2. (a)

stack	buffer	new dependency	transition
[ROOT]	[I, parsed, this, sentence, correctly]		Initial Configuration
[ROOT, I]	[parsed, this, sentence, correctly]		SHIFT
[ROOT, I, parsed]	[this, sentence, correctly]		SHIFT
[ROOT, parsed]	[this, sentence, correctly]	parsed $\to$I	LEFT-ARC
[ROOT, parsed, this]	[sentence, correctly]		SHIFT
[ROOT, parsed, this, sentence]	[correctly]		SHIFT
[ROOT, parsed, sentence]	[correctly]	sentence $\to$this	LEFT-ARC
[ROOT, parsed]	[correctly]	parsed $\to$sentence	RIGHT-ARC
[ROOT, parsed, correctly]	[]		SHIFT
[ROOT, parsed]	[]	parsed $\to$correctly	RIGHT-ARC
[ROOT]	[]	ROOT $\to$parsed	RIGHT-ARC

2. (b) A sentence containing n words will be parsed in $2n$ steps, because each word needs to be pushed into the stack, which needs n steps, and then popped out of the stack until only ROOT is in the stack, which needs another n steps.

3. (a) (i) To answer this question, we need to take into consideration that only one element of $\{\boldsymbol{y}\}$ is 1, and others 0. Assume that the ith element is 1. Then for the cross-entroy, we have

$$J = -y_i \ln \hat{y}_i = -\ln \hat{y}_i.$$

Correspondingly, the expression for the perplexity becomes

$$PP = \frac{1}{\hat{y}_i}.$$

Therefore,

$$\mathrm{e}^J = \mathrm{e}^{-\ln \hat{y}_i} = \mathrm{e}^{\ln \hat{y}_i^{-1}} = \frac{1}{\hat{y}_i} = PP.$$

3. (a) (ii) Apply logarithm to the geometric mean perplexity:

$$\begin{aligned}\ln\left(\prod_{t=1}^{T} PP^{(t)}\right)^{1/T} &= \frac{1}{T}\left(\ln PP^{(1)} + \ln PP^{(2)} + \cdots + \ln PP^{(T)}\right)\\ &= \frac{1}{T}\left(J^{(1)} + J^{(2)} + \cdots + J^{(T)}\right) = \frac{1}{T}\sum_{t=1}^{T} J^{(t)}.\end{aligned}$$

Since the logarithm function is an increasing one, minimizing the geometric mean perplexity is equivalent to minimizing the arithmetic mean cross-entropy loss.

3. (a) (iii) At any step, the probability of the model predicting the correct word is $\frac{1}{|V|}$. So the perplexity is $|V| = 100000$. The cross-entropy loss is $\ln|V| = 4\ln 10 \approx 9.21$.

3. (b) We already know that $\frac{\partial J}{\partial\left\{\boldsymbol{\theta}^{(t)}\right\}} = \left\{\hat{\boldsymbol{y}}^{(\boldsymbol{t})}\right\} - \left\{\boldsymbol{y}^{(t)}\right\}$, therefore

$$
\begin{aligned}
\frac{\partial J^{(t)}}{\partial [\boldsymbol{U}]} &= \left(\left\langle \hat{\boldsymbol{y}}^{(\boldsymbol{t})} \right\rangle - \left\langle \boldsymbol{y}^{(t)} \right\rangle\right) \frac{\partial \{\boldsymbol{\theta}\}}{\partial [\boldsymbol{U}]} \\
&= \begin{bmatrix}
\left(\left\langle \hat{\boldsymbol{y}}^{(\boldsymbol{t})} \right\rangle - \left\langle \boldsymbol{y}^{(t)} \right\rangle\right) \begin{bmatrix} \frac{\partial \theta_1}{\partial \langle \boldsymbol{u}_1 \rangle} \\ \vdots \\ \frac{\partial \theta_1}{\partial \langle \boldsymbol{u}_V \rangle} \end{bmatrix} \\
\left(\left\langle \hat{\boldsymbol{y}}^{(\boldsymbol{t})} \right\rangle - \left\langle \boldsymbol{y}^{(t)} \right\rangle\right) \begin{bmatrix} \frac{\partial \theta_2}{\partial \langle \boldsymbol{u}_1 \rangle} \\ \vdots \\ \frac{\partial \theta_2}{\partial \langle \boldsymbol{u}_V \rangle} \end{bmatrix} \\
\vdots \\
\left(\left\langle \hat{\boldsymbol{y}}^{(\boldsymbol{t})} \right\rangle - \left\langle \boldsymbol{y}^{(t)} \right\rangle\right) \begin{bmatrix} \frac{\partial \theta_V}{\partial \langle \boldsymbol{u}_1 \rangle} \\ \vdots \\ \frac{\partial \theta_V}{\partial \langle \boldsymbol{u}_V \rangle} \end{bmatrix}
\end{bmatrix} \\
&= \begin{bmatrix}
\left(\left\langle \hat{\boldsymbol{y}}^{(\boldsymbol{t})} \right\rangle - \left\langle \boldsymbol{y}^{(t)} \right\rangle\right) \begin{bmatrix} \left\langle \boldsymbol{h}^{(t)} \right\rangle \\ \langle \boldsymbol{0} \rangle \\ \vdots \\ \langle \boldsymbol{0} \rangle \end{bmatrix} \\
\left(\left\langle \hat{\boldsymbol{y}}^{(\boldsymbol{t})} \right\rangle - \left\langle \boldsymbol{y}^{(t)} \right\rangle\right) \begin{bmatrix} \langle \boldsymbol{0} \rangle \\ \left\langle \boldsymbol{h}^{(t)} \right\rangle \\ \vdots \\ \langle \boldsymbol{0} \rangle \end{bmatrix} \\
\vdots \\
\left(\left\langle \hat{\boldsymbol{y}}^{(\boldsymbol{t})} \right\rangle - \left\langle \boldsymbol{y}^{(t)} \right\rangle\right) \begin{bmatrix} \langle \boldsymbol{0} \rangle \\ \vdots \\ \left\langle \boldsymbol{h}^{(t)} \right\rangle \end{bmatrix}
\end{bmatrix} \\
&= \begin{bmatrix}
\left(\left\langle \hat{\boldsymbol{y}}^{(\boldsymbol{t})} \right\rangle - \left\langle \boldsymbol{y}^{(t)} \right\rangle\right)_{*1} \cdot \left\langle \boldsymbol{h}^{(t)} \right\rangle \\
\left(\left\langle \hat{\boldsymbol{y}}^{(\boldsymbol{t})} \right\rangle - \left\langle \boldsymbol{y}^{(t)} \right\rangle\right)_{*2} \cdot \left\langle \boldsymbol{h}^{(t)} \right\rangle \\
\vdots \\
\left(\left\langle \hat{\boldsymbol{y}}^{(\boldsymbol{t})} \right\rangle - \left\langle \boldsymbol{y}^{(t)} \right\rangle\right)_{*V} \cdot \left\langle \boldsymbol{h}^{(t)} \right\rangle
\end{bmatrix} \\
&= \left\{ \boldsymbol{h}^{(t)} \right\} \left(\left\langle \hat{\boldsymbol{y}}^{(\boldsymbol{t})} \right\rangle - \left\langle \boldsymbol{y}^{(t)} \right\rangle\right).
\end{aligned}
$$

$$
\frac{\partial \left\{ \boldsymbol{\theta}^{(t)} \right\}}{\partial \left\langle \boldsymbol{h}^{(t)} \right\rangle} = [\boldsymbol{U}],
$$

$$
\frac{\partial \left\{ \boldsymbol{h}^{(t)} \right\}}{\partial \left\langle \boldsymbol{z}^{(t)} \right\rangle} = \operatorname{diag}\left(\left\langle \boldsymbol{h}^{(t)} \right\rangle \odot \left(\langle \boldsymbol{1} \rangle - \left\langle \boldsymbol{h}^{(t)} \right\rangle \right) \right).
$$

$$\frac{\partial \left\{ \boldsymbol{z}^{(t)} \right\}}{\partial \left\langle \boldsymbol{e}^{(t)} \right\rangle} = [\boldsymbol{W}_e] \, .$$

Therefore,

$$\begin{aligned}
\frac{\partial J^{(t)}}{\partial \left\langle \boldsymbol{e}^{(t)} \right\rangle} &= \frac{\partial J^{(t)}}{\partial \left\langle \boldsymbol{\theta}^{(t)} \right\rangle} \frac{\partial \left\{ \boldsymbol{\theta}^{(t)} \right\}}{\partial \left\langle \boldsymbol{h}^{(t)} \right\rangle} \frac{\partial \left\{ \boldsymbol{h}^{(t)} \right\}}{\partial \left\langle \boldsymbol{z}^{(t)} \right\rangle} \frac{\partial \left\{ \boldsymbol{z}^{(t)} \right\}}{\partial \left\langle \boldsymbol{e}^{(t)} \right\rangle} \\
&= \left(\left\langle \hat{\boldsymbol{y}}^{(\boldsymbol{t})} \right\rangle - \left\langle \boldsymbol{y}^{(t)} \right\rangle \right) [\boldsymbol{U}] \operatorname{diag} \left(\left\langle \boldsymbol{h}^{(t)} \right\rangle \odot \left(\langle \mathbf{1} \rangle - \left\langle \boldsymbol{h}^{(t)} \right\rangle \right) \right) [\boldsymbol{W}_e] \, .
\end{aligned}$$

$$\frac{\partial J^{(t)}}{\partial \left\{ \boldsymbol{e}^{(t)} \right\}} = [\boldsymbol{W}_e]^\mathsf{T} \operatorname{diag} \left(\left(\{ \mathbf{1} \} - \left\{ \boldsymbol{h}^{(t)} \right\} \right) \odot \left\{ \boldsymbol{h}^{(t)} \right\} \right) [\boldsymbol{U}]^\mathsf{T} \left(\left\{ \hat{\boldsymbol{y}}^{(t)} \right\} - \left\{ \boldsymbol{y}^{(t)} \right\} \right) .$$

$$\begin{aligned}
\frac{\partial \left\{ \boldsymbol{z}^{(t)} \right\}}{\partial \left[\boldsymbol{W}_e \right]} &= \begin{bmatrix} \begin{bmatrix} \frac{\partial z_1}{\partial \langle \boldsymbol{W}_{e,1*} \rangle} \\ \vdots \\ \frac{\partial z_1}{\partial \langle \boldsymbol{W}_{e,D_h*} \rangle} \end{bmatrix} \\ \begin{bmatrix} \frac{\partial z_2}{\partial \langle \boldsymbol{W}_{e,1*} \rangle} \\ \vdots \\ \frac{\partial z_2}{\partial \langle \boldsymbol{W}_{e,D_h*} \rangle} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \frac{\partial z_{D_h}}{\partial \langle \boldsymbol{W}_{e,1*} \rangle} \\ \vdots \\ \frac{\partial z_{D_h}}{\partial \langle \boldsymbol{W}_{e,D_h*} \rangle} \end{bmatrix} \end{bmatrix} \\
&= \begin{bmatrix} \begin{bmatrix} \left\langle \boldsymbol{e}^{(t)} \right\rangle \\ \langle \mathbf{0} \rangle \\ \vdots \\ \langle \mathbf{0} \rangle \end{bmatrix} \\ \begin{bmatrix} \langle \mathbf{0} \rangle \\ \left\langle \boldsymbol{e}^{(t)} \right\rangle \\ \vdots \\ \langle \mathbf{0} \rangle \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \langle \mathbf{0} \rangle \\ \vdots \\ \left\langle \boldsymbol{e}^{(t)} \right\rangle \end{bmatrix} \end{bmatrix} .
\end{aligned}$$

$$\left.\frac{\partial J^{(t)}}{\partial [\boldsymbol{W}_e]}\right|_t = \frac{\partial J^{(t)}}{\partial \left\langle \boldsymbol{\theta}^{(t)} \right\rangle} \frac{\partial \left\{\boldsymbol{\theta}^{(t)}\right\}}{\partial \left\langle \boldsymbol{h}^{(t)} \right\rangle} \frac{\partial \left\{\boldsymbol{h}^{(t)}\right\}}{\partial \left\langle \boldsymbol{z}^{(t)} \right\rangle} \frac{\partial \left\{\boldsymbol{z}^{(t)}\right\}}{\partial [\boldsymbol{W}_e]}$$

$$= \left(\left(\left\langle \hat{\boldsymbol{y}}^{(\boldsymbol{t})} \right\rangle - \left\langle \boldsymbol{y}^{(t)} \right\rangle\right) [\boldsymbol{U}] \odot \left\langle \boldsymbol{h}^{(t)} \right\rangle \odot \left(\langle \mathbf{1} \rangle - \left\langle \boldsymbol{h}^{(t)} \right\rangle\right)\right)^{\mathsf{T}} \otimes \left\langle \boldsymbol{e}^{(t)} \right\rangle .$$

Similarly,

$$\left.\frac{\partial J^{(t)}}{\partial [\boldsymbol{W}_h]}\right|_t = \left(\left(\left\langle \hat{\boldsymbol{y}}^{(\boldsymbol{t})} \right\rangle - \left\langle \boldsymbol{y}^{(t)} \right\rangle\right) [\boldsymbol{U}] \odot \left\langle \boldsymbol{h}^{(t)} \right\rangle \odot \left(\langle \mathbf{1} \rangle - \left\langle \boldsymbol{h}^{(t)} \right\rangle\right)\right)^{\mathsf{T}} \otimes \left\langle \boldsymbol{h}^{(t-1)} \right\rangle .$$

$$\frac{\partial J^{(t)}}{\partial \left\{\boldsymbol{h}^{(t-1)}\right\}} = [\boldsymbol{W}_h]^{\mathsf{T}} \operatorname{diag}\left(\left(\{\mathbf{1}\} - \left\{\boldsymbol{h}^{(t)}\right\}\right) \odot \left\{\boldsymbol{h}^{(t)}\right\}\right) [\boldsymbol{U}]^{\mathsf{T}} \left(\left\{\hat{\boldsymbol{y}}^{(t)}\right\} - \left\{\boldsymbol{y}^{(t)}\right\}\right).$$

3. c

$$\frac{\partial J^{(t)}}{\partial \left\langle \boldsymbol{e}^{(t-1)} \right\rangle} = \frac{\partial J^{(t)}}{\partial \left\langle \boldsymbol{h}^{(t-1)} \right\rangle} \frac{\partial \left\{\boldsymbol{h}^{(t-1)}\right\}}{\partial \left\langle \boldsymbol{z}^{(t-1)} \right\rangle} \frac{\partial \left\{\boldsymbol{z}^{(t-1)}\right\}}{\partial \left\langle \boldsymbol{e}^{(t-1)} \right\rangle}$$

$$= \left\langle \boldsymbol{\gamma}^{(t-1)} \right\rangle \operatorname{diag}\left(\left\langle \boldsymbol{h}^{(t-1)} \right\rangle \odot \left(\langle \mathbf{1} \rangle - \left\langle \boldsymbol{h}^{(t-1)} \right\rangle\right)\right) [\boldsymbol{W}_e] .$$

$$\left.\frac{\partial J^{(t)}}{\partial [\boldsymbol{W}_e]}\right|_{t-1} = \frac{\partial J^{(t)}}{\partial \left\langle \boldsymbol{h}^{(t-1)} \right\rangle} \frac{\partial \left\{\boldsymbol{h}^{(t-1)}\right\}}{\partial \left\langle \boldsymbol{z}^{(t-1)} \right\rangle} \frac{\partial \left\{\boldsymbol{z}^{(t-1)}\right\}}{\partial [\boldsymbol{W}_e]}$$

$$= \left(\left\langle \boldsymbol{\gamma}^{(t-1)} \right\rangle \odot \left\langle \boldsymbol{h}^{(t-1)} \right\rangle \odot \left(\langle \mathbf{1} \rangle - \left\langle \boldsymbol{h}^{(t-1)} \right\rangle\right)\right)^{\mathsf{T}} \otimes \left\langle \boldsymbol{e}^{(t-1)} \right\rangle .$$

$$\left.\frac{\partial J^{(t)}}{\partial [\boldsymbol{W}_h]}\right|_{t-1} = \left(\left\langle \boldsymbol{\gamma}^{(t-1)} \right\rangle \odot \left\langle \boldsymbol{h}^{(t-1)} \right\rangle \odot \left(\langle \mathbf{1} \rangle - \left\langle \boldsymbol{h}^{(t-1)} \right\rangle\right)\right)^{\mathsf{T}} \otimes \left\langle \boldsymbol{h}^{(t-2)} \right\rangle .$$

3. d $\mathsf{O}\left(|V|\, D_h + dD_h + D_h^2\right)$.
3. e $\mathsf{O}\left(T\left(|V|\, D_h + dD_h + D_h^2\right)\right)$.
3. f $\mathsf{O}\left(|V|\, D_h\right)$ would be the dominant part.