1. Machine Learning & Neural Networks

(a) i. Since β_1 is often set to 0.9, so even if the gradient wrt θ in current minibatch is large, it has little influence to the whole m, so it makes the gradients more stable, which can prevent overshooting during learning process.

ii. Parameters with smaller gradients will get larger updates, so it can accelerate learning process.

(b) i. $\gamma = \frac{1}{1-p_{drop}}$ Since dropout is not used during evaluation, so γ is needed to keep the activations of neurons in the same scale both during training and evaluation. Assume one neuron's output is x during training process, with dropout, the expectation of its output is $p_{drop} * 0 + (1 - p_{drop}) * x$, while during evalution process its output is x, so the γ needs to be $\frac{1}{1-p_{drop}}$.

ii. If dropout is applied during evaluation, the output of our model can be unstable.

2. Neural Transition-Based Dependency Parsing

(a)

Stack	Buffer	New Dependency	Transition
[Root, parsed, this]	[sentence, correctly]		SHIFT
[Root, parsed, this, sentence]	[correctly]		SHIFT
[Root, parsed, sentence]	[correctly]	this->sentence	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) Totally n steps.