1 Exhaustive Inference

1.1 Test word node potential

Table 1 – Feature potentials for *test_word_1*

	Categories									
	0	1	2	3	4	5	6	7	8	9
1	−7.6443	18.4683	−6.3285	10.4224	−4.9671	−1.9340	−0.9451	−5.6571	5.3952	−6.8098
2	−4.0744	5.7448	1.1763	−1.7931	−1.2122	−1.7848	−8.2998	3.0951	6.8065	0.3416
3	−10.2081	0.8973	17.1910	−12.0176	5.5793	−0.5940	−21.4263	9.1489	9.4824	1.9471
4	6.4648	24.5312	−13.3429	5.8712	−10.9548	−11.4964	−5.4946	−7.1956	8.0456	3.5714

1.2 Energy Calculation

Table 2 – Energy for $test\_word_i$

Test Word	Energy
1	63.9793
2	89.6109
3	96.9406

1.3 Log Partition function

Table 3 – Log Partition for $test\_word_i$

Test Word	Log Partition
1	67.6019
2	89.6144
3	103.5276

1.4 Most Likely labels

Table 4 – Most Likely labels for $test\_word_i$

Test Word	Word	Probabitliy
1	trat	0.7958
2	hire	0.9965
3	riser	0.9370

1.5 Marginal Label Probabilities

Table 5 – Marginal label probabilities $test\_word_i$

Category	1	2	3	4
0	7.2227×10^{-12}	1.2658×10^{-5}	1.1321×10^{-12}	8.8683×10^{-9}
1	9.9952×10^{-1}	1.7247×10^{-1}	2.2945×10^{-8}	10.0000×10^{-1}
2	2.6262×10^{-11}	2.7314×10^{-3}	9.9946×10^{-1}	2.1357×10^{-17}
3	4.7272×10^{-4}	1.7528×10^{-4}	1.6119×10^{-13}	7.4054×10^{-9}
4	7.1555×10^{-11}	2.0074×10^{-4}	3.6976×10^{-6}	3.2900×10^{-16}
5	2.1138×10^{-9}	1.4005×10^{-4}	1.7611×10^{-8}	1.4410×10^{-16}
6	3.2960×10^{-9}	1.0646×10^{-7}	5.1721×10^{-18}	5.3711×10^{-14}
7	4.3493×10^{-11}	2.6735×10^{-2}	2.8353×10^{-4}	1.3178×10^{-14}
8	2.6281×10^{-6}	7.9660×10^{-1}	2.5376×10^{-4}	6.3940×10^{-8}
9	1.0694×10^{-11}	9.3629×10^{-4}	9.4638×10^{-8}	6.3736×10^{-10}

2 Sum-Product Message Passing

2.1 Log message values

Table 6 – Message values in log-space

	$m_{1\to 2}(Y_2)$	$m_{2\to 1}(Y_1)$	$m_{2\to 3}(Y_3)$	$m_{3\to 2}(Y_2)$
e	18.5893	49.5924	25.6511	41.8098
t	17.8153	49.1330	25.2369	42.2842
a	18.7494	49.5675	25.5984	41.7732
i	18.5227	49.5224	25.5779	42.2232
n	18.1808	49.2085	25.2716	42.1198
o	18.6773	49.5611	25.6012	41.8359
s	18.0913	49.0165	25.0715	41.7550
h	18.8341	49.4006	25.3880	42.0509
r	18.3634	49.3573	25.4145	42.2045
d	18.2164	49.1503	25.2026	42.0703

2.2 Marginal Probabilities

Table 7 – Marginal Probabilities

	Sequence			
char	0	1	2	3
e	7.2227×10^{-12}	1.2658×10^{-5}	1.1321×10^{-12}	8.8683×10^{-9}
t	9.9952×10^{-1}	1.7247×10^{-1}	2.2945×10^{-8}	10.0000×10^{-1}
a	2.6262×10^{-11}	2.7314×10^{-3}	9.9946×10^{-1}	2.1357×10^{-17}
i	4.7272×10^{-4}	1.7528×10^{-4}	1.6119×10^{-13}	7.4054×10^{-9}
n	7.1555×10^{-11}	2.0074×10^{-4}	3.6976×10^{-6}	3.2900×10^{-16}
o	2.1138×10^{-9}	1.4005×10^{-4}	1.7611×10^{-8}	1.4410×10^{-16}
s	3.2960×10^{-9}	1.0646×10^{-7}	5.1721×10^{-18}	5.3711×10^{-14}
h	4.3493×10^{-11}	2.6735×10^{-2}	2.8353×10^{-4}	1.3178×10^{-14}
r	2.6281×10^{-6}	7.9660×10^{-1}	2.5376×10^{-4}	6.3940×10^{-8}
d	1.0694×10^{-11}	9.3629×10^{-4}	9.4638×10^{-8}	6.3736×10^{-10}

2.3 Inference

2.3.1 Marginal Pair Probabilities

Table 8 – Marginal Pair Probabilities

1	t	h	a
t	1.7236×10^{-1}	2.6730×10^{-2}	2.7305×10^{-3}
h	1.5904×10^{-11}	5.3897×10^{-13}	7.2001×10^{-14}
a	7.4658×10^{-12}	3.3086×10^{-13}	2.7860×10^{-14}
2	**t**	**h**	**a**
t	2.2314×10^{-9}	0.0001	0.1724
h	1.2104×10^{-9}	0.0000	0.0267
a	1.4997×10^{-10}	0.0000	0.0027
3	**t**	**h**	**a**
t	2.2945×10^{-8}	1.0581×10^{-21}	2.0796×10^{-24}
h	2.8353×10^{-4}	2.8571×10^{-18}	7.3432×10^{-21}
a	9.9946×10^{-1}	1.3171×10^{-14}	2.1337×10^{-17}

2.3.2 Predictions

Actual	Predicted
that	trat
hire	hire
rises	riser
edison	edison
shore	shore

2.3.3 Character-level accuracy

Accuracy: `0.8991`

3 Maximum Likelihood Learning Derivation

3.1 Average Log likelihood

$P_W(y,x) = \frac{1}{Z(W)} \exp\left(\sum_{j=1}^{L_i}\sum_{f=1}^{F} W^F_{y_j f} x_{jf} + \sum_{j=1}^{L_i-1} W^T_{y_j y_{j+1}}\right)$

The average log likelihood is given by,

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\log P_W(y^{(i)},x^{(i)}) &= \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{Z(W,x^{(i)})}\exp\sum_{j=1}^{L_i}\sum_{f=1}^{F} W^F_{y_j^{(i)} f} x^{(i)}_{jf} + \sum_{j=1}^{L_i-1} W^T_{y_j^{(i)} y_{j+1}^{(i)}}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{L_i}\sum_{f=1}^{F} W^F_{y_j^{(i)} f} x^{(i)}_{jf} + \sum_{j=1}^{L_i-1} W^T_{y_j^{(i)} y_{j+1}^{(i)}} - \log Z(W,x^{(i)})\right)
\end{aligned} \tag{1}$$

3.2 Derivative of Log Likelihood w.r.t W^F_{cf}

Let the average likelihood be defined as,

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\log P_W(y^{(i)},x^{(i)}) \tag{2}$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial W_{c'f'}} &= \frac{\partial}{\partial W_{c'f'}} \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{L_i} \sum_{f=1}^{F} W^F_{y_j^{(i)} f} x_{jf}^{(i)} + \sum_{j=1}^{L_i - 1} W^T_{y_j^{(i)} y_{j+1}^{(i)}} - \log Z(W, x^{(i)}) \right) \\
&\text{Taking derivative inside the summations} \\
&= \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{L_i} \sum_{f=1}^{F} \frac{\partial}{\partial W_{c'f'}} W^F_{y_j^{(i)} f} x_{jf}^{(i)} + \sum_{j=1}^{L_i - 1} \frac{\partial}{\partial W_{c'f'}} W^T_{y_j^{(i)} y_{j+1}^{(i)}} - \frac{\partial}{\partial W_{c'f'}} \log Z(W, x^{(i)}) \right) \\
&\text{Since, } W^T \text{ is constant w.r.t } W^F \text{, it is 0} \\
&= \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{L_i} \sum_{f=1}^{F} \mathbb{I}[y_j^{(i)} = c', f = f'] x_{jf}^{(i)} - \frac{1}{Z(W,x)} \frac{\partial}{\partial W_{c'f'}} \sum_{\mathbf{y}} \exp \left(\sum_{j=1}^{L_i} \sum_{f=1}^{F} W^F_{y_j^{(i)} f} x_{jf}^{(i)} + \sum_{j=1}^{L_i - 1} W^T_{y_j^{(i)} y_{j+1}^{(i)}} \right) \right) \\
&= \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{L_i} \sum_{f=1}^{F} \mathbb{I}[y_j^{(i)} = c', f = f'] x_{jf}^{(i)} \right. \\
&\quad \left. - \frac{1}{Z(W,x)} \sum_{\mathbf{y}} \exp \left(\sum_{j=1}^{L_i} \sum_{f=1}^{F} W^F_{y_j^{(i)} f} x_{jf}^{(i)} + \sum_{j=1}^{L_i - 1} W^T_{y_j^{(i)} y_{j+1}^{(i)}} \right) \mathbb{I}[y_j^{(i)} = c', f = f'] x_{jf}^{(i)} \right) \\
&= \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{L_i} \sum_{f=1}^{F} \mathbb{I}[y_j^{(i)} = c', f = f'] x_{jf}^{(i)} - \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}) \mathbb{I}[y_j^{(i)} = c', f = f'] x_{jf}^{(i)} \right) \\
&= \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{L_i} \sum_{f=1}^{F} \mathbb{I}[y_j^{(i)} = c', f = f'] x_{jf}^{(i)} - \mathbb{E}_{P(y|x)} \left[\mathbb{I}[y_j^{(i)} = c', f = f'] x_{jf}^{(i)} \right] \right)
\end{aligned}
\tag{3}
$$

3.3 Derivative of Log Likelihood w.r.t $W^T_{cc'}$

$$\frac{\partial \mathcal{L}}{\partial W_{cc'}} = \frac{\partial}{\partial W_{cc'}} \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{L_i} \sum_{f=1}^{F} W^F_{y_j^{(i)} f} x_{jf}^{(i)} + \sum_{j=1}^{L_i - 1} W^T_{y_j^{(i)} y_{j+1}^{(i)}} - \log Z(W, x^{(i)}) \right)$$

Taking derivative inside the summation

$$= \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{L_i} \sum_{f=1}^{F} \frac{\partial}{\partial W_{cc'}} W^F_{y_j^{(i)} f} x_{jf}^{(i)} + \sum_{j=1}^{L_i - 1} \frac{\partial}{\partial W_{cc'}} W^T_{y_j^{(i)} y_{j+1}^{(i)}} - \frac{\partial}{\partial W_{cc'}} \log Z(W, x^{(i)}) \right)$$

Since W^F is constant w.r.t W^T,

$$= \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{L_i - 1} \mathbb{I}[y_j^{(i)} = c, y_{j+1}^{(i)} = c'] - \frac{1}{Z(W,x)} \frac{\partial}{\partial W_{cc'}} \sum_{\mathbf{y}} \exp \left(\sum_{j=1}^{L_i} \sum_{f=1}^{F} W^F_{y_j^{(i)} f} x_{jf}^{(i)} + \sum_{j=1}^{L_i - 1} W^T_{y_j^{(i)} y_{j+1}^{(i)}} \right) \right) \tag{4}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{L_i - 1} \mathbb{I}[y_j^{(i)} = c, y_{j+1}^{(i)} = c'] - \frac{1}{Z(W,x)} \sum_{\mathbf{y}} \exp \left(\sum_{j=1}^{L_i - 1} W^T_{y_j^{(i)} y_{j+1}^{(i)}} \right) \mathbb{I}[y_j^{(i)} = c, y_{j+1}^{(i)} = c'] \right)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{L_i - 1} \mathbb{I}[y_j^{(i)} = c, y_{j+1}^{(i)} = c'] - \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}) \mathbb{I}[y_j^{(i)} = c, y_{j+1}^{(i)} = c'] \right)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{L_i - 1} \mathbb{I}[y_j^{(i)} = c, y_{j+1}^{(i)} = c'] - \mathbb{E}_{P(y|x)} \left[\mathbb{I}[y_j^{(i)} = c, y_{j+1}^{(i)} = c'] \right] \right)$$

3.4 Using Sum-Product in likelihood

The Sum-Product method allows to compute the overall potential of a configuration. This potential is equivalent to the unnormalized probability. Formally, $P(\mathbf{y}, \mathbf{x}) \propto \prod_{j=1}^{L} \phi^F(y_j, x_j) \prod_{j=1}^{L-1} \phi^T(y_j, y_{j+1})$. From the sum-product method, we can re-write this as,

$$P(\mathbf{y}, \mathbf{x}) \propto \sum_{y_1} \phi^F(y_1, x_1) \mathbf{m}_{2\to 1}(y_1) \tag{5}$$

The message $\mathbf{m}_{2\to 1}(y_1)$, encodes the "happiness" of the sequence $\in (2, 3, ...)$. We can use this to calculate the log-partition function efficiently.

Now, while computing the single and pair-wise marginal probabilities, we multiply the forward $(\mathbf{m}_{i\to i+1})$ and backward messages $(\mathbf{m}_{i\to i-1})$ along with the feature potentials to obtain the single/marginal probabilities. Lastly, we obtain a distribution over the sequence length, which we can normalize over to get the *likelihood* of the sequence. Using the previous result, we can obtain an average log likelihood over N datapoints.

Similarly, to compute the derivatives, the conditional probability $P(y|x)$, can be expressed in terms of single and marginal probabilities. We can use the already pre-computed marginals to efficiently compute $P(y|x)$

3.5 Training Average Log Likelihood

`Average likelihood` of 50 train words: `-4.583959`

4 Numerical Optimization Warm-Up

4.1 Derivative of $f(x,y)$

$$
\begin{aligned}
f_w(x,y) &= -(1-x)^2 - 100(y-x^2)^2 \\
\frac{\partial f(x,y)}{\partial x} &= -\frac{\partial}{\partial x}(1-x)^2 - 100\frac{\partial}{\partial x}(y-x^2)^2 \\
&= -2(1-x)(-1) + 200(y-x^2)\frac{\partial}{\partial x}x^2 \\
&= 2(1-x) + 400x(y-x^2) \\
\frac{\partial f(x,y)}{\partial y} &= -\frac{\partial}{\partial y}(1-x)^2 - 100\frac{\partial}{\partial y}(y-x^2)^2 \\
&= 0 - 200(y-x^2)\frac{\partial}{\partial y}y \\
&= -200(y-x^2)
\end{aligned}
$$

4.2 Numerical Optimizer

I used the `scipy.optimize.minimize` using the L-BFGS-B solver.

Maximum location: `x = 1., y = 0.99999999`
Maximum value: `2.6436083956216185e-17`