Mirror descent strategies for regret minimization and approachability

Joon Kwon

THÈSE DE DOCTORAT

DE

MATHÉMATIQUES

PRÉSENTÉE PAR

Joon KWON,

PORTANT SUR LES

STRATÉGIES DE DESCENTE MIROIR POUR LA MINIMISATION DU REGRET ET L'APPROCHABILITÉ,

DIRIGÉE PAR

MM. Rida LARAKI & Sylvain SORIN

et soutenue le 18 octobre 2016 devant le jury composé de :

M. Gérard BIAU	Université Pierre-et-Marie-Curie	examinateur,
M. Rida LARAKI	CNRS & Université Paris–Dauphine	directeur,
M. Éric MOULINES	École polytechnique	examinateur,
M. Vianney PERCHET	École normale supérieure de Cachan	examinateur,
M. Sylvain SORIN	Université Pierre-et-Marie-Curie	directeur,
M. Gilles STOLTZ	CNRS & HEC Paris	rapporteur,

après avis des rapporteurs MM. Gábor LUGOSI (Universitat Pompeu Fabra) & Gilles STOLTZ (CNRS & HEC Paris).

Université Pierre-et-Marie-Curie
École doctorale de mathématiques de Paris–Centre

Institut de mathématiques de Jussieu–Paris rive gauche, CNRS UMR 7586
Boîte courrier 247, 4 place Jussieu, 75 252 Paris Cedex 05

Université Pierre-et-Marie-Curie
École doctorale de sciences mathématiques de Paris–Centre
Boîte courrier 290, 4 place Jussieu, 75 252 Paris Cedex 05

What I cannot create, I do not understand.

Richard P. Feynman

REMERCIEMENTS

Mes premières pensées vont évidemment à mes directeurs Rida Laraki et Sylvain Sorin, ainsi que Vianney Perchet. Je me sens très chanceux d'avoir effectué mes premiers pas dans la recherche sous leur tutelle. Ils m'ont laissé une grande liberté tout en maintenant un haut niveau d'exigence. Je les remercie aussi pour leur disponibilité, leur patience et leur tolérance. Rida a été le premier à me proposer le sujet de la minimisation du regret et de l'approche en temps continu. Il a été pendant ces années de thèse d'une grande bienveillance, a toujours été accessible, et m'a prodigué de nombreux conseils. De Sylvain, j'ai acquis une grande exigence de rigueur et de précision, et j'ai pu apprendre de sa très grande culture mathématique. Mais je souhaite aussi mentionner la chaleur humaine qui le caractérise : il partage toujours avec générosité sa passion pour la vie en général, et les huîtres en particulier. Enfin, l'accompagnement de Vianney a été décisif pour mon travail de thèse. En autres choses, il m'a fait découvrir les sujets de l'approchabilité et des jeux à observations partielles, lesquels ont donné lieu à un chapitre important de la thèse. J'espère continuer à travailler avec lui à l'avenir.

Je remercie chaleureusement Gilles Stoltz et Gábor Lugosi d'avoir accepté d'être les rapporteurs de cette thèse. C'est un honneur que m'ont fait ces deux grands spécialistes du domaine.

Merci aussi à Panayotis Mertikopoulos avec qui j'ai effectué ma toute première collaboration, laquelle correspond à un chapitre de la thèse.

J'ai également une pensée pour Yannick Viossat avec qui j'ai effectué mon stage de M1, et qui m'a ensuite encouragé à prendre contact avec Sylvain.

Je suis aussi redevable des enseignants que j'ai eu tout au long de ma scolarité. En particulier, je tiens à citer Serge Francinou et le regretté Yves Révillon.

Ce fut un grand bonheur d'effectuer ma thèse au sein de l'équipe Combinatoire et optimisation de l'Institut de mathématiques de Jussieu, où j'ai bénéficié de conditions de travail exceptionnelles, ainsi que d'une très grande convivialité. Je remercie tous les doctorants, présents ou passés, que j'y ai rencontrés : Cheng, Daniel, Hayk, Mario, Miquel, Pablo, Teresa, Xiaoxi, Yining ; ainsi que les chercheurs confirmés Arnau Padrol, Benjamin Girard, Daniela Tonon, Éric Balandraud, Hélène Frankowska, Ihab Haidar, Jean-Paul Allouche et Marco Mazzola.

Merci aux doctorants des laboratoires voisins, avec qui j'ai partagé d'excellents moments : Boum, Casimir, Éric, Olga, Pierre-Antoine, Sarah et Vincent.

Je salue bien sûr toute la communauté française de théorie des jeux ou plutôt son adhérence, que j'ai eu le plaisir de cotoyer lors des séminaires, conférences, et autres écoles d'été : Olivier Beaude, Jérome Bolte, Roberto Cominetti, Mathieu Faure, Gaëtan Fournier, Stéphane Gaubert, Fabien Gensbittel, Saeed Hadikhanloo, Antoine Hochard, Marie Laclau, Nikos Pnevmatikos, Marc Quincampoix, Jérome Renault, Ludovic Renou, Thomas Rivera, Bill Sandholm, Marco Scarsini, Tristan Tomala, Xavier Venel, Nicolas Vieille, Guillaume Vigeral et Bruno Ziliotto.

Je n'oublie pas mes amis matheux (ou apparentés) qui ne rentrent pas dans les catégories précédentes : Guillaume Barraquand, Frédéric Bègue, Ippolyti Dellatolas, Nicolas Flammarion, Vincent Jugé, Igor Kortchemski, Matthieu Lequesne et Arsène Pierrot.

Ces années de thèse auraient été difficiles à surmonter sans l'immortel groupe de Lyon/PES composé de BZ, JChevall, La Pétrides, PCorre et moi-même, lequel groupe se retrouve au complet à Paris pour cette année 2016–2017 !

Un très grand merci à mes amis de toujours : Antoine, Clément, David, Mathilde, Nicolas et Raphaël.

Et enfin, je remercie mes parents, mon frère, et bien sûr Émilie qui me donne tant.

RÉSUMÉ

Le manuscrit se divise en deux parties. La première est constituée des chapitres I à IV et propose une présentation unifiée de nombreux résultats connus ainsi que de quelques éléments nouveaux.

On présente dans le Chapitre I le problème d'*online linear optimization*, puis on construit les stratégies de descente miroir avec paramètres variables pour la minimisation du regret, et on établit dans le Théorème I.3.1 une borne générale sur le regret garantie par ces stratégies. Ce résultat est fondamental car la quasi-totalité des résultats des quatre premiers chapitres en seront des corollaires. On traite ensuite l'extension aux pertes convexes, puis l'obtention d'algorithmes d'optimisation convexe à partir des stratégies minimisant le regret.

Le Chapitre II se concentre sur le cas où le joueur dispose d'un ensemble fini dans lequel il peut choisir ses actions de façon aléatoire. Les stratégies du Chapitre I sont aisément transposées dans ce cadre, et on obtient également des garanties presque-sûres d'une part, et avec grande probabilité d'autre part. Sont ensuite passées en revue quelques stratégies connues : l'*Exponential Weights Algorithm*, le *Smooth Fictitious Play*, le *Vanishingly Smooth Fictitious Play*, qui apparaissent toutes comme des cas particuliers des stratégies construites au Chapitre I. En fin de chapitre, on mentionne le problème de bandit à plusieurs bras, où le joueur n'observe que le paiement de l'action qu'il a jouée, et on étudie l'algorithme *EXP3* qui est une adaptation de l'Exponential Weights Algorithm dans ce cadre.

Le Chapitre III est consacré à la classe de stratégies appelée *Follow the Perturbed Leader*, qui est définie à l'aide de perturbations aléatoires. Un récent survey [ALT16] mentionne le fait que ces stratégies, bien que définies de façon différente, appartiennent à la famille de descente miroir du Chapitre I. On donne une démonstration détaillée de ce résultat.

Le Chapitre IV a pour but la construction de stratégies de descente miroir pour l'*approchabilité de Blackwell*. On étend une approche proposée par [ABH11] qui permet de transformer une stratégie minimisant le regret en une stratégie d'approchabilité. Notre approche est plus générale car elle permet d'obtenir des bornes sur une très large classe de quantités mesurant l'éloignement à l'ensemble cible, et non pas seulement sur la distance euclidienne à l'ensemble cible. Le caractère unificateur de cette

démarche est ensuite illustrée par la construction de stratégies optimales pour le problème d'*online combinatorial optimization* et la minimisation du *regret interne/swap*. Par ailleurs, on démontre que la stratégie de Backwell peut être vue comme un cas particulier de descente miroir.

La seconde partie est constituée des quatre articles suivants, qui ont été rédigés pendant la thèse.

Le Chapitre V est tiré de l'article [KP16b] et étudie le problème de la minimisation du regret dans le cas où le joueur possède un ensemble fini d'actions, et avec l'hypothèse supplémentaire que les vecteurs de paiement possèdent au plus s composantes non-nulles. On établit, en information complète, que la borne optimale sur le regret est de l'ordre de $\sqrt{T \log s}$ (où T est le nombre d'étapes) lorsque les paiements sont des gains (c'est-à-dire lorsqu'ils sont positifs), et de l'ordre de $\sqrt{Ts\frac{\log d}{d}}$ (où d est le nombre d'actions) lorsqu'il s'agit de pertes (i.e. négatifs). On met ainsi en évidence une différence fondamentale entre les gains et les pertes. Dans le cadre bandit, on établit que la borne optimale pour les pertes est de l'ordre de $\sqrt{Ts}$ à un facteur logarithmique près.

Le Chapitre VI est issu de l'article [KP16a] et porte sur l'approchabilité de Blackwell avec *observations partielles*, c'est-à-dire que le joueur observe seulement des signaux aléatoires. On construit des stratégies garantissant des vitesses de convergence de l'ordre de $O(T^{-1/2})$ dans le cas de signaux dont les lois ne dépendent pas de l'action du joueur, et de l'ordre de $O(T^{-1/3})$ dans le cas général. Cela établit qu'il s'agit là des vitesses optimales car il est connu qu'on ne peut les améliorer sans hypothèse supplémentaire sur l'ensemble cible ou la structure des signaux.

Le Chapitre VII est tiré de l'article [KM14] et définit les stratégies de descente miroir en temps continu. On établit pour ces derniers une propriété de non-regret. On effectue ensuite une comparaison entre le temps continu et le temps discret. Cela offre une interprétation des deux termes qui constituent la borne sur le regret en temps discret : l'un vient de la propriété en temps continu, l'autre de la comparaison entre le temps continu et le temps discret.

Enfin, le Chapitre VIII est indépendant et est issu de l'article [Kwo14]. On y établit une borne universelle sur les variations des fonctions convexes bornées. On obtient en corollaire que toute fonction convexe bornée est lipschitzienne par rapport à la métrique de Hilbert.

[KP16b] Joon Kwon and Vianney Perchet. Gains and losses are fundamentally different in regret minimization : the sparse case. *arXiv :1511.08405*, 2016 (*à paraître dans* Journal of Machine Learning Research)

[KP16a] Joon Kwon and Vianney Perchet. Blackwell approachability with partial monitoring : Optimal convergence rates. 2016 (*en préparation*)

[KM14] Joon Kwon and Panayotis Mertikopoulos. A continuous-time approach to online optimization. *arXiv :1401.6956*, 2014 (*en préparation*)

[Kwo14] Joon Kwon. A universal bound on the variations of bounded convex functions. *arXiv :1401.2104*, 2014 (*à paraître dans* Journal of Convex Analysis)

ABSTRACT

The manuscript is divided in two parts. The first consists in Chapters I to IV and offers a unified presentation of numerous known results as well as some new elements.

We present in Chapter I the *online linear optimization* problem, then construct Mirror Descent strategies with varying parameters for regret minimization, and establish in Theorem I.3.1 a general bound on the regret guaranteed by the strategies. This result is fundamental, as most of the results from the first four chapters will be obtained as corollaries. We then deal with the extension to convex losses, and with the derivation of convex optimization algorithms from regret minimizing strategies.

Chapter II focuses on the case where the Decision Maker has a finite set from which he can pick his actions at random. The strategies from Chapter I are easily transposed to this framework and we also obtain high-probability and almost-sure guarantees. We then review a few known strategies: *Exponential Weights Algorithm*, *Smooth Fictitious Play*, and *Vanishingly Smooth Fictitious Play*, which all appear as special cases of the strategies constructed in Chapter I. At the end of the chapter, we mention the multi-armed bandit problem, where the Decision Maker only observes the payoff of the action he has played. We study the *EXP3* strategy, which is an adaptation of the Exponential Weights Algorithm to this setting.

Chapter III is dedicated to the family of strategies called *Follow the Perturbed Leader*, which is defined using random perturbations. A recent survey [ALT16] mentions the fact that those strategies, although defined differently, actually belong to the family of Mirror Descent strategies from Chapter I. We give a detailed proof of this result.

Chapter IV aims at constructing Mirror Descent strategies for Blackwell's approachability. We extend an approach proposed by [ABH11] that turns a regret minimizing strategy into an approachability strategy. Our construction is more general, as it provides bounds for a very large class of distance-like quantities which measure the "distance" to the target set and not only on the Euclidean distance to the target set. The unifying character of this approach is then illustrated by the construction of optimal strategies for *online combinatorial optimization* and *internal/swap regret* minimization. Besides, we prove that Blackwell's strategy can be seen as a special case of Mirror Descent.

The second part of the manuscript contains the following four papers.

Chapter V is from [KP16b] and studies the regret minimization problem in the case where the Decision Maker has a finite set of actions, with the additional assumption that payoff vectors have at most s nonzero components. We establish, in the full information setting, that the minimax regret is of order $\sqrt{T \log s}$ (where T is the number of steps) when payoffs are gains (i.e nonnegative), and of order $\sqrt{Ts\frac{\log d}{d}}$ (where d is the number of actions) when the payoffs are losses (i.e. nonpositive). This demonstrates a fundamental difference between gains and losses. In the bandit setting, we prove that the minimax regret for losses is of order $\sqrt{Ts}$ up to a logarithmic factor.

Chapter VI is extracted from [KP16a] and deals with Blackwell's approachability with partial monitoring, meaning that the Decision Maker only observes random signals. We construct strategies which guarantee convergence rates of order $O(T^{-1/2})$ in the case where the signal does not depend on the action of the Decision Maker, and of order $O(T^{-1/3})$ in the case of general signals. This establishes the optimal rates in those two cases, as the above rates are known to be unimprovable without further assumption on the target set or the signalling structure.

Chapter VII comes from [KM14] and defines Mirror Descent strategies in continuous time. We prove that they satisfy a regret minimization property. We then conduct a comparison between continuous and discrete time. This offers an interpretation of the terms found in the regret bounds in discrete time: one is from the continuous time property, and the other comes from the comparison between continuous and discrete time.

Finally, Chapter VIII is independent and is from [Kwo14]. We establish a universal bound on the variations of bounded convex function. As a byproduct, we obtain that every bounded convex function is Lipschitz continuous with respect to the Hilbert metric.

[KP16b] Joon Kwon and Vianney Perchet. Gains and losses are fundamentally different in regret minimization: the sparse case. *arXiv:1511.08405*, 2016 (*to appear in* Journal of Machine Learning Research)

[KP16a] Joon Kwon and Vianney Perchet. Blackwell approachability with partial monitoring: Optimal convergence rates. 2016 (*in preparation*)

[KM14] Joon Kwon and Panayotis Mertikopoulos. A continuous-time approach to online optimization. *arXiv:1401.6956*, 2014 (*in preparation*)

[Kwo14] Joon Kwon. A universal bound on the variations of bounded convex functions. *arXiv:1401.2104*, 2014 (*to appear in* Journal of Convex Analysis)

TABLE OF CONTENTS

INTRODUCTION

Online learning

Online learning deals with making decisions sequentially with the goal of obtaining good overall results. Such problems have originated and have been studied in many different fields such as economics, computer science, statistics and information theory. In recent years, the increase of computing power allowed the use of online learning algorithms in countless applications: advertisement placement, web ranking, spam filtering, energy consumption forecast, to name a few. This has naturally boosted the development of the involved mathematical theories.

Online learning can be modeled as a setting where a Decision Maker faces Nature repeatedly, and in which information about his performance and the changing state of Nature is revealed throughout the play. The Decision Maker is to use the information he has obtained in order to make better decisions in the future. Therefore, an important characteristic of an online learning problem is the type of feedback the Decision Maker has, in other words, the amount of information available to him. For instance, in the *full information* setting, the Decision Maker is aware of everything that has happened in the past; in the *partial monitoring* setting, he only observes, after each stage, a random signal whose law depends on his decision and the state of Nature; and in the *bandit* setting, he only observes the payoff he has obtained.

Concerning the behavior of Nature, we can distinguish two main types of assumptions. In *stochastic* settings, the successive states of Nature are drawn according to some fixed probability law, whereas in the *adversarial* setting, no such assumption is made and Nature is even allowed to choose its states strategically, in response to the previous choices of the Decision Maker. In the latter setting, the Decision Maker is aiming at obtaining worst-case guarantees. This thesis studies adversarial online problems.

To measure the performance of the Decision Maker, a quantity to minimize or a criterion to satisfy has to be specified. We present below two of those: regret minimization and approachability. Both are very general frameworks which have been successfully applied to a variety of problems.

Regret minimization

We present the adversarial regret minimization problem which has been used as a unifying framework for the study of many online learning problems: pattern recognition, portfolio management, routing, ranking, principal component analysis, matrix learning, classification, regression, etc. Important surveys on the topic are [CBL06, RT09, Haz12, BCB12, SS11].

We first consider the problem where the Decision Maker has a finite set of *actions* $\mathcal{I} = \{1, \dots, d\}$. At each stage $t \geqslant 1$, the Decision Maker chooses an action $i_t \in \mathcal{I}$, possibly at random, then observes a *payoff vector* $u_t \in [-1, 1]^d$, and finally gets a scalar payoff equal to $u_t^{i_t}$. We assume Nature to be adversarial, and the Decision Maker is therefore aiming at obtaining *some guarantee* against any possible sequence of payoff vectors $(u_t)_{t \geqslant 1}$ in $[-1, 1]^d$. Hannan [Han57] introduced the notion of regret, defined as

$$\mathrm{R_T} = \max_{i \in \mathcal{I}} \sum_{t=1}^{\mathrm{T}} u_t^i - \sum_{t=1}^{\mathrm{T}} u_t^{i_t},$$

which compares the cumulative payoff $\sum_{t=1}^{\mathrm{T}} u_t^{i_t}$ obtained by the Decision Maker to the cumulative payoff $\max_{i \in \mathcal{I}} \sum_{t=1}^{\mathrm{T}} u_t^i$ he could have obtained by playing the best fixed action in hindsight. Hannan [Han57] established the existence of strategies for the Decision Maker which guarantee that the average regret $\frac{1}{\mathrm{T}}\mathrm{R_T}$ is asymptotically non-positive. This problem is also called *prediction with expert advice* because it models the following situation. Imagine $\mathcal{I} = \{1, \dots, d\}$ as a set of *experts*. At each stage $t \geqslant 1$, the Decision Maker has to make a decision and each expert give a piece of advice as to which decision to make. The Decision Maker must then choose the expert i_t to follow. Then, the vector $u_t \in \mathbb{R}^d$ is observed, where u_t^i is the payoff obtained by expert i. The payoff obtained by the Decision Maker is therefore $u_t^{i_t}$. The regret then corresponds to the difference between the cumulative payoff of the Decision Maker and the cumulative payoff obtained by the best expert. The Decision Maker having a strategy which makes sure that the average regret goes to zero means that he is able to perform, asymptotically and in average, as well as any expert.

The theory of regret minimization has since been refined and developed in a number of ways—see e.g. [FV97, HMC00, FL99, Leh03]. An important direction was the study of the best possible guarantee on the expected regret, in other words the study of the following quantity:

$$\inf \sup \mathbb{E}\left[\mathrm{R_T}\right],$$

where the infimum is taken over all possible strategies of the Decision Maker, the supremum over all sequences $(u_t)_{t \geqslant 1}$ of payoff vectors in $[-1, 1]^d$, and the expectation with respect to the randomization introduced by the Decision Maker in choos-

ing its actions i_t. This quantity has been established [CB97, ACBG02] to be of order $\sqrt{T \log d}$, where T is the number of stages and d the number of actions.

An interesting variant is the *online convex optimization* problem [Gor99, KW95, KW97, KW01, Zin03]: the Decision Maker chooses actions z_t in a convex compact set $\mathcal{Z} \subset \mathbb{R}^d$, and Nature chooses loss functions $\ell_t : \mathcal{Z} \to \mathbb{R}$. The regret is then defined by

$$R_T = \sum_{t=1}^{T} \ell_t(z_t) - \min_{z \in \mathcal{Z}} \sum_{t=1}^{T} \ell_t(z).$$

The special case where the loss functions are linear is called *online linear optimization* and is often written with the help of payoff vectors $(u_t)_{t \geqslant 1}$:

$$R_T = \max_{z \in \mathcal{Z}} \sum_{t=1}^{T} \langle u_t | z \rangle - \sum_{t=1}^{T} \langle u_t | z_t \rangle \,. \tag{$*$}$$

This will be the base model upon which Part I of the manuscript will be built.

Until now, we have assumed that the Decision Maker observes all previous payoff vectors (or loss functions), in other words, that he has a *full information* feedback. The problems in which the Decision Maker only observes the payoff (or the loss) that he obtains are called *bandit* problems. The case where the set of actions is $\mathcal{I} = \{1, \dots, d\}$ is called the adversarial multi-armed bandit problem, for which the minimax regret is known to be of order $\sqrt{Td}$ [AB09, ACBFS02]. The bandits settings for online convex/linear optimization has also attracted much attention [AK04, FKM05, DH06, BDH+08] and we refer to [BCB12] for a recent survey.

Approachability

Blackwell [Bla54, Bla56] considered a model of repeated games between a Decision Maker and Nature with vector-valued payoffs. He studied the sets to which the Decision Maker can make sure his average payoff converges. Such sets are said to be *approachable* by the Decision Maker. Specifically, let $\mathcal{I}$ and $\mathcal{J}$ be finite action sets for the Decision Maker and Nature respectively,

$$\Delta(\mathcal{I}) = \left\{ x = (x^i)_{i \in \mathcal{I}} \in \mathbb{R}_+^{\mathcal{I}} \,\middle|\, \sum_{i \in \mathcal{I}} x^i = 1 \right\}$$

the set of probability distributions on $\mathcal{I}$, and $g : \mathcal{I} \times \mathcal{J} \to \mathbb{R}^d$ a vector-valued payoff function. For a given (closed) *target set* $\mathcal{C} \subset \mathbb{R}^d$, the question is whether there exists a strategy for the Decision Maker which guarantees that

$$\frac{1}{T} \sum_{t=1}^{T} g(i_t, j_t) \xrightarrow[T \to +\infty]{} \mathcal{C},$$

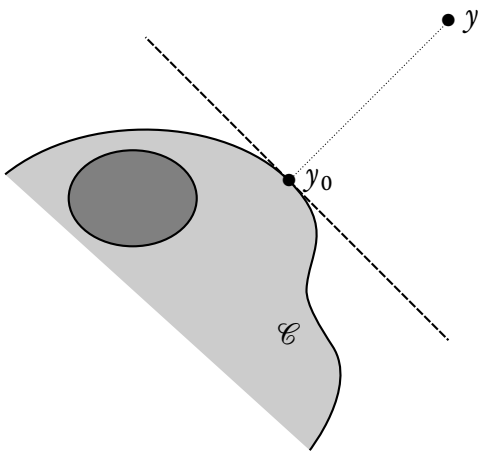


Figure 1. — The hyperplane $\langle y - y_0 | \cdot - y_0 \rangle = 0$ separates y and the set of all possible expected vector payoffs when the Decision Maker plays at random according to probability distribution $\mathbf{x}(y)$ (represented in dark gray).

where i_t and j_t denote the actions chosen at time t by the Decision Maker and Nature, respectively.

Blackwell provided the following sufficient condition for a closed set $\mathscr{C} \subset \mathbb{R}^d$ to be *approachable*: for all $y \in \mathbb{R}^d$, there exists an Euclidean projection y_0 of y onto $\mathscr{C}$, and a probability distribution $\mathbf{x}(y) \in \Delta(\mathscr{I})$ such that for all actions $j \in \mathscr{J}$ of Nature,

$$\langle \mathbb{E}_{i \sim \mathbf{x}(y)} [g(i, j)] - y_0 | y - y_0 \rangle \leqslant 0.$$

The above inequality is represented in Figure 1. $\mathscr{C}$ is then said to be a B-set. When this is the case, the Blackwell strategy is defined as

$$x_{t+1} = \mathbf{x}\left(\frac{1}{t} \sum_{s=1}^{t} g(i_s, j_s) \right) \quad \text{then draw} \quad i_{t+1} \sim x_{t+1},$$

which means that action $i_{t+1} \in \mathscr{I}$ is drawn according to probability distribution $x_{t+1} \in \Delta(\mathscr{I})$. This strategy guarantees the convergence of the average payoff $\frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} g(i_t, j_t)$ to the set $\mathscr{C}$. Later, [Spi02] proved that a closed set is approachable if and only if it contains a B-set. In the case of a convex set $\mathscr{C}$, Blackwell proved that it is approachable if and only if it is a B-set, which is then also equivalent to the following dual condition:

$$\forall y \in \Delta(\mathscr{J}),\ \exists x \in \Delta(\mathscr{I}), \quad \mathbb{E}_{\substack{i \sim x \\ j \sim y}} [g(i, j)] \in \mathscr{C}.$$

This theory turned out to be a powerful tool for constructing strategies for online learning, statistics and game theory. Let us mention a few applications. Many

variants of the regret minimization problem can be reformulated as an approachability problem, and conversely, regret minimization strategy can be turned into approachability strategy. Blackwell [Bla54] was already aware of this fundamental link between regret and approachability, which has since been much developed—see e.g. [HMC01, Per10, MPS11, ABH11, BMS14, Per15]. The statistical problem of *calibration* has also proved to be related to approachability [Fos99, MS10, Per10, RST11, ABH11, Per15]. We refer to [Per14] for a comprehensive survey on the relations between regret, calibration and approachability. Finally, Blackwell's theory has been applied to the construction of optimal strategies in zero-sum repeated games with incomplete information [Koh75, AM85].

Various techniques have been developed for constructing and analyzing approachability strategies. As shown above, Blackwell's initial approach was based on Euclidean projections. A potential-based approach was proposed to provide a wider and more flexible family of strategies [HMC01, CBL03, Per15]. In a somewhat related spirit, and building upon an approach with convex cones introduced in [ABH11], we define in Chapter IV a family of Mirror Descent strategies for approachability.

The approachability problem has also been studied in the partial monitoring setting [Per11a, MPS11, PQ14, MPS14]. In Chapter VI we construct strategies which achieve optimal convergence rates.

On the origins of Mirror Descent

In this section, we quickly present the succession of ideas which have led to the Mirror Descent algorithms for convex optimization and regret minimization. We do not aim at being comprehensive nor completely rigorous. We refer to [CBL06, Section 11.6], [Haz12], and to [Bub15] for a recent survey.

We first consider the unconstrained problem of optimizing a convex function $f : \mathbb{R}^d \to \mathbb{R}$ which we assume to be differentiable:

$$\min_{x \in \mathbb{R}^d} f(x).$$

We shall focus on the construction of algorithms based on first-order oracles—in other words, algorithms which have access to the gradient $\nabla f(x)$ at any point x.

Gradient Descent

The initial idea is to adapt the continuous-time gradient flow

$$\dot{x} = -\nabla f(x).$$

There are two basic discretizations. The first is the *proximal* algorithm, which starts at some initial point x_1 and iterates as

$$x_{t+1} = x_t - \gamma_t \nabla f(x_{t+1}), \tag{1}$$

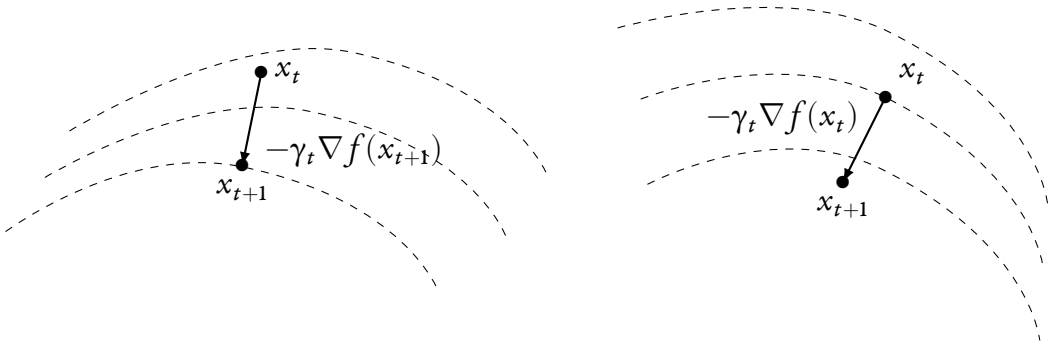


Figure 2. — The Proximal algorithm on the left and Gradient Descent on the right

where γ_t is a *step-size*. The algorithm is said to be *implicit* because one has to find a point x_{t+1} satisfying the above equality in which x_{t+1} implicitly appears in $\nabla f(x_{t+1})$. One can see that the above relation can be rewritten

$$x_{t+1} = \operatorname*{arg\,max}_{x \in \mathbb{R}^d} \left\{ f(x) + \frac{1}{2\gamma_t} \|x - x_t\|_2^2 \right\}. \tag{2}$$

Indeed, the function $x \longmapsto f(x) + \frac{1}{2\gamma_t} \|x - x_t\|_2^2$ having at point x_{t+1} a gradient equal to zero is equivalent to Equation (1). The above expression (2) guarantees the existence of x_{t+1} and provides the following interpretation: point x_{t+1} corresponds to a trade-off between minimizing f and being close to the previous iterate x_t. The algorithm can also be written in a variational form: x_{t+1} is characterized by

$$\langle \gamma_t \nabla f(x_{t+1}) + x_{t+1} - x_t | x - x_{t+1} \rangle \geqslant 0, \quad \forall x \in \mathbb{R}^d. \tag{3}$$

The second discretization is the *Euler scheme*, also called the *gradient descent* algorithm:

$$x_{t+1} = x_t - \gamma_t \nabla f(x_t), \tag{4}$$

which is said to be *explicit* because the point x_{t+1} follows from a direct computation involving x_t and $\nabla f(x_t)$, which are known to the algorithm. It can be rewritten

$$x_{t+1} = \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \left\{ \langle \nabla f(x_t) | x \rangle + \frac{1}{2\gamma_t} \|x - x_t\|_2^2 \right\}, \tag{5}$$

which can be seen as a modification of the proximal algorithm (2) where $f(x)$ has been replaced by its linearization at x_t. Its variational form is

$$\langle \gamma_t \nabla f(x_t) + x_{t+1} - x_t | x - x_{t+1} \rangle \geqslant 0, \quad \forall x \in \mathbb{R}^d. \tag{6}$$

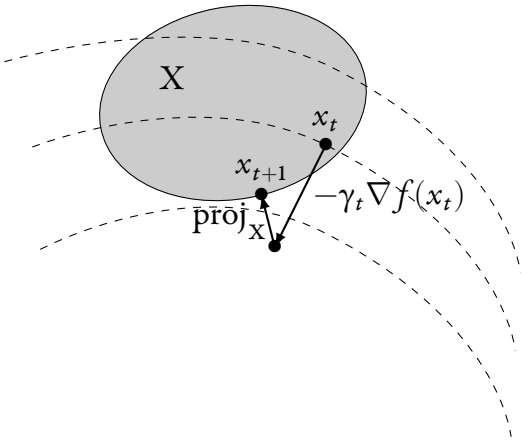


Figure 3. — Projected Subgradient algorithm

Projected Gradient Descent

We now turn to the constrained problem

$$\min_{x \in X} f(x),$$

where X is a convex compact subset of $\mathbb{R}^d$. The gradient descent algorithm (4) can be adapted for this problem by performing a Euclidean projection onto X after each gradient descent step, in order to have all iterates x_t in the set X. This gives the *projected gradient descent* algorithm [Gol64, LP66]:

$$x_{t+1} = \operatorname*{proj}_X \{x_t - \gamma_t \nabla f(x_t)\}, \tag{7}$$

which can rewritten as

$$x_{t+1} = \operatorname*{arg\,min}_{x \in X} \left\{ \langle \nabla f(x_t) | x \rangle + \frac{1}{2\gamma_t} \|x - x_t\|_2^2 \right\}, \tag{8}$$

and has variational characterization:

$$\langle \gamma_t \nabla f(x_t) + x_{t+1} - x_t | x - x_{t+1} \rangle \geqslant 0, \quad \forall x \in X,\ x_{t+1} \in X. \tag{9}$$

Typically, when the gradients of f are assumed to be bounded by $M > 0$ with respect to $\|\cdot\|_2$ (in other words, if f is M-Lipschitz continuous with respect to $\|\cdot\|_2$), the above algorithm with constant step-size $\gamma_t = \|X\|_2 / M\sqrt{T}$ provides a $M/\sqrt{T}$-optimal solution after T steps. When the gradients are bounded by some other norm, the above still applies but the dimension d of the space appears in the bound. For instance, if the gradients are bounded by M with respect to $\|\cdot\|_\infty$, due to the comparison

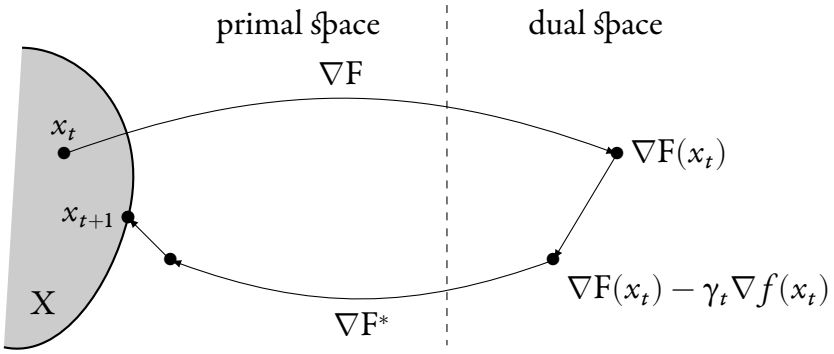


Figure 4. — Greedy Mirror Descent

between the norms, the above algorithm provides after T steps a $M\sqrt{d/T}$-optimal solution. Then, the following question arises: if the gradients are bounded by some other norm than $\|\cdot\|_2$, is it possible to modify the algorithm in order to get a guarantee that has a better dependency in the dimension? This motivates the introduction of Mirror Descent algorithms.

Greedy Mirror Descent

Let $F : \mathbb{R}^d \to \mathbb{R}$ be a differentiable convex function such that $\nabla F : \mathbb{R}^d \to \mathbb{R}^d$ is a bijection. Denote F^* its Legendre–Fenchel transform. Then, one can see that $(\nabla F)^{-1} = \nabla F^*$. We introduce the Bregman divergence associated with F:

$$D_F(x', x) = F(x') - F(x) - \langle \nabla F(x) | x' - x \rangle , \quad x, x' \in \mathbb{R}^d,$$

which is a quadratic quantity that can be interpreted as a generalized distance. It provides a new geometry which will replace the Euclidean structure used for the Projected Gradient Descent (7). The case of the Euclidean distance can be recovered by considering $F(x) = \frac{1}{2}\|x\|_2^2$ which gives $D_F(x', x) = \frac{1}{2}\|x' - x\|_2^2$. The *Greedy Mirror Descent* algorithm [NY83, BT03] is defined by replacing in the Projected Gradient Descent algorithm (8) the Euclidean distance $\frac{1}{2}\|x - x_t\|_2^2$ by the Bregman divergence $D_F(x, x_t)$:

$$x_{t+1} = \arg\min_{x \in X} \left\{ \langle \nabla f(x_t) | x \rangle + \frac{1}{\gamma_t} D_F(x, x_t) \right\}. \tag{10}$$

This algorithm can also be written with the help of a gradient descent and a projection:

$$x_{t+1} = \arg\min_{x \in X} D_F\left(x, \nabla F^*\left(\nabla F(x_t) - \gamma_t \nabla f(x_t)\right)\right). \tag{11}$$

The above expression of x_{t+1} can be decomposed and interpreted as follows. Since

we have forgotten about the Euclidean structure, point x_t belongs to the primal space whereas gradient $\nabla f(x_t)$ lives in the dual space. Therefore, we cannot directly perform the gradient descent $x_t - \gamma_t \nabla f(x_t)$ as in (7). Instead, we first use the map $\nabla \mathrm{F}$ to get from x_t in the primal space to $\nabla \mathrm{F}(x_t)$ in the dual space, and perform the gradient descent there: $\nabla \mathrm{F}(x_t) - \gamma_t \nabla f(x_t)$. We then use the inverse map $\nabla \mathrm{F}^* = (\nabla \mathrm{F})^{-1}$ to come back to the primal space: $\nabla \mathrm{F}^*(\nabla \mathrm{F}(x_t) - \gamma_t \nabla f(x_t))$. Since this point may not belong to the set X, we perform a projection with respect to the Bregman divergence D_F, and we get the expression of x_{t+1} from (11). Let us mention the variational expression of the algorithm, which is much more handy for analysis

$$\langle \gamma_t \nabla f(x_t) + \nabla \mathrm{F}(x_{t+1}) - \nabla \mathrm{F}(x_t) | x - x_{t+1} \rangle \geqslant 0, \quad \forall x \in \mathrm{X},\ x_{t+1} \in \mathrm{X}. \tag{12}$$

As initially wished, the Greedy Mirror Descent algorithm can adapt to different assumptions about the gradients of the objective function f. If f is assumed to be M-Lipschitz continuous with respect to a norm $\|\cdot\|$, the choice of a function F which is K-strongly convex with respect to $\|\cdot\|$ guarantees that the associated algorithm with constant step-size $\gamma_t = \sqrt{\mathrm{LK}}/\mathrm{M}\sqrt{\mathrm{T}}$ gives a $\mathrm{M}\sqrt{\mathrm{L}/\mathrm{KT}}$-optimal solution after T steps, where $\mathrm{L} = \max_{x,x' \in \mathrm{X}} \{\mathrm{F}(x) - \mathrm{F}(x')\}$.

There also exists a proximal version of Greedy Mirror Descent algorithm. It is called the *Bregman Proximal Minimization* algorithm and was introduced by [CZ92]. It is obtained by replacing in the proximal algorithm (2) the Euclidean distance by a Bregman divergence:

$$x_{t+1} = \operatorname*{arg\,min}_{x \in \mathrm{X}} \left\{ f(x) + \frac{1}{\gamma_t} \mathrm{D}_\mathrm{F}(x, x_t) \right\}.$$

Lazy Mirror Descent

We now introduce a variant of the Greedy Mirror Descent algorithm (10) by modifying it as follows. To compute x_{t+1}, instead of considering $\nabla \mathrm{F}(x_t)$, we perform the gradient descent starting from a point y_t (which will be defined in a moment) of the dual space: $y_t - \gamma_t \nabla f(x_t)$. We then map the latter point back to the primal space via $\nabla \mathrm{F}^*$ and then perform the projection onto X with respect to D_F. This gives the *Lazy Mirror Descent* algorithm, also called *Dual Averaging* [Nes09] which starts at some point $x_1 \in \mathrm{X}$ and iterates

$$x_{t+1} = \operatorname*{arg\,min}_{x \in \mathrm{X}} \mathrm{D}_\mathrm{F}\left(x, \nabla \mathrm{F}^*(y_t - \gamma_t \nabla f(x_t))\right). \tag{13}$$

Besides, we perform the update $y_{t+1} = y_t - \gamma_t \nabla f(x_t)$. If the algorithm is started with $y_1 = 0$, we have $y_t = -\sum_{s=1}^{t-1} \gamma_s \nabla f(x_s)$ for all $t \geqslant 1$. Then, one can easily check that

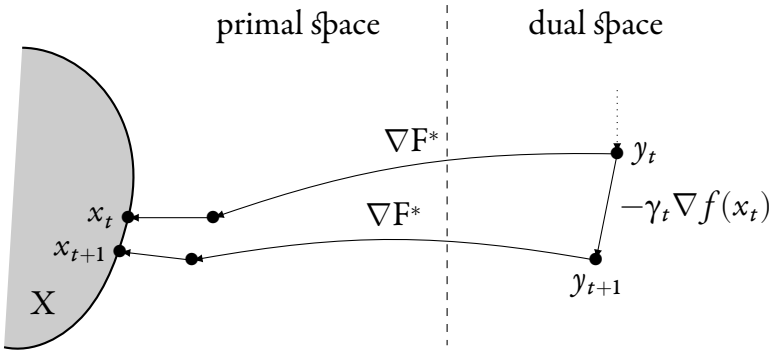


Figure 5. — Lazy Mirror Descent

(13) has the following simpler expression:

$$x_{t+1} = \underset{x \in X}{\arg\min} \left\{ \left\langle \sum_{s=1}^{t} \gamma_s \nabla f(x_s) \middle| x \right\rangle + F(x) \right\}, \tag{14}$$

as well as a variational characterization:

$$\langle \gamma_t \nabla f(x_t) + \nabla F(x_{t+1}) - y_t | x - x_{t+1} \rangle, \quad \forall x \in X,\ x_{t+1} \in X.$$

For the simple problem convex optimization that we are dealing with, this lazy algorithm provides similar guarantees as the greedy version (10)—compare [Nes09, Theorem 4.3] and [BT03, Theorem 4.1]. However, it has a computational advantage over the latter: the iteration in Equation (11) which gives x_{t+1} from x_t involves the successive computation of maps ∇F and ∇F^*, whereas iterating (13) only involves the computation of ∇F^* and the Bregman projection.

Online Mirror Descent

Interestingly, the above convex optimization algorithms can be used for the online convex optimization problem presented above. The first approach of this kind was proposed by [Zin03], who adapted algorithm (7) to the framework where the Decision Maker faces a sequence $(f_t)_{t \geqslant 1}$ of loss functions, instead of a function f that is constant over time. The *Greedy Online Gradient Descent* algorithm is obtained by simply replacing $\nabla f(x_t)$ in (7) by $\nabla f_t(x_t)$:

$$x_{t+1} = \underset{X}{\text{proj}} \{ x_t - \gamma_t \nabla f_t(x_t) \},$$

which can alternatively be written

$$x_{t+1} = \underset{x \in X}{\arg\min} \left\{ \langle \nabla f_t(x_t) | x \rangle + \frac{1}{2\gamma_t} \|x - x_t\|_2^2 \right\}.$$

By introducing a function F satisfying the same assumptions as in the previous section, we extend the above to a family of Greedy Online Mirror Descent algorithms [Bub11, BCB12]:

$$x_{t+1} = \operatorname*{arg\,min}_{x \in \mathrm{X}} \left\{ \langle \nabla f_t(x_t) | x \rangle + \frac{1}{\gamma_t} \mathrm{D}_{\mathrm{F}}(x, x_t) \right\}. \tag{15}$$

Similarly, we can also define a lazy version [SS07, SS11, KSST12, OCCB15]:

$$x_{t+1} = \operatorname*{arg\,min}_{x \in \mathrm{X}} \left\{ \left\langle \sum_{s=1}^{t} \gamma_s \nabla f_s(x_s) \middle| x \right\rangle + \mathrm{F}(x) \right\}. \tag{16}$$

More generally, we can define the above algorithms by replacing the gradients $\nabla f_t(x_t)$ by arbitrary vectors $u_t \in \mathbb{R}^d$ which need not be the gradients of some functions f_t. For instance, the Lazy Online Mirror Descent algorithm can be written:

$$x_{t+1} = \operatorname*{arg\,max}_{x \in \mathrm{X}} \left\{ \left\langle \sum_{s=1}^{t} u_s \middle| x \right\rangle - \mathrm{F}(x) \right\},$$

where F acts a *regularizer*. This motivates, for this algorithm, the alternative name: *Follow the Regularized Leader* [AHR08, RT09, AHR12]. This algorithm provides a guarantee on:

$$\max_{x \in \mathrm{X}} \sum_{t=1}^{\mathrm{T}} \langle u_t | x \rangle - \sum_{t=1}^{\mathrm{T}} \langle u_t | x_t \rangle,$$

which is the same quantity as in Equation (∗), i.e. the regret in the online linear optimization problem with *payoff vectors* $(u_t)_{t \geqslant 1}$. An important property is that payoff vector u_t is allowed to depend on x_t, as it is the case in (16) where $u_t = -\gamma_t \nabla f(x_t)$. This Lazy Online Mirror Descent family of algorithms will be our subject of study in Chapters I to IV. Throughout Part I of the manuscript, unless mentioned otherwise, *Mirror Descent* will designate the Lazy Online Mirror Descent algorithms.

FIRST PART

CHAPTER I

MIRROR DESCENT FOR REGRET MINIMIZATION

We present the regret minimization problem called *online linear optimization*. Some convexity tools are introduced, with a special focus on strong convexity. We then construct the family of Mirror Descent strategies with time-varying parameters and derive general regret guarantees in Theorem I.3.1. This result is central as most results in Part I will be obtained as corollaries. In Section I.4, we present the generalization to convex losses (instead of linear payoffs), and in Section I.5, we turn the aforementioned regret minimizing strategies into convex optimization algorithms.

I.1. Core model

The model we present here is called *online linear optimization*. It is a repeated play between a Decision Maker and Nature. Let $\mathcal{V}$ be a finite-dimensional vector space, $\mathcal{V}^*$ its dual space, and denote $\langle \cdot | \cdot \rangle$ the dual pairing. $\mathcal{V}^*$ will be called the *payoff space* [1]. Let $\mathcal{Z}$ be a nonempty convex compact subset of $\mathcal{V}$, which will be the set of actions of the Decision Maker. At each time instance $t \geqslant 1$, the Decision Maker

- chooses an action $z_t \in \mathcal{Z}$;
- observes a payoff vector $u_t \in \mathcal{V}^*$ chosen by Nature;
- gets a payoff equal to $\langle u_t | z_t \rangle$.

Formally, a strategy for the Decision Maker is a sequence of maps $\sigma = (\sigma_t)_{t \geqslant 1}$ where $\sigma_t : (\mathcal{Z} \times \mathcal{V}^*)^{t-1} \to \mathcal{Z}$. In a slight abuse of notation, σ_1 will be regarded as an element of $\mathcal{Z}$. For a given strategy σ and a given sequence $(u_t)_{t \geqslant 1}$ of payoff vectors, the sequence of play $(z_t)_{t \geqslant 1}$ is defined by

$$z_t = \sigma_t(z_1, u_1, \dots, z_{t-1}, u_{t-1}), \quad t \geqslant 1.$$

1. The dimension being finite, it would be good enough to work in $\mathbb{R}^d$. However, we believe that the theoretical distinction between the primal and dual spaces helps with the understanding of Mirror Descent strategies.

Concerning Nature, we assume it to be omniscient. Indeed, our main result, Theorem I.3.1, will provide guarantees that hold against any sequence of payoff vectors. Therefore, its choice of payoff vector u_t may depend on everything that has happened before he has to reveal it. In particular, payoff vector u_t may depend on action z_t.

The quantity of interest is the *regret* (up to time $T \geqslant 1$), defined by

$$\mathrm{Reg}_T\left\{\sigma, (u_t)_{t\geqslant 1}\right\} = \max_{z\in\mathcal{Z}} \sum_{t=1}^{T} \langle u_t | z\rangle - \sum_{t=1}^{T} \langle u_t | z_t\rangle, \quad T \geqslant 1.$$

In most situations, we simply write Reg_T since the strategy and the payoffs vectors will be clear from the context. In the case where Nature's choice of payoff vectors $(u_t)_{t\geqslant 1}$ does not depend on the actions of the Decision maker (Nature is then said to be *oblivious*), the regret can be interpreted as follows. It compares the cumulative payoff $\sum_{t=1}^{T} \langle u_t | z_t\rangle$ obtained by the Decision Maker to the best cumulative payoff $\sum_{t=1}^{T} \langle u_t | z\rangle$ that he could have obtained by playing a fixed action $z \in \mathcal{Z}$ at each stage. It therefore measures how much the Decision Maker *regrets* not having played the constant strategy that turned out to be the best. When Nature is not assumed to be oblivious (it is then said to be *adversarial*), in other words, when Nature can react to the actions $(z_t)_{t\geqslant 1}$ chosen by the Decision Maker, the regret is still well-defined and every result below will stand. The only difference is that the above *interpretation* of the regret is not valid.

The first goal is to construct strategies for the Decision Maker which guarantee that the average regret $\frac{1}{T}R_T$ is asymptotically nonpositive when the payoff vectors are assumed to be bounded. In Section I.3 we construct the Mirror Descent strategies and derive in Theorem I.3.1 general upper bounds on the regret which yield such guarantees.

One of the simplest strategies one can think of is called *Follow the Leader* or *Fictitious Play*. It consists in playing the action which would have given the highest cumulative payoff over the previous stages, had it been played at each stage:

$$z_t \in \arg\max_{z\in\mathcal{Z}} \left\langle \sum_{s=1}^{t-1} u_s \middle| z \right\rangle. \tag{I.1}$$

Unfortunately, this strategy does not guarantee the average regret to be asymptotically nonpositive, even in the following simple setting where the payoff vectors are bounded. Consider the framework where $\mathcal{V} = \mathcal{V}^* = \mathbb{R}^2$, $\mathcal{Z} = \Delta_2 = \{(z_1, z_2) \in \mathbb{R}^2_+ \,|\, z_1 + z_2 = 1\}$ and where the payoff vectors all belong to $[0,1]^2$. Suppose that Nature chooses payoff vectors

$$u_1 = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}, \quad u_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad u_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad u_4 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad u_5 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \ldots$$

Then, one can easily see that using the above strategy (I.1) gives for $t \geqslant 2$, $z_t = (1,0)$ if t is even, and $z_t = (0,1)$ if t is odd. As a result, the payoff $\langle u_t | z_t \rangle$ is zero as soon as $t \geqslant 2$. The Decision Maker is choosing at each stage, the action which gives the worst payoff. As far as the regret is concerned, since $\max_{z \in \mathcal{Z}} \sum_{t=1}^{T} \langle u_t | z \rangle$ is of order $T/2$, the regret grows linearly in T. Therefore, the average regret is not asymptotically nonpositive. This phenomenon is called *overfitting*: following too closely previous data may result in bad predictions. To overcome this problem, we can try modifying strategy (I.1) as

$$z_t = \operatorname*{arg\,max}_{z \in \mathcal{Z}} \left\{ \left\langle \sum_{s=1}^{t-1} u_s \middle| z \right\rangle - h(z) \right\},$$

where we introduced a function h in order to *regularize* the strategy. This is the key idea behind the *Mirror Descent* strategies (which are also called *Follow the Regularized Leader*) that we will define and study in Section I.3.

I.2. Regularizers

We here introduce a few tools from convex analysis needed for the construction and the analysis of the Mirror Descent strategies. These are classic (see e.g. [SS07, SS11, Bub11]) and the proofs are given for the sake of completeness. Again, $\mathcal{V}$ and $\mathcal{V}^*$ are finite-dimensional vectors spaces and $\mathcal{Z}$ is a nonempty convex compact subset of $\mathcal{V}$. We define regularizers, present the notion of strong convexity with respect to an arbitrary norm, and give three examples of regularizers along with their properties.

I.2.1. Definition and properties

We recall that the *domain* $\operatorname{dom} h$ of a function $h : \mathcal{V} \to \mathbb{R} \cup \{+\infty\}$ is the set of points where it has finite values.

Definition I.2.1. A convex function $h : \mathcal{V} \to \mathbb{R} \cup \{+\infty\}$ is a *regularizer* on $\mathcal{Z}$ if it is strictly convex, lower semicontinuous, and has $\mathcal{Z}$ as domain. We then denote $\delta_h = \max_{\mathcal{Z}} h - \min_{\mathcal{Z}} h$ the difference between its maximal and minimal values on $\mathcal{Z}$.

Proposition I.2.2. *Let h be a regularizer on $\mathcal{Z}$. Its Legendre–Fenchel transform $h^* : \mathcal{V}^* \to \mathbb{R} \cup \{+\infty\}$, defined by*

$$h^*(w) = \sup_{z \in \mathcal{V}} \{ \langle w | z \rangle - h(z) \}, \quad w \in \mathcal{V}^*,$$

satisfies the following properties.

(i) $\operatorname{dom} h^* = \mathcal{V}^*$;

(ii) h^* *is differentiable on* $\mathcal{V}^*$;

(iii) For all $w \in \mathcal{V}^*$, $\nabla h^*(w) = \arg\max_{z\in\mathcal{Z}} \{\langle w|z\rangle - h(z)\}$. *In particular,* ∇h^* *takes values in* $\mathcal{Z}$.

Proof. (i) Let $w \in \mathcal{V}^*$. The function $z \longmapsto \langle w|z\rangle - h(z)$ equals $-\infty$ outside of $\mathcal{Z}$, and is upper semicontinuous on $\mathcal{Z}$ which is compact. It thus has a maximum and $h^*(w) < +\infty$.

(ii, iii) Moreover, this maximum is attained at a unique point because h is strictly convex. Besides, for $z \in \mathcal{V}$ and $w \in \mathcal{V}^*$

$$z \in \partial h^*(w) \quad \Longleftrightarrow \quad w \in \partial h(z) \quad \Longleftrightarrow \quad z \in \arg\max_{z'\in\mathcal{Z}} \{\langle w|z'\rangle - h(z')\},$$

in other words, $\partial h^*(w) = \arg\max_{z'\in\mathcal{Z}} \{\langle w|z'\rangle - h(z')\}$. This argmax is a singleton as we noticed. It means that h^* is differentiable. □

Remark I.2.3. The above proposition demonstrates that h^* is a smooth approximation of $\max_{z\in\mathcal{Z}} \langle \cdot\,|z\rangle$ and that ∇h^* is an approximation of $\arg\max_{z\in\mathcal{Z}} \langle \cdot\,|z\rangle$. They will be used in Section I.3 in the construction and the analysis of the Mirror Descent strategies.

As soon as h is a regularizer, the Bregman divergence of h^* is well defined:

$$\mathrm{D}_{h^*}(w', w) = h^*(w') - h^*(w) - \langle \nabla h^*(w) | w' - w\rangle, \qquad w, w' \in \mathcal{V}^*.$$

This quantity will appear in the fundamental regret bound of Theorem I.3.1. As we will see below in Proposition I.2.8, by adding a strong convexity assumption on the regularizer h, the Bregman divergence can be bounded from above by a much more explicit quantity.

I.2.2. Strong convexity

Definition I.2.4. Let $h : \mathcal{V} \to \mathbb{R} \cup \{+\infty\}$ be a function, $\|\cdot\|$ a norm on $\mathcal{V}$, and $\mathrm{K} > 0$. h is K-strongly convex with respect to $\|\cdot\|$ if for all $z, z' \in \mathcal{V}$ and $\lambda \in [0, 1]$,

$$h(\lambda z + (1-\lambda)z') \leqslant \lambda h(z) + (1-\lambda)h(z') - \frac{\mathrm{K}\lambda(1-\lambda)}{2}\|z' - z\|^2. \tag{I.2}$$

Proposition I.2.5. *Let* $h : \mathcal{V} \to \mathbb{R} \cup \{+\infty\}$ *be a function,* $\|\cdot\|$ *a norm on* $\mathcal{V}$, *and* $\mathrm{K} > 0$. *The following conditions are equivalent.*

(i) h *is* K*-strongly convex with respect to* $\|\cdot\|$;

(ii) For all points $z, z' \in \mathcal{V}$ *and all subgradients* $w \in \partial h(z)$,

$$h(z') \geqslant h(z) + \langle w|z' - z\rangle + \frac{\mathrm{K}}{2}\|z' - z\|^2; \tag{I.3}$$

(iii) For all points $z, z' \in \mathscr{V}$ and all subgradients $w \in \partial h(z)$ and $w' \in \partial h(z')$,

$$\langle w' - w | z' - z \rangle \geqslant \mathrm{K} \|z' - z\|^2 . \tag{I.4}$$

Proof. (i) $\Longrightarrow$ (ii). We assume that h is K-strongly convex with respect to $\|\cdot\|$. In particular, h is convex. Let $z, z' \in \mathscr{V}$, $w \in \partial h(z)$, $\lambda \in (0,1)$, and denote $z'' = \lambda z + (1-\lambda)z'$. Using the convexity of h, we have

$$\begin{aligned}
\langle w | z' - z \rangle &= \frac{\langle w | z'' - z \rangle}{1-\lambda} \leqslant \frac{h(z'') - h(z)}{1-\lambda} \\
&\leqslant \frac{1}{1-\lambda} \left(\lambda h(z) + (1-\lambda) h(z') - \frac{\mathrm{K}\lambda(1-\lambda)}{2} \|z' - z\|^2 - h(z) \right) \\
&= h(z') - h(z) - \frac{\mathrm{K}\lambda}{2} \|z' - z\|^2 ,
\end{aligned}$$

and (I.3) follows from taking $\lambda \to 1$.

(ii) $\Longrightarrow$ (i). Let $z, z' \in \mathscr{V}$, $\lambda \in [0,1]$, denote $z'' = \lambda z + (1-\lambda)z'$. If $\lambda \in \{0,1\}$, inequality (I.2) is trivial. We now assume $\lambda \in (0,1)$. If z or z' does not belong to the domain of h, inequality (I.2) is also trivial. We now assume $z, z' \in \operatorname{dom} h$. Then, z'' belongs to $]z, z'[$ which is a subset of the relative interior of $\operatorname{dom} h$. Therefore, $\partial h(z'')$ is nonempty (see e.g. [Roc70, Theorem 23.4]). Let $w \in \partial h(z'')$. We have

$$\begin{aligned}
\langle w | z - z'' \rangle &\leqslant h(z) - h(z'') - \frac{\mathrm{K}}{2} \|z - z''\|^2 \\
\langle w | z' - z'' \rangle &\leqslant h(z') - h(z'') - \frac{\mathrm{K}}{2} \|z' - z''\|^2 .
\end{aligned}$$

By multiplying the above inequalities by λ and $1-\lambda$ respectively, and summing, we get

$$0 \leqslant \lambda h(z) + (1-\lambda) h(z') - h(z'') - \frac{\mathrm{K}}{2} \left(\lambda \|z - z''\|^2 + (1-\lambda) \|z' - z''\|^2 \right) .$$

Using the definition of z'', we have $z - z'' = (1-\lambda)(z' - z)$ and $z' - z'' = \lambda(z' - z)$. The last term of the above right-hand side is therefore equal to

$$\frac{\mathrm{K}}{2} \left(\lambda(1-\lambda)^2 \|z' - z\|^2 + (1-\lambda)\lambda^2 \|z' - z\|^2 \right) = \frac{\mathrm{K}\lambda(1-\lambda)}{2} \|z' - z\|^2 ,$$

and (I.2) is proved.

(ii) $\Longrightarrow$ (iii). Let $z, z' \in \mathscr{V}$, $w \in \partial h(z)$ and $w' \in \partial h(z')$. We have

$$h(z') \geqslant h(z) + \langle w | z' - z \rangle + \frac{\mathrm{K}}{2} \|z' - z\|^2 \tag{I.5}$$

$$h(z) \geqslant h(z') + \langle w' | z - z' \rangle + \frac{\mathrm{K}}{2} \|z' - z\|^2 . \tag{I.6}$$

Summing both inequalities and simplifying gives (I.4).

(iii) $\Longrightarrow$ (ii). Let $z, z' \in \mathscr{V}$. If $\partial h(z)$ is empty, condition (ii) is automatically satisfied. We now assume $\partial h(z) \neq \emptyset$. In particular, $z \in \operatorname{dom} h$. Let $w \in \partial h(z)$. If $h(z') = +\infty$, inequality (I.3) is satisfied. We now assume $z' \in \operatorname{dom} h$. Therefore, we have that $]z, z'[$ is a subset of the relative interior of $\operatorname{dom} h$. As a consequence, for all points $z'' \in]z, z'[$, we have $\partial h(z'') \neq \emptyset$ (see e.g. [Roc70, Theorem 23.4]). For all $\lambda \in [0,1]$, we define $z_\lambda = z + \lambda(z' - z)$. Using the convexity of h, we can now write, for all $n \geqslant 1$,

$$h(z') - h(z) = \sum_{k=1}^{n} h(z_{k/n}) - h(z_{(k-1)/n}) \geqslant \sum_{k=1}^{n} \left\langle w_{(k-1)/n} \middle| z_{k/n} - z_{(k-1)/n} \right\rangle ,$$

where $w_0 = w$ and $w_{k/n} \in \partial h(z_{k/n})$ for $k \geqslant 1$. Since $z_{k/n} - z_{(k-1)/n} = \frac{1}{n}(z' - z)$ for $k \geqslant 1$, subtracting $\langle w | z' - z \rangle$ we get

$$h(z') - h(z) - \langle w | z' - z \rangle \geqslant \frac{1}{n} \sum_{k=1}^{n} \left\langle w_{(k-1)/n} - w \middle| z' - z \right\rangle .$$

Note that the first term of the above sum is zero because $w = w_0$. Besides, for $k \geqslant 2$, we have $z' - z = \frac{n}{k-1}(z_{(k-1)/n} - z)$. Therefore, and this is where we use (iii),

$$\begin{aligned} h(z') - h(z) - \langle w | z' - z \rangle &\geqslant \sum_{k=2}^{n} \frac{1}{k-1} \left\langle w_{(k-1)/n} - w \middle| z_{(k-1)/n} - z \right\rangle \\ &\geqslant \mathrm{K} \sum_{k=2}^{n} \frac{1}{k-1} \left\| z_{(k-1)/n} - z \right\|^2 \\ &= \frac{\mathrm{K} \left\| z' - z \right\|^2}{n^2} \sum_{k=2}^{n} (k-1) \\ &\xrightarrow[n \to +\infty]{} \frac{\mathrm{K}}{2} \left\| z' - z \right\|^2 , \end{aligned}$$

and (ii) is proved. □

Similarly to usual convexity, there exists a strong convexity criterion involving the Hessian for twice differentiable functions.

Proposition I.2.6. *Let $\|\cdot\|$ be a norm on $\mathscr{V}$, $\mathrm{K} > 0$, and $\mathrm{F} : \mathscr{V} \to \mathbb{R}$ a twice differentiable function such that*

$$\left\langle \nabla^2 \mathrm{F}(z) u \middle| u \right\rangle \geqslant \mathrm{K} \left\| u \right\|^2 , \quad z \in \mathscr{V},\ u \in \mathscr{V}.$$

Then, F is K-strongly convex with respect to $\|\cdot\|$.

Proof. Let $z, z' \in \mathscr{V}$. Let us prove the condition (ii) from Proposition I.2.5. We define

$$\phi(\lambda) = F(z + \lambda(z' - z)), \quad \lambda \in [0,1].$$

By differentiating twice, we get for all $\lambda \in [0,1]$:

$$\phi''(\lambda) = \langle \nabla^2 F(z + \lambda(z' - z))(z' - z) | z' - z \rangle \geqslant K \|z' - z\|^2 .$$

There exists $\lambda_0 \in [0,1]$ such that $\phi(1) = \phi(0) + \phi'(0) + \phi''(\lambda_0)/2$. This gives

$$F(z') = \phi(1) = \phi(0) + \phi'(0) + \frac{\phi''(\lambda_0)}{2} \geqslant F(z) + \langle \nabla F(z) | z' - z \rangle + \frac{K}{2} \|z' - z\|^2 ,$$

and (I.3) is proved. □

Lemma I.2.7. *Let $\|\cdot\|$ a norm on $\mathscr{V}$, $K > 0$ and $h, F : \mathscr{V} \to \mathbb{R} \cup \{+\infty\}$ two convex functions such that for all $z \in \mathscr{V}$,*

$$h(z) = F(z) \quad \text{or} \quad h(z) = +\infty.$$

Then, if F is K-strongly convex with respect to $\|\cdot\|$, so is h.

Proof. Note that for all $z \in \mathscr{V}$, $F(z) \leqslant h(z)$. Let us prove that h satisfies the condition from Definition I.2.4. Let $z, z' \in \mathscr{V}$, $\lambda \in [0,1]$ and denote $z'' = \lambda z + (1-\lambda)z'$. Let us first assume that $h(z'') = +\infty$. By convexity of h, either $h(z)$ or $h(z')$ is equal to $+\infty$, and the right-hand side of (I.2) is equal to $+\infty$. Inequality (I.2) therefore holds. If $h(z'')$ is finite,

$$\begin{aligned} h(z'') = F(z'') &\leqslant \lambda F(z) + (1-\lambda) F(z') - \frac{K\lambda(1-\lambda)}{2} \|z' - z\|^2 \\ &\leqslant \lambda h(z) + (1-\lambda) h(z') - \frac{K\lambda(1-\lambda)}{2} \|z' - z\|^2 , \end{aligned}$$

and (I.2) is proved. □

For a given norm $\|\cdot\|$ on $\mathscr{V}$, the dual norm $\|\cdot\|_*$ on $\mathscr{V}^*$ is defined by

$$\|w\|_* = \sup_{\|z\| \leqslant 1} |\langle w | z \rangle| .$$

Proposition I.2.8. *Let $K > 0$ and $h : \mathscr{V} \to \mathbb{R} \cup \{+\infty\}$ be a regularizer which we assume to be K-strongly convex function with respect to a norm $\|\cdot\|$ on $\mathscr{V}$. Then,*

$$D_{h^*}(w', w) \leqslant \frac{1}{2K} \|w' - w\|_*^2 , \quad w, w' \in \mathscr{V}^* .$$

Proof. Let $w, w' \in \mathcal{Y}^*$ and denote $z = \nabla h^*(w)$ and $z' = \nabla h^*(w')$. Moreover, for $\lambda \in [0,1]$, we introduce $w_\lambda = w + \lambda(w' - w)$ and $z_\lambda = \nabla h^*(w_\lambda)$. Therefore, we have $w \in \partial h(z)$ and $w_\lambda \in \partial h(z_\lambda)$. h being strongly convex, condition (I.4) gives $\langle w_\lambda - w | z_\lambda - z \rangle \geqslant K \|z_\lambda - z\|^2$. Using the definition of $\|\cdot\|_*$ and dividing by $\|z_\lambda - z\|$ gives

$$\|z_\lambda - z\| \leqslant \frac{1}{K} \|w_\lambda - w\|_* .$$

Now consider $\phi(\lambda) = h^*(w_\lambda)$ defined for $\lambda \in [0,1]$. We have

$$\begin{aligned}
\phi'(\lambda) - \phi'(0) &= \langle w' - w | \nabla h^*(w_\lambda) - \nabla h^*(w) \rangle = \langle w' - w | z_\lambda - z \rangle \\
&\leqslant \|w' - w\|_* \|z_\lambda - z\| \leqslant \frac{1}{K} \|w_\lambda - w\|_* \|w' - w\|_* \\
&= \frac{\lambda}{K} \|w' - w\|_*^2 .
\end{aligned}$$

By integrating, we get

$$\phi(\lambda) - \phi(0) \leqslant \phi'(0)\lambda + \frac{\lambda^2}{2K} \|w' - w\|_*^2 ,$$

which for $\lambda = 1$ boils down to

$$h^*(w') - h^*(w) \leqslant \langle w' - w | \nabla h^*(w) \rangle + \frac{1}{2K} \|w' - w\|_*^2 .$$

In other words, $D_{h^*}(w', w) \leqslant \frac{1}{2K} \|w' - w\|_*^2$. □

I.2.3. The Entropic regularizer

Denote Δ_d the unit simplex of $\mathbb{R}^d$:

$$\Delta_d = \left\{ z \in \mathbb{R}_+^d \,\middle|\, \sum_{i=1}^d z^i = 1 \right\},$$

where $\mathbb{R}_+^d$ is the set of vectors in $\mathbb{R}^d$ with nonnegative components. We define the entropic regularizer $h_{\text{ent}} : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ as

$$h_{\text{ent}}(z) = \begin{cases} \sum_{i=1}^d z^i \log z^i & \text{if } z \in \Delta_d \\ +\infty & \text{otherwise,} \end{cases}$$

where $z^i \log z^i = 0$ when $z^i = 0$.

Proposition I.2.9. *(i) h_{ent} is a regularizer on Δ_d;*

(ii) $h^*_{\text{ent}}(w) = \log\left(\sum_{i=1}^{d} \exp w^i\right)$, *for all* $w \in \mathbb{R}^d$;

(iii) $\nabla h^*_{\text{ent}}(w) = \left(\frac{\exp w^i}{\sum_{j=1}^{d} \exp w^j}\right)_{1\leqslant j\leqslant d}$, *for all* $w \in \mathbb{R}^d$;

(iv) $\delta_{h_{\text{ent}}} = \log d$;

(v) h_{ent} *is* 1*-strongly convex with respect to* $\|\cdot\|_1$.

Proof. (i) is immediate, and (ii) and (iii) are classic (see e.g. [BV04, Example 2.25]).

(iv) h_{ent} being convex, its maximum on Δ_d is attained at one of the extreme points. At each extreme point, the value of h_{ent} is zero. Therefore, $\max_{\Delta_d} h_{\text{ent}} = 0$. As for the minimum, h_{ent} being convex and symmetric with respect to the components z^i, its minimum is attained at the centroid $(1/d, \dots, 1/d)$ of the simplex Δ_d, where its value is $-\log d$. Therefore, $\min_{\Delta_d} h_{\text{ent}} = -\log d$ and $\delta_{h_{\text{ent}}} = \log d$.

(v) Consider $\mathrm{F} : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ defined by

$$\mathrm{F}(z) = \begin{cases} \sum_{i=1}^{d} (z^i \log z^i - z^i) + 1 & \text{if } z \in \mathbb{R}^d_+ \\ +\infty & \text{otherwise.} \end{cases}$$

Let us prove that F is 1-strongly convex with respect to $\|\cdot\|_1$. By definition, the domain of F is $\mathbb{R}^d_+$. It is differentiable on the interior of the domain $(\mathbb{R}^*_+)^d$ and $\nabla\mathrm{F}(z) = (\log z^i)_{1\leqslant i\leqslant d}$ for $z \in (\mathbb{R}^*_+)^d$. Therefore, the norm of $\nabla\mathrm{F}(z)$ goes to $+\infty$ when z converges to a boundary point of $\mathbb{R}^d_+$. [Roc70, Theorem 26.1] then assures that the subdifferential $\partial\mathrm{F}(z)$ is empty as soon as $z \notin (\mathbb{R}^*_+)^d$. Therefore, condition (iii) from Proposition I.2.5, which we aim at proving, can be written

$$\langle \nabla\mathrm{F}(z') - \nabla\mathrm{F}(z) | z' - z \rangle \geqslant \|z' - z\|_1^2, \quad z, z' \in (\mathbb{R}^*_+)^d. \tag{I.7}$$

Let $z, z' \in (\mathbb{R}^*_+)^d$.

$$\langle \nabla\mathrm{F}(z') - \nabla\mathrm{F}(z) | z' - z \rangle = \sum_{i=1}^{d} \log \frac{(z')^i}{z^i} ((z')^i - z^i).$$

A simple study of function shows that $(s-1)\log s - 2(s-1)^2/(s+1) \geqslant 0$ for $s \geqslant 0$. Applied with $s = (z')^i/z^i$, this gives

$$\sum_{i=1}^{d} \log \frac{(z')^i}{z^i} ((z')^i - z^i) \geqslant \|z' - z\|_1^2,$$

and (I.7) is proved. F is therefore 1-strongly convex with respect to $\|\cdot\|_1$ and so is h_{ent} thanks to Lemma I.2.7. □

I.2.4. The Euclidean regularizer

Let $\mathcal{Z}$ be a nonempty convex compact subset of $\mathbb{R}^d$. We define the Euclidean regularizer on $\mathcal{Z}$ as

$$h_2(z) = \begin{cases} \frac{1}{2}\|z\|_2^2 & \text{if } z \in \mathcal{Z} \\ +\infty & \text{otherwise.} \end{cases}$$

Proposition I.2.10. *(i) h_2 is a regularizer on $\mathcal{Z}$;*

(ii) $\nabla h_2^*(w) = \operatorname{proj}_{\mathcal{Z}}(w)$ for all $w \in \mathbb{R}^d$ where $\operatorname{proj}_{\mathcal{Z}}$ is the Euclidean projection onto $\mathcal{Z}$;

(iii) h_2 is 1-strongly convex with respect to $\|\cdot\|_2$.

Proof. (i) is immediate.

(ii) For all $w \in \mathbb{R}^d$, using property (iii) from Proposition I.2.2,

$$\begin{aligned}\nabla h^*(w) &= \operatorname*{arg\,max}_{z\in\mathcal{Z}} \left\{ \langle w|z\rangle - \frac{1}{2}\|z\|_2^2 \right\} = \operatorname*{arg\,min}_{z\in\mathcal{Z}} \left\{ \frac{1}{2}\|z\|_2^2 - \langle w|z\rangle + \frac{1}{2}\|w\|_2^2 \right\} \\ &= \operatorname*{arg\,min}_{z\in\mathcal{Z}} \|w - z\|_2^2 = \operatorname*{proj}_{\mathcal{Z}}(w).\end{aligned}$$

(iii) We consider $F : \mathbb{R}^d \to \mathbb{R}$ defined by $F(z) = \frac{1}{2}\|z\|_2^2$ for all $z \in \mathbb{R}^d$. Its Hessian at all points $z \in \mathcal{V}$ is the identity matrix and for all vectors $u \in \mathbb{R}^d$, we have

$$\langle \nabla^2 F(z)u|u\rangle = \|u\|_2^2 .$$

Thanks to Proposition I.2.6, F is 1-strongly convex with respect to $\|\cdot\|_2$. Using Lemma I.2.7, we deduce that h_2 is also 1-strongly convex with respect to $\|\cdot\|_2$. □

I.2.5. The ℓ^p regularizer

For $p \in (1, 2)$, we define for any nonempty convex compact subset $\mathcal{Z}$ of $\mathbb{R}^d$:

$$h_p(z) = \begin{cases} \frac{1}{2}\|z\|_p^2 & \text{if } z \in \mathcal{Z} \\ +\infty & \text{otherwise.} \end{cases}$$

Proposition I.2.11. *(i) h_p is a regularizer on $\mathcal{Z}$;*

(ii) h_p is $(p-1)$-strongly convex with respect to $\|\cdot\|_p$.

Proof. (i) Since $p \geqslant 1$, $\|\cdot\|_p$ is a norm and is therefore convex. h_p then clearly is a regularizer on $\mathcal{Z}$.

(ii) We consider the function $F(z) = \frac{1}{2}\|z\|_p^2$ defined on $\mathbb{R}^d$ which is $(p-1)$-strongly convex with respect to $\|\cdot\|_p$ (see e.g. [Bub11, Lemma 3.21]). Then, so is h_p thanks to Lemma I.2.7. □

I.3. Mirror Descent strategies

We now construct the family of Mirror Descent strategies with time-varying parameters and derive in Theorem I.3.1 general regret bounds. A discussion on the origins of Mirror Descent is provided in the introduction of the manuscript. We consider the notation introduced in Section I.1. Let h be a regularizer on the action set $\mathcal{Z}$ and $(\eta_t)_{t\geqslant 1}$ a positive and nonincreasing sequence of parameters. The Mirror Descent strategy associated with h and $(\eta_t)_{t\geqslant 1}$ is defined by $U_0 = 0$ and for $t \geqslant 1$ by

$$\begin{aligned} \text{play action} \quad & z_t = \nabla h^*(\eta_{t-1} U_{t-1}), \\ \text{update} \quad & U_t = U_{t-1} + u_t, \end{aligned}$$

which implies $U_t = \sum_{s=1}^t u_s$. Since ∇h^* takes values in $\mathcal{Z}$ by Proposition I.2.2, z_t is indeed an action. Besides, z_t only depends on payoff vectors up to time $t-1$. Therefore, the above is a valid strategy. Using property (iii) from Proposition I.2.2, it can also be written

$$z_t = \operatorname*{arg\,max}_{z\in\mathcal{Z}} \left\{ \left\langle \sum_{s=1}^{t-1} u_s \middle| z \right\rangle - \frac{h(z)}{\eta_{t-1}} \right\}.$$

This expression clearly demonstrates that the strategy is a regularized version of Follow the Leader (I.1) which would give $\operatorname{arg\,max}_{z\in\mathcal{Z}} \left\langle \sum_{s=1}^{t-1} u_s \middle| z \right\rangle$ instead. Moreover, we see that the higher is parameter η_{t-1}, the closer z_t is to $\operatorname{arg\,max}_{z\in\mathcal{Z}} \left\langle \sum_{s=1}^{t-1} u_s \middle| z \right\rangle$. This intuition is in particular useful in Section II.7 where we compare the regret bounds given by different choices of parameters $(\eta_t)_{t\geqslant 1}$.

We now state the general regret bound guaranteed by this strategy. Similar statements with constant parameters have appeared in e.g. [RT09, Proposition 11], [SS11, Lemma 2.20] and [BCB12, Theorem 5.4].

Theorem I.3.1. *Let $T \geqslant 1$ an integer and $M, K > 0$.*

(i) Against any sequence $(u_t)_{t\geqslant 1}$ of payoff vectors, the above strategy guarantees

$$\mathrm{Reg}_T \leqslant \frac{\delta_h}{\eta_T} + \sum_{t=1}^T \frac{1}{\eta_{t-1}} D_{h^*}(\eta_{t-1} U_t, \eta_{t-1} U_{t-1}),$$

where we set $\eta_0 = \eta_1$.

(ii) Moreover, if h is K-strongly convex with respect to a norm $\|\cdot\|$, then

$$\mathrm{Reg}_T \leqslant \frac{\delta_h}{\eta_T} + \frac{1}{2K} \sum_{t=1}^T \eta_{t-1} \|u_t\|_*^2 .$$

(iii) Moreover, if $\|u_t\|_* \leqslant \mathrm{M}$ *(for all* $t \geqslant 1$*), the choice* $\eta_t = \sqrt{\delta_h \mathrm{K}/\mathrm{M}^2 t}$ *(for* $t \geqslant 1$*) guarantees*

$$\mathrm{Reg}_\mathrm{T} \leqslant 2\mathrm{M}\sqrt{\frac{\mathrm{T}\delta_h}{\mathrm{K}}}.$$

Proof. (i) Let $z \in \mathscr{Z}$. Using Fenchel's inequality, we write

$$\begin{aligned}\langle \mathrm{U}_\mathrm{T} | z \rangle = \frac{\langle \eta_\mathrm{T} \mathrm{U}_\mathrm{T} | z \rangle}{\eta_\mathrm{T}} &\leqslant \frac{h^*(\eta_\mathrm{T}\mathrm{U}_\mathrm{T})}{\eta_\mathrm{T}} + \frac{h(z)}{\eta_\mathrm{T}} \\ &\leqslant \frac{h^*(0)}{\eta_0} + \sum_{t=1}^{\mathrm{T}} \left(\frac{h^*(\eta_t \mathrm{U}_t)}{\eta_t} - \frac{h^*(\eta_{t-1}\mathrm{U}_{t-1})}{\eta_{t-1}} \right) + \frac{\max_\mathscr{Z} h}{\eta_\mathrm{T}}.\end{aligned} \tag{I.8}$$

Let us bound $h^*(\eta_t \mathrm{U}_t)/\eta_t$ from above. For all $z \in \mathscr{Z}$ we have

$$\frac{\langle \eta_t \mathrm{U}_t | z \rangle - h(z)}{\eta_t} = \frac{\langle \eta_{t-1} \mathrm{U}_t | z \rangle - h(z)}{\eta_{t-1}} - h(z) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right).$$

The maximum over $z \in \mathscr{Z}$ of the above left-hand side gives $h^*(\eta_t \mathrm{U}_t)/\eta_t$. As for the right-hand side, let us take the maximum over $z \in \mathscr{Z}$ for each of the two terms separately. This gives

$$\begin{aligned}\frac{h^*(\eta_t \mathrm{U}_t)}{\eta_t} &\leqslant \max_{z \in \mathscr{Z}} \left\{ \frac{\langle \eta_{t-1} \mathrm{U}_t | z \rangle - h(z)}{\eta_{t-1}} \right\} + \max_{z \in \mathscr{Z}} \left\{ -h(z) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \right\} \\ &= \frac{h^*(\eta_{t-1}\mathrm{U}_t)}{\eta_{t-1}} + \left(\min_\mathscr{Z} h \right) \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t} \right),\end{aligned}$$

where we used the fact that the sequence $(\eta_t)_{t \geqslant 0}$ is nonincreasing. Injecting this inequality in (I.8), we get

$$\langle \mathrm{U}_\mathrm{T} | z \rangle \leqslant \frac{h^*(0)}{\eta_0} + \sum_{t=1}^{\mathrm{T}} \frac{h^*(\eta_{t-1}\mathrm{U}_t) - h^*(\eta_{t-1}\mathrm{U}_{t-1})}{\eta_{t-1}} + (\min_\mathscr{Z} h) \sum_{t=1}^{\mathrm{T}} \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t} \right) + \frac{\max_\mathscr{Z} h}{\eta_\mathrm{T}}.$$

We can make the Bregman divergence appear in the first sum above by subtracting

$$\frac{\langle \eta_{t-1}\mathrm{U}_t - \eta_{t-1}\mathrm{U}_{t-1} | \nabla h^*(\eta_{t-1}\mathrm{U}_{t-1}) \rangle}{\eta_{t-1}} = \langle u_t | z_t \rangle .$$

Therefore,

$$\langle \mathrm{U}_\mathrm{T} | z \rangle \leqslant \frac{h^*(0)}{\eta_0} + \sum_{t=1}^{\mathrm{T}} \frac{\mathrm{D}_{h^*}(\eta_{t-1}\mathrm{U}_t, \eta_{t-1}\mathrm{U}_{t-1})}{\eta_{t-1}} + \sum_{t=1}^{\mathrm{T}} \langle u_t | z_t \rangle - \frac{\min_\mathscr{Z} h}{\eta_\mathrm{T}} + \frac{\min_\mathscr{Z} h}{\eta_0} + \frac{\max_\mathscr{Z} h}{\eta_\mathrm{T}}.$$

Since $h^*(0) = -\min_{\mathcal{Z}} h$, we get

$$\begin{aligned}\mathrm{Reg}_{\mathrm{T}} &= \max_{z\in\mathcal{Z}} \langle \mathrm{U}_{\mathrm{T}}|z\rangle - \sum_{t=1}^{\mathrm{T}} \langle u_t|z_t\rangle \\ &\leqslant \frac{\max_{\mathcal{Z}} h - \min_{\mathcal{Z}} h}{\eta_{\mathrm{T}}} + \sum_{t=1}^{\mathrm{T}} \frac{\mathrm{D}_{h^*}(\eta_{t-1}\mathrm{U}_t, \eta_{t-1}\mathrm{U}_{t-1})}{\eta_{t-1}}.\end{aligned}$$

(ii) The strong convexity of the regularizer h and Proposition I.2.8 let us bound the above Bregman divergences as follows:

$$\mathrm{D}_{h^*}(\eta_{t-1}\mathrm{U}_t, \eta_{t-1}\mathrm{U}_{t-1}) \leqslant \frac{1}{2\mathrm{K}} \|\eta_{t-1}\mathrm{U}_t - \eta_{t-1}\mathrm{U}_{t-1}\|_*^2 = \frac{\eta_{t-1}^2}{2\mathrm{K}} \|u_t\|_*^2, \quad t \geqslant 1,$$

which proves the result.

(iii) Set $\eta = \sqrt{\delta_h \mathrm{K}/\mathrm{M}^2}$ so that $\eta_t = \eta\, t^{-1/2}$ for $t \geqslant 1$. The regret bound then becomes

$$\frac{\delta_h\sqrt{\mathrm{T}}}{\eta} + \frac{\mathrm{M}^2}{2\mathrm{K}} \sum_{t=1}^{\mathrm{T}} \eta_{t-1}.$$

We bound the above sum as follows. Since $\eta_0 = \eta_1 = \eta$,

$$\begin{aligned}\sum_{t=1}^{\mathrm{T}} \eta_{t-1} &= \eta\left(2 + \sum_{t=2}^{\mathrm{T}-1} \frac{1}{\sqrt{t}}\right) \leqslant \eta\left(\int_0^1 \frac{1}{\sqrt{s}}\,\mathrm{d}s + \int_1^{\mathrm{T}-1} \frac{1}{\sqrt{s}}\,\mathrm{d}s\right) \\ &= \eta \int_0^{\mathrm{T}-1} \frac{1}{\sqrt{s}}\,\mathrm{d}s = 2\eta\sqrt{\mathrm{T}-1} \leqslant 2\eta\sqrt{\mathrm{T}}.\end{aligned}$$

Injecting the expression of η and simplifying gives

$$\mathrm{Reg}_{\mathrm{T}} \leqslant 2\mathrm{M}\sqrt{\frac{\mathrm{T}\delta_h}{\mathrm{K}}}.$$

□

An alternative proof of this result based on a continuous-time approach is given in Chapter VII and offers the following interpretation. The first term $\delta_h/\eta_{\mathrm{T}}$ in the above bound (i) is the regret guarantee of the continuous-time mirror descent algorithm, whereas the Bregman divergences $\mathrm{D}_{h^*}(\eta_{t-1}\mathrm{U}_t, \eta_{t-1}\mathrm{U}_{t-1})$ come from the discrepancy between the continuous-time and the discrete-time strategies.

I.4. Convex losses

We consider here a more general regret minimization problem, called *online convex optimization*, in which Nature chooses at time $t \geqslant 1$ a convex loss function $\ell_t : \mathscr{V} \to \mathbb{R}$ instead of payoff vectors. The play is as follows. At time instance $t \geqslant 1$, the Decision Maker

- chooses a point $z_t \in \mathscr{Z}$;
- observes a (negative) subgradient $u_t \in -\partial \ell_t(z_t)$;
- incurs a loss equal to $\ell_t(z_t)$.

The feedback offered to the Decision Maker is therefore (an element of) the subdifferential $\partial \ell_t(z_t)$. The regret to minimize is defined by

$$\sum_{t=1}^{T} \ell_t(z_t) - \min_{z \in \mathscr{Z}} \sum_{t=1}^{T} \ell_t(z).$$

The regret minimization with payoff vectors defined in Section I.1 can be seen as a special case where the loss functions are linear. As demonstrated by [KW97, CB97], the setting with convex losses can be reduced to a regret minimization problem with linear payoffs as follows. Using a convexity inequality, we can write

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t(z_t) - \min_{z \in \mathscr{Z}} \sum_{t=1}^{T} \ell_t(z) &= \max_{z \in \mathscr{Z}} \sum_{t=1}^{T} \left(\ell_t(z_t) - \ell_t(z)\right) \\
&\leqslant \max_{z \in \mathscr{Z}} \sum_{t=1}^{T} \langle -u_t | z_t - z \rangle \\
&= \max_{z \in \mathscr{Z}} \sum_{t=1}^{T} \langle u_t | z \rangle - \sum_{t=1}^{T} \langle u_t | z_t \rangle .
\end{aligned}$$

This last quantity is obviously the regret as defined in Section I.1 where $(u_t)_{t \geqslant 1}$ are seen as payoff vectors. We then naturally define the Mirror Descent strategies as follows. Let h be a regularizer on $\mathscr{Z}$, $(\eta_t)_{t \geqslant 1}$ a positive and nonincreasing sequence. Set $U_0 = 0$ and for $t \geqslant 1$,

$$\begin{aligned}
z_t &= \nabla h^*(\eta_{t-1} U_{t-1}) \\
U_t &\in U_{t-1} - \partial \ell_t(z_t).
\end{aligned}$$

Note that payoff vector u_t belongs by definition to $-\partial \ell_t(z_t)$. It therefore depends on z_t, which is indeed allowed—see Section I.1.

Theorem I.4.1. *Let* $T \geqslant 1$ *an integer and* $K, M > 0$.

(i) Assume that h is K*-strongly convex with respect to a norm $\|\cdot\|$. Then, against any sequence of loss functions $(\ell_t)_{t\geqslant 1}$, the above Mirror Descent strategy guarantees*

$$\sum_{t=1}^{T} \ell_t(z_t) - \min_{z\in\mathcal{Z}} \sum_{t=1}^{T} \ell_t(z_t) \leqslant \frac{\delta_h}{\eta_T} + \frac{1}{2K} \sum_{t=1}^{T} \eta_{t-1} \|\partial\ell_t(z_t)\|_*^2,$$

where $\eta_0 = \eta_1$.

(ii) Moreover, if the loss functions are M*-Lipschitz continuous with respect to $\|\cdot\|$, the choice of parameters $\eta_t = \sqrt{\delta_h K/M^2 t}$ (for $t \geqslant 1$) guarantees*

$$\sum_{t=1}^{T} \ell_t(z_t) - \min_{z\in\mathcal{Z}} \sum_{t=1}^{T} \ell_t(z) \leqslant 2M\sqrt{\frac{T\delta_h}{K}}.$$

Proof. The bounds follow from Theorem I.3.1 and the above discussion. □

One important special case where $\mathcal{V}$ is an Euclidean space and where the Euclidean regularizer h_2 from Section I.2.3 is chosen. As stated in Proposition I.2.10, the map ∇h_2^* is simply the Euclidean projection onto $\mathcal{Z}$:

$$z_t = \operatorname*{proj}_{\mathcal{Z}}(\eta_{t-1}U_{t-1})$$
$$U_t \in U_{t-1} - \partial\ell_t(z_t).$$

I.5. Convex optimization

Ordinary convex optimization problems can be seen as a regret minimization problem where the loss function remains constant over time. In what follows, we outline how regret minimizing strategies can be used for this purpose and discuss the performance gap incurred by using variable step-sizes instead of a variable parameters.

Let $f : \mathcal{V} \to \mathbb{R}$ be a convex function to minimize on a nonempty convex compact set $\mathcal{Z} \subset \mathcal{V}$, h a regularizer on $\mathcal{Z}$, $(\eta_t)_{t\geqslant 1}$, a positive and nonincreasing sequence $(\eta_t)_{t\geqslant 1}$ and $(\gamma_t)_{t\geqslant 1}$ a positive sequence. We consider the following general algorithm. Set $U_0 = 0$ and for $t \geqslant 1$,

$$z_t = \nabla h^*(\eta_{t-1}U_{t-1}),$$
$$U_t \in U_{t-1} - \gamma_t \partial f(z_t),$$

which corresponds to the Mirror Descent algorithm from Section I.3 associated with payoff vectors $u_t \in -\gamma_t \partial f(z_t)$. We call $(\eta_t)_{t\geqslant 1}$ the *parameters*, which in the definition of the strategy multiply the whole sum $U_{t-1} = \sum_{s=1}^{t-1} u_s$, and $(\gamma_t)_{t\geqslant 1}$ the *step-sizes*, whose purpose is to give different weights to the successive subdifferentials $\partial f(z_t)$.

We now state the general guarantees offered by the above algorithm, which are similar to those obtained in [BT03, Theorem 4.1] for the greedy version of Mirror Descent (see the introduction of the manuscript for a discussion on the difference between greedy and lazy Mirror Descent).

Theorem I.5.1. *Let* $T \geqslant 1$ *an integer and* $K, M > 0$.

(i) Suppose that the function f *is* M*-Lipschitz continuous with respect to a norm* $\|\cdot\|$*, and that* h *is a* K*-strongly regularizer with respect to* $\|\cdot\|$*. Denote* $z'_T \in \arg\min_{1\leqslant t\leqslant T} f(z_t)$*. Then,*

$$f(z'_T) - \min_{\mathcal{Z}} f \leqslant \left(\sum_{t=1}^{T} \gamma_t\right)^{-1} \left(\frac{\delta_h}{\eta_T} + \frac{M^2}{2K}\sum_{t=1}^{T} \eta_{t-1}\gamma_t^2\right).$$

(ii) The choice of constant parameters $\eta_t = 1$ *gives*

$$f(z'_T) - \min_{\mathcal{Z}} f \leqslant \left(\sum_{t=1}^{T} \gamma_t\right)^{-1} \left(\delta_h + \frac{M^2}{2K}\sum_{t=1}^{T} \gamma_t^2\right),$$

(iii) and the choice of constant step-sizes $\gamma_t = 1$ *and variable parameters* $\eta_t = \sqrt{\delta_h K / M^2 t}$ *gives*

$$f(z'_T) - \min_{\mathcal{Z}} f \leqslant 2M\sqrt{\frac{\delta_h}{TK}}.$$

Proof. We make the regret appear as follows:

$$\begin{aligned} f(z'_T) - \min_{z\in\mathcal{Z}} f(z) &\leqslant \left(\sum_{t=1}^{T} \gamma_t\right)^{-1} \left(\sum_{t=1}^{T} \gamma_t f(z_t) - \min_{z\in\mathcal{Z}} \sum_{t=1}^{T} \gamma_t f(z)\right) \\ &= \left(\sum_{t=1}^{T} \gamma_t\right)^{-1} \left(\max_{z\in\mathcal{Z}} \sum_{t=1}^{T} \gamma_t \left(f(z_t) - f(z)\right)\right) \\ &\leqslant \left(\sum_{t=1}^{T} \gamma_t\right)^{-1} \left(\max_{z\in\mathcal{Z}} \sum_{t=1}^{T} \langle u_t | z_t - z\rangle\right), \end{aligned}$$

where we used in the last line the fact that $u_t \in -\gamma_t \partial f(z_t)$. Besides, f being M-Lipschitz continuous with respect to $\|\cdot\|$ is equivalent to its subgradients being bounded from above by M with respect to $\|\cdot\|_*$. Therefore, injecting $\|u_t\|_* \leqslant \gamma_t M$ into Theorem I.3.1 gives the result. □

One can see that the best convergence rate that we get in (ii) with a constant parameter and step-sizes of the form $\gamma_t = t^{-\alpha}$ is of order $O(T^{-1/2} \log T)$ (for $\alpha = 1/2$) (and there is no straightforward choice of γ_t leading to a better convergence rate). On the other hand, by taking in (iii) a constant step-size and varying the algorithm's parameter $\eta_t = O(t^{-1/2})$, we do achieve an $O(T^{-1/2})$ rate of convergence.

CHAPTER II

EXPERTS SETTING

We dedicate this chapter to a variant of the model from Section I.1 where the Decision Maker has a finite set of actions from which he is allowed to choose at random. The Mirror Descent strategies introduced in Section I.3 and the corresponding regret bounds are easily adapted to this framework. Randomization being introduced by the Decision Maker, we also derive high probability and almost-sure regret guarantees. We then examine a few important special cases: the Exponential Weights Algorithm, the case of sparse payoff vectors, the Smooth Fictitious Play and the Vanishingly Smooth Fictitious Play.

II.1. Model

Let $\mathcal{I} = \{1, \dots, d\}$ be the set of *pure actions* of the Decision Maker. Denote Δ_d the unit simplex of $\mathbb{R}^d$ which can be seen as the set of probability distributions over $\mathcal{I}$:

$$\Delta_d = \left\{ z \in \mathbb{R}_+^d \,\middle|\, \sum_{i=1}^{d} z^i = 1 \right\}.$$

An element of Δ_d is called a *mixed action*. The play goes as follows. At each time instance $t \geqslant 1$, the Decision Maker

- chooses a mixed action $z_t \in \Delta_d$;
- draws pure action $i_t \in \mathcal{I}$ according to probability distribution z_t;
- observes payoff vector $u_t \in \mathbb{R}^d$;
- receives payoff $u_t^{i_t}$.

Unlike the core model of Section I.1, the choice by Nature of payoff vector u_t must not depend on pure action i_t (but can still depend on mixed action z_t). Let $(\mathcal{F}_t)_{t\geqslant 1}$ the filtration where $\mathcal{F}_t$ is generated by

$$(z_1, u_1, i_1, \dots, z_{t-1}, u_{t-1}, i_{t-1}, z_t, u_t).$$

We then have $\mathbb{E}\left[u_t^{i_t} \,\middle|\, \mathscr{F}_t\right] = \mathbb{E}_{i\sim z_t}\left[u_t^i\right] = \langle u_t | z_t\rangle$. A strategy for the Decision Maker is a sequence of measurable maps $\sigma = (\sigma_t)_{t\geqslant 1}$ where $\sigma_t : (\Delta_d \times \mathscr{I} \times \mathscr{U})^{t-1} \to \Delta_d$. For a given strategy σ and a sequence of payoff vectors $(u_t)_{t\geqslant 1}$, we have:

$$z_t = \sigma_t(z_1, i_1, u_1, \dots, z_{t-1}, i_{t-1}, u_{t-1}), \quad t \geqslant 1.$$

The *realized regret* up to time $\mathrm{T} \geqslant 1$ is the random variable defined as

$$\widetilde{\mathrm{Reg}}_{\mathrm{T}} = \max_{i\in\mathscr{I}} \sum_{t=1}^{\mathrm{T}} u_t^i - \sum_{t=1}^{\mathrm{T}} u_t^{i_t}.$$

We call *regret* the following quantity, where the payoff $u_t^{i_t}$ has been replaced by its conditional expectation $\langle u_t | z_t\rangle = \mathbb{E}\left[u_t^{i_t} \,\middle|\, \mathscr{F}_t\right]$. It corresponds to the regret from Section I.1:

$$\mathrm{Reg}_{\mathrm{T}} = \max_{i\in\mathscr{I}} \sum_{t=1}^{\mathrm{T}} u_t^i - \sum_{t=1}^{\mathrm{T}} \langle u_t | z_t\rangle = \max_{z\in\Delta_d} \sum_{t=1}^{\mathrm{T}} \langle u_t | z\rangle - \sum_{t=1}^{\mathrm{T}} \langle u_t | z_t\rangle .$$

The Mirror Descent strategies adapted from Section II.1 will provide upper bounds on the regret. With the help of concentration inequalities, those will in turn provide high probability and almost-sure guarantees on the realized regret.

We call this setting the *experts setting* because it models the problem of prediction with experts advice which can be described as follows. Let $\mathscr{I} = \{1, \dots, d\}$ be a set of *experts*. At each stage $t \geqslant 1$, the Decision Maker is to make a decision and each expert gives an advice as to which decision to make. The Decision Maker must then choose the expert i_t to follow. Then, the vector $u_t \in \mathbb{R}^d$ is observed, where u_t^i is the payoff obtained by expert i. The payoff obtained by the Decision Maker is therefore $u_t^{i_t}$. The regret then corresponds to the difference between the cumulative payoff of the Decision Maker and the cumulative payoff obtained by the best expert in hindsight.

An important direction of research is the study of the best possible guarantee on the regret, in other words, the study of the minimax regret

$$\min_{\sigma} \max_{(u_t)_{t\geqslant 1}} \mathrm{R}_{\mathrm{T}},$$

where the minimum is taken over the strategies of the Decision Maker, and the maximum over the possible sequences of payoff vectors. Without any assumption on the payoff vectors, it is easy to see that this quantity is equal to $+\infty$. It becomes finite and therefore relevant when, typically, the payoff vectors are assumed to belong to a bounded set $\mathscr{U} \subset \mathbb{R}^d$. However, we are usually unable to compute the value of the minimax regret exactly, and we simply establish its asymptotic *dependencies* in

the parameters of the problem. For instance, the most common assumption in this framework is that payoff vectors belong to $\mathcal{U} = [-1,1]^d$. In this case, the minimax regret is known to be *of order* $\sqrt{T \log d}$, which gives the dependency in the number of stages T and in the number of actions d. This result has been proved in two steps. First, the Exponential Weights Algorithm was shown to guarantee a regret bound of $\sqrt{T \log d}$ (up to a multiplicative constant) [CB97], which gives an *upper bound* on the minimax regret (this result will be presented in detail in Section II.3). Second, using a probabilistic argument, it has been established [CBFH+97] that the minimax regret is higher than $\sqrt{T \log d}$ (up to a multiplicative constant) when T and d are large. Stronger assumptions involving sparsity will be considered in Section II.4 and will lead to lower minimax regrets, achieved by well-chosen strategies.

II.2. Mirror Descent strategies

We adapt the Mirror Descent strategies from Section I.3 to this framework by simply seeing the simplex Δ_d as the convex compact set of actions. The strategy associated with a regularizer h on Δ_d and a positive and nonincreasing sequence of parameters $(\eta_t)_{t\geqslant 1}$ is therefore defined as follows. Set $U_0 = 0$ and for $t \geqslant 1$,

$$\begin{aligned} \text{choose} \quad & z_t = \nabla h^*(\eta_{t-1} U_{t-1}), \\ \text{draw} \quad & i_t \sim z_t, \\ \text{update} \quad & U_t = U_{t-1} + u_t. \end{aligned}$$

The results of Theorem I.3.1 hold. We are now aiming at deriving high probability and almost-sure results on the realized regret. The Hoeffding–Azuma inequality will make sure that the regret and the realized regret are close.

Lemma II.2.1. *Let $(z_t)_{t\geqslant 1}$ and $(i_t)_{t\geqslant 1}$ be sequences of mixed and pure actions respectively played by the Decision Maker against payoff vectors $(u_t)_{t\geqslant 1}$. Let $M > 0$ and assume that $\|u_t\|_\infty \leqslant M$ (for all $t \geqslant 1$).*

(i) Let $\delta \in (0,1)$. With probability higher than $1-\delta$, we have

$$\widetilde{\mathrm{Reg}}_T \leqslant \mathrm{Reg}_T + M\sqrt{8T \log(1/\delta)}.$$

(ii) Almost-surely,

$$\limsup_{T\to+\infty} \frac{1}{T}\left(\widetilde{\mathrm{Reg}}_T - \mathrm{Reg}_T\right) \leqslant 0.$$

Proof. (i) Let $(\mathcal{F}_t)_{t\geqslant 1}$ be the filtration introduced in Section II.1 and $X_t = \langle u_t | z_t \rangle - u_t^{i_t}$. Then, $(X_t)_{t\geqslant 1}$ is sequence of martingale differences with respect to $(\mathcal{F}_t)_{t\geqslant 1}$. In-

deed, $\mathbb{E}\left[\langle u_t|z_t\rangle - u_t^{i_t} \,\middle|\, \mathscr{F}_t\right] = \langle u_t|z_t\rangle - \langle u_t|z_t\rangle = 0$. Besides, $|X_t| \leqslant 2M$. Proposition A.0.1 applied with $\varepsilon = M\sqrt{8\log(1/\delta)/T}$ then gives

$$\mathbb{P}\left[\frac{1}{T}\sum_{t=1}^{T} X_t > \varepsilon\right] \leqslant \delta.$$

In other words, with probability higher than $1-\delta$, we have

$$\sum_{t=1}^{T} \langle u_t|z_t\rangle \leqslant \sum_{t=1}^{T} u_t^{i_t} + M\sqrt{8T\log(1/\delta)}.$$

Adding $\max_{i\in\mathscr{I}} \sum_{t=1}^{T} u_t^i$ to both sides and reorganizing the terms gives the result.

(ii) The second part of the statement follows from a standard Borel–Cantelli argument. □

We now state the high-probability and almost-sure guarantees offered by the Mirror Descent strategies in the case of a strongly convex regularizer and bounded payoff vectors.

Theorem II.2.2. *Let $T \geqslant 1$ an integer, $K, M > 0$ and $\delta \in (0,1)$. With notation from Section II.2, assume that h is K-strongly convex with respect to $\|\cdot\|_1$.*

(i) Against any sequence of payoff vectors $(u_t)_{t\geqslant 1}$ such that $\|u_t\|_\infty \leqslant M$ (for all $t \geqslant 1$), the strategy defined in Section II.2 guarantees with probability higher than $1-\delta$

$$\widetilde{\mathrm{Reg}}_T \leqslant \frac{\delta_h}{\eta_T} + \frac{M^2}{2K}\sum_{t=1}^{T}\eta_{t-1} + M\sqrt{8T\log(1/\delta)}.$$

(ii) In particular, the choice of parameters $\eta_t = \sqrt{\delta_h K/M^2 t}$ (for $t \geqslant 1$) gives with probability higher that $1-\delta$,

$$\widetilde{\mathrm{Reg}}_T \leqslant M\sqrt{T}\left(2\sqrt{\frac{\delta_h}{K}} + \sqrt{8\log(1/\delta)}\right),$$

and almost-surely,

$$\limsup_{T\to+\infty} \frac{1}{T}\widetilde{\mathrm{Reg}}_T \leqslant 0.$$

Proof. Simply combine Theorem I.3.1, Lemma II.2.1, and the fact that $\|\cdot\|_\infty$ is the dual norm of $\|\cdot\|_1$. □

The average realized regret being asymptotically nonpositive, as stated in the very last part of the above theorem, is the original definition of a strategy being consistent, as proposed by Hannan [Han57].

II.3. Exponential Weights Algorithm

The most important instance of Mirror Descent strategies in the experts setting is the *Exponential Weights Algorithm*, introduced by [LW94, Vov90] and further studied by [KW95, CB97, ACBG02, Sor09] among others. As proved below in Theorem II.3.1, it achieves a minimax regret guarantee of order $\sqrt{T \log d}$. The algorithm corresponds to the choice the entropic regularizer:

$$h_{\mathrm{ent}}(z) = \begin{cases} \sum_{i=1}^{d} z^i \log z^i & \text{if } z \in \Delta_d \\ +\infty & \text{otherwise.} \end{cases}$$

Proposition I.2.9 then gives the following explicit expression of the algorithm:

$$z_t^i = \frac{\exp\left(\eta_{t-1} U_{t-1}^i\right)}{\sum_{j=1}^{d} \exp\left(\eta_{t-1} U_{t-1}^j\right)}, \quad i \in \mathcal{I}.$$

The following regret bound achieved by the Exponential Weights Algorithm with time-varying parameters $\eta_t = \sqrt{\log d / t}$ was first established in [ACBG02].

Theorem II.3.1. *Let $T \geqslant 1$ an integer. Against any sequence of payoff vectors in $[-1,1]^d$, the Exponential Weights Algorithm with parameters $\eta_t = \sqrt{\log d / t}$ (for $t \geqslant 1$) guarantees*

$$\mathrm{Reg}_T \leqslant 2\sqrt{T \log d}.$$

Let $\delta \in (0,1)$. With probability higher that $1-\delta$, we have

$$\widetilde{\mathrm{Reg}}_T \leqslant \sqrt{T}\left(2\sqrt{\log d} + \sqrt{8 \log(1/\delta)}\right).$$

Almost-surely,

$$\limsup_{T \to +\infty} \frac{1}{T} \widetilde{\mathrm{Reg}}_T \leqslant 0.$$

Proof. From Proposition I.2.9, we know $\delta_{h_{\mathrm{ent}}} = \log d$ and that h_{ent} is 1-strongly convex with respect to $\|\cdot\|_1$; and since $u_t \in [-1,1]^d$ implies $\|u_t\|_\infty \leqslant 1$, the results follow from Theorem II.2.2 applied with $M = 1$ and $K = 1$. □

We now turn to more precise regret bounds which hold in the case of payoff vectors whose components are bounded from above. For simplicity, we state the following results in the case of losses, i.e. payoff vectors with nonpositive coefficients. They are obtained by a finer analysis of the Bregman divergence associated with h_{ent}^*.

Theorem II.3.2. *(i) Against payoff vectors $(u_t)_{t\geqslant 1}$ in $\mathbb{R}_-^d$, the Exponential Weights Algorithm with parameters $(\eta_t)_{t\geqslant 1}$ guarantees*

$$\mathrm{Reg}_T \leqslant \frac{\log d}{\eta_T} + \sum_{t=1}^{T} \eta_{t-1} \sum_{i=1}^{d} (u_t^i)^2 z_t^i,$$

where we set $\eta_0 = \eta_1$.

(ii) Against payoff vectors $(u_t)_{t\geqslant 1}$ in $[-1,0]^d$, the Exponential Weights Algorithm with constant parameter $\eta \in (0,1)$ guarantees

$$\mathrm{Reg}_T \leqslant \frac{1}{1-\eta}\left(\frac{\log d}{\eta} - \eta \max_{i\in\mathcal{I}} \sum_{t=1}^{T} u_t^i\right).$$

Proof. (i) Theorem I.3.1 together with the fact that $\delta_{h_{\mathrm{ent}}} = \log d$ gives

$$\mathrm{Reg}_T \leqslant \frac{\log d}{\eta_T} + \sum_{t=1}^{T} \frac{1}{\eta_{t-1}} \mathrm{D}_{h_{\mathrm{ent}}^*}(\eta_{t-1}U_t, \eta_{t-1}U_{t-1}).$$

We aim at bounding from above the Bregman divergence in the above sum. Proposition I.2.9 gives the following expression for h_{ent}^*:

$$h_{\mathrm{ent}}^*(y) = \log\left(\sum_{i=1}^{d} e^{y^i}\right), \qquad y \in \mathbb{R}^d.$$

For $t \geqslant 1$, we can then express the Bregman divergence as

$$\begin{aligned}
\mathrm{D}_{h_{\mathrm{ent}}^*}(\eta_{t-1}U_t, \eta_{t-1}U_{t-1}) &= h_{\mathrm{ent}}^*(\eta_{t-1}U_t) - h_{\mathrm{ent}}^*(\eta_{t-1}U_{t-1}) \\
&\qquad - \left\langle \nabla h_{\mathrm{ent}}^*(\eta_{t-1}U_{t-1}) \middle| \eta_{t-1}U_t - \eta_{t-1}U_{t-1} \right\rangle \\
&= \log\left(\sum_{i=1}^{d} e^{\eta_{t-1}U_t^i}\right) - \log\left(\sum_{i=1}^{d} e^{\eta_{t-1}U_{t-1}^i}\right) - \eta_{t-1}\left\langle z_t | u_t \right\rangle \\
&= \log\left(\sum_{i=1}^{d} \frac{e^{\eta_{t-1}u_t^i} e^{\eta_{t-1}U_{t-1}^i}}{\sum_{j=1}^{d} e^{\eta_{t-1}U_{t-1}^j}}\right) - \eta_{t-1}\left\langle z_t | u_t \right\rangle \\
&= \log\left(\sum_{i=1}^{d} z_t^i e^{\eta_{t-1}u_t^i}\right) - \eta_{t-1}\left\langle z_t | u_t \right\rangle.
\end{aligned}$$

Since $u_t^i \in [-1,0]$ by hypothesis, it is true that

$$e^{\eta_{t-1}u_t^i} \leqslant 1 + \eta_{t-1}u_t^i + \eta_{t-1}^2 (u_t^i)^2.$$

Substituting in the previous expression,

$$\begin{aligned}
\mathrm{D}_{h^*_{\mathrm{ent}}}(\eta_{t-1}\mathrm{U}_t, \eta_{t-1}\mathrm{U}_{t-1}) &\leqslant \log\left(\sum_{i=1}^{d} z_t^i\,(1+\eta_{t-1}u_t^i+\eta_{t-1}^2(u_t^i)^2)\right) - \eta_{t-1}\,\langle z_t|u_t\rangle \\
&= \log\left(1+\eta_{t-1}\,\langle u_t|z_t\rangle + \eta_{t-1}^2\sum_{i=1}^{d} z_t^i(u_t^i)^2\right) - \eta_{t-1}\,\langle u_t|z_t\rangle \\
&\leqslant \eta_{t-1}\,\langle u_t|z_t\rangle + \eta_{t-1}^2\sum_{i=1}^{d} z_t^i(u_t^i)^2 - \eta_{t-1}\,\langle z_t|u_t\rangle \\
&= \eta_{t-1}^2\sum_{i=1}^{d} z_t^i(u_t^i)^2,
\end{aligned}$$

which gives the result.

(ii) The second bound is a corollary of the first one. We restrict to the Exponential Weights Algorithm with a constant parameter $\eta \in (0,1)$. Since $u_t^i \in [-1, 0]$, we have $(u_t^i)^2 \leqslant -u_t^i$. This gives

$$\max_{i\in\mathcal{I}} \sum_{t=1}^{\mathrm{T}} u_t^i - \sum_{t=1}^{\mathrm{T}} \langle z_t|u_t\rangle \leqslant \frac{\log d}{\eta} - \eta\sum_{t=1}^{\mathrm{T}}\sum_{i=1}^{d} z_t^i u_t^i.$$

Since $\sum_{i=1}^{d} z_t^i u_t^i$ simply is $\langle z_t|u_t\rangle$, we can reorganize the above quantities to get

$$(1-\eta)\left(\max_{i\in\mathcal{I}} \sum_{t=1}^{\mathrm{T}} u_t^i - \sum_{t=1}^{\mathrm{T}} \langle z_t|u_t\rangle\right) \leqslant \frac{\log d}{\eta} - \eta\max_{i\in\mathcal{I}}\sum_{t=1}^{\mathrm{T}} u_t^i,$$

and the result follows by dividing by $1-\eta > 0$. □

Regret bounds similar to (i) have appeared in e.g. [ACBFS02], [CBLS05] and [SS11, Theorem 2.2] in the case of constant parameters. As for (ii), a bound of the same kind was already proposed in [LW94] and is called *improvement for small losses*.

II.4. Sparse payoff vectors

We here add a sparsity assumption on the payoff vectors: we assume that they has at most s nonzero components (for a given integer $1 \leqslant s \leqslant d$). We aim at constructing strategies which take advantage of this restriction to guarantee regret bounds that are

tighter than the bound of order $\sqrt{T \log d}$ guaranteed in Theorem II.3.1 by the Exponential Weights Algorithm. A thorough investigation of this subject will be conducted in Chapter V. We distinguish two cases: sparse gains and sparse losses. Denote

$$\mathcal{U}^{+,s,d} = \{u \in [0,1]^d \,|\, u \text{ has at most } s \text{ nonzero components}\},$$
$$\mathcal{U}^{-,s,d} = \{u \in [-1,0]^d \,|\, u \text{ has at most } s \text{ nonzero components}\}.$$

Let $p \in (0,1)$ and consider the following regularizer on Δ_d:

$$h_p(z) = \begin{cases} \frac{1}{2}\|z\|_p^2 & \text{if } z \in \Delta_d \\ +\infty & \text{otherwise.} \end{cases}$$

The associated Mirror Descent strategy guarantees a regret bound of order $\sqrt{T \log s}$ in the case of sparse gains.

Theorem II.4.1. *Let $T \geqslant 1$ and $s \geqslant 3$. Against payoff vectors in $\mathcal{U}^{+,s,d}$, the Mirror Descent strategy associated with regularizer h_p with $p = 1 + (2\log s - 1)^{-1}$ and parameters $\eta_t = (4et \log s)^{-1/2}$ (for $t \geqslant 1$) guarantees*

$$\mathrm{Reg}_T \leqslant 2\sqrt{eT \log s}.$$

Proof. According to Proposition I.2.11, regularizer h_p is $(p-1)$-strongly convex with respect to $\|\cdot\|_p$. Let $q > 0$ such that $1/p + 1/q = 1$. We use the assumption on the payoff vectors to bound their ℓ^q norms as follows. Let $u \in \mathcal{U}^{+,s,d}$. u has at most s nonzero components. Thus,

$$\|u\|_q = \left(\sum_{i=1}^d |u^i|^q\right)^{1/q} \leqslant \left(\sum_{s \text{ terms}} |u^i|^q\right)^{1/q} \leqslant s^{1/q}.$$

Theorem I.3.1 then gives

$$\mathrm{Reg}_T \leqslant \frac{\delta_h}{\eta_T} + \frac{s^{2/q}}{2(p-1)} \sum_{t=1}^T \eta_{t-1}.$$

We know that $\delta_{h_p} \leqslant 1/2$. Then, note that $p - 1 = (2\log s - 1)^{-1}$ and that

$$\frac{1}{q} = 1 - \frac{1}{p} = \frac{p-1}{p} = \frac{(2\log s - 1)^{-1}}{1 + (2\log s - 1)^{-1}} = \frac{1}{2\log s}.$$

Therefore the bound on the regret becomes

$$\mathrm{Reg}_{\mathrm{T}} \leqslant \frac{1}{2\eta_{\mathrm{T}}} + \frac{e^{2\log s/(2\log s)}(2\log s - 1)}{2} \sum_{t=1}^{\mathrm{T}} \eta_{t-1} \leqslant \frac{1}{2\eta_{\mathrm{T}}} + e\log s \sum_{t=1}^{\mathrm{T}} \eta_{t-1},$$

and the choice $\eta_t = (4et\log s)^{-1/2}$ for $t \geqslant 1$ gives

$$\mathrm{Reg}_{\mathrm{T}} \leqslant 2\sqrt{e\mathrm{T}\log s}.$$

□

We now turn to the case of sparse losses. The above result still holds, but we are able to guarantee a much better regret bound, of order $\sqrt{\mathrm{T}s\frac{\log d}{d}}$, by using a different strategy.

Theorem II.4.2. *Let* $\mathrm{T} \geqslant 1$. *Against payoff vectors in* $\mathcal{U}^{-,s,d}$, *the Exponential Weights Algorithm with constant parameter* $\eta = \sqrt{d\log d/s\mathrm{T}}$ *guarantees for* $\mathrm{T} > 4d\log d/s$,

$$\mathrm{Reg}_{\mathrm{T}} \leqslant 4\sqrt{\mathrm{T}\frac{s\log d}{d}}.$$

Proof. Let $\mathrm{T} > 4d\log d/s$. Since u_t belongs to $[-1,0]^d$ and have at most s nonzero components, we have

$$s\mathrm{T} \geqslant -\sum_{t=1}^{\mathrm{T}}\sum_{i=1}^{d} u_t^i = -\sum_{i=1}^{d}\sum_{t=1}^{\mathrm{T}} u_t^i \geqslant -d\cdot\max_{i\in\mathcal{I}}\sum_{t=1}^{\mathrm{T}} u_t^i.$$

Therefore, the above maximum is bounded from below by $-s\mathrm{T}/d$. Injecting this inequality in the regret bound (ii) from Theorem II.3.2, we get

$$\mathrm{Reg}_{\mathrm{T}} \leqslant \frac{1}{1-\eta}\left(\frac{\log d}{\eta} + \eta\frac{s\mathrm{T}}{d}\right).$$

We then choose $\eta = \sqrt{d\log d/s\mathrm{T}}$. The assumption on T assures that $\eta \in (0,1/2)$. The bound therefore becomes

$$\mathrm{Reg}_{\mathrm{T}} \leqslant 4\sqrt{\mathrm{T}\frac{s\log d}{d}}.$$

□

We will prove in Chapter V that the bounds from Theorems II.4.1 and II.4.2 are both minimax optimal. This demonstrates that gains and losses are fundamentally different in the case of sparse payoff vectors.

II.5. Smooth Fictitious Play

The Smooth Fictitious Play was introduced by [FL95, FL98, FL99] and further examined using the theory of stochastic approximations by [BHS06]. It corresponds to a Mirror Descent strategy with an arbitrary regularizer h on Δ_d and a sequence of parameters $\eta_t = \eta/t$ for some $\eta > 0$. η is called the parameter of the Smooth Fictitious Play strategy. It therefore writes

$$\begin{aligned} \text{choose} \quad & z_t = \nabla h^* \left(\frac{\eta}{t-1} U_{t-1} \right), \\ \text{draw} \quad & i_t \sim z_t, \\ \text{update} \quad & U_t = U_{t-1} + u_t. \end{aligned}$$

The qualitative analysis of [BHS06] does not require the regularizer h to be strongly convex. We here do make this assumption in order to obtain an explicit regret bound.

Theorem II.5.1. *Let $T \geqslant 1$ an integer and $K > 0$. Assume that h is K-strongly convex with respect to $\|\cdot\|_1$. Against any sequence of payoff vectors in $[-1,1]^d$, the Smooth Fictitious Play with parameter $\eta > 0$ guarantees*

$$\mathrm{Reg}_T \leqslant \frac{\delta_h T}{\eta} + \frac{\eta \log T}{2K} + \frac{\eta}{K}.$$

Let $\delta \in (0,1)$. With probability higher than $1-\delta$, we have

$$\widetilde{\mathrm{Reg}}_T \leqslant \frac{\delta_h T}{\eta} + \frac{\eta \log T}{2K} + \frac{\eta}{K} + \sqrt{8T \log(1/\delta)}.$$

Almost-surely,

$$\limsup_{T \to +\infty} \frac{1}{T} \widetilde{\mathrm{Reg}}_T \leqslant \frac{\delta_h}{\eta}.$$

Proof. Theorem I.3.1 gives

$$\mathrm{Reg}_T \leqslant \frac{\delta_h}{\eta_T} + \frac{1}{2K} \sum_{t=1}^{T} \eta_{t-1} \|u_t\|_\infty^2,$$

where $\eta_0 = \eta_1$. Injecting $\eta_t = \eta/t$ and $\|u_t\|_\infty \leqslant 1$ for $t \geqslant 1$, we obtain

$$\mathrm{Reg}_T \leqslant \frac{\delta_h}{\eta} + \frac{\eta}{2K} \left(1 + \sum_{t=1}^{T-1} \frac{1}{t} \right).$$

We then bound the sum from above:

$$\sum_{t=1}^{T-1} \frac{1}{t} \leqslant 1 + \int_1^{T-1} \frac{ds}{s} = 1 + \log(T-1) \leqslant 1 + \log T,$$

and the bound on the regret is proved. The rest of the statement follows from Lemma II.2.1. □

II.6. Vanishingly Smooth Fictitious Play

A variant of the Smooth Fictitious Play, called the Vanishingly Smooth Fictitious Play was introduced and studied by [BF13]. It corresponds to a Mirror Descent Strategy with a strongly convex regularizer h on Δ_d and a sequence of parameters $(\eta_t)_{t\geqslant 1}$ which satisfies

$$t\eta_t \xrightarrow[t\to+\infty]{} +\infty \quad \text{and} \quad \eta_t = O(t^{-\alpha}) \text{ for some } \alpha > 0. \tag{II.1}$$

Those conditions will make sure, in the following theorem, that the average realized regret is asymptotically and almost-surely nonpositive. Note that the analysis in [BF13] relied on differential inclusions and stochastic approximations and did not provide explicit regret bounds.

Theorem II.6.1. *Let $T \geqslant 1$ an integer and $K > 0$. Assume that h is K-strongly convex with respect to $\|\cdot\|_1$. Against any sequence of payoff vectors in $[-1,1]^d$, the Vanishingly Smooth Fictitious Play with parameters $(\eta_t)_{t\geqslant 1}$ satisfying conditions (II.1) guarantees*

$$\mathrm{Reg}_T \leqslant \frac{\delta_h}{\eta_T} + \frac{1}{2K}\sum_{t=1}^{T} \eta_{t-1},$$

where $\eta_0 = \eta_1$. Let $\delta \in (0,1)$. With probability higher than $1-\delta$, we have

$$\widetilde{\mathrm{Reg}}_T \leqslant \frac{\delta_h}{\eta_T} + \frac{1}{2K}\sum_{t=1}^{T} \eta_{t-1} + \sqrt{8T\log(1/\delta)}.$$

Almost-surely,

$$\limsup_{T\to+\infty} \frac{1}{T}\widetilde{\mathrm{Reg}}_T \leqslant 0.$$

Proof. The first bound is a paraphrase of Theorem I.3.1. The high probability bound follows from Lemma II.2.1. Then, conditions (II.1) give $\delta_h/T\eta_T \to 0$ as $T \to +\infty$ and

$$\frac{1}{2KT}\sum_{t=1}^{T} \eta_{t-1} = O\left(\frac{T^{-\alpha+1}}{T}\right) = O(T^{-\alpha}) \xrightarrow[T\to+\infty]{} 0,$$

and the last result follows. □

II.7. On the choice of parameters

We discuss how different decreasing rates of the parameters $(\eta_t)_{t\geqslant 1}$ affect the regret bound offered by the corresponding strategies, and more specifically, whether the *no-regret* property, which we define as

$$\limsup_{T\to+\infty} \frac{1}{T} R_T \leqslant 0,$$

is guaranteed. We restrict our attention to the experts setting and assume that regularizers are strongly convex and that payoff vectors are bounded. This will allow us to use Sections II.5 and II.6 as illustrations. However, the ideas presented below extend to the online linear optimization framework of Section I.1.

We have seen at the end of Section I.1 that the Follow the Leader strategy

$$z_t = \operatorname*{arg\,max}_{z\in\Delta_d} \left\langle \sum_{s=1}^{t-1} u_s \middle| z \right\rangle \tag{II.2}$$

fails to guarantee no-regret. This motivated the introduction of Mirror Descent strategies

$$z_t = \operatorname*{arg\,max}_{z\in\Delta_d} \left\{ \left\langle \eta_{t-1} \sum_{s=1}^{t-1} u_s \middle| z \right\rangle - h(z) \right\} \tag{II.3}$$

which can be seen as a *regularized* version of Follow the Leader, where parameter η_{t-1} controls the level of regularization: the higher is η_{t-1}, the closer (II.3) is to (II.2). In the case of a constant parameter $\eta_t = \eta$, Theorem I.3.1 gives

$$\frac{1}{T}\operatorname{Reg}_T \leqslant \frac{\delta_h}{\eta T} + \frac{\eta M^2}{2K} \quad \text{and therefore} \quad \limsup_{T\to+\infty} \frac{1}{T} R_T \leqslant \frac{\eta M^2}{2K}.$$

No-regret is not guaranteed, but parameter η (and therefore the above bound $\eta M^2/2K$) can still be chosen arbitrarily small. A similar situation occurs in the case where $\eta_t = \eta/t$, which corresponds to the Smooth Fictitious Play. As stated in Theorem II.5.1, the average regret is asymptotically bounded by η_t/η. Through the choice of h and/or η, the above bound can be made arbitrarily small, but not zero. Let us now turn to the case where η_t decreases faster than η but slower than η/t: this corresponds to the Vanishingly Smooth Fictitious Play. Then, as seen in Theorem II.6.1, no-regret *is* guaranteed.

The above can be interpreted as follows. In the case $\eta_t = \eta$, no-regret is not guaranteed because the parameters do not decrease quickly enough and the algorithm is *not regularized enough*. If $\eta_t = \eta/t$, no-regret is not guaranteed because the parameters decrease too quickly and the algorithm is *too regularized*. Finally, if the decreasing rate of the parameters are between those two edge-cases, it is just right for strategy to guarantee no-regret.

II.8. Multi-armed bandit problem

The multi-armed bandit problem was originally studied in a stochastic setting [Rob52, LR85]. The nonstochastic model we consider below was introduced by [ACBFS02] and is a regret minimization problem in the experts setting with the restriction that the Decision Maker only observes the payoff of the action that he has chosen. See [BCB12] for a recent survey.

We briefly describe the model and present the *EXP3* strategy. Its analysis is based on a regret bound that we established in Section II.3 for the Exponential Weights Algorithm. We assume that the payoff vectors $(u_t)_{t\geqslant 1}$ are chosen before the play begins [1]. At each time instance $t \geqslant 1$, the Decision Maker

- chooses a mixed action $z_t \in \Delta_d$;
- draws $i_t \sim z_t$;
- receives and observes payoff $u_t^{i_t}$.

Let $(\mathscr{F}_t)_{t\geqslant 1}$ be a filtration where $\mathscr{F}_t$ is generated by

$$(z_1, i_1, u_1^{i_1}, \dots, z_{t-1}, i_{t-1}, u_{t-1}^{i_{t-1}}, z_t).$$

It will be convenient to assume that the payoff vectors $(u_t)_{t\geqslant 1}$ are normalized in $[-1, 0]^d$. We are aiming at bounding the expectation of the realized regret:

$$\mathbb{E}\left[\max_{i\in\mathscr{I}} \sum_{t=1}^{T} u_t^i - \sum_{t=1}^{T} u_t^{i_t}\right].$$

The key idea is to use a strategy from the full information setting (i.e. where the Decision Maker observes the whole payoff vector u_t), by replacing the unobserved payoff vector u_t by an unbiased estimator $\hat{u}_t$ which is constructed as follows. Assume $z_t^i > 0$ for all $t \geqslant 1$. The Decision Maker can then compute

$$\hat{u}_t^i = \frac{\mathbb{1}_{\{i_t=i\}}}{z_t^i} u_t^i, \quad i \in \mathscr{I}.$$

$\hat{u}_t$ is an estimator of u_t in the sense that $\mathbb{E}[\hat{u}_t \mid \mathscr{F}_t] = u_t$. The following result links the expectation of the realized regret (which we aim at minimizing) with the expectation of the regret (as defined in Section II.1) with respect to $(\hat{u}_t)_{t\geqslant 1}$ seen as payoff vectors.

1. If Nature is allowed to choose the payoffs vectors as a function of the previous actions of the Decision Maker, the analysis below must be carried out with the pseudo-regret $\max_{i\in\mathscr{I}} \mathbb{E}\left[\sum_{t=1}^{T} u_t^i - \sum_{t=1}^{T} u_t^{i_t}\right]$ instead of the expected realized regret. See [BCB12] (Section 3) for a detailed discussion on this issue.

Lemma II.8.1.

$$\mathbb{E}\left[\max_{i\in\mathcal{I}}\sum_{t=1}^{T}u_t^i-\sum_{t=1}^{T}u_t^{i_t}\right]\leqslant\mathbb{E}\left[\max_{i\in\mathcal{I}}\sum_{t=1}^{T}\hat{u}_t^i-\sum_{t=1}^{T}\langle\hat{u}_t|z_t\rangle\right].$$

Proof. Using the fact that $\mathbb{E}\max\geqslant\max\mathbb{E}$,

$$\begin{aligned}
\mathbb{E}\left[\max_{i\in\mathcal{I}}\sum_{t=1}^{T}\hat{u}_t^i-\sum_{t=1}^{T}\langle\hat{u}_t|z_t\rangle\right]&\geqslant\max_{i\in\mathcal{I}}\mathbb{E}\left[\sum_{t=1}^{T}\hat{u}_t^i\right]-\mathbb{E}\left[\sum_{t=1}^{T}\langle\hat{u}_t|z_t\rangle\right]\\
&=\max_{i\in\mathcal{I}}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[\hat{u}_t^i\,\middle|\,\mathcal{F}_t\right]\right]-\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[\langle\hat{u}_t|z_t\rangle\mid\mathcal{F}_t\right]\right]\\
&=\max_{i\in\mathcal{I}}\mathbb{E}\left[\sum_{t=1}^{T}u_t^i\right]-\mathbb{E}\left[\sum_{t=1}^{T}\langle u_t|z_t\rangle\right]\\
&=\max_{i\in\mathcal{I}}\mathbb{E}\left[\sum_{t=1}^{T}u_t^i\right]-\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[u_t^{i_t}\,\middle|\,\mathcal{F}_t\right]\right]\\
&=\mathbb{E}\left[\max_{i\in\mathcal{I}}\sum_{t=1}^{T}u_t^i-\sum_{t=1}^{T}u_t^{i_t}\right],
\end{aligned}$$

where for the last equality, we used the fact that u_t^i is deterministic to swap the maximum and the expectation. □

The EXP3 strategy was introduced and first analyzed in [ACBFS02]. It consists in using the Exponential Weights Algorithm against estimators $(\hat{u}_t)_{t\geqslant1}$. Set $U_0=0$ and for $t\geqslant1$,

$$\begin{aligned}
z_t^i&=\frac{\exp\left(\eta_{t-1}U_{t-1}^i\right)}{\sum_{j=1}^{d}\exp(\eta_{t-1}U_{t-1}^j)},\qquad i\in\mathcal{I}\\
\hat{u}_t^i&=\frac{\mathbb{1}_{\{i_t=i\}}}{z_t^i}u_t^i,\qquad i\in\mathcal{I}\\
U_t&=U_{t-1}+\hat{u}_t.
\end{aligned}$$

Note that the estimator is well defined since z_t^i is always positive.

Theorem II.8.2. *Let $T\geqslant1$. Against any sequence of payoff vectors $(u_t)_{t\geqslant1}$ in $[-1,0]^d$, the EXP3 strategy with parameters $\eta_t=\sqrt{\log d/2dt}$ (for $t\geqslant1$) guarantees*

$$\mathbb{E}\left[\max_{i\in\mathcal{I}}\sum_{t=1}^{T}u_t^i-\sum_{t=1}^{T}u_t^{i_t}\right]\leqslant2\sqrt{2Td\log d}.$$

Proof. Since estimators $(\widehat{u}_t)_{t\geqslant 1}$ are in $\mathbb{R}^{\underline{d}}$, we can apply Theorem II.3.2 and take the expectation, which gives

$$\mathbb{E}\left[\max_{i\in\mathcal{I}}\sum_{t=1}^{\mathrm{T}}\widehat{u}_t^i-\sum_{t=1}^{\mathrm{T}}\langle\widehat{u}_t^i|z_t\rangle\right]\leqslant\frac{\log d}{\eta_{\mathrm{T}}}+\sum_{t=1}^{\mathrm{T}}\eta_{t-1}\mathbb{E}\left[\sum_{i=1}^{d}(\widehat{u}_t^i)^2z_t^i\right].$$

We deal with the expectation of the right-hand side as follows.

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{d}(\widehat{u}_t^i)^2z_t^i\right]&=\mathbb{E}\left[\sum_{i=1}^{d}\mathbb{E}\left[(\widehat{u}_t^i)^2z_t^i\,\middle|\,\mathcal{F}_t\right]\right]=\mathbb{E}\left[\sum_{i=1}^{d}\mathbb{E}\left[\frac{\mathbb{1}_{\{i_t=i\}}(u_t^i)^2}{z_t^i}\,\middle|\,\mathcal{F}_t\right]\right]\\&=\mathbb{E}\left[\sum_{i=1}^{d}\mathbb{E}\left[\mathbb{1}_{\{i_t=i\}}\,\middle|\,\mathcal{F}_t\right]\frac{(u_t^i)^2}{z_t^i}\right]=\mathbb{E}\left[\sum_{i=1}^{d}(u_t^i)^2\right]\leqslant d.\end{aligned}$$

Together with Lemma II.8.1, we get:

$$\mathbb{E}\left[\max_{i\in\mathcal{I}}\sum_{t=1}^{\mathrm{T}}u_t^i-\sum_{t=1}^{\mathrm{T}}u_t^{i_t}\right]\leqslant\frac{\log d}{\eta_{\mathrm{T}}}+d\sum_{t=1}^{\mathrm{T}}\eta_{t-1}.$$

Then the choice $\eta_t=\sqrt{\log d/2dt}$ gives the result. □

Note that bound (i) from Theorem II.3.2 was needed in this analysis. This bound holds for payoff vectors in $\mathbb{R}^{\underline{d}}$, or more generally, for payoff vectors whose components are *bounded from above*. This is why we needed to normalize payoff vectors $(u_t)_{t\geqslant 1}$ as losses (in e.g. $[-1,0]^d$), otherwise, the components $\mathbb{1}_{\{i_t=i\}}u_t^i/z_t^i$ of the estimators might have been positive and arbitrarily large since z_t^i can be arbitrarily small.

Theorem II.8.2 establishes an upper bound of order $\sqrt{\mathrm{T}d\log d}$. Besides, a lower bound of order $\sqrt{\mathrm{T}d}$ was given in [ACBFS02]. The (slight) gap between those two bounds was closed by [AB09], which introduced the *Implicitly Normalized Forecaster* strategy which provides an upper bound of order $\sqrt{\mathrm{T}d}$. This algorithm can be seen as part of a larger family of algorithms for bandit problems based on Greedy Online Mirror Descent—see [BCB12] Section 5 for a detailed presentation and applications. A well-chosen algorithm from this family is used in Chapter V to obtain an upper bound of order $\sqrt{\mathrm{T}s}$ in the case of s-sparse losses (i.e. payoff vectors in $\mathcal{U}^{-,s,d}$) and bandit feedback.

CHAPTER III

FOLLOW THE PERTURBED LEADER

In this chapter, we present the *Follow the Perturbed Leader* strategies (FTPL) and prove in Theorem III.3.1 that they actually belong to the family of Mirror Descent strategies from Section I.3 as soon as the law of the perturbation is absolutely continuous with respect to the Lebesgue measure.

III.1. Presentation

Like the Mirror Descent strategies, FTPL strategies were constructed as a modification of the FTL strategy mentioned in Section I.1. But instead of using a *deterministic* function h to regularize the map $\arg\max_{z\in\mathcal{Z}} \langle \cdot\,|z\rangle$, they involve a *random* perturbation. Specifically, let ξ be an integrable random variable in $\mathbb{R}^d$. Then, we define the FTPL strategy associated with ξ and parameters $(\eta_t)_{t\geqslant 1}$ as

$$z_t = \mathbb{E}\left[\arg\max_{z\in\mathcal{Z}} \left\langle \eta_{t-1}\sum_{s=1}^{t-1} u_s + \xi \middle| z \right\rangle\right],$$

where $(u_t)_{t\geqslant 1}$ are the payoff vectors in $\mathbb{R}^d$ and $\mathcal{Z}$ the set of actions of the Decision Maker.

From a computational perspective, the FTPL strategy has an advantage over the Mirror Descent strategy. The latter involves the computation of ∇h^* at a given point, i.e. solving a convex program on $\mathcal{Z}$, whereas the former may be computed in a Monte Carlo fashion by drawing samples of the random variable ξ, solving a linear programs over $\mathcal{Z}$, and then considering the average. This advantage is even more interesting in the experts setting from Section II.1 where the Decision Maker draws pure action $i_t \in \mathcal{I}$ according to probability distribution $z_t \in \Delta(\mathcal{I})$. Then,

$$i_t = \arg\max_{z\in\Delta_d} \left\langle \eta_{t-1}\sum_{s=1}^{t-1} u_s + \xi \middle| z \right\rangle$$

almost-surely belongs to one of the vertices of the simplex Δ_d, i.e. to $\mathcal{I}$, and its law is precisely z_t. Therefore, the explicit computation of z_t is unnecessary and only a single draw of the random perturbation ξ is needed.

III.2. Historical background

A strategy of this type was already proposed in Hannan's seminal paper [Han57]. FTPL was later rediscovered in [KV05] in which a random perturbation of density $y \longmapsto (\eta/2)^d e^{-\eta\|y\|_1}\,\mathrm{d}y$ was used to achieve a minimax optimal regret bound of order $\mathrm{O}(\sqrt{\mathrm{T}\log d})$ in the experts setting—see also [HP04]. An even simpler perturbation with independent components drawn according to the uniform distribution over $[0,1]$ has been shown to guarantee a $\mathrm{O}(\sqrt{\mathrm{T}d})$ regret bound in the experts setting (see e.g. Corollary 4.4 in [CBL06]). More recently, [ALST14] used a standard Gaussian perturbation $\xi \sim \mathcal{N}(0,\mathrm{I})$ to achieve minimax optimal regret bounds both in the experts setting and the ℓ_2–ℓ_2 setting (where both the actions of the Decision Maker and the payoff vectors belong to the Euclidean unit ball). [CH15] applied those techniques to online combinatorial optimization.

Applications of similar strategies to various settings include: [DLN13] where a Bernoulli coin flip is added to each component of each payoff vector, [NB13] which deals with the semi-bandit online combinatorial optimization problem, and [VEKW14] where the minimax optimal guarantee is achieved in the experts setting by setting each component of each payoff vector to zero or one with some probability.

FTPL and Mirror Descent strategies share many common properties and close links between those two families were long suspected. For instance, it is known that the Exponential Weights Algorithm studied in Section II.3 coincides with the FTPL strategy with a perturbation which follows the Gumbel distribution. [HS02] proved in the case $\mathcal{Z} = \Delta_d$ that FTPL strategies *are* Mirror Descent strategies. [ALST14] proposed a unifying framework which encompasses both FTPL and lazy Mirror Descent, and established in the one-dimensional case a bijection between the two families.

III.3. Reduction to Mirror Descent

The following theorem proves that a FTPL strategy is a Mirror Descent strategy as soon as the distribution of the random perturbation is absolutely continuous with respect to the Lebesgue measure. This result is quickly mentioned in the recent survey [ALT16]. We here give a detailed proof. One can see that a Mirror Descent strategy associated with a regularizer h and a FTPL strategy associated with a perturbation

ξ coincide as soon as

$$\nabla h^*(w) = \mathbb{E}\left[\arg\max_{z\in\mathcal{Z}} \langle w + \xi | z\rangle\right], \quad \text{for all } w \in \mathbb{R}^d.$$

Theorem III.3.1. *Let ξ be in integrable random variable in $\mathbb{R}^d$ whose distribution is absolutely continuous with respect to the Lebesgue measure. Let $\mathcal{Z}$ be a nonempty convex compact subset of $\mathbb{R}^d$. Then, there exists a regularizer h on $\mathcal{Z}$ such that*

$$\nabla h^*(w) = \mathbb{E}\left[\arg\max_{z\in\mathcal{Z}} \langle w + \xi | z\rangle\right], \quad w \in \mathbb{R}^d.$$

Moreover,

$$\delta_h \leqslant \mathbb{E}\left[\max_{z\in\mathcal{Z}} \langle \xi | z\rangle\right] - \max_{z\in\mathcal{Z}} \langle \mathbb{E}\left[\xi\right] | z\rangle .$$

Proof. Consider $\Phi : \mathbb{R}^d \to \mathbb{R}$ defined by

$$\Phi(w) = \mathbb{E}\left[\max_{z\in\mathcal{Z}} \langle w + \xi | z\rangle\right], \quad w \in \mathbb{R}^d.$$

The map $w \longmapsto \max_{z\in\mathcal{Z}} \langle w + y | z\rangle$ being convex for all $y \in \mathbb{R}^d$, [Ber73, Lemma 2.1] assures that Φ is convex. Besides, the distribution of ξ being absolutely continuous with respect to the Lebesgue measure, [Ber73, Proposition 2.4] guarantees that Φ is differentiable on $\mathbb{R}^d$. We now define $h := \Phi^*$ to be the Legendre–Fenchel transform of Φ. Φ being convex, lower semicontinuous and proper, Moreau's theorem gives $\Phi = \Phi^{**} = h^*$. Therefore, this function will be called h^* from now on.

Let us prove that h is a regularizer on $\mathcal{Z}$. h being a Legendre–Fenchel transform by definition, it is convex and lower semicontinuous. It is also strictly convex since h^* is differentiable. Let us prove that the domain of h is $\mathcal{Z}$. Using the definition of h, we can write:

$$\begin{aligned} h(z) &= \sup_{w\in\mathbb{R}^d} \left\{ \langle w | z\rangle - \mathbb{E}\left[\max_{z'\in\mathcal{Z}} \langle w + \xi | z'\rangle\right] \right\} \\ &= \sup_{w\in\mathbb{R}^d} \left\{ \mathbb{E}\left[\langle w | z\rangle - \max_{z'\in\mathcal{Z}} \langle w + \xi | z'\rangle\right] \right\}. \end{aligned}$$

If $z \in \mathcal{Z}$,

$$h(z) \leqslant \sup_{w\in\mathbb{R}^d} \mathbb{E}\left[\langle w | z\rangle - \langle w + \xi | z\rangle\right] = -\langle \mathbb{E}\left[\xi\right] | z\rangle < +\infty.$$

We now assume $z \notin \mathcal{Z}$ and write

$$h(z) = \sup_{w\in\mathbb{R}^d} \mathbb{E}\left[\min_{z'\in\mathcal{Z}} \left\{ \langle w | z - z'\rangle - \langle \xi | z'\rangle \right\}\right].$$

The second scalar product above is bounded as $\langle \xi | z' \rangle \leqslant \|\xi\|_1 \|\mathcal{Z}\|_\infty$. Therefore,

$$h(z) \geqslant \sup_{w \in \mathbb{R}^d} \min_{z' \in \mathcal{Z}} \langle w | z - z' \rangle - \|\mathcal{Z}\|_\infty \cdot \mathbb{E}\left[\|\xi\|_1\right].$$

The quantity $\langle w | z - z' \rangle$ is affine in w and in z', $\mathbb{R}^d$ is convex, and $\mathcal{Z}$ is convex and compact. We can therefore apply Sion's minimax theorem to get:

$$h(z) \geqslant \min_{z' \in \mathcal{Z}} \sup_{w \in \mathbb{R}^d} \langle w | z - z' \rangle - \|\mathcal{Z}\|_\infty \cdot \mathbb{E}\left[\|\xi\|_1\right].$$

Let $\lambda > 0$. We now choose particular vector $w = \lambda(z - z')$ instead of taking the supremum over $w \in \mathbb{R}^d$. This gives

$$\begin{aligned} h(z) &\geqslant \min_{z' \in \mathcal{Z}} \langle \lambda(z - z') | z - z' \rangle - \|\mathcal{Z}\|_\infty \cdot \mathbb{E}\left[\|\xi\|_1\right] \\ &= \lambda \cdot \min_{z' \in \mathcal{Z}} \|z - z'\|_2^2 - \|\mathcal{Z}\|_\infty \cdot \mathbb{E}\left[\|\xi\|_1\right]. \end{aligned}$$

The set $\mathcal{Z}$ being compact and z being outside of $\mathcal{Z}$, the distance from z to $\mathcal{Z}$ is positive. The above inequality being true for all $\lambda > 0$, and because $\mathbb{E}\left[\|\xi\|_1\right] < +\infty$, we have $h(z) = +\infty$. The domain of h is indeed $\mathcal{Z}$ and h is a regularizer on $\mathcal{Z}$.

Finally, let us prove the equality from the statement of the theorem. Let us fix $y \in \mathbb{R}^d$ and consider the convex function $\phi_y(w) = \max_{z \in \mathcal{Z}} \langle w + y | z \rangle$ defined for $w \in \mathbb{R}^d$. Then, for all $w \in \mathbb{R}^d$, we have the following inclusion:

$$\arg\max_{z \in \mathcal{Z}} \langle w + y | z \rangle \subset \partial \phi_y(w).$$

Indeed, for $z_* \in \arg\max_{z \in \mathcal{Z}} \langle w + y | z \rangle$ and for all $w' \in \mathbb{R}^d$, we have

$$\begin{aligned} \phi_y(w') - \phi_y(w) &= \max_{z \in \mathcal{Z}} \langle w' + y | z \rangle - \max_{z \in \mathcal{Z}} \langle w + y | z \rangle \\ &\geqslant \langle w' + y | z_* \rangle - \langle w + y | z_* \rangle \\ &= \langle w' - w | z_* \rangle, \end{aligned}$$

in other words, $z_* \in \partial \phi_y(w)$ and the inclusion is proved. We then replace y by random variable ξ and take the expectation on both sides to get

$$\mathbb{E}\left[\arg\max_{z \in \mathcal{Z}} \langle w + \xi | z \rangle\right] \subset \mathbb{E}\left[\partial \phi_\xi(w)\right] = \partial h^*(w),$$

where the last equality comes from [Ber73, Proposition 2.2]. But we know that h^* is differentiable. In other words, $\partial h^*(w)$ is a singleton for all $w \in \mathbb{R}^d$ and we have

$$\nabla h^*(w) = \mathbb{E}\left[\arg\max_{z \in \mathcal{Z}} \langle w + \xi | z \rangle\right], \quad w \in \mathbb{R}^d.$$

We now turn to $\delta_h = \max_{\mathcal{Z}} h - \min_{\mathcal{Z}} h$. First, we have

$$\min_{\mathcal{Z}} h = -h^*(0) = -\mathbb{E}\left[\max_{z\in\mathcal{Z}} \langle \xi | z \rangle\right].$$

Second, we have seen above that $h(z) \leqslant -\langle \mathbb{E}[\xi] | z \rangle$ for $y \in \mathcal{Z}$. Taking the maximum over $z \in \mathcal{Z}$ gives

$$\max_{\mathcal{Z}} h \leqslant -\min_{z\in\mathcal{Z}} \langle \mathbb{E}[\xi] | z \rangle,$$

and the result follows. □

III.4. Discussion

Theorem III.3.1 provides an alternative method for defining a Mirror Descent strategy, which makes the explicit choice of a regularizer h unnecessary. However, some properties of h must still be known in order to turn the general regret bound from Theorem I.3.1 into an explicit one. The first term δ_h/η_T (from the general bound) can be immediately taken care of, since Theorem III.3.1 provides an upper bound on δ_h. The second term, which involves the Bregman divergences, is more challenging. The probabilistic expression $h^*(w) = \mathbb{E}\left[\max_{z\in\mathcal{Z}} \langle w + \xi | z \rangle\right]$ does not seem to provide any handy expression for the Bregman divergence associated with h^*. One way of dealing with those is to establish strong convexity for regularizer h. In the case of a standard Gaussian perturbation $\xi \sim \mathcal{N}(0, I)$, [AHR12] used a characterization of the strongly convexity of h which involves the Hessian of h^*. An interesting direction of research would be the study of the strong convexity of regularizer h as a function of the properties of the distribution of perturbation ξ.

CHAPTER IV

MIRROR DESCENT FOR APPROACHABILITY

We do not aim in this chapter at giving an overview of the theory of approachability. We rather focus on a framework in which Mirror Descent strategies can be defined naturally. We then illustrate the unifying character of this approach by applying it to the construction of optimal strategies for online combinatorial optimization and internal/swap regret minimization.

The first notice of the link between regret minimization and approachability goes back to [Bla54, Han57]. More recently, [HMC01] constructed a wide class of potential-based approachability strategies and derived regret minimizing strategies using a reduction (of the regret minimization problem to an approachability problem) based on the negative orthant. Conversely, [Per15] adapted the Exponential Weights Algorithm to approachability. In a similar spirit, [ABH11] proposed a generic scheme based on convex cones for converting regret minimizing strategies into approachability strategies (see also [Shi15]).

We aim at providing a unified approach. We build upon the idea proposed by [ABH11] and further develop it: instead of restricting our attention to strategies which minimize the Euclidean distance of the average payoff to the target set, we allow for a much wider choice of distance-like quantities to be minimized (see the choice of *generators* in Section IV.2 below). This flexibility will allow the construction of tailored strategies for online combinatorial optimization and internal/swap regret minimization. The tools and ideas introduced in this chapter will also be used in Chapter VI for the construction of strategies with optimal convergence rates in the problem of approachability with partial monitoring.

IV.1. Model

Let $\mathcal{V}$ be a finite-dimensional vector space and $\mathcal{V}^*$ its dual. The latter will be the *payoff space*. Let $\mathcal{X}$ be the *action set* for the Decision Maker about which we assume no particular structure. Let $\mathcal{G}$ be a set of payoff functions $g : \mathcal{X} \to \mathcal{V}^*$. The play goes

as follows. At time $t \geqslant 1$, the Decision Maker

- chooses action $x_t \in \mathscr{X}$;
- observes *vector payoff* $u_t := g_t(x_t) \in \mathscr{V}^*$,

where $(g_t)_{t\geqslant 1}$ is a sequence of payoff functions in $\mathscr{G}$ chosen by Nature. Formally, a strategy σ for the Decision Maker is a sequence of maps $\sigma = (\sigma_t)_{t\geqslant 1}$ where $\sigma_t : (\mathscr{X} \times \mathscr{V}^*)^{t-1} \to \mathscr{X}$ so that for a given strategy σ and a given sequence of payoff functions $(g_t)_{t\geqslant 1}$ we have

$$x_t = \sigma_t(x_1, u_1, \dots, x_{t-1}, u_{t-1}).$$

Analogously to Section I.1, g_t may depend on anything that has happened before, including x_t, and we may assume that Nature is aware of the strategy used by the Decision Maker.

The problem involves a *target set* $\mathscr{C} \subset \mathscr{V}^*$ which we assume to be a closed convex cone [1]. Definitions and properties about closed convex cones are gathered in the next section. The goal is to construct strategies which guarantee that the average payoff $\bar{u}_T := \frac{1}{T}\sum_{t=1}^{T} u_t$ is *close* to the target $\mathscr{C}$ in a sense that will be made precise.

IV.2. Closed convex cones and support functions

IV.2.1. Closed convex cones

Throughout the paragraph, $\mathscr{W}$ will be a finite-dimensional vector space and $\mathscr{W}^*$ its dual.

Definition IV.2.1. A nonempty subset $\mathscr{C}$ of $\mathscr{W}$ is a *closed convex cone* if it is closed and if for all $w, w' \in \mathscr{C}$ and $\lambda \in \mathbb{R}_+$, we have $w + w' \in \mathscr{C}$ and $\lambda w \in \mathscr{C}$.

The following proposition gathers a few immediate properties.

Proposition IV.2.2. *(i) A closed convex cone is convex.*

(ii) An intersection of closed convex cones is a closed convex cone.

(iii) A Cartesian product of closed convex cones is a closed convex cone.

(iv) A half-space of the form $\{w \in \mathscr{W} \mid \langle z|w\rangle \leqslant 0\}$ *(for some* $z \in \mathscr{W}^*$*) is a closed convex cone.*

Definition IV.2.3. Let $\mathscr{A}$ be a subset of $\mathscr{W}$. The *polar cone* of $\mathscr{A}$ is a subset of the dual space $\mathscr{W}^*$ defined by

$$\mathscr{A}^\circ = \{z \in \mathscr{W}^* \mid \forall w \in \mathscr{A},\ \langle w|z\rangle \leqslant 0\}.$$

The following proposition is an immediate consequence of the bipolar theorem—see e.g. Theorem 3.3.14 in [BL10].

1. For the case where target set is a closed convex set but not a cone, we refer to [ABH11] where a conversion scheme into an auxiliary problem where the target is a cone is presented.

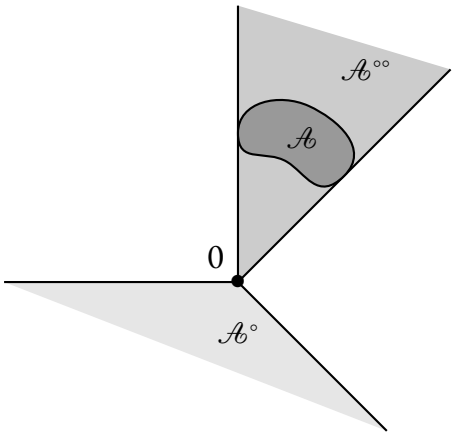


Figure IV.1. — The polar cone of a set $\mathscr{A}$ and the bipolar

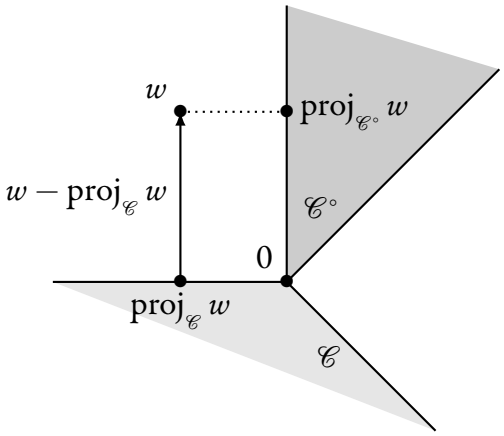


Figure IV.2. — Illustration of Proposition IV.2.5

Proposition IV.2.4. *Let $\mathscr{A}$ be a subset of $\mathscr{W}$.*

(i) $\mathscr{A}^{\circ\circ}$ is the smallest closed convex cone containing $\mathscr{A}$.

(ii) If $\mathscr{A}$ is closed and convex, then $\mathscr{A}^{\circ\circ} = \mathbb{R}_+\mathscr{A}$.

(iii) If $\mathscr{A}$ is a closed convex cone, then $\mathscr{A}^{\circ\circ} = \mathscr{A}$.

The following statement is a simpler version of Moreau's decomposition theorem [Mor62].

Proposition IV.2.5. *Assume that $\mathscr{W}$ is an Euclidean space. We identify $\mathscr{W}$ and its dual space $\mathscr{W}^*$. Let $\mathscr{C}$ be a closed convex cone in $\mathscr{W}$, and $w \in \mathscr{W}$. Then, $w - \operatorname{proj}_{\mathscr{C}} w = \operatorname{proj}_{\mathscr{C}^\circ} w$, where* proj *denotes the Euclidean projection. In particular, $w - \operatorname{proj}_{\mathscr{C}} w$ belongs to $\mathscr{C}^\circ$.*

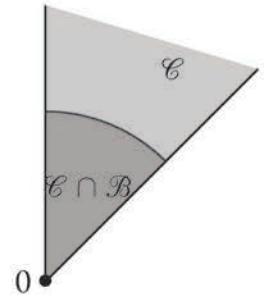


Figure IV.3. — $\mathscr{C} \cap \mathscr{B}$ is a generator of $\mathscr{C}$

Proposition IV.2.6. *Let $\phi : \mathscr{W} \to \tilde{\mathscr{W}}$ be a linear application between two finite-dimensional vector spaces $\mathscr{W}$ and $\tilde{\mathscr{W}}$, ϕ^* its transpose, $\mathscr{C}$ and $\tilde{\mathscr{C}}$ closed convex cones in $\mathscr{W}$ and $\tilde{\mathscr{W}}$ respectively.*

(i) $\phi(\mathscr{C})$ is a closed convex cone.

(ii) $\phi^{-1}(\tilde{\mathscr{C}}) = \phi^*(\tilde{\mathscr{C}}^\circ)^\circ$. In particular, $\phi^{-1}(\tilde{\mathscr{C}})$ is a closed convex cone.

Proof. Property (i) is obvious. We prove property (ii) as follows. For $w \in \mathscr{W}$,

$$\begin{aligned} w \in \phi^{-1}(\tilde{\mathscr{C}}) &\iff \phi(w) \in \tilde{\mathscr{C}} \iff \phi(w) \in \tilde{\mathscr{C}}^{\circ\circ} \\ &\iff \forall \tilde{z} \in \tilde{\mathscr{C}}^\circ, \quad \langle \tilde{z} | \phi(w) \rangle \leqslant 0 \\ &\iff \forall z \in \tilde{\mathscr{C}}^\circ, \quad \langle \phi^*(\tilde{z}) | w \rangle \leqslant 0 \\ &\iff w \in \phi^*(\tilde{\mathscr{C}}^\circ)^\circ. \end{aligned}$$

Therefore, $\phi^{-1}(\tilde{\mathscr{C}})$ is a closed convex cone because it is a polar cone. □

Definition IV.2.7. Let $\mathscr{C}$ be a closed convex cone. A set $\mathscr{Z}$ is a *generator* of $\mathscr{C}$ if it is convex, compact and if $\mathbb{R}_+ \mathscr{Z} = \mathscr{C}$.

Note that there always exists a generator: as illustrated in Figure IV.3, the set $\mathscr{B} \cap \mathscr{C}$ is one, where $\mathscr{B}$ is the closed unit ball of some norm $\|\cdot\|$. It is indeed nonempty, convex as the intersection of two convex sets, and for any point $z \in \mathscr{C}$, $z/\|z\|$ belongs to $\mathscr{B} \cap \mathscr{C}$, so that $\mathbb{R}_+(\mathscr{B} \cap \mathscr{C}) = \mathscr{C}$.

Proposition IV.2.8. *(i) If $\mathscr{W} = \mathscr{W}^* = \mathbb{R}^d$, the negative orthant $\mathbb{R}^d_-$ is a closed convex cone and $(\mathbb{R}^d_-)^\circ = \mathbb{R}^d_+$. Moreover, Δ_d is a generator of $\mathbb{R}^d_+$.*

(ii) If $\mathscr{Z}$ is a nonempty convex compact subset of $\mathscr{W}$, then $\mathscr{Z}$ is a generator of $\mathscr{Z}^{\circ\circ} = \mathbb{R}_+ \mathscr{Z}$.

IV.2.2. Support Functions

Definition IV.2.9. For a nonempty subset $\mathcal{Z} \subset \mathcal{V}$, the application $\mathrm{I}^*_{\mathcal{Z}} : \mathcal{V}^* \to \mathbb{R} \cup \{+\infty\}$ defined by

$$\mathrm{I}^*_{\mathcal{Z}}(w) = \sup_{z \in \mathcal{Z}} \langle w|z\rangle\,, \quad w \in \mathcal{V}^*,$$

is called the *support function* of $\mathcal{Z}$.

The support function can be written as the Legendre–Fenchel transform of the indicator function of $\mathcal{Z}$. It is therefore convex. Moreover, in the case where $\mathcal{Z}$ is a generator of the polar cone $\mathcal{C}^\circ$ of some closed convex cone $\mathcal{C} \subset \mathcal{V}^*$, the properties of $\mathrm{I}^*_{\mathcal{Z}}$ make it suitable for measuring how far a point of $\mathcal{V}^*$ is from $\mathcal{C}$. Indeed, it is easy to check that $\mathrm{I}^*_{\mathcal{Z}}$ is then real-valued, continuous, and that for all points $w \in \mathcal{V}^*$,

$$\mathrm{I}^*_{\mathcal{Z}}(w) \leqslant 0 \quad \Longleftrightarrow \quad w \in \mathcal{C}\,.$$

The following proposition shows that, in particular, the distance to a closed convex cone $\mathcal{C}$ with respect to an arbitraty norm can be written as a support function.

Proposition IV.2.10. *Let $\mathcal{C}$ be a closed convex cone in $\mathcal{V}^*$, $\|\cdot\|$ a norm on $\mathcal{V}$ and $\|\cdot\|_*$ its dual norm on $\mathcal{V}^*$. Then,*

$$\inf_{w' \in \mathcal{C}} \|w' - w\|_* = \mathrm{I}^*_{\mathcal{B} \cap \mathcal{C}^\circ}(w), \quad w \in \mathcal{V}^*,$$

where $\mathcal{B}$ is the closed unit ball for $\|\cdot\|$.

Proof. Let $w \in \mathcal{V}^*$. Using the definition of the dual norm and Sion's minimax theorem:

$$\inf_{w' \in \mathcal{C}} \|w' - w\|_* = \inf_{w' \in \mathcal{C}} \sup_{z \in \mathcal{B}} \langle w - w'|z\rangle = \sup_{z \in \mathcal{B}} \inf_{w' \in \mathcal{C}} \{\langle w|z\rangle - \langle w'|z\rangle\}\,.$$

Suppose z does not belong to $\mathcal{C}^\circ$. Then, there exists $w'_0 \in \mathcal{C}$ such that $\langle w'_0|z\rangle > 0$. $\mathcal{C}$ being stable by multiplication by $\mathbb{R}_+$, the quantity $\langle w'|z\rangle$ (with $w' \in \mathcal{C}$) can be made arbitrarily large, and thus the above infimum is equal to $-\infty$. Therefore, we can restrict the above supremum to $\mathcal{B} \cap \mathcal{C}^\circ$. We thus have

$$\inf_{w' \in \mathcal{C}} \|w' - w\|_* = \sup_{z \in \mathcal{B} \cap \mathcal{C}^\circ} \left\{ \langle w|z\rangle - \sup_{w' \in \mathcal{C}} \langle w'|z\rangle \right\}.$$

The above embedded supremum is zero because for $z \in \mathcal{B} \cap \mathcal{C}^\circ$ and $w' \in \mathcal{C}$ we obviously have $\langle w'|z\rangle \leqslant 0$, and 0 is attained with $w' = 0$. Finally,

$$\inf_{w' \in \mathcal{C}} \|w' - w\|_* = \sup_{z \in \mathcal{B} \cap \mathcal{C}^\circ} \langle w|z\rangle = \mathrm{I}^*_{\mathcal{B} \cap \mathcal{C}^\circ}(w).$$

□

IV.3. Mirror Descent strategies

We now construct the Mirror Descent strategies for the model introduced in Section IV.1 and derive guarantees using regret bounds from Theorem I.3.1. We will propose an intuitive description of these strategies in Section IV.4. Remember that $\mathcal{G}$ is the set of payoff functions. Let us state the all important Blackwell's condition which will be key in the construction and the analysis of the strategies.

Definition IV.3.1. A closed convex cone $\mathcal{C}$ of the payoff space $\mathcal{V}^*$ is a *$\mathcal{G}$-B-set* if

$$\forall z \in \mathcal{C}^\circ,\ \exists x(z) \in \mathcal{X},\ \forall g \in \mathcal{G}, \quad \langle g(x(z)) | z \rangle \leqslant 0.$$

Such an application $x : \mathcal{C}^\circ \to \mathcal{X}$ is called a *$(\mathcal{G}, \mathcal{C})$-oracle*.

Let $\mathcal{C}$ be a $\mathcal{G}$-B-set and $x : \mathcal{C}^\circ \to \mathcal{X}$ a $(\mathcal{G}, \mathcal{C})$-oracle. Let $\mathcal{Z} \subset \mathcal{V}$ be a generator of $\mathcal{C}^\circ$, h a regularizer on $\mathcal{Z}$, and $(\eta_t)_{t \geqslant 1}$ a positive and nonincreasing sequence of parameters. The associated strategy is then defined by $\mathrm{U}_0 = 0$ and for $t \geqslant 1$,

$$\begin{aligned} \text{compute} \quad & z_t = \nabla h^*(\eta_{t-1}\mathrm{U}_{t-1}) \\ \text{play} \quad & x_t = x(z_t) \\ \text{observe} \quad & u_t = g_t(x_t) \\ \text{update} \quad & \mathrm{U}_t = \mathrm{U}_{t-1} + u_t. \end{aligned}$$

Contrary to Section I.3, the set of actions of the Decision Maker is $\mathcal{X}$ and not $\mathcal{Z}$.

The following theorem provides upper bounds on $\mathrm{I}^*_{\mathcal{Z}}(\bar{u}_\mathrm{T})$ (where $\bar{u}_\mathrm{T} = \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} u_t$ is the average payoff) and not only the Euclidean distance from $\bar{u}_\mathrm{T}$ to $\mathcal{C}$, which is a special case—see Proposition IV.2.10. Therefore, the choice of generator $\mathcal{Z}$ determines the quantity that is minimized by the strategy. We present in Sections IV.7 and IV.8 examples of problems where a judicious choice of generator $\mathcal{Z}$ allows $\mathrm{I}^*_{\mathcal{Z}}(\bar{u}_\mathrm{T})$ to be actually equal to the quantity the Decision Maker aims at minimizing and therefore provides tailored strategies.

Theorem IV.3.2. *Let* $\mathrm{T} \geqslant 1$ *an integer and* $\mathrm{M}, \mathrm{K} > 0$.

(i) Against any sequence $(g_t)_{t \geqslant 1}$ of payoff functions in $\mathcal{G}$, the above strategy guarantees

$$\mathrm{I}^*_{\mathcal{Z}}(\bar{u}_\mathrm{T}) \leqslant \frac{\delta_h}{\mathrm{T}\eta_\mathrm{T}} + \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} \frac{1}{\eta_{t-1}} \mathrm{D}_{h^*}(\eta_{t-1}\mathrm{U}_t, \eta_{t-1}\mathrm{U}_{t-1}),$$

where $\eta_0 = \eta_1$.

(ii) Moreover, if h is K-strongly convex with respect to a norm $\|\cdot\|$, then

$$\mathrm{I}^*_{\mathcal{Z}}(\bar{u}_\mathrm{T}) \leqslant \frac{\delta_h}{\mathrm{T}\eta_\mathrm{T}} + \frac{1}{2\mathrm{KT}} \sum_{t=1}^{\mathrm{T}} \eta_{t-1} \|g_t(x_t)\|_*^2.$$

(iii) Moreover, if $\|g(x)\|_* \leqslant \mathrm{M}$ *(for all* $g \in \mathscr{G}$ *and* $x \in \mathscr{X}$*), the choice* $\eta_t = \sqrt{\delta_h \mathrm{K}/\mathrm{M}^2 t}$ *(for* $t \geqslant 1$*) guarantees*

$$\mathrm{I}^*_{\mathscr{Z}}(\bar{u}_{\mathrm{T}}) \leqslant 2\mathrm{M}\sqrt{\frac{\delta_h}{\mathrm{KT}}}.$$

Proof. The strategy can be interpreted as regret minimization Mirror Descent strategy from Section I.3 where $\mathscr{Z}$ would be the action set, z_t the actions of the Decision Maker, and $u_t = g_t(x_t)$ the payoff vectors. The corresponding regret is

$$\mathrm{Reg}_{\mathrm{T}} = \max_{z\in\mathscr{Z}} \sum_{t=1}^{\mathrm{T}} \langle u_t | z\rangle - \sum_{t=1}^{\mathrm{T}} \langle u_t | z_t\rangle .$$

The first term above can be written

$$\max_{z\in\mathscr{Z}} \sum_{t=1}^{\mathrm{T}} \langle u_t | z\rangle = \mathrm{T}\cdot \max_{z\in\mathscr{Z}} \left\langle \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} u_t \middle| z \right\rangle = \mathrm{T}\cdot \mathrm{I}^*_{\mathscr{Z}}(\bar{u}_{\mathrm{T}}),$$

whereas the second sum is nonpositive because each term is. Indeed, by definition of the strategy, and because x is a $(\mathscr{G},\mathscr{C})$-oracle.

$$\langle u_t | z_t\rangle = \langle g_t(x_t) | z_t\rangle = \langle g_t(x(z_t)) | z_t\rangle \leqslant 0.$$

Therefore $\mathrm{I}^*_{\mathscr{Z}}(\bar{u}_{\mathrm{T}}) \leqslant \frac{1}{\mathrm{T}}\mathrm{Reg}_{\mathrm{T}}$ and the regret bounds from Theorem I.3.1 give the results. □

IV.4. Smooth potential interpretation

We now propose an intuitive description of the strategies defined in Section IV.3 based on the idea that h^* serves as a potential, and provide an interpretation of the bound from Theorem IV.3.2. Similar ideas were used by [HMC01] for the construction of their so-called *potential-based* approachability strategies (see also [Per15]). Our approach provides more precise bounds thanks to the use of convex duality, variable parameters, and the large choice of generators $\mathscr{Z}$.

For simplicity, we restrict this discussion to Mirror Descent strategies with a constant parameter $\eta_t = \eta > 0$. The Decision Maker aims at minimizing $\mathrm{I}^*_{\mathscr{Z}}(\bar{u}_t)$. Instead of working directly with this quantity, we consider h^* which is a smooth approximation of $\arg\max_{z\in\mathscr{Z}} \langle \cdot | z\rangle$, as seen in Section I.3. We write

$$\mathrm{I}^*_{\mathscr{Z}}(\bar{u}_t) = \frac{1}{t\eta}\mathrm{I}^*_{\mathscr{Z}}(\eta \mathrm{U}_t) = \frac{1}{t\eta}\arg\max_{z\in\mathscr{Z}} \langle \eta \mathrm{U}_t | z\rangle \approx \frac{1}{t\eta} h^*(\eta \mathrm{U}_t).$$

We now ask the following question. Can the Decision Maker make sure that $h^*(\eta U_t)$ does not increase too much over time (at least when parameter η is small)? Let us write the following first-order Taylor approximation, which makes sense when η is small:

$$h^*(\eta U_{t+1}) - h^*(\eta U_t) \approx \eta \left\langle \nabla h^*(\eta U_t) | u_{t+1} \right\rangle .$$

The question then becomes: knowing vector ηU_t, can the Decision Maker play an action $x_{t+1} \in \mathcal{X}$ such that for all payoff functions g_{t+1} chosen by Nature,

$$\langle \nabla h^*(\eta U_t) | g_{t+1}(x_{t+1}) \rangle \leqslant 0 \quad ?$$

One can easily check that this condition is equivalent to $\mathscr{C}$ being a $\mathscr{G}$-B-set. When this is the case, and when the Decision Maker plays accordingly, we obtain after T stages:

$$\mathrm{I}^*_{\mathbb{Z}}(\bar{u}_T) \approx \frac{1}{t\eta} h^*(\eta U_t) \approx \frac{1}{t\eta} \sum_{t=1}^{T} (h^*(\eta U_t) - h^*(\eta U_{t-1})) \lesssim 0.$$

As a matter of fact, the first two approximations ("$\approx$") result in the first term $\delta_h/\eta T$ of the bound from Theorem IV.3.2, whereas the first-order Taylor approximation ("$\lesssim$") gives the second term involving the Bregman divergences.

IV.5. Blackwell's strategy

We recall the definition of Blackwell's strategy [Bla56] and show that it belongs to the family of Mirror Descent strategies defined in Section IV.3. We consider $\mathscr{V} = \mathscr{V}^* = \mathbb{R}^d$ equipped with its Euclidean structure. Let $\mathscr{C} \subset \mathbb{R}^d$ be a closed convex cone which we assume to be a $\mathscr{G}$-B-set and $x : \mathscr{C}^\circ \to \mathcal{X}$ a $(\mathscr{G}, \mathscr{C})$-oracle. It follows from Definition IV.3.1 that it is always possible to choose an oracle x which satisfies

$$z = \lambda z' \text{ for some } \lambda > 0 \quad \Longrightarrow \quad x(z) = x(z'), \quad z, z' \in \mathscr{C}^\circ. \tag{IV.1}$$

We assume in this section that oracle x satisfies this property.

The Blackwell strategy is defined by:

$$x_t = x\left(\bar{u}_{t-1} - \operatorname*{proj}_{\mathscr{C}} \bar{u}_{t-1}\right), \quad t \geqslant 1,$$

where $\operatorname{proj}_{\mathscr{C}}$ is the Euclidean projection onto $\mathscr{C}$. It can be rewritten, using Proposition IV.2.5, as

$$x_t = x\left(\operatorname*{proj}_{\mathscr{C}^\circ} \bar{u}_{t-1}\right), \quad t \geqslant 1.$$

Theorem IV.5.1. *Let $\mathcal{Z} = \mathcal{C}^\circ \cap \mathcal{B}$ where $\mathcal{B}$ denotes the Euclidean ball, and h_2 the Euclidean regularizer on $\mathcal{Z}$. The Blackwell strategy and the Mirror Descent strategy associated with h_2 and any sequence of positive and nonincreasing parameters $(\eta_t)_{t\geqslant 1}$ coincide. In other words,*

$$x\left(\bar{u}_{t-1} - \operatorname*{proj}_{\mathcal{C}} \bar{u}_{t-1}\right) = x\left(\nabla h_2^*(\eta_{t-1} U_{t-1})\right), \quad t \geqslant 1.$$

Proof. Recall that the Euclidean projection $\operatorname{proj}_{\mathcal{E}} w$ of a point w on a closed convex set $\mathcal{E}$ is the only point in $\mathcal{E}$ satisfying

$$\forall w' \in \mathcal{E}, \quad \left\langle w - \operatorname*{proj}_{\mathcal{E}} w \middle| w' - \operatorname*{proj}_{\mathcal{E}} w \right\rangle \leqslant 0. \tag{IV.2}$$

This characterization will be needed later.

Remember from Proposition I.2.10 that $\nabla h_2^* = \operatorname{proj}_{\mathcal{C}^\circ \cap \mathcal{B}}$. Since oracle x satisfies property (IV.1), it is enough to prove that for all $u \in \mathbb{R}^d$ and $\mu > 0$,

$$\operatorname*{proj}_{\mathcal{C}^\circ} u \in \mathbb{R}_+^* \operatorname*{proj}_{\mathcal{C}^\circ \cap \mathcal{B}} (\mu u).$$

Besides, $\mathcal{C}^\circ$ being a closed convex cone, $\operatorname{proj}_{\mathcal{C}^\circ}(\mu u) = \mu \operatorname{proj}_{\mathcal{C}^\circ} u$. It is therefore equivalent to prove that for all $w \in \mathbb{R}^d$,

$$\operatorname*{proj}_{\mathcal{C}^\circ} w \in \mathbb{R}_+^* \operatorname*{proj}_{\mathcal{C}^\circ \cap \mathcal{B}} w. \tag{IV.3}$$

Let $w \in \mathbb{R}^d$. If $\left\|\operatorname{proj}_{\mathcal{C}^\circ} w\right\|_2 \leqslant 1$, then obviously $\operatorname{proj}_{\mathcal{C}^\circ} w = \operatorname{proj}_{\mathcal{C}^\circ \cap \mathcal{B}} w$ as shown in Figure IV.4 and (IV.3) is true. We now assume that $\left\|\operatorname{proj}_{\mathcal{C}^\circ} w\right\|_2 > 1$. We define

$$w_0 := \frac{\operatorname{proj}_{\mathcal{C}^\circ} w}{\left\|\operatorname{proj}_{\mathcal{C}^\circ} w\right\|_2}.$$

Using characterization (IV.2), we aim at proving that $w_0 = \operatorname{proj}_{\mathcal{C}^\circ \cap \mathcal{B}} w$ (see Figure IV.5), which would prove (IV.3). First, w_0 belongs to $\mathcal{C}^\circ \cap \mathcal{B}$ by definition. Let $w' \in \mathcal{C}^\circ \cap \mathcal{B}$. For short, denote $w_1 = \operatorname{proj}_{\mathcal{C}^\circ} w$.

$$\begin{aligned}
\langle w - w_0 | w' - w_0 \rangle &= \langle w - w_1 + w_1 - w_0 | w' - w_0 \rangle \\
&= \langle w - w_1 | w' - w_0 \rangle + \langle w_1 - w_0 | w' - w_0 \rangle \\
&= \frac{1}{\|w_1\|} \langle w - w_1 | \|w_1\| w' - w_1 \rangle + \langle w_1 - w_0 | w' - w_0 \rangle .
\end{aligned}$$

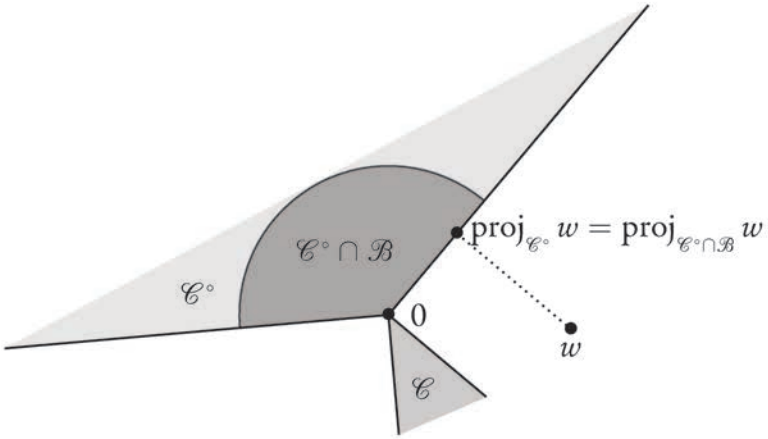


Figure IV.4. — In the case where $\left\|\mathrm{proj}_{\mathscr{C}^\circ} w\right\|_2 \leqslant 1$, we have $\mathrm{proj}_{\mathscr{C}^\circ} w = \mathrm{proj}_{\mathscr{C}^\circ \cap \mathscr{B}} w$

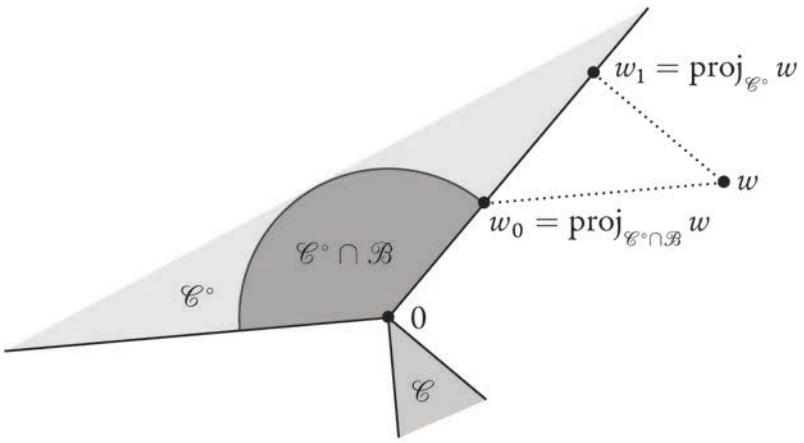


Figure IV.5. — In the case where $\left\|\mathrm{proj}_{\mathscr{C}^\circ} w\right\|_2 > 1$, we have $w_0 = \mathrm{proj}_{\mathscr{C}^\circ \cap \mathscr{B}} w$

The first scalar product above is nonpositive by characterization of $w_1 = \mathrm{proj}_{\mathscr{C}^\circ} w$, because $\|w_1\| w' \in \mathscr{C}^\circ$. Let us prove that the second scalar product is also nonpositive. For all $w'' \in \mathscr{C}^\circ \cap \mathscr{B}$, we have

$$\|w_1 - w''\| \geqslant |\|w_1\| - \|w''\|| \geqslant \|w_1\| - 1 = \|w_1 - w_0\|,$$

which means that $w_0 = \mathrm{proj}_{\mathscr{C}^\circ \cap \mathscr{B}} w_1$. Thus, $\langle w_1 - w_0 | w' - w_0 \rangle \leqslant 0$. Therefore, $\langle w - w_0 | w' - w_0 \rangle \leqslant 0$ and (IV.3) is proved. □

We can now recover via Theorem IV.3.2 the classic guarantee for the Blackwell strategy in the case where the vector payoffs are bounded with respect to the Euclidean norm.

Theorem IV.5.2. *Let* $\mathrm{T} \geqslant 1$ *an integer and* $\mathrm{M} > 0$. *Assume that* $\|g(x)\|_2 \leqslant \mathrm{M}$ *(for all* $g \in \mathscr{G}$ *and* $x \in \mathscr{X}$*). Then, against any sequence of payoff functions* $(g_t)_{t \geqslant 1}$ *in* $\mathscr{G}$*, the Blackwell strategy guarantees*

$$d_2(\bar{u}_\mathrm{T}, \mathscr{C}) \leqslant \frac{2\mathrm{M}}{\sqrt{\mathrm{T}}},$$

where d_2 *denotes the Euclidean distance.*

Proof. With notation from Theorem IV.5.1, we have $\delta_{h_2} = 1$, and h_2 is 1-strongly convex with respect to $\|\cdot\|_2$ by Proposition I.2.10. According to Theorem IV.5.1, the Blackwell strategy corresponds to the Mirror Descent strategy associated with h_2 and any sequence of parameters $(\eta_t)_{t \geqslant 1}$. We can therefore apply (iii) from Theorem IV.3.2 together with Proposition IV.2.10 and the result follows. □

IV.6. Finite action set

We now consider a variant of the model of Section IV.1, in which the Decision Maker has a finite set actions $\mathscr{I} = \{1, \dots, d\}$ from which he is allowed to choose at random. Let Δ_d be the set of mixed actions, and $\mathscr{G}$ a set of payoff functions $g : \mathscr{I} \to \mathscr{V}^*$. We linearly extend each payoff function $g \in \mathscr{G}$ from $\mathscr{I}$ to Δ_d:

$$g(x) := \mathbb{E}_{i \sim x}[g(i)] = \sum_{i=1}^{d} x^i g(i), \qquad x \in \Delta_d.$$

The play goes as follows. At time $t \geqslant 1$, the Decision Maker

- chooses mixed action $x_t \in \Delta_d$;
- draws pure action $i_t \sim x_t$;

- observes vector payoff $u_t = g_t(i_t)$,

where $(g_t)_{t\geqslant 1}$ is a sequence of payoff vectors chosen by Nature. Denote $(\mathscr{F}_t)_{t\geqslant 1}$ the filtration where $\mathscr{F}_t$ is generated by

$$(z_1, g_1, i_t, \dots, z_{t-1}, g_{t-1}, i_{t-1}, z_t, g_t).$$

A strategy for the Decision Maker is a sequence of maps $\sigma = (\sigma_t)_{t\geqslant 1}$ where $\sigma_t : (\Delta_d \times \mathscr{I} \times \mathscr{V}^*)^{t-1} \to \Delta_d$ so that

$$x_t = \sigma_t(x_1, i_1, u_1, \dots, x_{t-1}, i_{t-1}, u_{t-1}).$$

Concerning Nature, we assume that its choice of payoff function g_t does not depend on i_t, so that $\mathbb{E}\left[g_t(i_t) \mid \mathscr{F}_t\right] = \mathbb{E}_{i\sim x_t}\left[g_t(i)\right] = g_t(x_t)$.

Definition IV.6.1. A closed convex cone $\mathscr{C}$ of the payoff space $\mathscr{V}^*$ is a *$\mathscr{G}$-B-set* if

$$\forall z \in \mathscr{C}^\circ,\ \exists x(z) \in \Delta_d,\ \forall g \in \mathscr{G}, \quad \langle g(x(z)) | z\rangle \leqslant 0.$$

Such an application $x : \mathscr{C}^\circ \to \Delta_d$ is called a *$(\mathscr{G}, \mathscr{C})$-oracle*.

We can now define Mirror Descent strategies similarly as in Section IV.6. Let $\mathscr{C}$ be a closed convex cone of the payoff space $\mathscr{V}^*$ which is assumed to be a $\mathscr{G}$-B-set, $x : \mathscr{C}^\circ \to \Delta_d$ a $(\mathscr{G}, \mathscr{C})$-oracle, $\mathscr{Z}$ a generator of $\mathscr{C}^\circ$, h a regularizer on $\mathscr{Z}$, and $(\eta_t)_{t\geqslant 1}$ a positive and nonincreasing sequence. Then, set $\mathrm{U}_0 = 0$ and for $t \geqslant 1$,

$$\begin{aligned} \text{compute} \quad & z_t = \nabla h^*(\eta_{t-1}\mathrm{U}_{t-1}) \\ \text{compute} \quad & x_t = x(z_t) \\ \text{draw} \quad & i_t \sim x_t \\ \text{observe} \quad & u_t = g_t(i_t) \\ \text{update} \quad & \mathrm{U}_t = \mathrm{U}_{t-1} + u_t. \end{aligned}$$

Theorem IV.6.2. *Let* $\mathrm{T} \geqslant 1$ *an integer and* $\mathrm{K}, \mathrm{M} > 0$.

(i) Against any sequence $(g_t)_{t\geqslant 1}$ of payoff functions in $\mathscr{G}$, the above strategy guarantees

$$\mathbb{E}\left[\mathrm{I}^*_{\mathscr{Z}}(\bar{u}_\mathrm{T})\right] \leqslant \frac{\delta_h}{\mathrm{T}\eta_\mathrm{T}} + \frac{1}{\mathrm{T}}\sum_{i=1}^{d}\frac{1}{\eta_{t-1}}\mathbb{E}\left[\mathrm{D}_{h^*}(\eta_{t-1}\mathrm{U}_t, \eta_{t-1}\mathrm{U}_{t-1})\right],$$

where $\eta_0 = \eta_1$.

(ii) Moreover, if h is K*-strongly convex with respect to a norm* $\|\cdot\|$,

$$\mathbb{E}\left[\mathrm{I}^*_{\mathscr{Z}}(\bar{u}_\mathrm{T})\right] \leqslant \frac{\delta_h}{\mathrm{T}\eta_\mathrm{T}} + \frac{1}{2\mathrm{KT}}\sum_{t=1}^{\mathrm{T}}\eta_{t-1}\mathbb{E}\left[\|u_t\|_*^2\right].$$

(iii) Moreover, if $\|g(i)\|_* \leqslant \mathrm{M}$ *(for all* $g \in \mathscr{G}$ *and* $i \in \mathscr{I}$*), the choice of parameters* $\eta_t = \sqrt{\delta_h \mathrm{K}/\mathrm{M}^2 t}$ *(for* $t \geqslant 1$*) give*

$$\mathbb{E}\left[\mathrm{I}^*_{\mathscr{Z}}(\bar{u}_{\mathrm{T}})\right] \leqslant 2\mathrm{M}\sqrt{\frac{\delta_h}{\mathrm{KT}}}.$$

Let $\delta \in (0,1)$*. We have with probability higher than* $1-\delta$*,*

$$\mathrm{I}^*_{\mathscr{Z}}(\bar{u}_{\mathrm{T}}) \leqslant \frac{\mathrm{M}}{\sqrt{\mathrm{T}}}\left(2\sqrt{\frac{\delta_h}{\mathrm{K}}} + \|\mathscr{Z}\|\sqrt{2\log(1/\delta)}\right).$$

Almost-surely,

$$\limsup_{\mathrm{T}\to+\infty} \mathrm{I}^*_{\mathscr{Z}}(\bar{u}_{\mathrm{T}}) \leqslant 0.$$

Proof. Like in the proof of Theorem IV.3.2, Theorem I.3.1 gives:

$$\mathrm{I}^*_{\mathscr{Z}}(\bar{u}_{\mathrm{T}}) \leqslant \frac{1}{\mathrm{T}}\left(\sum_{t=1}^{\mathrm{T}}\langle u_t|z_t\rangle + \frac{\delta_h}{\eta_{\mathrm{T}}} + \sum_{t=1}^{\mathrm{T}}\frac{1}{\eta_{t-1}}\mathrm{D}_{h^*}(\eta_{t-1}\mathrm{U}_t, \eta_{t-1}\mathrm{U}_{t-1})\right). \tag{IV.4}$$

Consider $\mathrm{X}_t = \langle u_t|z_t\rangle$. Then, $(\mathrm{X}_t)_{t\geqslant 1}$ is a sequence of super-martingale differences with respect to filtration $(\mathscr{F}_t)_{t\geqslant 0}$:

$$\mathbb{E}\left[\langle u_t|z_t\rangle \mid \mathscr{F}_t\right] = \mathbb{E}\left[\langle g_t(i_t)|z_t\rangle \mid \mathscr{F}_t\right] = \langle \mathbb{E}\left[g_t(i_t) \mid \mathscr{F}_t\right]|z_t\rangle = \langle g_t(x_t)|z_t\rangle \leqslant 0,$$

because x is a $(\mathscr{G}, \mathscr{C})$-oracle. Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{\mathrm{T}}\langle u_t|z_t\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{\mathrm{T}}\mathbb{E}\left[\langle u_t|z_t\rangle \mid \mathscr{F}_t\right]\right] \leqslant 0.$$

Injecting this in Equation (IV.4) gives the bound (i):

$$\mathbb{E}\left[\mathrm{I}^*_{\mathscr{Z}}(\bar{u}_{\mathrm{T}})\right] \leqslant \frac{\delta_h}{\eta_{\mathrm{T}}} + \sum_{t=1}^{\mathrm{T}}\frac{1}{\eta_{t-1}}\mathbb{E}\left[\mathrm{D}_{h^*}(\eta_{t-1}\mathrm{U}_t, \eta_{t-1}\mathrm{U}_{t-1})\right].$$

We then deduce the bounds in expectation stated in (ii) and (iii) similarly as in Theorem I.3.1. We now turn to the high probability bound. Let $\delta \in (0,1)$. From Equation (IV.4), we deduce that under the assumptions (iii), we have

$$\mathrm{I}^*_{\mathscr{Z}}(\bar{u}_{\mathrm{T}}) \leqslant 2\mathrm{M}\sqrt{\frac{\delta_h}{\mathrm{KT}}} + \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\mathrm{X}_t.$$

Since we have $|X_t| = |\langle g_t(i_t)|z_t\rangle| \leqslant \|g_t(i_t)\|_* \|z_t\| \leqslant M\|\mathscr{Z}\|$ for all $t \geqslant 1$, Proposition A.0.1 assures that with probability higher than $1-\delta$,

$$\frac{1}{T}\sum_{t=1}^{T} X_t \leqslant M\|\mathscr{Z}\|\sqrt{\frac{2\log(1/\delta)}{T}}$$

and thus

$$I^*_{\mathscr{Z}}(\bar{u}_T) \leqslant \frac{M}{\sqrt{T}}\left(2\sqrt{\frac{\delta_h}{K}} + \|\mathscr{Z}\|\sqrt{2\log(1/\delta)}\right).$$

The almost-sure result follows from a standard Borel–Cantelli argument. □

IV.7. Online combinatorial optimization

We consider the online combinatorial optimization problem with full information feedback. It is a regret minimization problem in which the actions and the payoffs have a particular structure. Numerous papers were written on the topic, including [GW98, KW01, GLS01, TW03, KV05, WK08, HW09, HKW10]. A minimax optimal strategy was given in [KWK10]. We give below an alternative construction of such a strategy.

Let $d, m \geqslant 1$ be integers. Let $\mathcal{I} = \{1, \dots, d\}$ be a finite set. The set of pure actions of the Decision Maker is a set P which contains subsets of $\mathcal{I}$ of cardinality m. $\Delta(P)$ is the set of mixed actions over P. The play goes as follows. At time $t \geqslant 1$, the Decision Maker

- chooses mixed action $x_t \in \Delta(P)$;
- draws pure action $p_t \sim x_t$;
- observes payoff vector $v_t \in \mathbb{R}^d$;
- gets payoff $\sum_{i\in p_t} v_t^i$.

As usual, we assume that the choice by Nature of payoff vector $v_t \in \mathbb{R}^d$ does not depend on pure action p_t. The quantity to minimize is the following regret:

$$\max_{p\in P}\sum_{t=1}^{T}\sum_{i\in p} v_t^i - \sum_{t=1}^{T}\sum_{i\in p_t} v_t^i.$$

This problem can be seen as a basic regret minimization problem in the experts setting as in Section II.1 with pure action set P, and payoff vectors $(\sum_{i\in p} v^i)_{p\in P}$ which belong to $[-m, m]^P$ as soon as we assume $v \in [-1, 1]^d$. The Exponential Weights Algorithm would then guarantee (Theorem II.3.1) a regret bound of order

$m\sqrt{\mathrm{T}\log|\mathrm{P}|}$. However, our goal is to take advantage of the structure of the problem and to construct a strategy which guarantees a significantly tighter regret bound, of order $m\sqrt{\mathrm{T}\log(d/m)}$, which is known to be minimax optimal [KWK10]. To do so, we reduce this problem to a well-chosen approachability problem.

Let A be the $d \times |\mathrm{P}|$ matrix defined by $\mathrm{A} = (\mathbb{1}_{\{i\in p\}})_{\substack{i\in\mathcal{I}\\ p\in\mathrm{P}}}$, and for each $p \in \mathrm{P}$, let $e_p = (\mathbb{1}_{\{i\in p\}})_{i\in\mathcal{I}} \in \mathbb{R}^d$. The set of payoff functions we choose is the following:

$$\mathcal{G} = \left\{ g_v : p \in \mathrm{P} \longmapsto v - \frac{\langle v|e_p\rangle}{m}\mathbf{1} \in \mathbb{R}^d \right\}_{v\in\mathbb{R}^d},$$

where $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^d$. g_v is therefore the payoff function which corresponds to payoff vector v. For all $v \in \mathbb{R}^d$, the linear extension of g_v is given by

$$g_v(x) = v - \frac{\langle v|\mathrm{A}x\rangle}{m}\mathbf{1}, \quad x \in \Delta(\mathrm{P}).$$

We now choose the generator: let $\mathcal{Z} = \mathrm{A}(\Delta(\mathrm{P}))$ be the image of the simplex $\Delta(\mathrm{P})$ via A seen as a linear map from $\mathbb{R}^{\mathrm{P}}$ to $\mathbb{R}^d$. Its properties are gathered in the following proposition. In particular, property (v) demonstrates that this choice of $\mathcal{Z}$ makes $\mathrm{I}^*_{\mathcal{Z}}(\bar{u}_{\mathrm{T}})$ equal to the above defined regret.

Proposition IV.7.1. *(i) $\mathcal{Z}$ is the convex hull of the points e_p ($p \in \mathrm{P}$).*

(ii) $\mathcal{Z} \subset m\Delta_d$.

(iii) $\|\mathcal{Z}\|_1 = m$.

(iv) $\mathcal{Z}$ is a generator of $\mathcal{Z}^{\circ\circ} = \mathrm{A}(\Delta(\mathrm{P}))^{\circ\circ}$.

(v) Let $(p_t)_{t\geqslant 1}$ be a sequence of pure actions played against payoff vectors $(v_t)_{t\geqslant 1}$ and denote $u_t = g_{v_t}(p_t)$ for all $t \geqslant 1$. Then,

$$\mathrm{I}^*_{\mathcal{Z}}(\bar{u}_{\mathrm{T}}) = \frac{1}{\mathrm{T}}\left(\max_{p\in\mathrm{P}} \sum_{t=1}^{\mathrm{T}}\sum_{i\in p} v_t^i - \sum_{t=1}^{\mathrm{T}}\sum_{i\in p_t} v_t^i\right).$$

Proof. By definition, $\mathcal{Z}$ is the image of simplex $\Delta(\mathrm{P})$ via linear map A. It is therefore the convex hull of the image by A of the extreme points of $\Delta(\mathrm{P})$. And for $p_0 \in \mathrm{P}$, $\mathrm{A}(\mathbb{1}_{\{p=p_0\}})_{p\in\mathrm{P}} = e_p$. Hence (i). Each point e_p clearly belongs to $m\Delta_d$, and (ii) is true by convexity of $m\Delta_d$. For each element $z \in m\Delta_d$, we have $\|z\|_1 = m$, which implies (iii). $\mathcal{Z}$ is a nonempty convex compact set thanks to (i); Proposition IV.2.8 gives (iv).

As for the relation (v), we denote A^* the transpose of A and write

$$\begin{aligned}
\max_{p\in P}\sum_{t=1}^{T}\sum_{i\in p} v_t^i - \sum_{t=1}^{T}\sum_{i\in p_t} v_t^i &= \max_{p\in P}\sum_{t=1}^{T}\left((A^*v_t)^p - (A^*v_t)^{p_t}\right)\\
&= \max_{x\in\Delta(P)}\sum_{t=1}^{T}\left(\langle A^*v_t|x\rangle - \left\langle A^*v_t\middle|\left(\mathbb{1}_{\{p=p_t\}}\right)_{p\in P}\right\rangle\right)\\
&= \max_{x\in\Delta(P)}\sum_{t=1}^{T}\left(\langle v_t|Ax\rangle - \left\langle v_t\middle|A\left(\mathbb{1}_{\{p=p_t\}}\right)_{p\in P}\right\rangle\right)\\
&= \max_{z\in A(\Delta(P))}\sum_{t=1}^{T}\left(\langle v_t|z\rangle - \langle v_t|e_{p_t}\rangle\right)\\
&= \max_{z\in\mathcal{Z}}\sum_{t=1}^{T}\left\langle v_t - \frac{\langle v_t|e_{p_t}\rangle}{m}\mathbf{1}\middle|z\right\rangle\\
&= \max_{z\in\mathcal{Z}}\sum_{t=1}^{T}\langle g_{v_t}(p_t)|z\rangle\\
&= T\cdot I^*_{\mathcal{Z}}(\bar{u}_T),
\end{aligned}$$

where in the fifth line, we used the fact that for all $z \in \mathcal{Z}$, $\langle \mathbf{1}|z\rangle = m$, which is a consequence of (ii). □

Proposition IV.7.2. $A(\Delta(P))^\circ$ *is a* $\mathcal{G}$*-B-set.*

Proof. Since $\mathcal{Z}$ is a generator of $A(\Delta(P))^{\circ\circ}$, one can check that the condition that defines a B-set only needs to be verified for $z \in \mathcal{Z}$. Let $z \in \mathcal{Z}$. By definition of $\mathcal{Z}$, there exists $x \in \Delta(P)$ such that $z = Ax$. Then for $g \in \mathcal{G}$, there exists $v \in \mathbb{R}^d$ such that $g = g_v$ and

$$\langle g_v(x)|z\rangle = \left\langle v - \frac{\langle v|Ax\rangle}{m}\mathbf{1}\middle|Ax\right\rangle = \langle v|Ax\rangle - \langle v|Ax\rangle = 0,$$

which proves the result. □

As a consequence of Proposition IV.7.1, a point $z \in \mathcal{Z}$ only has nonnegative components. We can therefore define

$$h(z) = \begin{cases} \displaystyle\sum_{i=1}^{d}\frac{z^i}{m}\log\frac{z^i}{m} & \text{for } z\in\mathcal{Z}\\ +\infty & \text{otherwise.}\end{cases}$$

Proposition IV.7.3. *(i) h is a regularizer on $\mathcal{Z}$;*
(ii) $\delta_h \leqslant \log(d/m)$;
(iii) h is $1/m^2$-strongly convex with respect to $\|\cdot\|_1$.

Proof. For $z \in \mathcal{Z} \subset m\Delta_d$, we can write $h(z) = h_{\mathrm{ent}}(z/m) < +\infty$. The 1-strong convexity of h_{ent} with respect to $\|\cdot\|_1$ implies the $1/m^2$-strong convexity of h with respect to $\|\cdot\|_1$ and (iii) is proved. In particular, h is strictly convex. Besides, the domain of h is $\mathcal{Z}$ by definition and (i) is proved. As for (ii), h being convex, its maximum is attained at one of the extreme points e_p ($p \in \mathrm{P}$) of $\mathcal{Z}$:

$$\max_{z\in\mathcal{Z}} h(z) = \max_{p\in\mathrm{P}} h(e_p) = \max_{p\in\mathrm{P}} \sum_{i\in p} \frac{1}{m}\log\frac{1}{m} = -\log m.$$

As for the minimum,

$$\min_{z\in\mathcal{Z}} h(z) \geqslant \min_{z\in m\Delta_d} \sum_{i=1}^{d} \frac{z^i}{m}\log\frac{z^i}{m} = \min_{z\in\Delta_d} \sum_{i=1}^{d} z^i \log z^i = -\log d.$$

Therefore, $\delta_h \leqslant -\log m + \log d = \log(d/m)$. □

We can now consider the Mirror Descent strategy associated with regularizer h, a $(\mathcal{G},\mathcal{C})$-oracle x, and a positive nonincreasing sequence of parameters $(\eta_t)_{t\geqslant 1}$. Set $\mathrm{U}_0 = 0$ and for $t \geqslant 1$,

$$\begin{aligned}
\text{compute} \quad & z_t = \nabla h^*(\eta_{t-1}\mathrm{U}_{t-1}) \\
\text{choose} \quad & x_t = x(z_t) \\
\text{draw} \quad & p_t \sim x_t \\
\text{observe} \quad & u_t = g_{v_t}(p_t) = v_t - \frac{\langle v_t | \mathrm{A}e_{p_t}\rangle}{m}\mathbf{1} \\
\text{update} \quad & \mathrm{U}_t = \mathrm{U}_{t-1} + u_t.
\end{aligned}$$

Theorem IV.7.4. *Against any sequence of payoff vectors $(v_t)_{t\geqslant 1}$ in $[-1,1]^d$, the above strategy with parameter $\eta_t = \sqrt{\delta_h/4m^2t}$ (for $t \geqslant 1$) guarantees*

$$\mathbb{E}\left[\max_{p\in\mathrm{P}} \sum_{t=1}^{\mathrm{T}}\sum_{i\in p} v_t^i - \sum_{t=1}^{\mathrm{T}}\sum_{i\in p_t} v_t^i\right] \leqslant 4m\sqrt{\mathrm{T}\log(d/m)}.$$

For $\delta \in (0,1)$, we have with probability higher than $1-\delta$,

$$\max_{p\in\mathrm{P}} \sum_{t=1}^{\mathrm{T}}\sum_{i\in p} v_t^i - \sum_{t=1}^{\mathrm{T}}\sum_{i\in p_t} v_t^i \leqslant 2m\sqrt{\mathrm{T}}\left(2\sqrt{\log(d/m)} + \sqrt{2\log(1/\delta)}\right).$$

Almost-surely,

$$\limsup_{T\to+\infty} \frac{1}{T}\left(\max_{p\in P}\sum_{t=1}^{T}\sum_{i\in p} v_t^i - \sum_{t=1}^{T}\sum_{i\in p_t} v_t^i\right) \leqslant 0.$$

Proof. For all $v \in [-1,1]^d$ and $p \in P$,

$$\|g_v(p)\|_\infty = \left\| v - \frac{\langle v|Ae_p\rangle}{m}\mathbf{1}\right\|_\infty \leqslant \|v\|_\infty + \frac{\|\mathbf{1}\|_\infty}{m}\sum_{i\in p}|v^i| \leqslant 2.$$

The result then follows from Theorem IV.6.2 applied with $M = 2$, $K = 1/m^2$, the properties of the regularizer h given by Proposition IV.7.3, and the relation (v) from Proposition IV.7.1. □

Let us discuss the computational aspect of the above strategy compared to the Exponential Weights Algorithm played against payoff vectors $(\sum_{i\in p} v^i)_{p\in P}$. The Exponential Weights Algorithm has the advantage of having an explicit expression which can easily computed:

$$x_t^p = \frac{\exp\left(-\eta_{t-1}\sum_{s=1}^{t-1}\sum_{i\in p} v_s^i\right)}{\sum_{p'\in P}\exp\left(-\eta_{t-1}\sum_{s=1}^{t-1}\sum_{i\in p'} v_s^i\right)}, \qquad p \in P.$$

However, the drawback is that the above expression has to be computed for each of the $|P|$ components ($|P|$ being possibly much larger than d due to its combinatorial nature). In addition to that, the sum of the payoff vectors $(\sum_{i\in p} v_t^i)_{p\in P}$ has to be kept track of, which may be an issue regarding memory. The strategy constructed above has the advantage of working with vector payoffs u_t which have d components only. However, the difficulty lies in the computation of ∇h^* which consists of a convex program on the set $\mathcal{Z}$ which has $|P|$ vertices. Overall, it is unclear which strategy is computationally more efficient.

IV.8. Internal and swap regret

The notion of *internal regret* was introduced by [FV97]. It is an alternative quantity to minimize in the experts setting from Section II.1. [FV97] first established the existence of strategies which guarantees that the average internal regret is asymptotically nonpositive (see also [FL95, FL99, HMC00, HMC01, SL05]). [BM05] introduced the swap regret, which generalizes both the internal and the basic regret. The optimal bound on the swap regret is known since [BM05, SL05]. Later, [Per15] proposed an approachability-based optimal strategy. We present below the construction

of a similar strategy using the tools introduced in the previous sections. The internal regret is mentioned at the end of the section as a special case.

The *swap regret* is a variant of the regret minimization problem in the experts setting (Section II.1). The set of pure actions of the Decision Maker is $\mathcal{I} = \{1, \dots, d\}$. The play goes as follows. At time $t \geqslant 1$, the Decision Maker

- chooses mixed action $x_t \in \Delta_d$;
- draws pure action $i_t \sim x_t$;
- observes payoff vector $v_t \in \mathbb{R}^d$.

Let Φ be a nonempty subset of $\mathcal{I}^{\mathcal{I}}$. The quantity to minimize is the Φ-regret defined by:

$$\max_{\phi \in \Phi} \sum_{t=1}^{T} v_t^{\phi(i_t)} - \sum_{t=1}^{T} v_t^{i_t},$$

and can be interpreted as follows. For a given map $\phi \in \Phi$, $\sum_{t=1}^{T} v_t^{\phi(i_t)}$ is the cumulative payoff that the Decision Maker would have obtained if he had played pure action $\phi(i)$ each time he has actually played i (for all $i \in \mathcal{I}$). The Φ-regret therefore compares the actual cumulative payoff of the Decision Maker with the best such quantity (for $\phi \in \Phi$) in hindsight. The goal is to construct a strategy which guarantees on the Φ-regret a bound of order $\sqrt{T \log |\Phi|}$. To do so, we reduce this problem to a well-chosen approachability problem.

Let the payoff space be $\mathcal{V}^* = \mathbb{R}^{\Phi}$ and the target be $\mathbb{R}^{\Phi}_-$. We choose the following set of payoff functions:

$$\mathcal{G} = \left\{ g_v : i \in \mathcal{I} \longmapsto \left(v^{\phi(i)} - v^i\right)_{\phi \in \Phi} \in \mathbb{R}^{\Phi} \right\}_{v \in \mathbb{R}^d},$$

where g_v (for $v \in \mathbb{R}^d$) is the payoff function associated with payoff vector v. The linear extension of each payoff function is

$$g_v(x) = \left(\sum_{i \in \mathcal{I}} x^i (v^{\phi(i)} - v^i) \right)_{\phi \in \Phi}, \quad x \in \Delta_d,\ v \in \mathbb{R}^d.$$

Proposition IV.8.1. *$\mathbb{R}^{\Phi}_-$ is a $\mathcal{G}$-B-set.*

Proof. Let $z = (z^{\phi})_{\phi \in \Phi} \in (\mathbb{R}^{\Phi}_-)^\circ = \mathbb{R}^{\Phi}_+$. Let us prove that there exists $x \in \Delta(\mathcal{I})$ such that for all payoff function $g \in \mathcal{G}$, $\langle g(x) | z \rangle \leqslant 0$. First, the property is trivially true if $z = 0$. We assume from now on that $z \neq 0$.

Denote

$$\tilde{z}^{ij} = \sum_{\substack{\phi \in \Phi \\ \phi(i) = j}} z^{\phi}, \quad i, j \in \mathcal{I}$$

and let us first prove that there exists $x \in \Delta(\mathcal{I})$ such that:

$$\sum_{i\in\mathcal{I}} x^i \tilde{z}^{ij} = x^j \sum_{i\in\mathcal{I}} \tilde{z}^{ji}, \quad j \in \mathcal{I}. \tag{IV.5}$$

Notice that for all $i \in \mathcal{I}$ we have

$$\sum_{j\in\mathcal{I}} \tilde{z}^{ij} = \sum_{j\in\mathcal{I}} \sum_{\substack{\phi\in\Phi \\ \phi(i)=j}} z^\phi = \sum_{\phi\in\Phi} z^\phi = \|z\|_1 .$$

z being nonzero, the above quantity is also nonzero and the $d\times d$ matrix $(\tilde{z}^{ij}/\|z\|_1)_{i,j\in\mathcal{I}}$ is stochastic and therefore has an invariant measure $x \in \Delta(\mathcal{I})$:

$$\sum_{i\in\mathcal{I}} x^i \frac{\tilde{z}^{ij}}{\|z\|_1} = x^j, \quad j \in \mathcal{I}.$$

Multiplying on both sides by $\|z\|_1$, we get Equation (IV.5):

$$\sum_{i\in\mathcal{I}} x^i \tilde{z}^{ij} = x^j \|z\|_1 = x^j \sum_{i\in\mathcal{I}} \sum_{\substack{\phi\in\Phi \\ \phi(j)=i}} z^\phi = x^j \sum_{i\in\mathcal{I}} \tilde{z}^{ji}, \quad j \in \mathcal{J}.$$

Let $g \in \mathcal{G}$. By definition of $\mathcal{G}$, there exists a payoff vector $v \in \mathbb{R}^d$ such that

$$g(x) = \left(\sum_{i\in\mathcal{I}} x^i (v^{\phi(i)} - v^i) \right)_{\phi\in\Phi}, \quad x \in \Delta_d.$$

We now compute $\langle g(x)|z\rangle$:

$$\begin{aligned}
\langle g(x)|z\rangle &= \sum_{\phi\in\Phi} z^\phi \left(\sum_{i\in\mathcal{I}} x^i (v^{\phi(i)} - v^i) \right) = \sum_{i,j\in\mathcal{I}} x^i (v^j - v^i) \sum_{\substack{\phi\in\Phi \\ \phi(i)=j}} z^\phi \\
&= \sum_{i,j\in\mathcal{I}} x^i (v^j - v^i) \tilde{z}^{ij} = \sum_{j\in\mathcal{I}} v^j \sum_{i\in\mathcal{I}} x^i \tilde{z}^{ij} - \sum_{i,j\in\mathcal{I}} x^i v^i \tilde{z}^{ij} \\
&= \sum_{j\in\mathcal{I}} v^j x^j \sum_{i\in\mathcal{I}} \tilde{z}^{ji} - \sum_{i,j\in\mathcal{I}} x^i v^i \tilde{z}^{ij} = 0,
\end{aligned}$$

where we used Equation (IV.5) for the fifth equality. In particular, $\langle g(x)|z\rangle \leqslant 0$ and $\mathbb{R}^\Phi_-$ is indeed a $\mathcal{G}$-B-set. □

As for the generator, we choose $\mathcal{Z} = \Delta(\Phi)$ which is a generator of $(\mathbb{R}_-^\Phi)^\circ$ thanks to Proposition IV.2.8. Then the support function of $\Delta(\Phi)$ evaluated at the average payoff is equal to the (average) Φ-regret:

$$\begin{aligned}\mathrm{I}^*_{\Delta(\Phi)}(\bar{u}_\mathrm{T}) &= \frac{1}{\mathrm{T}}\mathrm{I}^*_{\Delta(\Phi)}\left(\sum_{t=1}^{\mathrm{T}} g_{v_t}(i_t)\right) = \frac{1}{\mathrm{T}}\max_{z\in\Delta(\Phi)}\left\langle \sum_{t=1}^{\mathrm{T}}\left(v_t^{\phi(i_t)} - v_t^{i_t}\right)_{\phi\in\Phi}\middle| z\right\rangle \\ &= \frac{1}{\mathrm{T}}\max_{\phi\in\Phi}\sum_{t=1}^{\mathrm{T}}\left(v_t^{\phi(i_t)} - v_t^{i_t}\right) = \frac{1}{\mathrm{T}}\left(\max_{\phi\in\Phi}\sum_{t=1}^{\mathrm{T}} v_t^{\phi(i_t)} - \sum_{t=1}^{\mathrm{T}} v_t^{i_t}\right).\end{aligned}$$

On the simplex $\Delta(\Phi)$, we choose the entropic regularizer presented in Section I.2.3:

$$h_{\mathrm{ent}}(z) = \begin{cases} \sum_{\phi\in\Phi} z^\phi \log z^\phi & \text{if } z \in \Delta(\Phi) \\ +\infty & \text{otherwise.}\end{cases}$$

Then, the strategy associated with regularizer h_{ent}, a $(\mathcal{G}, \mathbb{R}_-^\Phi)$-oracle x and a sequence of positive and nonincreasing parameters $(\eta_t)_{t\geqslant 1}$ is the following. Set $\mathrm{U}_0 = 0$ and for $t \geqslant 1$,

$$\begin{aligned}\text{compute} \quad & z_t^\phi = \frac{\exp\left(\eta_{t-1}\mathrm{U}_{t-1}^\phi\right)}{\sum_{\phi'\in\Phi}\exp\left(\eta_{t-1}\mathrm{U}_{t-1}^{\phi'}\right)}, \quad \phi\in\Phi \\ \text{choose} \quad & x_t = x(z_t) \\ \text{draw} \quad & i_t \sim x_t \\ \text{observe} \quad & u_t = g_{v_t}(i_t) = \left(v_t^{\phi(i_t)} - v_t^{i_t}\right)_{\phi\in\Phi} \\ \text{update} \quad & \mathrm{U}_t = \mathrm{U}_{t-1} + u_t.\end{aligned}$$

This strategy is computationally efficient. Indeed, the expression of z_t is explicit and straightforward. Then, the computation of mixed action $x_t = x(z_t)$ via oracle x consists, as shown in the proof of Proposition IV.8.1, in finding an invariant measure of a $d \times d$ stochastic matrix, which can be done efficiently.

Theorem IV.8.2. *Against payoff vectors $(v_t)_{t\geqslant 1}$ in $[-1,1]^d$, the above strategy with parameters $\eta_t = \sqrt{\log|\Phi|\,/4t}$ (for $t \geqslant 1$) guarantees*

$$\mathbb{E}\left[\max_{\phi\in\Phi}\sum_{t=1}^{\mathrm{T}} v_t^{\phi(i_t)} - \sum_{t=1}^{\mathrm{T}} v_t^{i_t}\right] \leqslant 4\sqrt{\mathrm{T}\log|\Phi|}.$$

Let $\delta \in (0,1)$. *With probability higher than* $1-\delta$, *we have*

$$\frac{1}{T}\left(\max_{\phi\in\Phi}\sum_{t=1}^{T} v_t^{\phi(i_t)} - \sum_{t=1}^{T} v_t^{i_t}\right) \leqslant \frac{1}{\sqrt{T}}\left(4\sqrt{\log|\Phi|} + 2\sqrt{2\log(1/\delta)}\right).$$

Almost-surely,

$$\limsup_{T\to+\infty} \frac{1}{T}\left(\max_{\phi\in\Phi}\sum_{t=1}^{T} v_t^{\phi(i_t)} - \sum_{t=1}^{T} v_t^{i_t}\right) \leqslant 0.$$

Proof. For every payoff vector $v \in [-1,1]^d$ and pure action $i \in \mathcal{I}$, we have

$$\|g_v(i)\|_\infty = \left\|(v^{\phi(i)} - v^i)_{\phi\in\Phi}\right\|_\infty \leqslant 2.$$

The result then follows from Theorem IV.6.2 applied with $M = 2$, $K = 1$ and the properties of regularizer h_{ent} given by Proposition I.2.9. □

An important special case is when Φ is the set of all transpositions of $\mathcal{I}$, in other words, the set of maps $\phi : \mathcal{I} \to \mathcal{I}$ such that there exists $i \neq j$ in $\mathcal{I}$ such that

$$\phi(i) = j, \quad \phi(j) = i, \quad \text{and} \quad \phi(k) = k \text{ for all } k \notin \{i, j\}.$$

The Φ-regret is then called the *internal regret* and can be written

$$\max_{i,j\in\mathcal{I}} \sum_{t=1}^{T} \mathbb{1}_{\{i_t=i\}}(v_t^j - v_t^i).$$

Since $|\Phi| = d(d-1)$ in this case, Theorem IV.8.2 assures that the corresponding strategy guarantees a bound on the internal regret of order $\sqrt{T\log d}$.

SECOND PART

CHAPTER V
SPARSE REGRET MINIMIZATION

This chapter is extracted from the paper *Gains and losses are fundamentally different in regret minimization: The sparse case*, in collaboration with Vianney Perchet, to appear in *Journal of Machine Learning Research*.

Abstract

We demonstrate that, in the classical non-stochastic regret minimization problem with d decisions, gains and losses to be respectively maximized or minimized are fundamentally different. Indeed, by considering the additional sparsity assumption (at each stage, at most s decisions incur a nonzero outcome), we derive optimal regret bounds of different orders. Specifically, with gains, we obtain an optimal regret guarantee after T stages of order $\sqrt{T \log s}$, so the classical dependency in the dimension is replaced by the sparsity size. With losses, we provide matching upper and lower bounds of order $\sqrt{Ts \log(d)/d}$, which is decreasing in d. Eventually, we also study the bandit setting, and obtain an upper bound of order $\sqrt{Ts \log(d/s)}$ when outcomes are losses. This bound is proven to be optimal up to the logarithmic factor $\sqrt{\log(d/s)}$.

V.1. Introduction

We consider the classical problem of regret minimization [Han57] that has been well developed during the last decade [CBL06, RT09, Bub11, SS11, Haz12, BCB12]. We recall that in this sequential decision problem, a decision maker (or agent, player, algorithm, strategy, policy, depending on the context) chooses at each stage a decision in a finite set (that we write as $[d] := \{1, \dots, d\}$) and obtains as an *outcome* a real number in $[0, 1]$. We specifically chose the word *outcome*, as opposed to *gain* or *loss*, as our results show that there exists a fundamental discrepancy between these two concepts.

The criterion used to evaluate the policy of the decision maker is the *regret*, i.e. the difference between the cumulative performance of the best stationary policy (that

always picks a given action $i \in [d]$) and the cumulative performance of the policy of the decision maker.

We focus here on the *non-stochastic* framework, where no assumption (apart from boundedness) is made on the sequence of possible outcomes. In particular, they are not i.i.d. and we can even assume, as usual, that they depend on the past choices of the decision maker. This broad setup, sometimes referred to as *individual sequences* (since a policy must be good against *any* sequence of possible outcomes) incorporates prediction with expert advice [CBL06], data with time-evolving laws, etc. Perhaps the most fundamental results in this setup are the upper bound of order $\sqrt{T \log d}$ achieved by the Exponential Weight Algorithm [LW94, Vov90, CB97, ACBG02] and the asymptotic lower bound of the same order [CBFH+97]. This general bound is the same whether outcomes are gains in $[0, 1]$ (in which case, the objective is to maximize the cumulative sum of gains) or losses in $[0, 1]$ (where the decision maker aims at minimizing the cumulative sum). Indeed, a loss ℓ can easily be turned into gain g by defining $g := 1 - \ell$, the regret being invariant under this transformation.

This idea does not apply anymore with structural assumption. For instance, consider the framework where the outcomes are limited to *s-sparse vectors*, i.e. vectors that have at most s nonzero coordinates. The coordinates which are nonzero may change arbitrarily over time. In this framework, the aforementioned transformation does not preserve the sparsity assumption. Indeed, if $(\ell_1, \dots, \ell_d)$ is a s-sparse loss vector, the corresponding gain vector $(1 - \ell_1, \dots, 1 - \ell_d)$ may even have full support. Consequently, results for loss vectors do not apply directly to sparse gains, and vice versa. It turns out that both setups are fundamentally different.

The sparsity assumption is actually quite natural in learning and have also received some attention in online learning [Ger13, CM12, AYPS12, DKC13]. In the case of gains, it reflects the fact that the problem has some hidden structure and that many options are irrelevant. For instance, in the canonical click-through-rate example, a website displays an ad and gets rewarded if the user clicks on it; we can safely assume that there are only a small number of ads on which a user would click.

The sparse scenario can also be seen through the scope of prediction with experts. Given a finite set of expert, we call the *winner of a stage* the expert with the highest revenue (or the smallest loss); ties are broken arbitrarily. And the objective would be to win as many stages as possible. The s-sparse setting would represent the case where s experts are designated as winners (or, non-loser) at each stage.

In the case of losses, the sparsity assumption is motivated by situations where rare failures might happen at each stage, and the decision maker wants to avoid them. For instance, in network routing problems, it could be assumed that only a small number of paths would lose packets as a result of a single, rare, server failure. Or a learner could have access to a finite number of classification algorithms that perform ideally most of the time; unfortunately, some of them makes mistakes on some examples and

the learner would like to prevent that. The general setup is therefore a number of algorithms/experts/actions that mostly perform well (i.e. find the correct path, classify correctly, optimize correctly some target function, etc.); however, at each time instance, there are rare mistakes/accidents and the objective would be to find the action/algorithm that has the smallest number (or probability in the stochastic case) of failures.

V.1.1. Summary of results

We investigate regret minimization scenarios both when outcomes are gains on the one hand, and losses on the other hand. We recall that our objectives are to prove that they are fundamentally different by exhibiting rates of convergence of different order.

When outcomes are gains, we construct an algorithm based on the Online Mirror Descent family [SS07, SS11, Bub11]. By choosing a regularizer based on the ℓ^p norm, and then tuning the parameter p as a function of s, we get in Theorem V.2.2 a regret bound of order $\sqrt{T \log s}$, which has the interesting property of being independent of the number of decisions d. This bound is trivially optimal, up to the constant.

If outcomes are losses instead of gains, although the previous analysis remains valid, a much better bound can be obtained. We build upon a regret bound for the Exponential Weight Algorithm [LW94, FS97] and we manage to get in Theorem V.3.1 a regret bound of order $\sqrt{\frac{Ts \log d}{d}}$, which is *decreasing* in d, for a given s. A nontrivial matching lower bound is established in Theorem V.3.3.

Both of these algorithms need to be tuned as a function of s. In Theorem V.4.1 and Theorem V.4.2, we construct algorithms which essentially achieve the same regret bounds without prior knowledge of s, by adapting over time to the sparsity level of past outcome vectors, using an adapted version of the doubling trick.

Finally, we investigate the bandit setting, where the only feedback available to the decision maker is the outcome of his decisions (and, not the outcome of all possible decisions). In the case of losses we obtain in Theorem V.5.1 an upper bound of order $\sqrt{Ts \log(d/s)}$, using the Greedy Online Mirror Descent family of algorithms [AB09, ABL13, Bub11]. This bound is proven to be optimal up to a logarithmic factor, as Theorem V.5.3 establishes a lower bound of order $\sqrt{Ts}$.

The rates of convergence achieved by our algorithms are summarized in Figure V.1.

V.1.2. General model and notation

We recall the classical non-stochastic regret minimization problem. At each time instance $t \geqslant 1$, the decision maker chooses a decision d_t in the finite set

	Full information		Bandit	
	Gains	Losses	Gains	Losses
Upper bound	$\sqrt{\mathrm{T}\log s}$	$\sqrt{\mathrm{T}s\frac{\log d}{d}}$	$\sqrt{\mathrm{T}d}$	$\sqrt{\mathrm{T}s\log\frac{d}{s}}$
Lower bound			$\sqrt{\mathrm{T}s}$	$\sqrt{\mathrm{T}s}$

Figure V.1. — Summary of upper and lower bounds.

$[d] = \{1, \dots, d\}$, possibly at random, according to $x_t \in \Delta_d$, where

$$\Delta_d = \left\{ x = (x^{(1)}, \dots, x^{(d)}) \in \mathbb{R}_+^d \,\middle|\, \sum_{i=1}^{d} x^{(i)} = 1 \right\}$$

is the the set of probability distributions over $[d]$. Nature then reveals an outcome vector $\omega_t \in [0,1]^d$ and the decision maker receives $\omega_t^{(d_t)} \in [0,1]$. As outcomes are bounded, we can easily replace $\omega_t^{(d_t)}$ by its expectation that we denote by $\langle \omega_t, x_t \rangle$. Indeed, Hoeffding–Azuma concentration inequality will imply that all the results we will state in expectation hold with high probability.

Given a time horizon $\mathrm{T} \geqslant 1$, the objective of the decision maker is to minimize his regret, whose definition depends on whether outcomes are *gains* or *losses*. In the case of gains (resp. losses), the notation ω_t is then changed to g_t (resp. ℓ_t) and the regret is:

$$\mathrm{R_T} = \max_{i\in[d]} \sum_{t=1}^{\mathrm{T}} g_t^{(i)} - \sum_{t=1}^{\mathrm{T}} \langle g_t, x_t \rangle \qquad \left(\text{resp. } \mathrm{R_T} = \sum_{t=1}^{\mathrm{T}} \langle \ell_t, x_t \rangle - \min_{i\in[d]} \sum_{t=1}^{\mathrm{T}} \ell_t^{(i)} \right).$$

In both cases, the well-known Exponential Weight Algorithm guarantees a bound on the regret of order $\sqrt{\mathrm{T}\log d}$. Moreover, this bound cannot be improved in general as it matches a lower bound.

We shall consider an additional structural assumption on the outcomes, namely that ω_t is s-sparse in the sense that $\|\omega_t\|_0 \leqslant s$, i.e. the number of nonzero components of ω_t is less than s, where s is a fixed known parameter. The set of components which are nonzero is not fixed nor known, and may change arbitrarily over time.

We aim at proving that it is then possible to drastically improve the previously mentioned guarantee of order $\sqrt{\mathrm{T}\log d}$ and that losses and gains are two fundamentally different settings with minimax regrets of different orders.

V.2. When outcomes are gains to be maximized

V.2.1. Online Mirror Descent algorithms

We quickly present the general Online Mirror Descent algorithm [SS11, Bub11, BCB12, KM14] and state the regret bound it incurs; it will be used as a key element in Theorem V.2.2.

A convex function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is called a *regularizer* on Δ_d if h is strictly convex and continuous on its domain Δ_d, and $h(x) = +\infty$ outside Δ_d. Denote $\delta_h = \max_{\Delta_d} h - \min_{\Delta_d} h$ and $h^* : \mathbb{R}^d \to \mathbb{R}^d$ the Legendre–Fenchel transform of h:

$$h^*(y) = \sup_{x \in \mathbb{R}^d} \{\langle y, x\rangle - h(x)\}, \quad y \in \mathbb{R}^d,$$

which is differentiable since h is strictly convex. For all $y \in \mathbb{R}^d$, it holds that $\nabla h^*(y) \in \Delta_d$.

Let $\eta \in \mathbb{R}$ be a parameter to be tuned. The Online Mirror Descent Algorithm associated with the regularizer h and parameter η is defined by:

$$x_t = \nabla h^* \left(\eta \sum_{k=1}^{t-1} \omega_k \right), \quad t \geqslant 1,$$

where $\omega_t \in [0,1]^d$ denote the vector of outcomes and x_t the probability distribution chosen at stage t. The specific choice $h(x) = \sum_{i=1}^d x^{(i)} \log x^{(i)}$ for $x = (x^{(1)}, \dots, x^{(d)}) \in \Delta_d$ (and $h(x) = +\infty$ otherwise) gives the celebrated Exponential Weight Algorithm, which can be written explicitly, component by component:

$$x_t^{(i)} = \frac{\exp\left(\eta \sum_{k=1}^{t-1} \omega_k^{(i)}\right)}{\sum_{j=1}^d \exp\left(\eta \sum_{k=1}^{t-1} \omega_k^{(j)}\right)}, \quad t \geqslant 1,\ i \in [d].$$

The following general regret guarantee for strongly convex regularizers is expressed in terms of the dual norm $\|\cdot\|_*$ of $\|\cdot\|$.

Theorem V.2.1 ([SS11] Th. 2.21; [BCB12] Th. 5.6; [KM14] Th. 5.1). *Let* $K > 0$ *and assume* h *to be* K*-strongly convex with respect to a norm* $\|\cdot\|$*. Then, for any sequence of outcome vectors* $(\omega_t)_{t \geqslant 1}$ *in* $\mathbb{R}^d$*, the Online Mirror Descent strategy associated with* h *and* η *(with* $\eta > 0$ *in cases of gains and* $\eta < 0$ *in cases of losses) guarantees, for* $T \geqslant 1$*, the following regret bound:*

$$R_T \leqslant \frac{\delta_h}{|\eta|} + \frac{|\eta|}{2K} \sum_{t=1}^T \|\omega_t\|_*^2 .$$

V.2.2. Upper bound on the regret

We first assume $s \geqslant 2$. Let $p \in (1, 2]$ and define the following regularizer:

$$h_p(x) = \begin{cases} \frac{1}{2}\|x\|_p^2 & \text{if } x \in \Delta_d \\ +\infty & \text{otherwise.} \end{cases}$$

One can easily check that h_p is indeed a regularizer on Δ_d and that $\delta_{h_p} \leqslant 1/2$. Moreover, it is $(p-1)$-strongly convex with respect to $\|\cdot\|_p$ (see [Bub11, Lemma 5.7] or [KSST12, Lemma 9]).

We can now state our first result, the general upper bound on regret when outcomes are s-sparse gains.

Theorem V.2.2. *Let $\eta > 0$ and $s \geqslant 3$. Against all sequences of s-sparse gain vectors g_t, i.e. $g_t \in [0,1]^d$ and $\|g_t\|_0 \leqslant s$, the Online Mirror Descent algorithm associated with regularizer h_p and parameter η guarantees:*

$$R_T \leqslant \frac{1}{2\eta} + \frac{\eta T s^{2/q}}{2(p-1)},$$

where $1/p + 1/q = 1$. In particular, the choices $\eta = \sqrt{(p-1)/Ts^{2/q}}$ and $p = 1 + (2\log s - 1)^{-1}$ give:

$$R_T \leqslant \sqrt{2eT\log s}.$$

Proof. h_p being $(p-1)$-strongly convex with respect to $\|\cdot\|_p$, and $\|\cdot\|_q$ being the dual norm of $\|\cdot\|_p$, Theorem V.2.1 gives:

$$R_T \leqslant \frac{\delta_{h_p}}{\eta} + \frac{\eta}{2(p-1)} \sum_{t=1}^{T} \|g_t\|_q^2 .$$

For each $t \geqslant 1$, the norm of g_t can be bounded as follows:

$$\|g_t\|_q^2 = \left(\sum_{i=1}^{d} \left| g_t^{(i)} \right|^q \right)^{2/q} \leqslant \left(\sum_{s \text{ terms}} \left| g_t^{(i)} \right|^q \right)^{2/q} \leqslant s^{2/q},$$

which yields

$$R_T \leqslant \frac{1}{2\eta} + \frac{\eta T s^{2/q}}{2(p-1)}.$$

We can now balance both terms by choosing $\eta = \sqrt{(p-1)/(\mathrm{T}s^{2/q})}$ and get:

$$\mathrm{R}_\mathrm{T} \leqslant \sqrt{\frac{\mathrm{T}s^{2/q}}{p-1}}.$$

Finally, since $s \geqslant 3$, we have $2\log s > 1$ and we set $p = 1 + (2\log s - 1)^{-1} \in (1, 2]$, which gives:

$$\frac{1}{q} = 1 - \frac{1}{p} = \frac{p-1}{p} = \frac{(2\log s - 1)^{-1}}{1 + (2\log s - 1)^{-1}} = \frac{1}{2\log s},$$

and thus:

$$\mathrm{R}_\mathrm{T} \leqslant \sqrt{\frac{\mathrm{T}s^{2/q}}{p-1}} = \sqrt{2\mathrm{T}\log s\, e^{2\log s/q}} = \sqrt{2e\,\mathrm{T}\log s}.$$

□

We emphasize the fact that we obtain, up to a multiplicative constant, the exact same rate as when the decision maker only has a set of s decisions.

Theorem V.2.2 was restricted to $s \geqslant 3$ to simplify the analysis. In the cases $s = 1, 2$, we can easily derive a bound of respectively $\sqrt{\mathrm{T}}$ and $\sqrt{2\mathrm{T}}$ using the same regularizer with $p = 2$.

V.2.3. Matching lower bound

For $s \in [d]$ and $\mathrm{T} \geqslant 1$, we denote $v_\mathrm{T}^{g,s,d}$ the minimax regret of the T-stage decision problem with outcome vectors restricted to s-sparse gains:

$$v_\mathrm{T}^{g,s,d} = \min_{\text{strat. } (g_t)_t} \max \mathrm{R}_\mathrm{T}$$

where the minimum is taken over all possible policies of the decision maker, and the maximum over all sequences of s-sparse gains vectors.

To establish a lower bound in the present setting, we can assume that only the s first coordinates of g_t may be positive (for all $t \geqslant 1$) and that the decision maker is aware of that. Therefore he has no interest in assigning positive probabilities to any decision but the first s ones. Indeed, for any mixed action x_t, the decision maker can construct alternative mixed action $x'_t = (x_t^{(1)}, \dots, x_t^{(s)} + \cdots + x_t^{(d)}, 0, \dots, 0)$ which obviously give a higher payoff:

$$\langle g_t, x_t \rangle \leqslant \langle g_t, x'_t \rangle$$

and therefore a lower regret:

$$\max_{i\in[d]} \sum_{t=1}^{\mathrm{T}} g_t^{(i)} - \sum_{t=1}^{\mathrm{T}} \langle g_t, x'_t \rangle \leqslant \max_{i\in[d]} \sum_{t=1}^{\mathrm{T}} g_t^{(i)} - \sum_{t=1}^{\mathrm{T}} \langle g_t, x_t \rangle .$$

Therefore, we can restrict the strategies of the decision maker to those which assign positive probability to the s first components only. That setup, which is simpler for the decision maker than the original one, is obviously equivalent to the basic regret minimization problem with only s decisions. Therefore, the classical lower bound [CBFH+97, Theorem 3.2.3] holds and we obtain the following.

Theorem V.2.3.

$$\liminf_{\substack{s\to+\infty \\ d\geqslant s}} \liminf_{\mathrm{T}\to+\infty} \frac{v_{\mathrm{T}}^{g,s,d}}{\sqrt{\mathrm{T}\log s}} \geqslant \frac{\sqrt{2}}{2}.$$

The same lower bound, up to the multiplicative constant actually holds non asymptotically, see [CBL06, Theorem 3.6].

An immediate consequence of Theorem V.2.3 is that the regret bound derived in Theorem V.2.2 is asymptotically minimax optimal, up to a multiplicative constant.

V.3. When outcomes are losses to be minimized

V.3.1. Upper bound on the regret

We now consider the case of losses, and the regularizer shall no longer depend on s (as with gains), as we will always use the Exponential Weight Algorithm. Instead, it is the parameter η that will be tuned as a function of s.

Theorem V.3.1. *Let $s \geqslant 1$. For any sequence of s-sparse loss vectors $(\ell_t)_{t\geqslant 1}$, i.e. $\ell_t \in [0,1]^d$ and $\|\ell_t\|_0 \leqslant s$, the Exponential Weight Algorithm with parameter $-\eta$ where $\eta := \log\left(1+\sqrt{2d\log d/s\mathrm{T}}\right) > 0$ guarantees, for $\mathrm{T} \geqslant 1$:*

$$\mathrm{R}_{\mathrm{T}} \leqslant \sqrt{\frac{2s\mathrm{T}\log d}{d}} + \log d.$$

We build upon the following regret bound for losses which is written in terms of the performance of the best action.

Theorem V.3.2 ([LW94]; [CBL06] Th 2.4)**.** *Let $\eta > 0$. For any sequence of loss vectors $(\ell_t)_{t\geqslant 1}$ in $[0,1]^d$, the Exponential Weight Algorithm with parameter $-\eta$ guarantees, for all $\mathrm{T} \geqslant 1$:*

$$\mathrm{R}_{\mathrm{T}} \leqslant \frac{\log d}{1-e^{-\eta}} + \left(\frac{\eta}{1-e^{-\eta}} - 1\right) \mathrm{L}_{\mathrm{T}}^* \ ,$$

where $\mathrm{L}_{\mathrm{T}}^* = \min_{i\in[d]} \sum_{t=1}^{\mathrm{T}} \ell_t^{(i)}$ is the loss of the best stationary decision.

Proof of Theorem V.3.1. Let $\mathrm{T} \geqslant 1$ and $\mathrm{L}_{\mathrm{T}}^* = \min_{i\in[d]} \sum_{t=1}^{\mathrm{T}} \ell_t^{(i)}$ be the loss of the best stationary policy. First note that since the loss vectors ℓ_t are s-sparse, we have $s \geqslant \sum_{i=1}^{d} \ell_t^{(i)}$. By summing over $1 \leqslant t \leqslant \mathrm{T}$:

$$s\mathrm{T} \geq \sum_{t=1}^{\mathrm{T}} \sum_{i=1}^{d} \ell_t^{(i)} = \sum_{i=1}^{d} \left(\sum_{t=1}^{\mathrm{T}} \ell_t^{(i)} \right) \geq d \left(\min_{i\in[d]} \sum_{t=1}^{\mathrm{T}} \ell_t^{(i)} \right) = d\mathrm{L}_{\mathrm{T}}^*,$$

and therefore, we have $\mathrm{L}_{\mathrm{T}}^* \leqslant \mathrm{T}s/d$.

Then, by using the inequality $\eta \leqslant (e^{\eta} - e^{-\eta})/2$, the bound from Theorem V.3.2 becomes:

$$\mathrm{R}_{\mathrm{T}} \leqslant \frac{\log d}{1 - e^{-\eta}} + \left(\frac{e^{\eta} - e^{-\eta}}{2(1 - e^{-\eta})} - 1 \right) \mathrm{L}_{\mathrm{T}}^* \ .$$

The factor of $\mathrm{L}_{\mathrm{T}}^*$ in the second term can be transformed as follows:

$$\frac{e^{\eta} - e^{-\eta}}{2(1 - e^{-\eta})} - 1 = \frac{(1 + e^{-\eta})(e^{\eta} - e^{-\eta})}{2(1 - e^{-2\eta})} - 1 = \frac{(1 + e^{-\eta})e^{\eta}}{2} - 1 = \frac{e^{\eta} - 1}{2} \ ,$$

and therefore the bound on the regret becomes:

$$\mathrm{R}_{\mathrm{T}} \leqslant \frac{\log d}{1 - e^{-\eta}} + \frac{e^{\eta} - 1}{2} \mathrm{L}_{\mathrm{T}}^* \leqslant \frac{\log d}{1 - e^{-\eta}} + \frac{(e^{\eta} - 1)\mathrm{T}s}{2d} \ ,$$

where we have been able to use the upper-bound on $\mathrm{L}_{\mathrm{T}}^*$ since $\frac{e^{\eta}-1}{2} \geqslant 0$. Along with the choice $\eta = \log(1 + \sqrt{2d \log d / \mathrm{T}s})$ and standard computations, this yields:

$$\mathrm{R}_{\mathrm{T}} \leqslant \sqrt{\frac{2\mathrm{T}s \log d}{d}} + \log d \ .$$

□

Interestingly, the bound from Theorem V.3.1 shows that $\sqrt{2s\mathrm{T} \log d / d}$, the dominating term of the regret bound, is *decreasing* when the number of decisions d increases. This is due to the sparsity assumptions (as the regret increases with s, the maximal number of decision with positive losses). Indeed, when s is fixed and d increases, more and more decisions are optimal at each stage, a proportion $1 - s/d$ to be precise. As a consequence, it becomes *easier* to find an optimal decisions when d increases. However, this intuition will turn out not to be valid in the bandit framework.

On the other hand, if the proportion s/d of positive losses remains constant then the regret bound achieved is of the same order as in the usual case.

V.3.2. Matching lower bound

When outcomes are losses, the argument from Section V.2.3 does not allow to derive a lower bound. Indeed, if we assume that only the first s coordinates of the loss vectors ℓ_t can be positive, and that the decision maker knows it, then he just has to take at each stage the decision $d_t = d$ which incurs a loss of 0. As a consequence, he trivially has a regret $R_T = 0$. Choosing at random, but once and for all, a fixed subset of s coordinates does not provide any interesting lower bound either. Instead, the key idea of the following result is to choose at random and at each stage the s coordinates associated with positive losses. And we therefore use the following classical probabilistic argument. Assume that we have found a probability distribution on $(\ell_t)_t$ such that the expected regret can be bounded from below by a quantity which does not depend on the strategy of the decision maker. This would imply that for any algorithm, there exists a sequence of $(\ell_t)_t$ such that the regret is greater than the same quantity.

In the following statement, $v_T^{\ell,s,d}$ stands for the minimax regret in the case where outcomes are losses.

Theorem V.3.3. *For all $s \geqslant 1$,*

$$\liminf_{d\to+\infty} \liminf_{T\to+\infty} \frac{v_T^{\ell,s,d}}{\sqrt{T\frac{s}{d}\log d}} \geqslant \frac{\sqrt{2}}{2}.$$

The main consequences of this theorem are that the algorithm described in Theorem V.3.1 is asymptotically minimax optimal (up to a multiplicative constant) and that gains and losses are fundamentally different from the point of view of regret minimization.

Proof. We define the sequence of i.i.d. loss vectors ℓ_t ($t \geqslant 1$) as follows. First, we draw a set $I_t \subset [d]$ of cardinality s uniformly among the $\binom{d}{s}$ possibilities. Then, if $i \in I_t$ set $\ell_t^{(i)} = 1$ with probability $1/2$ and $\ell_t^{(i)} = 0$ with probability $1/2$, independently for each component. If $i \notin I_t$, we set $\ell_t^{(i)} = 0$.

As a consequence, we always have that ℓ_t is s-sparse. Moreover, for each $t \geqslant 1$ and each coordinate $i \in [d]$, $\ell_t^{(i)}$ satisfies:

$$\mathbb{P}\left[\ell_t^{(i)} = 1\right] = \frac{s}{2d} \quad \text{and} \quad \mathbb{P}\left[\ell_t^{(i)} = 0\right] = 1 - \frac{s}{2d},$$

thus $\mathbb{E}\left[\ell_t^{(i)}\right] = s/2d$. Therefore we obtain that for any algorithm $(x_t)_{t\geqslant 1}$, $\mathbb{E}\left[\langle \ell_t, x_t\rangle\right] =$

$s/2d$. This yields that

$$\begin{aligned}\mathbb{E}\left[\frac{R_T}{\sqrt{T}}\right] &= \mathbb{E}\left[\frac{1}{\sqrt{T}}\left(\sum_{t=1}^{T}\langle \ell_t, x_t\rangle - \min_{i\in[d]}\sum_{t=1}^{T}\ell_t^{(i)}\right)\right]\\ &= \mathbb{E}\left[\max_{i\in[d]}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(\frac{s}{2d}-\ell_t^{(i)}\right)\right]\\ &= \mathbb{E}\left[\max_{i\in[d]}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}X_t^{(i)}\right],\end{aligned}$$

where $t \geqslant 1$, we have defined the random vector X_t by $X_t^{(i)} = s/2d - \ell_t^{(i)}$ for all $i \in [d]$. For $t \geqslant 1$, the X_t are i.i.d. zero-mean random vectors with values in $[-1, 1]^d$. We can therefore apply the comparison Lemma V.3.5 to get:

$$\liminf_{T\to+\infty}\mathbb{E}\left[\frac{R_T}{\sqrt{T}}\right] = \liminf_{T\to+\infty}\mathbb{E}\left[\max_{i\in[d]}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}X_t^{(i)}\right] \geqslant \mathbb{E}\left[\max_{i\in[d]}Z^{(i)}\right],$$

where $Z \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = (\operatorname{cov}(X_1^{(i)}, X_1^{(j)}))_{i,j}$.

We now make appeal to Slepian's lemma, recalled in Proposition V.3.4 below. Therefore, we introduce the Gaussian vector $W \sim \mathcal{N}(0, \tilde{\Sigma})$ where

$$\tilde{\Sigma} = \operatorname{diag}\left(\operatorname{Var} X_1^{(1)}, \ldots, \operatorname{Var} X_1^{(1)}\right).$$

As a consequence, the first two hypotheses of Proposition V.3.4 follow from the definitions of Z and W. Let $i \neq j$, then

$$\mathbb{E}\left[Z^{(i)}Z^{(j)}\right] = \operatorname{cov}(Z^{(i)}, Z^{(j)}) = \operatorname{cov}(\ell_1^{(i)}, \ell_1^{(j)}) = \mathbb{E}\left[\ell_1^{(i)}\ell_1^{(j)}\right] - \mathbb{E}\left[\ell_1^{(i)}\right]\mathbb{E}\left[\ell_1^{(j)}\right].$$

By definition of ℓ_1, $\ell_1^{(i)}\ell_1^{(j)} = 1$ if and only if $\ell_1^{(i)} = \ell_1^{(j)} = 1$ and $\ell_1^{(i)}\ell_1^{(j)} = 0$ otherwise. Therefore, using the random subset I_1 that appears in the definition of ℓ_1:

$$\begin{aligned}\mathbb{E}\left[Z^{(i)}Z^{(j)}\right] &= \mathbb{P}\left[\ell_1^{(i)} = \ell_1^{(j)} = 1\right] - \left(\frac{s}{2d}\right)^2\\ &= \mathbb{P}\left[\ell_1^{(i)} = \ell_1^{(j)} = 1 \,\middle|\, \{i, j\} \subset I_1\right]\mathbb{P}\left[\{i, j\} \subset I_1\right] - \left(\frac{s}{2d}\right)^2\\ &= \frac{1}{4}\cdot\frac{\binom{d-2}{s-2}}{\binom{d}{s}} - \left(\frac{s}{2d}\right)^2\\ &= \frac{1}{4}\left(\frac{s(s-1)}{d(d-1)} - \frac{s^2}{d^2}\right) \leqslant 0,\end{aligned}$$

and since $\mathbb{E}\left[W^{(i)}W^{(i)}\right] = 0$, the third hypothesis of Slepian's lemma is also satisfied. It yields that, for all $\theta \in \mathbb{R}$:

$$\mathbb{P}\left[\max_{i\in[d]} Z^{(i)} \leqslant \theta\right] = \mathbb{P}\left[Z^{(1)} \leqslant \theta, \dots, Z^{(d)} \leqslant \theta\right] \leqslant \mathbb{P}\left[W^{(1)} \leqslant \theta, \dots, W^{(d)} \leqslant \theta\right] = \mathbb{P}\left[\max_{i\in[d]} W^{(i)} \leqslant \theta\right].$$

This inequality between two cumulative distribution functions implies, the reverse inequality on expectations:

$$\mathbb{E}\left[\max_{i\in[d]} Z^{(i)}\right] \geqslant \mathbb{E}\left[\max_{i\in[d]} W^{(i)}\right] .$$

The components of the Gaussian vector W being independent, and of variance $\operatorname{Var} \ell_1^{(1)}$, we have

$$\mathbb{E}\left[\max_{i\in[d]} W^{(i)}\right] = \kappa_d \sqrt{\operatorname{Var} \ell_1^{(1)}} = \kappa_d \sqrt{\frac{s}{2d}\left(1 - \frac{s}{2d}\right)} \geqslant \kappa_d \sqrt{\frac{s}{4d}} ,$$

where κ_d is the expectation of the maximum of d Gaussian variables. Combining everything gives:

$$\liminf_{T\to+\infty} \frac{v_T^{\ell,s,d}}{\sqrt{T}} \geqslant \liminf_{T\to+\infty} \mathbb{E}\left[\frac{R_T}{\sqrt{T}}\right] \geqslant \mathbb{E}\left[\max_{i\in[d]} Z^{(i)}\right] \geqslant \mathbb{E}\left[\max_{i\in[d]} W^{(i)}\right] \geqslant \kappa_d \sqrt{\frac{s}{4d}} .$$

And for large d, since κ_d is equivalent to $\sqrt{2\log d}$, see e.g., [Gal78]

$$\liminf_{d\to+\infty} \liminf_{T\to+\infty} \frac{v_T^{\ell,s,d}}{\sqrt{T\frac{s}{d}\log d}} \geqslant \frac{\sqrt{2}}{2} .$$

□

Proposition V.3.4 (Slepian's lemma [Sle62]). *Let* $Z = (Z^{(1)}, \dots, Z^{(d)})$ *and* $W = (W^{(1)}, \dots, W^{(d)})$ *be Gaussian random vectors in* $\mathbb{R}^d$ *satisfying:*

(i) $\mathbb{E}[Z] = \mathbb{E}[W] = 0$;

(ii) $\mathbb{E}\left[(Z^{(i)})^2\right] = \mathbb{E}\left[(W^{(i)})^2\right]$ *for* $i \in [d]$;

(iii) $\mathbb{E}\left[Z^{(i)}Z^{(j)}\right] \leqslant \mathbb{E}\left[W^{(i)}W^{(j)}\right]$ *for* $i \neq j \in [d]$.

Then, for all real numbers $\theta_1, \dots, \theta_d$, *we have:*

$$\mathbb{P}\left[Z^{(1)} \leqslant \theta_1, \dots, Z^{(d)} \leqslant \theta_d\right] \leqslant \mathbb{P}\left[W^{(1)} \leqslant \theta_1, \dots, W^{(d)} \leqslant \theta_d\right] .$$

The following lemma is an extension of e.g. [CBL06, Lemma A.11] to random vectors with correlated components.

Lemma V.3.5 (Comparison lemma). *For $t \geqslant 1$, let $(\mathrm{X}_t)_{t\geqslant 1}$ be i.i.d. zero-mean random vectors in $[-1,1]^d$, Σ be the covariance matrix of X_t and $\mathrm{Z} \sim \mathcal{N}(0,\Sigma)$. Then,*

$$\liminf_{\mathrm{T}\to+\infty} \mathbb{E}\left[\max_{i\in[d]} \frac{1}{\sqrt{\mathrm{T}}} \sum_{t=1}^{\mathrm{T}} \mathrm{X}_t^{(i)}\right] \geqslant \mathbb{E}\left[\max_{i\in[d]} \mathrm{Z}^{(i)}\right].$$

Proof. Denote

$$\mathrm{Y}_\mathrm{T} = \max_{i\in[d]} \frac{1}{\sqrt{\mathrm{T}}} \sum_{t=1}^{\mathrm{T}} \mathrm{X}_t^{(i)}.$$

Let $\mathrm{A} \leqslant 0$ and consider the function $\phi_\mathrm{A} : \mathbb{R} \to \mathbb{R}$ defined by $\phi_\mathrm{A}(x) = \max(x, \mathrm{A})$.

$$\begin{aligned}
\mathbb{E}[\mathrm{Y}_\mathrm{T}] &= \mathbb{E}\left[\mathrm{Y}_\mathrm{T} \cdot \mathbb{1}_{\{\mathrm{Y}_\mathrm{T}\geqslant \mathrm{A}\}}\right] + \mathbb{E}\left[\mathrm{Y}_\mathrm{T} \cdot \mathbb{1}_{\{\mathrm{Y}_\mathrm{T}<\mathrm{A}\}}\right] \\
&= \mathbb{E}\left[\phi_\mathrm{A}(\mathrm{Y}_\mathrm{T}) \cdot \mathbb{1}_{\{\mathrm{Y}_\mathrm{T}\geqslant \mathrm{A}\}}\right] + \mathbb{E}\left[\mathrm{Y}_\mathrm{T} \cdot \mathbb{1}_{\{\mathrm{Y}_\mathrm{T}<\mathrm{A}\}}\right] \\
&= \mathbb{E}[\phi_\mathrm{A}(\mathrm{Y}_\mathrm{T})] - \mathbb{E}\left[\phi_\mathrm{A}(\mathrm{Y}_\mathrm{T}) \cdot \mathbb{1}_{\{\mathrm{Y}_\mathrm{T}<\mathrm{A}\}}\right] + \mathbb{E}\left[\mathrm{Y}_\mathrm{T} \cdot \mathbb{1}_{\{\mathrm{Y}_\mathrm{T}<\mathrm{A}\}}\right] \\
&= \mathbb{E}[\phi_\mathrm{A}(\mathrm{Y}_\mathrm{T})] - \mathbb{E}\left[(\mathrm{A}-\mathrm{Y}_\mathrm{T}) \cdot \mathbb{1}_{\{\mathrm{A}-\mathrm{Y}_\mathrm{T}>0\}}\right].
\end{aligned}$$

Let us estimate the second term. Denote $\mathrm{Z}_\mathrm{T} = (\mathrm{A}-\mathrm{Y}_\mathrm{T}) \cdot \mathbb{1}_{\mathrm{A}-\mathrm{Y}_\mathrm{T}>0}$. We clearly have, for all $u > 0$, $\mathbb{P}[\mathrm{Z}_\mathrm{T} > u] = \mathbb{P}[\mathrm{A}-\mathrm{Y}_\mathrm{T} > u]$. And Z_T being nonnegative, we can write:

$$\begin{aligned}
0 \leqslant \mathbb{E}\left[(\mathrm{A}-\mathrm{Y}_\mathrm{T}) \cdot \mathbb{1}_{\{\mathrm{A}-\mathrm{Y}_\mathrm{T}\}>0}\right] &= \mathbb{E}[\mathrm{Z}_\mathrm{T}] \\
&= \int_0^{+\infty} \mathbb{P}[\mathrm{Z}_\mathrm{T} > u]\, \mathrm{d}u \\
&= \int_0^{+\infty} \mathbb{P}[\mathrm{A}-\mathrm{Y}_\mathrm{T} > u]\, \mathrm{d}u \\
&= \int_{-\mathrm{A}}^{+\infty} \mathbb{P}[\mathrm{Y}_\mathrm{T} < -u]\, \mathrm{d}u \\
&= \int_{-\mathrm{A}}^{+\infty} \mathbb{P}\left[\max_{i\in[d]} \frac{1}{\sqrt{\mathrm{T}}} \sum_{t=1}^{\mathrm{T}} \mathrm{X}_t^{(i)} < u\right] \mathrm{d}u \\
&\leqslant \int_{-\mathrm{A}}^{+\infty} \mathbb{P}\left[\sum_{t=1}^{\mathrm{T}} \mathrm{X}_t^{(1)} < u\sqrt{\mathrm{T}}\right] \mathrm{d}u.
\end{aligned}$$

For $u > 0$, using Hoeffding's inequality together with the assumptions $\mathbb{E}\left[X_t^{(1)}\right] = 0$ and $X_t^{(1)} \in [-1, 1]$, we can bound the last integrand:

$$\mathbb{P}\left[\sum_{t=1}^{T} X_t^{(1)} < u\sqrt{T}\right] \leqslant e^{-u^2/2},$$

Which gives:

$$0 \leqslant \mathbb{E}\left[(A - Y_T) \cdot \mathbb{1}_{\{A - Y_T\} > 0}\right] \leqslant \int_{-A}^{+\infty} e^{-u^2/2}\, du \leqslant \frac{e^{-A^2/2}}{-A}.$$

Therefore:

$$\mathbb{E}[Y_T] \geqslant \mathbb{E}[\phi_A(Y_T)] + \frac{e^{-A^2/2}}{A}.$$

We now take the liminf on both sides as $t \to +\infty$. The left-hand side is the quantity that appears in the statement. We now focus on the second term of the right-hand side. The central limit theorem gives the following convergence in distribution:

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \xrightarrow[T \to +\infty]{\mathscr{L}} X.$$

The application $(x^{(1)}, \ldots, x^{(d)}) \longmapsto \max_{i \in [d]} x^{(i)}$ being continuous, we can apply the continuous mapping theorem:

$$Y_T = \max_{i \in [d]} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t^{(i)} \xrightarrow[n \to +\infty]{\mathscr{L}} \max_{i \in [d]} X^{(i)}.$$

This convergence in distribution allows the use of the portmanteau lemma: ϕ_A being lower semi-continuous and bounded from below, we have:

$$\liminf_{t \to +\infty} \mathbb{E}[\phi_A(Y_T)] \geqslant \mathbb{E}\left[\phi_A\left(\max_{i \in [d]} X^{(i)}\right)\right],$$

and thus:

$$\liminf_{t \to +\infty} \mathbb{E}[Y_T] \geqslant \mathbb{E}\left[\phi_A\left(\max_{i \in [d]} X^{(i)}\right)\right] + \frac{e^{-A^2/2}}{A}.$$

We would now like to take the limit as $A \to -\infty$. By definition of ϕ_A, for $A \leqslant 0$, we have the following domination:

$$\left|\phi_A\left(\max_{i \in [d]} X^{(i)}\right)\right| \leqslant \left|\max_{i \in [d]} X^{(i)}\right| \leqslant \max_{i \in [d]} \left|X^{(i)}\right| \leqslant \sum_{i=1}^{d} \left|X^{(i)}\right|,$$

where each $X^{(i)}$ is L^1 since it is a normal random variable. We can therefore apply the dominated convergence theorem as $A \to -\infty$:

$$\mathbb{E}\left[\phi_A\left(\max_{i\in[d]} X^{(i)}\right)\right] \xrightarrow[A\to-\infty]{} \mathbb{E}\left[\max_{i\in[d]} X^{(i)}\right],$$

and eventually, we get the stated result:

$$\liminf_{t\to+\infty} \mathbb{E}\left[Y_T\right] \geqslant \mathbb{E}\left[\max_{i\in[d]} X^{(i)}\right].$$

□

V.4. When the sparsity level s is unknown

We no longer assume in this section that the decision maker have the knowledge of the sparsity level s. We modify our algorithms to be adaptive over the sparsity level of the observed gain/loss vectors. The algorithms are proved to essentially achieve the same regret bounds as in the case where s is known. The constructions follow the same ideas behind the classical doubling trick.

Specifically, let $T \geqslant 1$ be the number of rounds and s^* the highest sparsity level of the gain/loss vectors chosen by Nature up to time T. In the following, we construct algorithms which achieve regret bounds of order $\sqrt{T \log s^*}$ and $\sqrt{T\frac{s^* \log d}{d}}$ for gains and losses respectively, without prior knowledge of s^*.

Boths algorithms need the foreknowledge of the time-horizon T for the tuning of the parameters. The use of time-varying parameters as in Theorem I.3.1 should provide any-time guarantees.

V.4.1. For losses

Let $(\ell_t)_{t\geqslant 1}$ be the sequence of loss vectors in $[0,1]^d$ chosen by Nature, and $T \geqslant 1$ the number of rounds. We denote $s^* = \max_{1\leqslant t\leqslant T} \|\ell_t\|_0$ the higher sparsity level of the loss vectors up to time T. The goal is to construct an algorithm which achieves a regret bound of order $\sqrt{\frac{Ts^* \log d}{d}}$ without any prior knowledge about the sparsity level of the loss vectors.

The time instances $\{1, \dots, T\}$ will be divided into several time intervals. On each of those, the previous loss vectors will be left aside, and a new instance of the Exponential Weight Algorithm with a specific parameter will be run. Let $M = \lceil \log_2 s^* \rceil$ and $\tau(0) = 0$. Then, for $1 \leqslant m < M$ we define

$$\tau(m) = \min\left\{1 \leqslant t \leqslant T \,\middle|\, \|\ell_t\|_0 > 2^m\right\} \quad \text{and} \quad \tau(M) = T.$$

In other words, $\tau(m)$ is the first time instance at which the sparsity level of the loss vector exceeds 2^m. $(\tau(m))_{1\leqslant m\leqslant \mathrm{M}}$ is thus a nondecreasing sequence. We can then define the time intervals $\mathrm{I}(m)$ as follows. For $1 \leqslant m \leqslant \mathrm{M}$, let

$$\mathrm{I}(m) = \begin{cases} \{\tau(m-1)+1, \dots, \tau(m)\} & \text{if } \tau(m-1) < \tau(m) \\ \emptyset & \text{if } \tau(m-1) = \tau(m). \end{cases}.$$

The sets $(\mathrm{I}(m))_{1\leqslant m\leqslant \mathrm{M}}$ clearly form a partition of $\{1, \dots, \mathrm{T}\}$ (some of the intervals may be empty). For $1 \leqslant t \leqslant \mathrm{T}$, we define $m_t = \min\{m \geqslant 1 \,|\, \tau(m) \geqslant t\}$ which implies $t \in \mathrm{I}(m_t)$. In other words, m_t is the index of the only interval t belongs to.

Let $\mathrm{C} > 0$ be a constant to be chosen later and for $1 \leqslant m \leqslant \mathrm{M}$, let

$$\eta(m) = \log\left(1 + \mathrm{C}\sqrt{\frac{d\log d}{2^m \mathrm{T}}}\right)$$

be the parameter of the Exponential Weight Algorithm to be used on interval $\mathrm{I}(m)$. In this section, h will be entropic regularizer on the simplex $h(x) = \sum_{i=1}^d x^{(i)} \log x^{(i)}$, so that $y \longmapsto \nabla h^*(y)$ is the *logit map* used in the Exponential Weight Algorithm. We can then define the played actions to be:

$$x_t = \nabla h^*\left(-\eta(m_t) \sum_{\substack{t'<t \\ t'\in \mathrm{I}(m_t)}} \ell_{t'}\right), \quad t = 1, \dots, \mathrm{T}.$$

Theorem V.4.1. *The above algorithm with* $\mathrm{C} = 2^{3/4}(\sqrt{2}+1)^{1/2}$ *guarantees*

$$\mathrm{R}_\mathrm{T} \leqslant 4\sqrt{\frac{\mathrm{T}s^* \log d}{d}} + \frac{\lceil \log s^* \rceil \log d}{2} + 5s^*\sqrt{\frac{\log d}{d\mathrm{T}}}.$$

Proof. Let $1 \leqslant m \leqslant \mathrm{M}$. On time interval $\mathrm{I}(m)$, the Exponential Weight Algorithm is run with parameter $\eta(m)$ against loss vectors in $[0,1]^d$. Therefore, the following regret bound derived in the proof of Theorem V.3.1 applies:

$$\begin{aligned} \mathrm{R}(m) &:= \sum_{t\in \mathrm{I}(m)} \langle \ell_t, x_t\rangle - \min_{i\in[d]} \sum_{t\in \mathrm{I}(m)} \ell_t^{(i)} \\ &\leqslant \frac{\log d}{1-e^{-\eta(m)}} + \frac{e^{\eta(m)}-1}{2} \min_{i\in[d]} \sum_{t\in \mathrm{I}(m)} \ell_t^{(i)} \\ &= \frac{1}{\mathrm{C}}\sqrt{\frac{2^m \mathrm{T}\log d}{d}} + \frac{\log d}{\mathrm{C}} + \frac{\mathrm{C}}{2}\sqrt{\frac{d\log d}{2^m \mathrm{T}}} \cdot \min_{i\in[d]} \sum_{t\in \mathrm{I}(m)} \ell_t^{(i)}. \end{aligned}$$

Algorithm 1: For losses in full information without prior knowledge about sparsity

input: $\mathrm{T} \geqslant 1$, $d \geqslant 1$ integers, and $\mathrm{C} > 0$.
$\eta \leftarrow \log(1 + \mathrm{C}\sqrt{d \log d / 2\mathrm{T}})$;
$m \leftarrow 1$;
for $i \leftarrow 1$ **to** d **do**
  $w^{(i)} \leftarrow 1/d$;
end
for $t \leftarrow 1$ **to** T **do**
  draw and play decision i with probability $w^{(i)} / \sum_{j=1}^{d} w^{(j)}$;
  observe loss vector ℓ_t;
  if $\|\ell_t\|_0 \leqslant 2^m$ **then**
    for $i \leftarrow 1$ **to** d **do**
      $w^{(i)} \leftarrow w^{(i)} e^{-\eta \ell_t^{(i)}}$;
    end
  else
    $m \leftarrow \lceil \log_2 \|\ell_t\|_0 \rceil$;
    $\eta \leftarrow \log(1 + \mathrm{C}\sqrt{d \log d / 2^m \mathrm{T}})$;
    for $i \leftarrow 1$ **to** d **do**
      $w^{(i)} \leftarrow 1/d$;
    end
  end
end

We now bound the "best loss" quantity from above, using the fact that ℓ_t is 2^m-sparse for $t \in \mathrm{I}(m) \setminus \{\tau(m)\}$ and that $\ell_{\tau(m)}$ is s^*-sparse:

$$\sum_{i=1}^{d} \sum_{t\in\mathrm{I}(m)} \ell_t^{(i)} = \sum_{t\in\mathrm{I}(m)} \sum_{i=1}^{d} \ell_t^{(i)} = \sum_{\substack{t<\tau(m)\\ t\in\mathrm{I}(m)}} \sum_{i=1}^{d} \ell_t^{(i)} + \sum_{i=1}^{d} \ell_{\tau(m)}^{(i)}$$
$$\leqslant (\tau(m) - \tau(m-1))2^m + s^*,$$

which implies:

$$\min_{i\in[d]} \sum_{t\in\mathrm{I}(m)} \ell_t^{(i)} \leqslant \frac{(\tau(m) - \tau(m-1))2^m + s^*}{d}.$$

Therefore, the regret on interval $\mathrm{I}(m)$, which we will denote $\mathrm{R}(m)$, is bounded by:

$$\begin{aligned}
\mathrm{R}(m) &:= \sum_{t\in\mathrm{I}(m)} \langle \ell_t, x_t\rangle - \min_{i\in[d]} \sum_{t\in\mathrm{I}(m)} \ell_t^{(i)} \\
&\leqslant \frac{1}{\mathrm{C}}\sqrt{\frac{2^m \mathrm{T}\log d}{d}} + \frac{\log d}{\mathrm{C}} + \frac{\mathrm{C}}{2}\sqrt{\frac{2^m \log d}{d\mathrm{T}}}\,(\tau(m) - \tau(m-1)) + \frac{\mathrm{C}}{2}\sqrt{\frac{\log d}{2^m d\mathrm{T}}}s^* \\
&\leqslant \frac{1}{\mathrm{C}}\sqrt{\frac{2^m \mathrm{T}\log d}{d}} + \frac{\log d}{\mathrm{C}} + \frac{\mathrm{C}}{2}\sqrt{\frac{2s^* \log d}{d\mathrm{T}}}\,(\tau(m) - \tau(m-1)) + \frac{\mathrm{C}}{2}\sqrt{\frac{\log d}{2^m d\mathrm{T}}}s^*,
\end{aligned}$$

where we used $2^m \leqslant 2^{\mathrm{M}} = 2^{\lceil \log_2 s^*\rceil} \leqslant 2^{\log_2 s^* + 1} = 2s^*$ for the third term of the last line.

We now turn the whole regret R_T from 1 to T. Since $(\mathrm{I}(m))_{1\leqslant m\leqslant \mathrm{M}}$ is a partition of $\{1, \dots, \mathrm{T}\}$, we obtain

$$\begin{aligned}
\mathrm{R}_\mathrm{T} &= \sum_{t=1}^{\mathrm{T}} \langle \ell_t, x_t\rangle - \min_{i\in[d]} \sum_{t=1}^{\mathrm{T}} \ell_t^{(i)} \leqslant \sum_{m=1}^{\mathrm{M}} \sum_{t\in\mathrm{I}(m)} \langle \ell_t, x_t\rangle - \sum_{m=1}^{\mathrm{M}} \min_{i\in[d]} \sum_{t\in\mathrm{I}(m)} \ell_t^{(i)} = \sum_{m=1}^{\mathrm{M}} \mathrm{R}(m) \\
&\leqslant \frac{1}{\mathrm{C}}\sqrt{\frac{\mathrm{T}\log d}{d}} \sum_{m=1}^{\mathrm{M}} \sqrt{2^m} + \mathrm{C}\sqrt{\frac{s^*\mathrm{T}\log d}{2d}} + \frac{\mathrm{M}\log d}{\mathrm{C}} + \frac{\mathrm{C}}{2}\sqrt{\frac{\log d}{d\mathrm{T}}}s^* \sum_{m=1}^{\mathrm{M}} 2^{-m/2}.
\end{aligned}$$

The sum in the first term above can be bounded as follows

$$\sum_{m=1}^{\mathrm{M}} \sqrt{2^m} \leqslant \sum_{m=1}^{\mathrm{M}} \sqrt{2}^m = \sqrt{2}\frac{\sqrt{2}^{\mathrm{M}} - 1}{\sqrt{2} - 1} \leqslant \sqrt{2}\frac{\sqrt{2^{\log_2 s^* + 1}}}{\sqrt{2} - 1} = 2\frac{\sqrt{s^*}}{\sqrt{2} - 1} = 2(\sqrt{2} + 1)\sqrt{s^*},$$

whereas the sum in the last term can be bounded by $\sqrt{2} + 1$. Eventually, the choice $\mathrm{C} = 2^{3/4}(\sqrt{2} + 1)^{1/2}$ give:

$$\mathrm{R}_\mathrm{T} \leqslant 2^{5/4}(\sqrt{2} + 1)^{1/2}\sqrt{\frac{\mathrm{T}s^*\log d}{d}} + \frac{\lceil \log s^*\rceil \log d}{2^{3/4}(\sqrt{2} + 1)^{1/2}} + 2^{1/4}(\sqrt{2} + 1)^{3/2}s^*\sqrt{\frac{\log d}{d\mathrm{T}}},$$

and the statement follows from numerical computation of the constant factors. □

V.4.2. For gains

The construction is similar to the case of losses, but the time intervals are slightly different. Let $(g_t)_{t\geqslant 1}$ be the sequence of gain vectors in $[0,1]^d$ chosen by Nature. We assume $s^* \geqslant 2$ and set $\mathrm{M} = \lceil \log_2 \log_2 s^* \rceil$ and $\tau(0) = 0$. For $1 \leqslant m \leqslant \mathrm{M}$ we define

$$\tau(m) = \min\left\{1 \leqslant t \leqslant \mathrm{T} \,\middle|\, \|g_t\|_0 > 2^{2^m}\right\} \quad \text{and} \quad \tau(\mathrm{M}) = \mathrm{T}.$$

We now define the time intervals $\mathrm{I}(m)$. For $1 \leqslant m \leqslant \mathrm{M}$,

$$\mathrm{I}(m) = \begin{cases} \{\tau(m-1)+1, \dots, \tau(m)\} & \text{if } \tau(m-1) < \tau(m) \\ \emptyset & \text{if } \tau(m-1) = \tau(m). \end{cases}$$

Therefore, for $1 \leqslant m \leqslant \mathrm{M}$ and $t < \tau(m)$, we have $\|g_t\|_0 \leqslant 2^{2^m}$. For $1 \leqslant t \leqslant \mathrm{T}$, we denote $m_t = \min\{m \geqslant 1 \,|\, \tau(m) \geqslant t\}$. Let $\mathrm{C} > 0$ be a constant to be chosen later and for $1 \leqslant m \leqslant \mathrm{M}$, let

$$p(m) = 1 + \frac{1}{\log 2 \cdot 2^{m+1} - 1},$$
$$q(m) = \left(1 - \frac{1}{p(m)}\right)^{-1},$$
$$\eta(m) = \mathrm{C}\sqrt{\frac{p(m)-1}{\mathrm{T}2^{2^{m+1}/q(m)}}}.$$

As in Section V.2.2, for $p \in (1,2]$, we denote h_p the regularizer on the simplex defined by:

$$h_p(x) = \begin{cases} \frac{1}{2}\|x\|_p^2 & \text{if } x \in \Delta_d \\ +\infty & \text{otherwise.} \end{cases}$$

The algorithm is then defined by:

$$x_t = \nabla h^*_{p(m_t)}\left(\eta(m_t) \sum_{\substack{t'<t \\ t' \in \mathrm{I}(m_t)}} g_{t'}\right), \quad t = 1, \dots, \mathrm{T}.$$

Theorem V.4.2. *The above algorithm with* $\mathrm{C} = (e\sqrt{2}(\sqrt{2}+1))^{1/2}$ *guarantees*

$$\mathrm{R}_\mathrm{T} \leqslant 7\sqrt{\mathrm{T}\log s^*} + \frac{4s^*}{\sqrt{\mathrm{T}}}.$$

Algorithm 2: For gains in full information without prior knowledge about sparsity.

input: $\mathrm{T} \geqslant 1$, $d \geqslant 1$ integers, and $\mathrm{C} > 0$.
$p \leftarrow 1 + (4 \log 2 - 1)^{-1}$;
$q \leftarrow (1 - 1/p)^{-1}$;
$\eta \leftarrow \mathrm{C}\sqrt{(p-1)/2^{4/q}\mathrm{T}}$;
$m \leftarrow 1$;
$y \leftarrow (0, \dots, 0) \in \mathbb{R}^d$;
for $t \leftarrow 1$ **to** T **do**
  draw and play decision $i \sim \nabla h_p^*(\eta \cdot y)$;
  observe gain vector g_t;
  if $\|g_t\|_0 \leqslant 2^{2^m}$ **then**
    $y \leftarrow y + g_t$;
  else
    $m \leftarrow \lceil \log_2 \log_2 \|g_t\|_0 \rceil$;
    $p \leftarrow 1 + (\log 2 \cdot 2^{m+1} - 1)^{-1}$;
    $q \leftarrow (1 - 1/p)^{-1}$;
    $\eta \leftarrow \mathrm{C}\sqrt{(p-1)/2^{2^{m+1}/q}\mathrm{T}}$;
    $y \leftarrow (0, \dots, 0)$;
  end
end

Proof. Let $1 \leqslant m \leqslant \mathrm{M}$. On time interval $\mathrm{I}(m)$, the algorithm boils down to an Online Mirror Descent algorithm with regularizer $h_{p(m)}$ and parameter $\eta(m)$. Therefore, using Theorem V.2.1, the regret on this interval is bounded as follows.

$$\begin{aligned}
\mathrm{R}(m) &:= \max_{i\in[d]} \sum_{t\in\mathrm{I}(m)} g_t^{(i)} - \sum_{t\in\mathrm{I}(m)} \langle g_t, x_t\rangle \\
&\leqslant \frac{1}{2\eta(m)} + \frac{\eta(m)}{2(p(m)-1)} \sum_{t\in\mathrm{I}(m)} \|g_t\|_{q(m)}^2 \\
&= \frac{1}{2\eta(m)} + \frac{\eta(m)}{2(p(m)-1)} \left(\sum_{\substack{t\in\mathrm{I}(m)\\ t<\tau(m)}} \|g_t\|_{q(m)}^2 + \|g_{\tau(m)}\|_{q(m)}^2 \right).
\end{aligned}$$

g_t being 2^{2^m}-sparse for $t < \tau(m)$ and $g_{\tau(m)}$ being s^*-sparse, the $q(m)$-norms can therefore bounded from above as follows:

$$\|g_t\|_{q(m)}^2 \leqslant 2^{2^{m+1}/q(m)} \quad \text{and} \quad \|g_{\tau(m)}\|_{q(m)}^2 \leqslant (s^*)^{2/q(m)}.$$

The bound on $\mathrm{R}(m)$ then becomes

$$\begin{aligned}
\mathrm{R}(m) &\leqslant \frac{1}{2\eta(m)} + \frac{\eta(m)(\tau(m)-\tau(m-1))2^{2^{m+1}/q(m)}}{2(p(m)-1)} + \frac{\eta(m)(s^*)^{2/q(m)}}{2(p(m)-1)} \\
&= \frac{1}{2\mathrm{C}}\sqrt{\mathrm{T}e(\log 2\cdot 2^{m+1}-1)} + \frac{\mathrm{C}}{2}\sqrt{\frac{e(\log 2\cdot 2^{m+1}-1)}{\mathrm{T}}}(\tau(m)-\tau(m-1)) \\
&\qquad + \frac{\mathrm{C}}{2}(s^*)^{1/(\log 2\cdot 2^m)}\sqrt{\frac{e(\log 2\cdot 2^{m+1}-1)}{\mathrm{T}}} \\
&\leqslant \frac{1}{2\mathrm{C}}\sqrt{\mathrm{T}e\log 2\cdot 2^{m+1}} + \mathrm{C}\sqrt{\frac{e\log s^*}{\mathrm{T}}}\,(\tau(m)-\tau(m-1)) \\
&\qquad + \frac{\mathrm{C}}{2}s^*\sqrt{\frac{e\log 2\cdot 2^{m+1}}{\mathrm{T}}},
\end{aligned}$$

where for the second term of the last expression we used:

$$\begin{aligned}
\log 2\cdot 2^{m+1}-1 &\leqslant \log 2\cdot 2^{\mathrm{M}+1} = \log 2\cdot\exp\left(\log 2\left(\lceil \log_2\log_2 s^*\rceil + 1\right)\right) \\
&\leqslant \log 2\cdot\exp\left(\log 2\left(\log_2\log_2 s^* + 2\right)\right) \\
&= \log 2\cdot e^{2\log 2}\exp\left(\log 2\cdot\log_2\log_2 s^*\right) \\
&= 4\log 2\cdot\exp\left(\log\log_2 s^*\right) \\
&= 4\log 2\cdot\log_2 s^* \\
&= 4\log s^*.
\end{aligned}$$

Then, the whole regret R_T is bounded by the sum of the regrets on each interval:

$$R_T \leqslant \sum_{m=1}^{M} R(m) \leqslant \frac{1}{2C}\sqrt{Te\log 2}\sum_{m=1}^{M}\sqrt{2^{m+1}} + C\sqrt{\frac{e\log s^*}{T}}\sum_{m=1}^{M}(\tau(m)-\tau(m-1)) + \frac{Cs^*}{2}\sqrt{\frac{e\log 2}{T}}\sum_{m=1}^{M}2^{-(m+1)/2}.$$

The second sum is equal to T and the third sum is bounded from above by $(\sqrt{2}+1)/\sqrt{2}$. Let us bound the first sum from above:

$$\begin{aligned}\sqrt{\log 2}\sum_{m=1}^{M}\sqrt{2^{m+1}} &= 2\sqrt{\log 2}\frac{2^{M/2}-1}{\sqrt{2}-1}\\ &\leqslant 2(\sqrt{2}+1)\sqrt{\log 2}\cdot\exp\left(\frac{\log 2}{2}\left(\log_2\log_2 s^*+1\right)\right)\\ &= 2(\sqrt{2}+1)\sqrt{\log 2}\cdot\sqrt{2e^{\log\log_2 s^*}}\\ &= 2\sqrt{2}(\sqrt{2}+1)\sqrt{\log 2\log_2 s^*}\\ &= 2\sqrt{2}(\sqrt{2}+1)\sqrt{\log s^*}.\end{aligned}$$

Therefore,

$$R_T \leqslant \frac{\sqrt{2}(\sqrt{2}+1)}{C}\sqrt{Te\log s^*} + C\sqrt{Te\log s^*} + \frac{C(\sqrt{2}+1)s^*}{2}\sqrt{\frac{e\log 2}{2T}}.$$

Choosing $C = (e\sqrt{2}(\sqrt{2}+1))^{1/2}$ balances the first two term and gives:

$$\begin{aligned}R_T &\leqslant 2(e\sqrt{2}(\sqrt{2}+1))^{1/2}\sqrt{T\log s^*} + 2^{-5/4}e\sqrt{\log 2}(\sqrt{2}+1)^{3/2}\frac{s^*}{\sqrt{T}}\\ &\leqslant 7\sqrt{T\log s^*} + \frac{4s^*}{\sqrt{T}}.\end{aligned}$$

□

V.5. The bandit setting

We now turn to the bandit framework (see for instance [BCB12] for a recent survey). Recall that the minimax regret [AB09] in the basic bandit framework (without sparsity) is of order $\sqrt{Td}$. In the case of losses, we manage to take advantage of the

sparsity assumption and obtain in Theorem V.5.1 an upper bound of order $\sqrt{\mathrm{T}s\log\frac{d}{s}}$, and a lower bound of order $\sqrt{\mathrm{T}s}$ in Theorem V.5.3. This establishes the order of the minimax regret up to a logarithmic factor. In the case of gains, the argument from Section V.2.3 can be adapted to get a lower bound of order $\sqrt{s\mathrm{T}}$; but the upper bound techniques from losses do not seem to work; this difficulty is discussed below in remark V.5.2.

For simplicity, we shall assume that the sequence of outcome vectors $(\omega_t)_{t\geqslant 1}$ is chosen before stage 1 by the environment, which is called *oblivious* in that case. We refer to [BCB12, Section 3] for a detailed discussion on the difference between oblivious and non-oblivious opponent, and between regret and pseudo-regret.

As before, at stage t, the decision maker chooses $x_t \in \Delta_d$ and draws decision $d_t \in [d]$ according to x_t. The main difference with the previous framework is that the decision maker only observes his own outcome $\omega_t^{d_t}$ before choosing the next decision d_{t+1}.

V.5.1. Upper bounds on the regret with sparse losses

We shall focus in this section on s-sparse losses. The algorithm we consider belongs to the family of Greedy Online Mirror Descent. We follow [BCB12, Section 5] and refer to it for the detailed and rigorous construction. Let $\mathrm{F}_q(x)$ be the Legendre function associated with the potential $\psi(x) = (-x)^{-q}$ $(q > 1)$, i.e.

$$\mathrm{F}_q(x) = -\frac{q}{q-1}\sum_{i=1}^{d}(x^i)^{1-1/q}.$$

The algorithm, which depends on a parameter $\eta > 0$ to be fixed later, is defined as follows. Set $x_1 = (\frac{1}{d},\dots,\frac{1}{d}) \in \Delta_d$. For all $t \geqslant 1$, we define the estimator $\widehat{\ell}_t$ of ℓ_t as usual:

$$\widehat{\ell}_t^{(i)} = \mathbb{1}_{\{d_t=i\}}\frac{\ell_t^{(i)}}{x_t^{(i)}}, \quad i \in [d],$$

which is then used to compute

$$z_{t+1} = \nabla\mathrm{F}_q^*(\nabla\mathrm{F}_q(x_t) - \eta\widehat{\ell}_t) \quad \text{and} \quad x_{t+1} = \operatorname*{arg\,min}_{x\in\Delta_d}\mathrm{D}_{\mathrm{F}_q}(x, z_{t+1}),$$

where $\mathrm{D}_{\mathrm{F}_q} : \bar{\mathcal{D}} \times \mathcal{D} \to \mathbb{R}$ is the Bregman divergence associated with F_q:

$$\mathrm{D}_{\mathrm{F}_q}(x', x) = \mathrm{F}_q(x') - \mathrm{F}_q(x) - \langle\nabla\mathrm{F}_q(x), x' - x\rangle.$$

Theorem V.5.1. *Let $\eta > 0$ and $q > 1$. For any sequence of s-sparse loss vectors, the above strategy with parameter η guarantees, for $T \geq 1$:*

$$R_T \leqslant q\left(\frac{d^{1/q}}{\eta(q-1)} + \frac{\eta T s^{1-1/q}}{2}\right).$$

In particular, if $d/s \geqslant e^2$, the choices

$$\eta = \sqrt{\frac{2d^{1/q}}{(q-1)Ts^{1-1/q}}} \quad \textit{and} \quad q = \log(d/s)$$

the following regret bound:

$$R_T \leqslant 2\sqrt{e}\sqrt{Ts\log\frac{d}{s}}.$$

Proof. [BCB12, Theorem 5.10] gives:

$$R_T \leqslant \frac{\max_{x\in\Delta_d} F(x) - F(x_1)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{d}\mathbb{E}\left[\frac{(\widehat{\ell}_t^{(i)})^2}{(\psi^{-1})'(x_t^{(i)})}\right],$$

with $(\psi^{-1})'(x) = (qx^{1+1/q})^{-1}$. Let us bound the first term.

$$\frac{1}{\eta}\max_{x\in\Delta_d} F_q(x) - F_q(x_1) \leqslant \frac{1}{\eta}\frac{q}{q-1}\left(0 + d\,(1/d)^{1-1/q}\right) = \frac{qd^{1/q}}{\eta(q-1)}.$$

We turn to the second term. Let $1 \leqslant t \leqslant T$.

$$\begin{aligned}
\sum_{i=1}^{d}\mathbb{E}\left[\frac{(\widehat{\ell}_t^{(i)})^2}{(\psi^{-1})'(x_t^{(i)})}\right] &= q\sum_{i=1}^{d}\mathbb{E}\left[(\widehat{\ell}_t^{(i)})^2(x_t^{(i)})^{1+1/q}\right] \\
&= q\sum_{i=1}^{d}\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{d_t=i\}}\frac{(\ell_t^{(i)})^2}{(x_t^i)^2}(x_t^i)^{1+1/q}\middle| x_t\right]\right] \\
&= q\sum_{i=1}^{d}\mathbb{E}\left[(\ell_t^{(i)})^2(x_k^{(i)})^{1/q}\right] \\
&= q\,\mathbb{E}\left[\sum_{s\text{ terms}}(\ell_t^{(i)})^2(x_t^{(i)})^{1/q}\right] \\
&\leqslant qs(1/s)^{1/q} = qs^{1-1/q},
\end{aligned}$$

where we used the assumption that ℓ_t has at most s nonzero components, and the fact that $x_t \in \Delta_d$. The first regret bound is thus proven. By choosing $\eta = \sqrt{\frac{2s^{1-1/q}}{(q-1)\mathrm{T}d^{1/q}}}$, we balance both terms and get:

$$\mathrm{R_T} \leqslant 2q\sqrt{\frac{\mathrm{T}d^{1/q}s^{1-1/q}}{2(q-1)}} = \sqrt{2q}\sqrt{\mathrm{T}s\left(\frac{d}{s}\right)^{1/q}\left(\frac{q}{q-1}\right)}\ .$$

If $d/s \geqslant e^2$ and $q = \log(d/s)$, then $q/(q-1) \leqslant 2$ and finally:

$$\mathrm{R_T} \leqslant 2\sqrt{e}\sqrt{\mathrm{T}s\log\frac{d}{s}}.$$

□

Remark V.5.2. The previous analysis cannot be carried in the case of gains because the bound from [BCB12, Theorem 5.10] that we use above only holds for nonnegative losses (and its proof strongly relies on this assumption). We are unaware of techniques which could provide a similar bound in the case of nonnegative gains.

V.5.2. Matching lower bound

The following theorem establishes that the bound from Theorem V.5.1 is optimal up to a logarithmic factor. We denote $\hat{v}_{\mathrm{T}}^{\ell,s,d}$ the minimax regret in the bandit setting with losses.

Theorem V.5.3. *For all $d \geqslant 2$, $s \in [d]$ and $\mathrm{T} \geqslant d^2/4s$, the following lower bound holds:*

$$\hat{v}_{\mathrm{T}}^{\ell,s,d} \geqslant \frac{1}{32}\sqrt{\mathrm{T}s}.$$

The intuition behind the proof is the following. Let us consider the case where $s = 1$ and assume that ℓ_t is a unit vector $e_{i_t} = (\mathbb{1}\{j = i_t\})_j$ where $\mathbb{P}(i_t = i) \simeq (1+\varepsilon)/d$ for all $i \in [d]$, except one fixed coordinate i^* where $\mathbb{P}(i_t = i^*) \simeq 1/d - \varepsilon$.

Since $1/d$ goes to 0 as d increases, the Kullback–Leibler divergence between two Bernoulli of parameters $(1+\varepsilon)/d$ and $1/d - \varepsilon$ is of order $d\varepsilon^2$. As a consequence, it would require approximately $1/d\varepsilon^2$ samples to distinguish between the two. The standard argument that one of the coordinates has not been chosen more than T/d times, yields that one should take $1/d\varepsilon^2 \simeq \mathrm{T}/d$ so that the regret is of order $\mathrm{T}\varepsilon$. This provides a lower bound of order $\sqrt{\mathrm{T}}$. Similar arguments with $s > 1$ give a lower bound of order $\sqrt{s\mathrm{T}}$.

We emphasize that one cannot simply assume that the s components with positive losses are chosen at the beginning once for all, and apply standard lower bound techniques. Indeed, with this additional information, the decision maker just has to choose, at each stage, a decision associated with a zero loss. His regret would then be uniformly bounded (or even possibly equal to zero).

V.5.3. Proof of Theorem V.5.3

Let $d \geqslant 1, 1 \leqslant s \leqslant d, \mathrm{T} \geqslant 1$, and $\varepsilon \in (0, s/2d)$. Denote $\mathfrak{P}_s([d])$ the set of subsets of $[d]$ of cardinality s, δ_{ij} the Kronecker symbol, and $\mathrm{B}(1, p)$ the Bernoulli distribution of parameter $p \in [0,1]$. If P, Q are two probability distributions on the same set, $\mathrm{D}(\mathrm{P} \,\|\, \mathrm{Q})$ will denote the relative entropy of P and Q.

Random s-sparse loss vectors ℓ_t and ℓ'_t. — For $t \geqslant 1$, define the random s-sparse loss vectors $(\ell_t)_{t\geqslant 1}$ as follows. Draw Z uniformly from $[d]$. We will denote $\mathbb{P}_i[\,\cdot\,] = \mathbb{P}[\,\cdot \mid \mathrm{Z} = i]$ and $\mathbb{E}_i[\,\cdot\,] = \mathbb{E}[\,\cdot \mid \mathrm{Z} = i]$. Knowing $\mathrm{Z} = i$, the random vectors ℓ_t are i.i.d and defined as follows. Draw I_t uniformly from $\mathfrak{P}_s([d])$. If $j \in \mathrm{I}_t$, define $\ell_t^{(j)}$ such that:

$$\mathbb{P}_i\left[\ell_t^{(j)} = 1\right] = 1 - \mathbb{P}_i\left[\ell_t^{(j)} = 0\right] = \frac{1}{2} - \frac{\varepsilon d}{s}\delta_{ij}.$$

If $j \notin \mathrm{I}_t$, set $\ell_t^{(j)} = 0$. Therefore, one can check that for each component $j \in [d]$ and all $t \geqslant 1$,

$$\mathbb{E}_i\left[\ell_t^{(j)}\right] = \frac{s}{2d} - \varepsilon\delta_{ij}.$$

For $t \geqslant 1$, define the i.i.d. random s-sparse loss vectors $(\ell'_t)_{t\geqslant 1}$ as follows. Draw I'_t uniformly from $\mathfrak{P}_s([d])$. Then if $j \in \mathrm{I}'_t$, set $(\ell'_t)^{(j)}$ such that:

$$\mathbb{P}\left[(\ell'_t)^{(j)} = 1\right] = \mathbb{P}\left[(\ell'_t)^{(j)} = 0\right] = 1/2.$$

And if $j \notin \mathrm{I}'_t$, set $(\ell'_t)^{(j)} = 0$. Therefore, one can check that for each component $j \in [d]$ and all $t \geqslant 1$,

$$\mathbb{E}_i\left[(\ell'_t)^{(j)}\right] = \frac{s}{2d}.$$

By construction, ℓ_t and ℓ'_t are indeed random s-sparse loss vectors.

A deterministic strategy σ for the player. — We assume given a deterministic strategy $\sigma = (\sigma_t)_{t\geqslant 1}$ for the player:

$$\sigma_t : ([d] \times [0,1])^{t-1} \longrightarrow [d].$$

Therefore,

$$d_t = \sigma_t(d_1, \omega_1^{(d_1)}, \ldots, d_{t-1}, \omega_{t-1}^{(d_{t-1})}),$$

where d_t denotes the decision chosen by the strategy at stage t and ω_t the outcome vector of stage t. But since d_t is determined by previous decisions and outcomes, we can consider that σ_t only depends on the received outcomes:

$$\sigma_t : [0,1]^{t-1} \longrightarrow [d],$$

$$d_t = \sigma_t(\omega_1^{(d_1)}, \ldots, \omega_{t-1}^{(d_{t-1})}).$$

We define d_t and d'_t to be the (random) decisions played by deterministic strategy σ against the random loss vectors $(\ell_t)_{t\geqslant 1}$ and $(\ell'_t)_{t\geqslant 1}$ respectively:

$$\begin{aligned} d_t &= \sigma_t(\ell_1^{(d_1)}, \dots, \ell_{t-1}^{(d_{t-1})}), \\ d'_t &= \sigma_t((\ell'_1)^{(d'_1)}, \dots, (\ell'_{t-1})^{(d'_{t-1})}). \end{aligned}$$

For $t \geqslant 1$ and $i \in [d]$, define $\mathrm{A}_t^{(i)}$ to be the set of sequences of outcomes in $\{0,1\}$ of the first $t-1$ stages for which strategy σ plays decision i at stage t:

$$\mathrm{A}_t^{(i)} = \left\{ (u_1, \dots, u_{t-1}) \in \{0,1\}^{t-1} \,\middle|\, \sigma_t(u_1, \dots, u_{t-1}) = i \right\},$$

and $\mathrm{B}_t^{(i)}$ the complement:

$$\mathrm{B}_t^{(i)} = \{0,1\}^{t-1} \quad \mathrm{A}_t^{(i)}.$$

Note that for a given $t \geqslant 1$, $(\mathrm{A}_t^{(i)})_{i\in[d]}$ is a partition of $\{0,1\}^{t-1}$ (with possibly some empty sets).

For $i \in [d]$, define $\tau_i(\mathrm{T})$ (resp. $\tau'_i(\mathrm{T})$) to be the number of times decision i is played by strategy σ against loss vectors $(\ell_t)_{t\geqslant 1}$ (resp. against $(\ell'_t)_{t\geqslant 1}$) between stages 1 and T:

$$\tau_i(\mathrm{T}) = \sum_{t=1}^{\mathrm{T}} \mathbb{1}_{\{d_t = i\}} \quad \text{and} \quad \tau'_i(\mathrm{T}) = \sum_{t=1}^{\mathrm{T}} \mathbb{1}_{\{d'_t = i\}}.$$

The probability distributions $\mathbb{Q}$ and $\mathbb{Q}_i$ ($i \in [d]$) on binary sequences. — We consider binary sequences $\vec{u} = (u_1, \dots, u_{\mathrm{T}}) \in \{0,1\}^{\mathrm{T}}$. We define $\mathbb{Q}$ and $\mathbb{Q}_i$ ($i \in [d]$) to be probability distributions on $\{0,1\}^{\mathrm{T}}$ as follows:

$$\begin{aligned} \mathbb{Q}_i[\vec{u}] &= \mathbb{P}_i\left[\ell_1^{(d_1)} = u_1, \dots, \ell_{\mathrm{T}}^{(d_{\mathrm{T}})} = u_{\mathrm{T}}\right], \\ \mathbb{Q}[\vec{u}] &= \mathbb{P}\left[(\ell'_1)^{(d'_1)} = u_1, \dots, (\ell'_{\mathrm{T}})^{(d'_{\mathrm{T}})} = u_{\mathrm{T}}\right]. \end{aligned}$$

Fix $(u_1, \dots, u_{t-1}) \in \{0,1\}^t$. The applications

$$u_t \longmapsto \mathbb{Q}[u_t \,|\, u_1, \dots, u_{t-1}] \quad \text{and} \quad u_t \longmapsto \mathbb{Q}_i[u_t \,|\, u_1, \dots, u_{t-1}],$$

are probability distributions on $\{0,1\}$, which we now aim at identifying. The first one is Bernoulli of parameter $s/2d$. Indeed,

$$\begin{aligned} \mathbb{Q}[1 \,|\, u_1, \dots, u_{t-1}] &= \mathbb{P}\left[(\ell'_t)^{(d'_t)} = 1 \,\middle|\, (\ell'_1)^{(d'_1)} = u_1, \dots, (\ell'_{t-1})^{(d'_{t-1})} = u_{t-1}\right] \\ &= \mathbb{P}\left[(\ell'_t)^{(d'_t)} = 1\right] \\ &= \mathbb{P}\left[d'_t \in \mathrm{I}'_t\right] \mathbb{P}\left[(\ell'_t)^{(d_t)} = 1 \,\middle|\, d'_t \in \mathrm{I}'_t\right] \\ &= \frac{s}{d} \times \frac{1}{2} \\ &= \frac{s}{2d}, \end{aligned}$$

where we used the independence of the random vectors $(\ell'_t)_{t\geqslant 1}$ for the second inequality. We now turn to the second distribution, which depends on $(u_1, \dots, u_{t-1})$. If $(u_1, \dots, u_{t-1}) \in \mathrm{A}_t^{(i)}$, it is a Bernoulli of parameter $s/2d - \varepsilon$:

$$\begin{aligned}
\mathbb{Q}_i\left[1 \,|\, u_1, \dots, u_{t-1}\right] &= \mathbb{P}_i\left[\ell_t^{(d_t)} = 1 \,\middle|\, \ell_1^{(d_1)} = u_1, \dots, \ell_{t-1}^{(d_{t-1})} = u_{t-1}\right] \\
&= \mathbb{P}_i\left[\ell_t^{(i)} = 1 \,\middle|\, \ell_1^{(d_1)} = u_1, \dots, \ell_{t-1}^{(d_{t-1})} = u_{t-1}\right] \\
&= \mathbb{P}_i\left[\ell_t^{(i)} = 1\right] \\
&= \mathbb{P}_i\left[i \in \mathrm{I}_t\right] \mathbb{P}_i\left[\ell_t^{(i)} = 1 \,\middle|\, i \in \mathrm{I}_t\right] \\
&= \frac{s}{d} \times \left(\frac{1}{2} - \frac{\varepsilon d}{s}\right) \\
&= \frac{s}{2d} - \varepsilon.
\end{aligned}$$

where for the third inequality, we used the assumption that the random vectors $(\ell_t)_{t\geqslant 1}$ are independent under $\mathbb{P}_i$, i.e. knowing $\mathrm{Z} = i$. On the other hand, if $(u_1, \dots, u_{t-1}) \in \mathrm{B}_t^{(i)}$, we can prove similarly that the distribution is a Bernoulli of parameter $s/2d$.

Computation the relative entropy of $\mathbb{Q}_i$ and $\mathbb{Q}$. — We apply iteratively the chain rule to the relative entropy of $\mathbb{Q}[\vec{u}]$ and $\mathbb{Q}_i[\vec{u}]$. Using the short-hand $\mathbb{D}_i[\,\cdot\,] := \mathrm{D}\left(\mathbb{Q}[\,\cdot\,] \,\|\, \mathbb{Q}_i[\,\cdot\,]\right)$,

$$\begin{aligned}
\mathrm{D}\left(\mathbb{Q}\left[\vec{u}\right] \| \mathbb{Q}_i\left[\vec{u}\right]\right) &= \mathbb{D}_i[\vec{u}] \\
&= \mathbb{D}_i\left[u_1\right] + \mathbb{D}_i\left[u_2, \dots, u_{\mathrm{T}} \,|\, u_1\right] \\
&= \mathbb{D}_i\left[u_1\right] + \mathbb{D}_i\left[u_2 \,|\, u_1\right] + \mathbb{D}_i\left[u_3, \dots, u_{\mathrm{T}} \,|\, u_1, u_2\right] \\
&= \sum_{t=1}^{\mathrm{T}} \mathbb{D}_i\left[u_t \,|\, u_1, \dots, u_{t-1}\right].
\end{aligned}$$

We now use the definition of the conditional relative entropy, and make the previously discussed Bernoulli distributions appear. For $1 \leqslant t \leqslant \mathrm{T}$,

$$\begin{aligned}
\mathbb{D}_i\left[u_t \,|\, u_1, \dots, u_{t-1}\right] &= \sum_{u_1,\dots,u_{t-1}} \mathbb{Q}\left[u_1, \dots, u_{t-1}\right] \\
&\qquad \times \sum_{u_t} \mathbb{Q}\left[u_t \,|\, u_1, \dots, u_{t-1}\right] \log \frac{\mathbb{Q}\left[u_t \,|\, u_1, \dots, u_{t-1}\right]}{\mathbb{Q}_i\left[u_t \,|\, u_1, \dots, u_{t-1}\right]} \\
&= \frac{1}{2^{t-1}} \sum_{u_1,\dots,u_{t-1}} \sum_{u_t} \mathbb{Q}\left[u_t \,|\, u_1, \dots, u_{t-1}\right] \log \frac{\mathbb{Q}\left[u_t \,|\, u_1, \dots, u_{t-1}\right]}{\mathbb{Q}_i\left[u_t \,|\, u_1, \dots, u_{t-1}\right]} \\
&= \frac{1}{2^{t-1}} \sum_{(u_1,\dots,u_{t-1}) \in \mathrm{A}_t^{(i)}} \mathrm{D}\left(\mathrm{B}\left(1, \frac{s}{2d}\right) \,\middle\|\, \mathrm{B}\left(1, \frac{s}{2d} - \varepsilon\right)\right)
\end{aligned}$$

$$+\frac{1}{2^{t-1}}\sum_{(u_1,\dots,u_{t-1})\in \mathrm{B}_t^{(i)}} \mathrm{D}\left(\mathrm{B}\left(1,\frac{s}{2d}\right)\middle\|\,\mathrm{B}\left(1,\frac{s}{2d}\right)\right)$$

$$=\frac{1}{2^{t-1}}\sum_{(u_1,\dots,u_{t-1})\in \mathrm{A}_t^{(i)}} \mathbb{B}\left(\frac{s}{2d},\varepsilon\right),$$

where we used the short-hand $\mathbb{B}\left(\frac{s}{2d},\varepsilon\right) := \mathrm{D}\left(\mathrm{B}\left(1,\frac{s}{2d}\right)\|\,\mathrm{B}\left(1,\frac{s}{2d}-\varepsilon\right)\right)$. Eventually:

$$\mathrm{D}\left(\mathbb{Q}[\vec{u}]\,\|\,\mathbb{Q}_i[\vec{u}]\right) = \mathbb{B}\left(\frac{m}{2d},\varepsilon\right)\sum_{t=1}^{\mathrm{T}}\frac{\left|\mathrm{A}_t^{(i)}\right|}{2^{t-1}}.$$

Upper bound on $\frac{1}{d}\sum_{i=1}^{d}\mathbb{E}_i\left[\tau_i(\mathrm{T})\right]$ using Pinsker's inequality. — In this step, we will make use of Pinsker's inequality to make the relative entropy appear.

Proposition V.5.4 (Pinsker's inequality). *Let* X *be a finite set, and* P, Q *probability distributions on* X. *Then,*

$$\frac{1}{2}\sum_{x\in\mathrm{X}}|\mathrm{P}(x)-\mathrm{Q}(x)| \leqslant \sqrt{\frac{1}{2}\mathrm{D}\left(\mathrm{P}\,\|\,\mathrm{Q}\right)}.$$

Immediate consequence:

$$\sum_{\substack{x\in\mathrm{X}\\ \mathrm{P}(x)>\mathrm{Q}(x)}}(\mathrm{P}(x)-\mathrm{Q}(x)) \leqslant \sqrt{\frac{1}{2}\mathrm{D}\left(\mathrm{P}\,\|\,\mathrm{Q}\right)}.$$

Let $i\in[d]$. If $(u_1,\dots,u_{\mathrm{T}})\in\{0,1\}^{\mathrm{T}}$ is given, since the decisions d_t and d'_t are determined by the previous losses $\ell_t^{(d_t)}$ and $(\ell'_t)^{(d'_t)}$ respectively, we have in particular:

$$\mathbb{E}_i\left[\tau_i(\mathrm{T})\,\middle|\,\ell_1^{(d_1)}=u_1,\dots,\ell_{\mathrm{T}}^{(d_{\mathrm{T}})}=u_{\mathrm{T}}\right] = \mathbb{E}\left[\tau'_i(\mathrm{T})\,\middle|\,(\ell'_1)^{(d'_1)}=u_1,\dots,(\ell'_{\mathrm{T}})^{(d'_{\mathrm{T}})}=u_{\mathrm{T}}\right].$$

Therefore,

$$\begin{aligned}\mathbb{E}_i\left[\tau_i(\mathrm{T})\right]-\mathbb{E}\left[\tau'_i(\mathrm{T})\right] &= \sum_{\vec{u}}\mathbb{Q}_i[\vec{u}]\cdot\mathbb{E}_i\left[\tau_i(\mathrm{T})\,\middle|\,\forall t,\ \ell_t^{(d_t)}=u_t\right]\\ &\qquad -\sum_{\vec{u}}\mathbb{Q}[\vec{u}]\cdot\mathbb{E}\left[\tau'_i(\mathrm{T})\,\middle|\,\forall t,\ (\ell'_t)^{d'_t}=u_t\right]\\ &= \sum_{\vec{u}}\left(\mathbb{Q}_i[\vec{u}]-\mathbb{Q}[\vec{u}]\right)\mathbb{E}_i\left[\tau_i(\mathrm{T})\,\middle|\,\forall t,\ \ell_t^{(d_t)}=u_t\right]\end{aligned}$$

$$
\begin{aligned}
&\leqslant \sum_{\substack{\vec{u} \\ \mathbb{Q}_i[\vec{u}] > \mathbb{Q}[\vec{u}]}} \left(\mathbb{Q}_i[\vec{u}] - \mathbb{Q}[\vec{u}]\right) \mathbb{E}_i\left[\tau_i(\mathrm{T}) \,\middle|\, \forall t,\ \ell_t^{(d_t)} = u_t\right] \\
&\leqslant \mathrm{T} \sum_{\substack{\vec{u} \\ \mathbb{Q}_i[\vec{u}] > \mathbb{Q}[\vec{u}]}} \left(\mathbb{Q}_i[\vec{u}] - \mathbb{Q}[\vec{u}]\right) \\
&\leqslant \mathrm{T}\sqrt{\frac{1}{2}\mathrm{D}\left(\mathbb{Q}[\vec{u}] \,\|\, \mathbb{Q}_i[\vec{u}]\right)} \\
&= \mathrm{T}\sqrt{\frac{\mathbb{B}(s/2d, \varepsilon)}{2}} \sqrt{\sum_{t=1}^{\mathrm{T}} \frac{\left|\mathrm{A}_t^{(i)}\right|}{2^{t-1}}},
\end{aligned}
$$

where we used Pinsker's inequality in the fifth line. Moreover, we have:

$$
\frac{1}{d}\sum_{i=1}^{d} \mathbb{E}\left[\tau_i'(\mathrm{T})\right] = \frac{1}{d}\mathbb{E}\left[\sum_{t=1}^{\mathrm{T}}\sum_{i=1}^{d} \mathbb{1}_{\{d_t'=i\}}\right] = \frac{1}{d}\mathbb{E}\left[\sum_{t=1}^{\mathrm{T}} 1\right] = \frac{\mathrm{T}}{d}.
$$

Combining this with the previous inequality gives:

$$
\begin{aligned}
\frac{1}{d}\sum_{i=1}^{d} \mathbb{E}_i\left[\tau_i(\mathrm{T})\right] &\leqslant \frac{1}{d}\sum_{i=1}^{d} \mathbb{E}\left[\tau_i'(\mathrm{T})\right] + \mathrm{T}\sqrt{\frac{\mathbb{B}(s/2d, \varepsilon)}{2}} \frac{1}{d}\sum_{i=1}^{d} \sqrt{\sum_{t=1}^{\mathrm{T}} \frac{\left|\mathrm{A}_t^{(i)}\right|}{2^{t-1}}} \\
&\leqslant \frac{\mathrm{T}}{d} + \mathrm{T}\sqrt{\frac{\mathbb{B}(s/2d, \varepsilon)}{2}} \sqrt{\frac{1}{d}\sum_{t=1}^{\mathrm{T}}\sum_{i=1}^{d} \frac{\left|\mathrm{A}_t^{(i)}\right|}{2^{t-1}}} \\
&= \frac{\mathrm{T}}{d} + \mathrm{T}\sqrt{\frac{\mathbb{B}(s/2d, \varepsilon)}{2}} \sqrt{\frac{1}{d}\sum_{t=1}^{\mathrm{T}} \frac{\left|\{0,1\}^{t-1}\right|}{2^{t-1}}} \\
&= \frac{\mathrm{T}}{d} + \mathrm{T}\sqrt{\frac{\mathbb{B}(s/2d, \varepsilon)}{2}} \sqrt{\frac{\mathrm{T}}{d}} \\
&= \frac{\mathrm{T}}{d} + \mathrm{T}^{3/2}\sqrt{\frac{\mathbb{B}(s/2d, \varepsilon)}{2d}}.
\end{aligned}
$$

where we used Jensen for the second inequality, and for the third line, we remembered that $(\mathrm{A}_t^{(i)})_{i\in[d]}$ is a partition of $\{0,1\}^{t-1}$.

An upper bound on $\mathbb{B}(s/2d, \varepsilon)$ for small enough ε. — We first write $\mathbb{B}(s/2d, \varepsilon)$ explicitely.

$$\begin{aligned}\mathbb{B}\left(\frac{s}{2d}, \varepsilon\right) &= \mathrm{D}\left(\mathrm{B}(1, s/2d) \,\|\, \mathrm{B}(1, s/2d - \varepsilon)\right) \\ &= \frac{s}{2d}\log\frac{s/2d}{s/2d-\varepsilon} + \left(1-\frac{s}{2d}\right)\log\frac{1-s/2d}{1-s/2d+\varepsilon} \\ &= -\frac{s}{2d}\log\left(1-\frac{2d\varepsilon}{s}\right) + \left(\frac{s}{2d}-1\right)\log\left(1+\frac{\varepsilon}{1-m/2d}\right).\end{aligned}$$

We now bound the two logarithms from above using respectively the two following easy inequalities:

$$\begin{aligned}-\log(1-x) &\leqslant x + x^2, \quad \text{for } x \in [0, 1/2] \\ -\log(1+x) &\leqslant -x + x^2, \quad \text{for } x \geqslant 0.\end{aligned}$$

This gives:

$$\begin{aligned}\mathbb{B}\left(\frac{s}{2d}, \varepsilon\right) &\leqslant \frac{s}{2d}\left(\frac{2d\varepsilon}{s} + \frac{4d^2\varepsilon^2}{s^2}\right) + \left(1-\frac{s}{2d}\right)\left(-\frac{\varepsilon}{1-s/2d} + \frac{\varepsilon^2}{(1-s/2d)^2}\right) \\ &= \frac{4d^2\varepsilon^2}{s(2d-s)},\end{aligned}$$

which holds for $2d\varepsilon/s \leqslant 1/2$, in other words, for $\varepsilon \leqslant s/4d$.

Lower bound on the expectation of the regret of σ against ℓ_t. — We can now bound from below the expected regret incurred when playing σ against loss vectors $(\ell_t)_{t\geqslant 1}$. For $\varepsilon \leqslant s/4d$,

$$\begin{aligned}\mathrm{R_T} &= \mathbb{E}\left[\sum_{t=1}^{\mathrm{T}} \ell_t^{(d_t)} - \min_{j\in[d]} \sum_{t=1}^{\mathrm{T}} \ell_t^{(j)}\right] \\ &= \frac{1}{d}\sum_{i=1}^{d} \mathbb{E}_i\left[\sum_{t=1}^{\mathrm{T}} \ell_t^{(d_t)} - \min_{j\in[d]} \sum_{t=1}^{\mathrm{T}} \ell_t^{(j)}\right] \\ &\geqslant \frac{1}{d}\sum_{i=1}^{d}\left(\mathbb{E}_i\left[\sum_{t=1}^{\mathrm{T}} \ell_t^{(d_t)}\right] - \min_{j\in[d]} \sum_{t=1}^{\mathrm{T}} \mathbb{E}_i\left[\ell_t^{(j)}\right]\right) \\ &= \frac{1}{d}\sum_{i=1}^{d}\left(\mathbb{E}_i\left[\sum_{t=1}^{\mathrm{T}} \mathbb{E}_i\left[\ell_t^{(d_t)} \,\middle|\, d_t\right]\right] - \mathrm{T}\min_{j\in[d]}\left(\frac{s}{2d} - \varepsilon\delta_{ij}\right)\right) \\ &= \frac{1}{d}\sum_{i=1}^{d}\left(\mathbb{E}_i\left[\sum_{t=1}^{\mathrm{T}}\left(\frac{s}{2d} - \varepsilon\delta_{id_t}\right)\right] - \mathrm{T}\left(\frac{s}{2d} - \varepsilon\right)\right)\end{aligned}$$

$$= \frac{1}{d}\sum_{i=1}^{d} \varepsilon \left(\mathrm{T} - \mathbb{E}_i\left[\tau_i(\mathrm{T})\right]\right)$$
$$= \varepsilon\left(\mathrm{T} - \frac{1}{d}\sum_{i} \mathbb{E}_i\left[\tau_i(\mathrm{T})\right]\right).$$

We now use the upper bound derived in Section V.5.3.

$$\begin{aligned}
\mathrm{R}_{\mathrm{T}} &\geqslant \varepsilon\left(\mathrm{T} - \frac{\mathrm{T}}{d} - \mathrm{T}^{3/2}\sqrt{\frac{\mathbb{B}(s/2d,\varepsilon)}{2d}}\right)\\
&\geqslant \varepsilon\left(\mathrm{T} - \frac{\mathrm{T}}{d} - \mathrm{T}^{3/2}\varepsilon\sqrt{\frac{2d}{s(2d-s)}}\right)\\
&\geqslant \varepsilon\left(\mathrm{T} - \frac{\mathrm{T}}{d} - 2\mathrm{T}^{3/2}\varepsilon\frac{1}{\sqrt{s}}.\right),
\end{aligned}$$

where in the penultimate, we used the upper bound on $\mathbb{B}(s/2d, \varepsilon)$ that we established above, and in the last line, the fact that $s \leqslant d$. Let $\mathrm{C} > 0$ and we choose $\varepsilon = \mathrm{C}\sqrt{s/\mathrm{T}}$. Then, for $\varepsilon \leqslant s/4d$,

$$\begin{aligned}
\mathrm{R}_{\mathrm{T}} &\geqslant \varepsilon\mathrm{T}\left(1 - \frac{1}{d} - 2\varepsilon\sqrt{\frac{\mathrm{T}}{s}}\right)\\
&= \mathrm{C}\sqrt{s\mathrm{T}}\left(1 - \frac{1}{d}\right) - 2\sqrt{s\mathrm{T}}\mathrm{C}^2\\
&\geqslant \sqrt{s\mathrm{T}}\left(\frac{\mathrm{C}}{2} - 2\mathrm{C}^2\right),
\end{aligned}$$

where in the last line, we used the assumption $d \geqslant 2$. The choice $\mathrm{C} = 1/8$ give:

$$\mathrm{R}_{\mathrm{T}} \geqslant \frac{1}{32}\sqrt{s\mathrm{T}},$$

which holds for $\varepsilon = \mathrm{C}\sqrt{s/\mathrm{T}} \leqslant s/4d$ i.e. for $\mathrm{T} \geqslant d^2/4s$.

The above inequality does not depend on σ. As it is a classic that a randomized strategy is equivalent to some random choice of deterministic strategies, this lower bound holds for any strategy of the player. In other words, for $\mathrm{T} \geqslant d^2/4s$,

$$\hat{v}_{\mathrm{T}}^{\ell,s,d} \geqslant \frac{1}{32}\sqrt{s\mathrm{T}}.$$

□

V.5.4. Discussion

If the outcomes are not losses but gains, then there is an important discrepancy between the upper and lower bounds we obtain. Indeed, obtaining small losses regret bound as in the first displayed equation of the proof of Theorem V.5.1 is still open. An idea for circumventing this issue would be to enforce exploration by perturbing x_t into $(1-\gamma)x_t + \gamma\mathcal{U}$ where $\mathcal{U}$ is the uniform distribution over $[d]$, but usual computations show that the only obtainable upper bounds are of order of $\sqrt{d\mathrm{T}}$. The aforementioned techniques used to bound the regret from below with losses would also work with gains, which would give a lower bound of order $\sqrt{s\mathrm{T}}$. Therefore, finding the optimal dependency in the dimension and/or the sparsity level is still an open question in that specific case.

CHAPTER VI

APPROACHABILITY WITH PARTIAL MONITORING

This chapter is extracted from the paper *Blackwell approachability with partial monitoring: Optimal convergence rates*, in collaboration with Vianney Perchet, in preparation.

Abstract

We study the approachability problem with partial monitoring and polytope target sets. When the target set is approachable, we construct, for the first time, approaching strategies with convergence rate of order $O(T^{-1/2})$ in the case of outcome-dependentoutcom signals and of order $O(T^{-1/3})$ in the case of general signals. Those rates are known to be unimprovable without further assumption on the target set or the signalling structure. It therefore establishes the optimal convergence rates for those two cases. Moreover, the proposed strategies are computationally efficient.

VI.1. Introduction

We study the following approachability problem with partial monitoring. The Decision Maker and Nature both have a finite set of pure actions. At each stage, the Decision Maker and Nature choose an action in their respective action sets, possibly at random. This determines a vector-valued payoff which is not observed by the Decision Maker. The latter only observes a random signal whose law depends on the pure actions of the Decision Maker and Nature. The Decision Maker is aiming at having the average payoff to converge to a given target set.

VI.1.1. Previous work

In the full information setting, both the regret minimization and approachability problems have a worst-case convergence of rate of order $O(T^{-1/2})$. The rate deals respectively with the average regret in regret minimization, and the distance of the average payoff to the target set in approachability.

In regret minimization with partial monitoring, depending on the signalling structure, the Decision Maker may or may not be able to guarantee an asymptotically non-positive average regret. This has given rise to two main directions of research.

The first one, initiated by [PS01] identifies the signalling structures which allow the average regret to be minimized and aims at constructing strategies in those cases: [PS01] constructed a strategy which guarantees a convergence rate of order $O(T^{-1/4})$ and [CBLS06] proposed an improved strategy with a $O(T^{-1/3})$ guarantee as well as a general lower bound of the same order. Later, [BPS10, BFP+14] gave a classification of signalling structures according to convergence rates: they established that the optimal convergence rate is either $O(T^{-1/2})$, $O(T^{-1/3})$ or $O(1)$—this last rate corresponds to the case where the average regret cannot be minimized.

The second line of research was proposed by [Rus99] who introduced a weaker variant of the regret, which involves the best performance that the Decision Maker could have achieved in hindsight (had he known the sequence of signal laws, but not the sequence of actions of Nature), for a given signalling structure. [Rus99] however did not provide an explicit strategy nor convergence rates. [MS03] constructed approachability-based algorithms in the special case where the law of the signal only depends on Nature's action (the so-called *outcome-dependentoutcom* case). [LMS08] proposed strategies with convergence rates of order $O(T^{-1/4}\sqrt{\log T})$ in the case of outcome-dependentoutcom signals and of order $O(T^{-1/5}\sqrt{\log T})$ in the case of general signals. The optimal rate of order $O(T^{-1/3})$ in the case of general signals (for both internal and external regret) was achieved by [Per11b] using calibration-based algorithms.

More recently, the problem of approachability with partial monitoring has been introduced by [Per11a]. The regret minimization problem from [Rus99] and the internal regret from [LS07, Per11b] turn out to be special cases of this very general framework. However, the convergence rate of the strategy provided in [Per11a] had the drawback of deteriorating quickly with the dimension of the payoff space. A strategy with dimension-free rate of order $O(T^{-1/5})$ was given in [MPS14]—see also [MPS13]. However, the optimal rate of convergence was conjectured to be of order $O(T^{-1/3})$, like for regret minimzation.

VI.1.2. Main contributions

We construct, for the first time, approachability strategies for polytope target sets with convergence rates of order $O(T^{-1/3})$ in the case of general signals and of order $O(T^{-1/2})$ in the case of outcome-dependentoutcom signals. Those rates are known to be unimprovable without further assumption on the target set or the signalling structure: in the case of general signals, a lower bound of order $O(T^{-1/3})$ was given in [CBLS06], and the $O(T^{-1/2})$ rate is already optimal in the full information setting. It therefore establishes the optimal convergence rates for those two cases. Moreover, the proposed strategies are computationally efficient.

VI.1.3. Outline

In Section VI.2, we present the model of two-player game with vector payoffs and with partial monitoring. In Section VI.3, we recall the dual characterizations of approachability, both in partial monitoring and in full information. In Section VI.4, we first construct an auxiliary full information game which we then use to define the strategy for the initial game. The efficiency of the strategy is discussed. In Section VI.5 we state and prove Theorem VI.5.1 which is our main result. It establishes an $O(T^{-1/3})$ rate of convergence for the strategy. In Section VI.6, we deal with the special case of outcome-dependentoutcom signals for which we propose a modified strategy which is proved in Theorem VI.6.2 to have an $O(T^{-1/2})$ rate of convergence.

VI.1.4. Notation

Exponents will be used to denote the components of a vector: for instance $x = (x^i)_{i\in\mathcal{I}} \in \mathbb{R}^{\mathcal{I}}$. Bold letters will denote maps and calligraphic letters will denote sets. $\langle \cdot \,|\, \cdot \rangle$ will denote the scalar product.

VI.2. The game

VI.2.1. Ingredients

We consider a repeated two-player game with vector-valued payoffs and partial monitoring between the *Decision Maker* and *Nature*. The Decision Maker (resp. Nature) has a finite set of pure actions $\mathcal{I}$ (resp. $\mathcal{J}$). Denote by

$$\Delta(\mathcal{I}) := \left\{ x = (x^i)_{i\in\mathcal{I}} \in \mathbb{R}_+^{\mathcal{I}} \,\middle|\, \sum_{i\in\mathcal{I}} x^i = 1 \right\}$$

the simplex which represents the set of probability distributions over $\mathcal{I}$. $\Delta(\mathcal{J})$ is defined similarly. Let $\mathbf{g} : \mathcal{I} \times \mathcal{J} \to \mathbb{R}^d$ be the vector-valued payoff function which we

bilinearly extend to $\mathbf{g} : \Delta(\mathcal{I}) \times \Delta(\mathcal{J}) \to \mathbb{R}^d$:

$$\mathbf{g}(x, y) := \mathbb{E}_{\substack{i\sim x \\ j\sim y}} [\mathbf{g}(i, j)] = \sum_{\substack{i\in\mathcal{I} \\ j\in\mathcal{J}}} x^i y^j \mathbf{g}(i, j)$$

$$\text{where} \quad x = (x^i)_{i\in\mathcal{I}} \in \Delta(\mathrm{I}) \quad \text{and} \quad y = (y^j)_{j\in\mathcal{J}} \in \Delta(\mathcal{J}).$$

Denote by $\|\mathbf{g}\|_2 := \max_{\substack{i\in\mathcal{I} \\ j\in\mathcal{J}}} \|\mathbf{g}(i, j)\|_2$ its Euclidean norm. Let $\mathcal{S}$ be a finite set of *signals* and $\mathbf{s} : \mathcal{I} \times \mathcal{J} \to \Delta(\mathcal{S})$ the signal distribution function, which we also bilinearly extend to $\Delta(\mathcal{I}) \times \Delta(\mathcal{J})$. All the above elements are assumed to be known to the Decision Maker. The special case where the law of the signal $\mathbf{s}(i, j)$ does not depend on i is called the *outcome-dependentoutcom signals* case, and will be treated in Section VI.6.

VI.2.2. The play

The game is played as follows. At time $t \geqslant 1$,

- the Decision Maker and Nature simultaneously choose pure actions $i_t \in \mathcal{I}$ and $j_t \in \mathcal{J}$, possibly at random according to mixed actions $x_t \in \Delta(\mathcal{I})$ and $y_t \in \Delta(\mathcal{J})$;
- the Decision Maker gets (but does not observe) vector payoff $g_t := \mathbf{g}(i_t, j_t) \in \mathbb{R}^d$;
- the Decision Maker observes signal $s_t \in \mathcal{S}$ which is drawn according to $\mathbf{s}(i_t, j_t) \in \Delta(\mathcal{S})$.

Formally, a strategy for the Decision Maker is a sequence of measurable maps $\sigma = (\sigma_t)_{t\geqslant 1}$ where $\sigma_t : (\Delta(\mathcal{I}) \times \mathcal{I} \times \mathcal{S})^{t-1} \to \Delta(\mathcal{I})$ indicates the mixed action x_t at time t as a function of the information available to the Decision Maker. In other words:

$$x_t = \sigma_t(x_1, i_1, s_1, \dots, x_{t-1}, i_{t-1}, s_{t-1}).$$

Similarly, a strategy for Nature is a sequence $(\tau_t)_{t\geqslant 1}$ where $\tau_t : (\Delta(\mathcal{I}) \times \mathcal{I} \times \mathcal{S} \times \Delta(\mathcal{J}) \times \mathcal{J})^{t-1} \to \Delta(\mathcal{J})$, so that

$$y_t = \tau_t(x_1, i_1, s_1, y_1, j_1, \dots, x_{t-1}, i_{t-1}, s_{t-1}, y_{t-1}, j_{t-1}).$$

For $\mathrm{T} \geqslant 1$, denote $\bar{g}_{\mathrm{T}} := \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} g_t$ the average vector payoff up to time T.

VI.2.3. Flags

The flag function $\mathbf{f} : \Delta(\mathcal{J}) \to \Delta(\mathcal{S})^{\mathcal{I}}$ is defined by

$$\mathbf{f}(y) = (\mathbf{s}(i, y))_{i\in\mathcal{I}}, \quad y \in \Delta(\mathcal{J}).$$

For $t \geqslant 1$, denote $f_t := \mathbf{f}(y_t)$ the flag associated with y_t. Denote $\mathscr{F} = \mathbf{f}(\Delta(\mathscr{J}))$ the set of all possible flags, which is a polytopial subset of $\mathbb{R}^{\mathscr{S}\times\mathscr{J}}$. The notion of flags is fundamental in games with partial monitoring. Although the Decision Maker does not directly observe it, he can, as will be shown, estimate it. As a matter of fact, it is the maximal information available to him. For $x \in \Delta(\mathscr{I})$ and $f \in \mathscr{F}$, let $\mathbf{m}(x, f) := \mathbf{g}(x, \mathbf{f}^{-1}(f))$ be the set of all payoffs that are compatible with mixed action x and flag f. The set-valued map $\mathbf{m} : \Delta(\mathscr{I}) \times \mathscr{F} \rightrightarrows \mathbb{R}^d$ will be essential in the statement of the characterization of approachable sets (Proposition VI.3.2) and in the construction of the strategies.

VI.3. Approachability

We recall the definition of approachability and the characterizations of approachable convex sets both in the partial monitoring and full information cases.

Definition VI.3.1. A closed convex $\mathscr{C} \subset \mathbb{R}^d$ is *approachable* if there exists a strategy of the Decision Maker which guarantees

$$\mathbb{E}\left[\mathbf{d}_2\left(\bar{g}_T,\ \mathscr{C}\right)\right] \xrightarrow[T\to+\infty]{} 0,$$

uniformly in the strategy τ of Nature, where $\mathbf{d}_2(\,\cdot\,,\ \mathscr{C})$ denotes the Euclidean distance to $\mathscr{C}$, and where the expectation corresponds to the randomization introduced by the strategies and the signals.

Proposition VI.3.2 (Characterization of approachable convex sets in games with partial monitoring [Per11a]). *A closed convex set $\mathscr{C} \subset \mathbb{R}^d$ is approachable if and only if*

$$\forall f \in \mathscr{F},\ \exists x \in \Delta(\mathscr{I}),\quad \mathbf{m}(x, f) \subset \mathscr{C}.$$

The construction of our strategies in Section VI.4 will involve an auxiliary full information game. We quickly review the characterizations of approachability in full information games with convex compact action sets and bilinear payoff functions. Let $\mathscr{X}$ and $\mathscr{Y}$ be convex compact action sets and $\mathbf{g} : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}^d$ a bilinear payoff function. The special case of target sets which are closed convex cones will be of particular importance in the subsequent sections. A few facts about closed convex cones are gathered in Section IV.2.

Proposition VI.3.3 (Characterization of approachability in full information games). *A closed convex set $\mathscr{C} \subset \mathbb{R}^d$ is approachable if and only if one of the following properties hold.*

(i) $\forall g \in \mathbb{R}^d,\ \exists x \in \mathscr{X},\ \forall y \in \mathscr{Y},\quad \langle \mathbf{g}(x, y) - \mathbf{P}_{\mathscr{C}}(g) | g - \mathbf{P}_{\mathscr{C}}(g)\rangle \leqslant 0$, *where $\mathbf{P}_{\mathscr{C}}$ denotes the Euclidean projection on $\mathscr{C}$;*

(ii) $\forall y \in \mathcal{Y},\ \exists x \in \mathcal{X},\quad \mathbf{g}(x, y) \in \mathcal{C}.$

Moreover, if $\mathcal{C}$ is a closed convex cone, the above is also equivalent to

(iii) $\forall z \in \mathcal{C}^\circ,\ \exists x \in \mathcal{X},\ \forall y \in \mathcal{Y},\quad \langle \mathbf{g}(x, y)|z\rangle \leqslant 0.$

Proof. The first two characterizations are classic [Bla56]. Let us assume that $\mathcal{C}$ is a closed convex cone. Let us prove that (ii) and (iii) are equivalent. $\mathcal{C}$ being a closed convex cone, $\mathbf{g}(x, y) \in \mathcal{C}$ is equivalent to $\max_{z\in\mathcal{C}^\circ} \langle \mathbf{g}(x, y)|z\rangle \leqslant 0$. Then, (ii) can be rewritten

$$\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} \max_{z\in\mathcal{C}^\circ} \langle \mathbf{g}(x, y)|z\rangle \leqslant 0.$$

$\mathcal{X}$ being compact and the quantity $\langle \mathbf{g}(x, y)|z\rangle$ being linear in x, y and z, we can apply Sion's minimax theorem twice to get

$$\max_{z\in\mathcal{C}^\circ} \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \langle \mathbf{g}(x, y)|z\rangle \leqslant 0,$$

which is exactly (iii). □

VI.4. Construction of the strategy

We study the case where the target set is the negative orthant $\mathbb{R}^d_-$ and we assume it to be approachable. Since a polytope can be represented as an orthant in a higher dimension space, the extension to polytope target sets can be easily carried out as in e.g. [MPS14, Section 5.4.2]. Most of the proofs are postponed to Section VI.8.

VI.4.1. Bi-piecewise affinity

We aim in this section at constructing a vector-valued map $\mathbf{r} : \Delta(\mathcal{I}) \times \mathcal{F} \to \mathbb{R}^d$ which can be seen as a simplified version of the set-valued map $\mathbf{m} : \Delta(\mathcal{I}) \times \mathcal{F} \rightrightarrows \mathbb{R}^d$. Its properties will be gathered at the end of the section in Proposition VI.4.4.

Definition VI.4.1. Let $\mathcal{U}$ be a convex set and $\mathcal{V}$ a vector space. Let $\mathbf{a} : \mathcal{U} \rightrightarrows \mathcal{V}$ be a set-valued function. $\mathbf{a}$ is *affine* if for all $u, u' \in \mathcal{U}$ and $\lambda \in [0, 1]$,

$$\mathbf{a}(\lambda u + (1-\lambda)u') = \lambda \mathbf{a}(u) + (1-\lambda)\mathbf{a}(u').$$

The map $\mathbf{f}$ being affine on $\mathcal{F}$ by definition, [RZ96, Proposition 2.4] guarantees the existence of a polytopial decomposition of $\mathcal{F}$ such that $\mathbf{f}^{-1}$ is affine on each of those polytopes. The decomposition can then be refined so that each point of $\mathcal{F}$ can be written as a unique convex combination of the vertices of the polytope to which it belongs. This is formalized by the following lemma.

Lemma VI.4.2. *There exists a finite family $(\mathcal{F}^k)_{k\in\mathcal{K}}$ of polytopes (denote $\mathcal{B}^k$ the set of vertices of $\mathcal{F}^k$ and $\mathcal{B} = \bigcup_{k\in\mathcal{K}} \mathcal{B}^k$) such that*

(i) $\mathscr{F} = \bigcup_{k\in\mathscr{K}} \mathscr{F}^k$;

(ii) for each $k \in \mathscr{K}$, $\mathbf{f}^{-1}$ *is affine on* $\mathscr{F}^k$;

(iii) for all $f \in \mathscr{F}$, *there exists a unique* $\mu = (\mu^b)_{b\in\mathscr{B}} \in \Delta(\mathscr{B})$ *such that*

(a) $f = \sum_{b\in\mathscr{B}} \mu^b \cdot b$;

(b) for $k \in \mathscr{K}$, $f \in \mathscr{F}^k \Longrightarrow \operatorname{supp}\mu \subset \mathscr{B}^k$.

From now on, we assume given such a decomposition.

We are going to construct the map $\mathbf{r} = (\mathbf{r}^n)_{1\leqslant n\leqslant d}$ component by component, and first on $\Delta(\mathscr{T})\times\mathscr{B}$ before extending it to $\Delta(\mathscr{T})\times\mathscr{F}$. Denote $(\mathbf{g}^n)_{1\leqslant n\leqslant d}$ the components of $\mathbf{g}$. For $x \in \Delta(\mathscr{T})$ and $b \in \mathscr{B}$, we set $\mathbf{r}^n(x, b)$ as being the maximum real number of the set $\mathbf{g}^n(x, \mathbf{f}^{-1}(b))$:

$$\mathbf{r}^n(x, b) := \max \mathbf{g}^n(x, \mathbf{f}^{-1}(b)). \tag{VI.1}$$

We then extend $\mathbf{r}$ to $\Delta(\mathscr{T}) \times \mathscr{F}$ as follows. Using property (iii) from Lemma VI.4.2, a given flag $f \in \mathscr{F}$ can be uniquely written

$$f = \sum_{b\in\mathscr{B}} \mu^b \cdot b,$$

with $\operatorname{supp}\mu$ contained in one of the polytopes $\mathscr{F}^k$. We then use the above coefficients $(\mu^b)_{b\in\mathscr{B}}$ to define

$$\mathbf{r}^n(x, f) := \sum_{b\in\mathscr{B}} \mu^b \cdot \mathbf{r}^n(x, b). \tag{VI.2}$$

This construction will lead to piecewise affinity of $\mathbf{r}(x, f)$ in f – see Proposition VI.4.4 below. We now turn to the piecewise affinity in x.

Lemma VI.4.3. *There exists a finite family of polytopes* $(\mathscr{X}^\ell)_{\ell\in\mathscr{L}}$ *such that*

(i) $\Delta(\mathscr{T}) = \bigcup_{\ell\in\mathscr{L}} \mathscr{X}^\ell$;

(ii) For each $\ell \in \mathscr{L}$ *and* $f \in \mathscr{F}$, $\mathbf{r}(\,\cdot\,, f)$ *is affine on* $\mathscr{X}^\ell$.

Let $\mathscr{A}$ be the set of the vertices of the polytopes $\mathscr{X}^\ell$ given the above lemma. The following proposition summarizes some properties of $\mathbf{r}$.

Proposition VI.4.4. *(i) For all* $x \in \Delta(\mathscr{T})$, $y \in \Delta(\mathscr{J})$ *and* $1 \leqslant n \leqslant d$, *we have* $\mathbf{g}^n(x, y) \leqslant \mathbf{r}^n(x, \mathbf{f}(y))$;

(ii) For all $f \in \mathscr{F}$, *there exists* $x \in \Delta(\mathscr{T})$ *such that* $\mathbf{r}(x, f) \in \mathbb{R}^d_-$;

(iii) For all $x \in \Delta(\mathscr{T})$, $\mathbf{r}(x, \,\cdot\,)$ *is affine on each* $\mathscr{F}^k$ $(k \in \mathscr{K})$;

(iv) For all $f \in \mathscr{F}$, $\mathbf{r}(\,\cdot\,, f)$ *is affine on each* $\mathscr{X}^\ell$ $(\ell \in \mathscr{L})$.

VI.4.2. From bi-piecewise affinity to linearity

In Section VI.4.1, we constructed a map $\mathbf{r} : \Delta(\mathcal{I}) \times \mathcal{F} \to \mathbb{R}^d$ which is bi-piecewise affine. In this section, we aim at constructing a *linear map* $\mathbf{R} : (\mathbb{R}^{\mathcal{S}\times\mathcal{I}})^{\mathcal{K}\times\mathcal{A}} \to \mathbb{R}^d$ which encodes the map $\mathbf{r}$ in the following sense. From all pairs $(x, f) \in \Delta(\mathcal{I}) \times \mathcal{F}$, there is a simple construction of a vector $\tilde{g} \in (\mathbb{R}^{\mathcal{S}\times\mathcal{I}})^{\mathcal{K}\times\mathcal{A}}$ such that $\mathbf{R}(\tilde{g}) = \mathbf{r}(x, f)$.

Lemma VI.4.5. *For every $k \in \mathcal{K}$, there exists a map $\mathbf{r}^{[k]} : \Delta(\mathcal{I}) \times \mathbb{R}^{\mathcal{S}\times\mathcal{I}} \to \mathbb{R}^d$ such that*

(i) for all $x \in \Delta(\mathcal{I})$, the map $\mathbf{r}^{[k]}(x, \cdot) : \mathbb{R}^{\mathcal{S}\times\mathcal{I}} \to \mathbb{R}^d$ is linear;

(ii) for all $x \in \Delta(\mathcal{I})$ and $f \in \mathcal{F}^k$, $\mathbf{r}^{[k]}(x, f) = \mathbf{r}(x, f)$.

Define $\mathrm{L}_{\mathbf{r}}$ as the maximal operator norm of the linear maps $\mathbf{r}^{[k]}(a, \cdot)$:

$$\mathrm{L}_{\mathbf{r}} := \max_{\substack{k\in\mathcal{K}\\ a\in\mathcal{A}}} \max_{\substack{f\in\mathbb{R}^{\mathcal{S}\times\mathcal{I}}\\ f\neq 0}} \frac{\left\|\mathbf{r}^{[k]}(a, f)\right\|_2}{\|f\|_2}.$$

Lemma VI.4.6. *$\mathrm{L}_{\mathbf{r}}$ is a common Lipschitz constant to $\mathbf{r}(a, \cdot)$ and $\mathbf{r}^{[k]}(a, \cdot)$ ($k \in \mathcal{K}$ and $a \in \mathcal{A}$). In other words, for all $k \in \mathcal{K}$ and $a \in \mathcal{A}$, we have*

(i) for all $f, f' \in \mathbb{R}^{\mathcal{S}\times\mathcal{I}}$, $\left\|\mathbf{r}^{[k]}(a, f) - \mathbf{r}^{[k]}(a, f')\right\|_2 \leqslant \mathrm{L}_{\mathbf{r}} \|f - f'\|_2$;

(ii) for all $f, f' \in \mathcal{F}$, $\left\|\mathbf{r}(a, f) - \mathbf{r}(a, f')\right\|_2 \leqslant \mathrm{L}_{\mathbf{r}} \|f - f'\|_2$.

For each $k \in \mathcal{K}$, define the linear map $\mathbf{R}_k : (\mathbb{R}^{\mathcal{S}\times\mathcal{I}})^{\mathcal{A}} \to \mathbb{R}^d$ as follows

$$\mathbf{R}_k\left((\tilde{g}^{ka})_{a\in\mathcal{A}}\right) := \sum_{a\in\mathcal{A}} \mathbf{r}^{[k]}(a, \tilde{g}^{ka}), \quad \text{for all } (\tilde{g}^{ka})_{a\in\mathcal{A}} \in (\mathbb{R}^{\mathcal{S}\times\mathcal{I}})^{\mathcal{A}}.$$

Then, define the linear map $\mathbf{R} : (\mathbb{R}^{\mathcal{S}\times\mathcal{I}})^{\mathcal{K}\times\mathcal{A}} \to \mathbb{R}^d$ by setting

$$\begin{aligned}\mathbf{R}(\tilde{g}) &:= \sum_{k\in\mathcal{K}} \mathbf{R}_k\left((\tilde{g}^{ka})_{a\in\mathcal{A}}\right)\\ &= \sum_{k\in\mathcal{K}} \sum_{a\in\mathcal{A}} \mathbf{r}^{[k]}(a, \tilde{g}^{ka}), \quad \text{for all } \tilde{g} = (\tilde{g}^{ka})_{\substack{k\in\mathcal{K}\\ a\in\mathcal{A}}} \in (\mathbb{R}^{\mathcal{S}\times\mathcal{I}})^{\mathcal{K}\times\mathcal{A}}.\end{aligned}$$

The following proposition shows that $\mathbf{R}$ does indeed encode $\mathbf{r}$.

Proposition VI.4.7. *Let $x \in \Delta(\mathcal{I})$, $f \in \mathcal{F}$, $\ell \in \mathcal{L}$ such that $x \in \mathcal{X}^\ell$, and $k_0 \in \mathcal{K}$ such that $f \in \mathcal{F}^{k_0}$. Moreover, let*

$$x = \sum_{a\in\mathcal{A}} \lambda^a \cdot a \quad \textit{where} \quad \begin{cases} (\lambda^a)_{a\in\mathcal{A}} \in \Delta(\mathcal{A}) \\ \mathrm{supp}(\lambda^a)_{a\in\mathcal{A}} \subset \mathcal{X}^\ell. \end{cases}$$

be an expression of x as a convex combination of the vertices of $\mathscr{X}^\ell$. Then,

$$\mathbf{R}\left(\left(\mathbb{1}_{\{k_0=k\}}\lambda^a \cdot f\right)_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}}\right) = \mathbf{r}(x,f).$$

Proof. Using the definition of $\mathbf{R}$,

$$\mathbf{R}\left(\left(\mathbb{1}_{\{k_0=k\}}\lambda^a \cdot f\right)_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}}\right) = \sum_{k\in\mathscr{K}}\sum_{a\in\mathscr{A}} \mathbf{r}^{[k]}(a, \mathbb{1}_{\{k_0=k\}}\lambda^a \cdot f) = \sum_{a\in\mathscr{A}} \lambda^a \cdot \mathbf{r}^{[k_0]}(a,f)$$
$$= \sum_{a\in\mathscr{A}} \lambda^a \cdot \mathbf{r}(a,f) = \mathbf{r}(x,f),$$

where the second equality holds because by linearity of $\mathbf{r}^{[k]}(a,\cdot)$ (property (i) in Lemma VI.4.5), the fourth because $\mathbf{r}^{[k_0]}(x,\cdot)$ and $\mathbf{r}(x,\cdot)$ coincide on $\mathscr{F}^{k_0}$ (property (ii) in Lemma VI.4.5), and the last by affinity of $\mathbf{r}(\cdot,f)$ on $\mathscr{X}^\ell$ (property (iv) in Proposition VI.4.4). □

VI.4.3. The auxiliary full information game

We now construct an auxiliary approachability game. The important point will be that the target set is approachable. This fact will be used in the construction and the analysis of the strategy for the initial game.

The payoff space for this auxiliary game is $(\mathbb{R}^{\mathscr{S}\times\mathscr{T}})^{\mathscr{K}\times\mathscr{A}}$. An element $\tilde{g} \in (\mathbb{R}^{\mathscr{S}\times\mathscr{T}})^{\mathscr{K}\times\mathscr{A}}$ will often be written as

$$\tilde{g} = (\tilde{g}^{ka})_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}}, \quad \text{where} \quad \tilde{g}^{ka} \in \mathbb{R}^{\mathscr{S}\times\mathscr{T}}.$$

Then, if $\tilde{z} = (\tilde{z}^{ka})_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}}$ also belongs to $(\mathbb{R}^{\mathscr{S}\times\mathscr{T}})^{\mathscr{K}\times\mathscr{A}}$, the scalar product $\langle\tilde{g}|\tilde{z}\rangle$ can obviously be written as the sum of the scalar products $\langle\tilde{g}^{ka}|\tilde{z}^{ka}\rangle$, and a similar expression holds for the square Euclidean norm:

$$\langle\tilde{g}|\tilde{z}\rangle = \sum_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}} \langle\tilde{g}^{ka}|\tilde{z}^{ka}\rangle \quad \text{and} \quad \|\tilde{g}\|_2^2 = \sum_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}} \|\tilde{g}^{ka}\|_2^2.$$

The auxiliary game is defined as follows. Let $\mathscr{K}\times\mathscr{A}$ be the set of pure actions for the Decision Maker and $\mathscr{F}$ the convex action set for Nature. The payoff function $\tilde{\mathbf{g}}$ takes values in $(\mathbb{R}^{\mathscr{S}\times\mathscr{T}})^{\mathscr{K}\times\mathscr{A}}$ and is defined as follows. For $(k,a)\in\mathscr{K}\times\mathscr{A}$ and $f\in\mathscr{F}$,

$$\tilde{\mathbf{g}}((k,a),f) := (\mathbb{1}_{\{k=k'\}}\mathbb{1}_{\{a=a'\}} \cdot f)_{\substack{k'\in\mathscr{K}\\ a'\in\mathscr{A}}} \quad \in (\mathbb{R}^{\mathscr{S}\times\mathscr{T}})^{\mathscr{K}\times\mathscr{A}}.$$

This payoff function is be bilinearly extended to $\Delta(\mathscr{K}\times\mathscr{A})\times\mathbb{R}^{\mathscr{S}\times\mathscr{T}}$. For each $k\in\mathscr{K}$, let $\mathscr{F}_c^k := \mathbb{R}_+\mathscr{F}^k = (\mathscr{F}^k)^{\circ\circ}$ be the smallest closed convex cone containing the convex

compact set $\mathscr{F}^k$ (see Section IV.2 for definitions and properties about closed convex cones), and consider the following subset of $(\mathbb{R}^{\mathcal{S}\times\mathcal{I}})^{\mathcal{A}}$:

$$\tilde{\mathscr{C}}^k := \mathbf{R}_k^{-1}(\mathbb{R}^d_-) \cap (\mathscr{F}_c^k)^{\mathcal{A}} \quad \subset (\mathbb{R}^{\mathcal{S}\times\mathcal{I}})^{\mathcal{A}}.$$

We then define the target set $\tilde{\mathscr{C}}$ as the Cartesian product of the sets $\tilde{\mathscr{C}}^k$:

$$\tilde{\mathscr{C}} := \prod_{k\in\mathcal{K}} \tilde{\mathscr{C}}^k \quad \subset (\mathbb{R}^{\mathcal{S}\times\mathcal{I}})^{\mathcal{A}\times\mathcal{K}}.$$

Lemma VI.4.8. *(i) The sets $\tilde{\mathscr{C}}^k$ and $\tilde{\mathscr{C}}$ are closed convex cones.*

(ii) $\tilde{\mathscr{C}} \subset \mathbf{R}^{-1}(\mathbb{R}^d_-) \cap \left(\prod_{k\in\mathcal{K}} (\mathscr{F}_c^k)^{\mathcal{A}} \right)$.

Proposition VI.4.9. *The set $\tilde{\mathscr{C}}$ is approachable in the auxiliary game. In other words, for all $\tilde{z} \in \tilde{\mathscr{C}}^\circ$, there exists $\tilde{x} := \tilde{\mathbf{x}}(\tilde{z}) \in \Delta(\mathcal{K}\times\mathcal{A})$ such that*

$$\forall f \in \mathscr{F}, \quad \langle \tilde{\mathbf{g}}(\tilde{x}, f) | \tilde{z} \rangle \leqslant 0.$$

Proof. This full information game has convex compact action sets and a bilinear payoff function. Thanks to Proposition VI.3.3, the statement of the proposition is then equivalent to Blackwell condition:

$$\forall f \in \mathscr{F}, \ \exists \tilde{x} \in \Delta(\mathcal{K}\times\mathcal{A}), \quad \tilde{\mathbf{g}}(\tilde{x}, f) \in \tilde{\mathscr{C}},$$

which we now aim at proving. Let $f \in \mathscr{F}$ and $k_0 \in \mathcal{K}$ such that $f \in \mathscr{F}^{k_0}$. According to property (ii) in Proposition VI.4.4, there exists $x \in \Delta(\mathcal{I})$ such that such that $\mathbf{r}(x, f) \in \mathbb{R}^d_-$. By Lemma VI.4.3, there exists $\ell \in \mathscr{L}$ such that $x \in \mathscr{X}^\ell$ and we can write x as a convex combination of the vertices of $\mathscr{X}^\ell$:

$$x = \sum_{a\in\mathcal{A}} \lambda^a \cdot a \quad \text{where} \quad \begin{cases} (\lambda^a)_{a\in\mathcal{A}} \in \Delta(\mathcal{A}) \\ \operatorname{supp}(\lambda^a)_{a\in\mathcal{A}} \subset \mathscr{X}^\ell. \end{cases}$$

Now consider the mixed action

$$\tilde{x} := \left(\mathbb{1}_{\{k=k_0\}} \lambda^a \right)_{\substack{k\in\mathcal{K} \\ a\in\mathcal{A}}} \quad \in \Delta(\mathcal{K}\times\mathcal{A})$$

and let us prove that $\tilde{\mathbf{g}}(\tilde{x}, f) \in \tilde{\mathscr{C}}$. We have by definition of $\tilde{\mathbf{g}}$:

$$\tilde{\mathbf{g}}(\tilde{x}, f) = \left(\mathbb{1}_{\{k=k_0\}} \lambda^a \cdot f \right)_{\substack{k\in\mathcal{K} \\ a\in\mathcal{A}}},$$

and since $\tilde{\mathscr{C}} = \prod_{k\in\mathscr{K}} \tilde{\mathscr{C}}^k$, we only have to check that $(\lambda^a f)_{a\in\mathscr{A}}$ belongs to $\tilde{\mathscr{C}}^{k_0} = \mathbf{R}_{k_0}^{-1}(\mathbb{R}^d_-) \cap (\mathscr{F}_c^{k_0})^{\mathscr{A}}$. First, because $f \in \mathscr{F}^{k_0}$, $\lambda^a f$ belongs to the closed convex cone $\mathscr{F}_c^{k_0} = \mathbb{R}_+\mathscr{F}^{k_0}$ and we have indeed $(\lambda^a f)_{a\in\mathscr{A}} \in (\mathscr{F}_c^{k_0})^{\mathscr{A}}$. Then, let us prove that $\mathbf{R}_{k_0}((\lambda^a f)_{a\in\mathscr{A}}) \in \mathbb{R}^d_-$. Using Proposition VI.4.7,

$$\mathbf{R}_{k_0}((\lambda^a f)_{a\in\mathscr{A}}) = \mathbf{R}\left(\left(\mathbb{1}_{\{k=k_0\}}\lambda^a \cdot f\right)_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}}\right) = \mathbf{r}(x, f) \in \mathbb{R}^d_-.$$

Therefore, we have proved that $(\lambda^a f)_{a\in\mathscr{A}}$ belongs to $\tilde{\mathscr{C}}^{k_0} = \mathbf{R}_{k_0}^{-1}(\mathbb{R}^d_-) \cap (\mathscr{F}_c^{k_0})^{\mathscr{A}}$, and thus, that $\tilde{\mathbf{g}}(\tilde{x}, f) \in \tilde{\mathscr{C}}$, which concludes the proof. □

VI.4.4. The strategy for the initial game

Let $\tilde{\mathscr{Z}} := \tilde{\mathscr{C}}^\circ \cap \mathscr{B}_2$ where $\mathscr{B}_2$ denotes the closed unit Euclidean ball on $(\mathbb{R}^{\mathscr{S}\times\mathscr{T}})^{\mathscr{K}\times\mathscr{A}}$. The strategy is defined as follows. Let $\eta > 0$ and $0 < \gamma \leqslant 1$ be parameters. For $t \geqslant 1$,

- compute $\tilde{z}_t := \mathbf{P}_{\tilde{\mathscr{Z}}}\left(\eta \sum_{s=1}^{t-1} \tilde{g}_s\right)$, where $\mathbf{P}_{\tilde{\mathscr{Z}}}$ denotes the Euclidean projection onto $\tilde{\mathscr{Z}}$;
- compute $\tilde{x}_t := \tilde{\mathbf{x}}(\tilde{z}_t) \in \Delta(\mathscr{K}\times\mathscr{A})$, where $\tilde{\mathbf{x}}$ is defined in Proposition VI.4.9;
- draw $(k_t, a_t) \sim \tilde{x}_t$ and then $i_t \sim (1-\gamma)a_t + \gamma u$, where $u := (\frac{1}{|\mathscr{T}|}, \dots, \frac{1}{|\mathscr{T}|})$ is the uniform distribution over $\mathscr{T}$;
- observe signal $s_t \sim \mathbf{s}(i_t, j_t)$ and compute estimator

$$\hat{f}_t = \left(\frac{\mathbb{1}_{\{i_t=i\}}}{\mathbb{P}[i_t = i \mid \mathscr{G}_t]}\delta_{s_t}\right)_{i\in\mathscr{T}} \in \mathbb{R}^{\mathscr{S}\times\mathscr{T}},$$

  where δ_{s_t} is the Dirac mass associated with $s_t \in \mathscr{S}$ and seen as an element of $\mathbb{R}^{\mathscr{S}}$;
- set $\tilde{g}_t = \tilde{\mathbf{g}}((k_t, a_t), \hat{f}_t)$.

Let $(\mathscr{G}_t)_{t\geqslant 1}$ be the filtration where for each $t \geqslant 1$,

$$\mathscr{G}_t \quad \text{is generated by} \quad (k_1, a_1, i_1, s_1, \dots, k_{t-1}, a_{t-1}, i_{t-1}, s_{t-1}, k_t, a_t).$$

The definition of the strategy implies that

$$\mathbb{P}[i_t = i \mid \mathscr{G}_t] = (1-\gamma)a_t^i + \frac{\gamma}{|\mathscr{T}|}, \quad i \in \mathscr{T}.$$

The following lemma gathers the properties of estimator $\hat{f}_t$.

Lemma VI.4.10. *For all $t \geqslant 1$,*

(i) $\mathbb{E}\left[\widehat{f}_t \,\middle|\, \mathscr{G}_t\right] = \mathbb{E}\left[f_t \,|\, \mathscr{G}_t\right]$;

(ii) $\mathbb{E}\left[\left\|\widehat{f}_t\right\|_2^2 \,\middle|\, \mathscr{G}_t\right] \leqslant \dfrac{|\mathscr{I}|^2}{\gamma}$;

(iii) $\left\|\widehat{f}_t\right\|_2^2 \leqslant \dfrac{|\mathscr{I}|^2}{\gamma^2}$.

VI.5. Main result

We now state our main result which establishes that the strategy defined in Section VI.4.4 guarantees that the average payoff $\bar{g}_T$ (of the initial game) converges in expectation to the negative orthant $\mathbb{R}^d_-$ at rate $O(T^{-1/3})$.

Theorem VI.5.1. *Let $T \geqslant 1$ be an integer. Against any strategy of Nature, the strategy defined in Section VI.4.4 run with*

$$\eta = \sqrt{\frac{\gamma}{T\,|\mathscr{I}|^2}} \quad \text{and} \quad \gamma = \min\left\{\left(\frac{11\,L_{\mathbf{r}}\,|\mathscr{I}|\,|\mathscr{H}|\,|\mathscr{A}|}{4\,\|\mathbf{g}\|_2}\right)^{2/3} T^{-1/3},\ 1\right\}$$

guarantees

$$\mathbb{E}\left[\mathbf{d}_2\left(\bar{g}_T,\ \mathbb{R}^d_-\right)\right] \leqslant \frac{12\,\|\mathbf{g}\|_2^{1/3}\,(L_{\mathbf{r}}\,|\mathscr{I}|\,|\mathscr{H}|\,|\mathscr{A}|)^{2/3}}{T^{1/3}} + \frac{2\sqrt{\pi}\,\|\mathbf{g}\|_2}{T^{1/2}} + \frac{6\,\|\mathbf{g}\|_2^{2/3}\,(L_{\mathbf{r}}\,|\mathscr{I}|\,|\mathscr{H}|\,|\mathscr{A}|)^{1/3}}{T^{2/3}},$$

where $\mathbf{d}_2(\cdot,\ \mathbb{R}^d_-)$ denotes the Euclidean distance to the negative orthant $\mathbb{R}^d_-$.

Remark VI.5.2. Since $L_{\mathbf{r}}$ scales linearly with $\|\mathbf{g}\|_2$, so does the dominant term of the above bound, as expected.

Let us introduce some notation. Let $\bar{\tilde{g}}_T$ be the average for $t = 1, \ldots, T$ of auxiliary payoffs $\tilde{g}_t$. In the analysis we will partition the set of stages $\{1, \ldots, T\}$ with respect to the realized values of $k_t \in \mathscr{K}$ and $a_t \in \mathscr{A}$. For $k \in \mathscr{K}$ and $a \in \mathscr{A}$, let $N_T(k, a)$ be the set of stages $t \in \{1, \ldots, T\}$ where $k_t = k$ and $a_t = a$, and $\lambda_T(k, a)$ the corresponding proportion of stages:

$$N_T(k, a) := \{1 \leqslant t \leqslant T \,|\, k_t = k,\ a_t = a\}$$
$$\lambda_T(k, a) := \frac{|N_T(k, a)|}{T}.$$

Then, for any sequence $(u_t)_{1\leqslant t\leqslant \mathrm{T}}$, we denote $\bar{u}_{\mathrm{T}}(k,a)$ its average over $t\in \mathrm{N}_{\mathrm{T}}(k,a)$:

$$\bar{u}_{\mathrm{T}}(k,a) := \begin{cases} \dfrac{1}{|\mathrm{N}_{\mathrm{T}}(k,a)|}\displaystyle\sum_{t\in \mathrm{N}_{\mathrm{T}}(k,a)} u_t & \text{if } \mathrm{N}_{\mathrm{T}}(k,a)\neq\emptyset \\ 0 & \text{otherwise.}\end{cases}$$

The proof is divided into the subsections below which are mostly independent. Here is a overview of the main steps:

$$\begin{aligned}
\bar{g}_{\mathrm{T}} \quad &\text{is close to} \quad \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} g(a_t,y_t) \quad \text{(Lemma VI.5.11)}\\
&\text{is equal to} \quad \sum_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}} \lambda_{\mathrm{T}}(k,a)\cdot \mathbf{g}(a,\bar{y}_{\mathrm{T}}(k,a)) \quad \text{(Lemma VI.5.10)}\\
\text{is closer to } \mathbb{R}^d_- \text{ than} \quad &\sum_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}} \lambda_{\mathrm{T}}(k,a)\cdot \mathbf{r}(a,\bar{f}_{\mathrm{T}}(k,a)) \quad \text{(Lemma VI.5.9)}\\
&\text{is close to} \quad \sum_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}} \lambda_{\mathrm{T}}(k,a)\cdot \mathbf{r}^{[k]}(a,\bar{\tilde{f}}_{\mathrm{T}}(k,a)) \quad \text{(Lemma VI.5.8)}\\
&\text{is equal to} \quad \mathbf{R}(\bar{\tilde{g}}_{\mathrm{T}}) \quad \text{(Lemma VI.5.5)}\\
&\text{is close to} \quad \mathbb{R}^d_- \quad \text{(Lemmas VI.5.4 and VI.5.3).}
\end{aligned}$$

VI.5.1. Average auxiliary payoff $\bar{\tilde{g}}_{\mathrm{T}}$ is close to auxiliary target set $\tilde{\mathscr{C}}$

Lemma VI.5.3.

$$\mathbb{E}\left[\mathbf{d}_2\left(\bar{\tilde{g}}_{\mathrm{T}},\,\tilde{\mathscr{C}}\right)\right] \leqslant \frac{1}{2\eta\mathrm{T}} + \frac{\eta\,|\mathscr{Y}|^2}{2\gamma}.$$

Proof. For $t\geqslant 1$, we can write

$$\begin{aligned}
\tilde{z}_t = \mathbf{P}_{\tilde{\mathscr{Z}}}\left(\eta\sum_{s=1}^{t-1}\tilde{g}_s\right) &= \operatorname*{arg\,min}_{\tilde{z}\in\tilde{\mathscr{Z}}}\left\|\tilde{z}-\eta\sum_{s=1}^{t-1}\tilde{g}_s\right\|_2^2\\
&= \operatorname*{arg\,max}_{\tilde{z}\in\tilde{\mathscr{Z}}}\left\{\left\langle \eta\sum_{s=1}^{t-1}\tilde{g}_s\middle|\tilde{z}\right\rangle - \frac{1}{2}\|\tilde{z}\|_2^2\right\}.
\end{aligned}$$

Then, Theorem I.3.1 together with the fact that $\left\|\tilde{\mathscr{Z}}\right\|_2 = \left\|\tilde{\mathscr{C}}^\circ\cap\mathscr{B}_2\right\|_2 \leqslant 1$ gives

$$\max_{\tilde{z}\in\tilde{\mathscr{Z}}}\sum_{t=1}^{\mathrm{T}}\langle\tilde{g}_t|\tilde{z}\rangle - \sum_{t=1}^{\mathrm{T}}\langle\tilde{g}_t|\tilde{z}_t\rangle \leqslant \frac{1}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{\mathrm{T}}\|\tilde{g}_t\|_2^2.$$

By taking the expectation and dividing by T, we get

$$\mathbb{E}\left[\max_{\tilde{z}\in\tilde{\mathcal{Z}}}\left\langle\tilde{\bar{g}}_{\mathrm{T}}\middle|\tilde{z}\right\rangle\right] \leqslant \frac{1}{2\eta\mathrm{T}} + \mathbb{E}\left[\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\langle\tilde{g}_t|\tilde{z}_t\rangle\right] + \frac{\eta}{2\mathrm{T}}\mathbb{E}\left[\sum_{t=1}^{\mathrm{T}}\|\tilde{g}_t\|_2^2\right].$$

We first analyze the first sum of the right-hand side. Let us prove that each scalar product $\langle\tilde{g}_t|\tilde{z}_t\rangle$ is nonpositive in expectation. For all $1 \leqslant t \leqslant \mathrm{T}$, we replace $\tilde{g}_t$ by its definition:

$$\mathbb{E}\left[\langle\tilde{g}_t|\tilde{z}_t\rangle\right] = \mathbb{E}\left[\left\langle\tilde{\mathbf{g}}((k_t,a_t),\widehat{f}_t)\middle|\tilde{z}_t\right\rangle\right].$$

We then consider the conditional expectation with respect to $\mathscr{G}_t$. The application $\tilde{\mathbf{g}}((k_t,a_t),\,\cdot\,)$ being linear, and the variables k_t, a_t and $\tilde{z}_t$ being measurable with respect to $\mathscr{G}_t$, we can make $\mathbb{E}\left[\widehat{f}_t\,\middle|\,\mathscr{G}_t\right]$ appear as follows:

$$\begin{aligned}\mathbb{E}\left[\langle\tilde{g}_t|\tilde{z}_t\rangle\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\langle\tilde{\mathbf{g}}((k_t,a_t),\widehat{f}_t)\middle|\tilde{z}_t\right\rangle\middle|\mathscr{G}_t\right]\right] = \mathbb{E}\left[\left\langle\tilde{\mathbf{g}}\left((k_t,a_t),\mathbb{E}\left[\widehat{f}_t\,\middle|\,\mathscr{G}_t\right]\right)\middle|\tilde{z}_t\right\rangle\right]\\ &= \mathbb{E}\left[\langle\tilde{\mathbf{g}}((k_t,a_t),\mathbb{E}\left[f_t\,|\,\mathscr{G}_t\right])|\tilde{z}_t\rangle\right] = \mathbb{E}\left[\langle\tilde{\mathbf{g}}((k_t,a_t),f_t)|\tilde{z}_t\rangle\right],\end{aligned}$$

where we used Lemma VI.4.10 to replace the conditional expectation of $\widehat{f}_t$ by the conditional expectation of f_t. Now consider the sigma-algebra $\mathscr{H}_t$ generated by

$$(k_1,a_1,i_1,s_1,\ldots,k_{t-1},a_{t-1},i_{t-1},s_{t-1}).$$

By definition of the strategy, the law of random variable (k_t,a_t) knowing $\mathscr{H}_t$ is $\tilde{x}_t$. We now resume the above computation by introducing the conditional expectation with respect to $\mathscr{H}_t$ and f_t:

$$\begin{aligned}\mathbb{E}\left[\langle\tilde{g}_t|\tilde{z}_t\rangle\right] &= \mathbb{E}\left[\langle\tilde{\mathbf{g}}((k_t,a_t),f_t)|\tilde{z}_t\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle\tilde{\mathbf{g}}((k_t,a_t),f_t)|\tilde{z}_t\rangle\,|\,\mathscr{H}_t,f_t\right]\right]\\ &= \mathbb{E}\left[\langle\tilde{\mathbf{g}}\left(\mathbb{E}\left[(k_t,a_t)\,|\,\mathscr{H}_t,f_t\right],f_t\right)|\tilde{z}_t\rangle\right] = \mathbb{E}\left[\langle\tilde{\mathbf{g}}\left(\mathbb{E}\left[(k_t,a_t)\,|\,\mathscr{H}_t\right],f_t\right)|\tilde{z}_t\rangle\right]\\ &= \mathbb{E}\left[\langle\tilde{\mathbf{g}}(\tilde{x}_t,f_t)|\tilde{z}_t\rangle\right].\end{aligned}$$

By definition of the strategy, $\tilde{x}_t = \tilde{\mathbf{x}}(\tilde{z}_t)$. In other words (see Proposition VI.4.9), for all $f \in \mathscr{F}$, the scalar product $\langle\tilde{\mathbf{g}}(\tilde{x}_t,f)|\tilde{z}_t\rangle$ is nonpositive. This is in particular true for $f = f_t$. Therefore, $\mathbb{E}\left[\langle\tilde{g}_t|\tilde{z}_t\rangle\right] \leqslant 0$.

We now turn to the second sum that involves the squared norms $\|\tilde{g}_t\|_2^2$. For $1 \leqslant t \leqslant \mathrm{T}$, using the definition of $\tilde{\mathbf{g}}$,

$$\begin{aligned}\|\tilde{g}_t\|_2^2 &= \left\|\tilde{\mathbf{g}}((k_t,a_t),\widehat{f}_t)\right\|_2^2 = \left\|\left(\mathbb{1}_{\{k=k_t\}}\mathbb{1}_{\{a=a_t\}}\widehat{f}_t\right)_{\substack{k\in\mathscr{K}\\a\in\mathscr{A}}}\right\|_2^2\\ &= \sum_{\substack{k\in\mathscr{K}\\a\in\mathscr{A}}}\left\|\mathbb{1}_{\{k=k_t\}}\mathbb{1}_{\{a=a_t\}}\widehat{f}_t\right\|_2^2 = \left\|\widehat{f}_t\right\|_2^2.\end{aligned}$$

Using (ii) from Lemma VI.5.3, we have

$$\mathbb{E}\left[\|\tilde{g}_t\|_2^2\right] = \mathbb{E}\left[\left\|\hat{f}_t\right\|_2^2\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\hat{f}_t\right\|_2^2 \middle| \mathcal{G}_t\right]\right] \leqslant \frac{|\mathcal{I}|^2}{\gamma}.$$

Putting everything together, we obtain in expectation the following bound on the distance from $\bar{\tilde{g}}_{\mathrm{T}}$ to $\tilde{\mathscr{C}}$:

$$\mathbb{E}\left[\mathbf{d}_2\left(\bar{\tilde{g}}_{\mathrm{T}}, \tilde{\mathscr{C}}\right)\right] = \mathbb{E}\left[\max_{\tilde{z}\in\tilde{\mathcal{Z}}} \left\langle \bar{\tilde{g}}_{\mathrm{T}} \middle| \tilde{z}\right\rangle\right] \leqslant \frac{1}{2\eta \mathrm{T}} + \frac{\eta |\mathcal{I}|^2}{2\gamma},$$

where the above equality comes from the expression of the Euclidean distance to $\tilde{\mathscr{C}}$ given by Proposition IV.2.10. □

VI.5.2. From $\bar{\tilde{g}}_{\mathrm{T}}$ in the auxiliary space to $\mathbf{R}(\bar{\tilde{g}}_{\mathrm{T}})$ in the initial space

Lemma VI.5.4.

$$\mathbf{d}_2\left(\mathbf{R}(\bar{\tilde{g}}_{\mathrm{T}}), \mathbb{R}_-^d\right) \leqslant \left(\mathrm{L}_{\mathbf{r}}\sqrt{|\mathcal{K}|\,|\mathcal{A}|}\right) \cdot \mathbf{d}_2\left(\bar{\tilde{g}}_{\mathrm{T}}, \tilde{\mathscr{C}}\right).$$

Proof. It follows from property (ii) in Lemma VI.4.8 that $\tilde{\mathscr{C}} \subset \mathbf{R}^{-1}(\mathbb{R}_-^d)$. Therefore, we can write

$$\begin{aligned}
\mathbf{d}_2(\mathbf{R}(\bar{\tilde{g}}_{\mathrm{T}}), \mathbb{R}_-^d) &= \min_{g'\in\mathbb{R}_-^d} \left\|\mathbf{R}(\bar{\tilde{g}}_{\mathrm{T}}) - g'\right\|_2 \leqslant \min_{\tilde{g}\in\mathbf{R}^{-1}(\mathbb{R}_-^d)} \left\|\mathbf{R}(\bar{\tilde{g}}_{\mathrm{T}}) - \mathbf{R}(\tilde{g})\right\|_2 \\
&\leqslant \min_{\tilde{g}\in\tilde{\mathscr{C}}} \left\|\mathbf{R}(\bar{\tilde{g}}_{\mathrm{T}}) - \mathbf{R}(\tilde{g})\right\|_2 \leqslant \|\mathbf{R}\| \cdot \min_{\tilde{g}\in\tilde{\mathscr{C}}} \left\|\bar{\tilde{g}}_{\mathrm{T}} - \tilde{g}\right\|_2 \\
&= \|\mathbf{R}\| \cdot \mathbf{d}_2\left(\bar{\tilde{g}}_{\mathrm{T}}, \tilde{\mathscr{C}}\right),
\end{aligned}$$

where $\|\mathbf{R}\|$ is the operator norm of $\mathbf{R}$. To conclude the proof, let us prove that the latter is bounded from above by $\mathrm{L}_{\mathbf{r}}\sqrt{|\mathcal{K}|\,|\mathcal{A}|}$. Let $\tilde{g} \in (\mathbb{R}^{\mathcal{S}\times\mathcal{I}})^{\mathcal{K}\times\mathcal{A}}$. By definition of $\mathbf{R}$, and using the Lipschitz constant $\mathrm{L}_{\mathbf{r}}$ from Lemma VI.4.6 which is common to the linear applications $\mathbf{r}^{[k]}(a, \cdot)$, we have

$$\begin{aligned}
\|\mathbf{R}(\tilde{g})\|_2 &= \left\|\sum_{\substack{k\in\mathcal{K}\\ a\in\mathcal{A}}} \mathbf{r}^{[k]}(a, \tilde{g}^{ka})\right\|_2 \leqslant \sum_{\substack{k\in\mathcal{K}\\ a\in\mathcal{A}}} \left\|\mathbf{r}^{[k]}(a, \tilde{g}^{ka})\right\|_2 \leqslant \sum_{\substack{k\in\mathcal{K}\\ a\in\mathcal{A}}} \mathrm{L}_{\mathbf{r}} \left\|\tilde{g}^{ka}\right\|_2 \\
&\leqslant \mathrm{L}_{\mathbf{r}} \sqrt{|\mathcal{K}|\,|\mathcal{A}| \sum_{\substack{k\in\mathcal{K}\\ a\in\mathcal{A}}} \|\tilde{g}^{ka}\|_2^2} = \mathrm{L}_{\mathbf{r}}\sqrt{|\mathcal{K}|\,|\mathcal{A}|} \cdot \|\tilde{g}\|_2,
\end{aligned}$$

which concludes the proof. □

VI.5.3. Decomposition of $\mathbf{R}(\bar{\tilde{g}}_{\mathrm{T}})$

We have the following expression of the image by $\mathbf{R}$ of the average auxiliary payoff $\bar{\tilde{g}}_{\mathrm{T}}$.

Lemma VI.5.5.

$$\mathbf{R}(\bar{\tilde{g}}_{\mathrm{T}}) = \mathbf{R}\left(\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\tilde{g}_t\right) = \sum_{\substack{k\in\mathcal{K}\\ a\in\mathcal{A}}} \lambda_{\mathrm{T}}(k,a)\cdot \mathbf{r}^{[k]}(a, \bar{\hat{f}}_{\mathrm{T}}(k,a)).$$

Proof. Using the definitions of $\mathbf{R}$, $\tilde{g}_t$, $\tilde{\mathbf{g}}$, and the linearity of $\mathbf{R}$ and $\mathbf{r}^{[k]}(a,\cdot)$, we can write

$$\begin{aligned}
\mathbf{R}\left(\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\tilde{g}_t\right) &= \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\mathbf{R}(\tilde{g}_t) = \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\sum_{\substack{k\in\mathcal{K}\\ a\in\mathcal{A}}}\mathbf{r}^{[k]}(a,\tilde{g}_t^{ka})\\
&= \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\sum_{\substack{k\in\mathcal{K}\\ a\in\mathcal{A}}}\mathbf{r}^{[k]}\left(a, \mathbb{1}_{\{k=k_t\}}\mathbb{1}_{\{a=a_t\}}\hat{f}_t\right)\\
&= \sum_{\substack{k\in\mathcal{K}\\ a\in\mathcal{A}}} \lambda_{\mathrm{T}}(k,a)\cdot \mathbf{r}^{[k]}(a, \bar{\hat{f}}_{\mathrm{T}}(k,a)).
\end{aligned}$$

□

VI.5.4. Average estimator $\bar{\hat{f}}_{\mathrm{T}}(k,a)$ is close to average flag $\bar{f}_{\mathrm{T}}(k,a)$

Lemma VI.5.6.

$$\mathbb{E}\left[\sum_{\substack{k\in\mathcal{K}\\ a\in\mathcal{A}}} \lambda_{\mathrm{T}}(k,a)\left\|\bar{\hat{f}}_{\mathrm{T}}(k,a) - \bar{f}_{\mathrm{T}}(k,a)\right\|_2\right] \leqslant |\mathcal{I}|\,|\mathcal{K}|\,|\mathcal{A}|\left(\frac{8}{\sqrt{\mathrm{T}\gamma}} + \frac{8}{3\mathrm{T}\gamma}\right).$$

Proof. Let $k\in\mathcal{K}$ and $a\in\mathcal{A}$. Consider the random process $(\mathrm{X}_t(k,a))_{t\geqslant 1}$ defined by

$$\mathrm{X}_t(k,a) := \mathbb{1}_{\{k_t=k,\, a_t=a\}}\left(\hat{f}_t - f_t\right),$$

and to which we are aiming at applying Corollary A.0.5. $(\mathrm{X}_t(k,a))_{t\geqslant 1}$ is a martingale difference sequence with respect to filtration $(\mathcal{G}_t)_{t\geqslant 1}$. Indeed, since $\mathbb{1}_{\{k_t=k,\, a_t=a\}}$ is measurable with respect to $\mathcal{G}_t$,

$$\mathbb{E}\left[\mathbb{1}_{\{k_t=k,\, a_t=a\}}\left(\hat{f}_t - f_t\right)\middle|\,\mathcal{G}_t\right] = \mathbb{1}_{\{k_t=k,\, a_t=a\}}\mathbb{E}\left[\hat{f}_t - f_t\middle|\,\mathcal{G}_t\right] = 0.$$

where the last equality follows from (i) in Lemma VI.4.10. Moreover, using (iii) from Lemma VI.4.10, we bound each $X_t(k,a)$ as follows.

$$\begin{aligned}\|X_t(k,a)\|_2 \leqslant \left\|\hat{f}_t - f_t\right\|_2 \leqslant \left\|\hat{f}_t\right\|_2 + \|f_t\|_2 \leqslant \frac{|\mathcal{I}|}{\gamma} + \left\|(\mathbf{s}(i,y_t))_{i\in\mathcal{I}}\right\|_2 \\ = \frac{|\mathcal{I}|}{\gamma} + \sqrt{\sum_{i\in\mathcal{I}} \|\mathbf{s}(i,y_t)\|_2^2} \leqslant \frac{|\mathcal{I}|}{\gamma} + \sqrt{|\mathcal{I}|} \leqslant \frac{2\,|\mathcal{I}|}{\gamma},\end{aligned}$$

where we used the fact that $\gamma \geqslant 1$ for the last inequality. As far as the conditional variances are concerned, we have

$$\begin{aligned}\mathbb{E}\left[\|X_t(k,a)\|_2^2 \,\middle|\, \mathcal{G}_t\right] = \mathbb{E}\left[\mathbb{1}_{\{k_t=k,\, a_t=a\}} \left\|\hat{f}_t - f_t\right\|_2^2 \,\middle|\, \mathcal{G}_t\right] \leqslant \mathbb{E}\left[\left\|\hat{f}_t - f_t\right\|_2^2 \,\middle|\, \mathcal{G}_t\right] \\ \leqslant \mathbb{E}\left[\left\|\hat{f}_t\right\|_2^2 \,\middle|\, \mathcal{G}_t\right] + \mathbb{E}\left[\|f_t\|_2^2 \,\middle|\, \mathcal{G}_t\right] \leqslant \frac{|\mathcal{I}|^2}{\gamma} + |\mathcal{I}| \leqslant \frac{2\,|\mathcal{I}|^2}{\gamma}.\end{aligned}$$

where the first term of the second line has been bounded using property (ii) from Lemma VI.4.10, whereas the second term is bounded by $|\mathcal{I}|$ since

$$\|f_t\|_2^2 = \left\|(\mathbf{s}(i,y_t))_{i\in\mathcal{I}}\right\|_2^2 = \sum_{i\in\mathcal{I}} \|\mathbf{s}(i,y_t)\|_2^2 \leqslant |\mathcal{I}|\,.$$

Therefore we have

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|X_t(k,a)\|_2^2 \,\middle|\, \mathcal{G}_t\right] \leqslant \frac{2\,|\mathcal{I}|^2}{\gamma}.$$

We can now apply Corollary A.0.5 with $M = 2\,|\mathcal{I}|\,/\gamma$ and $V = 2\,|\mathcal{I}|^2\,/\gamma$ to get:

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{T} X_t(k,a)\right\|_2\right] \leqslant \frac{8\,|\mathcal{I}|}{\sqrt{T\gamma}} + \frac{8\,|\mathcal{I}|}{3T\gamma}.$$

Besides, it follows from the definition of $X_t(k,a)$ that

$$\frac{1}{T}\sum_{t=1}^{T} X_t(k,a) = \lambda_T(k,a)\left(\bar{\hat{f}}_T(k,a) - \bar{f}_T(k,a)\right).$$

Finally, by summing over k and a, we obtain:

$$\mathbb{E}\left[\sum_{\substack{k\in\mathcal{K}\\ a\in\mathcal{A}}} \lambda_T(k,a)\left\|\left(\bar{\hat{f}}_T(k,a) - \bar{f}_T(k,a)\right)\right\|_2\right] \leqslant |\mathcal{I}|\,|\mathcal{K}|\,|\mathcal{A}|\left(\frac{8}{\sqrt{T\gamma}} + \frac{8}{3T\gamma}\right).$$

□

VI.5.5. Average estimator $\bar{\tilde{f}}_{\mathrm{T}}(k,a)$ is close to $\mathscr{F}_c^k$

Lemma VI.5.7.

$$\mathbb{E}\left[\sum_{\substack{k\in\mathscr{K}\\a\in\mathscr{A}}} \mathbf{d}_2\left(\bar{\tilde{g}}_{\mathrm{T}}^{ka},\ \mathscr{F}_c^k\right)\right] \leqslant \sqrt{|\mathscr{K}|\,|\mathscr{A}|}\left(\frac{1}{2\eta\mathrm{T}}+\frac{\eta\,|\mathscr{I}|^2}{2\gamma}\right)$$

Proof. Consider the set $\tilde{\mathscr{Z}}_0$ defined by

$$\tilde{\mathscr{Z}}_0 := \prod_{k\in\mathscr{K}} \left((\mathscr{F}_c^k)^\circ \cap \mathscr{B}_2\right)^{\mathscr{A}},$$

and let us assume for the moment that the following inclusion holds:

$$\tilde{\mathscr{Z}}_0 \subset \sqrt{|\mathscr{K}|\,|\mathscr{A}|}\cdot\tilde{\mathscr{Z}}. \tag{VI.3}$$

For each $k\in\mathscr{K}$ and $a\in\mathscr{A}$, $\mathscr{F}_c^k$ being a closed convex cone, Proposition IV.2.10 gives the following expression of the distance of $\bar{\tilde{g}}_{\mathrm{T}}^{ka}$ to $\mathscr{F}_c^k$:

$$\mathbf{d}_2\left(\bar{\tilde{g}}_{\mathrm{T}}^{ka},\ \mathscr{F}_c^k\right) = \max_{\tilde{z}^{ka}\in(\mathscr{F}_c^k)^\circ\cap\mathscr{B}_2} \left\langle \bar{\tilde{g}}_{\mathrm{T}}^{ka}\middle|\tilde{z}^{ka}\right\rangle.$$

By summing over k and a, we have:

$$\begin{aligned}\sum_{\substack{k\in\mathscr{K}\\a\in\mathscr{A}}} \mathbf{d}_2\left(\bar{\tilde{g}}_{\mathrm{T}}^{ka},\ \mathscr{F}_c^k\right) &= \sum_{\substack{k\in\mathscr{K}\\a\in\mathscr{A}}} \max_{\tilde{z}^{ka}\in(\mathscr{F}_c^k)^\circ\cap\mathscr{B}_2} \left\langle \bar{\tilde{g}}_{\mathrm{T}}^{ka}\middle|\tilde{z}^{ka}\right\rangle = \max_{\tilde{z}\in\tilde{\mathscr{Z}}_0} \sum_{\substack{k\in\mathscr{K}\\a\in\mathscr{A}}} \left\langle \bar{\tilde{g}}_{\mathrm{T}}\middle|\tilde{z}\right\rangle\\ &\leqslant \sqrt{|\mathscr{K}|\,|\mathscr{A}|}\cdot\max_{\tilde{z}\in\tilde{\mathscr{Z}}} \left\langle \bar{\tilde{g}}_{\mathrm{T}}\middle|\tilde{z}\right\rangle = \sqrt{|\mathscr{K}|\,|\mathscr{A}|}\cdot\mathbf{d}_2\left(\bar{\tilde{g}}_{\mathrm{T}},\ \tilde{\mathscr{C}}\right),\end{aligned}$$

where for the inequality we used inclusion (VI.3), and for the last equality Proposition IV.2.10 together with the fact that $\tilde{\mathscr{Z}} = \tilde{\mathscr{C}}^\circ\cap\mathscr{B}_2$ by definition. Taking the expectation and substituting distance $\mathbf{d}_2(\bar{\tilde{g}}_{\mathrm{T}},\ \tilde{\mathscr{C}})$ by the bound from Lemma VI.5.3 yields the result.

Let us now prove inclusion (VI.3). Let $\tilde{z} = (\tilde{z}^{ka})_{\substack{k\in\mathscr{K}\\a\in\mathscr{A}}} \in \tilde{\mathscr{Z}}_0$. First, let us prove that $\tilde{z}\in\tilde{\mathscr{C}}^\circ$. Let $\tilde{g}\in\tilde{\mathscr{C}}$. We can write

$$\langle\tilde{g}|\tilde{z}\rangle = \sum_{\substack{k\in\mathscr{K}\\a\in\mathscr{A}}} \left\langle \tilde{z}^{ka}\middle|\tilde{g}^{ka}\right\rangle.$$

But for each $k \in \mathcal{K}$ and $a \in \mathcal{A}$, by definition of $\tilde{\mathcal{Z}}_0$, we have $\tilde{z}^{ka} \in (\mathcal{F}_c^k)^\circ$, and since $\tilde{\mathcal{C}} \subset \prod_{k\in\mathcal{K}} (\mathcal{F}_c^k)^{\mathcal{A}}$ by definition, we also have $\tilde{g}^{ka} \in \mathcal{F}_c^k$. Therefore, $\langle \tilde{g}^{ka} | \tilde{z}^{ka} \rangle \leqslant 0$ and consequently, $\langle \tilde{g} | \tilde{z} \rangle \leqslant 0$. This proves $\tilde{\mathcal{Z}}_0 \subset \tilde{\mathcal{C}}^\circ$.

Let $\tilde{z} \in \tilde{\mathcal{Z}}_0$. By definition of $\tilde{\mathcal{Z}}_0$, we have $\left\|\tilde{z}^{ka}\right\|_2 \leqslant 1$ for all $k \in \mathcal{K}$ and $a \in \mathcal{A}$. Thus

$$\left\|\tilde{z}\right\|_2 = \sqrt{\sum_{\substack{k\in\mathcal{K}\\ a\in\mathcal{A}}} \left\|\tilde{z}^{ka}\right\|_2^2} \leqslant \sqrt{|\mathcal{K}|\,|\mathcal{A}|},$$

and therefore $\tilde{\mathcal{Z}}_0 \subset \sqrt{|\mathcal{K}|\,|\mathcal{A}|} \cdot \mathcal{B}_2$. Finally, we have

$$\tilde{\mathcal{Z}}_0 \subset \tilde{\mathcal{C}}^\circ \cap \sqrt{|\mathcal{K}|\,|\mathcal{A}|} \cdot \mathcal{B}_2 = \sqrt{|\mathcal{K}|\,|\mathcal{A}|} \cdot \tilde{\mathcal{Z}}.$$

□

VI.5.6. $\mathbf{r}^{[k]}(a, \bar{\hat{f}}_{\mathrm{T}}(k,a))$ is close to $\mathbf{r}(a, \bar{f}_{\mathrm{T}}(k,a))$

Lemma VI.5.8.

$$\mathbb{E}\left[\sum_{\substack{k\in\mathcal{K}\\ a\in\mathcal{A}}} \lambda_{\mathrm{T}}(k,a) \left\|\mathbf{r}(a, \bar{f}_{\mathrm{T}}(k,a)) - \mathbf{r}^{[k]}(a, \bar{\hat{f}}_{\mathrm{T}}(k,a))\right\|_2\right] \leqslant \mathrm{L}_{\mathbf{r}}\,|\mathcal{I}|\,|\mathcal{K}|\,|\mathcal{A}| \left(\frac{8}{\sqrt{\mathrm{T}\gamma}} + \frac{8}{3\mathrm{T}\gamma}\right) + \mathrm{L}_{\mathbf{r}}\sqrt{|\mathcal{K}|\,|\mathcal{A}|}\left(\frac{1}{\eta\mathrm{T}} + \frac{\eta\,|\mathcal{I}|^2}{\gamma}\right).$$

Proof. Let $(k,a) \in \mathcal{K} \times \mathcal{A}$ and denote $f := \bar{f}_{\mathrm{T}}(k,a)$ and $\hat{f} := \bar{\hat{f}}_{\mathrm{T}}(k,a)$ to alleviate notation. Denote $\mathbf{P}^{[k]}$ the Euclidean projection onto $\mathcal{F}_c^k$. Then of course $\mathbf{P}^{[k]}(\hat{f})$ belongs to $\mathcal{F}_c^k$, and since $\mathbf{r}(a,\,\cdot\,)$ and $\mathbf{r}^{[k]}(a,\,\cdot\,)$ coincide on $\mathcal{F}_c^k$ by Lemma VI.4.5, we can write

$$\mathbf{r}(a,f) - \mathbf{r}^{[k]}(a,\hat{f}) = \mathbf{r}(a,f) - \mathbf{r}(a,\hat{f}) + \mathbf{r}(a,\hat{f}) - \mathbf{r}(a,\mathbf{P}^{[k]}(\hat{f})) + \mathbf{r}^{[k]}(a,\mathbf{P}^{[k]}(\hat{f})) - \mathbf{r}^{[k]}(a,\hat{f}).$$

Thus, by taking the norm and using the triangle inequality and the Lipschitz constant $\mathrm{L}_{\mathbf{r}}$ which is common to $\mathbf{r}(a,\,\cdot\,)$ and $\mathbf{r}^{[k]}(a,\,\cdot\,)$ to get

$$\left\|\mathbf{r}(a,f) - \mathbf{r}^{[k]}(a,\hat{f})\right\|_2 \leqslant \mathrm{L}_{\mathbf{r}}\left(\left\|f-\hat{f}\right\|_2 + 2\cdot\mathbf{d}_2\left(\hat{f},\,\mathcal{F}_c^k\right)\right).$$

We now multiply by $\lambda_{\mathrm{T}}(k,a)$. The last term in the above right-hand side is transformed as

$$2\lambda_{\mathrm{T}}(k,a)\cdot\mathbf{d}_2\left(\hat{f},\,\mathcal{F}_c^k\right) = 2\cdot\mathbf{d}_2\left(\lambda_{\mathrm{T}}(k,a)\hat{f},\,\mathcal{F}_c^k\right) = 2\cdot\mathbf{d}_2\left(\bar{\tilde{g}}_{\mathrm{T}}^{ka},\,\mathcal{F}_c^k\right),$$

where used the fact that $\mathscr{F}_c^k$ is a convex cone to push the factor $\lambda_T(k,a)$ into the distance. Therefore,

$$\lambda_T(k,a)\left\|\mathbf{r}(a,f)-\mathbf{r}^{[k]}(a,\widehat{f})\right\|_2 \leqslant L_{\mathbf{r}}\cdot\lambda_T(k,a)\left\|f-\widehat{f}\right\|_2 + 2L_{\mathbf{r}}\cdot\mathbf{d}_2\left(\tilde{\bar{g}}_T^{ka},\ \mathscr{F}_c^k\right).$$

Finally, we get the result by taking the expectation, summing over k and a, and plugging Lemmas VI.5.6 and VI.5.7. □

VI.5.7. g is closer to $\mathbb{R}^d_-$ than r

Lemma VI.5.9.

$$\mathbf{d}_2\left(\sum_{\substack{k\in\mathscr{K}\\a\in\mathscr{A}}}\lambda_T(k,a)\cdot\mathbf{g}(a,\bar{y}_T(k,a)),\ \mathbb{R}^d_-\right) \leqslant \mathbf{d}_2\left(\sum_{\substack{k\in\mathscr{K}\\a\in\mathscr{A}}}\lambda_T(k,a)\cdot\mathbf{r}(a,\bar{f}_T(k,a)),\ \mathbb{R}^d_-\right).$$

Proof. Let $k\in\mathscr{K}$ and $a\in\mathscr{A}$. First note that $\mathbf{f}(\bar{y}_T(k,a)) = \bar{f}_T(k,a)$. Indeed, using the affinity of $\mathbf{f}$,

$$\begin{aligned}\mathbf{f}(\bar{y}_T(k,a)) &= \mathbf{f}\left(\frac{1}{|N_T(k,a)|}\sum_{t\in N_T(k,a)} y_t\right) = \frac{1}{|N_T(k,a)|}\sum_{t\in N_T(k,a)}\mathbf{f}(y_t)\\ &= \frac{1}{|N_T(k,a)|}\sum_{t\in N_T(k,a)} f_t = \bar{f}_T(k,a).\end{aligned}$$

For each component $n\in\{1,\dots,d\}$, we have $\mathbf{g}^n(a,\bar{y}_T(k,a)) \leqslant \mathbf{r}^n(a,\bar{f}_T(k,a))$ by property (i) in Proposition VI.4.4. Finally, using the explicit expression of the Euclidean distance to $\mathbb{R}^d_-$, we have

$$\begin{aligned}\mathbf{d}_2\left(\sum_{\substack{k\in\mathscr{K}\\a\in\mathscr{A}}}\lambda_T(k,a)\cdot\mathbf{g}(a,\bar{y}_T(k,a)),\ \mathbb{R}^d_-\right) &= \sqrt{\sum_{n=1}^d\left(\sum_{\substack{k\in\mathscr{K}\\a\in\mathscr{A}}}\lambda_T(k,a)\cdot\mathbf{g}^n(a,\bar{y}_T(k,a))\right)_+^2}\\ &\leqslant \sqrt{\sum_{n=1}^d\left(\sum_{\substack{k\in\mathscr{K}\\a\in\mathscr{A}}}\lambda_T(k,a)\cdot\mathbf{r}^n(a,\bar{f}_T(k,a))\right)_+^2}\\ &= \mathbf{d}_2\left(\sum_{\substack{k\in\mathscr{K}\\a\in\mathscr{A}}}\lambda_T(k,a)\cdot\mathbf{r}(a,\bar{f}_T(k,a)),\ \mathbb{R}^d_-\right).\end{aligned}$$

□

VI.5.8. Decomposition of $\mathbf{g}(a_t, y_t)$ with respect to the realized auxiliary action (k_t, a_t)

Lemma VI.5.10.

$$\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} \mathbf{g}(a_t, y_t) = \sum_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}} \lambda_{\mathrm{T}}(k,a)\cdot \mathbf{g}(a, \bar{y}_{\mathrm{T}}(k,a))$$

Proof. Using the definitions of $\mathrm{N}_{\mathrm{T}}(k,a)$ and $\lambda_{\mathrm{T}}(k,a)$, and the linearity of $\mathbf{g}(a,\,\cdot\,)$, we have

$$\begin{aligned}
\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} \mathbf{g}(a_t, y_t) &= \frac{1}{\mathrm{T}} \sum_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}} \sum_{t\in \mathrm{N}_{\mathrm{T}}(k,a)} \mathbf{g}(a, y_t)\\
&= \sum_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}} \frac{|\mathrm{N}_{\mathrm{T}}(k,a)|}{\mathrm{T}} \cdot \frac{1}{|\mathrm{N}_{\mathrm{T}}(k,a)|} \sum_{t\in \mathrm{N}_{\mathrm{T}}(k,a)} \mathbf{g}(a, y_t)\\
&= \sum_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}} \lambda_{\mathrm{T}}(k,a)\cdot \mathbf{g}(a, \bar{y}_{\mathrm{T}}(k,a)).
\end{aligned}$$

□

VI.5.9. From $\mathbf{g}(i_t, j_t)$ to $\mathbf{g}(a_t, y_t)$

Lemma VI.5.11.

$$\mathbb{E}\left[\left\|\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} \mathbf{g}(i_t, j_t) - \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} \mathbf{g}(a_t, y_t)\right\|_2\right] \leqslant \frac{2\sqrt{\pi}\,\|\mathbf{g}\|_2}{\sqrt{\mathrm{T}}} + 2\gamma\,\|\mathbf{g}\|_2 .$$

Proof. Consider the process $(\mathrm{X}_t)_{t\geqslant 1}$ defined by

$$\mathrm{X}_t = \mathbf{g}(i_t, j_t) - (1-\gamma)\mathbf{g}(a_t, y_t) - \gamma \mathbf{g}(u, y_t),$$

and the filtration $(\mathscr{G}'_t)_{t\geqslant 1}$ where $\mathscr{G}'_t$ is generated by

$$(k_1, a_1, y_1, i_1, s_1, \dots, k_{t-1}, a_{t-1}, y_{t-1}, i_{t-1}, s_{t-1}, k_t, a_t, y_t).$$

$(\mathrm{X}_t)_{t\geqslant 1}$ is martingale difference sequence with respect to filtration $(\mathscr{G}'_t)_{t\geqslant 1}$. Indeed, knowing $\mathscr{G}'_t$, the law of i_t is $(1-\gamma)a_t + \gamma u$ by definition of the strategy, and thus the law of (i_t, j_t) is $((1-\gamma)a_t + \gamma u)\otimes y_t$. We can then write, by bilinearity of $\mathbf{g}$:

$$\mathbb{E}\left[\mathbf{g}(i_t, j_t) \mid \mathscr{G}'_t\right] = (1-\gamma)\mathbf{g}(a_t, y_t) + \gamma \mathbf{g}(u, y_t).$$

Moreover, $\|X_t\|_2$ is always bounded by $2\|\mathbf{g}\|_2$:

$$\begin{aligned}\|X_t\|_2 &= \|(1-\gamma)(\mathbf{g}(i_t, j_t) - \mathbf{g}(a_t, y_t)) + \gamma(\mathbf{g}(i_t, j_t) - \mathbf{g}(u, y_t))\|_2 \\ &\leqslant (1-\gamma)\|\mathbf{g}(i_t, j_t) - \mathbf{g}(a_t, y_t)\|_2 + \gamma\|\mathbf{g}(i_t, j_t) - \mathbf{g}(u, y_t)\|_2 \\ &\leqslant 2\|\mathbf{g}\|_2 .\end{aligned}$$

We can thus apply Corollary A.0.3 with $M = 2\|\mathbf{g}\|_2$ to get

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{T} X_t\right\|_2\right] \leqslant \frac{2\sqrt{\pi}\|\mathbf{g}\|_2}{\sqrt{T}}.$$

Therefore,

$$\begin{aligned}\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{g}(i_t, j_t) - \frac{1}{T}\sum_{t=1}^{T}\mathbf{g}(a_t, y_t)\right\|_2 &= \left\|\frac{1}{T}\sum_{t=1}^{T}(X_t + \gamma(\mathbf{g}(u, y_t) - \mathbf{g}(a_t, y_t)))\right\|_2 \\ &\leqslant \left\|\frac{1}{T}\sum_{t=1}^{T} X_t\right\|_2 + \left\|\frac{\gamma}{T}\sum_{t=1}^{T}(\mathbf{g}(u, y_t) - \mathbf{g}(a_t, y_t))\right\|_2 \\ &\leqslant \left\|\frac{1}{T}\sum_{t=1}^{T} X_t\right\|_2 + 2\gamma\|\mathbf{g}\|_2 ,\end{aligned}$$

And taking the expectation:

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{g}(i_t, j_t) - \frac{1}{T}\sum_{t=1}^{T}\mathbf{g}(a_t, y_t)\right\|_2\right] \leqslant \frac{2\sqrt{\pi}\|\mathbf{g}\|_2}{\sqrt{T}} + 2\gamma\|\mathbf{g}\|_2 .$$

□

VI.5.10. Final bound

We now combine the above lemmas in the order specified at the beginning of the section to get:

$$\begin{aligned}\mathbb{E}\left[\mathbf{d}_2\left(\bar{g}_T, \mathbb{R}^d_-\right)\right] \leqslant \frac{2\sqrt{\pi}\|\mathbf{g}\|_2}{\sqrt{T}} + 2\gamma\|\mathbf{g}\|_2 + L_{\mathbf{r}}|\mathcal{I}||\mathcal{K}||\mathcal{A}|\left(\frac{8}{\sqrt{T\gamma}} + \frac{8}{3T\gamma}\right) \\ + \frac{3L_{\mathbf{r}}}{2}\sqrt{|\mathcal{K}||\mathcal{A}|}\left(\frac{1}{\eta T} + \frac{\eta|\mathcal{I}|^2}{\gamma}\right).\end{aligned}$$

Injecting the values of η and γ yields the result.

VI.6. Outcome-dependent signals

This section studies the special case where the law $\mathbf{s}(i, j)$ of the signal does not depend on the pure action i of the Decision Maker. In other words, we assume that

$$\mathbf{s}(\cdot, j) \quad \text{is constant,} \quad \text{for all } j \in \mathcal{J}.$$

We aim at constructing a strategy which achieves a $O(T^{-1/2})$ convergence rate. Again, we assume that the target set is the negative orthant $\mathbb{R}^d_-$ and that it is approachable. We will heavily rely on elements from the previous sections. To take advantage of the above assumption, the strategy from Section VI.4 will be modified in two ways. First, the estimator will be simpler since exploration is unnecessary, and second, the mixed action of the Decision Maker will not be perturbed with the uniform distribution. Unless stated otherwise, all previous notation and assumptions stand.

The modified strategy is defined as follows. Let $\eta > 0$ be a parameter. For $1 \leqslant t \leqslant T$;

- compute $\tilde{z}_t = \mathbf{P}_{\tilde{\mathcal{Z}}}\left(\eta \sum_{s=1}^{t-1} \tilde{g}_s\right)$ and $\tilde{x}_t := \tilde{\mathbf{x}}(\tilde{z}_t) \in \Delta(\mathcal{K} \times \mathcal{A})$.
- draw $(k_t, a_t) \sim \tilde{x}_t$ and then $i_t \sim a_t$;
- observe signal $s_t \in \mathcal{S}$ and compute estimator

$$\widehat{f}_t = \left(\delta_{s_t}\right)_{i \in \mathcal{J}} \qquad \in \mathbb{R}^{\mathcal{S} \times \mathcal{J}};$$

- set $\tilde{g}_t = \tilde{\mathbf{g}}((k_t, a_t), \widehat{f}_t)$.

The definition of the strategy implies that the law of i_t knowing $\mathcal{G}_t$ is a_t. Let us state the properties of the new estimator.

Lemma VI.6.1. *For $t \geqslant 1$,*

(i) $\mathbb{E}\left[\widehat{f}_t \,\middle|\, \mathcal{G}_t\right] = \mathbb{E}\left[f_t \,|\, \mathcal{G}_t\right]$;

(ii) $\left\|\widehat{f}_t\right\|_2^2 = |\mathcal{J}|$.

Theorem VI.6.2. *Let $T \geqslant 1$. Against any strategy of Nature, the above strategy with parameter $\eta = (T\,|\mathcal{J}|)^{-1/2}$ guarantees*

$$\mathbb{E}\left[\mathbf{d}_2\left(\bar{g}_T, \mathbb{R}^d_-\right)\right] \leqslant \frac{2\sqrt{\pi}\left(\|\mathbf{g}\|_2 + 2L_{\mathbf{r}}\sqrt{|\mathcal{J}|}\,|\mathcal{K}|\,|\mathcal{A}|\right)}{T^{1/2}}.$$

One can check that statements from Lemmas VI.5.4, VI.5.5, VI.5.9 and VI.5.10 still hold. We state and prove below new versions of the remaining lemmas, which were affected by the modifications of the estimator and the law of i_t. The analysis can be summarized as follows.

$$\begin{aligned}
\bar{g}_{\mathrm{T}} \quad \text{is close to} \quad & \frac{1}{\mathrm{T}} \sum_{t=1}^{\mathrm{T}} g(a_t, y_t) && \text{(Lemma VI.6.7)} \\
\text{is equal to} \quad & \sum_{\substack{k \in \mathcal{K} \\ a \in \mathcal{A}}} \lambda_{\mathrm{T}}(k,a) \cdot \mathbf{g}(a, \bar{y}_{\mathrm{T}}(k,a)) && \text{(Lemma VI.5.10)} \\
\text{is closer to } \mathbb{R}^d_- \text{ than} \quad & \sum_{\substack{k \in \mathcal{K} \\ a \in \mathcal{A}}} \lambda_{\mathrm{T}}(k,a) \cdot \mathbf{r}(a, \bar{f}_{\mathrm{T}}(k,a)) && \text{(Lemma VI.5.9)} \\
\text{is close to} \quad & \sum_{\substack{k \in \mathcal{K} \\ a \in \mathcal{A}}} \lambda_{\mathrm{T}}(k,a) \cdot \mathbf{r}^{[k]}(a, \bar{\bar{f}}_{\mathrm{T}}(k,a)) && \text{(Lemma VI.6.6)} \\
\text{is equal to} \quad & \mathbf{R}(\bar{\bar{g}}_{\mathrm{T}}) && \text{(Lemma VI.5.5)} \\
\text{is close to} \quad & \mathbb{R}^d_- && \text{(Lemmas VI.5.4 and VI.6.3).}
\end{aligned}$$

VI.6.1. Average auxiliary payoff $\bar{\bar{g}}_{\mathrm{T}}$ is close to auxiliary target set $\tilde{\mathscr{C}}$

Lemma VI.6.3.

$$\mathbb{E}\left[\mathbf{d}_2\left(\bar{\bar{g}}_{\mathrm{T}}, \tilde{\mathscr{C}}\right)\right] \leqslant \frac{1}{2\eta \mathrm{T}} + \frac{\eta |\mathcal{I}|}{2}.$$

Proof. We follow the proof of Lemma VI.5.4. The regret bound given by TheoremI.3.1 still holds:

$$\max_{\tilde{z} \in \tilde{\mathcal{Z}}} \sum_{t=1}^{\mathrm{T}} \langle \tilde{g}_t | \tilde{z} \rangle - \sum_{t=1}^{\mathrm{T}} \langle \tilde{g}_t | \tilde{z}_t \rangle \leqslant \frac{1}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{\mathrm{T}} \|\tilde{g}_t\|_2^2 .$$

In Lemma VI.5.4, the second sum was nonpositive in expectation thanks to the fact that $\mathbb{E}\left[\widehat{f}_t \,\middle|\, \mathscr{G}_t\right] = \mathbb{E}\left[f_t \,|\, \mathscr{G}_t\right]$. The same reasoning can be applied in the present case since the property of the estimator is guaranteed by Lemma VI.6.1. Therefore, we have

$$\mathbb{E}\left[\mathbf{d}_2\left(\bar{\bar{g}}_{\mathrm{T}}, \tilde{\mathrm{C}}\right)\right] \leqslant \frac{1}{2\eta \mathrm{T}} + \frac{\eta}{2\mathrm{T}} \mathbb{E}\left[\sum_{t=1}^{\mathrm{T}} \|\tilde{g}_t\|_2^2\right].$$

Then, for $1 \leqslant t \leqslant \mathrm{T}$, we have

$$\|\tilde{g}_t\|_2^2 = \left\|\tilde{\mathbf{g}}((k_t, a_t), \widehat{f}_t)\right\|_2^2 = \left\|\left(\mathbb{1}_{\{k_t = k,\, a_t = a\}} \widehat{f}_t\right)_{\substack{k \in \mathcal{K} \\ a \in \mathcal{A}}}\right\|_2^2 = \left\|\widehat{f}_t\right\|_2^2 = |\mathcal{I}|,$$

where we used property (ii) from Lemma VI.6.1 for the last equality. The result follows. □

VI.6.2. Average estimator $\bar{\hat{f}}_{\mathrm{T}}(k,a)$ is close to average flag $\bar{f}_{\mathrm{T}}(k,a)$

Lemma VI.6.4.

$$\mathbb{E}\left[\sum_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}}\lambda_{\mathrm{T}}(k,a)\left\|\bar{\hat{f}}_{\mathrm{T}}(k,a)-\bar{f}_{\mathrm{T}}(k,a)\right\|_2\right]\leqslant 2\,|\mathscr{K}|\,|\mathscr{A}|\sqrt{\frac{\pi\,|\mathscr{T}|}{\mathrm{T}}}.$$

Proof. Let $k\in\mathscr{K}$ and $a\in\mathscr{A}$. As in Lemma VI.5.6, we consider

$$\mathrm{X}_t(k,a):=\mathbb{1}_{\{k_t=k,\,a_t=a\}}\left(\hat{f}_t-f_t\right),$$

which is a sequence of martingale differences with respect to filtration $(\mathscr{G}_t)_{t\geqslant 1}$ thanks to property (i) from Lemma VI.6.1. But this time, we use Corollary A.0.3 instead of Corollary A.0.5. Each X_t is bounded as follows

$$\|\mathrm{X}_t(k,a)\|_2\leqslant\left\|\hat{f}_t\right\|_2+\|f_t\|_2=\sqrt{|\mathscr{T}|}+\sqrt{\sum_{i\in\mathscr{T}}\|\mathbf{s}(i,y_t)\|_2^2}\leqslant 2\sqrt{|\mathscr{T}|},$$

where we used property (ii) from Lemma VI.6.1. Corollary A.0.3 then gives

$$\mathbb{E}\left[\lambda_{\mathrm{T}}(k,a)\left\|\bar{\hat{f}}_{\mathrm{T}}(k,a)-\bar{f}_{\mathrm{T}}(k,a)\right\|_2\right]=\mathbb{E}\left[\left\|\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\mathrm{X}_t(k,a)\right\|_2\right]\leqslant 2\sqrt{\frac{\pi\,|\mathscr{T}|}{\mathrm{T}}}.$$

The result follows by summing over $k\in\mathscr{K}$ and $a\in\mathscr{A}$. □

VI.6.3. Average estimator $\bar{\hat{f}}_{\mathrm{T}}(k,a)$ is close to $\mathscr{F}_c^k$

Lemma VI.6.5.

$$\mathbb{E}\left[\sum_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}}\mathbf{d}_2\left(\bar{\tilde{g}}_{\mathrm{T}}^{ka},\ \mathscr{F}_c^k\right)\right]\leqslant\sqrt{|\mathscr{K}|\,|\mathscr{A}|}\left(\frac{1}{2\eta\mathrm{T}}+\frac{\eta\,|\mathscr{T}|}{2}\right).$$

Proof. The following inequality from the proof of Lemma VI.5.7 still holds

$$\sum_{\substack{k\in\mathscr{K}\\ a\in\mathscr{A}}}\mathbf{d}_2\left(\bar{\tilde{g}}_{\mathrm{T}}^{ka},\ \mathscr{F}_c^k\right)\leqslant\sqrt{|\mathscr{K}|\,|\mathscr{A}|}\cdot\mathbf{d}_2\left(\bar{\tilde{g}}_{\mathrm{T}},\ \tilde{\mathscr{C}}\right).$$

Then, taking the expectation and injecting the new bound on $\mathbb{E}\left[\mathbf{d}_2\left(\bar{\tilde{g}}_{\mathrm{T}},\ \tilde{\mathscr{C}}\right)\right]$ given by Lemma VI.6.3 yields the result. □

VI.6.4. $\mathbf{r}^{[k]}(a, \bar{\hat{f}}_T(k,a))$ is close to $\mathbf{r}(a, \bar{f}_T(k,a))$

Lemma VI.6.6. *For all $k \in \mathcal{K}$ and $a \in \mathcal{A}$,*

$$\mathbb{E}\left[\sum_{\substack{k\in\mathcal{K}\\a\in\mathcal{A}}} \lambda_T(k,a)\left\|\mathbf{r}(a,\bar{f}_T(k,a)) - \mathbf{r}^{[k]}(a,\bar{\hat{f}}_T(k,a))\right\|_2\right] \leqslant 2L_{\mathbf{r}}\,|\mathcal{K}|\,|\mathcal{A}|\sqrt{\frac{\pi\,|\mathcal{I}|}{T}} + L_{\mathbf{r}}\sqrt{|\mathcal{K}|\,|\mathcal{A}|}\left(\frac{1}{\eta T} + \eta\,|\mathcal{I}|\right).$$

Proof. Let $k \in \mathcal{K}$ and $a \in \mathcal{A}$. Using notation $f = \bar{f}_T(k,a)$ and $\hat{f} = \bar{\hat{f}}_T(k,a)$, the following inequality from the proof of Lemma VI.5.8 still holds

$$\lambda_T(k,a)\left\|\mathbf{r}(a,f) - \mathbf{r}^{[k]}(a,\hat{f})\right\|_2 \leqslant L_{\mathbf{r}}\cdot\lambda_T(k,a)\left\|f-\hat{f}\right\|_2 + 2L_{\mathbf{r}}\cdot\mathbf{d}_2\left(\bar{\tilde{g}}_T^{ka},\ \mathscr{F}_c^k\right).$$

The result follows from taking the expectation, summing over $k \in \mathcal{K}$ and $\mathcal{A}$, and injecting the bounds from Lemmas VI.6.4 and VI.6.5. □

VI.6.5. From $\mathbf{g}(i_t, j_t)$ to $\mathbf{g}(a_t, y_t)$

Lemma VI.6.7.

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{g}(i_t,j_t) - \frac{1}{T}\sum_{t=1}^{T}\mathbf{g}(a_t,y_t)\right\|_2\right] \leqslant \frac{2\sqrt{\pi}\,\|\mathbf{g}\|_2}{\sqrt{T}}.$$

Proof. The process $(\mathbf{g}(i_t,j_t) - \mathbf{g}(a_t,y_t))_{t\geqslant 1}$ is a martingale difference sequence with respect to filtration $(\mathscr{G}'_t)_{t\geqslant 1}$ introduced in the proof of Lemma VI.5.11. It is moreover bounded by $2\,\|\mathbf{g}\|_2$. Therefore, Corollary A.0.3 gives:

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{g}(i_t,j_t) - \frac{1}{T}\sum_{t=1}^{T}\mathbf{g}(a_t,y_t)\right\|_2\right] \leqslant \frac{2\sqrt{\pi}\,\|\mathbf{g}\|_2}{\sqrt{T}}.$$

□

VI.6.6. Final bound

Similarly to the proof of Theorem VI.5.1, the combination of the above lemmas gives:

$$\mathbb{E}\left[\mathbf{d}_2\left(\bar{g}_T,\ \mathbb{R}^d_-\right)\right] \leqslant \frac{2\sqrt{\pi}\,\|\mathbf{g}\|_2}{\sqrt{T}} + \frac{2\sqrt{\pi}L_{\mathbf{r}}\,|\mathcal{K}|\,|\mathcal{A}|\,\sqrt{|\mathcal{I}|}}{\sqrt{T}} + \frac{3L_{\mathbf{r}}\sqrt{|\mathcal{K}|\,|\mathcal{A}|}}{2}\left(\frac{1}{\eta T} + \eta\,|\mathcal{I}|\right).$$

Injecting the value of η yields the result.

VI.7. Discussion

VI.7.1. Computational efficiency

We discuss the computational efficiency of the strategies studied in Sections VI.5 and VI.6. The following arguments hold for both.

The first step of the strategy is the computation of $\tilde{z}_t$ which consists of an Euclidean projection onto $\tilde{\mathcal{Z}} := \tilde{\mathcal{C}}^\circ \cap \mathcal{B}_2$, which is efficient. Indeed, $\tilde{\mathcal{C}}^\circ$ being a closed convex cone, the Euclidean projection onto $\tilde{\mathcal{Z}}$ can be immediately deduced from the Euclidean projection onto $\tilde{\mathcal{C}}^\circ$. The latter projection can be efficiently computed since $\tilde{\mathcal{C}}^\circ$ is a polytope (as it can be easily checked). The second step is the computation of $\tilde{x}_t := \tilde{\mathbf{x}}(\tilde{z}_t)$ which, according to the definition of $\tilde{\mathbf{x}}$ in Proposition VI.4.9, can be computed by solving the following minimax problem:

$$\min_{\tilde{x} \in \Delta(\mathcal{H} \times \mathcal{A})} \max_{f \in \mathcal{F}} \langle \tilde{\mathbf{g}}(\tilde{x}, f) | \tilde{z}_t \rangle \,.$$

The sets $\Delta(\mathcal{H} \times \mathcal{A})$ and $\mathcal{F}$ being polytopes, this can be solved efficiently using e.g. linear programming. Then, the computations of estimator $\hat{f}_t$ and auxiliary payoff $\tilde{g}_t$ are easy.

Therefore, the whole strategy can be efficiently computed. Moreover, the per-step complexity is constant.

VI.7.2. Uniform guarantee over time

To achieve the guarantee given in Theorem VI.5.1 the time-horizon T must be known in advance in order to tune the parameters η and γ accordingly. Let us quickly explain how to obtain a strategy with a convergence guarantee of the same order that holds uniformly over time, without resorting to a doubling trick.

We first deal with parameter η. As explained in Section IV.5, it is always possible to choose an oracle $\mathbf{x}$ such that condition (IV.1) is satisfied. Let us assume that this is the case. As shown in the proof of Lemma VI.5.3, $\tilde{z}_t = \mathbf{P}_{\tilde{\mathcal{Z}}}(\eta \sum_{s=1}^{t-1} \tilde{g}_s)$ can also be written

$$z_t = \operatorname*{arg\,max}_{\tilde{z} \in \tilde{\mathcal{Z}}} \left\{ \left\langle \eta \sum_{s=1}^{t-1} \tilde{g}_s \middle| \tilde{z} \right\rangle - \frac{1}{2} \|\tilde{z}\|_2^2 \right\},$$

which corresponds, according to Theorem IV.5.1, to Blackwell's strategy associated with target set $\tilde{\mathcal{C}}$ (which is closed convex cone), oracle $\tilde{x}$, and vector payoffs $\tilde{g}_t$. The same theorem assures that $\tilde{z}_t$ does not depend on the value of parameter η. Thus, the strategy can be run with any fixed value of η, and the bound from Lemma VI.5.3 would

still hold for any value of $\eta > 0$. Therefore, the parameter η need not be tuned as a function of the time-horizon T, because it does not have be chosen at all.

We now turn to exploration parameter γ. We modify the strategy by making it time-dependent:

$$\gamma_t = \min\left\{\gamma_0 t^{-1/3}, 1\right\},$$

where $\gamma_0 > 0$ is to be chosen later. Lemmas VI.5.4, VI.5.5, VI.5.9 and VI.5.10 are unaffected by this modification. Lemma VI.4.10 can be immediately adapted by replacing in the bounds γ by γ_t. Using the fact that γ_t is nonincreasing, the statements of Lemmas VI.5.3, VI.5.6, VI.5.7 and VI.5.8 can be adapted by replacing γ by γ_T. Finally, in Lemma VI.5.11, γ, which appears in the numerator, is replaced by

$$\frac{1}{T}\sum_{t=1}^{T}\gamma_t \leqslant \frac{1}{T}\sum_{t=1}^{T}\gamma_0 t^{-1/3} \leqslant \frac{3\gamma_0}{2}T^{-1/3} = \frac{3\gamma_T}{2}, \quad \text{for T large enough.}$$

Overall, combining the modified lemmas as in Section VI.5.10, we obtain a bound in which each term already has the expected dependency in T. Therefore, γ_0 can be tuned independently of T to eventually obtain a bound identical to Theorem VI.5.1 up to multiplicative constants.

VI.7.3. High probability guarantee and almost-sure convergence

Theorem VI.5.1 only provides a convergence guarantee in expectation. We quickly describe how the analysis can be adapted to obtain, for the same strategy, a high probability guarantee as well as almost-sure convergence.

We do not modify Lemmas VI.5.4, VI.5.5, VI.5.9 and VI.5.10 as they do not involve expectations.

The proof of Lemma VI.5.3 is modified as follows in order to obtain a high probability guarantee on $\mathbf{d}_2(\bar{\tilde{g}}_T, \tilde{\mathscr{C}})$. We can easily see that $(\langle \tilde{g}_t | \tilde{z}_t \rangle)_{t\geqslant 1}$ is a bounded sequence of super-martingale differences with respect to filtration $(\mathscr{H}_t)_{t\geqslant 1}$ and that $(\|\tilde{g}_t\|_2^2 - (|\mathcal{I}|/\gamma)^2)_{t\geqslant 1}$ is a bounded sequence of super-martingale differences with respect to $(\mathscr{G}_t)_{t\geqslant 1}$. Applying the Hoeffding–Azuma inequality (Proposition A.0.1) then gives the high probability version of the lemma.

The modification of Lemmas VI.5.6 and VI.5.11 is straightforward. We simply apply the high probability version of the involved concentration inequalities instead of the bounds in expectation: Proposition A.0.4 instead of Corollary A.0.5 and Proposition A.0.2 instead of Corollary A.0.3, respectively.

The high probability versions of Lemmas VI.5.7 and VI.5.8 immediately follow from those of Lemma VI.5.3, and Lemmas VI.5.6 and VI.5.7, respectively.

Then, the almost-sure convergence follow from a standard Borel-Cantelli argument.

VI.7.4. Using other regret minimizing strategies

As explained in the proof of Lemma VI.5.3, the strategy defined in Section VI.4.4 is based on a regret minimizing strategy, specifically, the Mirror Descent strategy associated with the Euclidean regularizer on $\tilde{\mathcal{Z}}$ and constant parameter η. As detailed in the proof, this strategy guarantees the following regret bound:

$$\max_{\tilde{z}\in\tilde{\mathcal{Z}}} \sum_{t=1}^{T} \langle \tilde{g}_t | \tilde{z} \rangle - \sum_{t=1}^{T} \langle \tilde{g}_t | \tilde{z}_t \rangle \leqslant \frac{1}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\tilde{g}_t\|_2^2 .$$

We can easily see that any regret minimizing strategy which guarantees a regret bound of the form

$$\max_{\tilde{z}\in\tilde{\mathcal{Z}}} \sum_{t=1}^{T} \langle \tilde{g}_t | \tilde{z} \rangle - \sum_{t=1}^{T} \langle \tilde{g}_t | \tilde{z}_t \rangle \leqslant \frac{A}{\eta} + B\eta \sum_{t=1}^{T} \|\tilde{g}_t\|_2^2$$

could be used to construct an alternative approachability strategy for the initial game, with the same rate of convergence. In particular, any Mirror Descent strategy from Section I.3 associated with some strongly convex regularizer on $\tilde{\mathcal{Z}}$ would be appropriate.

An interesting question is whether the choice of another regularizer would help improve the dependency in $|\mathcal{I}|$, $|\mathcal{H}|$ and $|\mathcal{A}|$ of the bound from Theorem VI.5.1. Note however that a general regularizer would not *a priori* retain the computational efficiency of the Euclidean regularizer (see Section VI.7.1).

VI.7.5. Comparison with [MPS14]

The strategy proposed in [MPS14] is computationally efficient and has a dimension-independent convergence rate of $O(T^{-1/5})$. We here highlight a few ideas which were already present in [MPS14], and those we have introduced in the present work to achieve an optimal convergence rate of $O(T^{-1/3})$.

[MPS14] already used the single-valued map $\mathbf{r}$ which is a simpler version of the set-valued map $\mathbf{m}$, which retains the key property characterizing the approachability of the target set (see Proposition VI.3.2 and property (ii) in Proposition VI.4.4). Besides, the decomposition of $\mathcal{F}$ and $\Delta(\mathcal{I})$ into polytopes was considered to obtain the piecewise-affinity of $\mathbf{r}$. This fundamental property was then used in the averaging of the flag estimators. The proposed strategy is constructed by dividing time into blocks of the same length: the Decision Maker plays a constant mixed action on each time block, which is used to average the flag estimators; and the Decision Maker changes his mixed action from one block to the other in order to achieve the convergence to the target set.

The strategy constructed in Section VI.4.4 manages to average the estimators and to approach the target *at the same time*, resulting in an improved (and optimal) convergence rate of $O(T^{-1/3})$. We enumerate some of the main ideas used to achieve this. First, we introduce the linear map $\mathbf{R}$ which allows to easily relate the auxiliary game and the initial game. In particular, it gives a simple comparison between a) the distance of the average payoff to the target set in the initial game and b) the distance of the average auxiliary payoff to the auxiliary target set (Lemma VI.5.4). Moreover, it combines well with the use of convex cones. Those are used, in particular, to consider the distance $\mathbf{d}_2(\bar{\tilde{g}}_T^{ka}, \mathscr{F}_c^k)$ instead of $\mathbf{d}_2(\bar{\hat{f}}_T(k,a), \mathscr{F}^k)$: this avoids the difficulty of having a different estimator normalization for each couple (k,a), by simply considering working with sums. Finally, the auxiliary target set $\tilde{\mathscr{C}}$ is defined by

$$\tilde{\mathscr{C}} = \prod_{k\in\mathscr{K}} \tilde{\mathscr{C}}^k \quad \text{where} \quad \tilde{\mathscr{C}}^k = \mathbf{R}_k^{-1}(\mathbb{R}_-^d) \cap (\mathscr{F}_c^k)^{\mathscr{A}}.$$

The set $\mathbf{R}_k^{-1}(\mathbb{R}_-^d)$ corresponds to approaching the negative orthant in the initial game, whereas the set $(\mathscr{F}_c^k)^{\mathscr{A}}$ corresponds to making the sure the average estimator $\bar{\hat{f}}_T(k,a)$ is close to $\mathscr{F}^k$. Considering the intersection therefore allows to manage both *at the same time.*

VI.8. Proofs of technical lemmas

VI.8.1. Proof of Lemma VI.4.3

Let $1 \leqslant n \leqslant d$ and $b \in \mathscr{B}$. Let us first prove that $\mathbf{r}^n(\cdot, b)$ is piecewise affine. The map $\mathbf{f}$ being affine and defined on $\Delta(\mathscr{J})$, the set $\mathbf{f}^{-1}(b)$ is a polytope. Denote $y_{b,1}, \dots, y_{b,q}$ its vertices. Let $x \in \Delta(\mathscr{I})$. By linearity of $\mathbf{g}(x, \cdot)$, $\mathbf{r}^n(x,b)$ can then be written

$$\mathbf{r}^n(x,b) = \max \mathbf{g}^n(x, \mathbf{f}^{-1}(b)) = \max_{1\leqslant p\leqslant q} \mathbf{g}^n(x, y_{b,p}).$$

$\mathbf{r}^n(\cdot, b)$ now appears as the maximum of a finite family $(\mathbf{g}^n(\cdot, y_{b,p}))_{1\leqslant p\leqslant q}$ of linear functions. It is therefore piecewise affine and so is $\mathbf{r}(\cdot, b)$. Therefore, for each $b \in \mathscr{B}$ there exists a decomposition of $\Delta(\mathscr{I})$ into polytopes on each of which $\mathbf{r}(\cdot, b)$ is affine. $\mathscr{B}$ being finite, we can consider the decomposition $(\mathscr{X}^\ell)_{\ell\in\mathscr{L}}$ which refines all of them. $\mathbf{r}(\cdot, b)$ is therefore affine on each polytope $\mathscr{X}^\ell$ for all $b \in \mathscr{B}$. Let us now prove that $\mathbf{r}(\cdot, f)$ is affine on each polytope $\mathscr{X}^\ell$ for all $f \in \mathscr{F}$.

Let $f \in \mathscr{F}$, $\ell \in \mathscr{L}$, $x_1, x_2 \in \mathscr{X}^\ell$ and $\lambda \in [0,1]$. Using property (iii) from Lemma VI.4.2, we consider the unique decomposition $f = \sum_{b\in\mathscr{B}} \mu^b \cdot b$ and $k \in \mathscr{K}$ such that $\operatorname{supp}\mu \subset \mathscr{F}^k$. Using the definition of $\mathbf{r}$ and the affinity of $\mathbf{r}(\cdot, b)$ on $\mathscr{X}^\ell$, we

have

$$\begin{aligned}\mathbf{r}(\lambda x_1 + (1-\lambda)x_2, f) &= \sum_{b\in\mathcal{B}} \mu^b \cdot \mathbf{r}(\lambda x_1 + (1-\lambda)x_2, b)\\ &= \sum_{b\in\mathcal{B}} \mu^b \left(\lambda \mathbf{r}(x_1, b) + (1-\lambda)\mathbf{r}(x_2, b)\right)\\ &= \lambda \sum_{b\in\mathcal{B}} \mu^b \cdot \mathbf{r}(x_1, b) + (1-\lambda) \sum_{b\in\mathcal{B}} \mu^b \cdot \mathbf{r}(x_2, b)\\ &= \lambda \mathbf{r}(x_1, f) + (1-\lambda)\mathbf{r}(x_2, f),\end{aligned}$$

where the last equality stands because of the uniqueness of the decomposition of f lets us recognize the definitions of $\mathbf{r}(x_1, b)$ and $\mathbf{r}(x_2, b)$ from Equation (VI.2).

VI.8.2. Proof of Proposition VI.4.4

(i) Let $x \in \Delta(\mathcal{I})$ and $y \in \Delta(\mathcal{J})$. Denote $f = \mathbf{f}(y)$. We use property (iii) from Lemma VI.4.2 to get the unique decomposition $f = \sum_{b\in\mathcal{B}} \mu^b \cdot b$ and $k \in \mathcal{K}$ such that $\operatorname{supp}\mu \subset \mathcal{F}^k$. $\mathbf{f}^{-1}$ being affine on $\mathcal{F}^k$ by property (ii) in Lemma VI.4.2, we have

$$\begin{aligned}\mathbf{g}(x, y) \in \mathbf{g}(x, \mathbf{f}^{-1}(f)) &= \mathbf{g}\left(x, \mathbf{f}^{-1}\left(\sum_{b\in\operatorname{supp}\mu} \mu^b \cdot b\right)\right) = \mathbf{g}\left(x, \sum_{b\in\mathcal{B}} \mu^b \cdot \mathbf{f}^{-1}(b)\right)\\ &= \sum_{b\in\operatorname{supp}\mu} \mu^b \cdot \mathbf{g}(x, \mathbf{f}^{-1}(b)).\end{aligned}$$

Then for each $1 \leqslant n \leqslant d$,

$$\begin{aligned}\mathbf{g}^n(x, y) &\leqslant \max \sum_{b\in\operatorname{supp}\mu} \mu^b \cdot \mathbf{g}(x, \mathbf{f}^{-1}(b)) = \sum_{b\in\mathcal{B}} \mu^b \cdot \max \mathbf{g}^n(x, \mathbf{f}^{-1}(b))\\ &= \sum_{b\in\mathcal{B}} \mu^b \cdot \mathbf{r}^n(x, b) = \mathbf{r}^n(x, f),\end{aligned}$$

where for the second equality, we recognized the definition of $\mathbf{r}^n(x, b)$ from Equation (VI.1) on page 135, and the the last equality, the definition of $\mathbf{r}^n(x, f)$ from Equation (VI.2).

(ii) Let $f \in \mathcal{F}$. Thanks to the characterization of approachability from Proposition VI.3.2, there exists $x \in \Delta(\mathcal{I})$ such that $\mathbf{m}(x, f) \in \mathbb{R}^{\underline{d}}$. Let $f = \sum_{b\in\mathcal{B}} \mu^b \cdot b$ be the unique decomposition of f given by Lemma VI.4.2. With the same arguments as

above, we have for each $1 \leqslant n \leqslant d$,

$$\begin{aligned}\mathbf{r}^n(x,f) &= \sum_{b\in\mathcal{B}} \mu^b \cdot \mathbf{r}^n(x,b) = \sum_{b\in\mathcal{B}} \mu^b \cdot \max \mathbf{g}^n(x, \mathbf{f}^{-1}(b)) \\ &= \max \sum_{b\in\mathcal{B}} \mu^b \cdot \mathbf{g}^n(x, \mathbf{f}^{-1}(b)) = \max \mathbf{g}^n\left(x, \mathbf{f}^{-1}\left(\sum_{b\in\mathcal{B}} \mu^b \cdot b\right)\right) \\ &= \max \mathbf{g}^n(x, \mathbf{f}^{-1}(f)) = \max \mathbf{m}^n(x,f) \leqslant 0.\end{aligned}$$

Therefore, $\mathbf{r}(x,f) \in \mathbb{R}^d_-$.

(iii) Let $x \in \Delta(\mathcal{I})$, $k \in \mathcal{K}$, $f_1, f_2 \in \mathcal{F}^k$ and $\lambda \in [0,1]$. We use property (iii) from Lemma VI.4.2 to write $f_1 = \sum_{b\in\mathcal{B}} \mu_1^b \cdot b$ and $f_2 = \sum_{b\in\mathcal{B}} \mu_2^b \cdot b$ with $\operatorname{supp}\mu_1 \subset \mathcal{F}^k$ and $\operatorname{supp}\mu_2 \subset \mathcal{F}^k$. The unique decomposition of $\lambda f_1 + (1-\lambda) f_2$ given by Lemma VI.4.2 is then

$$\lambda f_1 + (1-\lambda) f_2 = \sum_{b\in\mathcal{B}} (\lambda\mu_1^b + (1-\lambda)\mu_2^b) \cdot b.$$

Therefore, using the definition of $\mathbf{r}$ and the affinity of $\mathbf{r}(x, \cdot)$ on $\mathcal{F}^k$,

$$\begin{aligned}\mathbf{r}(x, \lambda f_1 + (1-\lambda) f_2) &= \mathbf{r}\left(x, \sum_{b\in\mathcal{B}} (\lambda\mu_1^b + (1-\lambda)\mu_2^b) \cdot b\right) \\ &= \sum_{b\in\mathcal{B}} (\lambda\mu_1^b + (1-\lambda)\mu_2^b) \cdot \mathbf{r}(x,b) \\ &= \lambda \sum_{b\in\mathcal{B}} \mu_1^b \cdot \mathbf{r}(x,b) + (1-\lambda) \sum_{b\in\mathcal{B}} \mu_2^b \cdot \mathbf{r}(x,b) \\ &= \lambda \mathbf{r}(x, f_1) + (1-\lambda) \cdot \mathbf{r}(x, f_2).\end{aligned}$$

(iv) is already proved in Lemma VI.4.3.

VI.8.3. Proof of Lemma VI.4.5

Let $k \in \mathcal{K}$ and $x \in \Delta(\mathcal{I})$. Let us consider $\operatorname{span}(\mathcal{F}^k) \subset \mathbb{R}^{\mathcal{J}\times\mathcal{I}}$, the linear span of $\mathcal{F}^k$. There exists a basis $(f_1, \dots, f_q)$ of $\operatorname{span}(\mathcal{F}^k)$ such that f_p belongs to $\mathcal{F}^k$ for each $1 \leqslant p \leqslant q$. We now define $\mathbf{r}^{[k]}(x, \cdot)$ on $\operatorname{span}(\mathcal{F}^k)$ by setting

$$\mathbf{r}^{[k]}(x, f_p) := \mathbf{r}(x, f_p), \qquad \text{for each element } f_p \text{ of the basis,}$$

and extending linearly. $\mathbf{r}^{[k]}(x, \cdot)$ can then be further extended to the whole space $\mathbb{R}^{\mathcal{J}\times\mathcal{I}}$ by setting its value to zero on some complementary subspace of $\operatorname{span}(\mathcal{F}^k)$.

Let us now prove that $\mathbf{r}^{[k]}(x, \cdot)$ coincides with $\mathbf{r}(x, \cdot)$ on $\mathscr{F}^k$. Let $f \in \mathscr{F}^k$. In particular, f belongs to $\mathrm{span}(\mathscr{F}^k)$ and can be uniquely written

$$f = \sum_{p=1}^{q} \lambda_p f_p, \quad \text{where} \quad \lambda_1, \dots, \lambda_q \in \mathbb{R}.$$

The application $\mathbf{r}^{[k]}(x, \cdot)$ being linear by definition, we have

$$\mathbf{r}^{[k]}(x, f) = \sum_{p=1}^{q} \lambda_p \mathbf{r}(x, f_p).$$

We now aim at proving that the above sum is equal to $\mathbf{r}(x, f)$. This cannot be done by directly applying the affinity of $\mathbf{r}(x, \cdot)$ (property (iii) in Lemma VI.4.4) because some of the above coefficients λ_p may be negative. To overcome this, we first separate the terms according to the signs of the coefficients λ_p. We denote Λ^+ (resp. Λ^-) the sum of all positive (resp. negative) coefficients λ_p and write

$$\begin{aligned}\mathbf{r}^{[k]}(x, f) &= \sum_{\lambda_p > 0} \lambda_p \mathbf{r}(x, f_p) + \sum_{\lambda_p < 0} \lambda_p \mathbf{r}(x, f_p) \\ &= \Lambda^+ \sum_{\lambda_p > 0} \left(\frac{\lambda_p}{\Lambda^+} \right) \mathbf{r}(x, f_p) + \Lambda^- \sum_{\lambda_p < 0} \left(\frac{\lambda_p}{\Lambda^-} \right) \mathbf{r}(x, f_p).\end{aligned}$$

Since each of the above sum is now a convex combination, we can apply the affinity of $\mathbf{r}(x, \cdot)$:

$$\mathbf{r}^{[k]}(x, f) = \Lambda^+ \cdot \mathbf{r}\left(x, \sum_{\lambda_p > 0} \left(\frac{\lambda_p}{\Lambda^+} \right) f_p \right) + \Lambda^- \mathbf{r}\left(x, \sum_{\lambda_p < 0} \left(\frac{\lambda_p}{\Lambda^-} \right) f_p \right).$$

Let us prove that

$$\mathbf{r}(x, f) - \Lambda^- \mathbf{r}\left(x, \sum_{\lambda_p < 0} \left(\frac{\lambda_p}{\Lambda^-} \right) f_p \right) = \Lambda^+ \cdot \mathbf{r}\left(x, \sum_{\lambda_p > 0} \left(\frac{\lambda_p}{\Lambda^+} \right) f_p \right). \quad \text{(VI.4)}$$

This will prove that $\mathbf{r}^{[k]}(x, f) = \mathbf{r}(x, f)$.

$$\begin{aligned}
\mathbf{r}(x,f) - \Lambda^- \mathbf{r}\left(x, \sum_{\lambda_p<0}\left(\frac{\lambda_p}{\Lambda^-}\right) f_p\right) &= (1-\Lambda^-)\left(\frac{1}{1-\Lambda^-}\mathbf{r}(x,f)\right. \\
&\qquad \left. + \frac{-\Lambda^-}{1-\Lambda^-}\mathbf{r}\left(x, \sum_{\lambda_p<0}\left(\frac{\lambda_p}{\Lambda^-}\right) f_p\right)\right) \\
&= (1-\Lambda^-)\cdot \mathbf{r}\left(x, \frac{1}{1-\Lambda^-} f + \sum_{\lambda_p<0}\left(-\frac{\lambda_p}{1-\Lambda^-}\right) f_p\right) \\
&= (1-\Lambda^-)\cdot \mathbf{r}\left(x, \frac{1}{1-\Lambda^-}\left(f - \sum_{\lambda_p<0}\lambda_p f_p\right)\right) \\
&= (1-\Lambda^-)\cdot \mathbf{r}\left(x, \sum_{\lambda_p>0}\left(\frac{\lambda_p}{1-\Lambda^-}\right) f_p\right).
\end{aligned}$$

For relation (VI.4) to be true, it is now enough to prove that $\Lambda^+ + \Lambda^- = 1$. Since $\mathscr{F}^k \subset \mathscr{F} \subset \Delta(\mathscr{S})^{\mathscr{I}}$, for any $f_0 = (f_0^{is})_{\substack{s\in\mathscr{S}\\ i\in\mathscr{I}}} \in \mathscr{F}^k$, we have

$$\sum_{\substack{s\in\mathscr{S}\\ i\in\mathscr{I}}} f_0^{is} = \sum_{i\in\mathscr{I}}\sum_{s\in\mathscr{S}} f_0^{is} = \sum_{i\in\mathscr{I}} 1 = |\mathscr{I}|\,.$$

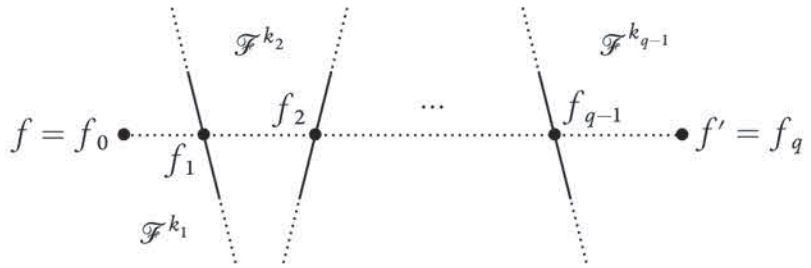
By applying the above to f and the f_p, we get

$$\begin{aligned}
|\mathscr{I}| = \sum_{\substack{s\in\mathscr{S}\\ i\in\mathscr{I}}} f^{is} &= \sum_{\substack{s\in\mathscr{S}\\ i\in\mathscr{I}}}\left(\sum_{\lambda_p>0}\lambda_p f_p^{is} + \sum_{\lambda_p<0}\lambda_p f_p^{is}\right) \\
&= \sum_{\lambda_p>0}\lambda_p \sum_{\substack{s\in\mathscr{S}\\ i\in\mathscr{I}}} f_p^{is} + \sum_{\lambda_p<0}\lambda_p \sum_{\substack{s\in\mathscr{S}\\ i\in\mathscr{I}}} f_p^{is} \\
&= \Lambda^+ |\mathscr{I}| + \Lambda^- |\mathscr{I}|\,,
\end{aligned}$$

and we indeed get $\Lambda^+ + \Lambda^- = 1$ by dividing by $|\mathscr{I}|$, which concludes the proof.

VI.8.4. Proof of Lemma VI.4.6

Property (i) follows from the definition of $\mathrm{L}_{\mathbf{r}}$ and the linearity of the map $\mathbf{r}^{[k]}(a, \cdot)$.



(ii) Let $k \in \mathscr{K}$, $a \in \mathscr{A}$ and $f, f' \in \mathscr{F}$. $(\mathscr{F}^k)_{k\in\mathscr{K}}$ being a finite decomposition of $\mathscr{F}$ into convex polytopes, there exists a finite sequence $(k_1, k_2, \dots, k_q)$ in $\mathscr{K}$ such that the k_p's are all different and a sequence $(f_0 = f, f_1, f_2, \dots, f_q = f')$ in the affine segment $[f, f']$ such that $[f_{p-1}, f_p] \subset \mathscr{F}^{k_p}$ for each $1 \leqslant p \leqslant q$. Therefore, using the fact that $\mathbf{r}^{[k']}(a, \cdot)$ and $\mathbf{r}(a, \cdot)$ coincide on $\mathscr{F}^{k'}$ for all $k' \in \mathscr{K}$, we can write

$$\begin{aligned}
\|\mathbf{r}(a, f) - \mathbf{r}(a, f')\|_2 &= \left\| \sum_{p=1}^{q} \big(\mathbf{r}(a, f_{p-1}) - \mathbf{r}(a, f_p)\big) \right\|_2 \\
&= \left\| \sum_{p=1}^{q} \mathbf{r}^{[k_p]}(a, f_{p-1}) - \mathbf{r}^{[k_p]}(a, f_p) \right\|_2 \\
&\leqslant \sum_{p=1}^{q} \left\| \mathbf{r}^{[k_p]}(a, f_{p-1}) - \mathbf{r}^{[k_p]}(a, f_p) \right\|_2 \\
&\leqslant \mathrm{L}_{\mathbf{r}} \sum_{p=1}^{q} \|f_{p-1} - f_p\|_2 \\
&= \mathrm{L}_{\mathbf{r}} \|f - f'\|_2 ,
\end{aligned}$$

where the last equality holds because the points $f_0, \dots, f_q$ are aligned and ordered.

VI.8.5. Proof of Lemma VI.4.8

(i) Let $k \in \mathscr{K}$. $\mathbf{R}_k^{-1}(\mathbb{R}^d_-)$ is a closed convex cone as the inverse image via a linear application of the closed convex cone $\mathbb{R}^d_-$ (Proposition IV.2.6). $\mathscr{F}_c^k$ is a closed convex cone by definition, and $(\mathscr{F}_c^k)^{\mathscr{A}}$ is thus a closed convex cone as a Cartesian product of closed convex cones. Therefore, $\tilde{\mathscr{C}}^k = \mathbf{R}_k^{-1}(\mathbb{R}^d_-) \cap (\mathscr{F}_c^k)^{\mathscr{A}}$ is also a closed convex cone as the intersection of two closed convex cones. Then, $\tilde{\mathscr{C}}$ is also a closed convex cone as a Cartesian product of closed convex cones.

(ii) Let $\tilde{g} = (\tilde{g}^{ka})_{\substack{k\in\mathscr{K} \\ a\in\mathscr{A}}} \in \tilde{\mathscr{C}}$. By definition of $\tilde{\mathscr{C}}$, for each $k \in \mathscr{K}$, $(\tilde{g}^{ka})_{a\in\mathscr{A}}$ belongs

to $\tilde{\mathscr{C}}^k$ and thus to $(\mathscr{F}_c^k)^{\mathscr{A}}$. Therefore, $\tilde{g} \in \prod_{k\in\mathscr{K}}(\mathscr{F}_c^k)^{\mathscr{A}}$. Moreover,

$$\mathbf{R}(\tilde{g}) = \sum_{k\in\mathscr{K}} \mathbf{R}_k\left((\tilde{g}^{ka})_{a\in\mathscr{A}}\right)$$

belongs to $\mathbb{R}^d_-$. Indeed, each term of the above sum belongs to $\mathbb{R}^d_-$ because for all $k \in \mathscr{K}$, $(\tilde{g}^{ka})_{a\in\mathscr{A}} \in \tilde{\mathscr{C}}^k \subset \mathbf{R}_k^{-1}(\mathbb{R}^d_-)$.

VI.8.6. Proof of Lemma VI.4.10

(i) Let $i \in \mathscr{I}$. Using the conditional expectation with respect to event $\{i_t = i\}$, we have

$$\begin{aligned}
\mathbb{E}\left[\widehat{f}_t^i \,\middle|\, \mathscr{G}_t\right] &= \mathbb{E}\left[\frac{\mathbb{1}_{\{i_t=i\}}}{\mathbb{P}\left[i_t = i \,|\, \mathscr{G}_t\right]}\delta_{s_t} \,\middle|\, \mathscr{G}_t\right] \\
&= \mathbb{P}\left[i_t = i \,|\, \mathscr{G}_t\right] \times \mathbb{E}\left[\frac{\delta_{s_t}}{\mathbb{P}\left[i_t = i \,|\, \mathscr{G}_t\right]} \,\middle|\, \mathscr{G}_t, \{i_t = i\}\right] \\
&= \mathbb{E}\left[\delta_{s_t} \,|\, \mathscr{G}_t, \{i_t = i\}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\delta_{s_t} \,|\, y_t, \mathscr{G}_t, \{i_t = i\}\right] \,|\, \mathscr{G}_t, \{i_t = i\}\right] \\
&= \mathbb{E}\left[\mathbf{s}(i, y_t) \,|\, \mathscr{G}_t, \{i_t = i\}\right] \\
&= \mathbb{E}\left[\mathbf{s}(i, y_t) \,|\, \mathscr{G}_t\right] \\
&= \mathbb{E}\left[f_t^i \,|\, \mathscr{G}_t\right],
\end{aligned}$$

hence the result.

(ii) We write

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{f}_t\right\|_2^2 \,\middle|\, \mathscr{G}_t\right] &= \mathbb{E}\left[\sum_{i\in\mathscr{I}}\left\|\frac{\mathbb{1}_{\{i_t=i\}}}{\mathbb{P}\left[i_t = i \,|\, \mathscr{G}_t\right]}\delta_{s_t}\right\|_2^2 \,\middle|\, \mathscr{G}_t\right] \\
&= \mathbb{P}\left[i_t = i \,|\, \mathscr{G}_t\right] \times \mathbb{E}\left[\sum_{i\in\mathscr{I}}\left\|\frac{\delta_{s_t}}{\mathbb{P}\left[i_t = i \,|\, \mathscr{G}_t\right]}\right\|_2^2 \,\middle|\, \mathscr{G}_t, \{i_t = i\}\right] \\
&= \sum_{i\in\mathscr{I}} \frac{1}{\mathbb{P}\left[i_t = i \,|\, \mathscr{G}_t\right]} \mathbb{E}\left[\left\|\delta_{s_t}\right\|_2^2 \,\middle|\, \mathscr{G}_t, \{i_t = i\}\right] \\
&= \sum_{i\in\mathscr{I}} \frac{1}{\mathbb{P}\left[i_t = i \,|\, \mathscr{G}_t\right]} \\
&\leqslant \frac{|\mathscr{I}|^2}{\gamma},
\end{aligned}$$

where the last inequality stands because $\mathbb{P}[i_t = i \,|\, \mathcal{G}_t] \geqslant \gamma/|\mathcal{I}|$ by definition of the strategy.

(iii) We have

$$\left\|\widehat{f}_t\right\|_2^2 = \sum_{i\in\mathcal{I}} \left\| \frac{\mathbb{1}_{\{i_t=i\}}}{\mathbb{P}[i_t = i \,|\, \mathcal{G}_t]} \delta_{s_t} \right\|_2^2 = \sum_{i\in\mathcal{I}} \mathbb{1}_{\{i_t=i\}} \frac{\|\delta_{s_t}\|_2^2}{\mathbb{P}[i_t = i \,|\, \mathcal{G}_t]^2}$$
$$\leqslant \frac{|\mathcal{I}|^2}{\gamma^2} \sum_{i\in\mathcal{I}} \mathbb{1}_{\{i_t=i\}} = \frac{|\mathcal{I}|^2}{\gamma^2}.$$

VI.8.7. Proof of Lemma VI.6.1

(i) For $i \in \mathcal{I}$, we write

$$\mathbb{E}\left[\widehat{f}_t^i \,\middle|\, \mathcal{G}_t\right] = \mathbb{E}\left[\mathbb{E}\left[\delta_{s_t} \,|\, \mathcal{G}_t, y_t\right] \,|\, \mathcal{G}_t\right] = \mathbb{E}\left[\mathbf{s}(i, y_t) \,|\, \mathcal{G}_t\right] = \mathbb{E}\left[f_t^i \,|\, \mathcal{G}_t\right].$$

(ii) The Euclidean norm of a Dirac being equal to 1,

$$\left\|\widehat{f}_t\right\|_2^2 = \left\|(\delta_{s_t})_{i\in\mathcal{I}}\right\|_2^2 = \sum_{i\in\mathcal{I}} \|\delta_{s_t}\|_2^2 = |\mathcal{I}|.$$

CHAPTER VII

CONTINUOUS-TIME MIRROR DESCENT

This chapter is extracted from the paper *A continuous-time approach to online optimization*, in collaboration with Panayotis Mertikopoulos, in preparation.

Abstract

We consider a family of learning strategies for online optimization problems that evolve in continuous time and we show that they lead to no regret. From a more traditional, discrete-time viewpoint, this continuous-time approach allows us to derive the no-regret properties of a large class of discrete-time algorithms including as special cases the exponential weight algorithm, online mirror descent, smooth fictitious play and vanishingly smooth fictitious play. In so doing, we obtain a unified view of many classical regret bounds, and we show that they can be decomposed into a term stemming from continuous-time considerations and a term which measures the disparity between discrete and continuous time. As a result, we obtain a general class of infinite horizon learning strategies that guarantee an $O(n^{-1/2})$ regret bound without having to resort to a doubling trick.

VII.1. Introduction

Online optimization focuses on decision-making in sequentially changing environments (the weather, the stock market, etc.). More precisely, at each stage of a repeated decision process, the agent/decision-maker obtains a payoff (or incurs a loss) based on the environment and his decision, and his long-term objective is to maximize his cumulative payoff via the use of past observations.

The worst-case scenario for the agent – and one which has attracted considerable interest in the literature – is when he has no Bayesian-like prior belief on the environment. In this context, the cumulative payoff difference between an oracle-like device (a decision rule which prescribes an action based on knowledge of the future) and a

learning strategy (a rule which only relies on past observations) can become arbitrarily large, even in very simple problems. As a result, in the absence of absolute payoff guarantees, the most widely used online optimization criterion is that of *regret minimization*, a notion which was first introduced by [Han57] and has since given rise to a vigorous literature at the interface of optimization, statistics and theoretical computer science – see e.g. [CBL06], [SS11] for a survey. Specifically, the *cumulative regret* of a strategy compares the payoff obtained by an agent that follows it to the payoff that he would have obtained by constantly choosing one action; accordingly, one of the main goals in online optimization is to devise strategies that lead to (vanishingly) small average regret against any fixed action, and irrespective of how the agent's environment evolves over time.

In this paper, we take a continuous-time approach to online optimization and we consider a class of strategies that lead to no regret in continuous time. From a more traditional, discrete-time viewpoint, the importance of this approach lies in that it provides a unifying view of the regret properties of a broad class of well-known online optimization algorithms. In particular, the discrete-time version of our family of strategies is an extension of the general class of online mirror descent (OMD) algorithms (themselves equivalent to "Following the Regularized Leader" (FtRL) in the case of linear payoffs; see e.g. [SS11], [Bub11], [Haz12]) with a time-varying parameter. As such, our analysis contains as special cases *a*) the exponential weight (EW) algorithm ([LW94], [Vov90]) and its decreasing parameter variant ([ACBG02]); *b*) smooth fictitious play (SFP) ([FL99], [BHS06]) and vanishingly smooth fictitious play (VSFP) ([BF13]); and *c*) the method of online gradient descent (OGD) introduced by [Zin03] (the Euclidean predecessor of OMD).

With regards to the OMD/FtRL family of algorithms, the vanishing regret bounds that we derive by using a time-varying parameter are not particularly new: bounds of the same order can be obtained by taking existing guarantees for learning with a finite horizon and then using the so-called "doubling trick" ([CBFH+97], [Vov98]).[1] That said, the introduction of a time-varying parameter has several advantages: *a*) it allows us to integrate SFP and VSFP into the fold and to derive explicit bounds for their regret; *b*) it provides a unified any-time analysis without needing to reboot the algorithm every so often (to the best of our knowledge, such an analysis only exists for the EW algorithm with a time-varying parameter ([Bub11], [ACBG02])); and *c*) in the case of ordinary convex optimization problems with an open-ended termination criterion (as opposed to a fixed number of steps), a variable parameter leads to more efficient value convergence bounds than a variable step-size.

1. In a nutshell, the doubling trick amounts to breaking up the learning timeline in blocks of exponentially increasing horizon, and then resetting the algorithm at the start of each block with an optimal parameter for the block's (finite) horizon.

Building on an idea that was introduced by [WJ97] in the framework of convex optimization and by [Sor09] in the study of the exponential weight algorithm, the key ingredient of our analysis is the descent from continuous to discrete time. More precisely, given an online optimization problem in discrete time, we construct a continuous-time interpolation where our continuous-time dynamics lead to no regret; then, by comparing the agent's payoffs in discrete and continuous time, we are able to deduce a bound for the agent's regret in the original discrete-time framework.

One of the main contributions of this approach is that it leads to a unified derivation of several existing regret bounds with disparate proofs; secondly, it allows us to decompose many classical bounds into two components, a term coming from continuous-time considerations and a comparison term which measures the disparity between discrete and continuous time (see also [PM13] for an alternative interpretation of such a decomposition). Each of these terms can be made arbitrarily small by itself, but their sum is coupled in a nontrivial way that induces a trade-off between continuous- and discrete-time considerations: in a sense, faster decay rates in continuous time lead to greater discrepancies in the discrete/continuous comparison – and hence, to slower regret decay bounds in discrete time.

Finally, we also give a brief account of how the derived regret bounds are related to classical convergence results for certain convex optimization and stochastic convex optimization algorithms—including the projected subgradient (PSG) method, mirror descent (MD), and their stochastic variants ([NY83], [NJLS09]), and we illustrate a (somewhat surprising) performance gap incurred by using an optimization algorithm with a decreasing parameter instead of a decreasing step-size.

VII.1.1. Paper outline

In Section VII.2, we present some basics of online optimization to fix notation and terminology; then, in Section VII.3, we define regularizer functions, choice maps and the class of variable-parameter OMD/FtRL strategies that we will focus on. The core of our paper consists of Sections VII.4 and VII.5: we first show that the corresponding class of continuous-time strategies leads to no regret in Section VII.4; this analysis is then translated to discrete time in Section VII.5 where we derive the no-regret properties of the class of algorithms under consideration. Finally, in Section VII.6, we establish several links with existing online learning and convex optimization algorithms, and we show how their properties can be derived as corollaries of our results.

VII.1.2. Notation and preliminaries

Let d be a positive integer and let $\mathrm{V} = \mathbb{R}^d$ be equipped with an arbitrary norm $\|\cdot\|$. The dual of V will be denoted by V^* and the induced dual norm on V^* will be given

by the familiar expression:

$$\|y\|_* = \sup_{\|x\|\leq 1} |\langle y|x\rangle|\,, \tag{VII.1}$$

where $\langle y|x\rangle$ denotes the canonical pairing between $y \in \mathrm{V}^*$ and $x \in \mathrm{V}$. For a nonempty subset $\mathrm{U} \subset \mathrm{V}$ will use the notation $\|\mathrm{U}\| = \sup_{x\in\mathrm{U}} \|x\|$.

In the rest of our paper, $\mathscr{C}$ will denote a nonempty compact convex subset of V; moreover, given a convex function $f\colon \mathrm{V} \to \mathbb{R}\cup\{+\infty\}$, its *effective domain* will be the convex set $\operatorname{dom} f = \{x \in \mathrm{V} : f(x) < \infty\}$. For convenience, if $f\colon \mathscr{C} \to \mathbb{R}$ is convex, we will treat f as a convex function on V by setting $f(x) = +\infty$ for $x \in \mathrm{V}\setminus\mathscr{C}$; conversely, if $f\colon \mathrm{V} \to \mathbb{R}\cup\{+\infty\}$ has domain $\operatorname{dom} f = \mathscr{C}$, we will also treat f as a real-valued function on $\mathscr{C}$ (in all cases, the ambient space V will be clear from the context). We will then say that $v \in \mathrm{V}^*$ is *a subgradient of f at $x \in \mathscr{C}$* if $f(x') - f(x) \geq \langle v|x'-x\rangle$ for all $x' \in \mathscr{C}$; likewise, the set $\partial f(x) = \{v \in \mathrm{V}^* : v \text{ is a subgradient of } f \text{ at } x\}$ will be called the *subdifferential of f at x* and f will be called *subdifferentiable* if $\partial f(x)$ is nonempty for all $x \in \operatorname{dom} f$.

If it exists, the minimum (resp. maximum) of a function $f\colon \mathrm{V} \to \mathbb{R}\cup\{+\infty\}$ will be denoted by $f_{\min}$ (resp. $f_{\max}$). Moreover, if $\mathscr{A} = \{a_1, \dots, a_d\}$ is a finite set, the set $\Delta(\mathscr{A})$ of probability measures on $\mathscr{A}$ will be identified with the standard $(d-1)$-dimensional simplex $\Delta_d = \{x \in \mathbb{R}^d_+ : \sum_{i=1}^d x_i = 1\}$ of $\mathbb{R}^d$; also, the elements of $\mathscr{A}$ will be identified with the corresponding vertices of $\Delta(\mathscr{A})$, i.e. the canonical basis vectors $\{e_i\}_{i=1}^d$ of $\mathbb{R}^d$. Finally, for $x, y \in \mathbb{R}$, we will let $\lfloor x\rfloor = \max\{k \in \mathbb{Z} : k \leq x\}$ and $\lceil x\rceil = \min\{k \in \mathbb{Z} : k \geq x\}$, and we will write $x \vee y = \max\{x, y\}$ and $x \wedge y = \min\{x, y\}$.

VII.2. The model

The heart of the online optimization model that we consider is as follows: at every discrete time instance $n \geq 1$, an agent (decision-maker) chooses an action from a nonempty convex action set $\mathscr{C} \subset \mathrm{V}$ and gains a payoff (or incurs a loss) determined by some time-dependent function. Information about this function is only revealed to the agent after he picks his action, and the agent's objective is to maximize his long-term payoff in an adaptive manner.

VII.2.1. The core model

Let $\mathscr{C} \subset \mathrm{V}$ denote the agent's action space. Then, at each stage $n \geq 1$, the process of play is as follows:

1. The agent chooses an action $x_n \in \mathscr{C}$.
2. Nature chooses and reveals the *payoff vector* $u_n \in \mathrm{V}^*$ of the n-th stage and the agent receives a payoff of $\langle u_n|x_n\rangle$.[2]

2. Nature may be adversarial, i.e. u_n may be chosen as a function of $x_1, \dots, x_n$.

3. The agent uses some decision rule to pick a new action $x_{n+1} \in \mathscr{C}$ and the process is repeated ad infinitum.

More precisely, define a *strategy* to be a sequence of maps $\sigma_n \colon (V^*)^{n-1} \to \mathscr{C}$, $n \geq 1$, such that σ_{n+1} determines the player's action at stage $n+1$ in terms of the payoff vectors $u_1, \dots, u_n \in V^*$ that have been revealed up to stage n (in a slight abuse of notation, σ_1 will be regarded as an element of $\mathscr{C}$). Then, given a sequence of payoff vectors $u = (u_n)_{n\geq 1}$ in V^*, the *sequence of actions generated by* σ will be

$$x_{n+1} \equiv \sigma_{n+1}(u_1, \dots, u_n), \tag{VII.2}$$

and the agent's *cumulative regret* with respect to $x \in \mathscr{C}$ is defined as:

$$\begin{aligned} \mathrm{Reg}_n^{\sigma,u}(x) &= \sum_{k=1}^{n} \langle u_k | x \rangle - \sum_{k=1}^{n} \langle u_k | x_k \rangle \\ &= \sum_{k=1}^{n} \langle u_k | x \rangle - \sum_{k=1}^{n} \langle u_k | \sigma_k(u_1, \dots, u_{k-1}) \rangle \,. \end{aligned} \tag{VII.3}$$

In what follows, we focus on strategies that lead to *no* (or, at worst, *small*) *regret*:

Definition VII.2.1. A strategy σ *leads to* ε*-regret* ($\varepsilon \geq 0$) if, for every sequence of payoff vectors $(u_n)_{n\geq 1}$ in V^* such that $\|u_n\|_* \leq 1$:

$$\limsup_{n\to\infty} \frac{1}{n} \max_{x\in\mathscr{C}} \mathrm{Reg}_n^{\sigma,u}(x) \leq \varepsilon. \tag{VII.4}$$

In particular, if (VII.4) holds with $\varepsilon = 0$, we will say that σ *leads to no regret*.

Remark VII.2.2. The definition of an ε-regret strategy depends on the dual norm $\|\cdot\|_*$ of V^* (and hence, on the original norm $\|\cdot\|$ on V); on the other hand, the definition of "no regret" is independent of the norm.

Remark VII.2.3. In our framework, we can easily see that a strategy leading to ε-regret against "any sequence" is equivalent to leading to ε-regret against "any strategy of nature". However, this may not be true in the randomized setting we present in the following paragraph.

Despite its simplicity, this online linear optimization model may be used to analyze more general online optimization models. In what follows, we summarize some examples of this kind.

VII.2.2. The case of the simplex and mixed actions

Consider a discrete decision process where, at each stage $n \geq 1$, the agent chooses an action a_n from a finite set of *pure* actions $\mathcal{A} = \{1, \dots, d\}$. To do so, the agent draws a_n according to some probability distribution $x_n \in \Delta(\mathcal{A})$; then, once a_n is drawn, the payoff vector $u_n \in [-1, 1]^d$ which prescribes the payoff $u_{n,a}$ of each action $a \in \mathcal{A}$ is revealed and the agent receives the payoff u_{n,a_n} that corresponds to his choice of action. Moreover, we assume that Nature's choice of payoff vector u_n does not depend on pure action a_n.

In this setting, a strategy is still defined as in the core model of Section VII.2.1 with the agent's action set replaced by the set of *mixed actions* $\Delta(\mathcal{A})$.[3] The agent's *realized* regret with respect to a pure action $a \in \mathcal{A}$ will then be

$$\sum_{k=1}^{n} (u_{k,a} - u_{k,a_k}), \tag{VII.5}$$

and we will say that a strategy σ leads to ε-*realized-regret* (resp. to *no realized regret* for $\varepsilon = 0$) if

$$\limsup_{n\to\infty} \frac{1}{n} \max_{a\in\mathcal{A}} \sum_{k=1}^{n} (u_{k,a} - u_{k,a_k}) \leq \varepsilon \quad \text{(a.s.)}, \tag{VII.6}$$

for every strategy of Nature choosing payoff vectors $(u_n)_{n\geqslant 1}$ in $\mathbb{R}^d$ such that $\|u_n\|_\infty \leq 1$.[4] Besides, consider the filtration $(\mathcal{F}_n)_{n\geqslant 1}$ where $\mathcal{F}_n$ is generated by

$$(x_1, u_1, i_1, \dots, x_{n-1}, u_{n-1}, i_{n-1}, x_n, u_n).$$

Then, then conditional expectation $\mathbb{E}\left[u_{n,a_n} \mid \mathcal{F}_n\right]$ is equal to $\langle u_n | x_n \rangle$. Using a classical argument based on Hœffding's inequality and the Borel–Cantelli lemma, the realized regret can be shown to be close with high probability to the regret as defined in Section VII.2.1 (see Lemma II.2.1). Therefore, the minimization of (VII.5) is then reduced to the core model of Section VII.2.1:

Proposition VII.2.4 ([CBL06], Corollary 4.3). *If a strategy σ leads to ε-regret with respect to the uniform norm on V^*, it also leads to ε-realized-regret.*

VII.2.3. Online convex optimization

We briefly discuss here a more general online convex optimization model where losses are determined by a sequence of convex functions. Formally, the only change

3. In a more general setting, the choice at each stage might depend not only on the past payoff vectors, but also on the agent's realized actions $a_1, \dots, a_n$.

4. This condition is also called external ε-consistency ([FL99], [BHS06]).

from Section VII.2.1 is that at each stage $n \geq 1$, the agent incurs a loss $\ell_n(x_n)$ determined by a subdifferentiable convex *loss function* $\ell_n : \mathscr{C} \to \mathbb{R}$. In this nonlinear setting, the information revealed to the agent after playing includes a (negative) subgradient $u_n \in -\partial\ell_n(x_n) \subset \mathrm{V}^*$ of ℓ_n at x_n, so the incurred cumulative regret with respect to a fixed action $x \in \mathscr{C}$ is:

$$\sum_{k=1}^{n} \ell_k(x_k) - \sum_{k=1}^{n} \ell_k(x). \tag{VII.7}$$

By convexity, $\ell_k(x') - \ell_k(x) \leq \langle v | x' - x \rangle$ for all $v \in \partial\ell_k(x')$ and for all $x \in \mathscr{C}$; in this way, (VII.7) readily yields:

$$\sum_{k=1}^{n} \ell_k(x_k) - \sum_{k=1}^{n} \ell_k(x) \leqslant -\sum_{k=1}^{n} \langle u_k | x_k - x \rangle = \sum_{k=1}^{n} \langle u_k | x \rangle - \sum_{k=1}^{n} \langle u_k | x_k \rangle \tag{VII.8}$$

where $u_k \in -\partial\ell_k(x_k)$. This last expression can obviously be interpreted as the regret incurred by an agent facing a sequence of payoff vectors $u_n \in \mathrm{V}^*$ (cf. the core model of Section VII.2.1), so a strategy which guarantees a bound on the right-hand side of (VII.8) will guarantee the same for (VII.7). Consequently, when the loss functions ℓ_n are uniformly Lipschitz continuous, results for the core model can be directly translated into this one.

VII.3. Regularizer functions, choice maps and learning strategies

VII.3.1. Regularizer functions and choice maps

We begin with the concept of a *regularizer function*:

Definition VII.3.1. A convex function $h : \mathrm{V} \to \mathbb{R} \cup \{+\infty\}$ will be called a *regularizer function on* $\mathscr{C}$ if $\operatorname{dom} h = \mathscr{C}$ and $h|_{\mathscr{C}}$ is strictly convex and continuous.

Remark VII.3.2. This definition is intimately related to the notion of a Legendre-type function (see e.g. [Roc70, Section 26]); however, as was recently noted by [SS07] (and in contrast to the analysis of e.g. [BF13], [Bub11] and [BHS06]), we will not require any differentiability or steepness assumptions.

A key tool in our analysis will be the *convex conjugate* $h^* : \mathrm{V}^* \to \mathbb{R} \cup \{+\infty\}$ of h defined as

$$h^*(y) = \sup_{x \in \mathrm{V}} \{\langle y | x \rangle - h(x)\}. \tag{VII.9}$$

Since h is equal to $+\infty$ on $\mathrm{V} \setminus \mathscr{C}$ and $h|_{\mathscr{C}}$ is continuous and strictly convex, the supremum in (VII.9) will be attained at a *unique* point in $\mathscr{C}$. This unique maximizer then defines our choice map as follows:

Definition VII.3.3. The *choice map* associated to a regularizer function h on $\mathscr{C}$ will be the map $Q_h : V^* \to \mathscr{C}$ defined as

$$Q_h(y) = \arg\max_{x \in \mathscr{C}} \{\langle y|x\rangle - h(x)\}, \quad y \in V^*. \tag{VII.10}$$

Example VII.3.4 (Entropy and logit choice). In the case of the simplex ($\mathscr{C} = \Delta_d$),[5] a classical example of a choice map is generated by the entropy function

$$h(x) = \begin{cases} \sum_{i=1}^d x_i \log x_i & \text{if } x \in \Delta_d, \\ +\infty & \text{otherwise.} \end{cases} \tag{VII.11}$$

A standard calculation then yields the so-called *logit choice map*:

$$Q_h(y) = \frac{1}{\sum_{j=1}^d e^{y_j}} \left(e^{y_1}, \dots, e^{y_d}\right). \tag{VII.12}$$

This map is used to define the exponential weight algorithm (cf. Section VII.6), and its importance stems from the well known fact that it leads to the optimal regret bound for $\mathscr{C} = \Delta_d$ ([CBL06, Theorems 2.2 and 3.7]).

Example VII.3.5 (Euclidean projection). Another important example arises by taking the squared Euclidean distance as a regularizer function; more precisely, we define the *Euclidean regularizer* on $\mathscr{C}$ as

$$h(x) = \begin{cases} \frac{1}{2}\|x\|_2^2 & \text{if } x \in \mathscr{C}, \\ +\infty & \text{otherwise.} \end{cases} \tag{VII.13}$$

The associated choice map $Q_h : \mathbb{R}^N \to \mathscr{C}$ corresponds to taking the orthogonal projection with respect to $\mathscr{C}$:

$$\begin{aligned} Q_h(y) &= \arg\max_{x \in \mathscr{C}} \{ \quad \langle y|x\rangle - \tfrac{1}{2}\|x\|_2^2 \} \\ &= \arg\min_{x \in \mathscr{C}} \{ \quad \tfrac{1}{2}\|x\|_2^2 - \langle y|x\rangle + \tfrac{1}{2}\|y\|_2^2 \} \quad = \arg\min_{x \in \mathscr{C}} \|y - x\|_2^2. \end{aligned} \tag{VII.14}$$

Example VII.3.6 (Bregman projections). The Euclidean example above is a special case of a class of projection mappings known as *Bregman projections* ([Bre67]).

Let $F : V \longrightarrow \mathbb{R} \cup \{+\infty\}$ be a proper convex function, differentiable on its domain. Let us denote $\mathscr{D} = \operatorname{dom} F$ and for $x, x' \in \mathscr{D}$, the *Bregman divergence* $D_F : \mathscr{D} \times \mathscr{D} \longrightarrow \mathbb{R}$ is defined as

$$D_F(x, x') = F(x) - F(x') - \langle \nabla F(x') | x - x' \rangle. \tag{VII.15}$$

5. In this setting, choice maps are more commonly known as *smooth best reply maps* ([FL98], [HS02], [BHS06], [BF13]).

Hence, given a compact set $\mathscr{C} \subset \mathscr{D}$, the associated *Bregman projection* of a point $x_0 \in \mathscr{D}$ onto $\mathscr{C}$ is given by

$$\mathrm{pr}_{\mathrm{F}}^{\mathscr{C}}(x_0) = \operatorname*{arg\,min}_{x \in \mathscr{C}} \mathrm{D}_{\mathrm{F}}(x, x_0). \tag{VII.16}$$

Now assume that F^* is also differentiable on its domain which we will denote $\mathscr{D}^*$. It is easy to check that for $y \in \mathscr{D}^*$, $\nabla \mathrm{F}^*(y) \in \mathscr{D}$ and $\nabla \mathrm{F}(\nabla \mathrm{F}^*(y)) = y$. Then, the process of mapping $y \in \mathscr{D}^*$ to $\nabla \mathrm{F}^*(y)$ and then projecting to $\mathscr{C}$ can be written as a choice map in the sense of (VII.10):

$$\begin{aligned} \mathrm{pr}_{\mathrm{F}}^{\mathscr{C}} \nabla \mathrm{F}^*(y) &= \operatorname*{arg\,min}_{x \in \mathscr{C}} \{\mathrm{F}(x) - \mathrm{F}(\nabla \mathrm{F}^*(y)) - \langle \nabla \mathrm{F}(\nabla \mathrm{F}^*(y)) | x - \nabla \mathrm{F}^*(y) \rangle\} \\ &= \operatorname*{arg\,min}_{x \in \mathscr{C}} \{\mathrm{F}(x) - \langle y | x \rangle\} = \operatorname*{arg\,max}_{x \in \mathbb{R}^d} \{\langle y | x \rangle - h(x)\} = \mathrm{Q}_h(y), \end{aligned} \tag{VII.17}$$

where $h|_{\mathscr{C}} = \mathrm{F}|_{\mathscr{C}}$ and $h(x) = +\infty$ for $x \in \mathbb{R}^d \ \{\mathscr{C}\}$.

VII.3.2. Strategies generated by regularizer functions

The class of strategies that we will consider in the rest of this paper is a variable-parameter extension of the so-called online mirror descent (OMD) method – itself equivalent to the family of algorithms known as Follow the Regularized Leader (FtRL) in the case of linear payoffs (see e.g. [SS11] and [Haz12]).

In a nutshell, this class of strategies may be described as follows: the agent aggregates his payoffs over time into a score vector $y \in \mathrm{V}^*$ and then uses a choice map to turn these scores into actions and continue playing. Formally, if h is a regularizer function on the agent's action space $\mathscr{C}$ and $(\eta_n)_{n \geqslant 1}$ is a positive nonincreasing sequence, *the strategy* $\sigma \equiv \left(\sigma_n^{h,\eta_n}\right)_{n \geqslant 1}$ *generated by* h *with parameter* η_n is defined as

$$\sigma_{n+1}(u_1, \dots, u_n) = \mathrm{Q}_h\left(\eta_n \sum_{k=1}^{n} u_k\right), \tag{VII.18}$$

with $\sigma_1 = \mathrm{Q}_h(0)$. The corresponding sequence of play $x_{n+1} = \sigma_{n+1}(u_1, \dots, u_n)$ will then be given by the recursion:

$$\begin{aligned} \mathrm{U}_n &= \mathrm{U}_{n-1} + u_n, \\ x_{n+1} &= \mathrm{Q}_h(\eta_n \mathrm{U}_n). \end{aligned}$$

In addition to the standard variants of OMD/FtRL, a list of examples of strategies and algorithms that can be expressed in this general form is given in Table VII.1. A more detailed analysis (including the regret properties of each algorithm) will also be provided in Section VII.6; we only mention here that the variability of η_n will be key for the no-regret properties of σ: when η_n is constant, the strategy (VII.18) does not guarantee a sublinear regret bound (see e.g. [SS11] and [Bub11]).

VII.3.3. Regularity of the choice map and the role of strong convexity

In this section, we derive some regularity properties of the choice map Q_h that will be needed in the analysis of the subsequent sections. We begin by showing that Q_h is continuous and equal to the gradient of h^*:

Proposition VII.3.7. *Let h be a regularizer function on $\mathscr{C}$. Then h^* is continuously differentiable on $\mathscr{C}$ and $\nabla h^*(y) = Q_h(y)$ for all $y \in V^*$.*

Proof. For $y \in V^*$, we have

$$x \in \partial h^*(y) \iff y \in \partial h(x) \iff x \in \arg\max_{x' \in \mathscr{C}} \{\langle y|x'\rangle - h(x')\}, \qquad \text{(VII.19)}$$

i.e. $\partial h^*(y) = \arg\max_{x' \in \mathscr{C}} \{\langle y|x'\rangle - h(x')\}$. However, since the latter set only consists of $Q_h(y)$, h^* will be differentiable with $\nabla h^*(y) = Q_h(y)$ for all $y \in V^*$. The continuity of ∇h^* then follows from [Roc70, Corollary 25.5.1]. □

In the discrete-time analysis of Section VII.5, (VII.18) will be shown to guarantee a regret bound of a simple form when Q_h is Lipschitz continuous. This last requirement is equivalent to h being *strongly convex*:

Definition VII.3.8. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function, let $\|\cdot\|$ be a norm on $\mathbb{R}^d$, and let $K > 0$.

(1) f is K-*strongly convex w.r.t.* $\|\cdot\|$ if, for all $w_1, w_2 \in \mathbb{R}^d$ and for all $\lambda \in [0,1]$:

$$f(\lambda w_1 + (1-\lambda)w_2) \leq \lambda f(w_1) + (1-\lambda) f(w_2) - \tfrac{1}{2} K \lambda(1-\lambda) \|w_2 - w_1\|^2. \qquad \text{(VII.20)}$$

(2) f is K-*strongly smooth w.r.t.* $\|\cdot\|$ if it is differentiable and, for all $w_1, w_2 \in \mathbb{R}^d$:

$$f(w_2) \leq f(w_1) + \langle \nabla f(w_1) | w_2 - w_1 \rangle + \tfrac{1}{2} K \|w_2 - w_1\|^2. \qquad \text{(VII.21)}$$

Strong convexity of a function was shown in [KSST12] to be equivalent to strong smoothness of its conjugate. In turn, this equivalence yields the following characterization of Lipschitz continuity:

Proposition VII.3.9. *Let $f : V \to \mathbb{R} \cup \{+\infty\}$ be proper and lower semi-continuous. Then, for $K > 0$, the following are equivalent:*

(i) f is K-strongly convex with respect to $\|\cdot\|$.

(ii) f^* is differentiable and ∇f^* is $1/K$-Lipschitz.

(iii) f^* is $1/K$-strongly smooth with respect to $\|\cdot\|_*$.

Hence, given that regularizer functions are proper and lower semi-continuous by definition, Proposition VII.3.9 leads to the following characterization:

Corollary VII.3.10. *Let h be a regularizer function $\mathscr{C}$ and $K > 0$. The associated choice map Q_h is K-Lipschitz continuous if and only if h is K-strongly convex with respect to $\|\cdot\|$.*

This characterization of the Lipschitz continuity of ∇f^* (which will be of particular interest to us) is a classical result in the case of the Euclidean norm – see e.g. [RW98, Proposition 12.60]. On the other hand, the implication (ii) $\Longrightarrow$ (iii) appears to be new in the case of an arbitrary norm (though the proof technique is fairly standard).

Proof of Proposition VII.3.9. We will show that (i) $\Longrightarrow$ (ii) $\Longrightarrow$ (iii) $\Longrightarrow$ (i).
(i) $\Longrightarrow$ (ii). — See e.g. [BT03, Proposition 3.1], [Nes09, Lemma 1] or [SS07, Lemma 15].

(ii) $\Longrightarrow$ (iii). — Fix $y_1, y_2 \in V^*$, let $z = y_2 - y_1$, and set $\phi(t) = f^*(y_1 + tz)$, $t \in [0,1]$. Identifying V with V^{**} and $\|\cdot\|_{**}$ with $\|\cdot\|$, we have:

$$\begin{aligned}\phi'(t) - \phi'(0) &= \langle \nabla f^*(y_1 + tz) - \nabla f^*(y_1) | z \rangle \\ &\leq \|z\|_* \|\nabla f^*(y_1 + tz) - \nabla f^*(y_1)\| \leq \frac{t}{K}\|z\|_*^2, \end{aligned} \tag{VII.22}$$

where the first inequality follows from the definition of the dual norm and the second from the assumed Lipschitz continuity of f^*. By integrating, we then get:

$$\phi(t) - \phi(0) \leq \phi'(0)t + \frac{1}{2K}t^2\|z\|_*^2, \tag{VII.23}$$

and hence, for $t = 1$:

$$f^*(y_2) - f^*(y_1) \leq \langle \nabla f^*(y_1) | y_2 - y_1 \rangle + \frac{1}{2K}\|y_2 - y_1\|_*^2, \tag{VII.24}$$

which shows that f^* is $1/K$-strongly smooth.

(iii) $\Longrightarrow$ (i). — Since f is proper and lower semi-continuous, it will also be closed. Our assertion then follows from e.g. [KSST12, Theorem 3]. □

We close this section by stating the strong convexity properties of the regularizer functions of Examples VII.3.4 and VII.3.5 (which thus imply the Lipschitz continuity of the corresponding choice maps):

Proposition VII.3.11. *With notation as in Examples VII.3.4 and VII.3.5, we have:*

(*i*) *The entropy $h\colon \Delta_d \to \mathbb{R}$ of (VII.11) is 1-strongly convex w.r.t. $\|\cdot\|_1$.*

(*ii*) *The Euclidean regularizer $h\colon \mathscr{C} \to \mathbb{R}$ of (VII.13) is 1-strongly convex w.r.t. $\|\cdot\|_2$.*

Proof. The strong convexity of the Euclidean regularizer is trivial; for the strong convexity of the entropy with respect to $\|\cdot\|_1$, see e.g. [BT03, Proposition 5.1]. □

VII.4. The continuous-time analysis

Motivated by a technique introduced by [Sor09] in the context of the exponential weight (EW) algorithm, we present in this section a continuous-time version of the class of strategies of Section VII.2 and we derive a bound for the induced regret in continuous time. This will then enable us to bound the actual discrete-time regret by comparing the continuous- and discrete-time variants of this and the previous section respectively.

In continuous time, instead of a sequence of payoff vectors $(u_n)_{n\geq 1}$ in V^*, the agent will be facing a measurable and locally integrable stream of payoff vectors $(u_t)_{t\in\mathbb{R}_+}$ in V^*. Hence, extending (VII.18) to continuous time, we will consider the process:

$$x_t^c = Q_h\left(\eta_t \int_0^t u_s\, ds\right), \tag{VII.25}$$

where $(\eta_t)_{t\in\mathbb{R}_+}$ is a positive, nonincreasing and piecewise continuous parameter, while $x_t^c \in \mathscr{C}$ denotes the agent's action at time t given the history of payoff vectors u_s, $0 \leq s < t$.[6]

Our main result in this section is the following regret bound for (VII.25):

Theorem VII.4.1. *If h is a regularizer function on $\mathscr{C}$ and $(\eta_t)_{t\in\mathbb{R}_+}$ is a positive, nonincreasing and piecewise continuous parameter, then, for every locally integrable payoff stream $(u_t)_{t\in\mathbb{R}_+}$ in V^*, we have:*

$$\max_{x\in\mathscr{C}} \int_0^t \langle u_s|x\rangle\, ds - \int_0^t \langle u_s|x_s^c\rangle\, ds \leq \frac{h_{\max} - h_{\min}}{\eta_t}. \tag{VII.26}$$

Proof. Assume first that η_t is of class C^1 and let $y_t = \eta_t \int_0^t u_s\, ds$. Then, for all $x \in \mathscr{C}$ and for all $t \geq 0$, Fenchel's inequality gives:

$$\int_0^t \langle u_s|x\rangle\, ds = \frac{\langle y_t|x\rangle}{\eta_t} \leq \frac{h^*(y_t) + h(x)}{\eta_t} \leq \frac{h^*(y_t)}{\eta_t} + \frac{h_{\max}}{\eta_t}. \tag{VII.27}$$

On the other hand, with $x_t^c = Q_h(y_t)$, we will also have by definition:

$$\frac{h^*(y_t)}{\eta_t} = \frac{\langle y_t|x_t^c\rangle - h(x_t^c)}{\eta_t} = \int_0^t \langle u_s|x_t^c\rangle\, ds - \frac{h(x_t^c)}{\eta_t}. \tag{VII.28}$$

Consider the function $\phi\colon (x,t) \mapsto \int_0^t \langle u_s|x\rangle\, ds - h(x)/\eta_t$. For fixed $t \geq 0$, one can check that x_t^c maximizes $\phi(x,t)$, so we can apply the envelope theorem (see

6. In the rest of the paper, we will consistently use n and k for discrete indices and $s, t, \ldots$ for continuous ones.

e.g. [MCWG95, Theorem M.L.1]) to differentiate $\phi(x_t^c, t)$ with respect to t:

$$\frac{d}{dt}\frac{h^*(y_t)}{\eta_t} = \frac{\partial\phi}{\partial t}(x_t^c, t) = \langle u_t | x_t^c\rangle + \frac{\dot\eta_t}{\eta_t^2} h(x_t^c) \leq \langle u_t | x_t^c\rangle + h_{\min}\frac{\dot\eta_t}{\eta_t^2}, \tag{VII.29}$$

where we used the fact that, by assumption, $\dot\eta \leqslant 0$. Integrating (VII.29) then yields

$$\frac{h^*(y_t)}{\eta_t} \leq \frac{h^*(y_0)}{\eta_0} + \int_0^t \langle u_s | x_s^c\rangle \, ds + h_{\min}\int_0^t \frac{\dot\eta_s}{\eta_s^2}\, ds = \int_0^t \langle u_s | x_s^c\rangle \, ds - \frac{h_{\min}}{\eta_t}, \tag{VII.30}$$

where we have used the fact that $h^*(y_0) = h^*(0) = -h_{\min}$ in the second step. Hence, by combining this last equation with (VII.27), we finally obtain:

$$\int_0^t \langle u_s | x\rangle \, ds \leq \int_0^t \langle u_s | x_s^c\rangle \, ds - \frac{h_{\min}}{\eta_t} + \frac{h_{\max}}{\eta_t}, \tag{VII.31}$$

and our claim follows by taking the maximum of the left-hand side over $x \in \mathscr{C}$.

If η_t is not smooth, let η_t^m, $m = 1, 2 \dots$, be a sequence of positive and nonincreasing parameters of class C^1 that converges pointwise to η_t. Then, if we let $y_t^m = \eta_t^m \int_0^t u_s \, ds$ and $x_t^m = Q_h(y_t^m)$, we will also have $x_s^m \to x_s^c$ pointwise for all $s \in [0, t]$ by the continuity of Q_h. By the dominated convergence theorem, this implies that $\int_0^t \langle u_s | x_s^m\rangle \, ds \to \int_0^t \langle u_s | x_s^c\rangle \, ds$ and our assertion follows by the bound (VII.31) for smoothly varying parameters. □

Remark VII.4.2. We should note here that the quantity $\delta_h = h_{\max} - h_{\min}$ in (VII.26) can be taken arbitrarily small so there is no "optimal" regret bound in continuous time. That said, we shall see in the following section that smaller values of δ_h result in greater disparities between continuous and discrete time, thus leading to a trade-off for the regret in discrete time.

VII.5. Regret minimization in discrete time

In this section, our aim will be to provide a bound for the regret incurred by the discrete-time strategy (VII.18). To that end, our approach will be as follows: first, given a positive nonincreasing parameter $(\eta_n)_{n\geq 1}$ and a sequence of payoff vectors $(u_n)_{n\geq 1}$, we construct their continuous-time counterparts by setting

$$u_t = u_{\lceil t\rceil} \tag{VII.32a}$$

and

$$\eta_t = \eta_{\lfloor t\rfloor \vee 1} \tag{VII.32b}$$

for all $t \in \mathbb{R}_+$ (i.e. $\eta_t = \eta_{\lfloor t \rfloor}$ if $t \geq 1$ and $\eta_t = \eta_1$ otherwise). Then, given a regularizer $h\colon \mathscr{C} \to \mathbb{R}$, we will compare the cumulative payoffs of the processes $(x_n)_{n\geq 1}$ and $(x^c_t)_{t\in\mathbb{R}_+}$ that are generated by (VII.18) and (VII.25) in discrete and continuous time respectively. In this way, the derived regret bound will consist of two terms: one coming from the continuous-time bound (VII.26), and a term coming from the discrete/continuous comparison. Formally:

Theorem VII.5.1. *Let h be a* K*-strongly convex regularizer on $\mathscr{C}$ and let $(\eta_n)_{n\geq 1}$ be a positive nonincreasing parameter. Then, for every sequence of payoff vectors $(u_n)_{n\geq 1}$ in* V^**, the sequence of play*

$$x_{n+1} = \mathrm{Q}_h\left(\eta_n \sum_{k=1}^{n} u_k\right) \tag{VII.33}$$

generated by the strategy $\sigma = (\sigma_n^{h,\eta_n})_{n\geq 1}$ of (VII.18) *guarantees the bound*

$$\max_{x\in\mathscr{C}} \mathrm{Reg}_n^{\sigma,u}(x) \leq \frac{h_{\max} - h_{\min}}{\eta_n} + \frac{1}{2\mathrm{K}} \sum_{k=1}^{n} \eta_{k-1}\|u_k\|_*^2, \tag{VII.34}$$

where we have set $\eta_0 = \eta_1$. In particular, if $\|u_n\|_* \leq \mathrm{M}$ for some $\mathrm{M} > 0$, then:

$$\max_{x\in\mathscr{C}} \mathrm{Reg}_n^{\sigma,u}(x) \leq \frac{h_{\max} - h_{\min}}{\eta_n} + \frac{\mathrm{M}^2}{2\mathrm{K}} \sum_{k=1}^{n} \eta_{k-1}. \tag{VII.35}$$

Proof. Define the continuous-time interpolations of u_n and η_n as in (VII.32) and let $y_t = \eta_t \int_0^t u_s\, ds$; Then, for the continuous-time process $x^c_t = \mathrm{Q}_h(y_t)$ generated by (VII.25), we will have:

$$x_n = \mathrm{Q}_h\left(\eta_{n-1} \sum_{k=1}^{n-1} u_k\right) = x^c_{n-1}, \tag{VII.36}$$

and hence, for $k \geq 1$ and $t \in (k-1, k)$, the payoffs corresponding to x^c_t and x_k will differ by at most

$$\begin{aligned} |\langle u_t | x^c_t\rangle - \langle u_k | x_k\rangle| &= |\langle u_k | x^c_t - x^c_{k-1}\rangle| \\ &\leq \|u_k\|_* \|\mathrm{Q}_h(y_t) - \mathrm{Q}_h(y_{k-1})\| \leq \frac{1}{\mathrm{K}}\|u_k\|_* \|y_t - y_{k-1}\|_*, \end{aligned} \tag{VII.37}$$

where the last inequality follows from the 1/K-Lipschitz continuity of Q_h (Corollary VII.3.10). On the other hand, the definition of y_t gives

$$\|y_t - y_{k-1}\|_* = \left\|\eta_{k-1} \int_{k-1}^{t} u_s\, ds\right\|_* \leq \eta_{k-1}\|u_k\|_*(t - k + 1), \tag{VII.38}$$

which leads to the estimate:

$$\left|\int_0^n \langle u_t|x_t^c\rangle - \sum_{k=1}^n \langle u_k|x_k\rangle\right| \le \sum_{k=1}^n \int_{k-1}^k |\,\langle u_t|x_t^c\rangle - \langle u_k|x_k\rangle\,|\,dt$$
$$\le \frac{1}{\mathrm{K}}\sum_{k=1}^n \eta_{k-1}\|u_k\|_*^2 \int_{k-1}^k (t-k+1)\,dt$$
$$= \frac{1}{2\mathrm{K}}\sum_{k=1}^n \eta_{k-1}\|u_k\|_*^2. \tag{VII.39}$$

In view of this discrete/continuous comparison, we thus obtain:

$$\max_{x\in\mathscr{C}} \sum_{k=1}^n \langle u_k|x\rangle = \max_{x\in\mathscr{C}} \int_0^t \langle u_t|x\rangle\ dt$$
$$\le \int_0^n \langle u_t|x_t^c\rangle\ dt + \frac{h_{\max}-h_{\min}}{\eta_n}$$
$$\le \sum_{k=1}^n \langle u_k|x_k\rangle + \frac{1}{2\mathrm{K}}\sum_{k=1}^n \eta_{k-1}\|u_k\|_*^2 + \frac{h_{\max}-h_{\min}}{\eta_n}, \tag{VII.40}$$

where the first inequality follows from Theorem VII.4.1 and the last one from (VII.39); the bounds (VII.34) and (VII.35) are then immediate. □

To get the optimal dependence of the bound (VII.35) in n, both terms should scale as $\sqrt{n}$ (otherwise, one would be slower than the other). In this case, we get a bound for the average regret which vanishes as $\mathrm{O}(n^{-1/2})$:

Corollary VII.5.2. *Let $(u_n)_{n\ge1}$ be a sequence of payoff vectors in V^*. Then, with notation as in Theorem VII.5.1, the sequence of play*

$$x_{n+1} = \mathrm{Q}_h\left(\sqrt{\frac{\mathrm{K}(h_{\max}-h_{\min})}{\mathrm{M}^2 n}}\sum_{k=1}^n u_k\right) \tag{VII.41}$$

guarantees the regret bound:

$$\max_{x\in\mathscr{C}} \mathrm{Reg}_n^{\sigma,u}(x) \le 2\mathrm{M}\sqrt{\frac{h_{\max}-h_{\min}}{\mathrm{K}}}\left(\frac{1}{4}+\sqrt{n}\right). \tag{VII.42}$$

Proof. Set $\delta_h = h_{\max}-h_{\min}$ and $\eta_n = \eta/\sqrt{n}$ with $\eta = \mathrm{M}^{-1}\sqrt{\mathrm{K}\delta_h}$. Then:

$$\sum_{k=1}^n \eta_{k-1} = \eta + \eta\sum_{k=1}^{n-1}\frac{1}{\sqrt{k}} \le \eta + \eta\int_0^{n-1}\frac{1}{\sqrt{t}}\,dt \le \eta\left(1+2\sqrt{n}\right), \tag{VII.43}$$

so the bound (VII.35) becomes:

$$\frac{\delta_h}{\eta_n} + \frac{M^2}{2K}\sum_{k=1}^{n}\eta_{k-1} \leq \frac{\delta_h}{\eta}\sqrt{n} + \frac{M^2\eta}{2K}\left(1+2\sqrt{n}\right) = 2M\sqrt{\frac{\delta_h}{K}}\left(\frac{1}{4}+\sqrt{n}\right).$$

□

Remark VII.5.3. We should stress here that regret guarantees of the same order as (VII.42) can be obtained for the OMD/FtRL family of algorithms by optimizing the choice of parameter over a finite learning horizon and then restarting the algorithm every so often, using the doubling trick ([CBFH+97], [Vov98]) to guarantee a sublinear regret bound in the long run. The doubling trick may thus be seen as a special case of a nonincreasing parameter; for the general case, the bounds (VII.34)/(VII.35) describe in a precise way the impact of the variability of η_n on the method's regret guarantees (see also Section VII.6 for a more detailed discussion).

Remark VII.5.4. The dependence of η on δ_h, K and M in (VII.42) has been chosen precisely so as to minimize the expression $(\delta_h/\eta + M^2\eta/K)$ over all $\eta > 0$.

Remark VII.5.5 (On the dependence on K and the choice of optimal h)**.** The dependence of the bound (VII.42) on K is clearly artificial: (VII.42) remains invariant if h is rescaled by a positive constant, so it suffices to consider regularizer functions that are 1-strongly convex over $\mathscr{C}$. This then leads to the following question: *given a norm* $\|\cdot\|$ *on* V *and a compact convex subset* $\mathscr{C} \subset$ V*, which* 1*-strongly convex function minimizes* $h_{\max} - h_{\min}$*?* With the exception of the Euclidean norm, this question does not seem to admit a trivial answer (cf. Section VII.7.1 for a more detailed discussion).

By expressing the cumulative payoff gap between discrete- and continuous-time *exactly*, Theorem VII.5.1 can be extended further to regularizer functions that are not strongly convex over $\mathscr{C}$. The only thing that changes in this case is that the comparison term of the bound (VII.35) is replaced by a term involving the Bregman divergence associated with the convex conjugate h^* of h.

The following result is a variable-parameter extension of Theorem 5.6 in [BCB12].

Theorem VII.5.6. *Let* h *be a regularizer function on* $\mathscr{C}$*. Then, with notation as in Theorem VII.5.1, the strategy* $\sigma = (\sigma_n^{h,\eta_n})_{n\geq 1}$ *of* (VII.18) *guarantees the regret bound:*

$$\max_{x\in\mathscr{C}} \mathrm{Reg}_n^{\sigma,u}(x) \leq \frac{h_{\max}-h_{\min}}{\eta_n} + \sum_{k=1}^{n}\frac{1}{\eta_{k-1}}D_{h^*}(y_k^-, y_{k-1}^+), \tag{VII.44}$$

where we have set $y_n^+ = \eta_n\sum_{k=1}^n u_k$, $y_n^- = \eta_{n-1}\sum_{k=1}^n u_k$ *and* $\eta_0 = \eta_1$.

Proof. With notation as in the proof of Theorem VII.5.1, the variables $y_n^{\pm}$ in the statement of the theorem may be expressed more concisely as:

$$y_n^{\pm} = \lim_{t \to n^{\pm}} y_t = \lim_{t \to n^{\pm}} \eta_t \int_0^t u_s \, ds, \tag{VII.45}$$

and hence, with η_t right-continuous, we get $x_n = \mathrm{Q}_h(y_{n-1}) = \mathrm{Q}_h(y_{n-1}^+)$. Accordingly, if $x_t^c = \mathrm{Q}_h(y_t)$ denotes the continuous-time process generated by (VII.25), then, for all $k \geq 1$ and for all $t \in (k-1, k)$, we will have:

$$\langle u_t | x_t^c \rangle - \langle u_k | x_k \rangle = \langle u_t | \mathrm{Q}_h(y_t) \rangle - \langle u_k | \mathrm{Q}_h(y_{k-1}^+) \rangle = \langle u_k | \nabla h^*(y_t) \rangle - \langle u_k | \nabla h^*(y_{k-1}^+) \rangle \,. \tag{VII.46}$$

In this way, noting that $\langle u_t | \nabla h^*(y_t) \rangle$ is simply the derivative of $h^*(y_t)/\eta_{k-1}$ for $t \in (k-1, k)$, we obtain the following comparison over $(k-1, k)$:

$$\begin{aligned} \int_{k-1}^k \langle u_t | x_t^c \rangle \, dt - \langle u_k | x_k \rangle &= \int_{k-1}^k \frac{1}{\eta_{k-1}} \frac{d}{dt} \left(h^*(y_t) \right) \, dt - \frac{1}{\eta_{k-1}} \langle \eta_{k-1} u_k | \nabla h^*(y_{k-1}^+) \rangle \\ &= \frac{1}{\eta_{k-1}} \left(h^*(y_k^-) - h^*(y_{k-1}^+) - \langle y_k^- - y_{k-1}^+ | \nabla h^*(y_{k-1}^+) \rangle \right) \\ &= \frac{1}{\eta_{k-1}} \mathrm{D}_{h^*} \left(y_k^-, y_{k-1}^+ \right) . \end{aligned} \tag{VII.47}$$

In view of the above, the claim follows by summing this bound over $k = 1, \dots, n$ and plugging the resulting expression in the first inequality of (VII.40) – which holds independently of any assumptions on h. □

VII.6. Links with existing results

In this section, we discuss how certain existing results in online optimization and (stochastic) convex programming can be obtained as corollaries of the general analysis of the previous sections.

VII.6.1. Links with known online optimization algorithms

The Exponential Weight Algorithm. — The exponential weight (EW) algorithm was introduced independently by [LW94] and [Vov90] as a learning strategy in discrete time. Motivated by the approach of [Sor09] who used a continuous-time variant to retrieve the algorithm's classical regret bounds, we show here how the same bounds can be obtained directly from Theorem VII.5.1.

The framework of the EW algorithm is that of randomized action selection as in Section VII.2.2. Specifically, let $\mathscr{A} = \{1, \dots, d\}$ be a finite set of *pure* actions, and let

the agent's action set be the unit simplex $\mathscr{C} = \Delta_d$ of $\mathbb{R}^d$ – the latter being endowed with the ℓ^1 norm $\|\cdot\|_1$. In this context, the EW algorithm is defined as:

$$\begin{aligned} \mathrm{U}_n &= \mathrm{U}_{n-1} + u_n, \\ x_{i,n+1} &= \frac{e^{\eta \mathrm{U}_{i,n}}}{\sum_{j=1}^d e^{\eta \mathrm{U}_{j,n}}} \end{aligned} \tag{EW}$$

where $\eta > 0$ is a (fixed) parameter and $(u_n)_{n\geq 1}$ is a sequence of payoff vectors in $[-1,1]^d$ (so that $\|u_n\|_\infty \leq 1$ in the induced dual norm).

Example VII.3.4 in Section VII.3.1 shows that (EW) corresponds to (VII.18) with $\eta_n = \eta$ and $h(x) = \sum_{i=1}^d x_i \log x_i$. Since $h_{\max} - h_{\min} = \log d$ and h is 1-strongly convex with respect to $\|\cdot\|_1$ (cf. Proposition VII.3.11), Theorem VII.5.1 readily yields the bound

$$\max_{a\in\mathscr{A}} \mathrm{Reg}_n(a) \leq \frac{\log d}{\eta} + \frac{n\eta}{2}. \tag{VII.48}$$

Additionally, if the time horizon n is known in advance, the optimal parameter choice $\eta = \sqrt{2\log d/n}$ leads to

$$\max_{a\in\mathscr{A}} \mathrm{Reg}_n(a) \leq \sqrt{2n\log d}, \tag{VII.49}$$

which, as far as the dependence on d and n is concerned, is the best possible bound a strategy can guarantee in this framework – see e.g. [CBL06, Theorem 3.7].

Remark VII.6.1. By taking $u_n \in [0,1]^d$ (as is often the case in the literature) and then shifting to $[-1/2, 1/2]^d$, Theorem VII.5.1 can be applied with $\mathrm{M} = 1/2$. This yields a factor of $1/8$ in the second term of (VII.48) and leads to the bound obtained by [CB97] and [CBL06].

The Exponential Weight Algorithm with $\eta_n = 1/\sqrt{n}$. — [ACBG02] considered the following variant of (EW)

$$\begin{aligned} \mathrm{U}_n &= \mathrm{U}_{n-1} + u_n, \\ x_{i,n+1} &= \frac{e^{\eta \mathrm{U}_{i,n}/\sqrt{n}}}{\sum_{j=1}^d e^{\eta \mathrm{U}_{j,n}/\sqrt{n}}}. \end{aligned} \tag{EW$'$}$$

In our context, a direct application of Corollary VII.5.2 with $\mathrm{M} = \mathrm{K} = 1$ then gives

$$\max_{a\in\mathscr{A}} \mathrm{Reg}_n(a) \leq 2\sqrt{n\log d} + \frac{1}{2}\sqrt{\log d}, \tag{VII.50}$$

a bound which, unlike (VII.49), has the advantage of holding uniformly in time.

Smooth Fictitious Play. — The smooth fictitious play (SFP) process was introduced by [FL95] (see also [FL98] and [FL99]), and its regret properties were examined further by [BHS06] using the theory of stochastic approximation – but without providing any quantitative bounds for the regret.

Just like the EW algorithm, SFP falls within the randomized actions framework of Section VII.2.2. In particular, SFP corresponds to the sequence of play generated by (VII.18) for an arbitrary regularizer on Δ_d and with parameter η/n for some $\eta > 0$; specifically:

$$x_{n+1} = \mathrm{Q}_h\left(\frac{\eta}{n}\sum_{k=1}^{n} u_k\right). \tag{SFP}$$

With regards to the regret induced by (SFP), [BHS06, Theorem 6.6] show that for every $\varepsilon > 0$, there exists some $\eta^* \equiv \eta^*(\varepsilon)$ such that the strategy (SFP) with parameter $\eta \geq \eta^*$ leads to ε-realized-regret. On the other hand, combining Proposition VII.2.4 with Theorem VII.5.1 yields the following more precise statement:

Proposition VII.6.2. *Let h be a K-strongly convex regularizer on the unit simplex $\Delta_d \subset \mathbb{R}^d$ endowed with the ℓ^1 norm. Then, for every sequence of payoff vectors $(u_n)_{n\geq 1}$ in $[-1,1]^d$, the strategy (SFP) with parameter $\eta > 0$ guarantees*

$$\max_{a\in\mathcal{A}} \mathrm{Reg}_n(a) \leq \frac{h_{\max} - h_{\min}}{\eta} n + \frac{\eta \log n}{2\mathrm{K}} + \frac{\eta}{\mathrm{K}}. \tag{VII.51}$$

In particular, (SFP) with parameter η leads to $(h_{\max} - h_{\min})/\eta$ (realized) regret.

Proof. Simply combine the logarithmic growth estimate $\sum_{k=1}^{n} k^{-1} < 1 + \log n$ for the harmonic series and Theorem VII.5.1 with $\eta_n = \eta/n$; the claim for the realized regret then follows from Proposition VII.2.4. □

Remark VII.6.3. It should be noted here that the qualitative analysis of [BHS06] does not require h to be strongly convex; that said, if h is strongly convex, Proposition VII.6.2 gives a quantitative bound on the regret.

Vanishingly Smooth Fictitious Play. — The variant of SFP known as vanishingly smooth fictitious play (VSFP) was introduced by [BF13], and its regret properties were established using sophisticated tools from the theory of differential inclusions and stochastic approximation – but, again, without providing explicit regret bounds.

Using the same notation as before, VSFP corresponds to the sequence of play

$$x_{n+1} = \mathrm{Q}_h\left(\eta_n \sum_{k=1}^{n} u_k\right), \tag{VSFP}$$

where h is a strongly convex regularizer on Δ_d and the sequence η_n satisfies:

(A1) $\lim_{n\to\infty} n\eta_n = +\infty$.

(A2) $\eta_n = \mathrm{O}(n^{-\alpha})$ for some $\alpha > 0$.

Under these assumptions, the main result of [BF13] is that (VSFP) leads to no realized regret; in our framework, this follows directly from Proposition VII.2.4 and Theorem VII.5.1 (which also gives a quantitative regret guarantee):

Proposition VII.6.4. *With notation as in Proposition VII.6.2, the strategy* (VSFP) *with* η_n *satisfying assumptions* (A1) *and* (A2) *guarantees the regret bound*

$$\max_{a\in\mathscr{A}} \frac{1}{n}\operatorname{Reg}_n(a) \leq \frac{h_{\max} - h_{\min}}{n\eta_n} + \frac{1}{2n\mathrm{K}}\sum_{k=1}^{n}\eta_{k-1}, \tag{VII.52}$$

and thus leads to no regret. In particular, if $\eta_n = \eta n^{-\alpha}$ *for some* $\alpha \in (0,1)$*, then:*

$$\max_{a\in\mathscr{A}} \frac{1}{n}\operatorname{Reg}_n(a) \leq \frac{h_{\max} - h_{\min}}{\eta n^{1-\alpha}} + \frac{\eta n^{-\alpha}}{2(1-\alpha)\mathrm{K}} + \frac{\eta}{2\mathrm{K}n}. \tag{VII.53}$$

Proof. The bound (VII.52) is an immediate corollary of Theorem VII.5.1; the no-regret property then follows from Assumptions (A1) and (A2). Finally, if $\eta_n = \eta n^{-\alpha}$, we get

$$\sum_{k=1}^{n}\eta_{k-1} = 1 + \sum_{k=1}^{n-1} k^{-\alpha} \leq 1 + \int_0^{n-1} t^{-\alpha}\,dt = 1 + \frac{n^{1-\alpha}}{1-\alpha}, \tag{VII.54}$$

and (VII.53) follows by substituting the above in (VII.52). □

Remark VII.6.5. If we take $h(x) = \sum_{i=1}^{d} x_i \log x_i$ and $\alpha = 1/2$, (VSFP) boils down to (EW′); the bound (VII.50) then also follows from (VII.53).

Online Gradient Descent. — The online gradient descent (OGD) algorithm was introduced by [Zin03] in the context of online convex optimization that we described in Section VII.2.3 – see also [Bub11, Section 4.1]. Here, we focus on a so-called *lazy* variant ([SS11, p. 144]) defined by means of the recursion

$$\begin{aligned} \mathrm{U}_n &\in \mathrm{U}_{n-1} - \eta\,\partial\ell_n(x_n), \\ x_{n+1} &= \operatorname*{arg\,min}_{x\in\mathscr{C}} \|x - \mathrm{U}_n\|^2, \end{aligned} \tag{OGD-L}$$

where $\ell_n \colon \mathscr{C} \to \mathbb{R}$ is a sequence of M-Lipschitz loss functions, $\eta > 0$ is a constant parameter, and the algorithm is initialized with $\mathrm{U}_0 = 0$.

In view of Example VII.3.5, (OGD-L) corresponds to the strategy $\sigma = (\sigma_n^{h,\eta})_{n\geqslant 1}$ generated by the Euclidean regularizer h on $\mathscr{C}$ – defined itself as in (VII.13). Theorem VII.5.1 thus yields the regret bound:

$$\max_{x\in\mathscr{C}} \frac{1}{n}\operatorname{Reg}_n(x) \leq \frac{\delta_{\mathscr{C}}^2}{2n\eta} + \frac{\eta M^2}{2} \tag{VII.55}$$

with $\delta_{\mathscr{C}}^2 = \max_{x\in\mathscr{C}} \|x\|_2^2 - \min_{x\in\mathscr{C}} \|x\|_2^2$. Accordingly, if the time horizon n is known in advance, the optimal choice for η is $\eta = \delta_{\mathscr{C}}/(M\sqrt{n})$, leading to a cumulative regret guarantee of $M\delta_{\mathscr{C}}\sqrt{n}$, which is essentially the bound derived by [SS11, Corollary. 2.7] (see also [Bub11, Theorem 3.1] for the greedy variant).[7]

Online Mirror Descent. — The family of (lazy) online mirror descent (OMD) algorithms studied by Shalev-Shwartz [SS07, SS11] is the most general family of strategies that we discuss in this section (see also [Bub11] for a greedy version). In particular, the OMD class of strategies contains EW and OGD as special cases, and it is also equivalent to the family of Follow the Regularized Leader (FtRL) algorithms in the case of linear payoffs ([SS11], [Haz12]).

Following [SS11] (and with notation as in Section VII.2.3), let $\ell_n : \mathscr{C} \to \mathbb{R}$ be a sequence of convex functions which are M-Lipschitz with respect to some norm $\|\cdot\|$ on $\mathbb{R}^d$. Then, given a regularizer function h on $\mathscr{C}$, the lazy OMD algorithm is defined by means of the recursion:

$$\begin{aligned} U_n &\in U_{n-1} - \eta\,\partial\ell_n(x_n), \\ x_{n+1} &= Q_h(U_n), \end{aligned} \tag{OMD-L}$$

where $\eta > 0$ is a *fixed* parameter and the algorithm is initialized with $U_0 = 0$. As a result, if h is taken K-strongly convex with respect to $\|\cdot\|$, Theorem VII.5.1 immediately yields the known regret bound for OMD:

$$\max_{x\in\mathscr{C}} \operatorname{Reg}_n(x) \leq \frac{h_{\max} - h_{\min}}{\eta} + \frac{\eta M^2 n}{2K}. \tag{VII.56}$$

VII.6.2. Links with convex optimization

Ordinary convex programs can be seen as online optimization problems where the loss function remains constant over time and the agent seeks to attain its minimum value. In what follows, we outline how regret-minimizing strategies can be used for this purpose and we describe the performance gap incurred by using a method with a variable step-size instead of a variable parameter.

7. For the difference between lazy and greedy variants, see Section VII.7.2.

Let $f\colon \mathscr{C} \to \mathbb{R}$ be a convex real-valued function on $\mathscr{C}$ and let $(\gamma_n)_{n\geq 1}$ be a positive sequence (which we will later interpret as a sequence of step-sizes); also, given a sequence $(x_n)_{n\geq 1}$ in $\mathscr{C}$, let

$$x_n^{\min} \in \operatorname*{arg\,min}_{1\leq k\leq n} f(x_k), \qquad x_n^{\gamma} = \frac{\sum_{k=1}^n \gamma_k x_k}{\sum_{k=1}^n \gamma_k}. \tag{VII.57}$$

If we use the notation $x_n' \in \{x_n^{\min}, x_n^{\gamma}\}$ to refer interchangeably to either $x_n^{\min}$ or x_n^{γ}, Jensen's inequality readily gives:

$$f(x_n') \leq \frac{\sum_{k=1}^n \gamma_k f(x_k)}{\sum_{k=1}^n \gamma_k}. \tag{VII.58}$$

Now consider the algorithm:

$$\begin{aligned} \mathrm{U}_n &\in \mathrm{U}_{n-1} - \gamma_n \partial f(x_n), \\ x_{n+1} &= \mathrm{Q}_h(\eta_n \mathrm{U}_n), \end{aligned} \tag{VII.59}$$

where γ_n is a sequence of step sizes and η_n is a sequence of parameters. In the case of a constant parameter $\eta_n = 1$, (VII.59) then becomes

$$\begin{aligned} \mathrm{U}_n &\in \mathrm{U}_{n-1} - \gamma_n \partial f(x_n), \\ x_{n+1} &= \mathrm{Q}_h(\mathrm{U}_n). \end{aligned} \tag{MD-L}$$

which is a lazy variant of the mirror descent (MD) algorithm ([NY83]). In particular, if h is the Euclidean regularizer on $\mathscr{C}$, the algorithm boils down to a lazy version of the standard projected subgradient (PSG) method:

$$\begin{aligned} \mathrm{U}_n &\in \mathrm{U}_{n-1} - \gamma_n \partial f(x_n), \\ x_{n+1} &= \operatorname*{arg\,min}_{x\in\mathscr{C}} \|x - \mathrm{U}_n\|_2. \end{aligned} \tag{PSG-L}$$

The following corollary shows that these lazy versions guarantee the same value convergence bounds as the corresponding greedy variants — see e.g. [BT03, Theorem 4.1].

Corollary VII.6.6 (Constant parameter, variable step size). *Let $f\colon \mathscr{C} \to \mathbb{R}$ be an* M*-Lipschitz convex function and let $(x_n)_{n\geq 1}$ be the sequence of play generated by* (MD-L) *for some* K*-strongly convex regularizer h on $\mathscr{C}$. Then, the adjusted iterates $x_n' \in \{x_n^{\min}, x_n^{\gamma}\}$ of x_n satisfy:*

$$f(x_n') \leq f_{\min} + \frac{h_{\max} - h_{\min} + \frac{1}{2}\mathrm{M}^2\mathrm{K}^{-1}\sum_{k=1}^n \gamma_k^2}{\sum_{k=1}^n \gamma_k}. \tag{VII.60}$$

Proof. With $\sigma = (\sigma_n^{h,\eta_n})_{n\geqslant 1}$, $u_k \in -\gamma_k \partial f(x_k)$ and $x'_n \in \{x_n^{\min}, x_n^{\gamma}\}$, we have:

$$\mathrm{Reg}_n^{\sigma,u}(x) = \sum_{k=1}^{n} \langle u_k | x - x_k \rangle \geq -\sum_{k=1}^{n} \gamma_k \left(f(x) - f(x_k)\right) \geq \sum_{k=1}^{n} \gamma_k \cdot \left(f(x'_n) - f(x)\right), \tag{VII.61}$$

where the last step follows from (VII.58). By taking $x \in \arg\min f$, we then obtain:

$$f(x'_n) - f_{\min} \leq \frac{\mathrm{Reg}_n^{\sigma,u}(x)}{\sum_{k=1}^{n} \gamma_k}. \tag{VII.62}$$

The result then follows by applying Theorem VII.5.1 and using the fact that $\|u_k\|_* \leq \|\gamma_k \partial f(x_k)\|_* \leq \gamma_k \mathrm{M}$ (recall that f is M-Lipschitz continuous). □

One can see that the best convergence rate that we get with constant η and step-sizes of the form $\gamma_n \propto n^{-\alpha}$ is $\mathrm{O}(\log n/\sqrt{n})$ for $\alpha = 1/2$ (and there is no straighforward choice of γ_n leading to a better convergence rate). On the other hand, by taking a *constant* step-size $\gamma_n = 1$ and *varying the algorithm's parameter* $\eta_n \propto n^{-1/2}$, we do achieve an $\mathrm{O}(n^{-1/2})$ rate of convergence.

Corollary VII.6.7 (Constant step size, variable parameter)**.** *With notation as in Corollary VII.6.6, let $(x_n)_{n\geq 1}$ be the sequence of play generated by (VII.59) with*

$$\eta_n = \frac{1}{\mathrm{M}} \sqrt{\frac{\mathrm{K}(h_{\max} - h_{\min})}{n}}, \tag{VII.63}$$

and constant $\gamma_n = 1$. Then, the adjusted iterates $x'_n \in \{x_n^{\min}, x_n^{\gamma}\}$ of x_n guarantee

$$f(x'_n) \leq f_{\min} + 2\mathrm{M} \sqrt{\frac{h_{\max} - h_{\min}}{\mathrm{K}}} \left(\frac{1}{\sqrt{n}} + \frac{1}{4n}\right). \tag{VII.64}$$

Proof. Similar to the proof of Corollary VII.6.6. □

VII.6.3. Noisy observations and links with stochastic convex optimization

Assume that at every stage $n = 1, 2, \ldots$ of the decision process, the agent does not observe the actual payoff vector $u_n \in \mathrm{V}^*$, but the realization of a random vector $\tilde{u}_n$ satisfying $\mathbb{E}[\tilde{u}_n | \mathcal{F}_n] = u_n$, where $\mathcal{F}_n$ is generated by

$$(\tilde{x}_1, u_1, \tilde{u}_1, i_1, \ldots, \tilde{x}_{n-1}, u_{n-1}, \tilde{u}_{n-1}, i_{n-1}, \tilde{x}_n, u_n).$$

In this case, a learning strategy σ can be used with the observed vectors $\tilde{u}_n$, thus leading to a (random) sequence of play $\tilde{x}_{n+1} = \sigma_{n+1}(\tilde{u}_1, \ldots, \tilde{u}_n)$ – see e.g. [SS11, Section 4.1] for a model of this kind.

In this framework, the agent's (maximal) cumulative regret, which is the quantity of interest, is given by

$$\max_{x\in\mathscr{C}} \sum_{k=1}^{n} \langle u_k | x\rangle - \sum_{k=1}^{n} \langle u_k | \tilde{x}_k\rangle \,. \tag{VII.65}$$

On the other hand,

$$\max_{x\in\mathscr{C}} \sum_{k=1}^{n} \langle \tilde{u}_k | x\rangle - \sum_{k=1}^{n} \langle \tilde{u}_k | \tilde{x}_k\rangle \,. \tag{VII.66}$$

can be interpreted as the agent's cumulative regret against the observed payoff sequence $(\tilde{u}_n)_{n\geq 1}$. The above two quantites can be related (in average) as follows. We assume that $\|\tilde{u}_k\|_* \leqslant \mathrm{M}$ (a.s.). As for the first term involving the maximum,

$$\begin{aligned}\max_{x\in\mathscr{C}} \frac{1}{n}\sum_{k=1}^{n} \langle u_k | x\rangle &= \max_{x\in\mathscr{C}} \frac{1}{n}\left\langle \sum_{k=1}^{n} \tilde{u}_k + \sum_{k=1}^{n}(u_k - \tilde{u}_k) \middle| x \right\rangle \\ &\leqslant \max_{x\in\mathscr{C}} \frac{1}{n}\sum_{k=1}^{n} \langle \tilde{u}_k | x\rangle + \left\| \frac{1}{n}\sum_{k=1}^{n}(u_k - \tilde{u}_k)\right\|_* \|\mathscr{C}\| \,,\end{aligned}$$

where the last term is small with high probability: indeed, since $\mathbb{E}\,[\tilde{u}_k - u_k \,|\, \mathscr{F}_k] = 0$, a classical argument based on bounded martingale differences can be used. We deal with the second sum similarly by noting that $\mathbb{E}\,[\langle \tilde{u}_k | \tilde{x}_k\rangle \,|\, \mathscr{F}_k] = \langle \mathbb{E}\,[\tilde{u}_k \,|\, \mathscr{F}_k] | \tilde{x}_k\rangle = \langle u_k | \tilde{x}_k\rangle$ and that:

$$\frac{1}{n}\sum_{k=1}^{n} \langle u_k | \tilde{x}_k\rangle = \frac{1}{n}\sum_{k=1}^{n} \langle \tilde{u}_k | \tilde{x}_k\rangle + \frac{1}{n}\sum_{k=1}^{n} \langle u_k - \tilde{u}_k | \tilde{x}_k\rangle \,.$$

The guarantees of Theorem VII.5.1 therefore translates to the present framwork with high probability.

The above can be adapted to the framework of stochastic convex optimization as follows: let $f\colon \mathscr{C} \to \mathbb{R}$ be a Lipschitz convex function on $\mathscr{C}$, let $(\gamma_n)_{n\geq 1}$ be a positive sequence of step sizes, and consider the strategy σ generated by (VII.18) with $\eta = 1$ and h a K-strongly convex regularizer on $\mathscr{C}$. Then, the sequence of play

$$\tilde{x}_{n+1} = \sigma_{n+1}(-\gamma_1\tilde{g}_1, \dots, -\gamma_n\tilde{g}_n) = \mathrm{Q}_h\left(-\sum_{k=1}^{n}\gamma_k\tilde{g}_k\right) \tag{VII.67}$$

where $\tilde{g}_n$ is a random vector with $\mathbb{E}\,[\quad \tilde{g}_n | \tilde{g}_{n-1}, \dots, \tilde{g}_1] \quad = g_n \in \partial f(\tilde{x}_n)$ may be written recursively as:

$$\begin{aligned}\widetilde{\mathrm{U}}_n &\in \widetilde{\mathrm{U}}_{n-1} - \gamma_n \partial f(\tilde{x}_n), \\ \tilde{x}_{n+1} &= \mathrm{Q}_h(\widetilde{\mathrm{U}}_n).\end{aligned} \tag{MDSA-L}$$

This algorithm may be seen as a lazy version of the so-called mirror descent stochastic approximation (MDSA) process of [NJLS09]; in particular, using the Euclidean regularizer leads to the lazy stochastic projected subgradient (SPSG) method:

$$\begin{aligned} \widetilde{U}_n &\in \widetilde{U}_{n-1} - \gamma_n \partial f(\tilde{x}_n), \\ \tilde{x}_{n+1} &= \operatorname*{arg\,min}_{x \in \mathscr{C}} \|x - \widetilde{U}_n\|_2. \end{aligned} \tag{SPSG-L}$$

Setting $u_n = -\gamma_n g_n$, $\tilde{u}_n = -\gamma_n \tilde{g}_n$ and taking $\tilde{x}'_n \in \{\tilde{x}_n^{\min}, \tilde{x}_n^{\gamma}\}$ as before, we can adapt Corollary VII.6.6 to we get, for all $x \in \mathscr{C}$,

$$\mathbb{E}\left[f(\tilde{x}'_n) - f(x)\right] \leqslant \mathbb{E}\left[\frac{1}{\sum_{k=1}^n \gamma_k} \sum_{k=1}^n \gamma_k (f(\tilde{x}_k) - f(x))\right] \tag{VII.68}$$

$$\leqslant \mathbb{E}\left[\frac{1}{\sum_{k=1}^n \gamma_k} \sum_{k=1}^n \langle u_k | x - \tilde{x}_k \rangle\right] \tag{VII.69}$$

$$= \mathbb{E}\left[\frac{1}{\sum_{k=1}^n \gamma_k} \sum_{k=1}^n \mathbb{E}\left[\langle \tilde{u}_k | x - \tilde{x}_k \rangle \mid \mathscr{F}_k\right]\right] \tag{VII.70}$$

$$= \mathbb{E}\left[\frac{1}{\sum_{k=1}^n \gamma_k} \sum_{k=1}^n \langle \tilde{u}_k | x - \tilde{x}_k \rangle\right] \tag{VII.71}$$

$$\leq \frac{h_{\max} - h_{\min} + \frac{1}{2}\mathrm{M}^2\mathrm{K}^{-1} \sum_{k=1}^n \gamma_k^2}{\sum_{k=1}^n \gamma_k}, \tag{VII.72}$$

which is essentially the same value guarantee as that of greedy MDSA ([NJLS09, Eq. 2.41]).

VII.7. Discussion

VII.7.1. On the optimal choice of h

As mentioned in the discussion after Corollary VII.5.2, the following open question arises: *given a norm $\|\cdot\|$ on* V *and a compact, convex subset $\mathscr{C} \subseteq$* V*, which* 1*-strongly convex regularizer on $h\colon \mathscr{C} \to \mathbb{R}$ has minimal depth $\delta_h = h_{\max} - h_{\min}$?*

As the following proposition shows, in the case of the Euclidean norm on V, this minimal depth is half the radius squared of the smallest enclosing sphere of $\mathscr{C}$:

Proposition VII.7.1. *Let $h\colon \mathscr{C} \to \mathbb{R}$ be a* 1*-strongly convex regularizer function on $\mathscr{C}$ with respect to the ℓ^2 norm $\|\cdot\|_2$ on* V*. Then:*

$$h_{\max} - h_{\min} \geq \frac{1}{2} \min_{x' \in \mathscr{C}} \max_{x \in \mathscr{C}} \|x' - x\|_2^2, \tag{VII.73}$$

ALGORITHM	$\mathscr{C}$	$h(x)$	η_n	INPUT	NORM
EW	Δ_d	$\sum_i x_i \log x_i$	CONSTANT	u_n	ℓ^1
EW′	Δ_d	$\sum_i x_i \log x_i$	$\eta/\sqrt{n}$	u_n	ℓ^1
SFP	Δ_d	ANY	η/n	u_n	ℓ^1
VSFP	Δ_d	ANY	$\eta n^{-\alpha}$ $(0<\alpha<1)$	u_n	ℓ^1
OGD-L	ANY	$\frac{1}{2}\|x\|_2^2$	CONSTANT	$-\nabla f_n(x_n)$	ℓ^2
OMD-L	ANY	ANY	CONSTANT	$-\nabla f_n(x_n)$	ANY
PSG-L	ANY	$\frac{1}{2}\|x\|_2^2$	1	$-\gamma_n \nabla f(x_n)$	ℓ^2
MD-L	ANY	ANY	1	$-\gamma_n \nabla f(x_n)$	ANY
MDSA-L	ANY	ANY	1	$-\gamma_n(\nabla f(x_n)+\xi_n)$	ANY
SPSG-L	ANY	$\frac{1}{2}\|x\|_2^2$	1	$-\gamma_n(\nabla f(x_n)+\xi_n)$	ℓ^2

Table VII.1. — Summary of the algorithms discussed in Section VII.6. The suffix "L" indicates a "lazy" variant; the INPUT column stands for the stream of payoff vectors which is used as input for the algorithm and the NORM column specifies the norm of the ambient space; finally, ξ_n represents a zero-mean stochastic process with values in $\mathbb{R}^d$.

and equality is attained by taking

$$h(x) = \begin{cases} \frac{1}{2}\|x - x_0\|_2^2 & \text{if } x \in \mathscr{C}, \\ +\infty & \text{otherwise}, \end{cases} \tag{VII.74}$$

where $x_0 \in \arg\min_{x' \in \mathscr{C}} \max_{x \in \mathscr{C}} \|x' - x\|_2^2$ *is the center of the smallest enclosing sphere of* $\mathscr{C}$.

Proof. Letting $x_1 \in \arg\min_{x \in \mathscr{C}} h(x)$ and $x_2 \in \arg\max_{x \in \mathscr{C}} \|x - x_1\|_2^2$, we readily get:

$$\begin{aligned} h_{\max} - h_{\min} &\geq h(x_2) - h(x_1) \\ &\geq \frac{1}{2}\|x_2 - x_1\|_2^2 = \frac{1}{2}\max_{x \in \mathscr{C}} \|x - x_1\|_2^2 \geq \frac{1}{2}\min_{x' \in \mathscr{C}} \max_{x \in \mathscr{C}} \|x - x'\|_2^2, \end{aligned} \tag{VII.75}$$

where the second inequality follows from the strong convexity of h and the fact that $\partial h(x_1) \ni 0$. That (VII.74) attains the bound (VII.73) is then a trivial consequence of its definition, as is its geometric characterization. □

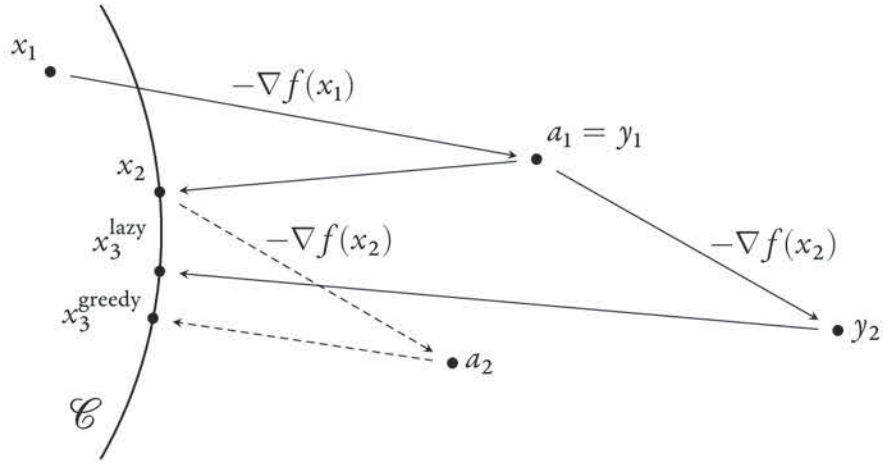


Figure VII.1. — Graphical illustration of the greedy (dashed) and lazy (solid) branches of the projected subgradient (PSG) method.

Despite the simplicity of the bound (VII.73), this analysis does not work for an arbitrary norm because $\frac{1}{2}\|x - x_0\|^2$ might fail to be 1-strongly convex with respect to $\|\cdot\|$ – for instance, $\|x - x_0\|_1^2$ is not even *strictly* convex.

VII.7.2. Greedy versus lazy

To illustrate the difference between *lazy* and *greedy* variants, we first focus on the PSG method run with constant step $\gamma = 1$ for a smooth function $f : \mathscr{C} \to \mathbb{R}$. The two variants may then be expressed by means of the recursions:

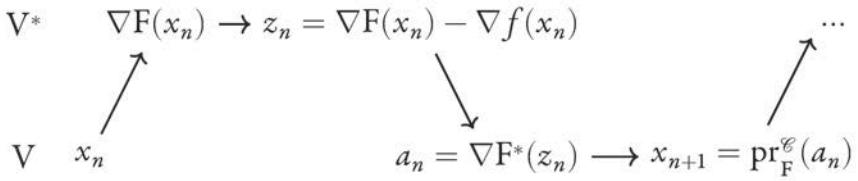
$$\begin{aligned} a_n &= x_n - \nabla f(x_n) \\ x_{n+1} &= \operatorname*{arg\,min}_{x \in \mathscr{C}} \|x - a_n\|_2 \end{aligned} \tag{VII.76a}$$

for the greedy version and:

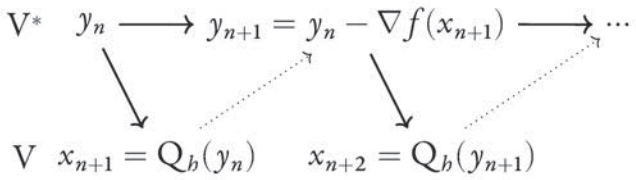
$$\begin{aligned} y_n &= y_{n-1} - \nabla f(x_n) \\ x_{n+1} &= \operatorname*{arg\,min}_{x \in \mathscr{C}} \|x - y_n\|_2 \end{aligned} \tag{VII.76b}$$

for the lazy one.

As can be seen in Fig. VII.1, the greedy variant is based on the classical idea of gradient descent, i.e. adding $-\nabla f(x_n)$ to x_n and projecting back to $\mathscr{C}$ if needed. On the other hand, in the lazy variant, the gradient term $-\nabla f(x_n)$ is *not* added to x_n, but to the "unprojected" iterate y_n; we only project to $\mathscr{C}$ in order to obtain the algorithm's next iterate. Owing to this modification, the lazy variant is thus driven by the sum $y_n = \sum_{k=1}^{n} \nabla f(x_n)$.



(a) Greedy Mirror Descent



(b) Lazy Mirror Descent

Figure VII.2. — Greedy and Lazy Mirror Descent with $\gamma_n = 1$.

In the case of mirror descent with an arbitrary regularizer function h, the lazy version has an implementation advantage over its greedy counterpart. Specifically, given a proper convex function F such that $F = h$ on $\mathscr{C}$ (cf. Example VII.3.6), greedy mirror descent is defined as:

$$\begin{aligned} a_n &= \nabla F^* \left(\nabla F(x_n) - \nabla f(x_n) \right), \\ x_{n+1} &= \mathrm{pr}_F^{\mathscr{C}}(a_n), \end{aligned} \tag{VII.77a}$$

where the Bregman projection $\mathrm{pr}_F^{\mathscr{C}}(a_n)$ is given by (VII.16); on the other hand, lazy MD is defined as

$$\begin{aligned} y_n &= y_{n-1} - \nabla f(x_n), \\ x_{n+1} &= Q_h(y_n). \end{aligned} \tag{VII.77b}$$

The computation steps for each variant are represented in Figure VII.2. The first step in the greedy version which consists in computing ∇F has no equivalent in the lazy version, which is thus computationally more lightweight.

CHAPTER VIII

A UNIVERSAL BOUND ON THE VARIATIONS OF BOUNDED CONVEX FUNCTIONS

This chapter is extracted from the paper *A universal bound on the variations of bounded convex functions*, to appear in *Journal of Convex Analysis*.

Abstract

Given a convex set C in a real vector space E and two points $x, y \in C$, we investigate which are the possible values for the variation $f(y) - f(x)$, where $f : C \longrightarrow [m, M]$ is a bounded convex function. We then rewrite the bounds in terms of the Funk weak metric, which will imply that a bounded convex function is Lipschitz-continuous with respect to the Thompson and Hilbert metrics. The bounds are also proved to be optimal. We also exhibit the maximal subdifferential of a bounded convex function at a given point $x \in C$.

VIII.1. The variations of bounded convex functions

Let C be a convex set of a real vector space E. Given two points $x, y \in C$, we define the following auxiliary quantity:

$$\tau_C(x, y) = \sup \{t \geqslant 1 \,|\, x + t(y - x) \in C\} .$$

Clearly, τ_C takes values in $[1, +\infty]$. Intuitively, it measures how far away x is from the boundary in the direction of y, taking the "distance" xy as unit. Clearly, $\tau_C(x, y) = +\infty$ if and only if $x + \mathbb{R}_+(y - x) \subset C$. Our first result is the following.

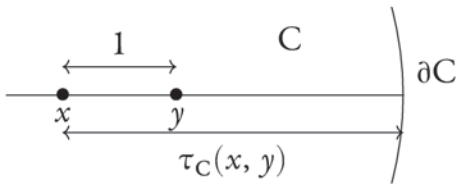


Figure VIII.1. — An intuitive representation of what $\tau_C(x, y)$ measures.

Theorem VIII.1.1. *Let $m \leqslant M$ be two real numbers. Let C be a convex set of a real vector space E and $f : C \longrightarrow [m, M]$ a convex function. For every couple of points $(x, y) \in C^2$, f satisfies:*

$$-\frac{M-m}{\tau_C(y,x)} \leqslant f(y) - f(x) \leqslant \frac{M-m}{\tau_C(x,y)}.$$

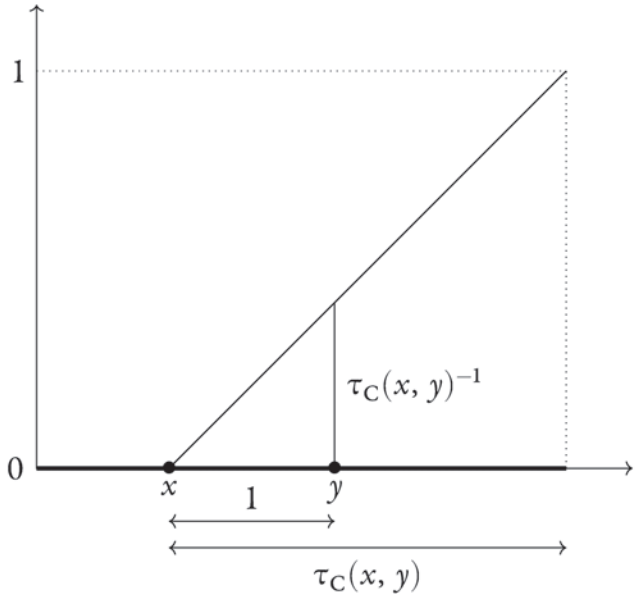


Figure VIII.2. — Illustration of the bound in the case $m = 0$ and $M = 1$. The thick horizontal line represents the cross section of C.

Proof. It is enough to prove the result for functions with values in $[0, 1]$, since we can consider $(M - m)^{-1}(f - m)$. Let x, y be two points in C. Let t be such that $1 \leqslant t < \tau_C(x, y)$. By definition of τ_C, and because C is convex, we have $x + t(y - x) \in C$. We can write y as a convex combination of $x + t(y - x)$ and x with coefficients $1/t$ and $(t - 1)/t$ respectively:

$$y = \frac{x + t(y - x) + (t - 1)x}{t}.$$

By convexity of f, we get:

$$\begin{aligned} f(y) - f(x) &\leqslant \frac{f(x + t(y-x)) + (t-1)f(x)}{t} - f(x) \\ &\leqslant \frac{f(x + t(y-x)) - f(x)}{t} \leqslant \frac{1}{t}, \end{aligned}$$

where the last inequality comes from the fact that f has values in $[0, 1]$. By taking the limit as $t \to \tau_C(x, y)$, we get:

$$f(y) - f(x) \leqslant \frac{1}{\tau_C(x, y)}.$$

The lower bound is obtained by exchanging the roles of x and y. □

VIII.2. The Funk, Thompson and Hilbert metrics

In this section, we rewrite the result from Theorem VIII.1.1 as a Lipschitz-like property in the framework of convex sets in normed spaces. But $1/\tau_C$ is far from being a distance. We thus consider the Funk, Thompson and Hilbert metrics (which were introduced in [Fun29], [Tho63] and [Hil95] respectively) and establish the link with τ_C.

We restrict our framework to the case where C is an open convex subset of a normed space $(E, \|\cdot\|)$. Let $x, y \in C$. If $\tau_C(x, y) < +\infty$, we can define $b(x, y)$ to be the following point:

$$b(x, y) = x + \tau_C(x, y)(y - x).$$

Note that since C is open, when $b(x, y)$ exists, it is necessarily different from y. This will be necessary to state the following definitions.

Definition VIII.2.1. Let C be an open convex subset of a normed space $(E, \|\cdot\|)$. We define

(i) the *Funk weak metric*:

$$F_C(x, y) = \begin{cases} \log \dfrac{\|x - b(x, y)\|}{\|y - b(x, y)\|} & \text{if } \tau_C(x, y) < +\infty \\ 0 & \text{otherwise} \end{cases};$$

(ii) the *Thompson pseudometric*:

$$T_C(x, y) = \max\left(F_C(x, y), F_C(y, x)\right);$$

(iii) the *Hilbert pseudometric*:

$$\mathrm{H}_{\mathrm{C}}(x, y) = \frac{1}{2}\left(\mathrm{F}_{\mathrm{C}}(x, y) + \mathrm{F}_{\mathrm{C}}(y, x)\right).$$

Remark VIII.2.2. Even if we will abusively call them *metrics*, they fail to satisfy the separation axiom in general. The Thompson and the Hilbert metrics are thus *pseudo-metrics*. Moreover, the Funk metric not being symmetric, it actually is a *weak* metric. The Thompson and the Hilbert metrics are respectively the *max-symmetrization* and *meanvalue-symmetrisation* of the Funk metric. For a detailed presentation of these notions, see e.g. [PT07].

We now establish the link between $\tau_{\mathrm{C}}(x, y)$ and $\mathrm{F}_{\mathrm{C}}(x, y)$.

Proposition VIII.2.3. *Let* C *be an open convex subset of a normed space* $(\mathrm{E}, \|\cdot\|)$. *For every points* $x, y \in \mathrm{C}$, *the following equality holds:*

$$\mathrm{F}_{\mathrm{C}}(x, y) = -\log\left(1 - \frac{1}{\tau_{\mathrm{C}}(x, y)}\right).$$

Proof. Let $x, y \in \mathrm{C}$. If $\tau_{\mathrm{C}}(x, y) = +\infty$, the right-hand side of the above equality is zero, as expected. If $\tau_{\mathrm{C}}(x, y) < +\infty$, $\tau_{\mathrm{C}}(x, y)$ can be expressed with the norm. Since by definition $b(x, y) = x + \tau_{\mathrm{C}}(x, y)(y - x)$, we have

$$\tau_{\mathrm{C}}(x, y) = \frac{\|x - b(x, y)\|}{\|x - y\|} \quad \text{and} \quad \tau_{\mathrm{C}}(x, y) - 1 = \frac{\|y - b(x, y)\|}{\|x - y\|}.$$

And thus:

$$\frac{\|x - b(x, y)\|}{\|y - b(x, y)\|} = \left(1 - \frac{1}{\tau_{\mathrm{C}}(x, y)}\right)^{-1}.$$

Therefore,

$$\mathrm{F}_{\mathrm{C}}(x, y) = -\log\left(1 - \frac{1}{\tau_{\mathrm{C}}(x, y)}\right).$$

□

By combining Theorem VIII.1.1 and the above proposition, we get the following corollary.

Corollary VIII.2.4. *Let* C *an open convex subset of a normed space* $(\mathrm{E}, \|\cdot\|)$ *and* $f : \mathrm{C} \longrightarrow [m, \mathrm{M}]$ *be a convex function. Then, for all* $x, y \in \mathrm{C}$, *the following bounds hold.*

(i) $-(\mathrm{M} - m)\left(1 - e^{-\mathrm{F}_{\mathrm{C}}(y,x)}\right) \leqslant f(y) - f(x) \leqslant (\mathrm{M} - m)\left(1 - e^{-\mathrm{F}_{\mathrm{C}}(x,y)}\right).$

(ii) $|f(y) - f(x)| \leqslant (\mathrm{M} - m)\left(1 - e^{-\mathrm{T}_{\mathrm{C}}(x,y)}\right).$

(iii) $|f(y) - f(x)| \leqslant (\mathrm{M} - m)\left(1 - e^{-2\mathrm{H}_{\mathrm{C}}(x,y)}\right)$.

Remark VIII.2.5. From (ii), by using the inequality $e^{-s} \geqslant 1 - s$, we get:

$$\begin{aligned} |f(x) - f(y)| &\leqslant (\mathrm{M} - m)\left(1 - e^{-\mathrm{T}_{\mathrm{C}}(x,y)}\right) \\ &\leqslant (\mathrm{M} - m)\mathrm{T}_{\mathrm{C}}(x, y), \end{aligned}$$

and similarly for (iii). Every convex function $f : \mathrm{C} \longrightarrow [m, \mathrm{M}]$ is thus $(\mathrm{M} - m)$-Lipschitz (resp. $2(\mathrm{M} - m)$-Lipschitz) with respect to the Thompson metric (resp. the Hilbert metric).

VIII.3. Optimality of the bounds

We show in this section that the bounds obtained in Theorem VIII.1.1 are optimal in the following sense. For a given convex set, and for a given couple a points, there is a function which attains the upper bound (resp. the lower bound). In other words, for $x, y \in \mathrm{C}$:

$$\begin{cases} \max\limits_{\substack{f:\mathrm{C}\longrightarrow[m,\mathrm{M}] \\ f \text{ convex}}} (f(y) - f(x)) = \dfrac{\mathrm{M} - m}{\tau_{\mathrm{C}}(x, y)} \\ \min\limits_{\substack{f:\mathrm{C}\longrightarrow[m,\mathrm{M}] \\ f \text{ convex}}} (f(y) - f(x)) = -\dfrac{\mathrm{M} - m}{\tau_{\mathrm{C}}(y, x)}. \end{cases}$$

In the proof of the following theorem, it will be very convenient to extend the notion of convexity to functions defined on C and taking values in $\mathbb{R} \cup \{-\infty\}$ (and not $\mathbb{R} \cup \{+\infty\}$). Obviously, the result according to which the upper envelope of two convex functions is also a convex function remains true.

Theorem VIII.3.1. *Let $m \leqslant \mathrm{M}$ be two real numbers. Let C be a convex set of a real vector space E. For every couple of points $(x, y) \in \mathrm{C}^2$, there exists a convex function $f : \mathrm{C} \longrightarrow [m, \mathrm{M}]$ (resp. $g : \mathrm{C} \longrightarrow [m, \mathrm{M}]$) such that the upper bound (resp. lower bound) of Theorem VIII.1.1 is attained; in other words:*

$$f(y) - f(x) = \frac{\mathrm{M} - m}{\tau_{\mathrm{C}}(x, y)} \qquad \left(\text{resp. } g(y) - g(x) = -\frac{\mathrm{M} - m}{\tau_{\mathrm{C}}(y, x)}\right).$$

Proof. Let x and y be two points in C, and let us construct a convex function $f : \mathrm{C} \longrightarrow [0, 1]$ satisfying the equality. If $\tau_{\mathrm{C}}(x, y) = +\infty$, the bound is zero, and $f = 0$ is adequate. From now on, we assume that $\tau_{\mathrm{C}}(x, y) < +\infty$. The idea of the construction is the following. Let us first consider the line through x and y. We want f to increase from 0 at x to 1 at the boundary in the direction of y, in an affine way; and to be equal to zero in the other direction. Then, we will have to extend f to all C in a convex way. Let

$\vec{u} = \tau_C(x, y)(y - x)$. For every $z \in C$, let us define $\sigma(z) = \sup\{t \geqslant 0 \mid z + t\vec{u} \in C\}$. σ clearly takes values in $[0, +\infty]$. Consider the following function.

$$\begin{array}{rccc} \phi : & C & \longrightarrow & [-\infty, 1] \\ & z & \longmapsto & 1 - \sigma(z) \end{array}.$$

Let us prove that ϕ is convex. Let z_1 and z_2 be two points in C and $z_3 = \lambda z_1 + (1 -$

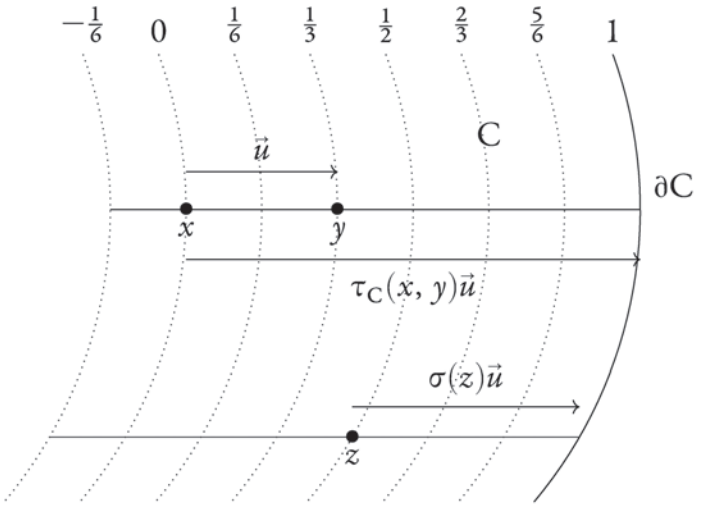


Figure VIII.3. — The construction of ϕ. The dotted curves are the level lines, whose corresponding values are specified above.

$\lambda)z_2$ (with $\lambda \in (0, 1)$) a convex combination. By definition of σ, if we take two real numbers s_1 and s_2 such that $0 \leqslant s_1 \leqslant \sigma(z_1)$ and $0 \leqslant s_2 \leqslant \sigma(z_2)$, we have:

$$\begin{cases} z_1 + s_1\vec{u} \in C \\ z_2 + s_2\vec{u} \in C. \end{cases}$$

And thus, the convex combination of these two points with coefficients λ and $1 - \lambda$ also belongs to C:

$$\lambda(z_1 + s_1\vec{u}) + (1 - \lambda)(z_2 + s_2\vec{u}) \in C.$$

This point can be rewritten with z_3:

$$z_3 + (\lambda s_1 + (1 - \lambda)s_2)\,\vec{u} \in C.$$

By definition of $\sigma(z_3)$, we have $\lambda s_1 + (1 - \lambda)s_2 \leqslant \sigma(z_3)$. This inequality is true for every $s_1 \leqslant \sigma(z_1)$ and $s_2 \leqslant z(s_2)$. Consequently:

$$\lambda\sigma(z_1) + (1 - \lambda)\sigma(z_2) \leqslant \sigma(z_3).$$

We can now prove the convexity inequality.

$$\begin{aligned}\phi(z_3) = 1 - \sigma(z_3) &\leqslant 1 - (\lambda\sigma(z_1) + (1-\lambda)\sigma(z_2)) \\ &= \lambda(1-\sigma(z_1)) + (1-\lambda)(1-\sigma(z_2)) \\ &= \lambda\phi(z_1) + (1-\lambda)\phi(z_2).\end{aligned}$$

We now choose $f = \max(\phi, 0)$. Since $\phi \leqslant 1$, f takes values in $[0,1]$. Let us prove that f satisfies the desired equality. Let us compute $f(x)$ and $f(y)$.

$$\begin{aligned}\sigma(x) &= \sup\{t \geqslant 0 \,|\, x + t\vec{u} \in \mathrm{C}\} \\ &= \sup\{t \geqslant 0 \,|\, x + t\tau_{\mathrm{C}}(x,y)(y-x) \in \mathrm{C}\} \\ &= \frac{1}{\tau_{\mathrm{C}}(x,y)} \sup\{t' \geqslant 0 \,|\, x + t'(y-x) \in \mathrm{C}\} \\ &= \frac{1}{\tau_{\mathrm{C}}(x,y)}\tau_{\mathrm{C}}(x,y) \\ &= 1.\end{aligned}$$

Thus $\phi(x) = 1 - \sigma(x) = 0$ and $f(x) = \max(0,0) = 0$. Similarly, we can prove:

$$\sigma(y) = \frac{\tau_{\mathrm{C}}(x,y) - 1}{\tau_{\mathrm{C}}(x,y)},$$

and thus, $\phi(y) = 1 - \sigma(y) = \tau_{\mathrm{C}}(x,y)^{-1}$ and $f(y) = \max(\tau_{\mathrm{C}}(x,y)^{-1}, 0) = \tau_{\mathrm{C}}(x,y)^{-1}$. We finally get:

$$f(y) - f(x) = \frac{1}{\tau_{\mathrm{C}}(x,y)}.$$

The construction of g is analogous. □

VIII.4. The maximal subdifferential

In the case of a nonempty convex subset $\mathrm{C} \subset \mathbb{R}^n$, and a given point $x_0 \in \mathrm{C}$, we wonder what is the maximal subdifferential at x_0 (in the sense of inclusion) for a function $f : \mathrm{C} \longrightarrow [m, \mathrm{M}]$. We will prove that there *is* a maximal one, and will express it in terms of the subdifferential of a translation of the Minkowski gauge. For each $x_0 \in \mathrm{C}$, we define $g_{\mathrm{C},x_0} : \mathrm{C} \longrightarrow [0,1]$ by

$$g_{\mathrm{C},x_0}(x) = \inf\{\lambda > 0 \,|\, x - x_0 \in \lambda(\mathrm{C} - x_0)\}.$$

This function is obviously well-defined, and can be seen as a Minkowski gauge centered in x_0 and restricted to C. It is well-known fact that the Minkowski gauge is a convex function. So is this one.

Theorem VIII.4.1. *Let* C *be a nonempty convex subset of* $\mathbb{R}^n$ *and* $x \in C$. *We have*

$$\max_{\substack{f:C\longrightarrow[m,M]\\ f\,convex}} \partial f(x) = (M-m)\partial g_{C,x}(x),$$

where the maximum is understood in the sense of inclusion.

Proof. Let us first relate g_{C,x_0} to τ. Let $x_0, x \in C$. We have

$$\begin{aligned} g_{C,x_0}(x) &= \inf\{\lambda > 0 \,|\, x - x_0 \in \lambda(C - x_0)\} \\ &= \sup\left\{t > 0 \,|\, x - x_0 \in \frac{1}{t}(C - x_0)\right\}^{-1} \\ &= \sup\{t > 0 \,|\, x_0 + t(x - x_0) \in C\}^{-1} \\ &= \frac{1}{\tau(x_0, x)}. \end{aligned}$$

Let us prove the result in the case $m = 0$ and $M = 1$, from which the general case follows immediately. Let $f : C \longrightarrow [0,1]$ be a convex function and $x_0 \in C$. Let us show that $\partial f(x_0) \subset \partial g_{C,x_0}(x_0)$. This is true if $\partial f(x_0)$ is empty. Otherwise, let $\zeta \in \partial f(x_0)$. For every $x \in C$, we have

$$\begin{aligned} \langle \zeta | x - x_0 \rangle \leqslant f(x) - f(x_0) &\leqslant \frac{1}{\tau(x_0, x)} \\ &= g_{C,x_0}(x) = g_{C,x_0}(x) - g_{C,x_0}(x_0), \end{aligned}$$

where we used Theorem VIII.1.1 for the second inequality. If $x \notin C$, the equality also holds, since $g_{C,x_0}(x) = +\infty$. We thus have $\partial f(x_0) \subset \partial g_{C,x_0}(x_0)$. We conclude by saying that g_{C,x_0} is a convex function on C with values in $[0,1]$. □

APPENDIX A
CONCENTRATION INEQUALITIES

Proposition A.0.1 (Hoeffding–Azuma for super-martingale differences [Hoe63, Azu67]). *Let $(X_t)_{t\geqslant 1}$ be a super-martingale difference sequence with respect to a filtration $(\mathcal{G}_t)_{t\geqslant 0}$:*

$$\mathbb{E}\left[X_t \mid \mathcal{G}_{t-1}\right] \leqslant 0, \quad t \geqslant 0.$$

Let $M > 0$ and we assume that $|X_t| \leqslant M$ almost-surely for all $t \geqslant 1$. Then, for all $\varepsilon > 0$ and $T \geqslant 1$,

$$\mathbb{P}\left[\frac{1}{T}\sum_{t=1}^{T} X_t > \varepsilon\right] \leqslant \exp\left(-\frac{\varepsilon^2 T}{2M^2}\right).$$

Proposition A.0.2 (Corollary 3.5 in [KS91]). *Let $(U_t)_{t\geqslant 1}$ be a sequence of martingale differences in $\mathbb{R}^d$, bounded almost-surely by $M > 0$:*

$$\forall t \geqslant 1, \quad \|U_t\|_2 \leqslant M, \quad a.s.$$

Then, for every $\varepsilon > 0$ and $T \geqslant 1$,

$$\mathbb{P}\left[\left\|\frac{1}{T}\sum_{t=1}^{T} U_t\right\|_2 \geqslant \varepsilon\right] \leqslant 2\exp\left(-\frac{T\varepsilon^2}{4M^2}\right).$$

Corollary A.0.3. *Under the assumptions of Proposition A.0.2, we have:*

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{T} U_t\right\|_2\right] \leqslant M\sqrt{\frac{\pi}{T}}.$$

Proof. The result follows from Proposition A.0.2 by integrating the tail of the distribution:

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{U}_T\right\|_2\right] &= \int_0^{+\infty} \mathbb{P}\left[\left\|\bar{U}_T\right\|_2 \geqslant \varepsilon\right] \mathrm{d}\varepsilon \leqslant \int_0^{+\infty} 2e^{-T\varepsilon^2/4M^2}\,\mathrm{d}\varepsilon \\ &= 2\int_0^{+\infty} e^{-\varepsilon^2(T/4M^2)}\,\mathrm{d}\varepsilon = M\sqrt{\frac{\pi}{T}}.\end{aligned}$$

□

The following Bernstein-like inequality is proved in [Pin94]—see also [TY14, Corollary A.2].

Proposition A.0.4. *Let $(X_t)_{t\geqslant 1}$ be a martingale difference sequence in a Hilbert space with respect to a filtration $(\mathcal{G}_t)_{t\geqslant 0}$. Suppose that $\|X_t\| \leqslant M$ almost-surely, and*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|X_t\|^2 \,\middle|\, \mathcal{G}_{t-1}\right] \leqslant V.$$

Then,

$$\mathbb{P}\left[\max_{1\leqslant t\leqslant T}\left\|\sum_{t'=1}^{t} X_{t'}\right\| \geqslant \varepsilon\right] \leqslant 2\exp\left(-\frac{\varepsilon^2}{2TV + 2M\varepsilon/3}\right).$$

Corollary A.0.5. *Under the assumptions of Proposition A.0.4,*

$$\mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{T} X_t\right\|\right] \leqslant 4\sqrt{2}\sqrt{\frac{V}{T}} + \frac{4M}{3T}.$$

Proof. Let $A \geqslant 0$ to be chosen later.

$$\begin{aligned}
\mathbb{E}\left[\|\bar{X}_T\|\right] &= \int_0^{+\infty} \mathbb{P}\left[\|\bar{X}_T\| \geqslant \varepsilon\right] \,\mathrm{d}\varepsilon \\
&\leqslant 2\int_0^{+\infty} \exp\left(-\frac{\varepsilon^2 T^2}{2VT + 2M\varepsilon T/3}\right) \mathrm{d}\varepsilon \\
&= 2\int_0^{+\infty} \exp\left(-\frac{\varepsilon^2 T}{2V + 2M\varepsilon/3}\right) \mathrm{d}\varepsilon \\
&\leqslant 2\left(A + \int_A^{+\infty} \exp\left(-\frac{\varepsilon^2 T}{2\varepsilon(V/A + M/3)}\right) \mathrm{d}\varepsilon\right) \\
&= 2\left(A + \int_A^{+\infty} \exp\left(-\frac{\varepsilon T}{2(V/A + M/3)}\right) \mathrm{d}\varepsilon\right) \\
&= 2\left(A + \left[-\frac{2}{T}\left(\frac{V}{A} + \frac{M}{3}\right)\exp\left(-\frac{\varepsilon T}{2(V/A + M/3)}\right)\right]_A^{+\infty}\right) \\
&\leqslant 2A + \frac{4}{T}\left(\frac{V}{A} + \frac{M}{3}\right).
\end{aligned}$$

Choosing $A = \sqrt{2V/T}$ gives:

$$\mathbb{E}\left[\|\bar{X}_T\|\right] \leqslant 4\sqrt{2}\sqrt{\frac{V}{T}} + \frac{4M}{3T}.$$

□

Bibliography

[AB09] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, pages 217–226, 2009.

[ABH11] Jacob Abernethy, Peter L. Bartlett, and Elad Hazan. Blackwell approachability and low-regret learning are equivalent. In *JMLR: Workshop and Conference Proceedings (COLT)*, volume 19, pages 27–46, 2011.

[ABL13] Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2013.

[ACBFS02] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[ACBG02] Peter Auer, Nicolo Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

[AHR08] Jacob Abernethy, Elad Hazan, and Alexander 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.

[AHR12] Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Interior-point methods for full-information and bandit online learning. *IEEE Transactions on Information Theory*, 58(7):4164–4175, 2012.

[AK04] Baruch Awerbuch and Robert D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, pages 45–53. ACM, 2004.

[ALST14] Jacob Abernethy, Chansoo Lee, Abhinav Sinha, and Ambuj Tewari. Online linear optimization via smoothing. In *JMLR: Workshop and Conference Proceedings (COLT)*, volume 35, pages 807–823, 2014.

[ALT16] Jacob Abernethy, Chansoo Lee, and Ambuj Tewari. Perturbation techniques in online learning and optimization. In Tamir Hazan, George Papandreou, and Daniel Tarlow, editors, *Perturbations, Optimization, and Statistics*, Neural Information Processing Series, chapter 8. MIT Press, 2016. to appear.

[AM85] Robert J. Aumann and Michael Maschler. Game theoretic analysis of a bankruptcy problem from the Talmud. *Journal of Economic Theory*, 36(2):195–213, 1985.

[AYPS12] Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *JMLR: Workshop and Conference Proceedings (AISTATS)*, volume 22, pages 1–9, 2012.

[Azu67] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

[BCB12] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, 5(1):1–122, 2012.

[BDH+08] Peter L. Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 2008.

[Ber73] Dimitri P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, 1973.

[BF13] Michel Benaïm and Mathieu Faure. Consistency of vanishingly smooth fictitious play. *Mathematics of Operations Research*, 38(3):437–450, 2013.

[BFP+14] Gábor Bartók, Dean P Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring – classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

[BHS06] Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. Part II: Applications. *Mathematics of Operations Research*, 31(4):673–695, 2006.

[BL10] Jonathan M. Borwein and Adrian S. Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer, 2010.

[Bla54] David Blackwell. Controlled random walks. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 336–338, 1954.

[Bla56] David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956.

[BM05] Avrim Blum and Yishay Mansour. From external to internal regret. In *Learning Theory*, pages 621–636. Springer, 2005.

[BMS14] Andrey Bernstein, Shie Mannor, and Nahum Shimkin. Opportunistic approachability and generalized no-regret problems. *Mathematics of Operations Research*, 39(4):1057–1083, 2014.

[BPS10] Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Toward a classification of finite partial-monitoring games. In *Proceedings of the 21st International Conference on Algorithmic Learning Theory (ALT)*, pages 224–238. Springer, 2010.

[Bre67] Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[BT03] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[Bub11] Sébastien Bubeck. *Introduction to Online Optimization: Lecture Notes*. Princeton University, 2011.

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[CB97] Nicolo Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory (COLT)*, pages 163–170. ACM, 1997.

[CBFH+97] Nicolo Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

[CBL03] Nicolo Cesa-Bianchi and Gábor Lugosi. Potential-based algorithms in on-line prediction and game theory. *Machine Learning*, 51(3):239–261, 2003.

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[CBLS05] Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005.

[CBLS06] Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.

[CH15] Alon Cohen and Tamir Hazan. Following the perturbed leader for online structured learning. In *JMLR: Workshop and Conference Proceedings (ICML)*, volume 37, pages 1034–1042, 2015.

[CM12] Alexandra Carpentier and Rémi Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 190–198, 2012.

[CZ92] Yair Censor and Stavros Andrea Zenios. Proximal minimization algorithm with D-functions. *Journal of Optimization Theory and Applications*, 73(3):451–464, 1992.

[DH06] Varsha Dani and Thomas P. Hayes. Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, pages 937–943. Society for Industrial and Applied Mathematics, 2006.

[DKC13] Josip Djolonga, Andreas Krause, and Volkan Cevher. High-dimensional gaussian process bandits. In *Advances in Neural Information Processing Systems (NIPS)*, volume 26, pages 1025–1033, 2013.

[DLN13] Luc Devroye, Gábor Lugosi, and Gergely Neu. Prediction by random-walk perturbation. In *JMLR: Workshop and Conference Proceedings (COLT)*, volume 30, pages 460–473, 2013.

[FKM05] Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394. Society for Industrial and Applied Mathematics, 2005.

[FL95] Drew Fudenberg and David K. Levine. Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19(5):1065–1089, 1995.

[FL98] Drew Fudenberg and David K. Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

[FL99] Drew Fudenberg and David K. Levine. Conditional universal consistency. *Games and Economic Behavior*, 29(1):104–130, 1999.

[Fos99] Dean P. Foster. A proof of calibration via Blackwell's approachability theorem. *Games and Economic Behavior*, 29(1):73–78, 1999.

[FS97] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[Fun29] Paul Funk. Über geometrien, bei denen die geraden die kürzesten sind. *Mathematische Annalen*, 101(1):226–237, 1929.

[FV97] Dean P. Foster and Rakesh V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1):40–55, 1997.

[Gal78] Janos Galambos. *The asymptotic theory of extreme order statistics*. John Wiley, New York, 1978.

[Ger13] Sébastien Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *The Journal of Machine Learning Research*, 14(1):729–769, 2013.

[GLS01] Adam J. Grove, Nick Littlestone, and Dale Schuurmans. General convergence results for linear discriminant updates. *Machine Learning*, 43(3):173–210, 2001.

[Gol64] Alan A. Goldstein. Convex programming in Hilbert space. *Bulletin of the American Mathematical Society*, 70(5):709–710, 1964.

[Gor99] Geoffrey J. Gordon. Regret bounds for prediction problems. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT)*, pages 29–40. ACM, 1999.

[GW98] Claudio Gentile and Manfred K. Warmuth. Linear hinge loss and average margin. In *Advances in Neural Information Processing Systems (NIPS)*, volume 11, pages 225–231, 1998.

[Han57] James Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3(97-139):2, 1957.

[Haz12] Elad Hazan. The convex optimization approach to regret minimization. In S. Nowozin S. Sra and S. Wrigh, editors, *Optimization for Machine Learning*, pages 287–303. MIT press, 2012.

[Hil95] David Hilbert. Über die gerade linie als kürzeste verbindung zweier punkte. *Mathematische Annalen*, 46(1):91–96, 1895.

[HKW10] Elad Hazan, Satyen Kale, and Manfred K. Warmuth. Learning rotations with little regret. In *Proceedings of the 23rd Conference on Learning Theory (COLT)*, pages 144–154, 2010.

[HMC00] Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68:1127–1150, 2000.

[HMC01] Sergiu Hart and Andreu Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98(1):26–54, 2001.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[HP04] Marcus Hutter and Jan Poland. Prediction with expert advice by following the perturbed leader for general weights. In *Proceedings of the 15th International Conference on Algorithmic Learning Theory (ALT)*, pages 279–293. Springer, 2004.

[HS02] Josef Hofbauer and William H. Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294, 2002.

[HW09] David P. Helmbold and Manfred K. Warmuth. Learning permutations with exponential weights. *The Journal of Machine Learning Research*, 10:1705–1736, 2009.

[KM14] Joon Kwon and Panayotis Mertikopoulos. A continuous-time approach to online optimization. *arXiv:1401.6956*, 2014.

[Koh75] E. Kohlberg. Optimal strategies in repeated games with incomplete information. *International Journal of Game Theory*, 4(1):7–24, 1975.

[KP16a] Joon Kwon and Vianney Perchet. Blackwell approachability with partial monitoring: Optimal convergence rates. 2016.

[KP16b] Joon Kwon and Vianney Perchet. Gains and losses are fundamentally different in regret minimization: the sparse case. *arXiv:1511.08405*, 2016.

[KS91] Olav Kallenberg and Rafal Sztencel. Some dimension-free features of vector-valued martingales. *Probability Theory and Related Fields*, 88(2):215–247, 1991.

[KSST12] Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, 13(1):1865–1890, 2012.

[KV05] Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

[KW95] Jyrki Kivinen and Manfred K. Warmuth. Additive versus exponentiated gradient updates for linear prediction. In *Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing*, pages 209–218. ACM, 1995.

[KW97] Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

[KW01] Jyrki Kivinen and Manfred K. Warmuth. Relative loss bounds for multidimensional regression problems. *Machine Learning*, 45(3):301–329, 2001.

[KWK10] Wouter M Koolen, Manfred K. Warmuth, and Jyrki Kivinen. Hedging structured concepts. In *Proceedings of the 23rd Conference on Learning Theory (COLT)*, pages 93–105, 2010.

[Kwo14] Joon Kwon. A universal bound on the variations of bounded convex functions. *arXiv:1401.2104*, 2014.

[Leh03] Ehud Lehrer. A wide range no-regret theorem. *Games and Economic Behavior*, 42(1):101–115, 2003.

[LMS08] Gábor Lugosi, Shie Mannor, and Gilles Stoltz. Strategies for prediction under imperfect monitoring. *Mathematics of Operations Research*, 33(3):513–528, 2008.

[LP66] Evgeny S. Levitin and Boris T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1–50, 1966.

[LR85] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[LS07] Ehud Lehrer and Eilon Solan. Learning to play partially-specified equilibrium. *Levine's Working Paper Archive*, 2007.

[LW94] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

[MCWG95] Andreu Mas-Colell, Michael Dennis Whinston, and Jerry R. Green. *Microeconomic theory*. Oxford University Press, 1995.

[Mor62] Jean-Jacques Moreau. Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires. *Comptes rendus de l'Académie des Sciences*, 255:238–240, 1962.

[MPS11] Shie Mannor, Vianney Perchet, and Gilles Stoltz. Robust approachability and regret minimization in games with partial monitoring. In *JMLR: Workshop and Conference Proceedings (COLT)*, volume 19, pages 515–536, 2011.

[MPS13] Shie Mannor, Vianney Perchet, and Gilles Stoltz. A primal condition for approachability with partial monitoring. *Journal of Dynamics and Games*, 1(3):447–469, 2013.

[MPS14] Shie Mannor, Vianney Perchet, and Gilles Stoltz. Set-valued approachability and online learning with partial monitoring. *The Journal of Machine Learning Research*, 15(1):3247–3295, 2014.

[MS03] Shie Mannor and Nahum Shimkin. On-line learning with imperfect monitoring. In *Learning Theory and Kernel Machines*, pages 552–566. Springer, 2003.

[MS10] Shie Mannor and Gilles Stoltz. A geometric proof of calibration. *Mathematics of Operations Research*, 35(4):721–727, 2010.

[NB13] Gergely Neu and Gábor Bartók. An efficient algorithm for learning with semi-bandit feedback. In *Proceedings of the 24th International Conference on Algorithmic Learning Theory (ALT)*, pages 234–248. Springer, 2013.

[Nes09] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

[NJLS09] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[NY83] Arkadi Nemirovski and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.

[OCCB15] Francesco Orabona, Koby Crammer, and Nicolo Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3):411–435, 2015.

[Per10] Vianney Perchet. *Approchabilité, calibration et regret dans les jeux à informations partielles*. PhD thesis, Université Pierre-et-Marie-Curie, 2010.

[Per11a] Vianney Perchet. Approachability of convex sets in games with partial monitoring. *Journal of Optimization Theory and Applications*, 149(3):665–677, 2011.

[Per11b] Vianney Perchet. Internal regret with partial monitoring: Calibration-based optimal algorithms. *The Journal of Machine Learning Research*, 12:1893–1921, 2011.

[Per14] Vianney Perchet. Approachability, regret and calibration: Implications and equivalences. *Journal of Dynamics and Games*, 1(2):181–254, 2014.

[Per15] Vianney Perchet. Exponential weight approachability, applications to calibration and regret minimization. *Dynamic Games and Applications*, 5(1):136–153, 2015.

[Pin94] Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

[PM13] Vianney Perchet and Shie Mannor. Approachability, fast and slow. In *JMLR: Workshop and Conference Proceedings (COLT)*, volume 30, pages 474–488, 2013.

[PQ14] Vianney Perchet and Marc Quincampoix. On a unified framework for approachability with full or partial monitoring. *Mathematics of Operations Research*, 40(3):596–610, 2014.

[PS01] Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *Proceedings of the 14th Annual Conference on Computational Learning Theory (COLT)*, pages 208–223. Springer, 2001.

[PT07] Athanase Papadopoulos and Marc Troyanov. Weak metrics on Euclidean domains. *JP Journal of Geometry and Topology*, 7(1):23–44, 2007.

[Rob52] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[Roc70] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[RST11] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Beyond regret. In *JMLR: Workshop and Conference Proceedings (COLT)*, volume 19, pages 559–594, 2011.

[RT09] Alexander Rakhlin and Ambuj Tewari. Lecture notes on online learning. 2009.

[Rus99] Aldo Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29(1):224–243, 1999.

[RW98] R. Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*. Springer-Verlag, Berlin, 1998.

[RZ96] Jörg Rambau and Günter M. Ziegler. Projections of polytopes and the generalized Baues conjecture. *Discrete & Computational Geometry*, 16(3):215–237, 1996.

[Shi15] Nahum Shimkin. An online convex optimization approach to Blackwell's approachability. *arXiv:1503.00255*, 2015.

[SL05] Gilles Stoltz and Gábor Lugosi. Internal regret in on-line portfolio selection. *Machine Learning*, 59(1-2):125–159, 2005.

[Sle62] David Slepian. The one-sided barrier problem for gaussian noise. *Bell System Technical Journal*, 41(2):463–501, 1962.

[Sor09] Sylvain Sorin. Exponential weight algorithm in continuous time. *Mathematical Programming*, 116(1-2):513–528, 2009.

[Spi02] Xavier Spinat. A necessary and sufficient condition for approachability. *Mathematics of Operations Research*, 27(1):31–44, 2002.

[SS07] Shai Shalev-Shwartz. *Online learning: Theory, algorithms, and applications*. PhD thesis, The Hebrew University of Jerusalem, 2007.

[SS11] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

[Tho63] A. C. Thompson. On certain contraction mappings in a partially ordered vector space. *Proceedings of the American Mathematical Society*, 14(3):438–443, 1963.

[TW03] Eiji Takimoto and Manfred K. Warmuth. Path kernels and multiplicative updates. *The Journal of Machine Learning Research*, 4:773–818, 2003.

[TY14] Pierre Tarres and Yuan Yao. Online learning as stochastic approximation of regularization paths: optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9):5716–5735, 2014.

[VEKW14] Tim Van Erven, Wojciech Kotlowski, and Manfred K. Warmuth. Follow the leader with dropout perturbations. In *JMLR: Workshop and Conference Proceedings (COLT)*, volume 35, pages 949–974, 2014.

[Vov90] Volodimir G. Vovk. Aggregating strategies. In *Proceedings of the Third Workshop on Computational Learning Theory (COLT)*, pages 371–383. Morgan Kaufmann, 1990.

[Vov98] Vladimir G. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 2(56):153–173, 1998.

[WJ97] Manfred K. Warmuth and Arun K. Jagota. Continuous and discrete-time nonlinear gradient descent: Relative loss bounds and convergence. In *Electronic Proceedings of the 5th International Symposium on Artificial Intelligence and Mathematics*, 1997.

[WK08] Manfred K. Warmuth and Dima Kuzmin. Randomized online PCA algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9(10):2287–2320, 2008.

[Zin03] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, 2003.

Index