

50.021 Artificial Intelligence

Theory Homework 6

Due: every Tuesday, 6PM

[Q1] The following is the AdaGrad algorithm for weight update.

$$\begin{aligned} \text{cache}_i &= \text{cache}_i + (\nabla_{w_i} L)^2 \\ w_i &= w_i - \frac{\eta}{\sqrt{\text{cache}_i} + \epsilon} \nabla_{w_i} L \end{aligned}$$

where w_i is the weight to be updated, $\nabla_{w_i} L$ is the gradient of the loss w.r.t w_i , ϵ is a hyperparameter between 10^{-8} and 10^{-4} and η is a hyperparameter similar to step size in SGD. List one difference between AdaGrad and SGD in terms of step size and **explain** what effects you expect from this difference.

[Q2] The following are the defining equations for a LSTM cell,

$$\begin{aligned} i_t &= \sigma(W^i x_t + U^i h_{t-1}) \\ f_t &= \sigma(W^f x_t + U^f h_{t-1}) \\ o_t &= \sigma(W^o x_t + U^o h_{t-1}) \\ \tilde{c}_t &= \tanh(W^c x_t + U^c h_{t-1}) \\ c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \\ h_t &= o_t \circ \tanh(c_t) \end{aligned}$$

The symbol \circ denotes element-wise multiplication and $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. Answer True/False to the following questions and give not more than 2 sentences explanation.

1. If $x_t = 0$ vector then $h_t = h_{t-1}$.
2. If f_t is very small or zero, then the error will not be back-propagated to earlier time steps.
3. The entries of f_t, i_t, o_t are non-negative.
4. f_t, i_t, o_t can be seen as probability distributions, which means that their entries are non-negative and their entries sum to 1.

[Q3] The defining equations for a GRU cell are,

$$\begin{aligned} z_t &= \sigma(W^z x_t + U^z h_{t-1}) \\ r_t &= \sigma(W^r x_t + U^r h_{t-1}) \\ \tilde{h}_t &= \tanh(W x_t + r_t \circ U h_{t-1}) \\ h_t &= z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h}_t \end{aligned}$$

1. Draw a diagram of this GRU cell.
2. Assume h_t and x_t are column vectors, with dimensions d_h and d_x respectively. What are the dimensions (rows \times columns) of the weight matrices W^z, W^r, W, U^z, U^r , and U ?
3. Like LSTM cells, GRU cells can tackle vanishing or exploding gradient problem too. By taking a look at the formula for LSTM in Q2, what is the main advantage of using GRU cells over LSTMs for some problems? Give an answer it at most 5 sentences.
Hint: We expect a qualitative answer (deep math proofs are not required) that comes with an explanation of the answer.