Update rule for Adam

$$m_t = \beta_1 * m_{t-1} + (1 - \beta_1) * \nabla w_t$$

$$v_t = \beta_2 * v_{t-1} + (1 - \beta_2) * (\nabla w_t)^2$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} * \hat{m}_t$$

- For convenience we will denote ∇w_t as g_t and β_1 as β

$$
\begin{aligned}
m_t &= \beta * m_{t-1} + (1-\beta) * g_t \\
m_0 &= 0 \\
m_1 &= \beta m_0 + (1-\beta) g_1 \\
&= (1-\beta) g_1 \\
m_2 &= \beta m_1 + (1-\beta) g_2 \\
&= \beta(1-\beta) g_1 + (1-\beta) g_2 \\
m_3 &= \beta m_2 + (1-\beta) g_3 \\
&= \beta(\beta(1-\beta) g_1 + (1-\beta) g_2) + (1-\beta) g_3 \\
&= \beta^2(1-\beta) g_1 + \beta(1-\beta) g_2 + (1-\beta) g_3 \\
&= (1-\beta) \sum_{i=1}^{3} \beta^{3-i} g_i
\end{aligned}
$$

- In general,

$$
m_t = (1-\beta) \sum_{i=1}^{t} \beta^{t-i} g_i
$$

$$E[m_t] = E[(1-\beta)\sum_{i=1}^{t}\beta^{t-i}g_i]$$

$$E[m_t] = (1-\beta)E[\sum_{i=1}^{t}\beta^{t-i}g_i]$$

$$E[m_t] = (1-\beta)\sum_{i=1}^{t}E[\beta^{t-i}g_i]$$

$$= (1-\beta)\sum_{i=1}^{t}\beta^{t-i}E[g_i]$$

- Assumption: All g_i's come from the same distribution i.e. $E[g_i] = E[g]\ \forall i$

$$
\begin{aligned}
E[m_t] &= (1-\beta)\sum_{i=1}^{t}(\beta)^{t-i}E[g_i] \\
&= E[g](1-\beta)\sum_{i=1}^{t}(\beta)^{t-i} \\
&= E[g](1-\beta)(\beta^{t-1}+\beta^{t-2}+\cdots+\beta^{0}) \\
&= E[g](1-\beta)\frac{1-\beta^t}{1-\beta}
\end{aligned}
$$

the last fraction is the sum of a GP with common ratio $= \beta$

$$
\begin{aligned}
E[m_t] &= E[g](1-\beta^t) \\
E[\frac{m_t}{1-\beta^t}] &= E[g] \\
E[\hat{m}_t] &= E[g](\because \frac{m_t}{1-\beta^t} = \hat{m}_t)
\end{aligned}
$$

Hence we apply the bias correction because then the expected value of $\hat{m}_t$ is the same as the expected value of g_t