MLT Homework set 8

Due 13 April 2023 before 10:00
via `elo.mastermath.nl`.
Rules about homework (allowed sources, collaboration, etc.)
are on the ELO as well.

1 Background

Definition 1 (Adversarial Bandit Setting)**.**

Protocol:

Adversary hides $\ell_t^k \in [0,1]$ for all $t \leq T, k \leq K$.

For $t = 1, 2, \ldots, T$

- Learner picks arm I_t (typically by sampling $I_t \sim \boldsymbol{w}_t$)
- Learner observes and incurs *loss* $\ell_t^{I_t}$

Objective: Expected regret w.r.t. best expert after T rounds:

$$\bar{R}_T = \mathop{\mathbb{E}}_{I_1,\ldots,I_T} \left\{ \sum_{t=1}^T \ell_t^{I_t} \right\} - \min_k \sum_{t=1}^T \ell_t^k$$

Definition 2 (Stochastic Bandit Setting)**.**

Protocol:

Environment: distributions $(\nu_1, \ldots, \nu_K)$ of arm rewards

For $t = 1, 2, \ldots, T$

- Learner picks arm I_t
- Learner observes and receives *reward* $X_t \sim \nu_{I_t}$

Objective: Pseudo-regret w.r.t. best arm after T rounds:

$$\bar{R}_T \;=\; T \max_k \mathop{\mathbb{E}}_{X \sim \nu_k}[X] - \mathop{\mathbb{E}}_{I_1, \ldots, I_T}\left\{\sum_{t=1}^{T} X_t\right\}$$

1.1 Confidence Intervals

You can use the following bound, known as Chernoff's bound for Gaussians. For $X_1, \ldots, X_t$ i.i.d. Gaussian random variables with mean μ and unit variance, the empirical estimate $\hat{\mu}_t \;=\; \frac{1}{t}\sum_{i=1}^{t} X_i$ satisfies

$$\mathbb{P}\left(\hat{\mu}_t - \mu \geq \epsilon\right) \;\leq\; e^{-t\frac{\epsilon^2}{2}} \qquad \text{and} \qquad \mathbb{P}\left(\hat{\mu}_t - \mu \leq -\epsilon\right) \;\leq\; e^{-t\frac{\epsilon^2}{2}}.$$

2 Exercises

1. [4 pt] **Importance Weighted Estimation with Shift**

   Fix $m \in \mathbb{R}$. Consider the estimator $\hat{\boldsymbol{\ell}}_t$ defined by

   $$\hat{\ell}_t^k \;=\; m + \frac{\ell_t^{I_t} - m}{w_t^{I_t}} \mathbf{1}_{I_t = k}$$

   (a) Show that $\hat{\boldsymbol{\ell}}$ is unbiased, i.e. $\mathbb{E}_{I_t \sim \boldsymbol{w}_t}[\hat{\boldsymbol{\ell}}_t] = \boldsymbol{\ell}_t$.

   (b) Recall that $\boldsymbol{\ell}_t$ in $[0,1]^K$ and $\boldsymbol{w}_t \in \triangle_K$. We used the estimator with $m = 0$ in the lecture. There, we used that the range of possible values of $\hat{\ell}_t^k$ is $[0, \infty)$. For $m = \frac{1}{2}$ determine the range of possible values of $\hat{\ell}_t^k$. Also determine the range for $m = 1$. You can assume that all entries of $\boldsymbol{w}_t$ are non-zero.

   Bigger Picture A variety of unbiased estimators can be defined, but we have to be careful in bounding the dot/mix loss relationship. This is where range assumptions come in.

2. [4 pt] **Adversarial Semi-bandit**

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 $b_t^i \in [0, 1]$ 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).

   (a) Consider running a single instance of EXP3 on all K^2 arms (with loss range $[0, 2]$). Show that the expected regret compared to the best arm (i^*, j^*) is bounded by
   $$\bar{R}_T \;\leq\; 2\sqrt{2TK^2 \ln(K^2)}.$$

   (b) Now we will use the observations a_t^i and b_t^j separately. Consider running two K-arm instances of EXP3, one with $i \mapsto a_t^i$ as the loss and one with $j \mapsto b_t^j$ as the loss. Have the first algorithm control I_t and the second J_t. Show that the overall expected regret is bounded by
   $$\bar{R}_T \;\leq\; 2\sqrt{2TK \ln K}.$$

Bigger Picture We see that we win a factor $\sqrt{K}$ by taking the structure of the observations into account. There are many interesting intermediate observation models where we see some interpolation between the full information regret $\sqrt{T \ln K}$ and the bandit regret $\sqrt{KT}$.

3. [4 pt] **ERM fails for Stochastic Bandits**

Consider 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 receives reward $X_t \sim \mathcal{N}(\mu_{I_t}, 1)$, where μ_i is the (unknown) mean 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, $I_t = t$ for $t \leq K$).

   (b) Then after each number $t \geq K$ of rounds, form the empirical estimates
   $$\hat{\mu}_i(t) \;=\; \frac{\sum_{s=1}^{t} \mathbf{1}_{\{I_s = i\}} X_s}{\sum_{s=1}^{t} \mathbf{1}_{\{I_s = i\}}}$$
   and play $I_{t+1} = \arg\max_i \hat{\mu}_i(t)$.

For $K = 2$, show that this algorithm has pseudo-regret
$$\bar{R}_T \;=\; T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t}\right]$$
that is *linear* in T.

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 T) 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 probability 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 first 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 forever. Deduce that the pseudo-regret is hence linear in T.

3 Training Exercises

4. [0 pt] **Deterministic fails for Adversarial Bandits** Show that any *deterministic* algorithm (UCB included) has linear regret in the adversarial bandit setting. Hint: consider the adversary that always gives maximal loss to the arm the learner picks.

5. [0 pt] **UCB with Ties**

   Consider a Gaussian K-armed bandit model where M of the K arms are tied for best arm ($1 < M < K$). Pick the correct answer below (only one is correct) and provide the argument:

   - The pseudo-regret can be linear in T. *Construct an example where linear regret happens.*
   - The pseudo-regret is $O(\ln T)$. *Sketch the steps of the UCB analysis, indicating where ties require care.*