Lecture #4 : Continuous and linear bandits

In many applications, the agent can observe a context c_t first (eg user information in the case of online recommendation)

Bandits with continuum of arms

Before that, let us consider the problem of continuous bandits

Setting (continuous bandits)

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

- the agent pulls an arm $a \in \mathcal{A} \subset [0,1]^d$
- the agent observes and receives the reward $Y_t = \mu(a_t) + \eta_t$ where η_t is 1 sub-Gaussian 0 mean independent noise

Goal: minimise regret $R_T = T\mu^* - \sum_{t=1}^{T} \mu(a_t)$ where $\mu^* = \sup_{a \in \mathcal{A}} \mu(a)$.

Without any assumption on μ or $\mathcal{A}$, we cannot do anything.
Otherwise, for $\mathcal{A} = [0,1]$ and any algorithm, we can choose x s.t.

$$\forall t \in \mathbb{N}, \quad \mathbb{P}_{\mu=0}(a_t = x) = 0$$

Then for $\mu(a) = \mathbb{1}_{a=x}$, the same algorithm would behave as if $\mu = 0$ and never pull x,

so that its regret is $R_T = T$.

Continuity is actually enough to control the regret (ie having $\mathbb{E}[R_T] = o(T)$).

To get precise bounds, we will consider a stronger assumption on μ.

μ is β-Hölder if there exists $c>0$ s.t $\forall a, a' \in \mathcal{A}$, $|\mu(a) - \mu(a')| \leq c\|a-a'\|^\beta$.

(meta) **Algo: Binning UCB**

Input: $\varepsilon > 0$

Let K be an ε-covering of minimal cardinal of $\mathcal{A}$.

Pull each arm in K once

Then for any $t \geq |K|$:

$$\text{pull } a_t \in \operatorname*{argmax}_{k \in K} \hat{\mu}_k(t-1) + 2\sqrt{\frac{2\ln(t)}{N_k(t-1)}}$$

run UCB on the set of arms K

(different constant because we have 1-sub-Gaussian rewards)

Theorem

Let $\beta > 0$ and $\varepsilon > 0$. Assume that μ is β-Hölder and $\mathcal{A} \subset [0,1]^d$.

Then running binning UCB with parameter ε yields the regret:

$$\mathbb{E}[R_T] \leq c\left(T\varepsilon^\beta + \sqrt{\frac{T\ln T}{\varepsilon^d}}\right)$$

for some universal constant $c>0$.

Taking ε of order $\left(\frac{\ln T}{T}\right)^{\frac{1}{2\beta+d}}$ yields a regret $\mathbb{E}[R_T] = O\left(T^{\frac{\beta+d}{2\beta+d}} \ln(T)^{\frac{\beta}{2\beta+d}}\right)$

Proof:

$$R_T = \sum_{t=1}^{T} \mu^* - \mu_{a_t}$$

$$= \sum_{t=1}^{T}\left(\mu^* - \max_{a \in K} \mu(a)\right) + \sum_{t=1}^{T} \max_{a \in K} \mu(a) - \mu_{a_t}.$$

$\underbrace{\leqslant c\,\varepsilon^{\beta}}$ by Hölder assumption $\qquad \underbrace{\text{regret of UCB run on } \mathcal{K}}$.

so for $K = \mathrm{card}(\mathcal{K})$, $\mathbb{E}[R_T] \leqslant cT\varepsilon^{\beta} + \underbrace{c\sqrt{KT\ln(T)}}_{\text{distribution free bound of UCB}}$

By minimisation of cardinal, a classical covering number bound yields $K \lesssim \varepsilon^{-d}$, which concludes $\square$

Contextual bandits

Motivation

Setting 1 (contextual bandits)

For each round $t = 1, \dots, T$:

- agent observes context $C_t \in \mathcal{C}$ (arbitrarily chosen by nature)
- agent chooses action $a_t \in [K]$, depending on C_t and past observations.
- agent observes and gets reward Y_t

  where $Y_t = r(a_t, C_t) + \eta_t$

  with η_t 1 sub-Gaussian $\quad \forall \lambda \in \mathbb{R},\ \mathbb{E}[e^{\lambda \eta_t}] \leqslant \exp\left(\frac{\lambda^2}{2}\right)$

  0 mean

  independent noise

$r : [K] \times \mathcal{C} \to \mathbb{R}$ is called the reward function $\leftarrow$ object to estimate

(pseudo)-regret defined as: $R_T = \sum_{t=1}^{T} \left\{ \max_{k \in [K]} r(k, C_t) - r(a_t, C_t) \right\}$

Without any assumption on r, independent bandit games for each context c.

- First possibility, r is "regular" (e.g. Lipschitz or Hölder)

In that case, we can again run a binning version of UCB to discretise the context set (instead of action set)

Binning UCB (contextual)

Input: $\varepsilon > 0$

Let $\mathcal{X}$ be an ε-covering of minimal cardinal of $\mathcal{C}$

For each time step $t \geq 1$:

- observe the context c_t.
- let $B \in \mathcal{X}$ a ball containing c_t : $c_t \in \mathcal{X}$
- pull the arm a_t following the UCB algorithm on the bin B, ie with

$$\begin{cases} N_k^B(t) = \sum_{s=1}^{t} \mathbb{1}_{a_s = k} \mathbb{1}_{c_s \in B} \\ \hat{\mu}_k^B(t) = \frac{1}{N_k^B(t)} \sum_{s=1}^{t} \mathbb{1}_{a_s = k} \mathbb{1}_{c_s \in B} Y_s. \end{cases}$$

Theorem:

Let $\beta > 0$ and $\varepsilon > 0$. Assume that $x \longmapsto \mu(k, x)$ is β-Hölder for any $k \in [K]$ and $\mathcal{C} \subset [0,1]^d$

The regret of (contextual) binning UCB is then bounded as:

$$\mathbb{E}[R_T] \leq c \left(T\varepsilon^\beta + \sqrt{\frac{KT \ln T}{\varepsilon^d}} \right)$$

Choosing ε of order $\left(\frac{K \log T}{T}\right)^{\frac{1}{2\beta+d}}$, we get $\mathbb{E}[R_T] = O\left(T^{\frac{\beta+d}{2\beta+d}} (K \ln T)^{\beta/2\beta+d} \right)$

Proof is not as direct as the continuous bandits case: UCB does not act with iid variables here, but random variables of the type: $X_k(t) = \mu_k + \eta_k(t)$ with $\eta_k(t)$ 1-sub-Gaussian $|\mathbb{E}[\eta_k(t)]| \leq c\varepsilon^\beta$ (left as exercise)

Remarks:

algo with an optimal distribution free $\sqrt{KT}$ bound

- In all the previous algos, we can replace UCB by MOSS to get rid of the $\log T$ terms.

- Instance dependent bounds? In the above results, we used the distribution free bound of UCB. Can we get something better, depending on the regularity level of μ?

Yes, with the α-margin assumption: for iid contexts c_t and all $\delta \in (0,1)$

The larger the α, the easier the problem.

$$\mathbb{P}\left(\min_{k,\, \Delta(k, c_t) > 0} \Delta(k, c_t) < \delta \right) \leq c\, \delta^{\alpha},$$

where $\Delta(k, c) = \max_{k'} \mu(k', c) - \mu(k, c)$

Theorem

Let $\beta \geq 0$ and $\alpha \in (0,1)$. Assume that $x \longmapsto \mu(k, x)$ is β-Hölder for any $k \in [K]$, $\mathcal{C} \subset [0,1]^d$ and the α-margin assumption.

Then the regret of (contextual) binning UCB is then bounded as:

$$\mathbb{E}[R_T] \leq c\, T^{\frac{\beta(1-\alpha)+d}{2\beta+d}} (K \ln T)^{\frac{\beta(1+\alpha)}{2\beta+d}}$$

for an optimised $\varepsilon > 0$

the proof is intricate.

Linear bandits

Another possible assumption is that r is linear with respect to a known feature map $\Psi : [K] \times \mathcal{C} \to \mathbb{R}^d$ and a parameter $\theta^* \in \mathbb{R}^d$ such that
to estimate

$$r(k, c) = \langle \theta^*, \Psi(k, c) \rangle \qquad \forall k, c.$$

This is equivalent to the following setting, with $\mathcal{A}_t = \{ \Psi(k, c_t) \mid k \in [K] \}$:

Setting 2 (linear bandits)

For each round $t = 1, \dots, T$:

- agent observes decision set $\mathcal{A}_t \subset \mathbb{R}^d$
- agent chooses action $a_t \in \mathcal{A}_t$
- agent observes and gets reward Y_t

  where $Y_t = \langle \theta^*, a_t \rangle + \eta_t$

  with η_t 1 sub-Gaussian
  0 mean
  independent noise

same defn of regret

Particular cases:

- $\mathcal{A}_t = \{e_1, \dots, e_d\} \to$ classical multi-armed bandits with d arms and $\mu_k = \theta^*_k$.
- $\mathcal{A}_t \subset \{0, 1\}^d \to$ combinatorial bandits.

We want to build an adaptation of UCB for linear bandits, called

LinUCB.

main question (

- build confidence regions $\mathcal{C}_t$ such that $\theta^* \in \mathcal{C}_t$ with high probability
- build confidence bounds for the arm means $U_a(t) = \underbrace{\max_{\theta \in \mathcal{C}_t} \langle a, \theta \rangle}_{\text{UCB score of arm } a}$

- be optimistic: pull $a_t \in \arg\max_{a \in \mathcal{A}_t} U_a(t)$

Before the confidence set, what is the estimate of θ^*? (ie "empirical mean")

Regularised least-squares estimator:

$$\boxed{\hat{\theta}_t = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{s=1}^{t} \left(Y_s - \langle \theta, a_s \rangle\right)^2 + \lambda \|\theta\|_2^2}$$

linear regression

$\lambda \geq 0$ is the penalty function (or regularization parameter)

$\lambda > 0$ ensures uniqueness of the minimiser

We can indeed easily check that:

$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} a_s Y_s \qquad \text{where } V_t = \lambda I_d + \sum_{s=1}^{t} a_s a_s^T$$

For any symmetric, positive definite matrix $M \in \mathbb{R}^{d \times d}$ and vector $u \in \mathbb{R}^d$, we denote

$$\|u\|_M^2 := (u^T M u)$$

Theorem (linear bandits concentration)

For any $\delta \in (0,1)$, $t \in \mathbb{N}$ and $\lambda > 0$, if for all s, $\max_{a \in \mathcal{A}_s} \|a\|_2 \leq 1$, then with probability at least $1-\delta$,

$$\|\hat{\theta}_t - \theta^*\|_{V_t} \leq \sqrt{\lambda} \|\theta^*\|_2 + \sqrt{2 \ln\left(\frac{1}{\delta}\right) + d \ln\left(1 + \frac{t}{\lambda d}\right)}$$

The proof relies on the following concentration lemma

Lemma

Let $S_t = \sum_{s=1}^{t} Y_s a_s$

For any $\lambda > 0, t \in \mathbb{N}$ and $\delta \in (0,1)$,

$$\mathbb{P}\left(\|S_t\|^2_{V_t^{-1}} \geq 2\ln\left(\frac{1}{\delta}\right) + \ln\left(\frac{\det(V_t)}{\lambda^d}\right)\right) \leq \delta$$

Proof of the Theorem (based on lemma)

Note that $\hat{\theta}_t = V_t^{-1}\left(S_t + \sum_{s=1}^{t} a_s a_s^T \theta^*\right)$

$$= V_t^{-1} S_t + V_t^{-1}(V_t - \lambda I_d)\theta^*$$

So $\|\hat{\theta}_t - \theta^*\|_{V_t} = \|V_t^{-1} S_t - \lambda V_t^{-1}\theta^*\|_{V_t}$

$$\leq \|V_t^{-1} S_t\|_{V_t} + \lambda \|V_t^{-1}\theta^*\|_{V_t}$$

$$= \|S_t\|_{V_t^{-1}} + \lambda \underbrace{\|\theta^*\|_{V_t^{-1}}}_{\sqrt{\theta^{*T} V_t^{-1} \theta^*}}$$

$\sqrt{\theta^{*T} V_t^{-1} \theta^*} \leq \|V_t^{-1}\|_{op}^{1/2} \|\theta^*\|_2$

$\leq \lambda_{min}(V_t)^{-1/2} \|\theta^*\|_2$

$\leq \lambda^{-\frac{1}{2}} \|\theta^*\|_2$

$$\leq \|S_t\|_{V_t^{-1}} + \sqrt{\lambda}\|\theta^*\|_2.$$

It remains to show that $\frac{\det(V_t)}{\lambda^d} \leq \left(1 + \frac{T}{\lambda}\right)^d$, ie $\det(V_t) \leq (\lambda + T)^d$.

geom vs arithmetic mean comparison

Indeed, we have: $\det(V_t) \leq \left(\frac{\mathrm{tr}(V_t)}{d}\right)^d \leq \left(\lambda + \frac{1}{d}\sum_{s=1}^{t} \mathrm{tr}(a_s a_s^T)\right)^d \leq \left(\lambda + \frac{t}{d}\right)^d$

□

(optional)

Proof of the lemma

For any $x \in \mathbb{R}^d$, define $M_t(x) = \exp\left(\langle x, S_t\rangle - \frac{1}{2}\|x\|^2_{V_t - \lambda I_d}\right)$

1) We show by induction that $M_t(x)$ is a supermartingale, so that

$$\mathbb{E}[M_t(x)] \leq M_0(x) = 1$$

$t \to t+1$

$$V_{t+1} = V_t + a_{t+1} a_{t+1}^T$$

$$M_{t+1}(x) = \exp\left(\langle x, S_{t+1}\rangle - \frac{1}{2}\left(x^T(V_{t+1} - \lambda I)x\right)\right)$$

$$= M_t(x) \cdot \exp\left(\langle x, a_{t+1}\rangle \eta_{t+1} - \frac{1}{2}\langle x, a_{t+1}\rangle^2\right)$$

$$\mathbb{E}[M_{t+1}(x) | \mathcal{F}_t] \leq M_t(x)$$

(η_{t+1} is 1 sub-Gaussian)

2) Let $\nu = \mathcal{N}(0, \lambda^{-1} I_d)$

$$\bar{M}_t = \int M_t(x)\, d\nu(x)$$

is also a supermartingale by Tonelli and

$$\bar{M}_t = \frac{1}{\sqrt{(2\pi)^d \lambda^{-d}}} \int_{\mathbb{R}^d} \exp\Big(\underbrace{\langle x, S_t\rangle - \frac{1}{2}\|x\|^2_{V_t - \lambda I} - \frac{1}{2}\|x\|^2_{\lambda I}}\Big)\, dx$$

$$= x^T S_t - \frac{1}{2} x^T V_t x$$

$$= -\frac{1}{2}\left(x - V_t^{-1} S_t\right)^T V_t \left(x - V_t^{-1} S_t\right) + \frac{1}{2} S_t^T V_t^{-1} S_t$$

$$= -\frac{1}{2}\|x - V_t^{-1}S_t\|^2_{V_t} + \frac{1}{2}\|S_t\|^2_{V_t^{-1}}$$

$$\bar{M}_t = \exp\left(\frac{1}{2}\|S_t\|^2_{V_t^{-1}}\right) \cdot \left(\frac{\lambda}{2\pi}\right)^{d/2} \int_{\mathbb{R}^d} \underbrace{\exp\left(-\frac{1}{2}\|x - V_t^{-1}S_t\|^2_{V_t}\right)}_{\text{up to rescaling, pdf of } \mathcal{N}(V_t^{-1}S_t, V_t)} dx$$

$$= \exp\left(\frac{1}{2}\|S_t\|^2_{V_t^{-1}}\right) \frac{\lambda^{d/2}}{\sqrt{\det(V_t)}}$$

$$\|S_t\|^2_{V_t^{-1}} = 2\ln(\bar{M}_t) - \ln\left(\frac{\lambda^d}{\det(V_t)}\right)$$

3)

$$\mathbb{P}\left(\|S_t\|^2_{V_t^{-1}} \geq 2\ln\left(\frac{1}{\delta}\right) + \ln\left(\frac{\det(V_t)}{\lambda^d}\right)\right) = \mathbb{P}\left(\ln(\bar{M}_t) \geq \ln\left(\frac{1}{\delta}\right)\right)$$

$$= \mathbb{P}\left(\bar{M}_t \geq \frac{1}{\delta}\right) \leq \mathbb{E}[\bar{M}_t]\,\delta \leq \delta. \qquad \square$$

Algo LinUCB: suppose we know m with $\|\theta^*\|_2 \leq m$

For each $t \in \mathbb{N}$

Play $a_t \in \underset{a \in \mathcal{A}_t}{\arg\max} \; \underset{\theta \in \mathcal{C}_{t-1}}{\max} \langle \theta, a_t \rangle$

can be computed efficiently for our specific form of $\mathcal{C}_t$ and nice $\mathcal{A}_t$.

with $\hat{\theta}_t = \underset{\theta \in \mathbb{R}^d}{\text{argmin}} \sum_{s=1}^{t} \left(Y_s - \langle \theta, a_s \rangle\right)^2 + \lambda \|\theta\|_2^2$ $\quad V_t = \lambda I + \sum_{s=1}^{t} a_s a_s^T$

and $\mathcal{C}_t = \left\{ \theta \in \mathbb{R}^d \;\middle|\; \|\hat{\theta}_t - \theta\|_{V_t} \leq \sqrt{\lambda}\, m + \sqrt{4 \ln(t) + d \ln\left(1 + \frac{t}{d\lambda}\right)} \right\}$

Theorem:

If $\|\theta^*\|_2 \leq 1$ and for any t, $\max_{a \in \mathcal{A}_t} \|a\|_2 \leq 1$, then the regret of LinUCB satisfies for any $\lambda > 0$:

$$\mathbb{E}[R_T] \leq C_\lambda\, d \sqrt{T} \ln T$$

where c_λ is a constant that only depends on λ.

Comments:

- distribution free bound.
- if $\mathcal{A}_t$ is finite, and the same for every t, we can get a $\log(t)$ instance dependent bound: $\mathbb{E}[R_T] \leq c \sqrt{Td \ln(TK)}$

- another possible improvement when $d \gg 1$ is to assume that θ^* is m_0-sparse. Then, we can get a regret of order $\tilde{O}(\sqrt{d m_0 T})$

Proof:

Let us bound the instantaneous regret first.

$$r_t = \langle \theta^*, A_t^* - a_t \rangle \qquad \text{when } A_t^* \in \arg\max_{a \in \mathcal{A}_t} \langle \theta^*, a \rangle.$$

Define the good event

$$\mathcal{E}_t = \left\{ \theta^* \in \mathcal{C}_{t-1} \right\}$$

Thanks to our concentration theorem, $\mathbb{P}(\neg \mathcal{E}_t) \leq \frac{1}{(t-1)^2}$

$$\text{So } \mathbb{E}[r_t] \leq 2 \cdot \mathbb{P}(\neg \mathcal{E}_t) + \mathbb{E}[r_t \mathbb{1}_{\mathcal{E}_t}]$$

$$\leq \frac{2}{(t-1)^2} + \mathbb{E}[r_t \mathbb{1}_{\mathcal{E}_t}].$$

if $\mathcal{E}_t$, $\theta^* \in \mathcal{C}_{t-1}$ so:

$$\langle \theta^*, A_t^* \rangle \leq \max_{\theta \in \mathcal{C}_{t-1}} \langle \theta, A_t^* \rangle$$

$$\leq \max_{\theta \in \mathcal{C}_{t-1}} \langle \theta, a_t \rangle \qquad \text{by defn of } a_t.$$

$$= \langle \tilde{\theta}_t, a_t \rangle \text{ for some } \tilde{\theta}_t \in \mathcal{C}_{t-1}.$$

Cauchy Schwarz gives:

$$r_r = \langle \theta^*, A_r^* - a_r \rangle \leq \langle \tilde{\theta}_r - \theta^*, a_r \rangle \leq \|\tilde{\theta}_r - \theta^*\|_{V_{r-1}} \|a_r\|_{V_{r-1}^{-1}}$$

$$\leq \|a_r\|_{V_{r-1}^{-1}} \left(\|\tilde{\theta}_r - \hat{\theta}_{r-1}\|_{V_{r-1}} + \|\theta^* - \hat{\theta}_{r-1}\|_{V_{r-1}} \right)$$

$$\leq 2 \|a_r\|_{V_{r-1}^{-1}} \cdot \underbrace{\left(\sqrt{\lambda} + \sqrt{4 \ln(r_{-1}) + d \ln\left(1 + \frac{T}{\lambda d}\right)} \right)}_{}$$

define $\alpha_r = \max(\hat{\cdot}, 1)$

also by assumption, $r_r \leq 2$, so

$$r_r \leq 2 \alpha_r \left(1 \wedge \|a_r\|_{V_{r-1}^{-1}} \right) \quad (\text{if } \mathcal{E}_r \text{ holds})$$

$x \wedge y = \min(x, y)$

overall:

$$R_T \leq \sum_{r=2}^{T} \mathbb{E}[r_r \mathbb{1}_{\mathcal{E}_r}] + \sum_{r=1}^{T} \left(\frac{1}{(r-1)^2} \wedge 1 \right)$$

$$\leq 2 \sum_{r=2}^{T} \alpha_r \left(1 \wedge \|a_r\|_{V_{r-1}^{-1}}\right) + c$$

$$\leq 2\sqrt{\sum_{r=1}^{T} \alpha_r^2}\ \sqrt{\sum_{r=1}^{T} \left(1 \wedge \|a_r\|^2_{V_{r-1}^{-1}}\right)} + c$$

$$\leq c_\delta \sqrt{\sum_{r=1}^{T} d \ln(T)}\ \sqrt{\sum_{r=1}^{T} \left(1 \wedge \|a_r\|^2_{V_{r-1}^{-1}}\right)} + c$$

$$\leq c_\delta \sqrt{dT\ln(T)}\ \sqrt{\sum_{r=1}^{T} \left(1 \wedge \|a_r\|^2_{V_{r-1}^{-1}}\right)} + c$$

Bound on $\sum_{r=1}^{T} \left(1 \wedge \|a_r\|^2_{V_{r-1}^{-1}}\right)$

$$u \wedge 1 \leq 2\ln(1+u)$$

so $$\sum_{r=1}^{T} \left(1 \wedge \|a_r\|^2_{V_{r-1}^{-1}}\right) \leq 2 \underbrace{\sum_{r=1}^{T} \ln\left(1 + \|a_r\|^2_{V_{r-1}^{-1}}\right)}_{= \ln\left(\det\left(\frac{V_T}{V_0}\right)\right)}$$

Indeed, $V_r = V_{r-1} + a_r a_r^T = V_{r-1}^{1/2}\left(I + V_{r-1}^{-1/2} a_r a_r^T V_{r-1}^{-1/2}\right) V_{r-1}^{1/2}$

so $\det(V_r) = \det(V_{r-1}) \cdot \det\left(I + \underbrace{V_{r-1}^{-1/2} a_r a_r^T V_{r-1}^{-1/2}}_{yy^T \text{ is a rank one matrix.}}\right)$

$I + yy^T$ has eigenvalues: $\left(\underbrace{1 + \|y\|^2}_{\text{eigenvector } y.}, 1, \dots, 1\right)$

$\det(V_r) = \det(V_{r-1}) \cdot \left(1 + \|V_{r-1}^{-1/2} a_r\|_2^2\right)$

$= \det(V_{r-1})\left(1 + \|a_r\|^2_{V_{r-1}^{-1}}\right)$

So by induction $\ln(\det(V_r)) = \ln(\det(V_0)) + \sum_{s=1}^{r} \ln\left(1 + \|a_r\|^2_{V_{r-1}^{-1}}\right)$

so $\sum_{r=1}^{T} \left(1 + \|a_r\|^2_{V_{r-1}^{-1}}\right) \leqslant 2\ln\left(\frac{\det(V_T)}{\underbrace{\det(V_0)}_{\lambda^d}}\right)$

$\leqslant 2d \ln\left(1 + \frac{T}{\lambda d}\right)$ Thanks to previous bound.

$\leqslant c_\lambda \, d \ln(T)$

In conclusion, gathering every thing we get.

$R_T \leqslant c_1 \, d \ln T \sqrt{T} + c.$

$\square$