Stochastic bandits (Part 2)

Pierre Gaillard

Reminder from last lectures

We recall the setting of stochastic multi-armed bandit in Setting 1.

Unknown parameters: K probability distributions $\nu_1, \dots, \nu_K$ on $[0,1]$
At each time step $t = 1, \dots, T$
- the player chooses an action $k_t \in \mathcal{X} = \{1, \dots, K\}$;
- given k_t, the environment draws the reward $X_{k_t,t} \sim \nu_{k_t}$;
- the player only observes the feedback $X_{k_t,t}$.

The player wants to minimize its pseudo-regret defined as

$$\bar{R}_T \overset{\text{def}}{=} T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{k_t}\right],$$

where we recall $\mu_k = \mathbb{E}[X_{k,t}]$ and $\mu^* = \max_{k=1,\dots,K} \mu_k$.

Setting 1: Setting of stochastic bandit with finitely many actions

During last lecture, we considered the finite-armed bandit setting and saw several algorithms (ETC, UCB, ε-greedy, Thomson sampling) that achieve sublinear pseudo regret. UCB achieves for instance

$$\bar{R}_T \lesssim \min\left\{ \sum_{k:\Delta_k>0}^{K} \frac{\log T}{\Delta_k}\,,\ \sqrt{TK\log T} \right\},$$

where $\Delta_k \overset{\text{def}}{=} \mu^* - \mu_k$ is the suboptimality gap of arm k. The first bound is distribution dependent (it depends on the gaps Δ_k) and is of order $O(\log T)$ while the second bound is distribution free but is of order $O(\sqrt{T})$.

In this lecture, we will consider the more practical setting of contextual bandits, in which the learner observes a context $c_t \in C$ before choosing the action k_t.

1 Contextual bandits

In most applications, before choosing an action k_t the player observes some context $c_t \in \mathcal{C}$.

For instance, consider a bandit problem in which the player needs to display ads on his website. At each new visitor, the player chooses an add to display and observes if the visitor click on it. The reward is one if there is a click and 0 otherwise. In this case, the player can see the cookie of the visitor before choosing the ad. A first step towards contextual bandits, is to consider continuous sets of actions $\mathcal{X}$, which may correspond to mapping between context and arms.

1.1 Continuous stochastic bandits

Let first generalize the finite-armed bandit setting to continuous set of arms in Setting 2.

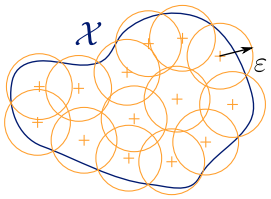
> *Unknown parameters:* $\nu(\theta)$, for each $\theta \in [0,1]^d$, a probability distribution on $[0,1]$ with expectation $\mu(\theta) \in [0,1]$.
>
> At each time step $t = 1, \dots, T$
> - the player chooses an action $\theta_t \in \Theta \subseteq [0,1]^d$;
> - given θ_t, the environment draws the reward $Y_t \sim \nu(\theta_t)$ independently from the past;
> - the player only observes the feedback Y_t.
>
> The player wants to minimize its pseudo-regret defined as
>
> $$\bar{R}_T \overset{\text{def}}{=} T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} \mu(\theta_t)\right],$$
>
> where $\mu^* = \sup_{\theta\in\Theta} \mu(\theta)$.

Setting 2: Setting of stochastic bandit with continuous set of actions

Similarly to what we did in the full-information setting with EWA, if the expectation function μ is β-Hölder: i.e., there exists $c > 0$

$$\forall \theta, \theta' \in \mathcal{X} \qquad \left|\mu(\theta) - \mu(\theta')\right| \le c\left\|\theta - \theta'\right\|^{\beta},$$

then we may discretize the action space Θ and run any discrete bandit algorithm (UCB, ε-greedy,...).



Theorem 1. *Let $\beta > 0$ and $\varepsilon > 0$. Assume that μ is β-Hölder. If UCB is run on an ε-covering of minimal cardinal of $\Theta \subset [0,1]^d$, then it satisfies*

$$\bar{R}_T \lesssim T\varepsilon^{\beta} + \sqrt{\frac{T\log(T)}{\varepsilon^d}}.$$

In particular for $\varepsilon \approx \left(\frac{\log T}{T}\right)^{\frac{1}{2\beta+d}}$, we have $\bar{R}_T \lesssim T\left(\frac{\log T}{T}\right)^{\frac{\beta}{2\beta+d}}$.

Proof. An optimal ε-covering of $[0,1]^d$ has cardinal of order $K \approx \varepsilon^{-d}$. Let $x^* \in \arg\max_{\theta\in\Theta} \mu(\theta)$ (we assume that it exists) and $\tilde{\theta}^*$ its ε-approximation, then the distribution-free upper-bound of UCB yields

$$\bar{R}_T \lesssim T\big(\mu(\theta^*) - \mu(\tilde{\theta}^*)\big) + \sqrt{KT\log T} \approx cT\varepsilon^{\beta} + \sqrt{\varepsilon^{-d}T\log T}.$$

The second part of the theorem is obtained by obtimizing ε. □

To build the discretization, both β and T need to be known in advance. The horizon T can be calibrated online through a "doubling trick" (left as exercise). The parameter β may be tuned through bandit with experts (or bandits where arms are bandit algorithms) that we may see in next lecture (see Exp4 algorithm).

Note that the per-round complexity of such an algorithm is of order $\varepsilon^{-d} \approx T^{\frac{d}{2\beta+d}}$. Quite surprinsingly it does not explodes with the dimension d and is always smaller than T. This is due to the fact that the higher the dimension d is, the worse will be the regret bound, and the cruder needs the discretization to be.

1.2 Contextual bandits through discretization

No we consider the following contextual bandit setting in which the player has a finite decision set $\Theta = \{1, \dots, K\}$ but observes a context $x_t \in \mathcal{X}$ before choosing his action.

Unknown parameters: $\nu(k,x)$, for each arm $k \in \{1,\dots,K\}$ and context $x \in \mathcal{X}$, a probability distribution on $[0,1]$ with expectation $\mu(k,x) \in [0,1]$.

At each time step $t = 1,\dots,T$
- the environment chooses $x_t \in \mathcal{X}$ and reveals it to the player;
- the player chooses an action $k_t \in \{1,\dots,K\}$;
- given k_t, the environment draws the reward $Y_t \sim \nu(k_t, x_t)$ independently from the past;
- the player only observes the feedback Y_t.

The player wants to minimize its pseudo-regret defined as

$$\bar{R}_T \overset{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^{T} \mu^*(x_t) - \sum_{t=1}^{T} Y_t\right],$$

where $\mu(k,x) = \mathbb{E}_{Y\sim\nu(k,x)}[Y]$ and $\mu^*(x) = \max_{k=1,\dots,K} \mu(k,x)$.

Setting 3: Setting of contextual stochastic bandit

Finite set of contexts If the set of context is finite $\mathcal{X} \overset{\text{def}}{=} \{1,\dots,|\mathcal{X}|\}$ we can denote

$$\bar{R}_T(c) \overset{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^{T} \big(\mu^*(x) - \mu(k_t,x)\big)\mathbb{1}_{x_t=x}\right]$$

the pseudo-regret due to context $x \in \mathcal{X}$. Then applying a separate instance of UCB (or any bandit algorithm) for each context $x \in \mathcal{X}$, we get by using the distribution-free upper-bound of UCB

$$\bar{R}_T(x) \lesssim \sqrt{T_x K \log T_x}, \qquad \text{where} \qquad T_x \overset{\text{def}}{=} \sum_{t=1}^{T} \mathbb{1}_{x_t=x}\,.$$

Note that because T_x are not known in advance it is important that the bound of UCB is anytime (i.e., that UCB does not need to know the horizon). The total pseudo-regret of UCB is then obtained by summing over all contexts

$$\bar{R}_T = \sum_{x\in\mathcal{X}} \bar{R}_T(x) \lesssim \sum_{x\in\mathcal{X}} \sqrt{T_x K \log T} \le \sqrt{|\mathcal{X}| T K \log T}\,,$$

where the last inequality is by Jensen's inequality using the concavity of the square root and $\sum_{x\in\mathcal{X}} T_x = T$.

Continuous set of contexts If the set of context is continuous $\mathcal{X} \subset [0,1]^d$, one needs again to make assumption on the distributions $\nu(k,x)$ which needs to vary smoothly in x. Doing so, one may discretize the set of context with an ε-covering of $\mathcal{X}$ of size $N \approx \varepsilon^{-d}$ and run an independent instance of UCB in each of the N bins.

Theorem 2. *Let $\beta > 0$ and $\varepsilon > 0$. Assume that $x \mapsto \mu(k,x)$ is β-Hölder for all $k \in \mathcal{X}$. If UCB is independently run in each bin of an optimal ε-covering of $\mathcal{X}$, then*

$$\bar{R}_T \lesssim T\varepsilon^{\beta} + \sqrt{\frac{KT\log T}{\varepsilon^d}}\,.$$

In particular for ε well-optimized, we have $\bar{R}_T \lesssim T\Big(\frac{K\log T}{T}\Big)^{\frac{\beta}{2\beta+d}}$.

Remark that in all these regret bounds, the suboptimal $\log T$ term can be removed by using MOSS (a minimax optimal variant of UCB).

Better rates using distribution-dependent bound? In the above results, we used the distribution-free regret bound of UCB. Because, if the function $\mu(\cdot, x)$ varies smoothly with x, there should be some context with zero suboptimality gaps. Yet, it is possible to get better rates by assuming the following α-margin assumption. It controls the suboptimality gap with high probability: the contexts x_t are i.i.d. and sastify for all $\delta \in (0, 1)$

$$\mathbb{P}\Big\{ \min_{k:\Delta(k,x_t)>0} \Delta(k, x_t) < \delta \Big\} \leq \square \delta^\alpha \tag{1}$$

where $\Delta(k, x_t) \overset{\text{def}}{=} \mu^*(x) - \mu(k, x)$ and $\square$ is some constant. Note that the larger the value of α is the easier is the problem.

Theorem 3 (Theorem 4.1, Perchet and Rigollet [2013])**.** *Let $\alpha \in (0, 1)$, $\beta > 0$ and $\varepsilon > 0$. Assume that $c \mapsto \mu(k, x)$ is β-Hölder for all $k \in \mathcal{X}$ and that the α-margin assumption* (1) *holds. Running a bandit algorithm (similar to UCB) independently in each bin of an optimal ε-covering of $\mathcal{X}$, we get*

$$\bar{R}_T \lesssim T \left(\frac{K \log K}{T} \right)^{\frac{\beta(\alpha+1)}{2\beta+d}} ,$$

for optimized ε.

The proof (for another algorithm then UCB) may be found in Perchet and Rigollet [2013]. We see that the factor α improves the rate of convergence with respect to the previous rate.

1.3 Stochastic Linear bandits

Contextual bandits that we just saw generalizes multi-armed bandits by allowing contexts. However, the dimension of the context space significantly worsen the regret rate from $\sqrt{T}$ to $T^{\frac{d+1}{d+2}}$ for Lipschitz rewards for instance (β-Hölder with $\beta = 1$). In this part, we will see *Stochastic linear bandits*, in which we assume the rewards to have a linear structure. This includes rich classes of models and allows better regret of order $O(\sqrt{T})$.

Unknown parameter: $\mu^* \in \mathbb{R}^d$.

At each time step $t = 1, \dots, T$

- the environment chooses $\Theta_t \subseteq \mathbb{R}^d$ the decision set;
- the player chooses an action $\theta_t \in \Theta_t$;
- given θ_t, the environment draws the reward

$$Y_t = \theta_t \cdot \mu^* + \varepsilon_t$$

  where ε_t is i.i.d. 1-subgaussian noise.
- the player only observes the feedback Y_t.

The player wants to minimize its pseudo-regret defined as

$$\bar{R}_T \overset{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^{T} \max_{\theta \in \Theta_t} \theta \cdot \mu^* - \sum_{t=1}^{T} Y_t \right] .$$

Setting 4: Setting of stochastic linear bandit

The setting of stochastic linear bandits is described in Setting 4. For simplicity, the noise ε_t is assumed to be i.i.d. and 1-subgaussian noise: i.e., $\mathbb{E}[\varepsilon_t] = 0$ and

$$\forall \lambda > 0, \qquad \mathbb{E}\big[\exp(\lambda \varepsilon_t) \big] \leq \exp(\lambda^2/2)$$

almost surely. Note that we could consider σ^2-subgaussian noise, or make it depend on the past $\mathcal{F}_t = \sigma(x_1, \varepsilon_1, \dots, x_t, \varepsilon_t)$ with $\mathbb{E}[\varepsilon_t | \mathcal{F}_t] = 0$.

Particular cases: why is this setting interesting? Different choices of decision sets $\mathcal{X}_t$ lead to different settings of stochastic bandits:

- *Finite-armed bandit*: if $\Theta_t = (e_1, \dots, e_d)$ where e_i are the unit vectors in $\mathbb{R}^d$ and $\mu^* = (\mu_1, \dots, \mu_d)$, we recover the setting of finite-armed bandit.
- *Contextual linear bandit*: we can recover a particular case of Setting 3, if $x_t \in \mathcal{X}$ is a context observed by the player and the reward function μ is of the form

  $$\mu(\theta, x) = \psi(\theta, x) \cdot \mu^*, \qquad \forall (\theta, x) \in [K] \times \mathcal{X}\,,$$

  for some unknown parameter $\mu^* \in \mathbb{R}^d$ and *feature map* $\psi : [K] \times \mathcal{X} \to \mathbb{R}^d$. For example, assume that you are a website which wants to display ads to visitors. The context x_t can be the cookie of the visitor containing information about what he likes, the actions are the possible ads to be displayed and the reward tells if there is a click. If the possible interests of the visitor are grouped in finite categories (such as traveling), so are the ads (in groups of products, such as flight tickets), the feature maps could contained all the combinations between interests and groups of products. The unknown vector θ^* would be tell which interests and groups of products are positively correlated. Of course the feature map could be created using any methods (such as deep-learning or splines).
- *Combinatorial bandit*: if $\Theta_t \subseteq \{0, 1\}^d$ yields to combinatorial bandit problems. For instance, the decision set corresponds to possible paths in a graph, the vector μ^* assigns to each edge a reward corresponding to its cost and the goal is to find the smallest path with smallest cost.

Algorithm: LinUCB As we saw earlier with UCB, the "optimism principle" is a good option for bandit problems to explore. The LinUCB algorithm is based on the same principle:

1. Build confidence set that contain μ^*: $\mu^* \in C_t$ with high probability
2. Build confidence upper-bound on the rewards: for all $\theta \in \Theta_t$

   $$\mathrm{UCB}_t(\theta) = \max_{\mu \in C_t} \ \theta \cdot \mu \tag{2}$$

3. Be optimistic: act as if the best possible rewards were the true rewards

   $$\theta_t \in \operatorname*{arg\,max}_{\theta \in \Theta_t} \mathrm{UCB_t}(\theta)\,. \tag{3}$$

Therefore the only remaining question is how to build the confidence set $C_t \subseteq \mathbb{R}^d$? They should contain μ^* with high probability but be as small as possible. Given the observed rewards the key is thus to estimate the parameter μ^*. Denoting by I_d the $d \times d$ identity matrix and picking $\lambda > 0$, we can estimate μ^* with *regularized least square*

$$\widehat{\mu}_t \overset{\text{def}}{=} \operatorname*{arg\,min}_{\mu \in \mathbb{R}^d} \left\{ \sum_{s=1}^{t} (Y_s - \theta_s \cdot \mu)^2 + \lambda \|\mu\|^2 \right\} = V_t^{-1} \sum_{s=1}^{t} \theta_s Y_s\,,$$

where $V_t \overset{\text{def}}{=} \lambda I_d + \sum_{s=1}^{t} \theta_s \theta_s^\top$. We have the following result whose proof can be found in Lattimore and Szepesvári [2020].

Lemma 1 (Theorem 20.2, **?**)**.** *Let $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$, if $\max_{\theta \in \Theta_t} \|\theta\|_2 \le 1$, for all $t \ge 1$*

$$\left\|\widehat{\mu}_t - \mu^*\right\|_{V_t} \le \sqrt{\lambda}\|\mu^*\| + \sqrt{2\log(1/\delta) + d\log\left(1 + \frac{T}{\lambda}\right)} \overset{def}{=} \beta(\delta)\,,$$

where $\|\mu\|_{V_t}^2 = \mu^\top V_t \mu$.

The above lemma, states that with probability $1-\delta$, for all $t \geq 1$,

$$\mu^* \in C_t, \quad \text{where} \quad C_t \stackrel{\text{def}}{=} \left\{ \theta \in \mathbb{R}^d : \left\| \mu - \widehat{\mu}_{t-1} \right\|_{V_{t-1}} \leq \beta(\delta/T) \right\}. \tag{4}$$

Proof of Lemma 1. The proof relies on Laplace's method on super-martingales which is a standard argument to provide confidence bounds on a self-normalized sum of conditionally centered random vectors. We have

$$\widehat{\mu}_t = V_t^{-1} \sum_{s=1}^{t} \theta_s Y_s = V_t^{-1} \sum_{s=1}^{t} \theta_s (\theta_s^\top \mu^* + \varepsilon_s) = V_t^{-1}(V_t - \lambda I_d)\mu^* + M_t) = \mu^* - \lambda V_t^{-1}\mu^* + V_t^{-1} M_t \,,$$

where we introduced $M_t = \sum_{s=1}^{t} \theta_s \varepsilon_s$, which is a martingale with respect to $\mathcal{F}_t = \sigma(\varepsilon_1, \ldots, \varepsilon_t)$. Therefore, by triangle inequality

$$\left\| V_t^{1/2}(\widehat{\mu}_t - \mu^*) \right\| = \left\| -\lambda V_t^{-1/2}\mu^* + V_t^{-1/2} M_t \right\| \leq \lambda \left\| V_t^{-1/2}\mu^* \right\| + \left\| V_t^{-1/2} M_t \right\| .$$

On the one hand, given that all eigenvalues of the symmetric matrix V_t are larger than λ, all eigenvalues of $V_t^{-1/2}$ are smaller than $1/\sqrt{\lambda}$ and thus

$$\lambda \left\| V_t^{-1/2}\mu^* \right\| \leq \lambda \frac{1}{\sqrt{\lambda}} \|\mu^*\| = \sqrt{\lambda} \|\mu^*\| .$$

We now prove, on the other hand, that with probability at least $1-\delta$

$$\left\| V_t^{-1/2} M_t \right\| \leq \sqrt{2 \log \frac{1}{\delta} + d \log \frac{1}{\lambda} + \log \det(V_t)} .$$

Upper-bounding $\log \det(V_t) \leq d \log(\lambda + t)$ (since all the eigenvalues of V_t are smaller than $\lambda + t$) will then conclude the proof of the Theorem.

Step 1: Introducing super-martingales. For all $\nu \in \mathbb{R}^d$, we consider

$$S_{t,\nu} = \exp\left(\nu^\top M_t - \frac{1}{2} \nu^\top V_t \nu \right)$$

and now show that it is an $\mathcal{F}_t$-super-martingale. First, note that since the common distribution of the $\varepsilon_1, \ldots, \varepsilon_t$ is 1-sub-Gaussian, the for all $\mathcal{F}_{t-1}$-measurable random variable ν_{t-1}

$$\mathbb{E}\left[e^{\nu_{t-1}\varepsilon_t} \middle| \mathcal{F}_{t-1} \right] \leq e^{\frac{\nu_{t-1}^2}{2}} .$$

Now,

$$\mathbb{E}\left[S_{t,\nu} \middle| \mathcal{F}_{t-1} \right] = S_{t-1,\nu} \mathbb{E}\left[\exp\left(\nu^\top \theta_t \varepsilon_t - \frac{1}{2} \nu^\top \theta_t \theta_t^\top \nu \right) \middle| \mathcal{F}_{t-1} \right] \leq S_{t-1,\nu} .$$

Note that rewriting $S_{t,\nu}$ in its vertex form is, with $m = V_t^{-1} M_t$:

$$S_{t,\nu} = \exp\left(\frac{1}{2} (\nu - m)^\top V_t (\nu - m) \right) \times \exp\left(\frac{1}{2} \left\| V_t^{-1/2} M_t \right\|^2 \right) .$$

Step 2: Laplace's method–integrating $S_{t,\nu}$ over $\nu \in \mathbb{R}^d$. The basic observation behind this method is that (given the vertex form) $S_{t,\nu}$ is maximal at $\nu = m = V_t^{-1} M_t$ and then equals $\exp\left(\frac{1}{2} \left\| V_t^{-1/2} M_t \right\|^2 \right)$, which is (a transformation of) the quantity to control. Now, because the exp function quickly vanishes, the integral over $\nu \in \mathbb{R}^d$ is close to its maximum. We therefore consider

$$\bar{S}_t = \int_{\mathbb{R}^d} S_{t,\nu} d\nu .$$

We will make repeated uses of the fact that the Gaussian density function

$$\nu \mapsto \frac{1}{\sqrt{\det(2\pi C)}} \exp\Big((\nu-m)^\top C^{-1}(\nu-m)\Big)\,,$$

where $m \in \mathbb{R}^d$ and C is a symmetric positive definite matrix, integrate to 1 over $\mathbb{R}^d$. This gives us the first rewriting

$$\bar{S}_t = \sqrt{\det(2\pi V_t^{-1})}\exp\Big(\frac{1}{2}\big\|V_t^{-1/2}M_t\big\|^2\Big)\,.$$

Second, by the Fubini-Tonelli theorem and the super-martingale property

$$\mathbb{E}\big[S_{t,\nu}\big] \le \mathbb{E}\big[S_{0,\nu}\big] = \exp(-\lambda\|\nu\|^2/2)\,,$$

we also have

$$\mathbb{E}\big[\bar{S}_t\big] \le \int_{\mathbb{R}^d}\exp(-\lambda\|\nu\|^2/2)d\nu = \sqrt{\det(2\pi\lambda^{-1}I_d)}\,.$$

Combining the two statements, we proved

$$\mathbb{E}\Big[\exp\Big(\frac{1}{2}\big\|V_t^{-1/2}M_t\big\|^2\Big)\Big] \le \sqrt{\frac{\det(V_t)}{\lambda^d}}\,.$$

Step 3: Markov-Chernov bound. For $u > 0$,

$$\begin{aligned}\mathbb{P}\Big(\big\|V_t{-}1/2M_t\big\| > u\Big) &= \mathbb{P}\Big(\frac{\big\|V_t{-}1/2M_t\big\|^2}{2} > \frac{u^2}{2}\Big)\\ &\le \exp\big(-\frac{1}{2}u^2\big)\mathbb{E}\Big[\exp\Big(\frac{1}{2}\big\|V_t^{-1/2}M_t\big\|^2\Big)\Big] \le \exp\Big(-\frac{u^2}{2}+\frac{1}{2}\log\frac{\det(V_t)}{\lambda^d}\Big) = \delta\,,\end{aligned}$$

for the claimed choice

$$u = \sqrt{2\log\frac{1}{\delta} + d\log\frac{1}{\lambda} + \log\det(V_t)}\,.$$

□

Theorem 4 (Corollary 19.2, **?**)**.** *Let $T \ge 1$ and $\mu^* \in \mathbb{R}^d$. Assume that for all $\theta \in \cup_{t=1}^T \Theta_t$, $|\theta\cdot\mu^*| \le 1$, $\|\mu^*\| \le 1$ and $\|\theta_t\| \le 1$, then LinUCB with C_t defined in* (4) *satisfies the pseudo-regret bound*

$$\bar{R}_T \le \square_\lambda d\sqrt{T}\log(T)\,,$$

where $\square_\lambda$ is a constant that may depend on λ.

Proof. Let $\delta = 1/T$. By Lemma 1, with probability $1-1/T$,

$$\forall t \ge 1, \qquad \mu^* \in C_t\,. \tag{5}$$

Step 1: Small instantaneous regrets under the event (5). Assume that (5) holds. Let

$$\theta_t^* \overset{\text{def}}{=} \max_{\theta\in\Theta_t}\theta\cdot\mu^* \qquad \text{and} \qquad r_t \overset{\text{def}}{=} (\theta_t^*-\theta_t)\cdot\mu^*$$

be respectively the optimal decision and the instantaneous regret at round t. We also define

$$\tilde{\mu}_t \in \operatorname*{arg\,max}_{\mu\in C_t}\big\{\theta_t\cdot\mu\big\}.$$

Since $\mu^* \in C_t$, we have

$$\theta_t^*\cdot\mu^* \;\le\; \max_{\mu\in C_t}\big\{\theta_t^*\cdot\mu\big\} \overset{(2)}{=} UCB_t(\theta_t^*) \overset{(3)}{\le} UCB_t(\theta_t) \;= \max_{\mu\in C_t}\big\{\theta_t\cdot\mu\big\} = \theta_t\cdot\tilde{\mu}_t\,,$$

which entails because μ^* and $\tilde{\mu}_t$ belong to C_t,

$$r_t \overset{\text{def}}{=} (\theta_t^* - \theta_t) \cdot \mu^* \leq \theta_t \cdot (\tilde{\mu}_t - \mu^*) \overset{\text{Cauchy-Schwarz}}{\leq} \|\theta_t\|_{V_{t-1}^{-1}} \left\|\tilde{\mu}_t - \mu^*\right\|_{V_{t-1}} \leq 2\|\theta_t\|_{V_{t-1}^{-1}} \beta(1/T^2)\,.$$

Therefore, summing over $t = 1, \ldots, T$ and using $r_t \leq 2$, we have

$$\begin{aligned}
R_T \overset{\text{def}}{=} \sum_{t=1}^{T} r_t &\leq \sqrt{T \sum_{t=1}^{T} r_t^2} && \leftarrow \text{ Jensen's inequality} \\
&\leq 2\sqrt{T \sum_{t=1}^{T} \min\left\{\|\theta_t\|_{V_{t-1}^{-1}}^2 \beta(1/T^2)^2, 1\right\}} \\
&\leq 2\beta(1/T^2)\sqrt{T \sum_{t=1}^{T} \min\left\{\|\theta_t\|_{V_{t-1}^{-1}}^2, 1\right\}} && \leftarrow \ \beta_T(1/T^2) \geq 1 \\
&\leq 2\beta(1/T^2)\sqrt{T \sum_{t=1}^{T} \log\left(1 + \|\theta_t\|_{V_{t-1}^{-1}}^2\right)} && \leftarrow \ \min\{u, 1\} \leq 2\log(1+u)\,.
\end{aligned}$$

But, we have

$$\begin{aligned}
1 + \|\theta_t\|_{V_{t-1}^{-1}}^2 &= \det\left(1 + \|\theta_t\|_{V_{t-1}^{-1}}^2\right) \\
&= \det\left(V_{t-1}^{-1}\left(V_{t-1} + V_{t-1}^{1/2}\|\theta_t\|_{V_{t-1}^{-1}}^2 V_{t-1}^{1/2}\right)\right) && \leftarrow \text{using} \det(I + AB) = \det(I + BA) \\
&= \det\left(V_{t-1}^{-1}\left(V_{t-1} + \theta_t\theta_t^\top\right)\right) && \leftarrow \|\theta_t\|_{V_{t-1}^{-1}} = V_{t-1}^{-1/2}\theta_t\theta_t^\top V_{t-1}^{-1/2} \\
&= \det\left(V_{t-1}^{-1} V_t\right) && \leftarrow V_t = V_{t-1} + \theta_t\theta_t^\top \\
&= \frac{\det(V_t)}{\det(V_{t-1})} && \leftarrow \det(A^{-1}B) = \frac{\det(B)}{\det(A)}\,.
\end{aligned}$$

Substituting into the regret bound, the sum telescopes and it entails

$$R_T \leq 2\beta(1/T^2)\sqrt{T \log\left(\frac{\det(V_T)}{\det(V_0)}\right)}\,.$$

Then, using $V_0 \overset{\text{def}}{=} \lambda I_d$ and since $V_T = \lambda I_d + \sum_{t=1}^{T} \theta_t\theta_t^\top$ with $\|\theta_t\| \leq 1$, all eigenvalues of V_T lie in $[\lambda, \lambda + T]$ which yields

$$\det(V_0) = \lambda^d \qquad \text{and} \qquad \det(V_T) \leq \left(\lambda + T\right)^d\,.$$

Plugging back into the previous upper-bound and using that $\beta(1/T^2) \leq \square_\lambda \sqrt{d \log T}$

$$R_T \leq 2\sqrt{dT\beta(1/T)\log\left(1 + \frac{T}{\lambda}\right)} \leq \square_\lambda d\sqrt{T}\log T\,.$$

Part 2: without the event (4) We because $r_t \leq 2$, almost surely $R_T \leq 2T$, and we have

$$\begin{aligned}
\bar{R}_T = \ \mathbb{E}\left[R_T\right] &\leq \ \mathbb{E}\left[R_T \middle|\ \text{Event (4)}\right]\mathbb{P}\{\text{Event (4)}\} + 2T\big(1 - \mathbb{P}\{\ \text{Event (4)}\}\big) \\
&\leq \ \square d\sqrt{T}\log T + 2\,.
\end{aligned}$$

This concludes the proof. □

Better regret with assumptions It is worth pointing out that if we make additional assumptions, it is possible to improve the regret bound $O(d\sqrt{T}\log T)$. A first setting corresponds to the case where the set of available actions at time t is fixed and finite; i.e., the learner needs to choose $\theta_t \in \Theta$ where $|\Theta| = K$. Then, it is possible to achieve

$$R_T \leq \square\sqrt{Td\log(TK)}\,,$$

which improves the previous bound by a factor $\sqrt{d}/\log(K)$ and improves the classical bound of UCB $O(\sqrt{TK\log T})$ by a factor $K/\sqrt{d}$. These improvements can be significant when $K \gg d \gg 1$. We refer the curious reader to [Lattimore and Szepesvári, 2020, Chapter 22].

Another possible improvement when $d \gg 1$ is to assume that μ^* is m_0-sparse (i.e., most of its components are zero). Then under assumptions, one can get a regret of order $\tilde{O}(\sqrt{dm_0T})$.

1.4 Other possible extensions of bandits

Note that there exist many different extensions of stochastic bandits to make it more realistic or with improved regret.

- *Bandit with delays:* For instance, consider the example of the website which wants to display ads. The website does not observe if there is no click, he needs to fix some time after which he consider that the visitor will not click, and if the visitor stays long on the webpage, the website may need to display ads to other visitors before getting the rewards. There is thus delayed feedback the website needs to deal with.
- *Non stationarity*
- *Combinatorial bandits*
- *Dueling bandits*
- ...

We refer the interested student to the monograph Lattimore and Szepesvári [2020] for more information on these settings. Next week, we will deal with adversarial bandits.

References

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, pages 693–721, 2013.