1 Justification for Construction 3.1

Set $A^{(1)}$ to be the (scaled) adjacency matrix of $\mathcal{G}$, i.e $A^{(1)}_{i,j} = \beta_1 \mathbf{1}(j = p(i))$, and $A^{(2)} = \beta_2 I_S$, where $\beta_1, \beta_2 \to \infty$. We will now show that the output of the disentangled transformer approximates $\hat{\pi}_{s_{1:T}}(\cdot \mid s_T)$.

First Attention. Note that by the construction of $\widetilde{A}^{(1)}$, $\tilde{X}\widetilde{A}^{(1)}\tilde{X}^\top = A^{(1)}$, which is the scaled adjacency matrix of $\mathcal{G}$. If i is not a root node (i.e. $i \in \overline{\mathcal{R}}$, $p(i) \neq \emptyset$), then

$$\mathcal{S}(\tilde{X}\widetilde{A}^{(1)}\tilde{X}^\top)_{i,j} = \mathbf{1}(j = p(i)) \tag{1}$$

so i attends to its parent $p(i)$. Therefore, the output of the first attention is the token at position $p(i)$, i.e. $\mathrm{attn}(\tilde{X}; \widetilde{A}^{(1)})_i = \tilde{x}_{p(i)}$. The transformer then appends $\tilde{x}_{p(i)}$ to the residual stream of token i.

When i is a root node (i.e. $i \in \mathcal{R}$, $p(i) = \emptyset$), then for all j, $(\tilde{X}\widetilde{A}^{(1)}\tilde{X}^\top)_{ij} = 0$. Therefore after the softmax, i will attend equally to all previous tokens:

$$\mathcal{S}(\tilde{X}\widetilde{A}^{(1)}\tilde{X}^\top)_{i,j} = \frac{1}{i} \quad \text{for all} \quad j \leq i. \tag{2}$$

Thus the first attention layer averages all of the tokens in the sequence: $\mathrm{attn}(\tilde{X}; \widetilde{A}^{(1)})_i = \frac{1}{i}\sum_{j \leq i} \tilde{x}_j$. It then copies this average into the residual stream.

Second Attention. We next show that the Tth token attends to all prior tokens whose parents tokens are equal to s_T. It then averages them and copies them into the residual stream.

After the first attention layer, the residual stream is $h^{(1)}_j = [\tilde{x}_j, \mathrm{attn}(\tilde{X}; \widetilde{A}^{(1)})_j]^\top$. The second attention layer compares the Tth token of the original sequence $\tilde{x}_T$ to the output of the first attention at all other positions. Explicitly, the attention pattern is equal to:

$$h^{(1)\top}_T \widetilde{A}^{(2)} h^{(1)}_j = \beta_2 \cdot \tilde{x}_T^\top \begin{bmatrix} A^{(2)} & 0_{S \times T} \\ 0_{T \times S} & 0_{T \times T} \end{bmatrix} \mathrm{attn}(\tilde{X}; \widetilde{A}^{(1)})_j = \beta_2 \cdot \begin{cases} \mathbf{1}(s_{p(i)} = s_T) & i \in \overline{\mathcal{R}} \\ \frac{1}{i}\sum_{j \leq i} \mathbf{1}(s_j = s_T) & i \in \mathcal{R}. \end{cases} \tag{3}$$

As $\beta_2 \to \infty$, the softmax converges to a hard max, and so the Tth token attends equally to all tokens i such that $s_{p(i)} = s_T$. The attention then averages all of these tokens, so the Tth token in the residual stream is equal to $h^{(2)}_T = \left[\tilde{x}_T, \frac{1}{T}\sum_{j \leq T} \tilde{x}_j, Z, \tilde{x}_T\right]$ where

$$Z := \frac{\sum_{s_{p(i)} = s_T} \tilde{x}_i}{|\{i \;:\; s_{p(i)} = s_T\}|} \tag{4}$$

is the average of the tokens whose parent is equal to s_T.

Output Layer. W_O reads from the third block in this stream, which we denoted by Z in (4) above. It then returns the token embedding of Z which is equal to:

$$f_{\widetilde{\theta}}(s_{1:T}) = \frac{\sum_{s_{p(i)}=s_T} e_{s_i}}{|\{i \ : \ s_{p(i)} = s_T\}|} = \hat{\pi}_{s_{1:T}}(\cdot|s_T), \tag{5}$$

as desired.

2 Justification for Equation 3

The output of the first attention layer is

$$\mathrm{attn}(\tilde{X}; \widetilde{A}^{(1)}) = \mathcal{S}(\mathrm{MASK}(\tilde{X}\widetilde{A}^{(1)}\tilde{X}^\top)\tilde{X} = \mathcal{S}(\mathrm{MASK}(A^{(1)}))\tilde{X}.$$

Next, we have that

$$\begin{aligned} h_T^{(1)^\top}\widetilde{A}^{(2)}h^{(1)^\top} &= \tilde{x}_T^\top \begin{bmatrix} A^{(2)} & 0_{S\times T} \\ 0_{T\times S} & 0_{T\times T}\end{bmatrix} \mathrm{attn}(\tilde{X}; \widetilde{A}^{(1)})^\top \\ &= \tilde{x}_T^\top \begin{bmatrix} A^{(2)} & 0_{S\times T} \\ 0_{T\times S} & 0_{T\times T}\end{bmatrix} \tilde{X}^\top \mathcal{S}(\mathrm{MASK}(A^{(1)}))^\top \\ &= \overline{x}_T^\top A^{(2)}\overline{X}^\top \mathcal{S}(\mathrm{MASK}(A^{(1)}))^\top. \end{aligned}$$

Thus the output of the second attention layer is

$$\begin{aligned} \mathrm{attn}(h^{(1)}; \widetilde{A}^{(2)})_T &= h^{(1)^\top}\mathcal{S}\left(h^{(1)}\left(\widetilde{A}^{(2)}\right)^\top h^{(T)}\right) \\ &= h^{(1)^\top}\mathcal{S}\Big(\mathcal{S}(\mathrm{MASK}(A^{(1)}))\overline{X}A^{(2)^\top}\overline{x}_T\Big) \end{aligned}$$

Finally, the output is

$$\begin{aligned} \widetilde{\mathrm{TF}}_{\widetilde{\theta}}(s_{1:T}) &= \widetilde{W}_O^\top h_T^{(2)} \\ &= \begin{bmatrix} I_S & 0_{S\times T} \mid 0_{S\times d}\end{bmatrix} \mathrm{attn}(h^{(1)}; \widetilde{A}^{(2)})_T \\ &= \begin{bmatrix} I_S & 0_{S\times T} \mid 0_{S\times d}\end{bmatrix} h^{(1)^\top}\mathcal{S}\Big(\mathcal{S}(\mathrm{MASK}(A^{(1)}))\overline{X}A^{(2)^\top}\overline{x}_T\Big) \\ &= \overline{X}^\top\mathcal{S}\Big(\mathcal{S}(\mathrm{MASK}(A^{(1)}))\overline{X}A^{(2)^\top}\overline{x}_T\Big), \end{aligned}$$

as desired.