CSCI 659 Lecture 7

Spring 2026

Instructor: Haipeng Luo

1 Stochastic Multi-Armed Bandits

In the last lecture, we discussed algorithms for the *adversarial* MAB problem where there is no assumption on how the loss vectors are generated, similar to all the problems we have analyzed so far. On the other hand, there is a also huge literature on the *stochastic* version of MAB, where each arm represents an unknown distribution and each pull of the arm generates an independent sample of that distribution. While this is clearly just a special case of adversarial MAB, the goal here is usually to derive regret bounds that are instance-dependent and in some situations better than the worst-case $\mathcal{O}(\sqrt{TK})$ bound. Moreover, unlike the full-information setting (where the stochastic assumption makes the problem much easier), stochastic MAB is still a highly non-trivial problem due to the partial information feedback.

Formally, in stochastic MAB, we assume that for each action a, there is an unknown distribution with support $[0, 1]$ and mean $\mu(a)$, such that $\ell_1(a), \ldots, \ell_T(a)$ are independent samples of this distribution. For this problem we usually care about a slightly different version of regret, called *pseudo-regret*, defined as

$$\bar{\mathcal{R}}_T = \max_{a\in[K]} \mathbb{E}\left[\sum_{t=1}^T \ell_t(a_t) - \sum_{t=1}^T \ell_t(a)\right],$$

where the expectation is over the randomness of both the environment and the algorithm. It is easy to see that pseudo-regret is a weaker measure compared to standard expected regret since the latter moves the "max" inside the expectation. However, pseudo-regret not only is more natural in stochastic settings but also allows us to ignore the deviation of the samples $\ell_1(a), \ldots, \ell_T(a)$ from their mean $\mu(a)$ and derive better regret bounds as we will see.

Let $a^\star \in \operatorname{argmin}_a \mu(a)$ be an optimal action in terms of the expected loss. Then, pseudo-regret for stochastic MAB can be simplified as

$$\bar{\mathcal{R}}_T = \mathbb{E}\left[\sum_{t=1}^T (\mu(a_t) - \mu(a^\star))\right] = \mathbb{E}\left[\sum_{t=1}^T \sum_{a=1}^K \Delta_a \mathbf{1}\{a_t = a\}\right] = \sum_{a:\Delta_a>0} \Delta_a \mathbb{E}[n_T(a)] \tag{1}$$

where $\Delta_a = \mu(a) - \mu(a^\star)$ is the *suboptimality gap* of action a and $n_t(a) = \sum_{s=1}^t \mathbf{1}\{a_s = a\}$ is the number of times action a has been pulled up to round t.

In stochastic MAB, the trade-off between exploration and exploitation is perhaps even more intuitive. To see this, let $\widehat{\mu}_t(a) = \frac{1}{n_t(a)} \sum_{s=1}^t \mathbf{1}\{a_s = a\}\ell_s(a)$ be the empirical average loss of action a up to round t. Since the environment is stochastic, $\widehat{\mu}_t(a)$ should be a good approximation of $\mu(a)$ as long as $n_t(a)$ is large enough. This is formally stated in the following concentration lemma.

Lemma 1. *For stochastic MAB, no matter what the learner's strategy is, we have with probability at least* $1 - 2K/T$*, for all action* $a \in [K]$ *and all time* $t \in [T]$*:*

$$|\widehat{\mu}_t(a) - \mu(a)| \le 2\sqrt{\frac{\ln T}{n_t(a)}}.$$

The lemma is essentially proven by the standard Hoeffding's inequality, except for the extra technicality needed to deal with the fact that $n_t(a)$ itself is also random. We omit the proof for simplicity.

Therefore, on the one hand, we want to exploit by picking the empirically best action $\mathrm{argmin}_a\, \widehat{\mu}_t(a)$, but on the other hand, we also need to explore so that all actions are picked frequently enough to make sure that $\widehat{\mu}_t(a)$ is indeed a good approximation of $\mu(a)$.

Question 1. *What if one ignores the exploration-exploitation trade-off and simply runs a greedy algorithm, that is,* $a_t = \mathrm{argmin}_{a\in[K]}\, \widehat{\mu}_{t-1}(a)$*? How large can the regret be?*

Explore-then-Exploit. Based on the intuition above, one naive approach would be "explore-then-exploit", that is, first spend a certain number of rounds to pick all the actions equally often, and then simply stick with the empirically best action for all remaining rounds. While simple, this approach is also intuitively wasteful and nonadaptive in the exploration phase, in the sense that every action is selected equally often, even if some of them look much worse than others. A quick analysis also shows that the algorithm is suboptimal: suppose that we want the estimation error for every action to be ϵ (with high probability), then based on Lemma 1 we need to select every action roughly $(\ln T)/\epsilon^2$ rounds; therefore, the total regret would be of order $\widetilde{\mathcal{O}}(\frac{K}{\epsilon^2} + \epsilon T)$ (since in the exploration phase, per round regret could be as large as a constant, while in the exploitation phase, per round regret can be at most $\mathcal{O}(\epsilon)$), which is at best $\widetilde{\mathcal{O}}(T^{2/3}K^{1/3})$ with the optimally-chosen ϵ.

2 The UCB Algorithm

To improve over this naive approach, we need a more adaptive way to trade off exploration and exploitation. The classic algorithm to do so is the UCB (Upper Confidence Bound) algorithm [Auer et al., 2002]. Since we consider "losses" instead of "rewards" (the original setting of [Auer et al., 2002]), the algorithm that we will discuss is actually LCB (Lower Confidence Bound). For convention, however, we will still call it UCB.

UCB applies a very powerful principle called "optimism in face of uncertainty", which is useful in many other stochastic problems with bandit feedback. The main idea of the principle is the following: among all plausible environments that are consistent with the data observed so far, *the learner should be optimistic and act as if the environment is the best possible one*.

Specifically, at each round t, Lemma 1 already tells us what the plausible environments are given the data observed so far. Among them, the most favorable one to the learner is the one with the smallest loss mean, that is, when the loss mean for each action a is the following lower confidence bound:

$$\mathrm{LCB}_t(a) \overset{\mathrm{def}}{=} \widehat{\mu}_{t-1}(a) - 2\sqrt{\frac{\ln T}{n_{t-1}(a)}}.$$

Therefore, an optimistic learner would wishfully believe that this is the true environment and naturally selects

$$a_t \in \operatorname*{argmin}_{a\in[K]} \mathrm{LCB}_t(a). \tag{2}$$

This is exactly the UCB algorithm.

A couple of remarks are in order. First, note that $n_{t-1}(a)$ is initially 0, leading to negative infinity for $\mathrm{LCB}_t(a)$, so the algorithm will be forced to pick each action exactly once for the first K rounds. Afterwards, the two terms in $\mathrm{LCB}_t(a)$ are essentially playing the role of exploitation and exploration respectively: the first term suggests picking actions with low empirical mean (exploitation), while the second term suggests picking actions that have not been selected often enough (exploration). Together, they achieve an adaptive trade-off between exploitation and exploration, which does not waste many rounds to explore an infrequently selected action if it already looks pretty bad. Also note that optimism is indeed important to derive exploration — think about what would happen if we adopt pessimism instead and pick the arm with the smallest upper confidence bound $\widehat{\mu}_{t-1}(a) + 2\sqrt{(\ln T)/n_{t-1}(a)}$. Moreover, in contrast to Exp3, UCB is a deterministic algorithm, that is, no randomness is used in deciding which actions to play.

Question 2. *In adversarial MAB, can a deterministic algorithm be no-regret?*

To analyze the algorithm, we start with the following key lemma.

Lemma 2 (per-round regret)**.** *Under the event stated in Lemma 1, UCB ensures* $\Delta_{a_t} \leq 4\sqrt{\frac{\ln T}{n_{t-1}(a_t)}}$ *for each time* t.

Proof. This is a direct consequence of optimism:

$$\begin{aligned}
\Delta_{a_t} = \mu(a_t) - \mu(a^\star) &\leq \mu(a_t) - \mathrm{LCB}_t(a^\star) && \text{(Lemma 1)} \\
&\leq \mu(a_t) - \mathrm{LCB}_t(a_t) && \text{(by Eq. (2), } a_t \text{ has the smallest LCB)} \\
&\leq 4\sqrt{\tfrac{\ln T}{n_{t-1}(a_t)}}, && \text{(by the definition of LCB and Lemma 1)}
\end{aligned}$$

which completes the proof. □

This simple fact immediately tells us how many times each action can be selected by UCB.

Theorem 1. *Under the event stated in Lemma 1, we have* $n_T(a) \leq \frac{16 \ln T}{\Delta_a^2} + 1$ *for every action* a*. Consequently, the pseudo-regret of UCB satisfies* $\bar{\mathcal{R}}_T = \mathcal{O}(\sum_{a:\Delta_a>0} \frac{\ln T}{\Delta_a})$.

Proof. For each action a, applying Lemma 2 with t being the last time a is selected shows: $\Delta_a \leq 4\sqrt{\frac{\ln T}{n_T(a)-1}}$, which, after rearranging, proves the first statement. Therefore, if we denote the event stated in Lemma 1 by E, then the pseudo-regret (recall Eq. (1)) of UCB is

$$\begin{aligned}
\bar{\mathcal{R}}_T &\leq \Pr(E) \times \sum_{a:\Delta_a>0} \Delta_a \mathbb{E}\left[n_T(a) \mid E\right] + \Pr(\neg E) \times T \\
&\leq \sum_{a:\Delta_a>0} \left(\frac{16 \ln T}{\Delta_a} + \Delta_a\right) + 2K = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\ln T}{\Delta_a}\right),
\end{aligned}$$

which proves the second statement. □

This regret bound is in sprit similar to that in Theorem 6 of Lecture 3 for the full information setting, in the sense that they both depend on the difficulty of the specific problem instance, measured by some suboptimality gaps, and in exchange enjoy much better dependence on T compared to the worst-case $\sqrt{T}$-bound. In fact, it can be shown that such a bound is *instance-optimal* for stochastic MAB, meaning that for each problem instance, this is the best pseudo-regret bound one can achieve using a "reasonable" algorithm.

Question 3. *The* $\Omega(\sqrt{TK})$ *lower bound construction for adversarial MAB in Lecture 6 is in fact also a stochastic environment, so why is that not a contradiction with this* $(\ln T)$*-bound? Is this due to the difference between expected regret and pseudo-regret?*

In this instance-optimal bound, the smaller the gaps, the larger the pseudo-regret, which makes sense to some degree because smaller gaps make it harder to distinguish the optimal actions from the rest. On the other than, however, if an action really has a tiny suboptimality gap, then by definition selecting it does not lead to large regret and thus there is really no point in distinguishing it from the optimal actions. While this intuition is not reflected in this instance-optimal regret bound, a different analysis below shows that UCB indeed still guarantees $\mathcal{O}(\sqrt{TK \ln T})$ regret at the same time, no matter how small the gaps are.

Theorem 2. *For any algorithm, if* t_0 *is the first time such that* $n_{t_0}(a) \geq 1$ *for all* a *(that is, all actions have been selected at least once), then* $\sum_{t=t_0+1}^{T} 1/\sqrt{n_{t-1}(a_t)} \leq 2\sqrt{TK}$*. Consequently, UCB ensures* $\bar{\mathcal{R}}_T = \mathcal{O}(\sqrt{TK \ln T})$.

Proof. The first statement can be shown by direct calculation:

$$\begin{aligned}
&\sum_{t=t_0+1}^{T} \frac{1}{\sqrt{n_{t-1}(a_t)}} = \sum_{a=1}^{K} \sum_{t=t_0+1}^{T} \frac{\mathbf{1}\left\{a = a_t\right\}}{\sqrt{n_{t-1}(a)}} \\
&\leq \sum_{a=1}^{K} \sum_{s=1}^{n_T(a)-1} \frac{1}{\sqrt{s}} \leq 2 \sum_{a=1}^{K} \sqrt{n_T(a)} \qquad \left(\textstyle\sum_{s=1}^{m} \frac{1}{\sqrt{s}} \leq 2 \sum_{s=1}^{m} (\sqrt{s} - \sqrt{s-1}) = 2\sqrt{m}\right)
\end{aligned}$$

$$\le 2\sqrt{\left(\sum_{a=1}^{K} n_T(a)\right) K} = 2\sqrt{TK}. \qquad \text{(Cauchy-Schwarz inequality)}$$

For UCB, t_0 is simply K as discussed. Thus, using Lemma 2 we can bound the pseudo-regret as

$$\bar{\mathcal{R}}_T \le (K + 8\sqrt{TK \ln T}) + \Pr(\text{event in Lemma 1 does not hold}) \times T = \mathcal{O}(\sqrt{TK \ln T}),$$

which proves the second statement. □

3 Best of Both Worlds

Given the two different types of results we have seen so far: $\mathcal{O}(\sqrt{TK})$ regret for adversarial MAB and $\mathcal{O}(\sum_{a:\Delta_a>0} \frac{\ln T}{\Delta_a})$ regret for stochastic MAB, it is natural to ask whether there is one single algorithm that achieves the best of these two worlds *simultaneously*, without knowing ahead of time whether the world is stochastic or adversarial. It can be verified either theoretically or empirically that neither UCB or vanilla Exp3 achieves such a best-of-both-worlds result. In fact, the possibility of such results had been highly unclear, until the seminal work of Bubeck and Slivkins [2012], where they propose to start with running UCB, and then switch to Exp3 if some sophisticated tests show that the environment is unlikely to be stochastic.

However, such algorithms are not only complicated and impractical, but also suboptimal in several aspects. Ideally, we would like to achieve such best-of-both-worlds results via a simple algorithm, or, even better, via a simple adaptive regret bound, independent of the concrete algorithm, similar to the result of Theorem 6 from Lecture 3 for the full information setting. Recent research has made significant progress in this direction and indeed provides a positive answer to this question. The following result, taken from [Zimmert and Seldin, 2019], shows that if a certain adaptive regret bound holds, then the algorithm not only automatically achieves the best of the both stochastic and adversarial worlds, but also smoothly bridges these two worlds with a natural performance guarantee for any environments in between.

Theorem 3. *Suppose that an MAB algorithm always ensures for any action $a^\star \in [K]$,*

$$\bar{\mathcal{R}}_T \le \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \sum_{a \ne a^\star} \sqrt{\frac{p_t(a)}{t}}\right]\right), \tag{3}$$

where $p_t(a)$ is the probability of picking action a at time t. Then the same algorithm

- *always ensures $\bar{\mathcal{R}}_T = \mathcal{O}(\sqrt{TK})$, even if the environment is adversarial;*[1]
- *ensures $\bar{\mathcal{R}}_T = \mathcal{O}\left(\sum_{a\ne a^\star} \frac{\ln T}{\Delta_a} + \sqrt{C \sum_{a \ne a^\star} \frac{\ln T}{\Delta_a}}\right)$ if the environment is such that*

$$\bar{\mathcal{R}}_T \ge \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a \ne a^\star} p_t(a)\Delta(a)\right] - C \tag{4}$$

  for some action $a^\star$, gap measures $\Delta(a) > 0$ for $a \ne a^\star$, and a constant $C > 0$.

The first statement is clear: the algorithm is always minimax optimal in the adversarial world. The second statement is a bit obscure, and let's unpack its meaning. First, note that stochastic MAB satisfies the required condition (4) with $a^\star$ being the optimal action (assumed to be unique here for simplicity), $\Delta(a)$ being the suboptimality gap defined earlier, and $C = 0$. In fact, in this case Eq. (4) holds with equality. Thus, the theorem says that for stochastic MAB, the algorithm automatically enjoys the instance-optimal pseudo-regret $\mathcal{O}(\sum_{a \ne a^\star} \frac{\ln T}{\Delta_a})$.

But the second statement says much more than that. For example, consider a *corrupted stochastic setting* that smoothly interpolates between the stochastic setting and the adversarial setting: the environment first follows the protocol of stochastic MAB and generates losses $\ell'_1(a), \ldots, \ell'_T(a)$

[1] Note that for an oblivious adversary, pseudo-regret is the same as regular expected regret.

for each action a according to a fixed distribution with mean $\mu(a)$, then, an adversary arbitrarily corrupts the losses to generate the final loss sequence $\ell_1(a), \ldots, \ell_T(a)$, subject to a "budget" C such that $2\sum_{t=1}^{T} \|\ell_t - \ell'_t\|_\infty \leq C$. This setting satisfies (4) again since

$$\bar{\mathcal{R}}_T \geq \max_a \mathbb{E}\left[\sum_{t=1}^{T} \ell'_t(a_t) - \sum_{t=1}^{T} \ell'_t(a)\right] - C = \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a \neq a^\star} p_t(a)\Delta(a)\right] - C,$$

where $a^\star$ and $\Delta(a)$ are respectively the optimal action and the suboptimality gap of the original stochastic environment. In this case, the bound stated in Theorem 3, $\bar{\mathcal{R}}_T = \mathcal{O}\left(\sum_{a\neq a^\star} \frac{\ln T}{\Delta_a} + \sqrt{C\sum_{a\neq a^\star} \frac{\ln T}{\Delta_a}}\right)$, naturally increases in C (the amount of corruption), smoothly interpolating between the instance-optimal bound for stochastic MAB with no corruption and the $\sqrt{T}$ type of bound for adversarial MAB (since C is at most $2T$). Importantly, the theorem also implicitly implies that the algorithm does not need the knowledge of C!

More generally, condition (4) can be even weaker than the settings discussed above, since no independence (across actions or time) is really required. Below, we provide the proof for this theorem.

Proof. The first statement is by a direct application of Cauchy-Schwarz inequality following Eq. (3):

$$\bar{\mathcal{R}}_T \leq \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \sum_{a\neq a^\star} \sqrt{\frac{p_t(a)}{t}}\right]\right) = \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \sqrt{\frac{K\sum_{a\neq a^\star} p_t(a)}{t}}\right]\right) = \mathcal{O}(\sqrt{TK}).$$

For the second statement, we let B be the constant hidden in the $\mathcal{O}(\cdot)$ notation of Eq. (3), and then bound the regret by *a faction of itself*:

$$\bar{\mathcal{R}}_T \leq \mathbb{E}\left[B\sum_{t=1}^{T} \sum_{a\neq a^\star} \sqrt{\frac{p_t(a)}{t}}\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a\neq a^\star} \frac{p_t(a)\Delta(a)}{2z} + \frac{zB^2}{2t\Delta(a)}\right] \leq \frac{\bar{\mathcal{R}}_T + C}{2z} + 2zU,$$

where the second step is by AM-GM inequality and holds for any $z \geq 0$, and the third step uses condition (4), the fact $\sum_{t=1}^{T} \frac{1}{t} \leq 2\ln T$, and a shorthand $U = \frac{B^2}{2}\sum_{a\neq a^\star} \frac{\ln T}{\Delta(a)}$. As long as $z \geq 1/2$, rearranging shows $\bar{\mathcal{R}}_T \leq \frac{4z^2}{2z-1}U + \frac{C}{2z-1}$. With $x = 2z - 1$, this can be rewritten as $\bar{\mathcal{R}}_T \leq 2U + Ux + \frac{U+C}{x}$, which is of order $\mathcal{O}(U + \sqrt{(U+C)U}) = \mathcal{O}(U + \sqrt{CU})$ after picking the optimal x to minimize the bound (note that x only appears in the analysis and is not a parameter of the algorithm). This completes the proof. □

The next question, of course, is what algorithms achieve this magical Eq. (3). In fact, we have seen one such example already: FTRL with Tsallis entropy regularizer $\psi(p) = -2\sum_{a=1}^{K} \sqrt{p(a)}$. The only modification needed is that at time t, we use a time-varying learning rate $\eta_t = 1/\sqrt{t}$. We leave the proof to HW3, and only point out that it is quite surprising that such a simple algorithm exhibits strong optimality and adaptivity to many different types of environments simultaneously.

4 Stochastic Linear Bandits and LinUCB

Next, we come back to the stochastic setting and study a generalization of stochastic MAB. One issue of stochastic MAB is that it completely ignores the possible dependence across different actions, while in practice they are often correlated one way or another. Taking such correlation into account is especially important when the number of actions is huge. Here, we consider a setting where such correlation is represented in the simplest linear manner, known as *stochastic linear bandits*.

The learning protocol is as follows: for each round $t = 1, \ldots, T$,

1. An *arbitrary* set of actions $A_t \subset \mathbb{R}^d$ is revealed to the learner;
2. the learner picks an action $a_t \in A_t$ and observe its loss $c_t = \langle a_t, \theta^\star\rangle + \epsilon_t$ where $\theta^\star \in \mathbb{R}^d$ is a fixed and unknown parameter and $\epsilon_t \sim \mathcal{N}(0,1)$ is independent standard Gaussian noise.

Let $a_t^\star \in \operatorname{argmin}_{a \in A_t} \langle a, \theta^\star \rangle$ be an optimal action at time t. The pseudo-regret of this problem is naturally defined as

$$\bar{\mathcal{R}}_T = \mathbb{E}\left[\sum_{t=1}^{T} \langle a_t - a_t^\star, \theta^\star \rangle\right].$$

Note that stochastic MAB is essentially a special case of this model, where $d = K$, A_t is always $\{e_1, \ldots, e_d\}$ (the standard basis vectors of $\mathbb{R}^d$), and $\theta^\star = (\mu(1), \ldots, \mu(K))$ is the vector of loss mean for each action. The only slight difference is that now we consider Gaussian noise for each observation of the loss, which is less general compared to the arbitrary bounded noise model in stochastic MAB, but more general in that the noise can be unbounded.

In general, the stochastic linear bandit model is much more powerful since 1) it allows each action to come with an arbitrary "feature"; 2) the number of actions can be arbitrary (even infinite); and 3) the set of available actions can be different at different time. This allows us to capture real-life problems such as building a personalized news recommendation system [Li et al., 2010]. In this application, each time t corresponds to a visit of some user to the website. The available news articles at that time as well as the user's information are then used to generate a feature vector for each article. Afterwards, a linear bandit algorithm selects an action and recommends the corresponding article to the user. The loss is then based on whether the user clicks on the recommended article or not. It is assumed that the expected loss of an action can be perfectly predicted by an unknown linear predictor $\theta^\star$ (but we will see generalization to nonlinear models in the future).

Note that because of the changing action sets, it only makes sense to define the pseudo-regret so that it compares the expected loss of the algorithm to the expected loss of the best action *at each time*. This relates to the notion of dynamic regret discussed before. However, while in general sublinear dynamic regret is impossible, due to the stochastic assumption, regret of order $\mathcal{O}(\sqrt{T})$ is in fact achievable here as we will see soon.

Finally, without loss of generality, we make two scaling assumptions: $\max_{a \in A_t} \|a\|_2 \le 1$ for all t and $\|\theta^\star\|_2 \le 1$.

4.1 The LinUCB Algorithm

To solve this problem, let's apply the same "optimism in face of uncertainty" principle. The first step is to come up with the set of plausible environments that are consistent with the observed data. Here, the only parameter of the environment is the linear predictor $\theta^\star$, so the first goal would be to come up with a confidence set Θ_t for time t based on the observation $a_1, c_1, \ldots, a_t, c_t$, so that $\theta^\star \in \Theta_t$ with high probability. With such a confidence set, similarly to the UCB algorithm, at time $t+1$ we optimistically assume that the loss for each action $a \in A_{t+1}$ is

$$\text{LCB}_{t+1}(a) = \min_{\theta \in \Theta_t} \langle a, \theta \rangle,$$

and finally pick action $a_{t+1} = \operatorname{argmin}_{a \in A_{t+1}} \text{LCB}_{t+1}(a)$.

It remains to come up with the confidence set Θ_t. First, we need to figure out what the "center" of this set is. For UCB, the center of the confidence set is simply and naturally the empirical average of observations. For linear bandit, note that we are observing $c_s \approx \langle a_s, \theta^\star \rangle$ for $s = 1, \ldots, t$. It is thus natural to perform least square regression to obtain an estimate of $\theta^\star$ as the center:

$$\widehat{\theta}_t = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t} \left(\langle a_s, \theta \rangle - c_s\right)^2.$$

By direct calculations, one can verify $\widehat{\theta}_t = M_t^{-1} \sum_{s=1}^{t} c_s a_s$ where $M_t = \sum_{s=1}^{t} a_s a_s^\top$ is called the empirical covariance matrix and is assumed to be invertible for now. Note that this is consistent with UCB: when $A_t = \{e_1, \ldots, e_d\}$, M_t is a diagonal matrix with $M_t(i,i)$ being the number of times action i has been selected, and thus $\widehat{\theta}_t$ is exactly the vector of empirical mean for each action.

Next, we need to figure out what Θ_t should look like around the center $\widehat{\theta}_t$, that is, to understand the distribution of $\widehat{\theta}_t$. To get an intuition, we first ignore the fact that each a_s is itself random variables and think of them as fixed vectors (all assumptions will be dropped eventually). Then, the only randomness in $\widehat{\theta}_t$ comes from the noises $\epsilon_1, \ldots, \epsilon_t$. Plugging in $c_s = \langle a_s, \theta^\star \rangle + \epsilon_s$, we rewrite $\widehat{\theta}_t$ as

$$\widehat{\theta}_t = \left(M_t^{-1} \sum_{s=1}^t (\langle a_s, \theta^\star \rangle + \epsilon_s) a_s \right) = M_t^{-1} M_t \theta^\star + M_t^{-1} \sum_{s=1}^t \epsilon_s a_s = \theta^\star + M_t^{-1} Z_t$$

where $Z_t = \sum_{s=1}^t \epsilon_s a_s$ is a zero-mean d-dimensional Gaussian variable with covariance matrix

$$\mathbb{E}\left[Z_t Z_t^\top \right] = \sum_{s_1=1}^t \sum_{s_2=1}^t \mathbb{E}\left[\epsilon_{s_1} \epsilon_{s_2} \right] a_{s_1} a_{s_2}^\top = \sum_{s=1}^t \mathbb{E}\left[\epsilon_s^2 \right] a_s a_s^\top = M_t.$$

Therefore, the random variable $M_t^{1/2}(\widehat{\theta}_t - \theta^\star)$ is distributed as $\mathcal{N}(0, I)$, the d-dimensional standard Gaussian. By standard concentration results (specifically, tail bounds of the χ_d^2 distribution), the following holds for any $\delta \in (0,1)$ (details omitted)

$$\Pr\left(\left\| M_t^{1/2}(\widehat{\theta}_t - \theta^\star) \right\|_2^2 \le d + 2\sqrt{d \ln \tfrac{1}{\delta}} + 2 \ln \tfrac{1}{\delta} \right) \ge 1 - \delta.$$

Inspired by this, the confidence set can be constructed as (recall the notation $\|v\|_M = \sqrt{v^\top M v}$)

$$\Theta_t = \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \widehat{\theta}_t \right\|_{M_t}^2 \le d + 2\sqrt{d \ln \tfrac{1}{\delta}} + 2 \ln \tfrac{1}{\delta} \right\},$$

which is in fact an ellipsoid centered at $\widehat{\theta}_t$. (The set defined by $\|v\|_M \le 1$ is the standard analytic form of an ellipsoid centered at the origin, where the eigenvectors of M define the principal axes of the ellipsoid and the corresponding eigenvalues are the reciprocals of the square of the semi-axes.)

In the derivation above, we made two assumptions. First, M_t is invertible, which is not true until $a_1, \ldots, a_t$ span $\mathbb{R}^d$. This can be resolved by adding an ℓ_2-regularization to the least square regression

$$\widehat{\theta}_t = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{s=1}^t (\langle a_s, \theta \rangle - c_s)^2 + \|\theta\|_2^2 = M_t^{-1} \sum_{s=1}^t c_s a_s$$

where M_t is redefined as $I_d + \sum_{s=1}^t a_s a_s^\top$ (I_d is the d by d identity matrix) and is always invertible now. However, what is more difficult to get rid of is the second assumption that each a_s is fixed and not random. Fortunately, with some probability tools, this can still be addressed. We omit the details and only show the final lemma taken from [Abbasi-Yadkori et al., 2011].

Lemma 3 (confidence ellipsoid)**.** *For any stochastic linear bandit algorithm we have with probability at least $1 - 1/T$, $\theta^\star \in \Theta_t$ holds for all t, where*

$$\Theta_t = \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \widehat{\theta}_t \right\|_{M_t} \le \beta \right\},$$

$M_t = I_d + \sum_{s=1}^t a_s a_s^\top$, $\widehat{\theta}_t = M_t^{-1} \sum_{s=1}^t c_s a_s$, and $\beta = 1 + \sqrt{2 \ln T + d \ln\left(1 + \frac{T}{d}\right)} = \widetilde{\mathcal{O}}(\sqrt{d})$.

Finally, having this confidence set Θ_t, we can further simplify the algorithm by noting

$$\begin{aligned}
\text{LCB}_{t+1}(a) &= \min_{\theta \in \Theta_t} \langle a, \theta \rangle = \min_{\left\| \theta - \widehat{\theta}_t \right\|_{M_t} \le \beta} \langle a, \theta \rangle \\
&= \min_{\|\theta'\|_2 \le \beta} \left\langle a, M_t^{-\frac{1}{2}} \theta' + \widehat{\theta}_t \right\rangle & \text{(define } \theta' = M_t^{\frac{1}{2}}(\theta - \widehat{\theta}_t)\text{)} \\
&= \left\langle a, \widehat{\theta}_t \right\rangle + \min_{\|\theta'\|_2 \le \beta} \left\langle M_t^{-\frac{1}{2}} a, \theta' \right\rangle \\
&= \left\langle a, \widehat{\theta}_t \right\rangle - \beta \|a\|_{M_t^{-1}} & (\min_{\|v\|_2 \le 1} \langle x, v \rangle = -\|x\|_2)
\end{aligned}$$

and thus

$$a_{t+1} = \operatorname*{argmin}_{a \in A_{t+1}} \text{LCB}_{t+1}(a) = \operatorname*{argmin}_{a \in A_{t+1}} \left(\langle a, \widehat{\theta}_t \rangle - \beta \|a\|_{M_t^{-1}} \right).$$

This algorithm is called LinUCB (or OFUL, shorf for Optimism in Face of Uncertainty for Linear bandit). Very similar to UCB, the term $\langle a, \widehat{\theta}_t \rangle$ represents exploitation while the term $-\beta_t \|a\|_{M_t^{-1}}$ drives exploration for unobserved directions (since $\|a\|_{M_t^{-1}}$ is large when a represents a direction that is very different from the observed ones $a_1, \ldots, a_t$). Indeed, when $A_t = \{e_1, \ldots, e_d\}$, one can verify that LinUCB recovers UCB (other than the difference in the coefficient β).

4.2 Regret Analysis

The analysis of LinUCB is also a generalization of that for UCB, starting with the following key lemma enabled by optimism, which is an analogue of Lemma 2.

Lemma 4 (per-round regret)**.** *With probability at least $1 - \frac{1}{T}$, LinUCB ensures $\langle a_t - a_t^\star, \theta^\star \rangle \leq 2\beta \left\|a_t\right\|_{M_{t-1}^{-1}}$ for all t.*

Proof. Recall that under the event stated in Lemma 3, $\text{LCB}_t(a) = \min_{\theta \in \Theta_{t-1}} \langle a, \theta \rangle = \langle a, \widehat{\theta}_{t-1} \rangle - \beta \left\|a\right\|_{M_{t-1}^{-1}}$ is indeed a lower bound on the true loss $\langle a, \theta^\star \rangle$ for any action a. Therefore,

$$\begin{aligned}
\langle a_t - a_t^\star, \theta^\star \rangle &\leq \langle a_t, \theta^\star \rangle - \text{LCB}_t(a_t^\star) \\
&\leq \langle a_t, \theta^\star \rangle - \text{LCB}_t(a_t) && (a_t \text{ has the smallest LCB}) \\
&= \left\langle a_t, \theta^\star - \widehat{\theta}_{t-1} \right\rangle + \beta \left\|a_t\right\|_{M_{t-1}^{-1}} \\
&\leq \left\|a_t\right\|_{M_{t-1}^{-1}} \|\theta^\star - \widehat{\theta}_{t-1}\|_{M_{t-1}} + \beta \left\|a_t\right\|_{M_{t-1}^{-1}} && \text{(Hölder's inequality)} \\
&\leq 2\beta \left\|a_t\right\|_{M_{t-1}^{-1}}, && (\theta^\star \in \Theta_t)
\end{aligned}$$

which finishes the proof. □

We now proceed with a statement analogous to Theorem 2, whose proof is purely an exercise of linear algebra and is independent of the algorithm.

Theorem 4. *For any algorithm, we have $\sum_{t=1}^T \left\|a_t\right\|_{M_{t-1}^{-1}} \leq \widetilde{\mathcal{O}}(\sqrt{dT})$. Consequently, LinUCB with ensures $\bar{\mathcal{R}}_T = \widetilde{\mathcal{O}}(d\sqrt{T})$.*

Proof. First, we apply Cauchy-Schwarz and obtain $\sum_{t=1}^T \left\|a_t\right\|_{M_{t-1}^{-1}} \leq \sqrt{T \sum_{t=1}^T \left\|a_t\right\|_{M_{t-1}^{-1}}^2}$. Note that a slightly different expression $\left\|a_t\right\|_{M_t^{-1}}^2$ can be bounded as

$$\left\|a_t\right\|_{M_t^{-1}}^2 = \left\langle M_t^{-1}, a_t a_t^\top \right\rangle = \left\langle M_t^{-1}, M_t - M_{t-1} \right\rangle \leq \ln\det(M_t) - \ln\det(M_{t-1})$$

where the last step uses the concavity of the $\ln\det(M)$ function and that its gradient is exactly M^{-1}. Therefore, $\sum_{t=1}^T \left\|a_t\right\|_{M_t^{-1}}^2 \leq \ln\det(M_T) - \ln\det(M_0) = \ln\det(M_T)$. To bound $\det(M_T)$, we apply AM-GM inequality and the assumption $a_t^\top a_t \leq 1$:

$$\det(M_T) \leq \left(\frac{\textsc{Tr}(M_T)}{d}\right)^d = \left(\frac{d + \sum_{t=1}^T \textsc{Tr}(a_t a_t^\top)}{d}\right)^d = \left(1 + \frac{\sum_{t=1}^T a_t^\top a_t}{d}\right)^d \leq \left(1 + \frac{T}{d}\right)^d.$$

We have thus shown $\sum_{t=1}^T \left\|a_t\right\|_{M_t^{-1}}^2 \leq d \ln\left(1 + \frac{T}{d}\right)$, and it remains to bound the difference between $\sum_{t=1}^T \left\|a_t\right\|_{M_t^{-1}}^2$ and $\sum_{t=1}^T \left\|a_t\right\|_{M_{t-1}^{-1}}^2$. Indeed, they are very close:

$$\begin{aligned}
&\sum_{t=1}^T \left\|a_t\right\|_{M_{t-1}^{-1}}^2 - \sum_{t=1}^T \left\|a_t\right\|_{M_t^{-1}}^2 = \sum_{t=1}^T \left\langle M_{t-1}^{-1} - M_t^{-1}, a_t a_t^\top \right\rangle \\
&\leq \sum_{t=1}^T \left\langle M_{t-1}^{-1} - M_t^{-1}, I_d \right\rangle \leq \left\langle M_0^{-1}, I_d \right\rangle = d
\end{aligned}$$

where the first inequality is by the fact that both $M_{t-1}^{-1} - M_t^{-1}$ and $I_d - a_t a_t^\top$ are PSD matrices. Combining all results above proves the first statement. The second statement is now clear since $\sum_{t=1}^T \langle a_t - a_t^\star, \theta^\star \rangle$ is at most $2\beta \sum_{t=1}^T \left\|a_t\right\|_{M_{t-1}^{-1}}$ with probability at least $1 - 1/T$ (Lemma 4), and at most $2T$ with probability $1/T$. □

Importantly, this regret bound has no dependence on the number of actions at all! Instead, it pays for linear dependence on the dimension d of the features. While this does not recover the $\widetilde{\mathcal{O}}(\sqrt{TK})$ bound of UCB, such linear dependence on d is known to be generally unavoidable for this problem.

Can we also derive an instance-dependent regret bound analogous to Theorem 1 using some suboptimality gap as a difficulty measure? The answer is yes, via a quite simple argument.

Theorem 5. *Define the minimal suboptimal gap as* $\Delta = \min_t \min_{a \in A_t : \langle a - a_t^\star, \theta^\star \rangle > 0} \langle a - a_t^\star, \theta^\star \rangle$. *Then LinUCB ensures* $\bar{\mathcal{R}}_T = \mathcal{O}\left(\frac{(d \ln T)^2}{\Delta}\right)$.

Proof. Simply bound the per-round regret $\langle a_t - a_t^\star, \theta^\star \rangle$ by $\frac{1}{\Delta} \langle a_t - a_t^\star, \theta^\star \rangle^2$, which, under the event stated in Lemma 3, is further bounded by $\frac{4\beta^2}{\Delta} \|a_t\|^2_{M_{t-1}^{-1}}$. In the proof of Theorem 4, we have already shown $\sum_{t=1}^T \|a_t\|^2_{M_{t-1}^{-1}} \leq d + d \ln \left(1 + \frac{T}{d}\right)$. The proof is finished after trivially bounding the regret by $2T$ when the high-probability event does not hold. □

While this looks very similar to the UCB result, unlike the case for stochastic MAB, such a gap-dependent bound is *not* instance-optimal. In fact, somewhat surprisingly, it has been shown that optimism-based approaches can never be instance-optimal for stochastic linear bandits; see [Lattimore and Szepesvari, 2017]. Therefore, while the "optimism in face of uncertainty" principle is very powerful, especially in obtaining $\mathcal{O}(\sqrt{T})$ type of regret bounds, it is by no means always the best approach.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, 2011.

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

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1, 2012.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 467–475. PMLR, 2019.