Offline Stochastic Shortest Path: Learning, Evaluation and Towards Optimality (Supplementary material)

Ming Yin[*1,2] **Wenjing Chen**[*3] **Mengdi Wang**[4] **Yu-Xiang Wang**[1]

[1]Department of Computer Science, UC Santa Barbara
[2]Department of Statistics and Applied Probability, UC Santa Barbara
[3]Department of Electrical and Computer Engineering, Texas A&M University
[4]Department of Electrical and Computer Engineering, Princeton University

1 GENERAL BELLMAN EQUATION FOR A FIXED POLICY

In this section, we prove Proposition **??**. In particular, the first part of the proposition for $V^\star$ has been covered in Bertsekas and Tsitsiklis [1991]. Therefore, we only consider the second part, which is a Bellman equation for fixed policy. Moreover, we do not constraint to proper policy and our result holds true for all the policies.

Lemma 1.1 (General Bellman equation for fixed policy π)**.** *Let π be a fixed policy, proper or improper and cost $c \geq 0$ for the SSP. Then the following Bellman equations hold:*

$$Q^\pi(s,a) = c(s,a) + P_{s,a}V^\pi, \quad V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^\pi(s,a)]. \tag{1}$$

Proof of Lemma 1.1. By definition of Q^π, we have

$$Q^\pi(s,a) = \lim_{T\to\infty} \mathbb{E}_\pi[\sum_{h=0}^{T} c(s_h,a_h)|s_0=s,a_0=a].$$

We can rewrite term $\mathbb{E}_\pi[\sum_{h=0}^{T} c(s_h,a_h)|s_0=s,a_0=a]$ as

$$\begin{aligned}\mathbb{E}_\pi[\sum_{h=0}^{T} c(s_h,a_h)|s_0=s,a_0=a] &= c(s,a) + \sum_{s'}\mathbb{P}(s'|s,a)\mathbb{E}_\pi[\sum_{h=1}^{T} c(s_h,a_h)|s_1=s'] \\ &= c(s,a) + \sum_{s'}\mathbb{P}(s'|s,a)\left\{\mathbb{E}_\pi[\sum_{h=0}^{T-1} c(s_h,a_h)|s_0=s']\right\},\end{aligned}$$

where the first equality is by law of total expectation. The second equality follows from the fact that the transition kernel P is **homogeneous** in SSP.

Define the sequence $V_T(s) := \left\{\mathbb{E}_\pi[\sum_{h=0}^{T-1} c(s_h,a_h)|s_0=s]\right\}$. Since for any state-action pair (s,a), $c(s,a)\geq 0$, we know that the sequence $\{V_T(s)\}_{T=1}^\infty$ is non-decreasing. It implies that $\lim_{T\to\infty} V_T(s)$ exists. ($\lim_{T\to\infty} V_T(s)$ either diverges to $+\infty$ or converges to a positive number.) It follows that (the following switching the order of limit and summation is valid since the summation is finite sum)

$$\lim_{T\to\infty}\sum_{s'}\mathbb{P}(s'|s,a)\left\{\mathbb{E}_\pi[\sum_{h=0}^{T-1} c(s_h,a_h)|s_0=s']\right\} = \sum_{s'}\mathbb{P}(s'|s,a)\lim_{T\to\infty}\left\{\mathbb{E}_\pi[\sum_{h=0}^{T-1} c(s_h,a_h)|s_0=s']\right\}. \tag{2}$$

[*]Equal contribution.

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

Combine the above two equalities together, we can get

$$
\begin{aligned}
Q^\pi(s,a) &= c(s,a) + \sum_{s'} \mathbb{P}(s'|s,a) \lim_{T\to\infty} \left\{ \mathbb{E}_\pi[\sum_{h=0}^{T-1} c(s_h,a_h)|s_0 = s'] \right\} \\
&= c(s,a) + \sum_{s'} \mathbb{P}(s'|s,a) V^\pi(s').
\end{aligned}
\tag{3}
$$

From the definition of value function, we have (where the second line uses law of total expectation)

$$
\begin{aligned}
V^\pi(s) &= \lim_{T\to\infty} \mathbb{E}_\pi[\sum_{h=0}^{T} c(s_h,a_h)|s_0 = s] \\
&= \lim_{T\to\infty} \mathbb{E}_{a_0}[\mathbb{E}_\pi[\sum_{h=0}^{T} c(s_h,a_h)|s_0 = s, a_0]|s_0 = s] \\
&= \lim_{T\to\infty} \sum_a \pi(a|s) \mathbb{E}_\pi[\sum_{h=0}^{T} c(s_h,a_h)|s_0 = s, a_0 = a]
\end{aligned}
$$

Similar to $\lim_{T\to\infty} \mathbb{E}_\pi[\sum_{h=0}^T c(s_h,a_h)|s_0 = s]$, we can prove that $\lim_{T\to\infty} \mathbb{E}_\pi[\sum_{h=0}^T c(s_h,a_h)|s_0 = s, a_0 = a]$ exists. Then we have

$$
\begin{aligned}
V^\pi(s) &= \lim_{T\to\infty} \sum_a \pi(a|s) \mathbb{E}_\pi[\sum_{h=0}^{T} c(s_h,a_h)|s_0 = s, a_0 = a] \\
&= \sum_a \pi(a|s) \lim_{T\to\infty} \mathbb{E}_\pi[\sum_{h=0}^{T} c(s_h,a_h)|s_0 = s, a_0 = a] = \sum_a \pi(a|s) Q^\pi(s,a).
\end{aligned}
$$

□

Remark 1.2. *Essentially, the above proof only requires $c(s,a) \geq 0$. Moreover, even if the general Bellman equation holds, it does not imply $c^\pi + P^\pi(\cdot)$ is a contraction (*i.e. *doing value iteration for general policy π might not converge to V^π).*

2 RESULTS FOR GENERAL STOCHASTIC SHORTEST PATH PROBLEM

Lemma 2.1. *For any two contraction mapping T_1 and T_2 that are monotone (i.e. for any vector greater than $V \geq V'$, it holds $T_1 V \geq T_1 V'$ and $T_2 V \geq T_2 V'$) on the metric space $\mathbb{R}^{\mathcal{S}'}$. Suppose V_1 and V_2 are the fixed points for T_1 and T_2 respectively. If we have $T_1(V)(s) \geq T_2(V)(s)$ for any $s \in \mathcal{S}'$, then we have $V_1(s) \geq V_2(s)$ for any $s \in \mathcal{S}'$.*

Proof. First we have $T_1 V_1 \geq T_2 V_1$. Since V_1 is the fixed point of T_1, we know $V_1 := T_1 V_1 \geq T_2 V_1$. By monotone property with recursion, we have that

$$
V_1 \geq (T_2)^k V_1. \tag{4}
$$

Since V_2 is the fixed point of T_2, we have

$$
\lim_{k\to\infty} (T_2)^k V_1 = V_2.
$$

Combine the above inequalities together we can get $V_1 \geq V_2$. □

3 CONVERGENCES FOR ALGORITHM ??

Lemma 3.1. *$\widehat{\mathcal{T}}^\pi : \mathbb{R}^{\mathcal{S}} \times \{0\} \to \mathbb{R}^{\mathcal{S}} \times \{0\}$ is a contraction mapping, i.e., $\forall V_1, V_2 \in \mathbb{R}^{\mathcal{S}'}$, $V_1(g) = V_2(g) = 0$, we have*

$$
\left\| \widehat{\mathcal{T}}^\pi V_1 - \widehat{\mathcal{T}}^\pi V_2 \right\|_\infty \leq \rho \left\| V_1 - V_2 \right\|_\infty, \tag{5}
$$

Here $\rho := \max_{\substack{s,a \\ s\neq g}} (\frac{n_{s,a}}{n_{s,a}+1}) < 1$ and $\widehat{\mathcal{T}}^\pi V(s) = \langle \pi(\cdot|s), \widehat{c}(s,\cdot) + \widetilde{P}_{s,\cdot} V \rangle$.

Proof of Lemma 3.1. We first prove the result for state g. Since g is a zero-cost absorbing state, we have for any $a \in \mathcal{A}$, $\widehat{c}(g,a) = 0$ and $\widetilde{P}_{g,a}V = V(g)$. Then for any $V \in \mathbb{R}^{\mathcal{S}} \times \{0\}$, $V(g) = 0$, we have

$$\widehat{\mathcal{T}}^{\pi} V(g) = \langle \pi(\cdot|g), \widehat{c}(g,\cdot) + \widetilde{P}_{g,\cdot} V \rangle = 0. \tag{6}$$

Therefore $\widehat{\mathcal{T}}^{\pi} V_1(g) - \widehat{\mathcal{T}}^{\pi} V_2(g) = 0 \le \rho \|V_1 - V_2\|_{\infty}$. Next we only need to prove for all state $\forall s \neq g$. Indeed,

$$\begin{aligned}
|\widehat{\mathcal{T}}^{\pi} V_1(s) - \widehat{\mathcal{T}}^{\pi} V_2(s)| &= |\langle \pi(\cdot|s), \widetilde{P}_{s,\cdot}(V_1 - V_2)\rangle| \\
&\le \max_a |\widetilde{P}_{s,a}(V_1 - V_2)| \\
&= \max_a |\sum_{s' \neq g} \widetilde{P}(s'|s,a)(V_1(s') - V_2(s'))| \\
&\le \max_a (\frac{n_{s,a}}{n_{s,a}+1}) |\sum_{s' \neq g} \widehat{P}(s'|s,a)(V_1(s') - V_2(s'))| \\
&\le \max_a (\frac{n_{s,a}}{n_{s,a}+1}) \|V_1 - V_2\|_{\infty}. \qquad (7)
\end{aligned}$$

where the second inequality is due to $V_1(g) = V_2(g) = 0$ and the third inequality is by the definition of $\widetilde{P}$. Take the supremum over s, we get

$$\left\|\widehat{\mathcal{T}}^{\pi} V_1 - \widehat{\mathcal{T}}^{\pi} V_2\right\|_{\infty} \le \max_{\substack{s,a \\ s \neq g}} (\frac{n_{s,a}}{n_{s,a}+1}) \|V_1 - V_2\|_{\infty} \tag{8}$$

□

Lemma 3.2. *$\forall \pi \in \Pi_{proper}$, define $\widehat{V}^{\pi} := \lim_{i \to \infty} V^{(i)}$ (Note by Lemma 3.1 this limit always exists since $\widehat{V}^{\pi}$ is the fixed point of $\widehat{\mathcal{T}}^{\pi}$ and $V^{(i+1)} = \widehat{\mathcal{T}}^{\pi} V^{(i)}$). Then (recall $\rho := \max_{\substack{s,a \\ s \neq g}} (\frac{n_{s,a}}{n_{s,a}+1}) < 1$)*

$$||\widehat{V}^{\pi}||_{\infty} \le \max_{\substack{s,a \\ s \neq g}} n(s,a) + 1.$$

Proof of Lemma 3.2. Recall the definition, $\widehat{V}^{\pi} = \widehat{\mathcal{T}}^{\pi} \widehat{V}^{\pi}$

$$\begin{aligned}
&\left\|\widehat{V}^{\pi}\right\|_{\infty} = \left\|\widehat{\mathcal{T}}^{\pi} \widehat{V}^{\pi}\right\|_{\infty} &(9)\\
&\le \max_{s, s \neq g} |\langle \pi(\cdot|s), \widehat{c}(s,\cdot)\rangle| + \max_{s, s \neq g} |\langle \pi(\cdot|s), \widetilde{P}_{s,\cdot} \widehat{V}^{\pi}\rangle| &(10)\\
&\le \max_{s, s \neq g} |\langle \pi(\cdot|s), \widehat{c}(s,\cdot)\rangle| + \max_{\substack{s,a \\ s \neq g}} |\widetilde{P}_{s,a} \widehat{V}^{\pi}| \\
&\le 1 + \max_{\substack{s,a \\ s \neq g}} |\widetilde{P}_{s,a} \widehat{V}^{\pi}| \\
&\le 1 + \max_{\substack{s,a \\ s \neq g}} \{(\frac{n_{s,a}}{n_{s,a}+1}) |\sum_{s' \neq g} \widehat{P}(s'|s,a) \widehat{V}^{\pi}(s')|\} \\
&\le 1 + \rho \left\|\widehat{V}^{\pi}\right\|_{\infty}. &(11)
\end{aligned}$$

□

The first inequality follows from $\widehat{\mathcal{T}}^{\pi} \widehat{V}^{\pi}(g) = 0$ and triangle inequality. Since $\rho < 1$, we can get $\left\|\widehat{V}^{\pi}\right\|_{\infty} \le \frac{1}{1-\rho}$. From the definition of ρ, we can conclude the proof.

Lemma 3.3. *$\left\|\widehat{V}^{\pi} - V^{(i)}\right\|_{\infty} \le \frac{\epsilon_{OPE}}{1-\rho}$, where $\rho := \max_{\substack{s,a \\ s \neq g}} (\frac{n_{s,a}}{n_{s,a}+1}) < 1$ as in Lemma 3.1 and $V^{(i)}$ is the output of Algorithm* **??**.

Proof of Lemma 3.3. Using definition $\widehat{V}^\pi := \lim_{j\to\infty} V^{(j)}$ and the telescoping sum we obtain

$$\left\|\widehat{V}^\pi - V^{(i)}\right\|_\infty \le \sum_{j=i}^{\infty}\left\|V^{(j+1)} - V^{(j)}\right\|_\infty \le \left\|V^{(i+1)} - V^{(i)}\right\|_\infty \sum_{j=0}^{\infty}\rho^j \le \frac{\epsilon_{\text{OPE}}}{1-\rho},$$

where the second inequality uses $\widehat{\mathcal{T}}^\pi$ is a ρ-contraction. ◻

Remark 3.4. *Throughout the paper, we denote the number of state-action visitation as either $n_{s,a}$ or $n(s,a)$. They represent the same quantity.*

Lemma 3.5. *For any $V(\cdot)\in\mathbb{R}^S$ satisfying $V(g)=0$,*

$$|(\widetilde{P}_{s,a} - \widehat{P}_{s,a})V| \le \frac{||V||_\infty}{n(s,a)+1}, \qquad |\mathrm{Var}(\widetilde{P}_{s,a},V) - \mathrm{Var}(\widehat{P}_{s,a},V)| \le \frac{2\left\|V\right\|_\infty^2 S}{n(s,a)+1}. \tag{12}$$

Proof. See Lemma 12 of Tarbouriech et al. [2021]. ◻

4 SOME KEY LEMMAS

4.1 HIGH PROBABILITY EVENT

We define the *"good property"* event $\mathcal{E} := \mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\cap\mathcal{E}_4\cap\mathcal{E}_5$ according the following (where $\iota := \log(10S^2A/\delta)$)

$$\begin{aligned}
\mathcal{E}_1 &:= \left\{\forall (s,a,s')\in(\mathcal{S}\times\mathcal{A}\times\mathcal{S}), \forall n(s,a)\ge 1 : |P(s'|s,a) - \widehat{P}(s'|s,a)| \le \sqrt{\frac{2P(s'|s,a)\iota}{n(s,a)}} + \frac{2\iota}{3n(s,a)}\right\}\\
\mathcal{E}_2 &:= \left\{\forall (s,a)\in(\mathcal{S}\times\mathcal{A}), \forall n(s,a)\ge 1 : |(P_{s,a} - \widehat{P}_{s,a})V| \le \sqrt{\frac{2\mathrm{Var}(P_{s,a},V)\iota}{n(s,a)}} + \frac{2||V||_\infty\iota}{3n(s,a)}\right\}\\
\mathcal{E}_3 &:= \left\{\forall (s,a)\in(\mathcal{S}\times\mathcal{A}), \forall n(s,a)\ge 1 : |(P_{s,a} - \widehat{P}_{s,a})V| \le \sqrt{\frac{2\mathrm{Var}(\widehat{P}_{s,a},V)\iota}{n(s,a)}} + \frac{7||V||_\infty\iota}{3n(s,a)}\right\}\\
\mathcal{E}_4 &:= \left\{\forall (s,a)\in(\mathcal{S}\times\mathcal{A}), \forall n(s,a)\ge 1 : |\widehat{c}(s,a) - c(s,a)| \le \sqrt{\frac{2\mathrm{Var}_c(s,a)\iota}{n(s,a)}} + \frac{2\iota}{3n(s,a)}\right\}\\
\mathcal{E}_5 &:= \left\{\forall (s,a)\in(\mathcal{S}\times\mathcal{A}), \forall n(s,a)\ge 1 : |\widehat{c}(s,a) - c(s,a)| \le \sqrt{\frac{2\widehat{c}(s,a)\iota}{n(s,a)}} + \frac{7\iota}{3n(s,a)}\right\}.
\end{aligned} \tag{13}$$

Lemma 4.1. *The event $\mathcal{E}$ holds for any V that is independent from $\widehat{P}$ with probability $1-\frac{\delta}{2}$.*

Proof. From the empirical Bernstein's inequality given in Lemma 12.4, we have that for each fixed (s,a), the event $|(P_{s,a} - \widehat{P}_{s,a})V| \le \sqrt{\frac{2\mathrm{Var}(\widehat{P}_{s,a},V)\iota}{n(s,a)}} + \frac{7||V||_\infty\iota}{3n(s,a)}$ holds with probability $1-\frac{\delta}{10S^2A}$. By taking a union bound, we have that event $\mathcal{E}_3$ holds with probability $1-\frac{\delta}{10S}$. Similarly, we have event $\mathcal{E}_5$ holds with probability $1-\frac{\delta}{10S}$. By applying the standard Bernstein's inequality in Lemma 12.3 and taking union bound over (s,a,s'), we can get that event $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_4$ holds with probability $1-\frac{\delta}{10}$. Since $\mathcal{E}$ is the intersection of the above events, we can prove the lemma by taking a union bound again over all of the five events. ◻

4.2 VALUE DECOMPOSITION LEMMA

Lemma 4.2. *Suppose $\widehat{V}^\pi := \lim_{j\to\infty} V^{(j)}$ where $V^{(j)} = \widehat{\mathcal{T}}^\pi V^{(j-1)}$ for all j, then we have the following suboptimality decomposition for any initial state $\bar{s}$:*

$$\widehat{V}^\pi(\bar{s}) - V^\pi(\bar{s}) = \sum_{h=0}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi^\pi_{h,\bar{s}}(s,a)\{(\widehat{c}-c)(s,a) + (\tilde{P}_{s,a} - P_{s,a})\widehat{V}^\pi\} \tag{14}$$

Proof of Lemma 4.2. We prove this lemma by recursion. First, we have for any $h \geq 0$,

$$
\begin{aligned}
\sum_{\substack{s \\ s\neq g}} \xi^{\pi}_{h,\bar{s}}(s)(\widehat{V}^{\pi}(s) - V^{\pi}(s)) &= \sum_{\substack{s \\ s\neq g}} \xi^{\pi}_{h,\bar{s}}(s) \sum_{a} \pi(a|s)(\widehat{Q}^{\pi}(s,a) - Q^{\pi}(s,a)) \\
&= \sum_{\substack{s,a \\ s\neq g}} \xi^{\pi}_{h,\bar{s}}(s,a)(\widehat{Q}^{\pi}(s,a) - Q^{\pi}(s,a)) \\
&= \sum_{\substack{s,a \\ s\neq g}} \xi^{\pi}_{h,\bar{s}}(s,a)\{(\widehat{c} - c)(s,a) + (\widetilde{P}_{s,a}\widehat{V}^{\pi} - P_{s,a}V^{\pi})\} \\
&= \sum_{\substack{s,a \\ s\neq g}} \xi^{\pi}_{h,\bar{s}}(s,a)\{(\widehat{c} - c)(s,a) + (\widetilde{P}_{s,a} - P_{s,a})\widehat{V}^{\pi} + P_{s,a}(\widehat{V}^{\pi} - V^{\pi})\} \\
&= \sum_{\substack{s,a \\ s\neq g}} \xi^{\pi}_{h,\bar{s}}(s,a)\{(\widehat{c} - c)(s,a) + (\widetilde{P}_{s,a} - P_{s,a})\widehat{V}^{\pi}\} + \sum_{\substack{s \\ s\neq g}} \xi^{\pi}_{h+1,\bar{s}}(s)(\widehat{V}^{\pi} - V^{\pi})(s),
\end{aligned}
$$

where the third equality uses both Bellman equations and empirical Bellman equations and the last equality follows from the fact that $\xi_{h+1}(s') = \sum_{s,a} \xi_h(s,a)P(s'|s,a)$ and $\widehat{V}^{\pi}(g) = V^{\pi}(g) = 0$. By recursion, we have that

$$
\widehat{V}^{\pi}(\bar{s}) - V^{\pi}(\bar{s}) = \sum_{h=0}^{H} \sum_{\substack{s,a \\ s\neq g}} \xi^{\pi}_{h,\bar{s}}(s,a)\{(\widehat{c} - c)(s,a) + (\tilde{P}_{s,a} - P_{s,a})\widehat{V}^{\pi}\} + \sum_{\substack{s \\ s\neq g}} \xi^{\pi}_{H+1,\bar{s}}(s)(\widehat{V}^{\pi}(s) - V^{\pi}(s)), \tag{15}
$$

for all H. Then we have

$$
|\sum_{\substack{s \\ s\neq g}} \xi^{\pi}_{H+1,\bar{s}}(s)(\widehat{V}^{\pi}(s) - V^{\pi}(s))| \leq \sum_{\substack{s \\ s\neq g}} \xi^{\pi}_{H+1,\bar{s}}(s) \cdot \left\|\widehat{V}^{\pi} - V^{\pi}\right\|_{\infty} \leq P^{\pi}_{\bar{s}}(s_{H+1} \neq g) \cdot \left\|\widehat{V}^{\pi} - V^{\pi}\right\|_{\infty}.
$$

Since π is proper, we have $||V^{\pi}||_{\infty} \leq \infty$ and by Lemma 4.3 $\lim_{H\to+\infty} P^{\pi}_{\bar{s}}(s_H \neq g) = 0$. From Lemma 3.2, we have $||\widehat{V}^{\pi}||_{\infty} \leq \infty$. It follows that

$$
\lim_{H\to+\infty} \sum_{\substack{s \\ s\neq g}} \xi^{\pi}_{H+1,\bar{s}}(s)(\widehat{V}^{\pi}(s) - V^{\pi}(s)) = 0 \tag{16}
$$

By taking H to infinity in Equation (15), we conclude the proof. □

4.3 KEY LEMMAS: ARRIVAL TIME DECOMPOSITION AND DEPENDENCE IMPROVEMENT FOR SSP

Below we present two lemmas for SSP problem, which is the key for obtaining tight instance-dependent bounds.

Lemma 4.3 (Arrival time decomposition)**.** *Let $T^{\pi}_{\bar{s}}$ be the expected time of arrival to goal state g when applying proper policy π and starting from $\bar{s}$, then $T^{\pi}_{\bar{s}} = \sum_{h=0}^{\infty} \sum_{\substack{s,a \\ s\neq g}} \xi^{\pi}_{h,\bar{s}}(s,a)$. Moreover, $T^{\pi}_{\bar{s}} < \infty$ for all $\bar{s}$.*

Proof of Lemma 4.3. Denote T to be the random variable of arrival time to goal state g when applying proper policy π, starting from $\bar{s}$. Then $\mathbb{E}[T] = T^{\pi}_{\bar{s}}$. Furthermore, since T is non-negative integral variable, it holds $\mathbb{E}[T] = \sum_{h=0}^{\infty} \mathbb{P}(T > h)$.

Then we have

$$
\begin{aligned}
T^{\pi}_{\bar{s}} &= \mathbb{E}_{P,\pi} T = \sum_{h=0}^{\infty} \mathbb{P}_{P,\pi}(T > h) \\
&= \sum_{h=0}^{\infty} \mathbb{P}_{P,\pi}(s_1 \neq g, s_2 \neq g, ..., s_h \neq g) \\
&\overset{\text{(i)}}{=} \sum_{h=0}^{\infty} \mathbb{P}_{P,\pi}(s_h \neq g) = \sum_{h=0}^{\infty} \sum_{\substack{s,a \\ s\neq g}} \xi^{\pi}_{h,\bar{s}}(s,a),
\end{aligned}
$$

where equality (i) follows from the fact that g is an absorbing state, so we can only reach a state which is not a goal state if all the previous steps are not goal state and vice versa.

Lastly, since π is proper, $T^{\pi}_{\bar{s}} < \infty$ for all $\bar{s}$ and this implies $\lim_{h\to\infty} \mathbb{P}_{P,\pi}(s_h \neq g) = 0$. □

The next lemma is the key for achieving optimal rate.

Lemma 4.4 (Dependency Improvement)**.** *For any probability transition matrix P, policy π, and any cost function $c \in [0,1]$, we use $\xi^{\pi}_h(s,a)$ to denote the probability of visiting (s,a) associated with $\widehat{SSP}(P,\pi)$. Suppose $V \in \mathbb{R}^{S+1}$ is any value function satisfying order property (where $V(g)=0$), i.e., $V(s) \geq \sum_a \pi(a|s) P_{s,a} V$ for all $s \in \mathcal{S}$, then we have*

$$\sum_{h=0}^{\infty} \sum_{\substack{s,a \\ s\neq g}} \xi^{\pi}_h(s,a) \mathrm{Var}(P_{s,a},V) \leq 2\left\|V\right\|_\infty \sum_{\substack{s \\ s\neq g}} \xi_0(s) V(s) \leq 2\left\|V\right\|^2_\infty . \tag{17}$$

Proof of Lemma 4.4.

$$\begin{aligned}
&\sum_{h=0}^{\infty} \sum_{\substack{s,a \\ s\neq g}} \xi^{\pi}_h(s,a) \mathrm{Var}(P_{s,a},V) = \sum_{h=0}^{\infty} \sum_{\substack{s,a \\ s\neq g}} \xi^{\pi}_h(s,a) \{P_{s,a}(V)^2 - (P_{s,a}V)^2\} \\
=&\sum_{h=0}^{\infty} \sum_{\substack{s \\ s\neq g}} \xi^{\pi}_{h+1}(s) V^2(s) - \sum_{h=0}^{\infty} \sum_{\substack{s,a \\ s\neq g}} \xi^{\pi}_h(s,a)(P_{s,a}V)^2 \\
\leq&\sum_{h=0}^{\infty} \sum_{\substack{s \\ s\neq g}} \xi^{\pi}_{h}(s) V^2(s) - \sum_{h=0}^{\infty} \sum_{\substack{s,a \\ s\neq g}} \xi^{\pi}_h(s,a)(P_{s,a}V)^2 \\
\overset{(i)}{=}&\sum_{h=0}^{\infty} \sum_{\substack{s \\ s\neq g}} \xi^{\pi}_{h}(s) \{V^2(s) - \sum_a \pi(a|s)(P_{s,a}V)^2\} \\
\overset{(ii)}{\leq}&\sum_{h=0}^{\infty} \sum_{\substack{s \\ s\neq g}} \xi^{\pi}_{h}(s) \{V^2(s) - (\sum_a \pi(a|s)P_{s,a}V)^2\} \\
=&\sum_{h=0}^{\infty} \sum_{\substack{s \\ s\neq g}} \xi^{\pi}_{h}(s) \{(V(s) - \sum_a \pi(a|s)P_{s,a}V)(V(s) + \sum_a \pi(a|s)P_{s,a}V)\} \\
\overset{(iii)}{\leq}&2\left\|V\right\|_\infty \sum_{h=0}^{\infty} \sum_{\substack{s \\ s\neq g}} \xi^{\pi}_{h}(s) (V(s) - \sum_a \pi(a|s)P_{s,a}V) \\
=&2\left\|V\right\|_\infty \left[\sum_{h=0}^{\infty} \sum_{\substack{s \\ s\neq g}} \xi^{\pi}_{h}(s) V(s) - \sum_{h=0}^{\infty} \sum_{\substack{s \\ s\neq g}} \xi^{\pi}_{h+1}(s) V(s)\right] \\
=&2\left\|V\right\|_\infty \sum_{\substack{s \\ s\neq g}} \xi_0(s)V(s) \leq 2\left\|V\right\|^2_\infty ,
\end{aligned}$$

where (i) follows from the fact that $\xi(s,a) = \xi(s)\pi(a|s)$, (ii) uses the Jensen's inequality and the fact that $f(x)=x^2$ is a convex function. (iii) uses the ordering condition. □

5 CRUDE EVALUATION BOUND

Theorem 5.1. *Denote $d_m := \min\{\sum_{h=1}^{\infty} \xi^{\mu}_h(s,a) : s.t. \sum_{h=1}^{\infty} \xi^{\mu}_h(s,a) > 0\}$, and T^{π}_s to be the expected time to hit g when starting from s. Define $\bar{T}^{\pi} = \max_{\bar{s}\in\mathcal{S}} T^{\pi}_{\bar{s}}$. Then when $n \geq \max\{\frac{49 S\iota}{9 d_m}, 64(\bar{T}^{\pi})^2 \frac{S\iota}{d_m}, C\cdot \log(SA/\delta)/\sum_{h=1}^{\infty} \xi^{\mu}_h(s,a)\}$, we* *have with probability* $1-\delta$, *(here* $\iota = O(\log(SA/\delta))$

$$\left\|\widehat{V}^\pi - V^\pi\right\|_\infty \le O\left(\frac{\bar{T}^\pi\sqrt{\max_{s,a}\mathrm{Var}_c(s,a)\iota} + \sqrt{\bar{T}^\pi\left\|V^\pi\right\|_\infty^2\iota}}{\sqrt{n\cdot d_m}}\right) + O\left(\frac{\bar{T}^\pi\left\|V^\pi\right\|_\infty\cdot\iota}{n\cdot d_m}\right).$$

Proof. We denote $\bar{s}\in\mathcal{S}$ to be any initial state. From Lemma 4.2, we have (here $\xi^\pi_{h,\bar{s}}(s,a)$ is the marginal state-action probability when starting from state $\bar{s}$ and following π)

$$\begin{aligned}
|V^\pi(\bar{s}) - \widehat{V}^\pi(\bar{s})| = &\left|\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi^\pi_{h,\bar{s}}(s,a)\{(\widehat{c}-c)(s,a) + (\tilde{P}_{s,a}-P_{s,a})\widehat{V}^\pi\}\right| \\
&\le \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi^\pi_{h,\bar{s}}(s,a)|(\widehat{c}-c)(s,a)| + \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi^\pi_{h,\bar{s}}(s,a)|(\tilde{P}_{s,a}-P_{s,a})(\widehat{V}^\pi - V^\pi)| \\
&\quad + \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi^\pi_{h,\bar{s}}(s,a)|(\tilde{P}_{s,a}-P_{s,a})V^\pi|
\end{aligned}$$

We bound the above three parts one by one. First of all, by Bernstein inequality, Lemma 12.6 and union bound, with probability $1-\delta$,

$$\begin{aligned}
&\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi^\pi_{h,\bar{s}}(s,a)|(\widehat{c}-c)(s,a)| \le \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi^\pi_{h,\bar{s}}(s,a)\left[2\sqrt{\frac{\mathrm{Var}_c(s,a)\iota}{n(s,a)}} + \frac{4\iota}{3n(s,a)}\right] \\
&\overset{(i)}{\le}\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi^\pi_{h,\bar{s}}(s,a)\left[2\sqrt{\frac{2\mathrm{Var}_c(s,a)\iota}{n\sum_{h=1}^\infty\xi^\mu_h(s,a)}} + \frac{8\iota}{3n\sum_{h=1}^\infty\xi^\mu_h(s,a)}\right] \\
&\le\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi^\pi_{h,\bar{s}}(s,a)\left[2\sqrt{\frac{2\max_{s,a}\mathrm{Var}_c(s,a)\iota}{n\cdot d_m}} + \frac{8\iota}{3n\cdot d_m}\right] \\
&\le T^\pi_{\bar{s}}\left[2\sqrt{\frac{2\max_{s,a}\mathrm{Var}_c(s,a)\iota}{n\cdot d_m}} + \frac{8\iota}{3n\cdot d_m}\right]
\end{aligned}$$

(i) uses Lemma 12.6 and the last inequality uses Lemma 4.3.

For the second part, note

$$\begin{aligned}
&\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi^\pi_{h,\bar{s}}(s,a)|(\tilde{P}_{s,a}-P_{s,a})(\widehat{V}^\pi - V^\pi)| \\
&\le \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi^\pi_{h,\bar{s}}(s,a)\left[\sqrt{\frac{2S\cdot\mathrm{Var}(P_{s,a},\widehat{V}^\pi - V^\pi)\iota}{n(s,a)}} + \frac{4\left\|\widehat{V}^\pi - V^\pi\right\|_\infty S\iota}{3n(s,a)} + \frac{\left\|\widehat{V}^\pi - V^\pi\right\|_\infty}{n(s,a)+1}\right] \\
&\le \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi^\pi_{h,\bar{s}}(s,a)\left[\sqrt{\frac{4S\cdot\mathrm{Var}(P_{s,a},\widehat{V}^\pi - V^\pi)\iota}{n\sum_{h=1}^\infty\xi^\pi_h(s,a)}} + \frac{14\left\|\widehat{V}^\pi - V^\pi\right\|_\infty S\iota}{3n\sum_{h=1}^\infty\xi^\pi_h(s,a)}\right] \\
&\le\sqrt{\sum_{\substack{s,a\\ s\neq g}}\frac{\sum_{h=0}^\infty\xi^\pi_{h,\bar{s}}(s,a)}{\sum_{h=0}^\infty\xi^\mu_h(s,a)}}\sqrt{4S\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_{h,\bar{s}}(s,a)\left\|\widehat{V}^\pi - V^\pi\right\|^2\frac{\iota}{n}} + \frac{14\left\|\widehat{V}^\pi - V^\pi\right\|_\infty S\iota}{3n}\sum_{s,a,s\neq g}\frac{d^\pi_{\bar{s}}(s,a)}{d^\mu(s,a)}
\end{aligned}$$

$$\leq \sqrt{\sum_{\substack{s,a\\ s\neq g}} \frac{d^{\pi}_{\bar{s}}(s,a)}{d^{\mu}(s,a)} \cdot 4ST^{\pi}_{\bar{s}} \left\| \widehat{V}^{\pi} - V^{\pi} \right\|_{\infty}^{2} \cdot \frac{\iota}{n}} + \frac{14 \left\| \widehat{V}^{\pi} - V^{\pi} \right\|_{\infty} S\iota}{3n} \sum_{s,a,s\neq g} \frac{d^{\pi}_{\bar{s}}(s,a)}{d^{\mu}(s,a)}$$

$$\leq 4\sqrt{\sum_{\substack{s,a\\ s\neq g}} \frac{d^{\pi}_{\bar{s}}(s,a)}{d^{\mu}(s,a)} \cdot ST^{\pi}_{\bar{s}} \left\| \widehat{V}^{\pi} - V^{\pi} \right\|_{\infty}^{2} \cdot \frac{\iota}{n}} \leq 4T^{\pi}_{\bar{s}} \sqrt{\frac{S\iota}{n \cdot d_m}} \left\| \widehat{V}^{\pi} - V^{\pi} \right\|_{\infty},$$

where the first inequality uses Lemma 10.3 and Lemma 3.5, the second inequality uses Lemma 12.6, the third inequality uses $\mathrm{Var}(\cdot) \leq \|\cdot\|_{\infty}^{2}$, the fourth and fifth inequality use Lemma 4.3 and the last inequality follows from the condition $n \geq \frac{49S\iota}{9d_m} \geq \frac{49S\iota}{9T^{\pi}_{\bar{s}}} \sum_{s,a,s\neq g} \frac{d^{\pi}_{\bar{s}}(s,a)}{d^{\mu}(s,a)}$.

For the third part, we have

$$\sum_{h=1}^{\infty} \sum_{\substack{s,a\\ s\neq g}} \xi^{\pi}_{h,\bar{s}}(s,a) |(\tilde{P}_{s,a} - P_{s,a})V^{\pi}|$$

$$\leq \sum_{h=1}^{\infty} \sum_{\substack{s,a\\ s\neq g}} \xi^{\pi}_{h,\bar{s}}(s,a) \left[2\sqrt{\frac{\mathrm{Var}(P_{s,a}, V^{\pi})\iota}{n(s,a)}} + \frac{4\left\|V^{\pi}\right\|_{\infty} \iota}{3n(s,a)} + \frac{\left\|V^{\pi}\right\|_{\infty}}{n(s,a)} \right]$$

$$\leq \sum_{h=1}^{\infty} \sum_{\substack{s,a\\ s\neq g}} \xi^{\pi}_{h,\bar{s}}(s,a) \left[2\sqrt{\frac{2\mathrm{Var}(P_{s,a}, V^{\pi})\iota}{n\sum_{h=1}^{\infty} \xi^{\mu}_{h}(s,a)}} + \frac{14\left\|V^{\pi}\right\|_{\infty} \iota}{3n\sum_{h=1}^{\infty} \xi^{\mu}_{h}(s,a)} \right]$$

$$= \sqrt{8\sum_{\substack{s,a\\ s\neq g}} \frac{\sum_{h=1}^{\infty} \xi^{\pi}_{h,\bar{s}}(s,a)}{\sum_{h=1}^{\infty} \xi^{\mu}_{h}(s,a)} \cdot \sum_{s,a,s\neq g} \sum_{h=1}^{\infty} \xi^{\pi}_{h,\bar{s}}(s,a) \mathrm{Var}(P_{s,a}, V^{\pi}) \frac{\iota}{n}} + \frac{14\left\|V^{\pi}\right\|_{\infty} \iota}{3n} \cdot \sum_{s,a,s\neq g} \frac{d^{\pi}_{\bar{s}}(s,a)}{d^{\mu}(s,a)}$$

$$\leq \sqrt{8\frac{T^{\pi}_{\bar{s}}}{d_m} \cdot \frac{\left\|V^{\pi}\right\|_{\infty}^{2} \iota}{n}} + \frac{14\left\|V^{\pi}\right\|_{\infty} \iota}{3n} \cdot \frac{T^{\pi}_{\bar{s}}}{d_m}.$$

where the first inequality uses Lemma 3.5 and Bernstein inequality, the second inequality uses Lemma 12.6 and the last one uses Lemma 4.3. Recall $\bar{T}^{\pi} = \max_{\bar{s}\in\mathcal{S}} T^{\pi}_{\bar{s}}$, then combine all the three parts together and take the max over $\bar{s}$, we can derive

$$\left(1 - 4\bar{T}^{\pi}\sqrt{\frac{S\iota}{nd_m}}\right) \left\| \widehat{V}^{\pi} - V^{\pi} \right\|_{\infty} \leq T^{\pi}_{\bar{s}} \left[2\sqrt{\frac{2\max_{s,a} \mathrm{Var}_c(s,a)\iota}{n \cdot d_m}} + \frac{8\iota}{3n \cdot d_m} \right] + \sqrt{8\frac{T^{\pi}_{\bar{s}}}{d_m} \cdot \frac{\left\|V^{\pi}\right\|_{\infty}^{2} \iota}{n}} + \frac{14\left\|V^{\pi}\right\|_{\infty} \iota}{3n} \frac{T^{\pi}_{\bar{s}}}{d_m}$$

$$\leq O\left(\frac{\bar{T}^{\pi}\sqrt{\max_{s,a} \mathrm{Var}_c(s,a)\iota} + \sqrt{\bar{T}^{\pi}\left\|V^{\pi}\right\|_{\infty}^{2} \iota}}{\sqrt{n \cdot d_m}} \right) + O\left(\frac{\bar{T}^{\pi}\left\|V^{\pi}\right\|_{\infty} \cdot \iota}{n \cdot d_m} \right),$$

therefore it implies (by applying the condition $n \geq 64(\bar{T}^{\pi})^2 \frac{S\iota}{d_m}$)

$$\left\| \widehat{V}^{\pi} - V^{\pi} \right\|_{\infty} \leq O\left(\frac{\bar{T}^{\pi}\sqrt{\max_{s,a} \mathrm{Var}_c(s,a)\iota} + \sqrt{\bar{T}^{\pi}\left\|V^{\pi}\right\|_{\infty}^{2} \iota}}{\sqrt{n \cdot d_m}} \right) + O\left(\frac{\bar{T}^{\pi}\left\|V^{\pi}\right\|_{\infty} \cdot \iota}{n \cdot d_m} \right)$$

□

6 PROOF OF THEOREM ??

Theorem 6.1 (Restatement of Theorem **??**). *Denote $d_m := \min\{\sum_{h=1}^{\infty} \xi^{\mu}_{h}(s,a) : s.t. \sum_{h=1}^{\infty} \xi^{\mu}_{h}(s,a) > 0\}$, and T^{π}_{s} to be the expected time to hit g when starting from s. Define $\bar{T}^{\pi} = \max_{\bar{s}\in\mathcal{S}} T^{\pi}_{\bar{s}}$. Then when $n \geq \max\{\frac{49S\iota}{9d_m}, 64(\bar{T}^{\pi})^2 \frac{S\iota}{d_m}, C\cdot\iota/d_m\}$,* *we have with probability* $1-\delta$*,*

$$|V^{(i)}(s_{\text{init}}) - V^{\pi}(s_{\text{init}})| \leq 4 \sum_{s,a,s\neq g} d^{\pi}(s,a)\sqrt{\frac{2\mathrm{Var}_{P_{s,a}}[V^{\pi}+c]}{n\cdot d^{\mu}(s,a)}} + \widetilde{O}(\frac{1}{n}) + \frac{\epsilon_{\text{OPE}}}{1-\rho}.$$

where the $\widetilde{O}$ *absorbs Polylog term and even higher order term.*

Proof. Recall that we start from the initial state s_{init}. Then by Lemma 3.3,

$$\left|V^{\pi}(s_{\text{init}}) - V^{(i)}(s_{\text{init}})\right| \leq \left|V^{\pi}(s_{\text{init}}) - \widehat{V}^{\pi}(s_{\text{init}})\right| + \left|\widehat{V}^{\pi}(s_{\text{init}}) - V^{(i)}(s_{\text{init}})\right| \leq \left|V^{\pi}(s_{\text{init}}) - \widehat{V}^{\pi}(s_{\text{init}})\right| + \frac{\epsilon_{\text{OPE}}}{1-\rho}, \tag{18}$$

it remains to bound $|V^{\pi}(s_{\text{init}}) - \widehat{V}^{\pi}(s_{\text{init}})|$.

From Lemma 4.2, we have

$$\begin{aligned}
\left|V^{\pi}(s_{\text{init}}) - \widehat{V}^{\pi}(s_{\text{init}})\right| &= \left|\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h^{\pi}(s,a)\{(\widehat{c}-c)(s,a) + (\tilde{P}_{s,a}-P_{s,a})\widehat{V}^{\pi}\}\right| \\
&\leq \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h^{\pi}(s,a)|(\widehat{c}-c)(s,a)| + \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h^{\pi}(s,a)|(\tilde{P}_{s,a}-P_{s,a})(\widehat{V}^{\pi}-V^{\pi})| \\
&\quad + \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h^{\pi}(s,a)|(\tilde{P}_{s,a}-P_{s,a})V^{\pi}|
\end{aligned}$$

We bound the above three parts one by one. First of all, by Bernstein inequality and Lemma 12.6 together with union bound, with probability $1-\delta$,

$$\begin{aligned}
&\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h^{\pi}(s,a)|(\widehat{c}-c)(s,a)| \leq \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h^{\pi}(s,a)\left[2\sqrt{\frac{\mathrm{Var}_c(s,a)\iota}{n(s,a)}} + \frac{4\iota}{3n(s,a)}\right] \\
&\overset{\text{(i)}}{\leq} \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h^{\pi}(s,a)\left[2\sqrt{\frac{2\mathrm{Var}_c(s,a)\iota}{n\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)}} + \frac{8\iota}{3n\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)}\right] \\
&\overset{\text{(ii)}}{\leq} \sqrt{\sum_{\substack{s,a\\ s\neq g}} \frac{\sum_{h=1}^{\infty}\xi_h^{\pi}(s,a)}{\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)}}\sqrt{\sum_{\substack{s,a\\ s\neq g}}\sum_{h=1}^{\infty}\xi_h^{\pi}(s,a)\mathrm{Var}_c(s,a)\frac{\iota}{n}} + \left(\sum_{s,a,s\neq g}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)}\right)\frac{8\iota}{3n} \\
&\overset{\text{(iii)}}{\leq} \sqrt{\sum_{\substack{s,a\\ s\neq g}} \frac{d^{\pi}(s,a)}{d^{\mu}(s,a)}\cdot\frac{V^{\pi}(s_{\text{init}})\cdot\iota}{n}} + \left(\sum_{s,a,s\neq g}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)}\right)\frac{8\iota}{3n},
\end{aligned} \tag{19}$$

(i) uses Lemma 12.6 and (ii) uses Cauchy-Schwartz inequality. (iii) uses the fact that $\mathrm{Var}_c(s,a) \leq \mathbb{E}C(s,a)^2 \leq c(s,a)$ since the realization $C(s,a) \in [0,1]$ and the definition of $V^{\pi}(s_{\text{init}})$.

For the second part, note

$$
\begin{aligned}
&\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h(s,a)|(\tilde{P}_{s,a}-P_{s,a})(\widehat{V}^{\pi}-V^{\pi})| \\
\leq&\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h(s,a)\left[\sqrt{\frac{2S\cdot\mathrm{Var}(P_{s,a},\widehat{V}^{\pi}-V^{\pi})\iota_{s,a}}{n(s,a)}}+\frac{4\left\|\widehat{V}^{\pi}-V^{\pi}\right\|_{\infty}S\iota_{s,a}}{3n(s,a)}+\frac{\left\|\widehat{V}^{\pi}-V^{\pi}\right\|_{\infty}}{n(s,a)+1}\right] \\
\leq&\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h(s,a)\left[\sqrt{\frac{4S\cdot\mathrm{Var}(P_{s,a},\widehat{V}^{\pi}-V^{\pi})\iota_{s,a}}{n\sum_{h=1}^{\infty}\xi_h^{\pi}(s,a)}}+\frac{14\left\|\widehat{V}^{\pi}-V^{\pi}\right\|_{\infty}S\iota_{s,a}}{3n\sum_{h=1}^{\infty}\xi_h^{\pi}(s,a)}\right] \\
\leq&\sqrt{\sum_{\substack{s,a\\ s\neq g}}\frac{\sum_{h=0}^{\infty}\xi_h^{\pi}(s,a)}{\sum_{h=0}^{\infty}\xi_h^{\mu}(s,a)}}\sqrt{4S\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h(s,a)\left\|\widehat{V}^{\pi}-V^{\pi}\right\|^2\frac{\iota}{n}}+\frac{14\left\|\widehat{V}^{\pi}-V^{\pi}\right\|_{\infty}S\iota}{3n}\sum_{s,a,s\neq g}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)} \\
\leq&\sqrt{\sum_{\substack{s,a\\ s\neq g}}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)}\cdot 4ST^{\pi}\left\|\widehat{V}^{\pi}-V^{\pi}\right\|_{\infty}^2\cdot\frac{\iota}{n}}+\frac{14\left\|\widehat{V}^{\pi}-V^{\pi}\right\|_{\infty}S\iota}{3n}\sum_{s,a,s\neq g}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)} \\
\leq&\,4\sqrt{\sum_{\substack{s,a\\ s\neq g}}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)}\cdot ST^{\pi}\left\|\widehat{V}^{\pi}-V^{\pi}\right\|_{\infty}^2\cdot\frac{\iota}{n}}
\end{aligned}
\tag{20}
$$

where the first inequality uses Lemma 10.3 and Lemma 3.5, the second inequality uses Lemma 12.6, the third inequality uses $\mathrm{Var}(\cdot)\leq\|\cdot\|_{\infty}^2$ and CS inequality, the fourth inequality use Lemma 4.3 and the last inequality follows from the condition $n\geq\frac{49S\iota}{9d_m}\geq\frac{49S\iota}{9T^{\pi}}\sum_{s,a,s\neq g}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)}$.

For the third part, we have

$$
\begin{aligned}
&\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^{\pi}(s,a)|(\widetilde{P}_{s,a}-P_{s,a})V^{\pi}| \\
\leq&\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^{\pi}(s,a)\left[2\sqrt{\frac{\mathrm{Var}(P_{s,a},V^{\pi})\iota}{n(s,a)}}+\frac{4\left\|V^{\pi}\right\|_{\infty}\iota}{3n(s,a)}+\frac{\left\|V^{\pi}\right\|_{\infty}}{n(s,a)}\right] \\
\leq&\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^{\pi}(s,a)\left[2\sqrt{\frac{2\mathrm{Var}(P_{s,a},V^{\pi})\iota}{n\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)}}+\frac{14\left\|V^{\pi}\right\|_{\infty}\iota}{3n\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)}\right] \\
(\leq&\sqrt{8\sum_{\substack{s,a\\ s\neq g}}\frac{\sum_{h=1}^{\infty}\xi_h^{\pi}(s,a)}{\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)}\cdot\sum_{s,a,s\neq g}\sum_{h=1}^{\infty}\xi_h^{\pi}(s,a)\mathrm{Var}(P_{s,a},V^{\pi})\frac{\iota}{n}}+\frac{14\left\|V^{\pi}\right\|_{\infty}\iota}{3n}\cdot\sum_{s,a,s\neq g}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)} \\
\leq&\sqrt{8\sum_{\substack{s,a\\ s\neq g}}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)}\cdot\frac{\left\|V^{\pi}\right\|_{\infty}^2\iota}{n}}+\frac{14\left\|V^{\pi}\right\|_{\infty}\iota}{3n}\cdot\sum_{s,a,s\neq g}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)} \\
\leq&4\sqrt{2\sum_{\substack{s,a\\ s\neq g}}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)}\cdot\frac{\left\|V^{\pi}\right\|_{\infty}^2\iota}{n}}),
\end{aligned}
\tag{21}
$$

where the first inequality uses Lemma 10.3 and Lemma 3.5, the second inequality uses Lemma 12.6. The third inequality uses the Cauchy-Schwartz inequality.

Combine Equation (19), (20) and (21) together, we obtain

$$\left|V^\pi(s_{\text{init}}) - \widehat{V}^\pi(s_{\text{init}})\right| \le \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h^\pi(s,a)\left[2\sqrt{\frac{2\mathrm{Var}(P_{s,a},V^\pi)\iota}{n\sum_{h=1}^\infty \xi_h^\mu(s,a)}} + \frac{14\left\|V^\pi\right\|_\infty \iota}{3n\sum_{h=1}^\infty \xi_h^\mu(s,a)}\right]$$

$$+\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h^\pi(s,a)\left[2\sqrt{\frac{2\mathrm{Var}_c(s,a)\iota}{n\sum_{h=1}^\infty \xi_h^\mu(s,a)}} + \frac{8\iota}{3n\sum_{h=1}^\infty \xi_h^\mu(s,a)}\right]$$

$$+4\sqrt{\sum_{\substack{s,a\\ s\neq g}} \frac{d^\pi(s,a)}{d^\mu(s,a)}\cdot ST^\pi \left\|\widehat{V}^\pi - V^\pi\right\|_\infty^2 \cdot \frac{\iota}{n}}$$

$$\le 4\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h^\pi(s,a)\sqrt{\frac{2[\mathrm{Var}(P_{s,a},V^\pi)+\mathrm{Var}_c(s,a)]\iota}{n\sum_{h=1}^\infty \xi_h^\mu(s,a)}} + \sum_{\substack{s,a\\ s\neq g}} \frac{22\sum_{h=1}^\infty \xi_h^\pi(s,a)\left\|V^\pi\right\|_\infty \iota}{3n\sum_{h=1}^\infty \xi_h^\mu(s,a)}$$

$$+4\sqrt{\sum_{\substack{s,a\\ s\neq g}} \frac{d^\pi(s,a)}{d^\mu(s,a)}\cdot ST^\pi \cdot \frac{(\bar{T}^\pi)^2\cdot \max_{s,a}\mathrm{Var}_c(s,a)\iota + \bar{T}^\pi\left\|V^\pi\right\|_\infty^2 \iota}{n\cdot d_m}\cdot\frac{\iota}{n}} + \widetilde{O}\left(\frac{1}{n^{3/2}}\right)$$

$$=4\sum_{s,a,s\neq g} d^\pi(s,a)\sqrt{\frac{2\mathrm{Var}_{P_{s,a}}[V^\pi + c]}{nd^\mu(s,a)}} + \widetilde{O}(\frac{1}{n})$$

where the only inequality uses Theorem 5.1. Combining this with (18) we finish the proof of Theorem **??**. □

7 PREPARATIONS FOR PROVING OFFLINE LEARNING SSP

Throughout the whole section, we denote $\iota = O(\log(SA/\delta))$. All the results apply to the construction of Algorithm **??**. In particular, we use $\bar{V}$ to denote the limit of $V^{(i)}$ (by letting $\epsilon_{\text{OPL}} = 0$). This limit exists, as guaranteed by Lemma 7.6.

7.1 AUXILIARY LEMMAS

Lemma 7.1. *Denote the limit of sequence $V^{(i)}$ in Algorithm* **??** *as $\bar{V}$, we have that*

$$\left\|\bar{V}\right\|_\infty \le \widetilde{B}$$

Proof. First of all, by Lemma 7.6, we know $\bar{V}$ exists. Next, from the Algorithm **??**, we can get that

$$Q^{(i+1)}(s,a) = \min\{\widetilde{c}(s,a) + \widetilde{P}'_{s,a}V^{(i)} + b_{s,a}(V^{(i)})\ ,\widetilde{B}\} \le \widetilde{B} \qquad \forall (s,a)\in\mathcal{S}\times\mathcal{A}, \forall i\in\mathbb{N}$$

and thus

$$V^{(i+1)}(s) = \min_a Q^{(i+1)}(s,a) \le \widetilde{B} \qquad \forall s\in\mathcal{S}\times\mathcal{A}, \forall i\in\mathbb{N}.$$

It implies that $\bar{V}(s) = \lim_{i\to\infty} V^{(i)}(s) \le \widetilde{B}$. □

Lemma 7.2. *For any $V(\cdot)\in\mathbb{R}^S$ satisfying $V(g)=0$,*

$$|(\widetilde{P}'_{s,a} - \widehat{P}_{s,a})V| \le \frac{||V||_\infty}{n_{max}+1}, \qquad |\mathrm{Var}(\widetilde{P}'_{s,a},V) - \mathrm{Var}(\widehat{P}_{s,a},V)| \le \frac{2\left\|V\right\|_\infty^2}{n_{max}+1}. \tag{22}$$

Proof. The proof is similar to Lemma 12 in Tarbouriech et al. [2021]. We include the proof for completeness. Since $V(g)=0$, for all state $s\neq g$, we have $\widetilde{P}'_{s,a}V = \sum_{s',s'\neq g}(\frac{n_{max}}{n_{max}+1})\widehat{P}(s'|s,a)V(s') = (\frac{n_{max}}{n_{max}+1})\widehat{P}_{s,a}V$

$$|(\widetilde{P}'_{s,a} - \widehat{P}_{s,a})V| = |(\frac{n_{max}}{n_{max}+1})\widehat{P}_{s,a}V - \widehat{P}_{s,a}V|$$

$$= \frac{|\sum_{s', s' \neq g} \widehat{P}(s'|s,a)V(s')|}{n_{max}+1} \leq \frac{||V||_\infty}{n_{max}+1}.$$

Then we prove the second inequality. Similarly,

$$\begin{aligned}
|\mathrm{Var}(\widetilde{P}'_{s,a}, V) - \mathrm{Var}(\widehat{P}_{s,a}, V)| &= |\widetilde{P}'_{s,a}V^2 - (\widetilde{P}'_{s,a}V)^2 - \widehat{P}_{s,a}V^2 + (\widehat{P}_{s,a}V)^2| \\
&= |(\frac{n_{max}}{n_{max}+1})\widehat{P}_{s,a}V^2 - (\frac{n_{max}}{n_{max}+1}\widehat{P}_{s,a}V)^2 - \widehat{P}_{s,a}V^2 + (\widehat{P}_{s,a}V)^2| \\
&= |-(\frac{1}{n_{max}+1})\{\widehat{P}_{s,a}V^2 - (\widehat{P}_{s,a}V)^2\} + \frac{n_{max}}{(n_{max}+1)^2}(\widehat{P}_{s,a}V)^2| \\
&= (\frac{1}{n_{max}+1})|\mathrm{Var}(\widehat{P}_{s,a}, V)| + |\frac{n_{max}}{(n_{max}+1)^2}(\widehat{P}_{s,a}V)^2| \leq \frac{2\|V\|_\infty^2}{n_{max}+1}.
\end{aligned}$$

□

Lemma 7.3. *Let* $T_{\max} = \max_i T_i$ *and* $n > O(T_{\max}^2 \log(SA/\delta)/d_m^2)$. *If in addition* $n \geq \frac{S\iota}{2d_m}$, *with probability at least* $1-\delta$, *we have that for any state action pair (s,a),*

$$|(P_{s,a} - \widetilde{P}'_{s,a})\bar{V}| \leq \frac{\widetilde{B}}{n(s,a)} + \frac{16\widetilde{B}\iota}{3n(s,a)} + 2\sqrt{\frac{\mathrm{Var}(\widetilde{P}', \bar{V})\iota}{n(s,a)}} + 6\sqrt{\frac{S\iota}{n(s,a)}}\left\|\bar{V} - V^\star\right\|_\infty$$

Proof. First, we can bound term $(P_{s,a} - \widetilde{P}'_{s,a})\bar{V}$.

$$|(P_{s,a} - \widetilde{P}'_{s,a})\bar{V}| \leq |(\widehat{P}_{s,a} - \widetilde{P}'_{s,a})\bar{V}| + |(P_{s,a} - \widehat{P}_{s,a})(\bar{V} - V^\star)| + |(P_{s,a} - \widehat{P}_{s,a})V^\star| \tag{23}$$

Then we bound the above three terms one by one. From Lemma 7.1 and Lemma 7.2, we have

$$|(\widehat{P}_{s,a} - \widetilde{P}_{s,a})\bar{V}| \leq \frac{\left\|\bar{V}\right\|_\infty}{n_{max}+1} \leq \frac{\widetilde{B}}{n_{max}+1}. \tag{24}$$

For the second term, we have

$$\begin{aligned}
|(P_{s,a} - \widehat{P}_{s,a})(\bar{V} - V^\star)| &\leq \sqrt{\frac{2S\mathrm{Var}(P_{s,a}, \bar{V} - V^\star)\iota}{n(s,a)}} + \frac{2\left\|\bar{V} - V^\star\right\|_\infty S\iota}{3n(s,a)} \\
&\leq \sqrt{\frac{2S\iota}{n(s,a)}}\left\|\bar{V} - V^\star\right\|_\infty + \frac{\left\|\bar{V} - V^\star\right\|_\infty S\iota}{n(s,a)},
\end{aligned} \tag{25}$$

where the first inequality holds because of lemma 10.3. For the last term, we have that

$$\begin{aligned}
|(P_{s,a} - \widehat{P}_{s,a})V^\star| &\overset{\text{(i)}}{\leq} \sqrt{\frac{2\mathrm{Var}(\widehat{P}_{s,a}, V^\star)\iota}{n(s,a)}} + \frac{7||V^\star||_\infty\iota}{3n(s,a)} \\
&\overset{\text{(ii)}}{\leq} 2\sqrt{\frac{\mathrm{Var}(\widehat{P}_{s,a}, V^\star - \bar{V})\iota}{n(s,a)}} + 2\sqrt{\frac{\mathrm{Var}(\widehat{P}_{s,a}, \bar{V})\iota}{n(s,a)}} + \frac{7\widetilde{B}\iota}{3n(s,a)} \\
&\leq 2\sqrt{\frac{\iota}{n(s,a)}})\left\|\bar{V} - V^\star\right\|_\infty + 2\sqrt{\frac{\mathrm{Var}(\widehat{P}_{s,a}, \bar{V})\iota}{n(s,a)}} + \frac{7\widetilde{B}\iota}{3n(s,a)} \\
&\overset{\text{(iii)}}{\leq} 2\sqrt{\frac{\iota}{n(s,a)}})\left\|\bar{V} - V^\star\right\|_\infty + 2\sqrt{\frac{\mathrm{Var}(\widetilde{P}'_{s,a}, \bar{V})\iota}{n(s,a)}} + \frac{2\sqrt{2\iota}||\bar{V}||_\infty}{n(s,a)} + \frac{7\widetilde{B}\iota}{3n(s,a)} \\
&\leq 2\sqrt{\frac{\iota}{n(s,a)}})\left\|\bar{V} - V^\star\right\|_\infty + 2\sqrt{\frac{\mathrm{Var}(\widetilde{P}'_{s,a}, \bar{V})\iota}{n(s,a)}} + \frac{16\widetilde{B}\iota}{3n(s,a)},
\end{aligned} \tag{26}$$

where (i) holds under event $\mathcal{E}_3$. (ii) holds because of $\mathrm{Var}(X+Y) \le 2\mathrm{Var}(X) + 2\mathrm{Var}(Y)$. (iii) comes from Lemma 7.2. Both (ii) and (iii) uses the result that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ when $a \ge 0$ and $b \ge 0$. Combine the above inequalities together, we can get

$$(P_{s,a} - \widetilde{P}'_{s,a})\bar{V} \le \frac{\widetilde{B}}{n(s,a)} + \frac{16\widetilde{B}\iota}{3n(s,a)} + 2\sqrt{\frac{\mathrm{Var}(\widetilde{P}', \bar{V})\iota}{n(s,a)}} + \left(\sqrt{\frac{2S\iota}{n(s,a)}} + \frac{S\iota}{n(s,a)} + 2\sqrt{\frac{\iota}{n(s,a)}}\right) \left\|\bar{V} - V^\star\right\|_\infty.$$

Since with probability $1-\delta$, by Lemma 12.2 $n(s,a) \ge \frac{1}{2} n d_m$. When $n \ge \frac{S\iota}{2d_m}$, we have

$$\begin{aligned}
\left(\sqrt{\frac{2S\iota}{n(s,a)}} + \frac{S\iota}{n(s,a)} + 2\sqrt{\frac{\iota}{n(s,a)}}\right) \left\|\bar{V} - V^\star\right\|_\infty &\le (\sqrt{2} + 2 + 2)\sqrt{\frac{S\iota}{n(s,a)}} \left\|\bar{V} - V^\star\right\|_\infty \\
&\le 6\sqrt{\frac{S\iota}{n(s,a)}} \left\|\bar{V} - V^\star\right\|_\infty
\end{aligned}$$

□

Lemma 7.4. *Define function* $f : \mathbb{R}^{S'} \times \mathbb{R}^{S'} \times \mathbb{R} \to \mathbb{R}$ *as* $f(p, v, n) = pv + \max\{2\sqrt{\frac{\mathrm{Var}(p,v)\iota}{n}}, 4\frac{\widetilde{B}\iota}{n}\}$*, if* $\|v\|_\infty \le \widetilde{B}$ *and* $v(g) = 0$*, then we have* $(\frac{\partial f}{\partial v})(s) \ge 0$ *and* $\sum_{s, s\neq g} (\frac{\partial f}{\partial v})(s) \le 1 - p(g)^2$.

Proof.

$$\begin{aligned}
(\frac{\partial f}{\partial v})(s) &= p(s) + \mathbb{I}\{2\sqrt{\frac{\mathrm{Var}(p,v)\iota}{n}} \ge 4\frac{\widetilde{B}\iota}{n}\} 2\sqrt{\frac{\iota}{n}} \frac{\partial(\sqrt{\mathrm{Var}(p,v)})}{\partial v(s)} \\
&= p(s) + \mathbb{I}\{2\sqrt{\frac{\mathrm{Var}(p,v)\iota}{n}} \ge 4\frac{\widetilde{B}\iota}{n}\} 2\sqrt{\frac{\iota}{n}} \frac{p(s)(v(s) - pv)}{\sqrt{\mathrm{Var}(p,v)}}
\end{aligned}$$

Simplifying the above equation, we can get

$$(\frac{\partial f}{\partial v})(s) \ge \min\{p(s), p(s) - \frac{p(s)(pv - v(s))}{\widetilde{B}}\} \tag{27}$$

Since $|pv - v(s)| \le \widetilde{B}$, we have $p(s) - \frac{p(s)(pv - v(s))}{\widetilde{B}} \ge 0$. Then we would have $(\frac{\partial f}{\partial v})(s) \ge 0$. For the second part, we have

1. Case I: $2\sqrt{\frac{\mathrm{Var}(p,v)\iota}{n}} \ge 4\frac{\widetilde{B}\iota}{n}$, we have

$$\begin{aligned}
\sum_{s,s\neq g} (\frac{\partial f}{\partial v})(s) &= \sum_{s,s\neq g} \{p(s) + 2\sqrt{\frac{\iota}{n\mathrm{Var}(p,v)}} p(s)(v(s) - pv)\} \\
&\le \sum_{s,s\neq g} p(s) + 2\sqrt{\frac{\iota}{n\mathrm{Var}(p,v)}} \{\sum_{s,s\neq g} p(s)v(s) - \sum_{s,s\neq g} p(s)(\sum_{s,s'\neq g} p(s')v(s'))\} \\
&\le \sum_{s,s\neq g} p(s) + 2\sqrt{\frac{\iota}{n\mathrm{Var}(p,v)}} [\sum_{s,s\neq g} p(s)v(s)](1 - \sum_{s,s\neq g} p(s)) \\
&\le \sum_{s,s\neq g} p(s) + \frac{[\sum_{s,s\neq g} p(s)v(s)](1 - \sum_{s,s\neq g} p(s))}{\widetilde{B}} \\
&\le \sum_{s,s\neq g} p(s) + (\sum_{s,s\neq g} p(s))(1 - \sum_{s,s\neq g} p(s)) = 1 - (p(g))^2
\end{aligned}$$

2. Case II: $2\sqrt{\frac{\mathrm{Var}(p,v)\iota}{n}} \le 4\frac{\widetilde{B}\iota}{n}$, we have

$$\sum_{s,s\neq g} (\frac{\partial f}{\partial v})(s) = \sum_{s,s\neq g} p(s) = 1 - p(g) \le 1 - p(g)^2$$

Combine this inequality with (27), we complete the proof. $\square$

Lemma 7.5. *When* $n \geq \max\{\frac{4B_\star - 2c_{min}}{c_{min} d_{max}}, \frac{26^2 \times 2S\iota(\bar{T}^\star)^2}{d_m}, \frac{10^6(\sqrt{\widetilde{B}}+1)^4 S\iota\bar{T}^\star\widetilde{T}}{B_\star(\sqrt{B_\star}+1)^2 d_m}, O(T_{\max}^2 \log(SA/\delta)/d_m^2)\}$*, $\bar{\pi}$ is a proper policy (Recall $\bar{\pi}$ is the output of Algorithm* **??***).*

Proof. By definition we need to show that $T^{\bar{\pi}}(s) < \infty$ for any $s \in \mathcal{S}$. We prove this by contradiction. Suppose $T^{\bar{\pi}}(s) = \infty$, then we have that there exists at least one state e such that the expected visiting times of state e is infinite, i.e., $\exists e \in \mathcal{S}$, such that $T^{\bar{\pi}}_{e,e} = \infty$. In this case, e is a (positive) recurrent state in the finite Markov Chain induced by policy $\bar{\pi}$. Denote the communication class which e belongs as $\mathcal{S}_0$. Since the state space is finite, we have that every state in the communication class $\mathcal{S}_0$ is recurrent. From the finite Markov Chain theory, we know that the communication class $\mathcal{S}_0$ is closed. In other words, $\forall x \in \mathcal{S}_0$, and $\forall y \in \mathcal{S} \backslash \mathcal{S}_0$, we have $P(y|x, \bar{\pi}(x)) = 0$. Thus with probability 1, we have $\sum_{i=1}^n \sum_{j=1}^{T_i} \mathbb{I}(s_j^{(i)} = x, a_j^{(i)} = \bar{\pi}(x), s_{j+1}^{(i)} = y) = 0$. This implies that $\widehat{P}(y|x, \bar{\pi}(x)) = \frac{\sum_{i=1}^n \sum_{j=1}^{T_i} \mathbb{I}(s_j^{(i)}=x,\, a_j^{(i)}=\bar{\pi}(x),\, s_{j+1}^{(i)}=y)}{n(x,\bar{\pi}(x))} = 0$. By definition of the estimated transition matrix $\widetilde{P}'$, we have $\widetilde{P}'(g|x, \bar{\pi}(x)) = \frac{1}{n_{max}+1}$. It follows that

$$\sum_{h=0}^{\infty} \sum_{s \in \mathcal{S}_0} \tilde{\xi}_{h,e}(s) = \sum_{h=0}^{\infty} (\frac{n_{max}}{n_{max}+1})^h = n_{max} + 1. \tag{28}$$

Then we have

$$\sum_{h=0}^{\infty} \sum_{s \in \mathcal{S}_0} \tilde{\xi}_{h,e}(s) = n_{max} + 1 \geq \frac{1}{2} n d_{max} + 1, \tag{29}$$

where the last inequality holds with probability $1-\delta$ by Lemma 12.6. Define $\widetilde{V}^{\bar{\pi}}(e) = \sum_{h=0}^{\infty} \mathbb{E}_{\widetilde{P}',\bar{\pi}}[\hat{c}(s_h, a_h)|s_0 = e]$, then

$$\begin{aligned}
\widetilde{V}^{\bar{\pi}}(e) &\geq \sum_{h=0}^{\infty} \mathbb{E}_{\widetilde{P}',\bar{\pi}}[c_{min}] \\
&\geq \sum_{h=0}^{\infty} \mathbb{E}_{\widetilde{P}',\bar{\pi}}[c_{min}\mathbb{I}(s_h \in \mathcal{S}_0)] = c_{min} \sum_{h=0} \sum_{s \in \mathcal{S}_0} \tilde{\xi}_{h,e}(s) \geq c_{min}(\frac{1}{2} n d_{max} + 1).
\end{aligned}$$

Apply Lemma 1.1 to the SSP problem with $M := \left\langle \mathcal{S}, \mathcal{A}, \widetilde{P}', \hat{c}, e, g \right\rangle$, we can get

$$\widetilde{V}^{\bar{\pi}}(s) = \widetilde{P}'_{s,\bar{\pi}(s)} \widetilde{V}^{\bar{\pi}} + \hat{c}(s, \bar{\pi}(s)) := \widetilde{\mathcal{T}}' \widetilde{V}^{\bar{\pi}}(s)$$

Since $\bar{V}(s) = \widetilde{P}'_{s,\bar{\pi}(s)} \bar{V} + \hat{c}(s, \bar{\pi}(s)) + b_{s,\bar{\pi}(s)}(\bar{V}) = \widetilde{\mathcal{T}} \bar{V}(s)$, from Lemma 2.1 we have $\bar{V}(s) \geq \widetilde{V}^{\bar{\pi}}(s)$ (since $b(\bar{V})$ is non-negative and both $\widetilde{\mathcal{T}}, \widetilde{\mathcal{T}}'$ are monotone operators). Then we get

$$\bar{V}(e) \geq \widetilde{V}^{\bar{\pi}}(e) \geq c_{min}(\frac{1}{2} n d_{max} + 1). \tag{30}$$

From Lemma 8.1 (note Lemma 8.1 only bounds $\bar{V}$ and $V^\star$ and has nothing to do with $\bar{\pi}$), we have that with probability $1-\delta$, when $n \geq \max\{\frac{26^2 \times 2S\iota(\bar{T}^\star)^2}{d_m}, \frac{10^6(\sqrt{\widetilde{B}}+1)^4 S\iota\bar{T}^\star\widetilde{T}}{B_\star(\sqrt{B_\star}+1)^2 d_m}\}$, which implies $n \geq \frac{900\bar{T}^\star\iota(\sqrt{B_\star}+1)^2}{B_\star d_m}$, we have $\bar{V}(e) \leq V^\star(e) + B_\star \leq 2B_\star$. Combine this inequality with (30), we can get $n \leq \frac{4B_\star - 2c_{\min}}{c_{\min} d_{max}}$, which contradicts with the assumptions in the lemma. $\square$

7.2 CONVERGENCE OF PESSIMISTIC VALUE ITERATION IN ALGORITHM ??

Define the operator $\widetilde{\mathcal{T}}$ as $\widetilde{\mathcal{T}}(V)(s) = \min_a \left\{ \min\{\widehat{c}(s,a) + \widetilde{P}_{s,a} V + b_{s,a}(V)\, ,\, \widetilde{B}\} \right\}$. First, we prove that $\widetilde{\mathcal{T}}$ is a contraction mapping.

Lemma 7.6. $\widetilde{\mathcal{T}} : \mathbb{R}^{\mathcal{S}} \times \{0\} \to \mathbb{R}^{\mathcal{S}} \times \{0\}$ *is a contraction mapping, i.e.,* $\forall V_1, V_2 \in \mathbb{R}^{\mathcal{S}'}$, $V_1(g) = V_2(g) = 0$*, we have*

$$\left\| \widetilde{\mathcal{T}} V_1 - \widetilde{\mathcal{T}} V_2 \right\|_\infty \leq \gamma \left\| V_1 - V_2 \right\|_\infty, \tag{31}$$

where $\gamma := 1 - \frac{1}{(1+\max_{s,a} n(s,a))^2}$.

Proof of Lemma 7.6. First, we prove the result for $s = g$. Since $b_{g,a}(V) = 0$, then we have $\widetilde{\mathcal{T}}(V)(g) = 0$ and thus $\widetilde{\mathcal{T}}(V_1)(g) - \widetilde{\mathcal{T}}(V_2)(g) = 0$. When $s \neq g$, we have

$$
\begin{aligned}
|\widetilde{\mathcal{T}}V_1(s) - \widetilde{\mathcal{T}}V_2(s)| &\leq \max_a |\min\{\widehat{c}(s,a) + \widetilde{P}'_{s,a}V_1 + b_{s,a}(V_1)\,,\widetilde{B}\} - \min\{\widehat{c}(s,a) + \widetilde{P}'_{s,a}V_2 + b_{s,a}(V_2)\,,\widetilde{B}\}| \\
&\overset{\text{(i)}}{\leq} \max_a |\{\widehat{c}(s,a) + \widetilde{P}'_{s,a}V_1 + b_{s,a}(V_1)\} - \{\widehat{c}(s,a) + \widetilde{P}'_{s,a}V_2 + b_{s,a}(V_2)\}| \\
&= \max_a |f(\widetilde{P}'_{s,a}, V_1) - f(\widetilde{P}'_{s,a}, V_2)| \\
&\overset{\text{(ii)}}{=} \max_a \sum_s |\langle (\frac{\partial f}{\partial v})(\widetilde{P}'_{s,a}, \theta(V_1 - V_2) + V_2), V_1 - V_2\rangle| \\
&\leq \max_a (\sum_{s', s\neq g} |(\frac{\partial f}{\partial v(s')})(\widetilde{P}'_{s,a}, \theta(V_1 - V_2) + V_2)|) \left\|V_1 - V_2\right\|_\infty \\
&\overset{\text{(iii)}}{\leq} \max_a \{1 - \widetilde{P}'_{s,a}(g)^2\} \left\|V_1 - V_2\right\|_\infty ,
\end{aligned}
$$

where (i) comes from Lemma 12.7. (ii) is due to the mean value theorem. (iii) uses the result in Lemma 7.4. Then we have

$$
\left\|\widetilde{\mathcal{T}}V_1(s) - \widetilde{\mathcal{T}}V_2(s)\right\|_\infty \leq \max_{s,a}\{1 - \widetilde{P}'_{s,a}(g)^2\} \left\|V_1 - V_2\right\|_\infty \leq \{1 - \frac{1}{(1 + \max_{s,a} n(s,a))^2}\} \left\|V_1 - V_2\right\|_\infty
$$

□

We then introduce the following two regret decomposition lemma.

7.3 REGRET DECOMPOSITION LEMMA FOR POLICY OPTIMIZATION

Lemma 7.7. *Suppose $\bar{V}$ is the limit of the sequence $V^{(i)}$ in Algorithm* **??**, *we have the following decomposition lemma.*

$$
\bar{V} - V^\star \leq \sum_{h=0}^{\infty} \sum_{\substack{s \\ s\neq g}} \xi_h^\star(s)\{(\widetilde{P}'_{s,\pi^\star(s)} - P_{s,\pi^\star(s)})\bar{V} + \hat{c}(s,\pi^\star(s)) - c(s,\pi^\star(s)) + b_{s,\pi^\star(s)}(\bar{V})\} \tag{32}
$$

Proof.

$$
\bar{V} - V^\star = \sum_{s, s\neq g} \xi_0^\star(s)(\bar{V}(s) - V^\star(s))
$$

Since for any $h \in \mathbf{N}$, we have

$$
\begin{aligned}
\sum_{s,s\neq g} \xi_h^\star(s)(\bar{V}(s) - V^\star(s)) &\overset{\text{(i)}}{\leq} \sum_{s,s\neq g} \xi_h^\star(s)(\bar{Q}(s,\pi^\star(s)) - Q^\star(s,\pi^\star(s))) \\
&\overset{\text{(ii)}}{=} \sum_{s,s\neq g} \xi_h^\star(s)\{\widetilde{P}'_{s,\pi^\star(s)}\bar{V} - P_{s,\pi^\star(s)}V^\star + \hat{c}(s,\pi^\star(s)) - c(s,\pi^\star(s)) + b_{s,\pi^\star(s)}(\bar{V})\} \\
&= \sum_{s,s\neq g} \xi_h^\star(s)\{(\widetilde{P}'_{s,\pi^\star(s)} - P_{s,\pi^\star(s)})\bar{V} + \hat{c}(s,\pi^\star(s)) - c(s,\pi^\star(s)) + b_{s,\pi^\star(s)}(\bar{V})\} \\
&\quad + \sum_{s,s\neq g} \xi_h^\star(s)P_{s,\pi^\star(s)}(\bar{V} - V^\star) \\
&= \sum_{s,s\neq g} \xi_h^\star(s)\{(\widetilde{P}'_{s,\pi^\star(s)} - P_{s,\pi^\star(s)})\bar{V} + \hat{c}(s,\pi^\star(s)) - c(s,\pi^\star(s)) + b_{s,\pi^\star(s)}(\bar{V})\} \\
&\quad + \sum_{s,s\neq g} \xi_{h+1}^\star(s)(\bar{V} - V^\star)(s)
\end{aligned}
$$

(i) follows from the fact that $\bar{V}(s) = \min_a \bar{Q}(s,a) \leq \bar{Q}(s, \pi^\star(s))$. (ii) uses the fact that $\bar{V}$ is the limit of $V^{(i)}$ and we have $\bar{Q}(s,a) = \hat{c}(s,a) + \widetilde{P}'_{s,a}\bar{V} + b_{s,a}(\bar{V})$. By recursion over time step h, we can get

$$\begin{aligned}\bar{V} - V^\star \leq &\sum_{h=0}^{H}\sum_{s,s\neq g} \xi_h^\star(s)\{(\widetilde{P}'_{s,\pi^\star(s)} - P_{s,\pi^\star(s)})\bar{V} + \hat{c}(s,\pi^\star(s)) - c(s,\pi^\star(s)) + b_{s,\pi^\star(s)}(\bar{V})\} \\ &+ \sum_{s,s\neq g} \xi_{H+1}^\star(s)(\bar{V} - V^\star), \end{aligned} \tag{33}$$

Since $\pi^\star$ is a proper policy, we have that for any $s \neq g$, $\lim_{H\to\infty} \xi_{H+1}^\star(s) = 0$. Also, $\bar{V}$ is bounded by $\widetilde{B}$ and $V^\star$ is bounded by $B_\star$. Thus let H goes to infinity, we can complete the proof. □

Lemma 7.8. *When $n \geq \max\{\frac{4B_\star - 2c_{\min}}{c_{\min} d_{max}}, \frac{26^2 \times 2S\iota(\bar{T}^\star)^2}{d_m}, \frac{10^6(\sqrt{\widetilde{B}}+1)^4 S\iota \bar{T}^\star \widetilde{T}}{B_\star(\sqrt{B_\star}+1)^2 d_m}, O(T_{\max}^2 \log(SA/\delta)/d_m^2)\}$, we have*

$$V^{\bar{\pi}} - \bar{V} = \sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}} \xi_h^{\bar{\pi}}(s)\{(P_{s,\bar{\pi}(s)} - \widetilde{P}'_{s,\bar{\pi}(s)})\bar{V} + c(s,\bar{\pi}(s)) - \hat{c}(s,\bar{\pi}(s)) - b_{s,\bar{\pi}(s)}(\bar{V})\}. \tag{34}$$

Proof. First of all, by the condition we have $\bar{\pi}$ is a proper policy by Lemma 7.5. We prove by recursion formula.

$$\begin{aligned}\sum_{s,s\neq g} \xi_h^{\bar{\pi}}(s)(V^{\bar{\pi}} - \bar{V}) &= \sum_{s,s\neq g} \xi_h^{\bar{\pi}}(s)\{P_{s,\bar{\pi}(s)}V^{\bar{\pi}} + c(s,\bar{\pi}(s)) - \widetilde{P}'_{s,\bar{\pi}(s)}\bar{V} - \hat{c}(s,\bar{\pi}(s)) - b_{s,\bar{\pi}(s)}(\bar{V})\} \\ &= \sum_{s,s\neq g} \xi_h^{\bar{\pi}}(s)\{(P_{s,\bar{\pi}(s)} - \widetilde{P}'_{s,\bar{\pi}(s)})\bar{V} + P_{s,\bar{\pi}(s)}(V^{\bar{\pi}} - \bar{V}) + c(s,\bar{\pi}(s)) - \hat{c}(s,\bar{\pi}(s)) - b_{s,\bar{\pi}(s)}(\bar{V})\} \\ &= \sum_{s,s\neq g} \xi_h^{\bar{\pi}}(s)\{(P_{s,\bar{\pi}(s)} - \widetilde{P}'_{s,\bar{\pi}(s)})\bar{V} + c(s,\bar{\pi}(s)) - \hat{c}(s,\bar{\pi}(s)) - b_{s,\bar{\pi}(s)}(\bar{V})\} \\ &\quad + \sum_{s,s\neq g} \xi_{h+1}^{\bar{\pi}}(s)(V^{\bar{\pi}} - \bar{V}),\end{aligned}$$

where the first inequality uses the Bellman equation for policy $\bar{\pi}$, which follows from Lemma 1.1. By recursion, we have

$$\begin{aligned}V^{\bar{\pi}} - \bar{V} = &\sum_{h=0}^{H}\sum_{s,s\neq g} \xi_h^{\bar{\pi}}(s)\{(P_{s,\bar{\pi}(s)} - \widetilde{P}'_{s,\bar{\pi}(s)})\bar{V} + c(s,\bar{\pi}(s)) - \hat{c}(s,\bar{\pi}(s)) - b_{s,\bar{\pi}(s)}(\bar{V})\} \\ &+ \sum_{s,s\neq g} \xi_H^{\bar{\pi}}(s)(V^{\bar{\pi}} - \bar{V}).\end{aligned}$$

From Lemma 7.5, we have that when$n \geq \max\{\frac{4B_\star - 2c_{\min}}{c_{\min} d_{max}}, \frac{26^2 \times 2S\iota(\bar{T}^\star)^2}{d_m}, \frac{10^6(\sqrt{\widetilde{B}}+1)^4 S\iota \bar{T}^\star \widetilde{T}}{B_\star(\sqrt{B_\star}+1)^2 d_m}, O(T_{\max}^2 \log(SA/\delta)/d_m^2)\}$, $\bar{\pi}$ is a proper policy. Thus $\|V^{\bar{\pi}}\|_\infty < +\infty$, and for any state $s, s \neq g$, $\lim_{H\to\infty} \xi_H^{\bar{\pi}}(s) = 0$. Let H goes to infinity, we can prove the lemma. □

8 CRUDE OPTIMIZATION BOUND

In this section, we give a rough bound for $\bar{V} - V^\star$.

Theorem 8.1. *Denote $d_m := \min\{\sum_{h=1}^\infty \xi_h^\mu(s,a) \,:\, s.t. \sum_{h=1}^\infty \xi_h^\mu(s,a) > 0\}$ and $T_{\max} = \max_i T_i$. Let $T_s^\star$ be the expected time to hit g when starting from s with the optimal policy and denote $\bar{T}^\pi = \max_s T_s^\pi$. Then when $n \geq \max\{\frac{26^2 \times 2S\iota(\bar{T}^\star)^2}{d_m}, \frac{10^6(\sqrt{\widetilde{B}}+1)^4 S\iota \bar{T}^\star \widetilde{T}}{B_\star(\sqrt{B_\star}+1)^2 d_m}, O(T_{\max}^2 \log(SA/\delta)/d_m^2)\}$, we have with probability $1-\delta$,*

$$\left\|\bar{V} - V^\star\right\|_\infty \leq 30\sqrt{\frac{\bar{T}^\star B_\star \iota}{n d_m}}(\sqrt{B_\star}+1) \tag{35}$$

Proof. From Lemma 7.7, we have (by choosing $\xi_0 = \mathbf{1}[s_0 = \bar{s}]$)

$$|\bar{V}(\bar{s}) - V^\star(\bar{s})| \le \sum_{h=0}^{\infty} \sum_{\substack{s \\ s \ne g}} \xi^\star_{h,\bar{s}}(s) \{ (\widetilde{P}_{s,\pi^\star(s)} - P_{s,\pi^\star(s)})\bar{V} + \widehat{c}(s,\pi^\star(s)) - c(s,\pi^\star(s)) + b_{s,\pi^\star(s)}(\bar{V}) \}$$

For the first term, we can bound it by Lemma 7.3

$$|(\widetilde{P}_{s,a} - P_{s,a})\bar{V}| \le \frac{\widetilde{B}}{n(s,a)} + \frac{16\widetilde{B}\iota}{3n(s,a)} + 2\sqrt{\frac{\mathrm{Var}(\widetilde{P}', \bar{V})\iota}{n(s,a)}} + 6\sqrt{\frac{S\iota}{n(s,a)}} \left\| V - V^\star \right\|_\infty .$$

Conditioned on the event $\mathcal{E}_5$, we have

$$|\widehat{c}(s,a) - c(s,a)| \le \sqrt{\frac{2\widehat{c}(s,a)\iota}{n(s,a)}} + \frac{7\iota}{3n(s,a)}$$

Combine the above inequalities together, we can get

$$\begin{aligned}
|\bar{V}(\bar{s}) - V^\star(\bar{s})| &\le \sum_{h=0}^{\infty} \sum_{\substack{s \\ s \ne g}} \xi^\star_{h,\bar{s}}(s) \{ \frac{2\widetilde{B}}{n(s,\pi^\star(s))} + \frac{32\widetilde{B}\iota}{3n(s,\pi^\star(s))} + 2\sqrt{\frac{\mathrm{Var}(\widetilde{P}', \bar{V})\iota}{n(s,\pi^\star(s))}} + 6\sqrt{\frac{S\iota}{n(s,\pi^\star(s))}} \left\| V - V^\star \right\|_\infty \\
&\quad + 2\sqrt{\frac{2\widehat{c}(s,\pi^\star(s))\iota}{n(s,\pi^\star(s))}} + \frac{14\iota}{3n(s,\pi^\star(s))} + \max\{2\sqrt{\frac{\mathrm{Var}(\widetilde{P}, V)\iota}{n(s,\pi^\star(s))}}, 4\frac{\widetilde{B}\iota}{n(s,\pi^\star(s))}\} \\
&\quad + 180\sqrt{\frac{3\widetilde{T}\widetilde{B}S}{2n(s,\pi^\star(s))n_{\min}}}(\sqrt{\widetilde{B}} + 1)\iota \} \\
&\overset{\text{(i)}}{\le} \sum_{h=0}^{\infty} \sum_{\substack{s \\ s \ne g}} \xi^\star_{h,\bar{s}}(s) \{ \frac{2\widetilde{B}}{n(s,\pi^\star(s))} + \frac{44\widetilde{B}\iota}{3n(s,\pi^\star(s))} + 4\sqrt{\frac{\mathrm{Var}(\widetilde{P}', \bar{V})\iota}{n(s,\pi^\star(s))}} + 6\sqrt{\frac{S\iota}{n(s,\pi^\star(s))}} \left\| \bar{V} - V^\star \right\|_\infty \\
&\quad + 2\sqrt{\frac{2\widehat{c}(s,\pi^\star(s))\iota}{n(s,\pi^\star(s))}} + \frac{14\iota}{3n(s,\pi^\star(s))} + 180\sqrt{\frac{3\widetilde{T}\widetilde{B}S}{2n(s,\pi^\star(s))n_{\min}}}(\sqrt{\widetilde{B}} + 1)\iota \},
\end{aligned}$$

where (i) uses the inequality that $\max\{a,b\} \le a + b$. For notation simplicity, we define

$$b_0(s,a) := 180\sqrt{\frac{3\widetilde{T}\widetilde{B}S}{2n(s,a)n_{\min}}}(\sqrt{\widetilde{B}} + 1)\iota.$$

First, we bound the variance term

$$\begin{aligned}
\mathrm{Var}(\widetilde{P}'_{s,a}, \bar{V}) &\overset{\text{(i)}}{\le} \mathrm{Var}(\widehat{P}_{s,a}, \bar{V}) + \frac{2\left\|\bar{V}\right\|_\infty^2}{n_{max}+1} \\
&\overset{\text{(ii)}}{\le} \frac{3}{2}\mathrm{Var}(P_{s,a}, \bar{V}) + \frac{2\left\|\bar{V}\right\|_\infty^2 S\iota}{n(s,a)} + \frac{2\left\|\bar{V}\right\|_\infty^2}{n_{max}+1} \\
&\overset{\text{(iii)}}{\le} 3\mathrm{Var}(P_{s,a}, \bar{V} - V^\star) + 3\mathrm{Var}(P_{s,a}, V^\star) + \frac{2\left\|\bar{V}\right\|_\infty^2 S(\iota+1)}{n(s,a)} \\
&\overset{\text{(iv)}}{\le} 3\left\|\bar{V} - V^\star\right\|_\infty^2 + 3\mathrm{Var}(P_{s,a}, V^\star) + \frac{2\widetilde{B}^2 S(\iota+1)}{n(s,a)},
\end{aligned}$$

where (i) follows from Lemma 7.2. (ii) uses Lemma 10.1. (iii) uses the fact that $\mathrm{Var}(X+Y) \le 2\mathrm{Var}(X) + 2\mathrm{Var}(Y)$. (iv) uses the fact that $\left\|\bar{V}\right\|_\infty \le \widetilde{B}$. Then we can have

$$|\bar{V}(\bar{s}) - V^\star(\bar{s})| \le \sum_{h=0}^{\infty} \sum_{\substack{s \\ s \ne g}} \xi^\star_{h,\bar{s}}(s) \{ \frac{2\widetilde{B}}{n(s,\pi^\star(s))} + \frac{44\widetilde{B}\iota}{3n(s,\pi^\star(s))} + 4\sqrt{\frac{3\iota}{n(s,\pi^\star(s))}} \left\| \bar{V} - V^\star \right\|_\infty$$

$$
\begin{aligned}
&+ 4\sqrt{\frac{3\mathrm{Var}(P_{s,\pi^\star(s)},V^\star)\iota}{n(s,\pi^\star(s))}} + 4\sqrt{\frac{2S\iota(\iota+1)}{n^2(s,\pi^\star(s))}}\widetilde{B}\\
&+ 6\sqrt{\frac{S\iota}{n(s,\pi^\star(s))}}\left\|\bar{V}-V^\star\right\|_\infty + 2\sqrt{\frac{2\widehat{c}(s,\pi^\star(s))\iota}{n(s,\pi^\star(s))}} + \frac{14\iota}{3n(s,\pi^\star(s))} + b_0(s,a)\}\\
\overset{\text{(i)}}{\leq}& \sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}}\xi^\star_{h,\bar{s}}(s)\{\frac{27\max\{\widetilde{B},1\}\sqrt{S}\iota}{n(s,\pi^\star(s))} + 13\sqrt{\frac{S\iota}{n(s,\pi^\star(s))}}\left\|\bar{V}-V^\star\right\|_\infty\\
&+ 4\sqrt{\frac{3\mathrm{Var}(P_{s,\pi^\star(s)},V^\star)\iota}{n(s,\pi^\star(s))}} + 2\sqrt{\frac{2\widehat{c}(s,\pi^\star(s))\iota}{n(s,\pi^\star(s))}} + b_0(s,a)\}\\
\overset{\text{(ii)}}{\leq}& \sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}}\xi^\star_{h,\bar{s}}(s)\{\frac{31\max\{\widetilde{B},1\}\sqrt{S}\iota}{n(s,\pi^\star(s))} + 13\sqrt{\frac{S\iota}{n(s,\pi^\star(s))}}\left\|\bar{V}-V^\star\right\|_\infty\\
&+ 4\sqrt{\frac{3\mathrm{Var}(P_{s,\pi^\star(s)},V^\star)\iota}{n(s,\pi^\star(s))}} + 2\sqrt{\frac{3c(s,\pi^\star(s))\iota}{n(s,\pi^\star(s))}} + b_0(s,a)\},
\end{aligned}
$$

where (i) uses the assumption $\iota \geq 1$ and that $S \geq 1$. (ii) holds because of Lemma 10.2.

Then we have

$$
\begin{aligned}
|\bar{V}(\bar{s}) - V^\star(\bar{s})| \overset{\text{(i)}}{\leq}& \sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}}\xi^\star_{h,\bar{s}}(s)\{\frac{62\max\{\widetilde{B},1\}\sqrt{S}\iota}{nd_m} + 13\sqrt{2}\sqrt{\frac{S\iota}{nd_m}}\left\|\bar{V}-V^\star\right\|_\infty\\
&+ 4\sqrt{\frac{6\mathrm{Var}(P_{s,\pi^\star(s)},V^\star)\iota}{nd_m}} + 2\sqrt{\frac{6c(s,\pi^\star(s))\iota}{nd_m}} + \bar{b}_0\}\\
\overset{\text{(ii)}}{\leq}& \frac{62T^\star_{\bar{s}}\max\{\widetilde{B},1\}\sqrt{S}\iota}{nd_m} + 13\sqrt{\frac{2S\iota}{nd_m}}\left\|\bar{V}-V^\star\right\|_\infty T^\star_{\bar{s}} + T^\star_{\bar{s}}\bar{b}_0\\
&+ \sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}}\xi^\star_{h,\bar{s}}(s)\{4\sqrt{\frac{6\mathrm{Var}(P_{s,\pi^\star(s)},V^\star)\iota}{nd_m}} + 2\sqrt{\frac{6c(s,\pi^\star(s))\iota}{nd_m}}\},
\end{aligned}
$$

where $\bar{b}_0 := 180\sqrt{\frac{6\widetilde{T}\widetilde{B}S}{n^2d_m^2}}(\sqrt{\widetilde{B}}+1)\iota$. (i) holds with probability $1-\delta$ because of Lemma 12.6. For any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have $n(s,a) \geq \frac{1}{2}nd(s,a) \geq \frac{1}{2}nd_m$. In particular, $n_{\min} \geq \frac{1}{2}nd_m$. Since

$$
\begin{aligned}
\sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}}\xi^\star_{h,\bar{s}}(s)\{4\sqrt{\frac{6\mathrm{Var}(P_{s,\pi^\star(s)},V^\star)\iota}{nd_m}}\} \overset{\text{(i)}}{\leq}& 4\sqrt{\sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}}\xi^\star_{h,\bar{s}}(s)}\sqrt{\frac{6\sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}}\xi^\star_{h,\bar{s}}(s)\mathrm{Var}(P_{s,\pi^\star(s)},V^\star)\iota}{nd_m}}\\
\overset{\text{(ii)}}{\leq}& 4\sqrt{T^\star_{\bar{s}}}\sqrt{\frac{12\iota}{nd_m}}\left\|V^\star\right\|_\infty,
\end{aligned}
$$

where (i) uses the Cauchy-Schwartz inequality. (ii) uses the result in Lemma 4.3 and Lemma 10.1. Similarly, we have

$$
\begin{aligned}
\sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}}\xi^\star_{h,\bar{s}}(s)\{2\sqrt{\frac{6c(s,\pi^\star(s))\iota}{nd_m}}\}\leq& 2\sqrt{\sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}}\xi^\star_{h,\bar{s}}(s)}\sqrt{\frac{6\sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}}\xi^\star_{h,\bar{s}}(s)c(s,\pi^\star(s))\iota}{nd_m}}\\
\leq& 2\sqrt{T^\star_{\bar{s}}}\sqrt{\frac{6\left\|V^\star\right\|_\infty\iota}{nd_m}}.
\end{aligned}
$$

Combine the above together, we get

$$
|\bar{V}(\bar{s}) - V^\star(\bar{s})| \leq \frac{62T^\star_{\bar{s}}\max\{\widetilde{B},1\}\sqrt{S}\iota}{nd_m} + 13\sqrt{\frac{2S\iota}{nd_m}}\left\|\bar{V}-V^\star\right\|_\infty T^\star_{\bar{s}} + T^\star_{\bar{s}}\bar{b}_0
$$

$$
\begin{aligned}
&+ 4\sqrt{T^\star_{\bar{s}}}\sqrt{\frac{12\iota}{nd_m}}\left\|V^\star\right\|_\infty + 2\sqrt{T^\star_{\bar{s}}}\sqrt{\frac{6\left\|V^\star\right\|_\infty\iota}{nd_m}}\\
\leq &\frac{62\bar{T}^\star\max\{\widetilde{B},1\}\sqrt{S}\iota}{nd_m} + 13\sqrt{\frac{2S\iota}{nd_m}}\left\|\bar{V}-V^\star\right\|_\infty\bar{T}^\star + \bar{T}^\star\bar{b}_0\\
&+ 8\sqrt{\bar{T}^\star}\sqrt{\frac{3\iota}{nd_m}}\left\|V^\star\right\|_\infty + 2\sqrt{\bar{T}^\star}\sqrt{\frac{6\left\|V^\star\right\|_\infty\iota}{nd_m}},
\end{aligned}
$$

It implies that

$$
\begin{aligned}
(1-13\sqrt{\frac{2S\iota}{nd_m}}\bar{T}^\star)\left\|\bar{V}-V^\star\right\|_\infty \leq &\frac{62\max\{\widetilde{B},1\}\sqrt{S}\iota\bar{T}^\star}{nd_m} + \bar{T}^\star\bar{b}_0\\
&+ 14\sqrt{\frac{\bar{T}^\star B_\star\iota}{nd_m}}(\sqrt{B_\star}+1).
\end{aligned}
\tag{36}
$$

Since $n \geq \frac{26^2\times 2S\iota(\bar{T}^\star)^2}{d_m}$,

$$
\begin{aligned}
\left\|\bar{V}(s)-V^\star(s)\right\|_\infty \leq &\frac{124\max\{\widetilde{B},1\}\sqrt{S}\iota\bar{T}^\star}{nd_m} + 2\bar{T}^\star\bar{b}_0\\
&+ 21\sqrt{\frac{\bar{T}^\star B_\star\iota}{nd_m}}(\sqrt{B_\star}+1)\\
\leq &\frac{124\max\{\widetilde{B},1\}\sqrt{S}\iota\bar{T}^\star}{nd_m} + 360\bar{T}^\star\sqrt{\frac{6\widetilde{T}\widetilde{B}S}{n^2d_m^2}}(\sqrt{\widetilde{B}}+1)\iota\\
&+ 28\sqrt{\frac{\bar{T}^\star B_\star\iota}{nd_m}}(\sqrt{B_\star}+1)\\
\leq &720\bar{T}^\star\sqrt{\frac{6\widetilde{T}S}{n^2d_m^2}}(\sqrt{\widetilde{B}}+1)^2\iota + 28\sqrt{\frac{\bar{T}^\star B_\star\iota}{nd_m}}(\sqrt{B_\star}+1)
\end{aligned}
$$

When $n \geq \frac{10^6(\sqrt{\widetilde{B}}+1)^4 S\iota\bar{T}^\star\widetilde{T}}{B_\star(\sqrt{B_\star}+1)^2 d_m}$, we have

$$
\left\|\bar{V}(s)-V^\star(s)\right\|_\infty \leq 30\sqrt{\frac{\bar{T}^\star B_\star\iota}{nd_m}}(\sqrt{B_\star}+1)
$$

□

9 PROOF OF THEOREM ??

In this section, we provide the proof of Theorem **??**. However, before that, we first present a lemma that guarantees pessimism.

Lemma 9.1. *When $n \geq \max\{\frac{26^2\times 2S\iota(\bar{T}^\star)^2}{d_m}, \frac{10^6(\sqrt{\widetilde{B}}+1)^4 S\iota\bar{T}^\star\widetilde{T}}{B_\star(\sqrt{B_\star}+1)^2 d_m}, O(T_{\max}^2\log(SA/\delta)/d_m^2)\}$ (where $T_{\max} = \max_i T_i$), with probability at least $1-\delta$, we have that for any state action pair (s,a),*

$$
c(s,a) - \widehat{c}(s,a) + (P_{s,a} - \widetilde{P}'_{s,a})\bar{V} - b_{s,a}(\bar{V}) \leq 0
$$

Proof. Applying the result in Theorem 8.1, we can get

$$
6\sqrt{\frac{S\iota}{n(s,a)}}\left\|\bar{V}-V^\star\right\|_\infty \leq 180\sqrt{\frac{\bar{T}^\star B_\star S}{n(s,a)nd_m}}(\sqrt{B_\star}+1)\iota.
$$

Combine the above inequality with Lemma 7.3 implies that

$$(P_{s,a}-\widetilde{P}'_{s,a})\bar{V} \leq \frac{\widetilde{B}}{n(s,a)} + \frac{16\widetilde{B}\iota}{3n(s,a)} + 2\sqrt{\frac{\mathrm{Var}(\widetilde{P}',\bar{V})\iota}{n(s,a)}} + 180\sqrt{\frac{\bar{T}^\star B_\star S}{n(s,a)nd_m}}(\sqrt{B_\star}+1)\iota.$$

Conditioned on the event $\mathcal{E}_5$, then we have

$$c(s,a)-\widehat{c}(s,a)+(P_{s,a}-\widetilde{P}'_{s,a})\bar{V}-b_{s,a}(\bar{V}) \leq 180\sqrt{\frac{\bar{T}^\star B_\star S}{n(s,a)nd_m}}(\sqrt{B_\star}+1)\iota - 180\sqrt{\frac{3\widetilde{T}\widetilde{B}S}{2n(s,a)n_{\min}}}(\sqrt{\widetilde{B}}+1)\iota.$$

Applying the Chernoff bound given in Lemma 12.6, we have that with probability $1-\delta$, $n(s,a) < \frac{3}{2}nd^\mu(s,a)$ for any state action pair (s,a). Thus $n_{\min} := \min_{s,a,n(s,a)>0} n(s,a) < \frac{3}{2}n(\min_{n(s,a)>0} d^\mu(s,a))$. For any $(s,a) \in \mathcal{S}\times\mathcal{A}$, if we have $d^\mu(s,a)>0$, by the Lemma 12.6 we have that with probability $1-\delta$, $n(s,a) > \frac{nd^\mu(s,a)}{2} > 0$, which implies

$$\{(s,a)\in\mathcal{S}\times\mathcal{A} : d^\mu(s,a)>0\} \subseteq \{(s,a)\in\mathcal{S}\times\mathcal{A} : n(s,a)>0\}.$$

Then we can get $\min_{n(s,a)>0} d^\mu(s,a) \leq \min_{d^\mu(s,a)>0} d^\mu(s,a) = d_m$ and thus $n_{\min} \leq \frac{3}{2}nd_m$. Because $\bar{T}^\star \leq \widetilde{T}$ and $B_\star \leq \widetilde{B}$, we can prove the result in the Lemma. □

Now we are ready to introduce the final proof.

Theorem 9.2. *Given Assumption* **??** *and Assumption* **??***. When $n \geq n_0$, the suboptimality bound of the output policy $\bar{\pi}$ can be upper bounded as follows with probability $1-\delta$ (where $\iota = O(\log(SA/\delta))$),*

$$V^{\bar{\pi}}(s_{\text{init}}) - V^\star(s_{\text{init}}) \leq 8\sum_{s,a,s\neq g} d^\star(s,a)\sqrt{\frac{3\mathrm{Var}_{P_{s,a}}[V^\star+c]\iota}{n\cdot d^\mu(s,a)}} + \widetilde{O}(\frac{1}{n}), \tag{37}$$

where we define $n_0 := n \geq \max\{\frac{4B_\star-2c_{min}}{c_{min}d_{max}}, \frac{26^2\times 2S\iota(\bar{T}^\star)^2(\sqrt{B_\star}+1)^2}{d_m}, \frac{10^6(\sqrt{\widetilde{B}}+1)^4 S\iota\bar{T}^\star\widetilde{T}}{B_\star(\sqrt{B_\star}+1)^2 d_m}, O(T_{\max}^2\log(SA/\delta)/d_m^2)\}$.

Proof.

$$V^{\bar{\pi}}(s) - V^\star(s) = (V^{\bar{\pi}}(s)-\bar{V}(s)) + (\bar{V}(s)-V^\star(s)). \tag{38}$$

From Lemma 7.8, we have with probability $1-\delta$ and when $n \geq \max\{\frac{4B_\star-2c_{\min}}{c_{\min}d_{max}}, \frac{26^2\times 2S\iota(\bar{T}^\star)^2}{d_m}, \frac{10^6(\sqrt{\widetilde{B}}+1)^4 S\iota\bar{T}^\star\widetilde{T}}{B_\star(\sqrt{B_\star}+1)^2 d_m}\}$,

$$V^{\bar{\pi}} - \bar{V} = \sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}} \xi_h^{\bar{\pi}}(s)\{(P_{s,\bar{\pi}(s)}-\widetilde{P}'_{s,\bar{\pi}(s)})\bar{V} + c(s,\bar{\pi}(s)) - \hat{c}(s,\bar{\pi}(s)) - b_{s,\bar{\pi}(s)}(\bar{V})\}. \tag{39}$$

Next by Lemma 9.1 with probability $1-\delta$, we have

$$(P_{s,\bar{\pi}(s)}-\widetilde{P}'_{s,\bar{\pi}(s)})\bar{V} + c(s,\bar{\pi}(s)) - \hat{c}(s,\bar{\pi}(s)) - b_{s,\bar{\pi}(s)}(\bar{V}) \leq 0. \tag{40}$$

Thus combine (39) and (40) we have $V^{\bar{\pi}} - \bar{V} \leq 0$ by pessimism. For the term $\bar{V} - V^\star$, we apply Lemma 7.7 to obtain:

$$\bar{V} - V^\star \leq \sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}} \xi_h^\star(s)\{(\widetilde{P}'_{s,\pi^\star(s)} - P_{s,\pi^\star(s)})\bar{V} + \hat{c}(s,\pi^\star(s)) - c(s,\pi^\star(s)) + b_{s,\pi^\star(s)}(\bar{V})\}. \tag{41}$$

From Lemma 7.3, we have

$$|(P_{s,a}-\widetilde{P}'_{s,a})\bar{V}| \leq \frac{\widetilde{B}}{n(s,a)} + \frac{16\widetilde{B}\iota}{3n(s,a)} + 2\sqrt{\frac{\mathrm{Var}(\widetilde{P}'_{s,a},\bar{V})\iota}{n(s,a)}} + 6\sqrt{\frac{S\iota}{n(s,a)}}\left\|\bar{V}-V^\star\right\|_\infty \tag{42}$$

Conditioned on the event $\mathcal{E}_5$, we have

$$|\widehat{c}(s,a) - c(s,a)| \leq \sqrt{\frac{2\widehat{c}(s,a)\iota}{n(s,a)}} + \frac{7\iota}{3n(s,a)}$$

Combine the above inequalities together, we can get

$$\begin{aligned}
&(\widetilde{P}'_{s,a} - P_{s,a})\bar{V} + \widehat{c}(s,a) - c(s,a) + b_{s,a}(\bar{V}) \\
&\leq 2\sqrt{\frac{2\widehat{c}(s,a)\iota}{n(s,a)}} + \frac{14\iota}{3n(s,a)} + \frac{2\widetilde{B}}{n(s,a)} + \frac{32\widetilde{B}\iota}{3n(s,a)} + 4\sqrt{\frac{\mathrm{Var}(\widetilde{P}'_{s,a},\bar{V})\iota}{n(s,a)}} \\
&\quad + \frac{4\widetilde{B}\iota}{n(s,a)} + 6\sqrt{\frac{S\iota}{n(s,a)}}\left\|\bar{V} - V^\star\right\|_\infty + 180\sqrt{\frac{3\widetilde{T}\widetilde{B}S}{2n(s,a)n_{\min}}}(\sqrt{\widetilde{B}}+1)\iota \\
&\leq 2\sqrt{\frac{2\widehat{c}(s,a)\iota}{n(s,a)}} + 4\sqrt{\frac{\mathrm{Var}(\widetilde{P}'_{s,a},\bar{V})\iota}{n(s,a)}} + \widetilde{O}(\frac{(\widetilde{B}+1)\iota}{n(s,a)}) + 6\sqrt{\frac{S\iota}{n(s,a)}}\left\|\bar{V} - V^\star\right\|_\infty + 180\sqrt{\frac{3\widetilde{T}\widetilde{B}S}{2n(s,a)n_{\min}}}(\sqrt{\widetilde{B}}+1)\iota \\
&\leq 2\sqrt{\frac{3c(s,a)\iota}{n(s,a)}} + 4\sqrt{\frac{3\mathrm{Var}(P_{s,a},V^\star)\iota}{n(s,a)}} + \widetilde{O}(\frac{(\widetilde{B}+1)\sqrt{S}\iota}{n(s,a)}) \\
&\quad + \widetilde{O}(\sqrt{\frac{S\iota}{n(s,a)}}\left\|\bar{V} - V^\star\right\|_\infty) + 180\sqrt{\frac{3\widetilde{T}\widetilde{B}S}{2n(s,a)n_{\min}}}(\sqrt{\widetilde{B}}+1)\iota
\end{aligned}$$

Plug the above into (41), then we bound all the terms one by one. First,

$$\sum_{h=1}^{\infty}\sum_{\substack{s,a\\s\neq g}}\xi_h^\star(s,a)\left(2\sqrt{\frac{3c(s,a)\iota}{n(s,a)}}\right) \leq \sum_{h=1}^{\infty}\sum_{\substack{s,a\\s\neq g}}\xi_h^\star(s,a)\left[2\sqrt{\frac{6c(s,a)\iota}{n\sum_{h=1}^{\infty}\xi_h^\mu(s,a)}}\right] = 2\sum_{\substack{s,a\\s\neq g}}d^\star(s,a)\sqrt{\frac{6c(s,a)\iota}{nd^\mu(s,a)}}. \tag{43}$$

For the second term, first we have

$$\begin{aligned}
\sum_{h=1}^{\infty}\sum_{\substack{s,a\\s\neq g}}\xi_h^\star(s,a)\left(4\sqrt{\frac{3\mathrm{Var}(P_{s,a},V^\star)\iota}{n(s,a)}}\right) &\leq \sum_{h=1}^{\infty}\sum_{\substack{s,a\\s\neq g}}\xi_h^\star(s,a)\left[4\sqrt{\frac{6\mathrm{Var}(P_{s,a},V^\star)\iota}{n\sum_{h=1}^{\infty}\xi_h^\mu(s,a)}}\right] \\
&= 4\sum_{\substack{s,a\\s\neq g}}d^\star(s,a)\sqrt{\frac{6\mathrm{Var}(P_{s,a},V^\star)\iota}{nd^\mu(s,a)}}
\end{aligned} \tag{44}$$

From Chernoff bound given in Lemma 12.6, we have with probability $1-\delta$, we have

$$\widetilde{O}(\sum_{h=1}^{\infty}\sum_{\substack{s,a\\s\neq g}}\xi_h^\star(s,a)\frac{(\widetilde{B}+1)\sqrt{S}\iota}{n(s,a)}) \leq \widetilde{O}(\sum_{\substack{s,a\\s\neq g}}\frac{d^\pi(s,a)}{d^\mu(s,a)}\cdot\frac{(\widetilde{B}+1)\sqrt{S}\iota}{n}). \tag{45}$$

Similarly, we have

$$\begin{aligned}
\widetilde{O}(\sum_{h=1}^{\infty}\sum_{\substack{s,a\\s\neq g}}\xi_h^\star(s,a)\sqrt{\frac{S\iota}{n(s,a)}}\left\|\bar{V} - V^\star\right\|_\infty) &\leq \widetilde{O}(\sum_{\substack{s,a\\s\neq g}}d^\star(s,a)\sqrt{\frac{S\iota}{nd^\mu(s,a)}}\left\|\bar{V} - V^\star\right\|_\infty) \\
&\overset{\text{(i)}}{\leq} \widetilde{O}(\sum_{\substack{s,a\\s\neq g}}d^\star(s,a)\sqrt{\frac{\bar{T}^\star B_\star S}{n^2d^\mu(s,a)d_m}}(\sqrt{B_\star}+1)\iota),
\end{aligned} \tag{46}$$

where (i) uses the Crude optimization bound given in Theorem 8.1. For the last term, we have

$$\widetilde{O}(\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^\star(s,a)\sqrt{\frac{\widetilde{T}\widetilde{B}S}{n(s,a)n_{\min}}}(\sqrt{\widetilde{B}}+1)\iota)\leq\widetilde{O}(\sum_{\substack{s,a\\ s\neq g}}d^\star(s,a)\sqrt{\frac{\widetilde{T}\widetilde{B}S}{n^2d^\mu(s,a)d_m}}(\sqrt{\widetilde{B}}+1)\iota),\tag{47}$$

where the inequality comes from Lemma 12.6 again. Combine the inequalities (43), (44), (45), (46) and (45) together, we have

$$\begin{aligned}\bar{V}(s_{\text{init}})-V^\star(s_{\text{init}})&\overset{\text{(i)}}{\leq}2\sum_{\substack{s,a\\ s\neq g}}d^\star(s,a)\sqrt{\frac{6c(s,a)\iota}{nd^\mu(s,a)}}+4\sum_{\substack{s,a\\ s\neq g}}d^\star(s,a)\sqrt{\frac{6\mathrm{Var}(P_{s,a},V^\star)\iota}{nd^\mu(s,a)}}\\&\quad+\widetilde{O}(\sum_{\substack{s,a\\ s\neq g}}d^\star(s,a)\sqrt{\frac{\widetilde{T}\widetilde{B}S}{n^2d^\mu(s,a)d_m}}(\sqrt{\widetilde{B}}+1)\iota)+\widetilde{O}(\sum_{\substack{s,a\\ s\neq g}}\frac{d^\pi(s,a)}{d^\mu(s,a)}\cdot\frac{(\widetilde{B}+1)\sqrt{S}\iota}{n})\\&\overset{\text{(ii)}}{\leq}2\sum_{\substack{s,a\\ s\neq g}}d^\star(s,a)\sqrt{\frac{6c(s,a)\iota}{nd^\mu(s,a)}}+4\sum_{\substack{s,a\\ s\neq g}}d^\star(s,a)\sqrt{\frac{6\mathrm{Var}(P_{s,a},V^\star)\iota}{nd^\mu(s,a)}}\\&\quad+\widetilde{O}(\sum_{\substack{s,a\\ s\neq g}}d^\star(s,a)\sqrt{\frac{\widetilde{T}S}{n^2d^\mu(s,a)d_m}}(\widetilde{B}+1)\iota)\\&\leq 8\sum_{\substack{s,a\\ s\neq g}}d^\star(s,a)\sqrt{\frac{3\mathrm{Var}(P_{s,a},V^\star+c)\iota}{nd^\mu(s,a)}}+\widetilde{O}(\sum_{\substack{s,a\\ s\neq g}}d^\star(s,a)\sqrt{\frac{\widetilde{T}S}{n^2d^\mu(s,a)d_m}}(\widetilde{B}+1)\iota),\end{aligned}\tag{48}$$

where the inequality (i) uses the assumption that $\bar{T}^\star\leq\widetilde{T}$ and $B_\star\leq\widetilde{B}$. (ii) uses the fact that $\frac{d^\pi(s,a)}{d^\mu(s,a)}\leq\frac{d^\pi(s,a)}{\sqrt{d^\mu(s,a)d_m}}$ and that $\widetilde{B}+\sqrt{\widetilde{B}}\leq 2(\widetilde{B}+1)$. The last inequality comes from $\sqrt{a}+\sqrt{b}\leq\sqrt{2a+2b}$. □

Based on Theorem 8.1, we can also get the Proposition below.

Proposition 9.3. *When $n\geq n_0$ (where n_0 is defined the same as Theorem 9.2), then the suboptimality incurred by the limit of the output policy $\bar{\pi}$ can be upper bounded as (with probability $1-\delta$)*

$$V^{\bar{\pi}}(s_{init})-V^\star(s_{init})\leq 8(\sqrt{\sum_{\substack{s,a\\ s\neq g}}\frac{d^\star(s,a)}{d^\mu(s,a)}\cdot\frac{6\iota}{n}})\cdot(B_\star+1)+\widetilde{O}(\frac{1}{n}).\tag{49}$$

Proof. By Theorem **??**,

$$\begin{aligned}V^{\bar{\pi}}(s_{\text{init}})-V^\star(s_{\text{init}})&\leq 8\sum_{s,a,s\neq g}d^\star(s,a)\sqrt{\frac{3\mathrm{Var}_{P_{s,a}}[V^\star+c]\iota}{n\cdot d^\mu(s,a)}}+\widetilde{O}(\frac{1}{n})\\&\overset{\text{(i)}}{\leq}8\sqrt{\sum_{\substack{s,a\\ s\neq g}}\frac{d^\star(s,a)}{n\cdot d^\mu(s,a)}}\cdot\sqrt{3\sum_{\substack{s,a\\ s\neq g}}d^\star(s,a)\mathrm{Var}_{P_{s,a}}[V^\star+c]\iota},\end{aligned}\tag{50}$$

where (i) uses the Cauchy-Schwartz inequality. Since

$$\begin{aligned}\sum_{\substack{s,a\\ s\neq g}}d^\star(s,a)\mathrm{Var}_{P_{s,a}}[V^\star+c]&=\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^\star(s,a)(\mathrm{Var}(P_{s,a},V^\star)+c(s,a))\\&\overset{\text{(i)}}{\leq}2\left\|V^\star\right\|_\infty^2+V^\star(s_0)\end{aligned}$$

$$\leq 2B_\star^2 + B_\star, \tag{51}$$

where (i) comes from Lemma 10.1 and the definition of value function. Plug (51) into (50), we obtain

$$\begin{aligned} V^{\bar\pi}(s_{\text{init}}) - V^\star(s_{\text{init}}) &\leq 8\sqrt{\sum_{\substack{s,a\\ s\neq g}} \frac{d^\star(s,a)}{n\cdot d^\mu(s,a)}} \cdot \sqrt{6(B_\star^2+B_\star)\iota} + \widetilde{O}(\frac{1}{n}) \\ &\leq 8\sqrt{\sum_{\substack{s,a\\ s\neq g}} \frac{6d^\star(s,a)\cdot\iota}{d^\mu(s,a)\cdot n}} \cdot (B_\star+1) + \widetilde{O}(\frac{1}{n}), \end{aligned} \tag{52}$$

which completes the proof. □

10 PROPERTIES OF TRANSITION MATRIX ESTIMATE $\widehat{P}$

Lemma 10.1. *For any $V(\cdot)\in\mathbb{R}^S$ satisfying $V(g)=0$, i.e. (13), and suppose event $\mathcal{E}_1$ holds. we have*

$$\begin{aligned} \mathrm{Var}(\widehat{P}_{s,a},V) &\leq \frac{3}{2}\mathrm{Var}(P_{s,a},V) + \frac{2\left\|V\right\|_\infty^2 S\iota}{n(s,a)} \\ \mathrm{Var}(P_{s,a},V) &\leq 2\mathrm{Var}(\widehat{P}_{s,a},V) + \frac{4\left\|V\right\|_\infty^2 S\iota}{n(s,a)} \end{aligned} \tag{53}$$

Proof. From the event $\mathcal{E}_1$, we have

$$|\widehat{P}(s'|s,a) - P(s'|s,a)| \leq \sqrt{\frac{2P(s'|s,a)\iota}{n(s,a)}} + \frac{2\iota}{3n(s,a)} \leq \frac{P(s'|s,a)}{2} + \frac{5\iota}{3n(s,a)},$$

where the second inequality uses $\sqrt{ab}\leq\frac{a+b}{2}$ with $a=\frac{2\iota}{n(s,a)}$, $b=P(s'|s,a)$. Thus we have

$$\begin{aligned} \widehat{P}(s'|s,a) &\leq \frac{3P(s'|s,a)}{2} + \frac{5\iota}{3n(s,a)} \leq \frac{3P(s'|s,a)}{2} + \frac{2\iota}{n(s,a)} \\ P(s'|s,a) &\leq 2\widehat{P}(s'|s,a) + \frac{10\iota}{3n(s,a)} \leq 2\widehat{P}(s'|s,a) + \frac{4\iota}{n(s,a)}. \end{aligned} \tag{54}$$

For the first inequality, we have

$$\begin{aligned} \mathrm{Var}(\widehat{P}_{s,a},V) &= \widehat{P}_{s,a}(V-\widehat{P}_{s,a}V)^2 \leq \widehat{P}_{s,a}(V-P_{s,a}V)^2 \\ &\leq \sum_{s'}\left(\frac{3P(s'|s,a)}{2} + \frac{2\iota}{n(s,a)}\right)(V(s')-P_{s,a}V)^2 \\ &\leq \frac{3}{2}\mathrm{Var}(P_{s,a},V) + \frac{2\left\|V\right\|_\infty^2 S\iota}{n(s,a)}, \end{aligned}$$

here the first inequality is due to $\widehat{P}_{s,a}V := \mathrm{argmin}_z \sum_{s'}\widehat{P}_{s,a}(s')(V(s')-z)^2$, and the last term has $S+1$ due to the extra state g. For the second part, we have

$$\begin{aligned} \mathrm{Var}(P_{s,a},V) &= P_{s,a}(V-P_{s,a}V)^2 \leq P_{s,a}(V-\widehat{P}_{s,a}V)^2 \\ &\leq \sum_{s'}\left(2\widehat{P}(s'|s,a) + \frac{4\iota}{n(s,a)}\right)(V(s')-\widehat{P}_{s,a}V)^2 \\ &\leq 2\mathrm{Var}(\widehat{P}_{s,a},V) + \frac{4\left\|V\right\|_\infty^2 (S+1)\iota}{n(s,a)}, \end{aligned}$$

□

Lemma 10.2. *With probability at least* $1-\delta$*, we have*

$$c(s,a) \le 2\widehat{c}(s,a) + \frac{10\iota}{3n(s,a)}$$
$$\widehat{c}(s,a) \le \frac{3}{2}c(s,a) + \frac{5\iota}{3n(s,a)}$$

Proof. Conditioned on event $\mathcal{E}_4$, we have

$$\begin{aligned}|c(s,a)-\widehat{c}(s,a)| &\le \sqrt{\frac{2c(s,a)\iota}{n(s,a)}} + \frac{2\iota}{3n(s,a)} \\ &\le \frac{\iota}{n(s,a)} + \frac{1}{2}c(s,a) + \frac{2\iota}{3n(s,a)} \\ &\le \frac{5\iota}{3n(s,a)} + \frac{1}{2}c(s,a),\end{aligned}$$

where the first inequality uses the assumption that $c(s,a) \in [0,1]$. The second inequality follows from the result that $\sqrt{ab} \le \frac{a+b}{2}$. Simplify the above inequality, we can conclude the proof. □

Lemma 10.3. *With probability* $1-\delta$*, for all* $V(\cdot) \in \mathbb{R}^{S'}$ *such that* $\|V\|_\infty < \infty$*, we have for all sate-action pair* (s,a)

$$(\widehat{P}_{s,a} - P_{s,a})V \le \sqrt{\frac{2S\mathrm{Var}(P_{s,a},V)\iota}{n(s,a)}} + \frac{2\|V\|_\infty S\iota}{3n(s,a)}, \tag{55}$$

where $\iota = O(\log(SA/\delta))$*.*

Proof. Suppose the event $\mathcal{E}_1$ holds. Then we have (deterministically)

$$\begin{aligned}|(\widehat{P}_{s,a} - P_{s,a})V| &\overset{\text{(i)}}{=} |(\widehat{P}_{s,a} - P_{s,a})(V - P_{s,a}V\mathbf{1}_S)| \\ &\le (\sqrt{\frac{2P(\cdot|s,a)\iota}{n(s,a)}} + \frac{2\iota}{3n(s,a)})|V - P_{s,a}V\mathbf{1}_S| \\ &\le \sqrt{\frac{2P_{s,a}\iota}{n(s,a)}}|V - P_{s,a}V\mathbf{1}_S| + \frac{2S\|V\|_\infty \iota}{3n(s,a)} \\ &\overset{\text{(ii)}}{\le} (\sqrt{S}\sqrt{\frac{2P_{s,a}(V-P_{s,a}V\mathbf{1}_S)^2\iota}{n(s,a)}}) + \frac{2S\|V\|_\infty \iota}{3n(s,a)} \\ &\le \sqrt{\frac{2S\mathrm{Var}(P_{s,a},V)\iota}{n(s,a)}} + \frac{2\|V\|_\infty S\iota}{3n(s,a)},\end{aligned}$$

where (i) follows from the fact that $P_{s,a}V$ is a scalar, which implies that $(\widehat{P}_{s,a} - P_{s,a})(P_{s,a}V)\mathbf{1}_S = (P_{s,a}V)\sum_{s'}(\widehat{P}(s'|s,a) - P(s'|s,a)) = 0$. (ii) uses the Cauchy-Schwartz inequality. Lastly, $\mathcal{E}_1$ fails with probability only δ (by Lemma 4.1).

□

11 MINIMAX LOWER BOUND FOR OFFLINE SSP

In this section, we provide the minimax lower bound for offline stochastic shortest path problem. Concretely, we consider the family of problems satisfying bounded partial coverage, *i.e.* $\max_{s,a,s\neq g} \frac{d^{\pi^\star}(s,a)}{d^\mu(s,a)} \le C^\star$, where $d^\pi(s,a) = \sum_{h=0}^\infty \xi_h^\pi(s,a) < \infty$ for all s,a (excluding g) for any proper policy π. Formally, we have the following result:

Theorem 11.1 (Restatement of Theorem **??**). *We define the following family of SSPs:*

$$\mathrm{SSP}(C^\star) = \{(s_{\mathrm{init}}, \mu, P, c) | \max_{s,a,s\neq g} \frac{d^{\pi^\star}(s,a)}{d^\mu(s,a)} \leq C^\star\},$$

where $d^\pi(s,a) = \sum_{h=0}^\infty \xi_h^\pi(s,a)$*. Then for any* $C^\star \geq 1$, $\|V^\star\|_\infty = B_\star > 1$*, it holds (for some universal constant c)*

$$\inf_{\widehat{\pi} \text{ proper}} \sup_{(s_{\mathrm{init}},\mu,P,c)\in\mathrm{SSP}(C^\star)} \mathbb{E}_{\mathcal{D}}[V^{\widehat{\pi}}(s_{\mathrm{init}}) - V^\star(s_{\mathrm{init}})] \geq c \cdot B_\star \sqrt{\frac{SC^\star}{n}}.$$

The proof of Theorem 11.1 relies on the hard instances construction that is similar to Rashidinejad et al. [2021]. However, we need to incorporate the absorbing state g and assign the transition of initial state s_{init} carefully to make sure the optimal proper policy exists.

Proof of Theorem 11.1. We create hard instances of SSPs as follows: we split $S-1$ states (except s_{init}) into $S' = (S-1)/2$ groups, and denote it as $\mathcal{S} = \{s_{\mathrm{init}}\} \cup \{s_1^j, s_+^j\}_{j=1}^{S'}$. For s_1^j, $j = 1, \ldots, S'$, there are two actions a_1, a_2 and for states s_{init}, s_+^j and goal state g there is only one default action a_d (therefore the only choice is always optimal for those states). Concretely,

- For state s_{init}, it transitions to s_1^j ($j = 1, \ldots, S'$) uniformly with probability $1/S'$, *i.e.* $P(s_1^j | s_{\mathrm{init}}, a_d) = 1/S'$;
- For each state s_1^j, it satisfies

  $$P(s_+^j | s_1^j, a_1) = P(g | s_1^j, a_1) = 1/2;\ P(s_+^j | s_1^j, a_2) = \frac{1}{2} + v_j \delta;\ P(g | s_1^j, a_2) = \frac{1}{2} - v_j \delta.$$

  where $v_j \in \{+1, -1\}$ and δ to be specified later.
- For s_+^j, it satisfies

  $$P(s_+^j | s_+^j, a_d) = q, P(g | s_+^j, a_d) = 1 - q,$$

  where $q = 1 - \frac{1}{B_\star}$ and g is absorbing.
- the cost function satisfies (regradless of actions):

  $$c(s_{\mathrm{init}}) = c(s_1^j) = c(s_+^j) = 1, c(g) = 0.$$

  It is easy to check this is a SSP. Moreover, it is clear when $v_j = 1$, the optimal action at s_1^j is a_1 and if $v_j = -1$ the optimal action is a_2. Note by straightforward calculation we have that $\|V^\star\|_\infty \leq 2B_\star$.

We consider the family of SSP instances $\mathcal{P}$ to satisfy Lemma 12.1, i.e. it satisfies $|\mathcal{P}| \geq e^{S'/8}$ and for any two instances in $\mathcal{P}$, $\|v_i - v_j\|_1 \geq S'/2$. Also, it suffices to consider all the deterministic learning algorithms, as stochastic output policies are randomized versions over deterministic ones (c.f. Krishnamurthy et al. [2016]). Then we have the following lemma:

Lemma 11.2. *For any (deterministic) policy* π *and any two different transition probabilities* $P_1, P_2 \in \mathcal{P}$*, it holds:*

$$V_{P_1}^\pi(s_{\mathrm{init}}) - V_{P_1}^{\pi^\star}(s_{\mathrm{init}}) + V_{P_2}^\pi(s_{\mathrm{init}}) - V_{P_2}^{\pi^\star}(s_{\mathrm{init}}) \geq \delta B_\star/2.$$

Proof of Lemma 11.2. Since s_{init} uniformly transitions to S' states (w.r.t the default action a_d), therefore for any policy π,

$$V_{P_1}^\pi(s_{\mathrm{init}}) - V_{P_1}^{\pi^\star}(s_{\mathrm{init}}) = 1 + \frac{1}{S'} \sum_{i=1}^{S'} V_{P_1}^\pi(s_1^i) - \left(1 + \frac{1}{S'} \sum_{i=1}^{S'} V_{P_1}^{\pi^\star}(s_1^i)\right) = \frac{1}{S'} \sum_{i=1}^{S'} (V_{P_1}^\pi(s_1^i) - V_{P_1}^{\pi^\star}(s_1^i)).$$

Case 1. If $v_j = 1$, then

$$P(s_+^i | s_1^i, a_2) = \frac{1}{2} + \delta,\ P(g | s_1^i, a_2) = \frac{1}{2} - \delta$$

and in this case $\pi^\star(s_1^i) = a_1$.

If $\pi(s_1^i) = a_2$, then

$$V^\pi(s_1^i) = 1 + (\frac{1}{2} + \delta)V^\pi(s_+^i) + (\frac{1}{2} - \delta)V^\pi(g) = 1 + (\frac{1}{2} + \delta)V^\pi(s_+^i),$$

and this implies

$$\begin{aligned} V^\pi(s_1^i) - V^{\pi^\star}(s_1^i) &= (\frac{1}{2} + \delta)V^\pi(s_+^i) - \frac{1}{2}V^{\pi^\star}(s_+^i) \\ &\geq \delta V^\pi(s_+^i) = \delta(1 + q + q^2 + \ldots) = \delta \cdot \frac{1}{1-q} = \delta B_\star. \end{aligned}$$

If $\pi(s_1^i) = a_1$, then $V^\pi(s_1^i) - V^{\pi^\star}(s_1^i) \geq 0$. Therefore, in this case, one has

$$V^\pi(s_1^i) - V^{\pi^\star}(s_1^i) \geq \delta B_\star \cdot \mathbf{1}[\pi(s_1^i) \neq \pi^\star(s_1^i)].$$

Case 2. If $v_j = -1$, then

$$P(s_+^i|s_1^i, a_2) = \frac{1}{2} - \delta,\ P(g|s_1^i, a_2) = \frac{1}{2} + \delta$$

and in this case $\pi^\star(s_1^i) = a_2$.

If $\pi(s_1^i) = a_1$, then

$$V^\pi(s_1^i) = 1 + \frac{1}{2} \cdot V^\pi(s_+^i) + \frac{1}{2}V^\pi(g) = 1 + \frac{1}{2}V^\pi(s_+^i),$$

and this implies

$$\begin{aligned} V^\pi(s_1^i) - V^{\pi^\star}(s_1^i) &= \frac{1}{2}V^\pi(s_+^i) - (\frac{1}{2} - \delta)V^{\pi^\star}(s_+^i) \\ &\geq \delta V^{\pi^\star}(s_+^i) = \delta(1 + q + q^2 + \ldots) = \delta \cdot \frac{1}{1-q} = \delta B_\star. \end{aligned}$$

If $\pi(s_1^i) = a_2$, then $V^\pi(s_1^i) - V^{\pi^\star}(s_1^i) \geq 0$. Therefore, in this case, we still have

$$V^\pi(s_1^i) - V^{\pi^\star}(s_1^i) \geq \delta B_\star \cdot \mathbf{1}[\pi(s_1^i) \neq \pi^\star(s_1^i)].$$

Combine the above two cases, we have

$$\begin{aligned} &V_{P_1}^\pi(s_{\text{init}}) - V_{P_1}^{\pi^\star}(s_{\text{init}}) + V_{P_2}^\pi(s_{\text{init}}) - V_{P_2}^{\pi^\star}(s_{\text{init}}) \\ \geq &\frac{1}{S'}\sum_{i=1}^{S'}(V_{P_1}^\pi(s_1^i) - V_{P_1}^{\pi^\star}(s_1^i)) + \frac{1}{S'}\sum_{i=1}^{S'}(V_{P_2}^\pi(s_1^i) - V_{P_2}^{\pi^\star}(s_1^i)) \\ \geq &\frac{1}{S'}\delta B_\star \sum_{i=1}^{S'}\left(\mathbf{1}[\pi(s_1^i) \neq \pi_{P_1}^\star(s_1^i)] + \mathbf{1}[\pi(s_1^i) \neq \pi_{P_2}^\star(s_1^i)]\right) \\ \geq &\frac{\delta B_\star}{S'}\sum_{i=1}^{S'}\mathbf{1}[\pi_{P_1}^\star(s_1^i) \neq \pi_{P_2}^\star(s_1^i)] \end{aligned} \tag{56}$$

Lastly, by Lemma 12.1, $\sum_{i=1}^{S'}\mathbf{1}[\pi_{P_1}^\star(s_1^i) \neq \pi_{P_2}^\star(s_1^i)] = \|v_{P_1} - v_{P_2}\|_1 \geq S'/2$, and plug this back to (56) we obtain the result.

□

Now we construct the behavior policy μ such that the data trajectories generated from the induced distribution $\mu \circ P$ suffice for the lower bound. Since only s_1^i has two actions, we specify below:

$$\mu(a_2|s_1^i) = 1/C^\star,\ \mu(a_1|s_1^i) = 1 - 1/C^\star,\ \ \forall i \in \{1, \ldots, S'\}.$$

First, we examine this choice belongs to SSP($C^\star$). Indeed, the only case where a_2 is the suboptimal action for all s_1^i ($i \in \{1, \ldots, S'\}$) is when $v_1, \ldots, v_{S'}$ all equal 1. We can eliminate this SSP from $\mathcal{P}$ and the property of Lemma 12.1 still holds. Then, for some i_0 such that $v_{i_0} = -1$ (a_2 is the optimal action for this state), we have

$$d^\star(s_1^{i_0}, a_2) = d^\star(s_1^{i_0}) \cdot 1 = \frac{1}{S'}, \quad d^\mu(s_1^{i_0}, a_2) = d^\mu(s_1^{i_0})\mu(a_2|s_1^{i_0}) = \frac{1}{S'C^\star},$$

therefore $d^\star(s_1^{i_0}, a_2)/d^\mu(s_1^{i_0}, a_2) = C^\star$ and this $(s_{\text{init}}, \mu, P, c) \in$SSP($C^\star$).

Recall n is the number of episodes. Now apply Fano's inequality (Lemma 12.5) (where each whole trajectory is considered one single data point over the distribution $\mu \circ P$ therefore $\mathcal{D} := \{(s_1^{(i)}, a_1^{(i)}, c_1^{(i)}, s_2^{(i)}, \ldots, s_{T_i}^{(i)})\}_{i=1,\ldots,n}$ consists of n i.i.d. samples) and Lemma 11.2, we have[1]

$$\inf_{\widehat{\pi}} \sup_{(s_{\text{init}},\mu,P,c)\in\mathcal{P}} \mathbb{E}_{\mathcal{D}}[V^{\widehat{\pi}}(s_{\text{init}}) - V^\star(s_{\text{init}})]] \geq \frac{\delta B_\star}{4}\left(1 - \frac{n \cdot \max_{i\neq j} \mathrm{KL}(\mu \circ P_i || \mu \circ P_j) + \log 2}{\log|\mathcal{P}|}\right).$$

Note by the choice of $\mathcal{P}$, $\log|\mathcal{P}| \geq S'/8$, therefore it remains to bound $\max_{i\neq j} \mathrm{KL}(\mu \circ P_i || \mu \circ P_j)$. By definition, we have

$$\mathrm{KL}(\mu \circ P_1 || \mu \circ P_2) = \frac{1}{S'} \sum_{i=1}^{S'} \sum_{\tau_{s_1^i}} \mathbb{P}_1(\tau_{s_1^i}) \log \frac{\mathbb{P}_1(\tau_{s_1^i})}{\mathbb{P}_2(\tau_{s_1^i})},$$

where $\tau_{s_1^i}$ corresponds to all the possible trajectories starting from s_1^i. Then there are the following several cases:[2]

- If $\tau_{s_1^i} = \{s_1^i \to a_1 \to g\}$, then $\mathbb{P}(s_1^i \to a_1 \to g) = (1 - \frac{1}{C^\star})\frac{1}{2}$;
- If $\tau_{s_1^i} = \{s_1^i \to a_2 \to g\}$, then $\mathbb{P}(s_1^i \to a_2 \to g) = \frac{1}{C^\star} \cdot (\frac{1}{2} - v_i\delta)$;
- If $\tau_{s_1^i} = \{s_1^i \to a_1 \to s_+^i \to g\}$, then $\mathbb{P}(s_1^i \to a_1 \to s_+^i \to g) = (1 - \frac{1}{C^\star})\frac{1}{2}(1-q)$;
- If $\tau_{s_1^i} = \{s_1^i \to a_2 \to s_+^i \to g\}$, then $\mathbb{P}(s_1^i \to a_2 \to s_+^i \to g) = \frac{1}{C^\star} \cdot (\frac{1}{2} + v_i\delta)(1-q)$;
- If $\tau_{s_1^i} = \{s_1^i \to a_1 \to s_+^i \to s_+^i \to g\}$, then $\mathbb{P}(s_1^i \to a_1 \to s_+^i \to s_+^i \to g) = (1 - \frac{1}{C^\star})\frac{1}{2}q(1-q)$;
- If $\tau_{s_1^i} = \{s_1^i \to a_2 \to s_+^i \to s_+^i \to g\}$, then $\mathbb{P}(s_1^i \to a_2 \to s_+^i \to s_+^i \to g) = \frac{1}{C^\star}(\frac{1}{2} + v_i\delta)q(1-q)$;
- If $\tau_{s_1^i} = \{s_1^i \to a_1 \to (s_+^i)_{\times k} \to g\}$, then $\mathbb{P}(s_1^i \to a_1 \to (s_+^i)_{\times k} \to g) = (1 - \frac{1}{C^\star})\frac{1}{2}q^{k-1}(1-q)$;
- If $\tau_{s_1^i} = \{s_1^i \to a_2 \to (s_+^i)_{\times k} \to g\}$, then $\mathbb{P}(s_1^i \to a_2 \to (s_+^i)_{\times k} \to g) = \frac{1}{C^\star}(\frac{1}{2} + v_i\delta)q^{k-1}(1-q)$;

Note for path $\tau_{s_1^i}$ that chooses action a_1, $\mathbb{P}_1(\tau_{s_1^i}) = \mathbb{P}_2(\tau_{s_1^i})$ which implies $\mathbb{P}_1(\tau_{s_1^i}) \log \frac{\mathbb{P}_1(\tau_{s_1^i})}{\mathbb{P}_2(\tau_{s_1^i})} = 0$, so we only need to sum over the paths that choose a_2. In particular, we have

$$\begin{aligned}
&\sum_{\tau_{s_1^i}} \mathbb{P}_1(\tau_{s_1^i}) \log \frac{\mathbb{P}_1(\tau_{s_1^i})}{\mathbb{P}_2(\tau_{s_1^i})} = \frac{1}{C^\star} \cdot (\frac{1}{2} - v_i^{P_1}\delta) \log \frac{\frac{1}{C^\star} \cdot (\frac{1}{2} - v_i^{P_1}\delta)}{\frac{1}{C^\star} \cdot (\frac{1}{2} - v_i^{P_2}\delta)} \\
&+ \sum_{k=0}^{\infty} \frac{1}{C^\star}(\frac{1}{2} + v_i^{P_1}\delta)q^{k-1}(1-q) \log \frac{\frac{1}{C^\star}(\frac{1}{2} + v_i^{P_1}\delta)q^{k-1}(1-q)}{\frac{1}{C^\star}(\frac{1}{2} + v_i^{P_2}\delta)q^{k-1}(1-q)} \\
&= \frac{1}{C^\star} \cdot (\frac{1}{2} - v_i^{P_1}\delta) \log \frac{\frac{1}{2} - v_i^{P_1}\delta}{\frac{1}{2} - v_i^{P_2}\delta} + \frac{1}{C^\star} \cdot (\frac{1}{2} + v_i^{P_1}\delta) \log \frac{\frac{1}{2} + v_i^{P_1}\delta}{\frac{1}{2} + v_i^{P_2}\delta} \\
&\leq \frac{1}{C^\star} \cdot (\frac{1}{2} - v_i^{P_1}\delta) \log \frac{\frac{1}{2} - v_i^{P_1}\delta}{\frac{1}{2} + v_i^{P_1}\delta} + \frac{1}{C^\star} \cdot (\frac{1}{2} + v_i^{P_1}\delta) \log \frac{\frac{1}{2} + v_i^{P_1}\delta}{\frac{1}{2} - v_i^{P_1}\delta} \\
&= \frac{1}{C^\star} 2\delta \log \frac{\frac{1}{2} + \delta}{\frac{1}{2} - \delta} = \frac{1}{C^\star} 2\delta \log(1 + \frac{2\delta}{\frac{1}{2} - \delta}) \leq \frac{4\delta^2}{C^\star},
\end{aligned}$$

[1] Note here we drop $\widehat{\pi}$ is proper as the theorem statement did. We can do this since, for all the instances in $\mathcal{P}$, any policy is proper.

[2] We omit the subscript j in P_j here and only uses $\mathbb{P}$ to denote $\mu \circ P$ for the moment.

where the first inequality comes from when $v_i^{P_1} = v_i^{P_2}$ then the term is simply 0 and the second to the last equality holds true regardless of whether $v_i = 1$ or $v_i = -1$. The last inequality comes from $\log(1+x) \leq x$ for all $x > -1$ (here $0 < \delta < \frac{1}{2}$).

Plug above back into the definition we obtain

$$\max_{i \neq j} \mathrm{KL}(\mu \circ P_i || \mu \circ P_j) \leq 4\delta^2 / C^\star,$$

and as long as

$$\frac{4n\delta^2}{C^\star S'/8} \leq \frac{1}{2}$$

e.g. if we choose $\delta = \frac{1}{16}\sqrt{\frac{C^\star S}{n}}$ (recall $S' = (S-1)/2$), then we have

$$\inf_{\widehat{\pi}} \sup_{(s_{\text{init}}, \mu, P, c) \in \mathcal{P}} \mathbb{E}_{\mathcal{D}}[V^{\widehat{\pi}}(s_{\text{init}}) - V^\star(s_{\text{init}})]] \geq \frac{\delta B_\star}{4} \frac{1}{2} = \frac{1}{128} B_\star \sqrt{\frac{C^\star S}{n}}.$$

This completes the proof.

□

12 TECHNICAL LEMMAS

Lemma 12.1 (Gilbert-Varshamov)**.** *There exists a subset $\mathcal{V}$ of $\{-1, 1\}^S$ such that*

- $|\mathcal{V}| \geq 2^{S/8}$;
- *for any two different $v_i, v_j \in \mathcal{V}$, it holds $\|v_i - v_j\|_1 \geq S/2$.*

Lemma 12.2 (Generalized Chernoff bound)**.** *Suppose $X_1, \ldots, X_n$ are independent random variables taking values in $[a, b]$. Let $X = \sum_{i=1}^n X_i$ denote their sum and let $\mu = E[X_i]$. Then for any $\delta > 0$,*

$$\mathbb{P}[X < (1-\theta)n\mu] \leq e^{-2\theta^2 n\mu^2/(b-a)^2} \quad \text{and} \quad \mathbb{P}[X \geq (1+\theta)pn] \leq e^{-2\theta^2 n\mu^2/(b-a)^2}.$$

This result can be found in Sums of independent bounded random variables Section of `https://en.wikipedia.org/wiki/Chernoff_bound`.

Lemma 12.3 (Bernstein's Inequality)**.** *Let $x_1, \ldots, x_n$ be independent bounded random variables such that $\mathbb{E}[x_i] = 0$ and $|x_i| \leq \xi$ with probability 1. Let $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}[x_i]$, then with probability $1-\delta$ we have*

$$\frac{1}{n}\sum_{i=1}^n x_i \leq \sqrt{\frac{2\sigma^2 \cdot \log(1/\delta)}{n}} + \frac{2\xi}{3n}\log(1/\delta)$$

Lemma 12.4 (Empirical Bernstein's Inequality [Maurer and Pontil, 2009])**.** *Let $x_1, \ldots, x_n$ be i.i.d random variables such that $|x_i| \leq \xi$ with probability 1. Let $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\widehat{V}_n = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$, then with probability $1-\delta$ we have*

$$\left|\frac{1}{n}\sum_{i=1}^n x_i - \mathbb{E}[x]\right| \leq \sqrt{\frac{2\widehat{V}_n \cdot \log(2/\delta)}{n}} + \frac{7\xi}{3n}\log(2/\delta).$$

Lemma 12.5 (Generalized Fano's inequality)**.** *Let $L : \Theta \times \mathcal{A} \to \mathbb{R}_+$ be any loss function, and there exist $\theta_1, \ldots, \theta_m \in \Theta$ such that*

$$L(\theta_i, a) + L(\theta_j, a) \geq \Delta, \quad \forall i \neq j \in [m], a \in \mathcal{A}.$$

Then it holds

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta L(\theta, \widehat{\theta}) \geq \frac{\Delta}{2}\left(1 - \frac{n \cdot \max_{i \neq j} \mathrm{KL}(\mathbb{P}_{\theta_i} || \mathbb{P}_{\theta_j}) + \log 2}{\log m}\right),$$

where n is the number of i.i.d. samples sampled from the distribution $\mathbb{P}_\theta$.

Proof of Lemma 12.5. The proof come from the combination of Lemma 1 and Lemma 3 of Han and Fischer-Hwang [2019]. ◻

Lemma 12.6 (Chernoff Bound for Stochastic Shortest Path)**.** *Recall by definition*

$$n(s,a)=\sum_{i=1}^{n}\sum_{h=1}^{T_i}\mathbf{1}[s_h^{(i)}=s,a_h^{(i)}=a]=\sum_{i=1}^{n}\sum_{h=1}^{\infty}\mathbf{1}[s_h^{(i)}=s,a_h^{(i)}=a].$$

Let $T_{\max}=\max_i T_i$ and recall $d_m:=\min\{\sum_{h=0}^{\infty}\xi_h^{\mu}(s,a):s.t.\sum_{h=0}^{\infty}\xi_h^{\mu}(s,a)>0\}$. When $n>C\cdot T_{\max}^2\log(SA/\delta)/d_m^2$, with probability $1-\delta$, for all $s,a\in\mathcal{S}\times\mathcal{A}$,

$$\frac{1}{2}n\cdot\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)\le n(s,a)\le\frac{3}{2}n\cdot\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a).$$

Proof of Lemma 12.6. Indeed, denote $n_t(s,a)=\sum_{i=1}^{n}\sum_{h=1}^{t}\mathbf{1}[s_h^{(i)}=s,a_h^{(i)}=a]$, then

$$\mathbb{E}[n_t(s,a)]=\sum_{i=1}^{n}\sum_{h=1}^{t}\mathbb{E}[\mathbf{1}[s_h^{(i)}=s,a_h^{(i)}=a]]=\sum_{i=1}^{n}\sum_{h=1}^{t}\xi_h^{\mu}(s,a)=n\sum_{h=1}^{t}\xi_h^{\mu}(s,a).$$

Now define $X_{i,t}=\sum_{h=1}^{t}\mathbf{1}[s_h^{(i)}=s,a_h^{(i)}=a]$, then by $T_{\max}=\max_i T_i$ we have $0\le X_{i,t}\le T_{\max}$ for all i,t since $T_{\max}$ denotes the maximum length of trajectory. Then apply Lemma 12.2 (where we pick $\theta=\frac{1}{2}$) to $n_t(s,a)$ and $\sum_{h=1}^{t}\xi_h^{\mu}(s,a)$ and union bound over s,a, we have with probability $1-\delta$, for any fixed t,

$$\mathbb{P}\left[\frac{1}{2}n\cdot\sum_{h=1}^{t}\xi_h^{\mu}(s,a)\le n_t(s,a)\le\frac{3}{2}n\cdot\sum_{h=1}^{t}\xi_h^{\mu}(s,a),\forall s,a\right]\ge 1-\delta$$

Next note $n_t(s,a)\to n(s,a)$ almost surely, and $\sum_{h=1}^{t}\xi_h^{\mu}(s,a)\to\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)$ almost surely, and that a.s. convergence implies convergence in distribution, we have

$$\begin{aligned}&\mathbb{P}\left[\frac{1}{2}n\cdot\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)\le n(s,a)\le\frac{3}{2}n\cdot\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a),\forall s,a\right]\\=&\lim_{t\to\infty}\mathbb{P}\left[\frac{1}{2}n\cdot\sum_{h=1}^{t}\xi_h^{\mu}(s,a)\le n_t(s,a)\le\frac{3}{2}n\cdot\sum_{h=1}^{t}\xi_h^{\mu}(s,a),\forall s,a\right]\ge 1-\delta\end{aligned}$$

◻

Lemma 12.7. *For any $a,b,c\in\mathbb{R}$, we have*

$$|\min\{a,b\}-\min\{a,c\}|\le|b-c|. \tag{57}$$

Proof.

1. Case I: $a\le b$ and $a\le c$, $|\min\{a,b\}-\min\{a,c\}|=0$.
2. Case II: $a\ge b$ and $a\ge c$, $|\min\{a,b\}-\min\{a,c\}|=|b-c|$.
3. Case III: $b<a<c$ or $c<a<b$, $|\min\{a,b\}-\min\{a,c\}|\le\max\{|a-b|,|a-c|\}\le|b-c|$.

◻

References

Dimitri P Bertsekas and John N Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.

Yanjun Han and Irena Fischer-Hwang. Multiple hypothesis testing: Tree, fano and assouad. 2019. URL https://theinformaticists.com/2019/09/16/lecture-8-multiple-hypothesis-testing-tree-fano-and-assoaud/.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2016.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

Jean Tarbouriech, Runlong Zhou, Simon S Du, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. *Advances in neural information processing systems*, 2021.