TTIC 31230 Fundamentals of Deep Learning
Quiz 2

Problem 1: Running Averages. Consider a sequence of vectors x_0, x_1, ... and two running averages y_t and z_t defined by as follows for $0 < \beta < 1$ and $\gamma > 0$.

$$\begin{aligned} y_0 &= 0 \\ y_{t+1} &= \beta y_t + (1-\beta) x_t \end{aligned}$$

$$\begin{aligned} z_0 &= 0 \\ z_{t+1} &= \beta z_t + \gamma x_t \end{aligned}$$

(a) Suppose that the values x_t are drawn IID from a distribution with mean vector $\overline{x} = E\, x_t$. Give values for

$$\overline{y} = \lim_{t \to \infty} E\, y_t$$

and

$$\overline{z} = \lim_{t \to \infty} E\, z_t$$

as functions of β, γ and $\overline{x}$
Hint: Solve for $E\, y_{t+1}$ as a function of $E\, y_t$ and assume that a limiting expectation exists.
(b) Express z_t as a function of y_t, β and γ.

Problem 2. Adaptive SGD. This problem considers the question of whether the convergence theorem hold for adaptive methods — in the limit as the learning rate goes to zero do adaptive methods converge to a local minimum of the loss.
Consider a generalization of RMSProp where we allow an arbitrary adaptation with with different learning rates for ifferent parameter values. More specifically consider the SGD update equation

$$(1) \quad \Phi_{t+1} = \Phi_t - \eta \left(A(\Phi_t, x_t, y_t) \odot \nabla_\Phi \mathcal{L}(\Phi_t, x_t, y_t) \right)$$

where $\langle x_t, y_t \rangle$ is the tth training pair, $A(\Phi_t, x_t, y_t)$ is an adaptation vector, and $\odot$ is the Haddamard product $(x \odot y)[i] = x[i]\, y[i]$.
Consider the special case given by

$$\begin{aligned} A(\Phi, x, y)[i] &= \frac{1}{\sqrt{s(\Phi, x, y)} + \epsilon} \\ s(\Phi, x, y) &= \frac{1}{d} ||\nabla_\Phi\, \mathcal{L}(\Phi, x, y)||^2 \end{aligned}$$

where d is the dimension of Φ.

(a) For the given interpretation of $A(\Phi, x, y)$, let Φ^* be a parameter setting that is a stationary point of the update equation (1) in the sense that expected update over a random draw from the population is zero. Write this stationary condition on Φ^* explicitly as an expectation equaling zero under the given interpretation of $A(\Phi, x, y)$.

(b) Is Φ^* as defined in part (b) a stationary point of the original loss — a point where the expected gradient of $\mathcal{L}(\Phi^*, x, y)$ equal to zero?

(c) Do these observations have implications for the adaptive methods described in this class. Explain your answer.