CS 224n Assignment #3

Yoshihiro Kumazawa

June 1, 2020

1. **Machine Learning & Neural Networks**

   (a) i. When β_1 is large, $\boldsymbol{m}$ relies more on the history of the past gradients rather than the new one. For example, if $\beta_1 = 0.9$, the contribution of the new gradient to weight update is only 10% of that without momentum.

   ii. The model parameters with smaller gradients will get larger updates. When the variance of the gradients is high, Adam might help the parameters go in the direction of small gradients, which the vanilla update strategy would ignore due to rounding errors.

   (b) i. γ must equal $1/(1 - p_{\text{drop}})$ to make $\mathbb{E}_{p_{\text{drop}}}[\boldsymbol{h}_{\text{drop}}]_i = h_i$.

   ii.

2. **Neural Transition-Based Dependency Parsing**

   (a)

   (b)

   (c)

   (d)

   (e)

   (f)