

Lecture 7: February 19, 2024

Lecturer: Yishay Mansour

Scribe¹: Idan Barnea, Amit Attia, Noam Razin

1 Linear Bandits - Introduction

1.1 Motivation - Adversarial Online Shortest Paths

Consider the following task: A communication network is given (represented as a graph) and at every time t we want to send a message from node u to node v . How do we pick the best path?

One way to model the above problem is by the following adversarial model. At each time t :

- The adversary picks a cost (delay) for each edge $e \in E$ in the graph - $c_t(e) \in [0, 1]$. Indexing the edges of the graph, we can represent the costs as a vector of size $|E|$, $c_t = (c_t(1), c_t(2), \dots, c_t(|E|))$.
- The agent picks a path a_t from u to v given the edges E . With a slight abuse of notation we treat the path as both a sequence of edges, and as a vector of size $|E|$ over $\{0, 1\}$, where 1 means the edge is included in the path and 0 means it is not included in the path.
- The agent suffers the cost of the path, $a_t \cdot c_t = \sum_{e \in a_t} c_t(e)$.

We consider three types of information models:

- Full information: the agent observes the costs for all edges in the graph, i.e., $c_t(e)$ for all $e \in E$.
- Semi-bandit information: the agent observes the costs for edges included in the path used at time t , that is, $c_t(e)$ for all $e \in a_t$.
- Full-bandit information: the agent observes only the total cost of the path picked at time t , i.e., $a_t \cdot c_t = \sum_{e \in a_t} c_t(e)$.

1.2 Linear Bandit Model

Next we introduce the Linear Bandit model, a general model which includes the problem of online shortest paths. The model consists of the following:

- A set of actions $A \subseteq \mathbb{R}^d$. For example $A = \{a : a \text{ is a path from } u \text{ to } v\}$.
- A set of costs/gains C . For example $C = \{0, 1\}^d$ or $C = [-1, +1]^d$.
- The regret we will consider is

$$\text{regret} = \sum_{t=1}^T a_t \cdot c_t - \min_{a \in A} a \cdot \sum_{t=1}^T c_t = \sum_{t=1}^T a_t \cdot c_t - \min_{a \in A} a \cdot \sum_{t=1}^T c_t.$$

In addition we make the following assumption:

- The diameter of A is bounded:

$$\forall a, a' \in A : \|a - a'\|_1 \leq D.$$

- The costs are bounded:

$$\forall a \in A, c \in C : |a \cdot c| \leq R.$$

¹Based on previous notes by Alon Mendelson and Yoav Galtzur.

- The norm of the losses is bounded:

$$\forall c \in C : \|c\|_1 \leq s.$$

Our focus is the ℓ_1 norm, though other norms can be used with appropriate care.

2 Linear Bandits With Full Information

In the full information setting [3], by the end of every round we observe the entire costs vector c_t .

2.0.1 Regret Analysis

In this setting, the best arm has the following cost

$$\text{min-cost}_T = \min_{a \in A} \sum_{t=1}^T a \cdot c_t = M \left(\sum_{t=1}^T c_t \right) \cdot \left(\sum_{t=1}^T c_t \right)$$

2.1 Follow The Leader

Follow the Leader will act choose the best action according to all previous costs. Since finding the best action may be a hard problem, we'll assume oracle access to M , which solves the static optimization problem.

Algorithm 1 Follow The Leader

Require: Oracle $M(c) := \text{argmin}_{a \in A} a \cdot c$

```

for  $t = 1$  to  $T$  do
   $a_t = M \left( \sum_{\tau=0}^{t-1} c_\tau \right)$ 
  Observe  $c_t$ 
  Pay cost  $a_t \cdot c_t$ 
end for

```

2.1.1 Linear Regret Example

As seen in the example below, Follow the Leader (i.e., picking the best action until now in a greedy manner), can lead to poor results. Specifically, linear regret in T . Consider the case in which there are two actions $A = \{(0,1), (1,0)\}$, and the costs are defined as follows:

$$c_t = \begin{cases} (1/3, 2/3) & t = 1 \\ (0, 1) & t > 1 \wedge t \pmod 2 = 1 \\ (1, 0) & t > 1 \wedge t \pmod 2 = 0 \end{cases}$$

If the first arm was $(0,1)$ then it will suffer cost of $2/3$ and will choose next round arm $(1,0)$. This arm will suffer in round $t = 2$ cost of 1, and the best arm would be $(0,1)$. At $t = 3$ it will get again cost of 1, when the other arm, $(1,0)$, would be the best arm. This alternation will continue until the end of the horizon. Similar analysis for starting with the second arm would lead to that the algorithm suffers cost $\geq T - 2/3$. Where as the best fixed action is $(1,0)$, with total cost $\leq \frac{T}{2}$. Leading to linear regret.

$$\text{regret} = \text{cost}_{FL} - \text{min-cost}_T \geq T - \frac{2}{3} - \frac{T}{2} = \Omega(T)$$

We've seen such examples for worst-case behavior in other cases where deterministic algorithms are used. Our solution then will be adding randomness to this algorithm.

2.2 Follow The Perturbed Leader - FPL

From now on we'll use the notation $C_{1:t}$ to represent $\sum_{\tau=1}^t c_\tau$.

2.2.1 Follow The Perturbed Leader - Algorithm and Regret

Algorithm 2 Follow The Perturbed Leader (FPL)

Require: Oracle M , perturbation parameter $\eta \leq 1$

```

for  $t = 1$  to  $T$  do
  Sample  $q_t$  uniformly from  $\left[0, \frac{1}{\eta}\right]^d$ 
   $a_t = M\left(q_t + \sum_{\tau=0}^{t-1} c_\tau\right)$ 
  Observe  $c_t$ 
  Pay cost  $a_t \cdot c_t$ 
end for

```

Theorem 1 (FPL regret) For any set of costs $c_1, \dots, c_T \in C$, the FPL algorithm with $\eta \leq 1$ satisfies

$$\mathbb{E}[\text{cost}_{FPL}] \leq \text{min-cost}_T + \eta R s T + \frac{D}{\eta},$$

where $\forall a, a' \in A$ ($\|a - a'\|_1 \leq D$), $\forall a \in A$ $c \in C$ ($|a \cdot c| \leq R$) and $\forall c \in C$ ($\|c\|_1 \leq s$)

Corollary 2 For $\eta = \sqrt{\frac{D}{R s T}}$, Regret is $2\sqrt{R s D T}$

Before proving the regret bound, we'll analyze an 'algorithm' which observes c_t before choosing a_t to get some intuition.

2.2.2 Intuition - Be The Leader

We will start with intuition for the FPL. Consider the illegal(!) algorithm 'Be The Leader', which first observe the current cost vector and only then picks a_t . This is not a legal algorithm since it depends on unobservable information, i.e., c_t .

Algorithm 3 Be The Leader

Require: Oracle M

```

for  $t = 1$  to  $T$  do
  Observe  $c_t$ 
   $a_t = M(C_{1:t})$ 
  Pay cost  $a_t \cdot c_t$ 
end for

```

We show that BTL has zero regret.

Theorem 3 (BTL has no regret) Be the Leader has no regret, that is

$$\sum_{t=1}^T M(C_{1:t}) \cdot c_t \leq M(C_{1:T}) \cdot C_{1:T}$$

Proof: By induction on T .

For $T = 1$, the two sides of the inequality are identical and the claim holds. Assume correctness for T .

$$\begin{aligned} \sum_{t=1}^{T+1} M(C_{1:t}) \cdot c_t &\leq M(C_{1:T}) \cdot C_{1:T} + M(C_{1:T+1}) \cdot c_{T+1} \\ &\leq M(C_{1:T+1}) \cdot C_{1:T} + M(C_{1:T+1}) \cdot c_{T+1} \\ &= M(C_{1:T+1}) \cdot C_{1:T+1} \end{aligned}$$

□

Where the first inequality holds because of linearity and the induction hypothesis, and the second inequality holds since $M(C_{1:T})$ is optimal w.r.t. cost $C_{1:T}$.

2.2.3 FPL Regret Analysis

In the lemma below we upper-bound the cost of Be The Leader when perturbations are added.

Lemma 4 For every set of perturbations q_0, \dots, q_T (where $q_0 = 0$)

$$\sum_{t=1}^T \mathbb{E}[M(C_{1:t} + q_t) \cdot c_t] \leq M(C_{1:T}) \cdot C_{1:T} + D \mathbb{E} \left[\sum_{t=1}^T \|q_t - q_{t-1}\|_\infty \right]$$

Proof: Consider the BTL algorithm, but with cost at time t being $c_t + q_t - q_{t-1}$. Note that for any t , the total cost from time 1 to t is:

$$\sum_{\tau=1}^t c_\tau + q_\tau - q_{\tau-1} = C_{1:t} + q_t$$

Now we'll look at the cost of this algorithm

$$\begin{aligned} \sum_{t=1}^T M(C_{1:t} + q_t) \cdot (c_t + q_t - q_{t-1}) &\leq M(C_{1:T} + q_T) \cdot (C_{1:T} + q_T) \\ &\leq M(C_{1:T}) \cdot (C_{1:T} + q_T) \\ &= M(C_{1:T}) \cdot C_{1:T} + \sum_{t=1}^T M(C_{1:T}) \cdot (q_t - q_{t-1}) \end{aligned}$$

Where the first inequality holds since BTL has zero regret, and the second since $M(C_{1:T} + q_T)$ is optimal w.r.t. cost $C_{1:T} + q_T$.

We'll rearrange the above expression and get

$$\begin{aligned} \sum_{t=1}^T M(C_{1:t} + q_t) \cdot c_t &\leq M(C_{1:T}) \cdot C_{1:T} + \sum_{t=1}^T (M(C_{1:T}) - M(C_{1:t} + q_t))(q_t - q_{t-1}) \\ &\leq M(C_{1:T}) \cdot C_{1:T} + \sum_{t=1}^T \underbrace{\|M(C_{1:T}) - M(C_{1:t} + q_t)\|_1}_{\leq D} \cdot \|q_t - q_{t-1}\|_\infty \\ &= M(C_{1:T}) \cdot C_{1:T} + D \sum_{t=1}^T \|q_t - q_{t-1}\|_\infty \end{aligned}$$

□

Corollary 5

$$\sum_{t=1}^T \mathbb{E}[M(C_{1:t} + q_t) \cdot c_t] \leq M(C_{1:T}) \cdot C_{1:T} + D/\eta = \text{min-cost}_T + D/\eta$$

Proof: Instead of sampling $q_t \leftarrow [0, 1/\eta]^d$ in every time step, we can sample $q_1 \leftarrow [0, 1/\eta]^d$ once and use it for every t . Since q_1 is taken from the same distribution as the other q_t , they are independent, and the lemma is true for every set of q_1, \dots, q_T , we can use the same analysis as in the lemma:

$$\sum_{t=1}^T \mathbb{E}[M(C_{1:t} + q_t) \cdot c_t] \leq M(C_{1:T}) \cdot C_{1:T} + D \mathbb{E}[\|q_1 - q_0\|_\infty] \leq M(C_{1:T}) \cdot C_{1:T} + D/\eta$$

Where the last inequality is due to the fact that $q_1 \in [0, 1/\eta]^d$ \square

Now we are ready to prove the FPL regret bound:

Proof:

Let $\text{cost}_{FPL}^t = \mathbb{E}[M(C_{1:t-1} + q_t)c_t]$ and $\text{cost}_{BTL}^t = \mathbb{E}[M(C_{1:t} + q_t)c_t]$ be the costs of FPL and BTL + perturbation at time t , respectively.

Note that the only difference between the costs is the argument passed to M : $C_{1:t-1} + q_t$ as oppose to $C_{1:t} + q_t$. These two sets of values only differ by c_t and as we'll show in the next lemma, lemma 6, their overlapping will help to bound the regret. Denote the probability of sampling q_t such that $C_{1:t-1} + q_t$ and $C_{1:t} + q_t$ are intersecting by f . Upper bound the cost of the other events by R and we'll get

$$\text{cost}_{FPL}^t = \mathbb{E}[M(C_{1:t-1} + q_t)c_t] \leq \mathbb{E}[M(C_{1:t} + q_t)c_t] + (1 - f)R$$

Following Lemma 6 below, where $z := c_t$, we get $1 - f \leq \eta \|c_t\|_1$. For one time-step we get:

$$\text{cost}_{FPL}^t = \mathbb{E}[M(C_{1:t-1} + q_t)c_t] \leq \mathbb{E}[M(C_{1:t} + q_t)c_t] + R\eta \|c_t\|_1 \leq \mathbb{E}[M(C_{1:t} + q_t)c_t] + R\eta s$$

Summing over the time-steps t :

$$\text{cost}_{FPL} \leq \sum_{t=1}^T \mathbb{E}[M(C_{1:t} + q_t)c_t] + R\eta s \leq \text{min-cost}_T + D/\eta + TR\eta s$$

\square

Lemma 6 For every $z \in \mathbb{R}^d$, the volume of the overlapping are between the cubes $[0, 1/\eta]^d$ and $z + [0, 1/\eta]^d$ is at least $1 - \eta \|z\|_1$

Proof: Pick some x uniformly from $[0, 1/\eta]^d$. If $x \notin z + [0, 1/\eta]^d$, then there exists some $i \in [d]$ such that $x_i \notin z_i + [0, 1/\eta]$. Since x is picked uniformly at random the event of $x_i \notin z_i + [0, 1/\eta]$ occurs with probability at most $\eta |z_i|$, and by union bound the event in which all coordinates of x are outside occurs with probability at most $\eta \|z\|_1$. The event that x is inside $z + [0, 1/\eta]^d$ happens with probability at least $1 - \eta \|z\|_1$. \square

Corollary 7 Let $D_1 \sim \text{Unif}([0, 1/\eta]^d)$, $D_2 \sim \text{Unif}(z + [0, 1/\eta]^d)$, then

$$\|D_1 - D_2\|_1 \leq 2\eta \|z\|_1$$

Proof:

$$\|D_1 - D_2\|_1 = 2\|D_1 - D_2\|_{TV} \leq 2\eta \|z\|_1$$

\square

3 Full Bandit Model

We proceed to the full bandit setting, in which the learner only observes the cost of the chosen action a_t at each round t .

3.1 Reduction to Full Information

Following is a reduction from the full bandit model to full information [1], similar to the reduction presented in lecture 6. The general reduction paradigm appears in Figure 1. For simplicity, we will assume that the standard basis of \mathbb{R}^d are valid actions (namely $e_1, \dots, e_d \in A$, where the i 'th index of e_i is 1 and the rest of the indices are 0). We will later show how to alleviate the assumption. Algorithm 4 use a full information algorithm, calling it once per block of B rounds. In each block it alternates between the action selected by the full information algorithm and a random exploration of the standard

basis actions, which are used to construct an unbiased estimator of the block's cost vector for updating the full information algorithm.

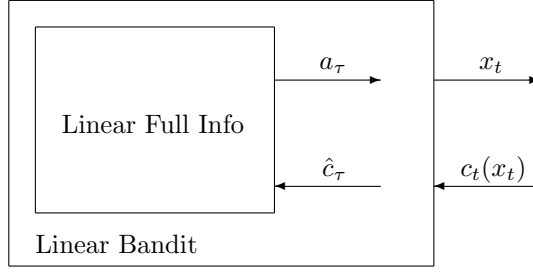


Figure 1: Paradigm for reducing full bandit feedback to full information.

Algorithm 4 Full MAB by reduction to full information

Require: Linear Bandit Full Information algorithm FI, block size B

```

for  $\tau = 1, \dots, \frac{T}{B}$  do
   $T_\tau \leftarrow \{(\tau - 1)B + 1, \dots, \tau B\}$ 
  Receive action  $a_i$  from FI
  For each  $e \in \{e_1, \dots, e_d\}$  sample  $s_\tau(e)$  from  $T_\tau$  (with  $s_\tau(e_i) \neq s_\tau(e_j)$  if  $i \neq j$ )
  for  $t \in T_\tau$  do
    if  $t = s_\tau(e_i)$  for some  $i \in [d]$  then
      Select  $x_t = e_i$ 
    else
      Select  $x_t = a_\tau$ 
    end if
    Observe  $c_t(x_t) = c_t \cdot x_t$ 
  end for
  Pass  $\hat{c}_\tau = (c_{s_\tau(e_1)}(e_1), \dots, c_{s_\tau(e_d)}(e_d))$  to FI
end for

```

Note that:

- In each block, \hat{c}_τ is an unbiased estimator of the cost vector in that block. This is due to the fact that the exploration of e_i is chosen randomly inside the block (as we saw in lecture 6).
- We call the full information algorithm once per block, using the fact that $e_i \cdot c_t$ extracts the i 'th coordinate of c_t and that \hat{c}_τ is unbiased to construct a full information cost vector.

Next is the regret guarantee of Algorithm 4.

Theorem 8 Using the FPL full information algorithm (Algorithm 2) and a block size $B > 0$, the expected regret of Algorithm 4 satisfy

$$\text{regret} \leq \sqrt{BRSDT} + \frac{TRd}{B}.$$

Setting $B = \frac{T^{1/3}R^{1/3}d^{2/3}}{S^{1/3}D^{1/3}}$, the regret satisfy

$$\text{regret} = O(T^{2/3}s^{1/3}d^{1/3}R^{2/3}D^{1/3}).$$

Proof: The number of rounds for the internal full information algorithm is $\frac{T}{B}$. By Theorem 1, for every $a \in A$,

$$\mathbb{E} \left[\sum_{\tau=1}^{T/B} \hat{c}_\tau \cdot a_\tau \right] \leq a \cdot \mathbb{E} \left[\sum_{\tau=1}^{T/B} \hat{c}_\tau \right] + 2\sqrt{RsD\frac{T}{B}}.$$

Suffering the maximal cost per exploration round of R , with d exploration rounds per block,

$$\mathbb{E} \left[\sum_{t=1}^T c_t \cdot x_t \right] \leq \mathbb{E} \left[\sum_{\tau=0}^{T/B-1} \sum_{i=1}^B c_{\tau B+i} \cdot a_\tau \right] + \frac{TRd}{B} \leq B \mathbb{E} \left[\sum_{\tau=1}^{T/B} \hat{c}_\tau \cdot a_\tau \right] + \frac{TRd}{B},$$

where the last inequality follows by \hat{c}_τ being an unbiased estimator of the cost at block τ . Plugging the regret bound of FPL,

$$\mathbb{E} \left[\sum_{t=1}^T c_t \cdot x_t \right] \leq Ba \cdot \mathbb{E} \left[\sum_{\tau=1}^{T/B} \hat{c}_\tau \right] + 2\sqrt{BRsDT} + \frac{TRd}{B}.$$

Again exploiting the fact that \hat{c}_τ is an unbiased estimator, we conclude that

$$\mathbb{E} \left[\sum_{t=1}^T c_t \cdot x_t \right] \leq a \cdot \sum_{t=1}^T c_t + 2\sqrt{BRsDT} + \frac{TRd}{B},$$

which yields the regret bound. Setting $B = \frac{T^{1/3}d^{2/3}R^{1/3}}{s^{1/3}D^{1/3}}$, we obtain the second regret bound. \square

3.2 Alleviating the Natural Basis Assumption

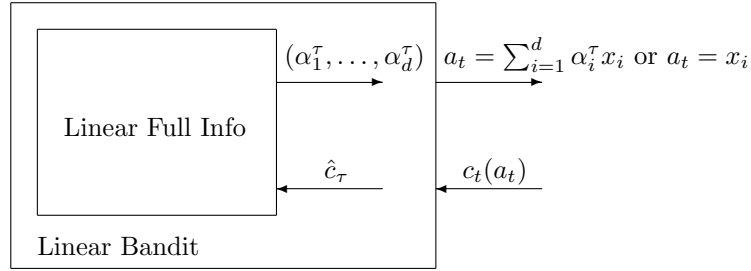


Figure 2: Modifying Algorithm 4 to accommodate a general basis x_1, \dots, x_d .

Algorithm 4 assumes that $e_1, \dots, e_d \in A$, which are used to provide the costs for the full information algorithm. While this assumption does not hold in general (for example in our communication network example), we can replace it with some basis $\{x_1, \dots, x_d\} \subseteq A$ of $\text{span}(A)$.² Each action $a \in A$ can be represented as $a = \sum_{i=1}^d \alpha_i x_i$ for $\alpha_1, \dots, \alpha_d \in \mathbb{R}$, and instead of exploring e_1, \dots, e_d the algorithm will explore x_1, \dots, x_d . To communicate with the full information algorithm, each action $a = \sum_{i=1}^d \alpha_i x_i$ will be presented to the full information algorithm as $\{\alpha_1, \dots, \alpha_d\}$. One challenge still remains, to show that the diameter of the actions introduced to the full information algorithm are bounded, that is, for any $a, a' \in A$ such that $a = \sum_{i=1}^d \alpha_i x_i$ and $a' = \sum_{i=1}^d \alpha'_i x_i$,

$$\sum_{i=1}^d |\alpha_i - \alpha'_i| \leq D$$

for some $D > 0$. While this is not true for any basis, next we will show that there is always a basis $\{x_1, \dots, x_d\} \in A$ such that for all $i \in [d]$, $\alpha_i \in [-1, 1]$.³ Such basis will ensure that $\sum_{i=1}^d |\alpha_i - \alpha'_i| \leq 2d$. Generally finding such basis is hard, but there is an efficient algorithm for finding a 2-approximation.

²Without loss of generality we assume that the dimension of $\text{span}(A)$ is d . If no such basis exists, we can simply represent A using a basis with size smaller than d .

³The costs and the norm of the costs vectors are bounded by the problem parameters for any basis.

Definition 9 (Barycentric spanner) A set $\{x_1, \dots, x_d\} \subseteq A \subseteq \mathbb{R}^d$ is a barycentric spanner if

$$\forall a \in A, \exists \alpha_1, \dots, \alpha_d \in [-1, 1] \text{ s.t. } \sum_{i=1}^d \alpha_i x_i = a$$

Lemma 10 For every compact set A there exists a barycentric spanner.

Proof: Let A be a compact set. Assume without loss of generality that $\text{span}(A) = \mathbb{R}^d$. Let $x_1, \dots, x_d \in A$ such that $\det(x_1, \dots, x_d)$ is maximal. Fix $a \in A$ and let $a = \sum_{i=1}^d \alpha_i x_i$ for some $\alpha_1, \dots, \alpha_d \in \mathbb{R}$ (if x_1, \dots, x_d is not a basis than the determinant above is zero, which contradicts the span assumption). For any $i \in [d]$,

$$\begin{aligned} |\det(x_1, \dots, x_{i-1}, a, x_{i+1}, \dots, x_n)| &= \left| \det \left(x_1, \dots, x_{i-1}, \sum_{j=1}^d \alpha_j x_j, x_{i+1}, \dots, x_n \right) \right| \\ &= \sum_{j=1}^d |\alpha_j| |\det(x_1, \dots, x_{i-1}, x_j, x_{i+1}, \dots, x_n)| \quad (\det \text{ is multilinear}) \\ &= |\alpha_i| |\det(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n)|. \quad (\text{if } i \neq j \text{ the determinant is } 0) \end{aligned}$$

By the definition of x_1, \dots, x_d (maximal determinant),

$$|\det(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n)| \geq |\det(x_1, \dots, x_{i-1}, a, x_{i+1}, \dots, x_n)|,$$

which implies that $|\alpha_i| \leq 1$. As this is true for any $i \in [d]$, $\{x_1, \dots, x_d\}$ is a barycentric spanner. \square

4 Stochastic Linear Bandits

We switch from an adversarial setting to a stochastic setting (see Chapter 19 in [4]). For convenience, we will consider maximizing rewards as opposed to minimizing costs.

4.1 Setting

Let $A \subseteq \mathbb{R}^d$ be the set of actions. We make the following assumptions regarding the stochastic linear bandits model.

- Each action $a \in A$ satisfies $\|a\|_2 \leq 1$.
- There exists $\theta^* \in \mathbb{R}^d$, satisfying $\|\theta^*\|_2 \leq 1$, such that the reward observed for an action $a \in A$ is given by $r(a) = \theta^* \cdot a + w$, where w is a 1-subgaussian random variable. Note that this implies $\mathbb{E}[r(a)] = \theta^* \cdot a$.

The regret under this model is defined by:

$$\text{regret} = \sum_{t=1}^T \max_{a \in A} a \cdot \theta^* - \sum_{t=1}^T a_t \cdot \theta^* = \sum_{t=1}^T \max_{a \in A} (a - a_t) \cdot \theta^*.$$

4.2 UCB Approach to Stochastic Linear Bandits

We tackle the stochastic linear bandits problem by generalizing the UCB algorithm. This approach goes by several different names, e.g.:

- LinRel - Linear Reinforcement Learning;
- LinUCB - Linear UCB; and

- OFUL - Optimism in face of uncertainty for linear.

We will refer to it by LinUCB. Similarly to UCB, the idea behind LinUCB is to maintain confidence sets $C_t \subseteq \mathbb{R}^d$ such that $\theta^* \in C_t$ with high probability, for every $t \in \{1, \dots, T\}$. The confidence sets are then used for deriving upper bounds on the expected reward of actions in A , based on which actions are chosen — see Algorithm 5. Below we describe how the confidence sets are constructed. Intuitively, we would like them to contain θ^* (with high probability) while being as small as possible.

Algorithm 5 LinUCB

Require: Set of actions $A \subseteq \mathbb{R}^d$ and vector $\theta^* \in \mathbb{R}^d$

for $t = 1$ to T **do**
 compute confidence set C_t
 $UCB_t(a) \leftarrow \max_{\theta \in C_t} \theta \cdot a$
 $a_t \leftarrow \arg \max_{a \in A} UCB_t(a)$
end for

4.2.1 Estimating θ^*

At time step $t \in \{1, \dots, T\}$, towards building the confidence set we compute an empirical estimate of θ^* . Given the information we have gathered after t steps — (a_τ, r_τ) for $1 \leq \tau \leq t$ — we do so by solving the following regularized least-squares problem:

$$\hat{\theta}_t = \arg \min_{\theta} \sum_{\tau=1}^t (r_\tau - \theta \cdot a_\tau)^2 + \lambda \|\theta\|_2^2,$$

where $\lambda > 0$. Note that λ ensures the objective has a unique minimizer, even when a_1, \dots, a_t do not span \mathbb{R}^d . The solution to this minimization problem can be written in closed form as:

$$\hat{\theta}_t = V_t^{-1} \sum_{\tau=1}^t a_\tau r_\tau,$$

where $V_0 = \lambda \cdot I \in \mathbb{R}^{d \times d}$ and $V_t = V_0 + \sum_{\tau=1}^t a_\tau a_\tau^\top$.

4.2.2 Constructing the Confidence Sets

We make two quick reminders that will be used in the construction process.

- For $x \in \mathbb{R}^d$ and positive semidefinite $M \in \mathbb{R}^{d \times d}$, the norm of x induced by M is denoted by $\|x\|_M^2 := x^\top M x$.
- For any $x, z \in \mathbb{R}^d$ it holds that $x^\top z \leq \|x\|_M \|z\|_{M^{-1}}$.

Now, at time t we take the estimate $\hat{\theta}_{t-1}$ to be the center of C_t and let:

$$C_t = \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}}^2 \leq \beta_t\}.$$

That is, the confidence set C_t is an ellipsoid centered at $\hat{\theta}_{t-1}$, whose principle axes are the eigenvectors of V_{t-1} . For $\delta \in (0, 1)$, we choose $\beta_t = \sqrt{\lambda} + \sqrt{2 \log \frac{1}{\delta} + d \log(\frac{d\lambda+t}{d\lambda})}$, meaning that β_t grows logarithmically with t , and $1 \leq \beta_1 \leq \beta_2 \leq \dots \leq \beta_T$. To better understand the intuition behind this choice of β_t , we can look at the definition of C_t and write the condition on $\theta \in C_t$ as:

$$(\theta - \hat{\theta}_{t-1})^\top V_{t-1} (\theta - \hat{\theta}_{t-1}) \leq \beta_t.$$

Typically, V_{t-1} grows linearly time while β_t grows only logarithmically. As a result, the size of the confidence set shrinks (and at a relatively fast rate).

4.2.3 Regret Analysis

Before bounding the regret of LinUCB, we prove the following lemma.

Lemma 11 $\sum_{t=1}^T \min(1, \|a_t\|_{V_{t-1}}^2) \leq 2 \log \frac{\det(V_T)}{\det(V_0)} \leq 2d \log \frac{d\lambda + T}{d\lambda}$

Proof: For any $u \in [0, 1]$ it holds that $\min(1, u) \leq 2 \ln(1 + u)$, and so:

$$\sum_{t=1}^T \min(1, \|a_t\|_{V_{t-1}}^2) \leq 2 \sum_{t=1}^T \log(1 + \|a_t\|_{V_{t-1}}^2).$$

We now show that:

$$\sum_{t=1}^T \log(1 + \|a_t\|_{V_{t-1}}^2) = \log \frac{\det(V_T)}{\det(V_0)}.$$

For $t \geq 1$:

$$V_t = V_{t-1} + a_t a_t^\top = V_{t-1}^{\frac{1}{2}} \left(I + V_{t-1}^{-\frac{1}{2}} a_t a_t^\top V_{t-1}^{-\frac{1}{2}} \right) V_{t-1}^{\frac{1}{2}}.$$

Since the determinant of a product of matrices is equal to the product of the respective determinants, we get that:

$$\det(V_t) = \det(V_{t-1}) \det \left(I + V_{t-1}^{-\frac{1}{2}} a_t a_t^\top V_{t-1}^{-\frac{1}{2}} \right).$$

The eigenvalues of the matrix $I + z z^\top \in \mathbb{R}^d$, for any $z \in \mathbb{R}^d$, are 1 and $1 + \|z\|_2^2$, with the multiplicity of the former being $d - 1$. Along with the fact that the determinant of a symmetric matrix is equal to the product of its eigenvalues, this implies that $\det \left(I + V_{t-1}^{-\frac{1}{2}} a_t a_t^\top V_{t-1}^{-\frac{1}{2}} \right) = 1 + \|a_t\|_{V_{t-1}}^2$ and:

$$\det(V_T) = \det(V_0) \prod_{t=1}^T (1 + \|a_t\|_{V_{t-1}}^2).$$

Dividing by $\det(V_0)$ and taking a log of both sides, we get:

$$\sum_{t=1}^T \log(1 + \|a_t\|_{V_{t-1}}^2) = \log \frac{\det(V_T)}{\det(V_0)},$$

which is the sought-after equality.

For the second part of the inequality, let $\lambda_1, \dots, \lambda_d$ be the eigenvalues of V_T . By the inequality of arithmetic and geometric means, we have that $\det(V_T) = \prod_{i=1}^d \lambda_i \leq \left(\frac{1}{d} \sum_{i=1}^d \lambda_i \right)^d = \left(\frac{1}{d} \text{tr}(V_T) \right)^d$. Since $\det(V_0) = \lambda^d$, and the trace of V_T can be bounded as follows:

$$\text{tr}(V_T) = \text{tr}(V_0) + \sum_{t=1}^T \text{tr}(a_t a_t^\top) \leq d\lambda + T,$$

we conclude:

$$\log \frac{\det(V_T)}{\det(V_0)} \leq \log \frac{\left(\frac{1}{d} \text{tr}(V_T) \right)^d}{\lambda^d} \leq d \log \frac{d\lambda + T}{d\lambda}.$$

□

With Lemma 11 in hand, we can upper bound the regret of LinUCB under the assumption that the confidence sets contain θ^* with high probability (see Chapter 20 in [4] for a proof that the confidence sets are valid).

Theorem 12 *Let $\delta \in (0, 1)$ and assume that with probability at least $1 - \delta$ for all $t \in \{1, \dots, T\}$ it holds that $\theta^* \in C_t$. Then, with probability at least $1 - \delta$ the regret of LinUCB satisfies:*

$$\text{regret} \leq \sqrt{8T\beta_T \log \frac{\det(V_T)}{\det(V_0)}} \leq \sqrt{8Td\beta_T \log \frac{d\lambda + T}{d\lambda}}.$$

Proof: Denote by $a^* \in \arg \min_{a \in A} a \cdot \theta^*$ an optimal action. If at time t the algorithm played action a_t , then there exists a vector $\tilde{\theta}_t \in C_t$ for which:

$$\theta^* \cdot a^* \leq UCB(a^*) \leq UCB(a_t) = \tilde{\theta}_t \cdot a_t.$$

Subtracting $\theta^* \cdot a_t$ from both sides of the inequality leads to:

$$\theta^* \cdot (a^* - a_t) \leq (\tilde{\theta}_t - \theta^*) \cdot a_t,$$

where the term on the left hand side is the regret at time t , which we denote by $regret_t$. Thus:

$$regret_t = \theta^* \cdot (a^* - a_t) \leq (\tilde{\theta}_t - \theta^*) \cdot a_t \leq \|a_t\|_{V_{t-1}^{-1}} \|\tilde{\theta}_t - \theta^*\|_{V_{t-1}},$$

where the last inequality is due to $x^\top z \leq \|x\|_M \|z\|_{M^{-1}}$ for any $x, z \in \mathbb{R}^d$ and positive definite $M \in \mathbb{R}^{d \times d}$. Since $\tilde{\theta}_t, \theta^* \in C_t$, by the triangle inequality and the definition of C_t we have that $\|\tilde{\theta}_t - \theta^*\|_{V_{t-1}} \leq 2\sqrt{\beta_t}$. Hence:

$$regret_t \leq 2\|a_t\|_{V_{t-1}^{-1}} \sqrt{\beta_t}.$$

Furthermore, recalling that $\|\theta^*\|_2^2 \leq 1$ and $\|a\|_2^2 \leq 1$, the Cauchy-Schwarz inequality implies that $|\theta^* \cdot a| \leq 1$. Consequently, $regret_t \leq 2$, which combined with the upper bound above for $regret_t$ yields:

$$regret_t \leq \min\left(2, 2\sqrt{\beta_t}\|a_t\|_{V_{t-1}^{-1}}\right) \leq 2\sqrt{\beta_t} \min\left(1, \|a_t\|_{V_{t-1}^{-1}}\right),$$

where the last transition is valid since $\beta_t \geq 1$. Now, by the Cauchy-Schwarz inequality:

$$regret = \sum_{t=1}^T regret_t \leq \sqrt{T \sum_{t=1}^T regret_t^2} \leq 2\sqrt{T \beta_T \sum_{t=1}^T \min\left(1, \|a_t\|_{V_{t-1}^{-1}}\right)},$$

where we used the fact that $\beta_t \leq \beta_T$ for all $t \in \{1, \dots, T\}$. Lastly, applying Lemma 11 concludes the proof. \square

As an immediate corollary of the above theorem we get a bound on the expected regret of LinUCB.

Corollary 13 *The expected regret of LinUCB is bounded by:*

$$\mathbb{E}[regret] = O\left(d\sqrt{T} \log T\right).$$

5 Further Reading

For more information see Chapter 7 in [5] and [2].

References

- [1] Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.
- [2] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.
- [3] Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.
- [4] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- [5] Aleksandrs Slivkins. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*, 2019.