Offline Stochastic Shortest Path: Learning, Evaluation and Towards Optimality

Abstract

Goal-oriented Reinforcement Learning, where the agent needs to reach the goal state while simultaneously minimizing the cost, has received significant attention in real-world applications. Its theoretical formulation, *stochastic shortest path* (SSP), has been intensively researched in the online setting. Nevertheless, it remains understudied when such an online interaction is prohibited and only historical data is provided. In this paper, we consider the *offline stochastic shortest path* problem when the state space and the action space are finite. We design the simple *value iteration*-based algorithms for tackling both *offline policy evaluation (OPE)* and *offline policy learning* tasks. Notably, our analysis over these simple algorithms yields strong instance-dependent bounds which can imply worst-case bounds that are near-minimax optimal. We hope our study could help illuminate the fundamental statistical limits of the offline SSP problem and motive its further studies beyond the scope of current consideration.

1 INTRODUCTION

Goal-oriented reinforcement learning aims at entering a goal state while minimizing its expected cumulative cost. The interplay between the agent and the environment keeps continuing when the target/goal state is not reached and this causes trajectories to have variable lengths among different trials, which makes it different from (or arguably more challenging than) the finite-horizon RL. In particular, this setting naturally subsumes the *infinite-horizon γ-discounted* case as one can make up a "ghost" goal state g and set $1-\gamma$ probability to enter g at each timestep for the latter.

The goal-oriented RL covers many popular reinforcement learning tasks, such as navigation problems (e.g., Mujoco mazes), Atari games (*e.g.* breakout) and Solving Rubik's cube [Akkaya et al., 2019] (also see Figure 1 for more examples). Parallel to its empirical popularity, the theoretical formulation, *stochastic shortest path* (SSP), has been studied from the control perspective (*i.e.* with known transition) since Bertsekas and Tsitsiklis [1991]. Recently, there is a surge of studying SSP from the data-driven aspects (*i.e.* with unknown transition) and existing literatures formulate SSP into the *online reinforcement learning* framework [Tarbouriech et al., 2020, Rosenberg et al., 2020, Cohen et al., 2021, Chen and Luo, 2021, Tarbouriech et al., 2021]. On the other hand, there exists no literature (to the best of our knowledge) formally study the *offline* behavior of stochastic shortest path problem.

Figure 1: Examples of Goal-oriented RL tasks in OpenAI-Gym environment. The robot can be asked to move-fetch to a position, orient a block or play with a pen.

In this paper, we study the offline counterpart of the stochastic shortest path (SSP) problem. Unlike its online version, we have no access to further explore new strategies (policies) and the data provided are historical trajectories. The goal is to come up with a cost-minimizing policy that can enter the goal state (*policy learning*) or to evaluate the performance of a target policy (*policy evaluation*).

Why should we study offline SSP? Online SSP provides a suitable learning framework for goal-oriented tasks with cheap experiments (*e.g.* Atari games). However, real-world applications usually have high-stake experiments which makes online interactions infeasible. For instance, in the application of logistic transportation, goods need to be delivered to their destinations. How to minimize the transporta-

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

tion cost should be decided/learned beforehand using the logged data. In the aircraft planning, changing flight routes instantaneously could be dangerous and designing routes based on history records is more appropriate for optimizing flying operation budget. In those scenarios, *offline SSP* suffices for treating the practical challenges as it only learns from historical data.

Our contributions. In this paper, we provide the first systematic study of the offline stochastic shortest path problem, and consider both *offline policy evaluation* (OPE) and *offline policy learning* tasks. As an initial attempt, we design the simple *value iteration*-based algorithms to tackle the problems and obtain strong statistical guarantees. Concretely, our contributions are four folds.

- For offline policy evaluation task, we design VI-OPE algorithm (Algorithm 1) under the coverage Assumption 2.4. In particular, our algorithm is *parameter-free* (requires no knowledge about T^π/B^π) and fully executed by the offline data. Theorem 3.1 provides the first statistical guarantee for offline SSP evaluation and nearly matches the statistical efficiency of its finite horizon counterpart (see discussion in Section 3);

- For offline learning task, we propose *pessimism*-based algorithm PVI-SSP (Algorithm 2) under the Assumption 2.5 and 2.6. Our result (Theorem 4.1) has several merits: it is instance-dependent (as opposed to the worst-case guarantees in the existing online SSP works), enjoys faster $\widetilde{O}(1/n)$ convergence when the system is deterministic, and is also minimax-rate optimal. We believe Theorem 4.1 is (in general) unimprovable for the current tabular setting.

- To understand the statistical limit of offline SSP, we prove the minimax lower bound $\Omega(B_\star\sqrt{\frac{SC^\star}{n}})$ (Theorem 5.1) under the marginal coverage concentrability $\max_{s,a,s\neq g}\frac{d^{\pi^\star}(s,a)}{d^\mu(s,a)}\leq C^\star$. Our Theorem 4.1 can match this rate (up to the logarithmic factor).

- Along the way for solving the problem, we highlight two new technical observations: Lemma 6.1 and Lemma 6.3. The first one depicts the connection between the expected time T^π and marginal coverage $d^\pi(s,a)$. As a result, we can express our result without using T^π but the ratio-based quantity $\frac{d^\pi(s,a)}{d^\mu(s,a)}$, which matches the flavor of previous finite-horizon RL studies (also see Remark 6.2). The second one is a general dependence improvement lemma that works with arbitrary policy π and is the key for guaranteeing minimax optimal rate (also see Remark 6.4). Both Lemmas are general and may be of independent interest.

1.1 RELATED WORKS.

Stochastic shortest path itself is a broad topic and we are not aiming for the exhaustive review. Here we discuss two aspects that are most relevant to us.

Online SSP. Previous literatures intensively focus on the online aspect of SSP learning. Earlier works consider two types of problems: online shortest path routing problem with deterministic dynamics, which can be solved using the combinatorial bandit technique (*e.g.* György et al. [2007], Talebi et al. [2017]); or SSP with stochastic transitions but adversarial feedbacks [Neu et al., 2012, Zimin and Neu, 2013, Rosenberg and Mansour, 2019, Chen and Luo, 2021, Chen et al., 2021b]. Recently, Tarbouriech et al. [2020] starts investigating general online SSP learning problem and introduce the UC-SSP algorithm to first achieve the no-regret bound $\widetilde{O}(DS\sqrt{ADK})$.[1] Rosenberg et al. [2020] improves this result to $\widetilde{O}(B_\star S\sqrt{AK})$ via a UCRL2-style algorithm with Bernstein-type bonus for exploration. Later, Cohen et al. [2021] eventually achieves the minimax rate $\widetilde{O}(B_\star\sqrt{SAK})$ by a reduction from SSP to finite-horizon MDP. However, the reduction technique requires the knowledge of $B_\star$ and $T_\star$. Most recently, Tarbouriech et al. [2021] proposes EB-SSP which recovers the minimax rate but gets rid of the parameter knowledge (*parameter-free*). When the parameters are known, their results can be *horizon-free*.

Other than the general tabular SSP learning, there are also other threads, *e.g.* Linear MDPs [Min et al., 2021, Vial et al., 2021, Chen et al., 2021a] and posterior sampling [Jafarnia-Jahromi et al., 2021]. Nevertheless, no analysis has been conducted for offline SSP yet.

Offline tabular RL. In the offline RL regime, there are fruitful results under different type of assumptions. Yin et al. [2021] first achieves the minimax rate $\widetilde{O}(\sqrt{H^3/nd_m})$ for non-stationary MDP with the strong uniform coverage assumption. Ren et al. [2021] improves the result to $\widetilde{O}(\sqrt{H^2/nd_m})$ for the stationary MDP setting. Later, Rashidinejad et al. [2021], Xie et al. [2021] use the weaker single concentrability assumption and achieve the minimax rate $\widetilde{O}\sqrt{H^3SC^\star/n}$ (or $\widetilde{O}\sqrt{(1-\gamma)^{-3}SC^\star/n}$). Recently, this is further subsumed by the tighter instance-dependent result [Yin and Wang, 2021]. For offline policy evaluation (OPE) task, statistical efficiency has been achieved in tabular [Yin and Wang, 2020], linear [Duan et al., 2020] and differentiable function approximation settings [Zhang et al., 2022].

2 PROBLEM SETUP

[1] Here the diameter of SSP is defined as $D := \max_{s\in\mathcal{S}}\min_{\pi\in\Pi}T_s^\pi$ and by Lemma 2 of Tarbouriech et al. [2020] $B_\star := \|V^\star\|_\infty \leq c_{\max}D$. In this paper, we consider the dependence on $B_\star$ only since our $c_{\max}=1$ and this implies $B_\star\leq D$.

Stochastic Shortest Path. An SSP problem consists of a *Markov decision process* (MDP) together with an initial state s_{init} and an extra goal state g and it is denoted by the tuple $M := \langle \mathcal{S}, \mathcal{A}, P, c, s_{\text{init}}, g \rangle$. In particular, we denote $\mathcal{S}' := \mathcal{S} \cup \{g\}$. Each state s and action a incurs a bounded random cost (within $[0, 1]$) drawn i.i.d. from a distribution with expectation $c(s, a)$ and will transition to the next state $s' \in \mathcal{S}'$ according to the probability distribution $P(\cdot|s, a)$. Here $\sum_{s' \in \mathcal{S}'} P(s' \mid s, a) = 1$. The goal state g is a termination state with absorbing property and has cost zero (*i.e.* $P(g|g, a) = 1, c(g, a) = 0$ for all $a \in \mathcal{A}$).

The optimal behavior of the agent is characterized by a stationary, deterministic and proper policy that minimizes the expected total cost of reaching the goal state from *any* state s. A stationary policy $\pi : \mathcal{S} \to \Delta^{\mathcal{A}}$ is a mapping from state s to a probability distribution over actions $\Delta^{\mathcal{A}}$. The definition of proper policy is defined as follows.

Definition 2.1 (Proper policies)**.** *A policy π is proper if playing π reaches the goal state with probability 1 when starting from any state. A policy is improper if it is not proper. Denote the set of proper policies as Π_{prop}.*

Value and Q-functions in SSP. Any policy π induces a *cost-to-go* value function $V^\pi : \mathcal{S} \mapsto [0, \infty]$ defined as

$$V^\pi(s) := \lim_{T \to \infty} \mathbb{E}^\pi \left[\sum_{t=0}^{T} c(s_t, a_t) | s_0 = s \right], \quad \forall s \in \mathcal{S}$$

and the Q-function is defined as $\forall s, a \in \mathcal{S} \times \mathcal{A}$,

$$Q^\pi(s, a) := \lim_{T \to \infty} \mathbb{E}^\pi \left[\sum_{t=0}^{T} c(s_t, a_t) | s_0 = s, a_0 = a \right],$$

where the expectation is taking w.r.t. the random trajectory of states generated by executing π and transitioning according to P. Also, we denote $T_s^\pi := \lim_{T \to \infty} \mathbb{E}[\sum_{t=0}^{T} \mathbf{1}[s_t \neq g] | s_0 = s] = \mathbb{E}[\sum_{t=0}^{\infty} \mathbf{1}[s_t \neq g] | s_0 = s]$ to be the expected time that π takes to enter g starting from s. By Definition 2.1, π is proper if $T_s^\pi < \infty$ for all s, and improper if $T_s^\pi = \infty$ for some state s. Moreover, by definition it follows $V^\pi(g) = Q^\pi(s, a) = 0$ for all π and action a. The next proposition is the Bellman equation for the SSP problem.

Proposition 2.2 (Bellman equations for SSP problem [Bertsekas and Tsitsiklis, 1991])**.** *Suppose there exists at least one proper policy and that for every improper policy π' there exists at least one state $s \in \mathcal{S}$ such that $V^{\pi'}(s) = +\infty$. Then the optimal policy $\pi^\star$ is stationary, deterministic, and proper. Moreover, $V^\star = V^{\pi^\star}$ is the unique solution of the equation $V^\star = \mathcal{L}V^\star$, where*

$$\mathcal{L}V(s) := \min_{a \in \mathcal{A}} \{c(s, a) + P_{s,a} V\} \quad \forall V \in \mathbb{R}^S.$$

Similarly, for a proper policy π, V^π is the unique solution of $V^\pi = \mathcal{L}^\pi V^\pi$ with $\mathcal{L}^\pi V(s) := \mathbb{E}_{a \sim \pi(\cdot|s)}[c(s, a) + P_{s,a} V]$, $\forall V \in \mathbb{R}^S$. Furthermore, it holds

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

We use $T_s^\star$ to denote the expected arriving time when coupled with the optimal policy $\pi^\star$ and the proof of Proposition 2.2 can be found in Appendix A.

The Offline SSP task. The goal of offline SSP is to reach the goal state but also minimize the cost using offline data $\mathcal{D} := \{(s_0^{(i)}, a_0^{(i)}, c_0^{(i)}, s_1^{(i)}, \ldots, s_{T_i}^{(i)})\}_{i=1,\ldots,n}$, which is collected by a proper (possibly stochastic) behavior policy μ. The optimal policy is a proper policy $\pi^\star$ (the existence of $\pi^\star$ is guaranteed by the Proposition 2.2) which minimizes the value function for all states, *i.e.*,

$$\pi^\star(s) = \arg \min_{\pi \in \Pi_{\text{prop}}} V^\pi(s). \tag{2}$$

The final learning objective is to come up with a (proper) policy $\widehat{\pi}$ using $\mathcal{D}$ such that the suboptimality gap $V^{\widehat{\pi}}(s_{\text{init}}) - V^\star(s_{\text{init}}) < \epsilon$ for a given accuracy $\epsilon > 0$.

Some Notations. In the paper, we may abuse the notation $V^\star$ with $V^{\pi^\star}$, and define $B_\star := \max_s \{V^\star(s)\}$. In addition, we denote $\xi_h^\pi(s, a)$ to be the marginal state-action occupancy at time step h under the policy π and $\xi_h^\pi(s)$ the marginal state occupancy at time h. Furthermore, we define the *marginal coverage* d^π as (given the initial state is s_{init}):

$$d^\pi(s, a) := \sum_{h=0}^{\infty} \xi_h^\pi(s, a), \quad \forall s, a \in \mathcal{S} \times \mathcal{A}. \tag{3}$$

Remark 2.3. *The notation of marginal coverage mirrors the marginal state-action occupancy in the infinite horizon γ-discounted setting but without normalization. Therefore, it is likely that $d^\pi(s, a) > 1$ (or even ∞) for the offline SSP problem. Nevertheless, the key Lemma 6.1 guarantees $d^\pi(s, a)$ is finite when π is a proper policy. This feature helps formalize the following assumptions in offline SSP.*

2.1 ASSUMPTIONS

Offline learning/evaluation in SSP is impossible without assumptions. We now present three required assumptions.

Assumption 2.4 (offline policy evaluation (OPE))**.** *We assume both the target policy π and behavior policy μ are proper. In this case, we have $\Pi_{prop} \neq \emptyset$. Moreover, we assume behavior policy μ can cover the exploration (state-action) space of π, i.e. $\forall s, a \in \mathcal{S} \times \mathcal{A}$ s.t. $d_{\bar{s}}^\pi(s, a) := \sum_{h=0}^{\infty} \xi_{h,\bar{s}}^\pi(s, a) > 0$, it implies $d_{\bar{s}}^\mu(s, a) := \sum_{h=0}^{\infty} \xi_{h,\bar{s}}^\mu(s, a) > 0$, where $d_{\bar{s}}^\pi(s, a)$ is the marginal coverage and $\xi_{h,\bar{s}}^\pi(s, a)$ the marginal state-action occupancy given the initial state $\bar{s}$. In particular, when $\bar{s} = s_{init}$, we suppress the subscript and use d^π, ξ_h^π only.*

There are two remarks that are in order.

Assumption 2.4 requires behavior policy μ can explore all the state-action locations that are explored by π and this mirrors the necessary OPE assumption made in the standard RL setting (*e.g.* Thomas and Brunskill [2016], Yin and Wang [2020], Uehara et al. [2020]). Otherwise, policy evaluation for SSP would incur constant suboptimality gap even when *infinite many* trajectories are collected.

Moreover, instead of making assumption only on $d^{\mu}_{s_{\text{init}}}$, 2.4 assumes μ can cover π when starting from any state $\bar{s}$ (*i.e.* $d^{\pi}_{\bar{s}} > 0$ implies $d^{\mu}_{\bar{s}} > 0$ for all $\bar{s}$). This extra requirement is reasonable since, by Definition 2.1, a proper policy can reach goal state g with probability 1 when starting from any state $\bar{s}$. Similarly, we need the assumptions for offline learning tasks.

Assumption 2.5 (offline policy learning)**.** *We assume there exists a deterministic proper policy and the behavior policy μ is (possible random) proper. Next, by Proposition 2.2, we know there exists a deterministic optimal proper policy $\pi^\star$. We assume behavior policy μ can cover the exploration (state-action) space of $\pi^\star$, i.e. $\forall s, a \in \mathcal{S} \times \mathcal{A}$ s.t. $d^{\pi^\star}_{\bar{s}}(s,a) := \sum_{h=0}^{\infty} \xi^{\pi^\star}_{h,\bar{s}}(s,a) > 0$, it implies $d^{\mu}_{\bar{s}}(s,a) := \sum_{h=0}^{\infty} \xi^{\mu}_{h,\bar{s}}(s,a) > 0$, where $d^{\pi^\star}_{\bar{s}}(s,a)$ and $\xi^{\pi^\star}_{h,\bar{s}}(s,a)$ is the same notion used in Assumption 2.4. In particular, when $\bar{s} = s_{init}$, we suppress the subscript and use $d^{\pi^\star}, \xi^{\pi^\star}_{h}$ only.*

2.5 provides the offline learning version of Assumption 2.4. It echos its offline RL counterpart assumed in Liu et al. [2019], Yin and Wang [2021], Uehara and Sun [2022]. Similar to the offline RL setting (*e.g.* see Yin and Wang [2021] for detailed explanations), this assumption is also required for the tabular offline SSP problem.

Assumption 2.6 (Positive cost [Rosenberg et al., 2020])**.** *There exists $c_{\min} > 0$ such that $c(s,a) \geq c_{\min}$ for every $(s,a) \in \mathcal{S} \times \mathcal{A}$.*[2]

This assumption guarantees there is no *"free-cost"* state. With 2.6 it holds that any policy does not reach the goal state has infinite cost, and this certifies the condition in Proposition 2.2 that for every improper policy π' there exists at least one state s such that $V^{\pi'}(s) = +\infty$. When $c_{\min}$ is 0, a simple workaround is to solve a perturbed SSP instance with all observed costs clipped to ϵ if they are below some $\epsilon > 0$, and in this case $c_{\min} = \epsilon > 0$. This will cause only an additive term of order $O(\epsilon)$ (see Tarbouriech et al. [2020] for online SSP). Therefore, as the first attempt for offline SSP problem, we stick to this assumption throughout the paper. Last but not least, Assumption 2.6 is only used in offline learning problem (Section 4) and our OPE analysis (Section 3) can work well with zero cost.

[2]Note this assumption only holds for $(s,a) \in \mathcal{S} \times \mathcal{A}$. For goal state g, it always has $c(g,a) = 0$ for all $a \in \mathcal{A}$.

3 OFF-POLICY EVALUATION IN SSP

Algorithm 1 VI-OPE (Value Iteration for OPE problem of Stochastic Shortest Path)

1: **Input:** ϵ_{OPE}, $\mathcal{D} := \{(s_1^{(i)}, a_1^{(i)}, c_1^{(i)}, s_2^{(i)}, \ldots, s_{T_i}^{(i)})\}_{i=1}^{n}$.

2: **for** $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}'$ **do**
3: Set $n(s,a) = \sum_{i=1}^{n} \sum_{j=1}^{T_i} \mathbb{I}(s_j^{(i)} = s, a_j^{(i)} = a)$.
4: **if** $n(s,a) > 0$ **then**
5: Calculate $\widehat{c}(s,a) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{T_i} \mathbb{I}(s_j^{(i)}=s,\, a_j^{(i)}=a) c_j^{(i)}}{n(s,a)}$
6: $\widehat{P}(s'|s,a) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{T_i} \mathbb{I}(s_j^{(i)}=s,\, a_j^{(i)}=a,\, s_{j+1}^{(i)}=s')}{n(s,a)}$,
7: **else**
8: $\widehat{c}(s,a) \leftarrow c_{\min}$, $\widehat{P}(s'|s,a) \leftarrow \mathbb{I}(s' = g)$.
9: **end if**
10: ⋄ Perturb the estimated transition kernel
11: $\widetilde{P}(s'|s,a) = \frac{n(s,a)}{n(s,a)+1} \widehat{P}(s'|s,a) + \frac{\mathbb{I}[s'=g]}{n(s,a)+1}$
12: **end for**
13: ⋄ Value Iteration for SSP problem
14: **Initialize:** $V^{(-1)}(\cdot) \leftarrow -\infty$, $V^{(0)}(\cdot) \leftarrow \mathbf{0}$, $i = 0$.
15: **while** $\left\| V^{(i)} - V^{(i-1)} \right\|_{\infty} > \epsilon_{\text{OPE}}$ **do**
16: **for** $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
17: $Q^{(i+1)}(s,a) = \widehat{c}(s,a) + \widetilde{P}_{s,a} V^{(i)}$
18: $V^{(i+1)}(s) = \langle \pi(\cdot|s), Q^{(i+1)}(s,\cdot) \rangle$
19: $i \leftarrow i + 1$
20: **end for**
21: **end while**
22: **Output**: $V^{(i)}(\cdot) \in \mathbb{R}^S$, $V^{(i)}(s_{\text{init}})$.

In this section, we assume that Assumption 2.4 holds and consider *offline policy evaluation* (OPE) for the *stochastic shortest path* (SSP) problem. Our algorithmic design follows the natural idea of *approximate value iteration* [Munos, 2005] and is named **VI-OPE** (Algorithm 1). Specifically, VI-OPE approximates (1) by solving the fixed point solution of the empirical Bellman equation associated with estimated cost $\widehat{c}$ and transition $\widetilde{P}$. One highlight is that we construct $\widetilde{P}$ to be the skewed version of the vanilla empirical estimation $\widehat{P}$ by injecting $\frac{1}{n(s,a)+1}$ probability to state g (Line 11 of Algorithm 1).[3] By such a shift, the empirical Bellman operator $\widehat{\mathcal{T}}^{\pi}(\cdot) := \widehat{c}^{\pi} + \widetilde{P}^{\pi}(\cdot)$ becomes a contraction with rate $\rho := \max_{\substack{s,a \\ s \neq g}} \left(\frac{n_{s,a}}{n_{s,a}+1}\right) < 1$ (see Lemma C.1 for details). Hence, *contraction mapping theorem* [Diaz and Margolis, 1968] guarantees the loop (Line15-21) will end after $O(\log(\epsilon_{\text{OPE}})/\rho)$ iterations for any $\epsilon_{\text{OPE}} > 0$. We have the following main result for VI-OPE, whose proof can be found in Appendix F.

Theorem 3.1 (Offline Policy Evaluation in SSP)**.** *Denote $d_m := \min\{\sum_{h=0}^{\infty} \xi_h^{\mu}(s,a) \,:\, s.t. \sum_{h=0}^{\infty} \xi_h^{\mu}(s,a) > 0\}$, and T_s^{π} to be the expected time to hit g when starting from s. Define $\bar{T}^{\pi} = \max_{\bar{s} \in \mathcal{S}} T_{\bar{s}}^{\pi}$. Then when $n \geq$*

[3]This treatment is also used in Tarbouriech et al. [2021].

$\max\{\frac{49S\iota}{9d_m}, 64(\bar{T}^\pi)^2\frac{S\iota}{d_m}, O(\iota/d_m)\}$, *we have with probability* $1-\delta$, *the output of Algorithm 1 satisfies* ($\iota = O(\log(SA/\delta))$)

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

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

On statistical efficiency. First of all, when VI-OPE converges exactly (*i.e.* $\epsilon_{\text{OPE}} = 0$), the output $\widehat{V}^\pi := \lim_{i\to\infty} V^{(i)}$ possesses no optimization error (*i.e.* $\epsilon_{\text{OPE}}/(1-\rho) = 0$) and the (non-squared) statistical rate achieved by VI-OPE is dominated by $O(\sum_{s,a} d^\pi(s,a)\sqrt{\frac{\text{Var}_{P_{s,a}}[V^\pi+c]}{n\cdot d^\mu(s,a)}})$. As a comparison, for the well-studied finite-horizon tabular MDP problem, the statistical limit $O(\sqrt{\sum_{h=1}^H\sum_{s,a} d_h^\pi(s,a)^2\frac{\text{Var}_{P_h}[V_{h+1}^\pi+c]}{n\cdot d_h^\mu(s,a)}}])$ has been achieved by Yin and Wang [2020], Duan et al. [2020], Kallus and Uehara [2020] which matches the previous proven lower bound [Jiang and Li, 2016]. Therefore, it is natural to conjecture that the statistical lower bound for SSP-OPE problem has the rate $O(\sqrt{\sum_{s,a} d^\pi(s,a)^2\frac{\text{Var}_{P_{s,a}}[V^\pi+c]}{n\cdot d^\mu(s,a)}})$. Our simple VI-OPE algorithm nearly matches this conjectured lower bound and only has the expectation outside of the square root. How to obtain the Carmer-Rao-style lower bound for SSP OPE problem and how to close the gap are beyond this initial attempt and we leave these as the future works.

Parameter-free. Different from the standard MDPs (*e.g.* finite-horizon, discounted), the SSP formulation generally has variable horizon length which yields no explicit bound on $\|V^\pi\|_\infty$. Consequently, most of the previous literature that study SSP problem requires the knowledge of expected running time $T^\pi/T^\star$ or $B^\pi/B_\star$, the upper bound on $\|V^\pi\|/\|V^\star\|$ (*e.g.* Tarbouriech et al. [2020], Rosenberg et al. [2020], Cohen et al. [2021], Chen and Luo [2021], Chen et al. [2021a]). In contrast, VI-OPE is fully parameter-free as it requires no prior information about neither T^π nor B^π and the main term of our bound does not explicitly scale with those parameters. Last but not least, VI-OPE does not reply on the positive cost Assumption 2.6.

4 OFFLINE LEARNING IN SSP

In this section, we consider the offline policy optimization problem. Similar to previous work, we assume the knowledge of an upper bound on the $B_\star := \|V^\star\|_\infty$, which is denoted as $\widetilde{B}$. How to deal with the case when $\widetilde{B}$ is unknown is discussed in Section 7.1. Throughout the section, we suppose Assumption 2.5 and Assumption 2.6 holds.

We introduce our algorithm in Algorithm 2. The main idea behind the algorithm is the pessimistic update of the value function via adding a bonus function to $V^{(i)}$. Here the **bonus function** $b_{s,a}(V) := \sqrt{\frac{2\widehat{c}(s,a)\iota}{n(s,a)}} + \frac{7\iota}{3n(s,a)} + \frac{\widetilde{B}}{n(s,a)} + \frac{16\widetilde{B}\iota}{3n(s,a)} + \max\{2\sqrt{\frac{\text{Var}(\widetilde{P}',V)\iota}{n(s,a)}}, 4\frac{\widetilde{B}\iota}{n(s,a)}\} + 180\sqrt{\frac{3\widetilde{T}\widetilde{B}S}{2n(s,a)n_{\min}}}(\sqrt{\widetilde{B}}+1)\iota$,[4] where $n_{\max} = \max_{s,a} n(s,a)$ and $n_{\min} = \min_{s,a}\{n(s,a): n(s,a)>0\}$.

Algorithm 2 PVI-SSP (Pessimistic Value Iteration for SSP)

Input: ϵ_{OPL}, $\mathcal{D} := \{(s_1^{(i)}, a_1^{(i)}, c_1^{(i)}, s_2^{(i)}, \ldots, s_{T_i}^{(i)})\}_{i=1}^n$. $\widetilde{B}$ and $\iota = O(\log(SA/\delta))$. $n_{\max}$ and $b_{s,a}$ see above.
for $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}'$ **do**
  Set $n(s,a) = \sum_{i=1}^n\sum_{j=1}^{T_i}\mathbb{I}(s_j^{(i)}=s, a_j^{(i)}=a)$.
  if $n(s,a)>0$ **then**
    Calculate $\widehat{c}(s,a) = \frac{\sum_{i=1}^n\sum_{j=1}^{T_i}\mathbb{I}(s_j^{(i)}=s, a_j^{(i)}=a)c_j^{(i)}}{n(s,a)}$
    $\widehat{P}(s'|s,a) = \frac{\sum_{i=1}^n\sum_{j=1}^{T_i}\mathbb{I}(s_j^{(i)}=s, a_j^{(i)}=a, s_{j+1}^{(i)}=s')}{n(s,a)}$,
  else
    $\widehat{c}(s,a) \leftarrow c_{\min}$, $\widehat{P}(s'|s,a) \leftarrow \mathbb{I}(s'=g)$.
  end if
  $\widetilde{P}'(s'|s,a) = \frac{n_{\max}}{n_{\max}+1}\widehat{P}(s'|s,a) + \frac{\mathbb{I}[s'=g]}{n_{\max}+1}$
end for
◇ Pessimistic Value Iteration for offline learning
Initialize: $V^{(-1)}(\cdot)\leftarrow\infty$, $V^{(0)}(\cdot)\leftarrow\widetilde{B}\cdot\mathbf{1}$, $i=0$.
while $\left\|V^{(i)}-V^{(i-1)}\right\|_\infty > 0(\epsilon_{\text{OPL}})$ **do**
  for $(s,a)\in\mathcal{S}\times\mathcal{A}$ **do**
    $Q^{(i+1)}(s,a) = \min\{\widehat{c}(s,a) + \widetilde{P}'_{s,a}V^{(i)} + b_{s,a}(V^{(i)}), \widetilde{B}\}$
    $V^{(i+1)}(s) = \min_a Q^{(i+1)}(s,a)$
    $i\leftarrow i+1$
  end for
end while
Calculate $\bar{\pi}(\cdot) = \text{argmin}_a Q^{(i)}(\cdot,a)$
Output: $\bar{\pi}$, $\bar{V}(\cdot) = \min_a Q^{(i)}(\cdot,a)$

The use of value iteration to approximate the underlying Bellman optimality equation $V^\star(s) = \max_{a\in\mathcal{A}}\{c(s,a) + P_{s,a}V^\star\}$, $\forall s\in\mathcal{S}'$ is natural when model components P, c are accurately estimated by $\widetilde{P}', \widehat{c}$. Moreover, comparing to VI-OPE, there are several differences for PVI-SSP. First, $\widetilde{P}'$ is chosen according to $n_{\max}$ (instead of $n(s,a)$), which makes $\widetilde{P}'$ "closer" to $\widehat{P}$ but preserves the positive one-step transition to g. More importantly, a pessimistic bonus $b_{s,a}$ is added to the value update differently at each state-action location which measures the uncertainty learnt so far from the offline data. Action with higher uncertainty are less likely to be chosen for the next update. Concretely, $\sqrt{\frac{\text{Var}(\widetilde{P}',V^{(i)})}{n}}$ measures the uncertainty of $V^{(i)}$ and $\sqrt{\frac{\widehat{c}}{n}}$ measures the un-

[4] Here we do point the design of $b_{s,a}$ requires $\widetilde{T}$ in addition to $\widetilde{B}$. However, this is not essential as (by Assumption 2.6) $\widetilde{T}$ can be bounded by $\widetilde{B}/c_{\min}$.

certainty of per-step cost $\widehat{c}$.[5] However, to guarantee proper pessimism, we require the knowledge of $\widetilde{B}$ in the design of $b_{s,a}$.

In addition, for analysis purpose we state our result under the regime where the iteration converges exactly and the output $\bar{V}$ (in Line 22) is fixed point of the operator $\widetilde{\mathcal{T}}$ (see Appendix G.2 for details). In practice, one can stop the iteration when the update difference is smaller than ϵ_{OPL}. We have the following offline learning guarantee for $\bar{\pi}$, which is our major contribution. The proof is deferred to Appendix I.

Theorem 4.1 (Offline policy learning in SSP)**.** *Denote* $d_m := \min\{\sum_{h=0}^{\infty} \xi_h^{\mu}(s,a) : s.t. \sum_{h=0}^{\infty} \xi_h^{\mu}(s,a) > 0\}$, *and* T_s^{π} *to be the expected time to hit* g *when starting from* s. *Define* $\bar{T}^{\pi} = \max_{\bar{s}\in\mathcal{S}} T_{\bar{s}}^{\pi}$. *Then when* $n \geq n_0$, *we have with probability* $1-\delta$, *the output* $\bar{\pi}$ *of Algorithm 2 is a proper policy and satisfies* ($\iota = O(\log(SA/\delta))$)

$$
\begin{aligned}
0 \leq & V^{\bar{\pi}}(s_{\text{init}}) - V^{\star}(s_{\text{init}}) \\
\leq & 4 \sum_{s,a,s\neq g} d^{\star}(s,a) \sqrt{\frac{2\text{Var}_{P_{s,a}}[V^{\star}+c]\iota}{n \cdot d^{\mu}(s,a)}} + \widetilde{O}(\frac{1}{n}),
\end{aligned}
$$

where $d_{max} = \max_{s,a} d^{\mu}(s,a)$ *and we define* $n_0 := \max\{2\frac{B_\star+1-c_{min}}{c_{min}d_{max}}, \frac{26^2\times 2S\iota(\bar{T}^\star)^2(\sqrt{B_\star}+1)^2}{d_m}, \frac{1000(\widetilde{B}+\sqrt{\bar{T}\widetilde{B}}(\sqrt{\widetilde{B}}+1))^2 S\iota\bar{T}^\star}{4B^\star(\sqrt{B^\star}+1)^2 d_m}\}$.

On guarantee for policy. Existing online SSP works measure the algorithm performance using *regret* $R_K^{\text{SSP}} := \sum_{k=1}^{K}\sum_{h=1}^{I^k} c_h^k - K \cdot \min_{\pi\in\Pi_{\text{proper}}} V^{\pi}(s_{\text{init}})$ (*e.g.* [Tarbouriech et al., 2021]) and is different from policy-based regret measurement $R_K := \sum_{k=1}^{K} V_1^{\star}(x_{k,1}) - V_1^{\pi_k}(x_{k,1})$ (*e.g.* Azar et al. [2017]) in standard RL. The notion of R_K^{SSP} provides the flexibility for policy update even within the episode (since it suffices to minimize $\sum_{h=1}^{I^k} c_h^k$), therefore unable to output a concrete stationary policy for the policy learning purpose. In contrast, Theorem 4.1 provides a policy learning result via bounding the performance of output policy $\bar{\pi}$ explicitly.

Instance-dependent bound. Prior online SSP studies focus on deriving better worst-case regret (*e.g.* the minimax rate is of order $\Theta(B_\star\sqrt{SAK})$) where the bounds are expressed by the parameters $B_\star/D, S, A$ that lack the characterization of individual instances. In offline SSP, the main term of PVI-SSP is fully expressed by the system quantities with marginal coverage $d^\star$ and d^μ, conditional variance over transition P and cost function c. This instance-adaptive result characterizes the hardness of learning better since the magnitude of the bounds change with the instances. It fully avoids the explicit use of worst-case parameters $B_\star, S, A$.

Faster convergence. When the SSP system is deterministic for both cost c and transition P, the conditional variances are always zero. In these scenarios, Theorem 4.1 automatically guarantees faster convergence rate $\widetilde{O}(1/n)$ in deterministic SSP learning. Such a feature is not enjoyed by the existing worst-case studies in online SSP as their regrets are dominated by the statistical rate $\widetilde{O}(\sqrt{K})$ even for deterministic systems.

On optimality. While instance-dependent, it is still of great interest to understand whether this result is optimal. We provide the affirmative answer by showing a (nearly) matching minimax lower bound under the single concentrability condition in the next section.

5 SSP MINIMAX LOWER BOUND

In this section, we study the statistical limit of offline policy learning in SSP. Concretely, we consider the family of problems satisfying bounded partial coverage, *i.e.* $\max_{s,a,s\neq g} \frac{d^{\pi^\star}(s,a)}{d^{\mu}(s,a)} \leq C^\star$, 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. This $C^\star$ formally defines the maximum ratio between $\pi^\star$ and μ in Assumption 2.5. Consequently, we have the following result (the full proof is in Appendix K):

Theorem 5.1. *We define the following family of SSPs:*

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

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)*

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

Theorem 5.1 reveals for the family with proper policy $\pi^\star$ and μ with bounded ratio $C^\star$, the minimax lower bound is $\Omega(B_\star\sqrt{\frac{SC^\star}{n}})$. In particular, the dominate term in Theorem 4.1 directly implies this rate (recall $\pi^\star$ is deterministic by 2.5) by the following calculation (assuming $B_\star > 1$ just like Theorem 5.1):

$$
\begin{aligned}
& \sum_{s,a,s\neq g} d^{\star}(s,a)\sqrt{\frac{\text{Var}_{P_{s,a}}[V^\star+c]}{n\cdot d^{\mu}(s,a)}} \\
= & \sum_{s,s\neq g} d^{\star}(s,\pi^\star(s))\sqrt{\frac{\text{Var}_{P_{s,\pi^\star(s)}}[V^\star+c]}{n\cdot d^{\mu}(s,\pi^\star(s))}} \\
\leq & \sqrt{\sum_{s,s\neq g}\frac{d^{\star}(s,\pi^\star(s))}{d^{\mu}(s,\pi^\star(s))}\cdot\sum_{s,s\neq g}\frac{d^{\star}(s,\pi^\star(s))\text{Var}_{P_{s,\pi^\star(s)}}[V^\star+c]}{n}} \\
\leq & \sqrt{\sum_{s,s\neq g} C^\star\cdot\frac{B_\star^2}{n}} = B_\star\sqrt{\frac{SC^\star}{n}} \quad \text{(also see Proposition I.3)}
\end{aligned}
\tag{4}
$$

where the first inequality uses CS inequality and the second one uses the key Lemma 6.1.[6] This verifies PVI-SSP is near-optimal up to the logarithmic and higher order terms.

[5] This is due to $\text{Var}(c) \leq E[c^2] \leq E[c]$ for r.v. $c \in [0,1]$.

[6] Here since $B_\star > 1$, when applying Lemma 6.1, $B_\star$ will dominate $c \in [0,1]$.

6 SKETCH OF THE ANALYSIS

In this section, we sketch the proofs of our main theorems. In particular, we focus on describing the procedure of offline policy learning Theorem 4.1. First of all, when the condition $n \geq n_0$ holds, the output $\bar{\pi}$ is proper with high probability and following this one can conduct standard decomposition:

$$V^{\bar{\pi}} - V^\star = (V^{\bar{\pi}} - \bar{V}) + (\bar{V} - V^\star)$$

where $V^\star$ is the solution of Bellman optimality operator $\mathcal{T}$ and $\bar{V}$ is the fixed point solution of the operator $\widetilde{\mathcal{T}}(V)(s) = \min_a \left\{ \min\{\widehat{c}(s,a) + \widetilde{P}_{s,a}V + b_{s,a}(V)\ , \widetilde{B}\} \right\}$ (Lem G.6). Also, $V^{\bar{\pi}}$ satisfies general Bellman equation (Lemma A.1) therefore we first decompose $V^{\bar{\pi}} - \bar{V}$ using a *simulation-lemma* style decomposition (Lemma G.8):

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

By the careful design of $b_{s,a}(\cdot)$, the pessimism guarantees $V^{\bar{\pi}} - \bar{V} \leq 0$ (Lemma I.1). For $\bar{V} - V^\star$, a similar *simulation-lemma* style SSP decomposition (Lemma G.7) follows:

$$\bar{V} - V^\star \leq \sum_{h=0}^{\infty} \sum_{\substack{s \\ s \neq g}} \xi_h^\star(s) \Big\{ (\widetilde{P}'_{s,\pi^\star(s)} - P_{s,\pi^\star(s)})\bar{V} + \widehat{c}(s,\pi^\star(s)) - c(s,\pi^\star(s)) + b_{s,\pi^\star(s)}(\bar{V}) \Big\}. \tag{5}$$

Before we proceed to explain about how to bound the residual summations, we present two new lemmas, which help characterize the key features of stochastic shortest path problem.

Lemma 6.1 (Informal version of Lemma D.3)**.** *Let T^π be the expected time of arrival to goal state g when applying proper policy π and starting from s_{init}, then*

$$T^\pi = \sum_{h=0}^{\infty} \sum_{\substack{s,a \\ s \neq g}} \xi_h^\pi(s,a) = \sum_{\substack{s,a \\ s \neq g}} d^\pi(s,a).$$

Remark 6.2. *Lemma 6.1 explicitly reflects the connection between the expected arriving time T^π and marginal coverage $d^\pi(s,a)$. Unlike finite horizon problem where d_h^π are probability measures (*e.g. *see Yin and Wang [2021]), for SSP $d^\pi(s,a)$ can be arbitrary large (for a general policy π) due to definition 3. Lemma 6.1 guarantees $d^\pi(s,a) < \infty$ for proper policy π since by Definition 2.1 $T^\pi < \infty$, and, as a result, make our bound in Theorem 4.1 valid. Note similar result is of less interests in the standard H-horizon RL since it holds trivially that $H = \sum_{h=1}^{H} \sum_{s,a} d_h^\pi(s,a)$ and, in SSP, this becomes important as we have undetermined horizon length. With Lemma 6.1, we can get away with estimating the aggregated measure $T^\pi / T^\star$ (like previous online SSP papers did) and use sub-component $d^\pi(s,a)/d^\star(s,a)$ to reflect the behaviors of individual state-action locations (to achieve more instance-dependent results).*

Lemma 6.3 (Informal version of Lemma D.4)**.** *For any probability transition matrix P, policy π, and any cost function $c \in [0,1]$ associated with $SSP(P,\pi)$. 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*

$$\sum_{h=0}^{\infty} \sum_{\substack{s,a \\ s \neq g}} \xi_h^\pi(s,a) \mathrm{Var}_{P_{s,a}}(V) \leq 2 \left\| V \right\|_\infty \cdot V(s_{\text{init}}) \leq 2 \left\| V \right\|_\infty^2 .$$

Remark 6.4. *Lemma 6.3 can be viewed as a dependence improvement result for SSP problem since it guarantees Theorem 4.1 to achieve the minimax rate via (4). More critically, it widely applies to arbitrary policies assuming the ordering condition holds for V. For instance, a direct upper bound using Lemma 6.1 will yield $T^\pi \left\| V \right\|_\infty^2$ and T^π could be very large or even ∞. In contrast, Lemma 6.3 always upper bounds by $2 \left\| V \right\|_\infty^2$ without no extra dependence. Similar result was previously derived in RL,* e.g. *Lemma 3.4 of Yin and Wang [2020] and also Ren et al. [2021], but their result only applies to V^π due the analysis via law of total variances and ours applies to all V (satisfying ordering condition) through only the telescoping sum.*

Now we go back to bounding (5). First of all, by leveraging Lemma 6.1, we are able to bound the ∞-norm of $\bar{V} - V^\star$ as (see Theorem H.1)

$$\left\| \bar{V} - V^\star \right\|_\infty \leq 30 \sqrt{\frac{\bar{T}^\star B_\star}{n d_m} \iota} (\sqrt{B_\star} + 1) \tag{6}$$

which is a crude/suboptimal bound that serves as an intermediate step for the final bound.

What give rise to instance-dependencies. Next, we apply *empirical Bernstein inequality* for structure $(\widetilde{P}'_{s,\pi^\star(s)} - P_{s,\pi^\star(s)})\bar{V}$ and $\widehat{c}(s,\pi^\star(s)) - c(s,\pi^\star(s))$ separately. In particular, since both $\bar{V}$ and $\widetilde{P}'_{s,\pi^\star(s)}$ depend on data, therefore Bernstein concentration cannot be directly applied. Informally, we can surpass this hurdle by decomposing

$$(\widetilde{P}' - P)\bar{V} = (\widetilde{P}' - P)(\bar{V} - V^\star) + (\widetilde{P}' - P)V^\star.$$

In this scenario, concentration can be readily applied to $(\widetilde{P}' - P)V^\star$ and crude bound (6) is leveraged here for bounding $(\widetilde{P}' - P)(\bar{V} - V^\star) \leq \left\| \widetilde{P}' - P \right\|_1 \left\| \bar{V} - V^\star \right\|_\infty$. As explained by Zanette and Brunskill [2019], the use of Bernstein concentration is the key for characterizing the structure of problem instance via the expression of conditional variance $\mathrm{Var}_{P_{s,a}}(V^\star)$.

On the proof for VI-OPE. At a high level, the proof for VI-OPE (Theorem 3.1) shares the same flavor as that of Theorem 4.1. Ideally, in finite horizon setting the tighter analysis could be conducted by following the pipeline of Section B.7 in Duan et al. [2020], where the dominate error of $\widehat{V}^\pi - V^\pi$ (where $\widehat{V}^\pi = \lim_{i\to\infty} V^{(i)}$ in Algorithm 1) can be decomposed as:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{h=0}^{\infty}\frac{\xi_h^\pi(s_h^{(i)},a_h^{(i)})}{\xi_h^\mu(s_h^{(i)},a_h^{(i)})}\left(Q^\pi(s_h^{(i)},a_h^{(i)}) - (c(s_h^{(i)},a_h^{(i)}) + V^\pi(s_h^{(i)},a_h^{(i)}))\right)$$

Applying Freedman's inequality for the above martingale structure, one can hope for a tighter rate $O(\sqrt{\sum_{s,a} d^\pi(s,a)^2 \frac{\mathrm{Var}_{P_{s,a}}[V^\pi + c]}{n\cdot d^\mu(s,a)}})$. However, such a procedure will have technical issue for SSP problem since: (1) SSP has stationary transition P and $n(s,a)$ is computed via collecting all the transitions that encounter s,a for tighter dependence. This breaks the sequential ordering that is needed for martingale.[7] (2) Even if we have a martingale, the martingale difference will incorporate an infinite sum that could be arbitrary large. Both facts indicate Freedman's inequality cannot be directly applied due to the technical hurdle.

Lastly, the lower bound proof uses a generalized Fano's argument (Lemma L.5), followed by reducing estimation problem to testing. The packing set of hard MDP instances is based on the modification of Rashidinejad et al. [2021] so that Gilbert-Varshamov Lemma L.1 can be applied.

7 DISCUSSIONS

7.1 THE KNOWLEDGE OF $B_\star/\widetilde{B}$

While VI-OPE (Algorithm 1) is parameter-free, our policy learning algorithm PVI-SSP (Algorithm 2) requires $\widetilde{B}$ in the pessimistic bonus design. Since $\widetilde{B}$ is a upper bound of $B_\star$, one natural idea is to use VI-OPE to provide an upper bound estimation given that a proper policy is provided. This idea is summarized as below.

Proposition 7.1 (Alternative offline learning algorithm VI-OPE+PVI-SSP)**.** *Suppose we are provided with an arbitrary proper policy π (*e.g. *some previously deployed strategy). In this scenario, one can equally halve the data $\mathcal{D}$ into $\mathcal{D}_1$ and $\mathcal{D}_2$, and use $\mathcal{D}_1$ to evaluate V^π. The ∞-norm of VI-OPE output serves as surrogate for $\widetilde{B}$ and uses as an input for computing $b_{s,a}$. Next, use $\mathcal{D}_2$ to run PVI-SSP (with calculated $b_{s,a}$).*

The above procedure will not deteriorate the theoretical guarantee since $\widetilde{B}$ is only used in $O(1/n)$ terms and the estimation error can only be higher order terms. This means we will end up with the same dominate term as Theorem 4.1.

[7] Note Duan et al. [2020] Corollary 1 considers time-inhomogeneous MDP and each P_t can be estimated stage-wisely so the decomposition forms a martingale.

7.2 ON LOWER ORDER TERMS.

In our analysis of Theorem 4.1, while the dominate $\widetilde{O}(\sqrt{1/n})$ term is near-optimal, the lower order term $\widetilde{O}(1/n)$ is not and depends on the parameters including $\widetilde{T}$, $\widetilde{B}$ and d_m (*e.g.* check the last line of (54)). In particular, if one can remove the polynomial dependence of $\widetilde{T}$, then the result is called *horizon-free* [Tarbouriech et al., 2021]. One potential approach for addressing the lower order dependence could be the recent development of robust estimation in RL [Wagenmaker et al., 2021]. As the initial attempt for offline SSP, this is beyond our scope and we leave it as the future work.

7.3 FUTURE DIRECTIONS

SSP under weaker conditions. Following previous works, we consider stochastic shortest path problem with a discrete action space $\mathcal{A}$ and non-negative cost bounded by $c \in [0,1]$. However, the convergence of SSP can hold under much weaker conditions. For instance, Bertsekas and Yu [2013] shows under *compactness and continuity condition*, *i.e.* for each state s the admissible action set $\mathcal{A}(s)$ is compact metric space and a subset of $\mathcal{A}$ where (for all s') transition $P(s'|s,\cdot)$ are continuous functions over $\mathcal{A}(s)$ and the cost function $c(s,\cdot)$ is lower semi-continuous over $\mathcal{A}(s)$, value iteration/policy iteration will still work under mild assumptions. This extends our setting (*e.g.* cost can even be negative) and how to conduct SSP learning in this case remains open.

Extension to linear MDP case. Another natural and promising generalization of the current study is the offline linear MDP for SSP problem. In the study of offline RL with linear MDPs, Jin et al. [2021] shows the provable efficiency, Zanette et al. [2021] improves the result in the *linear Bellman complete* setting and Yin et al. [2022] leverages variance-reweighting for least square objective to obtain the near-optimal result. Adopting their useful results in offline SSP problem is hopeful.

8 CONCLUSION

In this paper, we initiate the study of *offline stochastic shortest path* problem. We consider both *offline policy evaluation* (OPE) and *offline policy learning* tasks and propose the simple value-iteration-based algorithms (VI-OPE and PVI-SSP) that yield strong theoretical guarantees for both evaluation and learning tasks. To complement the discussion, we also provide an information-theoretical lower bound and it certifies PVI-SSP is minimax rate optimal. We hope our work can draw further attention for studying offline SSP setting.

References

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

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

Dimitri P Bertsekas and Huizhen Yu. Stochastic shortest path problems under weak conditions. *Lab. for Information and Decision Systems Report LIDS-P-2909, MIT*, 2013.

Liyu Chen and Haipeng Luo. Finding the stochastic shortest path with low regret: The adversarial cost and unknown transition case. In *International Conference on Machine Learning*, pages 1651–1660. PMLR, 2021.

Liyu Chen, Rahul Jain, and Haipeng Luo. Improved no-regret algorithms for stochastic shortest path with linear mdp. *arXiv preprint arXiv:2112.09859*, 2021a.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Minimax regret for stochastic shortest path with adversarial costs and known transition. In *Conference on Learning Theory*, pages 1180–1215. PMLR, 2021b.

Herman Chernoff et al. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

Alon Cohen, Yonathan Efroni, Yishay Mansour, and Aviv Rosenberg. Minimax regret for stochastic shortest path. *Advances in neural information processing systems*, 2021.

JB Diaz and Beatriz Margolis. A fixed point theorem of the alternative, for contractions on a generalized complete metric space. *Bulletin of the American Mathematical Society*, 74(2):305–309, 1968.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 8334–8342, 2020.

András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8 (10), 2007.

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-fan

Mehdi Jafarnia-Jahromi, Liyu Chen, Rahul Jain, and Haipeng Luo. Online learning for stochastic shortest path model via posterior sampling. *arXiv preprint arXiv:2106.05335*, 2021.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 652–661. JMLR. org, 2016.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *J. Mach. Learn. Res.*, 21(167):1–63, 2020.

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

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. In *Uncertainty in Artificial Intelligence*, 2019.

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

Yifei Min, Jiafan He, Tianhao Wang, and Quanquan Gu. Learning stochastic shortest path with linear function approximation. *arXiv preprint arXiv:2110.12727*, 2021.

Rémi Munos. Error bounds for approximate value iteration. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 1006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

Gergely Neu, Andras Gyorgy, and Csaba Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, pages 805–813. PMLR, 2012.

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.

Tongzheng Ren, Jialian Li, Bo Dai, Simon S Du, and Sujay Sanghavi. Nearly horizon-free offline reinforcement learning. *arXiv preprint arXiv:2103.14077*, 2021.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. *Advances in Neural Information Processing Systems*, 32, 2019.

Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pages 8210–8219. PMLR, 2020.

Mohammad Sadegh Talebi, Zhenhua Zou, Richard Combes, Alexandre Proutiere, and Mikael Johansson. Stochastic online shortest path routing: The value of feedback. *IEEE Transactions on Automatic Control*, 63(4):915–930, 2017.

Jean Tarbouriech, Evrard Garcelon, Michal Valko, Matteo Pirotta, and Alessandro Lazaric. No-regret exploration in goal-oriented reinforcement learning. In *International Conference on Machine Learning*, pages 9428–9437. PMLR, 2020.

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.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

Daniel Vial, Advait Parulekar, Sanjay Shakkottai, and R Srikant. Regret bounds for stochastic shortest path problems with linear function approximation. *arXiv preprint arXiv:2105.01593*, 2021.

Andrew Wagenmaker, Yifang Chen, Max Simchowitz, Simon S Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. *arXiv preprint arXiv:2112.03432*, 2021.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3948–3958. PMLR, 2020.

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34, 2021.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1567–1575. PMLR, 2021.

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. In *International Conference on Learning Representations*, 2022.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Ruiqi Zhang, Xuezhou Zhang, Chengzhuo Ni, and Mengdi Wang. Off-policy fitted q-evaluation with differentiable function approximators: Z-estimation and inference theory. *arXiv preprint arXiv:2202.04970*, 2022.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. *Advances in neural information processing systems*, 26, 2013.

A GENERAL BELLMAN EQUATION FOR A FIXED POLICY

In this section, we prove Proposition 2.2. 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 policies.

Lemma A.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:*

$$Q^\pi(s,a) = c(s,a) + P_{s,a}V^\pi, \quad V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^\pi(s,a)]. \tag{7}$$

Proof of Lemma A.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)

$$\lim_{T\to\infty} \sum_{s'} \mathbb{P}(s'|s,a)\left\{\mathbb{E}_\pi[\sum_{h=0}^{T-1} c(s_h,a_h)|s_0 = s']\right\} = \sum_{s'} \mathbb{P}(s'|s,a) \lim_{T\to\infty}\left\{\mathbb{E}_\pi[\sum_{h=0}^{T-1} c(s_h,a_h)|s_0 = s']\right\}. \tag{8}$$

Combine the above two equalities together, we can get

$$\begin{aligned}Q^\pi(s,a) &= c(s,a) + \sum_{s'} \mathbb{P}(s'|s,a) \lim_{T\to\infty}\left\{\mathbb{E}_\pi[\sum_{h=0}^{T-1} c(s_h,a_h)|s_0 = s']\right\} \\ &= c(s,a) + \sum_{s'} \mathbb{P}(s'|s,a)V^\pi(s').\end{aligned} \tag{9}$$

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 A.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).*

B RESULTS FOR GENERAL STOCHASTIC SHORTEST PATH PROBLEM

Lemma B.1. *For any two contraction mapping T_1 and T_2 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 $T_1V_1\geq T_2V_1$. Since V_1 is the fixed point of T_1, we know $V_1 := T_1V_1\geq T_2V_1$. By recursion, we have that

$$V_1\geq (T_2)^kV_1. \tag{10}$$

Since V_2 is the fixed point of T_2, we have

$$\lim_{k\to\infty}(T_2)^kV_1 = V_2$$

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

C CONVERGENCES FOR ALGORITHM 1

Lemma C.1. $\widehat{\mathcal{T}}^\pi:\mathbb{R}^{\mathcal{S}}\times\{0\}\to\mathbb{R}^{\mathcal{S}}\times\{0\}$ *is a contraction mapping, i.e.,* $\forall V_1,V_2\in\mathbb{R}^{\mathcal{S}'}$, $V_1(g)=V_1(g)=0$, *we have*

$$\left\|\widehat{\mathcal{T}}^\pi V_1-\widehat{\mathcal{T}}^\pi V_2\right\|_\infty\leq\rho\left\|V_1-V_2\right\|_\infty, \tag{11}$$

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 C.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

$$\widehat{\mathcal{T}}^\pi V(g)=\langle\pi(\cdot|g),\widehat{c}(g,\cdot)+\widetilde{P}_{g,\cdot}V\rangle=0. \tag{12}$$

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

$$\begin{aligned}|\widehat{\mathcal{T}}^\pi V_1(s)-\widehat{\mathcal{T}}^\pi V_2(s)| &= |\langle\pi(\cdot|s),\widetilde{P}_{s,\cdot}(V_1-V_2)\rangle|\\ &\leq \max_a|\widetilde{P}_{s,a}(V_1-V_2)|\\ &\leq \max_a|\sum_{s'\neq g}\widetilde{P}(s'|s,a)(V_1(s')-V_2(s'))|\\ &\leq \max_a(\frac{n_{s,a}}{n_{s,a}+1})|\sum_{s'\neq g}\widehat{P}(s'|s,a)(V_1(s')-V_2(s'))|\\ &\leq \max_a(\frac{n_{s,a}}{n_{s,a}+1})\left\|V_1-V_2\right\|_\infty.\end{aligned} \tag{13}$$

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

$$\left\|\widehat{\mathcal{T}}^{\pi} V_1 - \widehat{\mathcal{T}}^{\pi} V_2\right\|_{\infty} \leq \max_{\substack{s,a \\ s\neq g}} \left(\frac{n_{s,a}}{n_{s,a}+1}\right) \|V_1 - V_2\|_{\infty} \tag{14}$$

□

Lemma C.2. *$\forall \pi \in \Pi_{proper}$, define $\widehat{V}^{\pi} := \lim_{i\to\infty} V^{(i)}$ (Note by Lemma C.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 $\rho := \max_{\substack{s,a \\ s\neq g}} \left(\frac{n_{s,a}}{n_{s,a}+1}\right) < 1$)*

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

Proof of Lemma C.2. Since by definition, $\widehat{V}^{\pi} = \widehat{\mathcal{T}}^{\pi} \widehat{V}^{\pi}$

$$\begin{aligned}
\left\|\widehat{V}^{\pi}\right\|_{\infty} &= \left\|\widehat{\mathcal{T}}^{\pi} \widehat{V}^{\pi}\right\|_{\infty} && (15)\\
&\leq \max_{s,s\neq g} |\langle \pi(\cdot|s), \widehat{c}(s,\cdot)\rangle| + \max_{s,s\neq g} |\langle \pi(\cdot|s), \widetilde{P}_{s,\cdot} \widehat{V}^{\pi}\rangle| && (16)\\
&\leq \max_{s,s\neq g} |\langle \pi(\cdot|s), \widehat{c}(s,\cdot)\rangle| + \max_{\substack{s,a \\ s\neq g}} |\widetilde{P}_{s,a} \widehat{V}^{\pi}| \\
&\leq 1 + \max_{\substack{s,a \\ s\neq g}} |\widetilde{P}_{s,a} \widehat{V}^{\pi}| \\
&\leq 1 + \max_{\substack{s,a \\ s\neq g}} \{\left(\frac{n_{s,a}}{n_{s,a}+1}\right) | \sum_{s'\neq g} \widehat{P}(s'|s,a) \widehat{V}^{\pi}(s')|\} \\
&\leq 1 + \rho \left\|\widehat{V}^{\pi}\right\|_{\infty}. && (17)
\end{aligned}$$

□

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

Lemma C.3. *$\left\|\widehat{V}^{\pi} - V^{(i)}\right\|_{\infty} \leq \frac{\epsilon_{OPE}}{1-\rho}$, where $\rho := \max_{\substack{s,a \\ s\neq g}} \left(\frac{n_{s,a}}{n_{s,a}+1}\right) < 1$ as in Lemma C.1 and $V^{(i)}$ is the output of Algorithm 1.*

Proof of Lemma C.3. Using definition $\widehat{V}^{\pi} := \lim_{j\to\infty} V^{(j)}$ and the telescoping sum we obtain

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

□

Lemma C.4. *For any $V(\cdot) \in \mathbb{R}^S$ satisfying $V(g) = 0$,*

$$|(\widetilde{P}_{s,a} - \widehat{P}_{s,a})V| \leq \frac{||V||_{\infty}}{n(s,a)+1}, \qquad |\text{Var}(\widetilde{P}_{s,a}, V) - \text{Var}(\widehat{P}_{s,a}, V)| \leq \frac{2\,\|V\|_{\infty}^2\, S}{n(s,a)+1}. \tag{18}$$

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

D SOME KEY LEMMAS

D.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 $\iota := \log(10S^2A/\delta)$)

$$\mathcal{E}_1 := \left\{ \forall (s,a,s') \in (\mathcal{S} \times \mathcal{A} \times \mathcal{S}), \forall n(s,a) \geq 1 : |P(s'|s,a) - \widehat{P}(s'|s,a)| \leq \sqrt{\frac{2P(s'|s,a)\iota}{n(s,a)}} + \frac{2\iota}{3n(s,a)} \right\}$$

$$\mathcal{E}_2 := \left\{ \forall (s,a) \in (\mathcal{S} \times \mathcal{A}), \forall n(s,a) \geq 1 : |(P_{s,a} - \widehat{P}_{s,a})V| \leq \sqrt{\frac{2\mathrm{Var}(P_{s,a},V)\iota}{n(s,a)}} + \frac{2||V||_\infty \iota}{3n(s,a)} \right\}$$

$$\mathcal{E}_3 := \left\{ \forall (s,a) \in (\mathcal{S} \times \mathcal{A}), \forall n(s,a) \geq 1 : |(P_{s,a} - \widehat{P}_{s,a})V| \leq \sqrt{\frac{2\mathrm{Var}(\widehat{P}_{s,a},V)\iota}{n(s,a)}} + \frac{7||V||_\infty \iota}{3n(s,a)} \right\}$$

$$\mathcal{E}_4 := \left\{ \forall (s,a) \in (\mathcal{S} \times \mathcal{A}), \forall n(s,a) \geq 1 : |\widehat{c}(s,a) - c(s,a)| \leq \sqrt{\frac{2\mathrm{Var}_c(s,a)\iota}{n(s,a)}} + \frac{2\iota}{3n(s,a)} \right\}$$

$$\mathcal{E}_5 := \left\{ \forall (s,a) \in (\mathcal{S} \times \mathcal{A}), \forall n(s,a) \geq 1 : |\widehat{c}(s,a) - c(s,a)| \leq \sqrt{\frac{2\widehat{c}(s,a)\iota}{n(s,a)}} + \frac{7\iota}{3n(s,a)} \right\}. \tag{19}$$

Lemma D.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 L.4, we have that for each fixed (s,a), the event $|(P_{s,a} - \widehat{P}_{s,a})V| \leq \sqrt{\frac{2\mathrm{Var}(\widehat{P}_{s,a},V)\iota}{n(s,a)}} + \frac{7||V||_\infty \iota}{3n(s,a)}$ 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 L.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. □

D.2 VALUE DECOMPOSITION LEMMA

Lemma D.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 $\bar{s}$:*

$$\widehat{V}^\pi(\bar{s}) - V^\pi(\bar{s}) = \sum_{h=0}^{\infty} \sum_{\substack{s,a \\ s\neq g}} \xi^\pi_{h,\bar{s}}(s,a)\{(\widehat{c} - c)(s,a) + (\tilde{P}_{s,a} - P_{s,a})\widehat{V}^\pi\} \tag{20}$$

Proof of Lemma D.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

$$\widehat{V}^\pi(\bar{s}) - V^\pi(\bar{s}) = \sum_{h=0}^{H} \sum_{\substack{s,a \\ s\neq g}} \xi^\pi_{h,\bar{s}}(s,a)\{(\widehat{c}-c)(s,a) + (\tilde{P}_{s,a} - P_{s,a})\widehat{V}^\pi\} + \sum_{\substack{s \\ s\neq g}} \xi^\pi_{H+1,\bar{s}}(s)(\widehat{V}^\pi(s) - V^\pi(s)). \tag{21}$$

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 D.3 $\lim_{H\to+\infty} P^\pi_{\bar{s}}(s_H \neq g) = 0$. From Lemma C.2, we have $||\widehat{V}^\pi||_\infty \leq \infty$. It follows that

$$\lim_{H\to+\infty} \sum_{\substack{s \\ s\neq g}} \xi^\pi_{H+1,\bar{s}}(s)(\widehat{V}^\pi(s) - V^\pi(s)) = 0 \tag{22}$$

By taking H to infinity in Equation (21), we conclude the proof. □

D.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 D.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 D.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 D.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*

$$\sum_{h=0}^\infty \sum_{\substack{s,a \\ s\neq g}} \xi^\pi_h(s,a)\mathrm{Var}(P_{s,a}, V) \leq 2\left\|V\right\|_\infty \sum_{\substack{s \\ s\neq g}} \xi_0(s)V(s) \leq 2\left\|V\right\|^2_\infty. \tag{23}$$

Proof.

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

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. □

E CRUDE EVALUATION BOUND

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

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

Proof. We denote $\bar{s}\in\mathcal{S}$ to be any initial state. From Lemma D.2, we have (here $\xi_{h,\bar{s}}^{\pi}(s,a)$ is the marginal state-action probability when starting from state $\bar{s}$ and following π)

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

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

We bound the above three parts one by one. First of all, by Bernstein inequality, Lemma L.6 and union bound, with probability $1-\delta$,

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

(i) uses Lemma L.6 and the last inequality uses Lemma D.3.

For the second part, note

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

where the first inequality uses Lemma J.3 and Lemma C.4, the second inequality uses Lemma L.6, the third inequality uses $\mathrm{Var}(\cdot)\leq\|\cdot\|^2_{\infty}$, the fourth and fifth inequality use Lemma D.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

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

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

where the first inequality uses Lemma C.4 and Bernstein inequality, the second inequality uses Lemma L.6 and the last one uses Lemma D.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

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

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

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

□

F PROOF OF THEOREM 3.1

Theorem F.1 (Restatement of Theorem 3.1)**.** *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 $n\geq\max\{\frac{49S\iota}{9d_m},64(\bar{T}^{\pi})^2\frac{S\iota}{d_m},C\cdot\iota/d_m\}$, 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 C.3,

$$
\left|V^{\pi}(s_{\text{init}})-V^{(i)}(s_{\text{init}})\right|\leq\left|V^{\pi}(s_{\text{init}})-\widehat{V}^{\pi}(s_{\text{init}})\right|+\left|\widehat{V}^{\pi}(s_{\text{init}})-V^{(i)}(s_{\text{init}})\right|\leq\left|V^{\pi}(s_{\text{init}})-\widehat{V}^{\pi}(s_{\text{init}})\right|+\frac{\epsilon_{\text{OPE}}}{1-\rho},\quad(24)
$$

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

From Lemma D.2, we have

$$
\left|V^{\pi}(s_{\text{init}})-\widehat{V}^{\pi}(s_{\text{init}})\right|=\left|\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi^{\pi}_h(s,a)\{(\widehat{c}-c)(s,a)+(\tilde{P}_{s,a}-P_{s,a})\widehat{V}^{\pi}\}\right|
$$

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

We bound the above three parts one by one. First of all, by Bernstein inequality and Lemma L.6 together with union bound, with probability $1-\delta$,

$$
\begin{aligned}
&\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h^\pi(s,a)|(\widehat{c}-c)(s,a)| \leq \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h^\pi(s,a)\left[2\sqrt{\frac{\mathrm{Var}_c(s,a)\iota}{n(s,a)}} + \frac{4\iota}{3n(s,a)}\right] \\
&\overset{\text{(i)}}{\leq} \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h^\pi(s,a)\left[2\sqrt{\frac{2\mathrm{Var}_c(s,a)\iota}{n\sum_{h=1}^{\infty}\xi_h^\mu(s,a)}} + \frac{8\iota}{3n\sum_{h=1}^{\infty}\xi_h^\mu(s,a)}\right] \\
&\overset{\text{(ii)}}{\leq} \sqrt{\sum_{\substack{s,a\\ s\neq g}} \frac{\sum_{h=1}^{\infty}\xi_h^\pi(s,a)}{\sum_{h=1}^{\infty}\xi_h^\mu(s,a)}}\sqrt{\sum_{\substack{s,a\\ s\neq g}}\sum_{h=1}^{\infty}\xi_h^\pi(s,a)\mathrm{Var}_c(s,a)\frac{\iota}{n}} + \left(\sum_{s,a,s\neq g}\frac{d^\pi(s,a)}{d^\mu(s,a)}\right)\frac{8\iota}{3n} \\
&\overset{\text{(iii)}}{\leq} \sqrt{\sum_{\substack{s,a\\ s\neq g}} \frac{d^\pi(s,a)}{d^\mu(s,a)}\cdot\frac{V^\pi(s_{\text{init}})\cdot\iota}{n}} + \left(\sum_{s,a,s\neq g}\frac{d^\pi(s,a)}{d^\mu(s,a)}\right)\frac{8\iota}{3n},
\end{aligned}
\tag{25}
$$

(i) uses Lemma L.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

$$
\begin{aligned}
&\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h(s,a)|(\tilde{P}_{s,a}-P_{s,a})(\widehat{V}^\pi - V^\pi)| \\
&\leq \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h(s,a)\left[\sqrt{\frac{2S\cdot\mathrm{Var}(P_{s,a},\widehat{V}^\pi - V^\pi)\iota_{s,a}}{n(s,a)}} + \frac{4\left\|\widehat{V}^\pi - V^\pi\right\|_\infty S\iota_{s,a}}{3n(s,a)} + \frac{\left\|\widehat{V}^\pi - V^\pi\right\|_\infty}{n(s,a)+1}\right] \\
&\leq \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h(s,a)\left[\sqrt{\frac{4S\cdot\mathrm{Var}(P_{s,a},\widehat{V}^\pi - V^\pi)\iota_{s,a}}{n\sum_{h=1}^{\infty}\xi_h^\pi(s,a)}} + \frac{14\left\|\widehat{V}^\pi - V^\pi\right\|_\infty S\iota_{s,a}}{3n\sum_{h=1}^{\infty}\xi_h^\pi(s,a)}\right] \\
&\leq \sqrt{\sum_{\substack{s,a\\ s\neq g}} \frac{\sum_{h=0}^{\infty}\xi_h^\pi(s,a)}{\sum_{h=0}^{\infty}\xi_h^\mu(s,a)}}\sqrt{4S\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h(s,a)\left\|\widehat{V}^\pi - V^\pi\right\|^2\frac{\iota}{n}} + \frac{14\left\|\widehat{V}^\pi - V^\pi\right\|_\infty S\iota}{3n}\sum_{s,a,s\neq g}\frac{d^\pi(s,a)}{d^\mu(s,a)} \\
&\leq \sqrt{\sum_{\substack{s,a\\ s\neq g}} \frac{d^\pi(s,a)}{d^\mu(s,a)}\cdot 4ST^\pi\left\|\widehat{V}^\pi - V^\pi\right\|_\infty^2\cdot\frac{\iota}{n}} + \frac{14\left\|\widehat{V}^\pi - V^\pi\right\|_\infty S\iota}{3n}\sum_{s,a,s\neq g}\frac{d^\pi(s,a)}{d^\mu(s,a)} \\
&\leq 4\sqrt{\sum_{\substack{s,a\\ s\neq g}} \frac{d^\pi(s,a)}{d^\mu(s,a)}\cdot ST^\pi\left\|\widehat{V}^\pi - V^\pi\right\|_\infty^2\cdot\frac{\iota}{n}}
\end{aligned}
\tag{26}
$$

where the first inequality uses Lemma J.3 and Lemma C.4, the second inequality uses Lemma L.6, the third inequality uses $\mathrm{Var}(\cdot) \leq \|\cdot\|_\infty^2$ and CS inequality, the fourth inequality use Lemma D.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

$$
\begin{aligned}
&\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^{\pi}(s,a)|(\widetilde{P}_{s,a}-P_{s,a})V^{\pi}| \\
\leq&\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^{\pi}(s,a)\left[2\sqrt{\frac{\mathrm{Var}(P_{s,a},V^{\pi})\iota}{n(s,a)}}+\frac{4\left\|V^{\pi}\right\|_{\infty}\iota}{3n(s,a)}+\frac{\left\|V^{\pi}\right\|_{\infty}}{n(s,a)}\right] \\
\leq&\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^{\pi}(s,a)\left[2\sqrt{\frac{2\mathrm{Var}(P_{s,a},V^{\pi})\iota}{n\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)}}+\frac{14\left\|V^{\pi}\right\|_{\infty}\iota}{3n\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)}\right] \\
(\leq&\sqrt{8\sum_{\substack{s,a\\ s\neq g}}\frac{\sum_{h=1}^{\infty}\xi_h^{\pi}(s,a)}{\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)}\cdot\sum_{s,a,s\neq g}\sum_{h=1}^{\infty}\xi_h^{\pi}(s,a)\mathrm{Var}(P_{s,a},V^{\pi})\frac{\iota}{n}}+\frac{14\left\|V^{\pi}\right\|_{\infty}\iota}{3n}\cdot\sum_{s,a,s\neq g}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)} \\
\leq&\sqrt{8\sum_{\substack{s,a\\ s\neq g}}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)}\cdot\frac{\left\|V^{\pi}\right\|_{\infty}^{2}\iota}{n}}+\frac{14\left\|V^{\pi}\right\|_{\infty}\iota}{3n}\cdot\sum_{s,a,s\neq g}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)} \\
\leq&4\sqrt{2\sum_{\substack{s,a\\ s\neq g}}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)}\cdot\frac{\left\|V^{\pi}\right\|_{\infty}^{2}\iota}{n}}),
\end{aligned}
\tag{27}
$$

where the first inequality uses Lemma J.3 and Lemma C.4, the second inequality uses Lemma L.6. The third....

Combine Equation (25), (26) and (27) together, we obtain

$$
\begin{aligned}
&\left|V^{\pi}(s_{\text{init}})-\widehat{V}^{\pi}(s_{\text{init}})\right|\leq\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^{\pi}(s,a)\left[2\sqrt{\frac{2\mathrm{Var}(P_{s,a},V^{\pi})\iota}{n\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)}}+\frac{14\left\|V^{\pi}\right\|_{\infty}\iota}{3n\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)}\right] \\
&+\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^{\pi}(s,a)\left[2\sqrt{\frac{2\mathrm{Var}_c(s,a)\iota}{n\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)}}+\frac{8\iota}{3n\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)}\right] \\
&+4\sqrt{\sum_{\substack{s,a\\ s\neq g}}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)}\cdot ST^{\pi}\left\|\widehat{V}^{\pi}-V^{\pi}\right\|_{\infty}^{2}\cdot\frac{\iota}{n}} \\
\leq&4\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^{\pi}(s,a)\sqrt{\frac{2[\mathrm{Var}(P_{s,a},V^{\pi})+\mathrm{Var}_c(s,a)]\iota}{n\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)}}+\sum_{\substack{s,a\\ s\neq g}}\frac{22\sum_{h=1}^{\infty}\xi_h^{\pi}(s,a)\left\|V^{\pi}\right\|_{\infty}\iota}{3n\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)} \\
&+4\sqrt{\sum_{\substack{s,a\\ s\neq g}}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)}\cdot ST^{\pi}\cdot\frac{(\bar{T}^{\pi})^{2}\cdot\max_{s,a}\mathrm{Var}_c(s,a)\iota+\bar{T}^{\pi}\left\|V^{\pi}\right\|_{\infty}^{2}\iota}{n\cdot d_m}\cdot\frac{\iota}{n}}+\widetilde{O}\left(\frac{1}{n^{3/2}}\right) \\
=&4\sum_{s,a,s\neq g}d^{\pi}(s,a)\sqrt{\frac{2\mathrm{Var}_{P_{s,a}}[V^{\pi}+c]}{nd^{\mu}(s,a)}}+\widetilde{O}(\frac{1}{n})
\end{aligned}
$$

where the only inequality uses Theorem E.1. Combining this with (24) we finish the proof of Theorem 3.1. □

G 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 2. 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 G.6.

G.1 AUXILIARY LEMMAS

Lemma G.1. *Denote the limit of sequence $V^{(i)}$ in Algorithm 2 as $\bar{V}$, we have that*

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

Proof. First of all, by Lemma G.6, we know $\bar{V}$ exists. Next, from the Algorithm 2, we can get that

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

and thus

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

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

Lemma G.2. *For any $V(\cdot) \in \mathbb{R}^S$ satisfying $V(g) = 0$,*

$$|(\widetilde{P}'_{s,a} - \widehat{P}_{s,a})V| \leq \frac{||V||_\infty}{n_{max}+1}, \qquad |\mathrm{Var}(\widetilde{P}'_{s,a}, V) - \mathrm{Var}(\widehat{P}_{s,a}, V)| \leq \frac{2\left\|V\right\|_\infty^2}{n_{max}+1}. \tag{28}$$

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

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

Then we prove the second inequality. Similarly,

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

□

Lemma G.3. *When $n \geq \frac{S\iota}{2d_m}$, with probability at least $1-\delta$, we have that for any state action pair (s,a),*

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

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

$$|(P_{s,a} - \widetilde{P}'_{s,a})\bar{V}| \leq |(\widehat{P}_{s,a} - \widetilde{P}'_{s,a})\bar{V}| + |(P_{s,a} - \widehat{P}_{s,a})(\bar{V} - V^*)| + |(P_{s,a} - \widehat{P}_{s,a})V^*| \tag{29}$$

Then we bound the above three terms one by one. From Lemma G.2, we have

$$|(\widehat{P}_{s,a} - \widetilde{P}_{s,a})\bar{V}| \leq \frac{\left\|\bar{V}\right\|_\infty}{n_{max}+1} \leq \frac{\widetilde{B}}{n_{max}+1}. \tag{30}$$

For the second term, we have

$$
\begin{aligned}
|(P_{s,a}-\widehat{P}_{s,a})(\bar{V}-V^*)| &\le \sqrt{\frac{2S\mathrm{Var}(P_{s,a},\bar{V}-V^*)\iota}{n(s,a)}}+\frac{2\left\|\bar{V}-V^*\right\|_\infty S\iota}{3n(s,a)} \\
&\le \sqrt{\frac{2S\iota}{n(s,a)}}\left\|\bar{V}-V^*\right\|_\infty+\frac{\left\|\bar{V}-V^*\right\|_\infty S\iota}{n(s,a)},
\end{aligned} \tag{31}
$$

where the first inequality holds because of lemma J.3. For the last term, we have that

$$
\begin{aligned}
|(P_{s,a}-\widehat{P}_{s,a})V^*| &\overset{\text{(i)}}{\le} \sqrt{\frac{2\mathrm{Var}(\widehat{P}_{s,a},V^*)\iota}{n(s,a)}}+\frac{7||V^*||_\infty\iota}{3n(s,a)} \\
&\overset{\text{(ii)}}{\le} 2\sqrt{\frac{\mathrm{Var}(\widehat{P}_{s,a},V^*-\bar{V})\iota}{n(s,a)}}+2\sqrt{\frac{\mathrm{Var}(\widehat{P}_{s,a},\bar{V})\iota}{n(s,a)}}+\frac{7\widetilde{B}\iota}{3n(s,a)} \\
&\le 2\sqrt{\frac{\iota}{n(s,a)}})\left\|\bar{V}-V^*\right\|_\infty+2\sqrt{\frac{\mathrm{Var}(\widehat{P}_{s,a},\bar{V})\iota}{n(s,a)}}+\frac{7\widetilde{B}\iota}{3n(s,a)} \\
&\overset{\text{(iii)}}{\le} 2\sqrt{\frac{\iota}{n(s,a)}})\left\|\bar{V}-V^*\right\|_\infty+2\sqrt{\frac{\mathrm{Var}(\widetilde{P}'_{s,a},\bar{V})\iota}{n(s,a)}}+\frac{2\sqrt{2\iota}||\bar{V}||_\infty}{n(s,a)}+\frac{7\widetilde{B}\iota}{3n(s,a)} \\
&\le 2\sqrt{\frac{\iota}{n(s,a)}})\left\|\bar{V}-V^*\right\|_\infty+2\sqrt{\frac{\mathrm{Var}(\widetilde{P}'_{s,a},\bar{V})\iota}{n(s,a)}}+\frac{16\widetilde{B}\iota}{3n(s,a)},
\end{aligned} \tag{32}
$$

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 G.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

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

Since with probability $1-\delta$, the Chernoff bound holds, i.e., $n(s,a)\ge\frac{1}{2}nd_m$. When $n\ge\frac{S\iota}{2d_m}$, we have

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

□

Lemma G.4. *Define function $f:\mathbb{R}^S\times\mathbb{R}^S\to\mathbb{R}$ 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

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

(33)

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

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

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

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

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

Combine this inequality with (33), we complete the proof. □

Lemma G.5. *When* $n \geq \max\{2\frac{B_*+1-c_{min}}{c_{min}d_{max}}, \frac{26^2\times 2S\iota(\bar{T}^*)^2(\sqrt{B_*}+1)^2}{d_m}, \frac{1000(\widetilde{B}+\sqrt{\widetilde{T}\widetilde{B}}(\sqrt{\widetilde{B}}+1))^2 S\iota\bar{T}^*}{4B^*(\sqrt{B^*}+1)^2 d_m}\}$, $\bar{\pi}$ *is a proper policy (Recall* $\bar{\pi}$ *is the output of Algorithm 2).*

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

$$\sum_{h=0}^{\infty}\sum_{s\in\mathcal{S}_0}\tilde{\xi}_{h,e}(s) = \sum_{h=0}^{\infty}(\frac{n_{max}}{n_{max}+1})^h = n_{max}+1. \tag{34}$$

Then we have

$$\sum_{h=0}^{\infty}\sum_{s\in\mathcal{S}_0}\tilde{\xi}_{h,e}(s) = n_{max}+1 \geq \frac{1}{2}nd_{max}+1, \tag{35}$$

where the last inequality holds with probability $1-\delta$ by Lemma L.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

$$\widetilde{V}^{\bar{\pi}}(e) \geq \sum_{h=0}^{\infty}\mathbb{E}_{\widetilde{P}',\bar{\pi}}[c_{min}]$$

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

Apply Lemma A.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 B.1 we have $\bar{V}(s) \geq \widetilde{V}^{\bar{\pi}}(s)$ (since $b(\bar{V})$ is non-negative). Then we get

$$\bar{V}(e) \geq \widetilde{V}^{\bar{\pi}}(e) \geq c_{min}(\frac{1}{2}nd_{max}+1). \tag{36}$$

From Lemma H.1 (note Lemma H.1 only bounds $\bar{V}$ and $V^\star$ and has nothing to do with $\bar{\pi}$), we have that with probability $1-\delta$, when $n \geq \max\{\frac{26^2\times 2S\iota(\bar{T}^*)^2}{d_m}, \frac{1000(\widetilde{B}+\sqrt{\widetilde{\mathcal{T}}\widetilde{B}}(\sqrt{\widetilde{B}}+1))^2 S\iota\bar{T}^*}{4B^*(\sqrt{B^*}+1)^2 d_m}\}$, and when $n \geq \frac{\bar{T}^* B_*\iota(\sqrt{B^*}+1)^2}{d_m}$, we have $\bar{V}(e) \leq V^*(e)+1 \leq B_*+1$. Combine this inequality with (36), we can get $n \leq 2\frac{B_*+1-c_{\min}}{c_{\min}d_{max}}$, which contradicts with the assumptions in the lemma. □

G.2 CONVERGENCE OF PESSIMISTIC VALUE ITERATION IN ALGORITHM 2

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 $\widetilde{T}$ is a contraction mapping.

Lemma G.6. *$\widetilde{\mathcal{T}} : \mathbb{R}^{\mathcal{S}} \times \{0\} \to \mathbb{R}^{\mathcal{S}} \times \{0\}$ is a contraction mapping, i.e., $\forall V_1, V_2 \in \mathbb{R}^{\mathcal{S}'}$, $V_1(g) = V_1(g) = 0$, we have*

$$\left\|\widetilde{\mathcal{T}}V_1 - \widetilde{\mathcal{T}}V_2\right\|_\infty \leq \gamma \left\|V_1 - V_2\right\|_\infty, \tag{37}$$

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

Proof of Lemma G.6.

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

where (i) comes from Lemma L.7. (ii) is due to the mean value theorem. (iii) uses the result in Lemma G.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.

G.3 REGRET DECOMPOSITION LEMMA FOR POLICY OPTIMIZATION

Lemma G.7. *Suppose $\bar{V}$ is the limit of the sequence $V^{(i)}$ in Algorithm 2, we have the following decomposition lemma.*

$$\bar{V} - V^* \leq \sum_{h=0}^{\infty} \sum_{\substack{s \\ s \neq g}} \xi_h^*(s)\{(\widetilde{P}'_{s,\pi^*(s)} - P_{s,\pi^*(s)})\bar{V} + \hat{c}(s,\pi^*(s)) - c(s,\pi^*(s)) + b_{s,\pi^*(s)}(\bar{V})\} \tag{38}$$

Proof.

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

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

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

(i) follows from the fact that $\bar{V}(s) = \min_a \bar{Q}(s,a) \leq \bar{Q}(s,\pi^*(s))$. (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

$$\begin{aligned}
\bar{V} - V^* \leq \sum_{h=0}^{H} \sum_{s,s\neq g} &\xi_h^*(s)\{(\widetilde{P}'_{s,\pi^*(s)} - P_{s,\pi^*(s)})\bar{V} + \hat{c}(s,\pi^*(s)) - c(s,\pi^*(s)) + b_{s,\pi^*(s)}(\bar{V})\} \\
&+ \sum_{s,s\neq g} \xi_{H+1}^*(s)(\bar{V} - V^*),
\end{aligned} \tag{39}$$

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

Lemma G.8. *When $n \geq \max\{2\frac{B_*+1-c_{min}}{c_{min}d_{max}}, \frac{26^2\times 2S\iota(\bar{T}^*)^2}{d_m}, \frac{1000(\widetilde{B}+\sqrt{\widetilde{T}\widetilde{B}}(\sqrt{\widetilde{B}}+1))^2 S\iota\bar{T}^*}{4B^*(\sqrt{B^*}+1)^2 d_m}\}$, we have*

$$V^{\bar{\pi}} - \bar{V} = \sum_{h=0}^{\infty} \sum_{\substack{s \\ s \neq g}} \xi_h^{\bar{\pi}}(s)\{(P_{s,\bar{\pi}(s)} - \widetilde{P}'_{s,\bar{\pi}(s)})\bar{V} + c(s,\bar{\pi}(s)) - \hat{c}(s,\bar{\pi}(s)) - b_{s,\bar{\pi}(s)}(\bar{V})\}. \tag{40}$$

Proof. First of all, by the condition we have $\bar{\pi}$ is a proper policy by Lemma G.5. We prove by recursion formula.

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

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

where the first inequality uses the Bellman equation for policy $\bar\pi$, which follows from Lemma A.1. By recursion, we have

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

From Lemma G.5, we have that when $n \geq \max\{\frac{26^2\times 2S\iota(\bar T^*)^2}{d_m}, \frac{1000(\widetilde{B}+\sqrt{\widetilde{T}\widetilde{B}}(\sqrt{\widetilde{B}}+1))^2 S\iota\bar T^*}{4B^*(\sqrt{B^*}+1)^2 d_m}\}$, $\bar\pi$ 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. □

H CRUDE OPTIMIZATION BOUND

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

Theorem H.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_s^π to be the expected time to hit g when starting from s with $\bar T^\pi = \max_s T_s^\pi$. Then when $n \geq \max\{\frac{26^2\times 2S\iota(\bar T^*)^2}{d_m}, \frac{1000(\widetilde{B}+\sqrt{\widetilde{T}\widetilde{B}}(\sqrt{\widetilde{B}}+1))^2 S\iota\bar T^*}{4B^*(\sqrt{B^*}+1)^2 d_m}\}$, we have with probability $1-\delta$,*

$$\left\|\bar V - V^*\right\|_\infty \leq 30\sqrt{\frac{\bar T^* B_*\iota}{nd_m}}(\sqrt{B_*}+1) \tag{41}$$

Proof. From Lemma G.7, we have (by choosing $\xi_0 = \mathbf{1}[s_0 = \bar s]$)

$$|\bar V(\bar s) - V^*(\bar s)| \leq \sum_{h=0}^\infty \sum_{\substack{s\\ s\neq g}} \xi_{h,\bar s}^*(s)\{(\widetilde{P}_{s,\pi^*(s)} - P_{s,\pi^*(s)})\bar V + \widehat{c}(s,\pi^*(s)) - c(s,\pi^*(s)) + b_{s,\pi^*(s)}(\bar V)\}$$

For the first term, we can bound it by Lemma G.3

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

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

$$|\bar V(\bar s) - V^*(\bar s)| \leq \sum_{h=0}^\infty \sum_{\substack{s\\ s\neq g}} \xi_{h,\bar s}^*(s)\{\frac{2\widetilde{B}}{n(s,\pi^*(s))} + \frac{32\widetilde{B}\iota}{3n(s,\pi^*(s))} + 2\sqrt{\frac{\mathrm{Var}(\widetilde{P}',\bar V)\iota}{n(s,\pi^*(s))}} + 6\sqrt{\frac{S\iota}{n(s,\pi^*(s))}}\left\|V - V^*\right\|_\infty$$
$$+ 2\sqrt{\frac{2\widehat{c}(s,\pi^*(s))\iota}{n(s,\pi^*(s))}} + \frac{14\iota}{3n(s,\pi^*(s))} + \max\{2\sqrt{\frac{\mathrm{Var}(\widetilde{P},V)\iota}{n(s,\pi^*(s))}}, 4\frac{\widetilde{B}\iota}{n(s,\pi^*(s))}\}$$
$$+ 180\sqrt{\frac{3\widetilde{T}\widetilde{B}S}{2n(s,\pi^*(s))n_{\min}}}(\sqrt{\widetilde{B}}+1)\iota\}$$

$$\overset{\text{(i)}}{\leq} \sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}} \xi^*_{h,\bar{s}}(s)\{\frac{2\widetilde{B}}{n(s,\pi^*(s))}+\frac{44\widetilde{B}\iota}{3n(s,\pi^*(s))}+4\sqrt{\frac{\mathrm{Var}(\widetilde{P}',\bar{V})\iota}{n(s,\pi^*(s))}}+6\sqrt{\frac{S\iota}{n(s,\pi^*(s))}}\left\|V-V^*\right\|_\infty$$

$$+2\sqrt{\frac{2\widehat{c}(s,\pi^*(s))\iota}{n(s,\pi^*(s))}}+\frac{14\iota}{3n(s,\pi^*(s))}+180\sqrt{\frac{3\widetilde{T}\widetilde{B}S}{2n(s,\pi^*(s))n_{\min}}}(\sqrt{\widetilde{B}}+1)\iota\},$$

where (i) uses the inequality that $\max\{a,b\}\leq a+b$. 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

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

where (i) follows from Lemma G.2. (ii) uses Lemma J.1. (iii) uses the fact that $\mathrm{Var}(X+Y)\leq 2\mathrm{Var}(X)+2\mathrm{Var}(Y)$. (iv) uses the fact that $\left\|\bar{V}\right\|_\infty\leq\widetilde{B}$. Then we can have

$$\begin{aligned}
|\bar{V}(\bar{s})-V^*(\bar{s})| \leq& \sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}} \xi^*_{h,\bar{s}}(s)\{\frac{2\widetilde{B}}{n(s,\pi^*(s))}+\frac{44\widetilde{B}\iota}{3n(s,\pi^*(s))}+4\sqrt{\frac{3\iota}{n(s,\pi^*(s))}}\left\|\bar{V}-V^*\right\|_\infty\\
&+4\sqrt{\frac{3\mathrm{Var}(P_{s,\pi^*(s)},V^*)\iota}{n(s,\pi^*(s))}}+4\sqrt{\frac{2S\iota(\iota+1)}{n^2(s,\pi^*(s))}}\widetilde{B}\\
&+6\sqrt{\frac{S\iota}{n(s,\pi^*(s))}}\left\|V-V^*\right\|_\infty+2\sqrt{\frac{2\widehat{c}(s,\pi^*(s))\iota}{n(s,\pi^*(s))}}+\frac{14\iota}{3n(s,\pi^*(s))}+b_0(s,a)\}\\
\overset{\text{(i)}}{\leq}& \sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}} \xi^*_{h,\bar{s}}(s)\{\frac{30\max\{\widetilde{B},1\}\sqrt{S}\iota}{n(s,\pi^*(s))}+13\sqrt{\frac{S\iota}{n(s,\pi^*(s))}}\left\|V-V^*\right\|_\infty\\
&+4\sqrt{\frac{3\mathrm{Var}(P_{s,\pi^*(s)},V^*)\iota}{n(s,\pi^*(s))}}+2\sqrt{\frac{2\widehat{c}(s,\pi^*(s))\iota}{n(s,\pi^*(s))}}+b_0(s,a)\}\\
\overset{\text{(ii)}}{\leq}& \sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}} \xi^*_{h,\bar{s}}(s)\{\frac{34\max\{\widetilde{B},1\}\sqrt{S}\iota}{n(s,\pi^*(s))}+13\sqrt{\frac{S\iota}{n(s,\pi^*(s))}}\left\|\bar{V}-V^*\right\|_\infty\\
&+4\sqrt{\frac{3\mathrm{Var}(P_{s,\pi^*(s)},V^*)\iota}{n(s,\pi^*(s))}}+2\sqrt{\frac{3c(s,\pi^*(s))\iota}{n(s,\pi^*(s))}}+b_0(s,a)\},
\end{aligned}$$

where (i) uses the assumption $\iota\geq 1$ and that $S\geq 1$. (ii) holds because of Lemma J.2.

Then we have

$$|\bar{V}(\bar{s})-V^*(\bar{s})| \overset{\text{(i)}}{\leq} \sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}} \xi^*_{h,\bar{s}}(s)\{\frac{68\max\{\widetilde{B},1\}\sqrt{S}\iota}{nd_m}+13\sqrt{2}\sqrt{\frac{S\iota}{nd_m}}\left\|V-V^*\right\|_\infty$$

$$
\begin{aligned}
&+4\sqrt{\frac{6\mathrm{Var}(P_{s,\pi^*(s)},V^*)\iota}{nd_m}}+2\sqrt{\frac{6c(s,\pi^*(s))\iota}{nd_m}}+\bar{b}_0\}\\
&\overset{\text{(ii)}}{\leq}\frac{68T^*_{\bar{s}}\max\{\widetilde{B},1\}\sqrt{S}\iota}{nd_m}+13\sqrt{\frac{2S\iota}{nd_m}}\left\|V-V^*\right\|_\infty T^*_{\bar{s}}+T^*_{\bar{s}}\bar{b}_0\\
&\quad+\sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}}\xi^*_{h,\bar{s}}(s)\{4\sqrt{\frac{6\mathrm{Var}(P_{s,\pi^*(s)},V^*)\iota}{nd_m}}+2\sqrt{\frac{6c(s,\pi^*(s))\iota}{nd_m}}\},
\end{aligned}
$$

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 L.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

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

where (i) uses the Cauchy-Schwartz inequality. (ii) uses the result in Lemma D.3 and Lemma J.1. Similarly, we have

$$
\begin{aligned}
\sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}}\xi^*_{h,\bar{s}}(s)\{2\sqrt{\frac{6c(s,\pi^*(s))\iota}{nd_m}}\}&\leq2\sqrt{\sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}}\xi^*_{h,\bar{s}}(s)}\sqrt{\frac{6\sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}}\xi^*_{h,\bar{s}}(s)c(s,\pi^*(s))\iota}{nd_m}}\\
&\leq2\sqrt{T^*_{\bar{s}}}\sqrt{\frac{6\left\|V^*\right\|_\infty\iota}{nd_m}}.
\end{aligned}
$$

Combine the above together, we get

$$
\begin{aligned}
|\bar{V}(\bar{s})-V^*(\bar{s})|\leq&\frac{68T^*_{\bar{s}}\max\{\widetilde{B},1\}\sqrt{S}\iota}{nd_m}+13\sqrt{\frac{2S\iota}{nd_m}}\left\|V-V^*\right\|_\infty T^*_{\bar{s}}+T^*_{\bar{s}}\bar{b}_0\\
&+4\sqrt{T^*_{\bar{s}}}\sqrt{\frac{12\iota}{nd_m}}\left\|V^*\right\|_\infty+2\sqrt{T^*_{\bar{s}}}\sqrt{\frac{6\left\|V^*\right\|_\infty\iota}{nd_m}}\\
\leq&\frac{68\bar{T}^*\max\{\widetilde{B},1\}\sqrt{S}\iota}{nd_m}+13\sqrt{\frac{2S\iota}{nd_m}}\left\|V-V^*\right\|_\infty\bar{T}^*+\bar{T}^*\bar{b}_0\\
&+8\sqrt{\bar{T}^*}\sqrt{\frac{3\iota}{nd_m}}\left\|V^*\right\|_\infty+2\sqrt{\bar{T}^*}\sqrt{\frac{6\left\|V^*\right\|_\infty\iota}{nd_m}},
\end{aligned}
$$

It implies that

$$
\begin{aligned}
(1-13\sqrt{\frac{2S\iota}{nd_m}}\bar{T}^*)\left\|\bar{V}-V^*\right\|_\infty\leq&\frac{68\max\{\widetilde{B},1\}\sqrt{S}\iota\bar{T}^*}{nd_m}+\bar{T}^*\bar{b}_0\\
&+14\sqrt{\frac{\bar{T}^*B_*\iota}{nd_m}}(\sqrt{B_*}+1).
\end{aligned}
\tag{42}
$$

Since $n\geq\frac{26^2\times2S\iota(\bar{T}^*)^2}{d_m}$,

$$
\begin{aligned}
\left\|\bar{V}(s)-V^*(s)\right\|_\infty\leq&\frac{136\max\{\widetilde{B},1\}\sqrt{S}\iota\bar{T}^*}{nd_m}+2\bar{T}^*\bar{b}_0\\
&+28\sqrt{\frac{\bar{T}^*B_*\iota}{nd_m}}(\sqrt{B_*}+1)
\end{aligned}
$$

$$\leq \frac{136\max\{\widetilde{B},1\}\sqrt{S}\iota\bar{T}^*}{nd_m} + 360\bar{T}^*\sqrt{\frac{6\widetilde{T}\widetilde{B}S}{n^2d_m^2}}(\sqrt{\widetilde{B}}+1)\iota$$
$$+ 28\sqrt{\frac{\bar{T}^*B_*\iota}{nd_m}}(\sqrt{B_*}+1)$$

When $n \geq \frac{1000(\widetilde{B}+\sqrt{\widetilde{T}\widetilde{B}}(\sqrt{\widetilde{B}}+1))^2S\iota\bar{T}^*}{4B^*(\sqrt{B^*}+1)^2d_m}$, we have

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

□

I PROOF OF THEOREM 4.1

In this section, we provide the proof of Theorem 4.1. However, before that, we first present a lemma that guarantees pessimism.

Lemma I.1. *When $n \geq \max\{\frac{26^2\times 2S\iota(\bar{T}^*)^2}{d_m}, \frac{1000(\widetilde{B}+\sqrt{\widetilde{T}\widetilde{B}}(\sqrt{\widetilde{B}}+1))^2S\iota\bar{T}^*}{4B^*(\sqrt{B^*}+1)^2d_m}\}$, 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 H.1, we can get

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

This implies that

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

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

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

Applying the Chernoff bound given in Lemma L.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 L.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 $\bar{T}^* \leq \widetilde{T}$ and $B_* \leq \widetilde{B}$, we can prove the result in the Lemma. □

Now we are ready to introduce the final proof.

Theorem I.2. *Given Assumption 2.5 and Assumption 2.6. 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 $\iota = O(\log(SA/\delta))$),*

$$V^{\bar{\pi}}(s_{\text{init}}) - V^*(s_{\text{init}}) \leq 8\sum_{s,a,s\neq g} d^*(s,a)\sqrt{\frac{3\mathrm{Var}_{P_{s,a}}[V^*+c]\iota}{n\cdot d^\mu(s,a)}} + \widetilde{O}(\frac{1}{n}), \tag{43}$$

where we define $n_0 := n \geq \max\{\frac{4B_*-2c_{min}}{c_{min}d_{\max}}, \frac{26^2\times 2S\iota(\bar{T}^*)^2(\sqrt{B_*}+1)^2}{d_m}, \frac{1000(\sqrt{\widetilde{B}}+1)^4S\iota\bar{T}^*\widetilde{T}}{B^*(\sqrt{B^*}+1)^2d_m}\}$.

Proof. Define the operator $\widetilde{\mathcal{T}}$ as $\widetilde{\mathcal{T}}(V)(s) = \min_a \left\{\min\{\widehat{c}(s,a) + \widetilde{P}'_{s,a}V + b_{s,a}(V)\ ,\ \widetilde{B}\}\right\}$. Lemma G.6 guarantees that $\widetilde{\mathcal{T}}$ is a contraction mapping, and thus Algorithm 2 will converge. Denote the limit of $V^{(i)}$ as $\bar{V}$ which is the fixed point of $\widetilde{\mathcal{T}}$. We first decompose the regret in the following way,

$$V^{\bar{\pi}}(s) - V^*(s) = (V^{\bar{\pi}}(s) - \bar{V}(s)) + (\bar{V}(s) - V^*(s)). \tag{44}$$

From Lemma G.8, we have with probability $1-\delta$ and when $n \geq \max\{2\frac{B_*+1-c_{\min}}{c_{\min}d_{max}}, \frac{26^2\times 2S\iota(\bar{T}^*)^2}{d_m}, \frac{1000(\sqrt{\widetilde{B}}+1)^4 S\iota\bar{T}^*\widetilde{\mathcal{T}}}{B^*(\sqrt{B^*}+1)^2 d_m}\}$,

$$V^{\bar{\pi}} - \bar{V} = \sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}} \xi_h^{\bar{\pi}}(s)\{(P_{s,\bar{\pi}(s)} - \widetilde{P}'_{s,\bar{\pi}(s)})\bar{V} + c(s,\bar{\pi}(s)) - \hat{c}(s,\bar{\pi}(s)) - b_{s,\bar{\pi}(s)}(\bar{V})\}. \tag{45}$$

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

$$(P_{s,\bar{\pi}(s)} - \widetilde{P}'_{s,\bar{\pi}(s)})\bar{V} + c(s,\bar{\pi}(s)) - \hat{c}(s,\bar{\pi}(s)) - b_{s,\bar{\pi}(s)}(\bar{V}) \leq 0. \tag{46}$$

Thus combine (45) and (46) we have $V^{\bar{\pi}} - \bar{V} \leq 0$ by pessimism. For the term $\bar{V} - V^*$, we apply Lemma G.7 to obtain:

$$\bar{V} - V^* \leq \sum_{h=0}^{\infty}\sum_{\substack{s\\ s\neq g}} \xi_h^*(s)\{(\widetilde{P}'_{s,\pi^*(s)} - P_{s,\pi^*(s)})\bar{V} + \hat{c}(s,\pi^*(s)) - c(s,\pi^*(s)) + b_{s,\pi^*(s)}(\bar{V})\}. \tag{47}$$

From Lemma G.3, we have

$$|(P_{s,a} - \widetilde{P}'_{s,a})\bar{V}| \leq \frac{\widetilde{B}}{n(s,a)} + \frac{16\widetilde{B}\iota}{3n(s,a)} + 2\sqrt{\frac{\mathrm{Var}(\widetilde{P}'_{s,a},\bar{V})\iota}{n(s,a)}} + 6\sqrt{\frac{S\iota}{n(s,a)}}\left\|\bar{V} - V^*\right\|_\infty \tag{48}$$

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

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

Plug the above into (47), then we bound all the terms one by one. First,

$$\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h^*(s,a)\left(2\sqrt{\frac{3c(s,a)\iota}{n(s,a)}}\right) \leq \sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}} \xi_h^*(s,a)\left[2\sqrt{\frac{6c(s,a)\iota}{n\sum_{h=1}^{\infty}\xi_h^\mu(s,a)}}\right] = 2\sum_{\substack{s,a\\ s\neq g}} d^*(s,a)\sqrt{\frac{6c(s,a)\iota}{nd^\mu(s,a)}}. \tag{49}$$

For the second term, first we have

$$\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^*(s,a)\left(4\sqrt{\frac{3\mathrm{Var}(P_{s,a},V^*)\iota}{n(s,a)}}\right)\leq\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^*(s,a)\left[4\sqrt{\frac{6\mathrm{Var}(P_{s,a},V^*)\iota}{n\sum_{h=1}^{\infty}\xi_h^{\mu}(s,a)}}\right]$$
$$=4\sum_{\substack{s,a\\ s\neq g}}d^*(s,a)\sqrt{\frac{6\mathrm{Var}(P_{s,a},V^*)\iota}{nd^{\mu}(s,a)}} \tag{50}$$

From Chernoff bound given in Lemma L.6, we have with probability $1-\delta$, we have

$$\widetilde{O}(\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^*(s,a)\frac{(\widetilde{B}+1)\sqrt{S\iota}}{n(s,a)})\leq\widetilde{O}(\sum_{\substack{s,a\\ s\neq g}}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)}\cdot\frac{(\widetilde{B}+1)\sqrt{S\iota}}{n}). \tag{51}$$

Similarly, we have

$$\widetilde{O}(\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^*(s,a)\sqrt{\frac{S\iota}{n(s,a)}}\left\|\bar{V}-V^*\right\|_{\infty})\leq\widetilde{O}(\sum_{\substack{s,a\\ s\neq g}}d^*(s,a)\sqrt{\frac{S\iota}{nd^{\mu}(s,a)}}\left\|\bar{V}-V^*\right\|_{\infty})$$
$$\overset{\text{(i)}}{\leq}\widetilde{O}(\sum_{\substack{s,a\\ s\neq g}}d^*(s,a)\sqrt{\frac{\bar{T}^*B_*S}{n^2d^{\mu}(s,a)d_m}}(\sqrt{B_*}+1)\iota), \tag{52}$$

where (i) uses the Crude optimization bound given in Theorem H.1. For the last term, we have

$$\widetilde{O}(\sum_{h=1}^{\infty}\sum_{\substack{s,a\\ s\neq g}}\xi_h^*(s,a)\sqrt{\frac{\widetilde{T}\widetilde{B}S}{n(s,a)n_{\min}}}(\sqrt{\widetilde{B}}+1)\iota)\leq\widetilde{O}(\sum_{\substack{s,a\\ s\neq g}}d^*(s,a)\sqrt{\frac{\widetilde{T}\widetilde{B}S}{n^2d^{\mu}(s,a)d_m}}(\sqrt{\widetilde{B}}+1)\iota), \tag{53}$$

where the inequality comes from Lemma L.6 again. Combine the inequalities (49), (50), (51), (52) and (51) together, we have

$$\bar{V}(s_{\text{init}})-V^*(s_{\text{init}})\overset{\text{(i)}}{\leq}2\sum_{\substack{s,a\\ s\neq g}}d^*(s,a)\sqrt{\frac{6c(s,a)\iota}{nd^{\mu}(s,a)}}+4\sum_{\substack{s,a\\ s\neq g}}d^*(s,a)\sqrt{\frac{6\mathrm{Var}(P_{s,a},V^*)\iota}{nd^{\mu}(s,a)}}$$
$$+\widetilde{O}(\sum_{\substack{s,a\\ s\neq g}}d^*(s,a)\sqrt{\frac{\widetilde{T}\widetilde{B}S}{n^2d^{\mu}(s,a)d_m}}(\sqrt{\widetilde{B}}+1)\iota)+\widetilde{O}(\sum_{\substack{s,a\\ s\neq g}}\frac{d^{\pi}(s,a)}{d^{\mu}(s,a)}\cdot\frac{(\widetilde{B}+1)\sqrt{S\iota}}{n})$$
$$\overset{\text{(ii)}}{\leq}2\sum_{\substack{s,a\\ s\neq g}}d^*(s,a)\sqrt{\frac{6c(s,a)\iota}{nd^{\mu}(s,a)}}+4\sum_{\substack{s,a\\ s\neq g}}d^*(s,a)\sqrt{\frac{6\mathrm{Var}(P_{s,a},V^*)\iota}{nd^{\mu}(s,a)}}$$
$$+\widetilde{O}(\sum_{\substack{s,a\\ s\neq g}}d^*(s,a)\sqrt{\frac{\widetilde{T}S}{n^2d^{\mu}(s,a)d_m}}(\widetilde{B}+1)\iota)$$
$$\leq 8\sum_{\substack{s,a\\ s\neq g}}d^*(s,a)\sqrt{\frac{3\mathrm{Var}(P_{s,a},V^*+c)\iota}{nd^{\mu}(s,a)}}+\widetilde{O}(\sum_{\substack{s,a\\ s\neq g}}d^*(s,a)\sqrt{\frac{\widetilde{T}S}{n^2d^{\mu}(s,a)d_m}}(\widetilde{B}+1)\iota), \tag{54}$$

where the inequality (i) uses the assumption that $\bar{T}^*\leq\widetilde{T}$ and $B_*\leq\widetilde{B}$. (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 H.1, we can also get the Proposition below.

Proposition I.3. *When $n \geq n_0$ (where n_0 is defined the same as Theorem I.2), then the suboptimality incurred by the limit of the output policy $\bar{\pi}$ can be upper bounded as (with probability $1-\delta$)*

$$V^{\bar{\pi}}(s_{init}) - V^*(s_{init}) \leq 8\Big(\sqrt{\sum_{\substack{s,a \\ s\neq g}} \frac{d^*(s,a)}{d^\mu(s,a)} \cdot \frac{6\iota}{n}}\Big) \cdot (B_*+1) + \widetilde{O}(\frac{1}{n}). \tag{55}$$

Proof. By Theorem 4.1,

$$\begin{aligned} V^{\bar{\pi}}(s_{\text{init}}) - V^*(s_{\text{init}}) &\leq 8 \sum_{s,a,s\neq g} d^*(s,a)\sqrt{\frac{3\text{Var}_{P_{s,a}}[V^*+c]\iota}{n \cdot d^\mu(s,a)}} + \widetilde{O}(\frac{1}{n}) \\ &\overset{\text{(i)}}{\leq} 8\sqrt{\sum_{\substack{s,a \\ s\neq g}} \frac{d^*(s,a)}{n \cdot d^\mu(s,a)}} \cdot \sqrt{3\sum_{\substack{s,a \\ s\neq g}} d^*(s,a)\text{Var}_{P_{s,a}}[V^*+c]\iota}, \end{aligned} \tag{56}$$

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

$$\begin{aligned} \sum_{\substack{s,a \\ s\neq g}} d^*(s,a)\text{Var}_{P_{s,a}}[V^*+c] &= \sum_{h=1}^{\infty}\sum_{\substack{s,a \\ s\neq g}} \xi_h^*(s,a)(\text{Var}(P_{s,a},V^*) + c(s,a)) \\ &\overset{\text{(i)}}{\leq} 2\left\|V^*\right\|_\infty^2 + V^*(s_0) \\ &\leq 2B_*^2 + B_*, \end{aligned} \tag{57}$$

where (i) comes from Lemma J.1 and the definition of value function. Plug (57) into (56), we obtain

$$\begin{aligned} V^{\bar{\pi}}(s_{\text{init}}) - V^*(s_{\text{init}}) &\leq 8\sqrt{\sum_{\substack{s,a \\ s\neq g}} \frac{d^*(s,a)}{n \cdot d^\mu(s,a)}} \cdot \sqrt{6(B_*^2+B_*)\iota} + \widetilde{O}(\frac{1}{n}) \\ &\leq 8\sqrt{\sum_{\substack{s,a \\ s\neq g}} \frac{6d^*(s,a)\cdot\iota}{d^\mu(s,a)\cdot n}} \cdot (B_*+1) + \widetilde{O}(\frac{1}{n}), \end{aligned} \tag{58}$$

which completes the proof. □

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

Lemma J.1. *For any $V(\cdot) \in \mathbb{R}^S$ satisfying $V(g)=0$, i.e. (19), and suppose event $\mathcal{E}_1$ holds. we have*

$$\begin{aligned} \text{Var}(\widehat{P}_{s,a},V) &\leq \frac{3}{2}\text{Var}(P_{s,a},V) + \frac{2\left\|V\right\|_\infty^2 S\iota}{n(s,a)} \\ \text{Var}(P_{s,a},V) &\leq 2\text{Var}(\widehat{P}_{s,a},V) + \frac{4\left\|V\right\|_\infty^2 S\iota}{n(s,a)} \end{aligned} \tag{59}$$

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

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

$$P(s'|s,a) \le 2\widehat{P}(s'|s,a) + \frac{10\iota}{3n(s,a)} \le 2\widehat{P}(s'|s,a) + \frac{4\iota}{n(s,a)}. \tag{60}$$

For the first inequality, we have

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

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

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

□

Lemma J.2. *With probability at least $1-\delta$, we have*

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

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 J.3. *With probability $1-\delta$, for all $V(\cdot) \in \mathbb{R}^{S'}$ such that $\left\|V\right\|_\infty < \infty$, we have for all sate-action pair (s,a)*

$$(\widehat{P}_{s,a} - P_{s,a})V \le \sqrt{\frac{2S\mathrm{Var}(P_{s,a}, V)\iota}{n(s,a)}} + \frac{2\left\|V\right\|_\infty S\iota}{3n(s,a)}, \tag{61}$$

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

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

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

$$
\begin{aligned}
&\leq \sqrt{\frac{2P_{s,a}\iota}{n(s,a)}}\,|V - P_{s,a}V1_S| + \frac{2S\,\|V\|_\infty\,\iota}{3n(s,a)} \\
&\overset{\text{(ii)}}{\leq} (\sqrt{S}\sqrt{\frac{2P_{s,a}(V-P_{s,a}V)^2\iota}{n(s,a)}}) + \frac{2S\,\|V\|_\infty\,\iota}{3n(s,a)} \\
&\leq \sqrt{\frac{2S\mathrm{Var}(P_{s,a},V)\iota}{n(s,a)}} + \frac{2\,\|V\|_\infty\,S\iota}{3n(s,a)},
\end{aligned}
$$

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 D.1).

□

K 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.* $\max_{s,a,s\neq g}\frac{d^{\pi^\star}(s,a)}{d^\mu(s,a)} \leq C^\star$, 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 K.1 (Restatement of Theorem 5.1)**.** *We define the following family of SSPs:*

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

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)

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

The proof of Theorem K.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 K.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 $\mathcal{S} = \{s_{\text{init}}\}\cup\{s_1^j, s_+^j\}_{j=1}^{S'}$. 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.* $P(s_1^j|s_{\text{init}},a_d) = 1/S'$;
- For each state s_1^j, it satisfies

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

  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):

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

  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 L.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 K.2. *For any (deterministic) policy π and any two different transition probabilities $P_1, P_2 \in \mathcal{P}$, it holds:*

$$V^{\pi}_{P_1}(s_{\text{init}}) - V^{\pi^\star}_{P_1}(s_{\text{init}}) + V^{\pi}_{P_2}(s_{\text{init}}) - V^{\pi^\star}_{P_2}(s_{\text{init}}) \geq \delta B^\star/2.$$

Proof of Lemma K.2. Since s_{init} uniformly transitions to S' states (w.r.t the default action a_d), therefore for any policy π,

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

Case1. 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

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

Case2. 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

$$
\begin{aligned}
& V_{P_1}^{\pi}(s_{\text{init}}) - V_{P_1}^{\pi^\star}(s_{\text{init}}) + V_{P_2}^{\pi}(s_{\text{init}}) - V_{P_2}^{\pi^\star}(s_{\text{init}}) \\
\geq & \frac{1}{S'}\sum_{i=1}^{S'}(V_{P_1}^{\pi}(s_1^i) - V_{P_1}^{\pi^\star}(s_1^i)) + \frac{1}{S'}\sum_{i=1}^{S'}(V_{P_2}^{\pi}(s_1^i) - V_{P_2}^{\pi^\star}(s_1^i)) \\
\geq & \frac{1}{S'}\delta B^\star \sum_{i=1}^{S'}\left(\mathbf{1}[\pi(s_1^i) \neq \pi_{P_1}^\star(s_1^i)] + \mathbf{1}[\pi(s_1^i) \neq \pi_{P_2}^\star(s_1^i)]\right) \\
\geq & \frac{\delta B^\star}{S'}\sum_{i=1}^{S'}\mathbf{1}[\pi_{P_1}^\star(s_1^i) \neq \pi_{P_2}^\star(s_1^i)]
\end{aligned}
\tag{62}
$$

Lastly, by Lemma L.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 (62) 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 L.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 L.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 K.2, we have[8]

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

Note by the choice of $\mathcal{P}$, $\log|\mathcal{P}| \geq S'/2$, therefore it remains to bound $\max_{i \neq j} \text{KL}(\mu \circ P_i || \mu \circ P_j)$. By definition, we have

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

where $\tau_{s_1^i}$ corresponds to all the possible trajectories starting from s_1^i. Then there are the following several cases:[9]

- 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 $\mathbb{P}(s_1^i \to a_1 \to s_+^i \to s_+^i \to g) = (1 - \frac{1}{C^\star})\frac{1}{2}q(1 - q)$;

[8] 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.

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

- 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

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

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

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

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

$$\inf_{\widehat{\pi}} \sup_{(s_{\text{init}}, \mu, P, c) \in \mathcal{P}} \mathbb{E}_{\mathcal{D}}[V^{\widehat{\pi}}(s_{\text{ininit}}) - V^\star(s_{\text{init}})]] \geq \frac{\delta B^\star}{4} \frac{1}{2} = \frac{1}{32} B^\star \sqrt{\frac{C^\star S}{n}}.$$

This completes the proof.

□

L TECHNICAL LEMMAS

Lemma L.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 L.2 (Multiplicative Chernoff bound Chernoff et al. [1952])**.** *Let X be a Binomial random variable with parameter p, n. For any $1 \geq \theta > 0$, we have that*

$$\mathbb{P}[X < (1-\theta)pn] < e^{-\frac{\theta^2 pn}{2}} \quad \textit{and} \quad \mathbb{P}[X \geq (1+\theta)pn] < e^{-\frac{\theta^2 pn}{3}}.$$

Lemma L.3 (Bernstein's Inequality). *Let $x_1, ..., x_n$ be independent bounded random variables such that $\mathbb{E}[x_i] = 0$ and $|x_i| \le \xi$ with probability 1. Let $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}[x_i]$, then with probability $1-\delta$ we have*

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

Lemma L.4 (Empirical Bernstein's Inequality [Maurer and Pontil, 2009]). *Let $x_1, ..., x_n$ be i.i.d random variables such that $|x_i| \le \xi$ 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*

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

Lemma L.5 (Generalized Fano's inequality). *Let $L : \Theta \times \mathcal{A} \to \mathbb{R}_+$ be any loss function, and there exist $\theta_1, \dots, \theta_m \in \Theta$ such that*

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

Then it holds

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

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

Proof of Lemma L.5. The proof come from the combination of Lemma 1 and Lemma 3 of Han and Fischer-Hwang [2019]. □

Lemma L.6 (Chernoff Bound for Stochastic Shortest Path). *Recall by definition*

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

Then when $n > C \cdot \log(SA)/\sum_{h=1}^\infty \xi_h^\mu(s,a)$, 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. 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

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

Then apply Lemma L.2 (where we pick $\theta = \frac{1}{2}$) to $n_t(s,a)/t$ and $\sum_{h=1}^t \xi_h^\mu(s,a)/t$ and union bound, we have with probability $1-\delta$, for all s, a

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

Since above holds for all $s, a \in \mathcal{S} \times \mathcal{A}$, it implies with probability $1-\delta$,

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

This is because $n_t(s,a) \to n(s,a)$ 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 ($\lim_{t\to\infty} P[n_t(s,a) \le \frac{3}{2}n\sum_{h=1}^t \xi_h^\mu(s,a)] = P[n(s,a) \le \frac{3}{2}n\sum_{h=1}^\infty \xi_h^\mu(s,a)]$). □

Lemma L.7. *For any $a, b, c \in \mathbb{R}$, we have*

$$|\min\{a, b\} - \min\{a, c\}| \leq |b - c|. \tag{63}$$

Proof.

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

□