IOE 516

Stochastic Processes II

Winter Term, 2024

Prof. Xiuli Chao

Email: xchao@umich.edu

Recap

- In the last lecture, we discussed concentration inequalities for sub-Gaussian and sub-exponential random variables.

- We also started to discuss revenue management problem. The simplest model is a single-leg dynamic pricing optimization.

DP Formulation

- Recall that the DP formulation of the RM problem is

$$
\begin{aligned}
\max_{p_t} \quad & E\Big[\sum_{t=1}^{T} p_t D_t(p_t)\Big] \\
s.t. \quad & \sum_{t=1}^{T} D_t(p_t) \leq N, \\
& p_t \geq 0,\ t = 1, \ldots T.
\end{aligned}
$$

Fluid approximation model

- And the fluid model is, after changing the decision variable to d_t and using the inverse function $p(d_t)$ (assumed to be bounded),

$$
\begin{aligned}
\max_{p_t} \quad & \sum_{t=1}^{T} d_t \cdot p(d_t) \\
s.t. \quad & \sum_{t=1}^{T} d_t \le N, \\
& d_t \ge 0, \; t = 1, \ldots T.
\end{aligned}
$$

- Assuming concavity of $d_t p(d_t)$, this is a convex optimization problem whose optimal solution is easy to obtain.

Lemma

- Let $d^* = \arg\max_d dp(d)$ and $p^* = d(p^*)$.

- **Claim 1.** Optimal solution to the fluid model is

$$
\begin{aligned}
d^*(N/T) &:= \min\{d^*, N/T\}, \\
p^*(N/T) &:= \max\{p^*, p(N/T)\}.
\end{aligned}
$$

- **Claim 2.** The optimal objective value of fluid model,

$$Tp^*(N/T)d^*(N/T),$$

is an upper bound for the original problem.

How do we evaluate a proposed policy?

- The loss, or regret is defined as the difference between the value function of the said policy and that of the true optimal solution.

- Since we do not know the true optimal value function, we evaluate the performance of the policy by, if our policy has price p_t in period t,

$$R(T) = Tp^*(N/T)d^*(N/T) - E\Big[\sum_{t=1}^{T} p_t D_t(p_t)\Big].$$

- The loss per period is $R(T)/T$.

Theorem

- Let

$$p^*(N/T) := \max\{p^*, p(N/T)\}.$$

- Then, using the static policy $p^*(N/T)$ in each period has a total loss bounded by

$$L(T) = O(\sqrt{T \log T}).$$

- The average loss per period is $O\left(\sqrt{\frac{\log T}{T}}\right)$.

Analysis

- Consider the proposed static policy with $p^*(N/T)$ and expected demand per period is $d^*(N/T)$.

- Denote the demand under policy $p_t = p^*(N/T)$, $t = 1, \ldots, T$, by $D_1, D_2, \ldots, D_T$, which is bounded from above by $\bar{D}$. Then $E[D_t] = d^*(N/T)$. Let $S_T = \sum_{t=1}^{T} D_t$. By Concentration inequality, with appropriate choice of α, we have

$$P\Big(|S_T - Td^*(N/T)| > \sqrt{\alpha T \log T}\Big) \leq 1/T$$

Analysis

- Define good event $\mathcal{A} = \{|S_T - Td^*(N/T)| \leq \sqrt{\alpha T \log T}\}$. Then

$$P(\mathcal{A}) \geq 1 - 2/T, \quad P(\mathcal{A}^c) \leq 2/T.$$

- We have

$$\begin{aligned}
& |E\Big[\sum_{t=1}^{T} p^*(N/T)D_t\Big] - Tp^*(N/T)d^*(N/T)| \\
\leq \; & p^*(N/T)E\Big[|\sum_{t=1}^{T} D_t - Td^*(N/T)| \;\Big|\; \mathcal{A}\Big]P(\mathcal{A}) \\
& + p^*(N/T)E\Big[|\sum_{t=1}^{T} D_t - Td^*(N/T)| \;\Big|\; \mathcal{A}^c\Big]P(\mathcal{A}^c) \\
\leq \; & p^*(N/T)\sqrt{\alpha T \log T} + T\bar{D} \times \frac{2}{T} \\
= \; & O(\sqrt{T \log T}).
\end{aligned}$$

Multi-Armed Bandit (MAB) problem

- There are m arms. Arm a generates i.i.d. reward with mean $\mu(a)$ when playing arm a. There DM has no information about the reward from each arm, and needs to determine a strategy to maximize the expected total reward up to any time T. Assume that reward is bounded (and WLOG, by 1).

- Let $n_t(a)$ denote the number of plays of arm a before time t, then we can compute the sample mean $\bar{\mu}_t(a)$. Let

$$r_t(a) = \sqrt{\frac{\alpha \log n_t(a)}{n_t(a)}}.$$

- You can call $[\bar{\mu}_t(a) - r_t(a), \bar{\mu}_t(a) + r_t(a)]$ the confidence interval of true mean $\mu(a)$ by time t.

Upper Confident Bound Method (UCB)

- **UCB Algorithm**: First, play each arm once.

- After that, in each period t always play the arm that has has the highest UCB:

$$UCB_t(a) = \bar{\mu}_t(a) + r_t(a) = \bar{\mu}_t(a) + \sqrt{\frac{\alpha \log n_t(a)}{n_t(a)}},$$

where α is some parameter.

Remarks

- The algorithms in the literature typically use a larger bound, defined by

$$UCB_t(a) = \bar{\mu}_t(a) + \sqrt{\frac{\alpha \log t}{n_t(a)}}$$

or

$$UCB_t(a) = \bar{\mu}_t(a) + \sqrt{\frac{\alpha \log T}{n_t(a)}} \tag{2}$$

In the discussion below, we use (2).

- UCB uses the so-called "optimistic estimate" to avoid chance of under-play.

- The same approach has been used to develop algorithms for general RL problems.

Probability of good event

- Define event $\mathcal{A}_t(a)$ by

$$\mathcal{A}_t(a) \;\; = \;\; \{|\bar{\mu}_t(a) - \mu(a)| > r_t(a)\}.$$

- Then concentration inequality shows, for some choice of α,

$$P\Big(\mathcal{A}_t(a)\Big) \leq \frac{2}{T^4}.$$

- Let $\mathcal{A} = \bigcup_{t,a} \mathcal{A}_t(a)$, then by union bound (assuming $m \leq T$)

$$P(\mathcal{A}) \leq \frac{2}{T^2}$$

- We shall call $\mathcal{A}$ "bad event" and $\mathcal{A}^c$ "good event".

Regret

- Clearly, the optimal policy is to play $a^* = \arg\max_a \mu(a)$. WLOG, suppose $a^* = 1$. For convenience we let $\Delta(a) = \mu(1) - \mu(a)$.

- If a policy plays arm a_t in period t, then its regret is

$$R_T \; = \; E\Big[\sum_{t=1}^{T} \Big(\mu(1) - \mu(a_t)\Big)\Big]$$

- **Theorem.** We have two regret bounds for UCB algorithm:

$$\begin{aligned} R_T &\le \sum_{a:\Delta(a)>0} \frac{4\alpha}{\Delta(a)} \log T + 2, \\ R_T &\le (1+4\alpha)\sqrt{mT\log T} + 2. \end{aligned}$$

Remark

- The first bound is known as "instance-dependent regret", and the second regret bound that is "instance-independent regret".

Analysis

- **Question 1.** When will you play a wrong arm in period t?

- **Answer:** When $\bar{\mu}_t(a) + r_t(a) \geq \bar{\mu}_t(1) + r_t(1)$ for some $a \neq 1$.

- Under good event $\mathcal{A}^c$, we have, for each t and a,

$$\mu(a) + 2r_t(a) \geq \bar{\mu}_t(a) + r_t(a)$$

and

$$\bar{\mu}_t(1) + r_t(1) \geq \mu(1).$$

- Combining we obtain, if an arm $a \neq 1$ is played at t,

$$\mu(a) + 2r_t(a) \geq \mu(1).$$

- This implies $r_t(a) \geq \Delta(a)/2$, or

$$n_t(a) \leq \frac{4\alpha}{\Delta^2(a)} \log T.$$

- **Result**. The analysis above shows that, any arm $a \neq 1$ is played at most $4\alpha\Delta^{-2}(a) \log T$ times under $\mathcal{A}^c$.

First regret

- **Question 2.** What is the expected number of times $a \neq 1$ is played by T?

- **Answer:** It can be computed as follows:

$$
\begin{aligned}
& E[n_T(a)] \\
= \; & E[n_T|\mathcal{A}^c]P(\mathcal{A}^c) + E[n_T|\mathcal{A}]P(\mathcal{A}) \\
\leq \; & 4\alpha\Delta^{-2}(a)\log T + T \times 2/T^2 \\
\leq \; & 4\alpha\Delta^{-2}(a)\log T + 2/T
\end{aligned}
$$

- The first regret bound is obtained as follows.

$$\begin{aligned}
R(T) &= E\Big[\sum_{a:\Delta(a)\neq 0} \Delta(a) n_T(a)\Big] \\
&= E\Big[\sum_{a:\Delta(a)\neq 0} \Delta(a) n_T(a)|\mathcal{A}^c\Big] P(\mathcal{A}^c) \\
&\qquad + E\Big[\sum_{a:\Delta(a)\neq 0} \Delta(a) n_T(a)|\mathcal{A}\Big] P(\mathcal{A}) \\
&\leq \sum_{a:\Delta(a)\neq 0} 4\alpha\Delta^{-1}(a)\log T + 1,
\end{aligned}$$

where the first term in the inequality follows from

$$E[n_T(a)|\mathcal{A}^c] \leq 4\alpha\Delta^{-2}(a)\log T,$$

and the second term follows from $P(\mathcal{A}) \leq 2/T$, $\Delta(a) \leq 1$, and $\sum_a n_T(a) \leq T$.

Second regret

- Divide the arms into two categories:

$$
\begin{aligned}
G_1 &= \left\{ i : \Delta(i) < \sqrt{\frac{m}{T} \log T} \right\}, \\
G_2 &= \left\{ i : \Delta(i) \geq \sqrt{\frac{m}{T} \log T} \right\}.
\end{aligned}
$$

- The total regret can be written as

$$
\sum_{i \in G_1} E[n_T(i)]\Delta_i + \sum_{i \in G_2} E[n_T(i)]\Delta_i.
$$

- We shall evaluate these two parts separately.

First part

- This part is bounded as follows:

$$\begin{aligned}
\sum_{i \in G_1} E[n_T(i)]\Delta(i) &\leq \sqrt{\frac{m}{T}\log T} \sum_{i \in G_1} E[n_T(i)] \\
&\leq \sqrt{\frac{m}{T}\log T} \times T \\
&= \sqrt{mT \log T}
\end{aligned}$$

Second part

- For $i \in G_2$, we have $\Delta(i) \geq \sqrt{\frac{m}{T} \log T}$ thus

$$\Delta(i)^{-1} \leq \sqrt{\frac{T}{m \log T}}.$$

- Therefore,

$$\begin{aligned}
\sum_{i \in G_2} E[n_T(i)]\Delta(i) &\leq \sum_{i \in G_1} (4\alpha\Delta^{-2}(a) \log T + 1/T)\Delta(i) \\
&= \sum_{i \in G_1} (4\alpha\Delta^{-1}(a) \log T + \Delta(i)/T) \\
&\leq 4\alpha\sqrt{mT \log T} + 1
\end{aligned}$$

Summary

- Summarizing, we obtain

$$R(T) \leq (1+\alpha)\sqrt{mT\log T} + 1 = O(\sqrt{mT\log T})$$

Brief summary and look ahead

- That's the end of the first part of the course, in which we focused on independent stochastic sequence (process).

- We now move to dependent random variables. Two classes of dependent processes will be studied:

  – The first is martingale process (discrete time only)

  – Markov process (discrete time as well as continuous time)

Martingale

- The results we discussed up to now assume that the stochastic sequence is independent. Many of our applications do not satisfy the independence condition.

- The most important and useful extension is to the class of martingale process.

- **Definition.** A stochastic sequence $\{X_n; n \geq 0\}$ is called a martingale if (i) $E[|X_n|] < \infty$, and (ii)

$$E[X_{n+1} \mid X_0, X_1, \ldots, X_n] = X_n, \qquad n = 1, \ldots.$$

Remarks

- From the definition and law of total expectation, we immediate have $E[X_{n+1}] = E[X_n]$, and thus, $E[X_n] = E[X_0]$ for all $n \geq 1$.

- A martingale is a generalization of a **fair game**. If we interpret X_n as a gambler's fortunate after the n-th gamble, then the definition states that his expected fortune after the $(n+1)$-st game is equal to his fortune after the n-th gamble no mater what may have previously occurred.

- Martingale was introduced by Paul Levy in 1930's, and the theory mainly due to Joseph Doob in 1950's.

- In the future, I am going to replace the event $\{X_0, \ldots, X_n\}$ by $\mathcal{F}_n$, representing the σ-algebra generated by $\{X_0, X_1, \ldots, X_n\}$, or simply all the information up to time n. For example, for $s < t$, then $\mathcal{F}_s$ is a subset of $\mathcal{F}_t$.

- Martingale is more general than it appears, because ...

- In the definition of martingale, why consider the case of "$=$"? When it is "$\geq$", corresponding to increasing sequence on average, it is called sub-martingale, while if "$\leq$", corresponding to decreasing sequence on average, then it is called "super-martingale".

Submartingale and supermartingale

- **Definition.** A stochastic sequence $\{X_n; n \geq 0\}$ is called a submartingale (supermartingale) if (i) $E[|X_n|] < \infty$, and (ii)

$$E[X_{n+1} \mid X_0, X_1, \ldots, X_n] \geq (\leq) X_n, \qquad n = 1, \ldots.$$

- If X_n is a submartingale, then $E[X_{n+1}] \geq E[X_n]$, amd of X_n is supermartingale, then $E[X_{n+1}] \leq E[X_n]$.

Example 1

- Let X_n be a sequence of indendent r.v.'s with $E[X_n] = \mu_n$, then

$$S_n = \sum_{i=1}^{n} (X_i - \mu_i)$$

and $S_0 = 0$ is a martingale process.

Example 2

- Let X_n be a sequence of i.i.d. r.v.'s with mean μ and variance σ^2, then

$$Y_n = \Big(\sum_{i=1}^{n} X_i - n\mu\Big)^2 - n\sigma^2$$

and $Y_0 = 0$ is a martingale process.

Example 3: Wald's martingale

- Let X_n be a sequence of i.i.d. r.v.'s with MGF $\phi(\theta) = E[e^{\theta X_1}]$. Then,

$$Y_n = (\phi(\theta))^{-n} \cdot e^{\theta \sum_{i-1}^{n} X_i}$$

with $Y_0 = 1$ is a martingale process.