A Unified Framework for Conservative Exploration

Yunchang Yang\textsuperscript{1} Tianhao Wu\textsuperscript{2}\textsuperscript{*} Han Zhong\textsuperscript{3}\textsuperscript{*} Evrard Garcelon\textsuperscript{4} Matteo Pirotta\textsuperscript{4} Alessandro Lazaric\textsuperscript{4} Liwei Wang\textsuperscript{1,5} Simon S. Du\textsuperscript{4,6}

\textsuperscript{1}Center for Data Science, Peking University \hspace{1cm} \textsuperscript{2}Peking University
\textsuperscript{3}University of Science and Technology of China \hspace{1cm} \textsuperscript{4}Facebook AI Research
\textsuperscript{5}Key Laboratory of Machine Perception, MOE, School of EECS, Peking University \hspace{1cm} \textsuperscript{6}University of Washington

yangyc@pku.edu.cn, thwu@pku.edu.cn, zhonghan927@mail.ustc.edu.cn
{evrard, pirotta, lazaric}@fb.com, wanglw@cis.pku.edu.cn, ssdu@cs.washington.edu

Abstract

We study bandits and reinforcement learning (RL) subject to a conservative constraint where the agent is asked to perform at least as well as a given baseline policy. This setting is particularly relevant in real-world domains including digital marketing, healthcare, production, finance, etc. For multi-armed bandits, linear bandits, and tabular RL, specialized algorithms and theoretical analyses were proposed in previous work. In this paper, we present a unified framework for conservative bandits and RL, in which our core technique is to compute the necessary and sufficient budget obtained from running the baseline policy. For lower bounds, our framework gives a black-box reduction that turns a certain lower bound in the nonconservative setting into a new lower bound in the conservative setting. We strengthen the existing lower bound for conservative multi-armed bandits and obtain new lower bounds for conservative linear bandits, tabular RL, and low-rank MDP. For upper bounds, our framework turns a certain nonconservative upper-confidence-bound (UCB) algorithm into a conservative algorithm with a simple analysis. For multi-armed bandits, linear bandits, and tabular RL, our new upper bounds tighten or match existing ones with significantly simpler analyses. We also obtain a new upper bound for conservative low-rank MDP.

1 Introduction

This paper studies online sequential decision making problems such as bandits and reinforcement learning (RL) subject to a conservative constraint. Specifically, the agent is given a reliable \textit{baseline policy} that may not be optimal but still satisfactory. In conservative bandits and RL, the agent is asked to perform nearly as well (or better) as the baseline policy at all time. This setting is a natural formalization of many real-world problems such as digital marketing, healthcare, finance, etc. For example, a company may want to explore new strategies to maximize profit while simultaneously maintaining profit above a fixed baseline at any time. See [Wu et al., 2016] for more discussions on the motivation of the conservative constraint.

Existing work proposed provably efficient algorithms for different settings, including bandits [Wu et al., 2016, Kazerouni et al., 2016, Garcelon et al., 2020b, Katariya et al., 2019, Zhang et al., 2019, Du et al., 2020, Wang et al., 2021] and tabular RL [Garcelon et al., 2020a]. On the other hand, a lower bound exists only for the multi-armed bandit setting [Wu et al., 2016] and it illustrates the specific difficulties introduced by the conservative constraint compared to the non-conservative classical setting. In Section 2, we provide a more detailed discussion of the related work.

\textsuperscript{*}Equal contribution.

Preprint. Under review.
For each of the different settings considered in the literature (i.e., multi-armed bandits, linear bandits, tabular MDPs), existing approaches rely on ad-hoc algorithm design and analysis of the trade-off between the setting-specific regret analysis and the conservative constraint. Furthermore, it is hard to argue about the optimality of the proposed algorithms because it would require clever constructions of the hard instances to prove the non-trivial regret lower bounds under the conservative constraint.

1.1 Our Contributions

In this paper, we address these limitations and make significant progress in studying the general problem of online sequential decision-making with conservative constraint. We propose a unified framework that is applicable to several online sequential decision-making problems. The common theme underlying our framework is to calculate the necessary and sufficient budget required to enable non-conservative exploration. Such a budget is obtained by running the baseline policy (cf. Section 4).

Lower Bounds. For any specific problem (e.g., multi-armed bandits, linear bandits), our framework immediately turns a minimax lower bound of the non-conservative setting to a non-trivial lower bound for the conservative case (cf. Section 5). We list some examples to showcase the power of our framework for lower bounds. Full results are given in Table 1.

- We derive a novel lower bound for multi-armed bandits that works on a wider range of parameters than the one derived in [Wu et al., 2016]. In particular, our lower bound shows a more refined dependence on the value of the baseline policy.
- We derive the first regret lower bound for conservative exploration in linear bandits, tabular MDPs and low-rank MDPs. These results allow to establish or disprove the optimality of the algorithms currently available in the literature.

We emphasize our technique for deriving lower bounds is simple and generic, so we believe it can be used to obtain lower bounds for other problems as well.

Upper Bounds. Our novel view of conservative exploration can also be used to derive high probability regret upper-bounds. When the suboptimality gap $\Delta_0$ and the expected return $\mu_0$ of the baseline policy are known, we present the BudgetFirst algorithm (Alg. 1) which attains minimax optimal regret in a wide variety of sequential decision-making problems, when associated to any minimax optimal non-conservative algorithm specific to the problem at hand. In the more realistic (and challenging) scenario where $\Delta_0$ and $\mu_0$ are unknown, we show how to simply convert an entire class of algorithms with a sublinear non-conservative regret bound into a conservative algorithms with a sublinear regret bound. With this reduction, we obtain the following results, full details are given in Table 1.

- In the MAB setting, we obtain a regret upper-bound that matches our refined lower-bound, thus improving on existing analysis. In the linear bandit setting, we match existing bounds that are already minimax optimal.
- In the RL setting, we provide two novel results. First, we provide the first minimax optimal result for tabular MDPs, improving over [Garcelon et al., 2020a]. Second, we derive the first upper bound for conservative exploration in low rank MDPs. Our bound matches the rate of existing non-conservative algorithms though it is not minimax optimal. How to achieve minimax optimality in low rank MDPs is an open question even in non-conservative exploration.

Again, our reduction technique is simple and generic, and can be used to obtain new results in previously unstudied settings, like we did for low rank MDPs.

2 Related Work

Non-conservative exploration has been widely studied in bandits, and minimax optimal algorithms have been provided for the settings considered in this paper [e.g. Lattimore and Szepesvári, 2020].

---

\(^2\)We refer non-trivial regret lower bounds to lower bounds that have terms related to the conservative constraints. Note existing lower bounds in the non-conservative setting still hold in the conservative setting.

\(^3\)In [Garcelon et al., 2020a], the upper bound is scales with $r_b = \min_{s \in \mathcal{S}, p_0(s) > 0} V^*_{\pi_0}(s)$ (with $\rho_0$ the distribution of the starting state), the minimum of the baseline’s value function at the first step over the potential starting states. Here, we assume there is a unique starting state hence $r_b = \mu_0$. 


The objective of this section is to provide a unified view of the settings considered in this paper, i.e., multi-armed bandits, linear bandits, tabular Markov Decision Processes (MDPs) and low-rank MDPs. We use the RL formalism which encompasses the bandit settings.
**Notations.** We begin by introducing some basic notation. We use $\Delta(\cdot)$ to represent the set of all probability distributions on a set. For $n \in \mathbb{N}_+$, we denote $[n] = \{1, 2, \ldots, n\}$. We use $O(\cdot), \Theta(\cdot), \Omega(\cdot)$ to denote the big-O, big-Theta, big-Omega notations. We use $\tilde{O}(\cdot)$ to hide logarithmic factors. We denote $A \gtrsim (\lesssim) B$ if there exists a constant $c$ such that $A \geq (\leq) cB$.

**Tabular MDPs.** A tabular finite-horizon time-inhomogeneous MDP can be represent as a tuple $M = (\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h=1}^H, s_1, \{r_h\}_{h=1}^H)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $H$ is the length of each episode and $s_1$ is the initial state. At each stage, every state-action pair $(s, a)$ is characterized by a reward distribution with mean $r_h(s, a)$ and support in $[0, r_{\text{max}}]$, and a transition distribution $p_h(\cdot | s, a)$ over next states. We denote by $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$. A (randomized) policy $\pi \in \Pi$ is a set of functions $\{\pi_h : S \rightarrow \Delta(\mathcal{A})\}_{h \in [H]}$. For each stage $h \in [H]$ and any state-action pair $(s, a) \in S \times \mathcal{A}$, the value functions of a policy $\pi$ are defined as:

$$Q^*_h(s, a) = \mathbb{E}\left[ \sum_{h' = h}^H r_{h'} | s_h = s, a_h = a, \pi \right], \quad V^*_h(s) = \mathbb{E}\left[ \sum_{h' = h}^H r_{h'} | s_h = s, \pi \right].$$

For each policy $\pi$, we define $V^*_{h+1}(s) = 0$ and $Q^*_{h+1}(s, a) = 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. There exists an optimal policy $\pi^*$ such that $Q^*_{h+1}(s, a) = Q^*_{h+1}(s, a) = \max_{\pi} Q^*_{h}(s, a)$ satisfy the optimal Bellman equations $Q^*_{h}(s, a) = r_h(s, a) + \mathbb{E}_{s' \sim p_h(s, a)}[V^*_{h+1}(s')]$ and $V^*_{h} = \max_{\pi \in \mathcal{A}} \{Q^*_{h}(s, a)\}$. Then the optimal policy is the greedy policy $\pi^*_h(s) = \arg\max_{a \in \mathcal{A}} \{Q^*_{h}(s, a)\}$.

**Low-Rank MDPs.** We assume that $\mathcal{S}, \mathcal{A}$ are measurable spaces with possibly infinite number of elements. For algorithmic tractability, we shall restrict the attention to $\mathcal{A}$ being a finite set with cardinality $A$. When the state space is large or uncountable, value functions cannot be represented in tabular form. A standard approach is to use a parametric representation. Here, we assume that transitions and rewards are linearly representable [Jin et al., 2020].

**Assumption 1** (Low-rank 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 $\mu_h = \left(\mu^{(1)}_h, \ldots, \mu^{(d)}_h\right)$ 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

$$p_h(\cdot | x, a) = \langle \phi(x, a), \mu_h(\cdot) \rangle, \quad r_h(x, a) = \langle \phi(x, a), \theta_h \rangle. \quad (1)$$

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 $h \in [H]$.

Under certain technical conditions [e.g., Shreve and Bertsekas, 1978], all the properties of tabular MDPs extend to low-rank MDPs. In addition, the state-action value function of any policy $\pi$ is linearly representable in low-rank MDPs. Formally, for any policy $\pi$ and stage $h \in [H]$, there exists $\theta^\pi \in \mathbb{R}^d$ such that $Q^*_h(s, a) = \langle \phi(s, a), \theta^\pi_h \rangle$.

**Connection between RL and Bandits.** To have a unified view, we can represent a multi-armed bandit as a tabular MDP with $S = 1$, $A$ actions, $H = 1$ and self-loop transitions in $s_1$. In multi-armed bandits, we consider only deterministic policies so that $\Pi = \mathcal{A}$, then $V^*(s_1) = r(s_1, \pi(s_1))$ and the optimal policy is simply $\pi^* = \arg\max_{a \in \mathcal{A}} r(s_1, a)$. Similarly, a linear bandit can be modeled through low-rank MDPs with $H = 1$. For generality, we allow the action space to be possibly uncounted and we define the value of a deterministic policy $\pi = a$ ($\Pi = \mathcal{A}$) as $V^*_{1}(s_1) = r_1(s_1, a) = \langle \phi(s_1, a), \theta_1 \rangle$. The optimal policy $\pi^*$ is thus such that $\pi^* = \arg\max_{a \in \mathcal{A}} \langle \phi(s_1, a), \theta_1 \rangle$.

We refer the reader to Appendix A for details.

4 General Framework For Conservative Exploration

With the unified view provided in the previous section, we can consider a generic sequential decision-making problem $\mathcal{P}$ over $T \in \mathbb{N}^*$ episodes. We consider the standard online interaction protocol where, at each episode $t \in [T]$, the learning agent $\mathfrak{A}$ selects a policy $\pi_t$, observes and stores a

---

5When the initial state is sampled from an initial distribution, we can extend the MDP with a new stage with single state $s_1$, single action and transitions that match the initial state distribution.
trajectory \((s_t, a_t, r_t)_{t \in [H]}\), updates the policy and restart with the next episode. We evaluate the performance of the learner through the pseudo-regret. Let \(V^\pi = P^\pi(s_1)\) be the value function of a policy \(\pi\), then the regret is defined as:

\[
R_T(\mathcal{A}) = \sum_{t=1}^{T} V^* - V^{\pi_t}. \tag{2}
\]

In conservative exploration, the learner aims to minimize the regret while guaranteeing that, at any episode \(t\), their expected performance is (nearly) above the one of a baseline policy \(\pi_0\). Formally, given a possibly randomized baseline policy \(\pi_0 \in \Pi\) and a conservative level \(\alpha \in [0, 1]\), the learner should satisfy w.h.p. that

\[
\forall t \leq T, \quad \sum_{j=1}^{t} V^{\pi_j} \geq (1 - \alpha) t V^{\pi_0}. \tag{3}
\]

We assume that the value of conservative policy \(V^{\pi_0}\) is known to the agent, and playing \(\pi_0\) will not provide any information about the unexplored environment. This is often the case in practice, since a conservative policy is usually an old policy, and cannot imply any information about the new environment. In the following, we show that to satisfy this condition the learner has to collect a budget by playing the baseline policy. In particular, we show that there is an optimal number of selection of the baseline policy that allows the learner to satisfy the conservative condition and minimize the regret.

### 4.1 Budget of a Conservative Algorithm

Given the set of policies \(\{\pi_t\}_{t \in [T]}\) selected by a conservative algorithm \(\mathcal{A}\), we can divide the episodes into the set \(T_0 = \{ t \leq T \mid \pi_t = \pi_0 \}\) and its complement \(T_0^c = \{ t \leq T \mid \pi_t \neq \pi_0 \}\). The set \(T_0^c\) denotes the episode where the algorithm played an exploratory policy, i.e., it had enough budget to satisfy condition (3) through a policy \(\pi_t \neq \pi_0\), potentially non-conservative. This sequence of non-baseline policies \(\{\pi_t\}_{t \in T_0^c}\) defines a new algorithm \(\mathcal{A}^c\), that we refer as the non-conservative algorithm.

However, the algorithm \(\mathcal{A}\) is conservative therefore, for any \(\delta > 0\) and \(t \in [T]\), we have with probability at least \(1 - \delta\) that \(\sum_{t=1}^{t} V^{\pi_t} \geq (1 - \alpha) t V^{\pi_0}\). Hence, for any \(t \in [T]\) that:

\[
\alpha V^{\pi_0}|_{\mathcal{T}_{0,t}} \geq \sum_{t \in T_{0,c}} (1 - \alpha) V^{\pi_0} - V^{\pi_t}, \tag{4}
\]

where \(T_{0,t} = T_0 \cap [t]\) and \(T_{0,c} = T_0^c \cap [t]\). Taking the maximum over \(t\) in Eq. (4), we have that with high probability the conservative algorithm \(\mathcal{A}\) is such that

\[
\alpha V^{\pi_0}|_{\mathcal{T}_0} \geq \max_{t \leq T} \sum_{t \in T_{0,c}} (1 - \alpha) V^{\pi_0} - V^{\pi_t}. \tag{5}
\]

The quantity on the right of the previous equation is exactly the amount of reward that the conservative algorithm \(\mathcal{A}\) has to collect by playing the baseline policy. Hence this quantity acts as a conservative budget \(\mathcal{B}\). The higher it is, the more \(\mathcal{A}\) needs to play the baseline policy to satisfy the conservative condition. In other words, it is the least amount of reward that an algorithm needs to not violate the conservative constraint. We now extend this notion to any (non necessarily conservative) algorithm.

**Definition 1.** For any \(T \in \mathbb{N}^+\), set of episodes \(\mathcal{O} \subset [T]\) and arbitrary sequence of policies \(\{\pi_t\}_{t \in \mathcal{O}}\), the budget of this sequence of policies is defined as:

\[
\mathcal{B}_T(\mathcal{O}, \{\pi_t\}_{t \in \mathcal{O}}) = \max_{t \in \mathcal{O}} \sum_{t \in \mathcal{O} \cap [t]} (1 - \alpha) V^{\pi_0} - V^{\pi_t}. \tag{5}
\]
5 Regret Lower Bound for Conservative Exploration

In this section, we leverage the framework introduced in Section 4 to build lower bounds for several conservative sequential decision-making problems. Our result is based on the notion of budget defined in Section 4. This notion is used to build an algorithm whose regret is a lower bound for any conservative algorithm.

Theorem 1 (Conservative Exploration Regret Lower Bound). Let’s consider a decision-making problem $\mathcal{P}$ over $T$ steps, a conservative level $\alpha \in [0, 1]$, a baseline policy $\pi_0$, an algorithm $\mathfrak{A}$ and $\delta \in (0, 1)$. We assume that:

- Lower-bound for non-conservative exploration. There exists a $\xi \in \mathbb{R}_+$ and $T_0 \in \mathbb{N}$ such that for any algorithm $\mathfrak{A}'$ there exists an environment (instance of $\mathcal{P}$) such that with probability at least $1 - \delta$, $R_T(\mathfrak{P}, \mathfrak{A}') \geq \xi \sqrt{T}$ for $T \geq T_0$.
- $\mathfrak{A}$ is conservative. The algorithm $\mathfrak{A}$ is conservative, that is to say with probability at least $1 - \delta$ for any $t \leq T$, $\sum_{t=1}^{\xi} V^{\pi_t} \geq (1 - \alpha) t V^{\pi_0}$.

Then, there exists an environment (instance of problem $\mathfrak{P}$) and $T_0 \in \mathbb{N}$ such that with probability at least $1 - \delta$ and $T \geq T_0$:

$$R_T(\mathfrak{P}, \mathfrak{A}) \geq \max \left\{ \frac{\xi^2 \Delta_0}{\alpha V^{\pi_0} (\alpha V^{\pi_0} + \Delta_0) \bigg\} \right.$$

where $\Delta_0 = V^* - V^{\pi_0}$ is the sub-optimality gap of policy $\pi_0$.

Theorem 1 provides a general framework deriving lower-bounds for conservative exploration and highlights the impact of the baseline policy on the regret. In particular, it shows that in any sequential decision-making problem, after a sufficiently large number of episodes the conservative condition can be verified and the baseline policy has no impact anymore on the learning process. The only requirement is the knowledge of a lower-bound for the non-conservative case. Before instantiating the result in specific settings, we provide an intuition about how this result is derived and what is the role of the conservative budget $B$.

Sketch of proof. Let’s consider a conservative algorithm $\mathfrak{A} = \{\pi_t \mid t \leq T\}$, which is associated to a non-conservative algorithm $\mathfrak{A} = \{\pi_t \mid t \in T_0\}$ with $T_0$ and $\mathfrak{A}$ the set of non-conservative and conservative episodes as defined in Sec. 4. By the reasoning of the previous section, we have that $|T_0|/\alpha V^{\pi_0} \geq B_T(T_0^c, \mathfrak{A})$ with high-probability. But the regret of $\mathfrak{A}$ can be decomposed as:

$$R_T(\mathfrak{P}, \mathfrak{A}) = \sum_{t \in T_0^c} V^* - V^{\pi_t} + |T_0| \geq \sum_{t \in T_0^c} V^* - V^{\pi_t} + \frac{\Delta_0}{\alpha V^{\pi_0}} B_T(T_0^c, \mathfrak{A}).$$

However, the right-hand side of Eq. (6) is exactly the regret of an algorithm which would play the baseline policy for the first $\frac{B_T(T_0^c, \mathfrak{A})}{\alpha V^{\pi_0}}$ steps and then chose the policies of $\mathfrak{A}$. Of course, this algorithm can not be deployed since it depends on the budget (which is unknown) and also depends on the realization of the algorithm $\mathfrak{A}$. However, using the assumption on the regret lower bound for non-conservative algorithm in the problem $\mathfrak{P}$, we show that the budget can be lower-bounded yielding the result thanks to Eq. (6).

Example of Lower Bounds. Theorem 1 is a powerful result as it allows us to instantiate a regret lower bound for conservative exploration for the different settings of Section 3 (and many more). For instance, in the multi-armed bandits, by leveraging the lower-bound in [Thm. 15.2 Lattimore and Szepesvári, 2020], we can obtain the following corollary of Theorem 1.

Corollary 1. For any $K \in \mathbb{N}^+$, $\alpha \in [0, 1]$, $\mu_0 \in [0, 1]$, $\delta \in (0, 1)$ and a conservative algorithm $\mathfrak{A}$ then there exists $\mu \in [0, 1]^K$ such that $\sum_{i=1}^{K} \mu_{\pi_i} \geq (1 - \alpha) \mu_0 t$ with high probability for any $t \leq T$.

Then, for $T \geq \frac{A}{\alpha \mu_0 + \Delta_0} + \frac{\sqrt{T}}{\alpha \mu_0 + \Delta_0}$

$$R_T(\mu, \mathfrak{A}) \geq \max \left\{ \frac{\sqrt{AT}}{\alpha \mu_0 \cdot (\alpha \mu_0 + \Delta_0) \bigg\} \right.$$
less tight, by instantiating it in the “hard” instance of Wu et al. [2016, Thm. 9], we recover their lower-bound, since \( \frac{\Delta_0}{\alpha V_\pi_0} + \Delta_0 \geq 0.9 \) has a constant effect. However, the explicit dependence on the sub-optimality gap \( \Delta_0 \) makes our lower bound more adaptive and more informative about the shape of regret we should aim at when constructing conservative algorithms.

The generality of Theorem 1 allows us to derive lower-bounds for conservative exploration in many different problems, where the lower-bound was unknown. Table 1 reports the lower-bound obtained through Theorem 1. Please refer to Appendix B for lower-bounds for non-conservative exploration.

In linear bandits, the lower bound we obtain matches the result in Kazerouni et al., 2016, Garcelon et al., 2020b, showing the optimality of their algorithms. In tabular MDPs, our result shows that the dependence on \( S, A \) and \( H \) of CUCBVI [Garcelon et al., 2020a] is not optimal. Finally, by instantiating Theorem 1 in low-rank MDPs, we obtain the first lower bound for this setting.

6 Upper Bounds

In this section, we show how to leverage the framework of Sec. 4 to derive an algorithm for any conservative sequential decision-making problem. We first show that when knowing \( \Delta_0 \) a simple algorithm achieves a minimax regret, as prescribed by our lower bound of Sec. 5. Then, we show how to remove this knowledge without hurting the performance by combining our framework and the idea of lower confidence bound. This allows to dynamically adapt the budget and overcome the need of knowing \( \Delta_0 \). If the non-conservative algorithm is minimax optimal, then also the conservative algorithm is minimax optimal.

6.1 The BudgetFirst Algorithm

Given a non-conservative algorithm \( \tilde{\pi} \), 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 \( \tilde{\pi} \), for any set \( O \subset [T] \) the budget can be rewritten as:

\[
B_T(O, \{\tilde{\pi}_l \mid l \leq T\}) = \max_{t \in O} \sum_{l \in O \cap \llbracket t \rrbracket} (1 - \alpha) V^\pi_0 - V^\pi_l = \max_{t \in O} \sum_{l \in O \cap \llbracket t \rrbracket} \left( V^* - V^\pi_l - (\Delta_0 + \alpha V^\pi_0) |O \cap \llbracket t \rrbracket| \right).
\]

Let’s define \( R_{O \cap \llbracket t \rrbracket}(\tilde{\pi}) := \sum_{l \in O \cap \llbracket t \rrbracket} V^* - V^\pi_l \) the regret over the time steps in \( O \) of the non-conservative algorithm \( \tilde{\pi} \). For most non-conservative algorithms with minimax regret bound, \( R_T(\tilde{\pi}, O) = O(C \sqrt{|O \cap \llbracket t \rrbracket|}) \) w.h.p., where \( C \in \mathbb{R} \) is a problem-dependent quantity as in Theorem 1. For example, in multi-armed bandit \( C = \sqrt{A} \) for the UCB algorithm or \( C = \sqrt{H^3SA} \) for the UCBVI-BF algorithm [Azar et al., 2017]. This implies that the budget required by \( \tilde{\pi} \) is at least \( \frac{C^2}{\Delta_0 + \alpha V^\pi_0} \). Therefore, the simple algorithm playing the baseline policy for the first \( T_0 := O\left( \frac{C^2}{\Delta_0 + \alpha V^\pi_0} \right) \) steps and then running the non-conservative algorithm \( \tilde{\pi} \), is conservative. We call such algorithm BudgetFirst (see Alg. 1). This algorithm is not only conservative but it is also minimax optimal. Indeed, we can show (see Theorem 2) that the regret upper bound of BudgetFirst matches the lower bounds of Section 5. While knowing \( \Delta_0 \) in advance may be a
restrictive assumption, it is interesting that a two-stage algorithm structure (deploying a baseline policy and then a non-conservative policy) is enough to achieve minimax optimality.

**Algorithm 1: BudgetFirst**

**Input:** A non-conservative algorithm $\mathcal{A}$, conservative policy cumulative reward $V^{\pi_0}$, conservative level: $\alpha \in (0, 1)$, baseline action gap: $\Delta_0 = V^* - V^{\pi_0}$, and a constant $C$

1. Set $B = \frac{C^2}{\alpha V^{\pi_0} + \Delta_0}$ and $T_0 = \frac{B}{\alpha V^{\pi_0}}$.
2. for $t = 1, \ldots, T$
   3. if $t < T_0$ then
      4. Play $\pi_0$;
   5. else
      6. Play according to $\mathcal{A}$;
   7. end
8. end

**Theorem 2.** Consider an algorithm $\mathcal{A}$, $\delta \in (0, 1)$ and constant $C \in \mathbb{R}$ such that with probability at least $1 - \delta$, for any $T \geq 1$, $R_T(\mathcal{A}) \leq \tilde{O}(C \sqrt{T})$. Then for any $T \geq 1$, the regret of BudgetFirst is bounded with probability at least $1 - \delta$ by $\tilde{O}(C \sqrt{T} + \frac{C^2 \Delta_0}{\alpha V^{\pi_0} + \Delta_0})$.

Instantiating Thm. 1 with $\mathcal{A}$ being the UCB algorithm [Lattimore and Szepesvári, 2020], then $C = \sqrt{A}$ and the regret of BudgetFirst is bounded w.h.p. by $\tilde{O}(\sqrt{AT} + \frac{\Delta_0}{\alpha V^{\pi_0} + \Delta_0})$, that matches our novel lower bound introduced in Sec. 5. Similar results can be obtained for the other settings, see Table 1. In linear bandits we consider LinUCB as the non-conservative algorithm, leading to $C = d$. Similarly, in tabular MDP and low-ran MDPs, we get $C = \sqrt{H^3SA}$ and $C = \sqrt{d^2H^2}$ respectively using UCBVI-BF [Azar et al., 2017] and LSVI-UCB [Jin et al., 2020]. Refer to Table 1 for a complete comparison of the results.

### 6.2 The LCBCE Algorithm

When $\Delta_0$ is unknown, we aim to use the same idea as BudgetFirst, that is to say to play a policy different than the baseline one only if the budget is positive. To achieve this, we need to build an online estimate of the conservative budget which amounts to build a lower confidence bounds (w.h.p.) on the value function of any policy $\pi$. Therefore, assuming a non-conservative algorithm $\mathcal{A}$ builds such confidence bounds, for example by estimate the MDP as done by Garcelon et al. [2020a], we show how our budget framework helps to derive a conservative regret bound.

Let’s consider a non-conservative algorithm $\mathcal{A} = \{\pi_t \mid t \leq T\}$ able to construct a high probability lower bound on the set of selected policies. That is to say, for any time $t \leq T$ and $\delta \in (0, 1)$, $\mathcal{A}$ computes a sequence of real numbers $(\lambda_t^\pi(\delta))_{k \leq t}$ such that with probability at least $1 - \delta$, for all $k \leq t$, $\lambda_t^\pi(\delta) \leq V^{\pi_k}$. Using, those lower bounds, we can define a proxy to the budget for $B_{T, \delta}(O, \mathcal{A})$ for any subset $O \subseteq [T]$ by

$$B_{T, \delta}(O, \mathcal{A}) = \max_{\pi_t \in O} \sum_{i \in O \cap [t]} ((1 - \alpha)V^{\pi_i} - \lambda_t^\pi(\delta)), \quad (7)$$

with $(\pi_t)_{t \in O}$ the sequence of policies computed by the non-conservative algorithm $\mathcal{A}$. Then following from the definition of $(\lambda_t^\pi(\delta))_{k \leq t}$, we have that with probability at least $1 - \delta$ that $B_{T, \delta}(O, \mathcal{A}) \geq B_T(O, \mathcal{A})$. This shows that it is possible to compute $B_{T, \delta}(O, \mathcal{A})$ without knowledge of the environment and the baseline parameters. The idea of our algorithm is now to play a non-conservative policy $\pi_t$ at time $t$ only if the difference between the proxy to the budget of $\mathcal{A}$ and the reward accumulated by playing the baseline policy is negative. Formally, the condition is $B_{t, \delta}(S_t \cup t, \mathcal{A}) \leq \alpha V^{\pi_t}(t - 1 - |S_t|)$ where $S_t$ is the set of time step where a non-conservative policy was deployed in episodes before $t$. As a result, the minimum budget that $\mathcal{A}$ requires to be conservative is $\max_t B_{t, \delta}(S_t \cup t, \mathcal{A}) = \max_{\pi_t \in [T]} \sum_{i \in S_t} ((1 - \alpha)V^{\pi_i} - \lambda_t^\pi(\delta))$. The algorithm, called **Lower Confidence Bound for Conservative Exploration (LCBCE)**, is detailed in Alg. 2.
Algorithm 2: Lower Confidence Bound for Conservative Exploration

\textbf{Input:} A non-conservative algorithm \( \tilde{A} \), \( \delta \in (0, 1) \), lower confidence bounds \( \lambda_t^{\pi_k} \leq V^{\pi_k} \), conservative policy value \( V^{\pi_0} \), \( \alpha \in (0, 1) \)

1. Set \( B = 0 \);  // the accumulated budget
2. Set \( t' = 0 \);  // the number of steps in which the agent acts as \( \hat{L} \)
3. for \( t = 1, 2, ..., T \) do  
   4. \( \bar{A} \) gives lower bound \( \lambda_{t+1} \) and a policy \( \bar{\pi}_{t+1} \);
   5. Compute \( \lambda = \sum_{k=1}^{t'} \lambda_{t+1}^{\pi_k} + \bar{\pi}_{t+1}^{\pi_k} \); // lower bound of expected total reward
   6. if \( \lambda - (t' + 1)\alpha V^{\pi_0} < B \) then
      7. Play \( \pi_t = \pi_0 \);
      8. \( B = B + \alpha V^{\pi_0} \);
   else
      9. Play \( \pi_t = \bar{\pi}_{t+1} \);
   10. \( t' = t' + 1 \);
end

Next, we show that when LCBCE achieves the same regret bound of the non-conservative algorithm it is paired with.

**Theorem 3.** Consider an algorithm \( \bar{A} \), \( \delta \in (0, 1) \) and constant \( C \in \mathbb{R} \) such that with probability at least \( 1 - \delta \), for any \( T \geq 1 \), \( R_T(\bar{A}) \leq \tilde{O}(C \sqrt{T}) \). If \( \bar{A} \) computes lower confidence bound such that \( \sum_{k=1}^{t'} (V^{\pi_k} - \lambda_t^{\pi_k}) \leq \tilde{O}(C \sqrt{T}) \) with probability at least \( 1 - \delta \), then for any \( T \geq 1 \), the regret of LCBCE is bounded with probability at least \( 1 - \delta \) by \( \tilde{O}(C \sqrt{T} + \frac{C^2 \Delta_0^2}{\alpha V^{\pi_0}(\alpha V^{\pi_0} + \Delta_0)}) \).

In the MAB case, LCBCE paired with UCB is similar to Alg. 1 in [Wu et al., 2016] but achieves a better regret bound. The same observation holds for tabular RL with UCBVI-BF compared to CUCBVI of [Garcelon et al., 2020a]. In particular, we provide the first minimax optimal bound also for the case of unknown baseline parameters. For linear bandit, using LinUCB leads to the same result in [Kazerouni et al., 2016, Garcelon et al., 2020b], that are already minimax optimal. Finally, in low rank MDPs we recover the same rate, up to constant term, as in the case of known baseline. See Table 1 for a complete comparison. The improvements are due the use of the budget (as defined in Sec. 4). Indeed, we have that, the regret of LinUCB can be written as:

\[
R_T(\bar{A}) \lesssim R_S(\bar{A}) + \max_t \frac{\tilde{B}_{t, \delta}(S_t, \bar{A}) \Delta_0}{\alpha V^{\pi_0}}.
\]

Therefore, to bound the regret it is sufficient to bound the proxy of the budget. Under the assumption in Thm. 3, we have that:

\[
\max_t \tilde{B}_{t, \delta}(S_t, \bar{A}) = \max_t \sum_{t \leq T} (1 - \alpha) V^{\pi_0} - V^{\pi_k} + \sum_{t \leq T} V^{\pi_k} - \lambda_t^{\pi_k} \leq \max_t 2C \sqrt{t} - (\Delta_0 + \alpha V^{\pi_0}) t = \tilde{O}\left(\frac{C^2}{\alpha V^{\pi_0} + \Delta_0}\right).
\]

Hence, this leads to the result in Thm. 3.

7 Conclusion

We have presented a unified framework for conservative exploration in sequential decision-making problems. This framework can be leveraged to derive both minimax lower and upper bounds. In bandits, we provided novel lower bounds that highlighted the optimality of existing algorithms. In RL, we provided the first lower bound for tabular MDPs and a matching upper bounds. Finally, we provided the first analysis for low rank MDPs. An interesting question raised by our paper is whether is it possible to leverage this framework to derive problem-dependent logarithmic bounds for conservative exploration. Another direction could be to investigate the usage of model-free algorithms (e.g., Q-learning [Jin et al., 2018]) for conservative exploration.
References

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. In *NeurIPS*, pages 9252–9262, 2019.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Dongsheng Ding, Kaijing Zhang, Tamer Basar, and Mihailo R. Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. In *NeurIPS*, 2020.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo R. Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pages 3304–3312. PMLR, 2021.

Yihan Du, Siwei Wang, and Longbo Huang. A one-size-fits-all solution to conservative bandit problems. *CoRR*, abs/2012.07341, 2020.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *CoRR*, abs/2003.02189, 2020.

Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Conservative exploration in reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1431–1441. PMLR, 2020a.

Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Improved algorithms for conservative exploration in bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3962–3969, 2020b.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is q-learning provably efficient? In *NeurIPS*, pages 4868–4878, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Sumeet Katariya, Branislav Kveton, Zheng Wen, and Vamsi K. Potluru. Conservative exploration using interleaving. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pages 954–963. PMLR, 2019.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. *arXiv preprint arXiv:1611.06426*, 2016.

Romain Laroche, Paul Trichet, and Remi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 3652–3661. PMLR, 2019.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Pierre Ménard, Omar Darwiche Domingues, Xuendong Shang, and Michal Valko. UCB momentum q-learning: Correcting the bias without forgetting. *CoRR*, abs/2103.01312, 2021.

Aldo Pacchiano, Mohammad Ghavamzadeh, Peter L. Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pages 2827–2835. PMLR, 2021.

Steven E Shreve and Dimitri P Bertsekas. Alternative theoretical frameworks for finite horizon discrete-time stochastic optimal control. *SIAM Journal on control and optimization*, 16(6):953–978, 1978.
Thiago D. Simão and Matthijs T. J. Spaan. Safe policy improvement with baseline bootstrapping in factored environments. In AAAI, pages 4967–4974. AAAI Press, 2019.

P. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence off-policy evaluation. In Proceedings of the Twenty-Ninth Conference on Artificial Intelligence, 2015a.

P. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence policy improvement. In Proceedings of the Thirty-Second International Conference on Machine Learning, pages 2380–2388, 2015b.

Kun Wang, Canzhe Zhao, Shuai Li, and Shuo Shao. Conservative contextual combinatorial cascading bandit. CoRR, abs/2104.08615, 2021.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In International Conference on Machine Learning, pages 1254–1262. PMLR, 2016.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In ICML, volume 97 of Proceedings of Machine Learning Research, pages 7304–7312. PMLR, 2019.

Andrea Zanette, Alessandro Lazaric, Mykel J. Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In ICML, volume 119 of Proceedings of Machine Learning Research, pages 10978–10989. PMLR, 2020.

Xiaojin Zhang, Shuai Li, and Weiwen Liu. Contextual combinatorial conservative bandits. CoRR, abs/1911.11337, 2019.

Zihan Zhang, Xiangyang Ji, and Simon S Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. arXiv preprint arXiv:2009.13503, 2020a.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. In NeurIPS, 2020b.

Liyuan Zheng and Lillian J. Ratliff. Constrained upper confidence reinforcement learning. In L4DC, volume 120 of Proceedings of Machine Learning Research, pages 620–629. PMLR, 2020.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvári. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. CoRR, abs/2012.08507, 2020.
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 \( \{0, \ldots, K\} \). There is a reward \( X_{t,i} \) associated with each arm \( i \) at each round \( t \in \{1, 2, \ldots\} \). In each round \( t \), the agent pulls arm \( I_t \in \{0, \ldots, K\} \) 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:

\[
R_T = \max_{i \in \{0, \ldots, K\}} \sum_{t=1}^{T} X_{t,i} - X_{t,I_t} \tag{8}
\]

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 \( \eta_i \) is a random noise such that

**Assumption 2.** Each element \( \eta_i \) of the noise sequence \( \{\eta_i\}_{i=1}^\infty \) is conditionally 1-sub-Gaussian, i.e.

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

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

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:

\[
\sum_{s=1}^{t} X_{s,0} \geq (1 - \alpha) \sum_{s=1}^{t} X_{s,0} \quad \text{for all } t \in \{1, \ldots, T\} \tag{10}
\]

where \( \alpha \in (0, 1) \) is a predefined constant. The parameter \( \alpha \) controls how conservative the agent should be. Small values of \( \alpha \) show that only small losses are tolerated, and thus, the agent should be overly conservative, whereas large values of \( \alpha \) 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 \( A \), where each action \( a \in A \) is associated with a feature vector \( \phi_a \in \mathbb{R}^d \). At each round \( t \), the agent should select an action \( a_t \in A \). Upon selecting \( a_t \), the agent observes a random reward \( X_t \) generated as

\[
X_{t,a_t} = \langle \theta^*, \phi_{a_t} \rangle + \eta_t, \quad \tag{11}
\]

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 \( \eta_t \) is a random noise that satisfies Assumption 2.

We also make the following standard assumption on the unknown parameter \( \theta^* \) and feature vectors:
We consider conservative exploration in finite horizon tabular MDPs. An MDP can be represented as $M = (S, A, H, p, r)$, where $S$ is the state space, $A$ 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 $[0, r_{\text{max}}]$, and a transition distribution $p(\cdot | s, a)$ over next states. We denote by $S = |S|$ and $A = |A|$. 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 receives 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 $\pi$ is a set of functions $\{\pi_h : S \mapsto \Delta(A)\}_{h \in [H]}$. Given a policy $\pi$, a level $h \in [H]$ and a state-action pair $(s, a) \in S \times A$, the $Q$ function and the value function are defined as:

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

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

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

In conservative tabular MDPs, at the beginning of each episode $t$, the agent can choose to run a conservative policy $\pi_0$, 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 $\pi_h$, and will receive a total reward $V_1^{\pi_\tau}$. Our goal is to minimize the following regret

$$R_T = \sum_{t=1}^T V_t^\star(s_1) - V_t^{\pi_\tau}(s_1)$$

while satisfying the following conservative constraint

$$\sum_{j=1}^t V_1^{\pi_j}(s_1) \geq (1 - \alpha) t V_1^{\pi_0}(s_1), \quad \forall t \in [T].$$
A.4 Conservative Linear MDPs

The conservative linear MDP setting is nearly the same as tabular MDPs, except that $S$ is a measurable space with possibly infinite number of elements and $A$ 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 $(S, A, H, p, r)$ is a linear MDP with a feature map $\phi: S \times A \rightarrow \mathbb{R}^d$, if for any $h \in [H]$, there exist $d$ unknown (signed) measures $\mu_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})$ over $S$ and an unknown vector $\theta_h \in \mathbb{R}^d$, such that for any $(x, a) \in S \times A$, we have

$$P_h(\cdot | x, a) = \langle \phi(x, a), \mu_h(\cdot) \rangle, \quad r_h(x, a) = \langle \phi(x, a), \theta_h \rangle.$$  \hspace{1cm} (16)

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

$$E[R_T] \gtrsim \sqrt{KT}.$$  \hspace{1cm} (17)

*Proof.* See Theorem 15.2 of Lattimore and Szepesvári [2020] for a detailed proof. \hfill \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

$$E[R_T] \gtrsim d\sqrt{T}.$$  \hspace{1cm} (18)

*Proof.* See Theorem 24.2 of Lattimore and Szepesvári [2020] for a detailed proof. \hfill \square

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

$$E[R_T] \gtrsim \sqrt{S AH^3 T}.$$  \hspace{1cm} (19)

*Proof.* See Jaksch et al. [2010], Azar et al. [2017], Jin et al. [2018] for a detailed proof. \hfill \square

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

$$E[R_T] \gtrsim \sqrt{d^2 H^3 T}.$$  \hspace{1cm} (20)

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

C Detailed Proof for Lower Bounds

*Proof of Theorem 1.* Let’s consider any sequential decision making problem $\mathcal{A}$ (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 like the dimension in linear problems or the number of action in tabular problems), an instance of problem $\mathcal{A}$ where for a number of time steps $T$ large enough and any algorithm $\mathcal{A}$ we have that:

$$E[R^*_T(\mathcal{A})] \geq \xi \sqrt{T},$$  \hspace{1cm} (21)

with $R^*_T(\mathcal{A})$ the regret of algorithm $\mathcal{A}$ in problem $\mathcal{A}$ and the expectation is . 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 $\mathcal{A}$ with a baseline policy $\pi_0$. 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 algorithm selects policies \((\pi_t)_{t \in [T]}\) and let \(T_0\) denotes the set of rounds in \([1, \ldots, T]\) where \(A_c\) selects the conservative policy \(\pi_0\). Here \(T \geq \frac{\xi^2}{\alpha V^{\pi_0} - (\alpha V^{\pi_0} + \Delta_0)} + \frac{\xi^2}{4(\alpha V^{\pi_0} + \Delta_0)^2}\). We now distinguish two cases:

- If \(\mathbb{E}|T_0| \geq \frac{\xi^2}{\alpha V^{\pi_0} - (\alpha V^{\pi_0} + \Delta_0)}\), then the definition of the regret implies that:

\[
\mathbb{E}[R^T_{\mathcal{A}}(A_c)] \geq \mathbb{E}_t \sum_{i \in T_0} [V^* - V^{\pi_t}] = \mathbb{E}|T_0| \cdot \Delta_0 \geq \frac{\xi^2 \Delta_0}{\alpha V^{\pi_0} - (\alpha V^{\pi_0} + \Delta_0)}, \tag{18}
\]

- If \(\mathbb{E}|T_0| < \frac{\xi^2}{\alpha V^{\pi_0} - (\alpha V^{\pi_0} + \Delta_0)}\), then let’s note \(T_0^- = \{i_1, i_2, \ldots, i_{|T_0^-|}\}\) the set of time steps where \(A_c\) does not execute the conservative policy \(\pi_0\). Considering the budget as we have defined in Def. 1 we have:

\[
B_{T_0^-}(A_c) = \max_{t \in T_0^-} \mathbb{E} \sum_{k=1}^{t} [(1 - \alpha)V^{\pi_k} - V^{\pi_t}]
= \max_{t \in T_0^-} \mathbb{E} \sum_{k=1}^{t} [V^* - V^{\pi_k} - \alpha V^{\pi_k} - (V^* - V^{\pi_0})]
= \max_{t \in T_0^-} \mathbb{E}[R^T_{\mathcal{A}}(A_c)(t)] - (\alpha V^{\pi_0} + \Delta_0) t,
\]

where \(\Delta_0 = V^* - V^{\pi_0}\) is the difference between the optimal policy and the baseline policy and \(\mathbb{E}[R^T_{\mathcal{A}}(A_c)(t)]\) is the regret incurred by the rounds \(\{i_k\}_{k \in [t]}\). Therefore, for any \(t \in [|T_0^-|]\), 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 \(\mathbb{E}[R^T_{\mathcal{A}}(A_c)(t)] \geq \xi \sqrt{T}\). Let \(t_0 = \frac{\xi^2}{4(\alpha V^{\pi_0} + \Delta_0)^2}\), then there exists an instance such that

\[
B_{T_0^-}(A_c) \geq \xi \sqrt{t_0} - (\alpha V^{\pi_0} + \Delta_0) t_0 \geq \frac{\xi^2 \Delta_0}{\alpha V^{\pi_0} + \Delta_0}. \tag{20}
\]

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

\[
\mathbb{E}|T_0| \geq B_{T_0^-}(A_c) \geq \frac{\xi^2}{\alpha V^{\pi_0} - (\alpha V^{\pi_0} + \Delta_0)}.
\]

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

\[
\mathbb{E}[R^T_{\mathcal{A}}(A_c)] \geq \frac{\xi^2 \Delta_0}{\alpha V^{\pi_0} - (\alpha V^{\pi_0} + \Delta_0)}. \tag{21}
\]

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

\[
\mathbb{E}[R^T_{\mathcal{A}}(A)] \geq \max \left\{ \xi \sqrt{T}, \frac{\xi^2 \Delta_0}{\alpha V^{\pi_0} - (\alpha V^{\pi_0} + \Delta_0)} \right\}. \tag{22}
\]

Then we discuss the lower bound for different setups:

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

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

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

\[
\mathbb{E}[R_T] \geq \max \left\{ d \sqrt{T}, \frac{d^2 \Delta_0}{\alpha V^{\pi_0} - (\alpha V^{\pi_0} + \Delta_0)} \right\}.
\]
• For tabular RL, by Lemma 6, we choose $\xi = \sqrt{SAH^3}$. Then we have
\[
\mathbb{E}[R_T] \geq \max \left\{ \sqrt{SAH^3 T}, \frac{SAH^3 \Delta_0}{\alpha V^\pi_0 \cdot (\alpha V^\pi_0 + \Delta_0)} \right\}.
\]

• For low-rank MDP, by Lemma 7, we choose $\xi = \sqrt{d^2 H^3}$. Then we have
\[
\mathbb{E}[R_T] \geq \max \left\{ \sqrt{d^2 H^3 T}, \frac{d^2 H^3 \Delta_0}{\alpha V^\pi_0 \cdot (\alpha V^\pi_0 + \Delta_0)} \right\}.
\]

Therefore, we conclude the proof. \qed

D Detailed Proof for Upper Bounds

D.1 Proof of Theorem 2

Proof. Given a non-conservative algorithm $\hat{\pi}$, 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 $\{\hat{\pi}_t \mid t \leq T\}$ the sequence of non-conservative policies executed by $\hat{\pi}$, then for any set $O \subset [T]$ the budget can be rewritten as:
\[
\mathcal{B}_T(O, \{\hat{\pi}_t \mid t \leq T\}) = \max_{t \in O} \sum_{t \in O \cap [t]} (1 - \alpha) V^{\pi_0} - V^{\pi_t}
\]
\[
= \max_{t \in O} \sum_{t \in O \cap [t]} \left( V^* - V^{\pi_t} - (\Delta_0 + \alpha V^{\pi_0}) |O \cap [t]| \right).
\]

Let’s define $R_{O \cap [t]}(\hat{\pi}) := \sum_{t \in O \cap [t]} V^* - V^{\pi_t}$ the regret over the time steps in $O$ of the non-conservative algorithm $\hat{\pi}$. Since $R_t(\hat{\pi}, O) = O(C \sqrt{|O \cap [t]|})$ w.h.p., where $C \in \mathbb{R}$ is a problem-dependent quantity as in Theorem 1. Therefore, we have
\[
\mathcal{B}_T(O, \{\hat{\pi}_t \mid t \leq T\}) = \max_{t \in O} \sum_{t \in O \cap [t]} (1 - \alpha) V^{\pi_0} - V^{\pi_t}
\]
\[
= \max_{t \in O} \sum_{t \in O \cap [t]} \left( O(C \sqrt{|O \cap [t]|}) - (\Delta_0 + \alpha V^{\pi_0}) |O \cap [t]| \right).
\]

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

The regret of BudgetFirst can be bounded as
\[
Reg(T) \leq t_0 + R_t(\hat{\pi}, O) = O\left(\frac{\xi}{\alpha V^{\pi_0} + \Delta_0} + C \sqrt{|O \cap [t]|}\right).
\]

Thus we finish the proof. \qed

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

Lemma 8 (Upper Bound for Multi-Armed Bandit). The regret of UCB can be upper bounded by
\[
R_T \leq 8\sqrt{T k \log(T)} + 3 \sum_{i=1}^k \Delta_i
\]  

Proof. See Theorem 7.2 in Lattimore and Szepesvári [2020] for details. \qed
For linear bandits, the LinUCB algorithm [Lattimore and Szepesvári, 2020] gives us the following guarantee.

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

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

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

**Proof.** See Corollary 19.3 in Lattimore and Szepesvári [2020] for details. \hfill \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

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

**Proof.** See Azar et al. [2017] for details. \hfill \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

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

**Proof.** See Jin et al. [2020] for details. \hfill \square

**D.2 Proof of Theorem 3**

**Proof.** Given an LCB algorithm $\tilde{A}$, suppose it maintains lower confidence bound $\lambda^\pi_k(\delta) \leq V^\pi_k$ with probability at least $1 - \delta$ that satisfies $\sum_{k=1}^t (V^\pi_k - \lambda^\pi_k) \leq \tilde{O}(C\sqrt{T})$. Let $S_t$ 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:

$$B_T(S_T; \tilde{A}) = \max_{t \in [T]} \sum_{t' \in S_t} [(1 - \alpha) V^{\pi_0} - \lambda^\pi_{t'}(\delta)]$$

$$= \max_{t \in [T]} \sum_{t' \in S_t} \left( V^* - V^{\pi_1} + V^{x_1} - \lambda^\pi_{t'}(\delta) \right) - (\Delta_0 + \alpha V^{\pi_0}) |S_t|$$

$$\leq \max_{t \in [T]} \sum_{t' \in S_t} \left( V^* - V^{x_1} \right) + \tilde{O}(C \sqrt{|S_t|}) - (\Delta_0 + \alpha V^{\pi_0}) |S_t|$$

$$= \max_{t \in [T]} R_{S_t}(\tilde{A}) + \tilde{O}(C \sqrt{|S_t|}) - (\Delta_0 + \alpha V^{\pi_0}) |S_t|$$

Note that $R_{S_t}(\tilde{A}) \leq \tilde{O}(C \sqrt{|S_t|})$, 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 $g(x) \leq \tilde{O}(\frac{C^2}{\Delta_0 + \alpha V^{\pi_0}})$ as a result. In other words, we show that with high probability, LCBCE only need to accumulate $B_T(S_T; \tilde{A}) \leq \tilde{O}(\frac{C^2}{\Delta_0 + \alpha V^{\pi_0}})$. Since playing the baseline policy yields $\alpha V^{\pi_0}$ budget, LCBCE play the baseline policy for at most $\tilde{O}(\frac{C^2}{\Delta_0 + \alpha V^{\pi_0}})$ times. Hence, the total regret incurred can be written as:

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

Thus we finish the proof. \hfill \square
Below we discuss the lower confidence bound for different setups. For the MAB setting, we can calculate the lower confidence bound simultaneously with the upper confidence bound as

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

(27)

where $$\psi^\delta(s) = 2 \log \left( K s^3 / \delta \right)$$ 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 and Szepesvári [2020].

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

$$(a_t', \tilde{\theta}_t) \in \arg \max_{(a, \theta) \in A_k \times C_t} \langle \theta, \phi_a^t \rangle$$

(28)

where $$C_{t+1}$$ is the confidence set $$C_{t+1} = \left\{ \theta \in \mathbb{R}^d : \| \theta - \tilde{\theta}_t \|_{V_t} \leq \beta_{t+1} \right\}$$. Then, we calculate $$L_t = \min_{\theta \in C_t} \langle \theta, \tilde{z}_{t-1} + \phi_{a_t'} \rangle$$, where $$\tilde{z}_{t-1} = \sum_{i=1}^{t-1} \phi_{a_i}$$. Then $$L_t$$ is a lower confidence bound of action $$a_t'$$. For tabular MDP setting, the upper bound of the $$Q$$ function can be calculated as $$Q_h(s, a) = r(s, a) + \hat{P}_h V_{h+1}(x, a) + b_h(s, a)$$, where the bonus function is chosen to be $$b_h = \hat{O}(\sqrt{\sum_{k=1}^{H} \beta_k / K})$$ 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,

$$V_{h+1}(\cdot) = \max_{a} Q_{h+1}^l(\cdot, a)$$

$$Q_{h}^l(\cdot, \cdot) = r(\cdot, \cdot) + \hat{P}_h V_{h+1}^l(\cdot, \cdot) - b_h(\cdot, \cdot)$$,

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

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]:

$$A_h \leftarrow \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau) \phi(x_h^\tau, a_h^\tau)^T + \lambda I$$

$$w_h \leftarrow A_h^{-1} \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau) [r_h(x_h^\tau, a_h^\tau) + \max_{a} Q_{h+1}(x_{h+1}, a)]$$

$$Q_h(\cdot, \cdot) \leftarrow \max_{a} \{ w_h^T \phi(\cdot, \cdot) - \beta (\phi(\cdot, \cdot)^T A_h^{-1} \phi(\cdot, \cdot))^{1/2}, 0 \}$$

$$V_h(\cdot) \leftarrow \max_{a} Q_h(\cdot, a)$$

We note that for all these settings, we have \( \sum_{k=1}^{T} (V^\pi_k - Q^k_\pi) \leq \hat{O}(C \sqrt{T}) \) 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

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

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

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