Decentralized learning in control and optimization for networks and dynamic games

Part III: Bandits and adversarial optimization

Alexandre Proutiere
KTH
Outline of Part III

• Multi-armed bandit problems
  – Notion of regret
  – Stochastic bandits
  – Adversarial bandits

• Stochastic bandit problem
  – IID setting
  – Lower bound on regret
  – UCB policies, finite time analysis
  – Asymptotically optimal policies: KL-UCB

• Adversarial bandit problems
  – Models
  – Multiplicative update algorithms
Outline of Part III

• Online convex optimization
  – Full information model
  – Bandit setting
Multi-Armed Bandit (MAB)
MAB problem

• Known parameters: number $K$ of arms (or decisions), time horizon (or number of rounds) $T$

• Unknown parameters: how rewards are generated
  
  $X_{j,t}$ : reward of pulling arm $j$ at time $t$

• Objective: maximize the total expected reward at time $T$
Stochastic vs. Adversarial

• Stochastic: rewards sampled from an unknown distribution
  – Example: IID case,
    
    \[(X_{j,t}, t = 1, 2, \ldots)\] IID random variables with mean \(\mu_j\)

• Adversarial setting: rewards chosen by an adversary
  - Oblivious adversary:
    
    \[(X_{j,t}, t = 1, 2, \ldots)\] chosen initially (at time 0)
  
  - Adaptive adversary: rewards depend on the history (selected arms so far)
Applications

- Clinical trials (Thompson 1933)
- Ads placement on webpages
- Routing problems
- ...
Stochastic bandits
Stochastic MAB

• Robbins 1952
• IID rewards

\[(X_{j,t}, t = 1, 2, \ldots)\] IID random variables with mean \(\mu_j\)

• At a given time, an arm is selected and the corresponding random reward is observed
• Best arm: \(j^* = \arg\max_j \mu_j\)
• Under a given policy, the arm selected at time \(t\) is \(j(t)\)

Expected regret:

\[R(t) = t \times \mu_{j^*} - \sum_{n=1}^{t} \mu_{j(n)}\]
Parametric model

- Measure on $\mathbb{R} : \nu$
- Reward distributions parametrized by $\theta \in \mathbb{R}$
- Configuration: $\mathcal{C} = (\theta_1, \ldots, \theta_K)$
- Arm $j$ reward distribution: $X_{j,t} \sim f(x, \theta_j) d\nu(x)$
  \[ \int |x| f(x, \theta_j) d\nu(x) < \infty \]
  \[ \int x f(x, \theta_j) d\nu(x) = \mu(\theta_j) \]
- Kullback-Leibler divergence:
  \[ I(\theta, \lambda) = \int \log \left( \frac{f(x, \theta)}{f(x, \lambda)} \right) f(x, \theta) d\nu(x) \]
Assumptions

• $\mu(\theta)$ strictly increasing
• $I(\theta, \lambda)$ continuous in $\lambda$
  \[ \theta \neq \lambda \implies I(\theta, \lambda) > 0 \]

• Finally: $\forall \lambda, \forall \delta, \exists \lambda'$:
  \[ \mu(\lambda) < \mu(\lambda') < \mu(\lambda) + \delta \]

• Notation: permutation $\sigma$
  \[ \mu(\theta_{\sigma(1)}) \geq \cdots \geq \mu(\theta_{\sigma(K)}) \]
  \[ \mu(\theta_{\sigma(1)}) = \mu(\theta_{\sigma(i)}) > \mu(\theta_{\sigma(i+1)}) \]
Example: Bernoulli rewards

- Rewards take values in \{0,1\}
- Measure \( \nu : \nu = \delta_0 + \delta_1 \)
- We have: \( \theta \in [0, 1] \)

\[
\mu(\theta) = \theta
\]

\[
I(\theta, \lambda) = \theta \log \left( \frac{\theta}{\lambda} \right) + (1 - \theta) \log \left( \frac{1 - \theta}{1 - \lambda} \right)
\]
Regret and uniformly good rules

• Number of time arm $j$ selected up to time $t$: $T_t(j)$

• Expected regret:

$$R(t, C) = \sum_{j \notin \{\sigma(1), \ldots, \sigma(l)\}} (\mu(\theta_{\sigma(1)}) - \mu(\theta_j)) \mathbb{E}[T_t(j)]$$

• Uniformly good rule: for all configuration $C$

$$\mathbb{E}[T_t(j)] = o(t^\alpha), \quad \forall \alpha > 0, \forall j \notin \{\sigma(1), \ldots, \sigma(l)\}$$
Lower bound on regret
Lai and Robbins 1985

**Theorem** Consider any uniformly good rule.

Configuration:  \( C = (\theta_1, \ldots, \theta_K) \)

\[ \forall \epsilon > 0, \; \forall j \notin \{\sigma(1), \ldots, \sigma(l)\}, \]

\[ \lim_{t \to \infty} P_C \left[ T_t(j) \geq \frac{(1 - \epsilon) \log t}{I(\theta_j, \theta_{\sigma(1)})} \right] = 1. \]

Hence:

\[ \lim_{t \to \infty} \inf \frac{R(t, C)}{\log(t)} \geq \sum_{j \notin \{\sigma(1), \ldots, \sigma(l)\}} \frac{\mu(\theta_{\sigma(1)}) - \mu(\theta_j)}{I(\theta_j, \theta_{\sigma(1)})}. \]
Universality of the bound

• Similar bound can be derived for controlled Markov chains, i.e., for parametrized average reward MDP
Model

- Markov chain: $X_n, n \geq 0$
- Action space $A$
- Transition probabilities: $p(y|x, a, \theta)$
- Unknown parameter: $\theta$
- Stationary control laws: $G = (g_1, \ldots, g_K)$
- Under control law $g$, irreducible MC, with stationary distribution $\pi^g_\theta$
- Reward: $\mu_\theta(g) = \int r(x, g(x)) d\pi^g_\theta(x)$

$$\mu^* = \max_g \mu_\theta(g)$$
Lower bound on regret

• The regret can be shown to “look” like:

\[ R(t, \theta) = \sum_{g: \mu_\theta(g) < \mu^*} (\mu^* - \mu_\theta(g)) \mathbb{E}[T_t(g)] \]

• We have: \( \lim_{t \to \infty} \inf_{t} R(t, \theta) \geq c(\theta) \)

\[ c(\theta) = \inf \left\{ \frac{\sum_{j \notin J(\theta)} \alpha_j (\mu^* - \mu_\theta(g_j))}{\inf_{\lambda \in B(\theta)} \sum_{j \notin J(\theta)} \alpha_j I^{g_j}(\theta, \lambda)} : \sum_{j \notin J(\theta)} \alpha_j = 1 \right\} \]

\( J(\theta) \): set of optimal control laws for parameter \( \theta \)

\( B(\theta) \): set of parameters such the optimal control laws under \( \theta \) are not optimal, and cannot be “distinguished”
Upper Confidence Bound policies

- Algorithm: UCB1 (an index policy)

**Initialization:** Play each arm once.

At each step $t > K$:  
$$
\bar{X}_{i,T_{t-1}(i)} = \frac{1}{T_{t-1}(i)} \sum_{s=1}^{T_{t-1}(i)} X_{i,s},
$$

Index of arm $i$:  
$$
\bar{X}_{i,T_{t-1}(i)} + \sqrt{\frac{2 \log t}{T_{t-1}(i)}},
$$

Play the arm with the highest index.
Finite analysis of UCB1

**Theorem** At any time $t$: $\Delta_j = \theta^* - \theta_j$

$$R(t) \leq 8 \sum_{j: \Delta_j > 0} \frac{\log t}{\Delta_j} + \left(1 + \frac{\pi^2}{3}\right) \sum_j \Delta_j$$

Proof. Chernoff-Hoeffding bound

$$X_1, \ldots, X_n \in [0, 1] \text{ i.i.d. with mean } \mu,$$
$$S_n = X_1 + \ldots, X_n,$$
$$P[S_n \geq n\mu + a] \leq e^{-\frac{2a^2}{n}},$$
$$P[S_n \leq n\mu - a] \leq e^{-\frac{2a^2}{n}}$$

Greedy policy

- Algorithm: Greedy

  **Initialization**: Play each arm once.

  At each step $t > K$: $\bar{X}_{i,T_{t-1}(i)} = \frac{1}{T_{t-1}(i)} \sum_{s=1}^{T_{t-1}(i)} X_{i,s}$

  Play the best arm so far with probability $1 - \epsilon_t$,
  play a random arm with probability $\epsilon_t$

- For an appropriate choice of exploration rate, the algorithm is order-optimal
Regret under Greedy algorithm

**Theorem**

\[ \Delta = \min_{j: \Delta_j > 0} \Delta_j, \quad \epsilon_t = \min(1, \frac{6K}{\Delta^2 t}), \]

\[ R(t) \leq \sum_{j: \Delta_j > 0} \frac{C\Delta_j}{\Delta^2} \log t \]

Asymptotically optimal policies

• The lower regret bound solves the following optimization problem:

\[ \inf \sum_j c_j \Delta_j, \]

s.t. \( \forall j \neq j^*, c_j I(\theta_j, \theta^*) \geq \log t \)

• Principle: provide an online solution of the above problem
KL-UCB

- **Algorithm:**
  
  **Initialization:** Play each arm once.

  At each step $t > K$: $\bar{X}_{i,T_{t-1}(i)} = \frac{1}{T_{t-1}(i)} \sum_{s=1}^{T_{t-1}(i)} X_{i,s}$,

  Play the arm with the highest index.

  Index of arm $j$:

  $\max\{q \in [0, 1] : T_{t-1}(j)I(\bar{X}_{j,T_{t-1}(j)}, q) \leq \log t + c \log \log t\}$.

---

**Theorem**

$\forall \epsilon > 0$, $\limsup_{t \to \infty} \frac{R(t)}{\log t} \leq \sum_j \frac{\Delta_j}{I(\theta_j, \theta^*)}$

* The KL-UCB for bounded stochastic bandits and beyond, Garivier-Cappe, COLT, 2011.
Non-stochastic bandits
Model

- Adversarial setting: rewards chosen by an adversary
  - Oblivious adversary:

\[
(X_{j,t}, t = 1, 2, \ldots) \text{ chosen initially (at time 0)}
\]

- Goal: Maximize the cumulative gains obtained.

\[
\text{Regret: } R(t) = \max_j \sum_{s=1}^{t} X_{j,t} - \mathbb{E}\left[\sum_{s=1}^{t} X_{j_t,t}\right]
\]

- Full information: at time \( t \), the forecaster knows

\[
(X_{j,s}, j = 1, \ldots, K, s = 1, \ldots, t - 1)
\]

- Bandit setting: at time \( t \), the forecaster knows

\[
(X_{j_s,s}, s = 1, \ldots, t - 1)
\]
Full information

- Cumulative reward of arm $j$: $S_{j,t-1} = \sum_{s=1}^{t-1} X_{j,s}$

- Follow-the-leader policy does not work!

- Multiplicative update algorithm (Littlestone-Warmuth, 1994)

Play arm $j$ with probability $p_{j,t}$ where:

$$p_{j,t} = \frac{e^{\eta S_{j,t-1}}}{\sum_i e^{\eta S_{i,t-1}}}$$
Multiplicative update algorithms have zero-regret!

The algorithm can be extended when the time horizon is not known, with similar performance.

**Theorem**

\[
\forall t, \quad R(t) \leq \frac{t\eta}{8} + \frac{\log K}{\eta}
\]

For \( \eta = \sqrt{\frac{8 \log K}{t}} \),

\[
R(t) \leq \sqrt{t \frac{\log K}{2}}
\]
Bandit setting

- Cumulative reward of arm \( j \) cannot be observed
- Idea: estimate the cumulative rewards
  
  Unbiased estimator: 
  \[
  \hat{S}_{i,t} = \sum_{s=1} \hat{X}_{i,s}
  \]
  
  \[
  \hat{X}_{i,t} = 1 - \frac{(1 - X_{j_s,s})}{p_{j_s,s}} \times 1_{j_s=i}
  \]

  note that: 
  \[
  \mathbb{E}[\hat{X}_{i,t}] = 1 - \sum_k \frac{(1 - X_{k,s})}{p_{k,s}} \times 1_{k=i} = X_{i,s}
  \]
Bandit setting

- Multiplicative update algorithm:

  Play arm $j$ with probability $p_{j,t}$ where:

  $$p_{j,t} = \frac{e^{\eta \hat{S}_{j,t-1}}}{\sum_i e^{\eta \hat{S}_{i,t-1}}}$$

**Theorem**

$$\forall t, \quad R(t) \leq \frac{tK\eta}{2} + \frac{\log K}{\eta}$$

For $\eta = \sqrt{\frac{2 \log K}{Kt}}$, $R(t) \leq \sqrt{2tK \log K}$
Online convex optimization

Based on:
At the beginning of each year, Volvo has to select a vector $x$ (in a convex set) representing the relative efforts in producing various models (S60, V70, ...). The reward is an arbitrarily varying and unknown concave function of $x$. How to maximize reward over say 50 years?
Model

• Online convex optimization
  – A feasible convex set of actions $X$
  – A sequence of convex cost functions on $X$: $c_1, c_2, \ldots$

• Decision maker
  – Time horizon $N$
  – At step $t$, selected action $x_t$
  – Cost: $c_t(x_t)$
  – Feedback. Full information: $\nabla c_t(x_t)$
  – Bandit: $c_t(x_t)$
Regret

- Cumulative cost: $\sum_{t=1}^{N} c_t(x_t)$
- Cumulative cost of the best action: $\sum_{t=1}^{N} c_t(x^*)$
  $$x^* \in \arg \max_{x \in X} \sum_{t=1}^{N} c_t(x)$$
- Regret: $R(N) = \sum_{t=1}^{N} c_t(x_t) - \sum_{t=1}^{N} c_t(x^*)$
- Goal: minimize regret
Full information

- Online gradient descent

\[ w_{t+1} = x_t - \eta \nabla c_t(x_t) \]

\[ x_{t+1} = \arg \min_{x \in X} \| x - w_{t+1} \|_2^2 \]
**Theorem**

Assume that \( \text{diam}(X) \leq R \)

\[ \| \nabla c_t(x) \|_2^2 \leq G, \quad \forall x \in X, \forall t = 1, \ldots, N \]

Then under the online gradient descent algorithm:

\[ R(N) \leq RG\sqrt{N} \]
Bandit setting

- Online convex optimization
  - A feasible convex set of actions $X$
  - A sequence of convex cost functions on $X$: $c_1, c_2, \ldots$

- Decision maker
  - Time horizon $N$
  - At step $t$, selected action $x_t$
  - Cost: $c_t(x_t)$
Bandit setting

- Idea: one sample estimate of the gradient

\[ \hat{f}(x) = \mathbb{E}_{v \in B}[f(x + \delta v)] \]
\[ B = \{ x : \|x\|_2 \leq 1 \} \]

\[ \mathbb{E}_{u \in S}[f(x + \delta u)u] = \frac{\delta}{d} \nabla \hat{f}(x) \]
\[ S = \{ x : \|x\|_2 = 1 \} \]

- Simulated gradient descent algorithm

\[ u_t \text{ uniformly chosen in } B \]
\[ x_t = y_t + \delta u_t \]
\[ y_{t+1} = P_{(1-\alpha)}X(y_t - \nu c_t(x_t)u_t) \]
Theorem

Assume that $r \leq \text{diam}(X) \leq R$

$$\left\| \nabla c_t(x) \right\|_2^2 \leq G, \quad \forall x \in X, \forall t = 1, \ldots, N$$

$$c_t(x) \leq C, \quad \forall x \in X, \forall t$$

If $N \geq \left( \frac{3Rd}{2r} \right)^2$, \( \nu = \frac{R}{C\sqrt{N}} \), \( \delta = \left( \frac{rR^2d^2}{12N} \right)^{1/3} \), \( \alpha = \left( \frac{3Rd}{2r\sqrt{N}} \right)^{1/3} \)

Then under the online gradient descent algorithm:

$$\mathbb{E}[R(N)] \leq 3CN^{5/6}(dR/r)^{1/3}$$
Summary

• Zero-regret algorithms exist in general (MAB, online optimization)
  \[ \lim_{t \to \infty} \frac{R(t)}{t} = 0 \]

• We are able to identify the best action in the long run, and a bit more ...

• Regrets:

<table>
<thead>
<tr>
<th>Problem</th>
<th>Algorithm</th>
<th>Regret scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stochastic bandit</td>
<td>Optimal</td>
<td>C.log t</td>
</tr>
<tr>
<td></td>
<td>KL-UCB</td>
<td>C.log t</td>
</tr>
<tr>
<td>Non-stochastic bandit</td>
<td>Optimal</td>
<td>\sqrt{t}</td>
</tr>
<tr>
<td></td>
<td>MUA</td>
<td>\sqrt{t}</td>
</tr>
<tr>
<td>Online cx opt.</td>
<td>Full inf.</td>
<td>\sqrt{t}</td>
</tr>
<tr>
<td></td>
<td>Bandit</td>
<td>t^{5/6}</td>
</tr>
</tbody>
</table>