Stochastic Optimisation - Problem Sheet 2

Dom Hutchinson

November 12, 2020

Question 1.
Consider a bandit with two independent arms, where the rewards from arm i are i.i.d. with a Normal(i, i) distribution, $i \in \{1, 2\}$. 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.

Question 1. (a)
Let $\hat{\mu}_{1,n}$ and $\hat{\mu}_{2,n}$ denote the samples means of the first n plays of arm 1 and 2 respectively. Using the answer to `Problem Sheet 1 Q6 (b)`, obtain an upper bound on

$$\mathbb{P}(\hat{\mu}_{1,n} \geq \hat{\mu}_{2,n})$$

Answer 1. (a)
Let random variables $X_i(t) \sin \text{Normal}(i, i)$ for $i \in \{1, 2\}$ model the reward received from arm i at time t and $\hat{\mu}_{i,n}$ denote the sample mean from the first n times arm i is played.

Define random variable $X(t) := X_1(t) - X_2(t)$ which has distribution Normal$(1-2, 1+2) =$ Normal(-1,3) and $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} X(t) = \hat{\mu}_{1,n} - \hat{\mu}_{2,n}$. Thus

$$\begin{array}{rcl}
\mathbb{P}(\hat{\mu}_{1,n} \geq \hat{\mu}_{2,n}) & = & \mathbb{P}(\hat{\mu}_{1,n} - \hat{\mu}_{2,n} \geq 0) \\
& = & \mathbb{P}(\hat{\mu}_n \geq 0) \\
& = & \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq 0\right) \\
& = & \mathbb{P}\left(\sum_{i=1}^{n} X_i \geq 0\right)
\end{array}$$

`Problem Sheet 1 Q6 b)` states that for IID random variables $X_1, X_2, \dots$ with distribution Normal(μ, σ^2) and for $\gamma > \mu$, the following bound exists.

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq n\gamma\right) \leq \exp\left(-n\frac{(\gamma-\mu)^2}{2\sigma^2}\right)$$

By defining $\gamma = 0$ and noting that $\gamma > \mu = -1$ we can apply this result to the inequality above.

$$\begin{array}{rrcl}
& \mathbb{P}\left(\sum_{i=1}^{n} X_i \geq 0\right) & \leq & \exp\left(-n\frac{(0-(-1))^2}{2\cdot 3}\right) \\
& & = & \exp\left(-\frac{n}{6}\right) \\
\implies & \mathbb{P}(\hat{\mu}_{1,n} \geq \hat{\mu}_{2,n}) & \leq & \exp\left(-\frac{n}{6}\right)
\end{array}$$

Question 1. (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.

Answer 1. (b)
The algorithm here can be considered to have two stages: learning; and post-learning. During the learning phase we are guaranteed to play the sub-optimal arm n times, but we play the same arm throughout the post-learning phase and thus regret only increases if the wrong arm is made (ie if $\hat{\mu}_{1,n} \geq \hat{\mu}_{2,n}$).
Note that the post-learning phase consists of $(T-2n)$ rounds and the loss incurred from playing the sub-optimal arm is $2-1=1$.

$$\begin{aligned} \mathcal{R}_T &= \underbrace{1\cdot n}_{\text{learning}} + \underbrace{1\cdot(T-2n)\mathbb{P}(\hat{\mu}_{1,n} \geq \hat{\mu}_{2,n})}_{\text{wrong choice made}} \\ &= n+(T-2n)\mathbb{P}(\hat{\mu}_{1,n} \geq \hat{\mu}_{2,n}) \\ &\leq n+(T-2n)e^{-n/6} \text{ by } \texttt{1. (a)} \end{aligned}$$

Since $n \ll T$ we can approximate $(T-2n) \simeq T$ giving

$$\mathcal{R}_T \leq n + Te^{-n/6}$$

We want to find the n which minimises this expression.

$$\begin{aligned} \frac{\partial}{\partial n}\left(n+Te^{-n/6}\right) &= 1-\frac{1}{6}Te^{-n/6} \\ \frac{\partial^2}{\partial n^2}\left(n+Te^{-n/6}\right) &= \frac{1}{36}Te^{-n/6} \geq 0 \ \forall\ T,n\in\mathbb{N} \\ \text{Setting}\quad \frac{\partial}{\partial n}\left(n+Te^{-n/6}\right) &= 0 \\ \implies\quad 1-\frac{1}{6}Te^{-\hat{n}/6} &= 0 \\ \implies\quad e^{-\hat{n}/6} &= \frac{6}{T} \\ \implies\quad \frac{-\hat{n}}{6} &= \ln(6)-\ln(T) \\ \implies\quad \hat{n} &= 6[\ln(T)-\ln(6)] \\ &= 6\ln\left(\frac{T}{6}\right) \end{aligned}$$

Since the second derivative is positive, this $\hat{n}$ minimises the bound on regret. Giving

$$\mathcal{R}_T \leq 6\ln\left(\frac{T}{6}\right) + T\exp\left(-\ln\left(\frac{T}{6}\right)\right) = 6\ln\left(\frac{T}{6}\right) - \frac{T^2}{6}$$

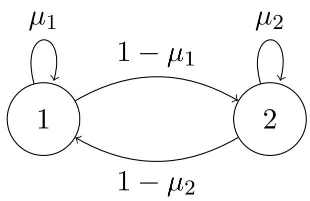
Question 2.
Consider a bandit with two independent Bernoulli arms, with parameters $\mu_1 > \mu_2$. Consider the following simple heuristic for this problem:

- 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.

Answer 2.
Let $\mu_1 > \mu_2$ and note that this algorithm can be summarised by the following automata



and transition matrix

$$P = \begin{pmatrix} \mu_1 & 1-\mu_1 \\ 1-\mu_2 & \mu_2 \end{pmatrix}$$

A stationary distribution π of the transition matrix P gives the proportion of times each arm is played in the long run. Let π be a stationary distribution for P

$$\begin{array}{rrcl} & \pi & = & \pi P \\ \implies & (\pi_1,\ \pi_2) & = & (\pi_1,\ \pi_2)\begin{pmatrix} \mu_1 & 1-\mu_1 \\ 1-\mu_2 & \mu_2 \end{pmatrix} \\ \implies & (\pi_1,\ \pi_2) & = & \big(\mu_1\pi_1 + \pi_2(1-\mu_2),\ \pi_1(1-\mu_1) + \pi_2\mu_2\big) \\ \implies & \pi_1 & = & \mu_1\pi_1 + \pi_2(1-\mu_2) \\ \implies & \pi_1(1-\mu_1) & = & \pi_2(1-\mu_2) \\ \implies & \pi_1 & = & \pi_2\dfrac{1-\mu_2}{1-\mu_1} \end{array}$$

By the definition of a stationary distribution $\pi_1 + \pi_2 = 1 \implies \pi_2 = 1 - \pi_1$. Substituting this result back in we can get explicit results for π_1 and π_2.

$$\begin{array}{rrcl} & \pi_1 & = & (1-\pi_1)\dfrac{1-\mu_2}{1-\mu_1} \\ \implies & \pi_1\left(1 + \dfrac{1-\mu_2}{1-\mu_1}\right) & = & \dfrac{1-\mu_2}{1-\mu_1} \\ \implies & \pi_1\left(\dfrac{2-\mu_1-\mu_2}{1-\mu_1}\right) & = & \dfrac{1-\mu_2}{1-\mu_1} \\ \implies & \pi_1 & = & \dfrac{1-\mu_2}{2-\mu_1-\mu_2} \\ & \pi_2 & = & 1-\pi_1 \\ \implies & \pi_2 & = & 1 - \dfrac{1-\mu_2}{2-\mu_1-\mu_2} \\ & & = & \dfrac{1-\mu_1}{2-\mu_1-\mu_2} \end{array}$$

We can now create an approximate expression for the regret $\mathcal{R}_T$ over time horizon T.

$$\begin{array}{rcl} \mathcal{R}_T & = & (\mu_1-\mu_2)\mathbb{E}(\text{times arm 2 played}) \\ & = & (\mu_1-\mu_2)[T\mathbb{P}(\text{arm 2 played})] \\ & = & (\mu_1-\mu_2)T\pi_2 \\ & = & T(\mu_1-\mu_2)\dfrac{1-\mu_1}{2-\mu_1-\mu_2} \end{array}$$

Question 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 $UCB(\alpha)$ algorithm is as follows:

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

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

  The index of arm 1 is always μ_1, which is known.
- In round $t+1$, play the arm with the greater index, breaking ties in favour of arm 2.

Assume that $\alpha > 1$

Question 3. (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.

i). $N_2(s) < \dfrac{2\alpha \ln(s)}{\Delta^2}$

ii). $\hat{\mu}_{2,N_2(s)} \geq \mu_2 + \sqrt{\dfrac{\alpha \ln(s)}{2N_2(s)}}$

Answer 3. (a)
This is a proof by contradiction.
Suppose $I(s+1) = 2$ but that none of the statements above hold. Then

$$\begin{array}{rrcll}
 & \hat{\mu}_{2,N_2(s)} - \sqrt{\dfrac{\alpha \ln(s)}{2N_2(s)}} & < & \mu_2 & \text{by not ii)} \\
 & & = & \mu_1 - \Delta & \text{by def. of } \Delta \\
 & & \leq & \mu_1 - \sqrt{\dfrac{2\alpha \ln(s)}{N_2(s)}} & \text{by not i)} \\
\implies & \hat{\mu}_{2,N_2(s)} + \sqrt{\dfrac{2\alpha \ln(s)}{N_2(s)}} - \sqrt{\dfrac{\alpha \ln(s)}{2N_2(s)}} & < & \mu_1 & \\
\implies & \hat{\mu}_{2,N_2(s)} + \left(\sqrt{2} - \dfrac{1}{\sqrt{2}}\right)\sqrt{\dfrac{\alpha \ln(s)}{N_2(s)}} & < & \mu_1 & \\
\implies & \hat{\mu}_{2,N_2(s)} + \sqrt{\dfrac{\alpha \ln(s)}{2N_2(s)}} & < & \mu_1 & \\
\implies & i_2(s) & < & \mu_1 &
\end{array}$$

This means $I(s+1) = 1$, which is a contradiction. Thus at least one of i) or ii) must be true.

Question 3. (b)

Recall that $N_2(t)=\sum_{s=1}^t \mathbb{1}\{I(S)=2\}$. For an arbitrary positive integer u and any $t\in\mathbb{N}$ explain why

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

Answer 3. (b)
Fix $t,u\in\mathbb{N}$. We have two possibilities

Case 1 $N_2(t)\leq u$ (i.e. Arm two has not been played u times yet). The result trivially holds in this case.

Case 2 $\exists\ s\in[1,t]$ such that $N(s)>u$ (i.e. Arm two has been played at least u times).
Let s^* denote the smallest such s. Then it must be true that $N(s^*-1)=u$ and $s^*\geq u+1$.
Hence

$$\begin{array}{rcll} N(t) &=& \sum_{s=1}^{s^*-1} I(s)+\sum_{s=s^*}^t I(s) & \\ &=& N(s^*-1)+\sum_{s=s^*}^t I(s)\,\underbrace{\mathbb{1}\{N(s-1)\geq u\}}_{\text{true for all in sum}} & \\ &\leq& u+\sum_{s=u+1}^t \mathbb{1}\{N(s-1)\geq u\} & \text{since } s^*\geq u+1 \end{array}$$

Thus the result holds in all cases.

Question 3. (c)
Define $u=\lceil(2\alpha\ln(t))/\Delta^2\rceil$. Using the answers to parts `(a)` and `(b)`, and relevant probability inequalities, show that

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

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

Answer 3. (c)
We have

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

Taking expectations of both sides

$$\begin{array}{rcl} \mathbb{E}[N_2(t)] &\leq& u+\sum_{s=u+1}^t \mathbb{P}\left(\{N_2(s-1)\geq u\}\text{ and }\{I(s)=2\}\right) \\ &\leq& u+\sum_{s=u}^{t-1} \mathbb{P}\left(\{N_2(s)\geq u\}\text{ and }\{I(s+1)=2\}\right) \end{array}$$

If $N_2(s)\geq u$ and $I(s+1)=2$ then

$$\hat{\mu}_{2,N_2(s)}\geq\mu_2+\sqrt{\frac{\alpha\ln(s)}{2N_2(s)}}\text{ by a)}$$

Thus

$$\mathbb{E}(N_2(t)) \leq u + \sum_{s=u}^{t-1} \mathbb{P}\left(\hat{\mu}_{2,N_2(s)} \geq \mu_2 + \sqrt{\frac{\alpha \ln(s)}{2N_2(s)}}\right) \quad (1)$$

Let $X_1, \dots, X_{N_2}$ be the random variables for each time arm 2 was played. Consider

$$\begin{array}{rcll}
\mathbb{P}\left(\hat{\mu}_{2,N_2(s)} \geq \mu_2 + \sqrt{\frac{\alpha \ln(s)}{2N_2(s)}}\right) & = & \mathbb{P}\left(\frac{1}{N_2}\sum_{i=1}^{N_2} X_i \geq \mu_2 + \sqrt{\frac{\alpha \ln(s)}{2N_2(s)}}\right) & \\
& = & \mathbb{P}\left(\sum_{i=1}^{N_2}(X_i - \mu_2) \geq N_2\sqrt{\frac{\alpha \ln(s)}{2N_2(s)}}\right) & \\
& \leq & \exp\left(-2 \cdot N_2 \cdot \frac{\alpha \ln(s)}{2N_2(s)}\right) & \text{by Hoeffding's Ineq.} \\
& = & \exp(-\alpha \ln(s)) & \\
\implies \qquad \mathbb{E}[N_2(t)] & \leq & u + \sum_{s=u+1}^{t} e^{-\alpha \ln(s)} & \text{by (1)}
\end{array}$$

Further

$$\begin{array}{rcl}
\mathbb{E}[N_2(t)] & \leq & u + \sum_{s=u+1}^{t} e^{-\alpha \ln(s)} \\
& = & u + \sum_{s=u+1}^{t} s^{-\alpha} \\
& \leq & u + \int_u^{\infty} s^{-\alpha} ds \quad \text{since } \alpha > 1 \\
& = & u + \left[\frac{s^{-\alpha+1}}{-\alpha+1}\right]_u^{\infty} \\
& = & u - \frac{u^{-\alpha+1}}{-\alpha+1} \\
& = & u + \frac{u^{-\alpha+1}}{\alpha+1}
\end{array}$$

By the definition of u, $u > 1$ thus $u^{-\alpha+1} < 1$ since $\alpha > 1$. Giving us

$$\mathbb{E}[N_2(t)] \leq u + \frac{1}{\alpha - 1}$$

Question 3. (d)
Use the answer to `(c)` to show that the regret of this algorithm is bounded above as

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

Answer 3. (d)

$$\begin{array}{rcll}
\mathcal{R}(T) & := & \Delta \mathbb{E}[N_2(t)] & \\
& \leq & \Delta\left(u + \frac{1}{\alpha - 1}\right) & \text{by 3. (c)} \\
& \leq & \Delta\left(\frac{2\alpha \ln(T)}{\Delta^2} + 1 + \frac{1}{\alpha - 1}\right) & \text{by def. of } u \\
& = & \frac{2\alpha \ln(T)}{\Delta} + \Delta\left(1 + \frac{1}{\alpha - 1}\right) & \\
& = & \frac{2\alpha \ln(T)}{\Delta} + \frac{\Delta\alpha}{\alpha - 1} &
\end{array}$$

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

Question 4. (a)
Let $\hat{\mu}_{i,n}$ denote the sample mean reward on arm i after n plays of this arms. Using a resulting from Homework 1, show that

$$\mathbb{P}\left(\hat{\mu}_{i,n} < \mu_i + \sqrt{\frac{\alpha \ln(t}{2n}}\right) \leq \exp\left(-\frac{\alpha \ln(t)}{4}\right)$$

Express the last quantity as power of t.

Answer 4. (a)
Let $\hat{\mu}_{i,n}$ be the sample mean reward on arm i after n plays of that arms.
From *Problem Sheet 1 6b)*, for $X_i \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$ and $\gamma > \mu_i$ we have that

$$\mathbb{P}(\hat{\mu} > \gamma) = \mathbb{P}\left(\sum_{i=1}^n X_i > n\gamma\right) \leq \exp\left(-n\frac{(\gamma - \mu)^2}{2\sigma^2}\right)$$

Applying this result to this scenario

$$\mathbb{P}(\hat{\mu}_{i,n} > \gamma) \leq \exp\left(-n\frac{(\gamma - \mu_i)^2}{2}\right)$$

By defining $\gamma = \mu_i + \sqrt{\frac{\alpha \ln(t)}{2n}}$ with $\alpha > 0$.
Note that $\gamma > \mu_i$ so we can use the above inequality

$$\begin{array}{rcl}\mathbb{P}\left(\hat{\mu}_{i,n} > \mu_i + \sqrt{\frac{\alpha \ln(t)}{2n}}\right) & \leq & \exp\left(-\frac{n}{2} \cdot \frac{\alpha \ln(t)}{2n}\right) \\ & = & \exp\left(-\frac{\alpha \ln(t)}{4}\right) \\ & = & t^{-\alpha/4}\end{array}$$

Question 4. (b)
Explain in a few sentences why the same bound holds the probability of the event that $\hat{\mu}_{i,n} < \mu_i - \sqrt{\frac{\alpha \ln(t)}{2n}}$

Answer 4. (b)
The result from *Problem Sheet 1 6b)* is derived from the Chernoff Bound for IID random variables when $\left\{\sum X_i \geq nc\right\}$ and considers $\inf_{\theta > 0} e^{-n\theta c}(\mathbb{E}[e^{\theta X}])^n$. The result requires $c > \mu_i$ in order to fulfil the restriction on the infimum (i.e. $\theta > 0$).

To derive a similar result to *Question 4. (a)* for the event $\left\{\hat{\mu}_{i,n} < \mu_i - \sqrt{\frac{\alpha \ln(t)}{2n}}\right\}$ we define $c = \mu_i - \sqrt{\frac{\alpha \ln(t)}{2n}}$, meaning $c < \mu_i$ and thus $\theta < 0$, for the θ in the infimum.

The Chernoff Bound for this complementary event considers the infimum of the same expression, except with the restriction that $\theta < 0$ (rather than $\theta > 0$). Given our definition of c and the resulting value of θ, the same value for the infimum is found. Meaning the same bound is derived for both the event and its compliment.

Question 4. (c)
Replicate the analysis of the UCB algorithm to obtain a regret bound of the form $\mathcal{R}(T) \leq c_1 + c_2 \ln(T)$ 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.

Answer 4. (c)
Assume WLOG $\mu_1 > \mu_2$ and define $\Delta = \mu_1 - \mu_2$. Let $N_2(t)$ be the number of times arm 2 is played in the first t steps. Define $u_t = \left\lceil \frac{2\alpha \ln(t)}{\Delta^2} \right\rceil$. We have

$$N_2(t) \leq u + \sum_{s=u-1}^{t} \mathbb{1}\left(\{N_2(s-1) \geq u_t\} \text{ and } \{I(s) = j\}\right)$$

Taking expectations of both side we get

$$\mathbb{E}[N_2(t)] \leq u_t + \sum_{s=u_t}^{t-1} \mathbb{P}\left(\{N_2(s-1) \geq u_t\} \text{ and } \{I(s) = j\}\right)$$

By considering the two cases where the sub-optimal arm is played: $\hat{\mu}_1$ is significantly lower than μ_1; or $\hat{\mu}_2$ is significantly higher than μ_2.

$$\begin{aligned}
\mathbb{E}[N_2(t)] &\leq u_t + \sum_{s=u_t}^{t-1} \left[\mathbb{P}\left(\hat{\mu}_{1,N_1(s)} \leq \mu_1 - \sqrt{\frac{\alpha \ln(s)}{2N_2(s)}}\right) + \mathbb{P}\left(\hat{\mu}_{2,N_2(s)} > \mu_2 - \sqrt{\frac{\alpha \ln(s)}{2N_2(s)}}\right)\right] \\
&\leq u_t + \sum_{s=u_t}^{t-1} 2t^{-\alpha/4} \text{ by } \textit{Question 4. (a)} \\
&\leq u + \int_{u_t-1}^{\infty} 2t^{-\alpha/4} dt \\
&= u_t + 2\left[\frac{t^{-\frac{\alpha}{4}+1}}{1-\frac{\alpha}{4}}\right]_{u_t-1}^{\infty} \\
&= u_t - \frac{2(u_t-1)^{-\frac{\alpha}{4}+1}}{-\frac{\alpha}{4}+1} \\
&\leq u_t + \frac{2}{\frac{\alpha}{4}-1} \\
&= u_t + \frac{8}{\alpha-4} \\
&\leq \frac{2\alpha \ln(t)}{\Delta^2} + 1 + \frac{8}{\alpha-4} \text{ by def. of } u_t \\
&= \frac{2\alpha \ln(t)}{\Delta^2} + \frac{\alpha+4}{\alpha-4}
\end{aligned}$$

In this scenario $\mathcal{R}(T) = \Delta \mathbb{E}[N_2(T)]$. Thus, using the results above

$$\mathcal{R}(T) \leq \frac{2\alpha \ln(T)}{\Delta} + \Delta \frac{\alpha+4}{\alpha-4}$$

This requires $\alpha > 4$.

Question 5.
Let $X \sim \text{Bern}(p)$ and $Y \sim \text{Bern}(q)$ with $p, q \in [0, 1]$. Recall that the KL-Divergence of a Bern(q) distribution wrt a Bern(p) distribution is defined as

$$KL(q;p) := q \ln\left(\frac{q}{p}\right) + (1-q)\ln\left(\frac{1-q}{1-p}\right)$$

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

$$KL(q;p) \geq 2\big(\text{Bern}(q), \text{Bern}(p)\big)^2$$

Answer 5.
Fix the value of p and define the consider the following function

$$\begin{array}{rcl} f(q) & := & KL(q;p) - 2(q-p)^2 \\ & = & q\ln\left(\frac{q}{p}\right) + (1-q)\ln\left(\frac{1-q}{1-p}\right) - 2(q-p)^2 \text{ by def of } KL \end{array}$$

I will show that this function is convex

$$\begin{array}{rcl} f'(q) & = & \ln\left(\frac{q}{p}\right) + q\cdot\frac{1/p}{q/p} - \ln\left(\frac{1-q}{1-p}\right) + (1-q)\frac{-1/(1-p)}{(1-q)/(1-p)} - 4(q-p) \\ & = & \ln\left(\frac{q}{p}\right) - \ln\left(\frac{1-q}{1-p}\right) - 4(q-p) \\ f''(q) & = & \frac{1/p}{q/p} - \frac{-1/(1-p)}{(1-q)/(1-p)} - 4 \\ & = & \frac{1}{q} + \frac{1}{1-q} - 4 \\ & = & \frac{1}{q(1-q)} - 4 \end{array}$$

Note $\min_{q\in(0,1)} \frac{1}{q(1-q)} = \frac{1}{\frac{1}{2}(1-\frac{1}{2})} = 4$. Thus $\frac{1}{q(1-q)} \geq 4\ \forall\ q \in (0,1)$. Further, $f''(q) \geq 0$ for the whole domain $q \in (0,1)$, meaning $f(q)$ is convex.
Now note that $f'(p) = 0$ (ie the minimum occurs when $q = p$) and that $f(p) = 0$ (ie $\min_{q\in(0,1)} f(q) = 0$), this means $f(q) \geq f(p) = 0\ \forall\ q \in (0,1)$.
Using this inequality we can finally derive *Pinsker's Inequality* for Bernoulli random variables

$$\begin{array}{rrcl} & f(q) & \geq & 0 \\ \implies & K(q;p) - 2(q-p)^2 & \geq & 0 \\ \implies & K(q;p) & \geq & 2(q-p)^2 = 2|q-p|^2 \\ & & = & 2d_{TV}(\text{Bern}(q), \text{Bern}(p))^2 \end{array}$$

This is *Pinsker's Inequality* for Bernoulli random variables.