Sequential learning – Lesson 4
Contextual and Linear Bandits

Rémy Degenne

February 10, 2023

Centre Inria de l'Université de Lille

Stochastic bandit

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

- the player observes a context $x_t \in \mathcal{X}$
- the player chooses an arm $k_t \in \Theta$ (compact decision/parameter set, often $\{1, \ldots, K\}$);
- the player observes
  - the rewards of every arm: $X_t^k \sim \nu_k$ for all $k \in \Theta$ $\quad\rightarrow\quad$ full-information feedback
  - the reward of the chosen arm only: $X_t^{k_t} \sim \nu_{k_t}$ $\quad\rightarrow\quad$ bandit feedback.

The goal of the player is to maximize their cumulative reward.

The main reference:

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

(online on Tor Lattimore's webpage)

UCB

Finitely many arms, no contexts.

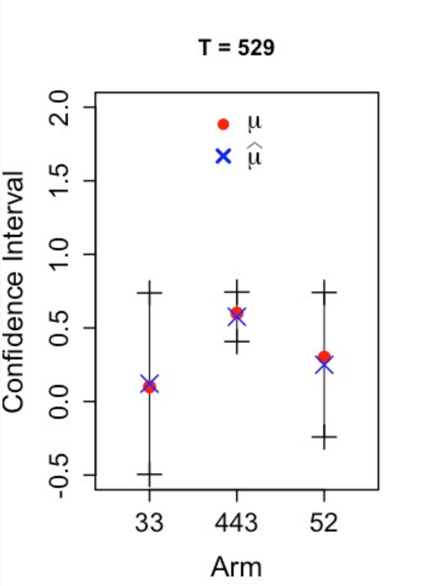
Initialization For rounds $t = 1, \ldots, K$ pull arm $k_t = t$.

For $t = K + 1, \ldots, T$, choose

$$k_t \in \underset{k \in [K]}{\arg\max} \left\{ \widehat{\mu}^k_{t-1} + \sqrt{\frac{2 \log t}{N^k_{t-1}}} \right\},$$

and get reward $X_t^{k_t}$.

UCB


T = 529
μ
$\hat{\mu}$
Confidence Interval
2.0
1.5
1.0
0.5
0.0
-0.5
33
443
52
Arm

UCB Regret Bounds

Theorem 1

If the distributions ν_k have supports all included in $[0,1]$ then for all k such that $\Delta_k > 0$

$$\mathbb{E}\left[N_T^k\right] \leqslant \frac{8 \log T}{\Delta_k^2} + 2\,.$$

In particular, this implies that the expected regret of UCB is upper-bounded as

$$\mathbb{E}[R_T] \leqslant 2K + \sum_{k:\Delta_k>0} \frac{8 \log T}{\Delta_k}\,.$$

Remarks :

- we can also prove $\mathbb{E}[R_T] \lesssim \sqrt{KT \log(T)}$. Close to the optimal $O(\sqrt{KT})$.
- Deals with multiple gaps, without any knowledge of the gaps.
- Anytime algorithm: does not depend on T.

Contextual Bandits

Continuous Bandits

Contextual Bandits with Continuous Contexts

Stochastic Linear Bandits

Contextual Bandits: Motivation

Online advertisement problem:

- A user connects to a website,
- The seller (website algorithm) observes a cookie $x_t \in \mathcal{X}$,
- The seller chooses an ad $k_t \in [K]$,
- The reward is 1 if the user clicks on the ad, 0 otherwise.

Setting and Regret

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

- the player observes a context $x_t \in \mathcal{X}$
- the player chooses an arm $k_t \in \Theta$
- the player observes the reward $X_t^{k_t} \sim \nu_{k_t}(x_t)$ (distribution with mean $\mu_{k_t}(x_t)$, with support in [0,1]).

The goal of the player is to maximize their cumulative reward.

Regret:

$$\boxed{R_T \stackrel{\text{def}}{=} \sum_{t=1}^{T} \mu^*(x_t) - \sum_{t=1}^{T} \mu_{k_t}(x_t)}$$

with $\mu^*(x) = \max_k \mu_k(x)$, mean of the best arm in context x.

Naive Algorithm for Finitely Many Contexts

Here $\mathcal{X}$ is finite.

Idea: treat contexts as independent $\to$ one algorithm per context.

Regret:

$$R_T = \sum_{x\in\mathcal{X}} R_T(x) = \sum_{x\in\mathcal{X}} \left(T_x \mu^*(x) - \sum_{t=1}^{T} \mu_{k_t}(x)\mathbb{I}\{x_t = x\} \right) .$$

where T_x is the number of times context x arises.

Theorem 2

The algorithm using one instance of UCB per context has regret

$$\mathbb{E}R_T \lesssim \sqrt{K|\mathcal{X}|T \log T} .$$

Proof

Issue: many/continuous contexts?

We have a $\sqrt{K|\mathcal{X}|T \log T}$ bound.

What if $|\mathcal{X}|$ is very large? Or $\mathcal{X}$ is continuous?

We treated the function $x \mapsto \nu_k(x)$ as an arbitrary function. Can we make sense of the following hypothesis: on contexts that are “similar”, the distributions (or their means) are “similar” ?

Continuous Bandit

No contexts here.

Continuous bandit setting: stochastic bandits with arm set $\Theta \subseteq \mathbb{R}^d$. To each arm $\theta \in \Theta$ corresponds a distribution $\nu(\theta)$.

Regret: $R_T = T\mu^* - \sum_{t=1}^{T} \mu_{\theta_t}$, with $\mu^* = \sup_{\theta \in \Theta} \mu_\theta$.

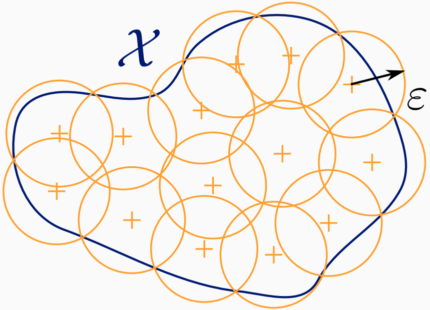
Idea to overcome infinity of arms: close arms have similar means.

Hölder Assumption

The expectation function $\theta \mapsto \mu_\theta$ is β-Hölder: i.e., there exists $c > 0$

$$\forall \theta, \theta' \in \mathcal{X}, \qquad \left|\mu_\theta - \mu_{\theta'}\right| \leqslant c\left\|\theta - \theta'\right\|^\beta,$$

Discretization



ε-covering of $\Theta \subseteq [0,1]^d$ with $O(\varepsilon^{-d})$ balls.

Then:

- arm set = {center of the balls},
- use UCB (for example) using that arm set.

UCB on Hölder Bandits

Theorem 3

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

$$\mathbb{E}R_T \lesssim T\varepsilon^\beta + \sqrt{\frac{T\log(T)}{\varepsilon^d}}\,.$$

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

Proof

Goal: $\mathbb{E}R_T \lesssim T\varepsilon^\beta + \sqrt{\frac{T\log(T)}{\varepsilon^d}}$.

UCB for continuous bandits: remarks

To build the discretization, both β and T need to be known in advance.

- T can be calibrated online through a “doubling trick”.
- β may be tuned through bandit with experts (or bandits where arms are bandit algorithms).

The per-round complexity of such an algorithm is of order $\varepsilon^{-d} \approx T^{\frac{d}{2\beta+d}}$.

$\rightarrow$ It does not explodes with the dimension d and is always smaller than T!

Explanation: higher dimension $d \Rightarrow$ worse regret bound $\Rightarrow$ cruder discretization needed.

Contextual Bandits with Continuous Contexts

Contextual bandit setting, this time with continuous contexts.

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

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

- the environment chooses $x_t \in \mathcal{X}$ and reveals it to the player;
- the player chooses an action $k_t \in \{1, \ldots, K\}$;
- given k_t, the environment draws the reward $X_t \sim \nu_{k_t}(x_t)$ independently from the past;
- the player only observes the feedback X_t.

The player wants to minimize its expected regret defined as

$$\mathbb{E}R_T \overset{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^{T} \mu^*(x_t) - \sum_{t=1}^{T} \mu_{k_t}(x_t)\right] ,$$

where $\mu_k(x) = \mathbb{E}_{X \sim \nu_k(x)}[X]$ and $\mu^*(x) = \max_{k=1,\ldots,K} \mu_k(x)$.

Discretization

Idea: discretize the context space using a regularity assumption on $\mathcal{X} \subset [0,1]^d$.

Theorem 4

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

$$\mathbb{E}R_T \lesssim T\varepsilon^\beta + \sqrt{\frac{KT\log T}{\varepsilon^d}}\,.$$

In particular for ε well-optimized, we have $\mathbb{E}R_T \lesssim T\left(\frac{K\log T}{T}\right)^{\frac{\beta}{2\beta+d}}$.

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

Distribution Dependent Bounds

We used the distribution-free regret bound of UCB. Why?

$\rightarrow$ if the function $\mu_k(x)$ varies smoothly with x, there should be some context with zero suboptimality gap for arm k. $\frac{1}{\Delta}$ is infinite.

Better rates are possible by assuming the following α-margin assumption: the contexts x_t are i.i.d. and satisfy for all $\delta \in (0,1)$

$$\mathbb{P}\left\{ \min_{k:\Delta_k(x_t)>0} \Delta_k(x_t) < \delta \right\} \leqslant \square \delta^\alpha \tag{1}$$

where $\Delta_k(x_t) \stackrel{\text{def}}{=} \mu^*(x) - \mu_k(x)$ and $\square$ is some constant. Larger $\alpha \Rightarrow$ easier problem.

Margin Assumption

α-margin assumption: the contexts x_t are i.i.d. and satisfy for all $\delta \in (0,1)$

$$\mathbb{P}\left\{ \min_{k:\Delta_k(x_t)>0} \Delta_k(x_t) < \delta \right\} \leqslant \square \delta^\alpha \tag{2}$$

where $\Delta_k(x_t) \overset{\text{def}}{=} \mu^*(x) - \mu_k(x)$ and $\square$ is some constant. Larger $\alpha \Rightarrow$ easier problem.

Theorem 5 (Theorem 4.1, Perchet and Rigollet, "The multi-armed bandit problem with covariates", 2013)

Let $\alpha \in (0,1)$, $\beta > 0$ and $\varepsilon > 0$. Assume that $x \mapsto \mu_k(x)$ is β-Hölder for all $k \in [K]$ and that the α-margin assumption (2) holds. Running a bandit algorithm (similar to UCB) independently in each bin of an optimal ε-covering of $\mathcal{X}$, we get

$$\mathbb{E}R_T \lesssim T \left(\frac{K \log K}{T} \right)^{\frac{\beta(\alpha+1)}{2\beta+d}},$$

for optimized ε.

Contextual Bandits

Continuous Bandits

Contextual Bandits with Continuous Contexts

Stochastic Linear Bandits

Stochastic Linear Bandits - Motivation

Main motivation: another way to use contexts.

For contextual bandits,

- we can successfully generalize multi-armed bandits to use contexts
- however, regret rate is worse:
  - $\sqrt{T}$ for non-contextual bandits
  - $T^{\frac{d+1}{d+2}}$ for Lipschitz rewards (β-Hölder with $\beta = 1$).

Goal in this part: use a linear model assumption to get better rates.

Stochastic Linear Bandits

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

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

- 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

$$X_t = \theta_t^\top \mu^* + \varepsilon_t$$

  where ε_t is i.i.d. 1-subgaussian noise. $(\forall \lambda > 0,\ \mathbb{E}\big[\exp(\lambda \varepsilon_t)\big] \leqslant \exp(\lambda^2/2))$
- the player only observes the feedback X_t.

The player wants to minimize its expected regret defined as

$$\mathbb{E}R_T \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^{T} \max_{\theta \in \Theta_t} \theta^\top \mu^* - \sum_{t=1}^{T} \theta_t^\top \mu^*\right].$$

Examples

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

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

  for some unknown parameter $\mu^* \in \mathbb{R}^d$ and feature map $\psi : [K] \times \mathcal{X} \to \mathbb{R}^d$.
- Combinatorial bandit: $\Theta_t \subseteq \{0,1\}^d \to$ combinatorial bandit problems. Example: decision set $=$ 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.

Algorithmic principle: optimism

Algorithm LinUCB - UCB for linear bandits.

- Build confidence region for the parameter: $\mathcal{C}_t$ such that $\mu^* \in \mathcal{C}_t$ with high probability.
- Build confidence bounds for the arm means: $U_t^\theta = \max_{\mu \in \mathcal{C}_t} \theta^\top \mu$.
- Be optimistic: pull $\theta_t = \arg\max_\theta U_t^\theta$.

Main question: how do we get $\mathcal{C}_t$?

Confidence region

After time t, the algorithm observed:

$$\begin{aligned} X_1 &= \theta_1^\top \mu^* + \varepsilon_1 \\ X_2 &= \theta_2^\top \mu^* + \varepsilon_2 \\ \dots \\ X_t &= \theta_t^\top \mu^* + \varepsilon_t \end{aligned}$$

The unknown parameter we want to estimate is μ^*.

Denoting by I_d the $d \times d$ identity matrix and picking $\lambda > 0$, we can estimate μ^* with regularized least square

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

where $V_t \stackrel{\text{def}}{=} \lambda I_d + \sum_{s=1}^{t} \theta_s \theta_s^\top$.

Confidence region

Lemma 1

Let $\delta \in (0,1)$. Then, with probability at least $1-\delta$, if $\max_{\theta \in \Theta_t} \|\theta\|_2 \leqslant 1$, for all $t \geqslant 1$

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

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

Conclusion: with probability $1-\delta$, for all $t \geqslant 1$,

$$\mu^* \in C_t, \quad \text{where} \quad C_t \stackrel{\text{def}}{=} \left\{\theta \in \mathbb{R}^d : \left\|\mu - \widehat{\mu}_{t-1}\right\|_{V_{t-1}} \leqslant \beta(\delta/T)\right\}. \tag{3}$$

Theorem 6

Let $T \geqslant 1$ and $\mu^* \in \mathbb{R}^d$. Assume that for all $\theta \in \cup_{t=1}^T \Theta_t$, $|\theta^\top \mu^*| \leqslant 1$, $\|\mu^*\| \leqslant 1$ and $\|\theta\| \leqslant 1$, then LinUCB with C_t defined as above satisfies the regret bound

$$\mathbb{E}R_T \leqslant \square_\lambda d\sqrt{T}\log(T)\,,$$

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

Remark:

- $O(\sqrt{T})$: the exponent does not depend on d.

Proof

With probability $1 - 1/T$, for all $t \geqslant 1$,

$$\mu^* \in C_t, \quad \text{where} \quad C_t \stackrel{\text{def}}{=} \left\{ \theta \in \mathbb{R}^d : \left\| \mu - \widehat{\mu}_{t-1} \right\|_{V_{t-1}} \leqslant \beta(1/T^2) \right\}.$$

Proof 2

Proof 3

Summary

LinUCB with C_t defined as above satisfies the regret bound

$$\mathbb{E} R_T \lesssim d\sqrt{T}\log(T)\,,$$

To prove it, we assumed the following lemma:

Lemma 2

Let $\delta \in (0,1)$. Then, with probability at least $1-\delta$, if $\max_{\theta\in\Theta_t} \|\theta\|_2 \leqslant 1$, for all $t \geqslant 1$

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

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

Improvements

Under additional assumptions, it is possible to improve the regret bound $O(d\sqrt{T}\log T)$.

- If the set of available actions at time t is fixed and finite; i.e., $\theta_t \in \Theta$ where $|\Theta| = K$. Then, it is possible to achieve

$$\mathbb{E}R_T \leqslant \square\sqrt{Td\log(TK)}\,,$$

  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 $K/\sqrt{d}$.
- 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 $\tilde{O}(\sqrt{dm_0T})$.

Combinatorial Semi-Bandits

Combinatorial semi-bandits: a linear bandit with $\Theta \subseteq \{0,1\}^d$.

When some $\theta = (0,1,1,\ldots,0) \in \Theta$ is pulled, the player observes X_t^k for all $k \in [d]$ for which $\theta_k = 1$.

The total reward is then (for example) $X_t = \sum_k X_t^k \mathbb{I}\{\theta_k = 1\}$.

Observation mechanisms:

- Full information: see X_t^k for all $k \in [d]$.
- Semi-bandit : see X_t^k for all $k \in [d]$ for which $\theta_k = 1$.
- Bandit: see only X_t, total reward.

Algorithm: use all available information in a LinUCB-like algorithm.

Confidence region proof

We want to prove: with probability at least $1-\delta$, if $\max_{\theta \in \Theta_t} \|\theta\|_2 \leqslant 1$, for all $t \geqslant 1$

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

Proof

Proof

Proof

Proof

Proof

References

Thank you!

- Cesa-Bianchi, Nicolo and Gábor Lugosi. Prediction, learning, and games. Cambridge university press, 2006.
- Hazan, Elad et al. “Introduction to online convex optimization”. In: Foundations and Trends® in Optimization 2.3-4 (2016), pp. 157–325.
- Lattimore, Tor and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.
- Perchet, Vianney and Philippe Rigollet. “The multi-armed bandit problem with covariates”. In: The Annals of Statistics (2013), pp. 693–721.
- Shalev-Shwartz, Shai et al. “Online learning and online convex optimization”. In: Foundations and Trends® in Machine Learning 4.2 (2012), pp. 107–194.

Whiteboard

Whiteboard

Whiteboard