CS6046: Multi-armed bandits
Homework - 1
Course Instructor : Prashanth L.A.
Due : Feb-12, 2018

Theory exercises

1. Suppose X_1, X_2 are σ_1 and σ_2-subgaussian random variables (r.v.s), respectively. (2+1 marks)

   (a) Show that $X_1 + X_2$ is $\sigma_1 + \sigma_2$-subgaussian.

   (b) If X_1 and X_2 are independent, then $X_1 + X_2$ is $\sqrt{\sigma_1^2 + \sigma_2^2}$-subgaussian.

2. True or False? (Justify your answer) (1+1+1.5+1.5 marks)

   (a) A r.v. X distributed as $N(\mu, \sigma^2)$ for some $\mu, \sigma > 0$ is subgaussian.

   (b) A r.v. X distributed as Unif$[5, 10]$ is subgaussian.

   (c) Consider a r.v. X satisfying $\mathbb{E}(\exp(\lambda X)) \leq \exp\left(\frac{\lambda^2\sigma^2}{2} + \lambda\mu\right)$ for any $\lambda \in \mathbb{R}$. Then, $EX = \mu$.

   (d) For the r.v. X as in the question above, Var$(X) = \sigma^2$.

3. For a K-armed stochastic bandit problem, with $m = n^{2/3}(\log n)^{1/3}$, show that the regret R_n of the explore-then-commit (ETC) algorithm satisfies

$$R_n \leq cn^{2/3}(K \log n)^{1/3},$$

   for some universal constant c. (5 marks)

4. Consider the following bandit algorithm:

   ϵ-greedy algorithm

   For $t = 1, 2, \dots, n$**, repeat**

   (1) Let i_t be the arm with the highest sample mean so far, i.e., $i_t = \underset{k=1,\dots,K}{\arg\max}\, \hat{\mu}_k(t-1)$, where $\hat{\mu}_k(t-1)$ is the average of rewards obtained from arm k upto time t.

   (2) With probability $1-\epsilon_t$, play arm i_t and with probability ϵ_t, play a random arm.

   For a two-armed bandit problem, show that the regret R_n incurred by the ϵ-greedy algorithm, with $\epsilon_t = 1/t^{1/3}$, satisfies

$$R_n \leq cn^{2/3}(\log n)^{1/3},$$

   for some universal constant c. (5 marks)

5. Consider the following game that proceeds over n rounds: In each round $t \in \{1, \dots, n\}$, you choose either to play or do nothing. If you do nothing, then your reward is $X_t = 0$. If you play, then your reward is $X_t = 1$ with probability p and $X_t = -1$ otherwise. You do not know p and we will assume it could take any value in $[0, 1]$.

   Answer the following: (1+1+2+2+2 marks)

(a) Formulate the game above as a stochastic bandit problem with horizon n.

(b) Write down the expression for the regret incurred by any algorithm $\mathcal{A}$.

(c) Describe an optimal way of choosing actions, i.e., the best algorithm, when p is known.

(d) For the unknown p case, apply ETC algorithm to the bandit problem formulated above and derive a bound on its regret.

(e) Does exploiting the fact that the reward is zero for "doing nothing" lead to an improved regret bound for ETC?

Simulation exercise

Consider a two-armed bandit problem, where each arm's distribution is Bernoulli. Consider the following three problem variants, with respective Bernoulli distribution parameters specified for each arm:

Problem	Arm 1	Arm 2
P1	0.9	0.6
P2	0.9	0.8
P3	0.55	0.45

Write a program (in your favorite language) to simulate each of the above bandit problems. In particular, do the following for each problem instance: (10 marks)

1. Choose the horizon n as 10000.
2. For each algorithm, repeat the experiment 100 times.
3. Store the number of times an algorithm plays the optimal arm, for each round $t = 1, \ldots, n$.
4. Store the regret in each round $m = 1, \ldots, n$.
5. Plot the percentage of optimal arm played and regret against the rounds $t = 1, \ldots, n$.
6. For each plot, add standard error bars.

Do the above for the following bandit algorithms:

- The explore-then-commit (ETC) algorithm with exploration parameter m chosen optimally so that the gap-dependent regret is minimum (this choice for m would require information about underlying gap).
- The ETC algorithm with a heuristic choice for exploration parameter m. Try different values for m and summarize your findings, say by tabulating regret for different m.

Interpret the numerical results and submit your conclusions. In particular, discuss the following: (2+3 marks)

1. Explain the results obtained for ETC with optimal m and correlate the results to the theoretical findings.

2. Explain the results obtained for ETC with a heuristic choice for m. In particular, how does ETC with a m that is far from the optimal, perform?

Here is what you have to submit:

Theory exercises (Q1-5): Hand-written (or typed) answer with concrete justification.

Simulation exercise: Include the following:

- Source code, preferably one that is readable with some comments;
- Plots/tabulated results in a document (or you could submit printouts of plots); and
- Discussion of the results - either hand-written or typed-up.