Task 1

a

We were not successful in proving the unweighted importance sampling is unbiased. Although we show our approaches below:

$$\begin{aligned}
v_\pi &= \mathbb{E}_\pi[g_k|s_k] \\
&= \mathbb{E}_{\mathbb{b}}\left[\frac{\sum_{k\in\mathcal{T}(s_k)}\rho_{k:T(k)}g_k}{|\mathcal{T}\left(s_k\right)|}|a_k\right] \\
&\text{we assume sample mean from MC and } \rho_{k:T(k)} \text{ is the factor that project from the b distrebution to the} \pi \text{distrebution} \\
&= \mathbb{E}_{\mathbb{b}}\left[\mathbb{E}_{\mathbb{g}}\left[g_k\cdot\rho_{k:T(k)}|a_k\right]\right] \\
&= \mathbb{E}_{\mathbb{b}}\left[g_k\cdot\frac{\sum_{i=k}^{T}\pi\left(a_k \mid s_k\right)}{|T(k)|}\right]
\end{aligned}$$

c

Weighted importance sampling is biased when only a small number of trajectories is sampled. In this case the state value can be dominated by a small number of samples. A example is the following scenario:

Take 3 states: s_1 the start state, s_2 a terminal state to the right and s_3 a state to the left of s_1 which is also a terminal state. The actions for s_1 are left (a_{left}) with reward -10 and right (a_{right}) with reward 10.

For the behavior policy $b(a_k|s_k)$ the probability of going left in s_1 is 95% and right is 5%. For $\pi(a_k|s_k)$ it is the opposite.

Given one sampled trajectory which denotes as follows: $s_1 \rightarrow a_{left} \rightarrow s_3$
The state value would be estimated as follows:

$$\begin{aligned}
V(s) &\doteq \frac{\sum_{t\in\mathcal{T}(s)}\rho_{t:T(t)-1}G_t}{\sum_{t\in\mathcal{T}(s)}\rho_{t:T(t)-1}} \\
&= \frac{\sum_{t\in\mathcal{T}(s)}\frac{\prod_{i=k}^{T}\pi(a_k|s_k)}{\prod_{i=k}^{T}b(a_k|s_k)}G_t}{\sum_{t\in\mathcal{T}(s)}\frac{\prod_{i=k}^{T}\pi(a_k|s_k)}{\prod_{i=k}^{T}b(a_k|s_k)}} \\
&= G_t
\end{aligned}$$

The fractions for one trajectory can be canceled out thus only G_t remains. In this case G_t is the unbiased estimation of the state value according to the behavior policy. Which we also use to estimate the state value of our policy π which is the bias.
In our example the estimated state value of the s_1 would be $-10 = \hat{v}_\pi(s_1)$ This is far from the true state value of $v_\pi(s_1)$ is: $0,95\cdot 10 + 0,05\cdot -10 = 9 = v_\pi(s_1)$ If we sample more trajectories the bias converges asymptotically to zero.