Homework 2 Solutions

2.

(a) We can ignore the first moment of Adam by setting $\beta_1 = 0$, which will result in RMSProp. **[1 mark]**

The final set of hyperparameters are $(\alpha_A = \alpha_R, \beta_1 = 0, \beta_2 = \gamma, \epsilon_A = \epsilon_R)$.

(b) Dependence of θ_t on the second moment v_t can be reduced by either of the following:

i. By setting ϵ_A and α_A to be extremely large values while maintaining the same ratio, we can approximately drown out the effect of v_t on the update to θ_t. This works for any value of β_2.

ii. By setting $\beta_2 = 1$ to have $v_t = 0$ for all t and then setting $\epsilon_A = 1$ will completely remove the effect of v_t on updates to θ_t. While this approach works for the version of Adam presented in this assignment, the actual Adam algorithm has an extra bias correction step "$\hat{v}_t \leftarrow v_t/(1-\beta_2^t)$" which would result in a division by zero if $\beta_2 = 1$.

Either approach for reducing the effect of v_t will result in full marks. **[1 mark]**

Then notice that m_t in Adam and p_t in SGD with momentum both accumulate the gradient using a running average but with different signs, so if we set $\beta_1 = \mu$, then $m_t = -p_t$. This works out since the update rules for θ_t are also different by a sign. **[1 mark; 0 if no explanation is made about the different signs; partial marks for a hand-wavy explanation.]**

Note that the above statement (if $\beta_1 = \mu$ then $m_t = -p_t$) can be verified by induction. Assume $\beta_1 = \mu$, then the base case is $m_0 = p_0 = 0$. For the induction step we assume $m_{t-1} = -p_{t-1}$, then

$$-m_t = -\mu m_{t-1} - (1-\mu)g_t = \mu p_{t-1} - (1-\mu)g_t = p_t.$$

The final set of hyperparameters are either

$$(\alpha_A = C \cdot \alpha_S, \beta_1 = \mu, \beta_2 \in [0,1), \epsilon_A = C)$$

for some very large constant C, or

$$(\alpha_A = \alpha_S, \beta_1 = \mu, \beta_2 = 1, \epsilon_A = 1).$$

(c) We can prove the statement $\tilde{m}_t = C \cdot m_t$, $\tilde{v}_t = C^2 \cdot v_t$, and $\tilde{\theta}_t = \theta_t$ for all $t \geq 0$.

Base case is satisfied at initialization, with $\tilde{m}_0 = m_0 = 0$, $\tilde{v}_0 = v_0 = 0$, and we start from the same initialization $\tilde{\theta}_0 = \theta_0$. The induction step assumes $C \cdot \tilde{m}_{t-1} = m_{t-1}$, $C \cdot \tilde{v}_{t-1} = v_{t-1}$, and $\tilde{\theta}_{t-1} = \theta_{t-1}$. Then

$$\begin{aligned}
\tilde{g}_t &= C \cdot \nabla J(\theta_0) = C \cdot g_t \\
\tilde{m}_t &= \tilde{m}_{t-1} + (1-\beta_1)\tilde{g}_1 = C \cdot m_{t-1} + (1-\beta_1)C \cdot g_t = m_t \\
\tilde{v}_t &= \tilde{v}_{t-1} + (1-\beta_2)\tilde{g}_1^2 = C^2 \cdot v_{t-1} + (1-\beta_2)C^2 \cdot g_1^2 = C^2 \cdot v_t \\
\tilde{\theta}_t &= \tilde{\theta}_{t-1} - \alpha_A \tilde{m}_t/(\sqrt{\tilde{v}_t}) = \theta_{t-1} - \alpha_A C m_t/(C\sqrt{v_t}) = \theta_t
\end{aligned}$$

We have proven $\tilde{\theta}_t = \theta_t$ for all $t = 0, 1, 2, \ldots$ as part of this statement.

Marking rubric for part (c):

- -0.5 if m_t and v_t are not part of the induction statement.
- -0.5 if not a formal proof.
- -0.5 for each other mistake.