On Gap-dependent Bounds for Offline Reinforcement Learning

Xinqi Wang
Institute for Interdisciplinary Information Sciences
Tsinghua University
wangxqkaxdd@gmail.com

Qiwen Cui
Paul G. Allen School of Computer Science Engineering
University of Washington
qwcui@cs.washington.edu

Simon S. Du
Paul G. Allen School of Computer Science Engineering
University of Washington
ssdu@cs.washington.edu

Abstract

This paper presents a systematic study on gap-dependent sample complexity in offline reinforcement learning. Prior work showed when the density ratio between an optimal policy and the behavior policy is upper bounded (the optimal policy coverage assumption), then the agent can achieve an $O\left(\frac{1}{\epsilon^2}\right)$ rate, which is also minimax optimal. We show under the optimal policy coverage assumption, the rate can be improved to $O\left(\frac{1}{\epsilon}\right)$ when there is a positive sub-optimality gap in the optimal Q-function. Furthermore, we show when the visitation probabilities of the behavior policy are uniformly lower bounded for states where an optimal policy's visitation probabilities are positive (the uniform optimal policy coverage assumption), the sample complexity of identifying an optimal policy is independent of $\frac{1}{\epsilon}$. Lastly, we present nearly-matching lower bounds to complement our gap-dependent upper bounds.

1 Introduction

Reinforcement Learning (RL) aims to learn a policy that maximizes the long-term reward in unknown environments [Sutton and Barto, 2018]. The success of reinforcement learning often relies on being able to deploy the algorithms that directly interact with the environment. However, such direct interactions with real environments can be expensive or even impossible in many real-world applications, e.g., health and medicine [Murphy et al., 2001, Gottesman et al., 2019], education [Mandel et al., 2014], conversational AI [Ghandeharioun et al., 2019] and recommendation systems [Chen et al., 2019]. Instead, we have access to a dataset generated from some past suboptimal policies. Offline reinforcement learning (offline RL) aims to find a near-optimal policy using the offline dataset, and has achieved promising empirical successes [Lange et al., 2012, Levine et al., 2020].

Recently, a line of works showed that under the single policy coverage assumption (Assumption 3.2), one can obtain a near-optimal policy with polynomial number of samples [Rashidinejad et al., 2021, Xie et al., 2021b, Yin and Wang, 2021, Xie et al., 2021a, Li et al., 2022]. In particular, for the tabular setting, recent works have obtained minimax optimal sample complexity bounds $\widetilde{\Theta}\left(\frac{H^3SC^*}{\epsilon^2}\right)$ where H is the planning horizon, S is the number of states, C^* is the constant for the single policy coverage assumption, and ϵ is the target accuracy [Xie et al., 2021b, Li et al., 2022]. The $O\left(1/\epsilon^2\right)$ is for the

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

worst case and in many benign settings, one may use much fewer samples to learn a (near-)optimal policy.

How the benign problem structures help reduce the sample complexity has been extensively studied in the online bandits and reinforcement learning [Simchowitz and Jamieson, 2019, Jonsson et al., 2020, Wagenmaker et al., 2021b, Xu et al., 2021]. In particular, when there is a suboptimality gap between the optimal policy and the rest, then one can obtain $\log T$-type regret in contrast to $\sqrt{T}$-type regret bounds in the worst case where T is the number of interactions. However, to our knowledge, how the gap structure helps reduce sample complexity in offline RL has not been thoroughly investigated. This paper presents a systematic study on gap-dependent bounds for offline RL in the canonical tabular setting, with nearly-matching upper and lower bounds in different regimes.

1.1 Main Contributions

We present novel analyses for the standard VI-LCB algorithm (Algorithm 2). Our main results are summarized in Table 1.1.

1. We develop a novel technique, *deficit thresholding*, to obtain gap-dependent bounds in offline RL. Different from the clip trick widely used in online RL [Simchowitz and Jamieson, 2019, Xu et al., 2021] to obtain gap-dependent bounds, our deficit thresholding technique is adaptive to the problem-instance. As will be shown in Section 6.2, this technique helps reduce the dependency on H by adapting to the variance of the estimate.
2. Using the deficit thresholding technique, we obtain the first gap-dependent bound under the optimal policy coverage assumption. Specifically, we obtain an $\widetilde{O}\left(\frac{H^4SC^*}{\epsilon \mathrm{gap}_{\min}}\right)$ bound where $\mathrm{gap}_{\min}$ is the minimum suboptimality gap between the optimal Q-value of the best action and that of the second-to-the-best action.[1] Notably, compared with the worst-case gap-indepdent, the rate improves from $1/\epsilon^2$ to $1/\epsilon$.
3. We also present the first gap-dependent lower bound for offline RL, $\Omega\left(\frac{H^2SC^*}{\epsilon \mathrm{gap}_{\min}}\right)$ to show the $O\left(1/\epsilon\right)$ is unimprovable even with the gap condition, and our upper bound is tight up to an H^2 factor. The main insight from our lower bound is that if there exists a state whose visitation probability of the behavior policy and the optimal policy is $O\left(\epsilon\right)$, then we cannot learn much of the state using the offline dataset and will inevitably incur an $O\left(\epsilon\right)$ error.
4. We further study what condition permits even faster rate than $O\left(1/\epsilon\right)$. Leveraging our lower bound mentioned above, we propose a new condition, *uniform optimal policy coverage*, which posits that the visitation probabilities of the behavior policy are uniformly lower bounded by P for states where an optimal policy's visitation probabilities are positive. Under this assumption, we obtain an $\widetilde{O}\left(\frac{H^3}{P\mathrm{gap}_{\min}^2}\right)$ bound. Importantly, this bound is *independent of* $1/\epsilon$ (not even $\log(1/\epsilon)$), a.k.a., one can identify an *exact optimal* policy. We also complement this upper bound with an $\Omega\left(\frac{H}{P\mathrm{gap}_{\min}^2}\right)$ lower bound to show our upper bound is tight up to an H^2 factor.

Lastly, we note that all of our bounds are obtained with the same algorithm, i.e., the algorithm automatically exploits the benign problem structure without any prior knowledge.

2 Related Work

We focus on existing theoretical results on gap-dependent bounds and offline RL in the tabular setting.

Theoretical Results on Offline Tabular RL. Theoretical analysis of offline RL can be traced back to Szepesvári and Munos [2005], under the uniform coverage assumption where *every state-action*

[1] The gap-dependent bounds in the online setting also depend on the suboptimality gaps of the actions other than the second-to-the-best action. This dependency is *not needed* for offline RL because even if the dataset does not have any information about the sub-optimal actions, the agent can still learn a near-optimal policy (as long as the dataset covers an optimal action). On the other hand, this dependency is needed in the online setting because the agent needs to explore the all actions.

Condition	Upper bound	Lower bound
C^*	$\widetilde{O}\left(\frac{H^3SC^*\log\frac{1}{\delta}}{\epsilon^2}\right)$	$\Omega\left(\frac{H^3SC^*}{\epsilon^2}\right)$
C^*, $\mathrm{gap}_{\min}$	$\widetilde{O}\left(\frac{H^4SC^*\log\frac{1}{\delta}}{\epsilon\mathrm{gap}_{\min}}\right)$	$\Omega\left(\frac{H^2SC^*}{\epsilon\mathrm{gap}_{\min}}\right)$
P, $\mathrm{gap}_{\min}$	$\widetilde{O}\left(\frac{H^3\log\frac{1}{\delta}}{P\mathrm{gap}_{\min}^2}\right)$	$\Omega\left(\frac{H}{P\mathrm{gap}_{\min}^2}\right)$

Table 1: Sample Complexity Bounds for different conditions about the sub-optimality and coverage. Cells in gray are the contributions of this work. The results in the first line without suboptimality gap assumptions were obtained in Xie et al. [2021b], Li et al. [2022]. C^* stands for the relative optimal policy coverage coefficient, i.e., $\max_{h,s}\frac{d_h^*(s)}{d_h^\mu(s)}$ where $d_h^*(s)$ is the visitation probability of the optimal policy for state s at level h and $d_h^\mu(s)$ is the visitation probability of the behavior policy μ for state s at level h. P stands for the uniform optimal policy coverage coefficient, i.e., $\min_{h,s|d_h^*(s)>0} d_h^\mu(s)$. $\mathrm{gap}_{\min}$ is the minimum non-zero suboptimality gap among all time-state-action tuples, i.e., $\min_{h,s,a|a \text{ is not optimal}} \mathbf{V}_h^*(s) - \mathbf{Q}_h^*(s,a)$.

pairs are visted by the behavior policy with a positive probability. Sharp sample complexity bounds have been obtained under this assumption [Xie and Jiang, 2021, Xie et al., 2019, Yin et al., 2020, 2021b, Ren et al., 2021]. Recently, a line of works showed that under a much weaker assumption, single policy coverage, one can design sample efficient algorithms with both model-based and model-free methods based on the pessimism principle, [Rashidinejad et al., 2021, Yin and Wang, 2021, Xie and Jiang, 2021, Jin et al., 2021, Uehara and Sun, 2021, Uehara et al., 2021, Zanette et al., 2021]. Recently, Yin and Wang [2021] obtained a problem-dependent sample complexity in terms of the variance.

Instance-dependent Sample Complexity in Online Learning and Generative Models. In online RL, a line of work studied the upper and lower bounds of gap-dependent sample complexity in both the regret and PAC settings [Simchowitz and Jamieson, 2019, Xu et al., 2021, Dann et al., 2021, Jonsson et al., 2020, Wagenmaker et al., 2021b,a, Tirinzoni et al., 2021]. Besides gap-dependent bounds, there are other problem-dependent bounds such as first-order and variance-dependent bounds Zanette and Brunskill [2019]. For generative models, Zanette et al. [2019] derived an upper bound depending on both variance and gap information.

Gap-dependent bounds in Offline Learning. The most related work is by Hu et al. [2021] who studied the convergence rate of Q-learning in the discounted MDPs under the uniform coverage assumption. They proved that one can obtain an ϵ-optimal policy with $O\left(S^3A^3\log(1/\epsilon)/(1-\gamma)^4P^2\mathrm{gap}_{\min}^2\right)$ samples for tabular MDP and $\widetilde{O}(1/\epsilon)$ (ignoring other parameters) for linear MDP. In comparison, we show that under the weaker uniform optimal policy coverage assumption, we can identify an exact optimal policy with $O\left(H^3/P\mathrm{gap}_{\min}^2\right)$ sample complexity, which has no dependency on $1/\epsilon$.

3 Preliminaries

Notations. We let $[n] = \{1, 2\cdots n\}$. For two vectors a, b of the same length k, we use $a\circ b$ to denote the Hadamard product $(a_1b_1, a_2b_2\cdots a_kb_k)$. We use the standard definitions of $O(\cdot), \Theta(\cdot), \Omega(\cdot)$ to hide absolute constants, and tilded notations $\widetilde{O}(\cdot), \widetilde{\Theta}(\cdot), \widetilde{\Omega}(\cdot)$ to hide absolute constants as well as poly-logarithmic factors except for $\log\frac{1}{\delta}$. $\mathbf{Var}_p(V) = p^\top V\circ V - (p^\top V)^2$ refers to the variance of V with respect to the weight p. $a \lesssim b$ means $a \le Cb$ for some positive absolute constant C, and similarly $a \gtrsim b$ means $a \ge Cb$. $\mathbb{I}\{\xi\}$ is the indicator function of the event ξ, which equals 1 when ξ is true and 0 otherwise. $\Delta(\mathcal{X})$ is the probability simplex over $\mathcal{X}$.

3.1 Markov Decision Processes

We consider tabular finite-horizon time-inhomogeneous MDPs described by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, p_0, r)$. Here $\mathcal{S}$ is the state space with cardinality S, $\mathcal{A}$ is the action space with cardinality A, $\mathcal{P} = \{p_{h,s,a}\}_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}}$ with $p_{h,s,a} \in \Delta(\mathcal{S})$ is the transition kernel at timestep h, state s and action a, $p_0 \in \Delta(\mathcal{S})$ is the initial state distribution of s_1, and $r = \{r_1, r_2, \ldots, r_H\}$ with $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function.[2] For each episode, the player will generate a trajectory $\{(s_h, a_h, r_h)\}_{h=1}^H$ where $s_1 \sim p_0$, $s_{h+1} \sim p_{h,s_h,a_h}$, and $r_h = r_h(s_h, a_h)$, by controlling the actions $\{a_h\}_{h=1}^H$. The target of the player is to maximize the total reward $\sum_{h=1}^H r_h$.

Policies. A policy $\pi = \{\pi_h(\cdot|s)\}_{(h,s)\in[H]\times\mathcal{S}}$ refers to a set of distributions over $\mathcal{A}$. With a slight abuse of the notations, when a policy π is deterministic, we use $\pi_h(s)$ to denote the action taken at timestep h and state s. We define $\mathbb{E}_{\pi,\mathcal{M}}[\cdot] \triangleq \mathbb{E}_{\phi\sim(\pi,\mathcal{M})}[\cdot]$, where $\phi = \{(s_h, a_h, r_h)\}_{h\in[H]}$ is a trajectory sampled using policy π in the MDP $\mathcal{M}$ and the expectation is over the randomness of both the policy and the transitions. $\mathcal{M}$ will be omitted when there is no confusion.

Value Functions, Q-Functions and Policy Distributions. For a given policy π and an MDP $\mathcal{M}$, we define the state value function and the state-action value function to be

$$\mathbf{V}_h^\pi(s) \triangleq \mathbb{E}_\pi\left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) \mid s_h = s\right], \mathbf{Q}_h^\pi(s,a) \triangleq \mathbb{E}_\pi\left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a\right].$$

We define the optimal Q-function as $\mathbf{Q}_h^*(s,a) \triangleq \sup_\pi \mathbf{Q}_h^\pi(s,a)$, and similarly $\mathbf{V}_h^*(s) \triangleq \sup_\pi \mathbf{V}_h^\pi(s)$ for all $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$. It is well known that there exists a deterministic optimal policy π^* that can achieve the above maximum for all $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $h \in [H]$ simultaneously.

We denote the value of a policy by $\mathbf{V}_0^\pi = \mathbb{E}_\pi\left[\sum_{h=1}^H r_h(s_h, a_h)\right]$, and the value of the optimal policy by $\mathbf{V}_0^* = \mathbf{V}_0^{\pi^*}$. A policy π is ϵ-optimal if

$$\text{Suboptimal}(\pi) \triangleq \mathbf{V}_0^* - \mathbf{V}_0^\pi \le \epsilon.$$

We use $d_h^\pi(\cdot)$ to denote the probability of reaching a state s under policy π:

$$d_h^\pi(s) \triangleq \mathbb{E}_\pi[\mathbb{I}\{s_h = s\}],\ d_h^\pi(s,a) \triangleq \mathbb{E}_\pi[\mathbb{I}\{(s_h, a_h) = (s,a)\}] = d_h^\pi(s)\pi_h(a \mid s).$$

Sub-optimality Gap. For a MDP instance $\mathcal{M}$, we define the gap at $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$ to be $\text{gap}_h(s,a) = \mathbf{V}_h^*(s) - \mathbf{Q}_h^*(s,a)$, which is always non-negative because of the definition of π^*. Our results will depend on the smallest positive gap: $\text{gap}_{\min} \triangleq \min_{(h,s,a)|\text{gap}_h(s,a)>0} \text{gap}_h(s,a)$, which quantifies the difficulty of learning the optimal policy in the MDP instance.[3]

3.2 Offline Reinforcement Learning

Offline Learning. For offline reinforcement learning, we want to solve an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$ with unknown transitions by utilizing a given dataset collected by some unknown behavior policy μ. Note that the algorithm is not allowed to perform any kind of additional sampling. The dataset is $\mathcal{D} = \{\{(s_{h,i}, a_{h,i}, r_{h,i})\}_{h\in[H]}\}_{i\in[N]}$, which contains N trajectories collected by the behavior policy μ independently. A (ϵ, δ)-PAC offline reinforcement learning algorithm is defined to output an ϵ-optimal policy π with probability at least $1 - \delta$.

Assumptions. We introduce two dataset assumptions that will be used in this paper.

Assumption 3.1 (Uniform optimal policy coverage)**.** *We define the uniform optimal policy coverage coefficient to be*

$$P \triangleq \min_{\pi^*} \min_{h,s:d_h^{\pi^*}(s,a)>0} d_h^\mu(s,a),$$

where π^* is an optimal policy. We assume that $P > 0$.

[2] We assume a known deterministic reward function as the main difficulty lies in learning the transition probability. All the conclusions in this paper can be proved for MDPs with unknown 1-subGuassian reward.

[3] We assume that at least one gap is positive. Otherwise all the actions are optimal and no learning is needed.

Algorithm 1: VI-LCB

input : Dataset $\mathcal{D}_0$, reward function r

Set $\underline{\mathbf{Q}}_{H+1}(s,a) = 0$
Set $\underline{\mathbf{V}}_{H+1}(s,a) = 0$
for $h \leftarrow H$ **to** 1 **do**
  Compute the empirical transition kernel $\hat{P}_h$
  $\hat{P}_{h,s,a}(s') = \frac{N_h(s,a,s')}{N_h(s,a)}$ with $0/0 = 0$
  for $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
    $b_h(s,a) \leftarrow C_b \sqrt{\frac{\mathbf{Var}_{\hat{P}_{h,s,a}}(\underline{\mathbf{V}}_{h+1})\iota}{N'_h(s,a)}} + C_b \frac{H\iota}{N'_h(s,a)}$, where $N'_h(s,a) = N_h(s,a) \vee \iota$
    $\underline{\mathbf{Q}}_h(s,a) \leftarrow \max\{0, r_h(s,a) + \hat{P}^\top_{h,s,a}\underline{\mathbf{V}}_{h+1} - b_h(s,a)\}$
  for $s \in \mathcal{S}$ **do**
    $\underline{\mathbf{V}}_h(s) \leftarrow \max_{a\in\mathcal{A}} \underline{\mathbf{Q}}_h(s,a)$
    $\underline{\pi}_h(s) \leftarrow \arg\max_{a\in\mathcal{A}} \underline{\mathbf{Q}}_h(s,a)$

output : policy $\underline{\pi}$

Algorithm 2: Subsampled VI-LCB

input : Dataset $\mathcal{D}$, reward function r

Split $\mathcal{D}$ into 2 halves containing same number of sample trajectories, $\mathcal{D}^{\text{main}}$ and $\mathcal{D}^{\text{aux}}$
$\mathcal{D}_0 = \{\}$
for $(h,s) \in [H] \times \mathcal{S}$ **do**
  $N_h^{\text{trim}}(s) \leftarrow \max\{0, N_h^{\text{aux}}(s) - 10\sqrt{N_h^{\text{aux}}(s) \log \frac{HS}{\delta}}\}$
  Randomly subsample $\min\{N_h^{\text{trim}}(s), N_h^{\text{main}}(s)\}$ samples of transition from (h,s) from $\mathcal{D}^{\text{main}}$ to add to $\mathcal{D}_0$
$\underline{\pi} \leftarrow$ VI-LCB$(\mathcal{D}_0, r)$

output : policy $\underline{\pi}$

Assumption 3.1 states that the behavior policy μ covers all the state-action pairs that some π^* will choose with positive probability. This is a natural assumption if we want to recover the optimal policy.

A closely related assumption is the uniform coverage assumption, i.e., all (h,s,a) tuples are covered by the dataset [Yin and Wang, 2020, Ren et al., 2021, Yin et al., 2021a, Hu et al., 2021]. Our assumption is significantly weaker as it only assumes covering the optimal policy. We will prove that under this assumption, we can identify the optimal policy with finite samples.

Assumption 3.2 (Optimal policy coverage)**.** *We define the relative optimal policy coverage coefficient to be*

$$C^* \triangleq \max_{\pi^*} \max_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \frac{d_h^{\pi^*}(s,a)}{d_h^{\mu}(s,a)}$$

with convention that $0/0 = 0$*, where* π^* *is an optimal policy. We assume that* $C^* < \infty$*.*

Similar coverage assumption has been widely adopted in Li et al. [2022], Shi et al. [2022], Yan et al. [2022], Jin et al. [2021], Rashidinejad et al. [2021]. Researchers have designed algorithms based on the pessimism principle to efficiently solve offline RL problems under this assumption. Assumption 3.2 is usually weaker than Assumption 3.1. For example, if there exists one unique optimal policy π^* and $\mu = \pi^*$, we have $C^* = 1$ while P can still be arbitrarily small. In addition, we always have $C^* \leq \frac{1}{P}$.

3.3 Subsampled VI-LCB

We briefly introduce the algorithm (Algorithm 2) that will be used in the analysis. Value Iteration with Lower Confidence Bound (VI-LCB) was first introduced by Rashidinejad et al. [2021] and improved by Li et al. [2022]. The main idea is to maintain a pessimistic estimate on the value functions so that the suboptimality of the output policy only depends on the uncertainty of the optimal policy. By utilizing the subsampling technique and Bernstein-style bonus, subsampled VI-LCB achieves the minimax sample complexity $\widetilde{O}(\frac{H^3SC^*}{\epsilon^2})$.

In the algorithm, $N_h(s,a)$ refers to the number of sample transitions starting from state s, taking action a at time step h of some given dataset, and $N_h(s) = \sum_{a\in\mathcal{A}} N_h(s,a)$. Superscripts stand for the dataset. See Li et al. [2022] for a more detailed description of the algorithm.

4 Finding an Exact Optimal Policy with Assumption 3.1

In this section, we show that we can identify the exact best policy with finite samples by utilizing the gap structure under Assumption 3.1. On the other hand, for the minimax sample complexity $\widetilde{O}(H^3SC^*\epsilon^{-2})$, it will become infinite when ϵ approaches 0. Note that directly setting $\epsilon < \mathrm{gap}_{\min}$ does not imply that the output policy π is optimal. Instead, we only have

$$\mathbf{V}_0^{\pi} \geq \mathbf{V}_0^* - \epsilon,$$

while π can still be suboptimal at states visited with low probability.

Theorem 4.1. *For an MDP $\mathcal{M}$ and a behavior policy μ with uniform optimal coverage coefficient P, if the number of sample trajectories satisfies*

$$N \geq \widetilde{O}\left(\frac{H^3 \log\frac{1}{\delta}}{P\mathrm{gap}_{\min}^2}\right),$$

then with probability at least $1-\delta$, Algorithm 2 returns an optimal policy.

Sketch of Proof. First, if for all $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$ satisfying $d_h^*(s,a) > 0$, we have

$$\underline{\mathbf{Q}}_h(s,a) > \mathbf{Q}_h^*(s,a) - \mathrm{gap}_{\min}, \tag{1}$$

then we have $\underline{\pi}$ is an optimal policy. This is because as $\underline{\mathbf{Q}}$ is a pessimistic estimate of Q^*, we have

$$\underline{\mathbf{Q}}_h(s,a) > \mathbf{Q}_h^*(s,a) - \mathrm{gap}_{\min} \geq \mathbf{Q}_h^*(s,a') \geq \underline{\mathbf{Q}}_h(s,a'),$$

for any action a' that is sub-optimal. As a result, $\underline{\pi}$ chooses the optimal action for all $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$ covered by the optimal policy, which implies $\underline{\pi}$ is an optimal policy.

Second, we show that (1) is satisfied with $N \geq \widetilde{O}\left(\frac{H^4}{P\mathrm{gap}_{\min}^2}\right)$. This is because if the number of samples at (h,s,a) exceeds $\widetilde{O}(H^4/\mathrm{gap}_{\min}^2)$, we can guarantee that the estimation error at that step is smaller than $\mathrm{gap}_{\min}/H$ by Hoeffding's inequality. Then the accumulated estimation error at $\underline{\mathbf{Q}}_h(s,a)$ can be bounded by $(H-h)\mathrm{gap}_{\min}/H < \mathrm{gap}_{\min}$, which implies (1). To further improve the dependence on H, we will use Bernstein's inequality and the proof is deferred to Appendix B.4. □

5 Gap-dependent Upper Bounds with Assumption 3.2

In the previous section, we show that the optimal policy can be identified if Assumption 3.1 is satisfied. However, the uniform optimal coverage coefficient P can be very small, which makes the bound $N \geq \widetilde{O}\left(\frac{H^3}{P\mathrm{gap}_{\min}^2}\right)$ useless. In this section, we present two results on learning an ϵ-optimal policy with assumption (Assumption 3.2) and we provide the proof sketch in the next section. The full proof can be found in Appendix B.

Theorem 5.1. *For an MDP $\mathcal{M}$ and behavior policy μ with relative optimal policy coverage coefficient C^*, if the number of sample trajectories satisfies*

$$N \geq \widetilde{O}\left(\frac{H^4SC^* \log\frac{1}{\delta}}{\epsilon\mathrm{gap}_{\min}}\right),$$

Algorithm 2 returns an ϵ-suboptimal policy with probability at least $1-\delta$.

Theorem 5.1 shows that to learn an ϵ-optimal policy, we only need $\widetilde{O}(1/\epsilon \mathrm{gap}_{\min})$ samples, which significantly improves the minimax sample complexity $\widetilde{O}(1/\epsilon^2)$ when $\epsilon \ll \mathrm{gap}_{\min}$.

Now we show that we can further improve the bound if both Assumption 3.1 and Assumption 3.2 are satisfied.

Theorem 5.2. *For an MDP $\mathcal{M}$ and behavior policy μ with uniform optimal policy coverage coefficient P and relative optimal policy coverage coefficient C^*, if the number of sample trajectories satisfies*

$$N \geq \widetilde{O}\left(\frac{H^3 S C^* \log \frac{1}{\delta}}{\epsilon \mathrm{gap}_{\min}} + \frac{H \log \frac{1}{\delta}}{P}\right),$$

Algorithm 2 returns an ϵ-suboptimal policy with probability at least $1-\delta$.

Theorem 5.2 improves an H factor compared with Theorem 5.1, with the cost of an extra H/P term. As the additional term is a constant with respect to ϵ and $\mathrm{gap}_{\min}$, the bound is improved when $\epsilon \mathrm{gap}_{\min}$ is small. We believe this H/P can be removed and we leave it to future works.

6 Main Proof Techniques

In this section, we will provide a proof sketch for Theorem 5.1 and Theorem 5.2.

6.1 Pessimistic Algorithms

First, we define pessimistic algorithms and imaginary MDPs $\underline{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \underline{r})$ determined by the pessimistic algorithms. Our analysis will generally hold for all pessimistic algorithms defined by Definition A.2 and we will show that VI-LCB (Algorithm 2) is a pessimistic algorithm. We use VI-LCB only to derive the final sample complexity guarantees.

Definition 6.1 (Pessimistic algorithms)**.** *An offline learning algorithm with output policy $\underline{\pi}$ is pessimistic if with probability at least $1-\delta$, the following arguments hold,*

1. *It maintains a pessimistic estimate $\underline{\mathbf{Q}}$ of the true Q^*.*
2. *$\underline{\mathbf{Q}}$ is the optimal Q function of an imaginary MDP $\underline{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, p_0, \underline{r})$, where $\underline{r}_h(s,a) \leq r_h(s,a)$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$.*[4]
3. *$\underline{\pi}$ is the greedy policy with respect to $\underline{\mathbf{Q}}$.*

Most of the existing offline RL algorithms are pessimistic algorithms and we will also prove it for VI-LCB [Li et al., 2022] in Appendix B.

Definition 6.2 (Deficit)**.** *For any pessimistic algorithm and the corresponding imaginary MDP $\underline{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, p_0, \underline{r})$, we define the deficit to be*

$$\mathbf{E}_h(s,a) = r_h(s,a) - \underline{r}_h(s,a),\ \forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}.$$

From the definition of pessimistic algorithms, we have $\mathbf{E}_h(s,a) \geq 0$ immediately. Intuitively, deficit stands for how pessimistic the estimates are. Note that the deficit is related to the algorithm itself and usually we can bound it by utilizing concentration inequalities.

6.2 Deficit Thresholding for Analysising LCB-style Algorithms

By defining $\underline{\mathbf{V}}^\pi$ to be the value function of policy π in $\underline{\mathcal{M}}$ and recalling that $\underline{\pi}$ is the optimal policy in $\underline{\mathcal{M}}$, we have $\underline{\mathbf{V}}_0^* \leq \underline{\mathbf{V}}_0^{\underline{\pi}} \leq \mathbf{V}_0^{\underline{\pi}} \leq \mathbf{V}_0^*$ and thus

$$\mathbf{V}_0^* - \underline{\mathbf{V}}_0^* = \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[r_h(s_h,a_h) - \underline{r}_h(s_h,a_h)] = \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\mathbf{E}_h(s_h,a_h)]$$

[4] Here we loosen the definition of MDP by allowing the reward function to have negative value. $\underline{r}$ maybe negative, but as will be shown in appendix, $\underline{\mathbf{V}}^{\underline{\pi}}$ and $\underline{\mathbf{Q}}^{\underline{\pi}}$ are still non-negative, thus does not affect our analysis.

can upper bound the suboptimality of $\underline{\pi}$. Surprisingly, we will show that even if we threshold the deficit, a similar upper bound still holds. Define the thresholded deficits and the corresponding reward functions as

$$\ddot{\mathbf{E}}_h(s,a) \triangleq \max\{0, \mathbf{E}_h(s,a) - \epsilon_h(s,a)\},\ \ddot{r}_h(s,a) \triangleq r_h(s,a) - \ddot{\mathbf{E}}_h(s,a),\ \forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A},$$

for any non-negative threshold function $\epsilon_h(s,a)$. Then we can define a thresholded MDP $\ddot{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \ddot{r})$ and we use $\ddot{\mathbf{V}}$ to denote the value function in $\ddot{\mathcal{M}}$. Now we present the key lemma showing that if the thresholding function satisfies certain conditions related to $\mathrm{gap}_{\min}$, then the suboptimality of $\underline{\pi}$ can still be bounded by the thresholded deficit.

Lemma 6.1. *Suppose for a thresholding function $\epsilon_h(s,a)$, for all $(h,s) \in [H] \times \mathcal{S}$, we have*

$$\underline{\mathbf{V}}_h^*(s) + \frac{\mathrm{gap}_{\min}}{2} \geq \ddot{\mathbf{V}}_h^*(s). \tag{2}$$

Then we can bound the suboptimality of $\underline{\pi}$ by the thresholded deficit:

$$\mathrm{suboptimal}(\underline{\pi}) = \mathbf{V}_0^* - \mathbf{V}_0^{\underline{\pi}} \leq \mathbf{V}_0^* - \underline{\mathbf{V}}_0^* \leq 2(\mathbf{V}_0^* - \ddot{\mathbf{V}}_0^*) = 2\sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\ddot{\mathbf{E}}_h(s_h, a_h)].$$

See Appendix A.2 for the rigorous proof, where this lemma is decomposed and restated as Theorem A.1 and Definition A.5. One way to satisfy (2) is to set $\epsilon_h(s,a) = \frac{\mathrm{gap}_{\min}}{2H}$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$. As a result, the reward at each timestep h is increased by at most $\frac{\mathrm{gap}_{\min}}{2H}$ after thresholding, so the overall increase of the value function is bounded by $\mathrm{gap}_{\min}/2$. This kind of $\epsilon_h(s,a)$ leads to the following corollary.

Corollary 6.1. *For any pessimistic algorithm, we have*

$$\mathbf{V}_0^* - \mathbf{V}_0^{\underline{\pi}} \leq 2\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\max\{0, \mathbf{E}_h(s_h, a_h) - \frac{\mathrm{gap}_{\min}}{2H}\}\right].$$

In addition, note that $\mathbf{E}_h(s_h, a_h)$ depends on variance of the transition p_{h,s_h,a_h} as if the transition variance is small, the estimate will be accurate and the deficit will be small. This inspires us to threshold adaptively based on the variance. We design the following adaptive threshold function:

$$\epsilon_h(s,a) \propto \frac{\mathbf{Var}_{\hat{P}_{h,s,a}}(\underline{\mathbf{V}}_{h+1})}{H^2}\mathrm{gap}_{\min}.$$

It turns out that $\widetilde{\Omega}(\frac{H}{P})$ sample complexity is enough to guarantee this kind of threshold function satisfies (2). (See Lemma B.8 in Appendix B.5.1)

Corollary 6.2. *For any pessimistic algorithm and number of samples $N \geq \widetilde{O}(\frac{H}{P})$, we have*

$$\mathbf{V}_0^* - \mathbf{V}_0^{\underline{\pi}} \leq 2\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\max\{0, \mathbf{E}_h(s_h, a_h) - \frac{\mathbf{Var}_{\hat{P}_{h,s,a}}(\underline{\mathbf{V}}_{h+1}^{\underline{\pi}})}{H^2}\mathrm{gap}_{\min} - \frac{\mathrm{gap}_{\min}}{4H}\}\right].$$

To proceed from Lemma 6.1, we will utilize the following inequality

$$\ddot{\mathbf{E}}_h(s,a) = \max\{0, \mathbf{E}_h(s,a) - \epsilon_h(s,a)\} \leq \frac{\mathbf{E}_h^2(s,a)}{\epsilon_h(s,a)}.$$

Then with different choices of $\epsilon_h(s,a)$ used in Corollary 6.1 and Corollary 6.2 and the bonus function used in the VI-LCB algorithm, we can derive Theorem 5.1 and Theorem 5.2. Here we briefly explain the $\widetilde{O}(1/\epsilon)$ dependence. We have $\mathbf{E}_h(s,a) \leq O(b_h(s,a))$, where $b_h(s,a)$ is the pessimistic bonus scaling as $O(1/\sqrt{n})$ (ignoring other dependences). Then by Lemma 6.1 and the above inequality, we can achieve an $\widetilde{O}(1/\epsilon)$ bound immediately. A more detailed proof is provided in Appendix B.

Our technique is similar to the clip techique [Simchowitz and Jamieson, 2019] in online MDP as both of them threshold the estimates (deficit here and surplus in their work) by some $\Theta(\mathrm{gap}_{\min})$ terms. Simchowitz and Jamieson [2019] further clip another $\frac{\mathrm{gap}_h(s,\bar{a})}{H}$ term, and we can actually achieve that as well, but it does not improve the sample complexity in the offline setting as we do not need to explore all the actions. In addition, we develop a new thresholding function based on the empirical variance, which can improve the H factor by utilizing the Bernstein's inequality and the total variance technique. We believe this new technique can also be applied to the online setting and improve the sample complexity there.

7 Gap-dependent Lower Bounds

In this section, we provide two lower bounds for uniform optimal coverage assumption (Assumption 3.1) and optimal policy coverage assumption (Assumption 3.2) respectively. Our lower bounds show that Theorem 4.1 and Theorem 5.1 are optimal up to H factors and logarithm terms. We begin with a general lower bound. Here the *offline learning algorithm*, $\mathbf{ALG}$, is defined as the algorithm that takes a dataset $\mathcal{D}$ as input and output a policy $\hat{\pi}$. Note that $\mathbf{ALG}$ can be stochastic.

Theorem 7.1. *There exists some absolute constant C, such that for any $A \geq 3, S \geq 2, H \geq 2, \tau < \frac{1}{2}, \lambda \leq \frac{1}{3}, \lambda_1 \geq 2$ and offline RL algorithm $\mathbf{ALG}$, if the number of sample trajectories satisfies*

$$N \leq C \cdot \frac{HS\lambda_1}{\lambda\tau^2},$$

there exists a MDP instance $\mathcal{M}$ and a behavior policy μ with $\mathrm{gap}_{\min} = \tau$, $P \geq \frac{\lambda}{eS\lambda_1}$, $C^* \leq \lambda_1$ such that the output policy $\hat{\pi}$ of $\mathbf{ALG}$ has expected suboptimality

$$\mathbb{E}_{\mathcal{M},\mu,\mathbf{ALG}}[\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}}] \geq \frac{\lambda H\tau}{12}.$$

By choosing $\lambda = 1/3$, $\lambda_1 = \frac{1}{3ePS}$ and $\tau = \mathrm{gap}_{\min}$, Theorem 7.1 indicates the following corollary.

Corollary 7.1. *For any given instance coefficients $(H, \mathrm{gap}_{\min}, P \leq \frac{1}{6eS})$, target suboptimality $\epsilon \lesssim H\mathrm{gap}_{\min}$ and any offline reinforcement learning algorithm $\mathbf{ALG}$, there exists an instance $(\mathcal{M}, \mu)$ such that if the number of trajectory samples satisfies*

$$N \leq C \cdot \frac{H}{P\mathrm{gap}_{\min}^2},$$

the output policy $\hat{\pi}$ would have expected suboptimality more than ϵ, i.e.,

$$\mathbb{E}_{\mathcal{M},\mu,\mathbf{ALG}}[\mathbf{V}_0^* - \mathbf{V}_0^{\pi^*}] \geq \frac{1}{36}H\mathrm{gap}_{\min} \geq \epsilon$$

By choosing $\lambda = \frac{12\epsilon}{H\mathrm{gap}_{\min}}$, $\lambda_1 = C^*$, $\tau = \mathrm{gap}_{\min}$, Theorem 7.1 indicates the following corollary.

Corollary 7.2. *For any given instance coefficients $(H, S, \mathrm{gap}_{\min}, C^* \geq 2)$, target suboptimality $\epsilon \lesssim H\mathrm{gap}_{\min}$ and any offline learning algorithm $\mathbf{ALG}$, there exists an instance $(\mathcal{M}, \mu)$ such that if the number of trajectory samples satisfies*

$$N \leq C \cdot \frac{H^2SC^*}{\epsilon\mathrm{gap}_{\min}},$$

the output policy $\hat{\pi}$ would have expected suboptimality more than ϵ, i.e.,

$$\mathbb{E}_{\mathcal{M},\mu,\mathbf{ALG}}[\mathbf{V}_0^* - \mathbf{V}_0^{\pi^*}] \geq \epsilon.$$

The proof of the corollaries can be found in Appendix C.1. We construct a family of MDPs that involve solving HS independent bandit problems, where each bandit requires $\widetilde{\Omega}(\frac{HA}{\tau^2})$ visits so that the estimation error can be bounded by τ. Similar constructions have been made in Yin et al. [2021a], Dann et al. [2017], Xie et al. [2021b] . Our key observation is that the initial state distribution can linearly determine the uniform optimal policy coverage coefficient P and the final suboptimality. Also, our proof can lead to another version of lower bound, where a $\Omega(\epsilon)$ suboptimality occurs with a constant probability. See Appendix C.2.3 for details.

8 Necessity of All-Optimal-Policy Coverage

As one may notice that Assumption 3.1 and Assumption 3.2 are not exactly the most widely adopted versions [Li et al., 2022, Xie et al., 2021b, Rashidinejad et al., 2021]. Here the assumptions require a coverage of all the optimal policies instead of only one optimal policy. This assumption can not be weaken as long as we are using naive LCB-style algorithm, and a counterexample can be found in Appendix D.

Intuitively, the problem lies in that we can not distinguish between different optimal policies. Gap information helps us distinguish optimal actions from sub-optimal ones when the estimation error is relatively small, but a naive LCB algorithm still has the probability of choosing an optimal action ((s, a_2) in the counterexample) without fully exploring the subsequent state-action pairs. In online gap-dependent analysis, this problem can be addressed by some sophisticated exploration techniques [Xu et al., 2021]. It is unclear if we can weaken the assumption to single policy coverage and we leave this to future work.

9 Conclusion

We presented a systematic study on gap-dependent upper and lower bounds for offline reinforcement learning. Depending on different assumptions, the rates can be improved from $\widetilde{O}(1/\epsilon^2)$ in the worst to $\widetilde{O}(1/\epsilon)$ or even independent of $1/\epsilon$. Also, we presented a counterexample to show that overall optimal policy coverage is necessary in gap-dependent analysis.

One open question is that there still exists a gap between the upper and lower bounds in terms of H. We note that this gap also exists in the online setting [Simchowitz and Jamieson, 2019, Xu et al., 2021]. Another direction is to generalize our results to the function approximation setting [He et al., 2021].

Acknowledgements

SSD acknowledges funding from NSF Award's CCF-2212261, IIS-2143493, DMS-2134106, CCF-2019844 and IIS-2110170.

References

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 456–464, 2019.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. Approximating interactive human evaluation with self-play for open-domain dialog systems. *arXiv preprint arXiv:1906.09308*, 2019.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR, 2021.

Yichun Hu, Nathan Kallus, and Masatoshi Uehara. Fast rates for the regret of offline reinforcement learning, 2021. URL `https://arxiv.org/abs/2102.00479`.

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.

Anders Jonsson, Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Edouard Leurent, and Michal Valko. Planning in markov decision processes with gap-dependent sample complexity. *Advances in Neural Information Processing Systems*, 33:1253–1263, 2020.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022.

Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, pages 1077–1084, 2014.

Susan A Murphy, Mark J van der Laan, James M Robins, and Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34, 2021.

Tongzheng Ren, Jialian Li, Bo Dai, Simon S Du, and Sujay Sanghavi. Nearly horizon-free offline reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. *arXiv preprint arXiv:2202.13890*, 2022.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Csaba Szepesvári and Rémi Munos. Finite time bounds for sampling based fitted value iteration. In *Proceedings of the 22nd international conference on Machine learning*, pages 880–887, 2005.

Andrea Tirinzoni, Matteo Pirotta, and Alessandro Lazaric. A fully problem-dependent regret lower bound for finite-horizon mdps. *arXiv preprint arXiv:2106.13013*, 2021.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

Andrew Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent pac reinforcement learning. *arXiv preprint arXiv:2108.02717*, 2021a.

Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Task-optimal exploration in linear dynamical systems. In *International Conference on Machine Learning*, pages 10641–10652. PMLR, 2021b.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021a.

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, 2021b.

Haike Xu, Tengyu Ma, and Simon Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In *Conference on Learning Theory*, pages 4438–4472. PMLR, 2021.

Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. The efficacy of pessimism in asynchronous q-learning. *arXiv preprint arXiv:2203.07368*, 2022.

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. *arXiv preprint arXiv:2007.03760*, 2020.

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, 2021a.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal offline reinforcement learning via double variance reduction. *Advances in neural information processing systems*, 34, 2021b.

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, Mykel J Kochenderfer, and Emma Brunskill. Almost horizon-free structure-aware best policy identification with a generative model. *Advances in Neural Information Processing Systems*, 32, 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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] Our results contain no direct social or anthropology element.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3.2
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Upper Bound with Gap-dependent Analysis

We begin with the proof of thresholding technique. In this section, definitions are restated for completeness.

A.1 Definitions

We first restate the notations.

Definition A.1 (Pessimistic algorithms)**.** *An offline learning algorithm with output policy $\underline{\pi}$ is pessimistic if with probability at least $1-\delta$, the following arguments hold,*

1. *It maintains a pessimistic estimate $\underline{\mathbf{Q}}$ of the true Q^*.*
2. *$\underline{\mathbf{Q}}$ is the optimal Q function of an imaginary MDP $\underline{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, p_0, \underline{r})$, where $\underline{r}_h(s,a) \leq r_h(s,a)$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$.*
3. *$\underline{\pi}$ is the greedy policy with respect to $\underline{\mathbf{Q}}$.*

Definition A.2 (Pessimistically estimated MDP)**.** *For a given successful pessimistic algorithm execution instance, where the arguments in Definition A.1 are simultaneously satisfied, we call $\underline{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, p_0, \underline{r})$ the pessimistically estimated MDP. At the same time, $\underline{\mathbf{V}}$, $\underline{\mathbf{Q}}$ are the corresponding value functions and Q functions. We use $\underline{\pi}$ to refer to the returned policy, which is optimal over $\underline{\mathcal{M}}$.*

And sometimes we use $\underline{\mathbf{Q}} = \underline{\mathbf{Q}}^{\underline{\pi}}$, $\underline{\mathbf{V}} = \underline{\mathbf{V}}^{\underline{\pi}}$ without superscript indicating the policy. We'll show that this notation matches the definition in Algorithm 1 so there is no need worrying about any possible confusion.

Definition A.3 (Deficit)**.** *For a pessimistically estimated MDP $\underline{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, p_0, \underline{r})$, we define deficit to be*

$$\mathbf{E}_h(s,a) \triangleq r_h(s,a) - \underline{r}_h(s,a).$$

Definition A.4 (Not-so-pessimistic MDP)**.** *For a given set of $\epsilon_h(s,a)$, a pessimistically estimated MDP $\underline{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, p_0, \underline{r})$, define*

$$\ddot{r}_h(s,a) \triangleq r_h(s,a) - \max\{0, \mathbf{E}_h(s,a) - \epsilon_h(s,a)\}.$$

Then we call $\ddot{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, p_0, \ddot{r})$ not-so-pessimistic MDP. At the same time, $\ddot{\mathbf{V}}$ is the corresponding value functions.

A.2 Main Theorem

For conciseness, we will use a^* and $\underline{a}$ to stand for $\pi_h^*(s)$ and $\underline{\pi}_h(s)$ respectively when it introduces no confusion. To formally present the deficit thresholding technique, we define a event ξ_{gap}.

Definition A.5 (Gap restriction event)**.** *For a given set of $\epsilon_h(s,a)$, event ξ_{gap} is defined to be the event such that for all optimal policy π^*, $h \in [H]$ and $s \in \mathcal{S}$,*

$$\ddot{\mathbf{V}}_h^*(s) \leq \underline{\mathbf{V}}_h^*(s) + \frac{\mathrm{gap}_{\min}}{2}.$$

Note that ξ_{gap} depends on the value of $\epsilon_h(s,a)$, and the definition of $\epsilon_h(s,a)$ may involve randomness. In the following proof of Corollary A.1, we will set $\epsilon_h(s,a) = \frac{\mathrm{gap}_{\min}}{2H}$.

Theorem A.1 (Deficit thresholding)**.** *When event ξ_{gap} happens, there exists an optimal policy π^*, such that replacing $\mathbf{V}_0^{\underline{\pi}}$ with $\ddot{\mathbf{V}}_0^*$ only harms the difference up to a constant factor,*

$$\mathbf{V}_0^* - \mathbf{V}_0^{\underline{\pi}} \leq 2(\mathbf{V}_0^* - \ddot{\mathbf{V}}_0^*).$$

Rigorous proof is deferred to Appendix A.4

Corollary A.1. *For a pessimistic algorithm running instance, there exists a deterministic optimal policy π^*, such that*

$$\mathbf{V}_0^* - \mathbf{V}_0^{\underline{\pi}} \leq 2\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\max\{0, \mathbf{E}_h(s_h,a_h) - \frac{\mathrm{gap}_{\min}}{2H}\}\right].$$

Proof. With Theorem A.1, we just need to prove that $\epsilon_h(s,a) = \frac{\text{gap}_{\min}}{2H}$ indicates ξ_{gap}.
Note that $\mathbf{E}_h(s,a^*) \le \ddot{\mathbf{E}}_h(s,a^*) + \epsilon_h(s,a^*) = \ddot{\mathbf{E}}_h(s,a^*) + \frac{\text{gap}_{\min}}{2H}$. Therefore, for all optimal policy π^*, we have

$$\begin{aligned}\ddot{\mathbf{V}}_h^*(s) - \underline{\mathbf{V}}_h^*(s) &= \sum_{h'=h}^{H} \mathbb{E}_{\pi^*, s_h=s}[-\ddot{\mathbf{E}}_{h'}(s_{h'}, a_{h'}) + \mathbf{E}_{h'}(s_{h'}, a_{h'})] \\ &\le \sum_{h'=h}^{H} \mathbb{E}_{\pi^*, s_h=s}[\frac{\text{gap}_{\min}}{2H}] \\ &\le \frac{\text{gap}_{\min}}{2}.\end{aligned}$$

□

A.3 Value/Q Function Ranking Lemma

The following lemmas will be frequently used throughout the proof of Theorem A.1 and upper bounds.

Lemma A.1 (Overall size relationships of value functions)**.** *When ξ_{gap} happens, different value functions satisfy that for any optimal policy π^*, we have*

$$\mathbf{V}^* \ge \begin{cases}\mathbf{V}^{\underline{\pi}} \ge \underline{\mathbf{V}}^{\underline{\pi}} \\ \ddot{\mathbf{V}}^*\end{cases} \ge \underline{\mathbf{V}}^* \ge \ddot{\mathbf{V}}^* - \frac{\text{gap}_{\min}}{2}.$$

Here $\mathbf{V} \ge \mathbf{V}'$ means $\mathbf{V}_h(s) \ge \mathbf{V}'_h(s)$ for all $(h,s) \in [H] \times \mathcal{S}$.

Proof. We study each inequality one by one:
$\mathbf{V}^* \ge \mathbf{V}^{\underline{\pi}}$: π^* is a optimal policy over $\mathcal{M}$.
$\mathbf{V}^{\underline{\pi}} \ge \underline{\mathbf{V}}^{\underline{\pi}}$: this follows from $\underline{r} \le r$,

$$\mathbf{V}_h^{\underline{\pi}}(s) = \sum_{h'=h}^{H} \mathbb{E}_{\underline{\pi}}[r_{h'}(s_{h'}, a_{h'}) \mid s_h = s] \ge \sum_{h'=h}^{H} \mathbb{E}_{\underline{\pi}}[\underline{r}_{h'}(s_{h'}, a_{h'}) \mid s_h = s] = \underline{\mathbf{V}}_h^{\underline{\pi}}(s).$$

$\underline{\mathbf{V}}^{\underline{\pi}} \ge \underline{\mathbf{V}}^*$: $\underline{\pi}$ is a optimal policy over $\underline{\mathcal{M}}$.
$\mathbf{V}^* \ge \ddot{\mathbf{V}}^*$: this follows from $\ddot{r} \le r$,

$$\mathbf{V}_h^*(s) = \sum_{h'=h}^{H} \mathbb{E}_{\pi^*}[r_{h'}(s_{h'}, a_{h'}) \mid s_h = s] \ge \sum_{h'=h}^{H} \mathbb{E}_{\pi^*}[\ddot{r}_{h'}(s_{h'}, a_{h'}) \mid s_h = s] = \ddot{\mathbf{V}}_h^*(s).$$

$\ddot{\mathbf{V}}^* \ge \underline{\mathbf{V}}^*$: this follows from $\underline{r} \le \ddot{r}$,

$$\ddot{\mathbf{V}}_h^*(s) = \sum_{h'=h}^{H} \mathbb{E}_{\pi^*}[\ddot{r}_{h'}(s_{h'}, a_{h'}) \mid s_h = s] \ge \sum_{h'=h}^{H} \mathbb{E}_{\pi^*}[\underline{r}_{h'}(s_{h'}, a_{h'}) \mid s_h = s] = \underline{\mathbf{V}}_h^*(s).$$

$\underline{\mathbf{V}}^* \ge \ddot{\mathbf{V}}^* - \frac{\text{gap}_{\min}}{2}$: this is just the definition of ξ_{gap}. □

Lemma A.2 (Overall size relationships of Q functions)**.** *When ξ_{gap} happens, different Q functions satisfy that for any optimal policy π^*, we have*

$$\mathbf{Q}_h^*(s,a^*) \ge \mathbf{Q}_h^*(s,\underline{a}) \ge \mathbf{Q}_h^{\underline{\pi}}(s,\underline{a}) \ge \underline{\mathbf{Q}}_h^{\underline{\pi}}(s,\underline{a}) \ge \underline{\mathbf{Q}}_h^{\underline{\pi}}(s,a^*) \ge \underline{\mathbf{Q}}_h^*(s,a^*).$$

Proof. We study each inequality one by one:
$\mathbf{Q}_h^*(s,a^*) \ge \mathbf{Q}_h^*(s,\underline{a})$: a^* is the optimal action at (h,s) over $\mathcal{M}$.
$\mathbf{Q}_h^*(s,\underline{a}) \ge \mathbf{Q}_h^{\underline{\pi}}(s,\underline{a})$: this follows from $\mathbf{V}^* \ge \mathbf{V}^{\underline{\pi}}$ in Lemma A.1,

$$\mathbf{Q}_h^*(s,\underline{a}) = \mathbb{E}_{s' \sim P_{h,s,\underline{a}}}[\mathbf{V}_{h+1}^*(s')] \ge \mathbb{E}_{s' \sim P_{h,s,\underline{a}}}[\mathbf{V}_{h+1}^{\underline{\pi}}(s')] = \mathbf{Q}_h^{\underline{\pi}}(s,\underline{a}).$$

$\mathbf{Q}_h^{\underline{\pi}}(s,\underline{a}) \ge \underline{\mathbf{Q}}_h^{\underline{\pi}}(s,\underline{a})$: this follows from $\mathbf{V}^{\underline{\pi}} \ge \underline{\mathbf{V}}^{\underline{\pi}}$ in Lemma A.1.
$\underline{\mathbf{Q}}_h^{\underline{\pi}}(s,\underline{a}) \ge \underline{\mathbf{Q}}_h^{\underline{\pi}}(s,a^*)$: $\underline{a}$ is the optimal action at (h,s) over $\underline{\mathcal{M}}$.
$\underline{\mathbf{Q}}_h^{\underline{\pi}}(s,a^*) \ge \underline{\mathbf{Q}}_h^*(s,a^*)$: this follows from $\underline{\mathbf{V}}^{\underline{\pi}} \ge \underline{\mathbf{V}}^*$ in Lemma A.1,

$$\underline{\mathbf{Q}}_h^{\underline{\pi}}(s,a^*) = \mathbb{E}_{s' \sim P_{h,s,a^*}}[\underline{\mathbf{V}}_{h+1}^{\underline{\pi}}(s')] \ge \mathbb{E}_{s' \sim P_{h,s,a^*}}[\underline{\mathbf{V}}_{h+1}^*(s')] = \underline{\mathbf{Q}}_h^*(s,a^*).$$

□

A.4 Proof of Theorem A.1

Proof. In this proof, we choose the π^* according to the given $\underline{\pi}$,

$$\pi^*_h(s) = \begin{cases} \underline{\pi}_h(s) & \underline{\pi}_h(s) \text{ is optimal,} \\ \text{arbitrary optimal action} & \underline{\pi}_h(s) \text{ is not optimal.} \end{cases}$$

So that every time π^* disagrees with $\underline{\pi}$, the choice made by $\underline{\pi}$ must be suboptimal. The intuition is that we only consider the cases where π^* and $\underline{\pi}$ have different opinions. To begin with, we define a set of prefix trajectories for any two given deterministic policies over MDPs that only differ in rewards:

$$\begin{aligned}\Psi(\pi_1, \pi_2) =& \{(s_1, a_1, \cdots, s_k) \mid \pi_{1,i}(s_i) = \pi_{2,i}(s_i) = a_i, \forall i = 1, 2, \cdots k-1, \\ & \pi_1(s_k) \neq \pi_2(s_k) \text{ or } k = H\}.\end{aligned}$$

And we use P^π_ψ to denote the probability that we can get a prefix trajectory $\psi = (s_1, a_1, \cdots, s_k)$ with a deterministic policy π,

$$P^\pi_\psi \triangleq p_0(s_1) \prod_{h=1}^{k-1} p_h(s_h, \pi_h(s_h), s_{h+1}).$$

Notice that for any given trajectory ξ and policy π_1, π_2, there is exactly one prefix trajectory $\psi \in \Psi(\pi_1, \pi_2)$ being the prefix of ξ, which ends at the first time π_1 disagrees with π_2. Denote the length and the last state of a trajectory to be $2h_\psi - 1$ and s_ψ, and set the cumulative reward over ψ under a given deterministic reward function $r_h(s, a)$ to be r_ψ, we can write the value function in the form

$$V^{\pi_1}_0 = \sum_{\psi \in \Psi(\pi_1, \pi_2)} P^{\pi_1}_\psi \left(r_\psi + \mathbf{V}^{\pi_1}_{h_\psi}(s_\psi) \right).$$

Also, notice that because π_1 and π_2 agrees on all the decisions in $\psi \in \Psi(\pi_1, \pi_2)$, we always have $P^{\pi_1}_\psi = P^{\pi_2}_\psi$ for $\psi \in \Psi(\pi_1, \pi_2)$. Now we have

$$\begin{aligned}\mathbf{V}^*_0 - \mathbf{V}^{\underline{\pi}}_0 &= \sum_{\psi \in \Psi(\pi^*, \underline{\pi})} P^{\pi^*}_\psi (\mathbf{V}^*_{h_\psi}(s_\psi) - \mathbf{V}^{\underline{\pi}}_{h_\psi}(s_\psi)) \\ &= \sum_{\psi \in \Psi(\pi^*, \underline{\pi})} P^{\pi^*}_\psi (\mathbf{Q}^*_{h_\psi}(s_\psi, \underline{a}) + \mathrm{gap}_{h_\psi}(s_\psi, \underline{a}) - \mathbf{V}^{\underline{\pi}}_{h_\psi}(s_\psi)).\end{aligned}$$

Then we prove a statement that $\forall \psi \in \Psi(\pi^*, \underline{\pi})$,

$$\mathbf{Q}^*_{h_\psi}(s_\psi, \underline{a}) + \mathrm{gap}_{h_\psi}(s_\psi, \underline{a}) - \ddot{\mathbf{V}}^*_{h_\psi}(s_\psi) \geq \mathbf{Q}^*_{h_\psi}(s_\psi, \underline{a}) + \frac{1}{2}\mathrm{gap}_{h_\psi}(s_\psi, \underline{a}) - \mathbf{V}^{\underline{\pi}}_{h_\psi}(s_\psi). \tag{3}$$

When $\underline{a} = a^*$, the only possibility is that $h_\psi = H$, then RHS=0. While the LHS is always non-negative because $\mathbf{Q}^*_{h_\psi}(s_\psi, \underline{a}) = \mathbf{V}^*_{h_\psi}(s_\psi) \geq \ddot{\mathbf{V}}^*_{h_\psi}(s_\psi)$ (Lemma A.1).
When $\underline{a} \neq a^*$, event ξ_{gap} guarantees that

$$\begin{aligned}\mathbf{Q}^*_{h_\psi}(s_\psi, \underline{a}) + \mathrm{gap}_{h_\psi}(s_\psi, \underline{a}) - \ddot{\mathbf{V}}^*_{h_\psi}(s_\psi) &\geq \mathbf{Q}^*_{h_\psi}(s_\psi, \underline{a}) + \mathrm{gap}_{h_\psi}(s_\psi, \underline{a}) - \underline{\mathbf{V}}^*_{h_\psi}(s_\psi) - \frac{\mathrm{gap}_{\min}}{2} \\ &\geq \mathbf{Q}^*_{h_\psi}(s_\psi, \underline{a}) + \frac{1}{2}\mathrm{gap}_{h_\psi}(s_\psi, \underline{a}) - \mathbf{V}^{\underline{\pi}}_{h_\psi}(s_\psi).\end{aligned}$$

The inequality uses that $\mathrm{gap}_{\min} \leq \mathrm{gap}_{h_\psi}(s_\psi, \underline{a})$ and that $\underline{\mathbf{V}}^* \leq \underline{\mathbf{V}}^{\underline{\pi}} \leq \mathbf{V}^{\underline{\pi}}$. At the same time, we can decompose $\mathbf{V}^*_0 - \ddot{\mathbf{V}}^*_0$ in a similar way,

$$\begin{aligned}\mathbf{V}^*_0 - \ddot{\mathbf{V}}^*_0 &= \sum_{\psi \in \Psi(\pi^*, \underline{\pi})} P^{\pi^*}_\psi (r_\psi - \ddot{r}_\psi + \mathbf{V}^*_{h_\psi}(s_\psi) - \ddot{\mathbf{V}}^*_{h_\psi}(s_\psi)) \\ &\geq \sum_{\psi \in \Psi(\pi^*, \underline{\pi})} P^{\pi^*}_\psi (\mathbf{Q}^*_{h_\psi}(s_\psi, \underline{a}) + \mathrm{gap}_{h_\psi}(s_\psi, \underline{a}) - \ddot{\mathbf{V}}^*_{h_\psi}(s_\psi)) \qquad (4) \\ &\geq \sum_{\psi \in \Psi(\pi^*, \underline{\pi})} P^{\pi^*}_\psi (\mathbf{Q}^*_{h_\psi}(s_\psi, \underline{a}) + \frac{1}{2}\mathrm{gap}_{h_\psi}(s_\psi, \underline{a}) - \mathbf{V}^{\underline{\pi}}_{h_\psi}(s_\psi)) \qquad (5) \\ &\geq \frac{1}{2} \sum_{\psi \in \Psi(\pi^*, \underline{\pi})} P^{\pi^*}_\psi (\mathbf{Q}^*_{h_\psi}(s_\psi, \underline{a}) + \mathrm{gap}_{h_\psi}(s_\psi, \underline{a}) - \mathbf{V}^{\underline{\pi}}_{h_\psi}(s_\psi)) \qquad (6) \\ &= \frac{1}{2}(\mathbf{V}^*_0 - \mathbf{V}^{\underline{\pi}}_0).\end{aligned}$$

(4) results from the fact that r is always larger than or equal to $\ddot{r}$. (5) just makes use of (3). (6) uses $\mathbf{Q}_h^*(s,\underline{a}) \geq \mathbf{Q}_h^{\pi}(s,\underline{a}) = \mathbf{V}_h^{\pi}(s)$(Lemma A.2). □

B VI-LCB based analysis

B.1 Algorithm Sketch and Notations

Algorithm used here is Lower Confidence Bound Value Iteration(VI-LCB)[Xie et al., 2021b] with subsampling trick and Berstein-style bonus. The basic idea of LCB is to pessimistically estimate the Q function so that the algorithm won't over estimate some hardly seen suboptimal actions in dataset. The subsampling trick introduced by Li et al. [2022] helps solve the independence problem between $\hat{P}_h$ and $\underline{\mathbf{V}}_{h+1}^{\pi}$, which avoid separating the dataset into H parts, resulting in one H dependency removed in final complexity.

Here we understand dataset as a set of transitions in the form (h,s,a,s') that allows duplicates. When we say that the dataset contains a trajectory $(s_1,a_1,\cdots,s_h,a_h)$, it means that the dataset contains all the decomposed transitions $\{(h,s_h,a_h,s_{h+1})\}_{h=1,\cdots H}$. Also note that $\mathcal{M}$ has deterministic rewards in our setting, so the reward function can be easily derived as long as the (h,s,a) tuple is visited for at least once. And if (h,s,a) is not contained in $\mathcal{D}$, the algorithm output wouldn't be influenced by the value of $r_h(s,a)$, and we can set $r_h(s,a)=0$. So we assume that the reward function is known from the beginning.

Algorithm 3: VI-LCB

input : Dataset $\mathcal{D}_0$, reward function r

set $\underline{\mathbf{Q}}_{H+1}^{\pi}(s,a)=0$
set $\underline{\mathbf{V}}_{H+1}^{\pi}(s,a)=0$
for $h \leftarrow H$ **to** 1 **do**
  compute the empirical transition kernel $\hat{P}_h$
  $\hat{P}_{h,s,a}(s') = \frac{N_h(s,a,s')}{N_h(s,a)}$ with $0/0=0$
  for $s\in\mathcal{S}, a\in\mathcal{A}$ **do**
    $b_h(s,a) \leftarrow C_b\sqrt{\frac{\mathbf{Var}_{\hat{P}_{h,s,a}}(\underline{\mathbf{V}}_{h+1}^{\pi})\iota}{N_h'(s,a)}} + C_b\frac{H\iota}{N_h'(s,a)}$, where $N_h'(s,a)=N_h(s,a)\vee\iota$
    $\underline{\mathbf{Q}}_h^{\pi}(s,a) \leftarrow \max\{0, r_h(s,a)+\hat{P}_{h,s,a}^{\top}\underline{\mathbf{V}}_{h+1}^{\pi} - b_h(s,a)\}$
  for $s\in\mathcal{S}$ **do**
    $\underline{\mathbf{V}}_h^{\pi}(s) \leftarrow \max_{a\in\mathcal{A}} \underline{\mathbf{Q}}_h^{\pi}(s,a)$
    $\underline{\pi}_h(s) \leftarrow \arg\max_{a\in\mathcal{A}} \underline{\mathbf{Q}}_h^{\pi}(s,a)$

output : policy $\underline{\pi}$

Algorithm 4: Subsampled VI-LCB

input : Dataset $\mathcal{D}$, reward function r

Split $\mathcal{D}$ into 2 halves containing same number of sample trajectories, $\mathcal{D}^{\text{main}}$ and $\mathcal{D}^{\text{aux}}$
$\mathcal{D}_0=\{\}$
for $(h,s)\in[H]\times\mathcal{S}$ **do**
  $N_h^{\text{trim}}(s) \leftarrow \max\{0, N_h^{\text{aux}}(s) - 10\sqrt{N_h^{\text{aux}}(s)\log\frac{HS}{\delta}}\}$
  Randomly subsample $\min\{N_h^{\text{trim}}(s), N_h^{\text{main}}(s)\}$ samples of transition from (h,s) from $\mathcal{D}^{\text{main}}$ to add to $\mathcal{D}_0$
$\underline{\pi} \leftarrow$ VI-LCB$(\mathcal{D}_0, r)$

output : policy $\underline{\pi}$

Notations in VI-LCB. In the algorithm, $N_h(s,a)$ refers to the number of sample transitions starting from state s, taking action a at time step h of some given dataset, and $N_h(s) = \sum_{a\in\mathcal{A}} N_h(s,a)$. Superscripts stand for the dataset. See Li et al. [2022] for a more detailed description of the algorithm.

The proof of independence between samples with different h in $\mathcal{D}_0$ is omitted here, and we will not need it directly because the proof of Lemma B.1 suggests it.

Note that different from the notation in Algorithm 1, we use $\underline{\mathbf{V}}^{\underline{\pi}}$ and $\underline{\mathbf{Q}}^{\underline{\pi}}$ instead of $\underline{\mathbf{V}}$ and $\underline{\mathbf{Q}}$ in Algorithm 4. We do this to emphasize that $\underline{\mathbf{V}}$ and $\underline{\mathbf{Q}}$ in Algorithm 1 directly satisfies the definitions of $\underline{\mathbf{V}}^{\underline{\pi}}$ and $\underline{\mathbf{Q}}^{\underline{\pi}}$ in thresholding technique (Definition A.2), which will be rigorously proved in Lemma B.3. To avoid unnecessary confusion or reading difficulty, $\underline{\mathbf{V}}$ and $\underline{\mathbf{Q}}$ without superscript stands for the true optimal Q/value functions of $\underline{\mathcal{M}}$, i.e., $\underline{\mathbf{V}}^{\underline{\pi}}$ and $\underline{\mathbf{Q}}^{\underline{\pi}}$, in the following proof.

B.2 Proof Preparation

To warm up, we first prove that VI-LCB perfectly matchs our definition of LCB-style algorithm. From the original paper of VI-LCB [Li et al., 2022], we quote a slightly modified version of their lemma 6, where constants and notations are changed, and V is replaced with $\underline{\mathbf{V}}^{\underline{\pi}}_{h+1}$.

Lemma B.1 (Transition estimation bound)**.** *For any $1 \le h \le H$, with probability at least $1-\frac{\delta}{2H}$, we have*

$$|(\hat{P}_{h,s,a}-P_{h,s,a})^\top \underline{\mathbf{V}}_{h+1}| \le b_n(s,a) = C_b\sqrt{\frac{\mathbf{Var}_{\hat{P}_{h,s,a}}(\underline{\mathbf{V}}^{\underline{\pi}}_{h+1})\iota}{N_h(s,a)}}+C_b\frac{H\iota}{N_h(s,a)}, \qquad \forall (s,a)\in\mathcal{S}\times\mathcal{A}.$$

Proof for Lemma B.1 is omitted here. With the union bound, we have the inequality in Lemma B.1 holds for all $h\in[H]$ with probability over $1-\frac{\delta}{2}$. Also, Lemma 1 from original paper helps with the guarantee of the sample number.

Lemma B.2. *With probability over $1-\frac{\delta}{2}$, we have*

$$N_h(s,a) \ge C_{data}(Nd_h^\mu(s,a) - \sqrt{Nd_h^\mu(s,a)\iota}), \qquad \forall (h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A},$$

for some positive constant C_{data}.

Proof of this lemma is also omitted, which is a direct result of Binomial concentration. Then we can prove the concentration lemma, which serves as the basis of following analysis.

Lemma B.3. *If we run VI-LCB on a offline learning instance $(\mathcal{M},\mu)$, with high probability (over $1-\delta$), the following event ξ_{conc} happens for some positive constant C_d:*

1. *the execution instance satisfies the three arguments in Definition A.1, and $0 \le \mathbf{E}_h(s,a) \le 2b_h(s,a)$ for all $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$.*
2. *$N'_h(s,a) = N_h(s,a) \vee \iota \ge C_d N d_h^\mu(s,a)$ for all $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$.*

Proof. We just need to prove that both statements are true with probability over $1-\frac{\delta}{2}$ respectively, and then we can finish the proof by applying union bound.

Proof of statement 1: $\underline{\mathbf{Q}}$ in the definition of pessimistic algorithm matches the $\underline{\mathbf{Q}}$ in VI-LCB. We first prove that $\underline{\mathcal{M}}$ exists. With $\underline{\mathbf{Q}}$ given, we can actually get a closed form of $\underline{r}$,

$$\begin{aligned}
\underline{r}_h(s,a) &= \underline{\mathbf{Q}}_h(s,a) - \mathbb{E}_{\underline{\pi}|s_h=h,a_h=a}[\underline{\mathbf{Q}}_h(s_{h+1},a_{h+1})] \\
&= \underline{\mathbf{Q}}_h(s,a) - \sum_{s'\in\mathcal{S}} P_{h,s,a}(s')\underline{\mathbf{Q}}_{h+1}(s',\underline{a}) \\
&= \underline{\mathbf{Q}}_h(s,a) - P_{h,s,a}^\top \underline{\mathbf{V}}_{h+1}.
\end{aligned}$$

Then we can find that the definition of $\underline{\mathbf{V}}$ in the algorithm agrees with the one in Definition A.2, and we won't distinguish between these two definitions in following induction. It remains to show that

$$0 \le \mathbf{E}_h(s,a) \triangleq r_h(s,a) - \underline{r}_h(s,a) \le 2b_h(s,a). \tag{7}$$

Both inequalities follow from Lemma B.1. Recall that $\underline{\mathbf{Q}}_h(s,a) = \max\{r_h(s,a) + \hat{P}_{h,s,a}^\top \underline{\mathbf{V}}_{h+1} - b_h(s,a), 0\}$,

$$\begin{aligned}\underline{\mathbf{Q}}_h(s,a) &= \max\{0, r_h(s,a) + \hat{P}_{h,s,a}^\top \underline{\mathbf{V}}_{h+1} - b_h(s,a)\} \\ &= \max\{0, P_{h,s,a}^\top \underline{\mathbf{V}}_{h+1} + r_h(s,a) + (\hat{P}_{h,s,a} - P_{h,s,a})^\top \underline{\mathbf{V}}_{h+1} - b_h(s,a)\} \\ &\leq \max\{0, P_{h,s,a}^\top \underline{\mathbf{V}}_{h+1} + r_h(s,a)\} = P_{h,s,a}^\top \underline{\mathbf{V}}_{h+1} + r_h(s,a).\end{aligned}$$

Simple transformation of above inequality leads to

$$r_h(s,a) - \underline{r}_h(s,a) \geq \underline{\mathbf{Q}}_h(s,a) - P_{h,s,a}^\top \underline{\mathbf{V}}_{h+1} - \underline{r}_h(s,a) = 0.$$

The second inequality is also straight forward. We first unfold the definitions of r and $\underline{r}$, then apply Lemma B.1 to get

$$\begin{aligned}r_h(s,a) - \underline{r}_h(s,a) &\leq \underline{\mathbf{Q}}_h(s,a) - \hat{P}_{h,s,a}^\top \underline{\mathbf{V}}_{h+1} + b_h(s,a) - (\underline{\mathbf{Q}}_h(s,a) - P_{h,s,a}^\top \underline{\mathbf{V}}_{h+1}) \\ &= b_h(s,a) - (\hat{P}_{h,s,a} - P_{h,s,a})^\top \underline{\mathbf{V}}_{h+1} \\ &\leq 2b_h(s,a).\end{aligned}$$

Proof of statement 2: We prove over the assumption of event: $N_h(s,a) \geq C_{data}(Nd_h^\mu(s,a) - \sqrt{Nd_h^\mu(s,a)\iota})$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, which is proved by Lemma B.2 to happen with probability over $1 - \frac{\delta}{2}$.
When $Nd_h^\mu(s,a) - \sqrt{Nd_h^\mu(s,a)\iota} \leq \frac{\iota}{C_{data}}$, simple calculation leads to

$$\sqrt{Nd_h^\mu(s,a)} \leq \frac{1 + \sqrt{1 + \frac{4}{C_{data}}}}{2}\sqrt{\iota} = \lambda\sqrt{\iota},$$

where $\lambda = \frac{1+\sqrt{1+\frac{4}{C_{data}}}}{2}$ is a constant larger than 1. Therefore

$$N_h(s,a) \vee \iota \geq \iota \geq \frac{1}{\lambda^2} Nd_h^\mu(s,a). \tag{8}$$

When $Nd_h^\mu(s,a) - \sqrt{Nd_h^\mu(s,a)\iota} \geq \frac{\iota}{C_{data}}$, simple calculation leads to

$$\begin{aligned}&\sqrt{Nd_h^\mu(s,a)} \geq \frac{1 + \sqrt{1 + \frac{4}{C_{data}}}}{2}\sqrt{\iota} = \lambda\sqrt{\iota} \\ &\Leftrightarrow Nd_h^\mu(s,a) - \sqrt{Nd_h^\mu(s,a)\iota} \geq (1 - \frac{1}{\lambda})Nd_h^\mu(s,a) \\ &\Rightarrow N_h(s,a) \vee \iota \geq C_{data}(Nd_h^\mu(s,a) - \sqrt{Nd_h^\mu(s,a)\iota}) \geq C_{data}(1 - \frac{1}{\lambda})Nd_h^\mu(s,a).\end{aligned} \tag{9}$$

Then together with (8) and (9), and letting $C_d = C_{data}(1 - \frac{1}{\lambda}) \wedge 1$, we finish the proof of statement 2. □

Lemma B.4. *When $0 \leq x \leq y$, for $\epsilon > 0$*

$$\max\{0, x - \epsilon\} \leq \frac{y^2}{\epsilon}.$$

Proof. When $x \leq \epsilon$, $\max\{0, x - \epsilon\} = 0 \leq \frac{y^2}{\epsilon}$.
When $x > \epsilon$,

$$\max\{0, x - \epsilon\} \leq y \leq y \cdot \frac{x}{\epsilon} \leq \frac{y^2}{\epsilon}.$$

□

B.3 Proof of Upper Bound with Relative Optimal Policy Coverage(Proof of Theorem 5.1)

This analysis is actually made with Hoeffding bonus for simplicity. Because Berstein bonus is larger than Hoeffding bonus up to log term,

$$b_h(s,a) = C_b\sqrt{\frac{\mathbf{Var}_{\hat{P}_{h,s,a}}(\underline{\mathbf{V}}^{\pi}_{h+1})\iota}{N'_h(s,a)}} + \frac{C_b H\iota}{N'_h(s,a)} \leq 2C_b\sqrt{\frac{H^2\iota^2}{N'_h(s,a)}}. \tag{10}$$

With Corollary A.1 and Lemma B.4,

$$\begin{aligned}
\mathbf{V}^*_0 - \mathbf{V}^{\pi}_0 &\leq 2\sum_{h=1}^{H}\mathbb{E}_{\pi^*}[\max\{0, \mathbf{E}_h(s,a) - \frac{\text{gap}_{\min}}{2H}\}] && \text{(Corollary A.1)}\\
&= 2\sum_{h,s} d^*_h(s)\max\{0, \mathbf{E}_h(s,a) - \frac{\text{gap}_{\min}}{2H}\}\\
&\lesssim \sum_{h,s} d^*_h(s)\frac{b^2_h(s,a^*)}{\frac{\text{gap}_{\min}}{2H}} && \text{(Lemma B.4 and } \xi_{\text{conc}})\\
&\lesssim \sum_{h,s} d^*_h(s)\frac{H^3\iota^2}{N_h(s,a^*)'\text{gap}_{\min}}\\
&\lesssim \sum_{h,s} d^*_h(s)\frac{H^3\iota^2}{N d^{\mu}_h(s,a)\text{gap}_{\min}} && (\xi_{\text{conc}})\\
&\lesssim \frac{1}{N}\sum_{h,s} d^*_h(s)\frac{H^3C^*\iota^2}{d^*_h(s)\text{gap}_{\min}} && \text{(relative optimal coverage assumption)}\\
&= \frac{1}{N}\frac{H^4SC^*\iota^2}{\text{gap}_{\min}}.
\end{aligned}$$

Therefore, under relative optimal policy coverage, the sample complexity bound can be

$$N = O(\frac{H^4SC^*\iota^2}{\epsilon\,\text{gap}_{\min}}).$$

A similar proof in Section B.5 can be applied to prove this result by replacing all the $NP \gtrsim H\iota$ requirements with $N_h(s,a)' \triangleq N_h(s,a) \vee \iota \geq \iota$. So the strict bound without extra ι can be derived. To avoid redundancy, we omit the proof.

$$N = O(\frac{H^4SC^*\iota}{\epsilon\,\text{gap}_{\min}}).$$

B.4 Proofs of Upper Bound with Uniform Optimal Policy Coverage(Proof of Theorem 4.1)

Proof of Theorem 4.1 does not necessarily involve the deficit thresholding technique introduced above. We just need to confirm that $\underline{\mathbf{Q}}^{\pi}_h(s,a^*) \geq \mathbf{V}^*_h(s) - \text{gap}_{\min} \geq \mathbf{V}^*_h(s,a') \geq \underline{\mathbf{Q}}^{\pi}_h(s,a')$, where a' is any suboptimal action, to get a optimal policy. We first present the proof applying this idea, and then present a simpler proof by applying the thresholding technique.

B.4.1 Proof without Deficit Thresholding Technique

First we introduce a new definition,

$d^*_{h\sim(h',s')}(s) \triangleq \mathbb{E}_{\pi^*}[\mathbb{I}\{s_h = s\} \mid s_{h'} = s']$,

$d^*_{h\sim(h',s')} \triangleq (d^*_{h\sim(h',s')}(s_1), \cdots, d^*_{h\sim(h',s')}(s_S))^{\top}$ for some certain order of states $s_1, s_2, \cdots s_S$.

And when there is no confusion, we use $d^{*'}_h$ to denote $d^*_{h\sim(h',s')}$.

Lemma B.5 (Part Decomposition). *$\forall (h', s') \in [H] \times \mathcal{S}$, if the event ξ_{conc} happens, and $P > 0$, then $\forall$ optimal policy π^*,*

$$\sum_{h=h'}^{H} \sum_{s} d^*_{h\sim(h',s')}(s) b_{h'}(s, a^*) \leq C_e \sqrt{\frac{H^3 \iota}{NP}} + C_e \frac{H^2 \iota}{NP},$$

where $C_e = \max\{4\frac{C_b}{\sqrt{C_d}}, \frac{16C_b^2+12C_b}{C_d}, 1\}$.

With this lemma, we can further limit $\underline{\mathbf{Q}}^{\underline{\pi}}_h(s, a^*)$.

$$\begin{aligned}
\mathbf{V}^*_{h'}(s') - \underline{\mathbf{Q}}^{\underline{\pi}}_{h'}(s', a^*) &\leq \mathbf{V}^*_{h'}(s') - \underline{\mathbf{Q}}^*_{h'}(s', a^*) && (\underline{\pi} \text{ is the optimal policy over } \underline{\mathcal{M}}) \\
&= \sum_{h=h'}^{H} \sum_{s} d^*_{h\sim(h',s')}(s) \mathbf{E}_h(s, a^*) \\
&\leq 2 \sum_{h,s} d^{*'}_h b_h(s, a^*) && (\xi_{\text{conc}}) \\
&\leq 2C_e \sqrt{\frac{H^3 \iota}{NP}} + 2C_e \frac{H^2 \iota}{NP}. && (\text{Lemma B.5})
\end{aligned}$$

When $N \geq \frac{4C_e^2 H^3 \iota}{\lambda^2 P}$ for some $\lambda \leq H$,

$$\begin{aligned}
&\mathbf{V}^*_{h'}(s') - \underline{\mathbf{Q}}^{\underline{\pi}}_{h'}(s', a^*) \\
&\leq \frac{\lambda}{2} + \frac{\lambda}{2} \frac{\lambda}{2C_e H} \\
&\leq \lambda.
\end{aligned}$$

Setting $\lambda = \text{gap}_{\min}$, we get the conclusion that

$$N = \frac{4C_e^2 H^3 \iota}{\text{gap}_{\min}^2 P}$$

can make sure that the returned policy is one of the optimal policies with probability over $1 - \delta$.

B.4.2 Proof with Deficit Thresholding Technique

By applying the Lemma B.8 which is orginally developed for the proof of Theorem 5.2, we shall directly prove that the suboptimality would be zero if $N > \frac{CH^3\iota}{P\text{gap}_{\min}}$. Lemma B.8 allows us to set $\epsilon_h(s, a) = C_{pac}\left(\frac{\mathbf{Var}_{\hat{P}_{h,s,a}}(\underline{\mathbf{V}}^{\underline{\pi}}_{h+1})}{H^2} + \frac{1}{H}\right)\text{gap}_{\min}$. When $N \geq C\frac{H^3\iota}{P\text{gap}_{\min}^2}$, for any optimal policy π^*,

$$\begin{aligned}
b_h(s, a^*) &= C_b \sqrt{\frac{\mathbf{Var}_{\hat{P}_{h,s,a^*}}(\underline{\mathbf{V}}^{\underline{\pi}}_{h+1})\iota}{N'_h(s, a^*)}} + C_b \frac{H\iota}{N_h(s, a^*)} \\
&\lesssim \sqrt{\frac{\mathbf{Var}_{\hat{P}_{h,s,a^*}}(\underline{\mathbf{V}}^{\underline{\pi}}_{h+1})\iota}{NP}} + \frac{H\iota}{NP} && (\xi_{\text{conc}}) \\
&\leq \sqrt{\frac{\mathbf{Var}_{\hat{P}_{h,s,a^*}}(\underline{\mathbf{V}}^{\underline{\pi}}_{h+1})\text{gap}_{\min}^2}{CH^3}} + \frac{\text{gap}_{\min}^2}{CH^2} && (N \geq C\frac{H^3\iota}{P\text{gap}_{\min}^2}) \\
&\leq \frac{\mathbf{Var}_{\hat{P}_{h,s,a^*}}(\underline{\mathbf{V}}^{\underline{\pi}}_{h+1})\text{gap}_{\min}}{2CH^2} + \frac{\text{gap}_{\min}}{2H} + \frac{\text{gap}_{\min}}{CH} \quad a + b \geq 2\sqrt{ab},\ \text{gap}_{\min} \leq H \\
&\lesssim \epsilon_h(s, a^*).
\end{aligned}$$

Therefore, with a large enough global constant C, we have $\mathbf{E}_h(s, a^*) \leq 2b_h(s, a^*) \leq \epsilon_h(s, a^*)$ holds for any time-state pair. Together with Theorem A.1,

$$\mathbf{V}^* - \mathbf{V}^{\underline{\pi}} \leq 2\mathbb{E}_{\pi^*}\left[\ddot{\mathbf{E}}_h(s, a)\right] = 0.$$

B.4.3 Tools for the Proof of Lemma B.5

We first introduce a modified version of Lemma from Li et al. [2022]. Note that the proof of this lemma didn't involve any assumption about the data coverage, and is a pure mathmatical analysis. So the original proof is valid, and to avoid redundancy, we won't prove this lemma again.

Lemma B.6. *$\forall h \in [H]$, and any vector $V \in \mathbb{R}^S$ independent of $\hat{P}_h$ obeying $\|V\|_\infty \leq H$. With probability at least $1-\delta$, one has*

$$\mathbf{Var}_{\hat{P}_{h,s,a}}(V) \leq 2\mathbf{Var}_{P_{h,s,a}}(V) + \frac{5H^2\iota}{3N'_h(s,a)}$$

simultaneously for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ obeying $N_h(s,a) > 0$

Modification lies in that we use $N'_h(s,a)$ to replace $N_h(s,a)$ in original version, for when $N_h(s,a) \leq \iota$, the inequalities hold trivially. Also we introduce the lemma needed to limit the overall variance. This lemma differs from Li et al. [2022]'s work from the definition of $d_h^{*'}$ to support our theorem.

Lemma B.7 (weighted variance sum)**.** *$\forall (h', s') \in [H] \times \mathcal{S}$, if the event ξ_{conc} happens, we have*

$$\sum_{h=h'}^{H}\sum_{s} d^*_{h\sim(h',s')}(s)\mathbf{Var}_{P_{h,s,a}}(\underline{\mathbf{V}}^{\pi}_{h+1}) \leq 4H\sum_{h=h'}^{H}\sum_{s} d^*_{h\sim(h',s')}(s)b_h(s,a^*) + 2H^2.$$

Proof. Here we use $P_h^* \in \mathbb{R}^{S\times S}$ to denote the transition kernel of optimal policy, where $P^*_{h,(m,n)}$ is the probability of transfer from s_m to s_n at step h while applying the optimal policy π^*. $A \circ B$ refers to the Hadamard product of A and B.

$$\begin{aligned}
&\sum_{h=h'}^{H}\sum_{s} d_h^{*'}(s)\mathbf{Var}_{P_{h,s,a^*}}(\underline{\mathbf{V}}^{\pi}_{h+1}) = \sum_{h=h'}^{H} d_h^{*'\top}(P_h^*\underline{\mathbf{V}}^{\pi}_{h+1}\circ\underline{\mathbf{V}}^{\pi}_{h+1} - (P_h^*\underline{\mathbf{V}}^{\pi}_{h+1})\circ(P_h^*\underline{\mathbf{V}}^{\pi}_{h+1}))\\
&= \sum_{h=h'}^{H} d_h^{*'\top}(P_h^*\underline{\mathbf{V}}^{\pi}_{h+1}\circ\underline{\mathbf{V}}^{\pi}_{h+1} - \underline{\mathbf{V}}^{\pi}_{h}\circ\underline{\mathbf{V}}^{\pi}_{h} + \underline{\mathbf{V}}^{\pi}_{h}\circ\underline{\mathbf{V}}^{\pi}_{h} - (P_h^*\underline{\mathbf{V}}^{\pi}_{h+1})\circ(P_h^*\underline{\mathbf{V}}^{\pi}_{h+1}))\\
&= \sum_{h=h'}^{H}\left(d_{h+1}^{*'\top}\underline{\mathbf{V}}^{\pi}_{h+1}\circ\underline{\mathbf{V}}^{\pi}_{h+1} - d_h^{*'\top}\underline{\mathbf{V}}^{\pi}_{h}\circ\underline{\mathbf{V}}^{\pi}_{h}\right) + \sum_{h=h'}^{H} d_h^{*'\top}(\underline{\mathbf{V}}^{\pi}_{h}\circ\underline{\mathbf{V}}^{\pi}_{h} - (P_h^*\underline{\mathbf{V}}^{\pi}_{h+1})\circ(P_h^*\underline{\mathbf{V}}^{\pi}_{h+1}))\\
&= 0 - d_{h'}^{*'\top}\underline{\mathbf{V}}^{\pi}_{h'}\circ\underline{\mathbf{V}}^{\pi}_{h'} + \sum_{h=h'}^{H} d_h^{*'\top}(\underline{\mathbf{V}}^{\pi}_{h} - P_h^*\underline{\mathbf{V}}^{\pi}_{h+1})\circ(\underline{\mathbf{V}}^{\pi}_{h} + P_h^*\underline{\mathbf{V}}^{\pi}_{h+1})\\
&\leq \sum_{h=h'}^{H} d_h^{*'\top}(\underline{\mathbf{V}}^{\pi}_{h} - P_h^*\underline{\mathbf{V}}^{\pi}_{h+1})\circ(\underline{\mathbf{V}}^{\pi}_{h} + P_h^*\underline{\mathbf{V}}^{\pi}_{h+1}). \qquad (11)
\end{aligned}$$

The above induction mainly uses the equality that $d_h^{*'}P_h^* = d_{h+1}^{*'}$ and non-negativity of $d_h^{*'}$. Because the concentration events ξ_{conc} guarantees that $b_h(s,a) \geq |(\hat{p}_h(s,a) - p_h(s,a))^\top\underline{\mathbf{V}}_{h+1}|$,

$$\begin{aligned}
&\underline{\mathbf{V}}^{\pi}_{h}(s) - p_h(s,a^*)^\top\underline{\mathbf{V}}^{\pi}_{h+1}\\
&= \underline{\mathbf{V}}^{\pi}_{h}(s) - \underline{\mathbf{Q}}^{\pi}_{h}(s,a^*) + r_h(s,a^*) - b_h(s,a^*) + (\hat{p}_h(s,a^*) - p_h(s,a^*))^\top\underline{\mathbf{V}}^{\pi}_{h+1}\\
&\geq 0 + 0 - b_h(s,a^*) - b_h(s,a^*) = -2b_h(s,a^*).
\end{aligned}$$

Then we can continue from (11) to get

$$
\begin{aligned}
&\sum_{h=h'}^{H}\sum_{s} d_h^{*'}(s)\mathbf{Var}_{P_{h,s,a}}(\underline{\mathbf{V}}_{h+1}^{\pi}) \leq \sum_{h,s} d_h^{*'\top}(\underline{\mathbf{V}}_{h}^{\pi} - P_h^*\underline{\mathbf{V}}_{h+1}^{\pi}) \circ (\underline{\mathbf{V}}_{h}^{\pi} + P_h^*\underline{\mathbf{V}}_{h+1}^{\pi}) \\
&\leq \sum_{h=h'}^{H} d_h^{*'\top}(\underline{\mathbf{V}}_{h}^{\pi} - P_h^*\underline{\mathbf{V}}_{h+1}^{\pi} + 2b_h(s,a^*)\mathbf{1}) \circ (\underline{\mathbf{V}}_{h}^{\pi} + P_h^*\underline{\mathbf{V}}_{h+1}^{\pi}) \\
&\leq 2H\sum_{h=h'}^{H} d_h^{*'\top}(\underline{\mathbf{V}}_{h}^{\pi} - P_h^*\underline{\mathbf{V}}_{h+1}^{\pi} + 2b_h(s,a^*)\mathbf{1}) \\
&= 2H(d_{h'}^{*'\top}\underline{\mathbf{V}}_{h'}^{\pi} - d_{H+1}^{*'\top}\underline{\mathbf{V}}_{H+1}^{\pi}) + 4H\sum_{h'=h}^{H}\sum_{s} d_{h\sim(h',s')}^{*}(s)b_h(s,a^*) \\
&\leq 2H^2 + 4H\sum_{h'=h}^{H}\sum_{s} d_{h\sim(h',s')}^{*}(s)b_h(s,a^*).
\end{aligned}
$$

□

B.4.4 Proof of Lemma B.5

Proof. This proof is similar to the one in Li et al. [2022]. The difference lies in that we generalize the conclusion to any part decomposition, while the original version only cares about the optimal policy distribution.
First, it follows from Lemma B.6 and inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ that

$$
\begin{aligned}
\frac{1}{C_b}b_h(s,a) &= \sqrt{\frac{\mathbf{Var}_{\hat{P}_{h,s,a}}(\underline{\mathbf{V}}_{h+1}^{\pi})\iota}{N_h'(s,a)}} + \frac{H\iota}{N_h'(s,a)} \\
&\leq \sqrt{\frac{2\mathbf{Var}_{P_{h,s,a}}(\underline{\mathbf{V}}_{h+1}^{\pi})\iota + \frac{5H^2\iota}{3N_h'(s,a)}\iota}{N_h'(s,a)}} + \frac{H\iota}{N_h'(s,a)} \\
&\leq \sqrt{\frac{2\mathbf{Var}_{P_{h,s,a}}(\underline{\mathbf{V}}_{h+1}^{\pi})\iota}{N_h'(s,a)}} + (1+\sqrt{\frac{5}{3}})\frac{H\iota}{N_h'(s,a)} \\
&\leq 2\sqrt{\frac{\mathbf{Var}_{P_{h,s,a}}(\underline{\mathbf{V}}_{h+1}^{\pi})\iota}{N_h'(s,a)}} + \frac{3H\iota}{N_h'(s,a)}. \qquad (12)
\end{aligned}
$$

Note that the concentration event ξ_{conc} guarantees that $N_h(s,a^*) \geq C_d N d_h^{\mu}(s,a^*) \geq C_d NP$. Then we can use Cauchy-Schwarz Inequality to limit the variance term,

$$
\begin{aligned}
&\sum_{h,s} d_h^{*'}(s)\sqrt{\frac{\mathbf{Var}_{P_{h,s,a^*}}(\underline{\mathbf{V}}_{h+1}^{\pi})\iota}{N_h'(s,a^*)}} \\
&\leq \sqrt{\frac{\iota}{C_d NP}}\sum_{h,s} d_h^{*'}(s)\sqrt{\mathbf{Var}_{P_{h,s,a^*}}(\underline{\mathbf{V}}_{h+1}^{\pi})} && (\xi_{\text{conc}}) \\
&\leq \sqrt{\frac{\iota}{C_d NP}}\sqrt{\sum_{h,s} d_h^{*'}(s)}\sqrt{\sum_{h,s} d_h^{*'}(s)\mathbf{Var}_{P_{h,s,a^*}}(\underline{\mathbf{V}}_{h+1}^{\pi})} && \text{(Cauchy-Schwarz's Inequality)} \\
&\leq \sqrt{\frac{H\iota}{C_d NP}}\sqrt{4H\sum_{h,s} d_h^{*'}(s)b_h(s,a^*) + 2H^2} && \text{(Lemma B.7)} \\
&\leq \sqrt{\frac{4H^2\iota}{C_d NP}\sum_{h,s} d_h^{*'}(s)b_h(s,a^*)} + \sqrt{\frac{2H^3\iota}{C_d NP}} && (\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}) \\
&\leq \frac{4C_b H^2\iota}{C_d NP} + \frac{1}{2C_b}\sum_{h,s} d_h^{*'}(s)b_h(s,a^*) + \sqrt{\frac{2H^3\iota}{C_d NP}}. && (\sqrt{2ab} \leq a+b)
\end{aligned}
$$

At the same time, we can limit the sum of $\frac{H\iota}{N_h'(s,a^*)}$,

$$
\sum_{h,s} d_h^{*'}(s)\frac{H\iota}{N_h'(s,a^*)} \leq \sum_{h,s} d_h^{*'}(s)\frac{H\iota}{C_d NP} = \frac{H^2\iota}{C_d NP}.
$$

By connecting these inequalities to (12), we get

$$
\begin{aligned}
&\sum_{h,s} d_h^{*'}(s)b_h(s,a^*) \leq C_b\sum_{h,s} 2d_h^{*'}(s)\sqrt{\frac{\mathbf{Var}_{P_{h,s,a}}(\underline{\mathbf{V}}_{h+1}^{\pi})\iota}{N_h'(s,a^*)}} + C_b\sum_{h,s} d_h^{*'}(s)\frac{3H\iota}{N_h'(s,a^*)} \\
&\leq \frac{1}{2}\sum_{h,s} d_h^{*'}(s)b_h(s,a^*) + (8C_b^2+6C_b)\frac{H^2\iota}{C_d NP} + 2C_b\sqrt{\frac{H^3\iota}{C_d NP}}.
\end{aligned}
$$

Rearranging the terms, we finish the proof of Lemma B.5. □

B.5 Proof of Upper bound with both assumptions (Proof of Theorem 5.2)

When we have access to both P and C^*, we can derive the bound

$$
N = O\left(\frac{H^3 SC^*\iota}{\epsilon \text{gap}_{\min}} + \frac{H\iota}{P}\right).
$$

To prove this, we need a specially designed $\epsilon_h(s)$ in Theorem A.1. By setting $\epsilon_h(s) = C_{pac}(\frac{\mathbf{Var}_{\hat{P}_{h,s,a^*}}(\underline{\mathbf{V}}_{h+1}^{\pi})}{H^2} + \frac{1}{H})\text{gap}_{\min}$, we will first prove that ξ_{gap} happenes, and then calculate the suboptimality gap.

B.5.1 Tools for the Proof of Theorem 5.2

Lemma B.8. *If we set* $\epsilon_h(s,a) = C_{pac}\left(\frac{\mathbf{Var}_{\hat{P}_{h,s,a}}(\underline{\mathbf{V}}_{h+1}^{\pi})}{H^2} + \frac{1}{H}\right)\text{gap}_{\min}$ *for some small enough constant* C_{pac}*, and* $N \geq C_3\frac{H\iota}{P}$ *for some constant* C_3*,* ξ_{conc} *indicates* ξ_{gap}*,*

$$
\forall (h,s) \in [H]\times\mathcal{S} \qquad \ddot{\mathbf{V}}_h^*(s) \leq \underline{\mathbf{V}}_h^*(s) + \frac{\text{gap}_{\min}}{2}.
$$

Proof. We have proved in the proof of Corollary A.1 that

$$\sum_{h,s} d_h^{*'}(s)\frac{\text{gap}_{\min}}{4H} \le \frac{\text{gap}_{\min}}{4},$$

where $d_h^{*'} = d^*_{h\sim(h',s')}$, which is the state distribution of time step h under π^* conditioned on having reached (h', s') before. So it remains to show that

$$\sum_{h,s} d_h^{*'}(s)\frac{\mathbf{Var}_{\hat{P}}(\underline{\mathbf{V}}^{\pi_{h+1}})}{H^2}\text{gap}_{\min} \lesssim \text{gap}_{\min}. \tag{13}$$

This follows from a similar analysis with the proof of uniform optimal policy coverage assumption case.

$$\begin{aligned}
&\sum_{h,s} d_h^{*'}(s)\mathbf{Var}_{\hat{P}}(\underline{\mathbf{V}}^{\pi}_{h+1}) \\
&\le \sum_{h,s} d_h^{*'}(s)\left(\mathbf{Var}_P(\underline{\mathbf{V}}^{\pi}_{h+1}) + \frac{5H^2\iota}{3N'_h(s,a^*)}\right) && \text{(Lemma B.6)}\\
&\lesssim \sum_{h,s} d_h^{*'}(s)\mathbf{Var}_P(\underline{\mathbf{V}}^{\pi}_{h+1}) + H^2 && (N'_h(s,a^*) \ge C_d NP \gtrsim H\iota)\\
&\le 4H\sum_{h,s} d_h^{*'}(s) b_h(s,a^*) + 3H^2 && \text{(Lemma B.7)}\\
&\lesssim 4H\left(\sqrt{\frac{H^3\iota}{NP}} + \frac{H^2\iota}{NP}\right) + 3H^2 && \text{(Lemma B.5)}\\
&\lesssim H^2. && (NP \gtrsim H\iota)
\end{aligned}$$

Then we can finish the proof

$$\begin{aligned}
\ddot{\mathbf{V}}^*_h(s) - \underline{\mathbf{V}}^*_h(s) &= \sum_{h,s} d_h^{*'}(s)(\ddot{\mathbf{E}}_h(s,a^*) - \mathbf{E}_h(s,a^*))\\
&\le \sum_{h,s} d_h^{*'}(s)\epsilon_h(s)\\
&\lesssim C_{cap}\sum_{h,s} d_h^{*'}\left(\frac{\mathbf{Var}_{\hat{P}}(\underline{\mathbf{V}}^{\pi}_{h+1})}{H^2} + \frac{1}{H}\right)\text{gap}_{\min}\\
&\lesssim C_{cap}\text{gap}_{\min}.
\end{aligned}$$

We can let C_{cap} be small enough to limit the difference between $\ddot{\mathbf{V}}^*_h(s)$ and $\underline{\mathbf{V}}^*_h(s)$ within $\frac{\text{gap}_{\min}}{2}$. □

Lemma B.9.

$$\ddot{\mathbf{E}}_h(s,a^*) \le 4C_b\sqrt{\frac{\mathbf{Var}_{\hat{P}}(\underline{\mathbf{V}}^{\pi}_{h+1})\iota}{N'_h(s,a^*)}}\left(\frac{b_h(s,a^*)}{\epsilon_h(s)}\right) + 2C_b\frac{H\iota}{N_h(s,a^*)}.$$

Proof. When $\mathbf{E}_h(s,a^*) < \epsilon_h(s)$, $\ddot{\mathbf{E}}_h(s,a^*) = 0$.
When $\mathbf{E}_h(s,a^*) \ge \epsilon_h(s)$,

$$\begin{aligned}
&2b_h(s,a^*) \ge \mathbf{E}_h(s,a^*) \ge \epsilon_h(s).\\
\Rightarrow \ddot{\mathbf{E}}_h(s,a^*) &\le \mathbf{E}_h(s,a^*)\\
&\le 2b_h(s,a^*)\\
&= 2C_b\sqrt{\frac{\mathbf{Var}_{\hat{P}}(\underline{\mathbf{V}}^{\pi}_{h+1})\iota}{N'_h(s,a^*)}} + 2C_b\frac{H\iota}{N'_h(s,a^*)}\\
&\le 4C_b\sqrt{\frac{\mathbf{Var}_{\hat{P}}(\underline{\mathbf{V}}^{\pi}_{h+1})\iota}{N'_h(s,a^*)}}\left(\frac{b_h(s,a^*)}{\epsilon_h(s)}\right) + 2C_b\frac{H\iota}{N'_h(s,a^*)}.
\end{aligned}$$

□

B.5.2 Main Proof

Proof. We treat the first term in the RHS of Lemma B.9. With basic inequality $a+b \geq 2\sqrt{ab}$, we can first lowe bound $\epsilon_h(s,a)$,

$$\epsilon_h(s,a) = C_{pac}\left(\frac{\mathbf{Var}_{\hat{P}_{h,s,a}}(\underline{\mathbf{V}}^{\pi}_{h+1})}{H^2} + \frac{1}{H}\right)\text{gap}_{\min} \gtrsim \max\left\{\frac{\mathbf{Var}_{\hat{P}_{h,s,a}}(\underline{\mathbf{V}}^{\pi}_{h+1})}{H^2}, \sqrt{\frac{\mathbf{Var}_{\hat{P}_{h,s,a}}(\underline{\mathbf{V}}^{\pi}_{h+1})}{H^3}}\right\}\text{gap}_{\min}. \tag{14}$$

Therefore we have,

$$\begin{aligned}
&\sqrt{\frac{\mathbf{Var}_{\hat{P}}(\underline{\mathbf{V}}^{\pi}_{h+1})\iota}{N'_h(s,a^*)}}\left(\frac{b_h(s,a^*)}{\epsilon_h(s,a^*)}\right) \\
&\lesssim \sqrt{\frac{\mathbf{Var}_{\hat{P}}(\underline{\mathbf{V}}^{\pi}_{h+1})\iota}{N'_h(s,a^*)}}\left(\frac{\sqrt{\frac{\mathbf{Var}_{\hat{P}}(\underline{\mathbf{V}}^{\pi}_{h+1})\iota}{N'_h(s,a^*)}}}{\frac{\mathbf{Var}_{\hat{P}}(\underline{\mathbf{V}}^{\pi}_{h+1})}{H^2}} + \frac{\frac{H\iota}{N_h(s,a)}}{\sqrt{\frac{\mathbf{Var}_{\hat{P}}(\underline{\mathbf{V}}^{\pi}_{h+1})}{H^3}}}\right)\frac{1}{\text{gap}_{\min}} \\
&= \frac{H^2\iota}{N'_h(s,a^*)\text{gap}_{\min}} + \frac{H^{5/2}\iota^{3/2}}{N_h^{'3/2}(s,a^*)\text{gap}_{\min}} \\
&\lesssim \frac{2H^2\iota}{N'_h(s,a^*)\text{gap}_{\min}}.
\end{aligned}$$

The first inequality is gained by expanding $b_h(s,a^*)$ and $\epsilon_h(s,a)$. The second inequality results from the inequality that $N'_h(s,a^*) \geq C_d NP \gtrsim H\iota$. Therefore, we can further write Lemma B.9 as

$$\ddot{\mathbf{E}}_h(s,a^*) \lesssim \frac{H^2\iota}{N'_h(s,a^*)\text{gap}_{\min}}.$$

Then with Lemma B.8, we have event ξ_{gap} hold. Then Theorem A.1 further indicates that for some deterministic optimal policy π^*,

$$\begin{aligned}
\mathbf{V}^*_0 - \mathbf{V}^{\pi}_0 &\lesssim \mathbf{V}^*_0 - \ddot{\mathbf{V}}^*_0 \\
&= \sum_{h,s} d^*_h(s)\ddot{\mathbf{E}}_h(s,a^*) \\
&\lesssim \sum_{h,s} \frac{d^*_h(s)H^2\iota}{N'_h(s,a^*)\text{gap}_{\min}} \\
&\lesssim \sum_{h,s} \frac{H^2C^*\iota}{N\text{gap}_{\min}} \\
&= \frac{H^3C^*S\iota}{N\text{gap}_{\min}} \lesssim \epsilon.
\end{aligned}$$

□

C Gap-dependent Lower Bounds

We begin by restating the formal version of lower bounds.

Definition C.1 (offline learning algorithm)**.** *For an algorithm* **ALG***, we call it an offline learning algorithm if*

1. **ALG** *takes a dataset $\mathcal{D}$ and optionally a reward function R as input,*
2. **ALG** *output a valid policy π.*

Notice that **ALG** can be stochastic.

C.1 Main Results

Theorem C.1. *There exists constant C_{lb}, s.t. for any $A \geq 3, S \geq 2, H \geq 2, \tau < \frac{1}{2}, \lambda < \frac{1}{3}, \lambda_1 \geq 2$ and algorithm* $\mathbf{ALG}$*, if the number of sample trajectories*

$$N \leq C_{lb} \cdot \frac{HS\lambda_1}{\lambda\tau^2},$$

there exists some MDP $\mathcal{M}$ and behavior policy μ with $\mathrm{gap}_{\min} = \tau$, $P \geq \frac{\lambda}{eS\lambda_1}$, $C^* \leq \lambda_1$ such that the output policy $\hat{\pi}$ suffers from a expected suboptimality

$$\mathbb{E}_{\mathcal{M},\mu,\mathbf{ALG}}[\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}}] \geq \frac{\lambda H\tau}{12}.$$

Corollary C.1 (lower bound for uniform optimal policy coverage)**.** *Given $A \geq 3, S \geq 2, H \geq 3, P \in (0, \frac{1}{6S}], \epsilon < 1/12, \mathrm{gap}_{\min} \in [\frac{24\epsilon}{H}, \frac{1}{2}]$ and any offline learning algorithm* $\mathbf{ALG}$ *returning a policy $\hat{\pi}$, there exists a constant C_1 such that if the number of offline sample trajectories*

$$N \leq C_1 \cdot \frac{H}{P\mathrm{gap}_{\min}^2},$$

then there exists a MDP instance M and behavior policy μ such that the output policy $\hat{\pi}$ suffers from expected ϵ-suboptimality

$$\mathbb{E}_{\mathcal{M},\mu,\mathbf{ALG}}[\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}}] \geq \epsilon.$$

Proof. Let $\lambda = 1/3$, $\lambda_1 = \frac{1}{3PS}$ and $\tau = \mathrm{gap}_{\min}$ in Theorem C.1, we get the proposition. □

Corollary C.2 (lower bound for relative optimal policy coverage)**.** *Given $A \geq 3, S \geq 2, H \geq 2, C^* \geq 2, \epsilon < 1/12, \mathrm{gap}_{\min} \in [\frac{24\epsilon}{H}, \frac{1}{2}]$ and any offline learning algorithm* $\mathbf{ALG}$ *returning a policy $\hat{\pi}$, there exists a constant C_2 such that if the number of offline sample trajectories*

$$N \leq C_2 \cdot \frac{H^2SC^*}{\mathrm{gap}_{\min}\epsilon},$$

then there exists a MDP instance $\mathcal{M}$ and behavior policy μ such that the output policy $\hat{\pi}$ suffers from expected ϵ-suboptimality

$$\mathbb{E}_{\mathcal{M},\mu,\mathbf{ALG}}[\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}}] \geq \epsilon.$$

Proof. Let $\lambda = \frac{12\epsilon}{H\mathrm{gap}_{\min}}$, $\lambda_1 = C^*$ and $\tau = \mathrm{gap}_{\min}$ in Theorem C.1, we get the conclusion. □

C.2 Proof of Theorem C.1

C.2.1 Construction of the MDP Family and Behavior Policy

We construct a MDP family and calculate the average minimum suboptimality.
First, we construct the prototype MDP $\mathcal{M}_0$ with $S + 2$ states, horizon of $2H + 1$ and A actions. There are 3 kind of states

1. good state s_g. An absorbing state. Reaching this state means a total reward of H.
2. bad state s_b. An absorbing state. Reaching this state means a total reward of 0.
3. true states $s_1, s_2, \cdots, s_S$. Actions chosen in these states determine the probability being transfered to s_g and s_b.

The initial state distribution $p_0(s)$ is

$$p_0(s) = \begin{cases} \frac{\lambda}{S} & s \in \{s_1, s_2, \cdots, s_S\}, \\ \frac{1-\lambda}{2} & s = s_b, \\ \frac{1-\lambda}{2} & s = s_g. \end{cases}$$

For any $\lambda \in [0, \frac{1}{3}]$. The avaliable action set is $\{a_i\}_{i=1}^A$. The only non-zero rewards in this MDP are $r_h(s_g, a) = 1$ for $h \geq H+2$ and any a. The transition probability of $\mathcal{M}_0$ in the first $H+1$ steps is,

$$
\begin{aligned}
p_h(s_i, a_j, s_i) &= 1 - \frac{1}{H} && \forall (h, i, j) \in [H] \times [S] \times [A], \\
p_h(s_i, a_j, s_g) &= p_h(s_i, a_j, s_b) = \frac{1}{2H} && \forall (h, i, j) \in [H] \times [S] \times [A], \\
p_{H+1}(s_i, a_j, s_g) &= p_{H+1}(s_i, a_j, s_b) = \frac{1}{2} && \forall (i, j) \in [S] \times [A].
\end{aligned}
$$

For all the other (h, s, a) tuples not mentioned, $p_h(s, a, s) = 1$, $p_h(s, a, s') = 0$, where s' is any state other than s.

Then we construct the MDP family M on the basis of $\mathcal{M}_0$. For each matrix $\phi \in [1, 2]^{H \times S}$, we define $\mathcal{M}_\phi$ to be the MDP almost the same as $\mathcal{M}_0$ except for that

$$
\begin{aligned}
p_h(s_i, a_{\phi_{h,i}}, s_g) &= \frac{1}{2H}(1 + 2\tau), \\
p_h(s_i, a_{\phi_{h,i}}, s_b) &= \frac{1}{2H}(1 - 2\tau).
\end{aligned}
$$

In other words, we make the action $a_{\Phi_{h,i}}$ the unique optimal action by lifting it's expected reward by τ. The behavior policy μ chooses a_1, a_2 with probability $1/\lambda_1$ respectively and choose a_3 with probability $1 - 2/\lambda_1$ at $\{s_i \mid i = 1, 2, \cdots S\}$, and always choose a_1 at s_g and s_b. We will prove the following lemmas in Section C.2.4

Lemma C.1. *For any MDP constructed above and μ, we have both assumptions hold with*

$$
C^* = \lambda_1, P \geq \frac{\lambda}{eS\lambda_1}, \mathrm{gap}_{\min} = \tau.
$$

Lemma C.2. *For a given algorithm* $\mathbf{ALG}$*, define the expectation of mistakes made by $\hat{\pi}$ at step h, state s_i over the uniform distribution ν of ϕ to be*

$$
l_{h,i}(\mathbf{ALG}) = \mathbb{E}_{\mathcal{D} \sim (\mathcal{M}_\phi, \mu), \mathbf{ALG}}[\mathbb{I}\{\hat{\pi}_h(s_i) \neq a_{\phi_{h,i}}\}].
$$

Then expected suboptimality with respect to the randomness of $\mathcal{M}_\phi$ and μ can be lower bounded by

$$
\mathbb{E}_{\mathcal{M}_\phi, \mu, \mathbf{ALG}}[\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}}] \geq \frac{\lambda}{eS}\tau \sum_{(h,i) \in [H] \times [S]} l_{h,i}(\mathbf{ALG}).
$$

Lemma C.3. *For any MDP constructed above, we have,*

$$
\max_\pi \mathbf{V}_0^* - \mathbf{V}_0^\pi \leq \lambda H \tau.
$$

C.2.2 Main Proof

Proof. To avoid making the proof prolix, we strengthen $\mathbf{ALG}$ by letting $\mathbf{ALG}$ know that the only thing influencing the value function of a state is the probabilities of transferring to s_g and s_b, and the total reward after getting to s_g is exactly H, which assumption is conventionally made for the lower bound proofs in MDPs. In this setting, any reasonable algorithm $\mathbf{ALG}$ would only consider the visitation counts $N_{h,i} = \{N_h(s_i, a, s') \mid a \in \mathcal{A}, s' \in \mathcal{S}\}$ at step h when determining the value of $\hat{\pi}_h(s_i)$.

Then we can rewrite $\bar{l}_{h,i}(\mathbf{ALG}) \triangleq \mathbb{E}_{\phi \sim \nu}[l_{h,i}(\mathbf{ALG})]$,

$$
\begin{aligned}
\bar{l}_{h,i}(\mathbf{ALG}) &= \mathbb{E}_{\phi \sim \nu} \mathbb{E}_{\mathcal{D} \sim (\mathcal{M}_\phi, \mu), \mathbf{ALG}}[\mathbb{I}\{\hat{\pi}_h(s_i) \neq a_{\phi_{h,i}}\}] \\
&= \mathbb{E}_{\phi \sim \nu} \mathbb{E}_{N_{h,i} \sim (\mathcal{M}_\phi, \mu), \mathbf{ALG}}[\mathbb{I}\{\hat{\pi}_h(s_i) \neq a_{\phi_{h,i}}\}].
\end{aligned}
$$

Because the KL divergence between the transition kernel $P_{h,s_i,a}^{\mathcal{M}_0}$ and $P_{h,s_i,a_{\phi_{h,i}}}^{\mathcal{M}_\phi}$ satisfies

$$
\begin{aligned}
&KL\left(\left(1 - \frac{1}{H}, \frac{1}{2H}, \frac{1}{2H}\right) \middle\| \left(1 - \frac{1}{H}, \frac{1}{H}(\frac{1}{2} + 2\tau), \frac{1}{H}(\frac{1}{2} - 2\tau)\right)\right) \\
&= \frac{1}{2H} \log \frac{1}{1 - 4\tau^2} \leq \frac{4\tau^2}{H}.
\end{aligned} \tag{15}
$$

We have

$$
\begin{aligned}
\bar{l}_{h,i}(\mathbf{ALG}) &= \mathbb{E}_{\phi\sim\nu}\mathbb{E}_{N_{h,i}\sim(\mathcal{M}_\phi,\mu),\mathbf{ALG}}[\mathbb{I}\{\hat{\pi}_h(s_i)\neq a_{\phi_{h,i}}\}] \\
&\geq \mathbb{E}_{\phi\sim\nu}\mathbb{E}_{N_{h,i}\sim(\mathcal{M}_0,\mu),\mathbf{ALG}}[\mathbb{I}\{\hat{\pi}_h(s_i)\neq a_{\phi_{h,i}}\}] \\
&\quad - \mathbb{E}_{\phi\sim\nu}[\mathrm{TV}(N_{h,i}\mid_{\mathcal{M}_0,\mu}, N_{h,i}\mid_{\mathcal{M}_\phi,\mu})] \\
&\geq \mathbb{E}_{N_{h,i}\sim(\mathcal{M}_0,\mu),\mathbf{ALG}}\mathbb{E}_{\phi\sim\nu}[\mathbb{I}\{\hat{\pi}_h(s_i)\neq a_{\phi_{h,i}}\}] \\
&\quad - \mathbb{E}_{\phi\sim\nu}\sqrt{\frac{1}{2}\mathrm{KL}(N_{h,i}\mid_{\mathcal{M}_0,\mu}\|N_{h,i}\mid_{\mathcal{M}_\phi,\mu})} && \text{(Pinsker's inequality)} \\
&\geq \mathbb{E}_{N_{h,i}\sim(\mathcal{M}_0,\mu)}[\frac{1}{2}] && (a_1 \text{ and } a_2 \text{ can not be distinguished in } \mathcal{M}_0) \\
&\quad - \mathbb{E}_{\phi\sim\nu}\sqrt{\frac{1}{2}\sum_{a\in\mathcal{A}}\mathbb{E}_{\mathcal{M}_0,\mu}[N_h(s_i,a)]\mathrm{KL}(P^{\mathcal{M}_0}_{h,s_i,a}\|P^{\mathcal{M}_\phi}_{h,s_i,a})} && \text{(KL decomposition)} \\
&= \frac{1}{2} - \mathbb{E}_{\phi\sim\nu}\sqrt{\frac{1}{2}\mathbb{E}_{\mathcal{M}_0,\mu}[N_h(s_i,a)]\mathrm{KL}(P^{\mathcal{M}_0}_{h,s_i,a_{\phi_{h,i}}}\|P^{\mathcal{M}_\phi}_{h,s_i,a_{\phi_{h,i}}})} \\
&\geq \frac{1}{2} - \mathbb{E}_{\phi\sim\nu}\sqrt{\mathbb{E}_{\mathcal{M}_0,\mu}[N_h(s_i,a_{\phi_{h,i}})]\frac{2\tau^2}{H}} && \text{(statement (15))} \\
&\geq \frac{1}{2} - \sqrt{\frac{1}{2}\sum_{a=a_1,a_2}\mathbb{E}_{\mathcal{M}_0,\mu}[N_h(s_i,a)]\frac{2\tau^2}{H}} && \text{(Jensen's inequality)} \\
&= \frac{1}{2} - \sqrt{\mathbb{E}_{\mathcal{M}_0,\mu}[N_h(s_i)]\frac{2\tau^2}{H\lambda_1}}. \\
&\qquad (\tfrac{\sum_{a=a_1,a_2}\mathbb{E}_{\mathcal{M}_0,\mu}[N_h(s_i,a)]}{\mathbb{E}_{\mathcal{M}_0,\mu}[N_h(s_i)]} = \mu_h(a_1\mid s_i)+\mu_h(a_2\mid s_i) = \tfrac{2}{\lambda_1})
\end{aligned}
$$

This further indicates that the expectation of overall mistakes can be lower bounded by

$$
\begin{aligned}
\sum_{(h,i)\in[H]\times[S]}\bar{l}_{h,i}(\mathbf{ALG}) &\geq \sum_{h,i}\left(\frac{1}{2} - \sqrt{\mathbb{E}_{\mathcal{M}_0,\mu}[N_h(s_i)]\frac{2\tau^2}{H\lambda_1}}\right) \\
&\geq \frac{HS}{2} - \sqrt{HS}\sqrt{\sum_{h,i}\mathbb{E}_{\mathcal{M}_0,\mu}[N_h(s_i)]\frac{2\tau^2}{H\lambda_1}} \\
&\qquad \text{(Cauchy Schwarz's Inequality)} \\
&= HS(\frac{1}{2} - \sqrt{\mathbb{E}_{\mathcal{M}_0,\mu}[\sum_{h,i}N_h(s_i)]\frac{2\tau^2}{H^2S\lambda_1}}). \qquad (16)
\end{aligned}
$$

Because state s_i can be reached at step h only when the initial state is s_i,

$$
\begin{aligned}
\mathbb{E}_{\mathcal{M}_0,\mu}[\sum_{(h,i)\in[H]\times[S]}N_h(s_i)] &= N\sum_{h,i}\frac{\lambda}{S}(1-\frac{1}{H})^{h-1} \\
&\leq N\sum_{h,i}\frac{\lambda}{S} = NH\lambda.
\end{aligned}
$$

Therefore, continue from inequality (16),

$$
\sum_{(h,i)\in[H]\times[S]}\bar{l}_{h,i}(\mathbf{ALG}) \geq HS\left(\frac{1}{2} - \sqrt{N\frac{2\lambda\tau^2}{HS\lambda_1}}\right),
$$

Now we can lower bound the suboptimality of $\hat{\pi}$ with Lemma C.2 by

$$
\mathbb{E}_{\phi\sim\nu}\mathbb{E}_{\mathcal{M}_\phi,\mu,\mathbf{ALG}}[\mathbf{V}^*_0 - \mathbf{V}^{\hat{\pi}}_0] \geq \mathbb{E}_{\phi\sim\nu}[\frac{\lambda\tau}{eS}\sum_{h,i}l_{h,i}(\mathbf{ALG})] \geq \frac{\lambda H\tau}{e}\left(\frac{1}{2} - \sqrt{N\frac{2\lambda\tau^2}{HS\lambda_1}}\right).
$$

Then we reach the conclusion that when

$$N \leq \frac{HS\lambda_1}{32\lambda\tau^2},$$

the average suboptimality of $\hat{\pi}$ must be large

$$\begin{aligned}
&\mathbb{E}_{\phi\sim\nu}\mathbb{E}_{\mathcal{M}_\phi,\mu,\mathbf{ALG}}[\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}}] \geq \frac{\lambda H\tau}{e}\cdot(\frac{1}{2}-\frac{1}{4}),\\
\Rightarrow &\exists\phi, s.t. \mathbb{E}_{\mathcal{M}_\phi,\mu,\mathbf{ALG}}[\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}}] \geq \frac{\lambda H\tau}{4e}.
\end{aligned}$$

□

C.2.3 Constant probability version of main theorem

Theorem C.1 is stated in the form of expectation, which is not directly consist with upper bound. Here we restate it in the language of probability,

Theorem C.2. *There exists constant C_{lb}, s.t. for any $A \geq 3, S \geq 2, H \geq 2, \tau < \frac{1}{2}, \lambda < \frac{1}{3}, \lambda_1 \geq 2$ and algorithm $\mathbf{ALG}$, if the number of sample trajectories*

$$N \leq C_{lb}\cdot\frac{HS\lambda_1}{\lambda\tau^2},$$

there exists some MDP $\mathcal{M}$ and behavior policy μ with $\mathrm{gap}_{\min} = \tau$, $P \geq \frac{\lambda}{eS\lambda_1}$, $C^* \leq \lambda_1$ such that the output policy $\hat{\pi}$ suffers from a expected suboptimality

$$\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}} \geq \frac{\lambda H\tau}{24},$$

with a probability over $\frac{1}{24}$.

Proof. From the last line of the proof of Theorem C.1, we know that there exists a MDP $\mathcal{M}$, such that

$$\mathbb{E}_{\mathcal{M}_\phi,\mu,\mathbf{ALG}}[\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}}] \geq \frac{\lambda H\tau}{12},$$

and it follows from Lemma C.3 that the random variable $\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}} \leq \lambda H\tau$. Therefore

$$\begin{aligned}
\frac{\lambda H\tau}{12} &\leq \mathbb{E}_{\mathcal{M}_\phi,\mu,\mathbf{ALG}}[\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}}]\\
&\leq \lambda H\tau\mathbb{P}[\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}} > \frac{\lambda H\tau}{24}] + \frac{\lambda H\tau}{24}\mathbb{P}[\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}} \leq \frac{\lambda H\tau}{24}]\\
&\leq \frac{\lambda H\tau}{12}(12\mathbb{P}[\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}} > \frac{\lambda H\tau}{24}] + \frac{1}{2})\\
&\Rightarrow \mathbb{P}[\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}} > \frac{\lambda H\tau}{24}] \geq \frac{1}{24}
\end{aligned}$$

□

C.2.4 Proof of Lemma C.1

Proof. From the construction we see that the policy doesn't influence the probability of reaching s_i at any time step. And the uniform random behavior policy makes sure that there is a chance of $1/A$ to visit action a_i at any state. For s_g and s_b, because each of them has a initial probability of $1/3$, the probability of reaching one of them at ant time step would be in $[\frac{1}{3}, \frac{2}{3}]$. By letting a optimal policy always choose a_1 in s_g and s_b as μ does, we make $\frac{d_h^*(s_g)}{d_h^\mu(s_g)} \leq \frac{2/3}{1/3} = 2 \leq C^*$. Therefore $C^* = \lambda_1$

As for P, we see that the probabality to reach s_g and s_d with behavior policy at step h is

$$\begin{aligned}
d_h^\mu(s_g, a) &\geq \frac{1-\lambda}{2} \geq \frac{\lambda}{2} \geq \frac{\lambda}{eS\lambda_1},\\
d_h^\mu(s_b, a) &\geq \frac{1-\lambda}{2} \geq \frac{\lambda}{2} \geq \frac{\lambda}{eS\lambda_1},\\
d_h^\mu(s_i, a) &= \frac{\lambda}{S}(1-\frac{1}{H})^{h-1}\frac{1}{\lambda_1} \geq \frac{\lambda}{eS\lambda_1}.
\end{aligned}$$

The part for $\mathrm{gap}_{\min}$ is direct calculation,

$$\mathrm{gap}_{\min} = \frac{1}{2H}(1+2\tau)H + 0 - \frac{1}{2H}H - 0 = \tau.$$

□

C.2.5 Proof of Lemma C.2

Proof. Because $\hat{\pi}$ only makes mistakes in s_i, and each mistake results in a expected τ decrease in final cumulative reward, we can directly calculate the expected loss with the performance difference lemma for finite-horizon MDP ,

$$\begin{aligned}
\mathbb{E}_{\mathcal{M}_\phi,\mu,\mathbf{ALG}}[\mathbf{V}_0^* - \mathbf{V}_0^{\hat{\pi}}] &= \mathbb{E}_{\hat{\pi},\mathcal{M}_\phi}[\mathrm{gap}_h(s_h,a_h)] \\
&= \sum_{i=1}^{S}\sum_{h=1}^{H} d_h^\mu(s_i) l_{h,i}(\mathbf{ALG})\tau \\
&= \sum_{i=1}^{S}\sum_{h=1}^{H} P_0(s_i)(1-\frac{1}{H})^{h-1}\tau l_{h,i}(\mathbf{ALG}) \\
&\geq \sum_{i=1}^{S}\sum_{h=1}^{H} \frac{\lambda}{eS}\tau l_{h,i}(\mathbf{ALG}) \\
&\geq \frac{\lambda}{eS}\tau \sum_{(h,i)\in[H]\times[S]} l_{h,i}(\mathbf{ALG}).
\end{aligned}$$

□

C.2.6 Proof of Lemma C.3

Proof. The proof is a direct result of performance decomposition lemma.

$$\begin{aligned}
\max_\pi \mathbf{V}_0^* - \mathbf{V}_0^\pi &= \max_\pi \sum_{h=1}^{2H} \mathbb{E}_{\pi^*}[\mathbf{V}_h^*(s) - \mathbf{V}_h^\pi(s)] \\
&= \max_\pi \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\mathbf{V}_h^*(s) - \mathbf{V}_h^\pi(s)] \\
&\leq \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\max_\pi \mathbf{V}_h^*(s) - \mathbf{V}_h^\pi(s)] \\
&= \sum_{h=1}^{H}\sum_{i=1}^{S} d_h^*(s)\max_\pi \mathbf{V}_h^*(s_i) - \mathbf{V}_h^\pi(s_i) \\
&= \sum_{h=1}^{H}\sum_{i=1}^{S} d_h^*(s)\tau \\
&= \lambda H\tau
\end{aligned}$$

□

D Proof of Necessity of Overall Data Coverage

One may wonder if the Assumption 3.1 has been too strong, because the minimax bound $O(\frac{H^3\iota}{P\epsilon^2})$ only requires the data coverage over a single optimal policy. Here we give a proof that to derive ϵ-irrevelant bounds for Algorithm 2, single optimal policy coverage is not sufficient.
We provide a hard instance to prove that if we only have data coverage over one of the optimal

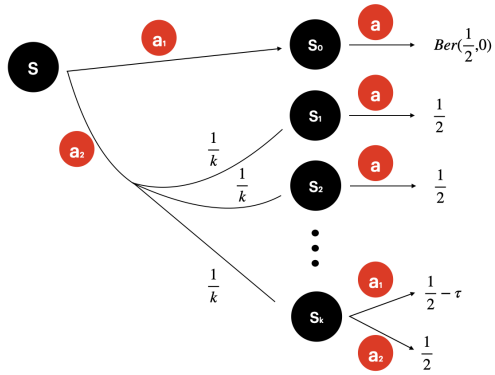

Figure 1: A hard instance with horizon 2, 2 actions and $k+2$ states. a_1 at s leads to s_0. The reward of both actions at s_0 is sampled from Bernoulli Distribution. a_2 at s leads to a uniformly random transition to $s_i, i = 1, \cdots, k$. The reward of both actions at $s_i, i = 1, \cdots, k-1$ are $\frac{1}{2}$. a_1 at s_k receives $\frac{1}{2} - \tau$ reward and a_2 at s_k receives $\frac{1}{2}$ reward. $\text{gap}_{\min} = \tau$ in this MDP.

policies, Algorithm 2 may output suboptimal policy with probability over $1/2$. We use P' to refer to the single policy coverage coefficient, i.e.,

$$P' = \max_{\pi^*} \min_{d_h^{\pi^*}(s,a)>0} d_h^{\mu}(s,a).$$

Here we consider the MDP with horizon length 2, 2 actions and k+2 states, which is illustrated in Figure 1. The initial state is s, and $P(s, a_1, s_0) = 1, P(s, a_2, s_i) = \frac{1}{k}$ for $i = 1, 2 \cdots, k$. The rewards of actions in $s_i, i = 1, 2 \cdots, k-1$ are all $\frac{1}{2}$, and the rewards of both actions in s_0 are sampled from $\mathcal{N}(\frac{1}{2}, 1)$. $r(s_k, a_1) = \frac{1}{2} - \tau$, $r(s_k, a_2) = \frac{1}{2}$.

In this MDP $\text{gap}_{\min} = \epsilon$, and the only suboptimal action is to take a_2 at s_k. Define μ,

$$\begin{aligned}
&\mu(a_1 \mid s) = \frac{1}{k+1}, \\
&\mu(a_2 \mid s) = \frac{k}{k+1}, \\
&\mu(a_1 \mid s_i) = 1 \qquad i = 0, 1, \cdots, k, \\
&\mu(a_2 \mid s_i) = 0 \qquad i = 0, 1, \cdots, k.
\end{aligned}$$

Then we can see that an optimal policy/route $s - a_1 - s_0 - a_1$ has been covered by μ with minimal coverage distribution $P' = \frac{1}{k+1}$. Then we show that for any constant C, the output policy of VI-LCB with $N = \frac{C}{P' \text{gap}_{\min}^2} = \frac{C(k+1)}{\tau^2}$ sample trajectories cannot be guaranteed to be optimal with high probability. For the conciseness of proof, we assume that $C > 10$, $C_b > 16$ and let $k > 10$.

Intuitively, this is because there exists the probability that some not-so-well covered optimal policy outperforms the covered one in execution process, and no optimality can be guaranteed over the not-so-well covered one. In this instance, (s, a_2) is also optimal, but as no information about s_k, a_2 is known by VI-LCB, it will choose a_1 following the principle of pessimism.

In the following proof, we omit the subscripts indicating the time step because the each state only appears in time step 1 or 2, which will not incur confusion. Rigorously, we define the event the $\{\hat{\pi}(s) = a_2\}$ as ξ_{bad},

$$\begin{aligned}
\mathbb{P}[\xi_{bad}] &= \mathbb{P}[\underline{\mathbf{Q}}(s, a_1) \leq \underline{\mathbf{Q}}(s, a_2)] \\
&\geq \mathbb{P}[\underline{\mathbf{Q}}(s, a_1) \leq \frac{1}{2} - \lambda\tau \leq \underline{\mathbf{Q}}(s, a_2)] \\
&\geq 1 - \mathbb{P}[\underline{\mathbf{Q}}(s, a_1) \geq \frac{1}{2} - \lambda\tau] - \mathbb{P}[\underline{\mathbf{Q}}(s, a_2) \leq \frac{1}{2} - \lambda\tau].
\end{aligned}$$

where λ can be any positive constant, which will be determined later. We limit these two terms respectively.

$$\mathbb{P}[\underline{\mathbf{Q}}(s,a_1) \geq \frac{1}{2} - \lambda\tau] \leq \mathbb{P}[N(s,a_1) \geq \frac{C_1 N}{k+1}] + \mathbb{P}[N(s,a_1) \leq \frac{C_2 N}{k+1}] \\ + \mathbb{P}[N(s,a_1) \in [\frac{C_2 C}{\tau^2}, \frac{C_1 C}{\tau^2}],\ \underline{\mathbf{Q}}(s,a_1) \geq \frac{1}{2} - \lambda\tau].$$

Because $N(s,a_1) \sim Bio(N, \frac{1}{k+1})$, it follows from the asymptotic feature of binomial distribution that $N(s,a_1) - \frac{C}{\tau^2} \sim subG(\frac{Ck}{\tau^2(k+1)})$, and then

$$\mathbb{P}[N(s,a_1) \geq \frac{C_1 C}{\tau^2}] \leq \exp\left(-\frac{(C_1-1)^2 C^2}{2(1-1/(k+1))^2}\right) \leq \exp(-\frac{C^2}{4}(C_1-1)^2),$$
$$\mathbb{P}[N(s,a_1) \leq \frac{C_2 C}{\tau^2}] \leq \exp\left(-\frac{(C_2-1)^2 C^2}{2(1-1/(k+1))^2}\right) \leq \exp(-\frac{C^2}{4}(C_2-1)^2).$$

Let $C_1 = 1.5, C_2 = 0.5$. Remember that we assume that $C > 10$, and this makes the sum of above two terms a small constant smaller than 0.1. Because $\underline{\mathbf{Q}}(s,a_1) = -b(s,a_1) + \hat{r}(s_0,a_1) - b(s,a_1) \leq -\frac{3C_b\sqrt{\iota}}{\sqrt{N(s,a_1)}} + \hat{r}(s_0,a_1)$, and the center limit theorem allow us to use $X \sim \mathcal{N}(\frac{1}{2}, \frac{1}{N(s_0,a_1)})$ to replace $\hat{h}(s_0,a_1)$,

$$\begin{aligned}\mathbb{P}[N(s,a_1) \in [\frac{C_2 C}{\tau^2}, \frac{C_1 C}{\tau^2}],\ \underline{\mathbf{Q}}(s,a_1) \geq \frac{1}{2} - \lambda\tau] &\leq \mathbb{P}[\hat{r}(s_0,a_1) - \frac{3C_b\sqrt{\iota}\tau}{\sqrt{C_1 C}} \geq \frac{1}{2} - \lambda\tau] \\ &\lesssim \mathbb{P}[X - \frac{1}{2} \geq \frac{3C_b\sqrt{\iota}\tau}{\sqrt{C_1 C}} - \lambda\tau] \\ &\leq \exp(-\frac{1}{2}(\frac{3C_b\sqrt{\iota}}{\sqrt{C_1 C}} - \lambda)^2 C C_2) \\ &\leq \exp(-(\frac{1}{2}C_b\sqrt{\iota} - \lambda)^2).\end{aligned}$$

The above term disappears quickly when $\iota = \Omega(\log\frac{k}{\delta})$ becomes larger. We will choose a $\lambda < \frac{1}{4}C_b\sqrt{\iota}$. Then we consider $\underline{\mathbf{Q}}(s,a_2)$,

$$\begin{aligned}\underline{\mathbf{Q}}(s,a_2) &= -b(s,a_2) - \sum_{i=1}^{k}\hat{P}(s,a_2,s_i)(r(s_i,a_1) - b(s_i,a_1)) \\ &= \frac{1}{2} - \frac{N(s,a_2,s_k)\tau}{N(s,a_2)} - \frac{(k+1)C_b\iota}{N(s,a_2)} - C_b\sqrt{\frac{\mathbf{Var}_{\hat{P}_{s,a_2}}(\underline{\mathbf{V}})\iota}{N(s,a_2)}}.\end{aligned}$$

With similar induction, we can prove that event $\{N(s,a_2) \geq \frac{Ck}{2\tau^2}\}\bigcap\{N(s,a_2,s_i) \in [\frac{C}{2\tau^2}, \frac{2C}{\tau^2}]\}_{i=1,\cdots,k}$ with probability over 0.8. When we have this concentration event true,

$$\mathbb{P}[\underline{\mathbf{Q}}(s,a_2) \leq \frac{1}{2} - \lambda\tau] \leq \mathbb{P}[-\frac{\tau}{4k} - C_b\tau^2\iota - \frac{C_b\tau^2\iota}{k} \leq -\lambda\tau].$$

By taking $\lambda = \frac{1}{4k} + C_b\tau\iota + \frac{C_b\tau^2}{k} + \epsilon$, where ϵ is a extremely small positive constant, $\mathbb{P}[\underline{\mathbf{Q}}(s,a_2) \leq \frac{1}{2} - \lambda\tau] = 0$. And further letting $\tau \leq \frac{1}{40\sqrt{\iota}}$, we have $\lambda \leq \frac{1}{4}C_b\sqrt{\iota}$. Putting the above inductions together,

$$\mathbb{P}[\xi_{bad}] \geq 1 - 0.1 - 0.2 - \exp(-\frac{1}{16}C_b^2\iota) \geq \frac{1}{2}.$$

This result points out that Algorithm 2 has a chance of over 1/2 to return a suboptimal policy. Therefore a overall coverage over all the optimal policies is necessary to derive a ϵ-irrelevant bound for VI-LCB.