Princeton University - Department of Operations Research and Financial Engineering

Introduction to Online Optimization

Sébastien Bubeck

December 14, 2011

Contents

Chapter 1. Introduction 5
- 1.1. Statistical learning theory 5
- 1.2. Online learning 7
- 1.3. General objective of the course 11
- 1.4. Different notions of convexity 11
- References 13

Chapter 2. Online optimization on the simplex 15
- 2.1. Exponentially weighted average forecaster (Exp strategy) 15
- 2.2. Bounded convex loss and expert regret 16
- 2.3. Exp-concave loss and expert regret 17
- 2.4. Lower bound 18
- 2.5. Anytime strategy 18
- 2.6. Subdifferentiable loss with bounded subgradient 20
- 2.7. Online finite optimization 20
- References 22

Chapter 3. Continuous exponential weights 23
- 3.1. Bounded convex loss 24
- 3.2. Exp-concave loss 25
- References 25

Chapter 4. Online gradient methods 27
- 4.1. Online Gradient Descent (OGD) 27
- 4.2. Strongly convex loss with bounded subgradient 28
- References 29

Chapter 5. Online mirror descent 31
- 5.1. Convex analysis 31
- 5.2. Online Mirror Descent (OMD) 33
- 5.3. Subgradient bounded in an arbitrary norm 35
- References 37

Chapter 6. Online combinatorial optimization 39
- 6.1. Examples 40
- 6.2. Lower bound 43
- 6.3. Expanded Exp (Exp2) 43
- 6.4. OMD with negative entropy 47
- 6.5. Follow the perturbated leader (FPL) 47
- References 49

Chapter 7. Limited feedback 51
7.1. Online Stochastic Mirror Descent (OSMD) 51
7.2. Online combinatorial optimization with semi-bandit feedback 53
7.3. Online linear optimization with bandit feedback 55
7.4. Lower bounds 62
7.5. Two-points bandit feedback 67
7.6. Online convex optimization with bandit feedback 69
References 69

Chapter 8. Online stochastic optimization 71
8.1. Optimism in face of uncertainty 71
8.2. Stochastic multi-armed bandit 72
References 76

Chapter 9. Open Problems 79

Bibliography 83

CHAPTER 1

Introduction

1.1. Statistical learning theory

In a world where automatic data collection becomes ubiquitous, statisticians must update their paradigms to cope with new problems. Whether we discuss the Internet network, consumer data sets, or financial market, a common feature emerges: huge amounts of dynamic data that need to be understood and quickly processed. This state of affair is dramatically different from the classical statistical problems, with many observations and few variables of interest. Over the past decades, learning theory tried to address this issue. One of the standard and thoroughly studied models for learning is the framework of statistical learning theory. We start by briefly reviewing this model.

1.1.1. Statistical learning protocol. The basic protocol of statistical learning is the following:

- Observe $Z_1, \dots, Z_n \in \mathcal{Z}$. We assume that it is an i.i.d. sequence from an unknown probability distribution $\mathbb{P}$.
- Make decision (or choose action) $a(Z_1, \dots, Z_n) \in \mathcal{A}$ where $\mathcal{A}$ is a given set of possible actions.
- Suffer an (average) loss $\mathbb{E}_{Z \sim \mathbb{P}}\, \ell(a(Z_1, \dots, Z_n), Z)$ where $\ell : \mathcal{A} \times \mathcal{Z} \to \mathbb{R}_+$ is a given loss function.

Objective: Minimize (and control) the excess risk:

$$r_n = \mathbb{E}_{Z \sim \mathbb{P}}\, \ell(a(Z_1, \dots, Z_n), Z) - \inf_{a \in \mathcal{A}} \mathbb{E}_{Z \sim \mathbb{P}}\, \ell(a, Z).$$

The excess risk represents how much extra (average) loss one suffers compared to the optimal decision.

REMARK 1.1. *Controlling the excess risk means finding an upper bound on r_n which holds either in expectation (with respect to the sequence $Z_1, \dots, Z_n$) or with probability at least $1 - \delta$. Usually the upper bound is expressed in terms of some complexity measure of $\mathcal{A}$ and ℓ. Moreover if the upper bound depends on $\mathbb{P}$ one says that it is a* distribution-dependent *bound, while if it is independent from $\mathbb{P}$ it is a* distribution-free *bound.*

This formulation is very general and encompasses many interesting problems. In the following we detail a few of them.

1.1.2. Example 1: Regression estimation.

- Here the observed data corresponds to pairs of points, that is $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$.
- The set of possible decisions is a set of functions from $\mathcal{X}$ to $\mathcal{Y}$, that is $\mathcal{A} \subset \{f : \mathcal{X} \to \mathcal{Y}\}$.

- The loss $\ell(a,(x,y))$ measures how well the function $a : \mathcal{X} \to \mathcal{Y}$ predicts y given x. For instance if $\mathcal{Y}$ is a normed space a typical choice is $\ell(a,(x,y)) = ||a(x) - y||$.

In other words, in this example we have a dataset of pairs (input, output) and we want to design a good function to predict outputs, given inputs, even for unseen inputs (this is the problem of *generalization*). A popular particular case of this setting is the linear regression problem for least squares loss. It corresponds to $\mathcal{X} = \mathbb{R}^d$; $\mathcal{Y} = \mathbb{R}$; $\mathcal{A}$ is the set of affine functions on $\mathbb{R}^d$, that is $a(x) = w_a^T x + b_a$ for some $(w_a, b_a) \in \mathbb{R}^d \times \mathbb{R}$; and $\ell(a,(x,y)) = (a(x) - y)^2$.

1.1.3. Example 2: Pattern recognition (or classification). This problem is a regression estimation task where $\mathcal{Y} = \{-1, 1\}$. The loss is the so-called zero-one loss, $\ell(a,(x,y)) = \mathbb{1}_{a(x) \neq y}$.

One often restricts to linear pattern recognition, which corresponds to decisions of the form $a(x) = \text{sign}(w_a^T x + b_a)$ (that is one chooses a decision hyperplane). In this setting it is common to consider *(convex) relaxations* of the zero-one loss such as:

- the logistic loss: $\ell(a,(x,y)) = \log(1 + \exp(-y(w_a^T x + b_a)))$,
- the hinge loss: $\ell(a,(x,y)) = \max(0, 1 - y(w_a^T x + b_a))$.

Note that the zero-one loss is not convex in (w_a, b_a), on the contrary to the logistic loss and the hinge loss.

1.1.4. Example 3: Density estimation. Here the goal is to estimate the probability distribution $\mathbb{P}$. In that case $\mathcal{A}$ is a set of probability distributions (or more precisely probability densities with respect to some fixed measure) over $\mathcal{Z}$, and $\mathcal{Z}$ is usually a finite set or $\mathcal{Z} = \mathbb{R}$ (in information theory $\mathcal{Z}$ is called the alphabet). The most common loss function for this problem is the log loss: $\ell(a, z) = -\log a(z)$. Note that if $\mathbb{P} \in \mathcal{A}$, then the excess risk is exactly the Kullback-Leibler divergence between a and $\mathbb{P}$.

1.1.5. Example 4: High-dimensional linear regression under sparsity scenario. A particular linear regression task has gained a lot of popularity in the recent years, it can be described as follows. We assume that there exists $w \in \mathbb{R}^d$ such that $Y_i = w^T X_i + \xi_i$ where (ξ_i) is an i.i.d. sequence of white noise (i.e. $\mathbb{E}(\xi_i | X_i) = 0$). The high-dimensional setting that we consider corresponds to the regime where $n \ll d$. While in general the regression problem is intractable in this regime, one can still hope to obtain good performances if the vector w is *sparse*, that is if the number s of non-zero components (or approximately non-zero) of w is small compared to the number of examples n. Thus we have $s < n \ll d$. In this setting one often proves excess risk bounds for the loss $\ell(a,(x,y)) = (a(x) - y)^2 + \lambda ||w_a||_0$ or for its convex relaxation $\ell(a,(x,y)) = (a(x) - y)^2 + \lambda ||w_a||_1$. The resulting bounds are called *sparsity oracle inequalities*.

1.1.6. Example 5: Compressed sensing. The problem of reconstruction in compressed sensing can also be viewed as a learning problem. We assume here that there is an unknown vector $x \in \mathbb{R}^d$, and that we have access to x through n noisy measurements $Z_i = W_i^T x + \xi_i$ where (ξ_i) is an i.i.d. sequence of white noise and (W_i) is a sequence of random variables with known distribution (typically a Gaussian distribution with a specific covariance matrix). The goal is to recover x from the measurements in the regime where $n \ll d$ (high-dimensional setting), under a sparsity assumption on x. Here the rules for the loss function are slightly different from the standard statistical learning protocol. Indeed we do not evaluate the reconstruction $a(Z_1, \dots, Z_n)$ on a new example Z, but rather we look at how far a is from the true signal x (this is the difference between *prediction* and *estimation*). Thus a typical choice is $\ell(a, x) = ||a - x||^2$.

1.1.7. Example 6: Matrix completion (or collaborative filtering). Matrix completion is yet another example of high-dimensional learning. Here $\mathbb{P}$ is the uniform distribution over the entries of an unknown matrix $M \in \mathbb{R}^{m \times d}$, and the goal is to reconstruct the matrix M. The high-dimensional setting corresponds to $n \ll m \times d$. To make the problem feasible, a natural assumption is that M has a rank k which is small compared to the number of examples n, thus we have $k < n \ll m \times d$. Several losses can be considered, either for the *prediction* version of the problem or for the *estimation* version.

1.1.8. Real world examples. Statistical learning theory had many successes in real world applications. We give here a few keywords for the most renowed applications: computer vision, spam detection, natural language processing (see in particular Watson for Jeopardy!), bioinformatics (functional genomics), collaborative filtering (see in particular the Netflix challenge), brain-computer interface, ect.

1.2. Online learning

Despite its many successes, the statistical learning theory fails at addressing one the key features of the new massive data: the dynamic aspect. Online learning is an attempt to overcome this shortcoming. In these notes we mostly use the name online optimization rather than online learning, which seems more natural for the protocol described below.

1.2.1. Online optimization protocol. Online learning is a natural extension of statistical learning. In some sense this model can be seen as pushing to its limit the agenda of distribution-free results. Indeed the essential difference between online learning and statistical learning, in addition to the sequential aspect, is the fact that no probabilistic assumption is made on the dataset $Z_1, \dots, Z_n$. By reducing the problem to its minimal struture, one can hope to attain the intrinsic difficulty of learning. As it turns out, this change of perspective indeed proved to be very fruitful. It had a profound impact and fundamentally changed the landscape of modern machine learning.

Formally the online learning protocol can be described as follows. At every time step $t = 1, 2, \dots, n$:

- Choose action $a_t \in \mathcal{A}$.
- Simultaneously an adversary (or Nature) selects $z_t \in \mathcal{Z}$.
- Suffer loss $\ell(a_t, z_t)$.
- Observe z_t.

Objective: Minimize (and control) the cumulative regret:

$$R_n = \sum_{t=1}^{n} \ell(a_t, z_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell(a, z_t).$$

In words the cumulative regret compares the cumulative loss of the player to the cumulative loss of the best action in hindsight. Here one strives for regret bounds (i.e. upper bounds on R_n) independent of the adversary's moves $z_1, \dots, z_n$.

Again this formulation is very general, and we detail now a few specific instances.

1.2.2. Example 1: Online regression, online classification. At each time step the player chooses a regression function $a_t : \mathcal{X} \to \mathcal{Y}$ and simultaneously the adversary selects an input/output pair $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$. The player suffers $\ell(a_t, (x_t, y_t))$ and observes the pair (x_t, y_t). Here one usually restricts $\mathcal{A}$ to a small class of regression functions, such as decision hyperplanes in the case of online pattern recognition.

A particularly interesting application of this example is the problem of dynamic pricing. Consider a vendor that serves a sequence of customer for a particular item. For every customer t, the vendor may have some information x_t about him. Based on this information he sets a price $a_t(x_t) \in \mathcal{Y}$. On the other hand the customer had a maximal price in mind $y_t \in \mathcal{Y}$ that he is willing to pay for the item. In this setting a natural loss could be $\ell(a, (x, y)) = -a(x)\mathbb{1}_{a(x) \leq y}$. Note that in this application it is rather unnatural to assume that the customer will reveal his maximal price y_t. This is a first example of the so-called *limited feedback* problems. More precisely, here one can assume that the feedback (i.e. the observation) corresponds to the incurred loss $\ell(a_t, (x_t, y_t))$ rather than the pair (x_t, y_t).

1.2.3. Example 2: Sequential investment. We consider here an idealized stock market with d assets that we model as follows: a market vector $z \in \mathbb{R}_+^d$ represents the price relatives for a given trading period. That is, if one invests $a \in \mathbb{R}_+^d$ in the market (i.e. $a(i)$ is invested in asset i), then the return at the end of the trading period is $\sum_{i=1}^d a(i)z(i) = a^T z$.

We consider now the problem of sequential investment in this stock market. At every trading period t, the current total capital is denoted by W_{t-1} and the player invests its total capital according to the proportions $a_t \in \mathcal{A} = \{a \in \mathbb{R}_+^d, \sum_{i=1}^d a(i) = 1\}$ (in other words the action set is the $(d-1)$-simplex). Simultaneously the market chooses the market vector $z_t \in \mathbb{R}_+^d$. The new wealth at the end of period t satisfies:

$$W_t = \sum_{i=1}^d a_t(i) W_{t-1} z_t(i) = W_{t-1} a_t^T z_t = W_0 \prod_{s=1}^t a_s^T z_s.$$

An important class of investment strategies for this problem is the set of constantly rebalanced portfolios, which correspond to static decisions, that is $a_t = a, \forall t \geq 1$. In other words the player rebalances, at every trading period t, his current wealth W_{t-1}, according to the proportions a.

We consider now the problem of being competitive with respect to the class of all constantly rebalanced portfolios. One can consider the competitive wealth ratio:

$$\sup_{a \in \mathcal{A}} \frac{W_n^a}{W_n},$$

where $W_n^a = W_0 \prod_{s=1}^n a^T z_s$ represents the wealth of the constantly rebalanced portfolio a at the end of the n^{th} trading period. The logarithmic wealth ratio is thus:

$$\sum_{t=1}^{n} - \log(a_t^T z_t) - \inf_{a \in \mathcal{A}} \sum_{t=1}^{n} - \log(a^T z_t),$$

which is exactly the cumulative regret for the online optimization problem on the $(d-1)$-simplex and with the log-loss $\ell(a, z) = -\log(a^T z)$.

In the case of a Kelly market (i.e. when z has exactly one non-zero component which is equal to one), this problem is exactly the online equivalent of density estimation (with alphabet $\{1, \dots, d\}$) for the log-loss.

1.2.4. Example 3: Prediction with expert advice. This example is particularly important, as it was the first framework proposed for online learning. Here we assume that the action set $\mathcal{A}$ is convex, and that at every time step t, the player receives a set of *advice* $(b_t(i))_{1 \le i \le d} \in \mathcal{A}^d$. The usual interpretation is that there is a set of *experts* playing the online optimization game simultaneously with the player. We make no restriction on the experts except that their decisions at time t are revealed to the player before he makes his own choice a_t. Note in particular that one can easily assume that the experts have access to external sources of information to make their decisions. In this setting the goal is to perform as well as the best expert. Thus we consider the following cumulative regret with respect to the experts:

$$R_n^E = \sum_{t=1}^{n} \ell(a_t, z_t) - \min_{1 \le i \le d} \sum_{t=1}^{n} \ell(b_t(i), z_t). \tag{1.1}$$

Here the goal is to obtain bounds independent of the expert advice and the adversary's moves. Any known strategy for this problem computes the action a_t by taking a convex combination of the expert advice $b_t(i)$. In that case, one can view the prediction with expert advice as an online optimization problem over the $(d-1)$-simplex, where one restricts to the vertices of the $(d-1)$-simplex for the comparison class in the definition of the cumulative regret. More precisely the reduction goes as follows: Let $\Delta_d = \{p \in \mathbb{R}_+^d, \sum_{i=1}^d p(i) = 1\}$ be the $(d-1)$-simplex, let $e_1, \dots, e_d$ be the canonical basis of $\mathbb{R}^d$ (which correspond to the vertices of the simplex), and let $\bar{\mathcal{Z}} = \mathcal{A}^d \times \mathcal{Z}$ be the set of expert advice and adversary's moves. We define now a new loss function $\bar{\ell}$ on $\Delta_d \times \bar{\mathcal{Z}}$ by

$$\bar{\ell}(p, (b, z)) = \ell\left(\sum_{i=1}^{d} p(i) b(i), z\right).$$

In words the loss $\bar{\ell}$ of a point in the simplex is the original loss ℓ of the corresponding convex combination of the expert advice. We can now consider the online optimization problem where the action set is Δ_d, the set of moves for the adversary is $\bar{\mathcal{Z}}$, the loss function is $\bar{\ell}$, and the cumulative regret with respect to the experts is defined as:

$$R_n^E = \sum_{t=1}^n \bar{\ell}(p_t, (b_t, z_t)) - \min_{1 \le i \le d} \sum_{t=1}^n \ell(e_i, (b_t, z_t)).$$

This restricted notion of cumulative regret for the online optimization problem coincides with the cumulative regret defined in (1.1). Importantly, note that if ℓ is convex in its first argument, then so is $\bar{\ell}$.

The prediction with expert advice is also a very general framework that encompasses many applications. In particular this setting can be used for meta-learning under very weak assumptions: one can for instance combine the predictions of different statistical procedures to obtain a prediction almost as good as the best statistical procedure in our set, even if there is an adversary that is choosing the data (this is some kind of robust model selection).

1.2.5. Example 4: Online linear optimization. Online linear optimization refers to online optimization problems where the loss function is linear in its first argument. We usually restrict to the finite dimensional case, where $\mathcal{A}, \mathcal{Z} \subset \mathbb{R}^d$ and $\ell(a, z) = a^T z$.

Many interesting and challenging applications fit in this framework. The problem of *path planning* is one of them:

- $\mathcal{A}$ represents a set of paths in a given graph with d edges (more precisely the elements of $\mathcal{A} \subset \{0, 1\}^d$ are the incidence vectors for the corresponding paths).
- $z \in \mathcal{Z}$ represents a weight on the graph.
- $a^T z$ is the total weight given by z on path a, it can be interpreted as the cost of sending a bit using path a when the delays in the graphs are given by z.
- The cumulative regret R_n corresponds to the total extra cost of sending n bits according to our strategy $a_1, \dots, a_n$ compared to the cost of the best path in hindsight.

1.2.6. Example 5: Online matrix completion. Here at every time step we predict a matrix a_t of size $m \times d$ with rank bounded by k, the adversary reveals the entry $z_t = (i, j)$ of an unknown (but fixed) matrix M. The loss can be the square loss $\ell(a, (i, j)) = (a_{i,j} - M_{i,j})^2$. A typical relaxation of this problem is to consider matrices with bounded trace norm rather than bounded rank.

1.2.7. Example 6: One-pass offline optimization. Consider the general problem of statistical learning theory described in Section 1.1.1 with a convex set $\mathcal{A}$ and a convex loss function, that is $\forall z \in \mathcal{Z}, a \mapsto \ell(a, z)$ is convex. Now consider an online optimization procedure run sequentially on the data set $Z_1, \dots, Z_n$. Clearly, by convexity of the loss, and by taking $a(Z_1, \dots, Z_n) = \frac{1}{n} \sum_{t=1}^n a_t$, one has:

$$r_n \le \frac{R_n}{n}.$$

In other words one can solve the statistical learning problem by doing a one-pass online optimization on the data set.

1.3. General objective of the course

The general goal of the course is to study the achievable growth rate for the cumulative regret in terms of the number of rounds n, the geometry of the action set $\mathcal{A}$, and under different assumptions on the loss function. The importance of an optimal regret bound can be justified as follows. If the regret is sublinear, then it means that asymptotically one performs almost as well as the best action in hindsight. In other words a regret bound gives an order of magnitude for the *size of the problem* (defined in terms of the complexity measure of $\mathcal{A}$ and ℓ that appears in the regret bound) that one can "solve" with an online learning algorithm. This is similar in spirit to computational complexity bounds, which give an order of magnitude for the input size that one can consider with a given budget of computational resources.

This characterization of the optimal regret in a given scenario is achieved through algorithms to obtain upper bounds on the achievable regret, and through information-theoretic arguments to obtain lower bounds. We also pay particular attention to the computational complexity of the proposed algorithms.

In these lecture notes we focus primarily on Online Convex Optimization, where $\mathcal{A}$ is a convex set and ℓ is a convex function[1]. More precisely we consider five different types of convex functions:

(1) bounded convex loss (with a different treatment for the prediction with expert advice and the general online optimization problem),
(2) exp-concave loss (with a different treatment for the prediction with expert advice and the general online optimization problem),
(3) subdifferentiable loss with bounded subgradients,
(4) strongly-convex loss with bounded subgradients,
(5) linear loss in a combinatorial setting.

Note that in the combinatorial setting $\mathcal{A}$ is a finite set, but we show how to reduce it to a convex problem. In addition we also consider the situation where the player receives only partial information on the adversary's move. A particularly interesting and challenging limited feedback case is the so-called *bandit* problem, where one does not observe z_t but only the suffered loss $\ell(a_t, z_t)$. In Chapter 7 we also consider different variants of the bandit problem.

In the next section we define precisely convexity, strong-convexity and exp-concavity. We also discuss the relation between these notions.

1.4. Different notions of convexity

Definition 1.1. *Let $f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X}$ is a convex set (which is a subset of an arbitrary vector space over the reals). Then one says that*

- *f is convex if $\forall (x, y, \gamma) \in \mathcal{X}^2 \times [0,1]$, $f((1-\gamma)x + \gamma y) \leq (1-\gamma) f(x) + \gamma f(y)$.*
- *f is concave if $-f$ is convex.*
- *f is σ-exp concave ($\sigma > 0$) if $x \mapsto \exp(-\sigma f(x))$ is a concave function.*

If $\mathcal{X} \subset \mathbb{R}^d$, then one says that

[1] Note that when we say that a loss function ℓ satisfies a certain property, we usually mean that $\forall z \in \mathcal{Z}$ the function $a \mapsto \ell(a, z)$ satisfies this property. For instance a convex loss function is such that is $\forall z \in \mathcal{Z}$, $a \mapsto \ell(a, z)$ is a convex function.

- *f is subdifferentiable if $\forall x \in \mathcal{X}$, there exists a subgradient $g \in \mathbb{R}^d$ such that*
$$f(x) - f(y) \leq g^T(x-y), \forall y \in \mathcal{X}.$$
With an abuse of notation, we note $\nabla f(x)$ for a subgradient of f at x.
- *f is α-strongly-convex if it is subdifferentiable and $\forall x \in \mathcal{X}$,*
$$f(x) - f(y) \leq \nabla f(x)^T(x-y) - \frac{\alpha}{2}||x-y||_2^2, \forall y \in \mathcal{X}.$$
- *f is α-strongly-convex with respect to a norm $||\cdot||$ if it is subdifferentiable and $\forall x \in \mathcal{X}$,*
$$f(x) - f(y) \leq \nabla f(x)^T(x-y) - \frac{\alpha}{2}||x-y||^2, \forall y \in \mathcal{X}.$$

The following Proposition is an easy exercise.

Proposition 1.1. *If a function is either α-strongly convex, or σ-exp-concave, or subdifferentiable, then it is convex. On the other hand if f is a convex function, $\mathcal{X} \subset \mathbb{R}^d$, and either f is differentiable or $\mathcal{X}$ is an open set, then f is subdifferentiable (in the former case the subgradient coincides with the gradient).*

The next lemma exhibits an equivalent definition for strong convexity.

Lemma 1.1. *Let f be a differentiable function. Then f is α-strongly convex with respect to a a norm $||\cdot||$ if and only if*
$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq \alpha||x-y||^2, \forall x, y \in \mathcal{X}.$$

Proof. One direction is trivial by simply summing the strong convexity condition for the pairs (x,y) and (y,x). For the other direction, let
$$h : t \in [0,1] \mapsto f(x + t(y-x)).$$
Note that
$$h'(t) - h'(0) = (y-x)^T \nabla f(x + t(y-x)) - \nabla f(x) \geq \alpha t||x-y||^2.$$
One has:
$$f(y) - f(x) = h(1) - h(0) = \int_0^1 h'(t)dt \geq \nabla f(x)^T(y-x) + \frac{\alpha}{2}||x-y||^2,$$
which concludes the proof. □

The next proposition relates the different convexity notions to properties of the Hessian of f and exhibits different relations between strong-convexity and exp-concavity.

Proposition 1.2. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex set, and $f : \mathcal{X} \to \mathbb{R}$ a twice continuously differentiable function. Then the following holds true.*

- *f is convex if and only if $\nabla^2 f(x) \succeq 0, \forall x \in \mathcal{X}$.*
- *f is α-strongly convex if and only if $\nabla^2 f(x) \succeq \alpha I_d, \forall x \in \mathcal{X}$.*
- *f is σ-exp concave if and only if $\nabla^2 \exp(-\sigma f(x)) \preceq 0, \forall x \in \mathcal{X}$.*

Moreover, if f is α-strongly convex, with bounded gradient $||\nabla f(x)||_2 \leq B, \forall x \in \mathcal{X}$, then f is $\frac{\alpha}{B^2}$-exp concave,

Finally we end this chapter with a description of the convexity properties of the log-loss.

PROPOSITION 1.3. *The log-loss* $(a, z) \in \mathbb{R}_+^d \times \mathbb{R}_+^d \mapsto -\log(a^T z)$ *has the following properties.*

- *It is* 1*-exp concave.*
- *It takes unbounded values and it has unbounded gradient, even when restricted to the* $(d-1)$*-simplex.*
- *It is not* α*-strongly convex, for any* $\alpha > 0$.

References

There are many references for statistical learning theory, see for example:

- V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995
- L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, 1996

The best reference for prediction with expert advice is:

- N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, 2006

Online convex optimization was introduced in:

- M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, 2003

The point of view presented in this chapter was inspired by Peter Bartlett's course at Institut Henri Poincaré in 2011. Example 6 was inspired by [11], and the corresponding averaging idea is sometimes called the Polyak-Ruppert averaging.

CHAPTER 2

Online optimization on the simplex

In this chapter we consider the online learning problem with the $(d-1)$-simplex $\Delta_d = \{p \in \mathbb{R}_+^d, \sum_{i=1}^d p(i) = 1\}$ as the action set. We recall that in online optimization the cumulative regret is defined as (with the notation a_t replaced by the more natural p_t for points in the simplex):

$$R_n = \sum_{t=1}^n \ell(p_t, z_t) - \inf_{q \in \Delta_d} \sum_{t=1}^n \ell(q, z_t), \tag{2.1}$$

while for the prediction with expert advice problem the comparison is restricted to the canonical basis $e_1, \dots, e_d$ of $\mathbb{R}^d$ and one obtains the expert regret:

$$R_n^E = \sum_{t=1}^n \ell(p_t, z_t) - \inf_{1 \le i \le d} \sum_{t=1}^n \ell(e_i, z_t). \tag{2.2}$$

The two notions coincide for linear losses, but they are different for general convex losses. We refer to upper bounds on R_n as *regret bounds* and on R_n^E as *expert regret bounds*.

In this chapter we first restrict our attention to prediction with expert advice, that is to the expert regret (2.2). Keep in mind that online linear optimization on the simplex is equivalent to prediction with expert advice with linear loss (since for linear losses $R_n^E = R_n$), thus by solving the latter we also obtain a solution for the former.

We start in Section 2.1 by describing the most renowed strategy of online learning, namely the exponentially weighted average forecaster. Then we proceed to prove an expert regret bound for this strategy in the setting of bounded convex losses in Section 2.2 (which implies in particular a regret bound for online bounded linear optimization) and exp-concave losses in Section 2.3. We also show in Section 2.4 that the expert regret bound for bounded convex losses is unimprovable (in fact we show that the regret bound for bounded linear losses is unimprovable). Then in Section 2.6 we show how to extend these results to the regret (2.1) for subdifferentiable losses with bounded subgradient.

Finally in Section 2.7 we show how to apply results from online optimization on the simplex to online optimization with a finite action set.

2.1. Exponentially weighted average forecaster (Exp strategy)

The idea of this fundamental strategy goes as follows. One wants to perform as well as the best vertex in the simplex (expert regret). A simple strategy to try to do this is to assign a weight to each vertex on the basis of its past performances, and then take the corresponding convex combination of the vertices as its decision p_t. The weight of a vertex should be a non-increasing function of its past cumulative loss. Here we choose the exponential function to obtain the following decision:

$$p_t = \sum_{i=1}^{d} \frac{w_t(i)}{\sum_{j=1}^{d} w_t(j)} e_i,$$

where

$$w_t(i) = \exp\left(-\eta \sum_{s=1}^{t-1} \ell(e_i, z_s)\right),$$

and $\eta > 0$ is a fixed parameter. In other words, $\forall i \in \{1, \dots, d\}$,

$$p_t(i) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell(e_i, z_s)\right)}{\sum_{j=1}^{d} \exp\left(-\eta \sum_{s=1}^{t-1} \ell(e_j, z_s)\right)}.$$

Note that $w_t(i) = w_{t-1}(i) \exp\left(-\eta \ell(e_{t-1}, z_{t-1})\right)$. Thus the computational complexity of one step of the Exp strategy is of order $O(d)$.

2.2. Bounded convex loss and expert regret

In this section we analyze the expert regret of the Exp strategy for bounded convex losses. Without loss of generality we assume $\ell(p, z) \in [0, 1], \forall (p, z) \in \Delta_d \times \mathcal{Z}$. Indeed if $\ell(p, z) \in [m, M]$ then one can work with a rescaled loss $\bar{\ell}(a, z) = \frac{\ell(a,z)-m}{M-m}$.

We use the following fundamental result from probability theory, known as Hoeffding's inequality:

Lemma 2.1. *Let X be a real random variable with $a \leq X \leq b$. Then for any $s \in \mathbb{R}$,*

$$\log\left(\mathbb{E}\exp(sX)\right) \leq s\mathbb{E}X + \frac{s^2(b-a)^2}{8}.$$

Now we can prove the following regret bound:

Theorem 2.1. *For any convex loss taking values in $[0, 1]$, the Exp strategy satisfies:*

$$R_n^E \leq \frac{\log d}{\eta} + \frac{n\eta}{8}.$$

In particular with $\eta = 2\sqrt{\frac{2 \log d}{n}}$ it satisfies:

$$R_n^E \leq \sqrt{\frac{n \log d}{2}}.$$

PROOF. Let $w_t(i) = \exp\left(-\eta \sum_{s=1}^{t-1} \ell(e_i, z_s)\right)$ and $W_t = \sum_{i=1}^{d} w_t(i)$ (by definition $w_1(i) = 1$ and $W_1 = d$). Then we have:

$$\begin{aligned}
\log \frac{W_{n+1}}{W_1} &= \log\left(\sum_{i=1}^{d} w_{n+1}(i)\right) - \log d \\
&\geq \log\left(\max_{1 \leq i \leq d} w_{n+1}(i)\right) - \log d \\
&= -\eta \min_{1 \leq i \leq d} \sum_{t=1}^{n} \ell(e_i, z_t) - \log d.
\end{aligned}$$

On the other hand, we have $\log \frac{W_{n+1}}{W_1} = \sum_{t=1}^{n} \log \frac{W_{t+1}}{W_t}$ and

$$\begin{aligned}
\log \frac{W_{t+1}}{W_t} &= \log\left(\sum_{i=1}^{d} \frac{w_t(i)}{W_t} \exp(-\eta \ell(e_i, z_t))\right) \\
&= \log\left(\mathbb{E} \exp(-\eta \ell(e_I, z_t)\right) \text{ where } \mathbb{P}(I = i) = \frac{w_t(i)}{W_t} \\
&\leq -\eta \mathbb{E} \ell(e_I, z_t) + \frac{\eta^2}{8} \text{ (Hoeffding's lemma)} \\
&\leq -\eta \ell(\mathbb{E} e_I, z_t) + \frac{\eta^2}{8} \text{ (Jensen's inequality)} \\
&= -\eta \ell(p_t, z_t) + \frac{\eta^2}{8}.
\end{aligned}$$

Thus we proved:

$$\sum_{t=1}^{n} \left(-\eta \ell(p_t, z_t) + \frac{\eta^2}{8}\right) \geq -\eta \min_{1 \leq i \leq d} \sum_{t=1}^{n} \ell(e_i, z_t) - \log d.$$

In other words:

$$R_n^E \leq \frac{\log d}{\eta} + \frac{n\eta}{8},$$

which concludes the proof. □

2.3. Exp-concave loss and expert regret

In this section we study another type of convex loss functions, namely the exp-concave losses. Recall that a loss function is σ-exp-concave if $\forall z \in \mathcal{Z}$, $p \mapsto \exp(-\sigma \ell(p, z))$ is a concave function. Note in particular that this definition does not require boundedness.

THEOREM 2.2. *For any σ-exp-concave loss, the Exp strategy with parameter $\eta = \sigma$ satisfies:*

$$R_n^E \leq \frac{\log d}{\sigma}.$$

PROOF. In the previous proof it suffices to replace Hoeffding's lemma followed by Jensen's inequality by a single Jensen's inequality applied to $p \mapsto \exp(-\eta \ell(p, z))$. □

2.4. Lower bound

In this section we prove that the expert regret bound for general convex and bounded losses is unimprovable. In fact we show that the regret bound for online bounded linear losses is unimprovable, and that it is even optimal up to the constant. The proof is based on the following result from probability theory.

Lemma 2.2. *Let $(\sigma_{i,t})_{1 \leq i \leq d, 1 \leq t \leq n}$ be i.i.d Rademacher random variables. The following holds true:*

$$\lim_{d \to +\infty} \lim_{n \to +\infty} \frac{\mathbb{E}\left(\max_{1 \leq i \leq d} \sum_{t=1}^{n} \sigma_{i,t}\right)}{\sqrt{2n \log d}} = 1.$$

Theorem 2.3. *Consider the loss $\ell : (p, z) \in \Delta_d \times \{0,1\}^d \mapsto p^T z \in [0,1]$. For any strategy, the following holds true:*

$$\sup_{n,d} \sup_{adversary} \frac{R_n}{\sqrt{(n/2) \log d}} \geq 1.$$

Proof. The proof relies on the probabilistic method. Instead of constructing explicitly a difficult adversary for a given strategy, we put a (uniform) distribution on the possible adversaries and show that the average regret (with respect to the drawing of the adversary) is large. In other words we use the inequality:

$$\sup_{\text{adversary}} R_n \geq \mathbb{E}_{\text{adversary}} \, R_n.$$

More precisely we consider an array $(\varepsilon_{i,t})_{1 \leq i \leq d, 1 \leq t \leq n}$ of i.i.d Bernoulli random variable with parameter $1/2$. The adversary corresponding to this array sets $z_t = (\varepsilon_{1,t}, \dots, \varepsilon_{d,t})$. We compute now the expected regret, where the expectation is taken with respect to the random draw of the Bernoulli array. Here a little bit of care is needed. Indeed the action p_t taken by the player is now a random variable. More precisely $p_t \in \sigma(z_1, \dots, z_{t-1})$, and thus:

$$\mathbb{E}\, p_t^T z_t = \mathbb{E}\left(\mathbb{E}\left(p_t^T z_t | z_1, \dots, z_{t-1}\right)\right) = \mathbb{E}\left(\sum_{i=1}^{d} p_t(i) \mathbb{E}\left(\varepsilon_{i,t} | z_1, \dots, z_{t-1}\right)\right) = \frac{1}{2}.$$

In particular one obtains that

$$\mathbb{E} R_n = \frac{n}{2} - \mathbb{E} \min_{1 \leq i \leq d} \sum_{t=1}^{n} \varepsilon_{i,t} = \frac{1}{2} \mathbb{E} \max_{1 \leq i \leq d} \sum_{t=1}^{n} \sigma_{i,t},$$

where $\sigma_{i,t} = 1 - 2\varepsilon_{i,t}$ is a Rademacher random variable. Using Lemma 2.2 ends the proof. □

2.5. Anytime strategy

One weakness of Exp is that the optimal parameter η depends on the horizon n. In many applications this horizon is unknown, thus it is of interest to obtain a strategy which admits a regret bound uniformly over time. We show here that this goal is easily achieved with a time-varying parameter η_t.

Theorem 2.4. *For any convex loss with values in $[0,1]$, the Exp strategy with time-varying parameter $\eta_t = 2\sqrt{\frac{\log d}{t}}$ satisfies for all $n \geq 1$:*

$$R_n^E \leq \sqrt{n \log d}.$$

PROOF. Let $w_t(i) = \exp\left(-\eta_t \sum_{s=1}^{t-1} \ell(e_i, z_s)\right)$ and $W_t = \sum_{i=1}^d w_t(i)$ (by definition $w_1(i) = 1$ and $W_1 = d$). Following the proof of Theoren 3.1, we focus on the quantity:

$$\zeta_t = \frac{1}{\eta_t} \log\left(\mathbb{E}\exp(-\eta_t \ell(e_I, z_t)\right) \text{ where } \mathbb{P}(I = i) = \frac{w_t(i)}{W_t}.$$

We already proved (thanks to Hoeffding's inequality and the fact that the loss is convex) that

$$\zeta_t \leq -\ell(p_t, z_t) + \frac{\eta_t}{8}.$$

On the other other hand, by defining $L_{i,t} = \sum_{s=1}^t \ell(e_i, z_t)$ and the function

$$\Phi_t(\eta) = \frac{1}{\eta} \log\left(\frac{1}{d}\sum_{i=1}^d \exp(-\eta L_{i,t})\right),$$

one also obtains

$$\zeta_t = \Phi_t(\eta_t) - \Phi_{t-1}(\eta_t).$$

A simple Abel's transform yields

$$\sum_{t=1}^n \left(\Phi_t(\eta_t) - \Phi_{t-1}(\eta_t)\right) = \Phi_n(\eta_n) - \Phi_0(\eta_1) + \sum_{t=1}^{n-1} \left(\Phi_t(\eta_t) - \Phi_t(\eta_{t+1})\right).$$

First note that $\Phi_0(\eta_1) = 0$, while (following the beginning of the proof of Theorem 3.1)

$$\Phi_n(\eta_n) \geq -\frac{\log d}{\eta_n} - \min_{1 \leq i \leq d} L_{i,t}.$$

Thus, up to straightforward computations such as

$$\sum_{t=1}^n \frac{1}{\sqrt{t}} \leq \int_0^n \frac{1}{\sqrt{t}} dt = 2\sqrt{n},$$

it suffices to show that $\Phi_t(\eta_t) - \Phi_t(\eta_{t+1}) \geq 0$, which is implied by $\Phi_t'(\eta) \geq 0$. We now prove the latter inequality:

$$\begin{aligned}
\Phi_t'(\eta) &= -\frac{1}{\eta^2} \log\left(\frac{1}{d}\sum_{i=1}^d \exp\left(-\eta L_{i,t}\right)\right) - \frac{1}{\eta} \frac{\sum_{i=1}^d L_{i,t} \exp\left(-\eta L_{i,t}\right)}{\sum_{i=1}^d \exp\left(-\eta L_{i,t}\right)} \\
&= \frac{1}{\eta^2} \frac{1}{\sum_{i=1}^d \exp\left(-\eta L_{i,t}\right)} \sum_{i=1}^d \exp\left(-\eta L_{i,t}\right) \times \left(-\eta L_{i,t} - \log\left(\frac{1}{d}\sum_{j=1}^d \exp\left(-\eta L_{j,t}\right)\right)\right) \\
&= \frac{1}{\eta^2} \sum_{i=1}^d p_{t+1}^\eta(i) \log\left(\frac{p_{t+1}^\eta(i)}{1/d}\right) \text{ where } p_{t+1}^\eta(i) = \frac{\exp(-\eta L_{i,t})}{\sum_{j=1}^d \exp(-\eta L_{j,t})}. \\
&= \frac{1}{\eta^2} \mathrm{KL}(p_{t+1}^\eta, \pi) \text{ where } \pi \text{ is the uniform distribution over } \{1, \ldots, d\}. \\
&\geq 0.
\end{aligned}$$

□

2.6. Subdifferentiable loss with bounded subgradient

We show now how to extend the results for the expert regret R_n^E to the regret R_n. We consider here a loss which is subdifferentiable on the simplex. Recall that $f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ is subdifferentiable if $\forall x \in \mathcal{X}$, there exists a subgradient $\nabla f(x) \in \mathbb{R}^d$ such that

$$f(x) - f(x_0) \leq \nabla f(x)^T (x - x_0).$$

In particular a differentiable convex function is subdifferentiable (and $\nabla f(x)$ corresponds to the usual gradient), and a convex function on an open set is also subdifferentiable.

Given a strategy which satisfies a regret bound for linear losses, it is easy to modify it to obtain a regret bound for subdifferentiable losses. Indeed, it suffices to run the strategy on the loss $\bar{\ell}(p,(q,z)) = \nabla \ell(q,z)^T q$ rather than on $\ell(p,z)$ (note that we artificially enlarged the adversary's move set to $\bar{\mathcal{Z}} = \Delta_d \times \mathcal{Z}$ and we consider that he plays (p_t, z_t) at time t), since we have the inequality:

$$\begin{aligned} \sum_{t=1}^n \left(\ell(p_t,z_t) - \ell(q,z_t)\right) &\leq \sum_{t=1}^n \nabla \ell(p_t,z_t)^T (p_t - q) \\ &= \sum_{t=1}^n \left(\bar{\ell}(p_t,(p_t,z_t)) - \bar{\ell}(q,(p_t,z_t))\right). \end{aligned} \tag{2.3}$$

In particular we call *subgradient-based Exp* the following strategy:

$$p_t = \sum_{i=1}^d \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \nabla \ell(p_s,z_s)^T e_i\right)}{\sum_{j=1}^d \exp\left(-\eta \sum_{s=1}^{t-1} \nabla \ell(p_s,z_s)^T e_j\right)} e_i.$$

The following theorem almost directly follows from (2.3) and Theorem 3.1.

THEOREM 2.5. *For any subdifferentiable loss with bounded subgradient* $||\nabla \ell(p,z)||_\infty \leq 1, \forall (p,z) \in \Delta_d \times \mathcal{Z}$, *the subgradient-based Exp strategy with parameter* $\eta = \sqrt{\frac{2 \log d}{n}}$ *satisfies:*

$$R_n \leq \sqrt{2n \log d}.$$

PROOF. First note that by Hölder's inequality $|\nabla \ell(q,z)^T p| \leq ||\nabla \ell(q,z)||_\infty ||p||_1 \leq 1$ for $(p,q,z) \in \Delta_d \times \Delta_d \times \mathcal{Z}$. Thus the modified loss $\bar{\ell}$ takes value in $[-1,1]$. But it easy to see that the regret bound of Theorem 3.1 becomes

$$\frac{\log d}{\eta} + \frac{n\eta}{2},$$

for losses with values in $[-1,1]$, which ends the proof thanks to (2.3). □

2.7. Online finite optimization

We consider here the online optimization problem over a finite action set $\mathcal{A} = \{1, \dots, d\}$. In this case the convexity assumptions of the previous sections do not make sense. Moreover it is clear that if the loss can take arbitrary large values, then no interesting regret bound can be derived. Thus in this section we focus on bounded losses, $\ell : \mathcal{A} \times \mathcal{Z} \to [0,1]$, with no further restriction.

First we observe that this problem is intractable (in the sense that no sublinear regret bound can be obtained) for the type of strategies considered so far (that were all deterministic). More precisely consider the case where $\mathcal{A} = \mathcal{Z} = \{0, 1\}$ with the zero-one loss $\ell(a, z) = \mathbb{1}_{a \neq z}$. Then if a_t is a deterministic function of $(z_1, \dots, z_{t-1})$, the adversary can set $z_t = 1 - a_t$ and thus $\ell(a_t, z_t) = 1$. On the other hand clearly we have $\min_{a \in \{0,1\}} \sum_{t=1}^n \ell(a, z_t) \leq \frac{n}{2}$. In other words for any deterministic strategy, there exists a sequence $z_1, \dots, z_n$ such that $R_n \geq \frac{n}{2}$.

The key to get around this impossibility is to add randomization in our decision to *surprise* the adversary. More precisely at every time step the player chooses a distribution $p_t \in \Delta_d$, based on the past (in particular the adversary could have access to p_t), and then draws his decision a_t at random from p_t. Note that the regret R_n is still well defined, but it is now a random variable. Thus one can prove an upper bounds on R_n which holds either with high probability or in expectation.

2.7.1. Linearization of the game. As we just saw, in online finite optimization the player chooses a point $p_t \in \Delta_d$. In other words, one could hope that online finite optimization boils down to online optimization over the simplex. This is indeed the case, and a general bounded loss in the finite case is equivalent to a bounded linear loss on the simplex. More precisey we consider the loss

$$\bar{\ell}(p, z) = \sum_{a=1}^{d} p(a) \ell(a, z).$$

Note that this loss is linear and takes values in $[0, 1]$. We show now that a regret bound with respect to this modified loss entails a regret bound for the original game. The proof is a direct application of the following result from probability theory, known as Hoeffding-Azuma's inequality for martingales.

Theorem 2.6. *Let $\mathcal{F}_1 \subset \dots \subset \mathcal{F}_n$ be a filtration, and $X_1, \dots, X_n$ real random variables such that X_t is $\mathcal{F}_t$-measurable, $\mathbb{E}(X_t | \mathcal{F}_{t-1}) = 0$ and $X_t \in [A_t, A_t + c_t]$ where A_t is a random variable $\mathcal{F}_{t-1}$-measurable and c_t is a positive constant. Then, for any $\varepsilon > 0$, we have*

$$\mathbb{P}\Big(\sum_{t=1}^{n} X_t \geq \varepsilon \Big) \leq \exp\left(- \frac{2\varepsilon^2}{\sum_{t=1}^n c_t^2} \right), \tag{2.4}$$

or equivalently for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\sum_{t=1}^{n} X_t \leq \sqrt{\frac{\log(\delta^{-1})}{2} \sum_{t=1}^{n} c_t^2}. \tag{2.5}$$

Lemma 2.3. *With probability at least $1 - \delta$ the following holds true:*

$$\sum_{t=1}^{n} \ell(a_t, z_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{n} \ell(a, z_t) \leq \sum_{t=1}^{n} \bar{\ell}(p_t, z_t) - \min_{q \in \Delta_d} \sum_{t=1}^{n} \bar{\ell}(q, z_t) + \sqrt{\frac{n \log \delta^{-1}}{2}}.$$

Proof. Apply Theorem 2.6 to the filtration $\mathcal{F}_t = \sigma(a_1, \dots, a_t)$ and to the random variables $X_t = \ell(a_t, z_t) - \sum_{a=1}^{d} p_t(a) \ell(a, z_t)$. □

2.7.2. Finite Exp. Given the preceding section, the natural strategy for online finite optimization is to apply the Exp strategy of Section 2.1 to the modified loss $\bar{\ell}(p,z)$. But recall that this strategy uses only the loss of the vertices in the simplex, which in this case corresponds exactly to the values $\ell(a,z), a \in \{1,\dots,d\}$. Thus we obtain the following strategy, $\forall a \in \{1,\dots,d\}$,

$$p_t(a) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell(a,z_s)\right)}{\sum_{i=1}^{d} \exp\left(-\eta \sum_{s=1}^{t-1} \ell(i,z_s)\right)}.$$

The following theorem directly follows from Theorem 3.1 and Lemma 2.3.

THEOREM 2.7. *For any loss with values in* $[0,1]$*, the finite Exp strategy with parameter* $\eta = 2\sqrt{2\frac{\log d}{n}}$ *satisfies with probability at least* $1-\delta$*:*

$$R_n \leq \sqrt{\frac{n \log d}{2}} + \sqrt{\frac{n \log \delta^{-1}}{2}}.$$

Note that Theorem 2.3 also implies that this bound is unimprovable.

References

The Exp strategy was introduced independently by:

- V. Vovk. Aggregating strategies. In *Proceedings of the third annual workshop on Computational learning theory (COLT)*, pages 371–386, 1990
- N. Littlestone and M. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994

This chapter also draws heavily on Chapter 2 and Chapter 4 (and the references therein) of:

- N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, 2006

CHAPTER 3

Continuous exponential weights

The Exp strategy proposed in the previous chapter relies heavily on the fact that in the simplex, it is sufficient to compete with the vertices, which are in finite number. This is clearly true for linear losses, and we were able to easily extend this reasoning to subdifferentiable losses with bounded subgradient. In this chapter we consider a more general scenario, where it is not enough to look at a finite number of points. This happens in two cases. Either the action set is not a polytope (e.g., the Euclidean ball), in which case even for linear losses one has to compare to an infinite number of points. Or the loss is convex but does not admit a bounded subgradient (e.g., the log-loss) in which case one cannot reduce the problem to linear losses.

We shall attack both issues simultaneously by considering the Continuous Exp strategy defined as follows. The idea is very simple, following Section 2.1, one defines a weight $w_t(a)$ for each point $a \in \mathcal{A}$, and compute the corresponding weighted average a_t. More precisely we assume that $\mathcal{A}$ is a convex subset of $\mathbb{R}^d$ and:

$$a_t = \int_{a \in \mathcal{A}} \frac{w_t(a)}{W_t} \; a \; da,$$

where

$$w_t(a) = \exp\left(-\eta \sum_{s=1}^{t-1} \ell(a, z_s)\right), \; W_t = \int_{a \in \mathcal{A}} w_t(a) \; da,$$

and $\eta > 0$ is a fixed parameter. Clearly computing a_t is much more difficult than computing the point given by the standard Exp strategy. However, quite miraculously, there exists efficient methods to compute the Continuous Exp strategy in most cases. Indeed, since the loss is convex, we have that $a \mapsto \frac{w_t(a)}{W_t}$ is a log-concave function. Then using random walks methods from [40], one can compute a_t up to a precision ε in $O\left(\text{poly}(d) \text{ polylog}(1/\varepsilon)\right)$ steps (see [40] for the details). Clearly with precision $\varepsilon = \text{poly}(1/n)$ the regret bounds for Continuous Exp and its approximate version will be similar, up to a constant. However note that the method described in [40] works only if $\mathcal{A}$ is a convex body (i.e., $\mathcal{A}$ is of full rank). When this assumption is not satisfied some care is needed, see [32] for the important example of the simplex.

In the following we prove a regret bound for Continuous Exp when the loss is convex and bounded, and then we propose an improved bound for the important case of exp-concave losses (which includes the log-loss).

3.1. Bounded convex loss

Similarly to Section 2.2, without loss of generality we assume here that $\ell(a, z) \in [0, 1], \forall (a, z) \in \mathcal{A} \times \mathcal{Z}$.

THEOREM 3.1. *For any convex loss taking values in* $[0, 1]$*, the Continuous Exp strategy satisfies* $\forall \gamma > 0$*:*

$$R_n \leq \frac{d \log \frac{1}{\gamma}}{\eta} + \frac{n\eta}{8} + \gamma n.$$

In particular with $\gamma = 1/n$*, and* $\eta = 2\sqrt{\frac{2d \log n}{n}}$*, it satisfies:*

$$R_n \leq 1 + \sqrt{\frac{dn \log n}{2}}.$$

PROOF. Let $\gamma > 0$, $a^* \in \operatorname{argmin}_{a \in \mathcal{A}} \sum_{t=1}^n \ell(a, z_t)$, $\mathcal{N}_\gamma = \{(1-\gamma)a^* + \gamma a, a \in \mathcal{A}\}$. Then

$$\begin{aligned}
\log \frac{W_{n+1}}{W_1} &= \log \left(\frac{\int_{a \in \mathcal{A}} w_{n+1}(a) \ da}{\int_{a \in \mathcal{A}} 1 \ da} \right) \\
&\geq \log \left(\frac{\int_{a \in \mathcal{N}_\gamma} w_{n+1}(a) \ da}{\int_{a \in \mathcal{A}} 1 \ da} \right) \\
&= \log \left(\frac{\int_{a \in \mathcal{N}_\gamma} \exp\left(-\eta \sum_{t=1}^n \ell(a, z_t)\right) \ da}{\int_{a \in \mathcal{A}} 1 \ da} \right) \\
&= \log \left(\frac{\int_{a \in \gamma\mathcal{A}} \exp\left(-\eta \sum_{t=1}^n \ell((1-\gamma)a^* + a, z_t)\right) \ da}{\int_{a \in \mathcal{A}} 1 \ da} \right) \\
&= \log \left(\frac{\int_{a \in \mathcal{A}} \exp\left(-\eta \sum_{t=1}^n \ell((1-\gamma)a^* + \gamma a, z_t)\right) \ \gamma^d da}{\int_{a \in \mathcal{A}} 1 \ da} \right) \\
&\geq \log \left(\frac{\int_{a \in \mathcal{N}_\gamma} \exp\left(-\eta \sum_{t=1}^n \left((1-\gamma)\ell(a^*, z_t) + \gamma \ell(a, z_t)\right)\right) \ \gamma^d da}{\int_{a \in \mathcal{A}} 1 \ da} \right) \\
&\geq \log \left(\frac{\int_{a \in \mathcal{N}_\gamma} \exp\left(-\eta \sum_{t=1}^n \left(\ell(a^*, z_t) + \gamma\right)\right) \ \gamma^d da}{\int_{a \in \mathcal{A}} 1 \ da} \right) \\
&= d \log \gamma - \eta \sum_{t=1}^n \ell(a^*, z_t) - \eta\gamma n.
\end{aligned}$$

On the other hand, we have $\log \frac{W_{n+1}}{W_1} = \sum_{t=1}^{n} \log \frac{W_{t+1}}{W_t}$ and

$$\begin{aligned}
\log \frac{W_{t+1}}{W_t} &= \log\left(\int_{\mathcal{A}} \frac{w_t(a)}{W_t} \exp(-\eta\ell(a,z_t))da\right) \\
&= \log\left(\mathbb{E}\exp(-\eta\ell(A,z_t)\right) \text{ where } \mathbb{P}(A=a) = \frac{w_t(a)}{W_t} \\
&\leq -\eta\mathbb{E}\ell(A,z_t) + \frac{\eta^2}{8} \text{ (Hoeffding's lemma)} \\
&\leq -\eta\ell(\mathbb{E}A,z_t) + \frac{\eta^2}{8} \text{ (Jensen's inequality)} \\
&= -\eta\ell(a_t,z_t) + \frac{\eta^2}{8}.
\end{aligned}$$

Thus we proved:

$$\sum_{t=1}^{n}\left(-\eta\ell(a_t,z_t) + \frac{\eta^2}{8}\right) \geq -\eta\sum_{t=1}^{n}\ell(a^*,z_t) - d\log\gamma - \eta\gamma n.$$

In other words:

$$R_n \leq \frac{d\log\frac{1}{\gamma}}{\eta} + \frac{n\eta}{8} + \gamma n,$$

which concludes the proof. □

3.2. Exp-concave loss

We study here the behavior of Continuous Exp with exp-concave losses.

Theorem 3.2. *For any σ-exp-concave loss, the Continuous Exp strategy with parameter $\eta = \sigma$ satisfies:*

$$R_n \leq 1 + \frac{d\log n}{\sigma}.$$

Proof. In the previous proof it suffices to use $\gamma = 1/n$ and to replace Hoeffding's lemma followed by Jensen's inequality by a single Jensen's inequality applied to $a \mapsto \exp(-\eta\ell(a,z))$. □

References

Originally Continuous Exp was derived for the sequential investment problem (i.e., the log-loss on the simplex) where it was called Universal Portfolio, see:

- T. Cover. Universal portfolios. *Math. Finance*, 1:1–29, 1991
- A. Kalai and S. Vempala. Efficient algorithms for universal portfolios. *Journal of Machine Learning Research*, 3:423–440, 2002

The presentation of this chapter is inspired by

- E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69:169–192, 2007

See also [42] for recent improvements regarding the computational complexity of Continuous Exp, and [18] for Continuous Exp is the context of aggregation under the squared loss.

CHAPTER 4

Online gradient methods

The Continuous Exp strategy described in the previous chapter gives an algorithm to deal with the general online convex optimization problem. However, from an algorithmic point of view the strategy is fairly involved (even if polynomial time implementations are possible). Moreover, while the dependency on n in the regret bound is optimal (given the lower bound of Section 2.4), it is not clear if the $\sqrt{d}$ factor is necessary. In this chapter we propose to take a radically different approach, which shall prove to be much simpler on the algorithmic side and also give better regret bounds (under stronger conditions however).

Consider for instance the case of online convex optimization (with a subdifferentiable loss) on the Euclidean ball:

$$B_{2,d} = \{x \in \mathbb{R}^d : ||x||_2 \leq 1\}.$$

Since, after all, we are trying to optimize a convex function on a continuous set, why not try a simple gradient descent. In other words, set

$$w_{t+1} = a_t - \eta \nabla \ell(a_t, z_t),$$

where η is the stepsize parameter of the gradient descent. The resulting point w_{t+1} might end up outside of the ball $B_{2,d}$, in which case one would need to normalize and set $a_{t+1} = \frac{w_{t+1}}{||w_{t+1}||_2}$. Otherwise one can simply take $a_{t+1} = w_{t+1}$.

4.1. Online Gradient Descent (OGD)

The approach described above is very general and can be applied to any closed convex set $\mathcal{A}$ and subdifferentiable loss ℓ. The resulting strategy is called Online Gradient Descent and can be described as follows: start at a point $a_1 \in \mathcal{A}$, then for $t \geq 1$,

$$w_{t+1} = a_t - \eta \nabla \ell(a_t, z_t), \tag{4.1}$$

$$a_{t+1} = \operatorname*{argmin}_{a \in \mathcal{A}} ||w_{t+1} - a||_2. \tag{4.2}$$

Note that in (4.1) it is enough to supply OGD with a subgradient.

Theorem 4.1. *For any closed convex action set $\mathcal{A}$ such that $||a||_2 \leq R, \forall a \in \mathcal{A}$, for any subdifferentiable loss with bounded subgradient $||\nabla \ell(a,z)||_2 \leq G, \forall (a,z) \in \mathcal{A} \times \mathcal{Z}$, the OGD strategy with parameters $\eta = \frac{R}{G\sqrt{n}}$ and $a_1 = 0$ satisfies:*

$$R_n \leq RG\sqrt{n}.$$

Proof. Let $g_t = \nabla \ell(a_t, z_t)$ and $a \in \mathcal{A}$. By definition of a subdifferentiable loss, we have:

$$\sum_{t=1}^{n} \Big(\ell(a_t, z_t) - \ell(a, z_t) \Big) \leq \sum_{t=1}^{n} g_t^T (a_t - a).$$

Moreover since $w_{t+1} = a_t - \eta g_t$, we have:

$$\begin{aligned} 2\eta g_t^T (a_t - a) &= 2(a_t - w_{t+1})^T (a_t - a) \\ &= ||a - a_t||_2^2 + ||a_t - w_{t+1}||_2^2 - ||a - w_{t+1}||_2^2 \\ &= \eta^2 ||g_t||_2^2 + ||a - a_t||_2^2 - ||a - w_{t+1}||_2^2. \end{aligned}$$

Now note that, by definition of a_{t+1} and since $\mathcal{A}$ is a convex set, one has:

$$||a - w_{t+1}||_2^2 \geq ||a - a_{t+1}||_2^2.$$

Thus by summing one directly obtains:

$$\begin{aligned} 2\eta \sum_{t=1}^{n} g_t^T (a_t - a) &\leq ||a - a_1||_2^2 + \eta^2 \sum_{t=1}^{n} ||g_t||_2^2 \\ &\leq R^2 + \eta^2 G^2 n, \end{aligned}$$

which ends the proof up to straightforward computations. □

4.2. Strongly convex loss with bounded subgradient

In this section we show how to improve the regret bound for strongly-convex losses. The key is to use a time-varying parameter η_t, like we did in [Section 2.5, Chapter 2].

Theorem 4.2. *For any closed convex action set $\mathcal{A} \subset B_{2,d}$, for any α-strongly loss with bounded subgradient $||\nabla \ell(a, z)||_2 \leq 1, \forall (a, z) \in \mathcal{A} \times \mathcal{Z}$, the OGD strategy with time-varying stepsize $\eta_t = \frac{1}{\alpha t}$ and $a_1 = 0$ satisfies:*

$$R_n \leq \frac{\log(en)}{2\alpha}.$$

Proof. Let $g_t = \nabla \ell(a_t, z_t)$ and $a \in \mathcal{A}$. By definition of a α-strongly convex loss, we have:

$$\sum_{t=1}^{n} \Big(\ell(a_t, z_t) - \ell(a, z_t) \Big) \leq \sum_{t=1}^{n} \Big(g_t^T (a_t - a) - \frac{\alpha}{2} ||a_t - a||_2^2 \Big).$$

Now following the exact same argument than in the proof of Theorem 4.1, one can prove:

$$g_t^T (a_t - a) \leq \frac{\eta_t}{2} + \frac{||a - a_t||_2^2 - ||a - a_{t+1}||_2^2}{2\eta_t}.$$

By summing these inequalities, and denoting $\zeta_t = ||a - a_t||_2^2$, one directly obtains (with a simple Abel's transform):

$$\begin{aligned} \sum_{t=1}^{n} \Big(\ell(a_t, z_t) - \ell(a, z_t) \Big) &\leq \frac{1}{2} \sum_{t=1}^{n} \eta_t + \frac{1}{2} \sum_{t=1}^{n} \left(\frac{\zeta_t - \zeta_{t+1}}{\eta_t} - \alpha \zeta_t \right) \\ &\leq \frac{\log(en)}{2\alpha} + \frac{1}{2} \left(\frac{1}{\eta_1} - \alpha \right) \zeta_1 + \frac{1}{2} \sum_{t=2}^{n} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \alpha \right) \zeta_t \\ &= \frac{\log(en)}{2\alpha}, \end{aligned}$$

which ends the proof. □

References

The OGD strategy was introduced by:

- M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, 2003

The analysis in the case of strongly convex losses was done in:

- E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69:169–192, 2007

Note that gradient descent ideas and analysis dates back to:

- B. Polyak. A general method for solving extremal problems. *Soviet Math. Doklady*, 174:33–36, 1967
- N. Shor. Generalized gradient descent with application to block programming. *Kibernetika*, 3:53–55, 1967. (In Russian)

CHAPTER 5

Online mirror descent

In this chapter we greatly generalize the online gradient method of the previous chapter. This generalization adds a lot of flexibility to the method, which in turns allows it to adapt to the geometry of the problem. To describe the idea, we first need a few notions from convex analysis.

5.1. Convex analysis

In this section we consider an open convex set $\mathcal{D} \subset \mathbb{R}^d$. We denote by $\overline{\mathcal{D}}$ the closure of $\mathcal{D}$. Let $||\cdot||$ be a norm on $\mathbb{R}^d$, and $||\cdot||_*$ the dual norm. Recall that

$$||u||_* = \sup_{x \in \mathbb{R}^d : ||x|| \leq 1} x^T u.$$

In particular, by definition, the Hölder's inequality holds: $|x^T u| \leq ||x||.||u||_*$.

Definition 5.1. *Let $f : \overline{\mathcal{D}} \to \mathbb{R}$ be a convex function. The Legendre-Fenchel transform of f is defined by:*

$$f^*(u) = \sup_{x \in \overline{\mathcal{D}}} \left(x^T u - f(x) \right).$$

The next proposition shows that the dual of a norm is related to the Legendre-Fenchel transform of that norm.

Proposition 5.1. *Let $f(x) = \frac{1}{2}||x||^2$. Then $f^*(u) = \frac{1}{2}||u||_*^2$.*

Proof. By Hölder's inequality, and a simple optimization of a quadratic polynomial, one has:

$$f^*(u) = \sup_{x \in \mathcal{D}} \left(x^T u - \frac{1}{2}||x||^2 \right) \leq \sup_{x \in \mathcal{D}} \left(||x||.||u||_* - \frac{1}{2}||x||^2 \right) = \frac{1}{2}||u||_*^2.$$

Moreover the inequality above is in fact an equality, by definition of the dual norm. □

Definition 5.2. *We call Legendre any continuous function $F : \overline{\mathcal{D}} \to \mathbb{R}$ such that*

(i) *F is strictly convex and admits continuous first partial derivatives on $\mathcal{D}$,*
(ii) $\lim_{x \to \overline{\mathcal{D}} \setminus \mathcal{D}} ||\nabla F(x)|| = +\infty$.[1]

The Bregman divergence $D_F : \overline{\mathcal{D}} \times \mathcal{D}$ associated to a Legendre function F is defined by

$$D_F(x, y) = F(x) - F(y) - (x - y)^T \nabla F(y).$$

Moreover we say that $\mathcal{D}^* = \nabla F(\mathcal{D})$ is the dual space of $\mathcal{D}$ under F.

[1] By the equivalence of norms in $\mathbb{R}^d$, this definition does not depend on the choice of the norm.

Note that, by definition, $D_F(x,y) > 0$ if $x \neq y$, and $D_F(x,x) = 0$.

LEMMA 5.1. *Let F be a Legendre function. Then $F^{**} = F$, and $\nabla F^* = (\nabla F)^{-1}$ (on the set $\mathcal{D}^*$). Moreover, $\forall x, y \in \mathcal{D}$,*

$$D_F(x,y) = D_{F^*}(\nabla F(y), \nabla F(x)). \tag{5.1}$$

The above lemma is the key to understand how a Legendre function act on the space. The mapping ∇F maps $\mathcal{D}$ to the dual space $\mathcal{D}^*$, and ∇F^* is the inverse mapping to go from the dual space to the original (primal) space. Moreover (5.1) shows that the Bregman divergence in the primal corresponds exactly to the Bregman divergence of the Legendre-transform in the dual. We examine now these phenomenons on a few examples.

EXAMPLE 5.1. *Let $F(x) = \frac{1}{2}||x||_2^2$ with $\mathcal{D} = \mathbb{R}^d$. F is Legendre, and it is easy to see that $D_F(x,y) = \frac{1}{2}||x-y||_2^2$. Moreover Proposition 5.1 shows that $F^* = F$. Thus here the primal and dual spaces are the same.*

EXAMPLE 5.2. *Let $F(x) = \sum_{i=1}^d x_i \log x_i - \sum_{i=1}^d x_i$ (generalized negative entropy) with $\mathcal{D} = (0, +\infty)^d$. F is Legendre, and it is easy to show that:*

$$\begin{aligned}
&\nabla F(x) = (\log x_1, \dots, \log x_d),\\
&D_F(x,y) = \sum_{i=1}^d x_i \log \frac{x_i}{y_i} - \sum_{i=1}^d (x_i - y_i) \ \textit{(generalized Kullback-Leibler divergence)},\\
&F^*(u) = \sum_{i=1}^d \exp(u_i),\\
&\nabla F^*(u) = (\exp(u_1), \dots, \exp(u_d)),\\
&D_{F^*}(u,v) = \sum_{i=1}^d \exp(v_i) \left(\exp(u_i - v_i) - 1 - (u_i - v_i)\right).
\end{aligned}$$

Here the primal space is $(0, +\infty)^d$, and the dual is $\mathbb{R}^d$.

EXAMPLE 5.3. *Let $F(x) = -2\sum_{i=1}^d \sqrt{x_i}$ with $\mathcal{D} = (0, +\infty)^d$. F is Legendre, and it is easy to show that:*

$$\begin{aligned}
\nabla F(x) &= -\left(\frac{1}{\sqrt{x_1}}, \dots, \frac{1}{\sqrt{x_d}}\right),\\
D_F(x,y) &= \sum_{i=1}^d \frac{(\sqrt{x_i} - \sqrt{y_i})^2}{\sqrt{y_i}},\\
F^*(u) &= \begin{cases} +\infty \ \textit{if}\ u \in \mathbb{R}^d \setminus (-\infty, 0)^d \\ -\sum_{i=1}^d \frac{1}{u_i} \ \textit{if}\ u \in (-\infty, 0)^d \end{cases}\\
\nabla F^*(u) &= \left(\frac{1}{u_1^2}, \dots, \frac{1}{u_d^2}\right) \ \textit{if}\ u \in (-\infty, 0)^d,\\
D_{F^*}(u,v) &= \sum_{i=1}^d v_i \left(\frac{1}{u_i} - \frac{1}{v_i}\right)^2 \ \textit{if}\ u, v \in (-\infty, 0)^d.
\end{aligned}$$

Here the primal space is $(0, +\infty)^d$, and the dual is $(-\infty, 0)^d$.

We now state a few property that will be useful when working with Bregman divergences. The first lemma is a trivial equality, but it is nonetheless useful to state it as we will use this result several times

LEMMA 5.2. *Let F be a Legendre function. Then $\forall (x,y,z) \in \overline{\mathcal{D}} \times \mathcal{D} \times \mathcal{D}$,*

$$D_F(x,y) + D_F(y,z) = D_F(x,z) + (x-y)^T(\nabla F(z) - \nabla F(y)).$$

The next lemma shows that the geometry induced by a Bregman divergence ressembles to the geometry of the squared euclidean distance.

LEMMA 5.3. *Let $\mathcal{A} \subset \overline{\mathcal{D}}$ be a closed convex set such that $\mathcal{A} \cap \mathcal{D} \neq \emptyset$. Then, $\forall x \in \mathcal{D}$,*

$$b = \operatorname*{argmin}_{a \in \mathcal{A}} D_F(a,x),$$

exists and is unique. Moreover $b \in \mathcal{A} \cap \mathcal{D}$, and $\forall\ a \in \mathcal{A}$,

$$D_F(a,x) \geq D_F(a,b) + D_F(b,x).$$

5.2. Online Mirror Descent (OMD)

The idea of OMD is very simple: first select a Legendre function F on $\overline{\mathcal{D}} \supset \mathcal{A}$ (such that $\mathcal{A} \cap \mathcal{D} \neq \emptyset$). Then perform an online gradient descent, where the update with the gradient is performed in the dual space $\mathcal{D}^*$ rather than in the primal $\mathcal{D}$, and where the projection step is defined by the Bregman divergence associated to F. More precisely the algorithm works as follows for a closed convex set $\mathcal{A}$ and a subdifferentiable loss ℓ: start at $a_1 \in \operatorname{argmin}_{a \in \mathcal{A}} F(a)$ (note that $a_1 \in \mathcal{A} \cap \mathcal{D}$), then for $t \geq 1$,

$$w_{t+1} = \nabla F^* \Big(\nabla F(a_t) - \eta \nabla \ell(a_t, z_t) \Big), \tag{5.2}$$

$$a_{t+1} = \operatorname*{argmin}_{a \in \mathcal{A}} D_F(a, w_{t+1}). \tag{5.3}$$

Note that (5.2) is well defined if the following consistency condition is satisfied:

$$\nabla F(a) - \eta \nabla \ell(a,z) \in \mathcal{D}^*, \forall (a,z) \in \mathcal{A} \cap \mathcal{D} \times \mathcal{Z}. \tag{5.4}$$

Note also that (5.2) can be rewritten as:

$$\nabla F(w_{t+1}) = \nabla F(a_t) - \eta \nabla \ell(a_t, z_t). \tag{5.5}$$

The projection step (5.3) is a convex program in the sense that $x \mapsto D_F(x,y)$ is always a convex function. This does not necessarily implies that (5.3) can be performed efficiently, since in some cases the feasible set $\mathcal{A}$ might only be described with an exponential number of constraints (we shall encounter examples like this in the next chapter).

In some cases it is possible to derive a closed form expression for the Bregman projection (5.3) using the following lemma.

LEMMA 5.4. *Let f be a convex and differentiable function on $\mathcal{X}$. Then $f(x) \leq f(y), \forall y \in \mathcal{X}$ if and only if $\nabla f(x)^T(y-x) \geq 0, \forall y \in \mathcal{X}$.*

PROOF. One direction is straightforward using the fact that a convex and differentiable function is subdifferentiable. For the other direction it suffices to note that if $\nabla f(x)^T(y-x) < 0$, then f is locally decreasing around x on the line to y (simply consider $h(t) = f(x + t(y-x))$ and note that $h'(0) < 0$). □

We detail now a few instances of OMD.

Example 5.4. *OMD with $F(x) = \frac{1}{2}||x||_2^2$ (defined on $\mathcal{D} = \mathbb{R}^d$) corresponds exactly to OGD. Note that here OMD is always well defined (in other words* (5.4) *is always satisfied) since $\mathcal{D}^* = \mathbb{R}^d$.*

Example 5.5. *Consider $\mathcal{A} = \Delta_d$, and $F(x) = \sum_{i=1}^d x_i \log x_i - \sum_{i=1}^d x_i$ (defined on $\mathcal{D} = (0, +\infty)^d$). Note that here OMD is always well defined (in other words* (5.4) *is always satisfied) since $\mathcal{D}^* = \mathbb{R}^d$. Using the computations that we did previously for this Legendre function, one can see that* (5.2) *rewrites:*

$$w_{t+1}(i) = a_t(i) \exp(-\eta \nabla \ell(a_t, z_t)^T e_i).$$

Moreover the projection step in (5.3) *is equivalent to a normalization. Indeed using Lemma 5.4, one can see that*

$$x = \operatorname*{argmin}_{a \in \Delta_d} \sum_{i=1}^d a_i \log \frac{a_i}{w_i} - \sum_{i=1}^d (a_i - w_i),$$

if and only $x \in \Delta_d$ and

$$\sum_{i=1}^d (y_i - x_i) \log \frac{x_i}{w_i} \geq 0, \forall y \in \Delta_d.$$

This latter condition is clearly satisfied for $x = \frac{w}{||w||_1}$.

Thus we proved that OMD on the simplex with the unnormalized negative entropy is equivalent to subgradient-based Exp.

We prove now a very general and powerful theorem to analyze OMD.

Theorem 5.1. *Let $\mathcal{A}$ be a closed convex action set, ℓ a subdifferentiable loss, and F a Legendre function defined on $\overline{\mathcal{D}} \supset \mathcal{A}$, such that* (5.4) *is satisfied. Then OMD satisfies:*

$$R_n \leq \frac{\sup_{a \in \mathcal{A}} F(a) - F(a_1)}{\eta} + \frac{1}{\eta} \sum_{t=1}^n D_{F^*}\Big(\nabla F(a_t) - \eta \nabla \ell(a_t, z_t), \nabla F(a_t)\Big).$$

Proof. Let $a \in \mathcal{A}$. Since ℓ is subdifferentiable we have:

$$\sum_{t=1}^n (\ell(a_t, z_t) - \ell(a, z_t)) \leq \sum_{t=1}^n \nabla \ell(a_t, z_t)^T (a_t - a).$$

Using (5.5), and applying Lemma 5.2, one obtains

$$\begin{aligned}\eta \nabla \ell(a_t, z_t)^T (a_t - a) &= (a - a_t)^T \big(\nabla F(w_{t+1}) - \nabla F(a_t)\big) \\ &= D_F(a, a_t) + D_F(a_t, w_{t+1}) - D_F(a, w_{t+1}).\end{aligned}$$

By Lemma 5.3, one gets $D_F(a, w_{t+1}) \geq D_F(a, a_{t+1}) + D_F(a_{t+1}, w_{t+1})$, hence

$$\eta \nabla \ell(a_t, z_t)^T (a_t - a) \leq D_F(a, a_t) + D_F(a_t, w_{t+1}) - D_F(a, a_{t+1}) - D_F(a_{t+1}, w_{t+1}).$$

Summing over t then gives

$$\begin{aligned}\sum_{t=1}^n \eta \nabla \ell(a_t, z_t)^T (a_t - a) \quad &\leq \quad D_F(a, a_1) - D_F(a, a_{n+1}) \\ &\quad + \sum_{t=1}^n \big(D_F(a_t, w_{t+1}) - D_F(a_{t+1}, w_{t+1})\big).\end{aligned}$$

By the nonnegativity of the Bregman divergences, we get

$$\sum_{t=1}^{n} \eta \nabla \ell(a_t, z_t)^T (a_t - a) \leq D_F(a, a_1) + \sum_{t=1}^{n} D_F(a_t, w_{t+1}).$$

From (5.1), one has

$$D_F(a_t, w_{t+1}) = D_{F^*}\big(\nabla F(a_t) - \eta \nabla \ell(a_t, z_t), \nabla F(a_t)\big).$$

Moreover since a_1 $\operatorname{argmin}_{a \in \mathcal{A}} F(a)$, Lemma 5.4 directly gives

$$D_F(a, a_1) \leq F(a) - F(a_1), \forall a \in \mathcal{A},$$

which concludes the proof. □

First note that for OGD, this theorem gives back the result of Theorem 4.1. On the other hand for subgradient-based Exp we recover Theorem 2.5, with a slightly worse constant: First note that the negative entropy is always negative, and minimized at the uniform distribution (i.e. $a_1 = (1/d, \dots, 1/d)$). Thus we get

$$\frac{F(a) - F(a_1)}{\eta} \leq \frac{\log d}{\eta}.$$

Moreover note that, using previous computations, one obtains:

$$D_{F^*}\bigg(\nabla F(a_t) - \eta \nabla \ell(a_t, z_t), \nabla F(a_t)\bigg) = \sum_{i=1}^{d} a_t(i) \Theta(-\eta \nabla \ell(a_t, z_t)^T e_i),$$

where $\Theta : x \in \mathbb{R} \mapsto \exp(x) - 1 - x$. One concludes using the following lemma:

Lemma 5.5. *If $x \in \mathbb{R}_-$, then $\Theta(x) \leq \frac{x^2}{2}$. Moreover if $x \leq 1$, then $\Theta(x) \leq x^2$.*

5.3. Subgradient bounded in an arbitrary norm

We already saw that Exp on the simplex has a bound naturally expressed in terms of $||\nabla \ell(a, z)||_\infty$, while OGD requires a bound on $||\nabla \ell(a, z)||_2$. Here we turn the table, and we ask for an algorithm for which the bound would express naturally in terms of some norm $|| \cdot ||$. The next theorem shows that this goal is achieved by computing a Legendre function F strongly convex (on $\mathcal{A}$) with respect to the dual norm.

Theorem 5.2. *For any closed convex action set $\mathcal{A}$, for any subdifferentiable loss with bounded subgradient $||\nabla \ell(a, z)||_* \leq G, \forall (a, z) \in \mathcal{A} \times \mathcal{Z}$, the OMD strategy with a Legendre function F on $\mathcal{D}$ such that $F(a) - F(a_1) \leq R^2, \forall a \in \mathcal{A}$, and its restriction to $\mathcal{A}$ is α-strongly convex with respect to $|| \cdot ||$, and with $\eta = \frac{G}{R}\sqrt{\frac{2}{n}}$ satisfies:*

$$R_n \leq RG\sqrt{\frac{2n}{\alpha}}.$$

Proof. Here we need to go back to the proof of Theorem 5.1, from which we extract that:

$$\begin{aligned} \sum_{t=1}^{n} \eta \nabla \ell(a_t, z_t)^T (a_t - a) \quad &\leq \quad D_F(a, a_1) - D_F(a, a_{n+1}) \\ &\qquad + \sum_{t=1}^{n} \big(D_F(a_t, w_{t+1}) - D_F(a_{t+1}, w_{t+1})\big). \end{aligned}$$

Now remark that, thanks to the strong convexity of F (on $\mathcal{A}$), the definition of OMD and Hölder's inequality:

$$\begin{aligned}
&\sum_{t=1}^{n} \big(D_F(a_t, w_{t+1}) - D_F(a_{t+1}, w_{t+1})\big) \\
&= F(a_t) - F(a_{t+1}) + \nabla F(w_{t+1})^T (a_{t+1} - a_t) \\
&\leq \nabla F(a_t)^T (a_t - a_{t+1}) - \frac{\alpha}{2} ||a_t - a_{t+1}||_*^2 + \nabla F(w_{t+1})^T (a_{t+1} - a_t) \\
&= -\eta \nabla \ell(a_t, z_t)^T (a_t - a_{t+1}) - \frac{\alpha}{2} ||a_t - a_{t+1}||^2 \\
&\leq \eta G ||a_t - a_{t+1}|| - \frac{\alpha}{2} ||a_t - a_{t+1}||^2 \\
&\leq \frac{(\eta G)^2}{2\alpha},
\end{aligned}$$

which concludes the proof. □

We show now a few examples where it is possible to design such a Legendre function.

Lemma 5.6. *Let Q be a symmetric $d \times d$ matrix such that $Q \succ 0$. Consider the norm defined by $||x|| = \sqrt{x^T Q x}$. Then $F(x) = \frac{1}{2} x^T Q x$ is 1-strongly-convex with respect to $|| \cdot ||$.*

Proof. The result follows from a simple application of Proposition 1.1, since here

$$(\nabla F(x) - \nabla F(y))^T (x - y) = \frac{1}{2} (x - y)^T Q (x - y).$$

□

Lemma 5.7. *Let $q \in [1, 2]$. Then $F(x) = \frac{1}{2} ||x||_q^2$ is $(q-1)$-strongly convex with respect to $|| \cdot ||_q$ on $\mathbb{R}^d$.*

Lemma 5.8. *Let $q \in [1, 2]$. Then $F(x) = \frac{1}{2} ||x||_q^2$ is $\frac{q-1}{d^{2\frac{q-1}{q}}}$-strongly convex with respect to $|| \cdot ||_1$ on $\mathbb{R}^d$. In particular with $q = \frac{2 \log d}{2 \log(d) - 1}$ it is $\frac{1}{2e \log d}$-strongly convex with respect to $|| \cdot ||_1$ on $\mathbb{R}^d$.*

Proof. First note that for any p,

$$||x||_p \leq (d ||x||_\infty^p)^{1/p} = d^{1/p} ||x||_\infty.$$

Thus by duality this implies that for q such that $\frac{1}{p} + \frac{1}{q} = 1$, one has:

$$||x||_q \geq \frac{1}{d^{1/p}} ||x||_1 = \frac{1}{d^{\frac{q-1}{q}}} ||x||_1,$$

which clearly concludes the proof thanks to Lemma 5.7. □

In some cases it is important to use the fact that one only need the restriction of F to the action set to be strongly convex. An important example is the case of the negative entropy on the simplex (or rescaled versions of the simplex).

Lemma 5.9. *$F(x) = \sum_{i=1}^d x_i \log x_i - x_i$ (Legendre on $\mathcal{D} = (0, +\infty)^d$) is 1-strongly convex with respect to $|| \cdot ||_1$ on Δ_d. Moreover it is also $\frac{1}{\alpha}$-strongly convex on $\alpha \Delta_d$.*

To prove this lemma we use Pinsker's inequality that we state without a proof.

LEMMA 5.10. *For any $x, y \in \Delta_d$,*

$$\frac{1}{2}||x-y||_1^2 \leq \mathrm{KL}(x,y) = \sum_{i=1}^{d} x_i \log \frac{x_i}{y_i}.$$

We can now prove easily the previous lemma.

PROOF. The result follows on Δ_d a simple application of Proposition 1.1 and Lemma 5.10, since here

$$(\nabla F(x) - \nabla F(y))^T (x-y) = \sum_{i=1}^{d} (x_i - y_i) \log \frac{x_i}{y_i} = \mathrm{KL}(x,y) + \mathrm{KL}(y,x).$$

The generalization to points such that $||x||_1 = \alpha$ is also easy, as in that case:

$$\sum_{i=1}^{d} (x_i - y_i) \log \frac{x_i}{y_i} = \alpha \left(\mathrm{KL}(x/\alpha, y/\alpha) + \mathrm{KL}(y/\alpha, x/\alpha) \right).$$

□

References

The idea of Mirror Descent dates back to:

- A. Nemirovski. Efficient methods for large-scale convex optimization problems. *Ekonomika i Matematicheskie Metody*, 15, 1979. (In Russian)
- A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley Interscience, 1983

A somewhat similar idea was rediscovered in the online learning community by:

- M. Herbster and M. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998
- A. Grove, N. Littlestone, and D. Schuurmans. General convergence results for linear discriminant updates. *Machine Learning*, 43:173–210, 2001
- J. Kivinen and M. Warmuth. Relative loss bounds for multidimensional regression problems. *Machine Learning*, 45:301–329, 2001

Recently these ideas have been flourishing in the online learning community, see for instance (with the references therein):

- S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications.* PhD thesis, The Hebrew University of Jerusalem, 2007

Two short surveys motivated the whole idea of these lecture notes:

- A. Rakhlin. Lecture notes on online learning. 2009
- E. Hazan. The convex optimization approach to regret minimization. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*, pages 287–303. MIT press, 2011

See also the following references for a modern introduction to mirror descent in standard optimization problems:

- A. Juditsky and A. Nemirovski. First-order methods for nonsmooth convex large-scale optimization, i: General purpose methods. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*, pages 121–147. MIT press, 2011

- A. Juditsky and A. Nemirovski. First-order methods for nonsmooth convex large-scale optimization, ii: Utilizing problem's structure. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*, pages 149–183. MIT press, 2011

The convex analysis part of this chapter was inspired by the following references:

- J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2001
- S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004
- Chapter 11 of N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006
- Appendix A of S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem, 2007

CHAPTER 6

Online combinatorial optimization

In this chapter we consider online linear optimization over a subset of $\{0,1\}^d$. As we shall see in the first section, many interesting and challenging problems fall into that framework.

More precisely, one is given a set of concept $\mathcal{C} = \{v_1, \dots, v_N\} \subset \{0,1\}^d$, and we consider online optimization with $\mathcal{A} = \mathcal{C}$, $\mathcal{Z} = [0,1]^d$, and $\ell(a,z) = a^T z$.[1] Just as in [Chapter 2, Section 2.7], by adding randomization this problem is equivalent to online optimization over the convex hull of $\mathcal{C}$ (since the loss is linear):

$$\mathcal{A} = Conv(\mathcal{C}) = \left\{ \sum_{i=1}^{N} p(i) v_i : p \in \Delta_N \right\}.$$

However note that here it is not obvious at all how to perform the randomization step in a computationally efficient way (in this chapter we consider as *efficient* a computational complexity polynomial in the dimension d). More precisely two issues arise:

- The decomposition problem: given $x \in Conv(\mathcal{C})$, find (efficiently) $p \in \Delta_N$ such that $\sum_{i=1}^N p(i) v_i = x$. Note that even writing down p might be impossible since N can be exponential in d. Carathéodory's Theorem below shows that in principle it it possible to avoid this issue and always get a p with at most $d+1$ non-zero coordinates.
- The sampling problem: given $p \in \Delta_N$, draw a point $V \in \mathcal{C}$ at random according to p. Again we may not able to write down p, so we have to assume some kind of oracle access to p (given v one can efficiently compute $p(v)$), or that p can factorize in some sense (we will see some examples below).

One way to solve both issues at the same time is to resort to the proof of Carathéodory's Theorem, which gives an efficient algorithm as soon as $\mathcal{A}$ can be described by a polynomial number of constraints (which shall be the case in several interesting and non-trivial examples). Let us recall that result.

THEOREM 6.1. *For any $x \in Conv(\mathcal{C})$, there exists $p \in \Delta_N$ such that $||p||_0 \leq d+1$ and $\sum_{i=1}^N p(i) v_i = x$.*

PROOF. Assume that $N > d+1$ (otherwise the proof is trivial). Let $x = \sum_{i=1}^k p(i) v_i$, with $k > d+1$ and $p(i) > 0, \forall i \in \{1, \dots, k\}$. We will show that we can define $q \in \Delta_N$ such that $x = \sum_{i=1}^k q(i) v_i$ and $||q||_0 \leq k-1$. This will clearly prove the theorem.

[1] Note that if all concepts have the same size, i.e. $||v||_1 = m, \forall v \in \mathcal{C}$, then one can reduce the case of $\mathcal{Z} = [\alpha, \beta]^d$ to $\mathcal{Z} = [0,1]^d$ via a simple renormalization.

First note that since we are in dimension d and $k > d+1$, there exists $\alpha \in \mathbb{R}^k$ such that

$$\sum_{i=1}^{k} \alpha(i) v_i = 0 \text{ and } \sum_{i=1}^{k} \alpha(i) = 0.$$

Now let

$$q(i) = p(i) - \left(\min_{j : \alpha(j) > 0} \frac{p(j)}{\alpha(j)} \right) \alpha_i .$$

By construction q satisfies the conditions described above. □

Note that in this chapter we set a very challenging problem. Indeed, even the offline optimization problem, i.e.

$$\text{given } z \in [0,1]^d, \text{ find } \operatorname*{argmin}_{v \in \mathcal{C}} v^T z,$$

might require sophisticated algorithms (or can even be intractable). Note in particular that if the convex hull of $\mathcal{C}$ can be described with a polynomial number of constraints, then the offline optimization probem can be solved efficiently by the ellipsoid method (though there might exist faster methods in some cases). As we will see, in that case one can also solve efficiently the online problem. However in some cases there exists efficient algorithms for the offline problem, even if the convex hull can only be described by an exponential number of constraints. We consider an algorithm taking advantage of that situation in the last section of this chapter.

6.1. Examples

We discuss here a few important examples of online combinatorial optimization.

6.1.1. Simplex. The simplest example is when $\mathcal{C} = \{e_1, \dots, e_d\}$. This corresponds to the finite optimization problem described in [Chapter 2, Section 2.7]. Here $Conv(\mathcal{C}) = \Delta_d$.

6.1.2. m-sets. Another important problem is finite optimization where at each round the player has to choose m alternatives out of the d possible choices. This corresponds to the set $\mathcal{C} \subset \{0,1\}^d$ of all vectors with exactly m ones. Note that here $N = \binom{d}{m}$. However the following lemma shows that the convex hull of m-sets is easily described.

LEMMA 6.1. *Let $\mathcal{C}$ be the set of m-sets, then*

$$Conv(\mathcal{C}) = \left\{ x \in [0,1]^d : \sum_{i=1}^{d} x(i) = m \right\}.$$

PROOF. Remark that, for two convex sets $C_1 \subset C_2 \subset \mathbb{R}^d$, if for any $z \in \mathbb{R}^d$, $\max_{x \in C_1} z^T x = \max_{x \in C_2} z^T x$ then $C_1 = C_2$ (otherwise just consider z to be the normal of a hyperplane separating a point $x \in C_2 \setminus C_1$ from C_1). Here we clearly have

$$Conv(\mathcal{C}) \subset \left\{ x \in [0,1]^d : \sum_{i=1}^{d} x(i) = m \right\},$$

and on both sets the maximum of a linear function is realized by putting all the mass on the m largest components of z. □

6.1.3. Bipartite matching. Consider the complete bipartite graph $K_{m,M}$ with $m \leq M$ (that is the set of vertices is composed of one set of size m and another set of size M, and the set of edges consist of all possible links from one set to another). Let $\mathcal{C}$ contain all matchings of size m (that is an injective mapping from $\{1,\dots,m\}$ to $\{1,\dots,M\}$). Here $N = \frac{M!}{(M-m)!}$, $d = m \times M$, and it is convenient to represent points in $\mathcal{C}$ as matrices in $\mathbb{R}^{m \times M}$ rather than vectors. Birkhoff's Theorem shows that the convex hull of matchings on bipartite graph is easily described.

Theorem 6.2. *Let $\mathcal{C}$ be the set of matchings of size m on $K_{m,M}$, then*

$$\begin{aligned} Conv(\mathcal{C}) &= \Bigg\{ x \in [0,1]^{m \times M} : \sum_{j=1}^{M} x(i,j) = 1,\ \forall i \in \{1,\dots,m\}, \\ &\qquad and \sum_{i=1}^{m} x(i,j) \in [0,1],\ \forall j \in \{1,\dots,M\} \Bigg\}. \end{aligned}$$

Proof. Let

$$\begin{aligned} \mathcal{D}_{m,M} &= \Bigg\{ x \in [0,1]^{m \times M} : \sum_{j=1}^{M} x(i,j) = 1,\ \forall i \in \{1,\dots,m\}, \\ &\qquad \text{and} \sum_{i=1}^{m} x(i,j) \in [0,1],\ \forall j \in \{1,\dots,M\} \Bigg\}. \end{aligned}$$

We will first prove that for $m = M$ we have $Conv(\mathcal{C}) = \mathcal{D}_{m,m}$.

Note that clearly the constraint $\sum_{i=1}^{m} x(i,j) \in [0,1]$ can be replaced by $\sum_{i=1}^{m} x(i,j) = 1$ since $m = M$ (just consider $\sum_{i=1}^{m}\sum_{j=1}^{M} x(i,j)$), that is

$$\begin{aligned} \mathcal{D}_{m,m} &= \Bigg\{ x \in [0,1]^{m \times m} : \sum_{j=1}^{m} x(i,j) = 1,\ \forall i \in \{1,\dots,m\}, \\ &\qquad \text{and} \sum_{i=1}^{m} x(i,j) = 1,\ \forall j \in \{1,\dots,M\} \Bigg\}. \end{aligned}$$

Now let x be a vertex (also called an extremal point) of $\mathcal{D}_{m,m}$. Let

$$F = \{(i,j) : x(i,j) \in (0,1)\},$$

be the set of edges in $K_{m,m}$ where the weight given by x is not an integer. We shall prove that $F = \emptyset$, which directly implies the statement of the theorem. First we prove that F contains no circuit (i.e. a closed path that visit every vertex at most once). Indeed suppose that F contains a circuit C. "Clearly" C is the disjoint union of two (partial) matchings of the same size C_1 and C_2. Denote by $\mathbb{1}_{C_1} \in [0,1]^{m \times m}$ (respectively $\mathbb{1}_{C_2} \in [0,1]^{m \times m}$) the adjacency matrix of C_1 (respectively C_2), and let

$$\varepsilon = \frac{1}{2} \min \left(\min_{(i,j) \in F} x(i,j); 1 - \max_{(i,j) \in F} x(i,j) \right).$$

Then clearly $x + \varepsilon(\mathbb{1}_{C_1} - \mathbb{1}_{C_2})$ and $x - \varepsilon(\mathbb{1}_{C_1} - \mathbb{1}_{C_2})$ are in the convex set under consideration, which contradicts the fact that x is a vertex of that set (since there is a small interval which contains x and that is contained in the convex set). Thus we proved that F is forest (i.e. a disjoint union of trees). Assume that F is not empty. Then there exists one node such that exactly one edge of $K_{m,m}$ contains that node and is in F. Using that the sum of the weights x on the edges that contains that node must be 1, and the fact that exactly one of those edge has a non integer weight, we arrive at a contradiction. Thus $F = \emptyset$, which concludes the proof of the case $m = M$.

We now prove the general case of $m \leq M$. To do this we enlarge our vector, from $x \in \mathcal{D}_{m,M}$ we build $y \in \mathcal{D}_{m+M,m+M}$ as follows:

$$\forall (i,j) \in \{1,\dots,m\} \times \{1,\dots,M\}, y(i,j) = x(i,j), \text{ and } y(m+j, M+i) = x(i,j)$$
$$\forall j \in \{1,\dots,M\}, y(m+j,j) = 1 - \sum_{i=1}^{m} x(i,j)$$
$$\text{for all other pairs}(i,j), y(i,j) = 0.$$

Now we just proved that y can be decomposed as a convex combination of matchings on $K_{m+M,m+M}$. By considering the restriction of this assertion to $K_{m,M}$ we obtain that x can be decomposed as a convex combination of matchings on $K_{m,M}$. □

6.1.4. Spanning trees. Consider the complete graph K_{m+1} with $m \leq M$ (that is the set of vertices is of size $m+1$ and all possible edges are present). Let $\mathcal{C}$ contain all spanning trees of K_{m+1} (a spanning tree is a tree that contains all the vertices). Note that all elements v have the same size m, $d = (m+1)m$, and Cayley's formula gives $N = (m+1)^{m-1}$. Here the convex hull can be described as follows [2]:

THEOREM 6.3. *Let $\mathcal{C}$ be the set of spanning trees on K_{m+1}, then*

$$\begin{aligned} Conv(\mathcal{C}) \quad = \quad & \Bigg\{ x \in [0,1]^d : \sum_{i=1}^{d} x(i) = m, \\ & \qquad \text{and} \sum_{i \in \mathcal{P}} x(i) \leq |\mathcal{P}| - 1, \ \forall\ \mathcal{P} \subset \{1,\dots,m+1\} \Bigg\}. \end{aligned}$$

Thus here the convex hull has an exponential number of facets. However, interestingly enough, there exists efficient algorithms for the offline problem (such as Kruskal's method).

6.1.5. Paths. Another important example is when $\mathcal{C}$ represents incidence vectors of paths in a graph with d edges. In the case of a directed acyclic graph one can write the convex hull of $s-t$ paths (that is paths which start at node s and end at node t) as follows. Consider the following sign function for any node α:

$$\varepsilon_\alpha(i) = \begin{cases} 1 & \text{if } \alpha \text{ is an end point of } i \\ -1 & \text{if } \alpha \text{ is a start point of } i \\ 0 & \qquad \text{otherwise} \end{cases}$$

[2] For a subset of vertices $\mathcal{P} \subset \{1,\dots,m+1\}$ one says that an edge $i \in \{1,\dots,d\}$ is contained in $\mathcal{P}$ (denoted $i \in \mathcal{P}$) if i connects two vertices that are in $\mathcal{P}$.

THEOREM 6.4. *Let $\mathcal{C}$ be the set of $s-t$ paths on a directed acyclic graph with d edges, then*

$$\begin{aligned} Conv(\mathcal{C}) \quad = \quad & \Bigg\{ x \in [0,1]^d : \sum_{i=1}^{d} \varepsilon_\alpha(i) x(i) = 0, \ \forall \text{ node } \alpha, \\ & \text{and} \sum_{i:\varepsilon_s(i)=-1} x(i) = 1, \sum_{i:\varepsilon_t(i)=1} x(i) = 1 \Bigg\}. \end{aligned}$$

Note that for general directed graphs (i.e. which may contain cycles), it is impossible to obtain such a simple representation since the shortest path problem is NP-complete.

6.2. Lower bound

We derive here an almost trivial lower bound on the attainable regret which will serve as a benchmark in the following sections.

THEOREM 6.5. *There exists a set $\mathcal{C} \subset \{0,1\}^d$ with $||v||_1 = m, \forall v \in \mathcal{C}$, such that for any strategy (playing on $Conv(\mathcal{C})$), the following holds true:*

$$\sup_{n,m,d} \sup_{adversary} \frac{R_n}{m\sqrt{(n/2)\log(d/m)}} \geq 1.$$

PROOF. The proof is straightforward using [Theorem 2.3, Chapter 2] and the following set of concepts (for d a multiple of m):

$$\mathcal{C} = \left\{ v \in \{0,1\}^d : \forall i \in \{1,\dots,m\}, \sum_{j=(i-1)m+1}^{im} v(j) = 1 \right\}.$$

□

6.3. Expanded Exp (Exp2)

A very simple strategy for online combinatorial optimization is to consider each point in $\mathcal{C}$ as an expert and play according to the Exp strategy. That is one select at time t, $V_t = \mathbb{E}_{v\sim p_t} v$ where

$$p_t(v) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} z_s^T v\right)}{\sum_{u\in\mathcal{C}} \exp\left(-\eta \sum_{s=1}^{t-1} z_s^T u\right)}.$$

Using the results of the previous chapter one can directly prove the following upper bound for the resulting strategy (which we call Exp2 for Expanded Exp).

THEOREM 6.6. *For any set $\mathcal{C} \subset \{0,1\}^d$ with $||v||_1 = m, \forall v \in \mathcal{C}$, Exp2 with $\eta = \sqrt{\frac{2\log\left(\frac{ed}{m}\right)}{nm}}$ satisfies:*

$$R_n \leq m^{3/2} \sqrt{2n \log\left(\frac{ed}{m}\right)}.$$

PROOF. Simply note that $\log|\mathcal{C}| \leq \log\binom{d}{m} \leq m \log\left(\frac{ed}{m}\right)$, and $\ell(v,z) = z_t^T v \in [0,m]$. □

Surprisingly one can see that there is a gap between this upper bound and the lower bound of the previous section. It is natural to ask whether one can improve the analysis of Exp2. The following theorem shows that in fact Exp2 is provably suboptimal for online combinatorial optimization.

THEOREM 6.7. *Let $n \geq d$. There exists a subset $\mathcal{C} \subset \{0,1\}^d$ such that in the full information game, for the* EXP2 *strategy (for any learning rate η), we have*

$$\sup_{adversary} R_n \geq 0.01\, d^{3/2}\sqrt{n}.$$

PROOF. For sake of simplicity we assume here that d is a multiple of 4 and that n is even. We consider the following subset of the hypercube:

$$\mathcal{C} = \Bigg\{ v \in \{0,1\}^d : \sum_{i=1}^{d/2} v_i = d/4 \ \text{ and}$$
$$\Big(v_i = 1, \forall i \in \{d/2+1; \dots, d/2+d/4\} \Big) \ \text{ or } \ \Big(v_i = 1, \forall i \in \{d/2+d/4+1, \dots, d\} \Big) \Bigg\}.$$

That is, choosing a point in $\mathcal{C}$ corresponds to choosing a subset of $d/4$ elements in the first half of the coordinates, and choosing one of the two first disjoint intervals of size $d/4$ in the second half of the coordinates.

We will prove that for any parameter η, there exists an adversary such that Exp2 (with parameter η) has a regret of at least $\frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right)$, and that there exists another adversary such that its regret is at least $\min\left(\frac{d \log 2}{12\eta}, \frac{nd}{12}\right)$. As a consequence, we have

$$\begin{aligned}
\sup R_n &\geq \max\left(\frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right), \min\left(\frac{d\log 2}{12\eta}, \frac{nd}{12}\right)\right) \\
&\geq \min\left(\max\left(\frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right), \frac{d\log 2}{12\eta}\right), \frac{nd}{12}\right) \geq \min\left(A, \frac{nd}{12}\right),
\end{aligned}$$

with

$$\begin{aligned}
A &= \min_{\eta \in [0,+\infty)} \max\left(\frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right), \frac{d\log 2}{12\eta}\right) \\
&\geq \min\left\{ \min_{\eta d \geq 8} \frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right), \min_{\eta d < 8} \max\left(\frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right), \frac{d\log 2}{12\eta}\right) \right\} \\
&\geq \min\left\{ \frac{nd}{16}\tanh(1), \min_{\eta d < 8} \max\left(\frac{nd}{16}\frac{\eta d}{8}\tanh(1), \frac{d\log 2}{12\eta}\right) \right\} \\
&\geq \min\left\{ \frac{nd}{16}\tanh(1), \sqrt{\frac{nd^3 \log 2 \times \tanh(1)}{128 \times 12}} \right\} \geq \min\left(0.04\, nd, 0.01\, d^{3/2}\sqrt{n}\right).
\end{aligned}$$

Let us first prove the lower bound $\frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right)$. We define the following adversary:

$$z_t(i) = \begin{cases} 1 & \text{if} \quad i \in \{d/2+1; \dots, d/2+d/4\} \ \text{ and } \ t \ \text{ odd}, \\ 1 & \text{if} \quad i \in \{d/2+d/4+1, \dots, d\} \ \text{ and } \ t \ \text{ even}, \\ 0 & \text{otherwise}. \end{cases}$$

This adversary always put a zero loss on the first half of the coordinates, and alternates between a loss of $d/4$ for choosing the first interval (in the second half of the coordinates) and the second interval. At the beginning of odd rounds, any vertex $v \in \mathcal{C}$ has the same cumulative loss and thus Exp2 picks its expert uniformly at random, which yields an expected cumulative loss equal to $nd/16$. On the other hand at even rounds the probability distribution to select the vertex $v \in \mathcal{C}$ is always the same. More precisely the probability of selecting a vertex which contains the interval $\{d/2 + d/4 + 1, \dots, d\}$ (i.e, the interval with a $d/4$ loss at this round) is exactly $\frac{1}{1+\exp(-\eta d/4)}$. This adds an expected cumulative loss equal to $\frac{nd}{8}\frac{1}{1+\exp(-\eta d/4)}$. Finally note that the loss of any fixed vertex is $nd/8$. Thus we obtain

$$R_n = \frac{nd}{16} + \frac{nd}{8}\frac{1}{1+\exp(-\eta d/4)} - \frac{nd}{8} = \frac{nd}{16}\tanh\left(\frac{\eta d}{8}\right).$$

We move now to the dependency in $1/\eta$. Here we consider the adversary defined by:

$$z_t(i) = \begin{cases} 1-\varepsilon & \text{if} & i \le d/4, \\ 1 & \text{if} & i \in \{d/4+1, \dots, d/2\}, \\ 0 & \text{otherwise.} \end{cases}$$

Note that against this adversary the choice of the interval (in the second half of the components) does not matter. Moreover by symmetry the weight of any coordinate in $\{d/4+1, \dots, d/2\}$ is the same (at any round). Finally remark that this weight is decreasing with t. Thus we have the following identities (in the big sums i represents the number of components selected in the first $d/4$ components):

$$\begin{aligned} R_n &= \frac{n\varepsilon d}{4}\frac{\sum_{v\in\mathcal{C}:v(d/2)=1}\exp(-\eta n z_1^T v)}{\sum_{v\in\mathcal{S}}\exp(-\eta n z_1^T v)} \\ &= \frac{n\varepsilon d}{4}\frac{\sum_{i=0}^{d/4-1}\binom{d/4}{i}\binom{d/4-1}{d/4-i-1}\exp(-\eta(nd/4 - in\varepsilon))}{\sum_{i=0}^{d/4}\binom{d/4}{i}\binom{d/4}{d/4-i}\exp(-\eta(nd/4 - in\varepsilon))} \\ &= \frac{n\varepsilon d}{4}\frac{\sum_{i=0}^{d/4-1}\binom{d/4}{i}\binom{d/4-1}{d/4-i-1}\exp(\eta in\varepsilon)}{\sum_{i=0}^{d/4}\binom{d/4}{i}\binom{d/4}{d/4-i}\exp(\eta in\varepsilon)} \\ &= \frac{n\varepsilon d}{4}\frac{\sum_{i=0}^{d/4-1}\left(1-\frac{4i}{d}\right)\binom{d/4}{i}\binom{d/4}{d/4-i}\exp(\eta in\varepsilon)}{\sum_{i=0}^{d/4}\binom{d/4}{i}\binom{d/4}{d/4-i}\exp(\eta in\varepsilon)} \end{aligned}$$

where we used $\binom{d/4-1}{d/4-i-1} = \left(1-\frac{4i}{d}\right)\binom{d/4}{d/4-i}$ in the last equality. Thus taking $\varepsilon = \min\left(\frac{\log 2}{\eta n}, 1\right)$ yields

$$R_n \ge \min\left(\frac{d\log 2}{4\eta}, \frac{nd}{4}\right)\frac{\sum_{i=0}^{d/4-1}\left(1-\frac{4i}{d}\right)\binom{d/4}{i}^2\min(2,\exp(\eta n))^i}{\sum_{i=0}^{d/4}\binom{d/4}{i}^2\min(2,\exp(\eta n))^i} \ge \min\left(\frac{d\log 2}{12\eta}, \frac{nd}{12}\right),$$

where the last inequality follows from Lemma 6.2 below. This concludes the proof of the lower bound. □

Lemma 6.2. *For any $k \in \mathbb{N}^*$, for any $1 \le c \le 2$, we have*

$$\frac{\sum_{i=0}^{k}(1-i/k)\binom{k}{i}^2 c^i}{\sum_{i=0}^{k}\binom{k}{i}^2 c^i} \ge 1/3.$$

Proof. Let $f(c)$ denote the left-hand side term of the inequality. Introduce the random variable X, which is equal to $i \in \{0, \dots, k\}$ with probability $\binom{k}{i}^2 c^i / \sum_{j=0}^k \binom{k}{j}^2 c^j$. We have $f'(c) = \frac{1}{c}\mathbb{E}[X(1 - X/k)] - \frac{1}{c}\mathbb{E}(X)\mathbb{E}(1 - X/k) = -\frac{1}{ck}\mathbb{V}\mathrm{ar}\, X \le 0$. So the function f is decreasing on $[1, 2]$, and, from now on, we consider $c = 2$. Numerator and denominator of the left-hand side (l.h.s.) differ only by the $1 - i/k$ factor. A lower bound for the left-hand side can thus be obtained by showing that the terms for i close to k are not essential to the value of the denominator. To prove this, we may use the Stirling formula: for any $n \ge 1$

$$\left(\frac{n}{e}\right)^n \sqrt{2\pi n} < n! < \left(\frac{n}{e}\right)^n \sqrt{2\pi n} e^{1/(12n)} \tag{6.1}$$

Indeed, this inequality implies that for any $k \ge 2$ and $i \in [1, k-1]$

$$\left(\frac{k}{i}\right)^i \left(\frac{k}{k-i}\right)^{k-i} \frac{\sqrt{k}}{\sqrt{2\pi i(k-i)}} e^{-1/6} < \binom{k}{i} < \left(\frac{k}{i}\right)^i \left(\frac{k}{k-i}\right)^{k-i} \frac{\sqrt{k}}{\sqrt{2\pi i(k-i)}} e^{1/12},$$

hence

$$\left(\frac{k}{i}\right)^{2i} \left(\frac{k}{k-i}\right)^{2(k-i)} \frac{ke^{-1/3}}{2\pi i(k-i)} < \binom{k}{i}^2 < \left(\frac{k}{i}\right)^{2i} \left(\frac{k}{k-i}\right)^{2(k-i)} \frac{ke^{1/6}}{2\pi i}$$

Introduce $\lambda = i/k$ and $\chi(\lambda) = \frac{2^\lambda}{\lambda^{2\lambda}(1-\lambda)^{2(1-\lambda)}}$. We have

$$[\chi(\lambda)]^k \frac{2e^{-1/3}}{\pi k} < \binom{k}{i}^2 2^i < [\chi(\lambda)]^k \frac{e^{1/6}}{2\pi\lambda}. \tag{6.2}$$

Lemma 6.2 can be numerically verified for $k \le 10^6$. We now consider $k > 10^6$. For $\lambda \ge 0.666$, since the function χ can be shown to be decreasing on $[0.666, 1]$, the inequality $\binom{k}{i}^2 2^i < [\chi(0.666)]^k \frac{e^{1/6}}{2 \times 0.666 \times \pi}$ holds. We have $\chi(0.657)/\chi(0.666) > 1.0002$. Consequently, for $k > 10^6$, we have $[\chi(0.666)]^k < 0.001 \times [\chi(0.657)]^k/k^2$. So for $\lambda \ge 0.666$ and $k > 10^6$, we have

$$\begin{aligned}\binom{k}{i}^2 2^i < 0.001 \times [\chi(0.657)]^k \frac{e^{1/6}}{2\pi \times 0.666 \times k^2} &< [\chi(0.657)]^k \frac{2e^{-1/3}}{1000\pi k^2} \\ &= \min_{\lambda \in [0.656, 0.657]} [\chi(\lambda)]^k \frac{2e^{-1/3}}{1000\pi k^2} \\ &< \frac{1}{1000k} \max_{i \in \{1,\dots,k-1\} \cap [0, 0.666k)} \binom{k}{i}^2 2^i. \end{aligned} \tag{6.3}$$

where the last inequality comes from (6.2) and the fact that there exists $i \in \{1, \dots, k-1\}$ such that $i/k \in [0.656, 0.657]$. Inequality (6.3) implies that for any $i \in \{1, \dots, k\}$, we have

$$\sum_{\frac{5}{6}k \le i \le k} \binom{k}{i}^2 2^i < \frac{1}{1000} \max_{i \in \{1,\dots,k-1\} \cap [0, 0.666k)} \binom{k}{i}^2 2^i < \frac{1}{1000} \sum_{0 \le i < 0.666k} \binom{k}{i}^2 2^i.$$

To conclude, introducing $A = \sum_{0 \le i < 0.666k} \binom{k}{i}^2 2^i$, we have

$$\frac{\sum_{i=0}^k (1 - i/k) \binom{k}{i}\binom{k}{k-i} 2^i}{\sum_{i=0}^k \binom{k}{i}\binom{k}{k-i} 2^i} > \frac{(1 - 0.666)A}{A + 0.001A} \ge \frac{1}{3}.$$

□

6.4. OMD with negative entropy

Here we show that OMD with the negative entropy attains the optimal rate. Note that, on the contrary to the simplex case, in general OMD with negative entropy on $Conv(\mathcal{C})$ and discrete Exp on $\mathcal{C}$ are two different strategies.

THEOREM 6.8. *For any set $\mathcal{C} \subset \{0,1\}^d$ with $||v||_1 = m, \forall v \in \mathcal{C}$, OMD with $F(x) = \sum_{i=1}^d x_i \log x_i - x_i$, and $\eta = \sqrt{\frac{2\log\left(\frac{d}{m}\right)}{nm}}$ satisfies:*

$$R_n \leq m\sqrt{2n\log\left(\frac{d}{m}\right)}.$$

PROOF. Using Theorem 5.2 and Lemma 5.9, it suffices to show that

$$F(v) - F(V_1) \leq m \log \frac{d}{m}, \forall v \in \mathcal{C}.$$

This follows from:

$$F(v) - F(V_1) \leq \sum_{i=1}^d V_1(i) \log \frac{1}{V_1(i)} \leq m \log \left(\sum_{i=1}^d \frac{V_1(i)}{m} \frac{1}{V_1(i)} \right) = m \log \frac{d}{m}.$$

□

6.5. Follow the perturbated leader (FPL)

In this section we consider a completely different strategy, called Follow the Perturbated Leader. The idea is very simple. It is clear that following the leader, i.e. choosing at time t:

$$\underset{v\in\mathcal{C}}{\operatorname{argmin}} \sum_{s=1}^{t-1} z_s^T v,$$

is a strategy that can be hazardous. In FPL, this choice is *regularized* by adding a small amount of noise. More precisely let $\xi_1, \dots, \xi_n$ be an i.i.d sequence of random variables uniformly drawn on $[0, 1/\eta]^d$. Then FPL corresponds to the decision:

$$\underset{v\in\mathcal{C}}{\operatorname{argmin}} \left(\xi_t + \sum_{s=1}^{t-1} z_s \right)^T v,$$

We analyze this strategy in a restrictive framework, namely we only consider oblivious adversaries (that is the sequence (z_t) is fixed and can not depend on the moves v_t of the player).

THEOREM 6.9. *For any oblivious adversary, the FPL strategy satisfies for any $u \in \mathcal{C}$:*

$$\mathbb{E}\left(\sum_{t=1}^n z_t^T (v_t - u) \right) \leq \frac{m}{2\eta} + \eta m d n.$$

In particular with $\eta = \sqrt{\frac{1}{2dn}}$ one obtains:

$$\mathbb{E}\left(\sum_{t=1}^n z_t^T (v_t - u) \right) \leq m\sqrt{2dn}$$

The first step of the proof is the so-called *Be-The-Leader* Lemma.

LEMMA 6.3. *Let*

$$a_t^* = \operatorname*{argmin}_{a \in \mathcal{A}} \sum_{s=1}^{t} \ell(a, z_t).$$

Then

$$\sum_{t=1}^{n} \ell(a_t^*, z_t) \leq \sum_{t=1}^{n} \ell(a_n^*, z_t).$$

PROOF. The proof goes by induction on n. For $n = 1$ it is clearly true. From n to $n+1$ it follows from:

$$\sum_{t=1}^{n+1} \ell(a_t^*, z_t) \leq \ell(a_{n+1}^*, z_{n+1}) + \sum_{t=1}^{n} \ell(a_n^*, z_t) \leq \sum_{t=1}^{n+1} \ell(a_{n+1}^*, z_t).$$

□

We can now prove the theorem.

PROOF. Let

$$v_t^* = \operatorname*{argmin}_{v \in \mathcal{C}} \left(\xi_1 + \sum_{s=1}^{t} z_s \right)^T v.$$

Using the BTL Lemma with

$$\ell(v, z_t) = \begin{cases} (\xi_1 + z_1)^T v_1 & \text{if } t = 1, \\ z_t^T v & \text{if } t > 1, \end{cases}$$

one obtains that for any $u \in \mathcal{C}$,

$$\xi_1^T v_1^* + \sum_{t=1}^{n} z_t^T v_t^* \leq \xi_1^T u + \sum_{t=1}^{n} z_t^T u.$$

In particular we get

$$\mathbb{E} \sum_{t=1}^{n} z_t^T (v_t^* - u) \leq \frac{m}{2\eta}.$$

Now let

$$\widetilde{v}_t = \operatorname*{argmin}_{v \in \mathcal{C}} \left(\xi_t + \sum_{s=1}^{t} z_s \right)^T v.$$

Since the adversary is oblivious, $\widetilde{v}_t$ has the same distribution than v_t^*, in particular we have $\mathbb{E} z_t^T v_t^* = \mathbb{E} z_t^T \widetilde{v}_t$, which implies

$$\mathbb{E} \sum_{t=1}^{n} z_t^T (\widetilde{v}_t - u) \leq \frac{m}{2\eta}.$$

We show now that $\mathbb{E} z_t^T (v_t - \widetilde{v}_t) \leq \eta m d$ which shall conclude the proof. Let

$$h(\xi) = z_t^T \left\{ \operatorname*{argmin}_{v \in \mathcal{C}} \left(\xi + \sum_{s=1}^{t-1} z_s \right)^T v \right\}.$$

Then one has

$$\begin{aligned}
\mathbb{E}z_t^T(v_t - \widetilde{v}_t) &= \mathbb{E}h(\xi_t) - \mathbb{E}h(\xi_t + z_t) \\
&= \eta^d \int_{\xi \in [0,1/\eta]^d} h(\xi) d\xi - \eta^d \int_{\xi \in z_t + [0,1/\eta]^d} h(\xi) d\xi \\
&\leq m\eta^d \int_{\xi \in [0,1/\eta]^d \setminus \{z_t + [0,1/\eta]^d\}} \\
&= m\mathbb{P}\left(\exists i \in \{1, \dots, d\} : \xi_1(i) \leq z_t(i)\right) \\
&\leq \eta m d.
\end{aligned}$$

□

Note that the bound we proved for FPL is suboptimal by a factor $\sqrt{d}$. It is likely that in fact FPL is provably suboptimal (a similar reasoning than for Exp2 should be possible). Nonetheless FPL has the advantage that it is computationally efficient as soon as there exists efficient algorithms for the offline problem. This is an important property, and the major open problem in online combinatorial optimization is to decide whether there exists a strategy with this property and optimal regret bounds.

References

A general and extremely useful reference for combinatorial optimization is:

- A. Schrijver. *Combinatorial Optimization*. Springer, 2003

Online combinatorial optimization was introduced in:

- N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 2011. To appear

Note however that many earlier works studied specific examples of online combinatorial optimization. See in particular the following list of papers [33, 57, 27, 26, 37, 56].

The proof that Exp is suboptimal in that framework is extracted from:

- J.-Y. Audibert, S. Bubeck, and G. Lugosi. Minimax policies for combinatorial prediction games. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, 2011

The important FPL strategy was introduced and studied in:

- A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71:291–307, 2005

CHAPTER 7

Limited feedback

In this chapter we attack the limited feedback case. In the so-called bandit version, one only observes $\ell(a_t, z_t)$ rather than the adversary's move z_t. Thus it is not possible to use OMD since one does not have access to the gradient $\nabla \ell(a_t, z_t)$. First we describe the general idea to attack this problem, namely to add randomization in order to build unbiased estimates of the gradients.

7.1. Online Stochastic Mirror Descent (OSMD)

The idea of stochastic gradient descent is very simple: assume that one wants to play a_t. Then one builds a random perturbation $\widetilde{a}_t$ of a_t, with the property that upon observing $\ell(\widetilde{a}_t, z_t)$, one can build an estimate $\widetilde{g}_t$ of $\nabla \ell(a_t, z_t)$. Then one feeds the gradient descent algorithm with $\widetilde{g}_t$ instead of $\nabla \ell(a_t, z_t)$ and follows the same scheme at point a_{t+1}. This strategy allows to bound the pseudo-regret defined by:

$$\overline{R}_n = \mathbb{E} \sum_{t=1}^{n} \ell(\widetilde{a}_t, z_t) - \min_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} \ell(a, z_t).$$

Indeed one can prove the following theorem.

Theorem 7.1. *For any closed convex action set $\mathcal{A}$, for any subdifferentiable loss with bounded subgradient $||\nabla \ell(a, z)||_* \leq G, \forall (a, z) \in \mathcal{A} \times \mathcal{Z}$, the OSMD strategy with a Legendre function F on $\mathcal{D}$ such that its restriction to $\mathcal{A}$ is α-strongly convex with respect to $||\cdot||$, and with loss estimate $\widetilde{g}_t$ such that $\mathbb{E}(\widetilde{g}_t | a_t) = \nabla \ell(a_t, z_t)$ satisfies*

$$\overline{R}_n \leq \frac{\sup_{a \in \mathcal{A}} F(a) - F(a_1)}{\eta} + \frac{\eta}{2\alpha} \sum_{t=1}^{n} \mathbb{E}||\widetilde{g}_t||_*^2 + G \sum_{t=1}^{n} \mathbb{E}||a_t - \widetilde{a}_t||.$$

Moreover if the loss is linear, that is $\ell(a, z) = a^T z$, then

$$\overline{R}_n \leq \frac{\sup_{a \in \mathcal{A}} F(a) - F(a_1)}{\eta} + \frac{\eta}{2\alpha} \sum_{t=1}^{n} \mathbb{E}||\widetilde{g}_t||_*^2 + G \sum_{t=1}^{n} \mathbb{E}||a_t - \mathbb{E}(\widetilde{a}_t | a_t)||.$$

Proof. Using Theorem 5.2 one directly obtains:

$$\sum_{t=1}^{n} \widetilde{g}_t^T (a_t - a) \leq \frac{F(a) - F(a_1)}{\eta} + \frac{\eta}{2\alpha} \sum_{t=1}^{n} ||\widetilde{g}_t||_*^2.$$

Moreover, using $\mathbb{E}(\widetilde{g}_t|a_t) = \nabla\ell(a_t, z_t)$, we have:

$$\begin{aligned}
\mathbb{E}\sum_{t=1}^{n}(\ell(\widetilde{a}_t, z_t) - \ell(a, z_t)) &= \mathbb{E}\sum_{t=1}^{n}(\ell(\widetilde{a}_t, z_t) - \ell(a_t, z_t) + \ell(a_t, z_t) - \ell(a, z_t)) \\
&\leq G\mathbb{E}\sum_{t=1}^{n}||a_t - \widetilde{a}_t|| + \mathbb{E}\sum_{t=1}^{n}\nabla\ell(a_t, z_t)^T(a_t - a) \\
&= G\mathbb{E}\sum_{t=1}^{n}||a_t - \widetilde{a}_t|| + \mathbb{E}\sum_{t=1}^{n}\widetilde{g}_t^T(a_t - a),
\end{aligned}$$

which concludes the proof of the first regret bound. The case of a linear loss follows very easily from the same computations. □

Unfortunately the above theorem is not strong enough to derive optimal regret bounds, because in most cases it is not possible to obtain a satisfactory bound on $\mathbb{E}||\widetilde{g}_t||_*^2$. In fact the key is to replace the rigid norm $||\cdot||_*$ by a local norm which depends on the current point a_t. This is achieved with the following theorem (whose proof follows the same line than Theorem 7.1, by using Theorem 5.1 instead of Theorem 5.2).

Theorem 7.2. *For any closed convex action set $\mathcal{A}$, for any subdifferentiable loss, the OSMD strategy with a Legendre function F on $\mathcal{D}$, and with loss estimate $\widetilde{g}_t$ such that $\mathbb{E}(\widetilde{g}_t|a_t) = \nabla\ell(a_t, z_t)$ satisfies*

$$\overline{R}_n \leq \frac{\sup_{a\in\mathcal{A}} F(a) - F(a_1)}{\eta} + \frac{1}{\eta}\sum_{t=1}^{n}\mathbb{E}D_{F^*}\Big(\nabla F(a_t) - \eta\widetilde{g}_t, \nabla F(a_t)\Big) + G\sum_{t=1}^{n}\mathbb{E}||a_t - \widetilde{a}_t||.$$

Moreover if the loss is linear, that is $\ell(a, z) = a^T z$, then

$$\overline{R}_n \leq \frac{\sup_{a\in\mathcal{A}} F(a) - F(a_1)}{\eta} + \frac{1}{\eta}\sum_{t=1}^{n}\mathbb{E}D_{F^*}\Big(\nabla F(a_t) - \eta\widetilde{g}_t, \nabla F(a_t)\Big) + G\sum_{t=1}^{n}\mathbb{E}||a_t - \mathbb{E}(\widetilde{a}_t|a_t)||.$$

The term $\mathbb{E}D_{F^*}\Big(\nabla F(a_t) - \eta\widetilde{g}_t, \nabla F(a_t)\Big)$ corresponds to a "local" norm of $\eta\widetilde{g}_t$ in the following sense.

Proposition 7.1. *If F is twice continuously differentiable, and if its Hessian $\nabla^2 F(x)$ is invertible $\forall x \in \mathcal{D}$, then $\forall x, y \in \mathcal{D}$, there exists $\zeta \in \mathcal{D}$ such that $\nabla F(\zeta) \in [\nabla F(x), \nabla F(y)]$ and:*

$$D_{F^*}(\nabla F(x), \nabla F(y)) = \frac{1}{2}||\nabla F(x) - \nabla F(y)||^2_{(\nabla^2 F(\zeta))^{-1}},$$

where $||x||_Q^2 = x^T Q x$.

Proof. Since F is twice continuously differentiable, the inverse function theorem implies that:

$$J_{(\nabla F)^{-1}}(\nabla F(x)) = (J_{\nabla F}(x))^{-1}.$$

In other words, since $\nabla F^* = (\nabla F)^{-1}$,

$$\nabla^2 F^*(\nabla F(x)) = (\nabla^2 F(x))^{-1}.$$

The proposition then follows from a simple Taylor's expansion. □

This proposition shows that for some ζ_t such that $\nabla F(\zeta_t) \in [\nabla F(a_t) - \eta \widetilde{g}_t, \nabla F(a_t)]$, we have:

$$D_{F^*}\Big(\nabla F(a_t) - \eta \widetilde{g}_t, \nabla F(a_t)\Big) = \frac{\eta^2}{2} ||\widetilde{g}_t||^2_{(\nabla^2 F(\zeta_t))^{-1}}.$$

The key to obtain optimal regret bounds will be to make the above equality more precise (in terms of ζ_t) for specific Legendre functions F (or in other words to take care of the third order error term in the Taylor's expansion).

7.2. Online combinatorial optimization with semi-bandit feedback

In this section we consider the online combinatorial optimization problem: the action set is $\mathcal{C} \subset \{0,1\}^d$, $\mathcal{Z} = [0,1]^d$, and $\ell(v,z) = v^T z$. We call *semi-bandit* feedback, the case when after playing $V_t \in \mathcal{C}$, one observes $(z_t(1)V_t(1), \dots, z_t(d)V_t(d))$. That is one observes only the coordinates of the loss that were *active* in the concept V_t that we choosed. It is thus a much weaker feedback than in the full information case, but it is also stronger than in the bandit version. Note that the semi-bandit setting includes the famous multi-armed bandit problem, which simply corresponds to $\mathcal{C} = \{e_1, \dots, e_d\}$.

Recall that in this setting one plays V_t at random from a probability $p_t \in \Delta_N$ (where $|\mathcal{C}| = N$) to which corresponds an average point $a_t \in Conv(\mathcal{C})$. Surprisingly, we show that this randomization is enough to obtain a good unbiased estimate of the loss and that it is not necessary to add further perturbations to a_t. Thus here we have $\widetilde{a}_t = V_t$, and in particular $\mathbb{E}(\widetilde{a}_t | a_t) = a_t$.

Note that $\nabla \ell(v,z) = z$, thus $\widetilde{g}_t$ should be an estimate of z_t. The following simple formula gives an unbiased estimate:

$$\widetilde{g}_t(i) = \frac{z_t(i)V_t(i)}{a_t(i)}, \forall i \in \{1, \dots, d\}. \tag{7.1}$$

Note that this is a valid estimate since it makes only use of $(z_t(1)V_t(1), \dots, z_t(d)V_t(d))$. Moreover it is unbiased with respect to the random drawing of V_t from p_t since by definition $\mathbb{E}_{V_t \sim p_t} V_t(i) = a_t(i)$. In other words $\mathbb{E}(\widetilde{g}_t | a_t) = \nabla \ell(a_t, z_t)$.

Using Theorem 7.2 we directly obtain:

$$\overline{R}_n \le \frac{\sup_{a \in \mathcal{A}} F(a) - F(a_1)}{\eta} + \frac{1}{\eta} \sum_{t=1}^n \mathbb{E} D_{F^*}\Big(\nabla F(a_t) - \eta \widetilde{g}_t, \nabla F(a_t)\Big). \tag{7.2}$$

We show now how to use this bound to obtain concrete performances for OSMD with the negative entropy. Then we show that one can improve the results by logarithmic factors, using a more subtle Legendre function.

7.2.1. Negative entropy.

THEOREM 7.3. *OSMD with the negative entropy $F(x) = \sum_{i=1}^d x_i \log x_i - \sum_{i=1}^d x_i$ satisfies:*

$$\overline{R}_n \le \frac{m \log \frac{d}{m}}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{i=1}^d a_t(i) \widetilde{g}_t(i)^2.$$

In particular with the estimate (7.1) *and* $\eta = \sqrt{2\frac{m \log dm}{nd}}$,

$$\overline{R}_n \leq \sqrt{2mdn \log \frac{d}{m}}.$$

Proof. We already showed in the proof of Theorem 6.8 that

$$F(a) - F(a_1) \leq m \log \frac{d}{m}.$$

Moreover we showed in Chapter 5 that:

$$D_{F^*}\Big(\nabla F(a_t) - \eta \widetilde{g}_t, \nabla F(a_t)\Big) = \sum_{i=1}^{d} a_t(i)\Theta(-\eta \widetilde{g}_t(i)),$$

where $\Theta : x \in \mathbb{R} \mapsto \exp(x) - 1 - x$. Thus Lemma 5.5 ends the proof of the first inequality (since $\widetilde{g}_t(i) \geq 0$). The second inequality follows from:

$$\mathbb{E} a_t(i) \widetilde{g}_t(i)^2 \leq \mathbb{E}\frac{V_t(i)}{a_t(i)} = 1.$$

□

7.2.2. Legendre function derived from a potential. We greatly generalize the negative entropy with the following definition.

Definition 7.1. *Let* $\omega \geq 0$. *A function* $\psi : (-\infty, a) \to \mathbb{R}_+^*$ *for some* $a \in \mathbb{R} \cup \{+\infty\}$ *is called an* ω*-potential if it is convex, continuously differentiable, and satisfies*

$$\lim_{x \to -\infty} \psi(x) = \omega \qquad \lim_{x \to a} \psi(x) = +\infty$$

$$\psi' > 0 \qquad \int_{\omega}^{\omega+1} |\psi^{-1}(s)| ds < +\infty.$$

To a potential ψ *we associate the function* F_ψ *defined on* $\mathcal{D} = (\omega, +\infty)^d$ *by:*

$$F_\psi(x) = \sum_{i=1}^{d} \int_{\omega}^{x_i} \psi^{-1}(s) ds.$$

In these lecture notes we restrict our attention to 0-potentials. A non-zero ω might be used to derive high probability regret bounds (instead of pseudo-regret bounds).

Note that with $\psi(x) = \exp(x)$ we recover the negative entropy for F_ψ.

Lemma 7.1. *Let* ψ *be a (*0*-)potential. Then* F_ψ *is Legendre, and for all* $u, v \in \mathcal{D}^* = (-\infty, a)^d$ *such that* $u_i \leq v_i, \forall i \in \{1, \ldots, d\}$,

$$D_{F^*}(u, v) \leq \frac{1}{2}\sum_{i=1}^{d} \psi'(v_i)(u_i - v_i)^2.$$

Proof. It is easy to check that F is a Legendre function. Moreover, since $\nabla F^*(u) = (\nabla F)^{-1}(u) = \big(\psi(u_1), \ldots, \psi(u_d)\big)$, we obtain

$$D_{F^*}(u, v) = \sum_{i=1}^{d}\bigg(\int_{v_i}^{u_i} \psi(s) ds - (u_i - v_i)\psi(v_i)\bigg).$$

From a Taylor expansion, we have

$$D_{F^*}(u,v) \le \sum_{i=1}^d \max_{s \in [u_i, v_i]} \frac{1}{2} \psi'(s)(u_i - v_i)^2.$$

Since the function ψ is convex, and $u_i \le v_i$, we have

$$\max_{s \in [u_i, v_i]} \psi'(s) \le \psi'\big(\max(u_i, v_i)\big) \le \psi'(v_i),$$

which gives the desired result. □

THEOREM 7.4. *Let ψ be a potential. OSMD with F_ψ and non-negative loss estimates $(\widetilde{g}_t)$ satisfies:*

$$\overline{R}_n \le \frac{\sup_{a \in \mathcal{A}} F_\psi(a) - F_\psi(a_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \sum_{i=1}^d \mathbb{E} \frac{\widetilde{g}_t(i)^2}{(\psi^{-1})'(a_t(i))}.$$

In particular with the estimate (7.1)*, $\psi(x) = (-x)^{-q}$ ($q > 1$) and $\eta = \sqrt{\frac{2}{q-1} \frac{m^{1-2/q}}{d^{1-2/q}}}$,*

$$\overline{R}_n \le q \sqrt{\frac{2}{q-1} mdn}.$$

With $q = 2$ this gives:

$$\overline{R}_n \le 2\sqrt{2mdn}.$$

PROOF. First note that since $\mathcal{D}^* = (-\infty, a)^d$ and $\widetilde{g}_t$ has non-negative coordinates, OSMD is well defined (that is (5.4) is satisfied).

The first inequality trivially follows from (7.2), Lemma 7.1, and the fact that $\psi'(\psi^{-1}(s)) = \frac{1}{(\psi^{-1})'(s)}$.

Let $\psi(x) = (-x)^{-q}$. Then $\psi^{-1}(x) = -x^{-1/q}$ and $F(x) = -\frac{q}{q-1} \sum_{i=1}^d x_i^{1-1/q}$. In particular note that by Hölder's inequality, since $\sum_{i=1}^d a_1(i) = m$:

$$F_\psi(a) - F_\psi(a_1) \le \frac{q}{q-1} \sum_{i=1}^d a_1(i)^{1-1/q} \le \frac{q}{q-1} m^{(q-1)/q} d^{1/q}.$$

Moreover note that $(\psi^{-1})'(x) = \frac{1}{q} x^{-1-1/q}$, and

$$\sum_{i=1}^d \mathbb{E} \frac{\widetilde{g}_t(i)^2}{(\psi^{-1})'(a_t(i))} \le q \sum_{i=1}^d a_t(i)^{1/q} \le q m^{1/q} d^{1-1/q},$$

which ends the proof. □

7.3. Online linear optimization with bandit feedback

In this section we consider online linear optimization over a finite set, that is $\mathcal{A} \subset \mathbb{R}^d$ is a finite set of points with $|\mathcal{A}| = N$, $\mathcal{Z} \subset \mathbb{R}^d$, and $\ell(a,z) = a^T z$. As far as regret bound goes, the restriction to finite set is not a severe one, since given a convex set one can always build a discretization $\mathcal{A}$ of that set such that the regret with respect to $\mathcal{A}$ is almost the same than the regret with respect to the convex set. However a strategy resulting from such a discretization is often not computationally efficient. Also note that typically in that case $\log N \sim d$.

Without loss of generality we restrict our attention to sets $\mathcal{A}$ of full rank, that is such that linear combinations of $\mathcal{A}$ span $\mathbb{R}^d$. If it is not the case then one can rewrite the elements of $\mathcal{A}$ in some lower dimensional vector space, and work there.

In this section we consider a *bounded scalar loss*, that is $\mathcal{A}$ and $\mathcal{Z}$ are such that $|\ell(a, z)| \leq G$. Note that so far we have been mainly working on *dual assumptions*, where $\mathcal{A}$ was bounded in some norm, and the gradient of the loss was bounded in the dual norm. Of course by Hölder's inequality this imply a bound on the scalar loss. However, it is important to note that in some cases one may exploit the dual assumption and prove better regret bounds than what would get by simply assuming a bounded scalar loss. This was in fact proved in Chapter 6, where we saw that Exp was provably suboptimal for online combinatorial optimization, while one can show that with only a bounded scalar loss assumption Exp is an optimal strategy. The main reason for restricting our attention to bounded scalar loss is simply that we do not know how to exploit a dual assumption in the bandit framework. More precisely the best strategy that we have (and that we will describe) is based on Exp (with a good perturbation and a good estimator for z_t), and we know that such strategies can not exploit (in general) dual assumptions.

The rest of this section is organized as follows. First we show a useful result from convex geometry, namely John's Theorem. Then we describe the strategy, called Exp2 with John's exploration, and prove its regret bound. Finally we show how this regret bound can be improved in the special case of the euclidean ball.

7.3.1. John's ellipsoid. John's theorem concerns the ellipsoid $\mathcal{E}$ of minimal volume enclosing a given convex set $\mathcal{K}$ (which we shall call John's ellipsoid of $\mathcal{K}$). Basically it states that $\mathcal{E}$ as many contact points with $\mathcal{K}$, and that those contact points are "nicely" distributed, that is they form almost an orthonormal basis.

Theorem 7.5. *Let $\mathcal{K} \subset \mathbb{R}^d$ be a convex set. If the ellipsoid $\mathcal{E}$ of minimal volume enclosing $\mathcal{K}$ is the unit ball in some norm derived from a scalar product $\langle \cdot, \cdot \rangle$, then there exists M (with $M \leq d(d+1)/2 + 1$) contact points $u_1, \dots, u_M$ between $\mathcal{E}$ and $\mathcal{K}$, and $q \in \Delta_M$, such that*

$$x = d \sum_{i=1}^{M} q(i) \langle x, u_i \rangle u_i, \forall x \in \mathbb{R}^d.$$

In fact John's theorem is a *if and only if*, but here we shall only need the direction stated in the theorem. Note also that the above theorem immediately gives a formula for the norm of a point:

$$\langle x, x \rangle = d \sum_{i=1}^{M} q(i) \langle x, u_i \rangle^2. \tag{7.3}$$

Proof. First note that it is enough to prove the statement for the standard scalar product $\langle x, y \rangle = x^T y$, since one can always rewrite everything in an orthonormal basis for $\langle \cdot, \cdot \rangle$.

We will prove that

$$\frac{I_d}{d} \in Conv\left(uu^T, u \text{ contact point between } \mathcal{K} \text{ and } \mathcal{E}\right).$$

This clearly implies the theorem by Carathéodory's Theorem.

Let us proceed by contradiction and assume that this is not the case. Then there exists a linear functional Φ on the space of $d \times d$ matrices such that:

$$\Phi\left(\frac{I_d}{d}\right) < \Phi\left(uu^T\right), \forall u \text{ contact point between } \mathcal{K} \text{ and } \mathcal{E}.$$

Now observe that Φ can be written as a $d \times d$ matrix $H = (h_{i,j})$ such that for any matrix $A = (a_{i,j})$:

$$\Phi(A) = \sum_{i,j} h_{i,j} a_{i,j}.$$

Since clearly $Tr\left(\frac{I_d}{d}\right) = 1$ and $Tr\left(uu^T\right) = 1$ (the only non-zero eigenvalue of uu^T is 1 because u is a unit vector since it lies on the surface on $\mathcal{E}$), one can remove a constant to the diagonal entries of H while not changing the inequality between $\Phi\left(\frac{I_d}{d}\right)$ and $\Phi\left(uu^T\right)$. Thus we can assume that $Tr(H) = 0$, or in other words that $\Phi\left(\frac{I_d}{d}\right) = 0$ (in particular we now have $u^T H u > 0$ for any contact point u). Finally note that $\frac{I_d}{d}$ and uu^T are symmetric matrices so we can also assume that H is symmetric (just consider $H + H^T$).

Now consider the following ellipsoid:

$$\mathcal{E}_\delta = \{x \in \mathbb{R}^d : x^T (I_d + \delta H)^{-1} x \leq 1\}.$$

First note that $\mathcal{E}_\delta$ tends to $\mathcal{E}$ when δ tends to 0, and that if u is a contact point between $\mathcal{E}$ and $\mathcal{K}$, then for δ small enough:

$$u^T (I_d + \delta H)^{-1} u = u^T (I_d - \delta H) u + o(\delta) = 1 - \delta u^T H u + o(\delta) < 1.$$

Thus, by continuity, one "clearly" has $\mathcal{K} \subset \mathcal{E}_\delta$. It remains to prove that $vol(\mathcal{E}_\delta) < vol(\mathcal{E})$. By definition this is equivalent to showing that the eigenvalues $\mu_1, \ldots, \mu_d$ of $(I_d + \delta H)$ satisfy $\prod_{i=1}^d \mu_i < 1$. This latter inequality is implied by the AM/GM inequality and the fact that $Tr(I_d + \delta H) = n = \sum_{i=1}^d \mu_i$. □

7.3.2. John's exploration. On the contrary to what happened in the semi-bandit case, here the randomization used to "play" points in $Conv(\mathcal{A})$ is not enough to build a good estimate of z_t. We propose here a perturbation based on John's ellipsoid for $Conv(\mathcal{A})$.

First we need to perform a preprocessing step:

- Find John's ellipsoid for $Conv(\mathcal{A})$: $\mathcal{E} = \{x \in \mathbb{R}^d : (x - x_0)^T H^{-1} (x - x_0) \leq 1\}$. The first preprocessing step is to translate everything by x_0. In other words we assume now that $\mathcal{A}$ is such that $x_0 = 0$.
- Consider the inner product: $\langle x, y \rangle = x^T H y$.
- We can now assume that we are playing on $\mathcal{A}' = H^{-1}\mathcal{A}$, and that $\ell(a', z) = \langle a', z \rangle$. Indeed: $\langle H^{-1} a, z \rangle = a^T z$. Moreover note that John's ellipsoid for $Conv(\mathcal{A}')$ is the unit ball for the inner product $\langle \cdot, \cdot \rangle$ (since $\langle H^{-1} x, H^{-1} x \rangle = x^T H^{-1} x$).
- Find the contact points $u_1, \ldots, u_M$ and $q \in \Delta_M$ that satisfy Theorem 7.5 for $Conv(\mathcal{A}')$. Note that the contact points are in $\mathcal{A}'$, thus they are valid points to play.

In the following we drop the prime on $\mathcal{A}'$. More precisely we play on a set $\mathcal{A}$ such that John's ellipsoid for $Conv(\mathcal{A})$ is the unit ball for some inner product $\langle \cdot, \cdot \rangle$, and the loss is $\ell(a, z) = \langle a, z \rangle$.

We can now describe John's exploration. Assume that we have a strategy that prescribes to play at random from probability distribution p_t. Then John's perturbation plays $\widetilde{a}_t \in \mathcal{A}$ as follows:

- with probability $1 - \gamma$, play a point at random from p_t,
- with probability $\gamma q(i)$, play u_i.

7.3.3. Exp2 with John's exploration. First we describe how to obtain an unbiased estimate of z_t, upon observing $\langle \widetilde{a}_t, z \rangle$ where $\widetilde{a}_t$ is drawn at random from a probability distribution $\widetilde{p}_t$ on $\mathcal{A}$ (with $\widetilde{p}_t(a) > 0$, $\forall a \in \mathcal{A}$). Note that we use $\widetilde{p}_t$ because it will correspond to a perturbation with John's exploration of some basic p_t.

Recall that the outer product $u \otimes u$ is defined as the linear mapping from $\mathbb{R}^d$ to $\mathbb{R}^d$ such that $u \otimes u(x) = \langle u, x \rangle u$. Note that one can also view $u \otimes u$ as a $d \times d$ matrix, so that the evaluation of $u \otimes u$ is equivalent to a multiplication by the corresponding matrix. Now let:

$$P_t = \sum_{a \in \mathcal{A}} \widetilde{p}_t(a) a \otimes a.$$

Note that this matrix is invertible, since $\mathcal{A}$ is of full rank and $\widetilde{p}_t(a) > 0$, $\forall a \in \mathcal{A}$. The estimate for z_t is given by:

$$\widetilde{g}_t = P_t^{-1} \left(\widetilde{a}_t \otimes \widetilde{a}_t \right) z_t. \tag{7.4}$$

Note that this is a valid estimate since $\left(\widetilde{a}_t \otimes \widetilde{a}_t \right) z_t = \langle \widetilde{a}_t, z_t \rangle \widetilde{a}_t$ and P_t^{-1} are observed quantities. Moreover it is also clearly an unbiased estimate.

Now Exp2 with John's exploration and estimate (7.4) corresponds to playing according to the following probability distribution:

$$\widetilde{p}_t(a) = (1 - \gamma) p_t(a) + \gamma \sum_{i=1}^{M} q_i \mathbb{1}_{a = u_i},$$

where

$$p_t(a) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \langle a, \widetilde{g}_t \rangle \right)}{\sum_{b \in \mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \langle b, \widetilde{g}_t \rangle \right)}.$$

Theorem 7.6. *Exp2 with John's exploration and estimate* (7.4) *satisfies for* $\frac{\eta G d}{\gamma} \leq 1$*:*

$$\overline{R}_n \leq 2 \gamma G n + \frac{\log N}{\eta} + \eta G^2 n d.$$

In particular with $\gamma = \eta G d$*, and* $\eta = \sqrt{\frac{\log N}{3ndG^2}}$*:*

$$\overline{R}_n \leq 2G \sqrt{3nd \log N}.$$

Note that for combinatorial optimization, $G = m$ and $\log N \leq m \log \frac{de}{m}$, thus the above result implies the bound:

$$\overline{R}_n \leq 2m^{3/2}\sqrt{3dn \log \frac{de}{m}}.$$

I conjecture that this bound is suboptimal, and that there exists an algorithm with a regret bound of order $m\sqrt{dn}$.

Proof. Using for instance Theorem 7.2 on Δ_N (with a slight modification to win in the constants), and the remarks after Theorem 5.1, one can easily show that, if:

$$\eta \langle a, \widetilde{g}_t \rangle \leq 1, \forall a \in \mathcal{A},$$

then

$$\begin{aligned}
\overline{R}_n &\leq 2\gamma G n + (1-\gamma)\left(\frac{\log N}{\eta} + \eta \mathbb{E} \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} p_t(a) \langle a, \widetilde{g}_t \rangle^2 \right) \\
&\leq 2\gamma G n + \frac{\log N}{\eta} + \eta \mathbb{E} \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} \widetilde{p}_t(a) \langle a, \widetilde{g}_t \rangle^2.
\end{aligned}$$

Thus it remains to bound $\langle a, \widetilde{g}_t \rangle$ and $\mathbb{E} \sum_{a \in \mathcal{A}} \widetilde{p}_t(a) \langle a, \widetilde{g}_t \rangle^2$. Let us start with the latter quantity:

$$\begin{aligned}
\sum_{a \in \mathcal{A}} \widetilde{p}_t(a) \langle a, \widetilde{g}_t \rangle^2 &= \sum_{a \in \mathcal{A}} \widetilde{p}_t(a) \langle \widetilde{g}_t, (a \otimes a) \widetilde{g}_t \rangle \\
&= \langle \widetilde{g}_t, P_t \widetilde{g}_t \rangle \\
&= \langle \widetilde{a}_t, z_t \rangle^2 \langle P_t^{-1} \widetilde{a}_t, P_t P_t^{-1} \widetilde{a}_t \rangle \\
&\leq G^2 \langle P_t^{-1} \widetilde{a}_t, \widetilde{a}_t \rangle.
\end{aligned}$$

Now we use a spectral decomposition of P_t in an orthonormal basis for $\langle \cdot, \cdot \rangle$ and write:

$$P_t = \sum_{i=1}^{d} \lambda_i v_i \otimes v_i.$$

In particular we have $P_t^{-1} = \sum_{i=1}^{d} \frac{1}{\lambda_i} v_i \otimes v_i$ and thus:

$$\begin{aligned}
\mathbb{E} \langle P_t^{-1} \widetilde{a}_t, \widetilde{a}_t \rangle &= \sum_{i=1}^{d} \frac{1}{\lambda_i} \mathbb{E} \langle (v_i \otimes v_i) \widetilde{a}_t, \widetilde{a}_t \rangle \\
&= \sum_{i=1}^{d} \frac{1}{\lambda_i} \mathbb{E} \langle (\widetilde{a}_t \otimes \widetilde{a}_t) v_i, v_i \rangle \\
&= \sum_{i=1}^{d} \frac{1}{\lambda_i} \langle P_t v_i, v_i \rangle \\
&= \sum_{i=1}^{d} \frac{1}{\lambda_i} \langle \lambda_i v_i, v_i \rangle \\
&= d.
\end{aligned}$$

This concludes the bound for $\mathbb{E}\sum_{a\in\mathcal{A}}\widetilde{p}_t(a)\langle a,\widetilde{g}_t\rangle^2$. We turn now to $\langle a,\widetilde{g}_t\rangle$:

$$\begin{aligned}\langle a,\widetilde{g}_t\rangle &= \langle\widetilde{a}_t,z_t\rangle\langle a,P_t^{-1}\widetilde{a}_t\rangle\\ &\leq G\langle a,P_t^{-1}\widetilde{a}_t\rangle\\ &\leq \frac{G}{\min_{1\leq i\leq d}\lambda_i},\end{aligned}$$

where the last inequality follows from the fact that $\langle a,a\rangle\leq 1$ for any $a\in\mathcal{A}$, since $\mathcal{A}$ is included in the unit ball. Now to conclude the proof we need to lower bound the smallest eigenvalue of P_t. This can be done as follows, using (7.3),

$$\begin{aligned}\min_{1\leq i\leq d}\lambda_i &= \min_{x\in\mathbb{R}^d:\langle x,x\rangle=1}\langle x,P_tx\rangle\\ &\geq \min_{x\in\mathbb{R}^d:\langle x,x\rangle=1}\gamma\sum_{i=1}^M\langle x,q(i)(u_i\otimes u_i)x\rangle\\ &= \min_{x\in\mathbb{R}^d:\langle x,x\rangle=1}\gamma\sum_{i=1}^M q(i)\langle x,u_i\rangle^2\\ &= \frac{\gamma}{d}.\end{aligned}$$

□

7.3.4. Improved strategies for specific action sets. We just saw that in general, under the bounded scalar loss assumption, one can obtain a regret bound of order $d\sqrt{n}$. As we will see in the next section, this is unimprovable in the sense that for some action set, one has a matching lower bound. However note that in Section 7.2.2 we proved that for the simplex, one can obtain a regret bound of order $\sqrt{dn}$ (we shall also prove that this is the optimal rate for this set in the next section). The key was to resort to OSMD, with a Legendre function adapted to the simplex (in some sense). Here we prove an improved regret bound for another action set: the euclidean ball. Unfortunately we get an extraneous logarithmic factor, we prove only a regret bound of order $\sqrt{dn\log n}$.

In the following $||\cdot||$ denotes the euclidean norm. We consider the online linear optimization problem on $\mathcal{A}=B_2=\{x\in\mathbb{R}^d:||x||\leq 1\}$. We perform the following perturbation of a point a_t in the interior of $\mathcal{A}$:

- Let $\xi_t\sim Ber(||a_t||)$, $I_t\sim unif(\{1,\dots,d\}$, and $\varepsilon_t\sim Rad$.
- If $\xi_t=1$, play $\widetilde{a}_t=\frac{a_t}{||a_t||}$,
- if $\xi_t=0$, play $\widetilde{a}_t=\varepsilon_t e_{I_t}$,

It is easy to check that this perturbation is unbiased, in the sense that:

$$\mathbb{E}(\widetilde{a}_t|a_t)=a_t. \tag{7.5}$$

We describe now how to build the unbiased estimate of the adversary's move:

$$\widetilde{g}_t=(1-\xi_t)\frac{d}{1-||a_t||}(z_t^T\widetilde{a}_t)\widetilde{a}_t. \tag{7.6}$$

Again it is easy to check that this is an unbiased estimate, that is:

$$\mathbb{E}(\widetilde{g}_t|a_t)=z_t \tag{7.7}$$

THEOREM 7.7. *Consider the online linear optimization problem on $\mathcal{A} = B_2$, and with $\mathcal{Z} = B_2$. Then OSMD on $\mathcal{A}' = \{x \in \mathbb{R}^d : ||x|| \leq 1 - \gamma\}$ with the estimate (7.6), $a_1 = 0$, and $F(x) = -\log(1 - ||x||) - ||x||$ satisfies for any η such that $\eta d \leq \frac{1}{2}$:*

$$\overline{R}_n \leq \gamma n + \frac{\log \gamma^{-1}}{\eta} + \eta \sum_{t=1}^{n} \mathbb{E}(1 - ||a_t||)||\widetilde{g}_t||^2. \tag{7.8}$$

In particular with $\gamma = \frac{1}{\sqrt{n}}$ and $\eta = \sqrt{\frac{\log n}{2nd}}$,

$$\overline{R}_n \leq 3\sqrt{dn \log n}. \tag{7.9}$$

PROOF. First it is clear that by playing on $\mathcal{A}'$ instead of $\mathcal{A} = B_2$, one incurs an extra γn regret. Second note that F is stricly convex (it is the composition of a convex and nondecreasing function with the euclidean norm) and:

$$\nabla F(x) = \frac{x}{1 - ||x||}, \tag{7.10}$$

in particular F is Legendre on $\mathcal{D} = \{x \in \mathbb{R}^d : ||x|| < 1\}$, and one has $\mathcal{D}^* = \mathbb{R}^d$, thus (5.4) is always satisfied and OSMD is well defined. Now the regret with respect to $\mathcal{A}'$ can be bounded as follows, thanks to Theorem 7.2, (7.5), and (7.7)

$$\frac{\sup_{a \in \mathcal{A}'} F(a) - F(a_1)}{\eta} + \frac{1}{\eta} \sum_{t=1}^{n} \mathbb{E} D_{F^*}\Big(\nabla F(a_t) - \eta \widetilde{g}_t, \nabla F(a_t)\Big).$$

The first term is clearly bounded by $\frac{\log \gamma^{-1}}{\eta}$. For the second term we need to do a few computations (the first one follows from (7.10)):

$$\begin{aligned}
\nabla F^*(u) &= \frac{u}{1 + ||u||}, \\
F^*(u) &= -\log(1 + ||u||) + ||u||, \\
D_{F^*}(u, v) &= \frac{1}{1 + ||v||}\left(||u|| - ||v|| + ||u|| \cdot ||v|| - v^T u - (1 + ||v||) \log\left(1 + \frac{||u|| - ||v||}{1 + ||v||}\right)\right).
\end{aligned}$$

Let $\Theta(u, v)$ such that $D_{F^*}(u, v) = \frac{1}{1+||v||}\Theta(u, v)$. First note that

$$\frac{1}{1 + ||\nabla F(a_t)||} = 1 - ||a_t||, \tag{7.11}$$

thus to prove (7.8) it remains to show that $\Theta(u, v) \leq ||u - v||^2$, for $u = \nabla F(a_t) - \eta \widetilde{g}_t$ and $v = \nabla F(a_t)$. In fact we shall prove that this inequality holds true as soon as $\frac{||u|| - ||v||}{1 + ||v||} \geq -\frac{1}{2}$. This is the case for the pair (u, v) under consideration, since by the triangle inequality, equations (7.6) and (7.11), and the assumption on η:

$$\frac{||u|| - ||v||}{1 + ||v||} \geq -\frac{\eta ||\widetilde{g}_t||}{1 + ||v||} \geq -\eta d \geq -\frac{1}{2}.$$

Now using that $\log(1+x) \geq x - x^2$, $\forall x \geq -\frac{1}{2}$, we obtain that for u, v such that $\frac{||u||-||v||}{1+||v||} \geq -\frac{1}{2}$,

$$\begin{aligned}
\Theta(u,v) &\leq \frac{(||u||-||v||)^2}{1+||v||} + ||u|| \cdot ||v|| - v^T u \\
&\leq (||u||-||v||)^2 + ||u|| \cdot ||v|| - v^T u \\
&= ||u||^2 + ||v||^2 - ||u|| \cdot ||v|| - v^T u \\
&= ||u-v||^2 + 2v^T u - ||u|| \cdot ||v|| - v^T u \\
&\leq ||u-v||^2,
\end{aligned}$$

which concludes the proof of (7.8). Now for the proof of (7.9) it suffices to note that:

$$\mathbb{E}(1-||a_t||)||\tilde{g}_t||^2 = (1-||a_t||)\sum_{i=1}^{d} \frac{1-||a_t||}{d} \frac{d^2}{(1-||a_t||)^2}(z_t^T e_i)^2 = d||z_t||^2 \leq d,$$

along with straightforward computations. □

7.4. Lower bounds

We prove here three lower bounds. First we consider online combinatorial optimization under both semi-bandit and bandit feedback. In the former case the lower bound matches the upper bound obtained in Section 7.2.2, while in the latter case there is a gap of $\sqrt{m \log \frac{de}{m}}$ as pointed out after Theorem 7.6. The third lower bound shows that Exp2 with John's exploration is essentialy optimal under the bounded scalar loss assumption, in the sense that for a given N, there exists a set $\mathcal{A}$ with $|\mathcal{A}| \leq N$ and such that the regret bound of Exp2 is unimprovable (up to a logarithmic factor).

Theorem 7.8. *Let $n \geq d \geq 2m$. There exists a subset $\mathcal{A} \subset \{0,1\}^d$ such that $||v||_1 = m, \forall v \in \mathcal{C}$, and for any strategy, under semi-bandit feedback:*

$$\sup_{adversaries\ s.t.\ z_t \in [0,1]^d} \overline{R}_n \geq 0.02\sqrt{mdn}, \tag{7.12}$$

and under bandit feedback:

$$\sup_{adversaries\ s.t.\ z_t \in [0,1]^d} \overline{R}_n \geq 0.02m\sqrt{dn}. \tag{7.13}$$

Moreover it also holds that $|\mathcal{A}| = (\lfloor d/m \rfloor)^m$, and for any strategy, under bandit feedback:

$$\sup_{adversaries\ s.t.\ |z_t^T a| \leq 1, \forall a \in \mathcal{A}} \overline{R}_n \geq 0.03\sqrt{mdn} \tag{7.14}$$

Proof. For sake of notation we assume here that d is a multiple of m, and we identify $\{0,1\}^d$ with $\{0,1\}^{m \times \frac{d}{m}}$. We consider the following set of actions:

$$\mathcal{A} = \{a \in \{0,1\}^{m \times \frac{d}{m}} : \forall i \in \{1,\ldots,m\}, \sum_{j=1}^{d/m} a(i,j) = 1\}.$$

In other words the player is playing in parallel m finite games with d/m actions.

We divide the proofs in five steps. From step 1 to 4 we restrict our attention to the bandit case, with $z_t \in [0,1]^{m \times \frac{d}{m}}$. Then in step 5 we show how to easily apply the same proof technique for semi-bandit and for bandit with bounded scalar loss. Moreover from step 1 to 3 we restrict our attention to the case of deterministic strategies for the player, and we show how to extend the results to arbitrary strategies in step 4.

First step: definitions.

We denote by $I_{i,t} \in \{1, \dots, m\}$ the random variable such that $a_t(i, I_{i,t}) = 1$. That is, $I_{i,t}$ is the action chosen at time t in the i^{th} game. Moreover let τ be drawn uniformly at random in $\{1, \dots, n\}$.

In this proof we consider random adversaries indexed by $\mathcal{A}$. More precisely, for $\alpha \in \mathcal{A}$, we define the α-adversary as follows: For any $t \in \{1, \dots, n\}$, $z_t(i, j)$ is drawn from a Bernoulli distribution with parameter $\frac{1}{2} - \varepsilon\alpha(i, j)$. In other words, against adversary α, in the i^{th} game, the action j such that $\alpha(i, j) = 1$ has a loss slightly smaller (in expectation) than the other actions. We note $\mathbb{E}_\alpha$ when we integrate with respect to the loss generation process of the α-adversary. We note $\mathbb{P}_{i,\alpha}$ the law of $\alpha(i, I_{i,\tau})$ when the player faces the α-adversary. Remark that we have $\mathbb{P}_{i,\alpha}(1) = \mathbb{E}_\alpha \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}_{\alpha(i,I_{i,t})=1}$, hence, against the α-adversary we have:

$$\overline{R}_n = \mathbb{E}_\alpha \sum_{t=1}^{n} \sum_{i=1}^{m} \varepsilon \mathbb{1}_{\alpha(i,I_{i,t}) \neq 1} = n\varepsilon \sum_{i=1}^{m} \left(1 - \mathbb{P}_{i,\alpha}(1)\right),$$

which implies (since the maximum is larger than the mean)

$$\max_{\alpha \in \mathcal{A}} \overline{R}_n \geq n\varepsilon \sum_{i=1}^{m} \left(1 - \frac{1}{(d/m)^m} \sum_{\alpha \in \mathcal{A}} \mathbb{P}_{i,\alpha}(1) \right). \tag{7.15}$$

Second step: information inequality.

Let $\mathbb{P}_{-i,\alpha}$ be the law of $\alpha(i, I_{i,\tau})$ against the adversary which plays like the α-adversary except that in the i^{th} game, the losses of all coordinates are drawn from a Bernoulli of parameter $1/2$ (we call it the $(-i, \alpha)$-adversary and we note $\mathbb{E}_{(-i,\alpha)}$ when we integrate with respect to its loss generation process). Now we use Pinsker's inequality (see Lemma 7.2 below) which gives:

$$\mathbb{P}_{i,\alpha}(1) \leq \mathbb{P}_{-i,\alpha}(1) + \sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}_{-i,\alpha}, \mathbb{P}_{i,\alpha})}.$$

Moreover note that by symmetry of the adversaries $(-i, \alpha)$,

$$\begin{aligned}
\frac{1}{(d/m)^m} \sum_{\alpha \in \mathcal{A}} \mathbb{P}_{-i,\alpha}(1) &= \frac{1}{(d/m)^m} \sum_{\alpha \in \mathcal{A}} \mathbb{E}_{(-i,\alpha)} \alpha(i, I_{i,\tau}) \\
&= \frac{1}{(d/m)^m} \sum_{\beta \in \mathcal{A}} \frac{1}{d/m} \sum_{\alpha : (-i,\alpha) = (-i,\beta)} \mathbb{E}_{(-i,\alpha)} \alpha(i, I_{i,\tau}) \\
&= \frac{1}{(d/m)^m} \sum_{\beta \in \mathcal{A}} \frac{1}{d/m} \mathbb{E}_{(-i,\beta)} \sum_{\alpha : (-i,\alpha) = (-i,\beta)} \alpha(i, I_{i,\tau}) \\
&= \frac{1}{(d/m)^m} \sum_{\beta \in \mathcal{A}} \frac{1}{d/m} \\
&= \frac{m}{d}, \qquad (7.16)
\end{aligned}$$

and thus, thanks to the concavity of the square root,

$$\frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathbb{P}_{i,\alpha}(1)\leq\frac{m}{d}+\sqrt{\frac{1}{2(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathrm{KL}(\mathbb{P}_{-i,\alpha},\mathbb{P}_{i,\alpha})}. \tag{7.17}$$

Third step: computation of $\mathrm{KL}(\mathbb{P}_{-i,\alpha},\mathbb{P}_{i,\alpha})$ *with the chain rule for Kullback-Leibler divergence.*

Note that since the forecaster is deterministic, the sequence of observed losses (up to time n) $W_n\in\{0,\dots,m\}^n$ uniquely determines the empirical distribution of plays, and in particular the law of $\alpha(i,I_{i,\tau})$ conditionally to W_n is the same for any adversary. Thus, if we note $\mathbb{P}^n_\alpha$ (respectively $\mathbb{P}^n_{-i,\alpha}$) the law of W_n when the forecaster plays against the α-adversary (respectively the $(-i,\alpha)$-adversary), then one can easily prove that $\mathrm{KL}(\mathbb{P}_{-i,\alpha},\mathbb{P}_{i,\alpha})\leq\mathrm{KL}(\mathbb{P}^n_{-i,\alpha},\mathbb{P}^n_\alpha)$. Now we use the chain rule for Kullback-Leibler divergence iteratively to introduce the laws $\mathbb{P}^t_\alpha$ of the observed losses W_t up to time t. More precisely, we have,

$$\begin{aligned}
&\mathrm{KL}(\mathbb{P}^n_{-i,\alpha},\mathbb{P}^n_\alpha)\\
&=\mathrm{KL}(\mathbb{P}^1_{-i,\alpha},\mathbb{P}^1_\alpha)+\sum_{t=2}^{n}\sum_{w_{t-1}\in\{0,\dots,m\}^{t-1}}\mathbb{P}^{t-1}_{-i,\alpha}(w_{t-1})\mathrm{KL}(\mathbb{P}^t_{-i,\alpha}(.|w_{t-1}),\mathbb{P}^t_\alpha(.|w_{t-1}))\\
&=\mathrm{KL}\left(\mathcal{B}_\emptyset,\mathcal{B}'_\emptyset\right)\mathbb{1}_{\alpha(i,I_{i,1})=1}+\sum_{t=2}^{n}\sum_{w_{t-1}:\alpha(i,I_{i,1})=1}\mathbb{P}^{t-1}_{-i,\alpha}(w_{t-1})\mathrm{KL}\left(\mathcal{B}_{w_{t-1}},\mathcal{B}'_{w_{t-1}}\right),
\end{aligned}$$

where $\mathcal{B}_{w_{t-1}}$ and $\mathcal{B}'_{w_{t-1}}$ are sums of m Bernoulli distributions with parameters in $\{1/2,1/2-\varepsilon\}$ and such that the number of Bernoullis with parameter $1/2$ in $\mathcal{B}_{w_{t-1}}$ is equal to the number of Bernoullis with parameter $1/2$ in $\mathcal{B}'_{w_{t-1}}$ plus one. Now using Lemma 7.3 (see below) we obtain,

$$\mathrm{KL}\left(\mathcal{B}_{w_{t-1}},\mathcal{B}'_{w_{t-1}}\right)\leq\frac{8\,\varepsilon^2}{(1-4\varepsilon^2)m}.$$

In particular this gives:

$$\mathrm{KL}(\mathbb{P}^n_{-i,\alpha},\mathbb{P}^n_\alpha)\leq\frac{8\,\varepsilon^2}{(1-4\varepsilon^2)m}\mathbb{E}_{-i,\alpha}\sum_{t=1}^{n}\mathbb{1}_{\alpha(i,I_{i,t})=1}=\frac{8\,\varepsilon^2 n}{(1-4\varepsilon^2)m}\mathbb{P}_{-i,\alpha}(1).$$

Summing and plugging this into (7.17) we obtain (again thanks to (7.16)), for $\varepsilon\leq\frac{1}{\sqrt{8}}$,

$$\frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathbb{P}_{i,\alpha}(1)\leq\frac{m}{d}+\varepsilon\sqrt{\frac{8n}{d}}.$$

To conclude the proof of (7.13) for deterministic players one needs to plug in this last equation in (7.15) along with straightforward computations.

Fourth step: Fubini's Theorem to handle non-deterministic players.

Consider now a randomized player, and let $\mathbb{E}_{rand}$ denote the expectation with respect to the randomization of the player. Then one has (thanks to Fubini's

Theorem),

$$\frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathbb{E}\sum_{t=1}^{n}(a_t^Tz_t-\alpha^Tz)=\mathbb{E}_{rand}\frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathbb{E}_\alpha\sum_{t=1}^{n}(a_t^Tz_t-\alpha^Tz).$$

Now remark that if we fix the realization of the forecaster's randomization then the results of the previous steps apply and in particular one can lower bound $\frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathbb{E}_\alpha\sum_{t=1}^{n}(a_t^Tz_t-\alpha^Tz)$ as before (note that α is the optimal action in expectation against the α-adversary).

Fifth step: Proof for semi-bandit, and bandit with bounded scalar loss.

The proof of (7.12) follows trivially from (7.13), using the same argument than in Theorem 6.5. On the other hand to prove (7.14) we need to work a little bit more. First we need to modifiy the α-adversary so that it satisfies the bounded scalar loss assumption. We do that as follows: at each turn the α-adversary selects uniformly at random $E_t\in\{1,\dots,m\}$, and sets to 0 the losses in all games but the E_t^{th} game where it sets the same losses than the original α-adversary described in the First step above. For this new set of adversaries, one has to do only two modifications in the above proof. First (7.15) is replaced by:

$$\max_{\alpha\in\mathcal{A}}\overline{R}_n\geq\frac{n\varepsilon}{m}\sum_{i=1}^{m}\left(1-\frac{1}{(d/m)^m}\sum_{\alpha\in\mathcal{A}}\mathbb{P}_{i,\alpha}(1)\right).$$

Second $\mathcal{B}_{w_{t-1}}$ is now a Bernoulli with mean $\mu_t\in\left[\frac{1}{2}-(m-1)\frac{\varepsilon}{m},\frac{1}{2}\right]$ and $\mathcal{B}'_{w_{t-1}}$ is a Bernoulli with mean $\mu_t-\frac{\varepsilon}{m}$, and thus we have (thanks to Lemma 7.2)

$$\mathrm{KL}\left(\mathcal{B}_{w_{t-1}},\mathcal{B}'_{w_{t-1}}\right)\leq\frac{4\varepsilon^2}{(1-4\varepsilon^2)m^2}.$$

The proof of (7.14) for deterministic players is then concluded again with straightforward computations. □

Lemma 7.2. *For any $p,q\in[0,1]$,*

$$2(p-q)^2\leq\mathrm{KL}(Ber(p),Ber(q))\leq\frac{(p-q)^2}{q(1-q)}.$$

Proof. The left hand side inequality is simply Lemma 5.10 (Pinsker's inequality) for Bernoulli distributions. The right hand side on the other hand comes from $\log x\leq x-1$ and the following computations:

$$\begin{aligned}\mathrm{KL}(Ber(p),Ber(q)) &= p\log\left(\frac{p}{q}\right)+(1-p)\log\left(\frac{1-p}{1-q}\right)\\ &\leq p\frac{p-q}{q}+(1-p)\frac{q-p}{1-q}\\ &= \frac{(p-q)^2}{q(1-q)}.\end{aligned}$$

□

Lemma 7.3. *Let ℓ and n be integers with $\frac{1}{2}\leq\frac{n}{2}\leq\ell\leq n$. Let $p,p',q,p_1,\dots,p_n$ be real numbers in $(0,1)$ with $q\in\{p,p'\}$, $p_1=\cdots=p_\ell=q$ and $p_{\ell+1}=\cdots=p_n$.*

Let $\mathcal{B}$ (resp. $\mathcal{B}'$) be the sum of $n+1$ independent Bernoulli distributions with parameters $p, p_1, \dots, p_n$ (resp. $p', p_1, \dots, p_n$). We have

$$\mathrm{KL}(\mathcal{B}, \mathcal{B}') \le \frac{2(p'-p)^2}{(1-p')(n+2)q}.$$

Proof. Let $Z, Z', Z_1, \dots, Z_n$ be independent Bernoulli distributions with parameters $p, p', p_1, \dots, p_n$. Define $S = \sum_{i=1}^{\ell} Z_i$, $T = \sum_{i=\ell+1}^{n} Z_i$ and $V = Z + S$. By slight abuse of notation, merging in the same notation the distribution and the random variable, we have

$$\begin{aligned}\mathrm{KL}(\mathcal{B}, \mathcal{B}') &= \mathrm{KL}\big((Z+S)+T, (Z'+S)+T\big)\\ &\le \mathrm{KL}\big((Z+S,T),(Z'+S,T)\big)\\ &= \mathrm{KL}\big(Z+S, Z'+S\big).\end{aligned}$$

Let $s_k = \mathbb{P}(S=k)$ for $k = -1, 0, \dots, \ell+1$. Using the equalities

$$s_k = \binom{\ell}{k} q^k (1-q)^{\ell-k} = \frac{q}{1-q}\frac{\ell-k+1}{k}\binom{\ell}{k-1} q^{k-1}(1-q)^{\ell-k+1} = \frac{q}{1-q}\frac{\ell-k+1}{k} s_{k-1},$$

which holds for $1 \le k \le \ell+1$, we obtain

$$\begin{aligned}\mathrm{KL}(Z+S, Z'+S) &= \sum_{k=0}^{\ell+1} \mathbb{P}(V=k) \log\left(\frac{\mathbb{P}(Z+S=k)}{\mathbb{P}(Z'+S=k)}\right)\\ &= \sum_{k=0}^{\ell+1} \mathbb{P}(V=k) \log\left(\frac{p s_{k-1} + (1-p)s_k}{p' s_{k-1} + (1-p')s_k}\right)\\ &= \sum_{k=0}^{\ell+1} \mathbb{P}(V=k) \log\left(\frac{p\frac{1-q}{q}k + (1-p)(\ell-k+1)}{p'\frac{1-q}{q}k + (1-p')(\ell-k+1)}\right)\\ &= \mathbb{E}\log\left(\frac{(p-q)V + (1-p)q(\ell+1)}{(p'-q)V + (1-p')q(\ell+1)}\right). \qquad (7.18)\end{aligned}$$

First case: $q = p'$.
By Jensen's inequality, using that $\mathbb{E}V = p'(\ell+1) + p - p'$ in this case, we then get

$$\begin{aligned}\mathrm{KL}(Z+S, Z'+S) &\le \log\left(\frac{(p-p')\mathbb{E}(V) + (1-p)p'(\ell+1)}{(1-p')p'(\ell+1)}\right)\\ &= \log\left(\frac{(p-p')^2 + (1-p')p'(\ell+1)}{(1-p')p'(\ell+1)}\right)\\ &= \log\left(1 + \frac{(p-p')^2}{(1-p')p'(\ell+1)}\right) \le \frac{(p-p')^2}{(1-p')p'(\ell+1)}.\end{aligned}$$

Second case: $q = p$.
In this case, V is a binomial distribution with parameters $\ell+1$ and p. From (7.18), we have

$$\begin{aligned}\mathrm{KL}(Z+S, Z'+S) &\le -\mathbb{E}\log\left(\frac{(p'-p)V + (1-p')p(\ell+1)}{(1-p)p(\ell+1)}\right)\\ &\le -\mathbb{E}\log\left(1 + \frac{(p'-p)(V-\mathbb{E}V)}{(1-p)p(\ell+1)}\right). \qquad (7.19)\end{aligned}$$

To conclude, we will use the following lemma.

Lemma 7.4. *The following inequality holds for any $x \geq x_0$ with $x_0 \in (0,1)$:*

$$-\log(x) \leq -(x-1) + \frac{(x-1)^2}{2x_0}.$$

Proof. Introduce $f(x) = -(x-1) + \frac{(x-1)^2}{2x_0} + \log(x)$. We have $f'(x) = -1 + \frac{x-1}{x_0} + \frac{1}{x}$, and $f''(x) = \frac{1}{x_0} - \frac{1}{x^2}$. From $f'(x_0) = 0$, we get that f' is negative on $(x_0, 1)$ and positive on $(1, +\infty)$. This leads to f nonnegative on $[x_0, +\infty)$. □

Finally, from Lemma 7.4 and (7.19), using $x_0 = \frac{1-p'}{1-p}$, we obtain

$$\begin{aligned}\mathrm{KL}(Z+S, Z'+S) &\leq \left(\frac{p'-p}{(1-p)p(\ell+1)}\right)^2 \frac{\mathbb{E}[(V-\mathbb{E}V)^2]}{2x_0} \\ &= \left(\frac{p'-p}{(1-p)p(\ell+1)}\right)^2 \frac{(\ell+1)p(1-p)^2}{2(1-p')} \\ &= \frac{(p'-p)^2}{2(1-p')(\ell+1)p}.\end{aligned}$$

□

7.5. Two-points bandit feedback

In this section we consider the general online convex optimization problem ($\mathcal{A}$ is a convex and compact set of full rank, and ℓ is convex) with the following modifications:

- The player chooses two points a_t and b_t and suffers the average loss $\frac{\ell(a_t,z_t)+\ell(b_t,z_t)}{2}$.
- The feedback is the bandit feedback for both points, that is the player observes $\ell(a_t, z_t)$ and $\ell(b_t, z_t)$.

Another interpretation of this scenario is that the adversary can update its move z_t only on odd rounds.

As one can expect, the fact that one observes two points allows for a good estimation of the gradient information. More precisely, assume that one wants to play a_t and to obtain the gradient $\nabla \ell(a_t, z_t)$. Also assume for the moment that there is an euclidean ball of radius γ around a_t that is contained in $\mathcal{A}$. Then one can play the following two perturbated points: $\widetilde{a}_t = a_t + \gamma v$ and $\widetilde{b}_t = a_t - \gamma v$, where v is drawn uniformly at random on $\mathbb{S}^{d-1}$. Using the feedback information one can build the following estimate of the gradient:

$$\widetilde{g}_t = d\frac{\ell(a_t+\gamma v, z_t) - \ell(a_t - \gamma v, z_t)}{2\gamma} v. \tag{7.20}$$

Note that, if $||\nabla \ell(a,z)||_2 \leq G$, then $||\widetilde{g}_t||_2 \leq dG$ (in other words the estimate can not be too big). Moreover the following lemma shows that it is an unbiased estimate of the gradient of a smoothed version of ℓ.

Lemma 7.5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function,*

$$\bar{f}(x) = \frac{1}{Vol_d(B_{2,d})} \int_{B_{2,d}} f(x+\gamma u) du,$$

and σ_{d-1} be the unnormalized spherical measure. Then $\bar{f}$ is differentiable and:

$$\nabla \bar{f}(x) = \frac{d}{\gamma \sigma_{d-1}(\mathbb{S}^{d-1})} \int_{\mathbb{S}^{d-1}} f(x+\gamma v) v d\sigma_{d-1}(v).$$

Proof. The proof of this result is an easy consequence of the Divergence Theorem and the fact that $\frac{Vol_d(B_{2,d})}{\sigma_{d-1}(\mathbb{S}^{d-1})} = \frac{1}{d}$. Indeed we have:

$$\begin{aligned} \nabla \bar{f}(x) &= \frac{1}{Vol_d(B_{2,d})} \int_{B_{2,d}} \nabla f(x+\gamma u) du \\ &= \frac{1}{Vol_d(B_{2,d})} \int_{\mathbb{S}^{d-1}} \frac{1}{\gamma} f(x+\gamma v) v d\sigma_{d-1}(v) \\ &= \frac{d}{\gamma \sigma_{d-1}(\mathbb{S}^{d-1})} \int_{\mathbb{S}^{d-1}} f(x+\gamma v) v d\sigma_{d-1}(v). \end{aligned}$$

□

Thus, if one defines (for $a \in \mathcal{A}$ such that the euclidean ball of radius γ around a is contained in $\mathcal{A}$)

$$\bar{\ell}(a,z) = \frac{1}{Vol_d(B_{2,d})} \int_{B_{2,d}} \ell(a+\gamma u, z) du,$$

then $\mathbb{E}(\widetilde{g}_t | a_t) = \nabla \bar{\ell}(a_t, z_t)$.

Moreover note that bounding the regret with respect to $\bar{\ell}$ directly gives regret for ℓ, since $|\bar{\ell}(a,z) - \ell(a,z)| \leq \gamma G$ (if $||\nabla \ell(a,z)||_2 \leq G$).

Given those properties, it is now easy to derive a regret bound for OGD with this unbiased estimate.

Theorem 7.9. *Consider a compact, convex, and full rank action set $\mathcal{A} \subset B_{2,d}(R)$, and a convex and differentiable loss ℓ with $||\nabla \ell(a,z)||_2 \leq G, \forall (a,z) \in \mathcal{A} \times \mathcal{Z}$. Then the two-points OSMD on $\mathcal{A}' = \{a \in \mathcal{A} : B_2(a,\gamma) \subset \mathcal{A}\}$ with $F(x) = \frac{1}{2}||x||_2^2$ satisfies*

$$\overline{R}_n \leq 4\gamma n G + \frac{R^2}{2\eta} + \frac{1}{2}\eta n d^2 G^2.$$

In particular with $\gamma \leq \frac{1}{4nG}$ and $\eta = \frac{R}{dG\sqrt{n}}$ we get

$$\overline{R}_n \leq RGd\sqrt{n} + 1.$$

Proof. First note that $\mathcal{A}'$ is clearly convex, and since $\mathcal{A}$ is of full rank, playing on $\mathcal{A}'$ instead of $\mathcal{A}$ only cost a regret of $\gamma n G$. Moreover we can bound the regret with respect to $\bar{\ell}$ instead of ℓ, which cost a regret of $2\gamma n G$. Now we can apply Theorem 7.1 and we directly obtain the regret bound above. □

Note that now one can play around with this theorem, and basically redo Chapter 5 under the assumption of two-points bandit feedback.

7.6. Online convex optimization with bandit feedback

The strategy described in the previous section can be directly applied to the bandit case. The only difference will be in the norm of the estimate. Indeed in that case one has:

$$\widetilde{g}_t = \frac{d}{\gamma} \ell(a_t + \gamma v, z_t),$$

and thus $||\widetilde{g}_t||_2 \leq \frac{d}{\gamma}$ (assuming that the loss takes values in $[-1, 1]$). Thus, on the contrary to what happened with two-points bandit feedback, here taking a small γ comes at a cost. In particular the regret bound for OGD with this estimate looks like (up to terms independent of n):

$$\gamma n + \frac{1}{\eta} + \frac{\eta n}{\gamma^2}.$$

Optimizing this bound gives a regret of order $n^{3/4}$. Of course the fact that we do not get a $\sqrt{n}$ should not be a surprise, since OGD does not use the local norm idea that was key to obtain good regret in the linear bandit case. However here this local norm idea can not be applied: indeed in linear bandit the variance of the estimate blows up only when one approaches the boundary of $\mathcal{A}$, while in the convex case the variance 'blows' up everywhere! As a result the $n^{3/4}$ regret bound is the best known bound for online convex optimization with bandit feedback. I see no reason why this bound would be optimal, and I conjecture that $\sqrt{n}$ is attainable in this case too. However it seems that fundamentally new ideas are required to attain this bound.

References

The study of bandit feedback in online (finite) optimization was initiated by:

- P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2003

An optimal strategy for online finite optimization was first given in:

- J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009

This latter strategy was first generalized in [4], and then in [5] (in the form given in Section 7.2.2).

Section 7.2 is based on:

- J.-Y. Audibert, S. Bubeck, and G. Lugosi. Minimax policies for combinatorial prediction games. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, 2011

In fact, several papers considered specific examples of online combinatorial optimization with semi-bandit feedback before the above paper: [23] for paths, [34] for matchings, and [52] for m-sets.

Section 7.3 is based on:

- V. Dani, T. Hayes, and S. Kakade. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, pages 345–352, 2008
- N. Cesa-Bianchi and S. Kakade. An optimal algorithm for linear bandits. arXiv:1110.4322v1, 2011
- K. Ball. An elementary introduction to modern convex geometry. In S. Levy, editor, *Flavors of Geometry*, pages 1–58. Cambridge University Press, 1997

Note that [16] specifically studied the case of online combinatorial optimization with bandit feedback, with Exp2 and a uniform exploration over $\mathcal{C}$. On the other hand the only paper that attains a $\sqrt{n}$ regret bound with OSMD is:

- J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pages 263–274, 2008

They achieve this with F being a self-concordant barrier for $\mathcal{A}$. Unfortunately this approach yields suboptimal bounds in terms of the dimension. However in some cases this approach is computationally efficient, while the discretization with Exp2 is not. Note that another useful reference for convex geometry, and in particular algorithmic convex geometry is [54].

Section 7.5 is based on:

- A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, 2010

Section 7.6 is based on:

- A. Flaxman, A. Kalai, and B. McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *In Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 385–394, 2005

Note also that two earlier references for online linear optimization with bandit feedback are:

- B. Awerbuch and R. Kleinberg. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. In *STOC '04: Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 45–53, 2004
- H. McMahan and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *In Proceedings of the 17th Annual Conference on Learning Theory (COLT)*, pages 109–123, 2004

Finally note that SGD goes back:

- H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951
- J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:462–466, 1952

See [51, 9] for a modern point view on SGD in a standard framework.

CHAPTER 8

Online stochastic optimization

In this section we assume that the sequence $z_1, \dots, z_n$ is i.i.d from some unknown probability distribution $\mathbb{P}$ over $\mathcal{Z}$. We restrict our attention to the bandit case, since the full information problem basically reduces to a classical statistical problem. Note that here the pseudo-regret can be defined as follows (since z_t is independent of a_t):

$$\begin{aligned} \overline{R}_n &= \mathbb{E} \sum_{t=1}^{n} \ell(a_t, z_t) - \min_{a \in \mathcal{A}} \mathbb{E} \sum_{t=1}^{n} \ell(a, z_t) \\ &= \mathbb{E} \sum_{t=1}^{n} \mathbb{E}_{Z \sim \mathbb{P}} \ell(a_t, Z) - \min_{a \in \mathcal{A}} \sum_{t=1}^{n} \mathbb{E}_{Z \sim \mathbb{P}} \ell(a, Z). \end{aligned}$$

In other words to minimize the pseudo-regret one should find the action with smallest *expected loss* and focus on this one once found. Note that the player faces an *exploration-exploitation dilemna*, since he has to choose between *exploiting* his knowledge to focus on the action that he believes to be the best, and *exploring* further the action set to identify with better precision which action is the best. As we shall see, the key is basically to not make this choice but do both exploration and exploitation simultaneously!

Note that, in terms of *distribution-free* inequalities, one cannot improve the bounds on $\overline{R}_n$ proved in the previous chapter. Indeed all lower bounds considered a stochastic adversary. Here we shall focus on *distribution-dependent* bounds, that is bounds on $\overline{R}_n$ that depend on $\mathbb{P}$. In that case one can propose tremendous improvements over the results of the previous chapter.

8.1. Optimism in face of uncertainty

Assume that one observed $\ell(a_1, z_1), \dots, \ell(a_t, z_t)$. The optimism in face of uncertainty corresponds to the following heuristic to choose the next action a_{t+1}. First, using the observed data, one builds a set $\mathcal{P}$ of probability distributions which are 'consistent' with the data. More precisely, given a set of possible probability distributions, and a threshold δ, one excludes all $\mathbb{P}$ such that:

$$\mathbb{P}(\text{observing } \ell(a_1, z_1), \dots, \ell(a_t, z_t)) < 1 - \delta.$$

Then the optimism in face of uncertainty says that one should play the optimal action for the 'best' environment in $\mathcal{P}$, that is:

$$a_{t+1} \in \operatorname*{argmin}_{a \in \mathcal{A}} \min_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{Z \sim \mathbb{P}} \ell(a, z).$$

This heuristic is a way to do both exploration and exploitation at the same time.

8.2. Stochastic multi-armed bandit

The stochastic multi-armed bandit corresponds to the case where $\mathcal{A}$ is a finite set. For historical reasons we consider gains rather than losses in this setting, and we modify the notation as follows. Let $d \geq 2$ be the number of actions (we call them arms from now on). For $i \in \{1, \dots, d\}$, let ν_i be the reward distribution of arm i and μ_i its mean. That is when one pulls arm i, one receives a reward drawn from ν_i (independently from the past). We also set $\mu^* = \max_{i \in \{1,\dots,d\}} \mu_i$ and $i^* \in \operatorname{argmax}_{i \in \{1,\dots,d\}} \mu_i$. Denote $I_t \in \{1, \dots, d\}$ the arm played at time t. Then the pseudo-regret is defined as:

$$\overline{R}_n = n\mu^* - \mathbb{E} \sum_{t=1}^{n} \mu_{I_t}.$$

Another form of the pseudo-regret shall be useful. Let $T_i(s) = \sum_{t=1}^{s} \mathbb{1}_{I_t = i}$ denote the number of times the player selected arm i on the first s rounds. Let $\Delta_i = \mu^* - \mu_i$ be the suboptimality parameter of arm i. Then the pseudo-regret can be written as:

$$\begin{aligned} \overline{R}_n &= n\mu^* - \mathbb{E} \sum_{t=1}^{n} \mu_{I_t} \\ &= \left(\sum_{i=1}^{d} \mathbb{E} T_i(n) \right) \mu^* - \mathbb{E} \sum_{i=1}^{d} T_i(n) \mu_i \\ &= \sum_{i=1}^{d} \Delta_i \mathbb{E} T_i(n). \end{aligned}$$

Finally we denote by $X_{i,s}$ the reward obtained by pulling arm i for the s^{th} time.

8.2.1. Upper Confidence Bounds (UCB).

Let $\psi : \mathbb{R}_+ \to \mathbb{R}$ be a convex function[1]. In this section we assume that the reward distributions satisfy the following conditions: For all $\lambda \geq 0$,

$$\log \mathbb{E} \exp(\lambda (X - \mathbb{E}X)) \leq \psi(\lambda) \text{ and } \log \mathbb{E} \exp(\lambda ((\mathbb{E}X) - X)) \leq \psi(\lambda). \tag{8.1}$$

We shall attack the stochastic multi-armed bandit with the optimism in face of uncertainty principle. To do so one needs to provide an upper bound estimate on the mean of each arm at some probability level. Given the assumption (8.1) it is easy to do so. Let $\widehat{\mu}_{i,s} = \frac{1}{s} \sum_{t=1}^{s} X_{i,t}$. Using Markov's inequality one obtains:

$$\mathbb{P}(\mu_i - \widehat{\mu}_{i,s} > \varepsilon) \leq \exp(-s\psi^*(\varepsilon)). \tag{8.2}$$

In other words, with probability at least $1 - \delta$,

$$\widehat{\mu}_{i,s} + (\psi^*)^{-1} \left(\frac{\log \delta^{-1}}{s} \right) > \mu_i.$$

Thus we consider the following strategy, called (α, ψ)-UCB (with $\alpha > 0$): at time t, select

$$I_t \in \operatorname*{argmax}_{i \in \{1,\dots,d\}} \widehat{\mu}_{i,T_i(t-1)} + (\psi^*)^{-1} \left(\frac{\alpha \log t}{T_i(t-1)} \right)$$

[1] One can easily generalize the discussion to functions ψ defined only on an interval $[0, b)$.

THEOREM 8.1. *Assume that the reward distributions satisfy* (8.1). *Then* (α, ψ)*-UCB with* $\alpha > 2$ *satisfies:*

$$\overline{R}_n \leq \sum_{i:\Delta_i>0} \left(\frac{\alpha\Delta_i}{\psi^*(\Delta_i/2)} \log n + 1 + \frac{2}{\alpha-2} \right).$$

For example for bounded random variables in $[0,1]$, thanks to Hoeffding's inequality one can take $\psi(\lambda) = \frac{\lambda^2}{8}$, which gives $\psi^*(\varepsilon) = 2\varepsilon^2$, and which in turns gives the following regret bound:

$$\overline{R}_n \leq \sum_{i:\Delta_i>0} \left(\frac{2\alpha}{\Delta_i} \log n + 1 + \frac{2}{\alpha-2} \right).$$

PROOF. First note that if $I_t = i$, then one the three following equations is true:

$$\widehat{\mu}_{i^*,T_{i^*}(t-1)} + (\psi^*)^{-1}\left(\frac{\alpha \log t}{T_{i^*}(t-1)}\right) \leq \mu^*, \tag{8.3}$$

or

$$\widehat{\mu}_{i,T_i(t-1)} > \mu_i + (\psi^*)^{-1}\left(\frac{\alpha \log t}{T_i(t-1)}\right), \tag{8.4}$$

or

$$T_i(t-1) < \frac{\alpha \log n}{\psi^*(\Delta_i/2)}. \tag{8.5}$$

Indeed, let us assume that the three equations are false, then we have:

$$\begin{aligned} \widehat{\mu}_{i^*,T_{i^*}(t-1)} + (\psi^*)^{-1}\left(\frac{\alpha \log t}{T_{i^*}(t-1)}\right) &> \mu^* \\ &= \mu_i + \Delta_i \\ &\geq \mu_i + 2\,(\psi^*)^{-1}\left(\frac{\alpha \log t}{T_i(t-1)}\right) \\ &\geq \widehat{\mu}_{i,T_i(t-1)} + (\psi^*)^{-1}\left(\frac{\alpha \log t}{T_i(t-1)}\right), \end{aligned}$$

which implies in particular that $I_t \neq i$. In other words, letting $u = \lceil \frac{\alpha \log n}{\psi^*(\Delta_i/2)} \rceil$, we proved:

$$\begin{aligned} \mathbb{E}T_i(n) = \mathbb{E}\sum_{t=1}^n \mathbb{1}_{I_t=i} &\leq u + \mathbb{E}\sum_{t=u+1}^n \mathbb{1}_{I_t=i \text{ and } (8.5) \text{ is false}} \\ &\leq u + \mathbb{E}\sum_{t=u+1}^n \mathbb{1}_{(8.3) \text{ or } (8.4) \text{ is true}} \\ &= u + \sum_{t=u+1}^n \mathbb{P}((8.3) \text{ is true}) + \mathbb{P}((8.4) \text{ is true}). \end{aligned}$$

Thus it suffices to bound the probability of the events (8.3) and (8.4). Using an union bound and (8.2) one directly obtains:

$$\begin{aligned}
\mathbb{P}((8.3) \text{ is true}) &\leq \mathbb{P}\left(\exists s \in \{1,\dots,t\} : \widehat{\mu}_{i^*,s} + (\psi^*)^{-1}\left(\frac{\alpha \log t}{s}\right) \leq \mu^*\right) \\
&\leq \sum_{s=1}^{t} \mathbb{P}\left(\widehat{\mu}_{i^*,s} + (\psi^*)^{-1}\left(\frac{\alpha \log t}{s}\right) \leq \mu^*\right) \\
&\leq \sum_{s=1}^{t} \frac{1}{t^\alpha} \\
&= \frac{1}{t^{\alpha-1}}.
\end{aligned}$$

The same upper bound holds true for (8.4), which concludes the proof up to straightforward computations. □

8.2.2. Lower bound. We show here that the result of the previous section is essentially unimprovable.

THEOREM 8.2. *Let us consider a strategy such that for any set of Bernoulli reward distributions, any arm i such that $\Delta_i > 0$ and any $a > 0$, one has $\mathbb{E}T_i(n) = o(n^a)$. Then for any set of Bernoulli reward distributions, the following holds true:*

$$\liminf_{n \to +\infty} \frac{\overline{R}_n}{\log n} \geq \sum_{i:\Delta_i > 0} \frac{\Delta_i}{\mathrm{KL}(\mu_i, \mu^*)}.$$

PROOF. We provide a proof in three steps.

First step: Notations.
Without loss of generality let us assume that arm 1 is optimal and arm 2 is suboptimal, that is $\mu_2 < \mu_1 < 1$. Let $\varepsilon > 0$. Since $x \mapsto \mathrm{KL}(\mu_2, x)$ is continuous one can find $\mu_2' \in (\mu_1, 1)$ such that

$$\mathrm{KL}(\mu_2, \mu_2') \leq (1+\varepsilon)\mathrm{KL}(\mu_2, \mu_1). \tag{8.6}$$

We note $\mathbb{E}', \mathbb{P}'$ when we integrate with respect to the modified bandit where the parameter of arm 2 is replaced by μ_2'. We want to compare the behavior of the forecaster on the initial and modified bandits. In particular we prove that with a fair probability the forecaster can not distinguish between the two problems. Then using the fact that we have a good forecaster (by hypothesis in the Theorem) we know that the algorithm does not make too much mistakes on the modified bandit where arm 2 is optimal, in other words we have a lower bound on the number of times the optimal arm is played. This reasoning implies a lower bound on the number of times arm 2 is played in the initial problem.

To complete this program we introduce a few notations. Recall that $X_{2,1}, \dots, X_{2,n}$ is the sequence of random variables obtained while pulling arm 2. For $s \in \{1, \dots, n\}$, let

$$\widehat{\mathrm{KL}}_s = \sum_{t=1}^{s} \log\left(\frac{\mu_2 X_{2,t} + (1-\mu_2)(1-X_{2,t})}{\mu_2' X_{2,t} + (1-\mu_2')(1-X_{2,t})}\right).$$

In particular note that with respect to the initial bandit, $\widehat{\mathrm{KL}}_{T_2(n)}$ is the (non renormalized) empirical estimation of $\mathrm{KL}(\mu_2, \mu_2')$ at time n since in that case (X_s) is i.i.d from a Bernoulli of parameter μ_2. Another important property is that for any event A one has:

$$\mathbb{P}'(A) = \mathbb{E}\, \mathbb{1}_A \exp\left(-\widehat{\mathrm{KL}}_{T_2(n)}\right). \tag{8.7}$$

Now to control the link between the behavior of the forecaster on the initial and modified bandits we introduce the event:

$$C_n = \left\{ T_2(n) < \frac{1-\varepsilon}{\mathrm{KL}(\mu_2, \mu_2')} \log(n) \text{ and } \widehat{\mathrm{KL}}_{T_2(n)} \leq (1-\varepsilon/2)\log(n) \right\}. \tag{8.8}$$

Second step: $\mathbb{P}(C_n) = o(1)$**.**

By (8.7) and (8.8) one has:

$$\mathbb{P}'(C_n) = \mathbb{E}\, \mathbb{1}_{C_n} \exp\left(-\widehat{\mathrm{KL}}_{T_2(n)}\right) \geq \exp\left(-(1-\varepsilon/2)\log(n)\right) \mathbb{P}(C_n),$$

which implies by (8.8) and Markov's inequality:

$$\begin{aligned} \mathbb{P}(C_n) &\leq n^{(1-\varepsilon/2)} \mathbb{P}'(C_n) \\ &\leq n^{(1-\varepsilon/2)} \mathbb{P}'\left(T_2(n) < \frac{1-\varepsilon}{\mathrm{KL}(\mu_2, \mu_2')} \log(n) \right) \\ &\leq n^{(1-\varepsilon/2)} \frac{\mathbb{E}'(n - T_2(n))}{n - \frac{1-\varepsilon}{\mathrm{KL}(\mu_2,\mu_2')} \log(n)}. \end{aligned}$$

Now remark that in the modified bandit arm 2 is the unique optimal arm, thus the assumption that for any bandit, any suboptimal arm i, any $a > 0$, one has $\mathbb{E}T_i(n) = o(n^a)$ implies that

$$\mathbb{P}(C_n) \leq n^{(1-\varepsilon/2)} \frac{\mathbb{E}'(n - T_2(n))}{n - \frac{1-\varepsilon}{\mathrm{KL}(\mu_2,\mu_2')} \log(n)} = o(1).$$

Third step: $\mathbb{P}\left(T_2(n) < \frac{1-\varepsilon}{\mathrm{KL}(\mu_2,\mu_2')} \log(n) \right) = o(1)$**.**

Remark that

$$\begin{aligned} &\mathbb{P}(C_n) \\ &\geq \mathbb{P}\left(T_2(n) < \frac{1-\varepsilon}{\mathrm{KL}(\mu_2, \mu_2')} \log(n) \text{ and } \max_{1 \leq s \leq \frac{1-\varepsilon}{\mathrm{KL}(\mu_2,\mu_2')} \log(n)} \widehat{\mathrm{KL}}_s \leq (1-\varepsilon/2)\log(n) \right) \\ &= \mathbb{P}\Bigg(T_2(n) < \frac{1-\varepsilon}{\mathrm{KL}(\mu_2, \mu_2')} \log(n) \\ &\qquad \text{and } \frac{\mathrm{KL}(\mu_2,\mu_2')}{(1-\varepsilon)\log(n)} \max_{1 \leq s \leq \frac{(1-\varepsilon)\log(n)}{\mathrm{KL}(\mu_2,\mu_2')}} \widehat{\mathrm{KL}}_s \leq \frac{1-\varepsilon/2}{1-\varepsilon} \mathrm{KL}(\mu_2, \mu_2') \Bigg). \end{aligned} \tag{8.9}$$

Now using Lemma 8.1 since $\mathrm{KL}(\mu_2, \mu_2') > 0$ and the fact that $\frac{1-\varepsilon/2}{1-\varepsilon} > 1$ we deduce that

$$\lim_{n \to +\infty} \mathbb{P}\left(\frac{\mathrm{KL}(\mu_2, \mu_2')}{(1-\varepsilon)\log(n)} \max_{1 \le s \le \frac{(1-\varepsilon)\log(n)}{\mathrm{KL}(\mu_2,\mu_2')}} \widehat{\mathrm{KL}}_s \le \frac{1-\varepsilon/2}{1-\varepsilon} \mathrm{KL}(\mu_2, \mu_2') \right) = 1,$$

and thus by the result of the second step and (8.9):

$$\mathbb{P}\left(T_2(n) < \frac{1-\varepsilon}{\mathrm{KL}(\mu_2, \mu_2')} \log(n) \right) = o(1).$$

Now using (8.6) we obtain:

$$\mathbb{E} T_2(n) \ge (1 + o(1)) \frac{1-\varepsilon}{1+\varepsilon} \frac{\log(n)}{\mathrm{KL}(\mu_2, \mu_1)}$$

which concludes the proof. □

Lemma 8.1 (A Maximal Law of Large Numbers). *Let $X_1, X_2, \ldots$ be a sequence of independent real random variables with positive mean and satisfying almost surely*

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{t=1}^{n} X_t = \mu. \tag{8.10}$$

Then we have almost surely:

$$\lim_{n \to +\infty} \frac{1}{n} \max_{1 \le s \le n} \sum_{t=1}^{s} X_t = \mu. \tag{8.11}$$

Proof. Let $S_n = \sum_{t=1}^{n} X_t$ and $M_n = \max_{1 \le i \le n} S_i$. We need to prove that $\lim_{n \to +\infty} \frac{M_n}{n} = \mu$. First of all we clearly have almost surely:

$$\liminf_{n \to +\infty} \frac{M_n}{n} \ge \liminf_{n \to +\infty} \frac{S_n}{n} = \mu.$$

Now we need to upper bound the lim sup. Let $\varphi : \mathbb{N} \to \mathbb{N}$ be an increasing function such that $\varphi(n)$ is the largest integer smaller than n satisfying $M_n = S_{\varphi(n)}$. Thus

$$\frac{M_n}{n} \le \frac{S_{\varphi(n)}}{\varphi(n)}.$$

If $\varphi(n) \to \infty$ then one can conclude from (8.10) that

$$\limsup_{n \to +\infty} \frac{S_{\varphi(n)}}{\varphi(n)} \le \mu.$$

On the other hand if $\varphi(n) \le N \ \forall n$ then for any $T > 0$ we have $\sum_{t=N+1}^{T} X_t < 0$ and this event has probability zero since $\mathbb{P}(X_t < 0) < 1$ (otherwise μ would not be positive). □

References

The two key papers for stochastic multi-armed bandits are:

- T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985
- P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning Journal*, 47(2-3):235–256, 2002

There exists a fantastic number of variants of the UCB strategy, both for the basic multi-armed bandit problem as well as for extensions. A number of these extensions are referenced in [Chapter 2, [13]].

CHAPTER 9

Open Problems

A summary of the results proved in these lecture notes can be found in Table 1. They suggest a list of very precise open problems, which may eventually lead to a better understanding of the intrinsic trade-offs involved between the geometric structure of the action sets $\mathcal{A}$ and $\mathcal{Z}$ (and their interplay), the amount of feedback received by the player, the rate of growth of the minimax regret, and the computational resources available.

Recall that the combinatorial setting corresponds to $\mathcal{Z} = [0,1]^d$ and $\mathcal{A} \subset \{0,1\}^d$ with $||a||_1 = m, \forall a \in \mathcal{A}$. The dual setting with some norm $|| \cdot ||$ corresponds to $\sup_{a,z} ||\nabla \ell(a,z)||_* \leq G$ and $\mathcal{A}$ such that $\sup_a ||a|| \leq R$. Finally the bounded assumption corresponds to $\sup_{a,z} |\ell(a,z)| \leq 1$.

Open Problem 1. *What is the minimax rate for the regret with a bounded convex loss (in the full information setting)? Continuous Exp gives an upper bound of order $\sqrt{dn \log n}$. On the other hand one can easily prove a lower bound of order $\sqrt{dn}$ (using $\mathcal{A} = \{-1,1\}^d$ and a linear loss, see the proof technique of Theorem 7.8, or [19]).*

Open Problem 2. *What is the minimax rate for the regret with a subdifferentiable loss, in a dual setting with $|| \cdot ||_2$ (say $\mathcal{A} = B_{2,d}$), and under two-points bandit feedback? The strategy described in Section 7.5 gives an upper bound of order $d\sqrt{n}$. It is not clear if the correct order of magnitude is $\sqrt{dn}$ or $d\sqrt{n}$.*

Open Problem 3. *What is the minimax rate for the regret with a linear loss, in a combinatorial setting, and under bandit feedback? The gap between the upper and lower bound is of order $\sqrt{m}$.*

Open Problem 4. *What is the minimax rate for the regret with a bounded linear loss, under bandit feedback? The gap between the upper and lower bound is of order $\sqrt{\log n}$.*

Open Problem 5. *What is the minimax rate for the regret with a linear loss, in a dual setting with $|| \cdot ||_2$ (say $\mathcal{A} = B_{2,d}$), and under bandit feedback? The gap between the upper and lower bound is of order $\sqrt{\log n}$. What about the rate for other dual settings (say $|| \cdot ||_p$ for example)?*

Open Problem 6. *What is the minimax rate for the regret with a bounded subdifferentiable loss, in a dual setting with $|| \cdot ||_2$ (say $\mathcal{A} = B_{2,d}$), and under bandit feedback? The gap between the upper and lower bound is of order $n^{1/4}$.*

Open Problem 7. *In the combinatorial setting with full information, is it possible to attain the optimal regret with only a linear (in d) number of calls to an oracle for the offline problem (i.e., an oracle minimizing linear functions on $\mathcal{A}$).*

The FPL strategy satisfies the constraint of number of calls, but attains a suboptimal regret compared to OMD with the negative entropy.

Open Problem 8. *Is it possible to design a polynomial time strategy (in d) with optimal regret under the assumption of a bounded linear loss?*

	Lower bound	Upper bound	Conjecture
bounded convex, expert regret	$\sqrt{n \log d}$	$\sqrt{n \log d}$	–
σ-exp concave, expert regret	–	$\frac{\log d}{\sigma}$	–
bounded convex	$\sqrt{dn}$	$\sqrt{dn \log n}$	–
σ-exp concave	–	$\frac{d \log n}{\sigma}$	–
subdifferentiable, dual setting with $\|\| \cdot \|\|_p$, $p \in [1, 2]$	–	$RG\sqrt{\frac{n}{p-1}}$	–
subdifferentiable, dual setting with $\|\| \cdot \|\|_1$ on the simplex	$G\sqrt{n \log d}$	$G\sqrt{n \log d}$	–
α-strongly convex, dual setting with $\|\| \cdot \|\|_2$	–	$RG\frac{\log n}{\alpha}$	–
combinatorial setting	$m\sqrt{n \log \frac{d}{m}}$	$m\sqrt{n \log \frac{d}{m}}$	–
combinatorial setting, semi-bandit	$\sqrt{mdn}$	$\sqrt{mdn}$	–
subdifferentiable, dual setting with $\|\| \cdot \|\|_2$, two-points bandit	–	$RGd\sqrt{n}$	–
bounded linear, bandit	$d\sqrt{n}$	$d\sqrt{n \log n}$	–
combinatorial setting, bandit	$m\sqrt{dn}$	$m^{3/2}\sqrt{dn \log \frac{d}{m}}$	$m\sqrt{dn}$
linear, dual setting with $\|\| \cdot \|\|_2$ on the Euclidean ball, bandit	–	$G\sqrt{dn \log n}$	$G\sqrt{dn}$
bounded subdifferentiable, dual setting with $\|\| \cdot \|\|_2$, bandit (dependencies other than n omitted)	$\sqrt{n}$	$n^{3/4}$	$\sqrt{n}$

Table 1. Summary of the results proved in these lecture notes.

Bibliography

[1] J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pages 263–274, 2008.

[2] A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, 2010.

[3] J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

[4] J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2635–2686, 2010.

[5] J.-Y. Audibert, S. Bubeck, and G. Lugosi. Minimax policies for combinatorial prediction games. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, 2011.

[6] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning Journal*, 47(2-3):235–256, 2002.

[7] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2003.

[8] B. Awerbuch and R. Kleinberg. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches. In *STOC '04: Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 45–53, 2004.

[9] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[10] K. Ball. An elementary introduction to modern convex geometry. In S. Levy, editor, *Flavors of Geometry*, pages 1–58. Cambridge University Press, 1997.

[11] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, 2008.

[12] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[13] S. Bubeck. *Bandits Games and Clustering Foundations.* PhD thesis, Université Lille 1, 2010.

[14] N. Cesa-Bianchi and S. Kakade. An optimal algorithm for linear bandits. arXiv:1110.4322v1, 2011.

[15] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, 2006.

[16] N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 2011. To appear.

[17] T. Cover. Universal portfolios. *Math. Finance*, 1:1–29, 1991.

[18] A. Dalalyan and A. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72:39–61, 2008.

[19] V. Dani, T. Hayes, and S. Kakade. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, pages 345–352, 2008.

[20] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[21] A. Flaxman, A. Kalai, and B. McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *In Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 385–394, 2005.

[22] A. Grove, N. Littlestone, and D. Schuurmans. General convergence results for linear discriminant updates. *Machine Learning*, 43:173–210, 2001.

[23] A. György, T. Linder, G. Lugosi, and G. Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8:2369–2403, 2007.

[24] E. Hazan. The convex optimization approach to regret minimization. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*, pages 287–303. MIT press, 2011.

[25] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69:169–192, 2007.

[26] E. Hazan, S. Kale, and M. Warmuth. Learning rotations with little regret. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, 2010.

[27] D. P. Helmbold and M. Warmuth. Learning permutations with exponential weights. *Journal of Machine Learning Research*, 10:1705–1736, 2009.

[28] M. Herbster and M. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998.

[29] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2001.

[30] A. Juditsky and A. Nemirovski. First-order methods for nonsmooth convex large-scale optimization, i: General purpose methods. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*, pages 121–147. MIT press, 2011.

[31] A. Juditsky and A. Nemirovski. First-order methods for nonsmooth convex large-scale optimization, ii: Utilizing problem's structure. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*, pages 149–183. MIT press, 2011.

[32] A. Kalai and S. Vempala. Efficient algorithms for universal portfolios. *Journal of Machine Learning Research*, 3:423–440, 2002.

[33] A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71:291–307, 2005.

[34] S. Kale, L. Reyzin, and R. Schapire. Non-stochastic bandit slate problems. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1054–1062, 2010.

[35] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:462–466, 1952.

[36] J. Kivinen and M. Warmuth. Relative loss bounds for multidimensional regression problems. *Machine Learning*, 45:301–329, 2001.

[37] W. Koolen, M. Warmuth, and J. Kivinen. Hedging structured concepts. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 93–105, 2010.

[38] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[39] N. Littlestone and M. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

[40] L. Lovasz and S. Vempala. Fast algorithms for logconcave functions: sampling, rounding, integration and optimization. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 57–68, 2006.

[41] H. McMahan and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *In Proceedings of the 17th Annual Conference on Learning Theory (COLT)*, pages 109–123, 2004.

[42] H. Narayanan and A. Rakhlin. Random walk approach to regret minimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[43] A. Nemirovski. Efficient methods for large-scale convex optimization problems. *Ekonomika i Matematicheskie Metody*, 15, 1979. (In Russian).

[44] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley Interscience, 1983.

[45] B. Polyak. A general method for solving extremal problems. *Soviet Math. Doklady*, 174:33–36, 1967.

[46] A. Rakhlin. Lecture notes on online learning. 2009.

[47] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[48] A. Schrijver. *Combinatorial Optimization.* Springer, 2003.

[49] S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications.* PhD thesis, The Hebrew University of Jerusalem, 2007.

[50] N. Shor. Generalized gradient descent with application to block programming. *Kibernetika*, 3:53–55, 1967. (In Russian).

[51] J. Spall. *Introduction to stochastic search and optimization. Estimation, simulation, and control.* Wiley Interscience, 2003.

[52] T. Uchiya, A. Nakamura, and M. Kudo. Algorithms for adversarial bandit problems with multiple plays. In *Proceedings of the 21st International Conference on Algorithmic Learning Theory (ALT)*, 2010.

[53] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.

[54] S. Vempala. Recent progress and open problems in algorithmic convex geometry. In K. Lodaya and M. Mahajan, editors, *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2010)*, volume 8 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 42–64. Schloss Dagstuhl–Leibniz-Zentrum fur Informatik, 2010.

[55] V. Vovk. Aggregating strategies. In *Proceedings of the third annual workshop on Computational learning theory (COLT)*, pages 371–386, 1990.

[56] M. Warmuth, W. Koolen, and D. Helmbold. Combining initial segments of lists. In *In Proceedings of the 22nd International Conference on Algorithmic*

Learning Theory (ALT), 2011.

[57] M. Warmuth and D. Kuzmin. Randomized online pca algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9:2287–2320, 2008.

[58] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, 2003.