Part2

a) let $V_t$ of Adam equal to $V_t$ of RMSprop

that is

$$\beta_2 V_{t-1} + (1-\beta_2) g_t^2 = \gamma V_{t-1} + (1-\gamma) g_t^2$$

so $\beta_2 = \gamma$

let $\theta_t$ of Adam equal to $\theta_t$ of RMSprop

then that is

$$\theta_{t-1} - \alpha_A m_t / (\sqrt{v_t} + \epsilon_A) = \theta_{t-1} - \alpha_R g_t / (\sqrt{v_t} + \epsilon_R)$$

so $m_t = g_t$, $\epsilon_A = \epsilon_R$, $\alpha_A = \alpha_R$,

By Adam, we know

$$m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$$

to make $m_t = g_t$, we need

$$\beta_1 = 0,$$

the we get hyperparameters $(\alpha_R, 0, \gamma, \epsilon_R)$

that matches $(\alpha_A, \beta_1, \beta_2, \epsilon_A)$

B) We want to make Adam approximatly equivalent to momentum SGD, we need to find a set of $(\alpha_A, \beta_1, \beta_2, \epsilon_A)$ such that

$$\cancel{\theta_{t-1}} - \alpha_A m_t / (\sqrt{v_t} + \epsilon_A) \approx \cancel{\theta_{t-1}} + \alpha_s p_t$$

then let $\alpha_A = \alpha_s$, we get

$$m_t / (\sqrt{v_t} + \epsilon_A) \approx -p_t .$$

According to the algorithms; we can get:

$$(\beta_1 m_{t-1} + (1-\beta_1) g_t / \sqrt{\beta_2 v_{t-1} + (1-\beta_2) g_t^2} + \epsilon_A \approx -(\mu p_{t-1} - (1-\mu) \nabla(\theta_{t-1}))$$

let $\beta_2 = 1$, $\epsilon_A = 1$ we can get

$$\frac{\beta_1 m_{t-1} + (1-\beta_1) g_t}{\sqrt{v_{t-1}} + 1} \quad \text{where } v_{t-1} = v_t .$$

By Adam, $v_0 = 0$, $v_t$ will be $0$ on all iterations.

then $\sqrt{v_{t-1}} + 1 = 1$

then we got

$$\underbrace{\beta_1 m_{t-1} + (1-\beta_1) g_t}_{= m_t} \approx \underbrace{-\mu p_{t-1} + (1-\mu) g_t}_{= -p_t}$$

let $\beta_1 = \mu$, if $\mu$ is small enough, then they will be closer

Then we can find hyperparameters

$(\alpha_s, \mu, 1, 1)$ where $\mu \to 0$.

c) According to the question, we can denote the quantities as $\tilde{g}_t$, $\tilde{m}_t$, $\tilde{v}_t$, $\tilde{\theta}_t$

WTS for $\epsilon_A = 0$, Adam is invariant to rescaling. that is $\tilde{\theta}_t = \theta_t$ for $\forall t \in N$, by hint we can use induction.

Base case:

let $m_0 = \tilde{m}_0 = v_0 = \tilde{v}_0 = 0$, $\tilde{\theta}_0 = \theta_0$

WTS $\tilde{\theta}_1 = \theta_1$

We know $\tilde{g}_1 = c \nabla J(\tilde{\theta}_0)$

then $= c \nabla J(\theta_0)$ since $\tilde{\theta}_0 = \theta_0$.

$= c g_1$

$\tilde{m}_1 = \beta_1 \tilde{m}_0 + (1-\beta_1)\tilde{g}_1$ ($\tilde{m}_0 = 0$)

$= (1-\beta_1) c g_1$

$m_1 = \beta_1 m_0 + (1-\beta_1) g_1$

$= (1-\beta_1) g_1$

then $\tilde{m}_1 = c m_1$

$\tilde{v}_1 = \beta_2 \tilde{v}_0 + (1-\beta_2)\tilde{g}_1^2$

$= (1-\beta_2) c^2 g_1^2$ ($\tilde{v}_0 = 0$)

$v_1 = \beta_2 v_0 + (1-\beta_2) g_1^2$

$= (1-\beta_2) g_1^2$

then $\tilde{v}_1 = c^2 v_1$

$\tilde{\theta}_1 = \theta_0 - \alpha_A \tilde{m}_1 / (\sqrt{\tilde{v}_1} + \epsilon_A)$ ($\epsilon_A = 0$)

$= \theta_0 - \alpha_A c m_1 / \sqrt{c^2 v_1}$ (By above)

$= \theta_0 - \alpha_A \cdot m_1 / \sqrt{v_1}$

$= \theta_0 - \alpha_A m_1 / (\sqrt{v_1} + \epsilon_A) = \theta_1$

I.S.

I.H $\tilde{\theta}_{t-1} = \theta_{t-1}$, $\tilde{m}_{t-1} = C m_{t-1}$, $\tilde{V}_{t-1} = C^2 V_{t-1}$

WTS. $\tilde{\theta}_t = \theta_t$

we get $\tilde{g}_t = C \nabla J(\tilde{\theta}_{t-1})$ $(\tilde{\theta}_{t-1} = \theta_{t-1})$

$= C \nabla J(\theta_{t-1})$

$= C g_t$

$\tilde{m}_t = \beta_1 \tilde{m}_{t-1} + (1-\beta_1)\tilde{g}_t$

$= \beta C \cdot m_{t-1} + (1-\beta_1) \cdot C \cdot g_t$ (By I.H)

$= C(\beta_1 m_{t-1} + (1-\beta_1) g_t)$

$= C m_t$

$\tilde{V}_t = \beta_2 \tilde{V}_{t-1} + (1-\beta_2)\tilde{g}_t^2$

$= \beta_2 C^2 V_{t-1} + (1-\beta_2) C^2 \cdot g_t^2$ (By I.H)

$= C^2 (\beta_2 V_{t-1} + (1-\beta_2) g^2_t)$

$= C^2 V_t$

then $\tilde{\theta}_t = \theta_{t-1} - \alpha_A \tilde{m}_t / (C\sqrt{\tilde{v}_t} + \epsilon_A)$ $(\epsilon_A = 0)$

$= \theta_{t-1} - \alpha_A \cancel{C} m_t / \sqrt{\cancel{C} V_t}$

$= \theta_{t-1} - \alpha_A m_t / (\sqrt{V_t} + \epsilon_A)$

$= \theta_t$

then we get Adam is invariant to this rescaling ∎