Stochastic Optimisation

Problem Sheet 2

$**$ **Please hand in solutions to question 4 on this sheet.** $**$

1. Consider a bandit with two independent arms, where the rewards from arm i are i.i.d. with a $N(i,i)$ distribution, $i = 1,2$. In other words, rewards from arm i are normally distributed with mean i and variance i, so that the second arm has the larger mean reward.

   Fix a time horizon T, and consider the heuristic which first plays each arm exactly n times, and subsequently plays the arm with the higher sample mean reward.

   (a) Let $\hat{\mu}_{1,n}$ and $\hat{\mu}_{2,n}$ denote the sample means of the first n plays of arms 1 and 2 respectively. Using the answer to Q6(b) from Problem Sheet 1, obtain an upper bound on $\mathbb{P}(\hat{\mu}_{1,n} \geq \hat{\mu}_{2,n})$.
   Hint. Let $X_i(t)$, $i = 1,2$ denote the reward observed on the t^{th} play of arm i. What can you say about the random variable $X_1(t) - X_2(t)$?

   (b) Using the answer to the last part, find an upper bound on the regret, $\mathcal{R}(T)$, of this heuristic. Optimize this upper bound over n, treating n as if it were a real number, and approximating quantities like $T - n$ by T, on the assumption that n is much smaller than T.

2. Consider a bandit with two independent Bernoulli arms, with parameters $\mu_1 > \mu_2$. Consider the following simple heuristic for this problem:

   - Play arm 1 in the first round.
   - If you obtained a reward of 1 in the previous round, play the same arm. Otherwise, switch to the other arm.

   Obtain an approximate expression for the regret of this heuristic up to some large time T.

   You do not need to be very precise in your calculations. I am looking for good intuition, and the correct scaling of the regret with T as T tends to infinity. Feel free to look up results you need, such as the means of well-known distributions. You do not need to calculate them from scratch.

3. Consider a bandit with two independent Bernoulli arms, with mean rewards $\mu_1 > \mu_2$. Define $\Delta = \mu_1 - \mu_2$. Let $N_i(t)$ denote the number of times that arm i has been played in the first t rounds, where $i \in \{1,2\}$ and $t \in \mathbb{N}$. Let $\hat{\mu}_{i,s}$ denote the empirical (or sample) mean reward obtained in the first s plays of arm i.

   Suppose a genie tells you the value of μ_1, the mean reward on arm 1 (but not that arm 1 is better). Then, the appropriate modification to the UCB(α) algorithm is as follows:

- Play arm 2 in the first round.
- At the end of round t, calculate the index of arm 2, defined as

$$\iota_2(t) = \hat{\mu}_{2,N_2(t)} + \sqrt{\frac{\alpha \log t}{2N_2(t)}}.$$

  The index of arm 1 is always μ_1, which is known (assuming we trust the genie).
- In round $t+1$, play the arm with the higher index, i.e., set $I(t+1) = 2$ if $\iota_2(t) \geq \mu_1$ and $I(t+1) = 1$ otherwise. (We have broken ties in favour of arm 2, but other ways of breaking ties are equally acceptable.)

We assume in the following that $\alpha > 1$.

(a) Show that, if arm 2 is played by the above algorithm in round $s+1$, i.e., $I(s+1) = 2$, then one of the following statements must be true:

$$N_2(s) < \frac{2\alpha \log s}{\Delta^2}, \tag{1}$$

$$\hat{\mu}_{2,N_2(s)} \geq \mu_2 + \sqrt{\frac{\alpha \log s}{2N_2(s)}}. \tag{2}$$

(b) Recall that $N_2(t) = \sum_{s=1}^{t} \mathbf{1}(I(s) = 2)$, where $\mathbf{1}(A)$ is the indicator of the event A. For an arbitrary positive integer u, and any $t \in \mathbb{N}$, explain why

$$N_2(t) \leq u + \sum_{s=u+1}^{t} \mathbf{1}(N_2(s-1) \geq u \text{ and } I(s) = 2).$$

A verbal explanation will suffice, but it should not leave out any essential details.

(c) Define $u = \lceil (2\alpha \log t)/\Delta^2 \rceil$. Using the answers to the last two parts, and relevant probability inequalities, show that

$$\mathbb{E}[N_2(t)] \leq u + \sum_{s=u+1}^{t} e^{-\alpha \log s}.$$

Use this to show that $\mathbb{E}[N_2(t)] \leq u + \frac{1}{\alpha - 1}$.

(d) Use the answer to the last part to show that the regret of this algorithm is bounded above as follows:

$$\mathcal{R}(T) \leq \frac{2\alpha \log T}{\Delta} + \frac{\alpha}{\alpha - 1}\Delta.$$

4. Consider a bandit with two independent Gaussian arms. Rewards on arm i constitute a sequence of iid $N(\mu_i, 1)$ random variables, i.e., normal with mean μ_i and variance 1.

(a) Let $\hat{\mu}_{i,n}$ denote the sample mean reward on arm i after n plays of this arm. Using a result from Homework 1, show that

$$\mathbb{P}\Big(\hat{\mu}_{i,n} > \mu_i + \sqrt{\frac{\alpha \log t}{2n}}\Big) \leq \exp\Big(-\frac{\alpha \log t}{4}\Big).$$

Express the last quantity as a power of t.

(b) Explain in a few sentences why the same bound holds for the probability of the event that $\hat{\mu}_{i,n} < \mu_i - \sqrt{\frac{\alpha \log t}{2n}}$.

(c) Replicate the analysis of the UCB algorithm to obtain a regret bound of the form $\mathcal{R}(T) \leq c_1 + c_2 \log T$, where c_1 and c_2 are constants that may depend on α, μ_1 and μ_2. Find explicit expressions for these constants.

The analysis will not work for all $\alpha > 1$. You will need α to be bigger than some other number. Find that number.

5. Let X and Y be Bernoulli random variables with parameters p and q respectively, where $p, q \in [0,1]$. Recall that the relative entropy or the KL-divergence of the Bern(q) distribution with respect to the Bern(p) distribution is defined as

$$K(q;p) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p},$$

with $x \log x$ defined to be zero if x is zero. Recall also that the total variation distance between these distributions, denoted $d_{TV}(\text{Bern}(q), \text{Bern}(p))$ is equal to $|p-q|$. Prove Pinsker's inequality, which states that

$$K(q;p) \geq 2\big(d_{TV}(\text{Bern}(q), \text{Bern}(p))\big)^2.$$

Hint. Fix p and show that the function $f(q) = K(q;p) - (q-p)^2$ is convex. Then show that $f(p) = 0$ and $f'(p) = 0$. For a convex function, the latter equality implies that p is a minimiser of the function f; you may use this fact without proof, but should look up a proof or convince yourself why it is true.

6. (optional hard problem)

Let X and Y be random variables with probability distributions P and Q respectively. Suppose P and Q have densities p and q with respect to a reference measure m; usually m is Lebesgue measure on the real line. Then, the relative entropy or KL-divergence of Q with respect to P is defined as

$$K(Q;P) = \int q(x) \log \frac{q(x)}{p(x)} dm(x).$$

If m is Lebesgue measure, we just write dx instead of $dm(x)$. In the following, you may use without proof the fact, which follows from Jensen's inequality, that $K(Q,P) \geq 0$ for all probability distributions P and Q.

The total variation distance between P and Q (which is symmetric) is defined as

$$d_{TV}(Q,P) = \sup_A |Q(A) - P(A)|,$$

where $P(A)$ and $Q(A)$ are the probabilities of the set A under the probability measures P and Q. In other words,

$$P(A) = \mathbb{P}(X \in A) = \int_A p(x) dm(x), \quad Q(A) = \mathbb{P}(Y \in A) = \int_A q(x) dm(x).$$

The supremum is taken over all (measurable) sets A.

(a) Let $A^* = \{x : q(x) \geq p(x)\}$. Explain why $d_{TV}(Q, P) = Q(A^*) - P(A^*)$. Maybe use Venn diagrams. You don't need to provide a formal proof.

(b) For any (measurable) set A such that $P(A)$ is not equal to zero or one, show that

$$K(Q, P) \geq K(Q(A), P(A)),$$

where the term on the right refers to the KL-divergence of a Bern($Q(A)$) distribution from a Bern$(P(A))$ distribution.

Hint. Split the integral in the definition of $K(Q, P)$ into an integral over A and an integral over A^c. Let Q^A and P^A respectively denote the probability distributions $Q(\cdot)/Q(A)$ and $P(\cdot)/P(A)$ on A. Express the integral over A as $K(Q^A; P^A)$ plus something. Do the same for A^c.

(c) Using the answers to the last two parts, prove the general version of Pinsker's inequality, which states that for any two probability measures P and Q (and not just for Bernoulli distributions),

$$K(Q; P) \geq 2\big(d_{TV}(Q, P)\big)^2.$$