MLT Homework 8

Ana Borovac
Bas Haver

November 12, 2018

Question 1

Let $\psi(\lambda) = \frac{\lambda^2}{2}$. The Legendre-Fenchel transform of ψ is given by

$$\psi^*(\epsilon) = \sup_{\lambda \in \mathbb{R}} \lambda\epsilon - \psi(\lambda).$$

Subquestion 1.1

$\psi^*(\epsilon) = \frac{\epsilon^2}{2}$.

Solution

Let define $\psi_1(\lambda)$:

$$\psi_1(\lambda) = \lambda\epsilon - \frac{\lambda^2}{2}$$

Furthermore:

$$\psi_1'(\lambda) = \epsilon - \lambda$$

Since $\psi_1(\lambda)$ is a parabola it has only one extreme; particularly it has just a maximum (negative sign before λ^2). So, the maximum is reached at:

$$\lambda = \epsilon \quad \Rightarrow \quad \psi_1(\epsilon) = \epsilon \cdot \epsilon - \frac{\epsilon^2}{2} = \frac{\epsilon^2}{2}$$

We can conclude:

$$\psi^*(\epsilon) = \psi_1(\epsilon) = \frac{\epsilon^2}{2}$$

Subquestion 1.2

$(\psi^*)^{-1}(z) = \pm\sqrt{2z}$.

Solution

From previous point we know:

$$\psi^*(\epsilon) = \frac{\epsilon^2}{2}$$

It follows:

$$\begin{aligned} z &= \frac{\epsilon^2}{2} \\ 2z &= \epsilon^2 \\ \epsilon &= \pm\sqrt{2z} \end{aligned}$$

So:

$$(\psi^*)^{-1}(z) = \pm\sqrt{2z}$$

Question 2

The Blooper Reel

Subquestion 2.1

Deterministic fails for Adversarial Bandits *Show that any deterministic algorithm (UCB included) has linear regret in the adversarial bandit setting. Hint: you can use the argument on the top of page 23.*

Solution

We were a bit confused by this question, since it states that we need to find a linear regret, which we did not find. Instead we found that it can be bounded between two linear functions, but it does not necessarily is linear itself.
As a counter-example of the linearity we have that for $n = 0$ no regret has been obtained. Therefore linearity only holds when $R_m + R_n = R_{n+m}$ for all n, m. But when we choose $n = 4$, four arms and choose to play on arms 1,2,3 and 4 succesively, we find $R_1 = 1$, $R_3 = 3$, $R_4 = 3$, so linearity dos not hold. It does however hold that it remains between our bounds $\frac{n}{K}$ and n for K the amount of arms.
We use the hint on the top of page 23, which tells us that for a deterministic forecaster, we can use the following sequence of losses:

$$\begin{aligned} &\text{if } I_t = 1, \quad \text{then } l_{2,t} = 0 \quad \text{and} \quad l_{i,t} = 1 \quad \text{for all } i \neq 2; \\ &\text{if } I_t \neq 1, \quad \text{then } l_{1,t} = 0 \quad \text{and} \quad l_{i,t} = 1 \quad \text{for all } i \neq 1 \end{aligned}$$

Of course this is just a worst-case feedback. For every choice of arm, we will get a loss of 1, which result in a regret that is as high as possible. Since it does the

trick, we will use it.
This sequence of losses now implies the following for the regret:

$$
\begin{aligned}
R_n &= \sum_{t=1}^{n} l_{t,I_t} - \min_k \sum_{t=1}^{n} l_n^k \\
&= n - \min_k \sum_{t=1}^{n} l_n^k
\end{aligned}
$$

But now for any choice of arm, with at least two arms, $\min_k \sum_{t=1}^{n} l_n^k$ is at most $\frac{n}{K}$. Therefore the regret is bounded from below by $\frac{n}{2}$ and since the loss function is nonnegative, the regret is also bounded from above by n.

Subquestion 2.2

C onsider a K-armed stochastic bandit model with unit-variance Gaussian rewards with means $\mu_1, \ldots, \mu_K$. In round t the learner chooses arm $I_t \in [K]$ and recieves reward $X_t \sim \mathcal{N}(\mu_{I_t}, 1)$, where μ_i is the (unknown) reward of arm i. Now let's fix the following algorithm, which is inspired by Empirical Risk Minimisation:

(a) First, pull every arm once (that is $T_i = t$ for $t \leq K$).

(b) Then after each number $t \geq K$ of rounds, from the empirical estimates

$$
\hat{\mu}_i(t) = \frac{\sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}} X_s}{\sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}}}
$$

and play $I_{t+1} = \arg\max_i \hat{\mu}_i(t)$.

For $K = 2$, show that this algorithm has pseudo-regret

$$
\bar{R} = n\mu^* - \mathbb{E}[\sum_{t=1}^{n} \mu_{I_t}]
$$

that is *linear* in n.
Hint: you can use the following outline. Assume $\mu_1 > \mu_2$. Pick some threshold $\epsilon > 0$ (which you will optimise in a later step).

- Argue that with constant probability (independent of n) the reward drawn from the best arm in the first phase is below $\mu_2 - \epsilon$.
- Bound the probability that for a single time step t we have $\hat{\mu}_2(t) < \mu_2 - \epsilon$ using Chernoff's bound.
- Use the union bound to bound the probaility that $\exists t \geq 2 : \hat{\mu}_2(t) < \mu_2 - \epsilon$.
- Now pick ϵ large enough so that the previous probability bound is non-trivial (i.e. is ¡1).

Conclude that with some small probability the sample from the best arm is very low, and the samples from the second-best arm are all typical, so the algorithm keeps pulling arm 2 pnly. Deduce that the pseudo-regret is hence linear in n.

Solution

Question 3

We consider an adversarial bandit model with K^2 arms indexed by $i \in [K]$ and $j \in [K]$. For each arm (i, j), the loss at time t is $a_t^i + b_t^j$, where $a_t^i \in [0, 1]$ and $b_t^j \in [0, 1]$ are chosen by the adversary before the start of the interaction. Then each round the learner picks an arm $(I_t, J_t) \in [K]^2$ and observes $a_t^{I_t}$ and $b_t^{J_t}$ separately (and incurs their sum as the loss).

Subquestion 3.1

Consider running a single instance of EXP3 on all K^2 arms (with loss range $[0, 2]$). Show that the expected pseudo-regret compared to the best arm (i^*, j^*) is bounded by

$$\bar{R}_n \leq 2\sqrt{2nK^2 \ln(K^2)}$$

Solution

Below we used the following facts:

- $\min x + y = \min x + \min y;\ \ x, y \geq 0$
- Linearity of expected value.
- Theorem from the lectures: $\bar{R}_n \leq \sqrt{2nK \ln K}$, where K is the number of arms; in our case we have K^2 arms.

$$\begin{aligned}
\bar{R}_n &= \mathbb{E}_{I_1,\ldots,I_n,J_1,\ldots,J_n}\left\{\sum_{t=1}^{n} a_t^{I_t} + b_t^{J_t}\right\} - \min_k \sum_{t=1}^{n} a_t^k + b_t^k \\
&= \left(\mathbb{E}_{I_1,\ldots,I_n}\left\{\sum_{t=1}^{n} a_t^{I_t}\right\} - \min_k \sum_{t=1}^{n} a_t^k\right) + \left(\mathbb{E}_{J_1,\ldots,J_n}\left\{\sum_{t=1}^{n} b_t^{J_t}\right\} - \min_k \sum_{t=1}^{n} b_t^k\right) \\
&\leq \sqrt{2nK^2 \ln K^2} + \sqrt{2nK^2 \ln K^2} \\
&= 2\sqrt{2nK^2 \ln K^2}
\end{aligned}$$

Subquestion 3.2

Now we will use the a_t^i and b_t^j observations separately. Consider running two K-arm instances of EXP3, one with $i \to a_t^i$ as the loss and one with $j \to b_t^j$ as the loss. Have the first algorithm control I_t and the second J_t. Show that the overall expected pseudo-regret is bounded by

$$\bar{R}_n \leq 2\sqrt{2nK \ln K}.$$

Solution

We do similar as before, just that now we choose arm in the set of K arms and not K^2.

$$\begin{aligned}
\bar{R}_n &= \mathbb{E}_{I_1,\ldots,I_n,J_1,\ldots,J_n}\left\{\sum_{t=1}^{n} a_t^{I_t} + b_t^{J_t}\right\} - \min_k \sum_{t=1}^{n} a_t^k + b_t^k \\
&= \left(\mathbb{E}_{I_1,\ldots,I_n}\left\{\sum_{t=1}^{n} a_t^{I_t}\right\} - \min_k \sum_{t=1}^{n} a_t^k\right) + \left(\mathbb{E}_{J_1,\ldots,J_n}\left\{\sum_{t=1}^{n} b_t^{J_t}\right\} - \min_k \sum_{t=1}^{n} b_t^k\right) \\
&\leq \sqrt{2nK\ln K} + \sqrt{2nK\ln K} \\
&= 2\sqrt{2nK\ln K}
\end{aligned}$$