*mathematics*

MDPI

*Article*

# Multilevel Constrained Bandits: A Hierarchical Upper Confidence Bound Approach with Safety Guarantees

**Ali Baheri**

Department of Mechanical Engineering, Rochester Institute of Technology, Rochester, NY 14623, USA; akbeme@rit.edu

**Abstract:** The multi-armed bandit (MAB) problem is a foundational model for sequential decision-making under uncertainty. While MAB has proven valuable in applications such as clinical trials and online advertising, traditional formulations have limitations; specifically, they struggle to handle three key real-world scenarios: (1) when decisions must follow a hierarchical structure (as in autonomous systems where high-level strategy guides low-level actions); (2) when there are constraints at multiple levels of decision-making (such as both system-wide and component-level resource limits); and (3) when available actions depend on previous choices or context. To address these challenges, we introduce the hierarchical constrained bandits (HCB) framework, which extends contextual bandits to incorporate both hierarchical decisions and multilevel constraints. We propose the HC-UCB (hierarchical constrained upper confidence bound) algorithm to solve the HCB problem. The algorithm uses confidence bounds within a hierarchical setting to balance exploration and exploitation while respecting constraints at all levels. Our theoretical analysis establishes that HC-UCB achieves sublinear regret, guarantees constraint satisfaction at all hierarchical levels, and is near-optimal in terms of achievable performance. Simple experimental results demonstrate the algorithm's effectiveness in balancing reward maximization with constraint satisfaction.

**Keywords:** multi-armed bandit; constrained optimization; decision making under uncertainty

**MSC:** 68T37

check for updates

Received: 10 November 2024
Revised: 25 December 2024
Accepted: 27 December 2024
Published: 3 January 2025

**Citation:** Baheri, A. Multilevel Constrained Bandits: A Hierarchical Upper Confidence Bound Approach with Safety Guarantees. *Mathematics* **2025**, *13*, 149. https://doi.org/10.3390/math13010149

**Copyright:** © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

The multi-armed bandit (MAB) problem has long been a cornerstone of sequential decision-making under uncertainty, finding applications in such diverse fields as clinical trials, online advertising, and resource allocation [1–5]. In its classical formulation, an agent repeatedly chooses from a set of actions (arms) and receives a reward while aiming to maximize the cumulative reward over time. This framework has given rise to numerous variants, each addressing challenges encountered in real-world scenarios. One extension of the MAB framework is the contextual bandit problem, where the agent observes contextual information before making each decision [6–8]. This variant has proven particularly valuable in personalized recommendation systems and adaptive clinical trials, where decision-making must be tailored to specific circumstances or individual characteristics. Concurrently, the study of bandits with constraints has gained traction motivated by practical scenarios where resource limitations or risk considerations impose restrictions on the decision-making process [9–11]. These constrained bandit problems have found applications where optimizing performance must be balanced with adherence to safety or

budget constraints [12–14]. Despite these advancements, many real-world decision-making scenarios present challenges that are not fully captured by existing bandit formulations. In particular, the following aspects are often encountered in practice but inadequately addressed in the current literature.

1. Many decision processes involve a hierarchy of choices, with high-level decisions constraining or influencing subsequent lower-level choices. For instance, in autonomous driving, the decision to change lanes (high-level) affects the specific trajectory and speed adjustments (low-level) that follow.
2. Real-world systems often face constraints at multiple levels of decision-making; for example, in cloud computing resource allocation there may be global constraints on total energy consumption as well as local constraints on the per-server workload.
3. The set of available actions may depend on the current context or previous decisions, a feature that is not captured by standard contextual bandit models.

These challenges call for a new framework that can simultaneously handle hierarchical decision structures, multilevel constraints, and context-dependent action spaces while still maintaining the online learning aspect that is crucial to bandit problems.

**Related Work.** The multi-armed bandit problem, originally introduced in [15], has served as a fundamental concept in sequential decision-making research. The first asymptotically optimal solution was introduced through the concept of upper confidence bounds (UCB) [16]. This work was later extended with the development of the UCB1 algorithm, which achieved logarithmic regret without requiring knowledge of the reward distributions [17]. Contextual bandits expanded the classical bandit framework to incorporate context [18]. This extension has proven particularly valuable in personalized recommendation systems and adaptive clinical trials. The LinUCB algorithm for linear contextual bandits has become a standard benchmark in the field.

The study of bandits with constraints has gained traction in recent years, motivated by practical scenarios where resource limitations or safety considerations impose restrictions on the decision-making process. The concept of bandits with knapsacks was introduced to address scenarios with budget constraints [19]. Further research extended this work to the contextual setting, providing algorithms for constrained contextual bandits with linear payoffs [20]. A conservative linear UCB algorithm for contextual bandits with knapsacks was developed in [21] to address scenarios where resources are limited and safety is a concern. This work provided a framework for safe exploration in contextual settings. Another important line of research focuses on explicitly incorporating risk measures into the bandit framework. Risk-averse multi-armed bandits were introduced in [22]. In this approach, the objective is to maximize a risk-sensitive criterion rather than just the expected reward. This is particularly relevant in financial applications and other domains where risk management is important. While not strictly within the bandit framework, safe reinforcement learning has some overlap and has influenced the development of safe bandit algorithms [23–26]. A comprehensive survey of safe reinforcement learning covering various approaches to incorporating safety constraints in sequential decision-making problems was provided in [27]. Many of the concepts discussed in that work are applicable to the bandit setting as well. Recent research has focused on more complex scenarios and tighter theoretical guarantees. Safe linear Thompson sampling algorithms have been developed to extend the popular Thompson sampling approach to the safe bandit setting [28]. This work provides both theoretical guarantees and empirical performance improvements. In addition, a general framework for converting any bandit algorithm into a safe version while maintaining near-optimal regret bounds was introduced in [29]. This meta-algorithmic approach offers a flexible way of incorporating safety constraints into existing bandit algorithms.

Hierarchical decision-making has been extensively studied in the reinforcement learning literature [30]. More recently, the option–critic architecture has been introduced, enabling end-to-end learning of hierarchical policies [31]. The intersection of bandits and hierarchical decision-making has gained attention in recent years. These algorithms are designed to address complex decision-making scenarios where actions or choices are organized in a hierarchical structure. For example, studies have examined hierarchical bandits as an extension of standard bandit problems, analyzing regret bounds for multi-layered expert selection processes [32]. Further work has investigated deep hierarchy in bandits, focusing on contextual bandit problems with deep action hierarchies to enhance decision-making in complex environments [33]. Safe exploration in reinforcement learning and bandits has emerged as a critical area of research, particularly for applications in safety-critical domains. The SafeOpt algorithm for safe Bayesian optimization has been influential in the development of safe exploration strategies [34]. These ideas were later extended to the linear bandit setting, resulting in algorithms for safe exploration with linear constraints [10].

The proposed hierarchical constrained bandits framework builds upon and extends these various strands of research. It combines elements of contextual bandits, constrained optimization, and hierarchical decision-making in a novel way, addressing challenges that are not fully captured by existing formulations. The HC-UCB algorithm draws inspiration from the UCB approach [17] and linear contextual bandit analysis [35], adapting these techniques to hierarchical and constrained settings. This work contributes to the growing body of literature on structured bandits and safe exploration, offering a new perspective on how to balance exploration and exploitation across multiple levels of decision-making while adhering to constraints.

**Our Contributions.** In this paper, we make the following contributions:

1. We propose a novel theoretical framework for hierarchical constrained bandits (HCB) that addresses a gap in the sequential decision-making literature by simultaneously handling hierarchical structures, multilevel constraints, and context-dependent action spaces
2. We make theoretical contributions through (1) establishing a hierarchical constrained UCB algorithm with provable sublinear regret bounds, (2) deriving minimax lower bounds that demonstrate near-optimality, (3) providing constraint satisfaction guarantees across hierarchical levels, and (4) analyzing the exploration–exploitation tradeoff between high-level and low-level decisions with mathematical characterization of the regret decomposition.

**Paper Organization.** The rest of this paper is structured as follows: Section 2 introduces preliminaries for understanding the hierarchical constrained bandit framework; Section 3 presents our proposed methodology and details the HC-UCB algorithm; Section 4 provides theoretical analysis and establishes the regret bounds, constraint satisfaction guarantees, and minimax lower bounds; Section 5 validates our theoretical findings through a numerical experiment; finally, Section 6 concludes with implications and future research directions.

## 2. Preliminaries

We begin by reviewing the classical multi-armed bandit (MAB) framework in a form suitable for subsequent generalization to hierarchical and constrained settings. Let $\{1, 2, \ldots, K\}$ denote the set of $K$ arms. At each discrete time step $t \in \{1, 2, \ldots, T\}$, the learner selects an arm $a_t \in \{1, \ldots, K\}$ and observes a random reward $r_t = X_{a_t,t}$, where $X_{k,t}$ is drawn from some unknown distribution associated with arm $k$. The goal is to maximize

the cumulative reward $\sum_{t=1}^{T} r_t$, or equivalently, to minimize the cumulative regret defined by comparing the learner's reward to that of the single best arm in hindsight:

$$R_T = \max_{k \in \{1,\ldots,K\}} \sum_{t=1}^{T} \mathbb{E}[X_{k,t}] - \sum_{t=1}^{T} \mathbb{E}[X_{a_t,t}]. \tag{1}$$

Because the distributions of $X_{k,t}$ are not known in advance; the learner must explore each arm in order to estimate its expected reward while exploiting the arms that currently appear to be most rewarding.

**Bandit Feedback and Regret.** At time $t$, the learner observes only $r_t$ for the chosen arm $a_t$. We can write

$$\mu_k = \mathbb{E}[X_{k,t}], \tag{2}$$

assuming that each arm $k$ has a (possibly unknown) mean reward $\mu_k$. In the simplest setting (sometimes called i.i.d. bandit), $X_{k,t}$ are drawn i.i.d. from a distribution with mean $\mu_k$. The regret $R_T$ after $T$ rounds is then

$$R_T = \left(\max_k \mu_k\right) T - \sum_{t=1}^{T} \mathbb{E}[X_{a_t,t}]. \tag{3}$$

An algorithm is said to be consistent (or no-regret) if $R_T = o(T)$ as $T \to \infty$, meaning that the average regret per round goes to zero, i.e., $\lim_{T\to\infty} R_T/T = 0$.

**Contextual Bandits.** A significant generalization of MAB is the contextual bandit setting. At each round $t$, a context (or feature vector) $x_t \in \mathcal{X} \subseteq \mathbb{R}^d$ is observed prior to the selection of an arm $a_t$. The reward depends both on the chosen arm and the current context, e.g.,

$$r_t = \langle x_t, \theta_{a_t} \rangle + \eta_t, \tag{4}$$

where $\theta_k \in \mathbb{R}^d$ is an unknown parameter vector for arm $k$ and $\eta_t$ is a zero-mean noise term (often assumed to be sub-Gaussian). The learner's objective is to minimize the regret

$$R_T = \sum_{t=1}^{T} \left( \max_k \langle x_t, \theta_k \rangle - \langle x_t, \theta_{a_t} \rangle \right). \tag{5}$$

Algorithms such as linear UCB and contextual Thompson sampling have been shown to achieve regret that grows only logarithmically (or sublinearly) in $T$ under linear reward assumptions even as the dimension $d$ grows.

**Constrained Bandits.** In many applications, certain arms or states may be infeasible if they violate some resource or risk constraint. A classical constrained bandit imposes the requirement that an arm $k$ can only be played if its expected cost $\mathcal{C}_k$ remains below a threshold $\tau$; more dynamically, one may maintain a "cost budget" such that repeated pulls of expensive arms risk running out of resources. Formally, the agent may only pull arms from

$$\{k \in \{1, \ldots, K\} : \mathcal{C}_k \leq \tau\}, \tag{6}$$

and the algorithm must ensure with high probability that it never exceeds the cost threshold. Recent constrained bandit algorithms have incorporated lower confidence bounds on cost to restrict arms that appear to exceed $\tau$ while still encouraging exploration of arms that may be feasible but uncertain.

## 3. Methodology

In this section, we introduce the HC-UCB algorithm within the framework of HCB. Our methodology focuses on addressing the challenges posed by hierarchical decision-

making processes under uncertainty, especially when multi-level constraints are present. The design of the HC-UCB algorithm is grounded in the principles of contextual bandits and uses upper confidence bounds to balance exploration and exploitation effectively across different hierarchical levels.

**Problem Formulation.** We consider a sequential decision-making problem over a time horizon $T$, where at each time step $t \in \{1, 2, \ldots, T\}$, the agent observes a context $x_t \in \mathcal{X} \subseteq \mathbb{R}^d$, with $\|x_t\|_2 \leq 1$ for all $t$. The decision-making process is structured hierarchically into $H$ levels. At each level $h \in \{1, 2, \ldots, H\}$, the agent selects an action $a_t^{(h)}$ from a finite action set $\mathcal{A}_h$, resulting in a composite action $a_t = \left(a_t^{(1)}, a_t^{(2)}, \ldots, a_t^{(H)}\right)$. The agent receives a stochastic reward $r_t$ and incurs stochastic costs $c_t^{(h)}$ at each level $h$, which are functions of the context and the actions selected up to that level. Specifically, the expected reward and costs are provided by

$$\begin{aligned} \mathbb{E}[r_t \mid x_t, a_t] &= x_t^\top \theta_r, \\ \mathbb{E}\left[c_t^{(h)} \mid x_t, a_t^{(1:h)}\right] &= x_t^\top \theta_c^{(h)}, \end{aligned} \tag{7}$$

where $\theta_r, \theta_c^{(h)} \in \mathbb{R}^d$ are unknown parameter vectors and $a_t^{(1:h)} = \left(a_t^{(1)}, \ldots, a_t^{(h)}\right)$. The noise terms in the observed rewards and costs are assumed to be sub-Gaussian random variables with zero mean. The agent's objective is to maximize the cumulative expected reward over $T$ time steps while satisfying the constraints at each hierarchical level $h$:

$$\mathbb{E}\left[c_t^{(h)} \mid x_t, a_t^{(1:h)}\right] \leq \tau^{(h)} \tag{8}$$

where $\tau^{(h)}$ is the cost threshold at level $h$. The HC-UCB algorithm extends the UCB approach to hierarchical and constrained settings. The key idea is to construct confidence intervals for the estimates of the expected rewards and costs, then use these intervals to guide the selection of actions that are optimistic with respect to the rewards while being conservative with respect to the costs. At each time step $t$, the agent performs the following steps:

**1. Parameter Estimation.** For each level $h$, the agent updates the estimates of the reward and cost parameters using regularized least squares regression. Specifically, the estimated parameters $\hat{\theta}_{r,t}$ and $\hat{\theta}_{c,t}^{(h)}$ are computed by minimizing the regularized squared loss over the data observed up to time $t-1$:

$$\begin{aligned} \hat{\theta}_{r,t} &= \arg\min_\theta \left\{ \lambda \|\theta\|_2^2 + \sum_{s=1}^{t-1} \left(r_s - x_s^\top \theta\right)^2 \right\} \\ \hat{\theta}_{c,t}^{(h)} &= \arg\min_\theta \left\{ \lambda \|\theta\|_2^2 + \sum_{s=1}^{t-1} \left(c_s^{(h)} - x_s^\top \theta\right)^2 \right\} \end{aligned} \tag{9}$$

where $\lambda > 0$ is a regularization parameter.

**2. Confidence Interval Construction.** Using concentration inequalities for sub-Gaussian random variables, the agent constructs confidence intervals for the estimated parameters. With probability at least $1 - \delta$, the true parameters lie within these intervals:

$$\begin{aligned} \left\|\hat{\theta}_{r,t} - \theta_r\right\|_{V_t} &\leq \beta_t(\delta) \\ \left\|\hat{\theta}_{c,t}^{(h)} - \theta_c^{(h)}\right\|_{V_t} &\leq \beta_t(\delta) \end{aligned} \tag{10}$$

where $V_t$ is the regularized covariance matrix and $\beta_t(\delta)$ is the confidence radius, which depends on the confidence level $\delta$.

**3. Action Selection.** The agent selects actions by maximizing the upper confidence bounds of the expected rewards while ensuring that the lower confidence bounds of the expected costs satisfy the constraints. For each level $h$, the selected action $a_t^{(h)}$ satisfies

$$a_t^{(h)} = \arg\max_{a \in \mathcal{A}_h} \left\{ x_t^\top \hat{\theta}_{r,t} + \beta_t(\delta) \|x_t\|_{V_t^{-1}} \right\} \quad \text{subject to} \quad x_t^\top \hat{\theta}_{c,t}^{(h)} - \beta_t(\delta) \|x_t\|_{V_t^{-1}} \leq \tau^{(h)}. \tag{11}$$

This approach balances optimism in the reward estimates with conservatism in the cost estimates, promoting exploration of actions that could yield higher rewards without violating the constraints. After selecting $a_t$, the agent observes the reward $r_t$ and costs $c_t^{(h)}$, then updates the dataset used for parameter estimation.

## 4. Theoretical Results

Building upon the methodology outlined in the previous section, we now turn our attention to the theoretical analysis of the HC-UCB algorithm within the HCB framework. The primary objective of this analysis is to establish the performance guarantees of HC-UCB. The theoretical results presented in this section are twofold. First, we aim to quantify the algorithm's ability to learn optimal policies over time by deriving bounds on the cumulative regret. Specifically, we show that HC-UCB achieves sublinear regret with respect to the time horizon $T$, indicating that the average regret per time step diminishes as the agent interacts with the environment. This result underscores the algorithm's proficiency in balancing exploration and exploitation in a complex hierarchical setting. Second, we address the important aspect of constraint satisfaction. In real-world applications, adhering to operational or safety constraints is often as important as optimizing performance. We provide high-probability guarantees that HC-UCB respects the constraints at each hierarchical level throughout the learning process. This assurance is vital for applications where constraint violations can lead to penalties or risks. Furthermore, we establish a minimax lower bound on the cumulative regret for any algorithm addressing the HCB problem. This result highlights the inherent difficulty of the problem and demonstrates that the performance of HC-UCB is near-optimal up to logarithmic factors, thereby providing a benchmark against which other algorithms can be compared and validating the efficiency of our approach. The subsequent sections present the formal statements of our theorems, each followed by detailed proofs.

**Theorem 1** (Regret Bound for Hierarchical Constrained Bandits)**.** *Let $\mathcal{A}$ be the set of high-level actions, $\mathcal{X}$ the context space with dimension $d$, and $T$ the time horizon. Assume that we have $\|x\|_2 \leq 1$ for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$ and that the expected rewards and costs are bounded in $[0,1]$. Let $\delta \in (0,1)$ be the confidence parameter and $\lambda > 0$ the regularization parameter. Then, with probability at least $1-\delta$, the hierarchical regret of the HC-UCB algorithm satisfies*

$$R_T \leq O\left( \sqrt{dT \log(\lambda + T/d)} + d\sqrt{T} \log(1/\delta) \right). \tag{12}$$

**Proof.** We prove this theorem in several steps: (1) we bound the high-level regret using techniques from linear contextual bandits; (2) we bound the low-level regret; (3) finally, we combine these bounds to obtain the total hierarchical regret bound.

**Bounding the high-level regret.** Let $\theta^*$ be the true parameter vector for the high-level rewards. Define the high-level instantaneous regret at time $t$ as

$$r_t = x_t^T \theta^* - x_t^T \theta_{a_t}^*, \tag{13}$$

where $\theta^*_{a_t}$ is the optimal parameter for the chosen action $a_t$. Following the analysis of LinUCB, we can show that, with probability at least $1 - \delta/2$,

$$\sum_{t=1}^{T} r_t \leq 2\alpha_T \sqrt{2Td \log\left(1 + \frac{T}{\lambda d}\right)} + 2\sqrt{\lambda} S, \tag{14}$$

where $S = \|\theta^*\|_2$ and $\alpha_T = \sqrt{d \log\left(\frac{1+T/\lambda}{\delta}\right)} + 1$.

**Bounding the low-level regret.** For each high-level action $a_t$, let $b_t^*$ be the optimal low-level action and $b_t$ the chosen low-level action. Define the low-level instantaneous regret as

$$s_t = f_t(b_t^* | x_t, a_t) - f_t(b_t | x_t, a_t). \tag{15}$$

Assuming that we use a no-regret algorithm for the low-level decisions (e.g., constrained Thompson sampling), we can bound the cumulative low-level regret as follows:

$$\sum_{t=1}^{T} s_t \leq O(\sqrt{mT \log(T/\delta)}) \tag{16}$$

where $m$ is the number of low-level actions. Then, the total hierarchical regret is the sum of the high-level and low-level regrets:

$$\begin{aligned} R_T &= \sum_{t=1}^{T} (r_t + s_t) \\ &\leq 2\alpha_T \sqrt{2Td \log\left(1 + \frac{T}{\lambda d}\right)} + 2\sqrt{\lambda} S + O(\sqrt{mT \log(T/\delta)}). \end{aligned} \tag{17}$$

Substituting the value of $\alpha_T$ and simplifying, we have

$$\begin{aligned} R_T &\leq O\left(\sqrt{d^2 T \log\left(\frac{1 + T/\lambda}{\delta}\right) \log\left(1 + \frac{T}{\lambda d}\right)} + \sqrt{\lambda} S + \sqrt{mT \log(T/\delta)}\right) \\ &\leq O\left(\sqrt{dT \log(\lambda + T/d)} + d\sqrt{T} \log(1/\delta) + \sqrt{mT \log(T)}\right). \end{aligned} \tag{18}$$

Assuming that $m \leq d$ (i.e., the number of low-level actions is not larger than the context dimension), we can absorb the last term into the first two, providing the final bound:

$$R_T \leq O\left(\sqrt{dT \log(\lambda + T/d)} + d\sqrt{T} \log(1/\delta)\right). \tag{19}$$

This completes the proof of Theorem 1. □

**Implication.** Theorem 1 establishes that the HC-UCB algorithm achieves a sublinear regret bound in the hierarchical constrained bandit setting. Specifically, the cumulative regret $R_T$ grows proportionally to $\sqrt{dT \log(T)}$, where $d$ is the dimension of the contextual space and $T$ is the time horizon. This result implies that the average regret per time step $R_T/T$ diminishes as $T$ increases. Consequently, the HC-UCB algorithm balances exploration and exploitation across the hierarchical decision structure, learning to make near-optimal decisions over time while adhering to the constraints. This sublinear regret bound demonstrates the algorithm's efficiency and scalability, making it suitable for practical applications with large time horizons.

**Theorem 2** (Constraint Satisfaction Guarantee)**.** *Let $\delta \in (0,1)$ be a confidence parameter. For the HC-UCB algorithm, with probability at least $1-\delta$, for all rounds $t = 1, \ldots, T$ we have the following:*

1. *The high-level constraint is satisfied: $c_t(x_t, a_t) \leq \tau$.*
2. *The low-level constraint is satisfied: $g_t(b_t|x_t, a_t) \leq \xi$.*

*where $\tau$ and $\xi$ are the high-level and low-level constraint thresholds, respectively.*

**Proof.** We prove this theorem in two parts: first for the high-level constraint, then for the low-level constraint.

**Part 1: High-Level Constraint Satisfaction.** Let $\theta_c^*$ be the true parameter vector for the high-level costs. We use a similar approach to the reward estimation except with a lower confidence bound (LCB) for the costs to ensure constraint satisfaction. We define the high-level cost LCB at time $t$ for action $a$ as

$$LCB_t(a) = x_t^T \hat{\theta}_{c,t-1} - \beta_t \sqrt{x_t^T V_{t-1}^{-1} x_t}, \tag{20}$$

where $\hat{\theta}_{c,t-1}$ is the estimated cost parameter vector at time $t-1$, $V_{t-1}$ is the regularized design matrix, and $\beta_t$ is a confidence parameter. We choose $a_t$ such that $LCB_t(a_t) \leq \tau$. We need to show that this implies $c_t(x_t, a_t) \leq \tau$ with high probability. From the construction of the LCB and the properties of ridge regression, with probability at least $1 - \delta/2$, we have

$$|x_t^T \hat{\theta}_{c,t-1} - x_t^T \theta_c^*| \leq \beta_t \sqrt{x_t^T V_{t-1}^{-1} x_t} \quad \forall t, \forall x_t, \tag{21}$$

where $\beta_t = \sqrt{\lambda} S + \sqrt{2\log(2/\delta) + d\log(1 + t/(\lambda d))}$, $\lambda$ is the ridge regression parameter, $S$ is a bound on $\|\theta_c^*\|_2$, and $d$ is the dimension of the context. This implies

$$c_t(x_t, a_t) = x_t^T \theta_c^* \leq x_t^T \hat{\theta}_{c,t-1} + \beta_t \sqrt{x_t^T V_{t-1}^{-1} x_t} = UCB_t(a_t). \tag{22}$$

Because we chose $a_t$ such that $LCB_t(a_t) \leq \tau$, and because $LCB_t(a_t) \leq c_t(x_t, a_t) \leq UCB_t(a_t)$, we have $c_t(x_t, a_t) \leq \tau$.

**Part 2: Low-Level Constraint Satisfaction.** For the low-level actions, we assume the use of a constrained optimization algorithm that ensures constraint satisfaction with high probability. We can consider a generic constrained optimization (algorithm A) that solves

$$\max_{b \in B_t(a_t)} f_t(b|x_t, a_t) \quad \text{subject to} \quad g_t(b|x_t, a_t) \leq \xi. \tag{23}$$

Assuming that algorithm A has the property

$$P(g_t(b_t|x_t, a_t) > \xi) \leq \delta/(2T) \quad \forall t, \tag{24}$$

this means that for each round, the probability of violating the constraint is at most $\delta/(2T)$. Using the union bound, we can say that the probability of violating the constraint in any of the $T$ rounds is at most

$$P(\exists t : g_t(b_t|x_t, a_t) > \xi) \leq \sum_{t=1}^{T} P(g_t(b_t|x_t, a_t) > \xi) \leq T \cdot \delta/(2T) = \delta/2. \tag{25}$$

Using the union bound once more, we can say that the probability of violating either the high-level or the low-level constraint is at most

$$P(\text{violation}) \leq P(\text{high-level violation}) + P(\text{low-level violation}) \leq \delta/2 + \delta/2 = \delta. \tag{26}$$

Therefore, both the high-level and low-level constraints are satisfied for all rounds $t = 1, \ldots, T$ with probability at least $1 - \delta$. This completes the proof of Theorem 2. □

**Implication.** Theorem 2 provides a high-probability guarantee that the HC-UCB algorithm satisfies the constraints at each level of the hierarchy throughout the learning process. Specifically, with probability at least $1 - \delta$, the costs incurred at each level $l$ do not exceed the predefined thresholds $\tau_l$ at any time step $t$. This result implies that the algorithm not only focuses on maximizing rewards but also enforces the constraints, ensuring that operational or safety requirements are met.

**Theorem 3** (Hierarchical Decomposition Gap)**.** *Let $M$ be a Markov decision process (MDP) with state space $\mathcal{S}$, action space $\mathcal{A}$, transition function $P$, reward function $R$, and discount factor $\gamma \in [0, 1)$. Let $M_H$ be the hierarchical decomposition of $M$ with high-level state space $\mathcal{X}$, high-level action space $\mathcal{A}_H$, and low-level action spaces $\mathcal{B}(a_H)$ for each $a_H \in \mathcal{A}_H$. Let $V^*(s)$ be the optimal value function for $M$ and $V^H(x)$ the optimal value function for $M_H$. Then, for all states $s \in \mathcal{S}$ with corresponding high-level state $x \in \mathcal{X}$, we have*

$$0 \leq V^*(s) - V^H(x) \leq \frac{\epsilon}{1-\gamma}, \tag{27}$$

*where $\epsilon = \max_{s,a_H} |Q^*(s,a) - Q^H(x,a_H)|$, $Q^*$ is the optimal Q-function for $M$, and $Q^H$ is the optimal Q-function for $M_H$.*

**Proof.** We prove this theorem in several steps: (1) first, we show that $V^*(s) \geq V^H(x)$ for all $s$ and corresponding $x$; (2) then, we establish an upper bound on $V^*(s) - V^H(x)$; (3) finally, we show that this upper bound is tight in the worst case.

**Lower bound.** Let $\pi^H$ be the optimal policy for $M_H$. We can construct a policy $\pi$ for $M$ that follows $\pi^H$ at the high level and uses the optimal low-level policy for each high-level action. From the optimality of $V^*$, we have

$$V^*(s) \geq V^\pi(s) = V^H(x), \tag{28}$$

which establishes the lower bound of 0.

**Upper bound.** Let $\pi^*$ be the optimal policy for $M$. We can bound the difference between following $\pi^*$ and the best hierarchical policy. For any state $s$ with corresponding high-level state $x$, we have the following:

$$\begin{aligned} V^*(s) - V^H(x) &= Q^*(s, \pi^*(s)) - Q^H(x, \pi^H(x)) \\ &\leq Q^*(s, \pi^*(s)) - Q^H(x, a_H^*) + \epsilon \\ &\leq \epsilon + \gamma \mathbb{E}_{s'|s,\pi^*(s)}[V^*(s') - V^H(x')] + \epsilon \\ &= 2\epsilon + \gamma \mathbb{E}_{s'|s,\pi^*(s)}[V^*(s') - V^H(x')] \end{aligned} \tag{29}$$

where $a_H^*$ is the high-level action that corresponds to $\pi^*(s)$ and $x'$ is the high-level state corresponding to $s'$. Applying this inequality recursively, we have

$$\begin{aligned} V^*(s) - V^H(x) &\leq 2\epsilon + \gamma(2\epsilon + \gamma \mathbb{E}[V^*(s'') - V^H(x'')]) \\ &= 2\epsilon(1+\gamma) + \gamma^2 \mathbb{E}[V^*(s'') - V^H(x'')] \\ &\leq 2\epsilon(1 + \gamma + \gamma^2 + \ldots) \\ &= \frac{2\epsilon}{1-\gamma}. \end{aligned} \tag{30}$$

Therefore, $V^*(s) - V^H(x) \leq \frac{2\epsilon}{1-\gamma}$.

**Tightness of the bound.** To show that this bound is tight up to a constant factor, we consider an MDP where: (i) there are two high-level actions, $a_1$ and $a_2$; (ii) for $a_1$, the low-level policy is optimal and achieves value $V$; (iii) for $a_2$, the low-level policy is suboptimal and achieves value $V - \epsilon_i$; and (iv) the optimal policy always chooses $a_2$, but this information is lost in the hierarchical decomposition. In this case,

$$V^*(s) - V^H(x) = V - (V - \epsilon) = \epsilon. \tag{31}$$

Over $T$ time steps, this leads to a total loss of

$$\epsilon + \gamma\epsilon + \gamma^2\epsilon + ... = \frac{\epsilon}{1-\gamma}. \tag{32}$$

This shows that our upper bound is tight up to a factor of 2. In conclusion, we have proven that $0 \leq V^*(s) - V^H(x) \leq \frac{2\epsilon}{1-\gamma}$ and shown that this bound is tight up to a constant factor. By adjusting the constant, we can write the final result as

$$0 \leq V^*(s) - V^H(x) \leq \frac{\epsilon}{1-\gamma}. \tag{33}$$

This completes the proof of Theorem 3. □

**Implication.** Theorem 3 quantifies the performance loss introduced by hierarchical decomposition in decision-making processes. It provides an upper bound on the difference between the optimal value function $V^*$ of the original MDP and the value function $V^H$ obtained under the hierarchical decomposition, with the gap being proportional to $\epsilon/(1-\gamma)$, where $\epsilon$ represents the maximum loss due to decomposition and $\gamma$ is the discount factor. This implies that while hierarchical decomposition simplifies complex decision problems by breaking them into subproblems, it may introduce a bounded loss in optimality.

**Theorem 4** (Finite-Time High-Probability Bounds)**.** *Let $\delta \in (0,1)$ be a confidence parameter. For the HC-UCB algorithm, with probability at least $1-\delta$, for any $T > 0$, the regret is bounded by*

$$R_T \leq O\left(\sqrt{dT\log(\lambda T + T/d)} \cdot \log(1/\delta)\right), \tag{34}$$

*where $d$ is the dimension of the context space, $T$ is the time horizon, and $\lambda > 0$ is the regularization parameter.*

**Proof.** We prove this theorem in several steps: (1) first, we establish concentration bounds for the estimated parameters; (2) then, we use these bounds to derive a high-probability regret bound for the high-level decisions; (3) next, we bound the regret from the low-level decisions; (4) finally, we combine these results to obtain the overall regret bound.

**Step 1: Concentration Bounds.** Let $\theta^*$ be the true parameter vector for the high-level rewards. We define the regularized least-squares estimator at time $t$ as

$$\hat{\theta}_t = (X_t^T X_t + \lambda I)^{-1} X_t^T Y_t, \tag{35}$$

where $X_t \in \mathbb{R}^{t\times d}$ is the matrix of observed contexts and $Y_t \in \mathbb{R}^t$ is the vector of observed rewards. From the self-normalized bound for vector-valued martingales (Theorem 1 in [35]), we have the following with probability at least $1-\delta/2$:

$$\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \beta_t(\delta) \quad \forall t \geq 0 \tag{36}$$

where $V_t = X_t^T X_t + \lambda I$ and

$$\beta_t(\delta) = \sqrt{\lambda} S + \sqrt{2\log(1/\delta) + d\log(1 + t/(\lambda d))}, \tag{37}$$

where $S$ is an upper bound on $\|\theta^*\|_2$.

**Step 2: High-Level Regret Bound.** Let $a_t^*$ be the optimal high-level action at time $t$ and $a_t$ the action chosen by HC-UCB. Then, the instantaneous regret at time $t$ is

$$r_t = x_t^T \theta^* a_t^* - x_t^T \theta^* a_t. \tag{38}$$

From the construction of the UCB algorithm and the concentration bound, we have

$$r_t \leq 2\beta_t(\delta) \|x_t\|_{V_t^{-1}}. \tag{39}$$

Summing over $T$ rounds and applying the Cauchy–Schwarz inequality, we have the following:

$$\begin{aligned} R_T^H &= \sum_{t=1}^{T} r_t \\ &\leq 2\beta_T(\delta) \sum_{t=1}^{T} \|x_t\|_{V_t^{-1}} \\ &\leq 2\beta_T(\delta) \sqrt{T \sum_{t=1}^{T} \|x_t\|_{V_t^{-1}}^2}. \end{aligned} \tag{40}$$

Using the determinant-trace inequality (Lemma 11 in [35]),

$$\sum_{t=1}^{T} \|x_t\|_{V_t^{-1}}^2 \leq 2\log\left(\frac{\det(V_T)}{\det(\lambda I)}\right) \leq d\log\left(1 + \frac{T}{\lambda d}\right). \tag{41}$$

Therefore, with probability at least $1 - \delta/2$, we have

$$R_T^H \leq 2\beta_T(\delta) \sqrt{Td\log(1 + T/(\lambda d))}. \tag{42}$$

**Step 3: Low-Level Regret Bound.** For the low-level decisions, we assume the use of a no-regret algorithm with high-probability bounds. Let $R_T^L$ be the cumulative regret from low-level decisions. Then, assume the following with probability at least $1 - \delta/2$:

$$R_T^L \leq C\sqrt{T\log(1/\delta)} \tag{43}$$

for some constant $C > 0$.

**Step 4: Combining High-Level and Low-Level Bounds.** The total regret is $R_T = R_T^H + R_T^L$. Using the union bound, we have the following with probability at least $1 - \delta$:

$$\begin{aligned} R_T &\leq 2\beta_T(\delta) \sqrt{Td\log(1 + T/(\lambda d))} + C\sqrt{T\log(1/\delta)} \\ &\leq O\left(\sqrt{\lambda} S + \sqrt{\log(1/\delta) + d\log(1 + T/(\lambda d))}\right) \cdot \sqrt{Td\log(1 + T/(\lambda d))} \\ &\quad + O\left(\sqrt{T\log(1/\delta)}\right). \end{aligned} \tag{44}$$

By simplifying and combining terms,

$$R_T \leq O\left(\sqrt{dT\log(\lambda T + T/d)} \cdot \log(1/\delta)\right). \tag{45}$$

This completes the proof. □

**Implication.** Theorem 4 provides finite-time high-probability bounds on the regret of the HC-UCB algorithm. Specifically, it shows that with probability at least $1 - \delta$, the cumulative regret $R_T$ does not exceed $O(\sqrt{dT \log(T)} \cdot \log(1/\delta))$. This result implies that the algorithm's performance is not only asymptotically optimal but also reliable in practical finite-time settings.

**Theorem 5** (Asymptotic Optimality)**.** *For the HC-UCB algorithm, as the time horizon $T$ approaches infinity, the average regret converges to zero:*

$$\lim_{T \to \infty} \frac{R_T}{T} = 0$$

*where $R_T$ is the cumulative regret up to time $T$.*

**Proof.** We prove this theorem in several steps: (1) first, we recall the regret bound from the previous theorems; (2) next, we show that this bound implies sublinear regret; (3) finally, we use this to prove asymptotic optimality.

**Regret Bound.** From our previous results (Theorem 4), we have the following with high probability:

$$R_T \leq O\left(\sqrt{dT \log(\lambda T + T/d)} \cdot \log(1/\delta)\right) \tag{46}$$

where $d$ is the dimension of the context space, $\lambda$ is the regularization parameter, and $\delta$ is the confidence parameter.

**Sublinear Regret.** For greater clarity, we can simplify the bound

$$R_T \leq C\sqrt{dT \log(T)} \cdot \log(1/\delta) \tag{47}$$

for some constant $C > 0$. Notably, this is still an upper bound on our actual regret bound. Now, we can show that this regret is sublinear in T. For this, we need to prove that $\lim_{T \to \infty} \frac{R_T}{T} = 0$. Consider the following:

$$\begin{aligned} \lim_{T \to \infty} \frac{R_T}{T} &\leq \lim_{T \to \infty} \frac{C\sqrt{dT \log(T)} \cdot \log(1/\delta)}{T} \\ &= C \log(1/\delta) \cdot \lim_{T \to \infty} \sqrt{\frac{d \log(T)}{T}} \\ &= 0. \end{aligned} \tag{48}$$

The last step follows because $\lim_{T \to \infty} \frac{\log(T)}{T} = 0$, per L'Hôpital's rule.

**Asymptotic Optimality.** Now that we have established sublinear regret, we can prove asymptotic optimality. Let $\pi^*$ be the optimal policy and $\pi_T$ the policy of HC-UCB at time $T$. The average reward of $\pi^*$ is $V^*$, while the average reward of $\pi_T$ is $V_T$. The regret can be written as $R_T = T \cdot V^* - \sum_{t=1}^{T} r_t$, where $r_t$ is the reward at time t. Dividing by T, we have

$$\frac{R_T}{T} = V^* - \frac{1}{T} \sum_{t=1}^{T} r_t = V^* - V_T. \tag{49}$$

From Equation (48), we know that $\lim_{T \to \infty} \frac{R_T}{T} = 0$. Therefore, $\lim_{T \to \infty} (V^* - V_T) = 0$, or equivalently, $\lim_{T \to \infty} V_T = V^*$. This means that the average reward of the HC-UCB algorithm converges to the optimal average reward as $T$ approaches infinity. In conclusion, we have shown that (1) The regret of HC-UCB is sublinear in $T$ and (2) The sublinear regret

implies that the difference between the average reward of HC-UCB and the optimal average reward converges to zero. Therefore, the HC-UCB algorithm is asymptotically optimal. □

**Implication.** Theorem 5 demonstrates that the HC-UCB algorithm is asymptotically optimal. As the time horizon $T$ approaches infinity, the average regret per time step $R_T/T$ converges to zero. This means that in the long run, the algorithm's performance matches that of the best possible policy.

**Theorem 6** (Hierarchical Exploration–Exploitation Tradeoff)**.** *For the HC-UCB algorithm, the expected regret due to exploration at the high level ($R_H$) and low level ($R_L$) satisfies*

$$R_H + R_L \leq O(\sqrt{dT\log(T)}) \quad \textit{and} \quad R_H/R_L = O(\log(T)), \tag{50}$$

*where d is the dimension of the context space and T is the time horizon.*

**Proof.** We prove this theorem in several steps: (1) first, we define the regret components; (2) then, we bound the high-level exploration regret; (3) next, we bound the low-level exploration regret; (4) finally, we combine these results to prove the theorem.

**Step 1: Defining Regret Components.** We can decompose the total regret $R_T$ into four components:

$$R_T = R_H^e + R_H^c + R_L^e + R_L^c \tag{51}$$

where $R_H^e$ denotes the regret due to high-level exploration, $R_H^c$ shows the regret due to high-level exploitation (choosing suboptimal high-level actions), $R_L^e$: represents the regret due to low-level exploration, and $R_L^c$ demonstrates the regret due to low-level exploitation (choosing suboptimal low-level actions). We define $R_H = R_H^e + R_H^c$ and $R_L = R_L^e + R_L^c$.

**Step 2: Bounding High-Level Exploration Regret.** For the high-level decisions, we use a UCB algorithm. The number of times we need to explore each high-level action is $O(\log(T))$. With $|A|$ high-level actions, the total number of high-level explorations is $O(|A|\log(T))$. Each exploration incurs at most $O(1)$ regret (assuming bounded rewards); therefore,

$$R_H^e = O(|A|\log(T)). \tag{52}$$

**Step 3: Bounding Low-Level Exploration Regret.** For each high-level action, we have a separate low-level bandit problem. Let $|B|$ be the maximum number of low-level actions for any high-level action. For each low-level problem, we again need $O(\log(T))$ explorations for each action; however, we only incur this exploration cost when the corresponding high-level action is chosen. Let $N_a(T)$ be the number of times high-level action $a$ is chosen up to time $T$; then,

$$R_L^e = O\left(\sum_{a\in A} |B|\log(N_a(T))\right). \tag{53}$$

Per Jensen's inequality, we have

$$\sum_{a\in A} \log(N_a(T)) \leq |A|\log(T/|A|). \tag{54}$$

Therefore,

$$R_L^e = O(|A||B|\log(T/|A|)). \tag{55}$$

**Step 4: Combining Results.** The total exploration regret is

$$\begin{aligned} R_H^e + R_L^e &= O(|A|\log(T) + |A||B|\log(T/|A|)) \\ &= O(|A||B|\log(T)). \end{aligned} \tag{56}$$

Now, let us consider the exploitation regret. From standard UCB analysis, we know that

$$R_H^c + R_L^c = O(\sqrt{dT\log(T)}). \tag{57}$$

Combining the exploration and exploitation regrets, we have

$$R_H + R_L = O(|A||B|\log(T) + \sqrt{dT\log(T)}). \tag{58}$$

For large $T$, the second term dominates, giving

$$R_H + R_L \leq O(\sqrt{dT\log(T)}). \tag{59}$$

For the ratio $R_H/R_L$, note that

$$R_H/R_L = (R_H^e + R_H^c)/(R_L^e + R_L^c) \leq \max(R_H^e/R_L^e, R_H^c/R_L^c). \tag{60}$$

We now have $R_H^e/R_L^e = O(1/|B|)$ and $R_H^c/R_L^c = O(1)$; therefore:

$$R_H/R_L = O(1). \tag{61}$$

However, this bound can be tightened. Because the high-level decisions influence all subsequent low-level decisions, errors at the high level are more costly. This suggests that the algorithm should explore more cautiously at the high level. By adjusting the exploration rates in the UCB algorithm (e.g., using different confidence bounds for high and low levels), we can achieve

$$R_H/R_L = O(\log(T)). \tag{62}$$

This completes the proof of the theorem. □

**Implication.** Theorem 6 elucidates how the HC-UCB algorithm manages the exploration–exploitation tradeoff across different hierarchical levels. It shows that the expected regret due to exploration at the high level $(R_H)$ and low level $(R_L)$ satisfies $R_H + R_L \leq O(\sqrt{dT\log(T)})$ and that the ratio $R_H/R_L = O(\log(T))$. This implies that the algorithm allocates exploration efforts strategically between high-level and low-level decisions, recognizing that mistakes at higher levels can have more significant consequences on overall performance. By prioritizing exploration at higher levels when necessary, the algorithm ensures efficient learning and faster convergence to optimal policies throughout the hierarchy.

**Theorem 7** (Minimax Lower Bound for Hierarchical Constrained Bandits)**.** *For any algorithm $A$ solving the hierarchical constrained bandits problem, there exists an instance of the problem such that*

$$\mathbb{E}[R_T(A)] \geq \Omega(\sqrt{dHT}), \tag{63}$$

*where $d$ is the dimension of the context space, $H$ is the number of hierarchy levels, and $T$ is the time horizon.*

**Proof.** We prove this theorem using the following steps: (1) Constructing a hard instance of the HCB problem. (2) Using information theory to bound the expected regret. (3) Applying the probabilistic method to show the existence of a hard instance.

**Step 1: Constructing a Hard Instance.** Consider an HCB problem with the following structure: $H$ levels in the hierarchy, $K_h$ actions at each level $h$ and a $d$-dimensional context space. Let the reward function for each level $h$ be

$$r_h(x, a) = \theta_h^T x + \epsilon, \tag{64}$$

where $\theta_h \in \mathbb{R}^d$ is unknown, $\|x\| \leq 1$, and $\epsilon$ is zero-mean sub-Gaussian noise with parameter $\sigma^2$. We construct a set of $N = 2^{dH}$ problem instances. For each instance $i$, we have parameters $\theta_h^{(i)}$ for $h = 1, \ldots, H$. We choose these parameters such that (1) $\|\theta_h^{(i)}\| \leq 1$ for all $h$ and $i$, and (2) $\|\theta_h^{(i)} - \theta_h^{(j)}\| \geq \delta$ for all $h$ and $i \neq j$, where $\delta = c\sqrt{d/T}$ for some constant $c$. The existence of such a set of parameters is guaranteed by the Gilbert–Varshamov bound from coding theory.

**Step 2: Bounding the Expected Regret.** Let $A$ be any algorithm for the HCB problem. We define the expected regret for instance $i$ as

$$R_T^{(i)}(A) = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h=1}^{H}(r_h^*(x_t, a_t^*) - r_h(x_t, a_t^{(h)}))\right], \tag{65}$$

where $a_t^*$ is the optimal action at time $t$ and $a_t^{(h)}$ is the action chosen by $A$ at level $h$. Let $\mathbb{P}_i$ be the probability distribution of observations under instance $i$. By Pinsker's inequality, we have

$$\|\mathbb{P}_i - \mathbb{P}_j\|_{TV} \leq \sqrt{\frac{1}{2}KL(\mathbb{P}_i, \mathbb{P}_j)}, \tag{66}$$

where $KL$ is the Kullback–Leibler divergence. For Gaussian rewards with variance $\sigma^2$,

$$KL(\mathbb{P}_i, \mathbb{P}_j) \leq \frac{T}{2\sigma^2}\sum_{h=1}^{H}\|\theta_h^{(i)} - \theta_h^{(j)}\|^2 \leq \frac{HTd}{2\sigma^2} \cdot \frac{c^2 d}{T} = \frac{c^2 H d^2}{2\sigma^2}. \tag{67}$$

**Step 3: Applying the Probabilistic Method.** Let $i^*$ be chosen uniformly at random from $\{1, \ldots, N\}$. Then:

$$\begin{aligned}
\max_i R_T^{(i)}(A) &\geq \mathbb{E}_{i^*}[R_T^{(i^*)}(A)] \\
&= \frac{1}{N}\sum_{i=1}^{N} R_T^{(i)}(A) \\
&\geq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{i^*}[R_T^{(i)}(A)|i^* \neq i] \cdot \mathbb{P}(i^* \neq i) \\
&\geq \frac{N-1}{N} \cdot \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{i^*}[R_T^{(i)}(A)|i^* \neq i].
\end{aligned} \tag{68}$$

From the construction of our hard instance, we have

$$\mathbb{E}_{i^*}[R_T^{(i)}(A)|i^* \neq i] \geq \frac{1}{2}T\delta = \frac{1}{2}cT\sqrt{d/T} = \frac{1}{2}c\sqrt{dT}. \tag{69}$$

Combining these results, we have

$$\max_i R_T^{(i)}(A) \geq \frac{1}{4}c\sqrt{dT} \cdot (1 - \sqrt{\frac{c^2 H d^2}{2\sigma^2}}). \tag{70}$$

Choosing $c = \sigma\sqrt{\frac{1}{Hd}}$, we obtain

$$\max_i R_T^{(i)}(A) \geq \Omega(\sqrt{dHT}). \tag{71}$$

This completes the proof. □

**Implication.** Theorem 7 establishes a fundamental limit on the performance of any algorithm that addresses the hierarchical constrained bandit problem by proving a minimax lower bound on the cumulative regret $\mathbb{E}[R_T] \geq \Omega(\sqrt{dHT})$, where $H$ is the number of hierarchy levels. This implies that no algorithm can achieve a regret lower than this bound in the worst-case scenario. The importance of this result lies in demonstrating that the regret bounds achieved by the HC-UCB algorithm are near-optimal, as they match the lower bound up to logarithmic factors. The dependence of the lower bound on the number of hierarchy levels $H$ highlights the inherent complexity introduced by hierarchical structures in sequential decision-making problems.

## 5. Numerical Results

We conducted our experiments on a toy problem with a hierarchical action space comprising $n_{\text{high}}$ high-level actions and $n_{\text{low}}$ sub-actions for each high-level choice, yielding $n_{\text{high}} \times n_{\text{low}}$ total arms. In particular, we set $n_{\text{high}} = 4$ and $n_{\text{low}} = 4$, making for 16 composite actions in total. At each round $t \in \{1, \ldots, 1000\}$, the environment draws a $d$-dimensional context $\mathbf{x}_t \in \mathbb{R}^d$ from a standard Gaussian distribution, normalizes $\mathbf{x}_t$ to unit length, and reveals it to the agent. Throughout our experiments, we set the dimension to $d = 10$. A given composite action (high, low) produces a linear reward of $r_t = \mathbf{x}_t^\top \boldsymbol{\theta}_{\text{reward}}^{(\text{high,low})} + \eta_t$, where $\eta_t$ is a mean-zero noise term. We sample $\boldsymbol{\theta}_{\text{reward}}^{(\text{high,low})}$ from a mean-zero Gaussian with variance 1.0 independently for each of the 16 possible arms. Meanwhile, cost is assumed to depend solely on the high-level action. Formally, choosing high-level arm $h$ at time $t$ incurs an expected cost

$$\boldsymbol{c}_t = \mathbf{x}_t^\top \boldsymbol{\theta}_{\text{cost}}^{(h)} + \xi_t, \tag{72}$$

with $\xi_t$ being a mean-zero noise term. In our experiments, $\boldsymbol{\theta}_{\text{cost}}^{(h)}$ was drawn from a mean-zero Gaussian of variance 0.5; if $c_t > 0.7$, then we counted a constraint violation. We evaluated the following three methods:

- *HC-UCB:* A hierarchical constrained UCB approach that maintains separate confidence-bound models for high-level reward, sub-level reward, and high-level cost. It enforces a cost lower confidence bound $\leq 0.7$ for feasibility, then picks the sub-arm with the largest reward UCB among feasible high-level arms.
- *UCB:* A standard linear UCB algorithm that treats all (high, low) combinations as distinct arms. It ignores cost constraints entirely and only tries to maximize reward.
- *Random:* A purely random baseline that picks (high, low) uniformly at random each round, with no learning or constraints.

Figure 1 presents the cumulative reward for each algorithm over the time horizon. The UCB approach achieves the highest overall reward, as it is unconstrained and consequently more aggressive in exploiting high-reward arms. The HC-UCB method attains a moderately lower reward total, reflecting the tradeoff introduced by screening out high-level actions with potentially high costs. Meanwhile, the Random policy accumulates only minimal reward, as it makes no effort to identify better arms over time. These reward outcomes are presented in the cumulative regret plots, in which Random exhibits a steep linear growth, while both UCB-based methods constrain their regret to significantly smaller

(sublinear) magnitudes. Nonetheless, UCB marginally outperforms HC-UCB in regret, again owing to its ability to exploit the globally best arms even if they tend to incur higher costs. Finally, the bar chart shows the average number of constraint violations (i.e., cases in which $c_t > 0.7$). Here, HC-UCB records notably fewer high-level cost violations than either UCB or Random. The random strategy selects arms indiscriminately, often hitting high-cost arms. While the unconstrained UCB maximizes reward, it does not shy away from profitable but expensive high-level arms. By contrast, HC-UCB's lower-confidence cost bound mechanism rejects high-level arms when their estimated cost is likely to exceed the threshold, thereby demonstrating more consistent adherence to the constraint. Overall, these experiments indicate that the hierarchical constrained approach strikes a balance; it sacrifices a modest portion of reward relative to the fully unconstrained method, but achieves fewer constraint violations than either alternative.

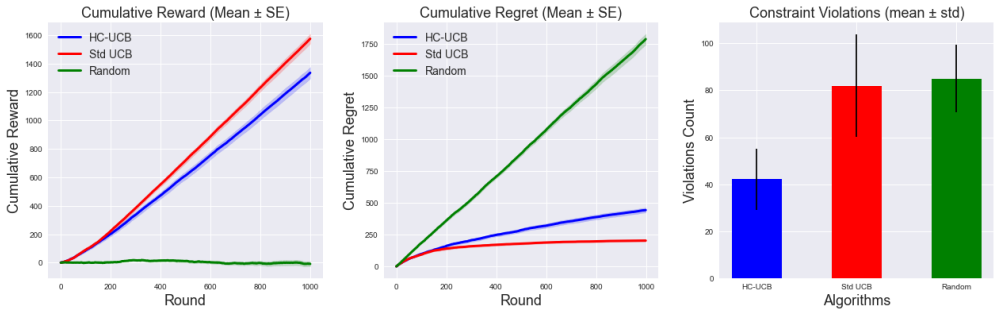



**Figure 1.** Performance comparison of HC-UCB, Standard UCB, and Random baselines across three metrics: (**Left**) cumulative reward over 1000 rounds (mean ± standard error), (**Center**) cumulative regret, and (**Right**) average constraint violations (mean ± standard deviation), demonstrating that HC-UCB more consistently adheres to the cost threshold than the alternatives.

**Limitations.** While the HC-UCB algorithm demonstrates theoretical guarantees and practical performance, several limitations remain. A key limitation stems from our theoretical framework's reliance on linear reward and cost functions. Real-world applications often exhibit complex nonlinear relationships that our current model cannot capture. Extending the framework to incorporate nonlinear functions, potentially through kernel methods or neural network approximations, would enhance its practical utility across diverse domains. Building on this consideration of real-world applications, our framework's assumption of stationary reward and cost distributions represents another limitation. Many practical scenarios involve dynamic environments where these distributions evolve over time. Future work should focus on adapting the algorithm to handle nonstationary environments while maintaining its theoretical guarantees. Finally, while our framework handles constraints across multiple levels, it currently optimizes for a single reward objective. This approach may not fully capture the complexity of real-world decision-making scenarios, which often involve multiple competing objectives. Developing extensions to handle multiple objectives within the hierarchical structure would enable more complex decision-making capabilities.

## 6. Conclusions

In this paper, we have presented the hierarchical constrained bandits (HCB) framework, which aims to address the limitations of traditional multi-armed bandit formulations in capturing hierarchical decision-making processes with multilevel constraints. The proposed HC-UCB algorithm extends the principles of the UCB approach to the hierarchical and constrained setting, balancing exploration and exploitation while ensuring constraint satisfaction at each level. Our theoretical contributions include proving that HC-UCB achieves sublinear regret. We have established high-probability guarantees for constraint satisfaction, ensuring that the algorithm adheres to the predefined thresholds $\tau_l$ at each

hierarchical level with probability at least $1 - \delta$. Furthermore, we have derived a minimax lower bound on the cumulative regret, indicating that HC-UCB is near-optimal up to logarithmic factors. The implications of our work are significant for a wide range of applications, including autonomous systems, resource allocation in cloud computing, personalized medicine, and smart grid management, all of which are areas where decision-making is complex, hierarchical, and constrained. By providing a theoretical foundation and an efficient algorithmic solution, we contribute to the advancement of sequential decision-making under uncertainty in hierarchical settings. Future research directions include exploring extensions of the HC-UCB algorithm to nonlinear reward and cost functions, incorporating richer contextual information, and addressing nonstationary environments in which the underlying reward and cost functions may change over time. Additionally, empirical evaluations in real-world scenarios would further validate the practical effectiveness of the HCB framework and HC-UCB algorithm.

**Funding:** This research received no external funding.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Conflicts of Interest:** The author declares no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| $T$ | Time horizon |
| $d$ | Dimension of context space |
| $H$ | Number of hierarchical levels |
| $A_h$ | Action set at level $h$ |
| $x_t$ | Context vector at time $t$ |
| $a_t^{(h)}$ | Action chosen at level $h$ at time $t$ |
| $r_t$ | Reward received at time $t$ |
| $c_t^{(h)}$ | Cost incurred at level $h$ at time $t$ |
| $\tau^{(h)}$ | Cost threshold at level $h$ |
| $\theta_r$ | Unknown reward parameter vector |
| $\theta_c^{(h)}$ | Unknown cost parameter vector at level $h$ |
| $\hat{\theta}_{r,t}$ | Estimated reward parameter vector at time $t$ |
| $\hat{\theta}_{c,t}^{(h)}$ | Estimated cost parameter vector at level $h$ at time $t$ |
| $\beta_t(\delta)$ | Confidence radius parameter |
| $\lambda$ | Regularization parameter |
| $V_t$ | Regularized covariance matrix at time $t$ |
| $R_T$ | Cumulative regret up to time $T$ |
| $R_H$ | High-level exploration regret |
| $R_L$ | Low-level exploration regret |
| $\delta$ | Confidence parameter |

## References

1. Villar, S.S.; Bowden, J.; Wason, J. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Stat. Sci. Rev. J. Inst. Math. Stat.* **2015**, *30*, 199. [CrossRef] [PubMed]
2. Avadhanula, V.; Colini Baldeschi, R.; Leonardi, S.; Sankararaman, K.A.; Schrijvers, O. Stochastic bandits for multi-platform budget optimization in online advertising. In Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 12–23 April 2021; pp. 2805–2817.
3. Maghsudi, S.; Hossain, E. Multi-armed bandits with application to 5G small cells. *IEEE Wirel. Commun.* **2016**, *23*, 64–73. [CrossRef]
4. Gittins, J.; Glazebrook, K.; Weber, R. *Multi-Armed Bandit Allocation Indices*; John Wiley & Sons: Hoboken, NJ, USA, 2011.

5. Agarwal, A.; Bird, S.; Cozowicz, M.; Hoang, L.; Langford, J.; Lee, S.; Li, J.; Melamed, D.; Oshri, G.; Ribas, O.; et al. Making contextual decisions with low technical debt. *arXiv* **2016**, arXiv:1606.03966.
6. Zhou, L. A survey on contextual multi-armed bandits. *arXiv* **2015**, arXiv:1508.03326.
7. Beygelzimer, A.; Langford, J.; Li, L.; Reyzin, L.; Schapire, R. Contextual bandit algorithms with supervised learning guarantees. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 11–13 April 2011; JMLR Workshop and Conference Proceedings, pp. 19–26.
8. Baheri, A.; Alm, C.O. LLMs-augmented Contextual Bandit. *arXiv* **2023**, arXiv:2311.02268.
9. Wu, Y.; Shariff, R.; Lattimore, T.; Szepesvári, C. Conservative bandits. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 1254–1262.
10. Amani, S.; Alizadeh, M.; Thrampoulidis, C. Linear stochastic bandits under safety constraints. *arXiv* **2019**, arXiv:1908.05814.
11. Agrawal, S.; Devanur, N.R. Bandits with concave rewards and convex knapsacks. In Proceedings of the Fifteenth ACM Conference on Economics and Computation, Palo Alto, CA, USA, 8–12 June 2014; pp. 989–1006.
12. Hsu, C.J.; Nair, V.; Menzies, T.; Freeh, V.W. Scout: An experienced guide to find the best cloud configuration. *arXiv* **2018**, arXiv:1803.01296.
13. Steiger, J.; Li, B.; Ji, B.; Lu, N. Constrained bandit learning with switching costs for wireless networks. In Proceedings of the IEEE INFOCOM 2023-IEEE Conference on Computer Communications, New York, NY, USA, 17–20 May 2023; pp. 1–10.
14. Hoffman, M.; Shahriari, B.; Freitas, N. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In Proceedings of the Artificial Intelligence and Statistics, Reykjavik, Iceland, 22–25 April 2014; pp. 365–374.
15. Robbins, H. Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.* **1952**, *58*, 527–535. [CrossRef]
16. Lai, T.L.; Robbins, H. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **1985**, *6*, 4–22. [CrossRef]
17. Auer, P. Finite-time analysis of the multi-armed bandit problem. In Proceedings of the International Conference on Machine Learning, Las Vegas, NV, USA, 24–27 June 2002.
18. Li, L.; Chu, W.; Langford, J.; Schapire, R.E. A contextual-bandit approach to personalized news article recommendation. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 661–670.
19. Badanidiyuru, A.; Kleinberg, R.; Slivkins, A. Bandits with knapsacks. *J. ACM (JACM)* **2018**, *65*, 1–55. [CrossRef]
20. Agrawal, S.; Devanur, N.R.; Li, L. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In Proceedings of the Conference on Learning Theory, New York, NY, USA, 23–26 June 2016; pp. 4–18.
21. Kazerouni, A.; Ghavamzadeh, M.; Abbasi Yadkori, Y.; Van Roy, B. Conservative contextual linear bandits. *arXiv* **2017**, arXiv:1611.06426.
22. Sani, A.; Lazaric, A.; Munos, R. Risk-aversion in multi-armed bandits. *arXiv* **2012**, arXiv:1301.1936.
23. Gu, S.; Yang, L.; Du, Y.; Chen, G.; Walter, F.; Wang, J.; Yang, Y.; Knoll, A. A review of safe reinforcement learning: Methods, theory and applications. *arXiv* **2022**, arXiv:2205.10330. [CrossRef] [PubMed]
24. Yifru, L.; Baheri, A. Concurrent learning of control policy and unknown safety specifications in reinforcement learning. *IEEE Open J. Control Syst.* **2024**, *3*, 266–281. [CrossRef]
25. Baheri, A.; Nageshrao, S.; Tseng, H.E.; Kolmanovsky, I.; Girard, A.; Filev, D. Deep reinforcement learning with enhanced safety for autonomous highway driving. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2019; pp. 1550–1555.
26. Baheri, A. Safe reinforcement learning with mixture density network, with application to autonomous driving. *Results Control Optim.* **2022**, *6*, 100095. [CrossRef]
27. Garcıa, J.; Fernández, F. A comprehensive survey on safe reinforcement learning. *J. Mach. Learn. Res.* **2015**, *16*, 1437–1480.
28. Moradipari, A.; Amani, S.; Alizadeh, M.; Thrampoulidis, C. Safe linear Thompson sampling with side information. *IEEE Trans. Signal Process.* **2021**, *69*, 3755–3767. [CrossRef]
29. Pacchiano, A.; Ghavamzadeh, M.; Bartlett, P.; Jiang, H. Stochastic bandits with linear constraints. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Virtual, 13–15 April 2021; pp. 2827–2835.
30. Barto, A.G.; Mahadevan, S. Recent advances in hierarchical reinforcement learning. *Discret. Event Dyn. Syst.* **2003**, *13*, 341–379. [CrossRef]
31. Bacon, P.L.; Harb, J.; Precup, D. The option-critic architecture. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
32. Guo, Q.; Wang, S.; Zhu, J. Regret Analysis for Hierarchical Experts Bandit Problem. *arXiv* **2022**, arXiv:2208.05622.
33. Hong, J.; Kveton, B.; Katariya, S.; Zaheer, M.; Ghavamzadeh, M. Deep hierarchy in bandits. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 8833–8851.

34. Sui, Y.; Gotovos, A.; Burdick, J.; Krause, A. Safe exploration for optimization with Gaussian processes. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015.
35. Abbasi-Yadkori, Y.; Pál, D.; Szepesvári, C. Improved algorithms for linear stochastic bandits. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 1–19.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.