CS6046: Multi-armed bandits
Homework - 2
Course Instructor : Prashanth L.A.
Due : Mar-2, 2018

Theory exercises

1. For distributions P and Q of a continuous random variable, the KL-divergence is defined to be the integral:
$$D(P, Q) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx,$$
where p and q denote the densities of P and Q, respectively.

   Answer the following: (2 + 3 marks)

   (a) Suppose that P and Q correspond to univariate Gaussian distributions with means μ_1 and μ_2 and a common variance σ^2. Show that
$$D(P, Q) \leq \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}.$$

   (b) Suppose that P and Q correspond to bivariate Gaussian distributions with zero mean and covariance matrices $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 & \rho^2 \\ \rho^2 & 1 \end{bmatrix}$, where $\rho \in (0, 1)$. Calculate $D(P, Q)$, upper bound it using the simplest possible function of ρ.

2. A regret upper bound of $O(\log n)$ was shown for UCB algorithm, while a lower bound of $O(\sqrt{n})$ (ignoring the dependence on number of arms K) was also derived. How does one resolve the apparent contradiction between these two bounds? (1 mark)

3. Suppose there are two coins. The first is a fair coin, while the second one is biased (i.e., it falls heads with probability $\frac{3}{4}$). Suppose n sample outcomes $X_1, \ldots, X_n$ are generated using one of the two coins and an algorithm, say $\mathcal{A}$, uses these samples to identify the source coin. Let $\hat{I}_n$ denote the index that the algorithm $\mathcal{A}$ returns as its estimate of the source coin. Let P_v (resp. $P_{v'}$) denote the law of the observed samples $(X_1, \ldots, X_n)$, when the underlying source is the fair (resp. biased) coin.

   If $n < 4 \log 2$, then show that no algorithm can ensure
$$\max(P_v(\hat{I}_n = 2), P_{v'}(\hat{I}_n = 1)) \leq \frac{1}{8}.$$
   Hint: Use the high-probability Pinsker's inequality. (5 marks)

4. Consider a stochastic K-armed bandit problem where the rewards of arms' distributions are bounded within $[\frac{1}{2}, \frac{1+\epsilon}{2}]$ for some $\epsilon \in (0, 1)$. Construct a variant of UCB algorithm that uses the knowledge of ϵ. Derive gap-dependent and gap-independent regret bounds for this UCB variant and discuss their dependence on ϵ.

   Hint: Use Hoeffding's inequality. (4 marks)

5. For each of the two-armed bandit algorithms listed below, answer if they achieve sub-linear regret. An intuitive justification will suffice. Notation: For $i = 1, 2$, let $\hat{\mu}_i(t)$ denote the sample mean of arm i from the rewards seen up to time t. (1+2+2 marks)

(a) Play arm $I_t = \arg\max_{i=1,2} \hat{\mu}_i(t-1)$.

(b) Fix two sequences $A_1 = \{1, 2, 4, 8, 16, \ldots\}$ and $A_2 = \{3, 9, 27, 81, \ldots\}$. If $t \in A_i$, then play arm i, else play $I_t = \arg\max_{i=1,2} \hat{\mu}_i(t-1)$.

(c) Suppose the means are in the set $S = \{\mu, \mu - \epsilon\}$ and the bandit algorithm is aware of the set S. Consider the following algorithm: At time t, if $\max \hat{\mu}_i(t-1) > \mu - \epsilon/2$, then pull the arm that has the maximum sample mean. Otherwise pull both arms once.

6. Consider a two-armed bandit problem with Bernoulli reward distributions. Show that the UCB algorithm, which was described in the class, satisfies the following: (5 marks)

$$\mathbb{P}\left(\hat{R}_n > \Delta\left(1 + \frac{32}{\Delta^2}\log n\right)\right) \leq \frac{a}{(\log n)^b},$$

where $b > 0$ is a problem independent constant and $\hat{R}_n = \sum_{k=1}^{2} \Delta_k T_k(n)$, with $T_k(n)$ denoting the number of times arm k was pulled upto time t.

Hint: For any $\tau \in \mathbb{R}$, any integer $u > 1$ and any sub-optimal arm k, we have

$$\mathbb{P}\left(T_k(n) > u\right) \leq \sum_{t=u+1}^{n} \mathbb{P}\left(\hat{\mu}_{k,u} + \sqrt{\frac{8\log t}{u}} > \tau\right) + \sum_{s=1}^{n-u} \mathbb{P}\left(\hat{\mu}_{k^*,s} + \sqrt{\frac{8\log(u+s)}{u}} \leq \tau\right).$$

In the above, $\hat{\mu}_{k,u}$ is the sample mean of u i.i.d. samples from arm k's distribution and k^* is the optimal arm.

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 UCB algorithm, which plays each arm once initially and then, in each round t, plays the arm I_t as follows:

$$I_t = \operatorname*{arg\,max}_{k=1,2} \hat{\mu}_k(t-1) + \sqrt{\frac{2\log t}{T_k(t-1)}}.$$

- A variant of the UCB algorithm, say UCB', where the horizon n is used in the confidence width as follows: $I_t = \operatorname*{arg\,max}_{k=1,2} \hat{\mu}_k(t-1) + \sqrt{\frac{2\log n}{T_k(t-1)}}$.

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

1. Explain the results obtained for UCB on each problem instance and correlate the results to the theoretical findings.
2. Explain the results obtained for UCB' that uses the horizon n in the confidence width and compare its results to that of regular UCB. Is there any advantage in using UCB' over UCB, when n is known.

Here is what you have to submit:

Theory exercises (Q1-6): 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.