Question 1: Machine Learning Neural Networks

Part a)

i) The formula for momentum updates the gradient by using an exponentially weighted moving average of the previous gradient value with weight β_1 and the latest calculated gradient with weight $(1-\beta_1)$. This causes the gradient to react much more slowly to changes in the gradient, since it only moves the gradient by some fraction towards the latest value. This reduced variance is effective in improving learning by keeping the optimization path moving in the direction where gradients agree over time, while giving little weight to new gradients that oppose the general gradient path. Therefore the optimization path tends to be less jittery and moves more consistently in the direction of minimization.

ii) Since the size of the gradient is in the denominator, smaller gradients will get a larger update while larger gradients get a smaller update. This makes sense as we want to take larger steps in directions where the optimization surface is shallow and there is less of a chance of overstepping while we take smaller steps around more extreme gradients. If we think of our surface as an oblong bowl, we want to take longer steps along the long axis and shorter steps down the more extreme short axis to ensure we don't miss the minimum. This is exactly what Adam incorporates.

Part b)

i) We need $\gamma = 1/(1-p_{drop})$:

$$\begin{aligned}\mathbb{E}_{p_{\text{drap}}}\ [\mathbf{h}_{\text{drop}}\ ]_i &= \mathbb{E}_{p_{\text{drap}}}\ [\gamma \mathrm{d} \odot \mathrm{h}]_i \\ &= \mathbb{E}_{p_{\text{drap}}}\ [\gamma \mathrm{d}_i \mathrm{h}_i] \\ &= \gamma(1-p_{drop})\mathrm{h}_i\end{aligned}$$

For $\mathbb{E}_{p_{\text{drap}}}\ [\mathbf{h}_{\text{drop}}\ ]_i = h_i$, we then see that this implies $\gamma(1-p_{drop})\mathrm{h}_i = h_i \implies \gamma = 1/(1-p_{drop})$

ii) Dropout is a form of regularization during training. The network cannot put too much weight on any given unit, since it may not be present in the next iteration of training. Weights become spread over more units, reducing the importance of any given unit. For each training batch, we are effectively training a smaller network, which restrict the ability to overfit. We also prevent feature co-adaptation, in which our training constructs unit weights that are only useful in the presence of other units. Since we drop units at random, model cannot count on units firing together.

However at test time, dropout would simply add noise to our predictions since it is a random process unrelated to the prediction task. If we see dropout as training an ensemble model each drawing on a portion of the units in the model, then at test time we want the output of that ensemble, not a sub-model composed of only some of those units.

Question 2. Neural Transition-Based Dependency Parsing

a)

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 $\rightarrow$ I	LEFT-ARC
[ROOT, parsed, this]	[sentence, correctly]		SHIFT
[ROOT, parsed, this , sentence]	[correctly]		SHIFT
[ROOT, parsed, sentence]	[correctly]	sentence $\rightarrow$ this	LEFT-ARC
[ROOT, parsed]	[correctly]	parsed $\rightarrow$ sentence	RIGHT-ARC
[ROOT, parsed, correctly]	[]		SHIFT
[ROOT, parsed]	[]	parsed $\rightarrow$ correctly	RIGHT-ARC
[ROOT]	[]	ROOT $\rightarrow$ parsed	RIGHT-ARC

b) Each word must be passed from the buffer to the stack and then each word gets a dependency arc applied. Thus a sentence of n words should have 2n operations and therefore the algorithm is linear in n (with an oracle determining actions).

e) Best UAS on Dev set: 88.42

Best UAS on Test set: 88.86

f) i) Error type: Verb Phrase Attachment Error
Incorrect dependency: wedding $\rightarrow$ fearing
Correct dependency: disembarked $\rightarrow$ fearing

ii) Error type: Coordination Attachment Error
Incorrect dependency: makes $\rightarrow$ rescue
Correct dependency: rush $\rightarrow$ rescue

iii) Error type: Prepositional Phrase Attachment Error
Incorrect dependency: named $\rightarrow$ Midland
Correct dependency: guy $\rightarrow$ Midland

iiii) Error type: Modifier Attachment Error
Incorrect dependency: elements $\rightarrow$ most
Correct dependency: crucial $\rightarrow$ most