Problem 1

A - False

Both losses include $\sum_{i=1}^{N} \log p(y|x)$, so $\frac{d}{d\theta}$ will be the same for both $L(\theta)$.

The derivative of θ_j will be the constant $C\theta_j$ as long as it's acknowledged θ_j can't be 0.

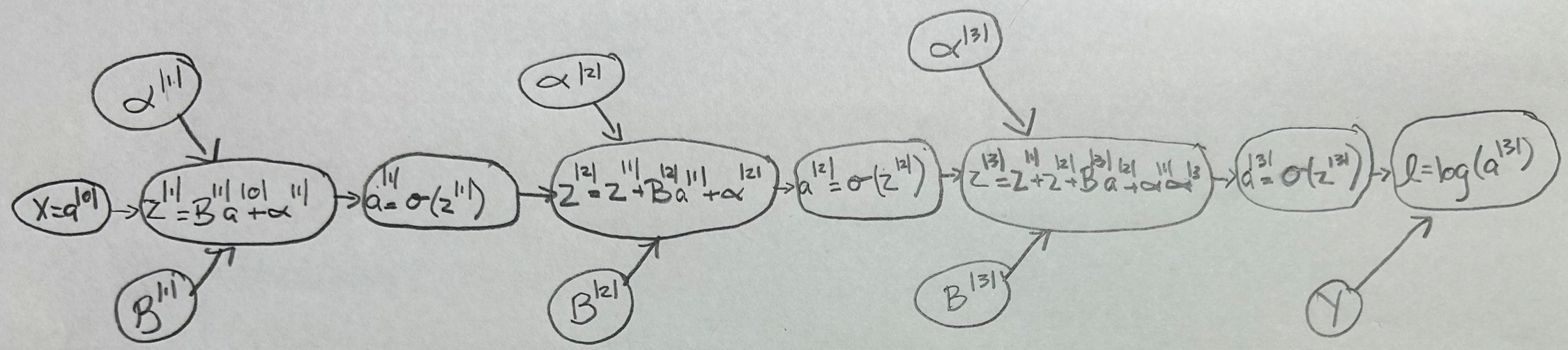
So subtracting a constant C from $L_2(\theta)$ means that

$$\frac{L_2(\theta)}{d\theta_j} < \frac{L_1(\theta)}{d\theta_j}$$

B - True, explanation same as above. In the case where $\log p(y|x)$ is the same for both $L(\theta)$ the regularized term would make $L_2(\theta) < L_1(\theta)$

C - True, $\log p(y^i|x^i)$ can diverge because $\log(0)$ is undefined

D - True, as is the case for log linear models. There are many cases where even with regularization they diverge



$$\frac{d\ell}{d\alpha^{|1|}} = \frac{d\ell}{da^{|3|}} \frac{da^{|3|}}{dz^{|3|}} \frac{dz^{|3|}}{a^{|2|}} \frac{da^{|2|}}{dz^{|2|}} \frac{dz^{|2|}}{da^{|1|}} \frac{da^{|1|}}{dz^{|1|}} \frac{dz^{|1|}}{d\alpha^{|1|}}$$

First calculate z, a, ℓ this is done by sequentially passing X, and α and B from the previous levels according to the equations in the graph starting from $X=a^{|0|}$. Once done, then according to the gradients shown above for $\frac{d\ell}{d\alpha^{|1|}}$, calculate the term.

Recall that $\frac{d}{d\sigma(z)} = \sigma(z)(1-\sigma(z))$.

Problem 3.

must sum to 1

$$\begin{cases} p(y|x) = \frac{9}{10} & \text{if } y = x \\ p(y|x) = \frac{1}{10} & \text{if } y \neq x \end{cases}$$

$$\boxed{\begin{array}{l} f_1(x,y) = 1_{\{x=y\}} \\ f_2(x,y) = 1_{\{x \neq y\}} \end{array}}$$

$$\frac{1}{10} = \frac{\alpha}{|V|-1}$$

$$\frac{1(|V|-1)}{10} = \frac{\alpha 10}{10}$$

$$\boxed{\alpha = \frac{|V|-1}{10}}$$

Problem 4

Interpretation -

$f_1(x^i, y) = i$ means it spits out number of index $i = 1, 2, \ldots, N$

(if you interpret it as $f_1(x^i, y) = x^i$, $\bar{x}$ would be the mean)

$$L(\theta) = \frac{\sum_{i=1}^{n} \theta^T f(x^i, y^i) - \sum_{i=1}^{n} \log \sum_{y} \exp \theta^T f(x^i, y)}{N}$$

$$\frac{dL}{d\theta_1} = \frac{\sum_{i=1}^{n} f_1(x^i, y^i)}{N} - \frac{1}{N} \sum_{i=1}^{n} \sum_{y} f_1(x^i, y) p(y|x^i) - \lambda\theta_1$$

$$\frac{dL}{d\theta_1} = 0$$

$$\frac{\sum_{i=1}^{n} f_1(x^i, y^i)}{N} = \frac{f_1(x^1) + f_1(x^2) + \ldots + f_1(x^N)}{N}$$

$$\bar{x} = \frac{1 + 2 + 3 + \ldots + N}{N}$$

$$\frac{1}{N} \sum_{i=1}^{n} \sum_{y} f_1(x^i, y) p(y|x^i) = \frac{1}{N} \sum_{i=1}^{n} \sum_{y} p(y|x^i)\, i$$

recall: $\sum_{y} p(y|x^i) = 1$

$$= \frac{1}{N} \sum_{i=1}^{n} i$$

$$\bar{x} = \frac{1 + 2 + 3 + \ldots + N}{N}$$

$$0 = \frac{\sum_{i=1}^{n} f_1(x^i, y^i)}{N} - \frac{1}{N} \sum_{i=1}^{n} \sum_{y} f_1(x^i, y) p(y|x^i) - \lambda\theta_1$$

$$0 = \bar{x} - \bar{x} - \lambda\theta_1$$

$$\boxed{0 = \theta_1^*}$$

Problem 5.

A- $p(y|x) = \prod_{t=1}^{T} \exp[\theta^T f(x, y_{t-1}, t, y_t)]$

<u>memm</u> $\dfrac{\prod_{t=1}^{T} \exp[\theta^T f(x, y_{t-1}, t, y_t)]}{\prod_{t=1}^{T} \sum_{y \in K} \exp[\theta^T f(x, y_{t-1}, t, y)]}$

<u>CRF</u>

$$\frac{\prod_{t=1}^{T} \exp \theta^T f(x, y_{t-1}, t, y_t)}{\sum_{y' \in K^T} \prod_{t=1}^{T} \exp \theta^T f(x, y'_{t-1}, t, y'_t)}$$

B - Memm denominator $O(TD|K|)$
CRF denominator $O(TD|K|^T)$
For CRF we have to consider that y_t depends not only on y_{t-1}, but also every other state in the sequence $x_1 \ldots x_T$

C- Forward Backward algorithm used

"it finds the optimal transition forward in time for a given state, over all the states while keeping track of the optimal states in a. It then finds the optimal transition for that state backward in time, over all the states in b.
These 2 probabilities are multiplied together in c to result in the probability of seeing a state first and then second

D - <u>CRF is used in training</u>: inference involves utilizing the dynamic programing tables created during training

Problem 6 -

A.) The model with the highest perplexity will be the <u>unigram model</u>. With no context other than $p(x) = \frac{count(x)}{*}$, it will perform worse than those that consider what comes before x at T-1 and T-2 because $p(x)_{unigram} \leq p(x)_{bigram/trigram}$

This will result in a lower perplexity considering $L = -\frac{\sum_{i=1}^{n} \log p(x^i)}{N}$ and perplexity $= 2^{\frac{NL}{T}}$

B.) Approximate perplexity of a uniform model:

$\log p(x^i) = \log \prod p(x^i)$

$T = \sum_{i=1}^{n} T_i$

$\log p(x^i) = T_i \log\left(\frac{1}{|V|}\right)$

$\frac{N}{T} L = \frac{-N}{T} \frac{\sum_{i=1}^{n} T_i}{N} \log\left(\frac{1}{|V|}\right)$

$= \log(|V|)$

$\boxed{\text{So we have perplexity } 2^{\log(|V|)} = O(|V|)}$

C.) Problems: will assign zero probability to any (u,v) not seen in the training set. Also, if count(u) is equal to zero, it is not well defined

Solutions:

1. Add one smoothing

$\frac{c^*(u,v)+1}{count(u)+|V|}$

⑥

a. $\theta(v|u) > 0$

i. $\min count(u,v) = 0$

ii. $\min count(u) = 0$

iii. $\theta(v|u) = \frac{0+1}{0+|V|} = \frac{1}{|V|}$

b. This is a probability distribution for any u

i $\sum_x \theta(x) = 1$

ii $\theta(v|u) = \frac{count(u,v)+1}{count(u)+|V|} =$

$$\sum_v \theta(v|u) = \frac{count(u,v1)+1}{count(u)+|V|} + \frac{count(u,v2)+1}{count(u)+|V|} + \dots \frac{count(u,vn)+1}{count(u)+|V|}$$

$$= \frac{count(u)+|V|}{count(u)+|V|} = 1$$

2. Add-K smoothing

a. $\theta(v|u) > 0$

i $\min count(u,v) = 0$

ii $\min count(u) = 0$

iii $\theta(v|u) = \frac{0+k}{0+k|V|} = \frac{1}{|V|}$

b. This is a probability distribution for any u

i $\sum_x \theta(x) = 1$

ii $\theta(v|u) = \frac{count(u,v)+k}{count(u)+k|V|} =$

$$\sum_{v \in (v_1, v_2, \dots, v_n)} \theta(v|u) = \frac{count(u,v1)+k}{count(u)+k|V|} + \frac{count(u,v2)+k}{count(u)+k|V|} + \dots \frac{count(u,vn)+k}{count(u)+k|V|}$$

$$= \frac{count(u)+k|V|}{count(u)+k|V|} = 1$$