

# Bandits

## Blitz Course

Odalric-Ambrym Maillard, Fabien Pesquerel and Patrick Saux

March 25, 2021

# Multi-armed bandit

- $K$  probability measures  $\nu_1, \dots, \nu_K$  with mean  $\mu_1, \dots, \mu_K$ .
- $\mu_{k^*} = \max_{k=1, \dots, K} \mu_k$ .
- Sequence of policies  $\pi_t \in \{1, \dots, K\}$  for  $t = 1, \dots, T$ .
- Rewards  $Y_t \sim \nu_{\pi_t}$ .
- Goal: maximise cumulative expected rewards, or equivalently

$$\min_{(\pi_t)} R_T^\pi = \mathbb{E} \left[ \sum_{t=1}^T \mu_{k^*} - \mu_{\pi_t} \right] = \sum_{k=1}^K (\mu_{k^*} - \mu_k) \underbrace{\mathbb{E}[N_k(t)]}_{:= \sum_{t=1}^T \mathbb{1}_{\pi_t=k}} .$$

- Partial feedback:
  - $\nu_1, \dots, \nu_K$  are not known.
  - Observe only  $Y_t$  at time  $t$ , *not* the rewards that other arms would have generated.

# Link with MDP

- Equivalent to a 1-state MDP (no transition matrix) with action space  $\mathcal{A} = \{1, \dots, K\}$  running for  $T$  episodes of length 1.
- Can be seen as a toy model for exploration-exploitation in model-free RL.
- Bandits are valuable on their own! Many applications: optimisation with scarce data, marketing, health, agriculture...

## Lower bound (instance dependent)

- If  $(\nu_1, \dots, \nu_K)$  are independent measures in family  $\mathcal{D}$ , then (under mild assumptions)

$$\liminf_{T \rightarrow +\infty} \frac{R_T}{\log T} \geq \sum_{k: \mu_k < \mu_{k^*}} \frac{\mu_{k^*} - \mu_k}{\mathcal{K}_{\text{inf}}(\nu_k, \mu_{k^*})}.$$

where

$$\mathcal{K}_{\text{inf}}(\nu_k, \mu_{k^*}) = \inf \left\{ KL(\nu_k \| \nu'_k) \mid \nu'_k \in \mathcal{D}, \int x d\nu'_k(x) > \mu_{k^*} \right\}.$$

- Intuition: hard identification problem when  $\nu_k$  resembles  $\nu_{k^*}$ , easy when the optimal arm is obvious.
- ↳ For  $\mathcal{N}(\theta_k, 1)$ ,  $\mathcal{B}(\theta_k)$ ...  $\mathcal{K}_{\text{inf}}$  can be replaced by  $KL(\theta_k \| \theta_{k^*})$ .

# Naive strategy: Explore-Then-Commit (ETC)

- Choose  $m \in \{1, \dots, T/K\}$ .
- Play each arm  $m$  times.
- Keep playing the best arm afterwards :

$$\pi_t \in \arg \max_{k=1, \dots, K} \underbrace{\frac{1}{N_k(t-1)} \sum_{s=1}^{t-1} Y_s \mathbb{1}_{\pi^s=k}}_{=:\hat{\mu}_{k,t}}.$$

- ✓  $R_T = \mathcal{O}(\log T)$  for a good choice of  $m$ ...
- ✗ ... if one knows all  $\mu_k$  in advance!
- ✗ ... if one knows  $T$ .
- ✗ Cannot rectify after exploration is over.

# Upper Confidence Bound (UCB)

- Compute a  $\delta \in (0, 1)$  upper confidence bound for arm  $k$  at time  $t$ :

$$\mathbb{P}\left(\mu_k \leq UCB_{k,t}\right) \geq 1 - \delta.$$

- Play  $\pi_t \in \arg \max_{k=1,\dots,K} UCB_{k,t}$ .

- Example:  $UCB_{k,t} = \hat{\mu}_{k,t} + R\sqrt{\frac{\log 1/\delta_t}{2N_k(t-1)}}$  with  $\delta_t = 1/t^3$  for  $R$ -sub-Gaussian distributions.

- $R = \frac{B}{2}$  if bounded in range of length  $B$  (Hoeffding lemma).
- $R = \sigma$  if Gaussian of variance  $\sigma^2$ .
- In practice, use it as a hyperparameter to tune.

✓  $R_T = \mathcal{O}(\log T)$ ...

✓ Balances exploration (low  $N_k(t)$ ) and exploitation (high  $\hat{\mu}_{k,t}$ ).

✗ ... good asymptotic rate, but suboptimal factor (does not match the  $\mathcal{K}_{\text{inf}}$ ) (sharper confidence intervals may help).

# Linear Contextual Bandit

- Set of vectors  $\mathcal{X}_t \subset \mathbb{R}^d$ .
- Sequence of policies  $X_t \in \mathcal{X}_t$  for  $t = 1, \dots, T$ .
- Rewards

$$Y_t = X_t^\top \theta^* + \eta_t$$

where  $\eta_t$  is a centred sub-Gaussian noise (think  $\mathcal{N}(0, \sigma^2)$ ).

- Goal: minimise

$$\min_{(\pi_t)} R_T^\pi = \sum_{t=1}^T \max_{x \in \mathcal{X}_t} x^\top \theta^* - X_t^\top \theta^*.$$

- Partial feedback:
  - $\theta^*$  is not known.
  - Observe only  $Y_t$  at time  $t$ , *not* the rewards that other vectors would have generated.

# Applications

- Multi-armed bandits are a special case of linear bandits.
  - $\mathcal{X}_t = \{e_1, \dots, e_K\}$  where  $e_k$  is the  $k$ -th vector of the canonical basis of  $\mathcal{R}^K$  ( $d = K$ ).
  - $\theta_k^* = \mu_k$ ,  $Y_t = \mu_{\pi_t} + \eta_{\pi_t} \sim \nu_k$  for  $k = 1, \dots, K$ .
- Personalised recommendation.
  - At time  $t$ , a user arrives with features  $X_t \in \mathbb{R}^p$  and we make a recommendation  $\pi_t$  among  $K$  options ( $d = pK$ ).
  - $Y_t = X_t^\top \theta_{\pi_t}^* + \eta_t$ .
  - Example:
    - $K$  movies.
    - $X_t$  vector of previously liked genres.
    - $Y_t$  likelihood that user watches movie  $\pi_t$ .



- Solve the regularised least-squares problem

$$\hat{\theta}_t \in \arg \min_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t-1} \|Y_s - X_s^\top \theta\|_2^2 + \lambda \|\theta\|_2^2.$$

- Compute for  $x \in \mathcal{X}_t$

$$UCB_t(x) = x^\top \hat{\theta}_t + \beta_t(x).$$

- Play  $\pi_t \in \arg \max_{x \in \mathcal{X}_t} UCB_t(x)$ .
- ✓  $R_T = \tilde{O}(d\sqrt{T})\dots$
- ✗ ... for a suitable choice of confidence bound  $\beta$ .

# Linear UCB in practice

- $V_t = \sum_{s=1}^{t-1} X_s X_s^\top + \lambda I_d$ .
- $V_t^{-1}$  obtained from  $V_{t-1}^{-1}$  in  $\mathcal{O}(d^2)$  (Sherman-Morrison).
- $\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t-1} Y_s X_s$ .
- $(\eta_t)_{t=1, \dots, T}$   $R$ -sub-Gaussian process.
- $\forall x \in \mathcal{X}_t, \|x\|_2 \leq L$ .
- $\|\theta^*\|_2 \leq S$ .
- $\lambda \geq 1$ .
- $\beta_t(x) = \left( \sqrt{\lambda} S + R \sqrt{d \log \left( 1 + \frac{tL^2}{d\lambda} \right) + 2 \log \frac{1}{\delta}} \right) \sqrt{x^\top V_t^{-1} x}$ .
- In practice :  $\beta_t(x) = \beta_0 \sqrt{x^\top V_t^{-1} x}$  where  $\beta_0$  is a hyperparameter to tune.