{\mathbb{P}}(\mathscr{O}_i,\mathscr{O}_{-i})\times_{\mathscr{O}_i}\widehat{\boldsymbol{\mathcal{U}}}_i) \geq (1-\varepsilon_0)\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i,\mathscr{O}_{-i}))$
3. $\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i,\mathscr{O}_{-i})\times_{\mathscr{O}_i}\widehat{\boldsymbol{\mathcal{U}}}_i) \geq \sqrt{1-\varepsilon_0}\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i,\mathscr{O}_{-i}))$
4. $\sigma_{\tau_i}(\widehat{\mathbb{P}}(\mathscr{O}_i|S_i)\times_{\mathscr{O}_i}\widehat{\boldsymbol{\mathcal{U}}}_i) \geq \sqrt{1-\varepsilon_0}\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i|S_i)))$

It follows that if $\epsilon(\mathscr{O}_i,\mathscr{O}_{-i}) \leq \sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i,\mathscr{O}_{-i}))/3$ then this implies that $\varepsilon_0 \leq \frac{1/9}{4/9} = \frac{1}{4}$. It then follows that,

1. $\sigma_{\tau_i}(\widehat{\mathbb{P}}(\mathscr{O}_i,\mathscr{O}_{-i})\times_{\mathscr{O}_i}\widehat{\boldsymbol{\mathcal{U}}}_i) \geq \frac{3}{4}\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i,\mathscr{O}_{-i}))$
2. $\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i,\mathscr{O}_{-i})\times_{\mathscr{O}_i}\widehat{\boldsymbol{\mathcal{U}}}_i) \geq \frac{\sqrt{3}}{2}\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i,\mathscr{O}_{-i}))$
3. $\sigma_{\tau_i}(\widehat{\mathbb{P}}(\mathscr{O}_i|S_i)\times_{\mathscr{O}_i}\widehat{\boldsymbol{\mathcal{U}}}_i) \geq \frac{\sqrt{3}}{2}\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i|S_i))$

3.5 Bounding the Transformed Quantities

Define,

1. **root:** $\breve{\mathcal{P}}(C_i) := \mathbb{P}(\mathscr{O}_{\alpha_i(1)},\ldots,\mathscr{O}_{\alpha_i(\gamma_i)})\times_{\mathscr{O}_{\alpha_i(1)}}\widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)}\ldots\times_{\mathscr{O}_{\alpha_i(\gamma_i)}}\widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)}$
2. **leaf**: $\breve{\mathcal{P}}_{r_i}(C_i) = \mathbb{P}(R_i=r_i,\mathscr{O}_{-i})\times_{\mathscr{O}_{-i}}(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i,\mathscr{O}_{-i}))^\dagger$
3. **internal:** $\breve{\mathcal{P}}(C_i) = \mathbb{P}(\mathscr{O}_{S_{\alpha_i(1)}},\ldots,\mathscr{O}_{\alpha_i(\gamma_i)},\mathscr{O}_{-i})\times_{\mathscr{O}_{-i}}(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i,\mathscr{O}_{-i}))^\dagger\times_{\mathscr{O}_{\alpha_i(1)}}\widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)}\times\ldots\times_{\mathscr{O}_{\alpha_i(\gamma_i)}}\widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)}$

(which are the observable quantities with the true probabilities, but empirical $\widehat{\boldsymbol{\mathcal{U}}}$'s). Similarly, we can define $\breve{\boldsymbol{\mathcal{F}}}_i = \mathbb{P}[\mathscr{O}_i|S_i]\times_{\mathscr{O}_i}\widehat{\boldsymbol{\mathcal{U}}}_i$. We have also abbreviated $\mathbb{P}(\mathscr{O}_i,\mathscr{O}_{-i})\times_{\mathscr{O}_i}\widehat{\boldsymbol{\mathcal{U}}}_i$ with $\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i,\mathscr{O}_{-i})$.

We seek to bound the following three quantities:

$$\delta_i^{root} := \|(\widehat{\mathcal{P}}(C_i)-\breve{\mathcal{P}}(C_i))\times_{\mathscr{O}_{\alpha_i(1)}}\breve{\boldsymbol{\mathcal{F}}}^{-1}_{\alpha_i(1)},...,\times_{\mathscr{O}_{\alpha_i(\gamma_i)}}\breve{\boldsymbol{\mathcal{F}}}^{-1}_{\alpha_i(\gamma_i)}\|_1 \quad (15)$$

$$\delta_i^{internal} := \left|\left|\left|(\widehat{\mathcal{P}}(C_i)-\breve{\mathcal{P}}(C_i))\times_{\mathscr{O}_{-i}}\breve{\boldsymbol{\mathcal{F}}}_i\times_{\mathscr{O}_{\alpha_i(1)}}\breve{\boldsymbol{\mathcal{F}}}_i^{-1},...,\times_{\mathscr{O}_{\alpha_i(\gamma_i)}}\breve{\boldsymbol{\mathcal{F}}}^{-1}_{\alpha_i(\gamma_i)}\right|\right|\right|_1^{S_i} \quad (16)$$

$$\xi_i^{leaf} := \sum_{r_i}\|(\widehat{\mathcal{P}}_{r_i}(C_i)-\breve{\mathcal{P}}_{r_i}(C_i))\times_{\mathscr{O}_{-i}}\breve{\boldsymbol{\mathcal{F}}}_i\|_1 \quad (17)$$

Lemma 3 *If* $\epsilon(\mathscr{O}_i, \mathscr{O}_{-i}) \leq \sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{-i}))/3$ *then*

$$\delta_i^{root} \leq \frac{2^{d_{\max}} k_h^{d_{\max}} \epsilon(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)})}{3^{d/2}\beta^d} \tag{18}$$

$$\delta_i^{internal} \leq \frac{2^{d_{\max}+3} k_h^{d_{\max}}}{3\sqrt{3}^d \beta^d} \left(\frac{\epsilon(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i})}{\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{-i}))} + \frac{\epsilon(\mathscr{O}_i, \mathscr{O}_{-i})}{(\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{-i})))^2} \right) \tag{19}$$

$$\xi_i^{leaf} \leq \frac{8 k_h^{d_{\max}}}{3} \left(\frac{\epsilon(\mathscr{O}_i, \mathscr{O}_{-i})}{\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{-i}))^2} + \frac{\sum_{r_i} \epsilon(R_i = r_i, \mathscr{O}_{-i})}{\sigma_{\tau_i}(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))} \right) \tag{20}$$

We define $\triangle = \max(\delta_i^{root}, \delta_i^{internal}, \xi_i^{leaf})$ *(over all i).*

Proof

Case δ_i^{root} :
For the root, $\widehat{\mathcal{P}}(C_i) = \widehat{\mathbb{P}}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}) \times_{\mathcal{O}_{\alpha_i(1)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)} \times \ldots \times_{\mathcal{O}_{\alpha_i(\gamma_i)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)}$ *and similarly* $\breve{\mathcal{P}}(C_i) = \mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}) \times_{\mathcal{O}_{\alpha_i(1)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)} \times \ldots \times_{\mathcal{O}_{\alpha_i(\gamma_i)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)}$.

$$\begin{aligned}
\delta^{root} &= \Big|\Big|\Big(\widehat{\mathbb{P}}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}) \times_{\mathcal{O}_{\alpha_i(1)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)} \times \ldots \times_{\mathcal{O}_{\alpha_i(\gamma_i)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)} \\
&\quad - \mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}) \times_{\mathcal{O}_{\alpha_i(1)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)} \times \ldots \times_{\mathcal{O}_{\alpha_i(\gamma_i)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)}\Big) \times_{\mathcal{O}_{\alpha_i(1)}} \breve{\boldsymbol{\mathcal{F}}}^{-1}_{\alpha_i(1)}, ..., \times_{\mathcal{O}_{\alpha_i(\gamma_i)}} \breve{\boldsymbol{\mathcal{F}}}^{-1}_{\alpha_i(1)}\Big|\Big|_1 \\
&\leq k_h^{d_{\max}} \Big|\Big|\Big|(\widehat{\mathbb{P}}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}) - \mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)})) \times_{\mathcal{O}_{\alpha_i(1)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)} \times \ldots \times_{\mathcal{O}_{\alpha_i(\gamma_i)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)}\Big|\Big|\Big|_{2R} \\
&\quad \times \Big|\Big|\Big|\breve{\boldsymbol{\mathcal{F}}}^{-1}_{\alpha_i(1)}\Big|\Big|\Big|_{2R} .... \Big|\Big|\Big|\breve{\boldsymbol{\mathcal{F}}}_{\alpha_i(1)})^{-1}\Big|\Big|\Big|_{2R} \\
&\leq \frac{k_h^{d_{\max}}}{\prod_{j=1}^{\gamma_i} \sigma_{\tau_{\alpha_i(j)}}(\breve{\boldsymbol{\mathcal{F}}}_{\alpha_i(j)})} \Big|\Big|\Big|(\widehat{\mathbb{P}}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}) - \mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)})) \times_{\mathcal{O}_{\alpha_i(1)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)} \times \ldots \times_{\mathcal{O}_{\alpha_i(\gamma_i)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)}\Big|\Big|\Big|_{2R} \\
&\leq \frac{k_h^{d_{\max}}}{\prod_{j=1}^{\gamma_i} \sigma_{\tau_{\alpha_i(j)}}(\breve{\boldsymbol{\mathcal{F}}}_{\alpha_i(j)})} \Big|\Big|\Big|(\widehat{\mathbb{P}}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}) - \mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}))\Big|\Big|\Big|_{2R} \times \\
&\qquad \Big|\Big|\Big|\widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)}\Big|\Big|\Big|_{2R} \Big|\Big|\Big|\widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(2)}\Big|\Big|\Big|_{2R} ... \Big|\Big|\Big|\widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)}\Big|\Big|\Big|_{2R} \\
&= \frac{k_h^{d_{\max}}}{\prod_{j=1}^{\gamma_i} \sigma_{\tau_{\alpha_i(j)}}(\breve{\boldsymbol{\mathcal{F}}}_{\alpha_i(j)})} \epsilon(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)})
\end{aligned}$$

Between the first and second line we use Lemma 8 to convert from elementwise one norm to spectral norm, and Lemma 6 (submultiplicativity). Lemma 6 (submultiplicativity) is applied again to get to the second-to-last line. We also use the fact that $\Big|\Big|\Big|\widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)}\Big|\Big|\Big|_{2R} = 1$.

Thus, by Lemma 2

$$\delta_i^{root} \leq \frac{2^{d_{max}} k_h^{d_{max}} \epsilon(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)})}{3^{d_{max}/2}\beta^{d_{max}}} \tag{21}$$

Case ξ_i^{leaf}***:***
We note that $\widehat{\mathcal{P}}_{r_i}(C_i) = \widehat{\mathbb{P}}(R_i = r_i, \mathscr{O}_{-i}) \times_{\mathscr{O}_{-i}} (\widehat{\mathbb{P}}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))^\dagger$ *and similarly* $\breve{\mathcal{P}}_{r_i}(C_i) = \mathbb{P}(R_i = r_i, \mathscr{O}_{-i}) \times_{\mathscr{O}_{-i}} (\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))^\dagger$.

Again, we use Lemma 8 to convert from the one norm to the spectral norm, and Lemma 6 for submulti-

plicativity.

$$\begin{aligned}
\xi_i^{leaf} &= \sum_{r_i} \left|\left|\left|(\widehat{\mathcal{P}}_{r_i}(C_i) - \breve{\mathcal{P}}_{r_i}(C_i)) \times_{\mathscr{O}_{-i}} \breve{\boldsymbol{\mathcal{F}}}_i\right|\right|\right|_1^{S_i} \\
&= \sum_{r_i} \left|\left|\left|(\widehat{\mathcal{P}}_{r_i}(C_i) - \breve{\mathcal{P}}_{r_i}(C_i)) \times_{\mathscr{O}_{-i}} \widehat{\boldsymbol{\mathcal{U}}}_i\right|\right|\right|_1^{S_i} |||\mathbb{P}[\mathscr{O}_i|S_i]|||_1^{S_i} \\
&\leq \sum_{r_i} k_h^{d_{\max}} \left|\left|\left|(\widehat{\mathcal{P}}_{r_i}(C_i) - \breve{\mathcal{P}}_{r_i}(C_i))\right|\right|\right|_{2R} \left|\left|\left|\widehat{\boldsymbol{\mathcal{U}}}_i\right|\right|\right|_{2R} \\
&\leq \sum_{r_i} k_h^{d_{\max}} \left|\left|\left|(\widehat{\mathcal{P}}_{r_i}(C_i) - \breve{\mathcal{P}}_{r_i}(C_i))\right|\right|\right|_{2R} \qquad (22)
\end{aligned}$$

Note that $|||\mathbb{P}[\mathscr{O}_i|S_i]|||_1^{S_i} = 1$, *and* $\left|\left|\left|\widehat{\boldsymbol{\mathcal{U}}}_i\right|\right|\right|_{2R} = 1$.

$$\begin{aligned}
\left|\left|\left|\widehat{\mathcal{P}}_{r_i}(C_i) - \breve{\mathcal{P}}_{r_i}(C_i)\right|\right|\right|_{2R} &= \left|\left|\left|\widehat{\mathbb{P}}(R_i = r_i, \mathscr{O}_{-i}) \times_{\mathscr{O}_{-i}} (\widehat{\mathbb{P}}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))^\dagger - \mathbb{P}(R_i = r_i, \mathscr{O}_{-i}) \times_{\mathscr{O}_{-i}} (\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))^\dagger\right|\right|\right|_{2R} \\
&\leq \left|\left|\left|\mathbb{P}(R_i = r_i, \mathscr{O}_{-i}) \times_{\mathscr{O}_{-i}} (\widehat{\mathbb{P}}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})^\dagger - \mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})^\dagger)\right|\right|\right|_{2R} + \\
&\quad \left|\left|\left|(\widehat{\mathbb{P}}(R_i = r_i, \mathscr{O}_{-i}) - \mathbb{P}(R_i = r_i, \mathscr{O}_{-i})) \times_{\mathscr{O}_{-i}} \mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})^\dagger\right|\right|\right|_{2R} \\
&\leq |||\mathbb{P}(R_i = r_i, \mathscr{O}_{-i})|||_{2R} \left|\left|\left|(\widehat{\mathbb{P}}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})^\dagger - \mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})^\dagger)\right|\right|\right|_{2R} + \\
&\quad |||\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})^\dagger|||_{2R} \left|\left|\left|(\widehat{\mathbb{P}}(R_i = r_i, \mathscr{O}_{-i}) - \mathbb{P}(R_i = r_i, \mathscr{O}_{-i}))\right|\right|\right|_{2R} \\
&\leq \mathbb{P}(R_i = r_i) \frac{1+\sqrt{5}}{2} \frac{\epsilon(\mathscr{O}_i, \mathscr{O}_{-i})}{\min(\sigma_{\tau_i}(\widehat{\mathbb{P}}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})), \sigma_{\tau_{ij}}(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})))^2} \\
&\quad + \frac{\epsilon(R_i = r_i, \mathscr{O}_{-i})}{\sigma_{\tau_i}(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))}
\end{aligned}$$

The last line follows from Eq. 35. We have also used the fact that $|||\mathbb{P}(R_i = r_i, \mathscr{O}_{-i})|||_{2R} \leq \|\mathbb{P}(R_i = r_i, \mathscr{O}_{-i})\|_F \leq \mathbb{P}[R_i = r]$ *by Lemma 7. Thus, using Lemma 2,*

$$\xi_i^{leaf} \leq \sum_{r_i} k_h^{d_{\max}} \left(\mathbb{P}(R_i = r_i) \frac{1+\sqrt{5}}{2} \frac{\epsilon(\mathscr{O}_i, \mathscr{O}_{-i})}{\min(\sigma_{\tau_i}(\widehat{\mathbb{P}}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})), \sigma_{\tau_i}(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})))^2} + \frac{\epsilon(R_i = r_i, \mathscr{O}_{-i})}{\sigma_{\tau_i}(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))} \right) \qquad (23)$$

$$\begin{aligned}
\xi_i^{leaf} &\leq \sum_{r_i} k_h^{d_{max}} \left(\frac{1+\sqrt{5}}{2} \frac{16\mathbb{P}(R_i = r_i)\epsilon(\mathscr{O}_i, \mathscr{O}_{-i})}{9\sigma_{\tau_i}(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))^2} + \frac{2\epsilon(R_i = r_i, \mathscr{O}_{-i})}{\sqrt{3}\sigma_{\tau_i}(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))} \right) \\
&\leq \frac{8k_h^{d_{max}}}{3} \left(\frac{\epsilon(\mathscr{O}_i, \mathscr{O}_{-i})}{\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{-i}))^2} + \frac{\sum_{r_i} \epsilon(R_i = r_i, \mathscr{O}_{-i})}{\sigma_{\tau_i}(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))} \right) \qquad (24)
\end{aligned}$$

3.5.1 $\delta_i^{internal}$

$\breve{\mathcal{P}}(C_i) = \mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i}) \times_{\mathscr{O}_{-i}} (\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))^\dagger \times_{\mathscr{O}_{\alpha_i(1)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)} \times \ldots \times_{\mathscr{O}_{\alpha_i(\gamma_i)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)}$.

Similarly, $\widehat{\mathcal{P}}(C_i) = \widehat{\mathbb{P}}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i}) \times_{\mathscr{O}_{-i}} (\widehat{\mathbb{P}}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))^\dagger \times_{\mathscr{O}_{\alpha_i(1)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)} \times \ldots \times_{\mathscr{O}_{\alpha_i(\gamma_i)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)}$.

Again, we use Lemma 8 to convert from one norm to spectral norm and Lemma 6 for submultiplicativity.

$$\begin{aligned}
\delta_i^{internal} &= \left|\left|\left|(\widehat{\mathcal{P}}(C_i) - \breve{\mathcal{P}}(C_i)) \times_{\mathscr{O}_{-i}} \breve{\boldsymbol{\mathcal{F}}}_i \times_{\mathscr{O}_{\alpha_i(1)}} \breve{\boldsymbol{\mathcal{F}}}^{-1}_{\alpha_i(1)}, ..., \times_{\mathscr{O}_{\alpha_i(\gamma_i)}} \breve{\boldsymbol{\mathcal{F}}}^{-1}_{\alpha_i(\gamma_i)}\right|\right|\right|_1^{S_i} \\
&\leq \left|\left|\left|(\widehat{\mathcal{P}}(C_i) - \breve{\mathcal{P}}(C_i)) \times_{\mathscr{O}_{-i}} \widehat{\boldsymbol{\mathcal{U}}}_i \times_{\mathscr{O}_{\alpha_i(1)}} \breve{\boldsymbol{\mathcal{F}}}^{-1}_{\alpha_i(1)}, ..., \times_{\mathscr{O}_{\alpha_i(\gamma_i)}} \breve{\boldsymbol{\mathcal{F}}}^{-1}_{\alpha_i(\gamma_i)}\right|\right|\right|_1^{S_i} |||\mathbb{P}[\mathscr{O}_i|S_i]||||_1^{S_i} \\
&\leq k_h^{d_{max}} \left|\left|\left|(\widehat{\mathcal{P}}(C_i) - \breve{\mathcal{P}}(C_i))\right|\right|\right|_{2R} \left|\left|\left|\widehat{\boldsymbol{\mathcal{U}}}_i\right|\right|\right|_{2R} \left|\left|\left|\breve{\boldsymbol{\mathcal{F}}}^{-1}_{\alpha_i(1)}\right|\right|\right|_{2R} ... \left|\left|\left|\breve{\boldsymbol{\mathcal{F}}}^{-1}_{\alpha_i(\gamma_i)}\right|\right|\right|_{2R} \\
&\leq \frac{k_h^{d_{max}}}{\prod_{j=1}^{\gamma_i} \sigma_{\tau_{\alpha_i(j)}}(\breve{\boldsymbol{\mathcal{F}}}_{\alpha_i(j)})} \left|\left|\left|(\widehat{\mathcal{P}}(C_i) - \breve{\mathcal{P}}(C_i))\right|\right|\right|_{2R} \qquad (25)
\end{aligned}$$

Note that $|||\mathbb{P}[\mathscr{O}_i|S_i]||||_1^{S_i} = 1$, *and* $\left|\left|\left|\widehat{\boldsymbol{\mathcal{U}}}_i\right|\right|\right|_{2R} = 1$.

$$\begin{aligned}
&\left|\left|\left|\widehat{\mathcal{P}}(C_i) - \breve{\mathcal{P}}(C_i)\right|\right|\right|_{2R} \\
= &\left|\left|\left|\widehat{\mathbb{P}}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i}) \times_{\mathscr{O}_{-i}} \left(\widehat{\mathbb{P}}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})\right)^{\dagger} \times_{\mathscr{O}_{\alpha_i(1)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)} \times \ldots \times_{\mathscr{O}_{\alpha_i(\gamma_i)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)} - \right.\right. \\
&\left.\left. \mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i}) \times_{\mathscr{O}_{-i}} \left(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})\right)^{\dagger} \times_{\mathscr{O}_{\alpha_i(1)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)} \times \ldots \times_{\mathscr{O}_{\alpha_i(\gamma_i)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)}\right|\right|\right|_{2R} \\
\leq &\left|\left|\left|(\widehat{\mathbb{P}}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i}) - \mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i})) \times_{\mathscr{O}_{-i}} \mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))^{\dagger} \right.\right. \\
&\quad \left.\left. \times_{\mathscr{O}_{\alpha_i(1)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)} \times \ldots \times_{\mathscr{O}_{\alpha_i(\gamma_i)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)}\right|\right|\right|_{2R} \\
&+ \left|\left|\left|\mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i}) \times_{\mathscr{O}_{-i}} \left(\left(\widehat{\mathbb{P}}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})\right)^{\dagger} - \left(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})\right)^{\dagger}\right)\right.\right. \\
&\quad \left.\left. \times_{\mathscr{O}_{\alpha_i(1)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)} \times \ldots \times_{\mathscr{O}_{\alpha_i(\gamma_i)}} \widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)}\right|\right|\right|_{2R} \\
\leq &\left|\left|\left|(\widehat{\mathbb{P}}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i}) - \mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i}))\right|\right|\right|_{2R} |||\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))^{\dagger}|||_{2R} \left|\left|\left|\widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)}\right|\right|\right|_{2R} ... \left|\left|\left|\widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)}\right|\right|\right|_{2R} \\
&+ |||\mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i})||||_{2R} \left|\left|\left|\widehat{\mathbb{P}}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))^{\dagger} - \mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))^{\dagger}\right|\right|\right|_{2R} \left|\left|\left|\widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(1)}\right|\right|\right|_{2R} ... \left|\left|\left|\widehat{\boldsymbol{\mathcal{U}}}_{\alpha_i(\gamma_i)}\right|\right|\right|_{2R} \\
\leq &\frac{\epsilon(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i})}{\sigma_{\tau_i}(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))} + \frac{1+\sqrt{5}}{2} \frac{\epsilon(\mathscr{O}_i, \mathscr{O}_{-i})}{\min(\sigma_{\tau_i}(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})), \sigma_{\tau_i}(\widehat{\mathbb{P}}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})))^2}
\end{aligned}$$

The last line follows from Eq. 35. We have also used the fact that $\left|\left|\left|\mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i})\right|\right|\right|_{2R} \leq \|\mathbb{P}(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i}))\|_F \leq 1$ *via Lemma 7. Using Lemma 2,*

$$\delta_i^{internal} \leq \frac{(2k_h)^{d_{max}}}{(\beta\sqrt{3})^{d_{max}}} \left(\frac{\epsilon(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i})}{\sigma_{\tau_i}(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i}))} + \frac{1+\sqrt{5}}{2} \frac{\epsilon(\mathscr{O}_i, \mathscr{O}_{-i})}{\min(\sigma_{\tau_i}(\mathbb{P}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})), \sigma_{\tau_i}(\widehat{\mathbb{P}}_{\widehat{\boldsymbol{\mathcal{U}}}}(\mathscr{O}_i, \mathscr{O}_{-i})))^2}\right) \qquad (26)$$

$$\begin{aligned}
\delta_i^{internal} &\leq \frac{(2k_h)^{d_{max}}}{(\beta\sqrt{3})^{d_{max}}} \left(\frac{2\epsilon(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i})}{\sqrt{3}\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{-i}))} + \frac{8\epsilon(\mathscr{O}_i, \mathscr{O}_{-i})}{3(\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{-i})))^2}\right) \\
&\leq \frac{8(2k_h)^{d_{max}}}{3(\beta\sqrt{3})^{d_{max}}} \left(\frac{\epsilon(\mathscr{O}_{\alpha_i(1)}, \ldots, \mathscr{O}_{\alpha_i(\gamma_i)}, \mathscr{O}_{-i})}{\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{-i}))} + \frac{\epsilon(\mathscr{O}_i, \mathscr{O}_{-i})}{(\sigma_{\tau_i}(\mathbb{P}(\mathscr{O}_i, \mathscr{O}_{-i})))^2}\right) \qquad (27)
\end{aligned}$$

■

3.6 Bounding the Propagation of Error

We now show all these errors propagate on the junction tree. For this section, assume the clique nodes are numbered $1, 2, ..., |\mathbb{C}|$ in breadth first order (such that 1 is the root). Moreover let $\Phi_{1:c}(\mathcal{C})$ be the transformed

factors accumulated so far if we computed the joint probability from the root down (instead of the bottom up). For example,

$$\Phi_{1:1}(\mathcal{C}) = \mathcal{P}(\mathcal{C}_1) \tag{28}$$
$$\Phi_{1:2}(\mathcal{C}) = \mathcal{P}(\mathcal{C}_1) \times_{S_2} \mathcal{P}(\mathcal{C}_2) \tag{29}$$
$$\Phi_{1:2}(\mathcal{C}) = \mathcal{P}(\mathcal{C}_1) \times_{S_2} \mathcal{P}(\mathcal{C}_2) \times_{S_3} \mathcal{P}(\mathcal{C}_3) \tag{30}$$

(Note how this is very computationally inefficient: the tensors get very large. However, it is useful for proving statistical properties). Then the modes of $\Phi_{1:c}$ can be partitioned into mode groups, $\mathscr{M}_1$,..., $\mathscr{M}_{d_c}$ (where each mode group consists of the variables on the corresponding separator edge). We now prove the following lemma,

Lemma 4 *Define* $\triangle = \max(\delta_i^{root}, \delta_i^{internal}, \xi_i^{leaf})$. *Then,*

$$\sum_{\boldsymbol{x}_{1:c}} \|(\breve{\Phi}_{1:c}(\mathcal{C}) - \widehat{\Phi}_{1:c}(\mathcal{C})) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_c}} \breve{\boldsymbol{\mathcal{F}}}_{d_c}^{-1}\|_1 \leq (1+\triangle)^{c-1}\delta_i^{root} + (1+\triangle)^{c-1} - 1 \tag{31}$$

$\boldsymbol{x}_{1:c}$ is all the observed variables in cliques 1 through c. Note that when $c = |\mathbb{C}|$ then this implies that $\widehat{\Phi}_{1:c}(\mathcal{C}) = \widehat{\mathbb{P}}[x_1, ..., x_O]$ and thus,

$$\sum_{\boldsymbol{x}} |\widehat{\mathbb{P}}[x_1, ..., x_O] - \mathbb{P}[x_1, ..., x_O]| \leq (1+\triangle)^{|\mathbb{C}|-1}\delta_i^{root} + (1+\triangle)^{|\mathbb{C}|-1} - 1 \tag{32}$$

Proof The proof is by induction on c. The base case follows trivially from the definition of δ_i^{root}. For the induction step, assume the claim holds for $c \geq 1$. Then we prove it holds for $c+1$.

$$\begin{aligned}
&\sum_{\boldsymbol{x}_{1:c+1}} \|(\breve{\Phi}_{1:c+1}(\mathcal{C}) - \widehat{\Phi}_{1:c+1}(\mathcal{C})) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_c}} \breve{\boldsymbol{\mathcal{F}}}_{d_{c+1}}^{-1}\|_1 \\
= &\sum_{\boldsymbol{x}_{1:c+1}} \Big|\Big|\Big(\breve{\Phi}_{1:c}(\mathcal{C}) \times (\widehat{\mathcal{P}}(\mathcal{C}_{c+1}) - \breve{\mathcal{P}}(\mathcal{C}_{c+1})) + (\breve{\Phi}_{1:c}(\mathcal{C}) - \widehat{\Phi}_{1:c}(\mathcal{C})) \times (\widehat{\mathcal{P}}(\mathcal{C}_{c+1}) - \breve{\mathcal{P}}(\mathcal{C}_{c+1})) + (\breve{\Phi}_{1:c}(\mathcal{C}) - \widehat{\Phi}_{1:c}(\mathcal{C})) \times \breve{\mathcal{P}}(\mathcal{C}_{c+1})\Big) \\
&\qquad \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_{c+1}}} \breve{\boldsymbol{\mathcal{F}}}_{d_{c+1}}^{-1}\Big|\Big|_1 \\
\leq &\sum_{\boldsymbol{x}_{1:c+1}} \|(\breve{\Phi}_{1:c}(\mathcal{C}) \times (\widehat{\mathcal{P}}(\mathcal{C}_{c+1}) - \breve{\mathcal{P}}(\mathcal{C}_{c+1}))) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_{c+1}}} \breve{\boldsymbol{\mathcal{F}}}_{d_{c+1}}^{-1}\|_1 + \\
&\sum_{\boldsymbol{x}_{1:c+1}} \|((\breve{\Phi}_{1:c}(\mathcal{C}) - \widehat{\Phi}_{1:c}(\mathcal{C})) \times (\widehat{\mathcal{P}}(\mathcal{C}_{c+1}) - \breve{\mathcal{P}}(\mathcal{C}_{c+1}))) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_{c+1}}} \breve{\boldsymbol{\mathcal{F}}}_{d_{c+1}}^{-1}\|_1 + \\
&\sum_{\boldsymbol{x}_{1:c+1}} \|((\breve{\Phi}_{1:c}(\mathcal{C}) - \widehat{\Phi}_{1:c}(\mathcal{C}) \times \breve{\mathcal{P}}(\mathcal{C}_{c+1})) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_{c+1}}} \breve{\boldsymbol{\mathcal{F}}}_{d_{c+1}}^{-1}\|_1
\end{aligned}$$

Now we consider two cases, when $c+1$ is a leaf clique and when it is an internal clique.

Case 1: Internal Clique

Note here the summation over $\boldsymbol{x}_{1:c+1}$ is irrelevant since there is no evidence. We use Lemma 5 to break up the three terms:

The first term,

$$\begin{aligned}
&\|(\breve{\Phi}_{1:c}(\mathcal{C}) \times (\widehat{\mathcal{P}}(\mathcal{C}_{c+1}) - \breve{\mathcal{P}}(\mathcal{C}_{c+1}))) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_{c+1}}} \breve{\boldsymbol{\mathcal{F}}}_{d_{c+1}}^{-1}\|_1 \\
\leq &\Big|\Big|\Big|(\widehat{\mathcal{P}}(\mathcal{C}_{c+1}) - \breve{\mathcal{P}}(\mathcal{C}_{c+1})) \times_{\mathscr{O}_{-(c+1)}} \breve{\boldsymbol{\mathcal{F}}}_{c+1} \times_{\mathscr{O}_{\alpha_{c+1}(1)}} \breve{\boldsymbol{\mathcal{F}}}_{\alpha_{c+1}(1)}^{-1}, ..., \times_{\mathscr{O}_{\alpha_{c+1}(\gamma_{c+1})}} \breve{\boldsymbol{\mathcal{F}}}_{\alpha_{c+1}(\gamma_{c+1})}^{-1}\Big|\Big|\Big|_1^{S_i} \\
\times &\|\breve{\Phi}_{1:c}(\mathcal{C}) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_c}} \breve{\boldsymbol{\mathcal{F}}}_{d_c}^{-1}\|_1 \\
\leq &\triangle \times 1
\end{aligned}$$

The first term above is simply $\delta_i^{internal} \leq \triangle$ while the second equals one since it is a joint distribution.

Now for the second term,

$$\begin{aligned}
& \|(\breve{\Phi}_{1:c}(\mathcal{C}) - \widehat{\Phi}_{1:c}(\mathcal{C})) \times (\widehat{\mathcal{P}}(\mathcal{C}_{c+1}) - \breve{\mathcal{P}}(\mathcal{C}_{c+1}))) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_{c+1}}} \breve{\boldsymbol{\mathcal{F}}}_{d_{c+1}}^{-1} \|_1 \\
\leq \; & \|(\breve{\Phi}_{1:c}(\mathcal{C}) - \widehat{\Phi}_{1:c}(\mathcal{C})) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_c}} \breve{\boldsymbol{\mathcal{F}}}_{d_c}^{-1} \|_1 \times \\
& \quad \left|\left|\left|(\widehat{\mathcal{P}}(\mathcal{C}_{c+1}) - \breve{\mathcal{P}}(\mathcal{C}_{c+1})) \times_{\mathscr{O}_{-(c+1)}} \breve{\boldsymbol{\mathcal{F}}}_{c+1} \times_{\mathscr{O}_{\alpha_{c+1}(1)}} \breve{\boldsymbol{\mathcal{F}}}_{\alpha_{c+1}(1)}^{-1}, ..., \times_{\mathscr{O}_{\alpha_{c+1}(\gamma_{c+1})}} \breve{\boldsymbol{\mathcal{F}}}_{\alpha_{c+1}(\gamma_{c+1})}^{-1} \right|\right|\right|_1^{S_{c+1}} \\
\leq \; & ((1+\triangle)^{c-1}\delta^{root} + (1+\triangle)^{c-1} - 1) \times \delta_{c+1}^{internal} \qquad \text{(via induction hypothesis)} \\
\leq \; & ((1+\triangle)^{c-1}\delta^{root} + (1+\triangle)^{c-1} - 1) \times \triangle
\end{aligned}$$

The third term,

$$\begin{aligned}
& \|((\breve{\Phi}_{1:c}(\mathcal{C}) - \widehat{\Phi}_{1:c}(\mathcal{C})) \times \breve{\mathcal{P}}(\mathcal{C}_{c+1})) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_{c+1}}} \breve{\boldsymbol{\mathcal{F}}}_{d_{c+1}}^{-1} \|_1 \\
& \leq \|(\breve{\Phi}_{1:c}(\mathcal{C}) - \widehat{\Phi}_{1:c}(\mathcal{C})) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_c}} \breve{\boldsymbol{\mathcal{F}}}_{d_c}^{-1} \|_1 \times \\
& \left|\left|\left|\breve{\mathcal{P}}(\mathcal{C}_{c+1}) \times_{\mathscr{O}_{-(c+1)}} \breve{\boldsymbol{\mathcal{F}}}_{c+1} \times_{\mathscr{O}_{\alpha_{c+1}(1)}} \breve{\boldsymbol{\mathcal{F}}}_{\alpha_{c+1}(1)}^{-1}, ..., \times_{\mathscr{O}_{\alpha_{c+1}(\gamma_{c+1})}} \breve{\boldsymbol{\mathcal{F}}}_{\alpha_{c+1}(\gamma_{c+1})}^{-1} \right|\right|\right|_1^{S_{c+1}} \\
& \leq ((1+\triangle)^{c-1}\delta^{root} + (1+\triangle)^{c-1} - 1) \times 1
\end{aligned}$$

Case 2: Leaf Clique

Again we use Lemma 5 to break up the three terms:

The first term,

$$\begin{aligned}
& \sum_{\boldsymbol{x}_{1:c+1}} \|(\breve{\Phi}_{1:c}(\mathcal{C}) \times (\widehat{\mathcal{P}}_{r_i}(\mathcal{C}_{c+1}) - \breve{\mathcal{P}}_{r_i}(\mathcal{C}_{c+1}))) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_{c+1}}} \breve{\boldsymbol{\mathcal{F}}}_{d_{c+1}}^{-1} \|_1 \\
\leq \; & \sum_{\boldsymbol{x}_{1:c+1}} \|\breve{\Phi}_{1:c}(\mathcal{C}) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_c}} \breve{\boldsymbol{\mathcal{F}}}_{d_c}^{-1} \|_1 \left|\left|\left|(\widehat{\mathcal{P}}_{r_i}(\mathcal{C}_{c+1}) - \breve{\mathcal{P}}_{r_i}(\mathcal{C}_{c+1})) \times_{\mathscr{O}_{-(c+1)}} \breve{\boldsymbol{\mathcal{F}}}_{c+1} \right|\right|\right|_1^{S_{c+1}} \\
\leq \; & \sum_{\boldsymbol{x}_{1:c}} \|\breve{\Phi}_{1:c}(\mathcal{C}) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_c}} \breve{\boldsymbol{\mathcal{F}}}_{d_c}^{-1} \|_1 \sum_{\boldsymbol{x}_{c+1}} \left|\left|\left|(\widehat{\mathcal{P}}_{r_i}(\mathcal{C}_{c+1}) - \breve{\mathcal{P}}_{r_i}(\mathcal{C}_{c+1})) \times_{\mathscr{O}_{-(c+1)}} \breve{\boldsymbol{\mathcal{F}}}_{c+1} \right|\right|\right|_1^{S_{c+1}} \\
\leq \; & 1 \times \triangle
\end{aligned}$$

The first term above equals 1 because it is a joint distribution and the second is the bound on the transformed quantity we had proved earlier (since $r_i = \boldsymbol{x}_{c+1}$).

The second term,

$$\begin{aligned}
& \sum_{\boldsymbol{x}_{1:c+1}} \|(\breve{\Phi}_{1:c}(\mathcal{C}) - \widehat{\Phi}_{1:c}(\mathcal{C})) \times (\widehat{\mathcal{P}}_{r_i}(\mathcal{C}_{c+1}) - \breve{\mathcal{P}}_{r_i}(\mathcal{C}_{c+1}))) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_{c+1}}} \breve{\boldsymbol{\mathcal{F}}}_{d_{c+1}}^{-1} \|_1 \\
\leq \; & \sum_{\boldsymbol{x}_{1:c}} \|(\breve{\Phi}_{1:c}(\mathcal{C}) - \widehat{\Phi}_{1:c}(\mathcal{C})) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_c}} \breve{\boldsymbol{\mathcal{F}}}_{d_c}^{-1} \|_1 \sum_{\boldsymbol{x}_{c+1}} \left|\left|\left|(\widehat{\mathcal{P}}_{r_i}(\mathcal{C}_{c+1}) - \breve{\mathcal{P}}_{r_i}(\mathcal{C}_{c+1})) \times_{\mathscr{O}_{-(c+1)}} \breve{\boldsymbol{\mathcal{F}}}_{c+1} \right|\right|\right|_1 \\
\leq \; & ((1+\triangle)^{c-1}\delta^{root} + (1+\triangle)^{c-1} - 1) \times \xi_{c+1}^{leaf} \\
\leq \; & ((1+\triangle)^{c-1}\delta^{root} + (1+\triangle)^{c-1} - 1) \times \triangle
\end{aligned}$$

The third term,

$$\begin{aligned}
& \sum_{\boldsymbol{x}_{1:c+1}} \|((\breve{\Phi}_{1:c}(\mathcal{C}) - \widehat{\Phi}_{1:c}(\mathcal{C}) \times \breve{\mathcal{P}}_{r_i}(\mathcal{C}_{c+1})) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_{c+1}}} \breve{\boldsymbol{\mathcal{F}}}_{d_{c+1}}^{-1} \|_1 \\
& \leq \sum_{\boldsymbol{x}_{1:c}} \|(\breve{\Phi}_{1:c}(\mathcal{C}) - \widehat{\Phi}_{1:c}(\mathcal{C})) \times_{\mathscr{M}_1} \breve{\boldsymbol{\mathcal{F}}}_1^{-1}, ..., \times_{\mathscr{M}_{d_c}} \breve{\boldsymbol{\mathcal{F}}}_{d_c}^{-1} \|_1 \sum_{\boldsymbol{x}_{c+1}} \left|\left|\left|\breve{\mathcal{P}}_{r_i}(\mathcal{C}_{c+1}) \times_{\mathscr{O}_{-(c+1)}} \breve{\boldsymbol{\mathcal{F}}}_{c+1} \right|\right|\right|_1^{S_{c+1}} \\
& \leq ((1+\triangle)^{c-1}\delta^{root} + (1+\triangle)^{c-1} - 1) \times 1
\end{aligned}$$

Combining these terms proves the induction step. ■

3.7 Putting it all together

We use the fact from HKZ [2] that $(1 + a/t)^t \leq 1 + 2a$ for $a \leq 1/2$. Now $\triangle$ is the main source of error. We set $\triangle \leq O(\epsilon_{total}/|\mathbb{C}|)$.

Note that

$$\triangle \leq \frac{2^{d_{\max}+3} k_h^{d_{\max}}}{3\sqrt{3}^{d_{max}} \beta^{d_{\max}}} \left(\frac{\sum_{\boldsymbol{o}_{d-e+1},...,\boldsymbol{o}_d} \epsilon(\mathscr{O}_1, ..., \mathscr{O}_{d-e}, \boldsymbol{o}_{d-e+1}, ..., \boldsymbol{o}_d)}{\alpha} + \frac{\sum_{\boldsymbol{o}_{d-e+1},...,\boldsymbol{o}_d} \epsilon(\mathscr{O}_1, ..., \mathscr{O}_{d-e}, \boldsymbol{o}_{d-e+1}, ..., \boldsymbol{o}_d)}{\alpha^2} \right)$$

This gives,

$$\frac{2^{d_{\max}+3} k_h^{d_{\max}}}{3\sqrt{3}^{d_{max}} \beta^{d_{max}}} \left(\frac{\sum_{\boldsymbol{o}_{d-e+1},...,\boldsymbol{o}_d} \epsilon(\mathscr{O}_1, ..., \mathscr{O}_{d-e}, \boldsymbol{o}_{d-e+1}, ..., \boldsymbol{o}_d)}{\alpha} + \frac{\sum_{\boldsymbol{o}_{d-e+1},...,\boldsymbol{o}_d} \epsilon(\mathscr{O}_1, ..., \mathscr{O}_{d-e}, \boldsymbol{o}_{d-e+1}, ..., \boldsymbol{o}_d)}{\alpha^2} \right) \leq K\epsilon_{total}/|\mathbb{C}|$$

where K is some constant.

This implies,

$$\sum_{\boldsymbol{o}_{d-e+1},...,\boldsymbol{o}_d} \epsilon(\mathscr{O}_1, ..., \mathscr{O}_{d-e}, \boldsymbol{o}_{d-e+1}, ..., \boldsymbol{o}_d) \leq K \frac{3^{d_{max}/2+1} \epsilon_{total} \alpha^2 \beta^{d_{max}}}{2^{d_{max}+3} k_h^{d_{max}} |\mathbb{C}|} \tag{33}$$

Now using the concentration bound (Lemma 1) will give,

$$K \frac{3^{d_{max}/2+1} \epsilon_{total} \alpha^2 \beta^{d_{max}}}{2^{d_{max}+3} k_h^{d_{max}} |\mathbb{C}|} \leq \sqrt{\frac{k_o^{e_{\max}}}{N} \ln \frac{2|\mathbb{C}|}{\delta}} + \sqrt{\frac{k_o^{e_{max}}}{N}}$$

Solving for N:

$$N \geq O\left(\left(\frac{4k_h^2}{3\beta^2} \right)^{d_{max}} \frac{k_o^{e_{max}} \ln \frac{|\mathbb{C}|}{\delta} |\mathbb{C}|^2}{\epsilon_{total}^2 \alpha^4} \right) \tag{34}$$

and this completes the proof.

4 Appendix

4.1 Matrix Perturbation Bounds

This is Theorem 3.8 from pg. 143 in Stewart and Sun, 1990 [4]. Let $A \in \boldsymbol{R}^{m \times n}$, with $m \geq n$ and let $\tilde{A} = A + E$. Then

$$\left|\left| \tilde{A}^+ - A^+ \right|\right|_2 \leq \frac{1 + \sqrt{5}}{2} \max(\left|\left| A^+ \right|\right|_2^2, \left|\left| \tilde{A} \right|\right|_2^2) \left|\left| E \right|\right|_2 \tag{35}$$

4.2 Tensor Norm Bounds

For matrices it is true that $\|\boldsymbol{M}\boldsymbol{v}\|_1 \leq |||\boldsymbol{M}|||_1 \|\boldsymbol{v}\|_1$. We prove the generalization to tensors.

Lemma 5 *Let $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ be tensors $\boldsymbol{\sigma}$ a set of (labeled) modes.*

$$\left|\left| \boldsymbol{T}_1 \times_{\boldsymbol{\sigma}} \boldsymbol{T}_2 \right|\right|_1 \leq |||\boldsymbol{T}_2|||_1^{\boldsymbol{\sigma}} \|\boldsymbol{T}_1\|_1 \tag{36}$$

Proof

$$\begin{aligned}
||\boldsymbol{T}_1 \times_{\boldsymbol{\sigma}} \boldsymbol{T}_2||_1 &= \sum_{\boldsymbol{i}_{1:N}\setminus\boldsymbol{\sigma}} \sum_{\boldsymbol{j}_{1:M\setminus\boldsymbol{\sigma}}} \sum_{\boldsymbol{x}} \boldsymbol{T}_1(\boldsymbol{i}_{1:N}\setminus\boldsymbol{\sigma}, \boldsymbol{\sigma}=\boldsymbol{x})\boldsymbol{T}_2(\boldsymbol{j}_{1:N}\setminus\boldsymbol{\sigma}, \boldsymbol{\sigma}=\boldsymbol{x}) \\
&= \sum_{\boldsymbol{i}_{1:N}\setminus\boldsymbol{\sigma}} \sum_{\boldsymbol{x}} \sum_{\boldsymbol{j}_{1:M\setminus\boldsymbol{\sigma}}} \boldsymbol{T}_1(\boldsymbol{i}_{1:N}\setminus\boldsymbol{\sigma}, \boldsymbol{\sigma}=\boldsymbol{x})\boldsymbol{T}_2(\boldsymbol{j}_{1:N}\setminus\boldsymbol{\sigma}, \boldsymbol{\sigma}=\boldsymbol{x}) \\
&= \sum_{\boldsymbol{i}_{1:N}\setminus\boldsymbol{\sigma}} \sum_{\boldsymbol{\sigma}} \boldsymbol{T}_1(\boldsymbol{i}_{1:N}\setminus\boldsymbol{\sigma}, \boldsymbol{\sigma}=\boldsymbol{x}) \sum_{\boldsymbol{j}_{1:M\setminus\boldsymbol{\sigma}}} \boldsymbol{T}_2(\boldsymbol{j}_{1:N}\setminus\boldsymbol{\sigma}, \boldsymbol{\sigma}=\boldsymbol{x}) \\
&\leq \max_{\boldsymbol{x}} \left(\sum_{\boldsymbol{j}_{1:M\setminus\boldsymbol{\sigma}}} \boldsymbol{T}_2(\boldsymbol{j}_{1:N}\setminus\boldsymbol{\sigma}, \boldsymbol{\sigma}=\boldsymbol{x}) \right) \|\boldsymbol{T}_1\|_1 \\
&= |||\boldsymbol{T}_2|||_1^{\boldsymbol{\sigma}} \, \|\boldsymbol{T}_1\|_1
\end{aligned}$$

■

We prove a restricted analog of the fact that spectral norm is submultiplicative for matrices i.e. $|||AB|||_2 \leq |||A|||_2 \, |||B|||_2$.

Lemma 6 *Let $\boldsymbol{T}$ be a tensor of order N and let $\boldsymbol{M}$ be a matrix. Then,*

$$|||\boldsymbol{T} \times_1 \boldsymbol{M}|||_2 \leq |||\boldsymbol{T}|||_2 \, |||\boldsymbol{M}|||_2 \tag{37}$$

Proof

$$\begin{aligned}
|||\boldsymbol{T} \times_1 \boldsymbol{M}|||_2 &= \sup_{\boldsymbol{v}_m, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N} \sum_{i_1,\ldots,i_N,m} \boldsymbol{T}(i_1, i_2, ..., i_N)\boldsymbol{M}(i_1, m)\boldsymbol{v}_m(m)\boldsymbol{v}_{i_2}(i_2)\boldsymbol{v}_{i_3}(i_3)....\boldsymbol{v}_{i_n}(i_n) \\
&= \sup_{\boldsymbol{v}_m, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N} \sum_{i_2,\ldots,i_N} \sum_{i_1} \boldsymbol{T}(i_1, i_2, ..., i_N) \sum_m \boldsymbol{M}(i_1, m)\boldsymbol{v}_m(m) \\
&\leq \sup_{\boldsymbol{v}_m, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N} \left(\sup_{i_1} \left|\left| \sum_m \boldsymbol{M}(i_1, m)\boldsymbol{v}_m(m) \right|\right|_2 \right) \times \\
&\quad \sum_{i_1,\ldots,i_N} \boldsymbol{T}(i_1, i_2, ..., i_N) \frac{1}{||\sum_m \boldsymbol{M}(i_1, m)\boldsymbol{v}_m(m)||_2} \left(\sum_m \boldsymbol{M}(i_1, m)\boldsymbol{v}_m(m) \right) \boldsymbol{v}_{i_2}(i_2)\boldsymbol{v}_{i_3}(i_3)....\boldsymbol{v}_{i_n}(i_n) \\
&\leq |||\boldsymbol{M}|||_2 \, |||\boldsymbol{T}|||_2
\end{aligned}$$

■

Lemma 7 *Let $\boldsymbol{T}$ be a tensor of order N. Then,*

$$|||\boldsymbol{T}|||_2 \leq ||\boldsymbol{T} \times_1 \boldsymbol{v}_1, ..., \times_{n-1}\boldsymbol{v}_{n-1}||_F \leq ||T \times_1 \boldsymbol{v}_1, ..., \times_{n-2}\boldsymbol{v}_{n-2}||_F \leq ... \leq ||T||_F \tag{38}$$

Proof *It suffices to show that $\sup_{v\, s.t.\, ||v||\leq 1} ||\boldsymbol{T} \times_1 \boldsymbol{v}||_F \leq ||\boldsymbol{T}||_F$. By submultiplicativity of the frobenius norm: $||\boldsymbol{T} \times_1 \boldsymbol{v}||_F \leq ||\boldsymbol{T}||_F \, ||\boldsymbol{v}||_F \leq ||\boldsymbol{T}||$, since $||\boldsymbol{v}||_F = ||\boldsymbol{v}||_2 \leq 1$.* ■

Lemma 8 *Let $\boldsymbol{T}$ be a tensor of order N, where each mode is of dimension k. Then,*

$$|||\boldsymbol{T}|||_1^{\boldsymbol{\sigma}} \leq k^N \, |||\boldsymbol{T}|||_2 \tag{39}$$

For any $\boldsymbol{\sigma}$.

Proof *We simply prove this for $\boldsymbol{\sigma} = \emptyset$ (which corresponds to elementwise one norm) since $|||\boldsymbol{T}|||_1^{\boldsymbol{\sigma}_1} \leq |||\boldsymbol{T}|||_1^{\boldsymbol{\sigma}_2}$ if $\boldsymbol{\sigma}_2 \subseteq \boldsymbol{\sigma}_1$. Note that $\|\boldsymbol{T}\|_1 \leq k^N \max(|\boldsymbol{T}|)$ (where $\max(\boldsymbol{T})$ is the maximum element of $|\boldsymbol{T}|$). Similarly, $\max(|\boldsymbol{T}|) \leq |||\boldsymbol{T}|||_2$ which implies that $\|\boldsymbol{T}\|_1 \leq k^N \, |||\boldsymbol{T}|||_2$.* ■

References

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

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

[3] 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.

[4] GW Stewart and J. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.