# Confidence-Budget Matching for Sequential Budgeted Learning

**Yonathan Efroni**[*1] **Nadav Merlis**[*2] **Aadirupa Saha**[1] **Shie Mannor**[2 3]

## Abstract

A core element in decision-making under uncertainty is the feedback on the quality of the performed actions. However, in many applications, such feedback is restricted. For example, in recommendation systems, repeatedly asking the user to provide feedback on the quality of recommendations will annoy them. In this work, we formalize decision-making problems with querying budget, where there is a (possibly time-dependent) hard limit on the number of reward queries allowed. Specifically, we consider multi-armed bandits, linear bandits, and reinforcement learning problems. We start by analyzing the performance of 'greedy' algorithms that query a reward whenever they can. We show that in fully stochastic settings, doing so performs surprisingly well, but in the presence of any adversity, this might lead to linear regret. To overcome this issue, we propose the Confidence-Budget Matching (CBM) principle that queries rewards when the confidence intervals are wider than the inverse square root of the available budget. We analyze the performance of CBM based algorithms in different settings and show that they perform well in the presence of adversity in the contexts, initial states, and budgets.

## 1. Introduction

In the past few decades, there have been great advances in the field of sequential decision making under uncertainty. From a practical perspective, recent algorithms achieve superhuman performance in problems that had been considered unsolvable (Mnih et al., 2015; Silver et al., 2017). From a theoretical perspective, algorithms with order-optimal performance were presented to various important settings (Garivier & Cappé, 2011; Azar et al., 2017, and others).

To solve such problems, most works share the same abstract interaction model. At each round, an agent (i) observes some information on the state of the environment, (ii) decides how to act, based on previous interactions, and, (iii) observes new feedback on the effect of its action. Finally, the environment changes its state based on the agent's action, and the cycle begins anew. Much effort had been devoted to study specific instances of this abstract model, e.g., multi-armed bandits (MABs) (Auer et al., 2002; Garivier & Cappé, 2011; Kaufmann et al., 2012; Agrawal & Goyal, 2012), linear bandits (Dani et al., 2008; Abbasi-Yadkori et al., 2011; Agrawal & Goyal, 2013; Abeille et al., 2017) and reinforcement learning (RL) settings (Azar et al., 2017; Jin et al., 2018; Dann et al., 2019; Zanette & Brunskill, 2019; Efroni et al., 2019; Simchowitz & Jamieson, 2019; Tarbouriech et al., 2020; Cohen et al., 2020; Zhang et al., 2020). However, there are (still) several gaps between theory and practice that hinder the application of these models in real-world problems.

One such evident gap is the need to act under a budget constraint that limits the amount of feedback from the environment. That is, receiving feedback on the quality of the agent's actions has an inherent cost. Consider, for example, an online recommendation system. There, asking for feedback from users negatively affects their experience, and feedback should be requested sparingly. Another example can be found in most large-scale RL domains, including autonomous driving. In many such cases, the reward should be labeled manually, and the resources for doing so are limited. Motivated by these problems, in this work, we aim to tackle the following question:

*How should an agent trade-off exploration and exploitation when the feedback is limited by a budget?*

In our efforts to answer this question, we study the effect of time-varying observation budget in various decision-making problems. Formally, we assume that at each round, the agent observes a non-decreasing, possibly adversarial, budget $B(t)$, which limits the number of queries for the reward of the problem. We first show that when the problem is stochastic and the budget is oblivious, greedily using any available budget leads to good performance. However, as soon as adversarial elements appear in the problem, or when the budget is controlled by an adaptive adversary, such an algorithm miserably fails. To tackle this problem, we suggest

[*]Equal contribution [1]Microsoft Research, New York [2]Technion, Israel [3]Nvidia Research, Israel. Correspondence to: Yonathan Efroni <jonathan.efroni@gmail.com>, Nadav Merlis <merlis.nadav@gmail.com>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. 

a simple, generic, scheme, that only samples rewards for actions with high uncertainty, in comparison to the budget. We call such a mechanism *confidence-budget matching* (CBM). We show how to apply CBM to MAB, linear bandit and RL problems. In all cases, the mechanism can be applied in the presence of adaptive adversarial budgets. For linear bandits and RL, we show that CBM can be applied even when the contexts and initial states are adversarial. Finally, we present lower bounds for MABs and linear bandits, which show that CBM leads to order-optimal regret bounds.

## 2. Preliminaries

We start by defining a general model for sequential decision-making under uncertainty. Then, we will explain its realization in each individual model. In the most general model, at each round $t$, the environment supplies the agent with a context $u_t$ that may either be stochastic or adversarially chosen. Then, the agent selects a policy $\pi_t \in \Pi(u_t)$ that can depend on $u_t$ and past observations. Finally, the environment generates two stochastic feedback variables, from fixed distributions conditioned on $u_t$ and $\pi_t$: feedback on the interaction with the environment $Z_t$ and reward feedback $R_t$. In RL, for example, $Z_t$ is the visited state-actions while $R_t$ is their respective rewards. We also assume that there exists a reward function $f$ such that the agent aims to maximize $f(R_t)$ throughout the interaction. Alternatively, algorithms aim to minimize its *pseudo-regret* (or regret), which is defined as

$$\mathrm{Reg}(T)=\sum_{t=1}^{T}\left(\max_{\pi\in\Pi(u_t)}\mathbb{E}[f(R_t)|u_t,\pi]-\mathbb{E}[f(R_t)|u_t,\pi_t]\right).$$

Note that the pseudo-regret is random, as the policy depend on random feedback from the environment and contexts might be stochastic. Thus, regret bounds for different algorithms hold either with expectation or with high probability.

To illustrate the generality of this model, we explain how it encompasses both MAB, linear bandit and RL problems:

**Contextual Multi-Armed Bandits (CMABs).** At the beginning of each round, a context $u_t \in \{1,\ldots,S\}$ is chosen, either stochastically or adversarially. Then, the agent chooses an action (arm) from a finite set of cardinality $A$, $\pi_t \triangleq a_t \in \mathcal{A}$ and the environment generates a reward $R_t \in [0,1]$ with an expectation $\mathbb{E}[R_t|u_t=u,a_t=a]=r(u,a)$. An optimal arm is denoted by $a^*(u) \in \arg\max_a r(u,a)$ and its value by $r^*(u)=\max_a r(u,a)$. The reward function is $f(R_t)=R_t$ and there is no additional feedback ($Z_t=\phi$). A specific case of interest is where a single context exists, which is the well known MAB problem. Then, we denote $r(a) \triangleq r(1,a)$.

**Linear Contextual Bandits.** In the stochastic setting, $u_t$ contains a set of $A$ vectors in $\mathbb{R}^d$, generated independently from a fixed distrubution. In the adversarial case, $u_t$ is an arbitrary set of vectors in $\mathbb{R}^d$. At each round $t$, the agent selects a single vector $\pi_t \triangleq x_t \in u_t$. Then, the environment generates a reward $R_t=\langle x_t,\theta\rangle+\eta_t$, where $\eta_t$ is zero-mean subgaussian noise and $\theta\in\mathbb{R}^d$ is unknown. As in the CMAB problem, the reward function is $f(R_t)=R_t$ and there are no additional observations ($Z_t=\phi$).

**Episodic Reinforcement Learning.** Let $\mathcal{S},\mathcal{A}$ be finite state and action sets with cardinalities of $S,A$, respectively. Before each episode $t$, an initial state $s_{t,1}$ is generated either stochastically or adversarially (and serves as a context $u_t$). Then, an agent selects a nonstationary policy $\pi_t:\mathcal{S}\times[H]\to\mathcal{A}$, for some $H\in\mathbb{N}$. The policy is evaluated for $H$ steps, and states are generated according to a transition kernel $P$; namely, for any $s'\in\mathcal{S}$ and $h\in\{1,\ldots,H\}$, $\Pr(s_{t,h+1}=s'|s_{t,h},\pi_{t,h}) = P_h(s'|s_{t,h},\pi_{t,h}(s_{t,h}))$. For brevity, we denote $a_{t,h}=\pi_{t,h}(s_{t,h})$. The agent observes the trajectory $Z_t=\{(s_{t,h},a_{t,h})\}_{h=1}^{H}$ and for each visited state, a reward $R_t=\{R_{t,h}\}_{h=1}^{H}\in[0,1]^H$ is generated such that $\mathbb{E}[R_{t,h}|s_{t,h}=s,a_{t,h}=a]=r(s,a)$. The reward function is then $f(R_t)=\sum_{h=1}^{H-1}R_{t,h}$.

**Sequential Budgeted Learning.** In most cases, it is natural to observe the effect of the policy on the environment; for example, it is reasonable to assume that the agent observes the visited states in RL, as it acts according to them. Thus, we assume that the agent always observes $Z_t$. On the other hand, many applications require specifically querying or labeling the reward. Then, oftentimes, such feedback is limited. Formally, let $\{B(t)\}_{t\geq1}$ be a non-negative budget sequence that might be adversarially chosen. We also assume that the budget is non-decreasing, that is, a budget that is given cannot be taken. At each round $t$, the agent observes $B(t)$ and selects whether to query $R_t$ or not, which we denote by $q_t=1$ and $q_t=0$, respectively. However, the agent can choose $q_t=1$ only if its budget was not exhausted. Throughout most of the paper, we assume that querying a reward incurs unit cost. Then, an agent can select $q_t=1$ only if $n^q_{t-1}\triangleq\sum_{k=1}^{t-1}1\{q_k=1\}\leq B(t)-1$. In some cases, we extend the cost to be action-dependent. Then, a reward can only be queried if

$$B^q(t-1)\triangleq\sum_{k=1}^{t-1}c(\pi_k)1\{q_k=1\}\leq B(t)-c(\pi_t).$$

Notice that when queries have unit costs, then $n^q_t=B^q(t)$. For the RL setting, we give access to more refined feedback from specific time steps, to avoid confusion we only discuss it in Section 5.3. In all cases, we allow $q_t$ to also depend on $Z_t$. Finally, and for ease of notations, we assume that the agent always observes $Y_t=R_t\cdot q_t$.

**General Notations** We let $\{F_t\}_{t\geq0}$ be a filtration, where $F_t$ is the $\sigma$-algebra that contains the random variables

$\{(u_k, \pi_k, Z_k, q_k, Y_k, B(k))_{k=0}^t, B(t+1), u_{t+1}\}$. In words, it contains the information on all *observed* rewards, actions, budget until the $t^{th}$ episode, the budget at the $(t+1)^{th}$ episode, and the context at the $(t+1)^{th}$ episode. We denote $[n] = \{1, \ldots, n\}$ for $n \in \mathbb{N}$ and also $x \vee 1 = \max\{x, 1\}$ for any $x \in \mathbb{R}$. We use $\mathcal{O}(X)$ and $\tilde{\mathcal{O}}(X)$ to refer to a quantity that depends on $X$ up to constants and poly-log and constant expressions in problem parameters, respectively. Lastly, $\lesssim, \gtrsim$ denote inequalities that hold up to poly-log and constant expressions in problem parameters.

## 3. Lower Bounds for Budgeted Problems

Before suggesting algorithms to the budgeted setting, it is of importance to understand how the new constraint affects the best-achievable regret. To this end, we study problem-independent lower bounds for budgeted MAB. By the end of the section, we also shortly discuss lower bounds for budgeted linear bandits. To derive the lower bounds, we require a more detailed description of the MAB model and additional notations. Moreover, we need to adapt the fundamental inequality of Garivier et al. (2019) to the case where the agent does not query all samples (Lemma 7). We refer the reader to Appendix A.1 for more details on the model and to Appendix A.2 for Lemma 7. Other proofs for this section can be found at Appendix A.3. Using Lemma 7, we can prove a lower bound for the following scenario in which $(i)$ sampling an arm requires a unit cost, $(ii)$ the budget constraint holds in expectation, and, $(iii)$ the budget is given to the learner at the initial interaction, i.e. $\forall t \in [T],\ B(t) = B$:

**Proposition 1.** *Let $T$ be some time horizon and let $\pi$ be some bandit strategy such that for any bandit instance, it holds that $\mathbb{E}[n_T^q] \leq B$. Then, for $A \geq 2$, there exists a bandit instance for which*

$$\mathbb{E}[\mathrm{Reg}(T)] \geq \frac{1}{140} \min\left\{T\sqrt{\frac{A}{B}}, T\right\} .$$

As expected, when the budget is linear ($B = T$), we get the standard $\Omega\left(\sqrt{AT}\right)$ lower bound. However, as we decrease the budget, the lower bound increases, up to the point of linear regret when the budget is not time-dependent. We also remark that the lower bound holds even if the budget constraint is only met *in expectation*. We will later present algorithms whose regret match these bounds, without *ever* violating the budget constraint. This implies that relaxing the budget requirement to hold in expectation cannot improve performance, from a worst-case perspective. Finally, note that when the budget is polynomial in $T$, .e.g., $B = T^\beta$, we get a lower bound of $\Omega\left(\sqrt{A}T^{1-\beta}\right)$. We will later prove upper bounds that match this rate.

Next, it is of interest to generalize the bound to the case of arm-dependent costs. In this case, we require a more subtle analysis that also costs in a $\sqrt{\log A}$ factor:

**Proposition 2.** *Let $T$ be the time horizon, and let $c(1), \ldots, c(A) \geq 0$ be arm-dependent querying costs. Also, let $\pi$ be some bandit strategy such that for any bandit instance, it holds that $\mathbb{E}[B^q(T)] \leq B$. Then, for $A \geq 2$, there exists a bandit instance for which*

$$\mathbb{E}[\mathrm{Reg}(T)] \geq \frac{1}{140} \min\left\{T\sqrt{\frac{\sum_{a=1}^A c(a)}{B(1+\log A)}}, T\right\} .$$

While both bounds deal with fixed budget, $B(t) = B$ for all rounds, one can easily reduce them to lower bounds for time-dependent budgets, by reducing the lower bound only at a logarithmic factor. This is done by lower bounding the regret by the bound of the 'worst-case' time horizon $\Omega\left(\max_{t\in[T]}\left\{\frac{t}{\sqrt{B(t)}}\right\}\right)$. We demonstrate how to do so for the case of arm-dependent costs in the following corollary:

**Corollary 3.** *Let $c(1), \ldots, c(A) \geq 0$ be arm-dependent querying costs and let $B(1), \ldots, B(T) > 0$ be an arbitrary non-decreasing budget sequence. Also, let $\pi$ be some bandit strategy such that for any bandit instance and any time index $t \in [T]$, it holds that $\mathbb{E}[B^q(t)] \leq B(t)$. Then, for any $A \geq 2$, there exists a bandit instance for which*

$$\mathbb{E}[\mathrm{Reg}(T)] \geq \frac{1}{140(1+\log T)} \sum_{t=1}^T \min\left\{\sqrt{\frac{\sum_{a=1}^A c(a)}{B(t)(1+\log A)}}, 1\right\} .$$

*Proof.* By Proposition 2, for any $t \in [T]$, there exists an instance such that

$$\mathbb{E}[\mathrm{Reg}(t)] \geq \frac{t}{140} \min\left\{\sqrt{\frac{\sum_{a=1}^A c(a)}{B(t)(1+\log A)}}, 1\right\} .$$

Let $t_m$ be the time index in which the r.h.s. is maximized and fix the bandit problem to the corresponding instance that leads to its lower bound. Then.

$$\mathbb{E}[\mathrm{Reg}(T)] \geq \mathbb{E}[\mathrm{Reg}(t_m)] = \max_{t\in[T]}\left\{\frac{t}{140} \min\left\{\sqrt{\frac{\sum_{a=1}^A c(a)}{B(t)(1+\log A)}}, 1\right\}\right\}.$$

Finally, by Hölder's inequality, if $x, y \in \mathbb{R}^T$ are such that $x_t, y_t \geq 0$ for all $t \in [T]$, then

$$\max_t x_t = \|x\|_\infty \geq \frac{\sum_{t=1}^T x_t y_t}{\|y\|_1} = \frac{\sum_{t=1}^T x_t y_t}{\sum_{t=1}^T y_t} .$$

Taking $x_t = \frac{t}{140}\min\left\{\sqrt{\frac{\sum_{a=1}^{A} c(a)}{B(t)(1+\log A)}}, 1\right\}$ and $y_t = \frac{1}{t}$ and recalling that $\sum_{t=1}^{T} \frac{1}{t} \leq 1+\log T$ concludes the proof. □

In the following sections, we derive regret upper bounds of similar budget-dependence, e.g., $\tilde{\mathcal{O}}\left(\sum_{t=1}^{T}\sqrt{A/B(t)}\right)$, if $c(a) = 1$ for all $a \in \mathcal{A}$ . It is therefore of interest to observe the behavior of such bounds as a function of different budget profiles.

**Example 1.** (Budget Profiles and Consequences).

- ***Linear Budget:*** *if $B(t) = \epsilon t$ for some $\epsilon > 0$, then $\mathrm{Reg}(T) \leq 2\sqrt{AT/\epsilon}$. Specifically, if $\epsilon = \Omega(1)$, then we get the standard rates of $\mathrm{Reg}(T) = \Theta(\sqrt{AT})$.*
- ***Polynomial Budget:*** *if $B(t) = t^c$ for some $c \in (0, 1]$, then the regret is also polynomial, i.e., $\mathrm{Reg}(T) = \Theta(\sqrt{A}T^{1-c/2})$.*
- ***Fixed Budget:*** *if $B(t) = B_0 > 0$ is an initial budget, then $\mathrm{Reg}(T) = \Theta(\sqrt{AT}/\sqrt{B_0})$. However, if the budget is given at the end of the game, namely, $B(t) = 0$ for $t \leq T - B_0$ and $B(t) = B_0$ for $t > T - B_0$, then $\mathrm{Reg}(T) = \Omega(T)$ for any $B_0 = o(T)$.*
- ***Periodically-replenished budget*** *if the budget is replenished by $B_0 > 0$ every $N \in \mathbb{N}$ steps, namely, $B(t) = B_0 \cdot \left(1 + \lfloor \frac{t}{N} \rfloor\right)$, then*

$$\mathrm{Reg}(T) = \mathcal{O}\left(\sum_{s=1}^{\lceil T/N \rceil} \frac{\sqrt{A}N}{\sqrt{B_0 s}}\right) = \mathcal{O}\left(\sqrt{\frac{ATN}{B_0}}\right) .$$

### 3.1. Lower Bounds for Linear Contextual Bandits

We end this section by presenting a lower bound for linear bandits. Here, we assume that the budget constraint is never violated (as we assume in the upper bounds). Then, for fixed budget and context space, we derive the following bound:

**Proposition 4.** *Let $T \in \mathbb{N}$ be some time horizon and let $\pi$ be a linear bandit policy such that $n_T^q \leq B$ a.s. for some fixed $B \leq T$. Then, there exists a $d$-dimensional linear bandit instance with arm set $[-1, 1]^d$ for which the expected regret of $\pi$ is lower bounded by $\frac{dT}{80\sqrt{B}}$.*

See Appendix A.4 for a proof. Importantly, this bound can be generalized to time-varying budgets, as in Corollary 3.

## 4. The Greedy Reduction: Gap Between Adversarial and Stochastic Contexts

We start by tackling the simpler case where the contexts are stochastic and the budget is oblivious. Formally, before the game starts, a sequence of budgets $\{B(t)\}_{t>1}$ is chosen, possibly adversarially. Later, at the beginning of each round $t$, a context $u_t$ is generated from a distribution $\mathcal{P}_u$, independently at random of other rounds. Then, the model continues as in Section 2. For this section, we also assume that queries have unit costs.

**Algorithm 1** Greedy Reduction

**Require:** Algorithm $\mathbb{A}$, initial budget $B(1) \geq 1$
**Initialize:** $l = 0$
**for** $t = 1, ..., T$ **do**
  Observe context $u_t \sim \mathcal{P}_u$, and current budget $B(t)$
  **if** $B(t) \geq B^q(t) + 1$ **then**
    // Query reward feedback, act with $\mathbb{A}$
    Advance $l \leftarrow l + 1$ and calculate $\pi_t \leftarrow \mathbb{A}_l(u_t)$
    Act with $\pi_t, q_t = 1$; observe $Z_t$ and $R_t$
  **else**
    // Don't query feedback, act with 'average' policy
    Sample $j \sim Uniform(\{1, \ldots, l\})$
    Act with $\pi_t \leftarrow \mathbb{A}_j(u_t)$and $q_t = 0$; ignore $Z_t$
  **end if**
**end for**

For this model, we suggest a simple greedy reduction (see Algorithm 1). Take an algorithm $\mathbb{A}$. If there is enough budget, query reward feedback and ask $\mathbb{A}$ for a policy $\pi_t$ to act with. Otherwise, when there is no available budget, pick uniformly at random a policy from past policies returned by $\mathbb{A}$, $\{\pi_t\}$, and act with it. We remark that $\mathbb{A}_k(u)$ denotes an output-policy of the algorithm at its $k^{th}$ iteration, with $u$ as the input context. Albeit simple, this algorithm performs surprisingly well, as we show in the following theorem:

**Theorem 1** (Black Box Reduction for Stochastic Contexts)**.** *Let $\mathbb{A}$ be an anytime algorithm with bounded regret $\mathbb{E}[\mathrm{Reg}(T)] \leq \alpha T^{\beta} + C$ for some $\alpha, C \in \mathbb{R}_+, \beta \in [0, 1]$ and any $T \in \mathbb{N}$. Moreover, assume the budget is chosen by an oblivious adversary such that it is non-decreasing, $B(1) \geq 1$ and $B(t) \in \mathbb{N}$ for all $t \geq 1$. Then, the expected regret of Algorithm 1 with base algorithm $\mathbb{A}$ and budget sequence $\{B(t)\}_{t\geq 1}$ is upper bounded by $\alpha T^{\beta} + C + \sum_{t=1}^{T} \frac{\alpha}{B^{1-\beta}(t)} + \frac{C}{B(t)}$.*

The proof of the theorem (and all other results in the section) can be found at Appendix B. One possible application of the theorem is in the MAB setting, combined with MOSS-anytime (Degenne & Perchet, 2016). This would result in a regret bound of $\mathcal{O}\left(\sum_{t=1}^{T}\sqrt{A/B(t)}\right)$, which matches the lower bound of Corollary 3 up to log-factors. For linear bandits, using OFUL (Abbasi-Yadkori et al., 2011) as the base algorithm implies a regret of $\tilde{\mathcal{O}}\left(\sum_{t=1}^{T} d/\sqrt{B(t)}\right)$, which matches the lower bound of Proposition 4 up to log-factors. In general, we believe that this reduction is tight in many interesting settings. One possible intuitive explanation for this can be found at the following proposition. In it,

we prove that in non-contextual problems, for any fixed horizon, any general algorithm can be converted to one that uses the budget at the beginning of the game. A reasonable adaptation for time-varying budget and anytime algorithm would be to use the budget whenever possible.

**Proposition 5.** *Assume that the decision-making problem is non-contextual ($u_t = \phi, \forall t$) with no environment feedback ($Z_t = \phi, \forall t$) and unit-querying costs. Then, for any $T, B \in \mathbb{N}$ such that $B \leq T$ and any policy $\pi$ under which $n_T^q \leq B$, there exists a policy $\pi'$ such that $q_t = 1$ for all $t \in [B]$ (and zero otherwise) and $\mathbb{E}[\mathrm{Reg}(T)|\pi'] = \mathbb{E}[\mathrm{Reg}(T)|\pi]$.*

We end this section by returning to its basic assumptions - stochastic contexts and oblivious budget. We show that when at least one of these assumptions do not hold, the greedy reduction suffers a linear regret in a very simple CMAB problem, even if the budget is linear in expectation:

**Proposition 6** (Greedy Reduction Degrades in the Presence of Adversary)**.** *If an adaptive adversary controls either (i) the contexts, or (ii) the budget, then for any base algorithm $\mathbb{A}$ used in Algorithm 1, there exists a contextual MAB problem with two contexts and two arms such that $\mathbb{E}[\mathrm{Reg}(T)] \geq \frac{T}{4}$, even if $\mathbb{E}[B(t)] = \frac{t}{2}$ for all $t \in [T]$.*

*Proof Sketch.* Consider a contextual multi-armed bandit instance with two contexts, $u \in \{1, 2\}$. Assume that querying a reward feedback costs 1 for all contexts and all arms. Furthermore, assume the budget increases in each episode by 1 with probability $1/2$.

If the adversary is adaptive to the history, it can choose $u = 1$ every round the budget increases and otherwise choose $u = 2$. The greedy reduction then only queries for feedback for $u = 1$. Thus, the regret for $u = 2$ is linear in $T$, since no information is gathered for this context, and the number of rounds $u = 2$ is $\Omega(T)$. Lastly, it can be shown that $\mathbb{E}[B(t)] = t/2$ in this construction. Equivalently, the same result holds if the contexts are uniformly distributed and an adaptive adversarial budget increases by a single unit only when $u = 1$. □

This emphasizes the need for developing non-greedy algorithms that store budget to face adversities in the problem.

## 5. The Confidence-Budget Matching Principle

In the previous section, we showed that a simple greedy query rule performs well for sequential budgeted learning with stochastic contexts and oblivious budget. That is, querying for feedback as long as a spare budget exists results in a well-performing approach. However, this 'greedy' approach can miserably fail in the presence of adversarial contexts or budget. In this section, we introduce an alternative approach we refer to as the Confidence-Budget Matching (CBM) principle. Unlike the greedy approach, CBM works well in the presence of adversities as it adequately preserves budget.

**Algorithm 2** Confidence-Budget Matching (CBM) Scheme

**Require:** Optimistic algorithm $\mathbb{A}$, $\{\alpha_t\}_{t\geq 1}$
**for** $t = 1, ..., T$ **do**
  Observe context $u_t$
  Act with $\pi_t$, acquired from $\mathbb{A}(F_{t-1})$ and observe $Z_t$
  Observe current budget $B(t)$
  **if** $CI_t(u_t, \pi_t) \geq \alpha_t \sqrt{1/B(t)}$ **then**
    Ask for feedback ($q_t = 1$) and observe $R_t$
  **end if**
**end for**

CBM is a generic algorithmic scheme that converts an unbudgeted optimistic algorithm to an algorithm that can be utilized in sequential budgeted learning. As evident in Algorithm 2, the agent follows a policy calculated by the baseline algorithm $\mathbb{A}$. Then, feedback on the reward of $\pi_t$ is queried if the confidence interval (CI) of the policy, given current context, $CI_t(u_t, \pi_t)$ is larger than $\alpha\sqrt{1/B(t)}$ for some $\alpha > 0$. As querying rewards decreases the CI, $CI_t(u_t, \pi_t)$ will gradually decrease. Then, if a policy is chosen frequently enough, reward querying will stop once its confidence matches $\alpha\sqrt{1/B(t)}$.

Unlike the greedy reduction, the performance of CBM does not degrade in the presence of adversarial contexts or budget, as we demonstrate later in this section. A crucial reason for this is that CBM stops querying rewards of policies with small CI. This somewhat conservative behavior leads to a more robust algorithm. To better understand the robustness of this querying rule, we consider the MAB problem. For this problem, we set $\alpha_t \sim \sqrt{A}$, thus, for the MAB problem, CBM queries reward feedback if $CI_t(a_t) \geq \tilde{\mathcal{O}}(\sqrt{A/B(t)})$. Denoting the number of queries from action $a$ before the $t^{th}$ episode by $n_{t-1}^q(a)$ and setting $CI_t(a_t) \sim 1/\sqrt{n_{t-1}^q(a_t)}$ (Hoeffding-based CI) leads to the following equivalent condition to CBM query rule for MAB: *ask for reward feedback if* $n_{t-1}(a_t) \lesssim B(t)/A$. Namely, query for feedback if $a_t$ was queried less than $B(t)/A$ times so far. Thus, this rule implicitly allocates $1/A$ of the current budget to each of the arms for possible use. This immediately implies the budget constraint is never violated, since there are $A$ arms in total.

**Remark 1.** *Notice that the CBM scheme plays actions selected by the optimistic baseline algorithm $\mathbb{A}$, which do not depend on the current budget $B(t)$. In particular, all our results also hold even if the budget is revealed after the agent selects an action, as depicted in Algorithm 2.*

Next, we study the performance of the CBM principle applied to MAB, linear bandits and RL problems. Importantly, we show that for all these settings, it matches the perfor-

mance of the greedy reduction for stochastic environments, while being able to face adversarial contexts and budgets.

**Remark 2** (Sufficient Initial Budget)**.** *For simplicity, we assume the initial budget $B(1)$ is large enough such that Algorithm 2 queries at the first round, that is $CI_1(u_1, \pi_1) \geq \alpha_1\sqrt{1/B(1)}$. If this condition does not hold, an extra term of $T_I$ should be added to the regret bounds where $T_I$ is the first time in which $CI_{T_I}(u_{T_I}, \pi_{T_I}) \geq \alpha_{T_I}\sqrt{1/B(T_I)}$.*

### 5.1. Multi-Armed Bandits

We start by studying the performance of CBM for the MAB problem, where the base algorithm is UCB1 (Auer et al., 2002). We call the resulting algorithm CBM-UCB, which follows Algorithm 2 with $\alpha_t = 4\sqrt{6\sum_a c(a)\log(At)}$. Although this setting is extremely simple, it highlights the central analysis technique, which is extended in the rest of this section to more challenging decision-making problems.

**Theorem 2** (Confidence Budget Matching for Multi Armed Bandits)**.** *For any querying costs $c(1), \ldots, c(A) \geq 0$, any adaptive non-decreasing adversarially chosen sequence $\{B(t)\}_{t\geq 1}$ and for any $T \geq 1$, the expected regret of CBM-UCB is upper bounded by $\tilde{\mathcal{O}}\left(\sqrt{AT} + \sqrt{\sum_a c(a)}\sum_{t=1}^{T}\mathbb{E}\left[\sqrt{\frac{1}{B(t)}}\right]\right)$.*

Full description of the algorithm, alongside the proof of Theorem 2, is supplied at Appendix C. We now present a proof sketch that highlights how the CBM principle affects the regret bounds.

*Proof Sketch.* We use UCB bonus of $b_t^r(a) \triangleq \sqrt{\frac{3\log(At)}{2n_{t-1}^q(a)\vee 1}}$, where $n_t^q(a)$ is the number of times arm $a$ was queried up to round $t$; namely, if $\bar{r}_t(a)$ is the empirical mean of $a$ then, $UCB_t(a) = \bar{r}_{t-1}(a) + b_t^r(a)$, $LCB_t(a) = \bar{r}_{t-1}(a) - b_t^r(a)$ and $CI_t(a) = UCB_t(a) - LCB_t(a) = 2b_t^r(a)$.

**Budget analysis.** We start the proof by establishing that the budget constraint is never violated, $B^q(T) \leq B(T)$ for all $T \geq 1$. For simplicity, we do so for unit querying costs (where $B^q(t) = n^q(t)$). By the CBM condition, if $q_t = 1$, then $CI_t(a_t) \geq \alpha_t/\sqrt{B(t)}$. Then, for any $T \geq 1$

$$n^q(T) = \sum_{t=1}^{T} 1\{q_t = 1\} \leq \sum_{t=1}^{T} \frac{CI_t(a_t)}{\alpha_t/\sqrt{B(t)}} 1\{q_t = 1\}$$
$$\lesssim \sqrt{B(T)}\sum_{t=1}^{T} \frac{1}{\sqrt{n_{t-1}^q(a)\vee 1}} 1\{q_t = 1\},$$

where in the last relation we substituted all parameters and used the fact that the budget is non-decreasing. Importantly, notice that when the reward of an arm is queried, its count increases, up to $n_t^q(a)$. Therefore, for any $T \geq 1$

$$n^q(T) \lesssim \sqrt{B(T)}\sum_{a=1}^{A}\sum_{i=0}^{n_T^q(a)} \frac{1}{\sqrt{i\vee 1}} \lesssim \sqrt{B(T)}\sqrt{n^q(T)}.$$

Reorganizing and choosing the right constants leads to the relation $n^q(T) \leq B(T)$, which deterministically holds. Importantly, this implies that CBM-UCB never tries to query reward without sufficient budget, so $q_t = 1$ if and only if the CBM condition holds, or, equivalently, $q_t = 0$ if and only if the CBM condition does not hold.

**Regret analysis**. Using standard concentration arguments, the expected regret $\mathbb{E}[\mathrm{Reg}(T)]$ is bounded by

$$\sum_{t=1}^{T}\mathbb{E}[(UCB_t(a_t) - LCB_t(a_t))1\{q_t = 1\}] \quad (1)$$
$$+\sum_{t=1}^{T}\mathbb{E}[(UCB_t(a_t) - LCB_t(a_t))1\{q_t = 0\}]. \quad (2)$$

For term (1), reward is always queried; therefore, the analysis closely follows standard analysis for UCB, which results with a bound of $\mathcal{O}\left(\sqrt{AT\log(AT)}\right)$. For (2), we know that reward was not queried, i.e, $q_t = 0$. Since $q_t = 0$ if and only if the CBM condition is not met, it implies that the CI is lower than the CBM-threshold, namely

$$\sum_{t=1}^{T}\mathbb{E}[(UCB_t(a_t) - LCB_t(a_t))1\{q_t = 0\}]$$
$$\lesssim \sum_{t=1}^{T}\mathbb{E}\left[\frac{\alpha_t}{\sqrt{B(t)}}\right] = \tilde{\mathcal{O}}\left(\sum_a c(a)\right)\sum_{t=1}^{T}\mathbb{E}\left[\sqrt{\frac{1}{B(t)}}\right].$$

Combining both bounds leads to the desired regret bound. □

### 5.2. Linear Bandits

Next, we focus on applying the CBM principle, i.e., Algorithm 2, for linear bandits. The base algorithm that we rely on is OFUL (Abbasi-Yadkori et al., 2011), and we set $\alpha_t = \tilde{\mathcal{O}}(d)$ (see Appendix D for the full description of the algorithm). We call the resulting algorithm CBM-OFUL. Importantly, and in contrast to the greedy reduction of Section 4, we allow both the contexts and the budget to be chosen by an adaptive adversary. Nonetheless, CBM-OFUL still achieve the same performance as the greedy reduction Theorem 1, while not suffering of performance degradation in the presence of adaptive adversary (for a complete proof see Appendix D):

**Theorem 3** (Confidence Budget Matching for Linear Bandits)**.** *For any adaptive adversarially chosen sequence of non-decreasing budget and context sets*

*$\{B(t), u_t\}_{t\geq 1}$ the regret of CBM-OFUL is upper bounded by $\tilde{\mathcal{O}}\left(d\left(\sqrt{T}+\sum_{t=1}^{T}\frac{1}{\sqrt{B(t)}}\right)\right)$ for any $T\geq 1$ with probability greater than $1-\delta$.*

Notice that this matches the lower bound of Proposition 4. Notably, the examples of Proposition 6 can be represented as a linear bandit problem with $d=4$. Thus, in contrast to the greedy reduction, which suffers linear regret, the regret of CBM-OFUL is $\tilde{\mathcal{O}}\left(\sqrt{T}\right)$.

### 5.3. Reinforcement Learning

In this section we apply the CBM principle to RL. For this setting, we relax the budget model presented in Section 2 and allow agents to query specific state action pairs along the trajectory observed at the $t^{th}$ episode $\{(s_{t,h}, a_{t,h})\}_{h\in[H]}$. Namely, at the $t^{th}$ episode, the agent acts with $\pi_t$, observes a trajectory $\{(s_{t,h}, a_{t,h})\}_{h\in[H]}$ and is allowed to query for reward feedback from any state-action pair along the trajectory. If the agent queries reward feedback in the $t^{th}$ episode at the $h^{th}$ time step it receives $R_{t,h}(s_{t,h}, a_{t,h})$. We denote this event as choosing $q_{t,h}=1$. For simplicity, we work with unit-budget costs, i.e., the total budget used by the agent is $B^q(t)=\sum_{k=1}^{t}\sum_{h=1}^{H}1\{q_{k,h}=1\}$ and must be smaller than $B(t)$. Observe that in the standard RL setting, the reward budget is $B(t)=Ht$ for all $t\geq 1$.

Notably, querying reward feedback from specific time steps allows us to derive regret bounds that depend on the *sparsity* of the reward function. Formally, let $\mathcal{L}_R$ be the set of tuples $(s,a,h)$ with $r_h(s,a)\neq 0$. Then, for any $(s,a,h)\notin\mathcal{L}_R$, $r_h(s,a)=0$, and since $R_t\in[0,1]$, it also implies that $R_{t,h}\equiv 0$. Assume that the algorithm knows the cardinality of this set $|\mathcal{L}_R|$ (or an upper bound on $|\mathcal{L}_R|$). Leveraging this knowledge, we set the CBM feedback query rule in Algorithm 2, line 6, as follows,

$$\text{Ask for reward feedback on } (s_{t,h},a_{t,h}) \text{ if}$$
$$CI^R_{t,h}(s_{t,h},a_{t,h}) \gtrsim \sqrt{\tfrac{|\mathcal{L}_R|}{B(t)}}+\tfrac{SAH}{B(t)}\ (q_{t,h}=1),$$

where $CI^R_{t,h}(s_{t,h},a_{t,h})$ is the CI of the reward estimation of $s_{t,h},a_{t,h}$ in the $h^{th}$ time step at the $t^{th}$ episode. Setting the reward bonus of the 'optimistic' model as in UCBVI-CH (Azar et al., 2017) leads to the following bound (see Appendix E for more details on the algorithm and proofs).

**Theorem 4** (CBM-UCBVI)**.** *For any adaptive adversarially chosen sequence of non-decreasing budget and initial state, $\{B(t), s_{t,1}\}_{t\geq 1}$, the regret of CBM-UCBVI is upper bounded by*

$$\tilde{\mathcal{O}}\left(\sqrt{SAH^4T}+H^3S^2A+\sum_{t=1}^{T}\sqrt{\frac{|\mathcal{L}_R|H^2}{B(t)}}+\frac{SAH^2}{B(t)}\right)$$

*for any $T\geq 1$ with probability greater than $1-\delta$.*

Notice that the last term of the regret is dominated by its first term when $B(t)=\Omega(\sqrt{T})$ and the remaining budget-dependent term only scales with the sparsity-level of the reward $|\mathcal{L}_R|$. Notably, this implies that when $B(t)\sim t\lceil|\mathcal{L}_R|/SAH\rceil$, the third term is of the same order as the first term. Differently put, if the query budget $B(t)$ increases by a single unit every $SAH/|\mathcal{L}_R|$ episodes, the worst case performance of CBM-UCBVI remains the same, while reducing the amount of reward feedback.

While CBM-UCBVI clearly demonstrates the analysis techniques and insights from applying the CBM principle to RL, it is of interest to combine it with an algorithm with order-optimal regret bounds of $\sqrt{SAH^3T}$ (e.g., (Jin et al., 2018)) when $B(t)=Ht$, that is, in the standard RL setting (notice that $T$ is the number of *episodes* and not the total number of time steps). We achieve this goal by performing a more refined analysis that uses tighter concentration results based on (Azar et al., 2017; Dann et al., 2019; Zanette & Brunskill, 2019). Indeed, doing so leads to tighter regret bounds by a $\sqrt{H}$ factor in the leading term (Full details on the algorithm and proofs can be found at Appendix F).

**Theorem 5** (CBM-ULCVI)**.** *For any adaptive adversarially chosen sequence of non-decreasing budget and initial state, $\{B(t), s_{t,1}\}_{t\geq 1}$, the regret of CBM-ULCBVI is upper bounded by*

$$\tilde{\mathcal{O}}\left(\sqrt{SAH^3T}+H^3S^2A+\sum_{t=1}^{T}\sqrt{\frac{|\mathcal{L}_R|H^2}{B(t)}}+\frac{SAH^2}{B(t)}\right)$$

*for any $T\geq 1$ with probability greater than $1-\delta$.*

This bound results in an interesting conclusion for general RL problems, i.e., when $|\mathcal{L}_R|=SAH$. Plugging this into Theorem 5, we observe that a budget of $B(t)=t$ – instead a budget of $B(t)=Ht$ as used in standard RL – results in order optimal regret bound. That is, it suffices for CBM-ULCVI to query reward feedback once per episode, without causing for performance degradation in a minimax sense.

### 5.4. General View on CBM for Optimistic Algorithms

The CBM principle queries for reward feedback (Algorithm 2, line 6) if the CI of the applied context-action is larger than a threshold, $CI_t(u_t,\pi_t)\geq\alpha_t\sqrt{1/B(t)}$, or more generally, if $CI_t(u_t,\pi_t)\geq\alpha_t f(B(t))$ for some $f:\mathbb{R}\to\mathbb{R}$. A natural question arises: how to choose $\alpha_t$ and $f$?

A useful rule of thumb to guide the choice of $\alpha_t$ and $f$ is the following: if the regret of the optimistic algorithm $\mathbb{A}$ is bounded by $\tilde{\mathcal{O}}(\alpha T^{\beta})$ then set $\alpha_t=\tilde{\mathcal{O}}(\alpha)$ and $f(x)=\tilde{\mathcal{O}}(x^{\beta-1})$. This matches the parameters chosen for both MAB and linear bandits. In RL, we relied on this rule but used a more complex function $f$, due to the application of an empirical Bernstein concentration argument (Maurer & Pontil, 2009).

The logic behind this choice is simple; it guarantees that the budget constraint is never violated, $B^q(T) \leq B(T)$ for all $T \geq 1$. Differently put, for any episode, reward feedback is not queried *if and only if* $CI_t(u_t, \pi_t) \leq \alpha_t f(B(t))$. This property can be proved via similar technique as in the proof sketch of Theorem 2 for CBM-MAB. An informal proof for the correctness of this statement for the general case goes as follows (for unit feedback-costs),

$$\begin{aligned} B^q(T) &\leq \sum_{t=1}^{T} 1\{q_t = 1\} CI_t(u_t, x_t)/(\alpha B(t)^{\beta-1}) \\ &\overset{(a)}{\leq} \big(B(T)^{1-\beta})/\alpha\big) \sum_{t=1}^{T} 1\{q_t = 1\} CI_t(u_t, x_t) \\ &\overset{(b)}{\lesssim} \big(B(T)^{1-\beta})/\alpha\big) \alpha B^q(T)^{\beta}, \end{aligned}$$

where$(a)$ holds since the budget is non-decreasing, and $(b)$ since $\sum_{t=1}^{T} 1\{q_t = 1\} CI_t(u_t, x_t) \sim \mathrm{Reg}(B^q(T))$ for optimistic algorithms. Rearranging yields that $B^q(T)^{1-\beta} \lesssim B(T)^{1-\beta}$ which implies that $B^q(T) \leq B(T)$ by the monotonicity of $x^{1-\beta}$. Although the analysis for CBM in linear bandits and RL is more subtle, the intuition supplied by this informal reasoning is of importance; we believe it can serve as a starting point for future analysis of CBM-based algorithms in sequential budgeted learning.

## 6. Related Work

**Multi-Armed Bandits with Paid Observations (Seldin et al., 2014).** Closely related to our work is the framework of MAB with paid observations. There, an agent plays with an arm $a_t$ and is allowed to query reward feedback on any subset of arms. Unlike in our case, there is no strict budget for observations, but, rather, each query comes at a cost that is subtracted from the reward. Notably, this requires translating the query costs to the same units as the reward, which is oftentimes infeasible. For example, in online recommendations, there is no clear way to quantify user dissatisfaction from feedback requests. In such cases, it is much more natural to enforce a (possibly time-varying) hard constraint on the number of feedback queries. Furthermore, the work of Seldin et al. (2014) focus on the MAB problem, whereas in this work, we focus on more involved contextual problems (i.e., linear bandits and RL). It is important to note that the analysis in (Seldin et al., 2014) holds for the adversarial reward model, whereas in this work, we focused on the stochastic reward model (with adversarial contexts and budget). We believe it is an interesting question what type of guarantees can be derived for the fully adversarial setting, i.e., when the rewards, budget and contexts are adversarially chosen. Finally, when applied to the stochastic case, the algorithm of Seldin et al. (2014) requires $B(T) = \Omega(T^{2/3})$. In contrast, our results hold for lower budgets, while achieving similar bounds when $B(T) = \Omega(T^{2/3})$.

**MABs with Additional Observations (Yun et al., 2018).** In this closely related MAB setting, observing the reward of arms that were not played is possible, at a certain cost, as long as a non-decreasing budget constraint is not violated. Nonetheless, a key difference from our work is that Yun et al. (2018) assume that the reward of the played arm is *always* observed and does not consume any budget. Therefore, there is no clear way to apply their results to our setting.

**Bandits with Knapsacks (BwK) (Badanidiyuru et al., 2013).** In the BwK model, a sampling budget is given prior to the game. At each round, the agent selects an arm and observes noisy samples of both the reward and the cost of the selected arm. That is, the agent always receives feedback on its actions. This comes in stark contrast to our model, where the budget restricts the amount of feedback an agent can obtain. Furthermore, in the BwK model, the game stops as soon as the cumulative cost exceeds the initial budget. In our model, where the budget serves as a constraint on the reward feedback, interaction continues even without an observation budget. When the budget is exhausted, the agent can still utilize its past information on the system to perform reasonably good actions. Notably, this forces the agent to sufficiently explore actions, even if they are costly, to identify high-rewarding ones.

We remark that there are additional extensions of the MAB setting in which arms incur costs (e.g., Sinha et al., 2021). There, the objective of an agent is to minimize a relaxed notion of cumulative regret and the cumulative cost. Unlike this work, we do not attribute cost to applying an action, but attribute a cost to *receiving feedback* on the reward.

**RL with trajectory feedback (Efroni et al., 2020).** Under this model, instead of observing a reward for each played state-action, the agent only observes the cumulative rewards of each episode. This serves two reasons: first, and similarly to our work, it aims to reduce the feedback that the algorithm requires (by a factor of $H$), and when rewards are manually labeled, reduce the labeling load. Second, for many applications, it is much more natural to label the reward for a full trajectory than to each state-action. However, this approach comes at a noticeable cost, both in performance and computational complexity. In contrast, by sampling specific state-action pairs, our approach allows reducing the amount of feedback while maintaining similar performance and computational complexity. Nonetheless, we believe that when trajectory feedback is more natural, our approach can also be applied to further reduce the feedback for this setting. We leave such an extension for future work.

## 7. Summary and Discussion

In this work, we presented a novel framework for sequential decision-making under time-varying budget constraints. We analyzed what can and cannot be achieved by greedily using querying whenever possible. Then, we presented the CBM principle, which only queries rewards for actions with high uncertainty, compared to the current budget. We demonstrated how to apply the principle to MAB, linear bandits and RL problems and proved that it performs well also in the presence of adversities. We believe that this model can be adapted to many real-world problems and leaves room for interesting extensions, which we leave for future work.

**Is there a value in knowing the future budget?** Throughout this work, we assume the agent only observes the current budget $B(t)$ at the beginning of each round and does not have knowledge on future values of the budget $B(t')$ for $t' > t$. Intuitively, one expects that knowing the future budget would result in an improved and less conservative behavior in terms of budget allocation. Surprisingly, our matching lower and upper bounds for MAB (Corollary 3 and Theorem 2) and linear bandits (Proposition 4 and Theorem 3) show that this intuition does not always hold. Nonetheless, understanding if or when information on future budget is of value remains an interesting open question.

**Monotonicity of the budget.** Throughout this work, we assume that the budget never decreases. Intuitively, it implies that once a budget is allocated, it does not matter when the algorithm decides to use it. Nonetheless, for some problems, different assumptions are sometimes more relevant. A budget might be given alongside an 'expiration date' or might expire probabilistically. Another possible assumption is that the spare (unused) budget is bounded. Finally, in some instances, the total budget might be characterized by a specific random process, e.g., a biased random walk.

**Problem-dependent bounds.** Throughout this work, we focused on problem-independent regret bounds, that is, bounds that do not depend on the specific problem instance. Bounds that depend on specific instances usually focus on sufficiently sampling suboptimal arms, while implicitly assuming that optimal arms are sufficiently sampled (Auer et al., 2002). In contrast, when rewards are not always observed, algorithms must also control the number of queries from optimal arms. This becomes much harder in the presence of multiple optimal arms; in this case, an algorithm can never know if an arm is optimal or has a small suboptimal gap and might 'waste' budget while trying to discern which is true. In some sense, we believe that the CBM principle is well-suited for this setting, as it prevents the agent from exhausting all budget on specific arms.

**Adaptivity to structure.** In Section 5.3, we proved that when rewards are sparse, our algorithm can query rewards according to the sparsity level, while maintaining the same regret bounds as the unbudgeted case. However, to do so, we required an upper bound on the sparsity of the problem. Therefore, a natural extension is to devise an algorithm that can adapt to an unknown sparsity level. Moreover, it is well known that structural assumptions can lead to improved regret bounds, and previous works proposed algorithms whose regret depends on nontrivial structural properties of the problem (Maillard et al., 2014; Zanette & Brunskill, 2019; Foster et al., 2019; 2020; Merlis & Mannor, 2019; 2020). Thus, it is interesting to understand what structural properties (beyond sparsity) affect the budgeted performance and how to design algorithms that adapt to such properties.

## Acknowledgments

This work was partially funded by the Israel Science Foundation under ISF grant number 2199/20. YE is partially supported by the Viterbi scholarship, Technion. NM is partially supported by the Gutwirth Scholarship.

## References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Abeille, M., Lazaric, A., et al. Linear thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

Agrawal, S. and Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1, 2012.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135, 2013.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. *arXiv preprint arXiv:1703.05449*, 2017.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 207–216. IEEE, 2013.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Four-*

*teenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26, 2011.

Bretagnolle, J. and Huber, C. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

Cohen, A., Kaplan, H., Mansour, Y., and Rosenberg, A. Near-optimal regret bounds for stochastic shortest path. *arXiv preprint arXiv:2002.09869*, 2020.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, pp. 355–366, 2008.

Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pp. 1507–1516, 2019.

Degenne, R. and Perchet, V. Anytime optimal algorithms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pp. 1587–1595. PMLR, 2016.

Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, pp. 12224–12234, 2019.

Efroni, Y., Merlis, N., and Mannor, S. Reinforcement learning with trajectory feedback. *arXiv preprint arXiv:2008.06036*, 2020.

Foster, D. J., Krishnamurthy, A., and Luo, H. Model selection for contextual bandits. *arXiv preprint arXiv:1906.00531*, 2019.

Foster, D. J., Rakhlin, A., Simchi-Levi, D., and Xu, Y. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

Garivier, A. and Cappé, O. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Conference on Learning Theory*, pp. 359–376, 2011.

Garivier, A., Ménard, P., and Stoltz, G. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pp. 4863–4873, 2018.

Jin, T. and Luo, H. Learning adversarial mdps with bandit feedback and unknown transition. *arXiv preprint arXiv:1912.01192*, 2019.

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pp. 199–213. Springer, 2012.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Maillard, O.-A., Mann, T. A., and Mannor, S. How hard is my mdp?" the distribution-norm to the rescue". *Advances in Neural Information Processing Systems*, 27: 1835–1843, 2014.

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

Merlis, N. and Mannor, S. Batch-size independent regret bounds for the combinatorial multi-armed bandit problem. In *Conference on Learning Theory*, pp. 2465–2489. PMLR, 2019.

Merlis, N. and Mannor, S. Tight lower bounds for combinatorial multi-armed bandits. In *Conference on Learning Theory*, pp. 2830–2857. PMLR, 2020.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529, 2015.

Seldin, Y., Bartlett, P., Crammer, K., and Abbasi-Yadkori, Y. Prediction with limited advice and multiarmed bandits with paid observations. In *International Conference on Machine Learning*, pp. 280–287. PMLR, 2014.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354, 2017.

Simchowitz, M. and Jamieson, K. G. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *Advances in Neural Information Processing Systems*, pp. 1153–1162, 2019.

Sinha, D., Sankararaman, K. A., Kazerouni, A., and Avadhanula, V. Multi-armed bandits with cost subsidy. In *International Conference on Artificial Intelligence and Statistics*, pp. 3016–3024. PMLR, 2021.

Tarbouriech, J., Garcelon, E., Valko, M., Pirotta, M., and Lazaric, A. No-regret exploration in goal-oriented reinforcement learning. In *International Conference on Machine Learning*, pp. 9428–9437. PMLR, 2020.

Yun, D., Proutiere, A., Ahn, S., Shin, J., and Yi, Y. Multi-armed bandit with additional observations. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(1):1–22, 2018.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *arXiv preprint arXiv:1901.00210*, 2019.

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