IEOR 8100-001: Learning and optimization for sequential decision making

Instructor: Shipra Agrawal

Industrial Engineering and Operations Research
Columbia University

...

Outline

In the regular MAB:

- At time t, pull an arm $I_t = i$, observe reward $r_t \in [0, 1]$, i.i.d. from distribution ν_i.
- maximize $\sum_t r_t$

Now, consider Reward + cost:

- At time t, pull an arm $I_t = i$, observe reward $r_t \in [0, 1]$ and cost $c_t \in [0, 1]$ i.i.d. from joint distribution D_i.
- Stop when $\sum_t c_t > B$.
- maximize $\sum_{t=1} r_t$ s.t. $\sum_t c_t \leq B$

Example: Dynamic pricing with limited supply

Assuming N discrete possible prices. Each price an arm i.

- At time t, pull an arm $I_t = i$ (price q_i)
- observe reward $r_t = q_i$, cost $c_t = 1$ if purchase happened (with probability $S(q_i)$); and reward $r_t = 0$, $c_t = 0$ otherwise (with probability $1 - S(q_i)$)
- mean reward $q_i S(q_i)$, mean cost $S(q_i)$
- (Given the arm (price), the rewards and costs were iid generated for every pull, but not independent of each other)
- Stop when supply ends.
- Goal: maximize $\sum_{t=1} r_t$ s.t. $\sum_t c_t \leq k$

More generally: Bandit with Knapsacks

At time t,

- pull an arm $I_t = i$, observe reward $r_t \in [0, 1]$ and cost $\mathbf{c}_t \in [0, 1]^d$.
- Given $I_t = i$ $(r_t, \mathbf{c}_t) \sim D_i$, $\mathbb{E}[r_t] = \mu_i$, $\mathbb{E}[c_{tj} | I_t = i] = C_{ij}$.
- Stop when any budget constraint is violated.
- Goal is to maximize $\sum_{t=1} r_t$ s.t. $\sum_t \mathbf{c}_t \leq \mathbf{B}$

Here, $B_j \geq 0$.

Regret

What does optimal policy look like? Policy = mapping from history to action.

- In regular stochastic MAB, playing I^* all the time achieved best expected regret. [Static pure optimal policy]
- Comparing to single best arm not enough: two resources, two types arms that consume $[0; 1]$ and $[1; 0]$. For each type, the highest mean reward is μ^*. Optimal decision would be to play best arm of each type half of the time. [mixed policy]
- In general, optimal policy could be dynamic! play type 1 until resource 1 is exhausted, then play type 2. [Dynamic policy]

Optimal static policy

Define optimal static mixed policy as a distribution $\{p_i\}$ over arms,

$$\bar{\text{OPT}} = \max \sum_i p_i \mu_i \text{ s.t. } \sum_i p_i C_{ij} \leq \frac{B_j}{T}; \sum_i p_i \leq 1. \quad (1)$$

- Sanity check: without constraints an optimal solution is $p_i = 1$ for $i = I^*$, and 0 elsewhere.
- Other example: costs $[0; 1]$ and $[1; 0]$, $B_1 = B_2 = B \leq T/2$, Two point distribution $p_i = B/T$ for best arm of each type.
- **Note:** only satisfies constraint in expectation.

Theorem: Assuming a constraint satisfying dynamic policy exists, then (1) is feasible and $\text{OPT} = T\bar{\text{OPT}}$ bounds the maximum expected reward achievable by optimal dynamic policy.

Therefore, *Sufficient to compare reward with that of a static policy*

Proof of Theorem: OPT bounds the optimal dynamic reward

- Count the average number of times that optimal policy plays arm i as X_i.
- Let $p_i = \mathbb{E}[X_i]$.
- Then, $\sum_i p_i \mu_i = \mathbb{E}[\sum_i X_i \mu_i] = \mathbb{E}[\frac{1}{T} \sum_t r_t]$, expected reward of optimal dynamic policy.
- $\sum_i p_i C_{ij} = \mathbb{E}[\sum_i X_i C_{ij}] = \mathbb{E}[\frac{1}{T} \sum_t c_{tj}] \leq B_j$, from feasibility of optimal dynamic policy.
- Therefore, $\{p_i\}$ is a feasible solution to (1), and its expected reward is more than the average expected reward of optimal dynamic policy.

Regret

Assuming the algorithm stops at time τ when a budget constraint is violated, define regret

$$R(T) = \mathrm{OPT} - \mathbb{E}[\sum_{t=1}^{\tau} r_t]$$

- By previous observations, bounding regret bounds the optimality gap: gap between expected reward of optimal dynamic policy and expected reward of the algorithm.

Stopping time τ needs special check of budget constraints.
Next : use Lagrange multiplier to combine reward and cost – **unconstrained** but **nonlinear** bandit problem.

Outline for solving

First, Reduce our constrained problem to unconstrained but non linear bandit problem

- Original MAB was unconstrained: maximize $\sum_{t=1}^{T} r_t$.
- We will reduce to unconstrained problem: maximize $f(\sum_{t=1}^{T} r_t, \sum_t \mathbf{c}_t)$, where f will be a concave 1-Lipschitz function.

Second, show a solution method for any concave L-Lipschitz function f. This is useful in itself.

Assume $B_j = B$ for all j. (w.l.o.g. $B = \min_j B_j$). Let $Z = 2\text{OPT}/B$ (Suppose we know value of OPT). Define function f as:

$$f(r, \mathbf{c}) = (r - Z \max_j(c_j - \frac{B}{T}, 0))$$

That is,

$$f(\frac{1}{T}\sum_t r_t, \frac{1}{T}\sum_t \mathbf{c}_t) = (\frac{1}{T}\sum_t r_t - Z \max_j(\frac{1}{T}\sum_t c_{tj} - \frac{B}{T}, 0))$$

Interpretation:

$$f(\text{reward, cost}) = (\text{reward - } Z\times \text{ constraint violation})$$

Consider a new *unconstrained bandit problem:*

$$\text{maximize } f(\frac{1}{T}\sum_t r_t, \frac{1}{T}\sum_t \mathbf{c}_t)$$

Unconstrained but non-linear bandit problem

At time t,

- pull arm $I_t = i$
- Observe $\mathbf{v}_t = (r_t, \mathbf{c}_t) \sim D_i$,
- Goal: maximize $f(\frac{1}{T}\sum_t r_t, \frac{1}{T}\sum_t \mathbf{c}_t)$

Regret:

$$R_f(T) = T\mathrm{O\bar{P}T}_f - T\mathbb{E}[f(\frac{1}{T}\sum_{t=1}^{T} r_t, \frac{1}{T}\sum_{t=1}^{T} \mathbf{c}_t)]$$

where

$$\mathrm{O\bar{P}T}_f = \max_{p:\sum_i p_i \leq 1} f(\sum_i p_i\mu_i, \sum_i p_i C_i) = \sum_i p_i\mu_i - Z \max_j(C_{ij} - \frac{B}{T}, 0)$$

Will this solve BwK?

Why $Z = 2\mathrm{OPT}/B$? If Z is too high, we are paying too much importance to constraints, might suffer on reward. If too low, might have good reward but violate the constraint. Is this Z the right multiplier

Theorem

Let $R'(T)$ be the regret of the constrained problem with larger budget

$$B' = B + \frac{1}{Z} R_f(T) + \tilde{O}(\sqrt{T}),$$

then

$$R'(T) \leq 2R_f(T) + Z\tilde{O}(\sqrt{T}),$$

Here $R'(T) = \mathrm{OPT} - \sum_{t=1}^{\tau} r_t$ is regret assuming you abort on time τ, the first time a budget B' is violated.

Proof outline

We prove the following two statements.

- We show that $\mathrm{OPT}_f \geq \mathrm{OPT} - R_f(T) - Z\tilde{O}(\sqrt{T})$: this is because the optimal solution for $\bar{\mathrm{OPT}}$ forms a feasible solution for $\bar{\mathrm{OPT}}_f$ with value $\bar{\mathrm{OPT}} - R_f(T) - Z\tilde{O}(\sqrt{T})$.
- We show that $\mathbb{E}[\sum_{t=1}^{T} \mathbf{c}_t] \leq B + \frac{1}{Z} R_f(T)$. This is proved using definition of Z. Intuitively, the most advantage you can get by violating budget B by $B\epsilon$ is $\epsilon\mathrm{OPT} = \frac{Z}{2}(\epsilon B)$. That is, advantage is only $\frac{Z}{2}\times$ budget violation, which is less than the penalty that f imposes on budget violation. Therefore, there is no benefit in violating budget, and by maximizing f we will be maximizing reward while trying to ensure that the budget is not violated.

From the second observation, we get that the budget constraints would not be violated with high probability if we had $B + \frac{1}{Z} R_f(T) + \tilde{O}(\sqrt{T})$ budget, so $\tau = T$, and $\sum_{t=1}^{\tau} r_t = \sum_{t=1}^{T} r_t$. And, therefore, $R'(T) \leq \mathrm{OPT}_f - \sum_{t=1}^{T} r_t + R_f(T) + Z\tilde{O}(\sqrt{T}) = 2R_f(T) + Z\tilde{O}(\sqrt{T})$.

Is this theorem enough?

Idea: If we can uniformly bound $R_f(T)$ by $\tilde{O}(Z\sqrt{NB})$, then start with smaller budget $B - \sqrt{NB}$. Show that this would only hurt by $Z\sqrt{NB}$.
Then,

- w.h.p, we are within budget, just abort in other cases to satisfy budget.
- And, $R(T) \leq R_f(T) \leq Z\sqrt{B} \leq \frac{2\text{OPT}}{B}\sqrt{B}$

Total regret $R(T) + Z\sqrt{NB} \leq O(Z\sqrt{NB})$.
Competitive ratio (multiplication approximation factor):

$$\frac{\mathbb{E}[\sum_{t=1}^{\tau} r_t]}{\text{OPT}} \geq \frac{\text{OPT} - \text{OPT}\sqrt{\frac{N}{B}}}{\text{OPT}} = 1 - O(\sqrt{\frac{N}{B}})$$

This actually matches the lower bound for this problem. Next, we will show an algorithm that achieves bound of $\tilde{O}(Z\sqrt{NT})$ on $R_f(T)$. Check the references for this lecture for the $\tilde{O}(Z\sqrt{NB})$ bound.

General unconstrained but non-linear bandit problem

Given a concave L-Lipschitz function $f(\mathbf{v})$, $f : \mathbb{R}^d \to [0, 1]$.
At time t,

- pull arm $I_t = i$
- Observe $\mathbf{v}_t \sim D_i$, $\mathbb{E}[\mathbf{v}_t] = V_i \in R^d$
- Goal: maximize $f(\frac{1}{T}\sum_t \mathbf{v}_t)$

Regret:

$$R_f(T) = T\mathrm{O\bar{P}T}_f - Tf(\frac{1}{T}\sum_{t=1}^{T}\mathbf{v}_t)$$

where

$$\mathrm{O\bar{P}T}_f = \max_{p:\sum_i p_i \leq 1} f(\sum_i p_i V_i)$$

Let's try to get $\tilde{O}(L\sqrt{NT})$ regret for this problem. Later, we will see how to get $\tilde{O}(Z\sqrt{NB})$ for the special case of *BwK*.

Full information "before" taking the decision [Agrawal, Devanur SODA 2015]

Given a concave L-Lipschitz function f. $f : [0,1]^d \to [0,1]$.
At time t,

- Observe $\mathbf{v}_{it}, i = 1, \ldots, N \sim D, \mathbb{E}[\mathbf{v}_{it}] = V_i \in R^d$
- pull arm I_t, let $\mathbf{v}_t := \mathbf{v}_{I_t,t}$
- Goal: maximize $f(\frac{1}{T}\sum_t \mathbf{v}_t)$

Regret:

$$R_f(T) = T\mathrm{O\bar{P}T}_f - Tf(\frac{1}{T}\sum_{t=1}^{T}\mathbf{v}_t)$$

where

$$\mathrm{O\bar{P}T}_f = \max_{p:\sum_i p_i \leq 1} f(\sum_i p_i V_i)$$

Overall idea: linearize

Easy problem in sum of rewards objective.

$$\max \sum_t r_{I_t,t} = \sum_t \max r_{I_t,t}$$

At every time t, just pick $\arg\max_i r_{it}$.

$$\max f(\sum_t r_{I_t,t}) \neq \sum_t f(\max r_{I_t,t})$$

Overall idea: linearize

- Fenchel duality: concave function as max of linear functions [PROVE]

$$f(\mathbf{v}) = \min_{\|\boldsymbol{\theta}\|_* \leq L} f^*(\boldsymbol{\theta}) - \boldsymbol{\theta} \cdot \mathbf{v}$$

where

$$f^*(\boldsymbol{\theta}) = \max_{\mathbf{y}} \boldsymbol{\theta}^T \mathbf{y} + f(\mathbf{y})$$

Fenchel conjugate is always convex.
$f^*(\boldsymbol{\theta})$: argument is slope of a linear function, value is function value at point where it becomes tangent to function.
$f(\mathbf{v})$: argument is a point, value is value of the linear function which is tangent.
For every point, there is linear function (tangent) which has same value as $f(\mathbf{v})$.

- In "**hindsight**", there exists a $\tilde{\theta}$ such that

$$f(\tfrac{1}{T}\sum_t \mathbf{v}_t) = f^*(\tilde{\theta}) - \tfrac{1}{T}\tilde{\theta} \cdot \sum_t \mathbf{v}_t$$

  - Separates out the contribution of each $\mathbf{v}_t$ and tells how to weigh different components of $\mathbf{v}_t$.
  - But, we cannot possibly know this $\tilde{\theta}$ before making the decisions.

Algorithm structure

Predict a $\boldsymbol{\theta}_t$ at time t, use it to decide which arm to play:

$$I_t := \arg\max_i f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \mathbf{v}_{it} = \arg\min_i \boldsymbol{\theta}_t \mathbf{v}_{it}$$

Predictions of $\boldsymbol{\theta}_t$ are made in such a way that they can compete with the " "best single $\boldsymbol{\theta}$", i.e., $\tilde{\boldsymbol{\theta}}$ in hindsight. That is,

$$\begin{aligned} \sum_t f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \mathbf{v}_t &\leq \sum_t (f^*(\tilde{\boldsymbol{\theta}}) - \tilde{\boldsymbol{\theta}} \cdot \mathbf{v}_t) + LD\sqrt{T} \\ &= Tf(\frac{1}{T}\sum_t \mathbf{v}_t) + LD\sqrt{T} \end{aligned}$$

where $D = ||x||$.

Is this enough? We will come back to this point.

Competing with the best single decision in hindsight

Recall: Adversarial linear bandits full information case . We used online gradient ascent to solve following problem: At time t,

- Pick $x_t \in A$,
- Observe w_t, get reward $w_t \cdot x_t$.

Regret was defined to compete with a single decision in hindsight:

$$\max_x \sum_t w_t x - \sum_t w_t x_t$$

Online gradient ascent achieved $O(DG\sqrt{T})$ bound for $D \geq ||x_t||, G \geq ||w_t||$.

Competing with the best single decision in hindsight

Can be extended to the following online convex optimization problem: At time t,

- Pick $x_t \in A$,
- Observe convex function h_t, get reward $h_t(x_t)$.

$$Regret(T) = \sum_t h_t(x_t) - \min_x \sum_t h_t(x) \leq O(DG\sqrt{T})$$

achieved on applying online gradient descent to gradients of h_t.

Is competing with best single θ enough?

We have

$$\begin{aligned}\sum_t f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \mathbf{v}_t &\leq \sum_t (f^*(\boldsymbol{\theta}^*) - \boldsymbol{\theta}^* \cdot \mathbf{v}_t) + LD\sqrt{T} \\ &= Tf(\frac{1}{T}\sum_t \mathbf{v}_t) + LD\sqrt{T}\end{aligned}$$

Why is approximating from one side enough? Intuition: any tangent is above the concave function, our decision is maximizing something that is already above the function, so it is better than the optimal function value.

Let's say that optimal static policy (using distribution p^*) chose $\mathbf{v}_t^*$ at time t. Because of our choice of $\mathbf{v}_t$:

$$\begin{aligned}
\sum_t f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \mathbf{v}_t &\geq \sum_t f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \mathbf{v}_t^* \\
\mathbb{E}[\sum_t f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \mathbf{v}_t | \boldsymbol{\theta}_t] &= \sum_t f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \mathbb{E}[\mathbf{v}_t^* | \boldsymbol{\theta}_t] \\
&= \sum_t f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot (p^* V) \\
&\geq \sum_t \min_{\boldsymbol{\theta}} f^*(\boldsymbol{\theta}) - \boldsymbol{\theta} \cdot (p^* V) \\
&= \sum_t f(p^* V) = T\mathrm{O}\bar{\mathrm{P}}\mathrm{T}_f
\end{aligned}$$

Combining

$$R_f(T) = T\mathrm{O}\bar{\mathrm{P}}\mathrm{T}_f - Tf(\frac{1}{T}\sum_t \mathbf{v}_t) \leq LD\sqrt{T}$$

Summary of algorithm in full information setting

Back and forth between online prediction and decision making

- Online learning/ Online convex optimization predicts a $\boldsymbol{\theta}_t$ gives it to the decision making algorithm.
- Algorithm picks $I_t = \arg\max_i \boldsymbol{\theta}_t \cdot \mathbf{v}_{it}$ sends $\mathbf{v}_t = \mathbf{v}_{I_t,t}$ to the online prediction algorithm.

How to handle bandit feedback?

Problem

- You get to see only $\mathbf{v}_t = \mathbf{v}_{i,t}$ "after" pulling arm $I_t = i$. You don't see any other $\mathbf{v}_{i',t}$.
- Cannot solve

$$I_t = \arg\max_i f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \mathbf{v}_{it} = \arg\min_i \boldsymbol{\theta}_t \mathbf{v}_{it}$$

Solution

- Use estimates of $\mathbb{E}[\mathbf{v}_{it}] = V_i$ instead of $\mathbf{v}_{it}$ for all i.
- Estimates are upper bounds in the following way: they satisfy $f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \mathbb{E}[\mathbf{v}_{it}] \leq f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \tilde{\mathbf{v}}_{it}$ with high probability.
- Easy to construct: use UCB for component V_{ij} where $\theta_{tj} < 0$, and LCB for V_{ij} where $\theta_{tj} > 0$.

Why does it work? Intutition: we used tangents which were "above" the function value: The approximation of the tangent we are using is still "above the function value", and the further error introduced is bounded by UCB bounds.

Summary of algorithm in full information setting

Back and forth between online prediction and decision making

- Online learning/ Online convex optimization predicts a $\boldsymbol{\theta}_t$ gives it to the decision making algorithm.
- Algorithm picks $I_t = \arg\max_i \boldsymbol{\theta}_t \cdot \tilde{\mathbf{v}}_{it}$,
- Algorithm observes $\mathbf{v}_t$, improves its estimates and sends it to online learning.

Online learning gets the correct $\mathbf{v}_t$, its guarantees still are the same. For the other side, we can only prove the guarantee for $\tilde{v}_t$

Is competing with best single θ enough?

We have

$$\begin{aligned}\sum_t f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \mathbf{v}_t &\leq \sum_t (f^*(\boldsymbol{\theta}^*) - \boldsymbol{\theta}^* \cdot \mathbf{v}_t) + LD\sqrt{T} \\ &= Tf(\frac{1}{T}\sum_t \mathbf{v}_t) + LD\sqrt{T}\end{aligned}$$

Why is approximating from one side enough? Intuition: any tangent is above the concave function, our decision is maximizing something that is already above the function, so it is better than the optimal function value.

Let's say that optimal static policy (using distribution p^*) chose $\mathbf{v}_t^*$ at time t. Because of our choice of l_t:

$$\begin{aligned}
\sum_t f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \tilde{\mathbf{v}}_t &\geq \sum_t f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \tilde{\mathbf{v}}_t^* \\
&\geq \sum_t f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \mathbf{v}_t^* \\
\sum_t \mathbb{E}[f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \tilde{\mathbf{v}}_t | \boldsymbol{\theta}_t] &= \sum_t f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \mathbb{E}[\mathbf{v}_t^* | \boldsymbol{\theta}_t] \\
&= \sum_t f^*(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot (p^* V) \\
&\geq \sum_t \min_{\boldsymbol{\theta}} f^*(\boldsymbol{\theta}) - \boldsymbol{\theta} \cdot (p^* V) \\
&= \sum_t f(p^* V) = T \mathrm{OPT}_f
\end{aligned}$$

Combining

$$R_f(T) = T\mathrm{OPT}_f - Tf\left(\frac{1}{T}\sum_t \mathbf{v}_t\right) \leq Tf\left(\frac{1}{T}\sum_t \tilde{\mathbf{v}}_t\right) - Tf\left(\frac{1}{T} + \sum_t \mathbf{v}_t\right) LD\sqrt{T}$$