Combinatorial Bandits with Linear Constraints: Beyond Knapsacks and Fairness

Qingsong Liu,† Weihang Xu, Siwei Wang, Zhixuan Fang

1 IIIS, Tsinghua University 2 Microsoft Research 3 Shanghai Qi Zhi Institute
{liu-qs19, xwh19}@mails.tsinghua.edu.cn
siweivang@microsoft.com
zfang@mail.tsinghua.edu.cn

Abstract

This paper proposes and studies for the first time the problem of combinatorial multi-armed bandits with linear long-term constraints. Our model generalizes and unifies several prominent lines of work, including bandits with fairness constraints, bandits with knapsacks (BwK), etc. We propose an upper-confidence bound LP-style algorithm for this problem, called UCB-LP, and prove that it achieves a logarithmic problem-dependent regret bound and zero constraint violations in expectation. In the special case of fairness constraints, we further provide a sharper constant regret bound for UCB-LP. Our regret bounds outperform the existing literature on BwK and bandits with fairness constraints simultaneously. We also develop another low-complexity version of UCB-LP and show that it yields \( \tilde{O}(\sqrt{T}) \) problem-independent regret and zero constraint violations with high-probability. Finally, we conduct numerical experiments to validate our theoretical results.

1 Introduction

In this paper, we study the problem of combinatorial bandits with long-term linear constraints. Our model captures important application scenarios like ad placement in online advertising systems [21], real-time traffic scheduling in wireless networks, and task assignment in crowdsourcing platforms [29], etc. Although being studied for the first time, our model subsumes several well-known problems in the Constrained Multi-Armed Bandit (CMAB) literature, including bandits with knapsacks, bandits with fairness constraints, etc. Details about these problems and how they fit into our framework are provided in Section 1.1.

Specifically, we consider an agent’s online decision problem faced with a fixed finite set of \( N \) arms labelled \( 1, 2, \ldots, N \), within the time horizon \( T \). At each round \( 1 \leq t \leq T \), every arm \( i \in [N] \) is associated with a random reward \( f_i(t) \in [0, 1] \) sampled from a time-invariant distribution \( P_i \). The reward \( f_i(t) \) and its distribution \( P_i \) are unknown to the agent a priori. The mean reward of distribution \( P_i \) is denoted as \( \mu_i \in [0, 1] \). We denote \( \mu = (\mu_1, \ldots, \mu_N) \top \in [0, 1]^N \) the mean reward vector, and define \( \mu^* := \max_{i \in [N]} \mu_i \) as the maximum mean reward value of all arms. At the beginning of each round \( t \), the agent is allowed to pull multiple, but no more than \( m \) arms. At round \( t \), the action taken by the agent is represented by an action vector \( a(t) = (a_1(t), \ldots, a_N(t)) \top \in \{0, 1\}^N \), where \( a_i(t) = 1 \) if and only if arm \( i \) is pulled. The set of all feasible action vectors is defined as \( \mathcal{A} = \{a \in \{0, 1\}^N, ||a||_1 \leq m\} \). After taking action \( a(t) \), the agent can observe reward from each pulled arm. Summing up the reward from the pulled arms, at round \( t \), the agent receives a total reward of \( R_t := \sum_{i=1}^{N} f_i(t)a_i(t) \).

†These authors contribute equally to this work.

Corresponding author: Zhixuan Fang (zfang@mail.tsinghua.edu.cn).

Beyond the standard combinatorial bandit setting above, we consider that the agent is subject to some constraints \(g(\cdot)\) at every round \(t\), defined as \(g(a(t)) := [g_1(a(t)), g_2(a(t)), \ldots, g_K(a(t))]^\top\), where \(g_1, g_2, \ldots, g_K : \mathbb{R}^N \to \mathbb{R}\) are linear functions. The goal of the agent is to maximize the accumulated expected reward up to the time horizon \(T\), while satisfying the constraints in the long term, i.e.,

\[
\max \sum_{t=1}^{T} R_t \quad \text{s.t.} \quad \sum_{t=1}^{T} g(a(t)) \leq 0.
\] (1)

(The comparison operator \(\leq\) is coordinate-wise) Define \(\text{OPT}(T)\) as the expected accumulated reward in \(T\) rounds of the optimal policy satisfying the long term constraints. The agent’s performance is measured in terms of regret and constraint violations defined respectively as

\[
\text{Regret}_T = \text{OPT}(T) - E[\sum_{t=1}^{T} R_t], \quad \text{Vio}(T) = \sum_{t=1}^{T} g(a(t)),
\]

where the expectation is taken w.r.t. the randomness of the reward and algorithm’s internal randomness. Consider the following linear programming problem (LP):

\[
\text{OPT}_{LP} = \max_{x \in \mathbb{R}^N} \mu^\top x \quad \text{s.t.} \quad g(x) \leq 0, \quad 0 \leq x \leq 1, \quad ||x||_1 \leq m.
\] (2)

\[\footnote{\[\text{4}\]}\] showed that \(\text{OPT}(T) \leq T \cdot \text{OPT}_{LP}\). We denote the set of optimal solutions to LP \(\[\text{2}\]\) as \(\mathcal{X}^*\), and define \(x^* \in \mathcal{X}^*\) as one of these optimal solutions. Following a common approach, we use the optimal randomized policy \(x^*\) as the benchmark in the regret:

\[
\text{Regret}_T \leq T \cdot \text{OPT}_{LP} - E[\sum_{t=1}^{T} R_t] = T(\mu, x^*) - E[\sum_{t=1}^{T} R_t].
\] (3)

### 1.1 Representative problems and related works

In this section, we outline several motivating and representative problems which fit into our general formulation and review the literature closely related to them (listed in Table [1]). Problem-dependent parameters \(\Delta_{\min}\) and \(\Delta\) will be formally defined in Section [2].

**Table 1: The comparison between our results and prior closely-related works.** In this table, \(r_{\min} = \min_i r_i\). “Single-arm” means \(m = 1\). “LP” means the algorithm should solve a linear program. “Complexity” refers to the computational-complexity of the algorithm at each round.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Setting</th>
<th>Assumptions</th>
<th>Regret</th>
<th>Violation</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>SINGLE-ARM, KNAPSACKS</td>
<td>N/A</td>
<td>(O(\min[N,K]^{4/3} \log^{2} T/\Delta_{\min}))</td>
<td>0</td>
<td>LP</td>
</tr>
<tr>
<td>19</td>
<td>SINGLE-ARM, KNAPSACKS</td>
<td>(</td>
<td>\mathcal{X}</td>
<td>= 1)</td>
<td>(O(\min[N,K]^{4/3} \log^{2} T/\Delta_{\min}))</td>
</tr>
<tr>
<td>20</td>
<td>SINGLE-ARM, KNAPSACKS</td>
<td>ONE RESOURCE ((N = 2))</td>
<td>(O(N \log^{2} T/\Delta_{\min}))</td>
<td>0</td>
<td>LP</td>
</tr>
<tr>
<td>21</td>
<td>SINGLE-ARM, KNAPSACKS</td>
<td>(O(\sqrt{m}))</td>
<td>(O(\log^{2} T/\Delta_{\min}))</td>
<td>0</td>
<td>LP</td>
</tr>
<tr>
<td>22</td>
<td>SINGLE-ARM, KNAPSACKS</td>
<td>(O(\log T/\Delta))</td>
<td>0</td>
<td>(O(\sqrt{m}))</td>
<td>(\sqrt{m})</td>
</tr>
<tr>
<td>23</td>
<td>SINGLE-ARM, FAIRNESS</td>
<td>N/A</td>
<td>(O(\sqrt{m} N \log T/\Delta_{\min}))</td>
<td>(O(T))</td>
<td>(O(N))</td>
</tr>
<tr>
<td>24</td>
<td>SINGLE-ARM, FAIRNESS</td>
<td>N/A</td>
<td>(O(\sqrt{m} N \log T))</td>
<td>(O(T))</td>
<td>(O(N))</td>
</tr>
<tr>
<td>25</td>
<td>SINGLE-ARM, FAIRNESS</td>
<td>(O(\sqrt{m} N \log T))</td>
<td>(O(\sqrt{m} N \log T))</td>
<td>(O(N^{2}))</td>
<td>(O(N^{2}))</td>
</tr>
<tr>
<td>26</td>
<td>SINGLE-ARM, FAIRNESS</td>
<td>N/A</td>
<td>(O(\sqrt{m} N \log T/\Delta_{\min}))</td>
<td>(O(T))</td>
<td>(O(N))</td>
</tr>
<tr>
<td>27</td>
<td>SINGLE-ARM, FAIRNESS</td>
<td>(O(\sqrt{m} N \log T))</td>
<td>(O(\sqrt{m} N \log T))</td>
<td>(O(N^{2}))</td>
<td>(O(N^{2}))</td>
</tr>
</tbody>
</table>

**Bandits with knapsacks.** The BwK (Bandits with Knapsacks) problem with deterministic costs studied in \[\text{18, 42, 40}\] assumes that there are \(K\) resources consumed over time, each with budgets \(B_1, B_2, \ldots, B_K\) respectively. Every resource \(i \in [K]\) is associated with a fixed consumption vector \(\lambda_i = (\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{iN}) \in \mathbb{R}^{N}_{\geq 0}\). The agent maximizes the accumulated reward subject to the budget constraints \(\sum_{i \in [T]} \lambda_i a(t) \leq B_i, \forall i \in [K]\). One can see that, this problem is equivalent with the special case of our setting where the linear constraint functions have the form \(g(a(t)) = (\lambda_1 a(t) - B_1/T, \ldots, \lambda_K a(t) - B_K/T)\).

The existing BwK literature \[\text{1, 4, 17}\] has derived algorithms that achieve the problem-independent regret bounds of the similar order \(O(\sqrt{KNT})\). Such order of the regret bound is also obtained
by [41, 42, 23] for combinatorial setting. However, the problem-dependent regret of BwK with deterministic costs is less explored. [18] achieved one regret bound of \( O \left( \min \{N, K\} \left( \frac{\log T}{\Delta_{\min}} \right) \right) \) with unsatisfying exponential dependence on \( K \) and \( N \). They also provide another regret bound of \( O((\min\{N, K\})^{\frac{3}{2}} \log T / \Delta_{\min}^{\frac{1}{2}}) \) with polynomial dependence on \( N, K \) but require an additional assumption of the optimal solution of LP (2) being unique. Later [40] improves this regret to \( O(N \frac{\log T}{\Delta_{\min}}) \) also under the assumption that LP (2) has an unique optimal solution. But their algorithm requires the knowledge of some parameters of the problem instance a priori (characterized by \( \Delta_{\min} \) and \( \mu^* \)), which is unpractical. With the “best-arm-optimality” assumption, i.e., there is an optimal policy that only pulls one arm, [42] achieved the regret of \( O(N \cdot \log T / G_{LAG}) \) (\( G_{LAG} \) is their defined Lagrangian gap) when there is only one resource (\( K = 2 \)), whose practicability is also restricted. Very recently, [15] derives a regret bound of \( O(\log^2 T) \) for combinatorial setting and one resource constraint, which is the best so far in the area of combinatorial BwK but still sub-optimal in terms of \( T \).

For BwK problems with stochastic costs, [18, 47, 46, 42, 30, 7] obtained logarithmic regrets under different restrictive assumptions, e.g., non-degeneracy of (2), \( K = 2 \) (single source) or “best-arm-optimality”. [18, 42] showed that it is impossible to achieve any problem-dependent regret bound of \( o(\sqrt{T}) \) without additional assumptions in general.

**Bandits with fairness constraints.** Recently, [29, 48, 39] studied the problem of bandits with fairness constraints. Under their setting, the agent maximizes the cumulative expected reward, and needs to ensure that each arm \( i \in [N] \) is pulled for at least \( \eta_i \in (0, 1) \) fraction of times at the end of \( T \) rounds. In our model, such fairness constraints are equivalent with the following special kind of linear long-term constraints: \( g_i(a(t)) = -a(t) + r, \) where \( r = (r_1, r_2, \ldots, r_N)^T \).

[29] first studied the combinatorial (sleeping) bandits with fairness constraints. Their algorithm LFG combines virtual queue technique and UCB learning. LFG yields \( O(\sqrt{T}) \) regret and sublinear \((o(T))\) constraint violations. [20] replaced the UCB learning with Thompson Sampling in LFG and obtained performance guarantees with the same order. Later [48] improved the constraint violations bound to \( O(\sqrt{T}) \) with the same regret order by using online convex optimization techniques and RRS rounding. A big advancement is made by recent works [12, 39] that they achieve \( O(\log T) \) and \( o(\log T) \) regret bounds, respectively, for MAB \((m = 1)\) with fairness constraints based on the modified UCB1 algorithm. And [16] achieved a “penalized” regret bound of \( O(\log T) \) in the single-arm setting.

**Bandits with group fairness.** In scenarios like ad-display optimization, the agent is subject to the group fairness constraint, e.g., arms belonging to one group should be pulled more frequently than arms belonging to another group [36], or the arm with higher average reward should be pulled more times than the arm with lower average reward [23], etc. This problem also fits into our formulation of linear constraints. Only problem-dependent regrets of \( O(\sqrt{T}) \) are obtained in related works.

**Other related literature.** A large body of literature (e.g., [10, 45]) derived \( O(\log T) \) regret bounds for unconstrained combinatorial bandits. [22, 24] studied the BwK problem under the adversarial setting. [2, 35, 5, 55] studied constrained linear bandits. Our framework is also related to online convex optimization with long-term constraints (e.g., [33, 11, 51, 50]), where the agent faces several convex constraints and these constraints need to be satisfied in the long term. We remark that, in this setting, the full reward function at each round would be revealed after the decision making, which is in contrast to our setting that we do not have such observation due to the semi-bandit feedback.

### 1.2 Discussion: significance of linear constraints

In this section, we discuss the significance of generalization to linear constraints. Previous literature on constrained bandits only studies constraints with specific forms, namely, the fairness constraint and the knapsacks constraint, etc. As we only require \( g_i(\cdot) \) to be linear, our model not only subsumes both of them as its special cases, but also solves a larger group of constraints, enabling broader applications. For example: 1. Our model for the first time addresses scenarios where both fairness and knapsack constraints exists simultaneously. 2. Our model solves the case of weighted fairness constraints, i.e., weighted sums of pulls of each arm are required to be larger than a threshold: \( \sum_{i=1}^N \kappa_{ij} h_i(T) \geq r_j T \), where \( \kappa_{ij} \) is the \( j^{th} \) weight of the \( i^{th} \) arm. In contrast, traditional fairness constraints treat each arm as unweighted and could be seen as a special case of weight fairness constraints.

In fact, there are many applications where combinatorial bandits with various complicated linear constraints are required. We show some concrete examples as follows. (a) In crowdsourcing, where a
group of workers are assigned with tasks to achieve high accuracy, a set of complex linear constraints
occurs: each tasks should have enough workers, while each worker should have a fair workload. (b) In
network routing where multiple paths (each path consists of a series of links) needs to be selected
to send the traffic within the time budget and the bandwidth constraints. (c) Case of Internet of Things
where a set of sensors need to be selected at each round to guarantee the QoS requirement [21] (e.g.,
throughout, mean-delay) under budget constraint on data collection cost (e.g., energy consumption).

1.3 Our contributions

Our main contributions are summarized below.

(a) We define a general formulation termed combinatorial bandits with linear long-term constraints.
In contrast to previously outlined pieces of work, we consider the problem in its full generality and do
not assume or require any prior knowledge of the problem instance. We design an upper-confidence
bound LP-style algorithm named UCB-LP, and develop a novel analytical technique to build a
relationship between LP solution (distribution support over arms) and “reward allocation”, with which
we show that UCB-LP guarantees a problem-dependent regret of $O(mN \sqrt{\frac{\log T}{\Delta m}})$ and no constraint
violation in expectation. To the best of our knowledge, this is the first logarithmic regret bound for
bandits with long term linear constraints under the combinatorial setting.

(b) For the special case of fairness constraints, we show that UCB-LP achieves $O(1)$ regret and
guarantees zero expected constraint violation at the same time. To the best of our knowledge, both
the regret and the constraint violations outperform all existing works on combinatorial bandits with
fairness constraints. We also show that UCB-LP has a low running time of $O(N)$ in this special case.

(c) To overcome the potentially time-consuming LP in UCB-LP, we further develop a low-complexity
version of UCB-LP, called UCB-PLLP, since it builds on the Lagrangian (L) of UCB-LP and pes-
simistically (P) tracks the constraint violations. We show that it yields $O(\sqrt{T})$ problem-independent
regret and guarantees zero constraint violations for any $\tau \leq T$ with high-probability. The computa-
tional complexity of UCB-PLLP is $O(N)$.

2 Main results

In this section, we present our algorithm and corresponding performance analysis for our general
formulation. All the proofs of listed lemmas, propositions and corollaries are deferred to the
supplementary material. Our experimental results are given in the supplementary material.

2.1 Preliminary: notations and existing techniques

Notations. For every arm $i$, define $h_i(t) := \sum_{\tau=1}^{t-1} a_{i}(\tau)$ as the number of pulls of it at the beginning
of round $t$, and $\bar{\mu}_i(t) := \frac{1}{h_i(t)} \sum_{\tau=1}^{t-1} a_{i}(\tau) f_i(\tau)$ as its empirical reward estimate at round $t$. Denote
the feasible region of LP (2) as $D := \{x | x \in [0, 1]^N, \|x\|_1 \leq m, g(x) \leq 0\}$. In this paper, for any vector
$v$, we use $v_i$ to denote its $i$th coordinate. For any event $E$, we use $\overline{E}$ to denote its negation.

Extreme points and general sub-optimality measure. Note that the feasible region $D$ to LP (2) is a
convex polytope, and we let $B$ be the set of its extreme points. An extreme point of $D$ is a point in $D
which does not lie in any open line segment joining two points of $D$. It is well-known in the theory
of LP that extreme points and basic feasible solutions are equivalent, and any LP attains its optimal
value at an extreme point [6]. Recall that $x^*$ is an optimal solution to LP (2), and we define the
sub-optimality gap for any $x \in B$ as $\Delta_x := \langle \mu, x^* \rangle - \langle \mu, x \rangle$. Define $\Delta_{\min} := \min_{x \in B, \forall \mu \in \Delta_x} \Delta_x$. Since
$B$ is finite, $\Delta_{\min}$ is well-defined and strictly positive. The same definition of the sub-optimality gap is
also used in [18]. We note that $\Delta_{\min}$ can be seen as a generalization of the minimum sub-optimality
gap in standard MAB problem defined as $\Delta := \min_{\mu, \forall \mu \neq \mu^*} |\mu^* - \mu|$. Specifically, under the standard
MAB setting, $m = 1, B = \{a | a \in [0, 1]^N, \|a\|_1 \leq 1\}, \Delta_{\min}$ coincides with $\Delta$. In this paper, we
will state our problem-dependent regret bounds in terms of $\Delta_{\min}$.

2.2 The general algorithm UCB-LP and its performance analysis

Now we introduce our algorithm UCB-LP for combinatorial bandits with long-term linear constraints.
UCB-LP is a generalization of SemiBwK algorithm [41] to the general linear constraints setting which
chooses arms through randomized arm selection. In our setting, one main challenge is that no super-arm is optimal across all rounds, but there exists an optimal allocation distribution over arms and the intuition behind UCB-LP is to identify such distribution and sampling arms based on it. At each round, UCB-LP consists of two stages. In the first stage, we first compute the truncated UCB estimate vector \( \hat{\mu}(t) \in \mathbb{R}^N \) defined as \( \hat{\mu}_i(t) = \min\left\{ \tilde{\mu}_i(t) + \frac{2\ln(t)}{h_i(t)} \right\}, \forall i \in [N] \). Then we solve the following LP and get an optimal solution \( x(t) \in B \):

\[
\max_{x \in \mathbb{R}^N} \langle \hat{\mu}(t), x \rangle \quad \text{s.t.} \quad g(x) \leq 0, \quad 0 \leq x \leq 1, \quad \|x\|_1 \leq m.
\]

(Note that LP (2) and LP (3) have the same feasible region \( D \), hence the same set of extreme points \( B \).) Here we require the solved optimal solution \( x(t) \) to be an extreme point of LP (3), i.e., \( x(t) \in B \). In fact, this is naturally satisfied by many LP algorithms, e.g., Simplex Method. Even if the optimal solution of LP is not an extreme point, it can be efficiently converted to one by tightening some slack constraints in [4] [57] [19].

The second stage is to construct a distribution \( \pi_t(\cdot) \) over \( A \) with expectation \( x(t) \), i.e., \( \sum_{a \in A} \pi_t(a) \cdot a = x(t) \), and sample super-arm \( a(t) \sim \pi_t \). Since \( x(t) \in D \subseteq \text{Conv}(A) \), such \( \pi_t \) exists and can be generated via a convex decomposition of \( x(t) \). Although there are many randomized rounding methods [32] with the idea of "regret allocation". The most natural idea to achieve this is to bound the number of times \( x(t) = \overline{x} \) for every sub-optimal policy \( \overline{x} \in B \setminus \mathcal{X}^* \). However, such an idea fails in the sense that the resulting bound would scale with \( |B| \), which is exponentially large. Addressing this technical challenge is nontrivial and previous works circumvented this problem by assuming special structures (e.g., [18] and [40] assumed \( |\mathcal{X}^*| = 1, [42] \) assumed only one resource and \( \|x^*\|_0 = 1 \), etc) or using the prior knowledge of some problem-dependent parameters [40]. In our proof, we overcome this difficulty directly by handling the case \( x(t) \notin \mathcal{X}^* \) with the idea of "regret allocation".

Define \( w(t) = (\sqrt{\frac{2\ln t}{h_t(1)}}, \ldots, \sqrt{\frac{2\ln t}{h_t(N)}})^T \). Since the regret only occurs when \( x \notin \mathcal{X}^* \), we claim that \( \langle \mu, x^* \rangle \leq \langle \hat{\mu}(t), x(t) \rangle, \langle \tilde{\mu}(t), x(t) \rangle \geq \langle \mu(t), x(t) \rangle + \Delta_x(t) \), and \( \langle \mu(t), x(t) \rangle \geq \Delta_x(t)/2 \) hold with high probability in such case since \( \langle \tilde{\mu}(t), x(t) \rangle \geq (\text{LP property}) \langle \hat{\mu}(t), x^* \rangle \geq (\text{high-prob}) \langle \mu, x^* \rangle \geq (\text{high-prob}) \Delta_x(t) \). With these properties, when \( x \notin \mathcal{X}^* \), we could allocate the incurred regret \( R_t = E[\langle \mu, x^* \rangle - \langle \mu, x(t) \rangle] = \Delta_x(t) \) to every base arm in the following way, and we will argue that this allocation is correct since:

\[
R_t = E[\langle \mu, x^* \rangle - \langle \mu, x(t) \rangle] = \Delta_x(t) \geq \Delta_x(t) - \Delta_x(t) = \Delta_x(t) - \Delta_x(t)
\]

\[
\leq (\text{high-prob})[\langle \tilde{\mu}(t) - \mu, x(t) \rangle] \leq (\text{high-prob})E[2w(t) x(t)] = 2E[\sum_{i \in [N]} x_i(t) \sqrt{\frac{2\ln t}{h_i(t)}}]
\]

Thus each arm contributes \( 2x_i(t) \sqrt{\frac{2\ln t}{h_i(t)}} \) to the regret \( \Delta_x(t) \). Next we claim that the sum above regret allocation on all arms is dominated by the arms in the set \( V(t) = \{ i : h_i(t) \leq \frac{32m^2 \ln t}{2\Delta_{\min}} \} \), i.e.,

\[
\sum_{i \in [N]} 2x_i(t) \sqrt{\frac{2\ln t}{h_i(t)}} \leq \sum_{i \in V(t)} 2x_i(t) \sqrt{\frac{2\ln t}{h_i(t)}}
\]

which is derived by:
\[ \sum_{i \in \{N\}/V(t)} x_i(t) \sqrt{\frac{2 \ln t}{h_i(t)}} \leq \sum_{i \in \{N\}/V(t)} x_i(t) \frac{\Delta_{\min}}{4} \leq \sum_{i \in \{N\}} x_i(t) \frac{\Delta_{\min}}{4} \leq \frac{\Delta_{\min}}{4}, \text{ and} \]

\[ \sum_{i \in \{N\}} x_i(t) \sqrt{\frac{2 \ln t}{h_i(t)}} = (w(t), x(t)) \geq \frac{\Delta_{\min}}{2}. \]

Therefore, the following total regret decomposition holds with high-probability:

\[ \text{Regret}_T \leq 2E \sum_{t=1}^{T} \sum_{i \in \{N\}} x_i(t) \sqrt{\frac{2 \ln t}{h_i(t)}} I\{x(t) \notin X^*\} \]

\[ \leq 4E \sum_{t=1}^{T} \sum_{i \in V(t)} x_i(t) \sqrt{\frac{2 \ln t}{h_i(t)}} I\{x(t) \notin X^*\} \leq 4E \sum_{t=1}^{T} \sum_{i \in V(t)} x_i(t) \sqrt{\frac{2 \ln t}{h_i(t)}}. \]

Define \( G(i) = \{ t : i \in V(t) \}, \text{ and } T_i = \arg \max_{\tau \in G(i)} \tau. \) Continuing from (7) we obtain

\[ \text{Regret}_T \leq 4E \sum_{t=1}^{T} \sum_{i \in V(t)} x_i(t) \sqrt{\frac{2 \ln t}{h_i(t)}} = 4E \sum_{t=1}^{T} \sum_{i \in V(t)} x_i(t) \sqrt{\frac{2 \ln t}{h_i(t)}} \leq 4E \sum_{t=1}^{T} \sum_{i = 1}^{T_i} x_i(t) \sqrt{\frac{2 \ln t}{h_i(t)}}. \]

The remaining problem is to bound \( 4E \sum_{t=1}^{T} \sum_{i = 1}^{T_i} x_i(t) \sqrt{\frac{2 \ln t}{h_i(t)}} \), which is solved by the following lemma.

**Lemma 1** \[ E \left[ \sum_{\tau=1}^{T_i} x_i(\tau) \sqrt{\ln \tau / h_i(\tau)} \right] \leq 3 \sqrt{\ln t \cdot E \left[ 1 / h_i(t) + 1 \right].} \]

Applying Lemma [1] yields \( 4E \sum_{t=1}^{T} \sum_{i \in \{N\}} x_i(t) \sqrt{\frac{2 \ln t}{h_i(t)}} \leq 12 \sum_{t=1}^{T} \sum_{i \in \{N\}} \sqrt{2 \ln T_i (h_i(T_i) + 1)} \leq 12 \sum_{t=1}^{T} \sum_{i \in \{N\}} \sqrt{2 \ln T \left( \frac{32m^2 \ln T}{\Delta_{\min}^2} + 1 \right)} = O(\sqrt{\ln T}). \) Putting all things together completes the proof of Theorem [1].

**Proof of Lemma 1:** Our proof idea is using the facts (a): \( 1 / \sqrt{x} \leq 3(\sqrt{1 + \sqrt{x}} - \sqrt{x}), \ x \geq 1; \)

(b): \( E \left[ 1 / h_i(t) + 1 \right] \neq 1 \), \( h_i(t) \neq 1 \);

(c): \( E \left[ 1 / h_i(t) + 1 \right] \neq 1 \), \( h_i(t) \neq 1 \);

(d): Almost all of them require additional assumptions or knowledge of some parameters of the problem instance a priori. For example, [18] explicitly admitted that their (poly-)logarithmic regret and the corresponding analysis only valid under the assumption of \( |x^*| = 1 \). (See Appendix A.1 and Assumption 9 in their arxiv version.) However, \( |x^*| = 1 \) does not always hold. When consider BwK problem with only one resource, if there exist arms \( p \), \( g \) such that \( \frac{\mu_e}{\lambda_e} = \frac{\mu_a}{\lambda_a} \), where \( \lambda_i \) is the amount of resource consumed by arm \( i \) if it gets pulled, then LP \( [2] \) has no-unique optimal solution, i.e., \( |x^*| = 1 \). Even consider problem of bandits with fairness constraints, if there exist two arms with the same required probability of being pulled, the optimal solution to LP \( [2] \) also may not be unique. Beyond assuming \( |x^*| = 1 \), the algorithms in [40] still require some prior knowledge of the problem instance (characterized by \( \mu^* \) and \( \Delta_{\min} \)). Substantial assumptions including only one resource and “best-arm-optimality” are also required in [42]. On the contrary, we do not require any assumptions or prior knowledge of the problem instance.

**Reduction to unconstrained (combinatorial) bandits setting.** When there are no constraints, i.e., \( g(t) = 0 \), UCB-LP reduces to the standard (Comb) UCB1 algorithm [25] and we could also recover the results of [25] in such case. This further justifies the tightness of our regret bound.

### 2.3 Achieving constant regret for fairness constraints

In this section, we consider the special case of fairness constraints. Following the same setting in the literature, e.g., [29, 43, 59], each arm \( i \in \{N\} \) is required to be pulled at least \( r_i \in (0, 1) \)
fraction of times, and \( \sum_{i=1}^{m} r_i < m \). Namely, the agent has to satisfy \( h_i(T) \geq r_i T, \forall i \in [N] \). Define \( r = (r_1, r_2, \ldots, r_N) \), then the equivalent linear long term constraint function \( g(\cdot) \) under our setting is \( g(a(t)) = -a(t) + r \). In the fairness constraints setting, the main challenge is that the time horizon \( T \) is unknown to the algorithm beforehand, and thus the “forced exploration” trick cannot be directly applied here. A lot of work (See Table 1) has sprung up recently to tackle this difficulty, but the best previous result is only \( \tilde{O}(\sqrt{T}) \) under the combinatorial setting. The following theorem shows that UCB-LP could guarantee a constant regret.

**Theorem 2** Define \( r_{\min} := \min_{i \in [N]} r_i \). In the case of fairness constraints, UCB-LP guarantees that
\[
\text{Regret}_{T} \leq \frac{32mN^2}{r_{\min}^2 \Delta_{\min}} \ln^2 \left( \frac{32N^2}{r_{\min}^2 \Delta_{\min}} \right) + \frac{mN^2}{2}.
\]

**Basic analysis of Theorem 2.** To derive a constant regret, our main idea is to show that the number of times \( E[h_i(T)] - r_i T \) is bounded for every sub-optimal arm \( i \) (whose mean reward is not \( \text{Top-m} \)). However, this is not the case in classic bandit analysis (e.g., for UCB-based algorithms) and thus cannot directly apply here. Our main proof idea is to show that UCB-LP only chooses a sub-optimal distribution over arms (i.e., \( x(t) \notin X^* \)) a limited number of times for fairness constraints. To achieve this, we first transform the event of \( x(t) \notin X^* \) to the events associated with arm’s UCB estimate error, i.e., we use the following lemma to characterize the case of \( x(t) \notin X^* \) by the sensitivity analysis of LP (\( \Omega(\text{log} T) \)).

**Lemma 2** If \( ||\hat{\mu}(t) - \mu||_1 < \Delta_{\min} \), then \( x(t) \in X^* \).

In other words, when \( x(t) \notin X^* \) we have \( ||\hat{\mu}(t) - \mu||_1 > \Delta_{\min} \). And then we just have to prove that \( ||\hat{\mu}(t) - \mu||_1 > \Delta_{\min} \Rightarrow \max_{i \in [N]} ||\hat{\mu}_i(t) - \mu_i||_1 > \Delta_{\min}/N \) only happens a finite number of times. This is a little counterintuitive for UCB-style algorithms, but it holds for UCB-LP in the fairness setting as UCB-LP guarantees that every arm has a positive probability of being pulled at each round which leads to the diminishing confidence width of all arms and the bonus of concentration inequality. In particular, by martingale analysis and extending the concentration bound to a random process that evolves over time, we prove that there exists a constant \( c > 0 \) (e.g., \( r_{\min}^2 \Delta_{\min} \ln^2 \left( \frac{32N^2}{r_{\min}^2 \Delta_{\min}} \right) \)) such that \( h_i(t) \geq \frac{s N^2}{\Delta_{\min}} \ln t \) holds with high-probability for each arm \( i \) when \( t > c \), which gives that \( ||\hat{\mu}_i(t) - \mu_i||_1 \leq \Delta_{\min}/N \) holds with high-probability when \( t > c \). It is worth noting that such properties do not hold for classic bandit algorithms (e.g., algorithms based on the UCB1 framework). And to the best of our knowledge there is no such results in the literature. Finally, to fit into our Lemma 2, we handle the regret as \( \text{Regret}_{T} \leq m \sum_{t=1}^{T} \Pr[x(t) \notin X^*] \leq mc + m \sum_{t=c+1}^{T} \Pr[x(t) \notin X^*] \leq mc + m \sum_{t=c+1}^{T} \Pr[||\hat{\mu}(t) - \mu||_1 \geq \Delta_{\min}] \). Put all things together then we complete the proof.

**Breaking the \( \Omega(\text{log} T) \) lower bound.** The constant regret bound given by Theorem 2 might seem counter-intuitive at first glance, since the lower bound of regret under classical MAB setting is \( \Omega(\log T) \). However, note that the fairness level \( r_i \) is required to be strictly positive, and the benchmark policy in the regret also needs to satisfy the long term constraints. This distinguishes our setting from the classical MAB setting. Thus, the traditional \( \Omega(\log T) \) lower bound no longer applies.

The reason why UCB-LP achieves a constant regret is: UCB-LP guarantees that every arm has a positive probability of being pulled at each round. This leads to more accurate reward estimates of arms and a diminishing gap between \( \hat{\mu}(t) \) and \( \mu \), which makes UCB-LP output the sub-optimal action distribution only a finite number of times. That said, it still requires non-trivial techniques to establish a constant regret bound. In fact, even \( O(\log T) \) regret bound has not yet been achieved for combinatorial setting in prior works.

Note that our regret bound has a quadratic dependence on \( 1/r_{\min} \), which goes to infinity as \( r \to 0 \). In fact, when \( r = 0 \), i.e., there are no fairness constraints, our problem degenerates to classical MAB setting where achieving constant regret bound is impossible. This indicates that such dependence on extra parameters is inevitable for any constant regret bound. Of course, the general \( O(\log T) \) regret bound developed in Theorem 2 still applies when \( r \to 0 \), since it does not depend on \( 1/r_{\min} \).

**\( \tilde{O}(N) \) running time.** The structural property of the fairness constraints allows us to obtain a closed form solution to LP (\( \Omega \)), as shown in the following proposition.

**Proposition 1** For each round \( t \), rearrange the coordinates of UCB estimate vector \( \hat{\mu}(t) \) in descending order such that \( \hat{\mu}_{\sigma_1}(t) \geq \hat{\mu}_{\sigma_2}(t) \geq \cdots \geq \hat{\mu}_{\sigma_N}(t) \), where \( \sigma_1, \ldots, \sigma_N \) is a permutation of \( 1, \ldots, N \). Then one of the optimal extreme points to LP (\( \Omega \)) (denoted as \( x(t) \)) has the following form:
With Corollary 1, we derive another constant regret bound of UCB-LP in the following theorem.

Although

Another constant regret bound. Proposition 1 further provides two important implications:

(a) Since LP (2) has the same form with LP (3), the optimal solution $x^*$ to LP (2) can also be written in closed form in the same manner to Proposition 1. Specifically, rearrange the coordinates of $\mu$ in descending order as $\mu_{s_1} \geq \mu_{s_2} \geq \cdots \geq \mu_{s_N}$, and define $k = \min\{q \in \mathbb{Z}^+ | \sum_{i=1}^q (1 - r_{s_i}) \geq m - \sum_{i=1}^N r_{s_i}\}$. Then $x^*$ has the following form:

$$x^*_{s_i} = 1, \forall i < k; \quad x^*_{s_i} = r_{s_i}, \forall i > k; \quad x^*_{s_k} = m + 1 - k - \sum_{i > k} r_{s_i}.$$ 

(b) The optimal solution $x(t)$ to LP (4) only depends on the relative order, not the absolute value of $\mu_{s_1}(t), \ldots, \mu_{s_N}(t)$. This motivates us to characterize the sensitivity of LP (4) from a new perspective. Intuitively, if $\hat{\mu}$ and $\mu$ are close enough such that the relative order between the coordinates is (at least partially) preserved, then the optimal solutions of LP (4) and LP (2) will coincide.

The above two implications motivate us to define a new parameter $\epsilon := \min_{i \neq k} |\mu_i - \mu_k|$, and propose the following corollary to characterize the sensitivity of LP (4).

**Corollary 1** If $||\hat{\mu}(t) - \mu||_{\infty} < \frac{\epsilon}{2}$, then $x(t) \in X^*$.

**Proof sketch of Corollary 1.** Proposition 1 shows that when $r$ is fixed, for $\forall i \neq s_k$, the value of $x_i(t)$ only depends on whether $\hat{\mu}_i(t) < \hat{\mu}_{s_k}(t)$ or not. Similarly, for $\forall i \neq s_k$, the value of $x_i^*$ only depends on whether $\mu_i < \mu_{s_k}$ or not. This implies that, if $k_1 = k, \sigma_{s_k} = \sigma_k$, and $\hat{\mu}_i(t) - \hat{\mu}_{s_k}(t)$ have the same sign with $\mu_i - \mu_{s_k}$, then $x_i(t) = x_i^*$. In other words, if $\mu_i > \mu_{s_k} \Rightarrow \hat{\mu}_i(t) > \hat{\mu}_{s_k}(t)$ and $\mu_i < \mu_{s_k} \Rightarrow \hat{\mu}_i(t) < \hat{\mu}_{s_k}(t)$ holds for $\forall i \in [N]$, then $x(t) = x^*$. When $||\hat{\mu}(t) - \mu||_{\infty} < \frac{\epsilon}{2}$, by the definition of $\epsilon$, for $\forall i$ we have $\mu_i > \mu_{s_k} \Rightarrow \mu_i = \mu_{s_k} + \epsilon \Rightarrow \hat{\mu}_i(t) = \mu_i + \frac{\epsilon}{2} \geq \mu_{s_k} + \frac{\epsilon}{2} \geq \hat{\mu}_{s_k}(t)$ and $\mu_i < \mu_{s_k} \Rightarrow \mu_i = \mu_{s_k} - \epsilon \Rightarrow \hat{\mu}_i(t) < \mu_i + \frac{\epsilon}{2} \leq \mu_{s_k} - \frac{\epsilon}{2} \leq \hat{\mu}_{s_k}(t)$. Then $x(t) = x^*$.

With Corollary 1, we derive another constant regret bound of UCB-LP in the following theorem.

**Theorem 3** UCB-LP also guarantees the following regret bound for fairness constraints,

$$\text{Regret}_{p_r} \leq \frac{64m}{r_{\min}^2} \ln^2 \left( \frac{64}{r_{\min}^2} \right) + \frac{mN\pi^2}{2}.$$ 

In the supplementary material, we show that $\Delta_{\min} = \epsilon = \Delta$ when our formulation reduces to the classical MAB setting, which suggests that $\Delta_{\min}$ and $\epsilon$ are both reasonable generalizations of $\Delta$. Although $\epsilon$ and $\Delta_{\min}$ are generally incomparable, we believe that the regret bound given by Theorem 3 is tighter than Theorem 2. To provide evidences for this, in supplementary material, we investigate the closed form expression of $\Delta_{\min}$. We show that $\epsilon = \Delta \geq \Delta_{\min} = \Delta(1 - \sum r_k)$ in the special case of $m = 1$. Intuitively, $\epsilon$ characterizes the structural properties of fairness constraints, while $\Delta_{\min}$ is more general and applies to all linear constraints.

**Remark 1** LP-sensitivity arguments are not new in bandit analyses and was firstly shown in [22] which used a technique based on LP-sensitivity to analyze the UcbBwK algorithm [1] for BwK problem in the single-arm setting. Here we would like to point out that the technique based on LP-sensitivity we used is completely different from theirs. Specifically, their analysis relies on the assumptions of single resource and that the best distribution over arms reduces to the best fixed arm, while we do not require any assumptions. Moreover, our technique is also non-standard as our sensitivity analysis (Lemma 2, Corollary 1) takes full advantage of the fairness structure (e.g., closed-form solution of LP). Thus, our result is sharper than standard results about LP sensitivity and leads to the $O(1)$ regret bound.
Algorithm 1 UCB-PLLP

1: Initialization: \( \mathcal{A} = \{\mathbf{x} | \mathbf{x} \in \{0, 1\}^N, ||\mathbf{x}||_1 \leq m\} \)
2: for round \( t = 1, \ldots, T - 1 \) do
3: \quad Compute UCBs: \( \hat{\mu}_i(t) = \min\{\bar{\mu}_i(t) + \sqrt{\frac{2 \ln t}{h_i(t)}}, 1\}, \forall i. \)
4: \quad Update the primal iterate: \( a(t) = \arg \max_{a \in \mathcal{A}} \langle \hat{\mu}(t) - \alpha_t \sum_{k=1}^K \nabla g_k(a(t-1))Q_k(t), a \rangle \)
5: \quad Play arm \( i \) and receive \( f_i(t) \) if \( a_i(t) = 1 \).
6: \quad Update the virtual queues: \( Q(t+1) = [Q(t) + g(a(t)) + \epsilon_t I]^{+} \).
7: \quad Update the statistics: \( h_i(t+1), \bar{\mu}_i(t+1), \forall i. \)
8: end for

Except [42] studies the BwK problem based on the LP methodology, [5] also develops an algorithm based on LP methodology to track the (contextual) blocking bandit problem wherein once an arm is pulled it cannot be played again for a fixed number of consecutive rounds. Here we argue that their LP-sensitivity-based proof techniques cannot obtain our regret bounds. Specifically, [5] utilizes techniques in [10] and [44], where [10] proposes the general CMAB framework and [44] improves upon their LP-sensitivity-based proof techniques cannot obtain our regret bounds. Specifically, [5] utilizes the well-established linear bandit algorithms like LinUCB [14] to track and identify the optimal arm for a given context. Remark 2 When the long-term constraints only need to be satisfied in expectation, someone may adopt the well-established linear bandit algorithms like LinUCB [14] to track and identify the optimal arm. However, this regret allocation leads to the arms in the support of LP solution associated with the same term while their probability of being triggered by the LP-solution is different. This coarseness makes the final regret bound to scale with the size of the support, transforming the regret bound from our result of \( \Theta(mN \ln T / \Delta_{min}) \) to \( \Theta(N^2 \ln T / \Delta_{min}) \).

### 2.4 \( \tilde{O}(N) \) running time version of UCB-LP and its theoretical performance

As mentioned earlier, UCB-LP might be computationally inefficient when constraint function \( g(\cdot) \) is complicated. In this section, we present the low computational-complexity version of UCB-LP, the UCB-PLLP algorithm, for problems of combinatorial bandits with linear long-term constraint. Note that the main computational bottleneck of UCB-LP is caused by the constraint \( g(x) \leq 0 \). To be computationally efficient, we optimize the (partial) Lagrangian of LP (4) at each round \( t \):

\[
\max_{\mathbf{x} \in \mathbb{R}^N} \mathcal{L}_t(\mathbf{x}, \lambda_t) = \langle \hat{\mu}(t), \mathbf{x} \rangle - \lambda_t^T \mathcal{G} \mathbf{x} \tag{8} \quad \text{s.t.} \quad 0 \leq \mathbf{x} \leq 1, ||\mathbf{x}||_1 \leq m.
\]

In (8) \( \lambda_t \) is the Lagrange multiplier associated with the constraints at round \( t \). The main challenge is to design \( \lambda_t \) to make a good balance between reward maximization and long-term constraints satisfaction. To address this challenge, we construct a virtual queue \( Q(t) \) to keep track of the “debt” of constraint violations up to round \( t \), i.e., \( Q(t) = [Q(t-1) + g(a(t-1))]^{+} \), and let \( \lambda_t = \alpha_t Q(t) \), where \( \alpha_t \) could be time-varying. However, this Lagrange multiplier desig may lead to large constraint violations as \( g(\cdot) \) becomes a “soft” constraint in LP (4) while in LP (4) it is a “hard” constraint.

To yield lower constraint violations, we incorporate the virtual queue update with “pessimistic” mechanism [35] so that the virtual queues overestimate the constraint violations, which is a novelty of our UCB-PLLP. Although the idea of using a “pessimistic” mechanism has been exploited in linear
(contextual) bandits with safety constraints (e.g., [2, 38], etc), the pessimism in our algorithm is achieved via adding a time-varying tightness constant to virtual queue update, i.e., $Q(t) = [Q(t-1) + g(a(t-1)) + \epsilon_t I]^+$, which is completely different from the pessimism in previous works. Since $g(\cdot)$ is linear, the optimization problem (8) has an integral closed-form solution and is equivalent to the following optimization problem:

$$
\max_x (\bar{\mu}(t) - \alpha_t \sum_{k \in [K]} \nabla g_k(a(t-1))Q_k(t), x) \quad s.t. \quad x \in [0, 1]^N, ||x||_1 \leq m.
$$

$$
\iff \max_a (\bar{\mu}(t) - \alpha_t \sum_{k \in [K]} \nabla g_k(a(t-1))Q_k(t), a) \quad s.t. \quad a \in \{0, 1\}^N, ||a||_1 \leq m.
$$

We illustrate this algorithmic approach in Algorithm 1. Note that if we set $\alpha_t = \eta$ and $\epsilon_t = 0$, the algorithm LBF in [29] is our special case as $\nabla g_k(\cdot) = -e_k$ and $K = N$ for fairness constraints setting. Apparently, the computational-complexity of UCB-PLL is essentially the same as choosing the top $m$ arms with maximum positive compound value, which is $\tilde{O}(N)$.

**Theoretical performance of UCB-PLL.** Here we present the regret bound and constraint violations for UCB-PLL. Our result relies on a mild assumption of Slater condition (Interior condition), i.e., there exists a $\delta > 0$ and $x \in D$ such that $g(x) \leq -\delta I$. Note that Slater condition automatically holds for fairness constraints as $r_i < 1, \forall i$, and $\sum_{i \in [N]} r_i < m$. It is also a default assumption in BwK literature (They all assume the null arm denoted by $0$ exists, i.e, $\lambda_i, 0 = 0, \forall i \in [K]$).

**Theorem 4** Set $\epsilon_t = O(\frac{\delta}{\sqrt{T}})$ and $\alpha_t = O(\frac{\sqrt{N}}{\sqrt{T}})$, then UCB-PLL achieves

$$
\text{Regret}_T \leq \tilde{O}(m \sqrt{T}), \quad \Pr(Viso_k(\tau) > 0) \leq O(e^{-\delta \sqrt{T}}), \forall k \in [K], \tau \leq T.
$$

We give the proof of Theorem 4 in supplementary material. Our proof technique is based on the Lyapunov-drift analysis for queueing systems. The bounds in Theorem 4 are sharp in the perspective that the regret bound matches the problem-independent regret of UCB1 algorithm in standard MAB literature (They all assume the null arm denoted by $0$ exists, i.e, $\lambda_i, 0 = 0, \forall i \in [K]$).

**Intuition behind Theorem 4.** Here we explain the intuition why UCB-PLL has such performance guarantees. Since $\alpha = O(1/\sqrt{T})$, the reward term dominates the whole term in (9) when $Q_k(t) = o(\sqrt{T})$. If $Q_k(t) = o(\sqrt{T})$, the term containing virtual queues dominates the reward term, and UCB-PLL tends to reduce the virtual queues $Q_k(t)$. Slater’s condition implies that there exists a policy that can reduce $Q_k(t)$ by a constant (related to $\delta$) in each round. Therefore, the algorithm takes at most $O(\sqrt{T})$ rounds to reduce $Q_k(t)$ to $o(\sqrt{T})$, which may lead to $O(\sqrt{T})$ increase of the regret. Thus, we could derive that $Q_k(t) = O(\sqrt{T})$. Recall that $Q_k(t+1) \geq \sum_{t=1}^T g_k(a(t)) + \sum_{t=1}^T \epsilon_t \Rightarrow Viso_k(T) = \sum_{t=1}^T g_k(a(t)) \leq Q_k(T+1) - \sum_{t=1}^T \epsilon_t$, we can obtain zero constraint violation via choosing proper $\epsilon_t$. Then, the high probability constraint violation guarantee is established by bounding the exponential moment of the virtual queues since $Pr(Viso_k(T) \geq 0) \leq Pr(Q_k(T+1) \geq \sum_{t=1}^T \epsilon_t) \leq E[e^{||Q(T+1)||_1}] / e^{\sum_{t=1}^T \epsilon_t}$.

**Remark 3** Although the virtual queue techniques have been used for various constrained online learning problems in the literature (e.g., [8, 9, 29, 43, 20, 49, 27, 35, 32, 37, 44]), our virtual queue update rule differs from theirs in the pessimistic mechanism via adding a time-varying tightness constant. Beyond the update rule of virtual queue, we also employ some new techniques in our Lyapunov analysis. For example, we establish a bound on the exponential moment of the virtual queue length (Lemma 6), which is the central focus of our high-probability guarantee on zero constraint violations. We also establish an upper bound on the $\epsilon_t$-tight term (incurred by our pessimistic mechanism) via comparing the optimal solution to the original LP problem and that to its $\epsilon_t$-tightened version based on LP-sensitivity (lemmas 3 and 4). These analysis techniques are not present in these works.

2.5 Experiments

The results of our numerical experiments are given in the supplementary material.

**Acknowledgments and Disclosure of Funding**

The work of Siwei Wang is supported in part by the National Natural Science Foundation of China Grant 62106122.
References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]