A. Adam

(i) • momentum increases the dimension whose gradient point in the same direction reduces the dimensions whose gradient change directions.

• Thus acelerate the convergence.

(ii) • Weights that receive high gradients will reduce the update

• Weights that receive low gradients will increase the update

• This can normalize the update step

B. Dropout

(i) $\gamma = \frac{1}{1-p_{drop}}$

↳ we want $E[h_{drop}] = h$

$\Rightarrow \gamma(1-p_{drop}) \odot h = h$

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

(ii) Dropout is a regularization technique. Regularization is used to cope with overfitting which only exists in training. Also, we will get random output if applying dropout during evaluation.

(a)

[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) $2n$

- Every word moves from Buffer to Stack need n steps
- Every dependency moves one word from Stack to Transition, needing n steps.

(f)

Problem ID	Error Type	Incorrect	Correct
i	Verb ...	wedding → fearing	heading → fearing
ii	Coordination ...	rescue → and	rescue → rush
iii	Prepositional ...	named → Midland	guy → Midland
iv	Modifier ...	elements → most	crucial → most