CS688 Assignment 02

Sourabh Kulkarni | 29572060

Problem 2

1. We parameterize the CRF in log-space, so that the operations involved in the inference and learning of these models, which are typically long chains of multiplications, do not cause underflow or overflow in limited precision computation. In log-space, multiplications become additions and additions are done by efficient log-sum-exponentiation operations.
   When we perform factor reduction on the CRF, we obtain a subgraph which is a chain. The factor reduction operation conditions the Y nodes in the chain on their corresponding Xs which can be denoted as the feature potential in log-space.
   Inference on chain structure can be efficiently done in linear time using the sum-product algorithm. This algorithm computes two partial factors which will be denoted as α and Ω which are computed using messages form the neighbors. Computing these partial factors also involves the feature potentials of the node.
   Using these partial factors, we can obtain log-probabilities and pairwise log-probabilities and also the partition function, which are all the components required for inference.

Problem 3

1. The average log likelihood function for the CRF given a data of N images x_i and label y_i is:

$$\mathcal{L}(W|\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} logP_W(\boldsymbol{y}_i|\boldsymbol{x}_i)$$

Where the conditional probability is given in terms of the energy function and the partition function:

$$P_W(\boldsymbol{y}_i|\boldsymbol{x}_i) = -\frac{\exp(-E_W(\boldsymbol{y}_i, \boldsymbol{x}_i))}{Z_W(\boldsymbol{x_i})}$$

The energy function is:

$$E_W(\boldsymbol{y}_i, \boldsymbol{x}_i) = -\left(\sum_{j=1}^{L_j}\sum_{c=1}^{C}\sum_{f=1}^{F} W_{cf}^{F}[y_{ij} = c]x_{ijf} + \sum_{j=1}^{L_{j-1}}\sum_{c=1}^{C}\sum_{c'=1}^{C} W_{cc'}^{T}[y_{ij} = c][y_{ij+1} = c']\right)$$

And the conditional partition function is:

$$Z_W(\boldsymbol{x_i}) = \sum_{\boldsymbol{y}} \exp(-E_W(\boldsymbol{y}, \boldsymbol{x}_i)$$

2. Now we derive the average conditional log likelihood function w.r.t W_{cf}^{F}. First, we apply chain rule:

$$\frac{\partial \mathcal{L}(W|\mathcal{D})}{\partial W_{cf}^{F}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{P_W(\boldsymbol{y}_i|\boldsymbol{x_i})}\frac{\partial P_W}{\partial W_{cf}^{F}}P_W(\boldsymbol{y}_i|\boldsymbol{x_i})$$

now we obtain the partial derivative of P_W w.r.t W_{cf}^{F}using the quotient rule:

$$\frac{\partial P_W}{\partial W_{cf}^{F}} = \frac{\exp(-E_W(\boldsymbol{y},\boldsymbol{x}_i)}{\sum_{\boldsymbol{y}'}\exp(-E_W(\boldsymbol{y}',\boldsymbol{x}'_i)}\frac{-\partial E_W(\boldsymbol{y},\boldsymbol{x}_i)}{\partial W_{cf}^{F}}$$
$$-\frac{\exp(-E_W(\boldsymbol{y},\boldsymbol{x}_i)}{[\sum_{\boldsymbol{y}'}\exp(-E_W(\boldsymbol{y}',\boldsymbol{x}'_i)]^{2}}\sum_{\boldsymbol{y}'}\exp(-E_W(\boldsymbol{y}',\boldsymbol{x}'_i)\ \frac{-\partial E_W(\boldsymbol{y}',\boldsymbol{x}'_i)}{\partial W_{cf}^{F}}$$

looks like we have some P_Ws in the above equation. Substituting those:

$$\frac{\partial P_W}{\partial W_{cf}^{F}} = P_W(\boldsymbol{y}_i|\boldsymbol{x_i})\frac{-\partial E_W(\boldsymbol{y_i},\boldsymbol{x}_i)}{\partial W_{cf}^{F}}$$
$$-P_W(\boldsymbol{y}_i|\boldsymbol{x_i})\sum_{\boldsymbol{y}'}P_W(\boldsymbol{y}'|\boldsymbol{x}'_i)\ \frac{-\partial E_W(\boldsymbol{y}',\boldsymbol{x}'_i)}{\partial W_{cf}^{F}}$$

now we compute partial derivative of E_Ww.r.t W_{cf}^{F}:

$$\frac{\partial E_W(\boldsymbol{y_i},\boldsymbol{x}_i)}{\partial W_{cf}^{F}} = \sum_{j=1}^{L_j}\sum_{c=1}^{C}\sum_{f=1}^{F}W_{cf}^{F}[y_{ij}=c]x_{ijf}$$

substituting back all of this in the original conditional log likelihood partial derivative expression:

$$\frac{\partial \mathcal{L}(W|\mathcal{D})}{\partial W_{cf}^{F}} = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j=1}^{L_j}\sum_{c=1}^{C}\sum_{f=1}^{F}\left[y_{ij}=c\right]x_{ijf}\right.$$
$$\left.-\sum_{j=1}^{L_j}\sum_{c=1}^{C}\sum_{f=1}^{F}P_W(y_{ij}|x_{ijf})\left[y_{ij}=c\right]x_{ijf}\right]$$

The expression inside the second summation can be interpreted as the conditional probability of $y_{ij}=c$ given x_{ijf} finally we have:

$$\frac{\partial \mathcal{L}(W|\mathcal{D})}{\partial W_{cf}^{F}} = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j=1}^{L_j}\sum_{c=1}^{C}\sum_{f=1}^{F}\left[y_{ij}=c\right]x_{ijf}\right.$$
$$\left.-\sum_{j=1}^{L_j}\sum_{c=1}^{C}\sum_{f=1}^{F}P_W\left(y_{ij}=c\middle|x_{ijf}\right).x_{ijf}\right]$$

3. Calculating partial derivative of the conditional log likelihood w.r.t $W_{cc'}^{T}$will be exactly similar until the point of the partial of the energy, which will be:

$$\frac{\partial E_W(\boldsymbol{y_i},\boldsymbol{x_i})}{\partial W^T_{cc'}} = \sum_{j=1}^{L_{j-1}} \sum_{c=1}^{C} \sum_{c'=1}^{C} [y_{ij} = c][y_{ij+1} = c']$$

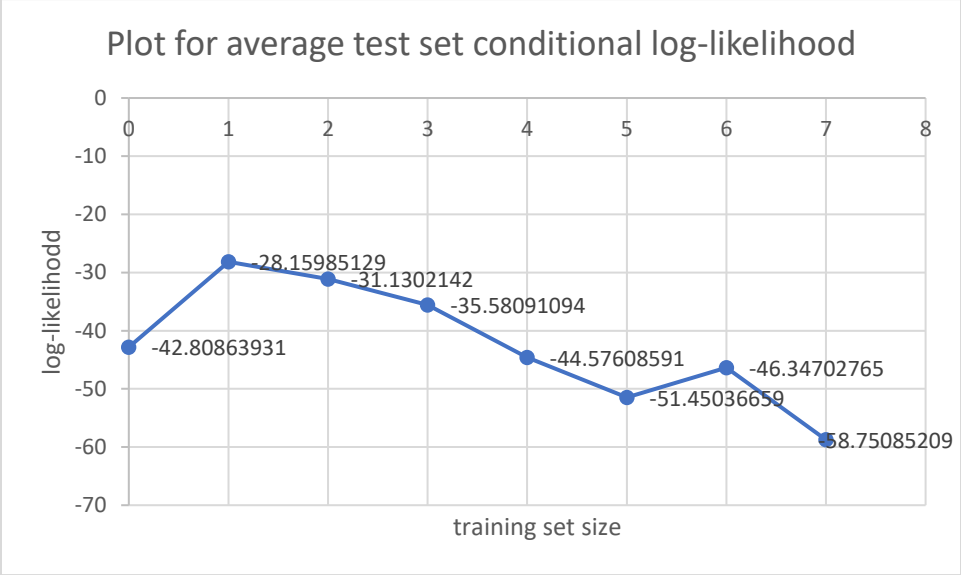
which makes the partial derivative of the conditional log likelihood:

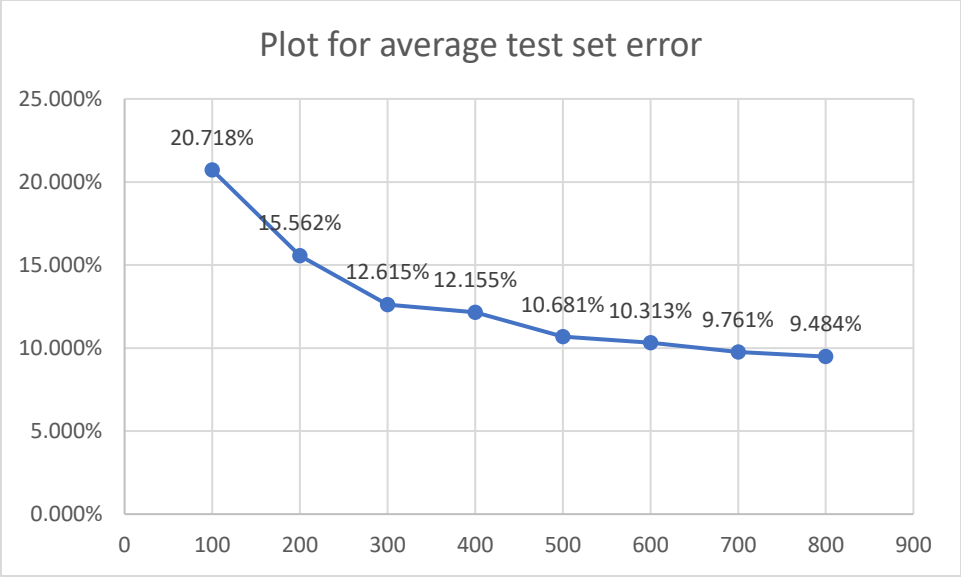
$$\frac{\partial \mathcal{L}(W|\mathcal{D})}{\partial W^F_{cc'}} = \frac{1}{N} \sum_{i=1}^{N} \left[\sum_{j=1}^{L_{j-1}} \sum_{c=1}^{C} \sum_{c'=1}^{C} \left[y_{ij} = c\right]\left[y_{ij+1} = c'\right] - \sum_{j=1}^{L_{j-1}} \sum_{c=1}^{C} \sum_{c'=1}^{C} P_W(y_{ij} = c, y_{ij+1} = c'|x_{ij}) \right]$$

4. As the equations above show, computation of gradients of the log likelihood w.r.t the parameters use the individual and pairwise marginals which were already computed in the predict_logprob method. The computation also involves summing over features and incrementation operations, which are not compute intensive. Hence with $O(LC^2)$ time we can simply assign those marginals as the gradients of the respective parameters.

Problem 5

1.





As the plot shows, the lowest error was 9.484%

2. To generate words of different sizes, I used the np.random.randint to generate Xs and np.random.choice over the CHARS to generate Ys over the range of lengths 1-20. The Xs and Ys were generated in the same shape as the ones used in training and testing, so they could be passed directly to the predict function. I then timed the predictions for a batch size of 50 per length and computed the average over those 50 predictions. As the plot suggests, the inference time is linear in word length.