Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits

Branislav Kveton
Adobe Research
San Jose, CA
kveton@adobe.com

Zheng Wen
Yahoo Labs
Sunnyvale, CA
zhengwen@yahoo-inc.com

Azin Ashkan
Technicolor Labs
Los Altos, CA
azin.ashkan@technicolor.com

Csaba Szepesvári
Department of Computing Science
University of Alberta
szepesva@cs.ualberta.ca

Abstract

A stochastic combinatorial semi-bandit is an online learning problem where at each step a learning agent chooses a subset of ground items subject to constraints, and then observes stochastic weights of these items and receives their sum as a payoff. In this paper, we close the problem of computationally and sample efficient learning in stochastic combinatorial semi-bandits. In particular, we analyze a UCB-like algorithm for solving the problem, which is known to be computationally efficient; and prove $O(KL(1/\Delta)\log n)$ and $O(\sqrt{KLn\log n})$ upper bounds on its n-step regret, where L is the number of ground items, K is the maximum number of chosen items, and Δ is the gap between the expected returns of the optimal and best suboptimal solutions. The gap-dependent bound is tight up to a constant factor and the gap-free bound is tight up to a polylogarithmic factor.

1 Introduction

A *stochastic combinatorial semi-bandit* [11, 10] is an online learning problem where at each step a learning agent chooses a subset of ground items subject to combinatorial constraints, and then observes stochastic weights of these items and receives their sum as a payoff. The problem can be viewed as a learning variant of combinatorial optimization with a linear objective function and binary variables. Many classical combinatorial optimization problems have linear objectives [16]. Therefore, stochastic combinatorial

semi-bandits have found many practical applications, such as learning spectrum allocations [11], shortest paths [11], routing networks [12], and recommendations [12, 13].

In our work, we study a variant of stochastic combinatorial semi-bandits where the learning agent has access to an *offline optimization oracle* that can find the optimal solution for any weights of the items. We say that the problem is a (L, K, Δ) instance when L is the cardinality of its ground set E, K is the maximum number of chosen items, and Δ is the gap between the expected returns of the optimal and best suboptimal solutions. We also say that the problem is a (L, K) instance if it is a (L, K, Δ) instance for some Δ. Based on the existing bandit work [6], it is relatively easy to propose a UCB-like algorithm for solving our problem [11], and we call this algorithm `CombUCB1`. `CombUCB1` is a variant of `UCB1` that calls the oracle to find the optimal solution with respect to the upper confidence bounds on the weights of the items. Chen *et al.* [10] recently showed that the n-step regret of `CombUCB1` in any (L, K, Δ) stochastic combinatorial semi-bandit is $O(K^2L(1/\Delta)\log n)$.

Our main contribution is that we derive two upper bounds on the n-step regret of `CombUCB1`, $O(KL(1/\Delta)\log n)$ and $O(\sqrt{KLn\log n})$. Both of these bounds are significant improvements over Chen *et al.* [10]. Moreover, we prove two novel lower bounds, $\Omega(KL(1/\Delta)\log n)$ and $\Omega(\sqrt{KLn})$, which match our upper bounds up to polylogarithmic factors. The consequence of these results is that `CombUCB1` is *sample efficient* because it achieves near-optimal regret. It is well known that `CombUCB1` is also *computationally efficient* [11], it can be implemented efficiently whenever the offline optimization oracle is computationally efficient. So we close the problem of computationally and sample efficient learning in stochastic combinatorial semi-bandits, by showing that `CombUCB1` has both properties. This problem is still open in the adversarial setting (Section 8).

Our analysis is novel. It is based on the idea that the event that "many" items in a chosen suboptimal solution are not observed "sufficiently often" does not happen "too often".

The reason is that this event happens for "many" items simultaneously. Therefore, when the event happens, the observation counters of "many" items increase. Based on this observation, we divide the regret associated with the event among "many" items, instead of attributing it separately to each item as in the prior work [11, 10]. This is the key step in our analysis that yields tight upper bounds.

Our paper is organized as follows. In Section 2, we introduce our learning problem and the algorithm for solving it. In Section 3, we summarize our results. In Section 4, we prove a $O(K^{\frac{4}{3}}L(1/\Delta)\log n)$ upper bound on the regret of CombUCB1. In Section 5, we prove a $O(KL(1/\Delta)\log n)$ upper bound on the regret of CombUCB1. In Section 6, we prove gap-dependent and gap-free lower bounds. In Section 7, we evaluate CombUCB1 on a synthetic problem and show that its n-step regret grows as suggested by our gap-dependent upper bound. In Section 8, we compare our results to prior work. In Section 9, we discuss extensions of our work. We conclude in Section 10.

2 Setting

Formally, a *stochastic combinatorial semi-bandit* is a tuple $B = (E, \Theta, P)$, where $E = \{1, \ldots, L\}$ is a finite set of L items, $\Theta \subseteq 2^E$ is a non-empty set of feasible subsets of E, and P is a probability distribution over a unit cube $[0,1]^E$. We borrow the terminology of combinatorial optimization and call E the *ground set*, Θ the *feasible set*, and $A \in \Theta$ a *solution*. The items in the ground set E are associated with a vector of stochastic *weights* $w \sim P$. The e-th entry of w, $w(e)$, is the weight of item e. The expected weights of the items are defined as $\bar{w} = \mathbb{E}_{w \sim P}[w]$. The *return* for choosing solution A under the realization of the weights w is:

$$f(A, w) = \sum_{e \in A} w(e).$$

The maximum number of chosen items is defined as $K = \max_{A \in \Theta} |A|$.

Let $(w_t)_{t=1}^n$ be an i.i.d. sequence of n weights drawn from P. At time t, the learning agents chooses solution $A_t \in \Theta$ based on its observations of the weights up to time t, gains $f(A_t, w_t)$, and observes the weights of all chosen items at time t, $\{(e, w_t(e)) : e \in A_t\}$. The learning agent interacts with the environment n times and its goal is to maximize its expected cumulative reward over this time. If the agent knew P a priori, the optimal action would be to choose the *optimal solution*[1]:

$$A^* = \arg\max_{A \in \Theta} f(A, \bar{w})$$

at all steps t. The quality of the agent's policy is measured

---

[1]For simplicity of exposition, we assume that the optimal solution is unique.

Algorithm 1 CombUCB1 for stochastic combinatorial semi-bandits.

Input: Feasible set Θ

// Initialization
$(\hat{w}_1, t_0) \leftarrow \text{Init}(\Theta)$
$T_{t_0-1}(e) \leftarrow 1 \qquad \forall e \in E$

for all $t = t_0, \ldots, n$ **do**
  // Compute UCBs
  $U_t(e) \leftarrow \hat{w}_{T_{t-1}(e)}(e) + c_{t-1,T_{t-1}(e)} \qquad \forall e \in E$

  // Solve the optimization problem
  $A_t \leftarrow \arg\max_{A \in \Theta} f(A, U_t)$

  // Observe the weights of chosen items
  Observe $\{(e, w_t(e)) : e \in A_t\}$, where $w_t \sim P$

  // Update statistics
  $T_t(e) \leftarrow T_{t-1}(e) \qquad \forall e \in E$
  $T_t(e) \leftarrow T_t(e) + 1 \qquad \forall e \in A_t$
  $\hat{w}_{T_t(e)}(e) \leftarrow \dfrac{T_{t-1}(e)\hat{w}_{T_{t-1}(e)}(e) + w_t(e)}{T_t(e)} \qquad \forall e \in A_t$

by its *expected cumulative regret*:

$$R(n) = \mathbb{E}\left[\sum_{t=1}^n R(A_t, w_t)\right],$$

where $R(A_t, w_t) = f(A^*, w_t) - f(A_t, w_t)$ is the regret of the agent at time t.

2.1 Algorithm

Gai *et al.* [11] proposed a simple algorithm for stochastic combinatorial semi-bandits. The algorithm is motivated by UCB1 [6] and therefore we call it CombUCB1. At each time t, CombUCB1 consists of three steps. First, it computes the *upper confidence bound (UCB)* on the expected weight of each item e:

$$U_t(e) = \hat{w}_{T_{t-1}(e)}(e) + c_{t-1,T_{t-1}(e)}, \qquad (1)$$

where $\hat{w}_s(e)$ is the average of s observed weights of item e, $T_t(e)$ is the number of times that item e is observed in t steps, and:

$$c_{t,s} = \sqrt{\frac{1.5\log t}{s}} \qquad (2)$$

is the radius of a confidence interval around $\hat{w}_s(e)$ at time t such that $\bar{w}(e) \in [\hat{w}_s(e) - c_{t,s}, \hat{w}_s(e) + c_{t,s}]$ holds with high probability. Second, CombUCB1 calls the *optimization oracle* to solve the combinatorial problem on the UCBs:

$$A_t = \arg\max_{A \in \Theta} f(A, U_t).$$

Finally, CombUCB1 chooses A_t, observes the weights of all chosen items, and updates the estimates of $\bar{w}(e)$ for these items. The pseudocode of CombUCB1 is in Algorithm 1.

Algorithm 2 `Init`: Initialization of `CombUCB1`.

Input: Feasible set Θ

$$\hat{w}(e) \leftarrow 0 \qquad\qquad\qquad \forall e \in E$$
$$u(e) \leftarrow 1 \qquad\qquad\qquad \forall e \in E$$
$$t \leftarrow 1$$
while $(\exists e \in E : u(e) = 1)$ **do**
  $A_t \leftarrow \arg\max_{A \in \Theta} f(A, u)$
  Observe $\{(e, w_t(e)) : e \in A_t\}$, where $w_t \sim P$
  for all $e \in A_t$ **do**
    $\hat{w}(e) \leftarrow w_t(e)$
    $u(e) \leftarrow 0$
  $t \leftarrow t + 1$

Output:
  Weight vector $\hat{w}$
  First non-initialization step t

2.2 Initialization

`CombUCB1` is initialized by calling procedure `Init` (Algorithm 2), which returns two variables. The first variable is a weight vector $\hat{w} \in [0, 1]^E$, where $\hat{w}(e)$ is a single observation from the e-th marginal of P. The second variable is the number of initialization steps plus one.

The weight vector $\hat{w}$ is computed as follows. `Init` repeatedly calls the oracle $A_t = \arg\max_{A \in \Theta} f(A, u)$ on a vector of auxiliary weights $u \in \{0, 1\}^E$, which are initialized to ones. When item e is observed, we set the weight $\hat{w}(e)$ to the observed weight of the item and $u(e)$ to zero. `Init` terminates when $u(e) = 0$ for all items e. Without loss of generality, let's assume that each item e is contained in at least one feasible solution. Then `Init` is guaranteed to terminate in at most L iterations, because at least one entry of u changes from one to zero after each call of the optimization oracle.

3 Summary of Main Results

We prove three upper bounds on the regret of `CombUCB1`. Two bounds depend on the gap Δ and one is gap-free:

Theorem 3 :	$O(K^{\frac{4}{3}} L(1/\Delta) \log n)$
Theorem 5 :	$O(KL(1/\Delta) \log n)$
Theorem 6 :	$O(\sqrt{KLn \log n})$.

Both gap-dependent bounds are major improvements over $O(K^2 L(1/\Delta) \log n)$, the best known upper bound on the n-step regret of `CombUCB1` [10]. The bound in Theorem 5 is asymptotically tighter than the bound in Theorem 3, but the latter is tighter for $K < (534/96)^3 < 173$.

One of the main contributions of our work is that we identify an algorithm for stochastic combinatorial semi-bandits

that is both computationally and sample efficient. The following are our definitions of computational and sample efficiency. We say that the algorithm is *computationally efficient* if it can be implemented efficiently whenever the offline variant of the problem can be solved computationally efficiently. The algorithm is *sample efficient* if it achieves optimal regret up to polylogarithmic factors. Based on our definitions, `CombUCB1` is both *computationally* and *sample efficient*. We state this result slightly more formally below.

Theorem 1. `CombUCB1` *is computationally and sample efficient in any* (L, K) *stochastic combinatorial semi-bandit where the offline optimization oracle* $\arg\max_{A \in \Theta} f(A, w)$ *can be implemented efficiently for any* $w \in (\mathbb{R}^+)^E$.

Proof. In each step t, `CombUCB1` calls the oracle once, and all of its remaining operations are polynomial in L and K. Therefore, `CombUCB1` is guaranteed to be computationally efficient when the oracle is computationally efficient.

`CombUCB1` is sample efficient because it achieves optimal regret up to polylogarithmic factors. In particular, the gap-dependent upper bound on the n-step regret of `CombUCB1` in Theorem 5 matches the lower bound in Proposition 1 up to a constant factor. In addition, the gap-free upper bound in Theorem 6 matches the lower bound in Proposition 2 up to a factor of $\sqrt{\log n}$. ∎

4 $O(K^{\frac{4}{3}})$ Upper Bounds

In this section, we prove two $O(K^{\frac{4}{3}} L(1/\Delta) \log n)$ upper bounds on the n-step regret of `CombUCB1`. In Theorem 2, we assume that the gaps of all suboptimal solutions are the same. In Theorem 3, we relax this assumption.

The *gap* of solution A is $\Delta_A = f(A^*, \bar{w}) - f(A, \bar{w})$. The results in this section are presented for their didactic value. Their proofs are simple. Yet they illustrate the main ideas that lead to the tight regret bounds in Section 5.

Theorem 2. *In any* (L, K, Δ) *stochastic combinatorial semi-bandit where* $\Delta_A = \Delta$ *for all suboptimal solutions* A, *the regret of* `CombUCB1` *is bounded as:*

$$R(n) \leq K^{\frac{4}{3}} L \frac{48}{\Delta} \log n + \left(\frac{\pi^2}{3} + 1 \right) KL.$$

The proof of Theorem 2 relies on two lemmas. In the first lemma, we bound the regret associated with the initialization of `CombUCB1` and the event that $\bar{w}(e)$ is outside of the high-probability confidence interval around $\hat{w}_{T_{t-1}(e)}(e)$.

Lemma 1. *Let:*

$$\mathcal{F}_t = \left\{ \Delta_{A_t} \leq 2 \sum_{e \in \bar{A}_t} c_{n, T_{t-1}(e)}, \; \Delta_{A_t} > 0 \right\} \qquad (3)$$

be the event that suboptimal solution A_t is "hard to distinguish" from A^* at time t, where $\tilde{A}_t = A_t \setminus A^*$. Then the regret of CombUCB1 is bounded as:

$$R(n) \le \mathbb{E}\left[\hat{R}(n)\right] + \left(\frac{\pi^2}{3} + 1\right) KL, \qquad (4)$$

where:

$$\hat{R}(n) = \sum_{t=t_0}^{n} \Delta_{A_t} \mathbb{1}\{\mathcal{F}_t\}. \qquad (5)$$

Proof. The claim is proved in Appendix A.1. ∎

Now we bound the regret corresponding to the events $\mathcal{F}_t$, the items in a suboptimal solution are not observed "sufficiently often" up to time t. To bound the regret, we define two events:

$$G_{1,t} = \Big\{ \text{at least } d \text{ items in } \tilde{A}_t \text{ were observed} \qquad (6)$$
$$\text{at most } \alpha K^2 \frac{6}{\Delta_{A_t}^2} \log n \text{ times} \Big\}$$

and:

$$G_{2,t} = \Big\{ \text{less than } d \text{ items in } \tilde{A}_t \text{ were observed} \qquad (7)$$
$$\text{at most } \alpha K^2 \frac{6}{\Delta_{A_t}^2} \log n \text{ times,}$$
$$\text{at least one item in } \tilde{A}_t \text{ was observed}$$
$$\text{at most } \frac{\alpha d^2}{(\sqrt{\alpha}-1)^2} \frac{6}{\Delta_{A_t}^2} \log n \text{ times} \Big\},$$

where $\alpha \ge 1$ and $d > 0$ are parameters, which are chosen later. The event $G_{1,t}$ happens when "many" chosen items, at least d, are not observed "sufficiently often" up to time t, at most $\alpha K^2 \frac{6}{\Delta_{A_t}^2} \log n$ times.

Events $G_{1,t}$ and $G_{2,t}$ are obviously mutually exclusive. In the next lemma, we prove that these events are exhaustive when event $\mathcal{F}_t$ happens. To prove this claim, we introduce new notation. We denote the set of items in $\tilde{A}_t$ that are not observed "sufficiently often" up to time t by:

$$S_t = \Big\{ e \in \tilde{A}_t : T_{t-1}(e) \le \alpha K^2 \frac{6}{\Delta_{A_t}^2} \log n \Big\}.$$

Lemma 2. *Let $\alpha \ge 1$, $d > 0$, and event $\mathcal{F}_t$ happen. Then either event $G_{1,t}$ or $G_{2,t}$ happens.*

Proof. By the definition of S_t, the following three events:

$$G_{1,t} = \{|S_t| \ge d\}$$
$$G_{2,t} = \Big\{ |S_t| < d, \ \Big[\exists e \in \tilde{A}_t : T_{t-1}(e) \le \frac{6\alpha d^2 \log n}{(\sqrt{\alpha}-1)^2 \Delta_{A_t}^2}\Big] \Big\}$$
$$\bar{G}_t = \Big\{ |S_t| < d, \ \Big[\forall e \in \tilde{A}_t : T_{t-1}(e) > \frac{6\alpha d^2 \log n}{(\sqrt{\alpha}-1)^2 \Delta_{A_t}^2}\Big] \Big\}$$

are exhaustive and mutually exclusive. Therefore, to prove that either $G_{1,t}$ or $G_{2,t}$ happens, it suffices to show that $\bar{G}_t$ does not happen. Suppose that event $\bar{G}_t$ happens. Then by the assumption that $\mathcal{F}_t$ happens and from the definition of $\bar{G}_t$, it follows that:

$$\Delta_{A_t} \le 2 \sum_{e \in \tilde{A}_t} \sqrt{\frac{1.5 \log n}{T_{t-1}(e)}}$$

$$= 2 \sum_{e \in \tilde{A}_t \setminus S_t} \sqrt{\frac{1.5 \log n}{T_{t-1}(e)}} + 2 \sum_{e \in S_t} \sqrt{\frac{1.5 \log n}{T_{t-1}(e)}}$$

$$< 2 \underbrace{|\tilde{A}_t \setminus S_t|}_{\le K} \sqrt{\frac{1.5 \log n}{\alpha K^2 \frac{6}{\Delta_{A_t}^2} \log n}} +$$

$$2 \underbrace{|S_t|}_{\le d} \sqrt{\frac{1.5 \log n}{\frac{\alpha d^2}{(\sqrt{\alpha}-1)^2} \frac{6}{\Delta_{A_t}^2} \log n}}$$

$$\le \frac{\Delta_{A_t}}{\sqrt{\alpha}} + \frac{\Delta_{A_t}(\sqrt{\alpha}-1)}{\sqrt{\alpha}} = \Delta_{A_t}.$$

This is clearly a contradiction. Therefore, event $\bar{G}_t$ cannot happen; and either $G_{1,t}$ or $G_{2,t}$ happens. ∎

Now we are ready to prove Theorem 2.

Proof. A detailed proof is in Appendix A.2. The key idea is to bound the number of times that events $G_{1,t}$ and $G_{2,t}$ happen in n steps. Based on these bounds, the regret associated with both events is bounded as:

$$\hat{R}(n) \le \left(\frac{\alpha}{d} K^2 + \frac{\alpha d^2}{(\sqrt{\alpha}-1)^2} \right) L \frac{6}{\Delta} \log n.$$

Finally, we choose $\alpha = 4$ and $d = K^{\frac{2}{3}}$, and substitute the above upper bound into inequality (4). ∎

Theorem 2 can be generalized to the problems with different gaps. Let $\Delta_{e,\min}$ be the minimum gap of any suboptimal solution that contains item $e \in \tilde{E}$:

$$\Delta_{e,\min} = \min_{A \in \Theta : e \in A, \Delta_A > 0} \Delta_A \qquad (8)$$
$$= f(A^*, \bar{w}) - \max_{A \in \Theta : e \in A, \Delta_A > 0} f(A, \bar{w}),$$

where $\tilde{E} = E \setminus A^*$ is the set of *subtimal items*, the items that do not appear in the optimal solution. Then the regret of CombUCB1 is bounded as follows.

Theorem 3. *In any (L, K) stochastic combinatorial semi-bandit, the regret of CombUCB1 is bounded as:*

$$R(n) \le \sum_{e \in \tilde{E}} K^{\frac{4}{3}} \frac{96}{\Delta_{e,\min}} \log n + \left(\frac{\pi^2}{3} + 1\right) KL.$$

Proof. A detailed proof is in Appendix A.3. The key idea is to define item-specific variants of events $G_{1,t}$ and $G_{2,t}$, $G_{e,1,t}$ and $G_{e,2,t}$; and associate $\frac{\Delta_{A_t}}{d}$ and Δ_{A_t} regret with $G_{e,1,t}$ and $G_{e,2,t}$, respectively. Then, for each item e, we order the events from the largest gap to the smallest, and show that the total regret is bounded as:

$$\hat{R}(n) < \sum_{e \in \tilde{E}} \left(\frac{\alpha}{d} K^2 + \frac{\alpha d^2}{(\sqrt{\alpha} - 1)^2} \right) \frac{12}{\Delta_{e,\min}} \log n \,.$$

Finally, we choose $\alpha = 4$ and $d = K^{\frac{2}{3}}$, and substitute the above upper bound into inequality (4). ∎

5 $O(K)$ **Upper Bounds**

In this section, we improve on the results in Section 4 and derive $O(KL(1/\Delta) \log n)$ upper bounds on the n-step regret of CombUCB1. In Theorem 4, we assume that the gaps of all suboptimal solutions are identical. In Theorem 5, we relax this assumption.

The key step in our analysis is that we define a cascade of infinitely-many mutually-exclusive events and then bound the number of times that these events happen when a suboptimal solution is chosen. The events are parametrized by two decreasing sequences of constants:

$$1 = \beta_0 > \beta_1 > \beta_2 > \ldots > \beta_k > \ldots \tag{9}$$
$$\alpha_1 > \alpha_2 > \ldots > \alpha_k > \ldots \tag{10}$$

such that $\lim_{i \to \infty} \alpha_i = \lim_{i \to \infty} \beta_i = 0$. We define:

$$m_{i,t} = \alpha_i \frac{K^2}{\Delta_{A_t}^2} \log n$$

and assume that $m_{i,t} = \infty$ when $\Delta_{A_t} = 0$. The events at time t are defined as:

$G_{1,t} = \{$at least $\beta_1 K$ items in $\tilde{A}_t$ were observed   (11)
  at most $m_{1,t}$ times$\}$,

$G_{2,t} = \{$less than $\beta_1 K$ items in $\tilde{A}_t$ were observed
  at most $m_{1,t}$ times,
  at least $\beta_2 K$ items in $\tilde{A}_t$ were observed
  at most $m_{2,t}$ times$\}$,

  $\vdots$

$G_{i,t} = \{$less than $\beta_1 K$ items in $\tilde{A}_t$ were observed
  at most $m_{1,t}$ times,
  $\ldots$,
  less than $\beta_{i-1} K$ items in $\tilde{A}_t$ were observed
  at most $m_{i-1,t}$ times,
  at least $\beta_i K$ items in $\tilde{A}_t$ were observed
  at most $m_{i,t}$ times$\}$,

  $\vdots$

The following lemma establishes a sufficient condition under which events $G_{i,t}$ are exhaustive. This is the key step in the proofs in this section.

Lemma 3. *Let (α_i) and (β_i) be defined as in* (9) *and* (10), *respectively; and let:*

$$\sqrt{6} \sum_{i=1}^{\infty} \frac{\beta_{i-1} - \beta_i}{\sqrt{\alpha_i}} \le 1 \,. \tag{12}$$

Let event $\mathcal{F}_t$ happen. Then event $G_{i,t}$ happens for some i.

Proof. We fix t such that $\Delta_{A_t} > 0$. Because t is fixed, we use shorthands $G_i = G_{i,t}$ and $m_i = m_{i,t}$. Let:

$$S_i = \left\{ e \in \tilde{A}_t : T_{t-1}(e) \le m_i \right\}$$

be the set of items in $\tilde{A}_t$ that are not observed "sufficiently often" under event G_i. Then event G_i can be written as:

$$G_i = \left(\bigcap_{j=1}^{i-1} \{|S_j| < \beta_j K\} \right) \cap \{|S_i| \ge \beta_i K\} \,.$$

As in Lemma 2, we prove that event G_i happens for some i by showing that the event that none of our events happen cannot happen. Note that this event can be written as:

$$
\begin{aligned}
\bar{G} &= \overline{\bigcup_{i=1}^{\infty} G_i} \\
&= \bigcap_{i=1}^{\infty} \left[\left(\bigcup_{j=1}^{i-1} \{|S_j| \ge \beta_j K\} \right) \cup \{|S_i| < \beta_i K\} \right] \\
&= \bigcap_{i=1}^{\infty} \{|S_i| < \beta_i K\} \,.
\end{aligned}
$$

Let $\bar{S}_i = \tilde{A}_t \setminus S_i$ and $S_0 = \tilde{A}_t$. Then by the definitions of $\bar{S}_i$ and S_i, $\bar{S}_{i-1} \subseteq \bar{S}_i$ for all $i > 0$. Furthermore, note that $\lim_{i \to \infty} m_i = 0$. So there must exist an integer j such that $\bar{S}_i = \tilde{A}_t$ for all $i > j$, and $\tilde{A}_t = \bigcup_{i=1}^{\infty} (\bar{S}_i \setminus \bar{S}_{i-1})$. Finally, by the definition of $\bar{S}_i$, $T_{t-1}(e) > m_i$ for all $e \in \bar{S}_i$. Now suppose that event $\bar{G}$ happens. Then:

$$
\begin{aligned}
\sum_{e \in \tilde{A}_t} \frac{1}{\sqrt{T_{t-1}(e)}} &< \sum_{i=1}^{\infty} \sum_{e \in \bar{S}_i \setminus \bar{S}_{i-1}} \frac{1}{\sqrt{m_i}} \\
&= \sum_{i=1}^{\infty} \frac{|\bar{S}_i \setminus \bar{S}_{i-1}|}{\sqrt{m_i}} \\
&< \sum_{i=1}^{\infty} \frac{(\beta_{i-1} - \beta_i) K}{\sqrt{m_i}} \,,
\end{aligned}
$$

where the last inequality is due to Lemma 4 (Appendix B).

In addition, let event $\mathcal{F}_t$ happen. Then:

$$
\begin{aligned}
\Delta_{A_t} &\leq 2 \sum_{e \in \tilde{A}_t} \sqrt{\frac{1.5 \log n}{T_{t-1}(e)}} \\
&< \sqrt{6 \log n} \sum_{i=1}^{\infty} \frac{(\beta_{i-1} - \beta_i) K}{\sqrt{m_i}} \\
&= \Delta_{A_t} \sqrt{6} \sum_{i=1}^{\infty} \frac{\beta_{i-1} - \beta_i}{\sqrt{\alpha_i}} \leq \Delta_{A_t},
\end{aligned}
$$

where the last inequality is due to our assumption in (12). The above is clearly a contradiction. As a result, $\bar{G}$ cannot happen, and event G_i must happen for some i. ■

Theorem 4. *In any (L, K, Δ) stochastic combinatorial semi-bandit where $\Delta_A = \Delta$ for all suboptimal solutions A, the regret of* CombUCB1 *is bounded as:*

$$
R(n) \leq KL \frac{267}{\Delta} \log n + \left(\frac{\pi^2}{3} + 1 \right) KL.
$$

Proof. A detailed proof is in Appendix A.4. The key idea is to bound the number of times that event $G_{i,t}$ happens in n steps for any i. Based on this bound, the regret due to all events $G_{i,t}$ is bounded as:

$$
\hat{R}(n) \leq KL \frac{1}{\Delta} \left[\sum_{i=1}^{\infty} \frac{\alpha_i}{\beta_i} \right] \log n,
$$

where $\hat{R}(n)$ is defined in (5). Finally, we choose (α_i) and (β_i), and apply the above upper bound in inequality (4). ■

Now we generalize Theorem 4 to arbitrary gaps.

Theorem 5. *In any (L, K) stochastic combinatorial semi-bandit, the regret of* CombUCB1 *is bounded as:*

$$
R(n) \leq \sum_{e \in \tilde{E}} K \frac{534}{\Delta_{e,\min}} \log n + \left(\frac{\pi^2}{3} + 1 \right) KL,
$$

where $\Delta_{e,\min}$ is the minimum gap of suboptimal solutions that contain item e, which is defined in (8).

Proof. A detailed proof is in Appendix A.5. The key idea is to introduce item-specific variants of events $G_{i,t}$, $G_{e,i,t}$, and associate $\frac{\Delta_{A_t}}{\beta_i K}$ regret with each of these events. Then, for each item e, we order the events from the largest gap to the smallest, and show that the total regret is bounded as:

$$
\hat{R}(n) < \sum_{e \in \tilde{E}} K \frac{2}{\Delta_{e,\min}} \left[\sum_{i=1}^{\infty} \frac{\alpha_i}{\beta_i} \right] \log n,
$$

where $\hat{R}(n)$ is defined in (5). Finally, we choose (α_i) and (β_i), and apply the above upper bound in inequality (4). ■

We also prove a gap-free bound.

Theorem 6. *In any (L, K) stochastic combinatorial semi-bandit, the regret of* CombUCB1 *is bounded as:*

$$
R(n) \leq 47 \sqrt{KLn \log n} + \left(\frac{\pi^2}{3} + 1 \right) KL.
$$

Proof. The proof is in Appendix A.6. The key idea is to decompose the regret of CombUCB1 into two parts, where the gaps are larger than ϵ and at most ϵ. We analyze each part separately and then set ϵ to get the desired result. ■

6 Lower Bounds

In this section, we derive two lower bounds on the n-step regret in stochastic combinatorial semi-bandits. One of the bounds is gap-dependent and the other one is gap-free.

Our bounds are derived on a K-*path semi-bandit* problem, which is illustrated in Figure 1a. The items in the ground set are L path segments $E = \{1, \ldots, L\}$. The feasible set Θ are L/K paths, each of which contains K unique items. Specifically, path j contains items $(j-1)K + 1, \ldots, jK$. Without loss of generality, we assume that L/K is an integer. The probability distribution P over the weights of the items is defined as follows. The weights of the items in the same path are equal. The weights of the items in different paths are distributed independently. The weight of item e is a Bernoulli random variable with mean:

$$
\bar{w}(e) = \begin{cases} 0.5 & \text{item } e \text{ belongs to path 1} \\ 0.5 - \Delta/K & \text{otherwise}, \end{cases}
$$

where $0 < \Delta/K < 0.5$. Note that our problem is designed such that $\Delta_{e,\min} = \Delta$ for any item e in path $j > 1$.

The key observation is that our problem is equivalent to a (L/K)-arm Bernoulli bandit whose returns are scaled by K, when the learning agent knows that the weights of the items in the same path are equal. Therefore, we can derive lower bounds for our problem based on the existing lower bounds for Bernoulli bandits [7, 8, 14].

Our gap-dependent lower bound is derived for the class of consistent algorithms, which is defined as follows. We say that the algorithm is *consistent* if for any stochastic combinatorial semi-bandit, any suboptimal A, and any $\alpha > 0$, $\mathbb{E}[T_n(A)] = o(n^\alpha)$, where $T_n(A)$ is the number of times that solution A is chosen in n steps. The restriction to the consistent algorithms is without loss of generality. In particular, an inconsistent algorithm is guaranteed to perform poorly on some semi-bandit, and therefore cannot achieve logarithmic regret on all semi-bandits, as CombUCB1.

Proposition 1. *For any L and K such that L/K is an integer, and any $0 < \Delta/K < 0.5$, the regret of any consistent algorithm on the K-path semi-bandit problem is bounded*
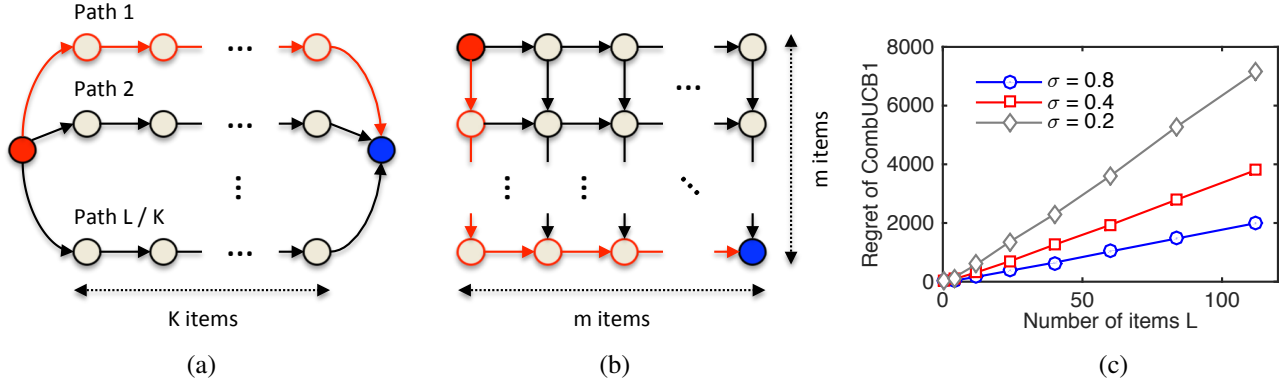
Figure 1: **a**. The K-path semi-bandit problem in Section 6. The red and blue nodes are the starting and end points of the paths, respectively. The optimal path is marked in red. **b**. The grid-path problem in Section 7. The red and blue nodes are the starting and end points of the paths, respectively. The optimal path is marked in red. **c**. The n-step regret of CombUCB1 on the grid-path problem.

from below as:

$$\liminf_{n \to \infty} \frac{R(n)}{\log n} \geq \frac{(L-K)K}{4\Delta} .$$

Proof. The proposition is proved as follows:

$$\liminf_{n \to \infty} \frac{R(n)}{\log n} \overset{(a)}{\geq} K \sum_{k=2}^{L/K} \frac{\Delta/K}{D_{\mathrm{KL}}(0.5 - \Delta/K \,\|\, 0.5)}$$
$$= \left(\frac{L}{K} - 1 \right) \frac{\Delta}{D_{\mathrm{KL}}(0.5 - \Delta/K \,\|\, 0.5)}$$
$$\overset{(b)}{\geq} \frac{(L-K)K}{4\Delta} ,$$

where $D_{\mathrm{KL}}(p \,\|\, q)$ is the Kullback-Leibler (KL) divergence between two Bernoulli random variables with means p and q. Inequality (a) follows from the fact that our problem is equivalent to a (L/K)-arm Bernoulli bandit whose returns are scaled by K. Therefore, we can bound the regret from below using an existing lower bound for Bernoulli bandits [14]. Inequality (b) is due to $D_{\mathrm{KL}}(p \,\|\, q) \leq \frac{(p-q)^2}{q(1-q)}$, where $p = 0.5 - \Delta/K$ and $q = 0.5$. ∎

We also derive a gap-free lower bound.

Proposition 2. *For any L and K such that L/K is an integer, and any horizon $n > 0$, there exists a K-path semi-bandit problem such that the regret of any algorithm is:*

$$R(n) \geq \frac{1}{20} \min(\sqrt{KLn}, Kn) .$$

Proof. The K-path semi-bandit problem is equivalent to a (L/K)-arm Bernoulli bandit whose payoffs are scaled by K. Therefore, we can apply Theorem 5.1 of Auer *et al.* [7] and bound the regret of any algorithm from below by:

$$\frac{K}{20} \min(\sqrt{(L/K)n}, n) = \frac{1}{20} \min(\sqrt{KLn}, Kn) .$$

Note that the bound of Auer *et al.* [7] is for the adversarial setting. However, the worst-case environment in the proof is stochastic and therefore it applies to our problem. ∎

7 Experiments

In this section, we evaluate CombUCB1 on a synthetic problem and demonstrate that its regret grows as suggested by our $O(KL(1/\Delta) \log n)$ upper bound. We experiment with a stochastic longest-path problem on a $(m+1) \times (m+1)$ square grid (Figure 1b). The items in the ground set E are the edges in the grid, $2m(m+1)$ in total. The feasible set Θ are all paths in the grid from the upper left corner to the bottom right corner that follow the directions of the edges. The length of these paths is $K = 2m$. The weight of edge e is drawn i.i.d. from a Bernoulli distribution with mean:

$$\bar{w}(e) = \begin{cases} 0.5 + \sigma/2 & e \text{ is a leftmost or bottommost edge} \\ 0.5 - \sigma/2 & \text{otherwise} , \end{cases}$$

where $0 < \sigma < 1$. The optimal solution A^* is a path along the leftmost and bottommost edges (Figure 1b).

The sample complexity of our problem is characterized by $|\tilde{E}| = 2m(m+1) - 2m$ gaps $\Delta_{e,\min}$ ranging from 2σ to $2m\sigma$. It is easy to show that the number of items e where $\Delta_{e,\min} = i\sigma$ is at most $2(i-1)$. Therefore, we can bound the $(\log n)$-term in Theorem 5 as:

$$\sum_{e \in \tilde{E}} K \frac{534}{\Delta_{e,\min}} \log n < 1068m \sum_{i=2}^{2m} \frac{2i}{i\sigma} \log n$$
$$< \frac{4272m^2 \log n}{\sigma} . \qquad (13)$$

Now we validate the dependence on m and σ empirically. We vary m and σ, and run CombUCB1 for $n = 10^5$ steps.

Our experimental results are reported in Figure 1c. We observe two trends. First, the regret of CombUCB1 is linear in

541

the number of items L, which depends quadratically on m since $L = 2m(m+1)$. Second, the regret is linear in $1/\sigma$. The dependence on m and $1/\sigma$ is the same as in our upper bound in (13).

8 Related Work

Gai *et al.* [11] proposed `CombUCB1` and analyzed it. Chen *et al.* [10] derived a $O(K^2L(1/\Delta)\log n)$ upper bound on the n-step regret of `CombUCB1`. In this paper, we show that the regret of `CombUCB1` is $O(KL(1/\Delta)\log n)$, a factor of K improvement over the upper bound of Chen *et al.* [10]. This upper bound is tight. We also prove a gap-free upper bound and show that it is nearly tight.

COMBAND [9], online stochastic mirror descent (OSMD) [4], and follow-the-perturbed-leader (FPL) with geometric resampling [15] are three recently proposed algorithms for adversarial combinatorial semi-bandits. In general, OSMD achieves optimal regret but is not guaranteed to be computationally efficient, in the same sense as in Section 3. FPL does not achieve optimal regret but is computationally efficient. It is an open problem whether adversarial combinatorial semi-bandits can be solved both computationally and sample efficiently. In this paper, we close this problem in the stochastic setting.

Matroid and polymatroid bandits [12, 13] are instances of stochastic combinatorial semi-bandits. The n-step regret of `CombUCB1` in these problems is $O(L(1/\Delta)\log n)$, a factor of K smaller than is suggested by our $O(KL(1/\Delta)\log n)$ upper bound. However, we note that the bound of Kveton *et al.* [12, 13] is less general, as it applies only to matroids and polymatroids.

Our problem can be viewed as a linear bandit [5, 1], where each solution A is associated with an indicator vector $\mathbf{x} \in \{0,1\}^E$ and the learning agent observes the weight of each non-zero entry of $\mathbf{x}$. This feedback model is clearly more informative than that in linear bandits, where the learning agent observes just the sum of the weights. Therefore, our learning problem has lower sample complexity. In particular, note that our $\Omega(\sqrt{KLn})$ lower bound (Proposition 2) is $\sqrt{K}$ smaller than that of Audibert *et al.* [4] (Theorem 5) for combinatorial linear bandits. The bound of Audibert *et al.* [4] is proved for the adversarial setting. Nevertheless, it applies to our setting because the worst-case environment in the proof is stochastic.

Russo and Van Roy [17], and Wen *et al.* [19], derived upper bounds on the *Bayes regret* of *Thompson sampling* in stochastic combinatorial semi-bandits. These bounds have a similar form as our gap-free upper bound in Theorem 6. However, they differ from our work in two aspects. First, the Bayes regret is a different performance metric from regret. From the frequentist perspective, it is a much weaker metric. Second, we also derive $O(\log n)$ upper bounds.

9 Extensions

The computational efficiency of `CombUCB1` depends on the computational efficiency of the offline optimization oracle. When the oracle is inefficient, we suggest resorting to approximations. Let `ALG` be a computationally-efficient oracle that returns an approximation. Then `CombUCB1` can be straightforwardly modified to call `ALG` instead of the original oracle. Moreover, it is easy to bound the regret of this algorithm if it is measured with respect to the best approximate solution by `ALG` in hindsight.

Thompson sampling [18] often performs better in practice than UCB1 [6]. It is straightforward to propose a variant of `CombUCB1` that uses Thompson sampling, by replacing the UCBs in Algorithm 1 with sampling from the posterior on the mean of the weights. The frequentist analysis of regret in Thompson sampling [2] resembles the analysis of UCB1. Therefore, we believe that our analysis can be generalized to Thompson sampling, and we hypothesize that the regret of the resulting algorithm is $O(KL(1/\Delta)\log n)$.

10 Conclusions

The main contribution of this work is that we derive novel gap-dependent and gap-free upper bounds on the regret of `CombUCB1`, a UCB-like algorithm for stochastic combinatorial semi-bandits. These bounds are tight up to polylogarithmic factors. In other words, we show that `CombUCB1` is sample efficient because it achieves near-optimal regret. It is well known that `CombUCB1` is also computationally efficient [11], it can be implemented efficiently whenever the offline variant of the problem can be solved computationally efficiently. Therefore, we indirectly show that stochastic combinatorial semi-bandits can be solved both computationally and sample efficiently, by `CombUCB1`.

Theorems 4 and 5 are proved quite generally, for any (α_i) and (β_i) subject to relatively mild constraints. At the end of the proofs, we choose (α_i) and (β_i) to be geometric sequences. This is sufficient for our purpose. But the choice is likely to be suboptimal and may lead to larger constants in our upper bounds than is necessary. We leave the problem of choosing better (α_i) and (β_i) for future work.

We leave open several questions of interest. For instance, our $\Omega(KL(1/\Delta)\log n)$ lower bound is derived on a problem where all suboptimal solutions have the same gaps. So technically speaking, our $O(KL(1/\Delta)\log n)$ upper bound is tight only on this family of problems. It is an open problem whether our upper bound is tight in general.

Our $O(\sqrt{KLn\log n})$ upper bound matches the $\Omega(\sqrt{KLn})$ lower bound up to a factor of $\sqrt{\log n}$. We believe that this factor can be eliminated by modifying the confidence radii in `CombUCB1` (2) along the lines of Audibert *et al.* [3]. We leave this for future work.

References

[1] Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320, 2011.

[2] Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceeding of the 25th Annual Conference on Learning Theory*, pages 39.1–39.26, 2012.

[3] Jean-Yves Audibert and Sebastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

[4] Jean-Yves Audibert, Sebastien Bubeck, and Gabor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

[5] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

[6] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

[7] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[8] Sebastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 2012.

[9] Nicolo Cesa-Bianchi and Gabor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

[10] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework, results and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pages 151–159, 2013.

[11] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478, 2012.

[12] Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: Fast combinatorial optimization with learning. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 420–429, 2014.

[13] Branislav Kveton, Zheng Wen, Azin Ashkan, and Michal Valko. Learning to act greedily: Polymatroid semi-bandits. *CoRR*, abs/1405.7752, 2014.

[14] T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[15] Gergely Neu and Gabor Bartok. An efficient algorithm for learning with semi-bandit feedback. In *24th International Conference on Algorithmic Learning Theory*, volume 8139 of *Lecture Notes in Computer Science*, pages 234–248. 2013.

[16] Christos Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization*. Dover Publications, Mineola, NY, 1998.

[17] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *CoRR*, abs/1403.5341, 2014.

[18] William. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[19] Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Branislav Kveton. Efficient learning in large-scale combinatorial semi-bandits. *CoRR*, abs/1406.7443, 2014.