Part #3 18 points

For the following two questions, write TRUE or FALSE below the question. **PLEASE GIVE JUSTIFICATION FOR YOUR ANSWERS: AT MOST 50% CREDIT WILL BE GIVEN FOR ANSWERS WITH NO JUSTIFICATION.**

For all questions in this section we assume as usual that a language model consists of a vocabulary $\mathcal{V}$, and a function $p(x_1 \ldots x_n)$ such that for all sentences $x_1 \ldots x_n \in \mathcal{V}^\dagger$, $p(x_1 \ldots x_n) \geq 0$, and in addition $\sum_{x_1 \ldots x_n \in \mathcal{V}^\dagger} p(x_1 \ldots x_n) = 1$. Here $\mathcal{V}^\dagger$ is the set of all sequences $x_1 \ldots x_n$ such that $n \geq 1$, $x_i \in \mathcal{V}$ for $i = 1 \ldots (n-1)$, and $x_n = \text{STOP}$.

We assume that we have a bigram log-linear language model, with

$$p(x_1 \ldots x_n) = \prod_{i=1}^{n} p(x_i | x_{i-1}; \theta)$$

where the bigram probabilities $p(x_i|x_{i-1};\theta)$ are defined using a log-linear model. Specifically, the model makes use of a feature vector definition $f(x, y)$, that maps each bigram (x, y) to a feature vector $f(x, y) \in \mathbb{R}^d$, and a parameter vector $\theta \in \mathbb{R}^d$, with

$$p(y|x;\theta) = \frac{\exp\left(\theta \cdot f(x, y)\right)}{\sum_{y' \in \mathcal{V} \cup \{\text{STOP}\}} \exp\left(\theta \cdot f(x, y')\right)}$$

Question 6 (4 points) Given a training corpus consisting of bigrams $(x^{(j)}, y^{(j)})$ for $j = 1 \ldots n$, the parameters are chosen to be

$$\theta^* = \arg\max L(\theta)$$

where

$$L(\theta) = \sum_{j=1}^{n} \log p(y^{(j)} | x^{(j)}; \theta) - \frac{\lambda}{2} \sum_{k=1}^{d} (\theta_k)^2$$

Here $\lambda > 0$ is a positive constant.

True or false? For any test corpus such that every word in the test corpus is in the set $\mathcal{V}$, the perplexity under the parameters θ^* is less than ∞.

Question 7 (4 points) True or false? For any test corpus such that every word in the test corpus is in the set $\mathcal{V}$, there are parameters θ such that the perplexity on the test corpus is $N+1$ where $N = |\mathcal{V}|$.

Question 8 (10 points) If we again define $N = |\mathcal{V}|$, show that it is possible to define a log-linear language model with a single feature (i.e., $d = 1$) such that

$$p(y|x;\theta) = 0.8 \quad \text{if } x = y$$

and

$$p(y|x;\theta) = \frac{0.2}{N} \quad \text{if } x \neq y$$

You should write down your definition for the single feature $f_1(x, y)$, and show the value for the parameter θ_1 that gives the above distribution.