Sequential Learning: Homework #1

Too long a homework? This homework statement comes with 4 exercises. This is perhaps a bit too much? I will decide based on what you collectively submit. Maybe solving 3 or 2 and 1/2 exercises will already appear a good performance.

What I care about. I care about well-written proofs: with sufficient details, with calculations worked out and leading to pleasant and readable bounds. I favor quality of the writing over the quantity of questions answered. I give bonus points for elegant solutions.

Format of your submission, deadline. I expect to receive a single PDF file, with answers either handwritten and neatly scanned (as I do for my weekly lecture notes) or typed in LaTeX. The PDF file must be named `HW1-YourName.pdf`. E.g., I would submit a PDF file named `HW1-Stoltz.pdf`. Deadline is Sunday, February 28, at 8pm. This is a strict deadline: submitting after this deadline will negatively impact your grade, with the impact depending on the delay.

Beware: Typos. Most likely the statement comes with typos. This is part of the job. Try to correct them on your own!

Exercise 1: Adversarial sparse losses

The aim of this exercise is to study what happens when both a non-negativity and a sparsity assumptions are made on the vectors of losses picked by the opponent.

More formally, we consider the setting of linear losses, with N components, where at most s components are positive while the other components are null. The parameter $s \in \{1, \ldots, N\}$ is fixed throughout the game but is unknown to the statistician. The online protocol is the following.

Protocol: For all rounds $t = 1, 2, \ldots$,

- The statistician picks a convex combination $(p_{j,t})_{1 \leqslant j \leqslant N}$ while the environment simultaneously picks a loss vector $(\ell_{j,t})_{1 \leqslant j \leqslant N} \in [0, +\infty)^N$, with at most s non-null components;
- The choices are publicly revealed.

The statistician aims to control the regret

$$R_T = \sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t} \, \ell_{j,t} - \min_{1 \leqslant i \leqslant N} \sum_{t=1}^{T} \ell_{i,t} \, .$$

The question is:
What is the optimal order of magnitude of the regret under the non-negativity and sparsity assumptions?

Upper bound on the regret

1. Recall first how, under the non-negativity assumption, i.e., assuming that the losses $\ell_{j,t}$ all lie in $[0, M]$, we could prove the bound

$$R_T \leqslant 13 M \ln N + 2 \sqrt{M \min_{j=1,\ldots,N} \sum_{t=1}^{T} \ell_{j,t} \ln N} \, ,$$

referred to as an "improvement for small cumulative losses."

More precisely, recall the algorithm at hand and the sketch of its performance bound above. (Answer in a about 10–15 lines only.)

2. Deduce a $13 M \ln N + 2M \sqrt{(Ts \ln N)/N}$ bound on the regret of this algorithm under the sparsity assumption.

Does the algorithm need to know s to ensure this bound? Explain and comment.

Lower bound on the regret

Consider the joint distribution over $\{0,1\}^N$ defined as the law of a random vector $\boldsymbol{L} = (L_1, \ldots, L_N)$ drawn in two steps. First, we pick s components uniformly at random among $\{1, \ldots, N\}$; we call them $K_1, \ldots, K_s$. Then, the components not picked ($k \neq K_j$ for all j) are associated with zero losses, $L_k = 0$. The losses L_k for picked components $K_1, \ldots, K_s$ are drawn according to a Bernoulli distribution with parameter $1/2$. The loss vector $\boldsymbol{L} \in [0,1]^N$ thus generated is indeed s–sparse and non-negative.

We fix an algorithm for the statistician, consider an i.i.d. sequence $\boldsymbol{L}_1, \boldsymbol{L}_2, \ldots$ of random vectors thus generated, and study the corresponding regret

$$R_T = \sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t} \, L_{j,t} - \min_{1 \leqslant i \leqslant N} \sum_{t=1}^{T} L_{i,t} \, .$$

3. Show that the expectation of the regret can be written as

$$\mathbb{E}\left[\frac{R_T}{\sqrt{T}}\right] = \mathbb{E}\left[\max_{1 \leqslant i \leqslant N} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t^{(i)}\right]$$

where the $\bigl(X_t^{(1)}, \ldots, X_t^{(N)}\bigr)$ are i.i.d. centered random vectors taking values in $[-1,1]^N$, with covariance matrix denoted by Γ: please give a closed-form definition of the $X_t^{(i)}$ based on the $L_{i,t}$.

4. Explain why

$$\mathbb{E}\left[\max_{1 \leqslant i \leqslant N} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t^{(i)}\right] \longrightarrow \mathbb{E}\Bigl[\max_{1 \leqslant i \leqslant N} Z_i\Bigr]$$

where $(Z_1, \ldots, Z_N)$ follows the normal distribution $\mathcal{N}(\mathbf{0}, \Gamma)$, i.e., the centered normal distribution with covariance matrix Γ.

5. Consider the Gaussian random vector $(W_1, \ldots, W_N)$ with i.i.d. components W_i with distribution $\mathcal{N}\bigl(0, \mathrm{Var}(Z_1)\bigr)$. Show that Slepian's lemma (stated below) is applicable and that it entails

$$\mathbb{E}\Bigl[\max_{1 \leqslant i \leqslant N} Z_i\Bigr] \geqslant \mathbb{E}\Bigl[\max_{1 \leqslant i \leqslant N} W_i\Bigr]$$

6. Conclude to an asymptotic lower bound of the order of $\sqrt{(Ts \ln N)/N}$; state it carefully and rigorously.

Slepian's lemma (1962): Let $(Z_1, \ldots, Z_N)$ and $(W_1, \ldots, W_N)$ be two centered Gaussian random vectors in $\mathbb{R}^N$. If

$$\forall i \in \{1, \ldots, N\}^2, \qquad \mathbb{E}\bigl[Z_i^2\bigr] = \mathbb{E}\bigl[W_i^2\bigr]$$

and

$$\forall (i,j) \in \{1, \ldots, N\}^2, \qquad i \neq j \quad \Rightarrow \quad \mathbb{E}\bigl[Z_i Z_j\bigr] \leqslant \mathbb{E}\bigl[W_i W_j\bigr]\,,$$

then for all $t \in \mathbb{R}$,

$$\mathbb{P}\Bigl\{\max_{1 \leqslant i \leqslant N} Z_i > t\Bigr\} \geqslant \mathbb{P}\Bigl\{\max_{1 \leqslant i \leqslant N} W_i > t\Bigr\}\,.$$

Exercise 2: Approachability of a closed convex set $\mathcal{C}$

A statistician plays against an opponent; the statistician wants her average loss to approach (converge to) a given closed convex set $\mathcal{C} \subseteq \mathbb{R}^d$, while the opponent aims to prevent this convergence. Formally, the statistician and the opponent have respective action sets $\{1,\ldots,N\}$ and $\{1,\ldots,M\}$ and a loss function

$$\ell : \{1,\ldots,N\} \times \{1,\ldots,M\} \longrightarrow \mathbb{R}^d$$

is given and known by both players. The learning protocol is the following.

Protocol: For all rounds $t = 1, 2, \ldots$,
- the statistician and the opponent simultaneously and independently pick actions $I_t \in \{1,\ldots,N\}$ and $J_t \in \{1,\ldots,M\}$, possibly at random, according to distributions denoted by $\boldsymbol{p}_t$ and $\boldsymbol{q}_t$, respectively;
- the statistician suffers the loss $\ell(I_t, J_t)$;
- both players observe I_t and J_t.

Respective aims: The statistician wants to ensure that

$$\frac{1}{T}\sum_{t=1}^{T} \ell(I_t,J_t) \longrightarrow \mathcal{C} \quad \text{a.s.}, \qquad \text{that is,} \qquad \min_{c\in\mathcal{C}} \left\| c - \frac{1}{T}\sum_{t=1}^{T} \ell(I_t,J_t) \right\| \longrightarrow 0 \quad \text{a.s.}, \tag{1}$$

while the opponent wants to prevent this convergence, i.e., ensure that

$$\mathbb{P}\left\{ \limsup_{T\to\infty} \min_{c\in\mathcal{C}} \left\| c - \frac{1}{T}\sum_{t=1}^{T} \ell(I_t,J_t) \right\| > 0 \right\} > 0 \tag{2}$$

A set $\mathcal{C}$ such that the statistician has a strategy ensuring (1) is called approachable by the statistician. Otherwise, in the case (2), we say that it is not approachable.

Blackwell's condition: We denote by $\mathcal{P}_N$ and $\mathcal{P}_M$ the sets of probability distributions over $\{1,\ldots,N\}$ and $\{1,\ldots,M\}$, respectively. We (bi-)linearly extend ℓ by defining, for all $\boldsymbol{p} = (p_1,\ldots,p_N) \in \mathcal{P}_N$, all $j \in \{1,\ldots,M\}$, and all $\boldsymbol{q} = (q_1,\ldots,q_M) \in \mathcal{P}_M$,

$$\ell(\boldsymbol{p}, j) = \sum_{i=1}^{N} p_i\, \ell(i,j) \qquad \text{and} \qquad \ell(\boldsymbol{p},\boldsymbol{q}) = \sum_{i=1}^{N}\sum_{j=1}^{M} p_i\, q_j\, \ell(i,j)$$

We consider Blackwell's condition:

$$\forall \boldsymbol{q} \in \mathcal{P}_M, \quad \exists \boldsymbol{p} \in \mathcal{P}_N \ \big| \quad \ell(\boldsymbol{p},\boldsymbol{q}) \in \mathcal{C}\,,$$

and will show that it is a necessary and sufficient condition for approachability.

Necessity

1. Show that when Blackwell's condition does not hold, then not only is $\mathcal{C}$ not approachable by the statistician, but we even have that there exists $\gamma > 0$ such that for all strategies of the statistician,

$$\liminf_{T\to\infty} \min_{c\in\mathcal{C}} \left\| c - \frac{1}{T}\sum_{t=1}^{T} \ell(I_t,J_t) \right\| \geqslant \gamma \quad \text{a.s.}$$

2. Rephrase the previous result in terms of approachability of some set for the opponent.

Hints: For Question 1, show that there exists $\boldsymbol{q}_0 \in \mathcal{P}_M$ such that

$$\min_{\boldsymbol{p}\in\mathcal{P}_N} \min_{c\in\mathcal{C}} \big\| c - \ell(\boldsymbol{p},\boldsymbol{q}) \big\| > 0$$

and carefully also explain why, for all strategies of the statistician and of the opponent,

$$\left\| \frac{1}{T}\sum_{t=1}^{T} \ell(I_t,J_t) - \frac{1}{T}\sum_{t=1}^{T} \ell(\boldsymbol{p}_t,\boldsymbol{q}_t) \right\| \longrightarrow 0 \quad \text{a.s.}$$

Sufficiency

We henceforth assume that Blackwell's condition holds and consider the following strategy for the statistician, where we denote by $\langle \cdot\,, \cdot \rangle$ the inner product in $\mathbb{R}^d$.

Strategy for the statistician:

- Play $\boldsymbol{p}_1 = (1/N, \dots, 1/N)$
- For $t \geqslant 2$,
  - Compute the current average loss $\overline{m}_{t-1} = \dfrac{1}{t-1} \displaystyle\sum_{s=1}^{t-1} \ell(\boldsymbol{p}_s, J_s)$
  - Project it onto $\mathcal{C}$ as $\overline{c}_{t-1} = \Pi_{\mathcal{C}}\big(\overline{m}_{t-1}\big)$
  - Pick $p_t \in \operatorname*{arg\,min}_{\boldsymbol{p} \in \mathcal{P}_N} \max_{\boldsymbol{q} \in \mathcal{P}_M} \big\langle \overline{m}_{t-1} - \overline{c}_{t-1},\, \ell(\boldsymbol{p}, \boldsymbol{q}) \big\rangle$
  - Draw I_t at random according to $\boldsymbol{p}_t$

We then analyze this strategy; we denote $L = \max_{i,j} \big|\ell(i,j)\big|$.

3. Recall thanks to a picture (no formal proof required) why for all $t \geqslant 2$,

$$\forall c \in \mathcal{C}, \qquad \big\langle \overline{m}_{t-1} - \overline{c}_{t-1},\, c - \overline{c}_{t-1} \big\rangle \leqslant 0$$

4. Deduce from this and from Sion's lemma (the fact that under some conditions, an inf sup equals a sup inf) that

$$\forall \boldsymbol{q} \in \mathcal{P}_M, \qquad \big\langle \overline{m}_{t-1} - \overline{c}_{t-1},\, \ell(\boldsymbol{p}_t, \boldsymbol{q}) - \overline{c}_{t-1} \big\rangle \leqslant 0$$

5. Show that the distance to $\mathcal{C}$ at round t, namely, $d_t = \inf_{c \in \mathcal{C}} \big\| \overline{m}_t - c \big\|$, satisfies, for all $t \geqslant 1$,

$$d_{t+1}^2 \leqslant \left(1 - \frac{2}{t+1}\right) d_t^2 + \frac{4L^2}{(t+1)^2}$$

Hint: consider $c = \overline{c}_t$ and upper bound d_{t+1} by $\big\| \overline{m}_{t+1} - \overline{c}_t \big\|$. Then "decompose" $\overline{m}_{t+1}$ into $\overline{m}_t$ and $\ell(\boldsymbol{p}_{t+1}, J_{t+1})$.

6. Prove that for all $T \geqslant 1$,

$$\min_{c \in \mathcal{C}} \left\| c - \frac{1}{T} \sum_{t=1}^{T} \ell(\boldsymbol{p}_t, J_t) \right\| \leqslant \frac{2L}{\sqrt{T}}\,.$$

7. Conclude. (Yes, there is a simple but final step to deal with.)

Exercise 3: Budgeted prediction

Ante-scriptum: we assume in this problem that the horizon T, the budget $m \in \{1, \ldots, T-1\}$ and the loss range $[0,1]$ are known.

We study a case of prediction of individual sequences when the statistician does not get to see the N–vector of losses at the end of each round, unless she asks for it, which she can only do m times during the T rounds. More formally, the prediction protocol is the following: for all rounds $t = 1, 2, \ldots, T$,

- the statistician picks a distribution $\boldsymbol{p}_t$ over $\{1, \ldots, N\}$ and draws a component I_t at random according to $\boldsymbol{p}_t$;
- simultaneously, the opponent picks a loss vector $(\ell_{1,t}, \ldots, \ell_{N,t}) \in [0,1]^N$;
- the statistician suffers the loss $\ell_{I_t,t}$ but does not observe it;
- the statistician decides whether she wants to observe the loss vector (and in this case, she observes all of its components); she may only do so if she performed less than $m-1$ observations so far;
- the opponent observes I_t and $\boldsymbol{p}_t$.

We will construct step by step a strategy for the statistician. We fix a confidence level $\delta \in (0,1)$.

Random observations and estimated losses

The statistician will make random decisions about observations. More precisely, she will set $\varepsilon \in (0,1)$, consider a sequence $Z_1, Z_2, \ldots, Z_T$ of i.i.d. random variables, distributed according to a Bernoulli distribution with parameter ε, and observe the t–th loss vector if and only if $Z_t = 1$.

To abide by the budget constraint, she wants to pick ε such that

$$\mathbb{P}\big\{Z_1 + Z_2 + \ldots + Z_T \leqslant m\big\} \geqslant 1 - \delta\,.$$

1. Show that $\varepsilon = m/T - (1/T)\sqrt{m/\delta}$ is a suitable choice when $\delta \geqslant 1/m$. You may use Chebychev's inequality to that end.

We define

$$\widehat{\ell}_{j,t} = \frac{\ell_{j,t}}{\varepsilon} Z_t\,.$$

2. Show that for a well-chosen filtration $\mathcal{F} = (\mathcal{F}_t)_{t \geqslant 0}$ to determine, we have

$$\mathbb{E}\Big[\widehat{\ell}_{j,t} \,\big|\, \mathcal{F}_{t-1}\Big] = \ell_{j,t}\,.$$

Strategy based on these estimated losses

3. Indicate a strategy that never asks for more than m observations and ensures that with probability at least $1-\delta$,

$$\sum_{t=1}^{T} \sum_{i=1}^{N} p_{i,t} \widehat{\ell}_{i,t} - \min_{j=1,\ldots,N} \sum_{t=1}^{T} \widehat{\ell}_{j,t} \leqslant 2\sqrt{\frac{1}{\varepsilon} \min_{j=1,\ldots,N} \sum_{t=1}^{T} \widehat{\ell}_{j,t} \ln N} + \frac{13}{\varepsilon} \ln N$$

4. Deduce from this a strategy that never asks for more than m observations and whose pseudo-regret

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{I_t,t}\right] - \min_{j=1,\ldots,N} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{j,t}\right]$$

is bounded by something of the order of $T\sqrt{(\ln N)/m}$. Please state a precise bound.

Hint: Of course you will take expectations in the bound of Question 3. But there are issues to take care of, like tuning δ and ε.

Note: one can show that $T\sqrt{(\ln N)/m}$ is the optimal order of magnitude of the pseudo-regret; when $m = T$, we are back to the classical case (same setting, same bound) discussed in our series of lectures.

Exercise 4: The polynomially weighted average forecaster

We consider the "vanilla" setting of linear losses, with N components: for all rounds $t = 1, 2, \ldots$,

- The statistician picks a convex combination $(p_{j,t})_{1 \leqslant j \leqslant N}$ while the environment simultaneously picks a loss vector $(\ell_{j,t})_{1 \leqslant j \leqslant N}$;
- The choices are publicly revealed.

The statistician aims to control the regret

$$R_T = \sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t}\, \ell_{j,t} - \min_{1 \leqslant i \leqslant N} \sum_{t=1}^{T} \ell_{i,t}$$

We will actually denote by

$$R_{i,T} = \sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t}\, \ell_{j,t} - \sum_{t=1}^{T} \ell_{i,t}$$

the regret associated with the component $i \in \{1, \ldots, N\}$. We also denote by $u_+ = \max\{u, 0\}$ the non-negative part of a real number u, and write $\boldsymbol{u}_+$ the vector based on $\boldsymbol{u} = (u_1, \ldots, u_N) \in \mathbb{R}^N$ with components $(u_j)_+$.

Strategy: The statistician considers the following strategy, with hyperparameter $p \geqslant 2$: for $t \geqslant 1$,

$$p_{j,t} = \frac{(R_{j,t-1})_+^{p-1}}{\sum_{k=1}^{N} (R_{k,t-1})_+^{p-1}} \qquad \text{if} \quad \sum_{k=1}^{N} (R_{k,t-1})_+^{p-1} > 0$$

and $p_{j,t} = 1/N$ otherwise (this is in particular the case when $t = 1$).

Analysis in the case $p = 2$

We consider the special case $p = 2$ to have a smooth start. We introduce the instantaneous regret vectors: for all $t \geqslant 1$,

$$\boldsymbol{r}_t = (r_{i,t})_{1 \leqslant i \leqslant N} = \left(\sum_{j=1}^{N} p_{j,t}\, \ell_{j,t} - \ell_{i,t} \right)_{1 \leqslant i \leqslant N}$$

We then define the cumulative regret vector $\boldsymbol{R}_T = \boldsymbol{r}_1 + \ldots + \boldsymbol{r}_T$.

1. Explain why $(u + v)_+ \leqslant \left| u_+ + v \right|$ for all real numbers $(u, v) \in \mathbb{R}^2$ and why we therefore have

$$\big\| (\boldsymbol{R}_t)_+ \big\| \leqslant \big\| (\boldsymbol{R}_{t-1})_+ + \boldsymbol{r}_t \big\|$$

2. Show that

$$\big\| (\boldsymbol{R}_{t-1})_+ + \boldsymbol{r}_t \big\|^2 = \big\| (\boldsymbol{R}_{t-1})_+ \big\|^2 + \|\boldsymbol{r}_t\|^2$$

3. Provide a regret bound for the algorithm considered, say, for losses $\ell_{j,t}$ all lying in some $[m, M]$ range; provide a closed-form regret bound only depending on m, M, T and N.

4. Does the algorithm need to know m, M and T? Are the dependencies in T and N optimal?

Analysis for $p > 2$

The general analysis of this strategy relies on a function Φ defined as: for all $\boldsymbol{u} = (u_1, \ldots, u_N) \in \mathbb{R}^N$,

$$\Phi(\boldsymbol{u}) = \left(\sum_{i=1}^{N} (u_i^+)^p \right)^{2/p}$$

5. Briefly explain why Φ is C^2–regular. Then show that there for all $t \geqslant 2$, there exists $\xi_t \in \mathbb{R}^N$ such that

$$\Phi(\boldsymbol{R}_t) \leqslant \Phi(\boldsymbol{R}_{t-1}) + \frac{1}{2} \sum_{i,j=1}^{N} \partial^2_{ij} \Phi(\xi_t) \, r_{i,t} \, r_{j,t}$$

6. Prove the bound

$$\sum_{i,j=1}^{N} \partial^2_{ij} \Phi(\xi_t) \, r_{i,t} \, r_{j,t} \leqslant 2(p-1) \|\boldsymbol{r}_t\|_p^2$$

You may do so by using that $\psi(x) = x^{2/p}$ is concave (thus $\psi'' \leqslant 0$) and by introducing $f(x) = x_+^p$ for the sake of more concise and more abstract calculations; Hölder's inequality may be useful as well.

7. Conclude to a $(M-m)\sqrt{(p-1)N^{2/p}T}$ regret bound.

8. Which value of p minimizes this bound?
Is the obtained upper bound optimal as far as its dependencies in T and N are concerned?