

Appendix

When the evaluation policy π_e is known, the DW estimator can be written as:

$$V_{DW}^{\pi_e} = \frac{1}{n} \sum_{k=1}^n \sum_{t=1}^T \gamma^t (\omega_{0:t}^{(k)} r_t^{(k)} + \omega_{0:t}^{(k)} (\hat{Q}(s_t^{(k)}, a_t^{(k)}) - \hat{Q}_{means}(s_t^{(k)}, a_t^{(k)}))), \quad (1)$$

where $\omega_{0:t}^{(k)} = \prod_0^t \frac{\pi_e(s_t^{(k)}, a_t^{(k)})}{\pi_b(s_t^{(k)}, a_t^{(k)})}$ and π_b is an uncertain policy. Then, we can compute $E_{P_\xi^{\pi_b}}[V_{DW}^{\pi_e}]$ as follows:

$$\begin{aligned} &= E_{P_\xi^{\pi_b}} \left[\sum_{t=1}^{T-1} \gamma^t (\omega_{0:t} r_t + \omega_{0:t} \hat{Q}^{\pi_e}(s_t, a_t) - \omega_{0:t} \frac{1}{t+1} \sum_{m=0}^t \hat{Q}^{\pi_e}(s_m, a_m)) \right] \\ &\stackrel{(a)}{=} E_{P_\xi^{\pi_b}} \left[\sum_{t=1}^{T-1} \gamma^t \left(\omega_{0:t} r_t + \omega_{0:t} \hat{Q}^{\pi_e}(s_0, a_0) - \right. \right. \\ &\quad \left. \left. \omega_{0:t} \frac{1}{t+1} \frac{(\gamma^t + \gamma^{t-1} + \dots + \gamma^0) \hat{Q}^{\pi_e}(s_0, a_0) - (t+1)V(s_t)}{\gamma^t} \right) \right] \\ &= E_{P_\xi^{\pi_b}} \left[\sum_{t=1}^{T-1} \gamma^t \left(\omega_{0:t} r_t + \omega_{0:t} \left(\frac{\hat{Q}^{\pi_e}(s_0, a_0) - V(s_0)}{\gamma^t} - \right. \right. \right. \\ &\quad \left. \left. \omega_{0:t} (\hat{Q}^{\pi_e}(s_0, a_0) \left(\frac{\gamma^t + \gamma^{t-1} + \dots + \gamma^0}{\gamma^t(t+1)} \right) - \frac{V(s_t)}{\gamma^t} \right) \right) \right] \\ &= E_{P_\xi^{\pi_b}} \left[\sum_{t=1}^{T-1} \gamma^t \omega_{0:t} r_t + \sum_{t=1}^{T-1} \left(\left(1 - \frac{\gamma^t + \gamma^{t-1} + \dots + \gamma^0}{t+1} \right) \right. \right. \\ &\quad \left. \left. \omega_{0:t} \hat{Q}^{\pi_e}(s_0, a_0) \right) + \sum_{t=1}^{T-1} \omega_{0:t} (V(s_t) - V(s_0)) \right] \\ &\stackrel{(b)}{=} E_{P_\xi^{\pi_b}} \left[\sum_{t=1}^{T-1} \gamma^t \omega_{0:t} r_t + \sum_{t=1}^{T-1} \left(\left(1 - \frac{\gamma^t + \gamma^{t-1} + \dots + \gamma^0}{t+1} \right) \right. \right. \\ &\quad \left. \left. \omega_{0:t} (Q^{\pi_e}(s_0, a_0) + \Delta^{\pi_e}(s_0, a_0)) \right) + \sum_{t=1}^{T-1} \omega_{0:t} V_t^{back} \right] \\ &\stackrel{(c)}{=} v_{T-1}^{\pi_e} + E_{P_\xi^{\pi_b}} \left[\sum_{t=1}^{T-1} \omega_{0:t} \underbrace{\phi^t}_{\approx 0} (Q^{\pi_e}(s_0, a_0) + \Delta^{\pi_e}(s_0, a_0)) + \right. \\ &\quad \left. \sum_{t=1}^{T-1} \omega_{0:t} V_t^{back} \right] \\ &= v_{T-1}^{\pi_e} + E_{P_\xi^{\pi_b}} \left[\sum_{t=1}^{T-1} \omega_{0:t} V_t^{back} \right], \end{aligned}$$

where $V_t^{back} = V(s_t) - V(s_0)$, $\phi^t = 1 - \frac{\gamma^t + \gamma^{t-1} + \dots + \gamma^0}{t+1}$ and $v_{T-1}^{\pi_e} = E_\xi^{\pi_b} [\sum_{t=1}^{T-1} \omega_{0:t} \gamma^t r_t]$. Therefore, the bias of the DW estimator can be written as: $Bias(V_{DW}^{\pi_e}) = E_{P_\xi^{\pi_b}}(V_{DW}^{\pi_e}) - v_{T-1}^{\pi_e} = E_{P_\xi^{\pi_b}} [\sum_{t=1}^{T-1} \omega_{0:t} V_t^{back}]$.

(a) This derivation is based on $\hat{Q}^{\pi_e}(s_t, a_t) = \gamma \hat{Q}^{\pi_e}(s_{t+1}, a_{t+1}) + r_t$, which is transformed into $\hat{Q}^{\pi_e}(s_{t+1}, a_{t+1}) = \frac{\hat{Q}^{\pi_e}(s_t, a_t) - r_t}{\gamma}$, and then $\hat{Q}^{\pi_e}(s_t, a_t) = \frac{\hat{Q}(s_0, a_0) - (r_t + \gamma r_{t-1} + \dots + \gamma^t r_0)}{\gamma^t}$.

(b) We denote $\Delta^{\pi_e}(s_t, a_t) = \hat{Q}^{\pi_e}(s_t, a_t) - Q^{\pi_e}(s_t, a_t)$.

(c) This derivation is based (i) $v_{T-1}^{\pi_e} = E_{s_1 \sim P_1}[V^{\pi_e}(s_1)]$, and

(ii) $\phi^t = 1 - \frac{\gamma^t + \gamma^{t-1} + \dots + \gamma^0}{t+1} \approx 0$ under the condition of $\gamma \approx 1$. Next, we give a proof of the variance $nVarP_{\xi}^{\pi_b}(\rho_{DW}^{\pi_e})$ of the DW estimator. For convenience, we omit the notation time t on the variance. Then, $nVarP_{\xi}^{\pi_b}(\rho_{DW}^{\pi_e})$ can be computed as follows:

$$\begin{aligned}
&= E_{P_{\xi}^{\pi_b}} \left[\frac{\pi_e(a|s)}{\pi_b(a|s)} \left(Q(s, a) + \Delta(s, a) - Q_{means}(s, a) - \right. \right. \\
&\quad \left. \left. \Delta_{means}(s, a) + r \right)^2 \right] - E_{P_{\xi}^{\pi_b}} \left[\frac{\pi_e(a|s)}{\pi_b(a|s)} (\Delta(s, a) + r + Q(s, a))^2 \right] \\
&= E_{P_{\xi}^{\pi_b}} \left[\left(\frac{\pi_e(a|s)}{\pi_b(a|s)} (Q(s, a) + r) \right)^2 \right] + E_{P_{\xi}^{\pi_b}} \left[\left(\frac{\pi_e(a|s)}{\pi_b(a|s)} (\Delta(s, a)) \right)^2 \right] \\
&\quad + E_{P_{\xi}^{\pi_b}} \left[\left(\frac{\pi_e(a|s)}{\pi_b(a|s)} (Q_{means}(s, a) + \Delta(s, a)) \right)^2 \right] + \\
&\quad 2E_{P_{\xi}^{\pi_b}} \left[\left(\frac{\pi_e(a|s)}{\pi_b(a|s)} \right)^2 (Q(s, a) + r) \Delta(s, a) \right] \\
&\quad - 2E_{P_{\xi}^{\pi_b}} \left[\left(\frac{\pi_e(a|s)}{\pi_b(a|s)} \right)^2 (Q_{means}(s, a) + \Delta_{means}(s, a)) \Delta(s, a) \right] \\
&\quad - 2E_{P_{\xi}^{\pi_b}} \left[\left(\frac{\pi_e(a|s)}{\pi_b(a|s)} \right)^2 (Q_{means}(s, a) + \Delta_{means}(s, a)) (Q(s, a) + r) \right] \\
&\quad - E_{P_{\xi}^{\pi_b}} \left[\left(\Delta(s, a) \frac{\pi_e(a|s)}{\pi_b(a|s)} \right)^2 \right] - E_{P_{\xi}^{\pi_b}} \left[(Q(s, a) + r) \left(\frac{\pi_e(a|s)}{\pi_b(a|s)} \right)^2 \right] \\
&\quad - 2E_{P_{\xi}^{\pi_b}} \left[(r + Q(s, a)) \Delta(s, a) \left(\frac{\pi_e(a|s)}{\pi_b(a|s)} \right)^2 \right]. \\
&= E_{P_{\xi}^{\pi_b}} \left[\left(\frac{\pi_e(a|s)}{\pi_b(a|s)} \right)^2 (Q_{means}(s, a) + \Delta(s, a))^2 \right] - 2E_{P_{\xi}^{\pi_b}} \left[\left(\frac{\pi_e(a|s)}{\pi_b(a|s)} \right)^2 \right. \\
&\quad \left. (Q_{means}(s, a) + \Delta_{means}(s, a)) \Delta(s, a) \right] - 2E_{P_{\xi}^{\pi_b}} \left[\left(\frac{\pi_e(a|s)}{\pi_b(a|s)} \right)^2 \right. \\
&\quad \left. (Q_{means}(s, a) + \Delta_{means}(s, a)) * (Q(s, a) + r) \right] \\
&\stackrel{(a)}{=} E_{P_{\xi}^{\pi_b}} \left[\left(\frac{\pi_e(a|s)}{\pi_b(a|s)} \right)^2 (\Delta_1(s, a)^2 + 2\Delta_2(s, a)(Q(s, a) + \Delta(s, a) + r) \right. \\
&\quad + 2\Delta_1(s, a)r - Q(s, a)^2 - 2Q(s, a)\Delta(s, a) - 2Q(s, a)r \\
&\quad \left. - 2\Delta(s, a)r - \Delta(s, a)^2) \right] \\
&\stackrel{(b)}{=} E_{P_{\xi}^{\pi_b}} \left[\left(\frac{\pi_e(a|s)}{\pi_b(a|s)} \right)^2 (\Delta_1(s, a)^2 + 2\Delta_2(s, a)(Q(s, a) + r) \right. \\
&\quad \left. + 2\Delta_1(s, a)r - Q(s, a)^2 - 2Q(s, a)r) \right] \\
&= E_{\xi}^{\pi_b} \left[\omega_{0:t}^2 (2(\Delta_2(s, a) - Q(s, a))(Q(s, a) + r) \right. \\
&\quad \left. + (\Delta_1(s, a) + r)^2 - r^2 - Q(s, a)^2) \right]
\end{aligned}$$

where $\omega_{0:t} = \prod_0^t \frac{\pi_e(s, a)}{\pi_b(s, a)}$. By adding the notation time t , this formula is equivalent to the result in Conclusion 1 in the main file.

(a) We denote $\Delta_1(s, a) = Q(s, a) - Q_{means}(s, a)$ and $\Delta_2(s, a) = \Delta(s, a) - \Delta_{means}(s, a)$.

(b) When the algorithm converges, we can conclude that $\hat{Q}(s_t, a_t)$ is an unbiased estimate of $Q(s_t, a_t)$, and thus $\Delta(s_t, a_t) = \hat{Q}(s_t, a_t) - Q(s_t, a_t) \approx 0$.