

A DETAILS ABOUT BANDITS AND RL

In this paper we consider conservative bandits and conservative reinforcement learning problems.

A.1 CONSERVATIVE MULTI-ARMED BANDIT

The multi-armed bandit problem is a sequential decision-making task in which a learning agent repeatedly chooses an action (called an arm) and receives a reward corresponding to that action. We assume there are $K + 1$ arms, denoted by $\{0, \dots, K\}$. There is a reward $X_{t,i}$ associated with each arm i at each round $t \in \{1, 2, \dots\}$. In each round t , the agent pulls arm $I_t \in \{0, \dots, K\}$ and receives a reward X_{t,I_t} corresponding to this arm. The agent does not observe the other rewards $X_{t,j}$ ($j \neq I_t$).

The learning performance of an agent over a time horizon T is usually measured by its regret, which is the difference between its reward and what it could have achieved by consistently choosing the single best arm in hindsight:

$$R_T = \max_{i \in \{0, \dots, K\}} \sum_{t=1}^T X_{t,i} - X_{t,I_t} \quad (8)$$

In conservative multi-armed bandits, we assume that the conservative default action is arm 0, and its reward is fixed and is known. That is, $X_{0,t} = \mu_0$ for all t . On the other hand, each arm $i > 0$ has a stochastic reward $X_{t,i} = \mu_i + \eta_{t,i}$, where $\mu_i \in [0, 1]$ is the expected reward of arm i and η_t is a random noise such that

Assumption 2. Each element η_t of the noise sequence $\{\eta_t\}_{t=1}^\infty$ is conditionally 1-sub-Gaussian, i.e.

$$\forall \zeta \in \mathbb{R}, \quad \mathbb{E} \left[e^{\zeta \eta_t} \mid a_{1:t}, \eta_{1:t-1} \right] \leq \exp \left(\frac{\zeta^2}{2} \right) \quad (9)$$

The sub-Gaussian assumption automatically implies that $\mathbb{E}[\eta_t \mid a_{1:t}, \eta_{1:t-1}] = 0$ and $\text{Var}[\eta_t \mid a_{1:t}, \eta_{1:t-1}] \leq 1$.

We denote the expected reward of the optimal arm by $\mu^* = \max_i \mu_i$ and the gap between it and the expected reward of the i th arm by $\Delta_i = \mu^* - \mu_i$.

In conservative multi-armed bandits, we constrain the learner to earn at least a $1 - \alpha$ fraction of the reward from simply playing arm 0 :

$$\sum_{s=1}^t X_{s,I_s} \geq (1 - \alpha) \sum_{s=1}^t X_{s,0} \quad \text{for all } t \in \{1, \dots, T\} \quad (10)$$

where $\alpha \in (0, 1)$ is a predefined constant. The parameter α controls how conservative the agent should be. Small values of α show that only small losses are tolerated, and thus, the agent should be overly conservative, whereas large values of α indicate that the manager is willing to take risk, and thus, the agent can explore more and be less conservative.

A.2 CONSERVATIVE LINEAR BANDITS

In the linear bandit setting, in each round t , the agent is given a set of (possibly) infinitely many actions/options \mathcal{A} , where each action $a \in \mathcal{A}$ is associated with a feature vector $\phi_a \in \mathbb{R}^d$. At each round t , the agent should select an action $a_t \in \mathcal{A}$. Upon selecting a_t , the agent observes a random reward X_t generated as

$$X_{t,a_t} = \langle \theta^*, \phi_{a_t} \rangle + \eta_t, \quad (11)$$

where $\theta^* \in \mathbb{R}^d$ is the unknown reward parameter, $\langle \theta^*, \phi_{a_t} \rangle = r_{a_t}$ is the expected reward of action a_t at time t , i.e., $r_{a_t} = \mathbb{E}[X_{t,a_t}]$, and η_t is a random noise that satisfies Assumption 2.

We also make the following standard assumption on the unknown parameter θ^* and feature vectors:

Assumption 3. *There exist constants $B, D \geq 0$ such that $\|\theta^*\|_2 \leq B$, $\|\phi_a\|_2 \leq D$, and $\langle \theta^*, \phi_a \rangle \in [0, 1]$, for all t and all $a \in \mathcal{A}$.*

We define $\mathcal{B} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq B\}$ and $\mathcal{F} = \{\phi \in \mathbb{R}^d : \|\phi\|_2 \leq D, \langle \theta^*, \phi \rangle \in [0, 1]\}$ to be the parameter space and feature space, respectively.

Similar to multi-armed bandits, the goal of the agent is to minimize the following regret:

$$R_T = \max_{a \in \mathcal{A}} \sum_{t=1}^T X_{t,a} - X_{t,a_t} \quad (12)$$

which is the difference between the cumulative reward of the optimal action and agent's strategies.

In the conservative linear bandit setting, at each round t , there exists a conservative action $b \in \mathcal{A}_t$ and selecting b incurs expected reward r_b . We assume that r_b is known, and the conservative action is not relevant to the underlying parameter θ_* . We constrain the learner to earn at least a $1 - \alpha$ fraction of the reward from simply playing arm b :

$$\sum_{i=1}^t r_{a_i} \geq (1 - \alpha) \sum_{i=1}^t r_b, \quad \forall t \in [T] \quad (13)$$

A.3 CONSERVATIVE TABULAR MDPs

We consider conservative exploration in finite horizon tabular MDPs. An MDP can be represent as $M = (\mathcal{S}, \mathcal{A}, H, p, r)$, where \mathcal{S} is the state space, \mathcal{A} is the action space, H is the length of each episode. Every state-action pair (s, a) is characterized by a reward distribution with mean $r(s, a)$ and support in $[0, r_{\max}]$, and a transition distribution $p(\cdot | s, a)$ over next states. We denote by $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$. In each episode, the agent starts from an initial state s_1 . At each step $h \in [H]$, the agent takes action a_h in state s_h and receive a random reward r_h with mean $r(s, a)$, and transits to state s_{h+1} according to the distribution $p(\cdot | s, a)$.

A (randomized) policy π is a set of functions $\{\pi_h : \mathcal{S} \mapsto \Delta(\mathcal{A})\}_{h \in [H]}$. Given a policy π , a level $h \in [H]$ and a state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, the Q function and the value function are defined as:

$$Q_h^\pi(s, a) = \mathbb{E} \left[\sum_{h'=h}^H r_{h'} | s_h = s, a_h = a, \pi \right],$$

$$V_h^\pi(s) = \mathbb{E} \left[\sum_{h'=h}^H r_{h'} | s_h = s, \pi \right].$$

We let $V_{H+1}(s) = 0$ and $Q_{H+1}(s, a) = 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. We use Q_h^* and V_h^* to denote the optimal Q -function and V -function at level $h \in [H]$ without corruptions, which satisfies $Q_h^*(s, a) = \max_\pi Q_h^\pi(s, a)$ and $V_h^*(s) = \max_a Q_h^*(s, a)$ respectively.

In conservative tabular MDPs, at the beginning of each episode t , the agent can choose to run a conservative policy π_0 , which will give the agent a fixed reward $V_1^{\pi_0}$ and ends the episode immediately, or choose to explore in the target MDP M with policy π_k , and will receive a total reward $V_1^{\pi_t}$. Our goal is to minimize the following regret

$$R_T = \sum_{t=1}^T V_1^*(s_1) - V_1^{\pi_t}(s_1) \quad (14)$$

while satisfying the following conservative constraint

$$\sum_{j=1}^t V_1^{\pi_j}(s_1) \geq (1 - \alpha)tV_1^{\pi_0}(s_1), \quad \forall t \in [T]. \quad (15)$$

A.4 CONSERVATIVE LINEAR MDPs

The conservative linear MDP setting is nearly the same as tabular MDPs, except that \mathcal{S} is a measurable space with possibly infinite number of elements and \mathcal{A} is a finite set with cardinality A . We assume that the transition kernels and the reward function are assumed to be linear (Jin et al., 2020).

Assumption 4 (Linear MDP). *An MDP $(\mathcal{S}, \mathcal{A}, H, p, r)$ is a linear MDP with a feature map $\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$, if for any $h \in [H]$, there exist d unknown (signed) measures $\mu_h = (\mu_h^{(1)}, \dots, \mu_h^{(d)})$ over \mathcal{S} and an unknown vector $\theta_h \in \mathbb{R}^d$, such that for any $(x, a) \in \mathcal{S} \times \mathcal{A}$, we have*

$$\mathbb{P}_h(\cdot | x, a) = \langle \phi(x, a), \mu_h(\cdot) \rangle, \quad r_h(x, a) = \langle \phi(x, a), \theta_h \rangle. \quad (16)$$

Without loss of generality, we assume $\|\phi(x, a)\| \leq 1$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$, and $\max\{\|\mu_h(\mathcal{S})\|, \|\theta_h\|\} \leq \sqrt{d}$ for all $h \in [H]$.

B LOWER BOUNDS FOR NON-CONSERVATIVE EXPLORATION

Lemma 4 (Lower Bound for Multi-Armed Bandit). *Let $K > 1$ and $T \geq k - 1$. Then for any multi-armed bandit algorithm, there exists a mean vector $\mu \in [0, 1]^K$ such that*

$$\mathbb{E}[R_T] \gtrsim \sqrt{KT}.$$

Proof. See Theorem 15.2 of Lattimore & Szepesvári (2020) for a detailed proof. \square

Lemma 5 (Lower Bound for Linear Bandit). *Let $d \leq 2T$. Then for any linear bandit algorithm, there exists a parameter $\theta \in \mathbb{R}^d$ such that*

$$\mathbb{E}[R_T] \gtrsim d\sqrt{T}.$$

Proof. See Theorem 24.2 of Lattimore & Szepesvári (2020) for a detailed proof. \square

Lemma 6 (Lower Bound for Tabular RL). *Let $T \geq SA$. Then for any bandit RL algorithm, there exists an MDP such that*

$$\mathbb{E}[R_T] \gtrsim \sqrt{SAH^3T}.$$

Proof. See Jaksch et al. (2010); Azar et al. (2017); Jin et al. (2018) for a detailed proof. \square

Lemma 7 (Lower Bound for Linear MDP). *Let $T \geq d$. Then for any bandit RL algorithm, there exists an MDP such that*

$$\mathbb{E}[R_T] \gtrsim \sqrt{d^2H^3T}.$$

Proof. This lower bound is obtained by extrapolating the lower bounds of linear bandit and tabular RL. \square

C DETAILED PROOF FOR LOWER BOUNDS

Proof of Theorem 1. Let's consider any sequential decision making problem \mathfrak{A} (for instance a multi-armed bandit problem, linear bandit, tabular RL or linear RL) such that there exists $\xi \in \mathbb{R}$ (a constant solely depending on the sequential decision making problem, e.g., the dimension in linear problems or the number of action in tabular problems), an instance of problem \mathfrak{A} where for a number of time steps T large enough and any algorithm \mathcal{A} we have that:

$$\mathbb{E}[R_{\mathfrak{A}}^T(\mathcal{A})] \geq \xi\sqrt{T}, \quad (17)$$

with $R_{\mathfrak{A}}^T(\mathcal{A})$ the regret of algorithm \mathcal{A} in problem \mathfrak{A} . For instance, in the MAB case $\xi = \sqrt{K} - 1/27$ with K the number of arms. Using this non-conservative lower bound, we show our lower bound for the conservative setting for the problem \mathfrak{A} with a baseline policy π_0 . To do so, let's consider any conservative algorithm (that is to say it satisfies Eq. (3)) noted as \mathcal{A}_c . We assume this algorithms

selects policies $(\pi^t)_{t \in [T]}$ and let \mathcal{T}_0 denotes the set of rounds in $\{1, \dots, T\}$ where \mathcal{A}_c selects the conservative policy π_0 . Here $T \geq \frac{\xi^2}{\alpha V^{\pi_0} \cdot (\alpha V^{\pi_0} + \Delta_0)} + \frac{\xi^2}{4(\alpha V^{\pi_0} + \Delta_0)^2}$.

We now distinguish two cases:

- If $\mathbb{E}|\mathcal{T}_0| \geq \frac{\xi^2}{\alpha V^{\pi_0} \cdot (\alpha V^{\pi_0} + \Delta_0)}$, then the definition of the regret implies that:

$$\mathbb{E}[R_{\mathcal{A}}^T(\mathcal{A}_c)] \geq \mathbb{E} \sum_{t \in \mathcal{T}_0} [V^* - V^{\pi^t}] = \mathbb{E}|\mathcal{T}_0| \cdot \Delta_0 \geq \frac{\xi^2 \Delta_0}{\alpha V^{\pi_0} \cdot (\alpha V^{\pi_0} + \Delta_0)}. \quad (18)$$

- If $\mathbb{E}|\mathcal{T}_0| < \frac{\xi^2}{\alpha V^{\pi_0} \cdot (\alpha V^{\pi_0} + \Delta_0)}$, then let's note $\mathcal{T}_0^c = \{i_1, i_2, \dots, i_{|\mathcal{T}_0^c|}\}$ the set of time steps where \mathcal{A}_c does not execute the conservative policy π_0 . Considering the budget as we have defined in Def. 1 we have:

$$\begin{aligned} B_{\mathcal{T}_0^c}(\mathcal{A}_c) &= \max_{t \in \mathcal{T}_0^c} \mathbb{E} \sum_{k=1}^t [(1 - \alpha)V^{\pi_0} - V^{\pi^k}] \\ &= \max_{t \in \mathcal{T}_0^c} \mathbb{E} \sum_{k=1}^t [V^* - V^{\pi^k} - \alpha V^{\pi_0} - (V^* - V^{\pi_0})] \\ &= \max_{t \in \mathcal{T}_0^c} \mathbb{E}[R_{\mathcal{A}}^{\mathcal{T}_0^c}(\mathcal{A}_c)(t)] - (\alpha V^{\pi_0} + \Delta_0)t, \end{aligned} \quad (19)$$

where $\Delta_0 = V^* - V^{\pi_0}$ is the difference between the optimal policy and the baseline policy and $\mathbb{E}[R_{\mathcal{A}}^{\mathcal{T}_0^c}(\mathcal{A}_c)(t)]$ is the regret incurred by the rounds $\{i_k\}_{k \in [t]}$. Therefore, for any $t \in [|\mathcal{T}_0^c|]$, by Eq. (17) we have that there exists an instance u (for instance in a bandit problem u is the means of each arm) of \mathcal{A} such that $\mathbb{E}[R_{\mathcal{A}}^{\mathcal{T}_0^c}(\mathcal{A}_c)(t)] \geq \xi\sqrt{t}$. Let $t_0 = \frac{\xi^2}{4(\alpha V^{\pi_0} + \Delta_0)^2}$, then there exists an instance such that

$$B_{\mathcal{T}_0^c}(\mathcal{A}_c) \geq \xi\sqrt{t_0} - (\alpha V^{\pi_0} + \Delta_0)t_0 \gtrsim \frac{\xi^2}{\alpha V^{\pi_0} + \Delta_0}. \quad (20)$$

Combining the conservative condition in Equation (3), we have

$$\mathbb{E}|\mathcal{T}_0| \geq \frac{B_{\mathcal{T}_0^c}(\mathcal{A}_c)}{\alpha V^{\pi_0}} \gtrsim \frac{\xi^2}{\alpha V^{\pi_0} \cdot (\alpha V^{\pi_0} + \Delta_0)}.$$

By the same derivation of Equation (18), we have

$$\mathbb{E}[R_{\mathcal{A}}^T(\mathcal{A}_c)] \gtrsim \frac{\xi^2 \Delta_0}{\alpha V^{\pi_0} \cdot (\alpha V^{\pi_0} + \Delta_0)}. \quad (21)$$

Combining Equations (17), (18), and (21), we obtain

$$\mathbb{E}[R_{\mathcal{A}}^T(\mathcal{A})] \gtrsim \max \left\{ \xi\sqrt{T}, \frac{\xi^2 \Delta_0}{\alpha V^{\pi_0} \cdot (\alpha V^{\pi_0} + \Delta_0)} \right\}. \quad (22)$$

Then we discuss the lower bound for different setups.

- For multi-armed bandits, by Lemma 4, we choose $\xi = \sqrt{K}$. Then we have

$$\mathbb{E}[R_T] \gtrsim \max \left\{ \sqrt{KT}, \frac{\xi^2 \Delta_0}{\alpha V^{\pi_0} \cdot (\alpha V^{\pi_0} + \Delta_0)} \right\}.$$

- For linear bandits, by Lemma 5, we choose $\xi = d$. Then we have

$$\mathbb{E}[R_T] \gtrsim \max \left\{ d\sqrt{T}, \frac{d^2 \Delta_0}{\alpha V^{\pi_0} \cdot (\alpha V^{\pi_0} + \Delta_0)} \right\}.$$

- For tabular RL, by Lemma 6, we choose $\xi = \sqrt{SAH^3}$. Then we have

$$\mathbb{E}[R_T] \gtrsim \max \left\{ \sqrt{SAH^3 T}, \frac{SAH^3 \Delta_0}{\alpha V^{\pi_0} \cdot (\alpha V^{\pi_0} + \Delta_0)} \right\}.$$

- For low-rank MDP, by Lemma 7, we choose $\xi = \sqrt{d^2 H^3}$. Then we have

$$\mathbb{E}[R_T] \gtrsim \max \left\{ \sqrt{d^2 H^3 T}, \frac{d^2 H^3 \Delta_0}{\alpha V^{\pi_0} \cdot (\alpha V^{\pi_0} + \Delta_0)} \right\}.$$

Therefore, we conclude the proof. \square

D DETAILED PROOF FOR UPPER BOUNDS

D.1 PROOF OF THEOREM 2

Proof. Given a non-conservative algorithm $\tilde{\mathfrak{A}}$, the minimum amount of rewards needed to play this non-conservative algorithm for T consecutive steps is the budget defined in Def. 1. Indeed, if we denote by $\{\tilde{\pi}_l \mid l \leq T\}$ the sequence of non-conservative policies executed by $\tilde{\mathfrak{A}}$, then for any set $\mathcal{O} \subset [T]$ the budget can be rewritten as:

$$\begin{aligned} \mathcal{B}_T(\mathcal{O}, \{\tilde{\pi}_l \mid l \leq T\}) &= \max_{t \in \mathcal{O}} \sum_{l \in \mathcal{O} \cap [t]} (1 - \alpha) V^{\pi_0} - V^{\pi_l} \\ &= \max_{t \in \mathcal{O}} \sum_{l \in \mathcal{O} \cap [t]} \left(V^* - V^{\pi_l} - (\Delta_0 + \alpha V^{\pi_0}) |\mathcal{O} \cap [t]| \right). \end{aligned}$$

Let's define $R_{\mathcal{O} \cap [t]}(\tilde{\mathfrak{A}}) := \sum_{l \in \mathcal{O} \cap [t]} V^* - V^{\pi_l}$ the regret over the time steps in \mathcal{O} of the non-conservative algorithm $\tilde{\mathfrak{A}}$. Since $R_t(\tilde{\mathfrak{A}}, \mathcal{O}) = \mathcal{O}(C\sqrt{|\mathcal{O} \cap [t]|})$ w.h.p., where $C \in \mathbb{R}$ is a problem-dependent quantity as in Theorem 1. Therefore, we have

$$\begin{aligned} \mathcal{B}_T(\mathcal{O}, \{\tilde{\pi}_l \mid l \leq T\}) &= \max_{t \in \mathcal{O}} \sum_{l \in \mathcal{O} \cap [t]} (1 - \alpha) V^{\pi_0} - V^{\pi_l} \\ &= \max_{t \in \mathcal{O}} \sum_{l \in \mathcal{O} \cap [t]} \left(\mathcal{O}(C\sqrt{|\mathcal{O} \cap [t]|}) - (\Delta_0 + \alpha V^{\pi_0}) |\mathcal{O} \cap [t]| \right). \end{aligned}$$

Let $f(x) = C\sqrt{x} - (\Delta_0 + \alpha V^{\pi_0})x$, then we have $f(x) \leq \frac{C^2}{\Delta_0 + \alpha V^{\pi_0}}$. This implies that the budget required by $\tilde{\mathfrak{A}}$ is at least $\frac{C^2}{\Delta_0 + \alpha V^{\pi_0}}$. Therefore, the simple algorithm playing the baseline policy for the first $t_0 := \mathcal{O}(\frac{\xi}{\alpha V^{\pi_0} + \Delta_0})$ steps and then running the non-conservative algorithm $\tilde{\mathfrak{A}}$, is conservative. This is actually the algorithm BudgetFirst.

The regret of BudgetFirst can be bounded as

$$\text{Reg}(T) \leq t_0 + R_t(\tilde{\mathfrak{A}}, \mathcal{O}) = \mathcal{O}\left(\frac{\xi}{\alpha V^{\pi_0} + \Delta_0} + C\sqrt{|\mathcal{O} \cap [t]|}\right)$$

Thus we finish the proof. \square

Now we discuss the regret upper bound for different setups. For multi-armed bandit, the UCB algorithm (Lattimore & Szepesvári, 2020) gives us the following guarantee.

Lemma 8 (Upper Bound for Multi-Armed Bandit). *The regret of UCB can be upper bounded by*

$$R_T \leq 8\sqrt{Tk \log(T)} + 3 \sum_{i=1}^k \Delta_i \quad (23)$$

Proof. See Theorem 7.2 in Lattimore & Szepesvári (2020) for details. \square

For linear bandits, the LinUCB algorithm (Lattimore & Szepesvári, 2020) gives us the following guarantee.

Lemma 9 (Upper Bound for Linear Bandit). *The regret of LinUCB can be upper bounded by*

$$R_T \leq Cd\sqrt{T} \log(TL) \quad (24)$$

where $C > 0$ is a suitably large universal constant.

Proof. See Corollary 19.3 in Lattimore & Szepesvári (2020) for details. \square

For tabular RL, the UCBVI-BF algorithm in Azar et al. (2017) gives us the following guarantee.

Lemma 10 (Upper Bound for Tabular RL). *The regret of UCBVI-BF can be upper bounded by*

$$R_T \leq O(\sqrt{H^3 SAT}) \quad (25)$$

Proof. See Azar et al. (2017) for details. \square

For linear MDP, the LSVI-UCB algorithm in Jin et al. (2020) gives us the following guarantee.

Lemma 11 (Upper Bound for Linear MDP). *the total regret of LSVI-UCB is upper bounded by*

$$R_T \leq \tilde{O}\left(\sqrt{d^3 H^4 T}\right). \quad (26)$$

Proof. See Jin et al. (2020) for details. \square

D.2 PROOF OF THEOREM 3

Proof. Given an LCB algorithm $\tilde{\mathfrak{A}}$, suppose it maintains lower confidence bound $\lambda_t^{\pi_k}(\delta) \leq V^{\pi_k}$ with probability at least $1 - \delta$ that satisfies $\sum_{k=1}^t (V^{\pi_k} - \lambda_t^{\pi_k}) \leq \tilde{O}(C\sqrt{t})$. Let S_t to be the set of time step where a non-conservative policy was deployed in episodes before t . The additional budget needed by the algorithm can be written as:

$$\begin{aligned} \tilde{B}_T(S_T, \mathfrak{A}) &= \max_{t \in [T]} \sum_{l \in S_t} [(1 - \alpha)V^{\pi_0} - \lambda_t^{\pi_l}(\delta)] \\ &= \max_{t \in [T]} \sum_{l \in S_t} \left(V^* - V^{\pi_l} + V^{\pi_l} - \lambda_t^{\pi_l}(\delta) \right) - (\Delta_0 + \alpha V^{\pi_0})|S_t| \\ &\leq \max_{t \in [T]} \sum_{l \in S_t} \left(V^* - V^{\pi_l} \right) + \tilde{O}(C\sqrt{|S_t|}) - (\Delta_0 + \alpha V^{\pi_0})|S_t| \\ &= \max_{t \in [T]} R_{S_t}(\tilde{\mathfrak{A}}) + \tilde{O}(C\sqrt{|S_t|}) - (\Delta_0 + \alpha V^{\pi_0})|S_t| \end{aligned}$$

Note that $R_{S_t}(\tilde{\mathfrak{A}}) \leq \tilde{O}(C\sqrt{|S_t|})$, so the last line can be upper bounded by $\max_{t \in [T]} \left(\tilde{O}(C\sqrt{|S_t|}) - (\Delta_0 + \alpha V^{\pi_0})|S_t| \right)$. This is a quadratic function $g(x) = \tilde{O}(C\sqrt{x}) - (\Delta_0 + \alpha V^{\pi_0})x$ with variable $x = \sqrt{|S_t|}$, we have $g(x) \leq \tilde{O}\left(\frac{C^2}{\Delta_0 + \alpha V^{\pi_0}}\right)$ as a result. In other words, we show that with high probability, LCBCE only need to accumulate $\tilde{B}_T(S_T, \mathfrak{A}) \leq \tilde{O}\left(\frac{C^2}{\Delta_0 + \alpha V^{\pi_0}}\right)$. Since playing the baseline policy yields αV^{π_0} budget, LCBCE play the baseline policy for at most $\tilde{O}\left(\frac{C^2}{\alpha V^{\pi_0}(\Delta_0 + \alpha V^{\pi_0})}\right)$ times. Hence, the total regret incurred can be written as:

$$R_T(\mathfrak{A}) = R_{S_T}(\tilde{\mathfrak{A}}) + \tilde{O}\left(\frac{C^2 \Delta_0}{\alpha V^{\pi_0}(\Delta_0 + \alpha V^{\pi_0})}\right) \leq \tilde{O}(C\sqrt{T} + \frac{C^2 \Delta_0}{\alpha V^{\pi_0}(\Delta_0 + \alpha V^{\pi_0})})$$

Thus we finish the proof. \square

Proof of Corollaries of Theorem 3 Below we discuss the lower confidence bound for different setups.

Multi-armed Bandits For the MAB setting, we can calculate the lower confidence bound simultaneously with the upper confidence bound as

$$\max \left\{ 0, \hat{\mu}_i(t-1) - \sqrt{\psi^\delta(T_i(t-1))/T_i(t-1)} \right\} \quad (27)$$

where $\psi^\delta(s) = 2 \log(Ks^3/\delta)$ and $T_i(t-1)$ is the times agent pulls arm i until time $t-1$. $\hat{\mu}_i(t-1)$ is the empirical reward. This is similar to the calculation of UCB in [Lattimore & Szepesvári \(2020\)](#).

Linear Bandits For the linear bandit setting, the lower confidence bound can be chosen as follows: first, we calculate the optimal action

$$(a'_t, \tilde{\theta}_t) \in \arg \max_{(a, \theta) \in \mathcal{A}_t \times \mathcal{C}_t} \langle \theta, \phi_a^t \rangle \quad (28)$$

where \mathcal{C}_{t+1} is the confidence set $\mathcal{C}_{t+1} = \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_{V_t} \leq \beta_{t+1} \right\}$. Then, we calculate $L_t = \min_{\theta \in \mathcal{C}_t} \langle \theta, z_{t-1} + \phi_{a'_t} \rangle$, where $z_{t-1} = \sum_{i=1}^{t-1} \phi_{a_i}$. Then L_t is a lower confidence bound of action a'_t .

Tabular MDP For tabular MDP setting, the upper bound of the Q function can be calculated as $Q_h(s, a) = r(s, a) + \hat{P}_h V_{h+1}(s, a) + b_h(s, a)$, where the bonus function is chosen to be $b_h = \tilde{O}(\sqrt{\frac{\text{Var}(V_{h+1})}{N(s, a)}} + \frac{H}{N(s, a)})$ in [Azar et al. \(2017\)](#). To obtain a high probability lower confidence bound, we substitute $b_h(s, a)$ with $-b_h(s, a)$. We use Q_h^l and V_h^l to denote the lower bound of Q_h and V_h respectively,

$$\begin{aligned} V_{h+1}^l(\cdot) &= \max_a Q_{h+1}^l(\cdot, a) \\ Q_h^l(\cdot, \cdot) &= r(\cdot, \cdot) + \hat{P}_h V_{h+1}^l(\cdot, \cdot) - b_h(\cdot, \cdot), \end{aligned}$$

then V_h^l is a lower confidence bound of V_h with high probability.

Linear MDP For linear MDP setting, the lower confidence bound can be obtained by reversing the sign of the bonus term of the upper confidence bound in [Jin et al. \(2020\)](#):

$$\begin{aligned} \Lambda_h &\leftarrow \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau) \phi(x_h^\tau, a_h^\tau)^T + \lambda \mathbf{I} \\ w_h &\leftarrow \Lambda_h^{-1} \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau) [r_h(x_h^\tau, a_h^\tau) + \max_a Q_{h+1}(x_{h+1}^\tau, a)] \\ Q_h(\cdot, \cdot) &\leftarrow \max \{ w_h^T \phi(\cdot, \cdot) - \beta [\phi(\cdot, \cdot)^T \Lambda_h^{-1} \phi(\cdot, \cdot)]^{1/2}, 0 \} \\ V_h(\cdot) &\leftarrow \max_a Q_h(\cdot, a) \end{aligned}$$

We note that for all these settings, we have $\sum_{k=1}^t (V^{\pi_k} - \lambda_t^{\pi_k}) \leq \tilde{O}(C\sqrt{t})$ with corresponding problem-dependent constant C . An easy way to see this is to use symmetry. For the above LCB algorithms, we reverse the sign of the bonus term of the upper confidence bound to obtain lower confidence bound. For example in the tabular MDP case, the regret can be bounded by

$$R_T \leq \sum_{k=1}^K V_{k,1}^u - V^{\pi_k} \leq \tilde{O}(\sum_{k=1}^K \sum_{h=1}^H b_{k,h}) \leq \tilde{O}(C\sqrt{T}).$$

Using the fact that $\sum_{k=1}^K V_{k,1}^u - V_{k,1}^l = O(\sum_{k=1}^K \sum_{h=1}^H b_{k,h})$, we have $\sum_{k=1}^K V^{\pi_k} - V_{k,1}^l \leq \tilde{O}(\sum_{k=1}^K \sum_{h=1}^H b_{k,h})$. therefore we can deduce that $\sum_{k=1}^K V^{\pi_k} - V_{k,1}^l \leq \tilde{O}(C\sqrt{T})$.

Using the same techniques, we can prove this property for the other settings.

E COMPARISON WITH WU ET AL. (2016)’S LOWER BOUND

First, we restate the lower bound of Wu et al. (2016) below.

Theorem 12 (Restatement of Theorem 9 in Wu et al. (2016)). *Suppose for any $\mu_i \in [0, 1] (i > 0)$ and V^{π_0} satisfying*

$$\min \{V^{\pi_0}, 1 - V^{\pi_0}\} \geq \max \{1/2\sqrt{\alpha}, \sqrt{e+1/2}\} \sqrt{K/T},$$

an algorithm satisfies $\mathbb{E}_\mu \sum_{t=1}^T X_{t,I_t} \geq (1 - \alpha)V^{\pi_0} T$. Then there is some $\mu \in [0, 1]^K$ such that its expected regret satisfies $\mathbb{E}_\mu R_n \geq B$ where

$$B = \max \left\{ \frac{K}{(16e+8)\alpha V^{\pi_0}}, \frac{\sqrt{KT}}{\sqrt{16e+8}} \right\}.$$

Here V^{π_0} is the reward of the conservative policy, K is the number of arms, T is the number of episodes. Compared with our result, the main difference is in the first term, where we have an additional coefficient $\frac{\Delta_0}{\alpha V^{\pi_0} + \Delta_0}$, which makes our result seems worse. However, as we will show below, in the hard instance of the proof in Wu et al. (2016), $\frac{\Delta_0}{\alpha V^{\pi_0} + \Delta_0}$ is **lower bounded by an absolute constant**. Therefore, our lower bound actually implies the result of Wu et al. (2016).

When proving the first term $\frac{K}{(16e+8)\alpha V^{\pi_0}}$ in the lower bound, Wu et al. (2016) requires that the parameters should satisfy the following conditions (see Case 2 in their proof):

$$\alpha < \frac{\sqrt{K}}{V^{\pi_0} \sqrt{(16e+8)T}}, \quad \Delta_0 = \frac{K}{4\alpha V^{\pi_0} T}.$$

With these conditions we immediately have

$$\frac{\alpha V^{\pi_0}}{\Delta_0} = \frac{4\alpha^2 (V^{\pi_0})^2 T}{K} < \frac{4}{16e+8},$$

which implies

$$1 > \frac{\Delta_0}{\alpha V^{\pi_0} + \Delta_0} > \frac{1}{\frac{4}{16e+8} + 1} > 0.9.$$

Therefore, this factor only has a constant effect, and we can recover the result of Wu et al. (2016).

F COMPARISON WITH WU ET AL. (2016)’S UPPER BOUND WITH KNOWN Δ_0

Here we discuss why the regret bound of BudgetFirst algorithm in Wu et al. (2016) is not tight and why our analysis improves theirs. In BudgetFirst, they require the number of times the agent plays π_0 to satisfy

$$(\forall t_0 \leq t \leq T) \quad tV^{\pi_0} - R_{\text{worst}} \geq (1 - \alpha)TV^{\pi_0} \quad (29)$$

where $R_{\text{worst}} = O\left(\sqrt{KT \log\left(\frac{\log(T)}{\delta}\right)}\right)$ is the worst case regret of the non-conservative algorithm in T steps. In other words, they accumulate budget by playing π_0 so that the budget can compensate for the T -step exploration of the non-conservative algorithm.

However, it is not necessary to have this much budget. Let us look at the analysis in our algorithm. In our Budget-Exploration, the budget needed can be written as

$$\begin{aligned} B_T(\mathcal{O}, \{\tilde{\pi}_l \mid l \leq T\}) &= \max_{t \in \mathcal{O}} \sum_{l \in \mathcal{O} \cap [t]} (1 - \alpha)V^{\pi_0} - V^{\pi_l} \\ &= \max_{t \in \mathcal{O}} \sum_{l \in \mathcal{O} \cap [t]} (V^* - V^{\pi_l}) - (\Delta_0 + \alpha V^{\pi_0})|\mathcal{O} \cap [t]|. \end{aligned}$$

Let us define $R_{\mathcal{O} \cap [t]}(\tilde{\mathfrak{A}}) := \sum_{l \in \mathcal{O} \cap [t]} V^* - V^{\pi_l}$ the regret over the time steps in \mathcal{O} of the non-conservative algorithm $\tilde{\mathfrak{A}}$. For UCB algorithm in MAB, $\tilde{R}_t(\tilde{\mathfrak{A}}, \mathcal{O}) = \mathcal{O}(\sqrt{K|\mathcal{O} \cap [t]|})$ w.h.p., where $K \in \mathbb{R}$ is the number of arms.

Now

$$\mathcal{B}_T(\mathcal{O}, \{\tilde{\pi}_l \mid l \leq T\}) = \max_{t \in \mathcal{O}} R_{\mathcal{O} \cap [t]}(\tilde{\mathfrak{A}}) - (\Delta_0 + \alpha V^{\pi_0}) |\mathcal{O} \cap [t]|$$

Note that the RHS is maximized when $|\mathcal{O} \cap [t]| = O(\frac{(\Delta_0 + \alpha V^{\pi_0})^2}{K})$, but not when $|\mathcal{O} \cap [t]| = T$. This means that we do not need to consider the T -step regret as in (Wu et al., 2016), which is an over-conservative estimate.