# Stochastic Optimisation

# Solutions to Problem Sheet 2

---

1. (a) Question 6(b) from Problem Sheet 1 gives us a tail bound on the probability of sums of iid normal random variables. Here, we have a difference of sums of normal random variables. But we can easily put this in the form we want. Notice that

$$\hat{\mu}_{1,n} \geq \hat{\mu}_{2,n} \iff \sum_{t=1}^{n} X_1(t) \geq \sum_{t=1}^{n} X_2(t) \iff \sum_{t=1}^{n} (X_1(t) - X_2(t)) \geq 0,$$

where the $X_i(t)$ are defined as in the hint. Now, $X_1(t)$ and $X_2(t)$ are independent normal random variables, with mean and variance 1, and mean and variance 2, respectively. Hence, $X_1(t) - X_2(t) \sim N(-1, 3)$, and these differences are mutually independent for distinct values of $t$. Hence, by Q6(b) from Problem Sheet 1,

$$\mathbb{P}\Big(\sum_{t=1}^{n} (X_1(t) - X_2(t)) \geq 0\Big) \geq \exp\Big(-n\frac{1^2}{2 \times 3}\Big) = e^{-n/6}.$$

(b) On the event that $\hat{\mu}_{1,n} < \hat{\mu}_{2,n}$, arm 1 is not played after the exploratory phase, so it is played only $n$ times up to time $T$. On each play, it incurs a regret of $\mu_2 - \mu_1 = 1$. Hence, the regret up to time $T$ is $n$. On the event that $\hat{\mu}_{1,n} > \hat{\mu}_{2,n}$, arm 1 is played in every time step after the exploratory phase, so the regret up to time $T$ is $(T - 2n + n)(\mu_2 - \mu_1) = T - n$. Combining these possibilities, and using the answer to part (a), we get

$$\begin{aligned}\mathcal{R}(T) &= n\mathbb{P}(\hat{\mu}_{1,n} < \hat{\mu}_{2,n}) + (T-n)\mathbb{P}(\hat{\mu}_{1,n} > \hat{\mu}_{2,n}) \\ &\leq n(1 - e^{-n/6}) + (T-n)e^{-n/6} = Te^{-n/6} + n(1 - 2e^{-n/6}) \\ &\approx Te^{-n/6} + n =: f(n).\end{aligned}$$

Treating $n$ as if it were continuous and differentiating $f(n)$ above with respect to $n$, we get

$$\frac{df}{dn} \approx \frac{-T}{6}e^{-n/6} + 1, \quad \frac{d^2 f}{dn^2} \approx \frac{T}{6^2}e^{-n/6}.$$

The first derivative vanishes at $n = 6\log(T/6)$ and the second derivative is positive, so $f$ achieves a local (and in fact, global) minimum at this value of $n$. Substituting in this value of $n$, we conclude that

$$\begin{aligned}\mathcal{R}(T) &\leq T\exp(-\log(T/6)) + 6\log\frac{T}{6} \\ &= 6 + 6\log T - 6\log 6 = 6\log T + \text{ const.}\end{aligned}$$

2. Denote by Geom($p$) a geometric distribution with parameter $p$, and mean $1/p$. From the description of the heuristic, arm 1 is played a random number of times before switching to arm 2, which is played a random number of times before switching back to arm 1, and so on.

Define $T_i^1$ to be the number of times arm 1 is played consecutively during the $i^{\text{th}}$ run of plays of this arm; define $T_i^2$ similarly. Thus, arm 1 is played $T_1^1$ times in a row, then arm 2 is played $T_1^2$ times, arm 1 is played $T_2^1$ times, and so on. Observe that the random variables $T_i^1, i \in \mathbb{N}$ and $T_i^2, i \in \mathbb{N}$ are all mutually independent, that $T_i^1$ have a Geom$(1-\mu_1)$ distribution and $T_i^2$ have a Geom$(1-\mu_2)$ distribution. Hence, by the law of large numbers,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} T_i^k = \frac{1}{1-\mu_k}, \quad k = 1, 2.$$

Now, up to any time $T$, the number of complete runs for which each arm has been played differ by at most one. Hence, if we denote by $N_1(T)$ and $N_2(T)$ the number of times that arms 1 and 2 have been played up to time $T$, we see from the law of large numbers result that

$$\lim_{T\to\infty} \frac{N_1(T)}{N_2(T)} = \frac{1-\mu_2}{1-\mu_1}.$$

Combining this with the fact that $N_1(T) + N_2(T) = T$, we conclude that

$$\lim_{T\to\infty} \frac{N_2(T)}{T} = \frac{1-\mu_1}{1-\mu_1+1-\mu_2} = \frac{1-\mu_1}{2-\mu_1-\mu_2}.$$

Taking expectations, we get

$$\lim_{T\to\infty} \frac{\mathbb{E}[N_2(T)]}{T} = \frac{1-\mu_1}{2-\mu_1-\mu_2}.$$

(The interchange of limit and expectation is justified since $N_2(T)/T$ is a bounded random variable. I do not necessarily expect students to justify this step - I only asked for an intuitive explanation.)

As arm 1 is better, a regret of $\mu_1 - \mu_2$ is incurred each time arm 2 is played. Hence, the regret up to time $T$ is given by $\mathcal{R}(T) = (\mu_1 - \mu_2)\mathbb{E}[N_2(T)]$. It follows that

$$\lim_{T\to\infty} \frac{\mathcal{R}(T)}{T} = \frac{(1-\mu_1)(\mu_1-\mu_2)}{2-\mu_1-\mu_2},$$

i.e., the regret scales linearly in $T$.

3. (a) Suppose neither of the claimed statements is true. If (1) is false, then we must have

$$\frac{\alpha \log s}{2N_2(s)} \le \frac{\Delta^2}{4},$$

and so

$$\mu_2 + \sqrt{\frac{\alpha \log s}{2N_2(s)}} \le \mu_2 + \frac{\Delta}{2}.$$

Hence, if (2) is also false, then we must have

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

But $\mu_1 = \mu_2 + \Delta$, so the above implies that $\hat{\mu}_{2,N_2(s)} < \mu_1$, and so arm 2 cannot be played in round $s+1$.

(b) Given a sequence $I(s), s \in \mathbb{N}$, we can define $\tau(u) = \inf\{s : N_2(s) = u\}$. We define $\tau(u) = +\infty$ if the set over which the infimum is taken is empty, i.e., if $N_2(s) < u$ for all $s \in \mathbb{N}$. The inequality asserted in the question holds trivially in this case, so we assume from now on that $\tau < \infty$.

Now, we can see that

$$N_2(t) \le N_2(\tau) + \sum_{s=\tau+1}^{t} \mathbf{1}(I(s) = 2), \tag{1}$$

where the latter sum is defined to be zero if the set of valid indices is empty, i.e., if $\tau + 1 > t$. The inequality holds with equality if $\tau \le t$, and is obvious if $\tau > t$ since $N_2(\cdot)$ is a non-decreasing function. For the sum on the RHS above, notice that for each $s \ge \tau + 1$, it holds that $N_2(s-1) \ge u$, by the definition of $\tau$ and the fact that $N_2(\cdot)$ is non-decreasing. In other words, for $s \ge \tau + 1$, the indicator $\mathbf{1}(N_2(s-1) \ge u)$ takes the value 1, so that

$$\mathbf{1}(I(s) = 2) = 1) = \mathbf{1}(N_2(s-1) \ge u \text{ and } I(s) = 2), \quad \forall\, s \ge \tau + 1. \tag{2}$$

Substituting (2) in (1), and noting that $N_2(\tau) = u$, we get

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

The inequality asserted in the question follows by noticing that $\tau \ge u$, since $N_2(\cdot)$ can increase by at most 1 in each time step.

(c) Taking expectations on both sides of the inequality in part (b). We get

$$\begin{aligned}
\mathbb{E}[N_2(t)] &\le u + \mathbb{E}\Big[\sum_{s=u+1}^{t} \mathbf{1}(N_2(s-1) \ge u \text{ and } I(s) = 2)\Big] \\
&= u + \sum_{s=u+1}^{t} \mathbb{E}[\mathbf{1}(N_2(s-1) \ge u \text{ and } I(s) = 2)] \\
&= u + \sum_{s=u+1}^{t} \mathbb{P}(N_2(s-1) \ge u \text{ and } I(s) = 2),
\end{aligned} \tag{3}$$

where the first equality follows from the linearity of expectation.
Let $u$ be defined as in the question. Then, on the event that $N_2(s-1) \ge u$, we must have

$$N_2(s-1) \ge \frac{2\alpha \log t}{\Delta^2} \ge \frac{2\alpha \log(s-1)}{\Delta^2}$$

for all $s \le t$. It follows from part (a) that, in order for arm 2 to be played at time $s$ (i.e., for $I(s) = 2$), we must have

$$\hat{\mu}_{2,N_2(s-1)} \ge \mu_2 + \sqrt{\frac{\alpha \log(s-1)}{2N_2(s-1)}}.$$

Hence, we obtain for all $s \in \{u+1, \dots, t\}$ that

$$\mathbb{P}(N_2(s-1) \ge u \text{ and } I(s) = 2) \le \mathbb{P}\Big(\hat{\mu}_{2,N_2(s-1)} \ge \mu_2 + \sqrt{\frac{\alpha \log(s-1)}{2N_2(s-1)}}\Big). \tag{4}$$

We now bound the RHS above using Hoeffding's inequality. Since the rewards from plays of arm 2 are Bernoulli random variables, they take values in $[0, 1]$ (in fact, in $\{0, 1\}$), and we denoted their mean by $\mu_2$. Hence, we have by Hoeffding's inequality that

$$\begin{aligned}
\mathbb{P}\Big(\hat{\mu}_{2,N_2(s-1)} \ge \mu_2 + \sqrt{\frac{\alpha \log(s-1)}{2N_2(s-1)}}\Big) &\le \exp\Big(-2N_2(s-1)\frac{\alpha \log(s-1)}{2N_2(s-1)}\Big) \\
&= \exp(-\alpha \log(s-1)).
\end{aligned}$$

Combining this with (3) and (4), we get

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

Approximating the latter sum by

$$\int_u^t x^{-\alpha} dx \leq \int_u^\infty x^{-\alpha} dx \leq \frac{u^{-\alpha+1}}{\alpha - 1} \leq \frac{1}{\alpha - 1},$$

we conclude that $\mathbb{E}[N_2(t)] \leq u + \frac{1}{\alpha-1}$, as required. Notice that the last inequality in the displayed equation above holds because $u \geq 1$.

(d) We now use the fact that a regret of $\Delta$ is incurred each time that arm 2 is played, while no regret is incurred when arm 1 is played. Hence, the regret up to time $T$ is $\mathcal{R}(T) = \Delta \mathbb{E}[N_2(T)]$. Using the answer to part (d), we get the bound

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

Now, by the definition of $u$,

$$u \leq \frac{2\alpha \log T}{\Delta^2} + 1.$$

Combining the two displayed equations above,

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

which is what we were required to show.

4. (a) It follows from Q6(b) in Homework 1 that

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

since the variance of the Gaussian random variables is $\sigma^2 = 1$. Thus,

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

(b) To see that the inequality can be reversed, note that if $X_i$ are iid with a $N(\mu, \sigma^2)$ distribution, then $-X_i$ are iid with a $N(-\mu, \sigma^2)$ distribution. Thus,

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

satisfies the same bound.

(c) Assume without loss of generality (wlog) that $\mu_1 > \mu_2$, and let $\Delta = \mu_1 - \mu_2$. In the analysis of the UCB algorithm, we showed that one of the following three things must hold in order for the sub-optimal arm 2 to be played in time step $t+1$:

$$\hat{\mu}_{1,N_1(t)} \leq \mu_1 - \sqrt{\frac{\alpha \log t}{2N_1(t)}}, \tag{5}$$

$$\hat{\mu}_{2,N_2(t)} > \mu_2 + \sqrt{\frac{\alpha \log t}{2N_2(t)}}, \tag{6}$$

$$N_2(t) < \frac{2\alpha \log t}{\Delta^2}, \tag{7}$$

where $N_1(t)$ and $N_2(t)$ denote the number of times that arms 1 and 2 have been played in the first $t$ time steps.

Next, defining $u = \lceil (2\alpha \log T)/\Delta^2 \rceil$, we bounded the number of plays of arm 2 in the first $T$ rounds as follows:

$$N_2(T) \leq u + \sum_{t=u}^{T-1} \mathbf{1}(N_2(t) \geq u) \text{ and arm 2 is played in round } t+1). \tag{8}$$

By definition of $u$, for the last indicator to be 1, one of the events in (5) or (6) needs to occur. Hence, taking expectations in (8),

$$\mathbb{E}[N_2(T)] \leq u + \sum_{t=u}^{T-1} \mathbb{P}\Big(\hat{\mu}_{1,N_1(t)} \leq \mu_1 - \sqrt{\frac{\alpha \log t}{2N_1(t)}}\Big) + \mathbb{P}\Big(\hat{\mu}_{2,N_2(t)} > \mu_2 + \sqrt{\frac{\alpha \log t}{2N_2(t)}}\Big).$$

Substituting the bounds on these probabilities from the first part of the question, we get

$$\mathbb{E}[N_2(T)] \leq u + \sum_{t=u}^{T-1} 2t^{-\alpha/4}.$$

Approximating the last sum by an integral,

$$\mathbb{E}[N_2(T)] \leq u + \int_{u-1}^{\infty} 2t^{-\alpha/4} dt.$$

If $\alpha > 4$, then the integral above converges, and we get

$$\mathbb{E}[N_2(T)] \leq u + \frac{2(u-1)^{-\alpha/4}}{\frac{\alpha}{4} - 1} \leq u + \frac{8}{\alpha - 4},$$

using the fact that $u \geq 2$. Substituting for $u$,

$$\mathbb{E}[N_2(T)] \leq \frac{2\alpha \log T}{\Delta^2} + 1 + \frac{8}{\alpha - 4} = \frac{2\alpha \log T}{\Delta^2} + \frac{\alpha + 4}{\alpha - 4}.$$

Finally, a regret of $\Delta = \mu_1 - \mu_2$ is incurred every time arm 2 is played. Hence, the regret up to time $T$ is bounded as follows:

$$\mathcal{R}(T) = \Delta \mathbb{E}[N_2(T)] \leq c_1 + c_2 \log T,$$

where

$$c_1 = \frac{\alpha + 4}{\alpha - 4}\Delta, \quad c_2 = \frac{2\alpha}{\Delta}.$$

5. Following the hint (which has a typo), we want to show that $f(q)$ defined as $K(q; p) - 2(q-p)^2$ is a convex function of $q$. Writing it out in full,

$$f(q = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p} - 2(q-p)^2.$$

We now differentiate it twice. We get

$$f'(q) = 1 + \log \frac{q}{p} - 1 - \log \frac{1-q}{1-p} - 4(q-p), \quad f''(q) = \frac{1}{q} + \frac{1}{1-q} - 4 = \frac{1}{q(1-q)} - 4.$$

Now, for $q \in (0,1)$, the quantity $q(1-q)$ achieves its maximum value of $\frac{1}{4}$ at $q = \frac{1}{2}$. Hence, $\frac{1}{q(1-q)} \geq 4$ for all $q \in (0,1)$, which implies that $f''(q) \geq 0$ for all $q \in (0,1)$. This shows that the function $f$ is convex on $(0,1)$.

Moreover, it is easy to see that $f'(p) = 0$, which means that $f$ attains its global minimum over $(0,1)$ at $q = p$. We also have $f(p) = 0$, which implies that $f(q) \geq f(p) = 0$ for all $q \in (0,1)$, i.e.,

$$K(q;p) \geq 2(q-p)^2 = 2d_{TV}(p,q)^2,$$

for all $q \in (0,1)$. For completeness, we also need to check that the inequality holds for $q = 0$ and $q = 1$. There are two ways to do this. The simpler is to notice that $K(q;p) - 2(q-p)^2$ is continuous on $[0,1]$; hence, the convexity proved on $(0,1)$ extends to $[0,1]$ and we are done.

Alternatively, you could check by hand that the claimed inequality $K(q,p) - 2(q-p)^2 \geq 0$ holds at $q = 0$ and $q = 1$ as well. If $p = 0$ and $q = 0$, or $p = 1$ and $q = 1$, then this expression is zero. We will check it for $q = 0$, $p \neq 0$; the case $q = 1$, $p \neq 1$ is similar. If $q = 0$, then $K(q,p) = -\log(1-p)$, so we need to show that $-\log(1-p) - 2p^2 \geq 0$ for all $p \in (0,1]$. Calling it $g(p)$, we notice that

$$g'(p) = \frac{1}{1-p} - 4p = \frac{1 - 4p(1-p)}{1-p} \geq 0 \; \forall \, p \in (0,1).$$

Hence, $g$ is a non-decreasing function on $(0,1)$. As $g(0) = 0$ and $g$ is continuous at $0$, it follows that $g(p) \geq 0$ for all $p \in (0,1)$. This also holds for $p = 1$, since $g(1) = +\infty$. This completes the proof.