1. Machine Learning & Neural Networks

(a) Adam Optimizer

i. Briefly explain in 2-4 sentences (you don't need to prove mathematically, just give an intuition) how using **m** stops the updates from varying as much and why this low variance may be helpful to learning, overall.

Answer: Momentum **m** is effectively an exponential moving average (EMA) of the real gradient. EMA can smooth a noisy series of gradients by exponentially weighting past gradients, and therefore stops the updates from varying too much. When the variance of the noisy gradient is large, a learning algorithm might spend much time bouncing around, leading to slower convergence and worse performance. Therefore, low variance updates can be helpful to find local minima.

ii. Since Adam divides the update by $\sqrt{\mathbf{v}}$, which of the model parameters will get larger updates? Why might this help with learning?

Answer: The quantity **v** is effectively the EMA of the elementwise square of the gradient. Then $\sqrt{\mathbf{v}}$ can be seen as the EMA of the absolute value or magnitude of the momentum **m** (but not exactly). The expression $\mathbf{m}/\sqrt{\mathbf{v}}$ can be seen as a normalization of **m**, which magnifies the small components of **m** ($\ll$ 1) and decreases the magnitude of large components. Therefore, this can give modal parameters with small gradients larger updates. This can help learning by giving quicker updates to model parameters with small gradients, and thus leading to a quicker convergence.

(b) Dropout

i. What must γ equal in terms of p_{drop}? Briefly justify your answer or show your math derivation using the equations given above.

Answer:

$$E_{p_{\text{drop}}}\left[h_{\text{drop}}\right]_i = p_{d_i=0}(0) + p_{d_i=1}\gamma h_i = \left(1 - p_{\text{drop}}\right)\gamma h_i = h_i.$$

Therefore,

$$\gamma = \frac{1}{1 - p_{\text{drop}}}$$

ii. Why should dropout be applied during training? Why should dropout NOT be applied during evaluation?

Answer: During training, without dropout, model parameters tend to overfit to some features and neighboring parameters can have high reliance on each other. In this way, the model is fragile and overfitting, and cannot handle out-of-distribution unseen data. Dropout can randomly cut of connections between parameters (weights) during training by zeroing out gradients. Therefore, dropout can reduce the reliance between parameters, making the trained model more robust and better capable of generalization. During evaluation, we need all the connections between parameters and consistent outputs, therefore, dropout should not be applied.

2. Neural Transition-Based Dependency Parsing

(a) At each step, give the configuration of the stack and buffer, as well as what transition was applied this step and what new dependency was added (if any).

Answer:

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 → I	LEFT-ARC
[ROOT, parsed, this]	[sentence, correctly]		SHIFT
[ROOT, parsed, this, sentence]	[correctly]		SHIFT
[ROOT, parsed, sentence]	[correctly]	sentence → this	LEFT-ARC
[ROOT, parsed]	[correctly]	parsed → sentence	RIGHT-ARC
[ROOT, parsed, correctly]	[ ]		SHIFT
[ROOT, parsed]	[ ]	parsed → correctly	RIGHT-ARC
[ROOT]	[ ]	ROOT → parsed	RIGHT-ARC

(b) A sentence containing n words will be parsed in how many steps (in terms of n)? Briefly explain in 1-2 sentences why.

Answer: $2n$

Since each word in the sentence needs two transitions before being removed from the stack: SHIFT and one of the ARCs, and each step during parsing must perform only one of these two transition for a word. Therefore, in general, a total of $2n$ steps are needed.

(e) Report the best UAS your model achieves on the dev set and the UAS it achieves on the test set.

Answer:

- The best UAS on the dev set: 88.73
- The best UAS on the test set: 89.15

(f) For each sentence, state the type of error, the incorrect dependency, and the correct dependency. While each sentence should have a unique error type, there may be multiple possible correct dependencies for some of the sentences.

i. I disembarked and was heading to a wedding fearing my death.
- Error type: Verb Phrase Attachment Error
- Incorrect dependency: wedding → fearing
- Correct dependency: heading → fearing

ii. It makes me want to rush out and rescue people from dilemmas of their own making.
- Error type: Coordination Attachment Error
- Incorrect dependency: makes → rescue
- Correct dependency: rush → rescue

iii. It is on loan from a guy named Joe O’Neill in Midland, Texas.

- Error type: Prepositional Phrase Attachment Error
- Incorrect dependency: named → Midland
- Correct dependency: O’Neill → Midland

iv. Brian has been one of the most crucial elements to the success of Mozilla software.

- Error type: Modifier Attachment Error
- Incorrect dependency: element → most
- Correct dependency: crucial → most