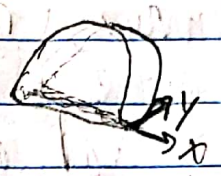
Assignment 3: Dependency Parsing

1. a) Momentum: $m \leftarrow \beta_1 m + (1-\beta_1) \nabla_\theta J_{\text{minibatch}}(\theta)$

i) $\theta \leftarrow \theta - \alpha m$

Consider the taco shape 

If this is a loss landscape, the y direction could have high but oscillating derivatives, but m will converge to the average of the oscillations, which will be much smaller and more stable. This helps to speed up convergence.

ii) Adaptive learning rates:

$$m \leftarrow \beta_1 m + (1-\beta_1) \nabla_\theta J(\theta)$$
$$v \leftarrow \beta_2 v + (1-\beta_2)(\nabla_\theta J(\theta) \odot \nabla_\theta J(\theta))$$
$$\theta \leftarrow \theta - \alpha \odot m / \sqrt{v}$$

v is a running average of gradient magnitude per parameter, so when dividing by $\sqrt{v}$, parameters with historically smaller gradients will get larger updates. This helps to accelerate convergence by preventing params w/ large gradients from exploding and prevents params w/ small gradients from converging very slowly.

6) i) Since each $d_i \sim \text{Bernoulli}(1-p_{drop})$,

$$E[h_{drop}] = \gamma k (1-p_{drop}) = k$$

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

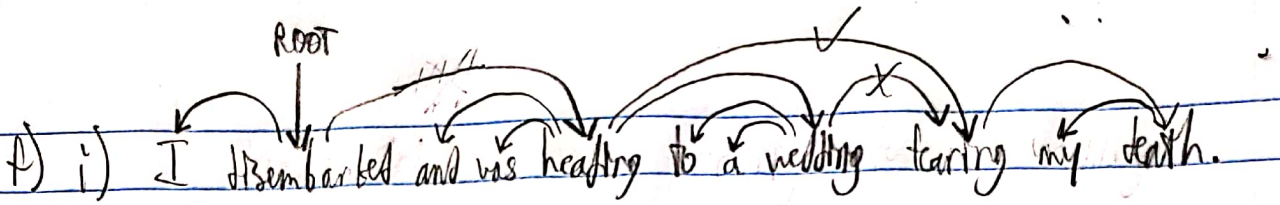
ii) Dropout introduces randomness, which is good for training a more generalizable network, but the very spirit out validation/testing is to have a deterministic evaluation of the network. Besides, dropout is a sort of bagging, and the ensembled learners perform better together.

2. a) ROOT I parsed this sentence correctly.

Stack	Buffer	New Dependency	Transition
[ROOT]	[I, parsed, this, sentence, correctly]		Initial Config
[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	RIGT-ARC

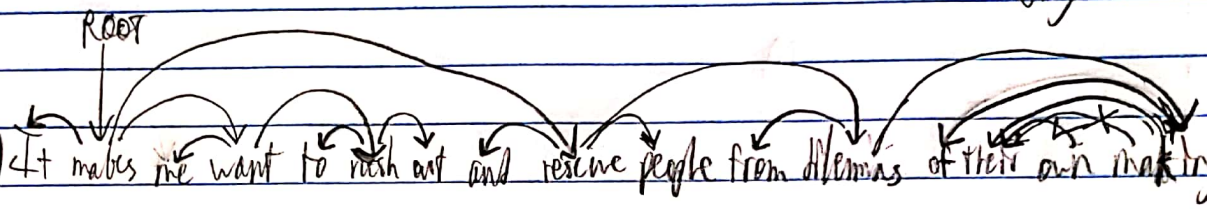
b) Since every word must go from the buffer to the stack $\Rightarrow n$ SHIFTS and since every word must be a dependent and removed $\Rightarrow n$ ARCS

$\Rightarrow 2n$ steps.

f) i) 

Error type: Verb phrase attachment error
Incorrect dependency: wedding → fearing
Correct dependency: heading → fearing

Since "fearing my death" doesn't modify "wedding", the verb phrase modifies the verb "heading" (modifies how I am heading to a wedding).

ii) 

Error type: Modifier attachment error
Incorrect dependency: making → their
Correct dependency: own → their