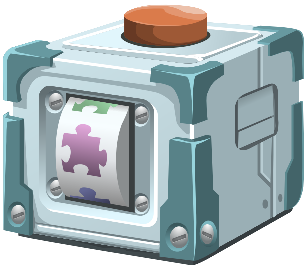
Inria Grenoble-Alpes,

Online Learning Lecture Notes



Pierre Gaillard

2023

Abstract

These notes explore fundamental ideas in online learning, where data are processed in real-time, and algorithms are updated dynamically. Topics include online linear and convex optimization, as well as multi-armed bandits. The main algorithms in the field will be introduced, and we will delve into regret minimization concepts for theoretical analysis. Online learning algorithms play a central role in recent advancements in reinforcement learning.

Useful information

- **Pierre Gaillard** (INRIA Grenoble-Alpes)
- Email: pierre.gaillard@inria.fr
- Relevant references: Cesa-Bianchi and Lugosi [2006], Shalev-Shwartz et al. [2012], Hazan et al. [2016], Lattimore and Szepesvári [2020]
- Content of the class: mostly theoretical (algorithms and proofs), sequential learning with adversarial data, stochastic bandits, adversarial bandits

Contents

1 Introduction

In many applications, the data set is not available from the beginning to learn a model but it is observed sequentially as a flow of data. Furthermore, the environment may be so complex that it is unfeasible to choose a comprehensive model and use classical statistical theory and optimization. A classic example is the spam detection which can be seen as a game between spammer and spam filters. Each trying to fool the other one. Another example, is the prediction of processes that depend on human behaviors such as the electricity consumption. These problems are often not adversarial games but cannot be modeled easily and are surely not i.i.d. There is a necessity to take a robust approach by using a method that learns as ones goes along, learning from experiences as more aspects of the data and the problem are observed. This is the goal of online learning. The curious reader can know more about online learning in the books Cesa-Bianchi and Lugosi [2006], Hazan et al. [2016], Shalev-Shwartz et al. [2012].

1.1 Setting of online learning

In online learning, a player sequentially makes decisions based on past observations. After committing the decision, the player suffers a loss (or receives a reward depending on the problem). Every possible decision incurs a (possibly different) loss. The losses are unknown to the player beforehand an may be arbitrarily chosen by some adversary. More formally, an online learning problem can be formalized as in Figure 1.1.

At each time step $t = 1, \dots, T$

- the player observes a context $x_t \in \mathcal{X}$ (optional step)
- the player chooses an action $\theta_t \in \Theta$ (compact decision/parameter set);
- the environment chooses a loss function $\ell_t : \Theta \to \mathbb{R}$;
- the player suffers loss $\ell_t(\theta_t)$ and observes
    - the losses of every actions: $\ell_t(\theta)$ for all $\theta \in \Theta$ → full-information feedback
    - the loss of the chosen action only: $\ell_t(\theta_t)$ → bandit feedback.

The goal of the player is to minimize his cumulative loss:

$$\widehat{L}_T \stackrel{\text{def}}{=} \sum_{t=1}^{T} \ell_t(\theta_t)\,.$$

Figure 1.1: Setting of an online learning problem/online convex optimization

Example 1.1 (Multi-armed bandit)**.** *In K-armed bandit, the decision set are K actions (or arms) $\Theta = \{1, \dots, K\}$ and the player only observes the performance of the chosen action (bandit feedback). In this problem, there is an exploration-exploitation trade-off: the player wants to select the best arm as often as possible but he also needs to explore all arms to estimate their performance.*

This problem takes his name from slot machines (also known as one-armed bandits because they were originally operated by one lever on the side of the machine) in which some player explores several slot *machines and tries to maximize his cumulative gain (or more likely minimize his loss!).*

Originally, multi-armed bandit setting was introduced by Thompson in 1933 and motivated by clinical trials. For the t-th patient in some clinical study, one needs to choose the treatment to assign to this patient and observe the response. The goal is to maximize the number of patients healed during the study.

Nowadays, multi-armed bandit is motivated by many applications coming from internet (recommender systems, online advertisements,…). We will see more on multi-armed bandit in next lectures.

Example 1.2 (Prediction with expert advice)**.** *In prediction with expert advice, there is some sequence of observations $y_1, \dots, y_T \in [0,1]$ to be predicted step by step with the help of expert forecasts. The setting can be formalized as follows: at each time step $t \geq 1$*

- *the environment reveals experts forecasts $x_t(k)$ for $k = 1, \dots, d$*
- *the player chooses a weight vector $p_t \in \Delta_d \overset{def}{=} \{p \in [0,1]^d : \sum_{k=1}^d p_k = 1\}$ (here θ_t is denoted p_t and $\Theta = \Delta_d$)*
- *the player forecasts $\widehat{y}_t = \sum_{k=1}^d p_t(k) x_t(k)$*
- *the environment reveals $y_t \in [0,1]$ and the player suffers loss $\ell_t(p_t) = \ell(\widehat{y}_t, y_t)$ where $\ell : [0,1]^2 \to [0,1]$ is a loss function.*

Considering $\Theta := \Delta_d$ and $\theta_t := p_t$, this setting can be recovered by the online learning setting of Figure 1.1. The inputs correspond to the expert advice $x_t(k)$ that are often revealed before the learner makes his decision p_t.

Player's performance is then measured via a loss function $\ell_t(p_t) = \ell(\widehat{y}_t, y_t)$ which measures the distance between the prediction $\widehat{y}_t$ and the output y_t. Typical loss functions are the squared loss $\ell(\widehat{y}_t, y_t) = (\widehat{y}_t - y_t)^2$, the absolute loss $\ell(\widehat{y}_t, y_t) = |\widehat{y}_t - y_t|$ or the absolute percentage of error $\ell(\widehat{y}_t, y_t) = |\widehat{y}_t - y_t|/|y_t|$. All these loss functions are convex, which will play an important role in the analysis.

1.2 How to measure the performance: the regret

Of course, if the environment chooses large losses $\ell_t(x)$ for all decisions $\theta \in \Theta$, it is impossible for the player to ensure small cumulative loss. Therefore, one needs a relative criterion: the regret of the player is the difference between the cumulative loss he incurred and that of the best fixed decision in hindsight.

Definition 1.1 **Regret**

The regret of the player with respect to a fixed parameter $\theta^* \in \Theta$ after T time steps is

$$R_T(\theta^*) \overset{def}{=} \sum_{t=1}^T \ell_t(\theta_t) - \sum_{t=1}^T \ell_t(\theta^*)\,.$$

The regret (or uniform regret) is defined as $R_T \overset{def}{=} \sup_{\theta^* \in \Theta} R_T(\theta^*)$.

We have some bias-variance decomposition:

$$\sum_{t=1}^T \ell_t(\theta_t) = \underbrace{\inf_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta)}_{\text{Approximation error = how good the possible actions are.}} + \underbrace{R_T}_{\text{Sequential estimation error of the best action}}$$

We will focus on the regret in these lectures. The goal of the player is to ensure a sublinear regret $R_T = o(T)$ as $T \longrightarrow \infty$ and this for any possible sequence of losses $\ell_1, \dots, \ell_T$. In this case, the average performance of the player will approach on the long term the one of the best decision.

Remarks Let us makes some remarks:

- Except in the stochastic bandit part, we will not make any random assumption on the process generating the losses ℓ_t. The latter are deterministic and may be chosen by some adversary. Typically, the problem can be seen as a game between the player who aims at optimizing with respect to $\theta_1, \dots, \theta_T$ against an environment who aims at mazimizing with respect to $\ell_t, \dots, \ell_T$ and θ^*. Player's goal is to approach the quantity:

$$\inf_{\theta_1} \sup_{\ell_1} \inf_{\theta_2} \sup_{\ell_2} \dots \inf_{\theta_T} \sup_{\ell_T} \sup_{\theta^* \in \Theta} R_T(\theta^*) \,.$$

- Note that the loss functions ℓ_t depend on the round t. This may be caused by many phenomena. We provide here some possible reasons. This may be because
  - of some observation to be predicted if $\ell_t(x) = \ell(x, y_t)$. For instance, if the goal is to predict the evolution of the temperature $y_1, \dots, y_T$, the latter changes over time and a prediction x is evaluated with $\ell_t(x) = (x - y_t)^2$.
  - the environment is stochastic and the variation over time t models some noise effect.
  - of a changing environment. For instance, if the player is playing a game against some adversary that evolves and adapts to its strategy. A typical example is the case of spam detections. If the player tries to detect spams, while some spammers (the environment) try at the same time to fool the player with new spam strategies.

Exercise 1.1. *Instead considering the regret with respect to a fixed $\theta^* \in \Theta$, one would be tempted to minimize the quantity*

$$R_T^* \overset{def}{=} \sum_{t=1}^{T} \ell_t(\theta_t) - \sum_{t=1}^{T} \inf_{\theta \in \Theta} \ell_t(\theta)$$

where the infimum is inside the sum. Show that the environment can ensure R_T^* to be linear in T by choosing properly the loss functions ℓ_t.

2 Online Linear Optimization

In this part, we assume that $\Theta \subset \mathbb{R}^d$ and that the loss functions $\ell_t : \Theta \to \mathbb{R}$ are linear

$$\forall \theta \in \Theta, \qquad \ell_t(\theta) = \langle \theta, g_t \rangle \tag{2.1}$$

where $g_t \in \mathbb{R}^d$ is a loss vector chosen by the environment at round t and which is revealed to the player at the end of the round. This setting might seem restrictive but we will see latter that it can easily be generalized to more complex frameworks.

2.1 Simplex decision set

Here, we start by presenting an algorithm when the decision set Θ is the d-dimensional simplex

$$\Delta_d \overset{\text{def}}{=} \left\{ p \in [0,1]^d : \sum_{k=1}^{d} p_k = 1 \right\}.$$

Since the decisions θ_t are probability distributions over $[d] \overset{\text{def}}{=} \{1, \dots, d\}$, in this part we will denote them by p_t instead of θ_t. The simplex is a versatile decision set that includes distributions, enables weighted averages for method combination, and can represent any convex hull, making it powerful.

2.1.1 The exponentially weighed average forecaster

At round t the player needs to choose a weight vector $p_t \in \Delta_d$. The question is how to choose it? The idea is to give more weight to actions that performed well in the past. But we should not give all the weight to the current best action, otherwise it would not work (see exercises). The exponentially weighted average forecaster (EWA) also called Hedge performs this trade-off by choosing a weight that decreases exponentially fast with the past errors.

```
Parameter: $\eta > 0$, $p_1 \in \Delta_d$
For $t = 1, \dots, T$
    – select $p_t$; incur loss $\ell_t(p_t) = \langle p_t, g_t \rangle$ and observe $g_t \in \mathbb{R}^d$;
    – update for all $k \in \{1, \dots, d\}$

        $$p_{t+1}(k) = \frac{p_1(k) e^{-\eta \sum_{s=1}^{t} g_s(k)}}{\sum_{j=1}^{d} p_1(j) e^{-\eta \sum_{s=1}^{t} g_s(j)}}.$$
```

Algorithm 2.1: The Exponentially weighted average forecaster (EWA)

Exercise 2.1. *Consider the strategy, called "Follow The Leader" (FTL) that puts all the mass on the best action so far:*

$$p_t \in \operatorname*{arg\,min}_{p \in \Theta} \sum_{s=1}^{t-1} \ell_s(p). \tag{FTL}$$

1. *Show that $p_t(k) > 0$ implies that $k \in \arg\min_j \sum_{s=1}^{t-1} g_s(j)$*
2. *Show that the regret of FTL might be linear: i.e., there exists a sequence $g_1, \dots, g_T \in [-1,1]^d$ such that $R_T \geq (1 - 1/d)T$.*

The following theorem proves that EWA, which is a smoothed version of FTL (it performs a soft-max), achieves sublinear regret.

Theorem 2.1

Let $T \geq 1$. For all sequences of loss vectors $(g_t) \in \mathbb{R}^d$ with $\|g_t\|_\infty \leq G_\infty$, if $\eta \leq G_\infty^{-1}$, and $p_1 = (1/d, \dots, 1/d)$, EWA achieves the regret upper bound

$$R_T \overset{def}{=} \sum_{t=1}^T \ell_t(p_t) - \min_{p \in \Delta_d} \sum_{t=1}^T \ell_t(p) \leq \eta \sum_{t=1}^T \sum_{k=1}^d p_t(k) g_t(k)^2 + \frac{\log d}{\eta}, \tag{2.2}$$

where we recall $\ell_t : p \in \Delta_d \mapsto \langle p, g_t \rangle$. Therefore, for the choice $\eta = \frac{1}{G_\infty}\sqrt{\frac{\log d}{T}}$, EWA satisfies the regret bound $R_T \leq 2G_\infty \sqrt{T \log d}$.

This regret bound is optimal up to multiplicative constants (see Cesa-Bianchi and Lugosi [2006]).

Proof. First, we remark that if the losses are linear $\ell_t(p) = \langle p, g_t \rangle$, then

$$\min_{p \in \Delta_d} \sum_{t=1}^T \ell_t(p) = \min_{1 \leq k \leq d} \sum_{t=1}^T g_t(k).$$

Indeed, for any $p \in \Delta_d$,

$$\sum_{t=1}^T \ell_t(p) = \sum_{t=1}^T \langle p, g_t \rangle = \sum_{k=1}^d p(k) \sum_{t=1}^T g_t(k) \geq \sum_{k=1}^d p(k) \min_{1 \leq k \leq d} \sum_{t=1}^T g_t(k) = \min_{1 \leq k \leq d} \sum_{t=1}^T g_t(k).$$

Therefore

$$R_T = \sum_{t=1}^T \langle p_t, g_t \rangle - \min_{1 \leq k \leq d} \sum_{t=1}^T g_t(k).$$

We denote $W_t(j) = e^{-\eta \sum_{s=1}^t g_t(j)}$ and $W_t = \sum_{j=1}^d W_t(j)$. The proof will consist in upper-bounding and lower-bounding W_T. We have

$$\begin{aligned}
W_t &= \sum_{j=1}^d W_{t-1}(j) e^{-\eta g_t(j)} && \leftarrow \quad W_t^{(j)} = W_{t-1}(j) e^{-\eta g_t(j)} \\
&= W_{t-1} \sum_{j=1}^d \frac{W_{t-1}(j)}{W_{t-1}} e^{-\eta g_t(j)} && \\
&= W_{t-1} \sum_{j=1}^d p_t(j) e^{-\eta g_t(j)} && \leftarrow \quad p_t(j) = \frac{e^{-\eta \sum_{s=1}^{t-1} g_s(j)}}{\sum_{k=1}^d e^{-\eta \sum_{s=1}^{t-1} g_s(k)}} = \frac{W_{t-1}(j)}{W_{t-1}} \\
&\leq W_{t-1} \sum_{j=1}^d p_t(j) \big(1 - \eta g_t(j) + \eta^2 g_t(j)^2\big) && \leftarrow \quad e^x \leq 1 + x + x^2 \text{ for } x \leq 1
\end{aligned}$$

$$= \quad W_{t-1}\big(1-\eta\langle p_t, g_t\rangle + \eta^2\langle p_t, g_t^2\rangle\big)\,,$$

where we used in the inequality $-\eta g_t(j) \leq -g_t(j)/G_\infty \leq 1$ and where we denote $g_t = (g_t(1),\dots,g_t(d))$, $g_t^2 = \big(g_t(1)^2,\dots,g_t(d)^2\big)$ and $p_t = (p_t(1),\dots,p_t(d))$. Now, using $1+x \leq e^x$, we get:

$$W_t \leq W_{t-1}\exp\big(-\eta\langle p_t, g_t\rangle + \eta^2\langle p_t, g_t^2\rangle\big)\,.$$

By induction on $t = 1,\dots,T$, this yields using $W_0 = d$

$$W_T \leq d\exp\Big(-\eta\sum_{t=1}^T\langle p_t, g_t\rangle + \eta^2\sum_{t=1}^T\langle p_t, g_t^2\rangle\Big)\,. \tag{2.3}$$

On the other hand, upper-bounding the maximum with the sum,

$$\exp\Big(-\eta\min_{j\in[d]}\sum_{t=1}^T g_t(j)\Big) \leq \sum_{j=1}^d \exp\Big(-\eta\sum_{t=1}^T g_t(j)\Big) \leq W_T\,.$$

Combining the above inequality with Inequality (2.3) and taking the log, we get

$$-\eta\min_{j\in[d]}\sum_{t=1}^T g_t(j) \leq -\eta\sum_{t=1}^T\langle p_t, g_t\rangle + \eta^2\sum_{t=1}^T\langle p_t, g_t^2\rangle + \log d\,. \tag{2.4}$$

Dividing by η and reorganizing the terms proves the first inequality:

$$R_T = \sum_{t=1}^T\langle p_t, g_t\rangle - \min_{1\leq j\leq d}\sum_{t=1}^T g_t(j) \leq \eta\sum_{t=1}^T\langle p_t, g_t^2\rangle + \frac{\log d}{\eta}$$

Optimizing η and upper-bounding $\langle p_t, g_t^2\rangle \leq G_\infty^2$ concludes the second inequality. □

Anytime algorithm (the doubling trick) The previous algorithm EWA depends on a parameter $\eta > 0$ that needs to be optimized according to d, G_∞, and T. For instance, for EWA using the value

$$\eta = \frac{1}{G_\infty}\sqrt{\frac{\log d}{T}}\,.$$

the bound of Theorem 2.1 is only valid for horizon T. However, the learner might not know G_∞ or the time horizon T in advance and one might want an algorithm with guarantees valid simultaneously for all $T \geq 1$. We can avoid the assumption that T is known in advance, at the cost of a constant factor, by using the so-called *doubling trick*. The general idea is the following. Whenever we reach a time step t which is a power of 2, we restart the algorithm (forgetting all the information gained in the past) setting η to $G_\infty^{-1}\sqrt{\log d/t}$. Let us denote EWA-doubling this algorithm.

Theorem 2.2 **Anytime bound on the regret**

For all $T \geq 1$, the pseudo-regret of EWA-doubling is then upper-bounded as:

$$R_T \leq 7G_\infty\sqrt{T\log d}\,.$$

The same trick can be used to turn most online algorithms into anytime algorithms (even in more general settings: bandits, general loss,...). We can use the *doubling trick* whenever we have an algorithm with a regret of order $O(T^\alpha)$ for some $\alpha > 0$ with a known horizon T to turn it into an algorithm with a regret $O(T^\alpha)$ for all $T \geq 1$.

Another solution is to use time-varying parameters η_t replacing T with the current value of t. The analysis is however less straightforward.

Exercise 2.2. *Prove a regret bound for the time-varying choice* $\eta_t = \sqrt{\log d/(1 + \sum_{s=1}^t \|g_t\|_\infty^2)}$ *in EWA.*

Proof of Theorem 2.2. For simplicity we assume $T = 2^{M+1} - 1$. The regret of EWA-doubling is then upper-bounded as:

$$\begin{aligned} R_T &= \sum_{t=1}^T \ell_t(p_t) - \min_{p \in \Delta_d} \sum_{t=1}^T \ell_t(p) \\ &\leq \sum_{t=1}^T \ell_t(p_t) - \sum_{m=0}^M \min_{p \in \Delta_d} \sum_{t=2^m}^{2^{m+1}-1} \ell_t(p) \\ &= \sum_{m=0}^M \underbrace{\sum_{t=2^m}^{2^{m+1}-1} \ell_t(p_t) - \min_{p \in \Delta_d} \sum_{t=2^m}^{2^{m+1}-1} \ell_t(p)}_{R_m} \,. \end{aligned}$$

Now, we remark that each term R_m corresponds to the expected regret of an instance of EWA over the 2^m rounds $t = 2^m, \dots, 2^{m+1} - 1$ and run with the optimal parameter $\eta = \sqrt{\log d/2^m}$. Therefore, using Theorem 2.1, we get $R_m \leq 2\sqrt{2^m \log d}$, which yields:

$$R_T \leq \sum_{m=0}^M 2G_\infty \sqrt{2^m \log d} \leq 2(1 + \sqrt{2})G_\infty \sqrt{2^{M+1} \log d} \leq 7G_\infty \sqrt{T \log d} \,.$$

□

Improvement for small losses The first inequality in Theorem 2.1 is sometimes called improvement for small losses when losses are non-negative. Let's define $\widehat{L}_T \overset{\text{def}}{=} \sum_{t=1}^T \ell_t(p_t)$ the loss of the algorithm and $L_T^* \overset{\text{def}}{=} \min_{p \in \Delta_d} \sum_{t=1}^T \ell_t(p)$ the optimal loss. Then, optimizing in $\eta = \big(\log(d)/\sum_{t=1}^T \langle p_t, g_t^2 \rangle\big)^{1/2}$, the regret is upper-bounded by

$$R_T \overset{\text{def}}{=} \widehat{L}_T - L_T^* \leq \frac{\log d}{\eta} + \eta \sum_{t=1}^T \langle p_t, g_t^2 \rangle = 2\sqrt{(\log d) \textstyle\sum_{t=1}^T \langle p_t, g_t^2 \rangle} \leq 2\sqrt{(\log d) G_\infty \widehat{L}_T} \,.$$

Therefore, using that $x^2 \leq a + cx$ implies $x \leq \sqrt{a} + c$ when $a, c \geq 0$, we get with $x = \sqrt{\widehat{L}_T}$ that

$$\sqrt{\widehat{L}_T} \leq \sqrt{L_T^*} + 2\sqrt{G_\infty \log d} \,,$$

which yields

$$R_T \leq 4\sqrt{G_\infty \log(d) L_T^*} + 4G_\infty \log d \,,$$

which is small whenever the optimal loss L_T^* is small.

2.1.2 Application to prediction with expert advice

The preceding section considers linear loss functions. Yet, it can yield non-trivial regret bounds for general convex losses. We consider here an application to the setting of prediction with expert advice detailed in Example 1.2. The goal is to minimize the regret with respect to the best expert

$$R_T^{\text{expert}} \overset{\text{def}}{=} \sum_{t=1}^T \ell(\widehat{y}_t, y_t) - \min_{1 \le k \le d} \sum_{t=1}^T \ell(x_t(k), y_t)\,,$$

where $\widehat{y}_t = \langle p_t, x_t \rangle$ are the predictions of the algorithm and y_t the observations to be predicted sequentially.

Convex loss function We state bellow a corrolary to Theorem 2.1 when the loss functions $\ell(\cdot,\cdot)$ are convex in their first argument.

> **Corollary 2.3** **Regret of EWA for prediction with expert advice and convex loss**
>
> *Let $T \geq 1$, $B > 0$. Assume that the loss function $\ell : (x,y) \in \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ is convex and takes values in $[-B,B]$. Then, EWA applied with the vector vectors $g_t = \big(\ell(x_t(1), y_t), \dots, \ell(x_t(d), y_t)\big) \in [-B,B]^d$ has a regret upper-bounded by*
>
> $$R_T^{expert} \le 2B\sqrt{T \log d}$$
>
> *where $\widehat{y}_t = \langle p_t, x_t \rangle$ and were $\eta > 0$ is well-tuned.*

Therefore, the average error of the algorithm will converge to the average error of the best expert. This is the case for the square loss, the absolute loss or the absolute percentage of error.

Proof. It suffices to remark that by convexity of $\ell(\cdot,\cdot)$ in its first argument

$$\begin{aligned} R_T^{\text{expert}} &= \sum_{t=1}^T \ell(\langle p_t, x_t \rangle, y_t) - \min_{1 \le k \le d} \sum_{t=1}^T \ell(x_t(k), y_t) \\ &\le \sum_{t=1}^T \langle p_t, g_t \rangle - \min_{1 \le k \le d} \sum_{t=1}^T g_t(k) \overset{\text{def}}{=} R_T\,. \end{aligned}$$

The result is then obtained by Theorem 2.1. □

Exp-concave loss function Here, we show that a faster rate can be obtained (with EWA) if the loss function are exp-concave.

> **Definition 2.1** **η-exp-concavity**
>
> *For $\eta \in \mathbb{R}$, a function f is said to be η-exp-concave if $x \mapsto e^{-\eta f(x)}$ is concave.*

Exp-concavity is stronger than convexity but weaker than strong convexity. Indeed, exp-concave functions are convex because $-\log$ is convex and decreasing. Furthermore, any η-exp-concave function is also η'-exp-concave for $0 \le \eta' \le \eta$.

In prediction with expert advice, if the loss are generated from a fixed loss function $\ell_t(p) = \ell(\langle p, x_t \rangle, y_t)$, then ℓ_t are η-expconcave if $\widehat{y} \mapsto \ell(\widehat{y}, y_t)$ are η-exp-concave for all y_t. We can compute η for some common loss functions:

- *the squared loss*: $\ell : (\widehat{y}, y) \in [0,1]^2 \mapsto (\widehat{y} - y)^2$, then ℓ_t are $1/2$-exp-concave. Indeed, let $y \in [0,1]$ and denote $G : \widehat{y} \mapsto \exp\big(-\eta(\widehat{y} - y)^2\big)$. Then, $G''(\widehat{y}) = G(\widehat{y})\big(4\eta^2(\widehat{y} - y)^2 - 2\eta\big)$. Thus G is concave if and only if $(\widehat{y} - y)^2 \leq 1/(2\eta)$ which is satisfied for $\eta = 1/2$. This is also the case in higher dimensions with $\ell(\widehat{y}, y) = \|\widehat{y} - y\|^2$. If the observations and prediction $\widehat{y}, y \in [0, B]$, then the ℓ_t are $1/(2B^2)$-exp-concave
- *the relative entropy* (or Kullback–Leibler divergence): $\ell : (\widehat{y}, y) \in [0,1]^2 \mapsto y \log(y/\widehat{y}) - (1 - y)\log((1-y)/(1-\widehat{y}))$. Then the functions ℓ_t are 1-exp-concave. This loss can for instance used for density estimation of the sequence $y_1, \dots, y_T$.
- the linear loss $\ell(\widehat{y}, y) = \widehat{y} \cdot y$, the absolute loss $\ell(\widehat{y}, y) = |\widehat{y} - y|$ or the absolute percentage of error are however not η-exp-concave for any $\eta > 0$.

Corollary 2.4 Regret of EWA for prediction with expert advice and exp-concave loss

In the setting of prediction with expert advice, if the loss functions $\ell(\cdot, y_t)$ are η-exp-concave for all y_t, then EWA run with vectors $g_t = \big(\ell(x_t(1), y_t), \dots, \ell(x_t(d), y_t)\big) \in \mathbb{R}^d$ with parameter $\eta > 0$ and $p_1 = (1/d, \dots, 1/d)$ satisfies

$$R_T^{expert} \leq \frac{\log d}{\eta},$$

for all $T \geq 1$.

The worst-case regret does not increase with T but grows logarithmically in the dimension d.

Proof. The proof is similar to the original proof of EWA. We define $W_t(i) = e^{-\eta \sum_{s=1}^t g_s(i)}$ and $W_t = \sum_{i=1}^d W_t(i)$. We have

$$\begin{aligned}
W_t &= \sum_{j=1}^N W_{t-1}(j) e^{-\eta g_t(j)} && \leftarrow \quad W_t(j) = W_{t-1}(j) e^{-\eta g_t(j)} \\
&= W_{t-1} \sum_{j=1}^N \frac{W_{t-1}(j)}{W_{t-1}} e^{-\eta g_t(j)} \\
&= W_{t-1} \sum_{j=1}^N p_t(j) e^{-\eta g_t(j)} && \leftarrow \quad p_t(j) = \frac{e^{-\eta \sum_{s=1}^{t-1} g_s(j)}}{\sum_{k=1}^N e^{-\eta \sum_{s=1}^{t-1} g_s(k)}} = \frac{W_{t-1}(j)}{W_{t-1}} \\
&\leq W_{t-1} \exp\big(-\eta \ell(\langle p_t, x_t \rangle, y_t)\big) && \leftarrow \quad \eta\text{-exp-concavity}
\end{aligned}$$

Now, by induction on $t = 1, \dots, T$, this yields using $W_0 = d$

$$W_T \leq d \exp\left(-\eta \sum_{t=1}^T \ell(\widehat{y}_t, y_t)\right). \tag{2.5}$$

On the other hand, upper-bounding the maximum with the sum,

$$\exp\Big(-\eta \min_{j \in [d]} \sum_{t=1}^T g_t(j)\Big) \leq \sum_{j=1}^d \exp\Big(-\eta \sum_{t=1}^T g_t(j)\Big) \leq W_T.$$

Combining the above inequality with Inequality (2.5) and taking the log concludes the proof. □

2.1.3 Non-convexity: Linearizing the loss through randomization

Setting: Θ finite, non-convex loss functions $\ell_t : \Theta \to [-B, B]$.

In this section, we consider a *finite set of decision* $\Theta = \{1, \dots, d\}$ and we assume that the player is restricted to play an action in Θ. In other words, the player cannot play convex combinations of the actions as it was done for prediction with expert advice. For instance, we may want to build a recommender system to recommend movies to customers. The loss function are *arbitrary bounded loss functions* $\ell_t : \Theta \to [-B, B]$.

Need of a random strategy The following proposition shows that the choice θ_t cannot be deterministic in this setting. Otherwise, the adversary may fool the player by taking ℓ_t depending on θ_t.

> **Proposition 2.5**
>
> *Any deterministic algorithm may incur a linear regret. In other words, we can find some sequence of losses ℓ_t such that $R_T \gtrsim T$.*

Proof. Since θ_t is deterministic, the loss function ℓ_t can depend on θ_t. We then choose $\ell_t(\theta_t) = 1$ and $\ell_t(\theta) = 0$ for $\theta \neq \theta_t$. Then one of the chosen actions was picked less then T/d times so that $\max_{1 \le k \le d} \ell_t(k) \le T/d$. Therefore, $R_T \ge (1 - 1/d)T$. □

From the above proposition, we see that the strategy of the learner needs to be random. Therefore, instead of choosing an action in $\{1, \dots, d\}$, the player chooses a probability distribution $p_t \in \Delta_d := \{p \in [0,1]^d : \sum_k p_k = 1\}$ and draws $\theta_t \sim p_t$. And we recover the setting with actions played in the simplex Δ_d.

A random regret The regret R_T will be here a random quantity that depends on the randomness of the algorithm (and eventually of the data). We will thus focus on upper-bounding the regret:

- with high-probability: $R_T \le \varepsilon$ with probability at least $1 - \delta$;
- in expectation: $\mathbb{E}[R_T] \le \varepsilon$.

From high-probability bound to expected bound. Note that since the losses are bounded in $[0, 1]$ a bound in high probability entails a bound in expectation. If $R_T \le \varepsilon$ with probability at least $1 - \delta$, then

$$\mathbb{E}[R_T] \le \mathbb{E}\big[R_T | R_T \le \varepsilon\big]\mathbb{P}(R_T \le \varepsilon) + \mathbb{E}\big[R_T | R_T \ge \varepsilon\big]\mathbb{P}(R_T \ge \varepsilon) \le \varepsilon + T\delta\,. \tag{2.6}$$

Another useful (and often better) tool to transform a high-probability bound into a bound in expectation is the inequality $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \ge \varepsilon) d\varepsilon$ for nonnegative random variable X.

From expected bound to high-probability bound. On the other hand, since the losses are bounded, using Hoeffding's inequality a bound in expectation entails a bound in high probability at the cost of an additive term of order $\sqrt{T \log(1/\delta)}$ in the regret.

> **Proposition 2.6**
>
> *Let Θ be finite of cardinal d and $\ell_1, \dots, \ell_T : \Theta \to [-B, B]$ be an arbirary sequence of losses. Then,*

applying EWA with a well-chosen η and $p_1 = (1/d, \dots, 1/d)$ and sampling $\theta_t \sim p_t$ at every round, satisfies the expected regret

$$\mathbb{E}\big[R_T\big] = \mathbb{E}\left[\sum_{t=1}^T \ell_t(\theta_t) - \min_{\theta\in\Theta} \sum_{t=1}^T \ell_t(\theta)\right] \leq 2B\sqrt{T\log d}$$

for η well tuned.

Exercise 2.3. *Using Hoeffding's inequality, provide a bound on the regret R_T with probability* $1-\delta$*.*

Proof. Using $g_t = \big(\ell_t(1), \dots, \ell_t(d)\big) \in [-B, B]^d$, from Theorem 1 of last class, we have

$$\sum_{t=1}^T \langle p_t, g_t\rangle - \min_{\theta\in\Theta} \sum_{t=1}^T \ell_t(\theta) \leq 2B\sqrt{T\log d}\,.$$

It suffices then to take the expectation and remark that

$$\mathbb{E}\big[\ell_t(\theta_t)\big] = \mathbb{E}\big[\mathbb{E}[\ell_t(\theta_t)|p_t]\big] = \mathbb{E}\big[\langle p_t, g_t\rangle\big]\,.$$

□

It is worth pointing out that we did not make any assumption on the loss function ℓ_t beside boundedness. In particular, it can be non-convex.

Example 2.1 (Online classification)**.** *Assume that you may want to predict a sequence of labels $y_1, \dots, y_T \in \{0,1\}$ (such as spams) based on expert advice $x_t(k) \in \{0,1\}$ (such as different spam detectors). Then, using the losses $\ell_t(k) = \mathbb{1}_{x_t(k)\neq y_t}$, EWA ensures*

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}_{\theta_t\neq y_t} - \min_{1\leq k\leq d} \sum_{t=1}^T \mathbb{1}_{x_t(k)\neq y_t}\right] \leq 2\sqrt{T\log d}\,.$$

Hence, the expected number of mistakes of the algorithms will not be much larger than the one of the best expert. This is valid though the loss function is nonconvex.

2.2 Convex and compact decision set

Setting: linear loss function, convex and compact decision set Θ.

In this section, we generalize the preceding sections to any compact convex decision set $\Theta \subset \mathbb{R}^d$. We still assume that the loss functions take a linear form

$$\forall\theta\in\Theta, \qquad \ell_t(\theta) = \langle\theta, g_t\rangle\,,$$

for some $g_t \in \mathbb{R}^d$. We provide a few well-known algorithms in this setting.

2.2.1 Online Gradient Descent

We introduce, Online Gradient Descent, and is due to Zinkevich [2003] in the online learning setting. It is an online variant of the well-known Gradient Descent algorithm in optimization.

```
Parameter: $\eta > 0$, $\theta_1 \in \Theta$
For $t = 1, \dots, T$
    – select $\theta_t$; incur loss $\ell_t(\theta_t) = \langle \theta_t, g_t \rangle$ and observe $g_t \in \mathbb{R}^d$;
    – update
                    $\theta_{t+1} = \Pi_\Theta\big(\theta_t - \eta g_t\big)$,
      where $\Pi_\Theta$ is the Euclidean projection onto $\Theta$.
```

Algorithm 2.2: Online Gradient Descent (OGD) for linear losses $\ell_t(\theta) = \langle \theta, g_t \rangle$.

Theorem 2.7

Then for any sequence $g_1, \dots, g_T \in \mathbb{R}^d$, if the looses are linear $\ell_t(\theta) = \langle \theta, g_t \rangle$, the regret of OGD satisfies

$$R_T = \sum_{t=1}^{T} \ell_t(\theta_t) - \ell_t(\theta^*) \le \frac{\|\theta^* - \theta_1\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|^2 ,$$

where $\theta^* \in \arg\min_{\theta \in \Theta} \sum_{t=1}^{T} \ell_t(\theta)$. In particular, if $\max_{\theta \in \Theta} \|\theta - \theta_1\| \le D_2$ and $\|g_t\| \le G_2$ for all t, for $\eta = \frac{D_2}{G_2\sqrt{T}}$, we have $R_T \le D_2 G_2 \sqrt{T}$.

Exercise 2.4. *Prove an upper-bound on the regret of OGD*

a) when η is calibrated with a doubling trick.
b) when η is calibrated using a time-varying parameter η_t

Exercise 2.5. *Prove an upper-bound on the regret of OGD with respect to any sequence of points $\theta_1^*, \dots, \theta_t^* \in \Theta$ such that $\sum_{t=2}^{T} \|\theta_t^* - \theta_{t-1}^*\| \le X$*

$$\sum_{t=1}^{T} \ell_t(\theta_t) - \sum_{t=1}^{T} \ell_t(\theta_t^*) \le \quad \dots$$

Remark. *Assume that $\Theta = \Delta_d$ is the simplex and the loss functions are of the form $\ell_t(\theta) = \langle \theta, g_t \rangle$ where $\|g_t\|_\infty \le G_\infty$. Then both EWA and OGD are possible algorithms (see Theorems 2.1 and 2.7). We saw in Theorem 2.1 that EWA has a regret bound $R_T \le 2G_\infty \sqrt{T \log d}$. In this case, for all $p, p' \in \Delta_d$*

$$\|p - p'\| = \sum_{k=1}^{d} \big(p(i) - p'(i)\big)^2 \le \sum_{i=1}^{d} \big|p(i) - p'(i)\big| \le \sum_{i=1}^{d} p(i) + p'(i) = 2 ,$$

and $\|g_t\| \le \sqrt{d}\|g_t\|_\infty \le \sqrt{d}G_\infty$. Therefore, the regret of OGD is upper-bounded by $R_T \le G_\infty\sqrt{2dT}$. To summarize

$$\text{EWA:} \quad R_T \le 2G_\infty\sqrt{T \log d} \qquad \text{and} \qquad \text{OGD:} \quad R_T \le \sqrt{2dT} .$$

The dependence on d of OGD is suboptimal in this case. This is solved by OMD, a generalization of both algorithms.

Proof of Theorem 2.7. Let $\theta^* \in \arg\min_{\theta \in \Theta} \sum_{t=1}^{T} \ell_t(\theta)$ and denote $z_{t+1} = \theta_t - \eta g_t$ so that by definition of θ_{t+1} in the algorithm, we have $\theta_{t+1} = \Pi_\Theta(z_{t+1})$. By convexity, the regret can be upper-bounded as

$$R_T \overset{\text{def}}{=} \sum_{t=1}^{T} \ell_t(\theta_t) - \ell_t(\theta^*) \quad = \quad \sum_{t=1}^{T} \langle g_t, \theta_t - \theta^* \rangle$$

$$= \frac{1}{\eta} \sum_{t=1}^{T} \langle z_{t+1} - \theta_t, \theta_t - \theta^* \rangle \qquad \leftarrow g_t = \frac{z_{t+1} - \theta_t}{\eta} \,.$$

Then, we use the equality $\|x - y\|^2 = \|x\|^2 + \|y\|^2 - 2\langle x, y \rangle$ for all $x, y \in \Theta$ so that

$$\langle x, y \rangle = \frac{\|x\|^2 + \|y\|^2 - \|x - y\|^2}{2} \,.$$

Applying it with $x = z_{t+1} - \theta_t$ and $y = \theta_t - \theta^*$ en substituting into the above regret bound, this yields

$$R_T \leq \frac{1}{2\eta} \sum_{t=1}^{T} \left(\|z_{t+1} - \theta_t\|^2 + \|\theta^* - \theta_t\|^2 - \|z_{t+1} - \theta^*\|^2 \right)$$

Then, using $\|z_{t+1} - \theta_t\| = \eta \|g_t\|$ and $\|\theta^* - \theta_t\| \leq \|\theta^* - z_t\|$ because Θ is convex and $\theta_t = \Pi_\Theta(z_t)$, we get

$$R_T \leq \frac{1}{2\eta} \sum_{t=1}^{T} \left(\eta^2 \|g_t\|^2 + \|\theta^* - z_t\|^2 - \|z_{t+1} - \theta^*\|^2 \right) .$$

The last terms telescope, therefore summing over t concludes the proof

$$R_T \leq \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|^2 + \frac{\|\theta^* - \theta_1\|^2}{2\eta} \,.$$

□

2.2.2 Regularized Follow the Leader

In the preceding section, EWA and OGD were introduced. Here, we will present more versatile algorithms that rely on a regularization function $R : \Theta \to \mathbb{R}$, appearing to be generalizations of the earlier algorithms. We consider the RFTL algorithm defined in Algorithm 2.3, which depends on a strongly convex, smooth, and twice differentiable regularization function $R : \Theta \to \mathbb{R}$.

```
Input: $\eta > 0$, regularization function $R > 0$
Let $\theta_1 = \arg\min_{\theta \in \Theta} \{R(\theta)\}$
For $t = 1$ to $T$ do
    – Play $\theta_t$ incur loss $\ell_t(\theta_t) = \langle \theta_t, g_t \rangle$ and and observe $g_t \in \mathbb{R}^d$
    – Update
        $\theta_{t+1} = \arg\min_{\theta \in \Theta} \left\{ \eta \sum_{s=1}^{t} \langle \theta, g_s \rangle + R(\theta) \right\}$
end for
```

Algorithm 2.3: Regularized Follow the Leader (RFTL)

Before stating the theorem, let us first recall a few basic definitions useful for the analysis.

Definition 2.2 **Strong convexity**

We say that a differentiable function $f : \Theta \to \mathbb{R}$ is α-strongly convex with respect to a norm $\|\cdot\|_f$ if for all $\theta, \theta' \in \Theta$

$$f(\theta) \geq f(\theta') + \langle \nabla f(\theta'), \theta - \theta' \rangle + \frac{\alpha}{2} \|\theta - \theta'\|_f^2 \,.$$

Theorem 2.8 **RFTL regret upper-bound**

Let $T \geq 1$. Let $R : \Theta \to \mathbb{R}$ be a regularization which is α-strongly convex with respect to some norm $\|\cdot\|_R$ and let $D_R \geq \sqrt{\max_\theta R(\theta) - \min_\theta R(\theta)}$. Then, for a well-chosen $\eta > 0$, the regret of RFTL is upper-bounded as

$$R_T \leq 2D_R \sqrt{\frac{2}{\alpha} \sum_{t=1}^{T} \|g_t\|_{R,*}^2}\,,$$

where $\|\cdot\|_{R,*} \overset{def}{=} \sup_{\|\theta\| \leq 1} \langle \cdot, \theta \rangle$ is the dual norm of $\|\cdot\|_R$.

Proof. Let $T \geq 1$. Define for all $t \geq 1$ the function Φ_t such that for all $\theta \in \Theta$

$$\Phi_t(\theta) = \eta \sum_{s=1}^{t} \langle \theta, g_s \rangle + R(\theta)\,.$$

Note that by definition $\theta_{t+1} = \arg\min_\theta \Phi_t(\theta)$. Let $\theta \in \Theta$. Then,

$$\begin{aligned}
R_T(\theta) &\overset{\text{def}}{=} \sum_{t=1}^{T} \langle \theta_t, g_t \rangle - \sum_{t=1}^{T} \langle \theta, g_t \rangle \\
&= \sum_{t=1}^{T} \langle \theta_t, g_t \rangle - \frac{\Phi_T(\theta)}{\eta} + \frac{R(\theta)}{\eta} \\
&\leq \sum_{t=1}^{T} \langle \theta_t, g_t \rangle - \frac{\Phi_T(\theta_{T+1})}{\eta} + \frac{R(\theta)}{\eta} \\
&= \sum_{t=1}^{T} \langle \theta_t, g_t \rangle - \langle \theta_{T+1}, g_T \rangle - \frac{\Phi_{T-1}(\theta_{T+1})}{\eta} + \frac{R(\theta)}{\eta} \\
&\leq \sum_{t=1}^{T} \langle \theta_t, g_t \rangle - \langle \theta_{T+1}, g_T \rangle - \frac{\Phi_{T-1}(\theta_T)}{\eta} + \frac{R(\theta)}{\eta} \\
&\leq \sum_{t=1}^{T} \langle \theta_t, g_t \rangle - \sum_{t=T-1}^{T} \langle \theta_{t+1}, g_t \rangle - \frac{\Phi_{T-2}(\theta_{T-1})}{\eta} + \frac{R(\theta)}{\eta} \\
&\leq \sum_{t=1}^{T} \langle \theta_t - \theta_{t+1}, g_t \rangle + \frac{R(\theta) - R(\theta_1)}{\eta} && \leftarrow \text{by induction} \\
&\leq \sum_{t=1}^{T} \|\theta_t - \theta_{t+1}\|_R \|g_t\|_{R,*} + \frac{R(\theta) - R(\theta_1)}{\eta} && \leftarrow \text{by Cauchy-Schwarz} \\
&\leq \sum_{t=1}^{T} \|\theta_t - \theta_{t+1}\|_R \|g_t\|_{R,*} + \frac{D_R^2}{\eta} && \leftarrow \text{by definition of } D_R \qquad (2.7)
\end{aligned}$$

It now remains to upper-bound $\|\theta_t - \theta_{t+1}\|_R$ for all $t \geq 1$. Let $t \geq 1$. Using that R is α-strongly convex, we have

$$\begin{aligned}
\frac{\alpha}{2} \|\theta_{t+1} - \theta_t\|_R^2 &\leq R(\theta_t) - R(\theta_{t+1}) + \langle \nabla R(\theta_{t+1}), \theta_{t+1} - \theta_t \rangle \\
&= \langle \nabla \phi_t(\theta_{t+1}), \theta_{t+1} - \theta_t \rangle + \Phi_t(\theta_t) - \Phi_t(\theta_{t+1})
\end{aligned}$$

But since $\theta_{t+1} = \arg\min_\theta \Phi_t(\theta)$, from the optimality condition, we have $\langle \nabla\phi_t(\theta_{t+1}), \theta_{t+1} - \theta_t \rangle \leq 0$, which yields

$$\begin{aligned}
\frac{\alpha}{2}\|\theta_{t+1} - \theta_t\|_R^2 &\leq \Phi_t(\theta_t) - \Phi_t(\theta_{t+1}) \\
&= \Phi_{t-1}(\theta_t) + \Phi_{t-1}(\theta_{t+1}) + \eta\langle \theta_t - \theta_{t+1}, g_t \rangle \\
&\leq \eta\langle \theta_t - \theta_{t+1}, g_t \rangle \\
&\leq \eta\|\theta_t - \theta_{t+1}\|_R\|g_t\|_{R,*}^2 \,.
\end{aligned}$$

Thus,

$$\|\theta_{t+1} - \theta_t\|_R \leq \frac{2\eta}{\alpha}\|g_t\|_{R,*} \,.$$

Plugging back into (2.7) and optimizing η concludes the regret bound

$$R_T \leq \frac{D_R^2}{\eta} + \frac{2\eta}{\alpha}\sum_{t=1}^T \|g_t\|_{R,*}^2 \,.$$

□

Let us first see a few special cases of RFTL.

Euclidean regularization. Choosing $R(\theta) = \frac{1}{2}\|\theta - \theta_1\|^2$, Theorem 2.8 recovers up to a factor $\sqrt{2}$ the same regret upper-bound as the one of OGD in Theorem 2.7. Indeed, here, R is 1-strongly convex with respect to the Euclidean norm $\|\cdot\|$ whose dual is also $\|\cdot\|_* = \|\cdot\|$. In this case, for any θ,

$$R(\theta) - R(\theta_1) = \frac{1}{2}\|\theta - \theta_1\| \,,$$

thus $D_R = D_2/2$, which gives the regret upper-bound $R_T \leq D_2 G_2\sqrt{2T}$. Note that in this case, RFTL is close but different from OGD.

Entropic regularization. Let $\theta_1 \in \Theta := \Delta_d$. Then, we recover the exponentially weighted average forecaster by choosing

$$R(\theta) = KL(\theta\|\theta_1) := \left\langle \theta, \log\left(\frac{\theta}{\theta_1}\right)\right\rangle$$

where by abuse of notation, we write $\log x = (\log x(1), \dots, \log x(d))$ and $\log\frac{x}{y} = \log x - \log y$. In this case,

$$D_R \leq -\log\left(\min_i \theta_1(i)\right)$$

and we can show that R is 1-strongly convex with respect to the ℓ_1-norm. Indeed, $\nabla R(\theta) = 1 + \log(\theta/\theta_1)$ and for any $x, y \in \Delta_d$,

$$\begin{aligned}
R(x) - R(y) - \langle \nabla R(y), x - y\rangle = \langle x, \log\frac{x}{\theta_1}\rangle - \langle y, \log\frac{y}{\theta_1}\rangle - \langle 1 + \log\frac{y}{\theta_1}, x - y\rangle \\
= \left\langle x, \log\frac{x}{y}\right\rangle = KL(x\|y) \geq \frac{\|x - y\|_1^2}{2}
\end{aligned}$$

where the last inequality is by Pinsker's inequality. Thus, Theorem 2.8 provides the regret upper-bound $R_T \leq 2\sqrt{2T\log d}$ if $\theta_1 = (1/d, \dots, 1/d)$ which is similar to the one of the exponentially weighted average forecaster proved in Theorem 2.1 up to a factor $\sqrt{2}$. Note that in this case, RFTL exactly matches with EWA as shown by the following proposition.

Proposition 2.9

Let $g \in \mathbb{R}^d$ then, the solution of $\theta_* = \arg\min_{\theta \in \Delta_d} \left\{ \eta\langle \theta, g\rangle + KL(\theta||\theta_1) \right\}$ has a closed-form solution defined component-wise by

$$\theta_*(k) = \frac{\theta_1(k)\exp(-\eta g_k)}{\sum_{i=1}^d \theta_1(k)\exp(-\eta g_j)} \,. \tag{2.8}$$

Proof. Let $\theta \in \mathbb{R}^d$, then

$$\begin{aligned} -\eta\langle\theta, g\rangle - KL(\theta||\theta_1) &= \log\left(e^{-\eta\langle\theta,g\rangle + KL(\theta||\theta_1)}\right) \\ &= \log\left(e^{\left\langle \theta, -\eta g + \log\frac{\theta_1}{\theta}\right\rangle}\right) \\ &\leq \log\left(\left\langle \theta, e^{-\eta g + \log\frac{\theta_1}{\theta}}\right\rangle\right) \qquad \leftarrow \text{Jensen's inequality} \\ &\leq \log\left(\left\langle \theta_1, e^{-\eta g}\right\rangle\right) \qquad \leftarrow \text{The ineq. is strict if } \theta_1 \text{ has smaller suport than } \theta. \end{aligned}$$

But for θ_* satisfying (2.8) we have

$$-\eta\langle\theta_*, g\rangle - KL(\theta_*||\theta_1) = -\eta\frac{\langle e^{-\eta g}\theta_1, g\rangle}{\langle\theta_1, e^{-\eta g}\rangle} - \frac{\langle e^{-\eta g}\theta_1, -\eta g - \log\langle\theta_1, e^{-\eta g}\rangle\rangle}{\langle\theta_1, e^{-\eta g}\rangle} = \log\left(\left\langle\theta_1, e^{-\eta g}\right\rangle\right).$$

This concludes the proof. □

The above proposition can also be proved using Lagragian and classical tools from constrained convex optimization (left as exercise).

2.2.3 Online Mirror Descent

Online Mirror Descent (OMD) is another generalization of OGD to better exploit the geometry of the decision space Θ. OMD is the online counterpart of the *Mirror Descent* algorithm from convex optimization. The generality of OMD comes from the updates being performed into a dual space which is defined by a convex differentiable regularization function $R : \Theta \to \mathbb{R}$.

Before stating the algorithm, we need to define the Bregman divergence.

Definition 2.3 **Bregman divergence**

For any continuously differentiable convex function R, the Bregman divergence with respect to R is defined as

$$D_R(x||y) \leq R(x) - R(y) - \nabla R(y)\cdot(x-y) \quad \forall x, y \in \Theta\,.$$

It is the difference between the value of the regularization function at x and the value of its first order Taylor approximation. It is nonnegative but not symmetric. Online Mirror Descent is then defined as follows.

Theorem 2.10 **Regret of OMD**

Let $t \geq 1$. Let Θ be a compact and convex set. Then, for all sequences $(g_t) \in \mathbb{R}^d$, the regret of OMD is upper-bounded as

$$\sum_{t=1}^T \ell_t(\theta_t) - \min_{\theta\in\Theta}\sum_{t=1}^T \ell_t(\theta) \leq \frac{D}{\eta} + \frac{1}{\eta}\sum_{t=1}^T D_{R^*}\big(\nabla R(\theta_t) - \eta g_t||\nabla R(\theta_t)\big)$$

```
Parameters: $\eta > 0$, regularization function $R$
Initialize: $z_1 \in \mathbb{R}^d$ such that $\nabla R(z_1) = 0$ and $\theta_1 = \arg\min_{\theta\in\Theta} D_R(\theta||y_1)$
For $t = 1, \dots, T$
    – select $\theta_t$; incur loss $\ell_t(\theta_t)$ and observe $g_t \in \mathbb{R}^d$
    – update $z_t$ such that
          $\nabla R(z_{t+1}) = \nabla R(\theta_t) - \eta g_t$ .
    – project according to the Bregman divergence
          $\theta_{t+1} \in \arg\min_{\theta\in\Theta} D_R(\theta||z_{t+1})$ .
```

Algorithm 2.4: Online Mirror Descent (OMD)

where $D \geq \max_{\theta\in\Theta} |R(\theta)|$ and R^* is the Fenchel conjugate of R defined as $R^*(z) \overset{def}{=} \max_{\theta\in\Theta}\big\{\theta\cdot z - R(\theta)\big\}$.

The proof can be found for instance in Bubeck et al. [2012]. EG and OGD are two particular cases of Online Mirror Descent.

Example 2.2 (Balls in $\mathbb{R}^d$ = OGD)**.** *If $\Theta \subset \mathbb{R}^d$, we can choose $R(x) = \frac{1}{2}\|x\|^2$. Then $\nabla R(x) = x$ and $D_R(x||y) = \frac{1}{2}\|x-y\|^2$. Therefore, the update of OMD becomes $y_{t+1} = \theta_t - \eta\nabla\ell_t(\theta_t)$ and $\theta_{t+1} = \Pi_\Theta(y_{t+1})$. We recover the online gradient descent algorithm.*

Example 2.3 (Simplex = EWA)**.** *If $\Theta = \Delta_d$. We can choose the negative entropy*

$$R(x) = \sum_{i=1}^{d} x(i)\log x(i)\,.$$

In this case, $\nabla R(x)_i = 1 + \log x(i)$ and the Bregman Divergence is $D_R(x||y) = \sum_{i=1}^d x(i)\log(x(i)/y(i))$ also known as the Kullback-Leibler divergence. The update of OMD is then

$$1 + \log(y_{t+1}(i)) = 1 + \log\theta_t(i) - \eta g_t(i)\,,$$

where $g_t = \nabla\ell_t(\theta_t) \in \mathbb{R}^d$. This can be rewritten

$$y_{t+1}(i) = \theta_t(i)e^{-\eta[\nabla\ell_t(\theta_t)]_i}.$$

The projection to the simplex is a simple renormalization (left as exercise), we thus get

$$\theta_{t+1}(i) = \frac{\theta_t(i)e^{-\eta g_t(i)}}{\sum_{k=1}^d \theta_t(k)e^{-\eta g_t(k)}}\,,$$

and we recover the update of EWA.

3 Online Convex Optimization

In this section, we aim at generalizing the previous algorithms beyond linear losses $\ell_t : \Theta \to \mathbb{R}$.

3.1 Variants of Exponential Weights for continuous action spaces

Here, we present simple solutions via discretization or integration of EWA that yield strong theoretical baselines (for the regret) when the action space is continuous but often yields prohibitive complexity.

3.1.1 Continuous EWA

Let $\Theta \in \mathbb{R}^d$ be a compact and convex subset of $\mathbb{R}^d$. We consider the following continous variant of EWA, that predicts

$$\theta_t = \frac{\int_\Theta \theta e^{-\eta \sum_{s=1}^{t-1} \ell_s(\theta)} d\mu(\theta)}{\int_\Theta e^{-\eta \sum_{s=1}^{t-1} \ell_s(\theta')} d\mu(\theta')} ,$$

where μ is the Lebesgue measure on Θ.

Theorem 3.1 **Regret of continuous EWA**

Let $T \geq 1$. For all sequences of η-exp-concave losses $\ell_1, \dots, \ell_t$ the continuous EWA forecaster satisfies

$$R_T \overset{def}{=} \sum_{t=1}^{T} \ell_t(\theta_t) - \inf_{\theta \in \Theta} \sum_{t=1}^{T} \ell_t(\theta) \leq \frac{1 + d \log(T+1)}{\eta} .$$

Note that if $\Theta = \Delta_d$, then Θ is of dimension $d-1$ and not d. As an exercise, prove a regret upper-bound without the expconcavity assumption of order $O(\sqrt{T})$.

Proof. The proof starts similarly to the one of Theorem 2.4. Let us denote $W_t(\theta) = e^{-\eta \sum_{s=1}^{t} \ell_s(\theta)}$, $W_t = \int_\Theta W_t(\theta) d\mu(\theta)$ and $d\widehat{\mu}_t(\theta) = W_t(\theta) d\mu(\theta) / W_t$. Then,

$$\begin{aligned}
W_T &= \int_\Theta e^{-\eta \sum_{t=1}^{T} \ell_t(\theta)} d\mu(\theta) \\
&= W_{T-1} \int_\Theta \frac{W_{T-1}(\theta)}{W_{T-1}} e^{-\eta \ell_T(\theta)} d\mu(\theta) \\
&= W_{T-1} \int_\Theta e^{-\eta \ell_T(\theta)} d\widehat{\mu}_{T-1}(\theta) && \leftarrow \quad \theta_T = \int_\Theta \theta d\widehat{\mu}_{T-1}(\theta) \\
&\leq W_{T-1} \exp\big(-\eta \ell_T(\theta_T)\big) && \leftarrow \quad \eta\text{-exp-concavity} \\
&\leq W_0 \exp\left(-\eta \sum_{t=1}^{T} \ell_t(\theta_t)\right) , && \leftarrow \quad \text{induction}
\end{aligned} \tag{3.1}$$

The second part of the proof to lower-bound W_T is however less straightforward. For simplicity, let us assume that ℓ_t are continuous on Θ (do the general case as exercise). Therefore the infimum is a minimum and let $\theta^* \in \arg\min_{\theta\in\Theta} \sum_{t=1}^T \ell_t(\theta)$ and define

$$\Theta_\varepsilon \stackrel{\text{def}}{=} \Big\{(1-\varepsilon)\theta^* + \varepsilon\theta, \quad \theta \in \Theta\Big\}, \qquad \varepsilon \in (0,1)\,.$$

Note that by convexity of Θ, $\Theta_\varepsilon \subseteq \Theta$. By expconcavity of ℓ_t, we have for all t and all $\theta = (1-\varepsilon)\theta^* + \varepsilon q$

$$e^{-\eta\ell_t(\theta)} \geq (1-\varepsilon)e^{-\eta\ell_t(\theta^*)} + \varepsilon e^{-\eta\ell_t(q)} \geq (1-\varepsilon)e^{-\eta\ell_t(\theta^*)}\,.$$

Therefore, for all $\theta \in \Theta_\varepsilon$

$$e^{-\eta\sum_{t=1}^T \ell_t(\theta)} \geq (1-\varepsilon)^T e^{-\eta\sum_{t=1}^T \ell_t(\theta^*)}$$

Integrating both parts over Θ_ε and using

$$\mu(\Theta_\varepsilon) = \int_{\Theta_\varepsilon} d\mu(\theta) = \int_\Theta d\mu((1-\varepsilon)\theta^* + \varepsilon\theta) \geq \int_\Theta \det(\varepsilon I_d) d\mu(\theta) = \int_\Theta \varepsilon^d d\mu(\Theta) = \varepsilon^d\mu(\Theta)$$

we get

$$W_T \geq \int_{\Theta_\varepsilon} e^{-\eta\sum_{t=1}^T \ell_t(\theta)} d\mu(\theta) \geq \mu(\Theta)\varepsilon^d(1-\varepsilon)^T e^{-\eta\sum_{t=1}^T \ell_t(\theta^*)}\,.$$

Combining with (3.1), using $W_0 = \mu(\Theta)$, taking the log and reorganizing the terms yields

$$R_T \stackrel{\text{def}}{=} \sum_{t=1}^T \ell_t(p_t) - \sum_{t=1}^T \ell_t(\theta^*) \leq \frac{d\log\frac{1}{\varepsilon} + T\log\frac{1}{1-\varepsilon}}{\eta}\,.$$

Optimizing $\varepsilon = 1/(T+1)$ concludes the proof since

$$T\log\frac{1}{1-\varepsilon} = T\log\left(1+\frac{1}{T}\right) \leq 1\,.$$

□

Though the nice theoretical result, this algorithm is complicated to implement because of the integral. In practice, θ_t can be computed by using $(1/T)$-discretization grid of Θ (bad complexity of order T^d!) or by using Monte-Carlo methods to approximate the integral of log-concave distributions (polynomial time algorithm). We will see in next lectures efficient algorithms with similar guarantees.

Example: Portfolio selection

This algorithm was introduced first for the problem of portfolio selection by **?** and also known as Univeresal Portfolio in this framework. In the latter, given a initial capital Cap_0 a trader repeatedly distributes her capital over d assets with the goal of maximizing the total return. At each round $t = 1, \ldots, T$, the trader chooses an allocation $p_t \in \Delta_d$. Here, $p_t(i)$ represents the share of capital innvested into asset $i \in [d]$ at this round. At the end of the round, the returns–the ratios of the closing and opening pricese in this round–are revealed in the form of $x_t \in \mathbb{R}_+^d$ and the trader's capital is updated as

$$\text{Cap}_t = \text{Cap}_{t-1}\langle p_t, x_t\rangle\,.$$

By Cover, the performance of a strategy that selected portfolios (p_t) is quantified by comparing the final capital $\text{Cap}_T = \text{Cap}_0 \prod_{t=1}^T \langle p_t, x_t\rangle$ against

$$\text{Cap}_T^* = \text{Cap}_0 \max_{p\in\Delta_d}\langle p_t, x_t\rangle\,,$$

the "idealized" final capital attained by the best "static" strategy constrained to select the same portfolio in all rounds. Due to the multiplicative structure, to maximize the capital it is natural to maximize the ratio $\mathrm{Cap}_T/\mathrm{Cap}_T^*$, which is equivalent to minimizing the regret

$$R_T = \sum_{t=1}^{T} \ell_t(p_t) - \min_{p \in \Delta_d} \sum_{t=1}^{T} \ell_t(p)$$

where $\ell_t(p) \overset{\text{def}}{=} -\log\langle p, x_t \rangle$. Noting that the loss is 1-exp-concave, continuous EWA achives the regret upper-bound of Thm 3.1, which yields

$$\mathrm{Cap}_T \geq \frac{\mathrm{Cap}_T^*}{e(T+1)^d},$$

which is optimal in the worst-case.

3.1.2 Discretized EWA

Setting: general compact decision set, β-Hölder loss functions

In this section, we aim at designing a procedure for general compact decision set Θ. We will assume for simplicity that $\Theta \subset \mathbb{R}^d$ with $\max_{\theta,\theta' \in \Theta} \|\theta - \theta'\| \leq D$, where $\|\cdot\|$ denotes the Euclidean norm. If the loss functions ℓ_t are β-Hölder, i.e.,

$$\left|\ell_t(\theta) - \ell_t(\theta')\right| \leq c\|\theta - \theta'\|^{\beta}$$

there exists a simple solution: approximate Θ with a finite discretization grid Θ_ε and apply EWA on Θ_ε. If Θ or the losses are non-convex, one needs to use the random EWA (see Section 2.1.3) and bound the regret with high-probability. For convenience, we will assume Θ and the loss functions ℓ_t to be convex so that the algorithm can play convex combinations of points in Θ_ε and all quantities are deterministic.

Lemma 3.2

Let $\varepsilon > 0$. Let $\Theta \subset \mathbb{R}^d$ such that $\max_{\theta,\theta' \in \Theta} \|\theta - \theta'\| \leq D$. Then, there exists $\Theta_\varepsilon \subset \Theta$ such that

$$\mathrm{Card}(\Theta_\varepsilon) \lesssim \left(\frac{D}{\varepsilon}\right)^d \qquad \textit{and} \qquad \forall x \in \Theta, \exists x' \in \Theta_\varepsilon \quad \|\theta - \theta'\| \leq \varepsilon,$$

where $\lesssim$ denotes a rough inequality (up to multiplicative constants and logarithmic terms).

Remark. *Remark that a set finite Θ_ε which approximate Θ at radius ε, is called an ε-covering of Θ. The cardinal of the smallest ε-covering is called the* covering number *of Θ. This cardinal is heavily used in theory to analyze the complexity of general spaces Θ. It heavily differentiates parametric spaces with covering number of order $(1/\varepsilon)^d$ with nonparametric spaces (spaces of functions) for which the logarithm of the covering number (or metric entropy) is of order $(1/\varepsilon)^d$.*

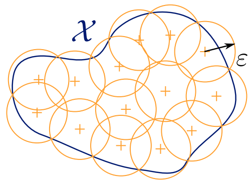
Proof sketch. We only provide the high-level idea of the proof. First, by properties of the Lebesgue measure in d-dimension, denoting $\mathcal{B}_2(r)$ is the ℓ_2-ball of radius $r > 0$, we have

$$\mathrm{Vol}\big(\mathcal{B}_2(r)\big) = \frac{\pi^{d/2}}{\Gamma(n/2+1)} r^d ,$$

where Γ is the Euler's gamma function. Therefore,

$$\mathrm{Vol}(\Theta) \le \mathrm{Vol}\big(\mathcal{B}_2(D/2)\big) = \left(\frac{D}{2\varepsilon}\right)^d \mathrm{Vol}\big(\mathcal{B}_2(\varepsilon)\big) ,$$

and thus approximatively $\left(\frac{D}{2\varepsilon}\right)^d$ balls of radius ε are sufficient to cover Θ.



□

Theorem 3.3 **Discretized EWA**

Let $T \ge 1$, $\varepsilon, D > 0$. Let Θ be a compact convex subset of $\mathbb{R}^d$ such that $\max_{\theta,\theta' \in \Theta} \|\theta - \theta'\| \le D$. Let Θ_ε be an ε-covering of Θ with smallest cardinal. Then, for all sequences of β-Hölder convex losses $\ell_1, \dots, \ell_T : \Theta \longrightarrow [0,1]$, EWA played on the finite set of action Θ_ε with optimized η satisfies the regret bound

$$R_T \overset{def}{=} \sum_{t=1}^T \ell_t(\theta_t) - \min_{\theta\in\Theta} \sum_{t=1}^T \ell_t(\theta) \lesssim \sqrt{Td\Big(\log D + \frac{1}{\beta}\log(cT)\Big)} .$$

Exercise 3.1. *Provide a bound on the expected regret for random EWA when the losses and the decision set are non-convex.*

Proof. Let $d = \mathrm{Card}(\Theta_\varepsilon)$. Let us order the elements of $\Theta_\varepsilon = \{\theta(1), \dots, \theta(d)\}$. Therefore, at time $t \ge 1$, EWA chooses a weight vector $p_t \in \Delta_d$ and predict the weighted average $\theta_t = \sum_{k=1}^d p_t(k)\theta(k) \in \Theta$. Applying the regret bound of EWA, we get

$$\sum_{t=1}^T \sum_{k=1}^d p_t(k)\ell_t(\theta(k)) - \min_{1\le j\le d} \sum_{t=1}^T \ell_t(\theta(j)) \le 2\sqrt{T\log d} . \tag{3.2}$$

Let $\theta^* \in \Theta$ and $\theta(k^*) \in \Theta_\varepsilon$ such that $\|\theta^* - \theta(k^*)\| \le \varepsilon$. Because the losses are β-Hölder and convex, we have

$$\begin{aligned}
R_T \quad &= \quad \sum_{t=1}^T \ell_t(\theta_t) - \ell_t(\theta^*) \\
&\overset{\text{Convexity}}{\le} \quad \sum_{t=1}^T \sum_{k=1}^d p_t(k)\ell_t(\theta(k)) - \ell_t(\theta^*) \qquad \leftarrow \theta_t = \sum_{k=1}^d p_t(k)\theta(k) \\
&\le \quad \sum_{t=1}^T \sum_{k=1}^d p_t(k)\ell_t(\theta(k^*)) - \ell_t(\theta(k^*)) + \sum_{t=1}^T \big|\ell_t(\theta(k^*)) - \ell_t(\theta^*)\big| \\
&\overset{(3.2)}{\le} \quad 2\sqrt{T\log d} \qquad + \qquad T \max_{1\le t\le T} \big|\ell_t(\theta^*) - \ell_t(\theta(k^*))\big| \\
&\overset{\beta\text{-Hölder}}{\le} \quad 2\sqrt{T\log d} + cT\varepsilon^\beta \\
&\overset{\text{Lem. 3.2}}{\lesssim} \quad \sqrt{Td\log\left(\frac{D}{\varepsilon}\right)} + cT\varepsilon^\beta .
\end{aligned}$$

Optimizing $\varepsilon^\beta = 1/cT$, hence $\varepsilon = (cT)^{-1/\beta}$, we get

$$R_T \lesssim \sqrt{Td\Big(\log D + \frac{1}{\beta}\log(cT)\Big)}\,.$$

□

Though this algorithm is theoretically convenient since it can deals with general compact sets Θ and general loss functions (which can be non-convex and non-differentiable). It suffers two considerable drawbacks:

- *computational complexity*: the algorithm needs to consider a discretization space of cardinal $(X/\varepsilon)^d$ which is of order $O(T^{d/\beta})$. This is prohibitive in practice.
- *bad regret dependence on the dimension*: the regret bound is of order $O(\sqrt{dT\log T})$. We will see how to have no dependence on d when Θ is bounded in ℓ_2-norm.

3.2 The Gradient Trick (from linear to convex losses)

Setting: compact convex decision set Θ, convex and sub-differentiable loss functions.

Here, we show how to generalize Chapter 2 to sub-differentiable loss functions $\ell_t : \Theta \subset \mathbb{R}^d \to \mathbb{R}$.

Definition 3.1 **Sub-gradient**

A sub-gradient of a convex function ℓ_t at point $\theta \in \Theta$ is a point in $\mathbb{R}^d$, denoted $\nabla\ell_t(\theta)$, that sastifies the convexity inequality:

$$\forall \theta' \in \Theta, \qquad \ell_t(\theta) - \ell_t(\theta') \le \langle \nabla\ell_t(\theta), \theta - \theta'\rangle\,.$$

To do so, we introduce the gradient trick that consist in linearizing the losses to apply the results of the previous section. The idea is to apply previous algorithms on the linear losses $\langle g_t, \theta\rangle$ with $g_t = \nabla\ell_t(\theta_t)$. In this case, we have from the definition of the sub-gradients,

$$R_T = \sum_{t=1}^T \ell_t(\theta_t) - \ell_t(\theta^*) \le \sum_{t=1}^T \langle g_t, \theta_t\rangle - \langle g_t, \theta^*\rangle\,,$$

where $\theta^* = \arg\min_{\theta\in\Theta} \sum_{t=1}^T \ell_t(\theta)$. Hence, upper-bouding the regret with linear losses $\theta \mapsto \langle g_t, \theta\rangle$ also upper-bound the regret with the true losses ℓ_t. The following theorem follows.

Theorem 3.4 **Gradient trick**

Let $\mathcal{A}$ be an algorithm that satisfies the regret-upper bound

$$\sum_{t=1}^T \langle g_t, \theta_t\rangle - \langle g_t, \theta^*\rangle \le R(T, (g_t), \Theta)$$

for some real-valued function R. Then, for any sequence of sub-differentiable losses $\ell_t : \Theta \to \mathbb{R}$, applying $\mathcal{A}$ with $g_t = \nabla\ell_t(\theta_t)$ achieves the regret upper-bound

$$R_T \overset{def}{=} \sum_{t=1}^T \ell_t(\theta) - \min_{\theta\in\Theta} \sum_{t=1}^T \ell_t(\theta) \le R\big(T, (\nabla\ell_t(\theta_t)), \Theta\big)\,.$$

This theorems shows that all results (Theorems 2.1, 2.7, 2.8 and 2.10) of Chapter 2 can be transposed to convex loss functions $\ell_t : \Theta \to \mathbb{R}$. Applyed with $g_t = \nabla \ell_t(\theta_t)$, we get the following regret upper-bounds:

$$
\begin{aligned}
R_T &\leq 2G_\infty \sqrt{T \log d} && \text{(EWA, Thm 2.1)}\\
R_T &\leq D_2 G_2 \sqrt{T} && \text{(OGD, Thm. 2.7)}\\
R_T &\leq 2D_R G_{R,*} \sqrt{2\alpha^{-1} T} && \text{(RFTL, Thm. 2.8)}
\end{aligned}
$$

The EG Algorithm When the decision space is the simplex $\Theta = \Delta_d$, applying the above trick to EWA results in an algorithm called Exponentiated Gradient forecaster (EG). Recall that in this case we denote the decision $\theta_t = p_t$.

Example 3.1 (Prediction with expert advice (continued)). *In prediction with expert advice, a sequence of observations $y_1, \dots, y_T \in [0,1]$ is to be predicted with the help of d expert advice $x_t(k) \in [0,1]$ for $1 \leq k \leq d$. The learner predict $\widehat{y}_t = \sum_{k=1}^d p_t(k) x_t(k)$ and suffers a loss $\ell(\widehat{y}_t, y_t)$. If the loss function is convex and Lipschitz in its first argument we can apply Theorem ?? with $\ell_t : p \mapsto \ell(p \cdot x_t, y_t)$. For instance, with the absolute loss, $G = 1$ and EG satisfies a bounded regret with respect to any fixed convex combination of experts:*

$$
\sum_{t=1}^T |\widehat{y}_t - y_t| - \min_{p \in \Theta} \sum_{t=1}^T \Big| p \cdot x_t - y_t \Big| \leq 2\sqrt{T \log d}\,.
$$

Hence, on the long run we perform as good as the best convex combination of the experts which may outperform the best expert. This may leads to much better performance than a simple EWA on the experts if

$$
\min_{p \in \Theta} \sum_{t=1}^T \Big| p \cdot x_t - y_t \Big| \ll \min_{k \in [d]} \sum_{t=1}^T \Big| x_t(k) - y_t \Big|.
$$

Convex hull of finite point set It is worth pointing out that the simplex decision set Δ_d can be generalized with any convex hull of a finite point set $S = \{\theta(1), \dots, \theta(d)\}$:

$$
\mathrm{Conv}(S) = \left\{ \sum_{i=1}^d p_i \theta(i) : \forall i, p_i > 0 \text{ and } \sum_{i=1}^d p_i = 1 \right\}.
$$

Transforming the loss functions, EWA can be applied to compete with such sets as shown by the theorem bellow.

Theorem 3.5

Let $T \geq 1$. Let $\Theta \subset \mathbb{R}^d$ be a convex set and $S = \{\theta(1), \dots, \theta(d)\} \in \Theta^d$ with diameter $D_1 \geq \max_{i,j} \|\theta(i) - \theta(j)\|_1$. Let $\ell_1, \dots, \ell_T : \Theta \to \mathbb{R}$ be an arbitrary sequence of convex differentiable losses with bounded gradient $\max_{\theta \in \Theta} \|\nabla \ell_t(\theta)\|_\infty \leq G_\infty$. Then, EWA applied with $g_t = \nabla \tilde{\ell}_t$ where $\tilde{\ell}_t : p \mapsto \ell_t \left(\sum_{i=1}^d p(i)\theta(i) \right)$ achieves the regret bound

$$
R_T \overset{def}{=} \sum_{t=1}^T \ell_t(\theta_t) - \min_{\theta \in Conv(S)} \sum_{t=1}^T \ell_t(\theta) \leq 2G_\infty D_1 \sqrt{T \log d}\,,
$$

where $\theta_t = \sum_{k=1}^d p_t(k)\theta(k)$

Such a trick can be used for instance to compete with the ℓ_1-balls using $S = \{\theta \in \mathbb{R}^d : \|\theta\|_1 = R, \|x\|_0 = 1\}$. Since ℓ_p-balls are contained into the ℓ_1-ball (of possibly larger radius depending on p) this can also be used to compete against any ℓ_p-ball for $p \geq 1$. This trick was introduced by Kivinen and Warmuth [1997] for the EG± forecaster.

4 Adversarial Bandits

In previous chapters, we considered the full-information feedback and the bandit feedback with stochastic loss functions. In *full information with finite decision set* $\Theta = [K] \overset{\text{def}}{=} \{1, \dots, K\}$, we saw the Random Exponentially Weighted Average (EWA) forecaster. It is defined as

$$p_t(k) = \frac{e^{-\eta \sum_{s=1}^{t-1} \ell_s(k)}}{\sum_{j=1}^{K} e^{-\eta \sum_{s=1}^{t-1} \ell_s(j)}} . \tag{EWA}$$

and draws $\theta_t = k$ with probability $p_t(k)$. If $-\eta \ell_t(j) \leq 1$ (see the proof of EWA in first lecture), it satisfies the upper-bound:

$$\sum_{t=1}^{T} p_t \cdot \ell_t - \min_{1 \leq j \leq K} \sum_{t=1}^{T} \ell_t(j) \leq \eta \sum_{t=1}^{T} \sum_{k=1}^{K} p_t(k) \ell_t(k)^2 + \frac{\log K}{\eta} . \tag{$*$}$$

Since the decision θ_t is random, we assume that ℓ_t cannot depend on θ_t but may depend on past information $\sigma(p_1, \ell_1, x_1, \dots, x_{t-1}, p_t)$. The above bound can be converted into a bound on the expected regret for well-calibrated learning rate η

$$\mathbb{E}\big[R_T\big] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\theta_t) - \min_{k \in [K]} \sum_{t=1}^{T} \ell_t(k)\right] \leq 2\sqrt{T \log K} .$$

In this chapter, we will see adversarial bandits: that is bandit feedback (only $\ell_t(\theta_t)$ is observed at the end of round t by the player) with an adversarial sequence of loss function ℓ_t (i.e., no stochastic assumptions). Note that we turn back to losses instead of rewards but we will come back to rewards whenever it makes the proof easier. Remember that the lower-bound on the regret in the worst-case is of order $O(\sqrt{TK})$.

4.1 Adversarial multi-armed bandits

We consider Setting 1.1 with bandit feedback, finite decision space $\Theta = [K] \overset{\text{def}}{=} \{1, \dots, K\}$ and adversarial losses. To emphasize that the action is in $[K]$, we denote by k_t the action chosen by the player (instead of θ_t). We do not assume the loss functions ℓ_t to be linear nor convex (the decision space is not). Similarly to Random EWA the chosen action $k_t \in [K]$ is sampled randomly from a distribution p_t chosen at round t by the player. We will provide an algorithm called Exp3 inspired by EWA.

4.1.1 Pseudo-regret bound

Let us denote the regret with respect to action $k \in [K]$ by

$$R_T(k) \overset{\text{def}}{=} \sum_{t=1}^{T} \ell_t(k_t) - \sum_{t=1}^{T} \ell_t(k) .$$

Instead of minimizing the *expected regret* $\mathbb{E}[R_T] = \mathbb{E}[\max_k R_T(k)]$, we will start with an easier objective, the *pseudo-regret* defined as

$$\bar{R}_T \overset{\text{def}}{=} \max_{k\in[K]} \mathbb{E}\big[R_T(k)\big] = \max_{k\in[K]} \mathbb{E}\left[\sum_{t=1}^T \ell_t(k_t) - \sum_{t=1}^T \ell_t(k)\right]. \qquad \text{(pseudo regret)}$$

It is worth pointing out that the expectations are taken with respect to the randomness of the algorithm: the decisions k_t are random. We can distinguish two types of adversaries:

- *oblivious adversary*: all the loss functions $\ell_1, \dots, \ell_t$ are chosen in advance before the game starts and do not depend on the past player decisions $k_1, \dots, k_T$. In this case, the losses $\ell_t(k)$ are determinist and there is thus equality: $\bar{R}_T = \mathbb{E}[R_T]$.
- *adaptive adversary*: the loss function ℓ_t at round $t \geq 1$ may depend on past information $\sigma(k_1, \dots, k_{t-1})$. It is thus random. By Jensen's inequality $\max_{k\in[K]} \mathbb{E}\big[R_T(k)\big] \leq \mathbb{E}\big[\max_{k\in[K]} R_T(k)\big]$ and thus $\bar{R}_T \leq \mathbb{E}[R_T]$.

The EXP3 algorithm Ideally, we would like to reuse our algorithm EWA that assigned weights

$$\forall k \in [K], \qquad p_t(k) = \frac{e^{-\eta \sum_{s=1}^{t-1} \ell_s(k)}}{\sum_{j=1}^K e^{-\eta \sum_{s=1}^{t} \ell_s(j)}} . \qquad \text{(EWA)}$$

Unfortunately this is not possible since the player does not observe $\ell_t(k)$ for $k \neq k_t$. The high-level idea of Exp3 is to replace $\ell_t(k)$ with an unbiased estimate that is observed by the player. A first idea would be to use $\ell_t(k)$ if we observe it and 0 otherwise:

$$g_t(k) = \begin{cases} \ell_t(k) & \text{if } k = k_t \quad \leftarrow \text{i.e., decision } k \text{ is observed} \\ 0 & \text{otherwise} \end{cases} .$$

However, this estimate is biased:

$$\mathbb{E}_{k_t\sim p_t}\big[g_t(k_t)\big] = p_t(k)\ell_t(k) \;\neq \ell_t(k) .$$

In other words, the actions that are less likely to be chosen by the algorithm (small weight $p_t(k)$) are more likely to be unobserved and incur 0 loss. We need to correct this phenomenon. Therefore we choose

$$g_t(k) = \frac{\ell_t(k)}{p_t(k)} \mathbb{1}\{k = k_t\} , \qquad (4.1)$$

which leads to the algorithm EXP3 detailed below.

```
EXP3
Parameter: $\eta > 0$
Initialize: $p_1 = \big(\frac{1}{K}, \dots, \frac{1}{K}\big)$
For $t = 1, \dots, T$
    – draw $k_t \sim p_t$; incur loss $\ell_t(k_t)$ and observe $\ell_t(k_t) \in [0,1]$;
    – update for all $k \in \{1, \dots, K\}$

        $p_{t+1}(k) = \frac{e^{-\eta \sum_{s=1}^t g_s(k)}}{\sum_{j=1}^K e^{-\eta \sum_{s=1}^t g_s(j)}}$ ,   where $g_s(k) = \frac{\ell_s(k)}{p_s(k)} \mathbb{1}\{k = k_s\}$
```

Then applying the Inequality (*) for EWA with the substituted losses g_t, we get the following theorem.

Theorem 4.1

Let $T \geq 1$. The pseudo-regret of EXP3 run with $\eta = \sqrt{\frac{\log K}{KT}}$ is upper-bounded as:

$$\bar{R}_T \leq 2\sqrt{KT \log K}\,.$$

Proof. Apply EWA to the estimated losses $g_t(j)$ that are completely observed (nonnegative but not bounded), we get from Inequality (*) and taking the expectation:

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle p_t, g_t \rangle - \min_{j \in [K]} \sum_{t=1}^{T} g_t(j)\right] \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^{T} \mathbb{E}\Big[\langle p_t, g_t \rangle^2\Big]\,. \tag{4.2}$$

Now we compute the expectations. Denote by $\mathcal{F}_{t-1} \overset{\text{def}}{=} \sigma(p_1, \ell_1, k_1, \ldots, k_{t-1}, p_t, \ell_t)$ the past information available at round t for the adversary (which cannot use the randomness of k_t but can use p_t). Note that ℓ_t and p_t are $\mathcal{F}_{t-1}$-measurable by assumption. We have

$$\forall j \in [K] \qquad \mathbb{E}\Big[g_t(j)\Big|\mathcal{F}_{t-1}\Big] = \mathbb{E}\left[\frac{\ell_t(j)}{p_t(j)}\mathbb{1}\{j = k_t\}\middle|\mathcal{F}_{t-1}\right] = \sum_{k=1}^{K} p_t(k)\frac{\ell_s(j)}{p_t(j)}\mathbb{1}\{j = k\} = \ell_t(j)$$

thus the estimated losses are unbiased $\mathbb{E}\big[g_t(j)\big] = \mathbb{E}\big[\ell_t(j)\big]$ and

$$\begin{aligned}
\mathbb{E}\Big[\langle p_t, g_t \rangle\Big] = \mathbb{E}\left[\sum_{j=1}^{K} p_t(j) g_t(j)\right] &= \mathbb{E}\left[\sum_{j=1}^{K} p_t(j)\mathbb{E}\Big[g_t(j)\Big|\mathcal{F}_{t-1}\Big]\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{K} p_t(j)\ell_t(j)\right] = \mathbb{E}\Big[\mathbb{E}\big[\ell_t(k_t)\big|\mathcal{F}_{t-1}\big]\Big] = \mathbb{E}\big[\ell_t(k_t)\big]\,.
\end{aligned}$$

Therefore, we can lower-bound the left-hand side:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \langle p_t, g_t \rangle - \min_{j \in [K]} \sum_{t=1}^{T} g_t(j)\right] &\geq \max_{j \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t, g_t \rangle - \sum_{t=1}^{T} g_t(j)\right] \\
&= \max_{j \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(k_t) - \sum_{t=1}^{T} \ell_t(j)\right] = \bar{R}_T\,.
\end{aligned}$$

On the other hand, the expectation of the right-hand side satisfies

$$\begin{aligned}
\mathbb{E}\big[\langle p_t, g_t \rangle^2\big] = \mathbb{E}\left[\sum_{j=1}^{K} p_t(j) g_t(j)^2\right] &= \mathbb{E}\left[\sum_{j=1}^{K} p_t(j)\,\mathbb{E}\Big[g_t(j)^2\,\Big|\,\mathcal{F}_{t-1}\Big]\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{K}\sum_{k=1}^{K} p_t(j) p_t(k)\left(\frac{\ell_t(j)}{p_t(j)}\mathbb{1}\{j = k\}\right)^2\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{K}\sum_{k=1}^{K} p_t(k)\frac{\ell_t(j)^2}{p_t(j)}\mathbb{1}\{j = k\}\right]
\end{aligned}$$

$$= \mathbb{E}\left[\sum_{j=1}^{K} \ell_t(j)^2\right] \le K\,.$$

Substituting into Inequality (4.2) yields

$$\bar{R}_T \le \frac{\log K}{\eta} + \eta K T\,.$$

and optimizing $\eta = \sqrt{KT/(\log K)}$ concludes. □

The issue with the above regret bound is that it bounds the pseudo-regret and not the expected regret. This is because we have

$$\mathbb{E}\left[\min_j \sum_{t=1}^{T} g_t(j)\right] \le \min_j \mathbb{E}\left[\sum_{t=1}^{T} g_t(j)\right] = \min_{j\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(j)\right]$$

but not

$$\mathbb{E}\left[\min_j \sum_{t=1}^{T} g_t(j)\right] \not\le \mathbb{E}\left[\min_j \sum_{t=1}^{T} \ell_t(j)\right]. \tag{4.3}$$

Hence, controlling the cumulative loss agains the best estimated action only controls the pseudo regret and not the true regret.

4.1.2 High probability bound on the regret

Gains versus losses In this part, we will switch the analysis from losses $\ell_t(k)$ to gains $g_t(k) = 1 - \ell_t(k) \in [0,1]$ because the core idea of the next algorithm is easier to see with gains. Remark that the loss and gain versions are symmetric via the transformation $g_t(k) = 1 - \ell_t(k)$. The regret in terms of gains is defined as

$$R_T \overset{\text{def}}{=} \max_{k\in[K]} \sum_{t=1}^{T} g_t(k) - \sum_{t=1}^{T} g_t(k_t)\,.$$

Using EWA with full information from (*), if $\eta g_t(k) \le 1$, we also have for gains the inequality

$$\max_{1\le j\le K} \sum_{t=1}^{T} g_t(j) - \sum_{t=1}^{T} \langle p_t, g_t\rangle \le \eta \sum_{t=1}^{T} p_t \cdot g_t^2 + \frac{\log K}{\eta}\,, \quad \text{where} \quad p_t(k) = \frac{e^{\eta \sum_{s=1}^{t-1} g_s(k)}}{\sum_{j=1}^{K} e^{\eta \sum_{s=1}^{t-1} g_s(j)}}\,. \tag{4.4}$$

High-level idea of EXP3.P The high-level idea of the next algorithm is to ensure that the estimators $\widehat{g}_t(k)$ of the gains satisfy

$$\mathbb{E}\left[\max_j \sum_{t=1}^{T} \widehat{g}_t(j)\right] \ge \mathbb{E}\left[\max_j \sum_{t=1}^{T} g_t(j)\right] \tag{4.5}$$

so that controlling the performance with respect to the estimated gains (left-hand side) also controls the performance with respect to the true gains (right-hand side). This was not the case of the estimators used for EXP3 (see (4.3)). To ensure (4.5), we add a bias term β to the estimators $\widehat{g}_t(k)$ as follows:

$$\widehat{g}_t(k) \overset{\text{def}}{=} \frac{g_t(k)\mathbb{1}\{k = k_t\} + \beta}{p_t(k)} \tag{4.6}$$

In contrary to (4.1), the estimator is indeed biased

$$\mathbb{E}\big[\widehat{g}_t(k)\big|\mathcal{F}_{t-1}\big] = g_t(k) + \frac{\beta}{p_t(k)}\,, \tag{4.7}$$

where we recall that $\mathcal{F}_{t-1} \overset{\text{def}}{=} \sigma\big(p_1, k_1, g_1, \dots, k_{t-1}, p_t, g_t\big)$ contains the information up to time t available to the environment. We have the following Lemma:

Lemma 4.2

For any $\delta > 0$, with probability $1-\delta$ and $\beta \in (0,1)$,

$$\sum_{t=1}^{T} \widehat{g}_t(j) \geq \sum_{t=1}^{T} g_t(j) - \frac{\log(1/\delta)}{\beta}\,.$$

Proof. Let $\beta \in (0,1)$, from Markov's inequality, we have

$$\begin{aligned}\mathbb{P}\bigg(\sum_{t=1}^{T} \widehat{g}_t(j) \geq \sum_{t=1}^{T} g_t(j) - \frac{\log(1/\delta)}{\beta}\bigg) &= \mathbb{P}\bigg(\exp\Big(\beta\sum_{t=1}^{T}\big(g_t(j) - \widehat{g}_t(j)\big)\Big) \geq \delta^{-1}\bigg)\\ &\leq \delta\mathbb{E}\bigg[\exp\Big(\beta\sum_{t=1}^{T}\big(g_t(j) - \widehat{g}_t(j)\big)\Big)\bigg]\,.\end{aligned}$$

It only remains to upper-bound the expectation in the right-hand side by 1, which we do now. Since $\beta \in (0,1)$ and $\widehat{g}_t(j) \geq \beta/p_t(j)$, we have $\beta(g_t(j) - \widehat{g}_t(j) + \beta/p_t(j)) \leq 1$. Therefore, we can use the inequality $e^x \leq 1 + x + x^2$ for $x \leq 1$, which entails

$$\begin{aligned}&\mathbb{E}\bigg[\exp\Big(\beta\big(g_t(j) - \widehat{g}_t(j)\big)\Big)\bigg|\mathcal{F}_{t-1}\bigg] = \mathbb{E}\bigg[\exp\Big(\beta\Big(g_t(j) - \widehat{g}_t(j) + \frac{\beta}{p_t(j)}\Big)\Big)\bigg|\mathcal{F}_{t-1}\bigg]\exp\Big(-\frac{\beta^2}{p_t(j)}\Big)\\ &\leq \mathbb{E}\bigg[\bigg(1 + \cancel{\beta\Big(g_t(j) - \widehat{g}_t(j) + \frac{\beta}{p_t(j)}\Big)} + \beta^2\Big(g_t(j) - \widehat{g}_t(j) + \frac{\beta}{p_t(j)}\Big)^2\bigg)\bigg|\mathcal{F}_{t-1}\bigg]e^{-\frac{\beta^2}{p_t(j)}}\\ &\overset{(4.7)}{=} \bigg(1 + \beta^2\mathbb{E}\bigg[\Big(g_t(j) - \widehat{g}_t(j) + \frac{\beta}{p_t(j)}\Big)^2\bigg|\mathcal{F}_{t-1}\bigg]\bigg)e^{-\frac{\beta^2}{p_t(j)}}\end{aligned}$$

where the last equality is by (4.7) and because $p_t(j)$ is $\mathcal{F}_{t-1}$-measurable. Now,

$$\begin{aligned}\mathbb{E}\bigg[\Big(g_t(j) - \widehat{g}_t(j) + \frac{\beta}{p_t(j)}\Big)^2\bigg|\mathcal{F}_{t-1}\bigg] &= \mathrm{Var}\Big(\widehat{g}_t(j)\Big|\mathcal{F}_{t-1}\Big) = \mathrm{Var}\Big(\frac{g_t(j)\mathbb{1}\{j = k_t\}}{p_t(j)}\Big|\mathcal{F}_{t-1}\Big)\\ &\leq \mathbb{E}\bigg[\bigg(\frac{g_t(j)\mathbb{1}\{j = k_t\}}{p_t(j)}\bigg)^2\bigg|\mathcal{F}_{t-1}\bigg] \leq \mathbb{E}\bigg[\frac{\mathbb{1}\{j = k_t\}}{p_t(j)^2}\bigg|\mathcal{F}_{t-1}\bigg] = \sum_{k=1}^{K}\frac{p_t(k)\mathbb{1}\{j = k\}}{p_t(j)^2} = \frac{1}{p_t(j)}\,.\end{aligned}$$

Substituting into the previous inequality and using $1 + x \leq e^x$, it yields

$$\mathbb{E}\bigg[\exp\Big(\beta\big(g_t(j) - \widehat{g}_t(j)\big)\Big)\bigg|\mathcal{F}_{t-1}\bigg] \leq \bigg(1 + \frac{\beta^2}{p_t(j)}\bigg)e^{-\beta^2/p_t(j)} \leq 1\,.$$

The proof is concluded by induction

$$\mathbb{E}\left[\exp\left(\beta\sum_{t=1}^{T}\big(g_t(j)-\widehat{g}_t(j)\big)\right)\right]=\mathbb{E}\left[\underbrace{\mathbb{E}\left[\exp\Big(\beta\big(g_T(j)-\widehat{g}_T(j)\big)\Big)\middle|\mathcal{F}_{T-1}\right]}_{\leq 1}\exp\left(\beta\sum_{t=1}^{T-1}\big(g_t(j)-\widehat{g}_t(j)\big)\right)\right]$$

$$\leq\mathbb{E}\left[\exp\left(\beta\sum_{t=1}^{T-1}\big(g_t(j)-\widehat{g}_t(j)\big)\right)\right]\leq\ldots\leq 1\,.$$

□

The issue with the estimators $\widehat{g}_t(j)\in(0,+\infty)$ defined in Equation (4.6) is that they might be unbounded if the weights $p_t(j)$ are close to zero. The condition $\eta\widehat{g}_t(j)\leq 1$ which appeared in the proof of EWA cannot hold for any $\eta>0$. Remark that this was not a problem for EXP3 with the preceding choice (4.1) because $-\eta g_t(j)\leq 1$ (see the proof of EWA for details).

The next algorithm called EXP3.P, is close to EXP3 but ensures the weights do not vanish to zero by adding an exploration parameter $\gamma>0$.

```
EXP3.P
Parameters: $\eta > 0, \beta \in (0,1), \gamma \in (0,1)$
Initialize: $p_1 = \left(\frac{1}{K},\ldots,\frac{1}{K}\right)$
For $t = 1,\ldots,T$
    – draw $k_t \sim p_t$; receive gain $g_t(k_t) = 1-\ell_t(k_t)$ and observe $g_t(k_t) \in [0,1]$;
    – update for all $k \in \{1,\ldots,K\}$
          $p_{t+1}(k) = (1-\gamma)\frac{e^{\eta\sum_{s=1}^t \widehat{g}_s(k)}}{\sum_{j=1}^K e^{\eta\sum_{s=1}^t \widehat{g}_s(j)}} + \frac{\gamma}{K}$,
      where $\widehat{g}_s(k) = \frac{g_s(k)\mathbb{1}\{k=k_s\}+\beta}{p_s(k)}$.
```

The weights $p_t(k)$ of EXP3.P are necessary larger than γ/K and thus $|\eta g_t(j)|\leq 1$ as soon as $\eta(1+\beta)K/\gamma\leq 1$. We get the following high-probability bound on the regret.

Theorem 4.3

For well-chosen parameters $\gamma\in(0,1)$, $\beta\in(0,1)$ and $\eta>0$ satisfying $\eta(1+\beta)K/\gamma\leq 1$, for any $\delta>0$, the EXP3.P algorithm achieves

$$R_T\leq 6\sqrt{TK\log K}+\sqrt{\frac{TK}{\log K}}\log(1/\delta)\,.$$

with probability at least $1-\delta$.

Remark that the above bound leads to a bound on the expected regret, with the choice $\delta=1/T$ it yields

$$\mathbb{E}\big[R_T\big]\leq 6\sqrt{TK\log K}+\sqrt{\frac{TK}{\log K}}\log(T)+1$$

The logarithmic dependency on T can even be removed using $\mathbb{E}[X]=\int_0^\infty\mathbb{P}(X\geq\varepsilon)d\varepsilon$.

Proof of Theorem 4.3. Defining the weights that would assign EXP3,

$$q_t(j) \overset{\text{def}}{=} \frac{e^{\eta \sum_{s=1}^{t-1} \widehat{g}_s(j)}}{\sum_{k=1}^{K} e^{\eta \sum_{s=1}^{t-1} \widehat{g}_s(k)}},$$

we get from Inequality (4.4) applied with $\widehat{g}_t(j)$,

$$\max_{j\in[K]} \sum_{t=1}^{T} \widehat{g}_t(j) \leq \sum_{t=1}^{T} q_t \cdot \widehat{g}_t + \eta \sum_{t=1}^{T} q_t \cdot \widehat{g}_t^2 + \frac{\log K}{\eta}.$$

where we used $\eta \widehat{g}_t(j) \leq 1$ because $\eta(1+\beta)K/\gamma \leq 1$. Now, we use that $p_t \overset{\text{def}}{=} (1-\gamma)q_t + \gamma/K$, which entails $q_t = (p_t - \gamma/K)/(1-\gamma) \leq p_t/(1-\gamma)$. Substituting into the above inequality

$$(1-\gamma) \max_{j\in[K]} \sum_{t=1}^{T} \widehat{g}_t(j) \leq \sum_{t=1}^{T} p_t \cdot \widehat{g}_t + \eta \sum_{t=1}^{T} p_t \cdot \widehat{g}_t^2 + \frac{\log K}{\eta}. \tag{4.8}$$

But by definition of $\widehat{g}_t$,

$$p_t \cdot \widehat{g}_t = \sum_{j=1}^{K} p_t(j)\widehat{g}_t(j) = \sum_{j=1}^{K} \big(g_t(j)\mathbb{1}\{j = k_t\} + \beta\big) = g_t(k_t) + K\beta.$$

and since $p_t(j)\widehat{g}_t(j) \leq (1+\beta)$,

$$\sum_{t=1}^{T} p_t \cdot \widehat{g}_t^2 \leq (1+\beta) \sum_{j=1}^{K} \sum_{t=1}^{T} \widehat{g}_t(j) \leq K(1+\beta) \max_{j\in[K]} \sum_{t=1}^{T} \widehat{g}_t(j) \leq \frac{\gamma}{\eta} \max_{j\in[K]} \sum_{t=1}^{T} \widehat{g}_t(j).$$

Therefore, substituting into Inequality (4.8) gives

$$(1-\gamma) \max_{j\in[K]} \sum_{t=1}^{T} \widehat{g}_t(j) \leq \sum_{t=1}^{T} g_t(k_t) + K\beta T + \gamma \max_{j\in[K]} \sum_{t=1}^{T} \widehat{g}_t(j) + \frac{\log K}{\eta},$$

where we used $(1+\beta)K \leq \gamma/\eta$. Reorganizing, we get

$$(1-2\gamma) \max_{j\in[K]} \sum_{t=1}^{T} \widehat{g}_t(j) \leq \sum_{t=1}^{T} g_t(k_t) + K\beta T + \frac{\log K}{\eta}.$$

Using Lemma 4.2 together with a union bound (to have it for all $j \in [K]$), we have with probability $1-\delta$

$$(1-2\gamma)\left(\max_{j\in[K]} \sum_{t=1}^{T} g_t(j) - \frac{\log(K/\delta)}{\beta} \right) \leq \sum_{t=1}^{T} g_t(k_t) + K\beta T + \frac{\log K}{\eta},$$

and thus reorganizing and choosing $\gamma \overset{\text{def}}{=} 2\eta K \geq \eta(1+\beta)K$,

$$\max_{j\in[K]} \sum_{t=1}^{T} g_t(j) - \sum_{t=1}^{T} g_t(k_t) \leq K\beta T + \frac{\log K}{\eta} + \frac{\log(K/\delta)}{\beta} + 4\eta K T.$$

The proof is concluded by optimizing $\eta \overset{\text{def}}{=} (1/2)\sqrt{(\log K)/KT}$ and $\beta \overset{\text{def}}{=} \sqrt{(\log K)/(KT)}$. □

4.2 Contextual adversarial multi-armed bandits

Here, we consider the adversarial bandit framework in which the learner has access to some external information (context) before making its decision. We turn back to the loss version of the game.

4.2.1 Adversarial multi-armed bandits with experts

We now consider prediction with expert advice in the bandit framework. The setting is the same as the one described in Figure 1.1, but at the beginning of each round $t \geq 1$, some experts $i = 1, \dots, N$ propose recommendations $h_t(i) \in [K]$. These recommendations may be random and may depend on past actions k_s, $s \leq t-1$ and past observations $\ell_s(k_s)$. The loss of each expert is given by the loss of the chosen decision $\ell_t\big(h_t(i)\big)$ but only $\ell_t(k_t)$ is observed by the learner. The goal of the learner is then to be competitive with the best expert on a long run. To do so, it minimizes the pseudo-regret

$$R_T^{\text{exp}} \stackrel{\text{def}}{=} \max_{i=1,\dots,N} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(k_t) - \sum_{t=1}^{T} \ell_t\big(h_t(i)\big)\right]$$

with respect to the experts. In order to bound the pseudo-regret, one could consider experts as the set of arms and use EXP3. This would give a bound of order $\sqrt{TN \log N}$. However it does not take into account the information on the reward of all experts that choose the same action $h_t(i) = k_t$.

```
EXP4
Parameter: $\eta > 0$
Initialize: $q_1 = \big(\frac{1}{N}, \dots, \frac{1}{N}\big)$.
For each round $t = 1, \dots, n$
    1. Get expert advice $h_t(1), \dots, h_t(N) \in [K]$
    2. Draw an expert $i_t$ with probability distribution $q_t \in \Delta_N$
    3. Choose decision $k_t = h_t(i_t)$
    4. Compute the estimated loss for each decision
           $$g_t(k) = \frac{\ell_t(k)}{p_t(k)} \mathbb{1}\{k = k_t\},$$
       where $p_t \stackrel{\text{def}}{=} \sum_{i=1}^{N} q_t(i) \delta_{\ell_t(i)} \in \Delta_K$.
    5. Compute the estimated loss of the experts component-wise $g_t(h_t(i))$
    6. Update the probability distribution over the experts component-wise
           $$q_{t+1}(i) = \frac{\exp\Big(-\eta \sum_{s=1}^{t} g_s\big(h_s(i)\big)\Big)}{\sum_{j=1}^{N} \exp\Big(\eta \sum_{s=1}^{t} g_s\big(h_s(j)\big)\Big)}, \qquad \forall 1 \leq i \leq N.$$
```

Theorem 4.4

EXP4 with $\eta = \sqrt{\log N/(KT)}$ satisfies $R_T^{exp} \leq 2\sqrt{TK \log N}$.

Similarly to the variant EXP3.P, we can define a variant EXP4.P to bound the regret with high probability (and thus the expected regret). Furthermore, the above algorithm (and theorem) can be extended to the case where expert advice are distributions $h_t(i) \in \Delta_K$. The algorithm is the same by sampling k_t according to $h_t(i_t)$ and assigning to expert i the loss $\sum_{k=1}^{K} h_t(i)(k) g_t(k)$.

Proof. We can apply the analysis of EXP to a learner using distribution q_t over N actions (here experts) with (full-information) losses $g_t\big(h_t(i)\big)$ for $i \in \{1,\dots,N\}$. We get from Inequality (*)

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} q_t(i)\cdot g_t\big(h_t(i)\big) - \min_{1\le j\le N}\sum_{t=1}^{T} g_t\big(h_t(j)\big)\right] \le \eta\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\Big[q_t(i)g_t\big(h_t(i)\big)^2\Big] + \frac{\log N}{\eta}\,. \tag{4.9}$$

Remark that $k_t = h_t(i)$ with probability $q_t(i)$ so that, k_t follows the distribution $p_t \overset{\text{def}}{=} \sum_{i=1}^{N} q_t(i)\delta_{h_t(i)}$ knowing the past information $\mathcal{F}_{t-1} \overset{\text{def}}{=} \sigma(h_1, i_1, k_1, \dots, i_{t-1}, k_{t-1}, h_t)$. Now, similarly to the proof of EXP3, we compute the expectations. We have for all $k \in [K] \overset{\text{def}}{=} \{1,\dots,K\}$

$$\mathbb{E}\Big[g_t(k)\Big|\mathcal{F}_{t-1}\Big] = \mathbb{E}\left[\frac{\ell_t(k)}{p_t(k)}\mathbb{1}\{k = k_t\}\middle|\mathcal{F}_{t-1}\right] = \sum_{j=1}^{K} p_t(j)\frac{\ell_t(k)}{p_t(k)}\mathbb{1}\{k = j\} = \ell_t(k)\,,$$

and thus for all $i \in \{1,\dots,N\}$

$$\mathbb{E}\Big[g_t\big(h_t(i)\big)\Big|\mathcal{F}_{t-1}\Big] = \ell_t\big(h_t(i)\big)\,, \tag{4.10}$$

and

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{N} q_t(i)\cdot g_t\big(h_t(i)\big)\middle|\mathcal{F}_{t-1}\right] &= \sum_{i=1}^{N} q_t(i)\mathbb{E}\Big[g_t\big(h_t(i)\big)\Big|\mathcal{F}_{t-1}\Big] = \sum_{i=1}^{N} q_t(i)\ell_t\big(h_t(i)\big)\\ &= \mathbb{E}\big[\ell_t\big(h_t(i_t)\big)\big|\mathcal{F}_{t-1}\big] = \mathbb{E}\big[\ell_t(k_t)\big|\mathcal{F}_{t-1}\big]\,. \end{aligned}\tag{4.11}$$

Furthermore,

$$\mathbb{E}\Big[g_t\big(h_t(i)\big)^2\Big|\mathcal{F}_{t-1}\Big] = \sum_{k=1}^{K} p_t(k)\left(\frac{\ell_t\big(h_t(i)\big)}{p_t\big(h_t(i)\big)}\right)^2 \mathbb{1}\{k = h_t(i)\} = \frac{\ell_t\big(h_t(i)\big)^2}{p_t\big(h_t(i)\big)} \le \frac{1}{p_t\big(h_t(i)\big)}\,,$$

and

$$\sum_{i=1}^{N} q_t(i)\mathbb{E}\Big[g_t\big(h_t(i)\big)^2\Big|\mathcal{F}_{t-1}\Big] \le \sum_{i=1}^{N}\frac{q_t(i)}{p_t\big(h_t(i)\big)} = \mathbb{E}\left[\frac{1}{p_t\big(h_t(i_t)\big)}\middle|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[\frac{1}{p_t(k_t)}\middle|\mathcal{F}_{t-1}\right] = \sum_{k=1}^{K}\frac{p_t(k)}{p_t(k)} = K\,. \tag{4.12}$$

Substituting (4.10), (4.11), and (4.12) into Inequality (4.9) and lower-bounding the expected regret with the pseudo-regret, we get

$$\begin{aligned}\bar{R}_T^{\text{exp}} &\overset{\text{def}}{=} \max_{1\le i\le N}\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(k_t) - \ell_t\big(h_t(i)\big)\right]\\ &\overset{(4.10),(4.11)}{=} \max_{1\le i\le N}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} q_t(i)g_t\big(h_t(i)\big) - g_t\big(h_t(i)\big)\right]\\ &\overset{\text{Jensen}}{\le} \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} q_t(i)g_t\big(h_t(i)\big) - \min_{1\le i\le N} g_t\big(h_t(i)\big)\right]\\ &\overset{(4.9)}{\le} \eta\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\Big[q_t(i)g_t\big(h_t(i)\big)^2\Big] + \frac{\log N}{\eta}\\ &\overset{(4.12)}{\le} \eta KT + \frac{\log N}{\eta}\,.\end{aligned}$$

Optimizing η concludes the proof. □

4.2.2 Adversarial multi-armed bandits with side information

A natural extension of the previous setting is by adding side (or contextual) information: this is called contextual bandits. It arises in most applications such as recommendation systems or online advertisement. The side information can then be the cookies of a new user to which we need to recommend a product.

Assume that for each time step $t \geq 1$, before doing its prediction k_t the learner observes a context x_t in a finite set $\mathcal{X}$ of contexts. The learner must than learn the best mapping $g : \mathcal{X} \to [K]$ and is evaluated with the contextual pseudo-regret:

$$R_T^{\mathcal{X}} \overset{\text{def}}{=} \max_{g:\mathcal{X}\to[K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(k_t) - \sum_{t=1}^{T} \ell_t\big(g(x_t)\big)\right].$$

Similarly, to the stochastic setting, if $\mathcal{X}$ is finite, a simple algorithm consists in running a different copy $EXP3(c)$ of EXP3 for each context $c \in \mathcal{X}$. We denote by $\mathcal{X}$-EXP3 this algorithm. At each time step $t \geq 1$, the learner uses $EXP(x_t)$ to make the prediction. The following theorem follows from Theorem 4.1.

Theorem 4.5

The contextual pseudo-regret of $\mathcal{X}$-EXP3 is upper-bounded as:

$$R_T^{\mathcal{X}} \leq 2\sqrt{T|\mathcal{X}|K \log K}\,.$$

Proof. Applying the proof of the pseudo-regret bound of EXP3 for each instance $x \in \mathcal{X}$:

$$\max_{j\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} \big(\ell_t(k_t) - \ell_t(j)\big)\mathbb{1}\{x_t = x\}\right] \leq 2\sqrt{K(\log K)T_x}\,,$$

where $T_x = \sum_{t=1}^{T} \mathbb{1}\{x_t = x\}$. Summing over $x \in \mathcal{X}$,

$$\sum_{x\in\mathcal{X}} \max_{j\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} \big(\ell_t(k_t) - \ell_t(j)\big)\mathbb{1}\{x_t = x\}\right] \leq 2\sum_{x\in\mathcal{X}} \sqrt{K(\log K)T_x} \overset{\text{Jensen}}{\leq} 2\sqrt{|\mathcal{X}|K(\log K)T}$$

where the last inequality is by using the concavity of the square root together with $\sum_{s\in\mathcal{X}} T_s = T$. The proof is concluded by remarking that the left-hand side is the contextual pseudo-regret. □

Similarly to the classical lower-bound $O(\sqrt{TK})$, a lower-bound of order $\sqrt{|\mathcal{X}|KT}$ holds under the assumption that a significant proportion of the contexts are used at least a constant fraction of the T rounds. The above bound is nice but the dependency $|\mathcal{X}|$ might be annoying if $\mathcal{X}$ is large.

Exercise 4.1. *Generalize the above algorithm and upper-bound when the context-space is continuous and the loss functions are β-Hölder in the contexts.*

Competing against the best context set

In some cases, one may want to combine bandit algorithms. For example, we could have in hand different context set $\mathcal{X}$. For each of these sets $\mathcal{X}$, we can bound the pseudo-regret $R_T^{\mathcal{X}}$ using Theorem 4.5 with $\mathcal{X}$-EXP3 of Section 4.2.2, but we would like to find the best set $\mathcal{X}$. To do so, we may want to combine with EXP4 different instances of $\mathcal{X}$-EXP3, each using its own context set $\mathcal{X}$. We can then combine the bounds of Theorem 4.4 and Theorem 4.5 to ensure we are competing with the best possible context set $\mathcal{X}$. In this case, each instance of $\mathcal{X}$-EXP3 does not observe their own choice of action but the action chosen by EXP4 which follows a different distribution. The bound of Theorem 4.4 is valid but the regrets of the experts cannot be bounded using Theorem 4.5. It is however possible to use a variant of EXP4 to combine bandit algorithms by adding an exploration parameter. We then lose however in the rate of the regret bound which is then of order

$$\max_{\mathcal{X}} R_T^{\mathcal{X}} \leq O\Big(T^{2/3}\big(\max|\mathcal{X}|K\log K\big)^{1/3}\sqrt{\log M}\Big)$$

where M is the number of context sets $\mathcal{X}$. We refer to Section 4.2.1 of Bubeck et al. [2012] for more details on this application.

4.3 Online convex optimization with bandit feedback

In the previous sections, we saw how to deal with bandit feedback when the decision set is finite $\Theta = [K]$. The goal of this section is to extend to bandit feedback the general framework of online convex optimization of Figure 1.1 for any compact and convex decision space $\Theta \subset \mathbb{R}^d$. Here, at each round $t \geq 1$, the learner picks $\theta_t \in \Theta$ and observes $\ell_t(\theta)$. When gradients are observed, one may choose θ_t by following online gradient descent (OGD):

$$\theta_{t+1} = \Pi_\Theta(\theta_t - \eta\nabla\ell_t(\theta_t)) \,. \tag{OGD}$$

Theorem 2.7 showed that the regret of OGD for well-chosen learning rate $\eta > 0$ is upper-bounded as:

$$R_T = \sum_{t=1}^{T}\ell_t(\theta_t) - \min_{\theta\in\Theta}\sum_{t=1}^{T}\ell_t(\theta) \leq DG\sqrt{T}\,,$$

where $D \geq \max_{\theta,\theta'\in\Theta}\|\theta-\theta'\|$ and $G \geq \max_{\theta\in\Theta}\|\nabla\ell_t(\theta)\|$. Here, we aim at answering the following question:

> “How to adapt OGD to the bandit setting, when $\ell_t(\theta_t)$ is observed but not $\nabla\ell_t(\theta_t)$?”

Similarly to EXP3, the idea is to replace the gradient in OGD with estimators. That is to try to find a random variable g_t that satisfies $\mathbb{E}\big[g_t\big] \approx \nabla\ell_t(\theta_t)$ and which can be computed with the observation at hand.

4.3.1 Estimating gradients from value observations

One-dimensional example. Let us first, consider the one-dimensional case, let $\ell : \mathbb{R} \longrightarrow \mathbb{R}$ be a one-dimensional diffentiable loss function. Is it possible to design an estimator of $\ell'(\theta)$ by evaluating ℓ in a single point? It turns out that this is the case. Indeed, the derivative in θ can be writen as

$$\ell'(\theta) = \lim_{\delta\to 0}\frac{\ell(\theta+\delta)-\ell(\theta-\delta)}{2\delta}\,.$$

Thus, denoting ξ a Rademacher random variable (which equals 1 with probability 1/2 and −1 otherwise) and defining $g(\theta) = \frac{1}{\delta}\xi\ell(\theta+\xi\delta)$, we have

$$\mathbb{E}\big[g(\theta)\big] = \frac{1}{2}\times\frac{\ell_t(\theta+\delta)}{\delta} + \frac{1}{2}\times\frac{-\ell_t(\theta-\delta)}{\delta} \approx \ell'(\theta)$$

if δ is sufficiently small. Therefore, for a small value of δ, $g(\theta)$ is approximatively an unbiased estimator of $\ell'(\theta)$. Note that $g(\theta)$ can be computed by evaluating the function ℓ a single time at a random point close to θ. On the other side, on may compute

$$\operatorname{Var}\big(g(\theta)\big) = \left(\frac{\ell(\theta+\delta) + \ell(\theta-\delta)}{2\delta}\right)^2 \approx \frac{\ell(\theta)^2}{\delta^2} \,,$$

which may explode as $\delta \longrightarrow 0$. Thus, δ will need to be chosen carefully to optimize a bias-variance trade-off.

Multi-dimensional case. We now formally extend the above point-wise estimator to the multi-dimensional case. To do so, we first need to define $\widehat{\ell_t}$ a smoothed version of the loss by: for all $\theta \in \Theta$

$$\widehat{\ell_t}(\theta) = \mathbb{E}_v\big[\ell_t(\theta + \delta v)\big] \,,$$

where $v \sim \mathrm{Unif}(\mathbb{B})$ is a uniform random-variable over the Euclidean unit ball $\mathbb{B} := \{x \in \mathbb{R}^d : \|x\| \le 1\}$. The following lemma follows from Lipschitzness of ℓ_t, and states that ℓ_t is a δG-approximation of ℓ_t.

Lemma 4.6

Let $\delta > 0$. Then, $\big|\widehat{\ell_t}(\theta) - \ell_t(\theta)\big| \le \delta G$ for all $\theta \in \mathbb{R}^d$.

Now, define the estimator

$$g_t = \frac{d}{\delta}\ell_t(\theta_t + \delta u_t)u_t \quad \text{with } u_t \sim \mathrm{Unif}(\mathbb{S}) \tag{4.13}$$

where $\mathbb{S} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ is the Euclidean unit sphere. Then, the following lemma states that g_t is an unbiased estimator of $\widehat{\ell_t}$.

Lemma 4.7

Let $\widehat{\ell_t}(\theta_t) = \mathbb{E}_v\big[\ell_t(\theta_t + \delta v)\big]$ with $v \sim \mathrm{Unif}(\mathbb{B})$ and $g_t = \frac{d}{\delta}\ell_t(\theta_t + \delta u_t)u_t$ with $u_t \sim \mathrm{Unif}(\mathbb{S})$. Then,

$$\mathbb{E}_u\big[g_t\big] = \nabla\widehat{\ell_t}(\theta_t)\,.$$

Proof. The proof is a consequence of Stokes' theorem that states that for any continuously differentiable function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$,

$$\int_{\mathbb{B}} \nabla f(x)dx = \int_{\mathbb{S}} f(x)xdx \,,$$

applied with $f : x \mapsto \ell_t(\theta_t + \delta x)$ and $\nabla f(x) = \delta\nabla\ell_t(\theta_t + \delta_x)$. Denoting by μ the Lebesgue measure, we then have

$$\mathbb{E}_u[g_t] = \frac{1}{\mu(\mathbb{S})}\int_{\mathbb{S}} \frac{d}{\delta}\ell_t(\theta_t{+}\delta u)udu = \frac{d}{\mu(\mathbb{S})}\int_{\mathbb{B}} \nabla\ell_t(\theta_t{+}\delta v)dv = \frac{d\mu(\mathbb{B})}{\mu(\mathbb{S})}\mathbb{E}_v\big[\nabla\ell_t(\theta_t{+}\delta v)\big] = \frac{d\mu(\mathbb{B})}{\mu(\mathbb{S})}\nabla\widehat{\ell_t}(\theta_t)\,,$$

where the last inequality is by intechanging the derivative and the integral. The proof is then completed by noting that $d\mu(\mathbb{B}) = \mu(\mathbb{S})$. □

4.3.2 OGD with bandit feedback

Noting that the above estimator g_t can be computed with a single evaluation of ℓ_t at the random point $\theta_t + \delta u_t$, we may define the Online Gradient Descent algorithm with bandit feedback.

```
Online Gradient Descent with bandit feedback (OGD without gradient)
Parameters: $\eta > 0$ $\delta > 0$
Initialize: $\theta_1 \in \Theta_\delta$ arbitrarily chosen in $\Theta_\delta := \big\{\theta \in \Theta \text{ s.t. } \forall u \in \mathbb{S}, \theta + \delta u \in \Theta\big\}$
For $t = 1, \dots, T$
    – Draw $u_t$ uniformly at random in the unit sphere.
    – Set $\widehat{\theta}_t := \theta_t + \delta u_t$ a random perturbation of the current point $\theta_t$
    – Play $\widehat{\theta}_t$
    – Incur and observe loss $\ell_t(\widehat{\theta}_t) \in [-1, 1]$
    – Estimate the gradient with
          $g_t := \frac{d}{\delta} \ell_t(\widehat{\theta}_t) u_t$
    – Update
          $\theta_{t+1} = \Pi_\Theta\big(\theta_t - \eta g_t\big)$ .
      where $\Pi_{\Theta_\delta}$ is the Euclidean projection onto $\Theta_\delta$.
```

We have the following theorem.

Theorem 4.8

If the losses ℓ_t are G-Lipschitz and take values in $[-B, B]$, then OGD without gradient with parameters $\delta = \min\big\{D, \frac{1}{2}\sqrt{\frac{dBD}{G}}T^{-1/4}\big\}$ and $\eta = \delta D/(dB\sqrt{T})$ satisfies the expected regret bound

$$\mathbb{E}\Bigg[\sum_{t=1}^T \ell_t(\widehat{\theta}_t) - \min_{\theta\in\Theta} \sum_{t=1}^T \ell_t(\theta)\Bigg] \leq 2dB\sqrt{T} + 4\sqrt{dBGD}T^{3/4} \,.$$

The above regret bound is sub-optimal and more sophisticated algorithms achieve $O(\sqrt{T})$ but with worse dependency on d and higher computational cost.

Proof. We define

$$\theta^* = \arg\min_{\theta\in\Theta} \sum_{t=1}^T \ell_t(\theta) \qquad \text{and} \qquad \theta^*_\delta = \Pi_{\Theta_\delta}(\theta^*) \,.$$

Then, by definition of $\Theta_\delta := \big\{\theta \in \Theta \text{ s.t. } \theta + \delta u \in \Theta \text{ for all } u \in \mathbb{S}\big\}$, we have $\big\|\theta^* - \theta^*_\delta\big\| \leq \delta$ (left as exercise). Thus, if the losses are G-Lipschitz

$$\begin{aligned} R_T &:= \mathbb{E}\Bigg[\sum_{t=1}^T \ell_t(\widehat{\theta}_t) - \sum_{t=1}^T \ell_t(\theta^*)\Bigg] \\ &\leq \mathbb{E}\Bigg[\sum_{t=1}^T \ell_t(\widehat{\theta}_t) - \sum_{t=1}^T \ell_t(\theta^*_\delta)\Bigg] + \delta T G \qquad \leftarrow \text{because } \|\theta^* - \theta^*_\delta\| \leq \delta \end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}\Bigg[\sum_{t=1}^{T} \ell_t(\theta_t) - \sum_{t=1}^{T} \ell_t(\theta_\delta^*)\Bigg] + 2\delta T G \qquad \leftarrow \text{because } \|\widehat{\theta}_t - \theta_t\| \leq \delta \\
&\leq \mathbb{E}\Bigg[\sum_{t=1}^{T} \widehat{\ell}_t(\theta_t) - \sum_{t=1}^{T} \widehat{\ell}_t(\theta_\delta^*)\Bigg] + 4\delta T G \qquad \leftarrow \text{because } \big|\widehat{\ell}_t(\theta_t) - \ell_t(\theta_t)\big| \leq \sup_{\|v\|\leq 1} |\ell_t(\theta_t + \delta v) - \ell_t(\theta_t)| \leq \delta G
\end{aligned}
\tag{4.14}
$$

where $\widehat{\ell}_t(\theta) = \mathbb{E}_v\big[\ell_t(\theta + \delta v)\big]$ with $v \sim Unif(\mathbb{B})$ are the smoothed versions of the losses.

Now, recall that the algorithm runs OGD with g_t in place of the gradients:

$$
\theta_{t+1} \leftarrow \Pi_{\Theta_\delta}(\theta_t - \eta g_t)
$$

Defining the pseudo-loss $h_t(\theta) = \widehat{\ell}_t(\theta) + (g_t - \nabla\widehat{\ell}_t(\theta_t))^\top \theta$, we can see that

$$
\nabla h_t(\theta_t) = \nabla\widehat{\ell}_t(\theta_t) + g_t - \nabla\widehat{\ell}_t(\theta_t) = g_t.
$$

Therefore, the algorithm actually runs OGD on the losses h_t and thus satisfies the OGD regret bound (see Theorem 2.7)

$$
\sum_{t=1}^{T} h_t(\theta_t) - \sum_{t=1}^{T} h_t(\theta_\delta^*) \leq \frac{D^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|g_t\|^2 .
$$

Furthermore, by construction of the gradient estimator, we have $\mathbb{E}_{u_t}\big[g_t\big] = \nabla\widehat{\ell}_t(\theta_t)$, which yields

$$
\mathbb{E}_{u_t}\big[h_t(\theta_t)\big] = \widehat{\ell}_t(\theta_t) \quad \text{and} \quad \mathbb{E}_{u_t}\big[h_t(\theta_\delta^*)\big] = \widehat{\ell}_t(\theta_\delta^*) .
$$

Thus taking the expectation in the previous regret bound entails

$$
\mathbb{E}\Bigg[\sum_{t=1}^{T} \widehat{\ell}_t(\theta_t) - \sum_{t=1}^{T} \widehat{\ell}_t(\theta_\delta^*)\Bigg] = \mathbb{E}\Bigg[\sum_{t=1}^{T} h_t(\theta_t) - \sum_{t=1}^{T} h_t(\theta_\delta^*)\Bigg] \leq \frac{D^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \mathbb{E}\big[\|g_t\|^2\big] . \tag{4.15}
$$

Combining the two bounds (4.14) and (4.15) that we have proved, we get

$$
R_T \leq \frac{D^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \mathbb{E}\big[\|g_t\|^2\big] + 4\delta T G .
$$

Then, since $|\ell_t(\theta)| \leq B$ for all $\theta \in \Theta$,

$$
\|g_t\|^2 = \left(\frac{d}{\delta}\ell_t(\widehat{\theta}_t)\right)^2 \leq \frac{d^2 B^2}{\delta^2} .
$$

This finally yields the regret

$$
R_T \leq \frac{D^2}{2\eta} + \frac{\eta d^2 B^2 T}{2\delta^2} + 4\delta G T = \frac{dBD}{\delta}\sqrt{T} + 4\delta G T \leq 2dB\sqrt{T} + 4\sqrt{dBGD}T^{3/4}
$$

for the choices of δ and η. □

Coordinate gradient descent. In some situations, the full gradient $\nabla \ell_t(\theta_t)$ is too costly to be computed at each step, and one only observes the gradient of a single coordinate $k_t \sim \mathrm{Unif}(d)$. In this case, one may use OGD (or OMD, or EG) with the gradient estimator

$$g_t = d\nabla \ell_t(\theta_t)_{k_t} e_{k_t} ,$$

where $\{e_k\}$ are the elements of the canonical basis in $\mathbb{R}^d$. Here,

$$\mathbb{E}_{k_t}[g_t] = \sum_{k=1}^{d} d\nabla \ell_t(\theta_t)_k e_k \mathbb{P}(k_t = k) = \nabla \ell_t(\theta_t) ,$$

and

$$\mathbb{E}_{k_t}\left[\|g_t\|^2\right] = \sum_{k=1}^{d} d^2 \nabla \ell_t(\theta_t)_k^2 \|e_k\|^2 \mathbb{P}(k_t = k) = d\left\|\nabla \ell_t(\theta_t)\right\|^2 \leq dG^2 .$$

Then, following the proof of OGD, for the update rule $\theta_{t+1} = \Pi_\Theta(\theta_t - \eta g_t)$, we get the regret bound in expectation

$$\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\theta_t) - \min_{\theta\in\Theta} \sum_{t=1}^{T} \ell_t(\theta)\right] \leq \frac{\eta d^2 G^2 T}{2} + \frac{D^2}{2\eta} = GD\sqrt{dT} ,$$

where we optimized over $\eta = D/G(dT)^{-1/2}$ in the last equality. The regret bound of OGD is deteriorated by a factor $\sqrt{d}$ compared to OGD with full gradient observation.

5 Stochastic Multi-armed Bandits

During the last few chapters, we have reviewed the framework for comprehensive information. We designed algorithms to minimize regret for the different decision spaces Θ and loss assumptions f_t. Most of the algorithms were based on variations in the exponentially weighted average forecaster or online gradient descent. We also found some links with game theory, including the Blackwell approach, two-player zero-sum games, and calibration.

In this chapter, we consider the bandit setting, when the player only observes the performance of $f_t(\theta_t)$ but not $f_t(\theta)$ for $\theta \neq \theta_t$. We will start by providing fundamental results for stochastic bandits with finitely many actions, also called K-armed bandits which basically corresponds to $\Theta = \{1, \dots, K\}$ and i.i.d. loss functions f_t. For a thorough introduction to stochastic bandits we refer the interested student to the monographs Bubeck et al. [2012] or Lattimore and Szepesvári [2020].

5.1 Setting: stochastic bandit with finitely many actions

We state here the setting of stochastic bandits with finitely many actions (also called multi-armed bandit) and fix notations that we will use.

A multi-armed bandit problem is a sequential decision problem defined by a finite set of actions $\Theta \overset{\text{def}}{=} \{1, \dots, K\}$ also called *arms*. We assume that there are K unknown sequences $X_{i,1}, X_{i,2}, \dots$ of rewards in $[0,1]$ associated with each arm $i = 1, \dots, K$. At each round, the player makes a decision by pulling an arm k_t in Θ and observes the corresponding reward[1] $X_{k_t,t}$. The objective of the player is to minimize his cumulative regret:

$$R_T \overset{\text{def}}{=} \max_{k=1,\dots,K} \sum_{t=1}^{T} X_{k,t} - \sum_{t=1}^{T} X_{k_t,t} \,.$$

In stochastic bandits, we generally assume the sequences to be i.i.d. Each arm $k = 1, \dots, K$ is associated an unknown probability distribution ν_k over $[0,1]$ and $X_{k,t} \sim \nu_k$. We also denote

$$\mu_k \overset{\text{def}}{=} \mathbb{E}[X_{k,t}], \qquad \text{and} \qquad \mu^* \in \operatorname*{arg\,max}_{k=1,\dots,K} \{\mu_k\} \,.$$

The player aims at finding the arm with the highest mean reward μ_k as quickly as possible. The setting is summarized in Setting 5.1. Note that we retrieve the setting of online optimization (Setting 1.1) with the notation $X_{k,t} \overset{\text{def}}{=} 1 - f_t(k)$ with i.d.d. loss functions.

[1] In the bandit community, it is more common to consider rewards rather than losses.

Unknown parameters: K probability distributions $\nu_1, \dots, \nu_K$ on $[0,1]$

At each time step $t = 1, \dots, T$

- the player chooses an action $k_t \in \Theta = \{1, \dots, K\}$;
- given k_t, the environment draws the reward $X_{k_t,t} \sim \nu_{k_t}$;
- the player only observes the feedback $X_{k_t,t}$.

Setting 5.1: Setting of stochastic bandit with finitely many actions

Multi-armed bandits have several concrete historical applications in a variety of fields, including ad placement, clinical trials, source routing or game AI. The name bandit refers to the "slot machine" in casinos, and the bandit problem corresponds to a player that inserts coins into different machines and tries to maximize his payoff. The finite arms bandit settings we consider are simple enough to be analyzed and the algorithms can often be generalized to more realistic settings including for example contextual bandits.

Remark. *Assume that all arms $\nu_k \sim \mathcal{B}(1/2)$ for $k = 1, \dots, K$. Then, $\mathbb{E}[X_{k,t}] = 1/2$ and $\mathbb{E}[X_{k_t,t}] = 1/2$. But because of fluctuations of random walks, the expected magnitude of the maximum rewards is of order*

$$\mathbb{E}\left[\max_{k=1,\dots,K} \sum_{k=1}^{T} X_{k,n}\right] \approx \sqrt{T \log K}\,.$$

Therefore, in this case though all arms are optimal, the expected regret is of order $\sqrt{T \log K}$. We will thus consider a more quantity in the stochastic framework called the pseudo-regret which corresponds to competing with the best action in expectation, rather than the optimal action on the sequence of realized rewards.

Definition 5.1 **Pseudo-regret**

The pseudo-regret is defined as

$$\bar{R}_T \overset{def}{=} T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{k_t}\right],$$

where we recall $\mu_k = \mathbb{E}[X_{k,t}]$.

Remark that the pseudo-regret is upper-bounded by the expected regret $\bar{R}_T \leq \mathbb{E}\big[R_T\big]$. It is thus harder to design algorithms for the true regret but we will focus here on the pseudo-regret.

Useful notation In the following, we will denote by $\widehat{\mu}_k(s)$ the empirical mean of rewards obtained after pulling arm k s times. Let us also denote for all arms $k = 1, \dots, K$ by

$$\Delta_k \overset{\text{def}}{=} \mu^* - \mu_k \qquad \text{and} \qquad N_k(t) \overset{\text{def}}{=} \sum_{s=1}^{t} \mathbb{1}_{k_t = k}\,,$$

respectively the suboptimal gap of arm k and the number of times the arm k was pulled by the player before time t. Then, the pseudo-regret can be rewritten

$$\bar{R}_T = \left(\sum_{k=1}^{K} \mathbb{E}\big[N_k(t)\big]\right)\mu^* - \mathbb{E}\left(\sum_{k=1}^{K} N_k(t)\mu_k\right) = \sum_{k=1}^{K} \Delta_k \mathbb{E}\big[N_k(t)\big]\,. \tag{5.1}$$

We recall Hoeffding's inequality that will be used in the proofs. We will often use Azuma-Hoeffding's inequality which is a generalization of Hoeffding's inequality to martingals.

Proposition 5.1 **Hoeffding's Inequality**

If $X_1, \dots, X_n$ are independent random variables almost surely in $[a, b]$ then for all $\delta \in (0,1)$ we have

$$\mathbb{P}\left\{\sum_{t=1}^{n} X_k - \mathbb{E}\Big[\sum_{t=1}^{n} X_k\Big] \geq (b-a)\sqrt{\frac{n}{2}\log\frac{1}{\delta}}\right\} \leq \delta\,,$$

or equivalently for all $\varepsilon > 0$

$$\mathbb{P}\left\{\sum_{t=1}^{n} X_k - \mathbb{E}\Big[\sum_{t=1}^{n} X_k\Big] \geq \varepsilon\right\} \leq \exp\left(-\frac{2\varepsilon^2}{n(b-a)^2}\right).$$

5.2 Explore-Then-Commit (ETC)

Contrary to the full information we examined earlier, the player only observes the rewards of the chosen actions. He must therefore make a trade-off between exploration and exploitation. The first bandit algorithm that we consider is Explore Then Commit (ETC). It consists in first performing an exploration phase of mK length in which each arm is pulled $m \geq 1$ times. Then it exploits by pulling the arm with the best empirical reward for the remaining rounds. It is summarized in Algorithm 5.1.

Parameter: $m \geq 1$.

1. Exploration

- For rounds $t = 1, \dots, mK$ explore by drawing each arm m times.
- Compute for each arm k its empirical mean of rewards obtained by pulling arm k m times

$$\widehat{\mu}_k(m) = \frac{1}{m}\sum_{s=1}^{Km} X_{k,t}\mathbb{1}\{k_t = k\}\,.$$

2. Exploitation: keep playing the best arm $\arg\max_k \widehat{\mu}_k(m)$ for the remaining rounds $t = mK + 1, \dots, T$.

Algorithm 5.1: Explore-Then-Commit (ETC)

Theorem 5.2 **Thm 6.1, ?**

If $1 \leq m \leq T/K$ then

$$\bar{R}_T \leq m\sum_{k=1}^{K}\Delta_k + (T - mK)\sum_{k=1}^{K}\Delta_k \exp\big(-m\Delta_k^2\big)\,.$$

Proof. Assume without loss of generality that the first arm is optimal, i.e., $\mu_1 = \mu^*$ and $\Delta_1 = 0$. From (5.1), we have

$$\bar{R}_T = \sum_{k=1}^{K}\Delta_k \mathbb{E}\big[N_k(t)\big]\,.$$

Let $k \geq 2$ be a suboptimal arm. Then, the arm k is selected m times during the exploration phase, and $T - mK$ times during the exploitation if k is selected, which implies $\widehat{\mu}_k(m) \geq \widehat{\mu}_1(m)$. Therefore,

$$\mathbb{E}\big[N_k(t)\big] \leq m + (T - mK)\mathbb{P}\big(\widehat{\mu}_k(m) \geq \widehat{\mu}_1(m)\big)$$

Now, we can use Hoeffding's inequality to control the probability in the right-hand side. Indeed $\widehat{\mu}_k(m)$ and μ_1 are respectively the empirical averages of m i.i.d. random variables in $[0,1]$ of mean μ_k and $\mu_1 = \mu^*$. Therefore,

$$\begin{aligned}
\mathbb{P}\big(\widehat{\mu}_k(m) - \widehat{\mu}_1(m) \geq 0\big) &= \mathbb{P}\big(\widehat{\mu}_k(m) - \widehat{\mu}_1(m) - \mu_k + \mu_1 \geq -\mu_k + \mu_1\big) \\
&= \mathbb{P}\big(\widehat{\mu}_k(m) - \widehat{\mu}_1(m) - \mu_k + \mu_1 \geq \Delta_k\big) \\
&= \mathbb{P}\big(m\widehat{\mu}_k(m) - m\widehat{\mu}_1(m) - m\mu_k + m\mu_1 \geq m\Delta_k\big) \\
&\leq \exp\big(-m\Delta_k^2\big)\,.
\end{aligned}$$

This implies

$$\bar{R}_T \leq m \sum_{k=1}^{K} \Delta_k + (T - mK) \sum_{k=1}^{K} \Delta_k \exp\big(-m\Delta_k^2\big)\,.$$

□

The bound in Theorem 5.2 illustrates the trade-off between exploration and exploitation. If m is large, the exploration is too long and the first term $m\sum_{k=1}^{K}\Delta_k$ yields a large regret. On the other hand, for small m, there is a large probability to choose a suboptimal arm during the exploitation and the other term might lead to a large regret. The question is which value of m should we choose?

To have an idea, we will consider the case $K = 2$, in which case the bound is

$$\bar{R}_T \leq m\Delta_2 + T\Delta_2 \exp\big(-m\Delta_2^2\big)\,.$$

Corollary 5.3

If $K = 2$ and $m = \max\big\{1, \lceil \log(T\Delta_2^2)/\Delta_2^2 \rceil\big\}$, then

$$\bar{R}_T \leq \Delta_2 + \frac{1 + \log\big(T\Delta_2^2\big)}{\Delta_2}\,.$$

The above bound is of order $O((\log T)/\Delta_2)$. Such bounds are called distribution-dependent because they heavily depend on the distributions ν_k via Δ_k. If $\Delta_2 \longrightarrow 0$, it explodes. However, we also have from (5.1) that $\bar{R}_T \leq \Delta_2 T$. Therefore, in the worst case for any value of Δ_2, Corollary 5.3 yields to the worst-case bound

$$\bar{R}_T \leq \min\left\{T\Delta_2, \Delta_2 + \frac{1 + \log\big(T\Delta_2^2\big)}{\Delta_2}\right\} \lesssim \sqrt{T\log T}\,. \tag{5.2}$$

The above bound is close to be optimal. Yet, the issue is that the parameter m depends on Δ_2 and T. If the dependence on T can be dealt with a doubling-trick it is harder to optimize it in Δ_2. Furthermore, when there are more than two arms, one might want to explore differently the arms. The upper-confidence-bound algorithm that we will see next does not have these issues.

Exercise 5.1. *Show that it is possible to achieve the worst-case bound on the pseudo-regret of order $O(T^{2/3})$ by optimizing m independently of Δ (only with T).*

Exercise 5.2. *Generalize the results of Theorem 5.2 and 5.3 when the rewards are not-bounded but σ^2-sub-Gaussian, i.e., for all $\lambda > 0$*

$$\mathbb{E}\left[\exp\left(\lambda(X_{k,t} - \mathbb{E}[X_{k,t}])\right)\right] \le \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

5.3 Upper-Confidence-Bound (UCB)

The UCB (Upper-Confidence-Bound) algorithm is a very popular bandit algorithm that has several advantages over ETC. It does not rely on an initial exploration phase but explores on the fly as rewards are observed. It explores and exploits sequentially throughout the experience. Unlike ETC, it does not require knowledge of gaps and behaves well when there are more than two arms.

To perform exploration and face uncertainty, the UCB algorithm is based on the *optimism principle*.

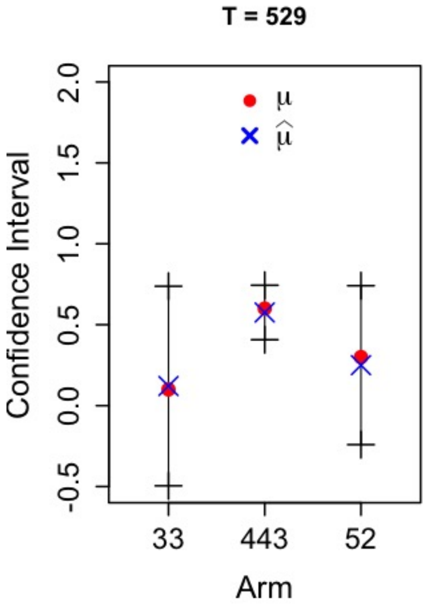
For each arm k, it builds a confidence interval on its expected reward based on past observation

$$I_k(t) = \left[LCB_k(k), UCB_k(t)\right]$$

where LCB is the Lower-Confidence-Bound and UCB is the Upper-Confidence-Bound. Then it is *optimistic* acting as if the best possible rewards are the real rewards and chooses the next arm accordingly

$$k_t \in \mathop{\arg\max}_{k \in \{1,\dots,K\}} \; UCB_k(t)\,.$$

In other words, it pulls the arm with the higher upper-confidence-bound. An example of how UCB works with three arms of means $\mu_1 = 0.1$, $\mu_2 = 0.6$ and $\mu_3 = 0.3$ is plotted in the Figure on the right. The best arm is pulled more often (see x-axis for number of times arms are selected) and his confidence interval is smaller.



The only question is how to design the upper-confidence-bounds. This is based on Hoeffding's inequality. Since the rewards are i.i.d. the distribution of $\widehat{\mu}_k(s)$ is equal to the distribution of

$$\frac{1}{s}\sum_{s'=1}^{s} X_{k,s'}\,,$$

with mean μ_k. Therefore, from Hoeffding's inequality, we have for all arms $k \in \{1,\dots,K\}$, for all $s \ge 1$ and all $\delta \in (0,1)$

$$\mathbb{P}\left\{\mu_k \ge \widehat{\mu}_k(s) + \sqrt{\frac{\log\frac{1}{\delta}}{2s}}\right\} \le \delta\,. \tag{5.3}$$

where $\widehat{\mu}_k(t)$ is the empirical reward of arm k after pulling it t times. Therefore, it is reasonable to choose the upper-confidence bound

$$UCB_t(k) = \begin{cases} \infty & \text{if } N_k(t-1) = 0 \\ \widehat{\mu}_k(N_k(t-1)) + \sqrt{\frac{2\log t}{N_k(t-1)}} & \text{otherwise} \end{cases}$$

The UCB algorithm is described in Algorithm 5.2.

Initialization For rounds $t = 1, \dots, K$ pull arm $k_t = t$

For $t = K + 1, \dots, T$, choose

$$k_t \in \operatorname*{arg\,max}_{k \in \{1,\dots,K\}} \left\{ \widehat{\mu}_k(N_k(t-1)) + \sqrt{\frac{2 \log t}{N_k(t-1)}} \right\} ,$$

and get reward $X_{k_t,t}$.

Algorithm 5.2: Upper-Confidence-Bound (UCB)

Theorem 5.4

If the distributions ν_k have supports all included in $[0,1]$ then for all k such that $\Delta_k > 0$

$$\mathbb{E}\big[N_k(T)\big] \leq \frac{8 \log T}{\Delta_k^2} + 2 \, .$$

In particular, this implies that the pseudo-regret of UCB is upper-bounded as

$$\bar{R}_T \leq 2K + \sum_{k:\Delta_k > 0} \frac{8 \log T}{\Delta_k} \, .$$

Remark. *Let us make some remarks about the about upper-bound on the pseudo-regret.*

- *UCB has a regret bound of order*
$$\bar{R}_T \lesssim \frac{K \log T}{\Delta} ,$$
where $\Delta = \min_{i:\Delta_i > 0} \Delta_i$. Once again, using that the regret incurred from pulling arm k cannot be larger than $T\Delta_k$, this distribution-dependent upper-bound can be transformed into a distribution-free bound of order $\bar{R}_T \lesssim \sqrt{TK \log T}$. We leave this proof as an exercise.
- *This bound is close to optimal since the lower bound is of order $O(\sqrt{KT})$. There exists modification of UCB to get rid of the extra logarithmic term. For instance, the MOSS algorithm (Minimax Optimal Strategy in the Stochastic Case) achieves*
$$\bar{R}_T \lesssim \min \left\{ \sqrt{TK}, \frac{K}{\Delta} \log \frac{T\Delta^2}{K} \right\} ,$$
however it depends on the smallest gap Δ only and not on all gaps Δ_i.
- *The assumption that the rewards are independent between arms can be relaxed.*
- *The assumption that the rewards are in $[0,1]$ can be relaxed to a sub-Gaussian assumption.*
- *While a bound on the pseudo-regret is interesting, one would actually want a bound with high probability on*
$$\widehat{R}_T \overset{def}{=} T\mu^* - \sum_{t=1}^{T} \mu_{k_t,t} \, .$$
Using Hoeffding’s inequality to control $\widehat{R}_T$ with $\bar{R}_T = \mathbb{E}[\widehat{R}_T]$ would yield an additional term of order $\sqrt{T}$ due to fluctuations which would dominate $O(K \log T/\Delta)$. Obtaining a bound of order

$O(K \log T/\Delta)$ on $\widehat{R}_T$ is a challenging problem and not achieved by UCB. Some strategies using the knowledge of T can satisfy it.

Proof. Without loss of generality let us assume that the first arm is optimal, i.e., $\mu_1 = \mu^*$ and $\Delta_1 = 0$. We show below that if $k_t = k$, then at least one of the following three inequalities must be satisfied

$$\mu^* > \widehat{\mu}_1\big(N_1(t-1)\big) + \sqrt{\frac{2\log t}{N_1(t-1)}} \qquad \leftarrow \mu^* \text{ larger than UCB} \tag{i}$$

$$\mu_k < \widehat{\mu}_k\big(N_k(t-1)\big) - \sqrt{\frac{2\log t}{N_k(t-1)}} \qquad \leftarrow \mu_k \text{ smaller than LCB} \tag{ii}$$

$$N_k(t-1) \leq \frac{8\log t}{\Delta_k^2} \qquad \leftarrow k \text{ not played enough yet} \tag{iii}$$

Indeed, otherwise assume that the three inequalities are all false than

$$\begin{aligned}
\widehat{\mu}_1(N_1(t-1)) + \sqrt{\frac{2\log t}{N_1(t-1)}} &\geq \mu^* && \leftarrow \quad \text{not (i)}\\
&\geq \mu_k + \Delta_k && \leftarrow \quad \text{Def of } \Delta_k\\
&> \mu_k + 2\sqrt{\frac{2\log t}{N_k(t-1)}} && \leftarrow \quad \text{not (iii)}\\
&\geq \widehat{\mu}_k(N_k(t-1)) + \sqrt{\frac{2\log t}{N_k(t-1)}} && \leftarrow \quad \text{not (ii)}\,.
\end{aligned}$$

This contradicts the fact that $k_t = k$ (see Algorithm 5.2). Therefore, denoting $u = \lfloor \frac{8\log T}{\Delta_k^2} \rfloor$, we have

$$\begin{aligned}
\mathbb{E}\big[N_k(T)\big] = \sum_{t=1}^T \mathbb{E}\left[\mathbb{1}_{k_t=k}\right] &= u + \sum_{t=u+1}^T \mathbb{P}\Big\{k_t = k \text{ and (iii) is false }\Big\}\\
&= u + \sum_{t=u+1}^T \Big(\mathbb{P}\big\{\text{(i) or (ii)}\big\}\Big)\\
&\leq u + \sum_{t=u+1}^T \Big(\mathbb{P}\big\{\text{(i)}\big\} + \mathbb{P}\big\{\text{(ii)}\big\}\Big)\,.
\end{aligned} \tag{5.4}$$

Therefore, it suffices to control the probabilities of (i) and (ii), which we do now. At round $t \geq 1$,

$$\begin{aligned}
\mathbb{P}\big\{\text{(i)}\big\} &\leq \mathbb{P}\left\{\exists s \in \{1,\dots,t-1\}, \text{ such that } \mu^* > \widehat{\mu}_1(s) + \sqrt{\frac{2\log t}{s}}\right\}\\
&\leq \sum_{s=1}^t \mathbb{P}\left\{\mu^* > \widehat{\mu}_1(s) + \sqrt{\frac{\log(1/t^{-4})}{2s}}\right\}\\
&\overset{(5.3)}{\leq} \sum_{s=1}^t t^{-4} = t^{-3}\,.
\end{aligned}$$

By symmetry, the same applies for $\mathbb{P}\big\{\text{(ii)}\big\} \leq t^{-3}$. Combining into (5.4), it concludes the proof of the first inequality

$$\mathbb{E}\big[N_k(T)\big] \leq \frac{8\log T}{\Delta_k^2} + 2\sum_{t=u+1}^{T} t^{-3} \leq \frac{8\log T}{\Delta_k^2} + 2\,.$$

The upper-bound on the pseudo-regret follows from (5.1). □

5.4 Other algorithms

Other algorithms exist in the literature. The best known are ε-greedy and Thompson sampling.

5.4.1 ε-greedy

The idea of ε-greedy is very simple: first choose a parameter $\varepsilon \in (0,1)$, then at each round, select the arm with the highest empirical mean with probability ε (i.e., be greedy), and explore by playing a random arm with probability ε. It works quite well in practice and is used in many application because of its simple implementation (in particular in reinforcement learning). Choosing $\varepsilon \approx K/(\Delta^2 T)$ yields to an upper-bound of order $K \log T/\Delta^2$. However it requires the knowledge of Δ.

5.4.2 Thompson Sampling

Thomson sampling was the first algorithm proposed for bandits by Thomson in 1933. It assumes a uniform prior over the expected rewards $\mu_i \in (0,1)$, then at each round $t \geq 1$, for each arm $\pi_{k,t}$, it

- computes $\widehat{\nu}_{k,t}$ the posterior distribution of the rewards of arm k given the rewards observed so far;
- samples $\theta_{k,t} \sim \widehat{\nu}_{k,t}$ independently;
- selects $k_t \in \arg\max_{k\in\{1,\dots,K\}} \theta_{k,t}$.

Thomson sampling has a similar upper-bound of order $O(K \log T/\Delta)$ than the one achieved by UCB. An advantage over UCB is the possibility of incorporating easily prior knowledge on the arms.

5.5 Lower bounds for multi-armed bandit

In this section, we essentially state that the regret bound of UCB

$$\bar{R}_T \lesssim \min\left\{\sqrt{KT\log T}, \sum_{k:\Delta_k>0} \frac{\log T}{\Delta_k}\right\}$$

is close to optimal regret for multi-armed bandit.

5.5.1 Distribution-free lower bound

The next theorem shows that the previous results are not improvable (up to log factors).

Theorem 5.5 **Lower bound**

For any forecaster, there exists distributions $\nu_1, \dots, \nu_K$ such that

$$\bar{R}_T \gtrsim \sqrt{KT}\,.$$

The complete proof can be found in Bubeck et al. [2012]. We only present here the high-level idea of the proof. We design the adversary as follows: it generates i.i.d. Bernoulli rewards such that $\mathbb{E}[X_{k,t}] = \frac{1}{2}$ for all $k \in \{1, \dots, K\}$ except for one arm k^* where $\mathbb{E}[X_{k^*,t}] = \frac{1}{2} + \varepsilon$.

- Fact 1: to distinguish between a Bernoulli of parameter $1/2$ and a Bernoulli of parameter $1/2 + \varepsilon$, one needs $1/\varepsilon^2$ samples. This result can be proved formally by using Pinsker's inequality. The intuition goes as follows. From the Central Limit Theorem (or the distribution of a Binomial random variable), after T_k observations of an arm k, on can estimate its mean with an error of order $1/\sqrt{T_k}$. In other words, to estimate it with an error smaller than ε, one needs $T_k \approx \varepsilon^{-2}$ observations.
- Fact 2: at least one arm is sampled less than T/K times.

Assume that this arm is k^*, than the learner cannot distinguish it with other arms as soon as $T_{k^*} \leq T/K \leq \varepsilon^{-2}$, which corresponds to $\varepsilon \leq \sqrt{K/T}$. Choosing $\varepsilon = \sqrt{K/T}$, the pseudo-regret is than at least $(1 - 1/K)T\varepsilon \approx T\varepsilon \approx \sqrt{KT}$.

5.5.2 Distribution-dependent lower bound

Here, we show that the distribution dependent upper bound is not also optimal in the case of Bernoulli rewards.

A caveat with distribution dependent lower bounds is that for any distribution, there exists an algorithm with no-regret. For instance, consider a distribution $\nu_1, \dots, \nu_K$ such that ν_1 is optimal (i.e., $\mu_1 = \max_k \mu_k$), the the algorithm that pull always the first arm will have zero regret. Yet such an algorithm will incure linear regret for some other distributions.

Hence, the following theorem states that any algorithm that incure sublinear regret for all distributions, achieves at best a pseudo regret of the same order of the one satisfied by UCB. The proof can be found in Bubeck et al. [2012].

Theorem 5.6 **Thm 2.2. Bubeck et al. [2012]**

Consider a strategy such that $\mathbb{E}\big[N_k(T)\big] = O(T^a)$ for any Bernoulli distributions, all suboptimal arms k and some $a > 0$. Then, for any Bernoulli distributions with means μ_k, we have

$$\liminf_{T \to \infty} \frac{\bar{R}_T}{\log T} \geq \sum_{k:\Delta_k > 0} \frac{\mu^*(1-\mu^*)}{\Delta_k}\,.$$

Note that the only difference with UCB is the factor $\mu^*(1-\mu^*)$ which corresponds to the variance of the best arm. In the case of Bernoulli noise, the KL-UCB algorithm can take advantage of the knowledge that the rewards are Bernoulli to close this gap.

6 Stochastic Contextual bandits

During last chapter, we considered the finite-armed bandit setting and saw several algorithms (ETC, UCB, ε-greedy, Thomson sampling) that achieve sublinear pseudo regret. UCB achieves for instance

$$\bar{R}_T \lesssim \min \left\{ \sum_{k:\Delta_k>0}^{K} \frac{\log T}{\Delta_k}, \sqrt{TK \log T} \right\} ,$$

where $\Delta_k \overset{\text{def}}{=} \mu^* - \mu_k$ is the suboptimality gap of arm k. The first bound is distribution dependent (it depends on the gaps Δ_k) and is of order $O(\log T)$ while the second bound is distribution free but is of order $O(\sqrt{T})$. In this chapter, we consider the more practical setting of contextual bandits, in which the learner observes a context $c_t \in C$ before choosing the action k_t.

In most applications, before choosing an action k_t the player observes some context $c_t \in \mathcal{C}$.

For instance, consider a bandit problem in which the player needs to display ads on his website. At each new visitor, the player chooses an add to display and observes if the visitor click on it. The reward is one if there is a click and 0 otherwise. In this case, the player can see the cookie of the visitor before choosing the ad. A first step towards contextual bandits, is to consider continuous sets of actions $\mathcal{X}$, which may correspond to mapping between context and arms.

6.1 Continuous stochastic bandits

Let first generalize the finite-armed bandit setting to continuous set of arms in Setting 6.1.

Unknown parameters: $\nu(\theta)$, for each $\theta \in [0,1]^d$, a probability distribution on $[0,1]$ with expectation $\mu(\theta) \in [0,1]$.

At each time step $t = 1, \dots, T$

- the player chooses an action $\theta_t \in \Theta \subseteq [0,1]^d$;
- given θ_t, the environment draws the reward $Y_t \sim \nu(\theta_t)$ independently from the past;
- the player only observes the feedback Y_t.

The player wants to minimize its pseudo-regret defined as

$$\bar{R}_T \overset{\text{def}}{=} T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} \mu(\theta_t) \right],$$

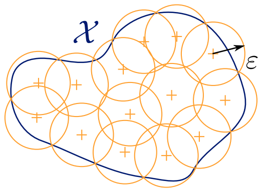
where $\mu^* = \sup_{\theta \in \Theta} \mu(\theta)$.

Setting 6.1: Setting of stochastic bandit with continuous set of actions

Similarly to what we did in the full-information setting with EWA, if the expectation function μ is β-Hölder: i.e., there exists $c > 0$

$$\forall \theta, \theta' \in \mathcal{X} \qquad \big|\mu(\theta) - \mu(\theta')\big| \leq c\big\|\theta - \theta'\big\|^{\beta},$$

then we may discretize the action space Θ and run any discrete bandit algorithm (UCB, ε-greedy,...).



Theorem 6.1

Let $\beta > 0$ and $\varepsilon > 0$. Assume that μ is β-Hölder. If UCB is run on an ε-covering of minimal cardinal of $\Theta \subset [0,1]^d$, then it satisfies

$$\bar{R}_T \lesssim T\varepsilon^{\beta} + \sqrt{\frac{T\log(T)}{\varepsilon^d}}\,.$$

In particular for $\varepsilon \approx \left(\frac{\log T}{T}\right)^{\frac{1}{2\beta+d}}$, we have $\bar{R}_T \lesssim T\left(\frac{\log T}{T}\right)^{\frac{\beta}{2\beta+d}}$.

Proof. An optimal ε-covering of $[0,1]^d$ has cardinal of order $K \approx \varepsilon^{-d}$. Let $x^* \in \arg\max_{\theta\in\Theta} \mu(\theta)$ (we assume that it exists) and $\tilde{\theta}^*$ its ε-approximation, then the distribution-free upper-bound of UCB yields

$$\bar{R}_T \lesssim T\big(\mu(\theta^*) - \mu(\tilde{\theta}^*)\big) + \sqrt{KT\log T} \approx cT\varepsilon^{\beta} + \sqrt{\varepsilon^{-d}T\log T}\,.$$

The second part of the theorem is obtained by obtimizing ε. □

To build the discretization, both β and T need to be known in advance. The horizon T can be calibrated online through a "doubling trick" (left as exercise). The parameter β may be tuned through bandit with experts (or bandits where arms are bandit algorithms) that we may see in next lecture (see Exp4 algorithm).

Note that the per-round complexity of such an algorithm is of order $\varepsilon^{-d} \approx T^{\frac{d}{2\beta+d}}$. Quite surprinsingly it does not explodes with the dimension d and is always smaller than T. This is due to the fact that the higher the dimension d is, the worse will be the regret bound, and the cruder needs the discretization to be.

6.1.1 Contextual bandits through discretization

No we consider the following contextual bandit setting in which the player has a finite decision set $\Theta = \{1, \dots, K\}$ but observes a context $x_t \in \mathcal{X}$ before choosing his action.

Finite set of contexts If the set of context is finite $\mathcal{X} \overset{\text{def}}{=} \{1, \dots, |\mathcal{X}|\}$ we can denote

$$\bar{R}_T(c) \overset{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^{T} \big(\mu^*(x) - \mu(k_t, x)\big)\mathbb{1}_{x_t = x}\right]$$

the pseudo-regret due to context $x \in \mathcal{X}$. Then applying a separate instance of UCB (or any bandit algorithm) for each context $x \in \mathcal{X}$, we get by using the distribution-free upper-bound of UCB

$$\bar{R}_T(x) \lesssim \sqrt{T_x K \log T_x}, \qquad \text{where} \qquad T_x \overset{\text{def}}{=} \sum_{t=1}^{T} \mathbb{1}_{x_t = x}\,.$$

Unknown parameters: $\nu(k, x)$, for each arm $k \in \{1, \dots, K\}$ and context $x \in \mathcal{X}$, a probability distribution on $[0, 1]$ with expectation $\mu(k, x) \in [0, 1]$.

At each time step $t = 1, \dots, T$

- the environment chooses $x_t \in \mathcal{X}$ and reveals it to the player;
- the player chooses an action $k_t \in \{1, \dots, K\}$;
- given k_t, the environment draws the reward $Y_t \sim \nu(k_t, x_t)$ independently from the past;
- the player only observes the feedback Y_t.

The player wants to minimize its pseudo-regret defined as

$$\bar{R}_T \overset{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^{T} \mu^*(x_t) - \sum_{t=1}^{T} Y_t\right],$$

where $\mu(k, x) = \mathbb{E}_{Y\sim\nu(k,x)}[Y]$ and $\mu^*(x) = \max_{k=1,\dots,K} \mu(k, x)$.

Setting 6.2: Setting of contextual stochastic bandit

Note that because T_x are not known in advance it is important that the bound of UCB is anytime (i.e., that UCB does not need to know the horizon). The total pseudo-regret of UCB is then obtained by summing over all contexts

$$\bar{R}_T = \sum_{x\in\mathcal{X}} \bar{R}_T(x) \lesssim \sum_{x\in\mathcal{X}} \sqrt{T_x K \log T} \leq \sqrt{|\mathcal{X}| T K \log T},$$

where the last inequality is by Jensen's inequality using the concavity of the square root and $\sum_{x\in\mathcal{X}} T_x = T$.

Continuous set of contexts If the set of context is continuous $\mathcal{X} \subset [0, 1]^d$, one needs again to make assumption on the distributions $\nu(k, x)$ which needs to vary smoothly in x. Doing so, one may discretize the set of context with an ε-covering of $\mathcal{X}$ of size $N \approx \varepsilon^{-d}$ and run an independent instance of UCB in each of the N bins.

Theorem 6.2

Let $\beta > 0$ and $\varepsilon > 0$. Assume that $x \mapsto \mu(k, x)$ is β-Hölder for all $k \in \mathcal{X}$. If UCB is independently run in each bin of an optimal ε-covering of $\mathcal{X}$, then

$$\bar{R}_T \lesssim T\varepsilon^{\beta} + \sqrt{\frac{KT \log T}{\varepsilon^d}}.$$

In particular for ε well-optimized, we have $\bar{R}_T \lesssim T\left(\frac{K \log T}{T}\right)^{\frac{\beta}{2\beta+d}}$.

Remark that in all these regret bounds, the suboptimal $\log T$ term can be removed by using MOSS (a minimax optimal variant of UCB).

Better rates using distribution-dependent bound? In the above results, we used the distribution-free regret bound of UCB. Because, if the function $\mu(\cdot, x)$ varies smoothly with x, there should be some context with zero suboptimality gaps. Yet, it is possible to get better rates by assuming the following α-margin assumption. It controls the suboptimality gap with high probability: the contexts x_t are i.i.d. and sastify for all $\delta \in (0,1)$

$$\mathbb{P}\Big\{ \min_{k:\Delta(k,x_t)>0} \Delta(k,x_t) < \delta \Big\} \le \square \delta^{\alpha} \tag{6.1}$$

where $\Delta(k,x_t) \overset{\text{def}}{=} \mu^*(x) - \mu(k,x)$ and □ is some constant. Note that the larger the value of α is the easier is the problem.

> **Theorem 6.3** **Theorem 4.1, Perchet and Rigollet [2013]**
>
> *Let $\alpha \in (0,1)$, $\beta > 0$ and $\varepsilon > 0$. Assume that $c \mapsto \mu(k,x)$ is β-Hölder for all $k \in \mathcal{X}$ and that the α-margin assumption (6.1) holds. Running a bandit algorithm (similar to UCB) independently in each bin of an optimal ε-covering of $\mathcal{X}$, we get*
>
> $$\bar{R}_T \lesssim T \left(\frac{K \log K}{T} \right)^{\frac{\beta(\alpha+1)}{2\beta+d}} ,$$
>
> *for optimized ε.*

The proof (for another algorithm then UCB) may be found in Perchet and Rigollet [2013]. We see that the factor α improves the rate of convergence with respect to the previous rate.

6.1.2 Stochastic Linear bandits

Contextual bandits that we just saw generalizes multi-armed bandits by allowing contexts. However, the dimension of the context space significantly worsen the regret rate from $\sqrt{T}$ to $T^{\frac{d+1}{d+2}}$ for Lipschitz rewards for instance (β-Hölder with $\beta = 1$). In this part, we will see *Stochastic linear bandits*, in which we assume the rewards to have a linear structure. This includes rich classes of models and allows better regret of order $O(\sqrt{T})$.

Unknown parameter: $\mu^* \in \mathbb{R}^d$.

At each time step $t = 1, \dots, T$

- the environment chooses $\Theta_t \subseteq \mathbb{R}^d$ the decision set;
- the player chooses an action $\theta_t \in \Theta_t$;
- given θ_t, the environment draws the reward

  $$Y_t = \theta_t \cdot \mu^* + \varepsilon_t$$

  where ε_t is i.i.d. 1-subgaussian noise.
- the player only observes the feedback Y_t.

The player wants to minimize its pseudo-regret defined as

$$\bar{R}_T \overset{\text{def}}{=} \mathbb{E}\left[\sum_{t=1}^{T} \max_{\theta \in \Theta_t} \theta \cdot \mu^* - \sum_{t=1}^{T} Y_t \right] .$$

Setting 6.3: Setting of stochastic linear bandit

The setting of stochastic linear bandits is described in Setting 6.3. For simplicity, the noise ε_t is assumed to be i.i.d. and 1-subgaussian noise: i.e., $\mathbb{E}[\varepsilon_t] = 0$ and

$$\forall \lambda > 0, \qquad \mathbb{E}\big[\exp(\lambda \varepsilon_t)\big] \leq \exp(\lambda^2/2)$$

almost surely. Note that we could consider σ^2-subgaussian noise, or make it depend on the past $\mathcal{F}_t = \sigma(x_1, \varepsilon_1, \dots, x_t, \varepsilon_t)$ with $\mathbb{E}[\varepsilon_t | \mathcal{F}_t] = 0$.

Particular cases: why is this setting interesting? Different choices of decision sets $\mathcal{X}_t$ lead to different settings of stochastic bandits:

- *Finite-armed bandit*: if $\Theta_t = (e_1, \dots, e_d)$ where e_i are the unit vectors in $\mathbb{R}^d$ and $\mu^* = (\mu_1, \dots, \mu_d)$, we recover the setting of finite-armed bandit.
- *Contextual linear bandit*: we can recover a particular case of Setting 6.2, if $x_t \in \mathcal{X}$ is a context observed by the player and the reward function μ is of the form

$$\mu(\theta, x) = \psi(\theta, x) \cdot \mu^*, \qquad \forall (\theta, x) \in [K] \times \mathcal{X},$$

  for some unknown parameter $\mu^* \in \mathbb{R}^d$ and *feature map* $\psi : [K] \times \mathcal{X} \to \mathbb{R}^d$. For example, assume that you are a website which wants to display ads to visitors. The context x_t can be the cookie of the visitor containing information about what he likes, the actions are the possible ads to be displayed and the reward tells if there is a click. If the possible interests of the visitor are grouped in finite categories (such as traveling), so are the ads (in groups of products, such as flight tickets), the feature maps could contained all the combinations between interests and groups of products. The unknown vector θ^* would be tell which interests and groups of products are positively correlated. Of course the feature map could be created using any methods (such as deep-learning or splines).
- *Combinatorial bandit*: if $\Theta_t \subseteq \{0, 1\}^d$ yields to combinatorial bandit problems. For instance, the decision set corresponds to possible paths in a graph, the vector μ^* assigns to each edge a reward corresponding to its cost and the goal is to find the smallest path with smallest cost.

Algorithm: LinUCB As we saw earlier with UCB, the "optimism principle" is a good option for bandit problems to explore. The LinUCB algorithm is based on the same principle:

1. Build confidence set that contain μ^*: $\mu^* \in C_t$ with high probability
2. Build confidence upper-bound on the rewards: for all $\theta \in \Theta_t$

$$\mathrm{UCB}_t(\theta) = \max_{\mu \in C_t} \ \theta \cdot \mu \tag{6.2}$$

3. Be optimistic: act as if the best possible rewards were the true rewards

$$\theta_t \in \operatorname*{arg\,max}_{\theta \in \Theta_t} \mathrm{UCB_t}(\theta)\,. \tag{6.3}$$

Therefore the only remaining question is how to build the confidence set $C_t \subseteq \mathbb{R}^d$? They should contain μ^* with high probability but be as small as possible. Given the observed rewards the key is thus to estimate the parameter μ^*. Denoting by I_d the $d \times d$ identity matrix and picking $\lambda > 0$, we can estimate μ^* with *regularized least square*

$$\widehat{\mu}_t \overset{\text{def}}{=} \operatorname*{arg\,min}_{\mu \in \mathbb{R}^d} \left\{ \sum_{s=1}^t (Y_s - \theta_s \cdot \mu)^2 + \lambda \|\mu\|^2 \right\} = V_t^{-1} \sum_{s=1}^t \theta_s Y_s\,,$$

where $V_t \overset{\text{def}}{=} \lambda I_d + \sum_{s=1}^t \theta_s \theta_s^\top$. We have the following result whose proof can be found in Lattimore and Szepesvári [2020].

Lemma 6.4 **Theorem 20.2, Lattimore and Szepesvári [2020]**

Let $\delta \in (0,1)$. Then, with probability at least $1-\delta$, if $\max_{\theta \in \Theta_t} \|\theta\|_2 \le 1$, for all $t \ge 1$

$$\big\|\widehat{\mu}_t - \mu^*\big\|_{V_t} \le \sqrt{\lambda}\|\mu^*\| + \sqrt{2\log(1/\delta) + d\log\left(1+\frac{T}{\lambda}\right)} \overset{def}{=} \beta(\delta)\,,$$

where $\|\mu\|_{V_t}^2 = \mu^\top V_t \mu$.

The above lemma, states that with probability $1-\delta$, for all $t \ge 1$,

$$\mu^* \in C_t, \quad \text{where} \quad C_t \overset{\text{def}}{=} \Big\{\theta \in \mathbb{R}^d : \big\|\mu - \widehat{\mu}_{t-1}\big\|_{V_{t-1}} \le \beta(\delta/T)\Big\}\,. \tag{6.4}$$

Proof of Lemma 6.4. The proof relies on Laplace's method on super-martingales which is a standard argument to provide confidence bounds on a self-normalized sum of conditionally centered random vectors. We have

$$\widehat{\mu}_t = V_t^{-1}\sum_{s=1}^t \theta_s Y_s = V_t^{-1}\sum_{s=1}^t \theta_s(\theta_s^\top \mu^* + \varepsilon_s) = V_t^{-1}(V_t - \lambda I_d)\mu^* + M_t) = \mu^* - \lambda V_t^{-1}\mu^* + V_t^{-1}M_t\,,$$

where we introduced $M_t = \sum_{s=1}^t \theta_s \varepsilon_s$, which is a martingale with respect to $\mathcal{F}_t = \sigma(\varepsilon_1, \dots, \varepsilon_t)$. Therefore, by triangle inequality

$$\big\|V_t^{1/2}(\widehat{\mu}_t - \mu^*)\big\| = \big\| - \lambda V_t^{-1/2}\mu^* + V_t^{-1/2}M_t\big\| \le \lambda\big\|V_t^{-1/2}\mu^*\big\| + \big\|V_t^{-1/2}M_t\big\|\,.$$

On the one hand, given that all eigenvalues of the symmetric matrix V_t are larger than λ, all eigenvalues of $V_t^{-1/2}$ are smaller than $1/\sqrt{\lambda}$ and thus

$$\lambda\big\|V_t^{-1/2}\mu^*\big\| \le \lambda\frac{1}{\sqrt{\lambda}}\|\mu^*\| = \sqrt{\lambda}\|\mu^*\|\,.$$

We now prove, on the other hand, that with probability at least $1-\delta$

$$\big\|V_t^{-1/2}M_t\big\| \le \sqrt{2\log\frac{1}{\delta} + d\log\frac{1}{\lambda} + \log\det(V_t)}\,.$$

Upper-bounding $\log\det(V_t) \le d\log(\lambda + t)$ (since all the eigenvalues of V_t are smaller than $\lambda + t$) will then conclude the proof of the Theorem.

Step 1: Introducing super-martingales. For all $v \in \mathbb{R}^d$, we consider

$$S_{t,v} = \exp\left(v^\top M_t - \frac{1}{2}v^\top V_t v\right)$$

and now show that it is an $\mathcal{F}_t$-super-martingale. First, note that since the common distribution of the $\varepsilon_1, \dots, \varepsilon_t$ is 1-sub-Gaussian, the for all $\mathcal{F}_{t-1}$-measurable random variable v_{t-1}

$$\mathbb{E}\Big[e^{v_{t-1}\varepsilon_t}\Big|\mathcal{F}_{t-1}\Big] \le e^{\frac{v_{t-1}^2}{2}}\,.$$

Now,

$$\mathbb{E}\Big[S_{t,v}\Big|\mathcal{F}_{t-1}\Big] = S_{t-1,v}\mathbb{E}\Big[\exp\Big(v^\top\theta_t\varepsilon_t - \frac{1}{2}v^\top\theta_t\theta_t^\top v\Big)\Big|\mathcal{F}_{t-1}\Big] \le S_{t-1,v}\,.$$

Note that rewriting $S_{t,v}$ in its vertex form is, with $m = V_t^{-1}M_t$:

$$S_{t,v} = \exp\Big(\frac{1}{2}(v-m)^\top V_t(v-m)\Big) \times \exp\Big(\frac{1}{2}\big\|V_t^{-1/2}M_t\big\|^2\Big)\,.$$

Step 2: Laplace's method–integrating $S_{t,v}$ over $v \in \mathbb{R}^d$. The basic observation behind this method is that (given the vertex form) $S_{t,v}$ is maximal at $v = m = V_t^{-1}M_t$ and then equals $\exp\big(\frac{1}{2}\|V_t^{-1/2}M_t\|^2\big)$, which is (a transformation of) the quantity to control. Now, because the exp function quickly vanishes, the integral over $v \in \mathbb{R}^d$ is close to its maximum. We therefore consider

$$\bar{S}_t = \int_{\mathbb{R}^d} S_{t,v}dv\,.$$

We will make repeated uses of the fact that the Gaussian density function

$$v \mapsto \frac{1}{\sqrt{\det(2\pi C)}}\exp\Big((v-m)^\top C^{-1}(v-m)\Big)\,,$$

where $m \in \mathbb{R}^d$ and C is a symmetric positive definite matrix, integrate to 1 over $\mathbb{R}^d$. This gives us the first rewriting

$$\bar{S}_t = \sqrt{\det(2\pi V_t^{-1})}\exp\Big(\frac{1}{2}\big\|V_t^{-1/2}M_t\big\|^2\Big)\,.$$

Second, by the Fubini-Tonelli theorem and the super-martingale property

$$\mathbb{E}\big[S_{t,v}\big] \le \mathbb{E}\big[S_{0,v}\big] = \exp(-\lambda\|v\|^2/2)\,,$$

we also have

$$\mathbb{E}\big[\bar{S}_t\big] \le \int_{\mathbb{R}^d}\exp(-\lambda\|v\|^2/2)dv = \sqrt{\det(2\pi\lambda^{-1}I_d)}\,.$$

Combining the two statements, we proved

$$\mathbb{E}\Big[\exp\Big(\frac{1}{2}\big\|V_t^{-1/2}M_t\big\|^2\Big)\Big] \le \sqrt{\frac{\det(V_t)}{\lambda^d}}\,.$$

Step 3: Markov-Chernov bound. For $u > 0$,

$$\begin{aligned}\mathbb{P}\Big(\big\|V_t{-}1/2M_t\big\| > u\Big) &= \mathbb{P}\Big(\frac{\big\|V_t{-}1/2M_t\big\|^2}{2} > \frac{u^2}{2}\Big)\\ &\le \exp\big(-\frac{1}{2}u^2\big)\mathbb{E}\Big[\exp\Big(\frac{1}{2}\big\|V_t^{-1/2}M_t\big\|^2\Big)\Big] \le \exp\Big(-\frac{u^2}{2} + \frac{1}{2}\log\frac{\det(V_t)}{\lambda^d}\Big) = \delta\,,\end{aligned}$$

for the claimed choice

$$u = \sqrt{2\log\frac{1}{\delta} + d\log\frac{1}{\lambda} + \log\det(V_t)}\,.$$

□

Theorem 6.5 **Corollary 19.2, ?**

Let $T \geq 1$ and $\mu^* \in \mathbb{R}^d$. Assume that for all $\theta \in \cup_{t=1}^T \Theta_t$, $|\theta \cdot \mu^*| \leq 1$, $\|\mu^*\| \leq 1$ and $\|\theta_t\| \leq 1$, then LinUCB with C_t defined in (6.4) satisfies the pseudo-regret bound

$$\bar{R}_T \leq \square_\lambda d\sqrt{T}\log(T)\,,$$

where $\square_\lambda$ is a constant that may depend on λ.

Proof. Let $\delta = 1/T$. By Lemma 6.4, with probability $1 - 1/T$,

$$\forall t \geq 1, \qquad \mu^* \in C_t\,. \tag{6.5}$$

Step 1: Small instantaneous regrets under the event (6.5). Assume that (6.5) holds. Let

$$\theta_t^* \overset{\text{def}}{=} \max_{\theta \in \Theta_t} \theta \cdot \mu^* \qquad \text{and} \qquad r_t \overset{\text{def}}{=} (\theta_t^* - \theta_t) \cdot \mu^*$$

be respectively the optimal decision and the instantaneous regret at round t. We also define

$$\tilde{\mu}_t \in \arg\max_{\mu \in C_t} \big\{\theta_t \cdot \mu\big\}.$$

Since $\mu^* \in C_t$, we have

$$\theta_t^* \cdot \mu^* \ \leq\ \max_{\mu \in C_t} \big\{\theta_t^* \cdot \mu\big\} \overset{(6.2)}{=} UCB_t(\theta_t^*) \overset{(6.3)}{\leq} UCB_t(\theta_t) \ = \max_{\mu \in C_t} \big\{\theta_t \cdot \mu\big\} = \theta_t \cdot \tilde{\mu}_t\,,$$

which entails because μ^* and $\tilde{\mu}_t$ belong to C_t,

$$r_t \overset{\text{def}}{=} (\theta_t^* - \theta_t) \cdot \mu^* \leq \theta_t \cdot (\tilde{\mu}_t - \mu^*) \overset{\text{Cauchy-Schwarz}}{\leq} \|\theta_t\|_{V_{t-1}^{-1}} \big\|\tilde{\mu}_t - \mu^*\big\|_{V_{t-1}} \leq 2\|\theta_t\|_{V_{t-1}^{-1}} \beta(1/T^2)\,.$$

Therefore, summing over $t = 1, \ldots, T$ and using $r_t \leq 2$, we have

$$\begin{aligned}
R_T \overset{\text{def}}{=} \sum_{t=1}^T r_t &\leq \sqrt{T \sum_{t=1}^T r_t^2} && \leftarrow \text{ Jensen's inequality} \\
&\leq 2\sqrt{T \sum_{t=1}^T \min\Big\{\|\theta_t\|^2_{V_{t-1}^{-1}} \beta(1/T^2)^2, 1\Big\}} \\
&\leq 2\beta(1/T^2)\sqrt{T \sum_{t=1}^T \min\Big\{\|\theta_t\|^2_{V_{t-1}^{-1}}, 1\Big\}} && \leftarrow\ \beta_T(1/T^2) \geq 1 \\
&\leq 2\beta(1/T^2)\sqrt{T \sum_{t=1}^T \log\Big(1 + \|\theta_t\|^2_{V_{t-1}^{-1}}\Big)} && \leftarrow\ \min\{u, 1\} \leq 2\log(1+u)\,.
\end{aligned}$$

But, we have

$$1 + \|\theta_t\|^2_{V_{t-1}^{-1}} = \det\big(1 + \|\theta_t\|^2_{V_{t-1}^{-1}}\big)$$

$$
\begin{aligned}
&= \det\left(V_{t-1}^{-1}\big(V_{t-1} + V_{t-1}^{1/2}\|\theta_t\|^2_{V_{t-1}^{-1}} V_{t-1}^{1/2}\big)\right) && \leftarrow \text{using} \det(I+AB) = \det(I+BA) \\
&= \det\left(V_{t-1}^{-1}\big(V_{t-1} + \theta_t\theta_t^\top\big)\right) && \leftarrow \|\theta_t\|_{V_{t-1}^{-1}} = V_{t-1}^{-1/2}\theta_t\theta_t^\top V_{t-1}^{-1/2} \\
&= \det\left(V_{t-1}^{-1}V_t\right) && \leftarrow V_t = V_{t-1} + \theta_t\theta_t^\top \\
&= \frac{\det(V_t)}{\det(V_{t-1})} && \leftarrow \det(A^{-1}B) = \frac{\det(B)}{\det(A)}\,.
\end{aligned}
$$

Substituting into the regret bound, the sum telescopes and it entails

$$
R_T \le 2\beta(1/T^2)\sqrt{T\log\left(\frac{\det(V_T)}{\det(V_0)}\right)}\,.
$$

Then, using $V_0 \overset{\text{def}}{=} \lambda I_d$ and since $V_T = \lambda I_d + \sum_{t=1}^T \theta_t\theta_t^\top$ with $\|\theta_t\| \le 1$, all eigenvalues of V_T lie in $[\lambda, \lambda+T]$ which yields

$$
\det(V_0) = \lambda^d \qquad \text{and} \qquad \det(V_T) \le \big(\lambda+T\big)^d\,.
$$

Plugging back into the previous upper-bound and using that $\beta(1/T^2) \le \square_\lambda\sqrt{d\log T}$

$$
R_T \le 2\sqrt{dT\beta(1/T)\log\left(1+\frac{T}{\lambda}\right)} \le \square_\lambda d\sqrt{T}\log T\,.
$$

Part 2: without the event (6.4) We because $r_t \le 2$, almost surely $R_T \le 2T$, and we have

$$
\begin{aligned}
\bar{R}_T = \mathbb{E}\left[R_T\right] &\le \mathbb{E}\left[R_T \middle| \text{Event (6.4)}\right]\mathbb{P}\{\text{Event (6.4)}\} + 2T\big(1-\mathbb{P}\{\text{ Event (6.4)}\}\big) \\
&\le \square d\sqrt{T}\log T + 2\,.
\end{aligned}
$$

This concludes the proof. □

Better regret with assumptions It is worth pointing out that if we make additional assumptions, it is possible to improve the regret bound $O(d\sqrt{T}\log T)$. A first setting corresponds to the case where the set of available actions at time t is fixed and finite; i.e., the learner needs to choose $\theta_t \in \Theta$ where $|\Theta| = K$. Then, it is possible to achieve

$$
R_T \le \square\sqrt{Td\log(TK)}\,,
$$

which improves the previous bound by a factor $\sqrt{d}/\log(K)$ and improves the classical bound of UCB $O(\sqrt{TK\log T})$ by a factor $K/\sqrt{d}$. These improvements can be significant when $K \gg d \gg 1$. We refer the curious reader to [Lattimore and Szepesvári, 2020, Chapter 22].

Another possible improvement when $d \gg 1$ is to assume that μ^* is m_0-sparse (i.e., most of its components are zero). Then under assumptions, one can get a regret of order $\tilde{O}(\sqrt{dm_0T})$.

6.2 Other possible extensions of bandits

Note that there exist many different extensions of stochastic bandits to make it more realistic or with improved regret.

- *Bandit with delays:* For instance, consider the example of the website which wants to display ads. The website does not observe if there is no click, he needs to fix some time after which he consider that the visitor will not click, and if the visitor stays long on the webpage, the website may need to display ads to other visitors before getting the rewards. There is thus delayed feedback the website needs to deal with.
- *Non stationarity*
- *Combinatorial bandits*
- *Dueling bandits*
- ...

We refer the interested student to the monograph Lattimore and Szepesvári [2020] for more information on these settings. Next week, we will deal with adversarial bandits.

Bibliography

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.* Cambridge university press, 2006.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132(1):1–63, 1997.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms.* Cambridge University Press, 2020.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, pages 693–721, 2013.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 928–936, 2003.