Theorem 9 (Performance Guarantees with Pure Offline Queries). *Suppose (1) $Q^\star \in \mathcal{Q}, \pi^\star \in \Pi$, and (2) $\mathbb{T}^\pi q \in \mathcal{Q}, \forall \pi \in \Pi, q \in \mathcal{Q}$. Also we suppose the difference of return functions has a finite Eluder dimension $d_{Elu}(\Delta R, \alpha)$ and the underlying distribution of the offline dataset admits a finite coverage coefficient $C^\dagger$. Let $\beta_k = c_1\sqrt{\log(K|\Delta\mathcal{R}|)/K}$ and $\epsilon = c_2\sqrt{\log(N|\Pi||\mathcal{Q}|)/N}$, where c_1, c_2 are universal constants. Then the expected suboptimality of $\bar{\pi}$ from Algorithm 2 with pure offline queries is upper bounded by*

$$\mathrm{SubOpt}(\bar{\pi}) \leq \mathcal{O}\left(\sqrt{\frac{C^\dagger \log(N|\mathcal{Q}||\Pi|)}{N(1-\gamma)^2}} + \sqrt{\frac{d_{Elu}(\Delta\mathcal{R}, 1/K)\log(K|\Delta\mathcal{R}|)}{K(1-\gamma)}} + \sqrt{\frac{C\log(N|\Delta R|)}{N(1-\gamma)}}\right), \tag{35}$$

where N is the size of the offline dataset, K is the number of queries and $C = \max_s \frac{d^{\pi^*}(s)}{\mu(s)}$, where μ is the distribution that generates the dataset $\mathcal{D}$.

Proof. The main difference between using pure offline queries and using online queries is that we have to use trajectories sampled from the dataset $\widehat{\tau}^{k,1}, \widehat{\tau}^{k,2}$ instead of online sampled trajectories $\tau^{k,1}, \tau^{k,2}$. This incurs an additional performance gap of $\mathcal{E}_{\text{gap}} = \sqrt{\frac{C\log(N|\Delta R|)}{N(1-\gamma)}}$, since we need to refine our query policies within the covered policy set

$$\Pi_{\text{covered}} = \left\{\pi \mid \max_s \frac{d^\pi(s)}{\mu(s)} \leq C\right\}.$$

The proof for $\mathcal{E}_{\text{gap}}$ is the same as standard offline guarantees, and are omitted for simplicity. Then similar to the proof of Theorem 6, the regret can be bounded as

$$\begin{aligned}
&\mathrm{Reg}(K)\\
&\quad\leq \sum_{k=1}^{K} \max_{R_1, R_2 \in \widehat{\mathcal{C}}_k(\mathcal{R})} \left((R_1(\widehat{\tau}^{k,1}) - R_1(\widehat{\tau}^{k,2})) - (R_2(\widehat{\tau}^{k,1}) - R_2(\widehat{\tau}^{k,2}))\right)\\
&\qquad + K\mathcal{E}_{\text{gap}} + 16\sqrt{\frac{K}{1-\gamma}\log\left(\frac{4}{\delta}\right)} + 5K\mathcal{E}_{\text{off}}.
\end{aligned} \tag{36}$$

Then following the proof of Theorem 6, we have

$$\mathrm{SubOpt}(\bar{\pi}) \leq c_0 \cdot \sqrt{\frac{C^\dagger \log(N|\mathcal{V}||\Pi|)}{N(1-\gamma)^2}} + c_1 \cdot \sqrt{\frac{d_{\text{Elu}}(\Delta\mathcal{R}, 1/K)\log(K|\Delta\mathcal{R}|)}{K(1-\gamma)}} + c_2 \cdot \sqrt{\frac{C\log(N|\Delta R|)}{N(1-\gamma)}}.$$

□