Necessary and Sufficient Conditions for Achieving Sub-linear Regret in Stochastic Multi-armed Bandits

Shivaram Kalyanakrishnan

November 2019

Abstract

In this note, we identify a condition that is necessary and sufficient for an algorithm to achieve sub-linear regret on stochastic multi-armed bandits. The condition is a conjunction of two conditions, $C1$ and $C2$. Informally, $C1$ is that the algorithm pull each arm infinitely often in the limit. $C2$ is that the fraction of the total pulls performed on empirically inferior arms vanish in the limit. We state these conditions formally and prove that their conjunction is both necessary and sufficient for achieving sub-linear regret.

1 Stochastic Multi-armed Bandits

We consider n-armed bandits, $n \geq 2$, in which each arm a from the set of arms A yields Bernoulli (0-1) rewards. A bandit instance is fully specified by the means corresponding to each arm; each instance I is an element of the universe of instances $\mathcal{I} = [0, 1]^n$. The bandit instance is sampled in sequence, with the *pull* of arm a^t returning a reward r^t, for $t = 0, 1, \ldots$. Thus, an algorithm's interaction with the bandit generates a sequence

$$a^0, r^0, a^1, r^1, \ldots.$$

Formally, an algorithm L is a mapping from the set of such interaction sequences, also known as *histories*, to the set of probability distributions over arms. Thus, given any history h, algorithm L must specify $q_L(a|h)$, the probability of pulling each arm $a \in A$.

For a horizon of $T \geq 1$ pulls, the expected cumulative regret (or simply "regret") of L on an instance $I \in \mathcal{I}$ is given by

$$R^T_{L,I} = p_\star(I)T - \sum_{t=0}^{T-1} \mathbb{E}_{L,I}[r^t],$$

where $p_\star(I)$ is the largest mean among the arms of I. If we denote the mean of each arm $a \in A$ in instance I by $p_a(I)$, we have $p_\star(I) = \max_{a \in A} p_a(I)$. If $p_a(I) = p_\star(I)$, then arm a is referred to as an *optimal* arm.

This note considers conditions on L so it achieves sub-linear regret.

2 Notation and Background

In this section, we present notation and some basic results that our subsequent analysis will utilise.

2.1 Random Variables and Constants

First, we define some relevant random variables and constants.

- For arm $a \in A$, let u_a^t denote the number of pulls of arm a after $t-1$ pulls of the bandit (that is, before the t-th pull of the bandit is performed).
- For $t \geq 0$, let $\mathit{exploit}^t$ be a Boolean random variable that is 1 if on round t,
  - all the arms have a valid empirical mean (that is, have been pulled at least once), and
  - the arm that is pulled—that is, a^t—has the highest empirical mean (possibly tied with other arms);

  otherwise $\mathit{exploit}^t$ is 0. For $T \geq 1$, let $\mathit{exploit}(T)$ denote the total number of "exploit" rounds performed in the first T pulls: that is, $\mathit{exploit}(T) = \sum_{t=0}^{T-1} \mathit{exploit}^t$.
- For $t \geq 0$, let $\mathit{separated}^t$ be a Boolean random variable that is 1 if on round t,
  - all the arms have a valid empirical mean (that is, have been pulled at least once), and
  - if arm $a \in A$ has the highest empirical mean on round t, then it is an optimal arm (that is, $p_a = p_\star$);

  otherwise $\mathit{separated}^t$ is 0. For $T \geq 1$, let $\mathit{separated}(T)$ denote the total number of rounds up to horizon T in which the arms have been "separated": that is, $\mathit{separated}(T) = \sum_{t=0}^{T-1} \mathit{separated}^t$.
- For every instance $I \in \mathcal{I}$ containing at least one non-optimal arm, define

  $$\Delta_{\min}(I) = \min_{a \in A, p_a \neq p_\star} (p_\star - p_a), \text{ and } \Delta_{\max}(I) = \max_{a \in A, p_a \neq p_\star} (p_\star - p_a).$$

  These quantities denote the expected loss from pulling the "best" and the "worst" non-optimal arms, respectively, when compared to pulling an optimal arm.

Observe that $\Delta_{\min}(I)$ and $\Delta_{\max}(I)$ are fixed constants for every instance $I \in \mathcal{I}$. On the other hand, u_a^t, $\mathit{exploit}^t$, $\mathit{exploit}(T)$ $\mathit{separated}^t$, and $\mathit{separated}(T)$ are random variables whose distribution depends on both the bandit instance and the algorithm. Also note the fact that for every $t \geq 0$,

$$(\mathit{separated}^t \wedge \mathit{exploit}^t) \implies (a^t \textit{ is an optimal arm}).$$

We shall show that successful algorithms must explore sufficiently to get the arms separated, and that they must also exploit sufficiently. As a consequence, they will pull optimal arms a sufficient number of times.

2.2 Probability of a History

For $T \geq 1$, let $\mathcal{H}^T$ denote the set of T-length histories (the sole 0-length history is denoted $\emptyset$). Suppose algorithm L is applied on bandit instance $I \in \mathcal{I}$, what is $\mathbb{P}_{L,I}\{h\}$, the probability that history $h \in \mathcal{H}^T$ is generated? If we take h to be $a^0, r^0, a^1, r^1, \dots a^{T-1}, r^{T-1}$, we observe

$$\begin{aligned}\mathbb{P}_{L,I}\{h\} =& q_L(a^0|\emptyset)\mathbb{P}_I\{r^0|a^0\}\times \\ & q_L(a^1|a^0r^0)\mathbb{P}_I\{r^1|a^1\}\times \\ & \vdots \\ & \times q_L(a^{T-1}|a^0r^0a^1r^1\dots a^{T-2}r^{T-2})\mathbb{P}_I\{r^{T-1}|a^{T-1}\}.\end{aligned}$$

Notice that $\mathbb{P}_{L,I}\{h\}$ factorises into "q_L" terms that only depend on h and the algorithm L, and "$\mathbb{P}_I$" terms that only depend on h and the bandit instance I. We find it convenient to club the algorithm-specific terms into a single term

$$Q_L(h) = q_L(a^0|\emptyset) \times q_L(a^1|a^0r^0) \times \cdots \times q_L(a^{T-1}|a^0r^0a^1r^1 \dots a^{T-2}r^{T-2}).$$

Now, $\mathbb{P}_I\{r^t|a^t\}$ is $p_{a^t}(I)$ if $r^t = 0$, and $1 - p_{a^t}(I)$ if $r^t = 0$. If h contains exactly $s_a(h)$ 1-rewards and $f_a(h)$ 0-rewards for arm $a \in A$, we get

$$\mathbb{P}_{L,I}(h) = Q_L(h) \prod_{a \in A} (p_a(I))^{s_a(h)}(1 - p_a(I))^{f_a(h)}. \tag{1}$$

Note that the total number of occurrences of arm a in h is $s_a(h) + f_a(h)$, which we denote $u_a(h)$.

2.3 Probability of an Event

When we refer to the probability of an event E, we really mean: *what is the probability of encountering a history in which E is true?* Consider, as an example, the event that $r^{10} = 1$. This even could happen in many possible ways, the probability of each way depending both on the sampling algorithm and the bandit instance. Suppose E can be verified (to have happened or not happened) within T rounds, we may write

$$\mathbb{P}_{L,I}\{E\} = \sum_{h \in \mathcal{H}^T, E \text{ happens in } h} \mathbb{P}_{L,I}\{h\}. \tag{2}$$

Put otherwise, an event is equivalent to a set of histories.

3 Necessary and Sufficient Conditions

In this section we present necessary and sufficient algorithms for an algorithm L to achieve sub-linear regret.

3.1 Target Family of Bandit Instances

To make our goal precise, we must first state *which* bandit instances are under consideration. As we already know, algorithms such as UCB and Thompson Sampling achieve sub-linear regret on *every* bandit instance from the universe $\mathcal{I} = [0, 1]^n$. Nevertheless, we shall discover in our upcoming discussion that $\mathcal{I}$ contains some exceptional bandit instances on which successful algorithms need not even pull every arm once—these are instances I for which $p_\star(I) = 1$.

The crux of our argument in this section will be that if h is a finite-length history generated by an algorithm L on instance I, there is a non-zero probability that the same history will be generated by L on a different instance I'. If we reflect for a moment, we observe that this claim cannot be valid in general if I' contains arms whose means are exactly 0 or 1: such arms can only generate all 0's or all 1's as rewards—they have a zero probability of generating sequences containing both 0- and 1-rewards. Since our discussion is in the context of *minimising* regret, it happens that arms with mean 0 will not affect our argument, whereas arms with mean 1 will interfere.

Hence, for now, we shall remove from our consideration all instances with arms whose mean is 1. We shall only consider instances from the set of instances $\overline{\mathcal{I}} = \{I \in \mathcal{I} : p_\star(I) \neq 1\}$: that is, $\overline{\mathcal{I}} = [0, 1)^n$. In Section 4, we shall revisit the question of bandit instances whose optimal mean is 1.

3.2 Infinite Exploration

Our first condition, denoted $C1$, expresses that the algorithm in question pulls each arm *infinitely often* in the limit. Here is the formal statement, applied to algorithm L and instance $I \in \overline{\mathcal{I}}$.

Definition 1 ($C1(L, I)$)**.** *For every $u \geq 0$ and $\delta > 0$, there exists $T_0(u, \delta) \geq u$ such that*

$$\mathbb{P}_{L,I}\{\forall a \in A : u_a^{T_0(u,\delta)} > u\} \geq 1 - \delta.$$

The following lemma shows that $C1$ is necessary for L to achieve sub-linear regret.

Lemma 1. *If there exists $I \in \overline{\mathcal{I}}$ such that $\neg C1(L, I)$, then there exists $I' \in \overline{\mathcal{I}}$ such that $R^T_{L,I'} = \Omega(T)$.*

Proof. If $C1(L, I)$ is false for some $I \in \overline{\mathcal{I}}$, it means there exist (1) $u_0 > 0$, (2) $\delta_0 > 0$, and (3) $a_0 \in A$ such that for all $T \geq u_0$, $\mathbb{P}_{L,I}\{u^T_{a_0} > u_0\} < 1 - \delta_0$, or equivalently,

$$\mathbb{P}_{L,I}\{u^T_{a_0} \leq u_0\} > \delta_0. \tag{3}$$

Now, consider a bandit instance I' such that for every arm $a \in A$:

$$p_a(I') = \begin{cases} p_a(I), & \text{if } a \neq a_0, \\ \frac{p_\star(I)+1}{2}, & \text{if } a = a_0. \end{cases}$$

The idea behind the construction is that I' is identical to I except for arm a_0, which for I' is the *sole optimal arm* (in fact we could take the mean of this arm in I' to be any arbitrary element in $(p_\star(I), 1)$). We establish that with a non-zero probability, L will also pull a_0 fewer than u_0 times on I'. We use (1) and (2) to write

$$\begin{aligned}
&\mathbb{P}_{L,I'}\{u^T_{a_0} \leq u_0\} \\
&= \sum_{h \in \mathcal{H}^T, u_{a_0}(h) \leq u_0} \mathbb{P}_{L,I'}\{h\} \\
&= \sum_{h \in \mathcal{H}^T, u_{a_0}(h) \leq u_0} Q_L(h) \prod_{a \in A} (p_a(I'))^{s_a(h)} (1 - p_a(I'))^{f_a(h)} \\
&= \sum_{h \in \mathcal{H}^T, u_{a_0}(h) \leq u_0} Q_L(h) \left(\prod_{a \in A} (p_a(I))^{s_a(h)} (1 - p_a(I))^{f_a(h)} \right) \left(\frac{p_{a_0}(I')}{p_{a_0}(I)} \right)^{s_{a_0}(h)} \left(\frac{1 - p_{a_0}(I')}{1 - p_{a_0}(I)} \right)^{f_{a_0}(h)} \\
&\geq \sum_{h \in \mathcal{H}^T, u_{a_0}(h) \leq u_0} Q_L(h) \left(\prod_{a \in A} (p_a(I))^{s_a(h)} (1 - p_a(I))^{f_a(h)} \right) \left(\frac{1 - p_{a_0}(I')}{1 - p_{a_0}(I)} \right)^{u_0} \\
&= \left(\frac{1 - p_{a_0}(I')}{1 - p_{a_0}(I)} \right)^{u_0} \mathbb{P}_{L,I}\{u^T_{a_0} \leq u_0\}.
\end{aligned}$$

From (3), we infer that

$$\mathbb{P}_{L,I'}\{u^T_{a_0} \leq u_0\} > \left(\frac{1 - p_{a_0}(I')}{1 - p_{a_0}(I)} \right)^{u_0} \delta_0.$$

Notice that the right hand side does not depend on T: it is fixed by I, I', a_0, δ_0, and u_0. If we call the right hand side δ_1, which is clearly positive, we essentially have the following: if $C1(L, I)$ is false for $I \in \overline{\mathcal{I}}$, then there exist (1) an instance $I' \in \overline{\mathcal{I}}$, (2) $u_0 > 0$, (3) $\delta_1 > 0$, and (4) $a_0 \in A$ such that for all $T \geq u_0$,

$$\mathbb{P}_{L,I'}\{u^T_{a_0} \leq u_0\} > \delta_1.$$

In other words, there is at least a δ_1-probability that L plays a_0, which is the sole optimal arm in I', no more than u_0 times. If we consider any horizon $T > 2u_0$, we get

$$R^T_{L,I'} \geq \delta_1(T - u_0)\Delta_{\min}(I') > \delta_1 \frac{\Delta_{\min}(I')}{2} T,$$

implying that $R^T_{L,I'} = \Omega(T)$. □

The crux of the proof of Lemma 1 was that a *finite* number of pulls can always be misleading (with a positive, even if small, probability). Hence, in order to progressively get closer to the true means of the arms (which shall reveal an optimal arm), an algorithm must continue to pull each arm for ever. The following lemma guarantees that continuous sampling of all arms results in progress.

Lemma 2. *Suppose $C1(L, I)$ is true for some algorithm L and instance $I \in \overline{\mathcal{I}}$. Then for all $\delta_0 > 0$, there is some $T_0(\delta_0) > 0$ such that for all $t \geq T_0(\delta_0)$,*

$$\mathbb{P}_{L,I}\{\mathit{separated}^t\} \geq 1 - \delta_0.$$

Proof. For $\epsilon > 0$, Hoeffding's Inequality guarantees that if each arm is pulled exactly u times, the probability that its empirical mean is ϵ or more away from the true mean is at most $e^{-2u\epsilon^2}$. Hence, if an arm has been pulled u_0 or more times, the probability that its empirical mean is ϵ or more away from the true mean is at most $\sum_{u=u_0}^{\infty} e^{-2u\epsilon^2} = e^{-2u_0\epsilon^2}/(1 - e^{-2\epsilon^2})$. Now, observe that if each arm's empirical mean is less than $\frac{\Delta_{\min}(I)}{2}$ away from its true mean, then an arm with the highest empirical mean must be an optimal arm: that is, an optimal arm must have "separated". Therefore, if we take $\epsilon = \frac{\Delta_{\min}(I)}{2}$ and $u_0 = \lceil \frac{2}{(\Delta_{\min}(I))^2} \ln \frac{2n}{\delta_0(1-e^{-2\epsilon^2})} \rceil$, we can make the following claim. If t is a round at which each arm has been pulled at least u_0 times, then the probability of an optimal arm *not* being separated—which can happen only if at least one arm is inaccurate—is at most

$$\sum_{a \in A} \frac{e^{-2u_0\epsilon^2}}{1 - e^{-2\epsilon^2}} \leq n\frac{\delta_0}{2n} = \frac{\delta_0}{2}.$$

In short, $\mathbb{P}_{L,I}\{\neg \mathit{separated}^t \wedge (\forall a \in A : u^t_a \geq u_0)\} \leq \frac{\delta_0}{2}$. Now, apply Definition 1 with u_0 as defined, with probability parameter $\delta_2 = \frac{\delta_0}{2}$. The definition gives us that there exists $T_0(\delta_2) \geq u_0$ such that for all $t \geq T_0(\delta_2)$, $\mathbb{P}_{L,I}\{\forall a \in A : u^t_a \leq u_0\} \leq \delta_2$. We get:

$$\begin{aligned}
\mathbb{P}_{L,I}\{\neg \mathit{separated}^t\} &= \mathbb{P}_{L,I}\{\neg \mathit{separated}^t \wedge (\forall a \in A : u^t_a \leq u_0)\} + \mathbb{P}_{L,I}\{\neg \mathit{separated}^t \wedge (\forall a \in A : u^t_a > u_0)\} \\
&\leq \mathbb{P}_{L,I}\{\forall a \in A : u^t_a \leq u_0\} + \mathbb{P}_{L,I}\{\neg \mathit{separated}^t \wedge \forall a \in A : u^t_a \geq u_0\} \\
&\leq \frac{\delta_0}{2} + \frac{\delta_0}{2} = \delta_0,
\end{aligned}$$

which completes the proof. □

The lemma establishes that an algorithm exploring sufficiently will eventually get an optimal arm separated out. We find it convenient to convert this result, which makes an observation about each round t, to one that aggregates rounds up to a specified horizon.

Corollary 1. *Suppose $C1(L, I)$ is true for some algorithm L and instance $I \in \overline{\mathcal{I}}$. Then for all $\epsilon_0 > 0$, there is some $T_0(\epsilon_0) > 0$ such that for all $T \geq T_0(\epsilon_0)$,*

$$\frac{\mathbb{E}_{L,I}\{\mathit{separated}(T)\}}{T} \geq 1 - \epsilon_0.$$

Proof. Apply Lemma 2 with probability parameter $\delta_0 = \epsilon_0/2$. There exists $T_1 > 0$ such that for all $t \geq T_1$, $\mathbb{P}\{separated^t\} \geq 1 - \delta_0$. Take $T_0(\epsilon_0) \geq T_1/\delta_0$. For $T \geq T_0(\epsilon_0)$, we have

$$\begin{aligned}\frac{\mathbb{E}_{L,I}\{separated(T)\}}{T} &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{P}\{separated^t\} \geq \frac{1}{T}\sum_{t=T_1}^{T-1}\mathbb{P}\{separated^t\} \\ &\geq \frac{(T-T_1)(1-\delta_0)}{T} = 1 - \delta_0 - \frac{T_1(1-\delta_0)}{T} \\ &\geq 1 - \delta_0 - \delta_0(1-\delta_0) > 1 - \epsilon_0.\end{aligned}$$

□

In summary, we have shown that infinite exploration is necessary for achieving sub-linear regret, and that it helps to get the arms' means separated. We now move to the next logical step: exploiting arms with high empirical means.

3.3 Greed in the Limit

Since algorithms do not directly know which arms are optimal, the best they can do is trust empirical data (and be greedy with respect to *empirical* means). Recall that $exploit(T) = \sum_{t=0}^{T-1} exploit^t$ denotes the number of rounds on which an *empirically* optimal arm is pulled. Our second condition, $C2$, is about $exploit(T)$.

Definition 2 ($C2(L, I)$)**.**

$$\lim_{T\to\infty}\frac{\mathbb{E}_{L,I}[exploit(T)]}{T} = 1.$$

In the two following lemmas, we shall assume that $C1$ is true. Under this assumption, we show that achieving sub-linear regret depends on $C2$. Taken together with Lemma 1, the implication is that $C1 \wedge C2$ is a necessary and sufficient condition for the achievement of sub-linear regret.

Lemma 3. *Assume that for algorithm L and instance $I \in \overline{\mathcal{I}}$, $C1(L, I)$ is true and $C2(L, I)$ is false. Then $R^T_{L,I} = \Omega(T)$.*

Proof. Since $C2(L, I)$ is false, there exist $\epsilon_0 > 0$ and $T_0 > 0$ such that for all $T \geq T_0$,

$$\frac{\mathbb{E}_{L,I}[exploit(T)]}{T} < 1 - \epsilon_0.$$

Since $C1(L, I)$ is true, we can apply Corollary 1 with parameter $\delta_0 = \epsilon_0/2$. We have that there exists $T_1 > 0$ such that $\frac{\mathbb{E}_{L,I}\{separated^T\}}{T} \geq 1 - \delta_0$. Now fix some $T_2 \geq \max(T_0, T_1)$. We observe that for all $T > T_2$,

$$\mathbb{E}_{L,I}[exploit(T)] < (1-\epsilon_0)T, \text{ and}$$

$$\mathbb{E}_{L,I}[separated(T)] \geq (1-\delta_0)T.$$

We observe the the expected number of non-optimal pulls up to horizon T is

$$\begin{aligned}\sum_{t=0}^{T-1} \mathbb{P}_{L,I}\{a^t \text{ is not an optimal arm}\} &\geq \sum_{t=0}^{T-1} \mathbb{P}_{L,I}\{\mathit{separated}^t \wedge \neg \mathit{exploit}^t)\} \\ &= \sum_{t=0}^{T-1} \left(\mathbb{P}_{L,I}\{\mathit{separated}^t\} + \mathbb{P}_{L,I}\{\neg \mathit{exploit}^t\} - \mathbb{P}_{L,I}\{\mathit{separated}^t \vee \neg \mathit{exploit}^t\}\right) \\ &\geq \sum_{t=0}^{T-1} \left(\mathbb{P}_{L,I}\{\mathit{separated}^t\} + \mathbb{P}_{L,I}\{\neg \mathit{exploit}^t\} - 1\}\right) \\ &= \sum_{t=0}^{T-1} \left(\mathbb{P}_{L,I}\{\mathit{separated}^t\} - \mathbb{P}_{L,I}\{\mathit{exploit}^t\}\right) \\ &= \mathbb{E}_{L,I}[\mathit{separated}(T)] - \mathbb{E}_{L,I}[\mathit{exploit}(T)] \\ &\geq (1-\delta_0)T - (1-\epsilon_0)T \\ &= \delta_0 T.\end{aligned}$$

Clearly $R^T_{L,I}$ is at least $\delta_0 \Delta_{\min}(I)T$ for all $T > T_2$, which means $R^T_{L,I} = \Omega(T)$. □

Lemma 4. *Assume that for algorithm L and instance $I \in \overline{\mathcal{I}}$, $C1(L,I)$ and $C2(L,I)$ are both true. Then $R^T_{L,I} = o(T)$.*

Proof. To prove the result, we show that for every $\gamma_0 > 0$, there exists $T_0 > 0$ such that for all $T \geq T_0$, $R^T_{L,I}/T \leq \gamma_0$.

Take $\delta_0 = \frac{\gamma_0}{2\Delta_{\max}(I)}$. From Corollary 1, we get that there is $T_1 > 0$ such that for all $T \geq T_1$,

$$\mathbb{E}[\mathit{separated}(T)] \geq (1-\delta_0)T.$$

Take $\epsilon_0 = \frac{\gamma_0}{2}$. From Definition 2, we see that there is $T_2 > 0$ such that for all $T \geq T_2$,

$$\mathbb{E}[\mathit{exploit}(T)] \geq (1-\epsilon_0)T.$$

It follows that for $T > T_0 \geq \max(T_1, T_2)$,

$$\begin{aligned}\frac{R^T_{L,I}}{T} &\leq \sum_{t=0}^{T-1} \Delta_{\max}(I)\mathbb{P}\{a^t \text{ is not an optimal arm}\} \\ &\leq \frac{1}{T}\sum_{t=0}^{T-1} \Delta_{\max}(I)\mathbb{P}\{\neg \mathit{separated}^t \vee \neg \mathit{exploit}^t\} \\ &\leq \frac{\Delta_{\max}(I)}{T}\sum_{t=0}^{T-1} \left(\mathbb{P}\{\neg \mathit{separated}^t\} + \mathbb{P}\{\neg \mathit{exploit}^t\}\right) \\ &= \frac{\Delta_{\max}(I)}{T}\left((T - \mathbb{E}[\mathit{separated}(T)]) + (T - \mathbb{E}[\mathit{exploit}(T)])\right) \\ &\leq \frac{\Delta_{\max}(I)}{T}\left(T\delta_0 + T\epsilon_0\right) \\ &= \gamma_0.\end{aligned}$$

□

3.4 Final Result

The results proven so far already establish necessary and sufficient conditions for achieving sub-linear regret. For clarity, we state our final result in concise form below and give a detailed working that combines the previous results. In the theorem and proof that follow, there is a "for all learning algorithms L" quantifier applying to all the statements; we leave out this quantifier to reduce clutter.

Theorem 1.

$$\forall I \in \overline{\mathcal{I}} : C1(L, I) \wedge C2(L, I) \implies \forall I \in \overline{\mathcal{I}} : R^T_{L,I} = o(T). \tag{4}$$

$$\forall I \in \overline{\mathcal{I}} : R^T_{L,I} = o(T) \implies \forall I \in \overline{\mathcal{I}} : C1(L, I) \wedge C2(L, I). \tag{5}$$

Proof. Lemma 4 gives that $\forall I \in \bar{\mathcal{I}} : (C1(L, I) \wedge C2(L, 1) \implies R^T_{L,I} = o(T))$, which implies (4). To show (5), we begin by considering the conjunction of lemmas 1 and 4 and proceeding in logical sequence. We have:

$$(\exists I \in \bar{\mathcal{I}} : \neg C1(L, I) \implies \exists I \in \bar{\mathcal{I}} : R^T_{L,I} = \Omega(T)) \wedge (\forall I \in \bar{\mathcal{I}} : C1(L, I) \wedge \neg C2(L, I)) \implies R^T_{L,I} = \Omega(T))$$

$$\implies$$

$$(\exists I \in \bar{\mathcal{I}} : \neg C1(L, I) \implies \exists I \in \bar{\mathcal{I}} : R^T_{L,I} = \Omega(T)) \wedge (\exists I \in \bar{\mathcal{I}} : C1(L, I) \wedge \neg C2(L, I) \implies \exists I \in \bar{\mathcal{I}} : R^T_{L,I} = \Omega(T))$$

$$\implies$$

$$\exists I \in \bar{\mathcal{I}} : \neg C1(L, I) \vee (C1(L, I) \wedge \neg C2(L, I)) \implies \exists I \in \bar{\mathcal{I}} : R^T_{L,I} = \Omega(T)$$

$$\iff$$

$$\exists I \in \bar{\mathcal{I}} : \neg C1(L, I) \vee \neg C2(L, I) \implies \exists I \in \bar{\mathcal{I}} : R^T_{L,I} = \Omega(T)$$

$$\iff$$

$$\neg(\exists I \in \bar{\mathcal{I}} : R^T_{L,I} = \Omega(T)) \implies \neg(\exists I \in \bar{\mathcal{I}} : \neg C1(L, I) \vee \neg C2(L, I))$$

$$\iff$$

$$\forall I \in \bar{\mathcal{I}} : R^T_{L,I} = o(T) \implies \forall I \in \bar{\mathcal{I}} : C1(L, I) \wedge C2(L, I),$$

which is the statement of (5). □

4 Discussion

Having completed our presentation, we have two points to discuss.

4.1 Arms with Mean 1

First, why did we exclude bandit instances with an optimal mean of 1 in our working? Recall that the universe of bandit instances considered in our working was $\overline{\mathcal{I}} = [0, 1)^n$. What would happen if we tried replacing it with $\overline{\mathcal{I}} = [0, 1]^n$? Do you think the following statements are correct?

$$\forall I \in \mathcal{I} : C1(L, I) \wedge C2(L, I) \implies \forall I \in \mathcal{I} : R^T_{L,I} = o(T). \tag{6}$$

$$\forall I \in \mathcal{I} : R^T_{L,I} = o(T) \implies \forall I \in \mathcal{I} : C1(L, I) \wedge C2(L, I). \tag{7}$$

Following the same proof structure as we presented in Section 3, it is not hard to show that (6) is true. However, (7) is not true. To see why, consider any successful algorithm L—concretely let us take ϵ^t-greedy sampling with $\epsilon^t = 1/(t+1)$. We know that L achieves sub-linear regret on every instance $I \in \mathcal{I}$. Now make one minor change to L: whenever an arm returns a reward of 1, continue pulling that arm until it returns a reward of 0. Let the resulting algorithm be L'. In short, L' performs "extended pulls" of each arm that last until a 0-reward is returned; the arm to pull next is decided by the ϵ^t-greedy strategy.

It is not hard to show that L' will achieve sub-linear regret on every instance $I \in \mathcal{I}$. However, observe that on an instance I_{opt} in which the optimal mean is 1, L' will never stop pulling an optimal arm once it is encountered. In this event, other arms (some of which might not have been sampled even once!) will not be explored infinitely often: that is, $C1(L', I_{\text{opt}})$ is false and consequently (7) is false.

4.2 Weaker Form of $C1$

In class, the instructor had presented (so-called) necessary and sufficient conditions with a weaker form of $C1$, which we define here as $C1W$.

Definition 3 ($C1W(L, I)$)**.** *For every $u \geq 0$, there exists $T_0 \geq u$ such that for all $T \geq T_0$ and for all $a \in A$,*

$$\mathbb{E}_{L,I}\{u_a^T\} \geq u.$$

Take a moment to consider whether the following statements are indeed true.

$$\forall I \in \overline{\mathcal{I}} : C1W(L, I) \land C2(L, I) \implies \forall I \in \overline{\mathcal{I}} : R_{L,I}^T = o(T). \tag{8}$$

$$\forall I \in \overline{\mathcal{I}} : R_{L,I}^T = o(T) \implies \forall I \in \overline{\mathcal{I}} : C1W(L, I) \land C2(L, I). \tag{9}$$

First, convince yourself that for all algorithms L and instances $I \in \overline{\mathcal{I}}$, $C1(L, I) \implies C1W(L, I)$. Hence, (9) follows from (5), which we have already proven to be true.

On the other hand, (8) is not true, and so the claim made by the instructor in class is not correct. The error is regretted. To see why (8) fails, consider the following three algorithms.

- $L1$. Algorithm $L1$ pulls each arm exactly once, and thereafter greedily pulls an arm with the highest empirical mean.
- $L2$. Algorithm $L2$ is any successful algorithm—for example ϵ^t-greedy sampling with $\epsilon^t = 1/(t+1)$.
- $L3$. Algorithm $L3$ is a randomised algorithm that with probability $1/2$, behaves through its entire run as $L1$, and with probability $1/2$, behaves through its entire run as $L2$. Note that $L1$, $L2$, and $L3$ can be suitably defined so the choice made by $L3$ before pulling an arm can be encoded in its first pull: that is, by observing the very first pull made by $L3$, we know whether it is implementing $L1$ or $L2$ (and so $L3$ is a legitimate algorithm).

We leave it to the student to verify that $L3$ satisfies $C1W$ and $C2$ on all instances $I \in \overline{\mathcal{I}}$, but it still achieves sub-linear regret on some instances $I \in \overline{\mathcal{I}}$.

Acknowledgement

The author is grateful to Aditya Gopalan for suggesting several improvements to the draft.