Supplementary Material for: Spectral Unsupervised Parsing with Additive Tree Metrics

Ankur P. Parikh
School of Computer Science
Carnegie Mellon University
apparikh@cs.cmu.edu

Shay B. Cohen
School of Informatics
University of Edinburgh
scohen@inf.ed.ac.uk

Eric P. Xing
School of Computer Science
Carnegie Mellon University
epxing@cs.cmu.edu

[ ]

The primary purpose of the supplemental is to provide the theoretical arguments that our algorithm is correct. We first give the proof that our proposed tree metric is indeed tree additive. We then analyze the consistency of Algorithm 1.

1 Path Additivity

We first prove that our proposed tree metric is path additive based on the proof technique in Song et al. (2011).

Lemma 1. *If Assumption 1 in the main paper holds then, d^{spectral} is an additive metric.*

Proof. For conciseness, we simply prove the property for paths of length 2. The proof for more general cases follows similarly (e.g. see Anandkumar et al. (2011)).

First note that the relationship between eigenvalues and singular values allows us to rewrite the distance metric as

$$d^{\text{spectral}}(i,j) = -\frac{1}{2}\log\Lambda_m(\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)^\top) + \tfrac{1}{4}\log\Lambda_m(\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,i)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,i)^\top) + \tfrac{1}{4}\log\Lambda_m(\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)^\top)$$

Furthermore, $\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)^\top$ is rank m by Assumption 1 and the conditional independence statements implied by the latent tree model. Thus $\Lambda_m(\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)^\top)$ is equivalent to taking the pseudo-determinant $|\cdot|_+$ of $(\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)^\top)$, which is the product of the non-zero eigenvalues. The pseudo-determinant can be alternatively defined in the following limit form:

$$|\boldsymbol{B}|_+ = \lim_{\alpha\to 0}\frac{|\boldsymbol{B}+\alpha\boldsymbol{I}|}{\alpha^{p-m}}$$

where $\boldsymbol{B}$ is a $p\times p$ matrix of rank m and $\boldsymbol{I}$ is the $p\times p$ identity and $|\cdot|$ equals the standard determinant. Note that if $m=p$ then $|\boldsymbol{B}|_+ = |\boldsymbol{B}|$.

Thus, our distance metric can be further rewritten as:

$$d^{\text{spectral}}(i,j) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)^\top|_+ + \tfrac{1}{4}\log|\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,i)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,i)^\top|_+ + \tfrac{1}{4}\log|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)^\top|_+ \quad (1)$$

We now proceed with the proof. For paths of length 2, $i-j-k$, there are three cases[1]:

- $i \leftarrow j \rightarrow k$
- $i \leftarrow j \leftarrow k$
- $i \rightarrow j \rightarrow k$

[1] although the additive distance metric is undirected, $\boldsymbol{A}$ and $\boldsymbol{C}$ in Assumption 1 are defined with respect to parent-child relationships, so we must consider direction

Case 1: ($i \leftarrow j \rightarrow k$) Note that in this case j must be latent but i and k can be either observed or latent. We assume below that both i and k are observed. The same proof strategy works for the other cases.

Due to Assumption 1,

$$\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k) = \mathbb{E}[v_i v_k^\top | \boldsymbol{x}] = \boldsymbol{C}_{i|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j) \boldsymbol{C}_{k|j,\boldsymbol{x}}^\top$$

Thus,

$$\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k)^\top = \boldsymbol{C}_{i|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j) \boldsymbol{C}_{k|j,\boldsymbol{x}}^\top \boldsymbol{C}_{k|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j) \boldsymbol{C}_{i|j,\boldsymbol{x}}^\top$$

Combining this definition with Sylvester's Determinant Theorem (Akritas et al., 1996), gives us that:

$$\begin{aligned} |\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k)^\top|_+ &= |\boldsymbol{C}_{i|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j) \boldsymbol{C}_{k|j,\boldsymbol{x}}^\top \boldsymbol{C}_{k|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j) \boldsymbol{C}_{i|j,\boldsymbol{x}}^\top|_+ \\ &= |\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top \boldsymbol{C}_{i|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j) \boldsymbol{C}_{k|j,\boldsymbol{x}}^\top \boldsymbol{C}_{k|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|_+ \end{aligned}$$

(i.e. we can move $\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top$ the to the front).

Now $\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top \boldsymbol{C}_{i|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j) \boldsymbol{C}_{k|j,\boldsymbol{x}}^\top \boldsymbol{C}_{k|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)$ is $m \times m$ and has rank m. Thus, the pseudo-determinant equals the normal determinant in this case. Using the fact that $|\boldsymbol{A}\boldsymbol{B}| = |\boldsymbol{A}||\boldsymbol{B}|$ if $\boldsymbol{A}$ and $\boldsymbol{B}$ are square, we get

$$\begin{aligned} |\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k)^\top|_+ &= |\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top \boldsymbol{C}_{i|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j) \boldsymbol{C}_{k|j,\boldsymbol{x}}^\top \boldsymbol{C}_{k|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)| \\ &= |\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top \boldsymbol{C}_{i|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)||\boldsymbol{C}_{k|j,\boldsymbol{x}}^\top \boldsymbol{C}_{k|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)| \\ &= \frac{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)||\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top \boldsymbol{C}_{i|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|}{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|} \times \frac{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)||\boldsymbol{C}_{k|j,\boldsymbol{x}}^\top \boldsymbol{C}_{k|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|}{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|} \\ &= \frac{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top \boldsymbol{C}_{i|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|}{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|} \times \frac{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)\boldsymbol{C}_{k|j,\boldsymbol{x}}^\top \boldsymbol{C}_{k|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|}{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|} \end{aligned}$$

Furthermore, note that

$$\begin{aligned} |\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top \boldsymbol{C}_{i|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)| &= |\boldsymbol{C}_{i|j,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j) \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j) \boldsymbol{C}_{i|j,\boldsymbol{x}}^\top|_+ \\ &= |\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)^\top|_+ \end{aligned}$$

This gives,

$$|\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k)^\top|_+ = \frac{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)^\top|_+}{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|} \times \frac{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,j)^\top|_+}{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|}$$

Substituting back into Eq. 1 proves that

$$\begin{aligned} d^{\text{spectral}}(i,k) &= -\frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)^\top|_+ - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,k)^\top|_+ + \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)^\top|_+ \\ &+ \tfrac{1}{4}\log|\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,i)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,i)^\top|_+ + \tfrac{1}{4}\log|\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,k)^\top|_+ \\ &= d^{\text{spectral}}(i,j) + d^{\text{spectral}}(j,k) \end{aligned}$$

Case 2: $i \leftarrow j \leftarrow k$ The proof is similar to above. Here since only leaf nodes can be observed,j and k must be latent but i can be either observed or latent. We assume it is observed, the latent case follows similarly.

Due to Assumption 1,

$$\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k) = \mathbb{E}[v_i v_k^\top | x] = \boldsymbol{C}_{i|j,\boldsymbol{x}} \boldsymbol{A}_{j|k,\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}}(k,k)$$

Thus, via Sylvester's Determinant Theorem as before,

$$\begin{aligned}|\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k)^\top|_+ &= |\boldsymbol{C}_{i|j,\boldsymbol{x}}\boldsymbol{A}_{j|k,\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,k)\boldsymbol{A}_{j|k,\boldsymbol{x}}^\top\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top|_+ \\ &= |\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top\boldsymbol{C}_{i|j,\boldsymbol{x}}\boldsymbol{A}_{j|k,\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,k)\boldsymbol{A}_{j|k,\boldsymbol{x}}^\top|_+\end{aligned}$$

Now $\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top\boldsymbol{C}_{i|j,\boldsymbol{x}}\boldsymbol{A}_{j|k,\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,k)\boldsymbol{A}_{j|k,\boldsymbol{x}}^\top$ is $m \times m$ and has rank m. Thus, the pseudo-determinant equals the normal determinant in this case. Using the fact that $|\boldsymbol{A}\boldsymbol{B}| = |\boldsymbol{A}||\boldsymbol{B}|$ if $\boldsymbol{A}$ and $\boldsymbol{B}$ are square, we get

$$\begin{aligned}|\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k)^\top|_+ &= |\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top\boldsymbol{C}_{i|j,\boldsymbol{x}}\boldsymbol{A}_{j|k,\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,k)\boldsymbol{A}_{j|k,\boldsymbol{x}}^\top| \\ &= |\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top\boldsymbol{C}_{i|j,\boldsymbol{x}}||\boldsymbol{A}_{j|k,\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,k)\boldsymbol{A}_{j|k,\boldsymbol{x}}^\top| \\ &= |\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top\boldsymbol{C}_{i|j,\boldsymbol{x}}||\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,k)^\top| \\ &= \frac{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)||\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top\boldsymbol{C}_{i|j,\boldsymbol{x}}||\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|}{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)||\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|} \times |\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,k)^\top| \\ &= \frac{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top\boldsymbol{C}_{i|j,\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|}{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)||\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|} \times |\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,k)^\top|\end{aligned}$$

Furthermore, note that

$$\begin{aligned}|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top\boldsymbol{C}_{i|j,\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)| &= |\boldsymbol{C}_{i|j,\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)\boldsymbol{C}_{i|j,\boldsymbol{x}}^\top)|_+ \\ &= |\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)^\top|_+\end{aligned}$$

This gives,

$$|\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,k)^\top|_+ = \frac{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)^\top|_+}{|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)||\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)|} \times |\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,k)^\top|_+$$

Substituting back into Eq. 1 proves that

$$\begin{aligned}d^{\text{spectral}}(i,k) &= -\frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,j)^\top|_+ - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,k)^\top|_+ + \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,j)^\top|_+ \\ &+ \tfrac{1}{4}\log|\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,i)\boldsymbol{\Sigma}_{\boldsymbol{x}}(i,i)^\top|_+ + \tfrac{1}{4}\log|\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,k)\boldsymbol{\Sigma}_{\boldsymbol{x}}(k,k)^\top|_+ \\ &= d^{\text{spectral}}(i,j) + d^{\text{spectral}}(j,k)\end{aligned}$$

Case 3: $i \to j \to k$ The same argument as case 2 holds here.

□

2 Theoretical Guarantees For Algorithm 1

Our main theoretical guarantee is that the learning Algorithm 1 will recover the correct undirected tree $u \in \mathcal{U}$ with high probability, if the given top bracket is correct and if we obtain a sufficient number of examples $(\boldsymbol{w}^{(i)}, \boldsymbol{x}^{(i)})$ being generated from the model in §2.

Our kernel is controlled by its "bandwidth" γ, a typical kernel parameter that appears when using kernel smoothing. The larger this positive number is, the more inclusive the kernel will be with respect to examples in $\mathcal{D}$ in order to estimate a given covariance matrix.

In order for the learning algorithm to be consistent the kernel must be able to effectively manage the bias-variance trade-off with the bandwidth parameter. Intuitively, if the sample size is small, then the bandwidth should be relatively large to control the variance. As the sample size increases, the bandwidth should be decreased at a certain rate to reduce the bias.

In kernel density estimation in $\mathbb{R}^p$, one can put smoothness assumptions on the density being estimated, such as bounded second derivatives. With these conditions, it can be shown that setting $\gamma = \mathcal{O}(N^{-1/5})$ will optimally trade-off the bias and the variance in a way that leads to a consistent estimator.

However, our space of possible sequences is discrete and thus it is much more difficult to define these analogous smoothness conditions. Therefore, giving an asymptotic rate to set the bandwidth as a function of the sample size is difficult. To solve this issue, we consider a specific theoretical kernel where the bandwidth directly relates to the bias of the estimator. Define the following γ-ball

$$B_{j,k,\boldsymbol{x}}(\gamma) = \{(j',k',\boldsymbol{x}') : \|\boldsymbol{\Sigma}_{\boldsymbol{x}'}(j',k') - \boldsymbol{\Sigma}_{\boldsymbol{x}}(j,k)\|_F \leq \gamma\}$$

where $\|A\|_F$ for a matrix A is the Frobenius norm of that matrix, i.e.: $\|A\|_F = \sqrt{\sum_{j,k} A_{jk}^2}$.

We then define the following kernel:

$$\begin{aligned} & K_\gamma(j,k,j',k'|\boldsymbol{x},\boldsymbol{x}') \\ & = \begin{cases} 1 & : (j',k',\boldsymbol{x}') \in B_{(j,k,\boldsymbol{x})}(\gamma) \\ 0 & : (j',k',\boldsymbol{x}') \notin B_{(j,k,\boldsymbol{x})}(\gamma) \end{cases} \end{aligned} \tag{30}$$

Furthermore, to quantify the expected effective sample size using kernel smoothing we define the following quantity:

$$\begin{aligned} & \nu_{j,k,\boldsymbol{x}}(\gamma) = \\ & \left(\sum_{\boldsymbol{x}' \in T^*} \frac{p(\boldsymbol{x}')}{\ell(\boldsymbol{x}')^2} \sum_{j',k' \in [\ell(\boldsymbol{x}')]} K_\gamma(j,k,j',k'|\boldsymbol{x},\boldsymbol{x}') \right) \end{aligned}$$

where $p(\boldsymbol{x})$ the prior distribution over tag sequences. Let $\nu_{\boldsymbol{x}}(\gamma) = \min_{j,k} \nu_{j,k,\boldsymbol{x}}(\gamma)$. Intuitively, $\nu_{j,k,\boldsymbol{x}}(\gamma)$ represents the probability of finding contexts that are "similar" to $(j,k,\boldsymbol{x})$ as defined by the kernel. Consequently, $N\nu_{\boldsymbol{x}}(\gamma)$ represents a lower bound on the expected effective sample size for tag sequence $\boldsymbol{x}$.

Denote $\sigma_{\boldsymbol{x}}(j,k)^{(r)}$ as the r^{th} singular value of $\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,k)$. Let $\sigma^*(x) := \min_{j,k \in \ell(\boldsymbol{x})} \min \left(\sigma_{\boldsymbol{x}}(j,k)^{(m)} \right)$. Finally, define ϕ as the difference of the maximum possible entry in $\boldsymbol{w}$ and the minimum possible entry in $\boldsymbol{w}$ (i.e. we assume that the embeddings are bounded vectors).

Using the above definitions and leveraging the proof technique in Zhou et al. (2010), we establish that our strategy will result in a consistent estimation of the word-word distance subblock $\boldsymbol{D}_{WW}$.

Lemma 2. *Assume the kernel used is that in Eq.* 30 *with bandwidth* $\gamma = \frac{C_1 \epsilon \sigma^*(x)}{m}$ *and that*

$$N \geq \frac{m^2 \phi^2}{C_3 \epsilon^2 \sigma^*(\boldsymbol{x})^2 \nu_{\boldsymbol{x}}(\gamma)^2} \log \left(\frac{C_2 p^2 \ell(\boldsymbol{x})^2}{\delta} \right)$$

Then, for a fixed tag sequence $\boldsymbol{x}$, *and any* $\epsilon < \frac{\sigma^*(\boldsymbol{x})}{2}$:

$$|d^{spectral}(j,k) - d^{spectral}(j,k)| \leq \epsilon \quad \forall j \neq k \in \ell(\boldsymbol{x})$$

with probability $1 - \delta$.

Then we have the following theorem.

Theorem 1. *Define*

$$\triangle(\boldsymbol{x}) := \frac{\min_{u' \in \mathcal{U}: u' \neq u(\boldsymbol{x})} (c(u(\boldsymbol{x})) - c(u'))}{8|\ell(\boldsymbol{x})|}$$

where $u(\boldsymbol{x})$ *is the correct tree and* u' *is any other tree.*

Let $\hat{u}$ *be the estimated tree for tag sequence* $\boldsymbol{x}$. *Assume that the kernel in Eq.* 30 *is used with bandwidth* $\gamma = \frac{\triangle(x)\sigma^*(x)}{C_4 m}$ *and that*

$$N \geq \frac{C_5 m^2 \phi^2 \log \left(\frac{C_2 p^2 \ell(\boldsymbol{x})^2}{\delta} \right)}{\min(\sigma^*(\boldsymbol{x})^2 \triangle(\boldsymbol{x})^2, \sigma^*(\boldsymbol{x})^2) \nu_{\boldsymbol{x}}(\gamma)^2}$$

Then with probability $1 - \delta$, $\hat{u} = u(\boldsymbol{x})$.

Proof. By Lemma 1, and this choice of N and γ,

$$|d^{\text{spectral}}(j,k) - d^{\text{spectral}}(j,k)| \leq \triangle(\boldsymbol{x}) \quad \forall j \neq k \in \ell(\boldsymbol{x}) \tag{31}$$

Now for any tree $u \in \mathcal{U}$, we can compute $c(u)$ by summing over the distances $d(i,j)$ where i and j are adjacent in u.

Recall the formulas for $d(i,j)$ (from the main paper):

- **leaf edge:**

$$d(i,j) = \frac{1}{2}\left(d(j,a^*) + d(j,b^*) - d(a^*,b^*)\right).$$

- **internal edge:**

$$d(i,j) = \frac{1}{4}\Big(d(a^*,g^*) + d(a^*,h^*) + d(b^*,g^*) + d(b^*,h^*) - 2d(a^*,b^*) - 2d(g^*,h^*)\Big).$$

where a^*, b^*, g^*, h^* are leaves (i.e. word nodes). Thus, Eq. 31 implies that

$$|\hat{d}(i,j) - d(i,j)| \leq 2\triangle(\boldsymbol{x}) \quad \forall j \neq k \in [M]$$

Since there are $\leq 2|\ell(\boldsymbol{x})|$ edges in the tree, this is sufficient to guarantee that $|\hat{c}(u) - c(u)| \leq 4|\ell(\boldsymbol{x})|\triangle(\boldsymbol{x})\ \forall u \in \mathcal{U}$. Thus,

$$\hat{c}(u(\boldsymbol{x})) - \hat{c}(u') < 0 \quad \forall u' \in \mathcal{U} : u' \neq u(\boldsymbol{x})$$

Thus, the correct tree is the one with minimum estimated cost and Algorithm 1 will return the correct tree. □

3 Lemma Proofs

Instead of proving Lemma 1 directly, we divide it into two stages. First, we show that our strategy yields consistent estimation of the covariance matrices (Lemma 3). We then show that a concentration bound on the covariance matrices implies that the empirical distance is close to the true distance (Lemma 4). Putting the results together proves Lemma 2.

Lemma 3. *Assume the kernel used is that in Eq.* 30 *with bandwidth* $\gamma = \epsilon/2$ *and that*

$$N \geq \frac{\log(\frac{C_1 p^2 \ell(\boldsymbol{x})^2}{\delta})\phi^2}{C_2 \epsilon^2 \nu_{\boldsymbol{x}}(\epsilon/2)^2}$$

Then for a fixed tag sequence $\boldsymbol{x}$ *and* $\forall j,k \in \ell(\boldsymbol{x})$*,*

$$\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(j,k) - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(j,k)\|_F \leq \epsilon$$

with probability $1 - \delta$*.*

Proof. Define the quantity,

$$\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(j,k) = \frac{\sum_{i=1}^{M} \sum_{j',k' \in [\ell(\boldsymbol{x}_i)]} K_\gamma(j,k,j',k'|\boldsymbol{x},\boldsymbol{x}_i)\boldsymbol{\Sigma}_{\boldsymbol{x}^i}(j',k')}{\sum_{i=1}^{N} \sum_{j',k' \in [\ell(\boldsymbol{x}_i)]} K_\gamma(j,k,j',k'|\boldsymbol{x},\boldsymbol{x}_i)}$$

Note that via triangle inequality,

$$\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(j,k) - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(j,k)\|_F \leq \|\boldsymbol{\Sigma}_{\boldsymbol{x}}(j,k) - \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(j,k)\|_F + \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(j,k) - \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(j,k)\|_F$$

The first term (the bias) is bounded by $\epsilon/2$ using the definition of the kernel in Eq. 30 with bandwidth $\epsilon/2$. The proof for bound for the second term (the variance) can be derived using the technique of (Zhou et al., 2010) and is in Utility Lemma 1. □

Lemma 3 can then be used to prove that with high probability the estimated mutual information is close to the true mutual information.

Lemma 4. *Assume that*

$$\|\widehat{\mathbf{\Sigma}}_{\boldsymbol{x}}(j,k) - \mathbf{\Sigma}_{\boldsymbol{x}}(j,k)\|_F \leq \epsilon \leq \frac{\sigma^*(x)}{2} \, \forall j,k \in [\ell(\boldsymbol{x})]$$

Then,

$$|d^{spectral}(j,k) - d^{spectral}(j,k)| \leq \frac{C_3 m\epsilon}{\sigma^*(x)} \quad \forall j \neq k \in [\ell(\boldsymbol{x})]$$

Proof. By triangle inequality,

$$\begin{aligned}
&|d^{\text{spectral}}(j,k) - d^{\text{spectral}}(j,k)| \\
&\quad \leq |\log \Lambda_m(\mathbf{\Sigma}_{\boldsymbol{x}}(j,k)) - \log \Lambda_m(\widehat{\mathbf{\Sigma}}_{\boldsymbol{x}}(j,k))| \\
&\quad + \frac{1}{2}|\log \Lambda_m(\mathbf{\Sigma}_{\boldsymbol{x}}(j,j)) - \log \Lambda_m(\widehat{\mathbf{\Sigma}}_{\boldsymbol{x}}(j,j))| \\
&\quad + \frac{1}{2}|\log \Lambda_m(\widehat{\mathbf{\Sigma}}_{\boldsymbol{x}}(k,k)) - \log \Lambda_m(\widehat{\mathbf{\Sigma}}_{\boldsymbol{x}}(k,k))| \qquad (34)
\end{aligned}$$

We show the bound on $|\log \Lambda_m(\mathbf{\Sigma}_{\boldsymbol{x}}(j,j)) - \log \Lambda_m(\widehat{\mathbf{\Sigma}}_{\boldsymbol{x}}(j,j))|$, the others follow similarly. By definition of $\Lambda_m(\cdot)$, and triangle inequality we have that,

$$|\log \Lambda_m(\mathbf{\Sigma}_{\boldsymbol{x}}(j,j)) - \log \Lambda_m(\widehat{\mathbf{\Sigma}}_{\boldsymbol{x}}(j,j))| \leq \sum_{r=1}^{m} |\log(\sigma_{\boldsymbol{x}}(j,j)^{(r)}) - \log(\widehat{\sigma}_{\boldsymbol{x}}(j,j)^{(r)})|$$

To translate our concentration bounds on Frobenius norm to concentration of eigenvalues, we apply Weyl's Theorem. Assume first that $\widehat{\sigma}_{\boldsymbol{x}}(j,j)^{(r)} \geq \sigma_{\boldsymbol{x}}(j,j)^{(r)}$. Then,

$$\begin{aligned}
|\log(\sigma_{\boldsymbol{x}}(j,j)^{(r)}) - \log(\widehat{\sigma}_{\boldsymbol{x}}(j,j)^{(r)})| &\leq |\log(\sigma_{\boldsymbol{x}}(j,j)^{(r)}) - \log(\sigma_{\boldsymbol{x}}(j,j)^{(r)} + \epsilon)| \\
&= |\log(\sigma_{\boldsymbol{x}}(j,j)^{(r)}) - \log(\sigma_{\boldsymbol{x}}(j,j)^{(r)}(1 + \frac{\epsilon}{\sigma_{\boldsymbol{x}}(j,j)^{(r)}}))| \\
&\leq \log(1 + \frac{\epsilon}{\sigma_{\boldsymbol{x}}(j,j)^{(r)}}) \leq \frac{\epsilon}{\sigma_{\boldsymbol{x}}(j,j)^{(r)}}
\end{aligned}$$

Similarly, when $\widehat{\sigma}_{\boldsymbol{x}}(j,j)^{(r)} < \sigma_{\boldsymbol{x}}(j,j)^{(r)}$, then

$$\begin{aligned}
&|\log(\sigma_{\boldsymbol{x}}(j,j)^{(r)}) - \log(\widehat{\sigma}_{\boldsymbol{x}}(j,j)^{(r)})| \\
&\leq \frac{\epsilon}{\widehat{\sigma}_{\boldsymbol{x}}(j,j)^{(r)}} = \frac{\epsilon}{\sigma_{\boldsymbol{x}}(j,j)^{(r)} - \epsilon}
\end{aligned}$$

Using the fact that $\epsilon \leq \frac{\sigma^*(x)}{2}$ we obtain that,

$$|\log(\sigma_{\boldsymbol{x}}(j,j)^{(r)}) - \log(\widehat{\sigma}_{\boldsymbol{x}}(j,j)^{(r)})| \leq \frac{2\epsilon}{\sigma_{\boldsymbol{x}}(j,j)^{(r)}}$$

As a result,

$$\begin{aligned}
&|\log \Lambda_m(\mathbf{\Sigma}_{\boldsymbol{x}}(j,j)) - \log \Lambda_m(\widehat{\mathbf{\Sigma}}_{\boldsymbol{x}}(j,j))| \\
&\leq m\frac{2\epsilon}{\sigma_{\boldsymbol{x}}(j,j)^{(m)}}
\end{aligned}$$

Repeating this for the other terms in Eq. 34 gives the bound. □

Utility Lemma 1.

$$P(\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(j,k) - \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(j,k)\|_F \geq \xi) \leq C_1 p^2 \ell(\boldsymbol{x})^2 \exp\left(-\frac{CN\nu_{\boldsymbol{x}}(\gamma)^2\xi^2}{\phi^2}\right) \quad \forall j,k \in [\ell(\boldsymbol{x})]$$

Proof. The proof uses the technique of Zhou et al. (2010).

Define, $N_{\boldsymbol{x};j,k}(\gamma) = \sum_{i=1}^{N}\sum_{j',k'\in[\ell(\boldsymbol{x}_i)]} K_\gamma(j,k,j',k'|\boldsymbol{x},\boldsymbol{x}_i)$ which is the empirical effective sample size under the smoothing kernel.

We first show that the $N_{\boldsymbol{x};j,k}(\gamma)$ is close to the expected effective sample size $N\nu_{\boldsymbol{x}}(\gamma)$.

Using Hoeffding's Inequality, we obtain that

$$P\left(|N_{\boldsymbol{x};j,k}(\gamma) - N\nu_{\boldsymbol{x}}(\gamma)| \geq N\nu_{\boldsymbol{x}}(\gamma)/2\right) \leq C\exp(-N\nu_{\boldsymbol{x}}(\gamma)^2/2) \tag{36}$$

Note that

$$\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(j,k) - \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(j,k)\|_F \leq p^2 \max_{a,b} |\widehat{\varsigma}_{\boldsymbol{x}}(j,k;a,b) - \widetilde{\varsigma}_{\boldsymbol{x}}(j,k;a,b)|$$

where $\widehat{\varsigma}_{\boldsymbol{x}}(j,k;a,b)$ is the element on the a^{th} row and b^{th} column of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(j,k)$. Thus, it suffices to bound $|\widehat{\varsigma}_{\boldsymbol{x}}(j,k;a,b) - \widetilde{\varsigma}_{\boldsymbol{x}}(j,k;a,b)|$.

Define the boolean variable $E = \mathbb{I}[N_{\boldsymbol{x};j,k}(\gamma) > N\nu_{\boldsymbol{x}}(\gamma)/2]$. Then,

$$\begin{aligned}
& P(\|\widehat{\varsigma}_{\boldsymbol{x}}(j,k;a,b) - \widetilde{\varsigma}_{\boldsymbol{x}}(j,k;a,b)\|_F \geq \xi) = \\
& \quad P(\|\widehat{\varsigma}_{\boldsymbol{x}}(j,k;a,b) - \widetilde{\varsigma}_{\boldsymbol{x}}(j,k;a,b)\|_F \geq \xi | E=1)P(E=1) \\
& \quad + P(\|\widehat{\varsigma}_{\boldsymbol{x}}(j,k;a,b) - \widetilde{\varsigma}_{\boldsymbol{x}}(j,k;a,b)\|_F \geq \xi | E=0)P(E=0) \\
\leq \; & P(\|\widehat{\varsigma}_{\boldsymbol{x}}(j,k;a,b) - \widetilde{\varsigma}_{\boldsymbol{x}}(j,k;a,b)\|_F \geq \xi | E=1) + P(E=0)
\end{aligned}$$

The second term is bounded by Eq. 36. We prove the first term below.

For shorthand, refer to $P(\|\widehat{\varsigma}_{\boldsymbol{x}}(j,k;a,b) - \widetilde{\varsigma}_{\boldsymbol{x}}(j,k;a,b)\|_F \geq \xi | E=1)$ as $P_E(\|\widehat{\varsigma}_{\boldsymbol{x}}(j,k;a,b) - \widetilde{\varsigma}_{\boldsymbol{x}}(j,k;a,b)\|_F \geq \xi | E=1)$. Define the following quantities:

For convenience define the following quantity, where $w^{(i)}_{j,a}$ is the a^{th} element of the vector $w^{(i)}_j$

$$\delta_{\boldsymbol{x}^{(i)}}(j,k;a,b) = w^{(i)}_{j,a}(w^{(i)}_{k,b})^\top - \varsigma_{\boldsymbol{x}^{(i)}}(j,k;a,b)$$

For every $\kappa > 0$, by Markov's inequality,

$$\begin{aligned}
& P_E\left(\sum_{i\in[N]}\sum_{j',k'\in\ell(\boldsymbol{x}^{(i)})} K_\gamma(j,k,j',k'|\boldsymbol{x},\boldsymbol{x}^{(i)})\delta_{\boldsymbol{x}^{(i)}}(j,k;a,b) > N_{\boldsymbol{x};j,k}(\gamma)\xi\right) \\
= \; & P_E\left(\exp\left(\kappa\sum_{i\in[N]}\sum_{j',k'\in\ell(\boldsymbol{x}^{(i)})} K_\gamma(j,k,j',k'|\boldsymbol{x},\boldsymbol{x}^{(i)})\delta_{\boldsymbol{x}^{(i)}}(j,k;a,b)\right) > \exp\left(\kappa N_{\boldsymbol{x};j,k}(\gamma)\xi\right)\right) \\
\leq \; & \frac{\mathbb{E}\left[\exp\left(\kappa\sum_{i\in[N]}\sum_{j',k'\in\ell(\boldsymbol{x}^{(i)})} K_\gamma(j,k,j',k'|\boldsymbol{x},\boldsymbol{x}^{(i)})\delta_{\boldsymbol{x}^{(i)}}(j,k;a,b)\right)\right]}{\exp\left(\kappa N_{\boldsymbol{x};j,k}(\gamma)\xi\right)}
\end{aligned}$$

We now bound the numerator. Using the fact that the samples are independent, we have that

$$\begin{aligned}
& \mathbb{E}\left[\exp\left(\kappa\sum_{i\in[N]}\sum_{j',k'\in\ell(\boldsymbol{x}^{(i)})} K_\gamma(j,k,j',k'|\boldsymbol{x},\boldsymbol{x}^{(i)})\delta_{\boldsymbol{x}^{(i)}}(j,k;a,b)\right)\right] \\
= \; & \prod_{i\in[N]}\prod_{j',k'\in\ell(\boldsymbol{x}^{(i)})} \mathbb{E}\left[\exp\left(\kappa K_\gamma(j,k,j',k'|\boldsymbol{x},\boldsymbol{x}^{(i)})\delta_{\boldsymbol{x}^{(i)}}(j,k;a,b)\right)\right] \\
= \; & \prod_{i\in[N]}\prod_{j',k'\in\ell(\boldsymbol{x}^{(i)})} \mathbb{I}[(j',k',\boldsymbol{x}^{(i)}) \in B_\gamma(j,k,\boldsymbol{x})]\mathbb{E}\left[\exp(\kappa\delta_{\boldsymbol{x}^{(i)}}(j,k;a,b))\right]
\end{aligned}$$

From Utility Lemma 2, we can conclude that

$$\begin{aligned}
& \prod_{i\in[N]} \prod_{j',k'\in\ell(\boldsymbol{x}^{(i)})} \mathbb{I}[(j',k',\boldsymbol{x}^{(i)}) \in B_\gamma(j,k,\boldsymbol{x})]\mathbb{E}\left[\exp(\kappa\delta_{\boldsymbol{x}^{(i)}}(j,k;a,b))\right] \\
\leq & \prod_{i\in[N]} \prod_{j',k'\in\ell(\boldsymbol{x}^{(i)})} \mathbb{I}[(j',k',\boldsymbol{x}^{(i)}) \in B_\gamma(j,k,\boldsymbol{x})]\exp(\kappa^2\phi^2/8) \\
\leq & \exp(N_{\boldsymbol{x};j,k}(\gamma)\kappa^2\phi^2/8)
\end{aligned}$$

$$\begin{aligned}
& P_E\left(\sum_{i\in[N]}\sum_{j',k'\in\ell(\boldsymbol{x}^{(i)})} K_\gamma(j,k,j',k'|\boldsymbol{x},\boldsymbol{x}^{(i)})\delta_{\boldsymbol{x}^{(i)}}(j,k;a,b) > N_{\boldsymbol{x},j,k}(\gamma)\xi\right) \\
\leq & \exp(-\kappa N_{\boldsymbol{x};j,k}(\gamma)\xi)\exp(N_{\boldsymbol{x};j,k}(\gamma)\kappa^2\phi^2/8)
\end{aligned}$$

Setting $\kappa = \frac{4\xi}{\phi^2}$, gives us that

$$\begin{aligned}
& P_E\left(\sum_{i\in[N]}\sum_{j',k'\in\ell(x^{(i)})} K_\gamma(j,k,j',k'|\boldsymbol{x},\boldsymbol{x}^{(i)})\delta_{\boldsymbol{x}^{(i)}}(j,k;a,b) > N_{\boldsymbol{x},j,k}(\gamma)\xi\right) \\
\leq & \exp(-2N_{\boldsymbol{x};j,k}(\gamma)\xi^2/\phi^2) \\
\leq & \exp(-N\nu_{\boldsymbol{x}}(\gamma)\xi^2/\phi^2)
\end{aligned}$$

Combining the above with Eq. 36 and taking some union bounds over a, b, j, k proves the lemma. □

Helper Lemmas

We make use of the following helper lemma that is standard in the proof of Hoeffding's Inequality, e.g. Casella and Berger (1990).

Utility Lemma 2. *Suppose that* $\mathbb{E}(X) = 0$ *and that* $a \leq X \leq b$. *Then*

$$\mathbb{E}[e^{tX}] \leq e^{t^2(b-a)^2/8}$$

References

A. G. Akritas, E. K. Akritas, and G. I. Malaschonok. 1996. Various proofs of Sylvester's (determinant) identity. *Mathematics and Computers in Simulation*, 42(4):585–593.

A. Anandkumar, K. Chaudhuri, D. Hsu, S. M. Kakade, L. Song, and T. Zhang. 2011. Spectral methods for learning multivariate latent tree structure. *arXiv preprint arXiv:1107.1283*.

G. Casella and R. L. Berger. 1990. *Statistical inference*, volume 70. Duxbury Press Belmont, CA.

L. Song, A.P. Parikh, and E.P. Xing. 2011. Kernel embeddings of latent tree graphical models. In *Proceedings of NIPS*.

S. Zhou, J. Lafferty, and L. Wasserman. 2010. Time varying undirected graphs. *Machine Learning*, 80(2-3):295–319.