

Regret bounds

*and their applications to online learning,
optimization and games*

Joon Kwon

February 18, 2024

Contents

1	Convexity tools	4
1.1	Preliminaries	4
1.2	Convexity	5
1.3	Subgradients	7
1.4	Legendre–Fenchel transform	8
1.5	Bregman divergences	9
1.6	Strong convexity and smoothness	10
2	UMD theory	13
2.1	Introduction	13
2.2	Regularizers	16
2.3	UMD iterates	17
2.4	Regret bounds	19
2.5	Time-dependent regularizers	20
3	Online optimization	22
3.1	Introduction to regret minimization	22
3.2	Dual averaging	26
3.3	Online mirror descent	29

Foreword

As of 2024, this document contains lecture notes from a course given in Master 2 in *Université Paris–Saclay*. These are highly incomplete and constantly updated as the lectures are given.

The course proposes a unified presentation of regret minimization for adversarial online learning problems, and its application to various problems such as Blackwell’s approachability, optimization algorithms (GD, Nesterov, SGD, AdaGrad), variational inequalities with monotone operators (Extra-gradient, Mirror-Prox, Dual Extrapolation), fixed-point iterations (Krasnoselskii–Mann), and games. The presentation aims at being modular, so that introduced tools and techniques could easily be used to define and analyze new algorithms.

The central notion of this presentation is the *regret*, which will be analyzed using the Legendre–Fenchel transform and Bregman divergences. An excellent recent monograph on the topic of online learning is the following:

- Francesco Orabona. A modern introduction to online learning. *arXiv:1912.13213*, 2023.

Additional notable references on the topic include:

- H. Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *The Journal of Machine Learning Research*, 18(1):3117–3166, 2017,
- Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016,
- Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011,
- Sébastien Bubeck. *Introduction to Online Optimization: Lecture Notes*. Princeton University, 2011,
- Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Regarding convex analysis, we refer to the following classical book:

-
- R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Acknowledgements

These notes highly benefited from discussions with Sylvain Sorin, Gilles Stoltz, Jaouad Mourtada, and the encouragements from Liliane Bel.

Chapter 1

Convexity tools

We present the basic convexity notions and tools that will be used in the subsequent chapters. Most of them are classical and are given without proof.

1.1 Preliminaries

Let $d \geq 1$. Throughout the chapter, we consider Euclidean space \mathbb{R}^d equipped with its canonical inner product denoted $\langle \cdot, \cdot \rangle$. For a set $A \subset \mathbb{R}^d$, $\text{int } A$ and $\text{cl } A$ denote its interior and closure, respectively.

Definition 1.1.1 (Domain of a function). The *domain* of a function $f : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ is the set

$$\text{dom } f = \left\{ x \in \mathbb{R}^d, f(x) < +\infty \right\}.$$

f is said to be *proper* if its domain is nonempty.

Definition 1.1.2 (Dual norm). Let $\|\cdot\|$ be a norm in \mathbb{R}^d . Its *dual norm* is defined as

$$\|y\|_* = \max_{\|x\| \leq 1} \langle y, x \rangle, \quad y \in \mathbb{R}^d.$$

Remark 1.1.3. The above maximum is indeed attained because for a given $y \in \mathbb{R}^d$, function $x \mapsto \langle y, x \rangle$ is continuous on the closed unit ball, which is compact. Besides, one can check that the dual norm is indeed a norm.

Proposition 1.1.4. Let $\|\cdot\|$ be a norm in \mathbb{R}^d . Then, $\|\cdot\|_{**} = \|\cdot\|$.

Example 1.1.5 (Common dual norms). In \mathbb{R}^d , ℓ_2 is its own dual norm, ℓ_p and ℓ_q (with $p, q \geq 1$ such that $1/p + 1/q = 1$) are dual of each other, and ℓ_1 and ℓ_∞ are dual of each other. If A is a positive definite matrix, the dual norm of the associated Mahalanobis norm $x \mapsto \sqrt{\langle x, Ax \rangle}$ is the Mahalanobis norm associated with A^{-1} .

Remark 1.1.6. It follows from the definition of the dual norm that for all $x, y \in \mathbb{R}^d$, $\langle y, x \rangle \leq \|y\|_* \|x\|$, which, together with the above examples recovers Cauchy-Schwarz and Hölder's inequalities.

For a function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ and $x \in \text{int dom } f$, if f is differentiable in x (resp. twice differentiable), we denote $\nabla f(x)$ its gradient at x (resp. $\nabla^2 f(x)$ its Hessian matrix at x).

1.2 Convexity

Definition 1.2.1. A set $\mathcal{X} \subset \mathbb{R}^d$ is *convex* if for all $x, x' \in \mathcal{X}$ and $\lambda \in [0, 1]$, $\lambda x + (1 - \lambda)x' \in \mathcal{X}$.

Example 1.2.2 (Unit balls). For all norms, as an immediate consequence of the triangle inequality, the unit ball is convex.

Example 1.2.3 (Simplex). Denote Δ_d the simplex in \mathbb{R}^d :

$$\Delta_d = \left\{ x \in \mathbb{R}_+^d, \sum_{i=1}^d x_i = 1 \right\},$$

which is a closed convex set. Note that it is contained in a hyperplane and therefore has empty interior.

Proposition 1.2.4 (Euclidean projection on a closed convex set). *Let $\mathcal{X} \subset \mathbb{R}^d$ be a closed convex set and $x \in \mathbb{R}^d$. Then, the Euclidean projection of x onto \mathcal{X} exists and is unique. In other words, $x' \mapsto \|x' - x\|_2^2$ admits a unique minimizer on \mathcal{X} .*

Definition 1.2.5 (Convex functions). A function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ is *convex* if for all $x, x' \in \mathbb{R}^d$ and $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)x') \leq \lambda f(x) + (1 - \lambda)f(x').$$

f is *strictly convex* if the above inequality is strict for $\lambda \in (0, 1)$.

Remark 1.2.6. The convexity of a function f is closely related to the convexity of sets, as the former can be equivalently defined as the epigraph

$$\text{epi } f = \left\{ (a, x) \in \mathbb{R} \times \mathbb{R}^d, a \geq f(x) \right\}$$

being convex.

Example 1.2.7. The following functions are convex: linear functions, quadratic functions of the form $x \mapsto \langle x, Ax \rangle$ where A is a positive semi-definite matrix, the exponential, the negative logarithm, convex combinations of convex functions, the point-wise supremum of convex functions. Let $\|\cdot\|$ be a norm in \mathbb{R}^d and $a \geq 1$. Function $x \mapsto \|x\|^a$ is convex.

Proposition 1.2.8 (Jensen's inequality). *Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a convex function, and X a random variable with values in $\text{dom } f$ so that $\mathbb{E}[X]$ exists. Then, f is measurable, $\mathbb{E}[f(X)]$ exists in $\mathbb{R} \cup \{+\infty\}$ and*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

In particular, for $n \geq 1$, $x_1, \dots, x_n \in \mathbb{R}^n$ and $\lambda_1, \dots, \lambda_n \geq 0$ such that $\lambda_1 + \dots + \lambda_n > 0$,

$$f\left(\frac{\sum_{i=1}^n \lambda_i x_i}{\sum_{i=1}^n \lambda_i}\right) \leq \frac{\sum_{i=1}^n \lambda_i f(x_i)}{\sum_{i=1}^n \lambda_i}.$$

Proposition 1.2.9 (First and second order characterizations of convexity). *Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ a function with open domain.*

(i) *If f is differentiable on its domain, f is convex if, and only if, for all $x, x' \in \text{dom } f$,*

$$f(x') \geq f(x) + \langle \nabla f(x), x' - x \rangle.$$

(ii) *If f is twice differentiable on its domain, f is convex if, and only if, for all $x \in \text{dom } f$, $\nabla^2 f(x)$ is positive semi-definite.*

Proposition 1.2.10. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a nonempty convex set, and $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a convex function, differentiable on an open set containing \mathcal{X} . Then, $x_* \in \mathcal{X}$ is a minimizer of f on \mathcal{X} if, and only if,*

$$\forall x \in \mathcal{X}, \langle \nabla f(x_*), x - x_* \rangle \geq 0.$$

Definition 1.2.11 (Lower semicontinuity). *A function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous if for all $a \in \mathbb{R}^d$, the set $\{x \in \mathbb{R}^d, f(x) \leq a\}$ is closed.*

Example 1.2.12. Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex set. The *convex indicator* of \mathcal{X} is the convex function defined as

$$I_{\mathcal{X}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{X} \\ +\infty & \text{otherwise,} \end{cases}$$

which is lower semicontinuous if, and only if \mathcal{X} is closed.

Example 1.2.13 (Negative entropy on the simplex). The function $h_{\text{ent}} : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ defined as

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

with convention $0 \log 0 = 0$, is a lower semicontinuous convex function. It will also be called the *entropic regularizer*.

Proposition 1.2.14. *Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function and $\mathcal{X}_0 \subset \mathbb{R}^d$ a compact set such that $\text{dom } f \cap \mathcal{X}_0 \neq \emptyset$. Then, f attains a minimum on \mathcal{X}_0 .*

1.3 Subgradients

Definition 1.3.1 (Subgradients). Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ and $x, y \in \mathbb{R}^d$. y is a *subgradient* of f at x if for all $x' \in \mathbb{R}^d$,

$$f(x') \geq f(x) + \langle y, x' - x \rangle.$$

The set of all subgradients of f at x is called the *subdifferential* of f at x and is denoted as $\partial f(x)$.

Example 1.3.2 (Absolute value). For $f : x \mapsto |x|$ defined on \mathbb{R} , the subdifferential is given by

$$\partial f(x) = \begin{cases} \{-1\} & \text{if } x < 0 \\ [0, 1] & \text{if } x = 0 \\ \{1\} & \text{if } x > 0. \end{cases}$$

Proposition 1.3.3. Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$. A point $x_* \in \mathbb{R}^d$ is a global minimizer of f if, and only if $0 \in \partial f(x_*)$.

Proof. x_* being a global minimizer can be written

$$\forall x \in \mathbb{R}^d, \quad f(x) \geq f(x_*) + \langle 0, x - x_* \rangle,$$

in other words $0 \in \partial f(x_*)$. □

Proposition 1.3.4. Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a convex function and $x \in \text{int dom } f$. f is differentiable in x if, and only if, $\partial f(x)$ is a singleton. When this is the case, $\partial f(x) = \{\nabla f(x)\}$.

Remark 1.3.5. Even in the case of a point in the domain of a convex function, the subdifferential may be empty. Consider for instance $f : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ defined as:

$$f(x) = \begin{cases} -\sqrt{x} & \text{if } x \geq 0 \\ +\infty & \text{if } x < 0, \end{cases}$$

which is a proper lower semicontinuous convex function. 0 belongs to the domain of f and yet, $\partial f(0) = \emptyset$.

Proposition 1.3.6 (see e.g. Theorem 23.4 in [Roc70]). Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper convex function. If $x \notin \text{dom } f(x)$, then $\partial f(x) = \emptyset$ and if $x \in \text{int dom } f$, then $\partial f(x) \neq \emptyset$.

Proposition 1.3.7. Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper convex function, $\|\cdot\|$ a norm on \mathbb{R}^d , and $L > 0$. Then, f is L -Lipschitz in $\text{int dom } f$ with respect to $\|\cdot\|$ if, and only if:

$$\forall x \in \text{int dom } f, \quad \forall y \in \partial f(x), \quad \|y\|_* \leq L.$$

1.4 Legendre–Fenchel transform

Definition 1.4.1. Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper function. The *Legendre–Fenchel transform* (or *convex conjugate*) of f is a function $f^* : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ defined as

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \{ \langle y, x \rangle - f(x) \}, \quad y \in \mathbb{R}^d.$$

Remark 1.4.2. In the above definition, for a given $y \in \mathbb{R}^d$, because f is assumed proper, quantity $\langle y, x \rangle - f(x)$ is not $-\infty$ for at least some point $x \in \mathbb{R}^d$, and therefore, the supremum is indeed a value in $\mathbb{R} \cup \{+\infty\}$.

Remark 1.4.3 (Fenchel’s inequality). It follows from the above definition that for all $x, y \in \mathbb{R}^d$, $\langle y, x \rangle \leq f(x) + f^*(y)$.

The above definition is somewhat abstract, and it may be insightful to informally examine a simple example in dimension 1 by decomposing the transformation into simpler steps. Consider $f(x) = \frac{a}{2}(x - b)^2$ for some $a, b \neq 0$, which is a differentiable convex function with finite values on \mathbb{R} . Its derivative is given by $f'(x) = a(x - b)$, which is (strictly) increasing (as would be the case as soon as f is differentiable and strictly convex), and is a bijection from its domain to its range (from \mathbb{R} to \mathbb{R} in this case). Then, the inverse $(f')^{-1}$ is also an increasing function: $(f')^{-1}(y) = y/a + b$. We then consider the following primitive function:

$$f^*(y) = \int_0^y (f')^{-1} - \min_{x \in \mathbb{R}} f(x), \quad y \in \mathbb{R},$$

which can be proved to be an alternative definition of the Legendre–Fenchel transform in this special case (although the proof is somewhat involved). As the primitive of an increasing function, f^* is also convex. Because of the inverse relation between the derivatives, this transformation $f \mapsto f^*$ is involutorial up a constant: $f^{**} = f + a$, ($a \in \mathbb{R}$). Then, it can be proved that $a = 0$, which can also be noticed graphically. An intuition that appears in this example is that the more f is curved, the less f^* is so, and vice-versa. The derivatives being inverses of each other can be interpreted as follows: f has slope y at point x if, and only if, f^* has slope x at point y . In higher dimension, this builds on the usual duality between points and hyperplanes. The Legendre–Fenchel transform indeed generalizes the above to higher dimensions, and for nondifferentiable functions. The neatest properties of e.g. Propositions 1.4.5 and 1.4.6 below are obtained for the class of proper lower semicontinuous convex functions. The derivative, which is a *function*, is then replaced by the subdifferential, which is a *correspondence*.

Proposition 1.4.4. *If $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ is a proper function, f^* is lower semicontinuous and convex.*

Theorem 1.4.5 (Fenchel–Moreau). *Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper function. Then, f is lower semicontinuous and convex if, and only if $f = f^{**}$. In this case, f^* is proper.*

Proposition 1.4.6. *Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous convex function and $x, y \in \mathbb{R}^d$. The following statements are equivalent:*

- (i) $x \in \partial f^*(y)$,
- (ii) $y \in \partial f(x)$,
- (iii) $\langle y, x \rangle = f(x) + f^*(y)$,
- (iv) $x \in \text{Arg max}_{x' \in \mathbb{R}^d} \{\langle y, x' \rangle - f(x')\}$,
- (v) $y \in \text{Arg max}_{y' \in \mathbb{R}^d} \{\langle y', x \rangle - f^*(y')\}$.

Example 1.4.7 (Norms and squared norms). Let $\|\cdot\|$ be a norm in \mathbb{R}^d and denote B its closed unit ball. Then, I_B is then a proper lower semicontinuous convex function and $I_B^* = \|\cdot\|_*$. Therefore, the involutonal property of the Legendre–Fenchel transform is an extension of the involutonal property for dual norms. Besides, if $f : x \mapsto \frac{1}{2} \|x\|^2$, then for $y \in \mathbb{R}^d$, $f^*(y) = \frac{1}{2} \|y\|_*^2$.

1.5 Bregman divergences

We now define a large class of similarity measures in \mathbb{R}^d called Bregman divergences, which in general are not distances because they may fail to be symmetric. They contain the squared Euclidean norm and the Kullback–Leibler divergence as special cases. Bregman divergences are used as an alternative geometry to the Euclidean one when it comes to e.g. defining and analyzing iterative algorithms. We present the classical definition which involves a gradient.

Definition 1.5.1. Let $\mathcal{X} \subset \mathbb{R}^d$, $f : \mathcal{X} \rightarrow \mathbb{R}$, $x \in \text{int } \mathcal{X}$ and $x' \in \mathcal{X}$ such that f is differentiable in x . Then, the *Bregman divergence* from x to x' is defined as

$$D_f(x', x) = f(x') - f(x) - \langle \nabla f(x), x' - x \rangle.$$

Remark 1.5.2. $D_f(x, x')$ is the remainder of the first order Taylor’s expansion from x to x' , and is a measure of the curvature of f between those two points. The Bregman divergence is nonnegative as soon as f is convex. In the case of a linear function f , the Bregman divergence is zero, which corresponds to the linear function having no curvature.

The above definition which requires the differentiability at starting point x is the most common. For our purposes however, we also consider the following generalization (proposed in [JKM23]) which involves a subgradient instead of a gradient.

Definition 1.5.3 (Generalized Bregman divergences). Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a function, $x, x', y \in \mathbb{R}^d$ such that $x \in \text{dom } f$ and $y \in \partial f(x)$. The *Bregman divergence* from x to x' with subgradient y is then defined as

$$D_f(x', x; y) = f(x') - f(x) - \langle y, x' - x \rangle.$$

Remark 1.5.4. The above generalized Bregman divergence may not exist even when x belongs to the domain of f , as the subdifferential may be empty. When it exists, because $x \in \text{dom } f$, it belongs to $\mathbb{R} \cup \{+\infty\}$ and is nonnegative when f is convex. When f is convex and differentiable at point x , the only subgradient at x is $\nabla f(x)$ according to Proposition 1.3.4, and the two previous definitions coincide.

Proposition 1.5.5. Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous convex function and $x, x', y, y' \in \mathbb{R}^d$ such that $y \in \partial f(x)$ and $y' \in \partial f(x')$. Then,

$$D_f(x', x; y) = D_{f^*}(y, y'; x').$$

Proof.

$$\begin{aligned} D_f(x', x; y) - D_{f^*}(y, y'; x') &= f(x') - f(x) - \langle y, x' - x \rangle \\ &\quad - f^*(y) + f^*(y') + \langle x', y - y' \rangle \\ &= \langle x', y' \rangle - \langle x, y \rangle - \langle y, x' - x \rangle + \langle x', y - y' \rangle \\ &= 0, \end{aligned}$$

where for the second equality, we applied Fenchel's identity from property (iii) in Proposition 1.4.6. \square

Example 1.5.6 (Squared Euclidean norm). Consider the squared Euclidean norm $h_2 : x \mapsto \frac{1}{2} \|x\|_2^2$, which is differentiable in \mathbb{R}^d . Then for all $x, x' \in \mathbb{R}^d$, $D_h(x', x) = \frac{1}{2} \|x' - x\|_2^2$.

1.6 Strong convexity and smoothness

We now introduce strongly convexity and smoothness which intuitively correspond to the curvature of a function being respectively bounded from below (by a positive number), and bounded from above in absolute value.

Definition 1.6.1 (Strong convexity). Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$, $\|\cdot\|$ a norm in \mathbb{R}^d and $K > 0$. f is *K-strongly convex* with respect to $\|\cdot\|$ if for all $x, x' \in \mathbb{R}^d$ and $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)x') \leq \lambda f(x) + (1 - \lambda)f(x') - \frac{K\lambda(1 - \lambda)}{2} \|x' - x\|^2.$$

Definition 1.6.2 (Smoothness). Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a differentiable function, $\|\cdot\|$ a norm in \mathbb{R}^d and $L > 0$. f is L -smooth with respect to $\|\cdot\|$ if for all $x, x' \in \mathbb{R}^d$, $|D_f(x', x)| \leq \frac{L}{2} \|x' - x\|^2$, in other words,

$$|f(x') - f(x) - \langle \nabla f(x), x' - x \rangle| \leq \frac{L}{2} \|x' - x\|^2.$$

Remark 1.6.3. If f is convex, the above definition reduces to $D_f(x', x) \leq \frac{L}{2} \|x' - x\|_2^2$ ($x, x' \in \mathbb{R}^d$).

Proposition 1.6.4 (Duality between strong convexity and smoothness). Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper, lower semicontinuous convex function, $\|\cdot\|$ a norm in \mathbb{R}^d and $K > 0$. The following statements are equivalent.

(i) f is K -strongly convex with respect to $\|\cdot\|$.

(ii) For all $x, x', y \in \mathbb{R}^d$ such that $y \in \partial f(x)$, $D_f(x', x; y) \geq \frac{K}{2} \|x' - x\|^2$, in other words

$$f(x') \geq f(x) + \langle y, x' - x \rangle + \frac{K}{2} \|x' - x\|^2.$$

(iii) For all $x, x', y, y' \in \mathbb{R}^d$ such that $y \in \partial f(x)$ and $y' \in \partial f(x')$,

$$\langle y' - y, x' - x \rangle \geq K \|x' - x\|^2.$$

(iv) For all $x, x', y, y' \in \mathbb{R}^d$ such that $y \in \partial f(x)$ and $y' \in \partial f(x')$,

$$\langle y' - y, x' - x \rangle \leq \frac{1}{K} \|y' - y\|_*^2.$$

(v) For all $x, x', y, y' \in \mathbb{R}^d$ such that $y \in \partial f(x)$ and $y' \in \partial f(x')$, $D_f(x', x; y) \leq \frac{1}{2K} \|y' - y\|_*^2$, in other words

$$f(x') \leq f(x) + \langle y, x' - x \rangle + \frac{1}{2K} \|y' - y\|_*^2.$$

(vi) f^* is differentiable on \mathbb{R}^d and $1/K$ -smooth with respect to $\|\cdot\|_*$.

Corollary 1.6.5. Let $K > 0$, $\|\cdot\|$ a norm on \mathbb{R}^d and $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ a proper lower semicontinuous function which we assume K -strongly convex with respect to $\|\cdot\|$. Then, for all $y, y' \in \mathbb{R}^d$,

$$D_{h^*}(y', y) \leq \frac{1}{2K} \|y' - y\|_*^2.$$

Proposition 1.6.6 (Second order characterization of strong convexity). Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a twice differentiable function, $\|\cdot\|$ a norm in \mathbb{R}^d and $K > 0$. Then, f is K -strongly convex with respect to $\|\cdot\|$ if, and only if,

$$\forall x \in \mathbb{R}^d, \forall u \in \mathbb{R}^d, \quad \langle u, \nabla^2 f(x) u \rangle \geq K \|u\|^2.$$

Proposition 1.6.7 (First and second order characterizations of smoothness). *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a differentiable function, $\|\cdot\|$ a norm in \mathbb{R}^d and $L > 0$.*

- (i) *f is L -smooth if, and only if, ∇f is L -Lipschitz with respect to $\|\cdot\|$.*
- (ii) *Moreover, if f is twice differentiable, f is L -smooth with respect to $\|\cdot\|$ if, and only if,*

$$\forall x \in \mathbb{R}^d, \forall u \in \mathbb{R}^d, \quad |\langle u, \nabla^2 f(x)u \rangle| \leq L \|u\|^2.$$

Corollary 1.6.8. *The squared Euclidean norm $h_2 : x \mapsto \frac{1}{2} \|x\|_2^2$ is 1-strongly convex and 1-smooth with respect to $\|\cdot\|_2$.*

Proposition 1.6.9. *For $p \in (1, 2)$, the squared ℓ_p norm $h_p : x \mapsto \frac{1}{2} \|x\|_p^2$ is $(p-1)$ -strongly convex with respect to ℓ_p .*

Proposition 1.6.10. *The negative entropy h_{ent} is 1-strongly convex with respect to ℓ_1 .*

Proposition 1.6.11. *A proper lower semicontinuous strongly convex function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ admits a unique minimizer on \mathbb{R}^d .*

Chapter 2

UMD theory

Throughout the chapter, \mathcal{X} is a nonempty closed convex set of \mathbb{R}^d .

2.1 Introduction

The goal of this chapter is to introduce a general scheme for defining a sequence of iterates $(x_t)_{t \geq 0}$ in \mathcal{X} based on another sequence $(u_t)_{t \geq 0}$ in \mathbb{R}^d , where for each $t \geq 0$, vector u_t is used in the update from x_t to x_{t+1} . General properties are then established. All algorithms and guarantees from the following chapters will be derived using this general approach, called UMD for *unified mirror descent*. To get some taste and intuition, we first examine a few simple special cases before presenting our general theory.

The simplest update is given by

$$x_{t+1} = x_t + u_t, \quad t \geq 0,$$

and is already of great interest, as it contains as special cases gradient descent (where $u_t = -\gamma_t \nabla f(x_t)$ is then a step in the opposite direction of the gradient of some objective function), as well as its stochastic counterpart (SGD). Such a sequence satisfies the following elementary result.

Proposition 2.1.1. *For all $t \geq 0$ and $x \in \mathbb{R}^d$,*

$$\langle u_t, x - x_t \rangle = \frac{1}{2} \|x - x_t\|_2^2 - \frac{1}{2} \|x - x_{t+1}\|_2^2 + \frac{1}{2} \|u_t\|_2^2.$$

Consequently, for all $T \geq 0$,

$$\sum_{t=0}^T \langle u_t, x - x_t \rangle = \frac{1}{2} \|x - x_0\|_2^2 - \frac{1}{2} \|x - x_{T+1}\|_2^2 + \frac{1}{2} \sum_{t=0}^T \|u_t\|_2^2.$$

Proof. Let $t \geq 1$. Using the definition of x_{t+1} ,

$$\|x_{t+1} - x\|_2^2 = \|x_t + u_t - x\|_2^2 = \|x_t - x\|_2^2 + 2 \langle u_t, x_t - x \rangle + \|u_t\|_2^2,$$

and the result follows. \square

For instance, the classical convergence guarantees about (stochastic) gradient descent in various settings, are consequences of the above identity. The quantity $\sum_{t=0}^T \langle u_t, x - x_t \rangle$ is called the *regret*, but the corresponding interpretation will be presented in the next chapter only. Some intuition about the above can be obtained through a continuous-time counterpart:

$$\text{if } \frac{d\tilde{x}_t}{dt} = \tilde{u}_t, \quad \text{then} \quad \frac{d}{dt} \left(\frac{1}{2} \|x - \tilde{x}_t\|_2^2 \right) = \langle \tilde{u}_t, \tilde{x}_t - x \rangle.$$

Therefore, going back to discrete-time, one can interpret the difference

$$\frac{1}{2} \|x - x_t\|_2^2 - \frac{1}{2} \|x - x_{t+1}\|_2^2$$

as a discrete-time derivative, and the term $\frac{1}{2} \|u_t\|_2^2$ —which does not appear in continuous-time—as a discretization error.

The quantity $\frac{1}{2} \|x - x_t\|_2^2$ also appears in the following alternative expression which will inspire generalizations.

Proposition 2.1.2. *For $t \geq 0$, $x_{t+1} = x_t + u_t$ if, and only if,*

$$x_{t+1} = \arg \min_{x \in \mathbb{R}^d} \left\{ -\langle u_t, x \rangle + \frac{1}{2} \|x - x_t\|_2^2 \right\}.$$

The above expression is an *incremental* point of view, in the sense that the next iterate x_{t+1} is written as a function of previous iterate x_t and vector u_t only. Another equivalent formulation is the *cumulative* one:

$$x_{t+1} = x_0 + \sum_{s=0}^t u_s,$$

where the next iterate x_{t+1} is only a function of the sum of the previous vectors u_s ($1 \leq s \leq t$) (*plus* the initial point x_0). These two points of view will yield different extensions below.

We now turn to a constrained setting where we need the iterates to all lie in a given nonempty closed convex set $\mathcal{X} \subset \mathbb{R}^d$. Then, the definition of the iterates can be adapted by adding a projection step onto \mathcal{X} with respect to the Euclidean distance. The projected gradient descent algorithm is a special case. Then, the iterates satisfy the following *regret bound*, where the comparison point x must lie in \mathcal{X} .

Proposition 2.1.3. *For $t \geq 0$, $x_{t+1} = \arg \min_{x \in \mathcal{X}} \|x_t + u_t - x\|_2^2$ if, and only if,*

$$x_{t+1} = \arg \min_{x \in \mathcal{X}} \left\{ -\langle u_t, x \rangle + \frac{1}{2} \|x - x_t\|_2^2 \right\}.$$

In that case, for all $x \in \mathcal{X}$,

$$\langle u_t, x - x_t \rangle \leq \frac{1}{2} \|x - x_t\|_2^2 - \frac{1}{2} \|x - x_{t+1}\|_2^2 + \frac{1}{2} \|u_t\|_2^2.$$

Another possibility for constraining the iterates in \mathcal{X} is the following, where the vector u_t is not added to the current iterate x_t as above, but to the point before the projection onto \mathcal{X} .

Proposition 2.1.4. *If $(x_t)_{t \geq 0}$ and $(y_t)_{t \geq 0}$ satisfy*

$$y_{t+1} = y_t + u_t \quad \text{and} \quad x_{t+1} = \arg \min_{x \in \mathcal{X}} \|y_{t+1} - x\|_2^2, \quad t \geq 0,$$

then for all $t \geq 0$,

$$\langle u_t, x - x_t \rangle = \frac{1}{2} \|x - x_t\|_2^2 - \frac{1}{2} \|x - x_{t+1}\|_2^2 + \frac{1}{2} \|u_t\|_2^2.$$

We now go back to the unconstrained case ($\mathcal{X} = \mathbb{R}^d$) and consider the following generalization.

Proposition 2.1.5. *For a positive definite matrix $A \in \mathbb{R}^{d \times d}$ and $t \geq 0$, $x_{t+1} = x_t + A^{-1}u_t$ if, and only if*

$$x_{t+1} = \arg \min_{x \in \mathbb{R}^d} \left\{ -\langle u_t, x \rangle + \frac{1}{2} \|x - x_t\|_A^2 \right\}.$$

In that case, for all $x \in \mathbb{R}^d$,

$$\langle u_t, x - x_t \rangle = \frac{1}{2} \|x - x_t\|_A^2 - \frac{1}{2} \|x - x_{t+1}\|_A^2 + \frac{1}{2} \|u_t\|_{A^{-1}}^2,$$

where $\|x\|_A = \sqrt{\langle x, Ax \rangle}$ and $\|y\|_{A^{-1}} = \sqrt{\langle y, A^{-1}y \rangle}$.

The last example considers the simplex $\mathcal{X} = \Delta_d$, and appears at first sight to be quite different from the above.

Proposition 2.1.6. *If for $t \geq 0$,*

$$x_{t+1} = \left(\frac{x_{t,i} \exp(u_{t,i})}{\sum_{j=1}^d x_{t,j} \exp(u_{t,j})} \right)_{1 \leq i \leq d},$$

then for all $x \in \Delta_d$,

$$\langle u_t, x - x_t \rangle = \text{KL}(x, x_t) - \text{KL}(x, x_{t+1}) + \log \left(\sum_{i=1}^d x_{t,i} \exp(u_{t,i}) - \langle u_t, x_t \rangle \right),$$

where $\text{KL}(x', x) = \sum_{i=1}^d x'_i \log(x'_i/x_i)$ denotes the Kullback–Leibler divergence.

2.2 Regularizers

Definition 2.2.1. A function $h : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ is an *pre-regularizer* on \mathcal{X} if it is strictly convex, lower semicontinuous, and if $\text{cl dom } h = \mathcal{X}$. Moreover, if $\text{dom } h^* = \mathbb{R}^d$, then h is said to be an *regularizer* on \mathcal{X} .

Remark 2.2.2. A regularizer is proper (because \mathcal{X} is nonempty), convex, and lower semicontinuous. In particular, Proposition 1.4.6 applies.

The following proposition gives several sufficient conditions for the condition $\text{dom } h^* = \mathbb{R}^d$ to be satisfied.

Proposition 2.2.3. *Let h be an pre-regularizer on \mathcal{X} .*

- (i) *If \mathcal{X} is compact, then h is a regularizer on \mathcal{X} .*
- (ii) *If h is differentiable on $\mathcal{D}_h := \text{int dom } h$ and $\nabla h(\mathcal{D}_h) = \mathbb{R}^d$, then h is a regularizer on \mathcal{X} .*
- (iii) *If h is strongly convex, then h is a regularizer on \mathcal{X} .*

Proof. Let $y \in \mathbb{R}^d$. For each of the three assumptions, let us prove that $h^*(y)$ is finite. This will prove that $\text{dom } h^* = \mathbb{R}^d$.

- (i) Because $\text{cl dom } h = \mathcal{X}$ by definition of a pre-regularizer, we have:

$$h^*(y) = \sup_{x \in \mathbb{R}^d} \{\langle y, x \rangle - h(x)\} = \sup_{x \in \mathcal{X}} \{\langle y, x \rangle - h(x)\}.$$

Besides, the function $x \mapsto \langle y, x \rangle - h(x)$ is upper semicontinuous and therefore, according to Proposition 1.2.14, attains a maximum on \mathcal{X} because \mathcal{X} is assumed to be compact. Therefore $h^*(y) < +\infty$.

- (ii) Because $\nabla h(\mathcal{D}_h) = \mathbb{R}^d$ by assumption, there exists $x \in \mathcal{D}_h$ such that $\nabla h(x) = y$. Then, by Proposition 1.4.6, $h^*(y) = \langle y, x \rangle - h(x) < +\infty$.
- (iii) The function $x \mapsto \langle y, x \rangle - h(x)$ is the opposite of a strongly convex lower semicontinuous function on \mathbb{R}^d and therefore admits a maximum by Proposition 1.6.11. Therefore, $h^*(y) < +\infty$.

□

Proposition 2.2.4 (Differentiability of h^*). *Let h be a regularizer on \mathcal{X} . Then, h^* is differentiable on \mathbb{R}^d , ∇h^* takes values in $\text{dom } h \subset \mathcal{X}$, and*

$$\nabla h^*(y) = \arg \max_{x \in \mathbb{R}^d} \{\langle y, x \rangle - h(x)\}.$$

Proof. Let $y \in \mathbb{R}^d$. Because $\text{dom } h^* = \mathbb{R}^d$, the subdifferential $\partial h^*(y)$ is nonempty by Proposition 1.3.6, $\partial h^*(y)$ is the set of maximizers of function $x \mapsto \langle y, x \rangle - h(x)$, which is strictly concave. Therefore, the maximizer belongs to $\text{dom } h$, is unique, and thus h^* is differentiable at y by Proposition 1.3.4. \square

Proposition 2.2.5 (Euclidean regularizer). *The Euclidean regularizer on \mathcal{X} , defined as*

$$h_2(x) = \frac{1}{2} \|x\|_2^2 + I_{\mathcal{X}}(x), \quad x \in \mathbb{R}^d,$$

is a regularizer on \mathcal{X} and ∇h^ is the Euclidean projection onto \mathcal{X} , in other words:*

$$\nabla h_2^*(y) = \arg \min_{x \in \mathcal{X}} \|y - x\|.$$

In particular, in the unconstrained case $\mathcal{X} = \mathbb{R}^d$, $\nabla h_2^(y) = y$ for all $y \in \mathbb{R}^d$.*

Proof. \square

Proposition 2.2.6 (Entropic regularizer). *The entropic regularizer*

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

is a regularizer on Δ_d and

$$\nabla h_{\text{ent}}^*(y) = \left(\frac{\exp(y_i)}{\sum_{j=1}^d \exp(y_j)} \right)_{1 \leq i \leq d}, \quad y \in \mathbb{R}^d.$$

Proof. \square

2.3 UMD iterates

Definition 2.3.1. Let h be a regularizer on \mathcal{X} and $(u_t)_{t \geq 0}$ a sequence in \mathbb{R}^d . A sequence $((x_t, y_t))_{t \geq 0}$ in $\mathbb{R}^d \times \mathbb{R}^d$ is a sequence of *UMD iterates* associated with regularizer h and *dual increments* $(u_t)_{t \geq 0}$ if for all $t \geq 0$,

- (i) $y_t \in \partial h(x_t)$,
- (ii) $x_{t+1} = \nabla h^*(y_t + u_t)$.

Remark 2.3.2. By Proposition 1.4.6, property (ii) is equivalent to $y_t + u_t \in \partial h(x_{t+1})$.

Remark 2.3.3. For each $t \geq 0$, $x_t \in \text{dom } h \subset \mathcal{X}$, because otherwise $\partial h(x_t)$ would be empty by Proposition 1.3.6 and could not contain y_t .

Definition 2.3.4. Let h be a regularizer on \mathcal{X} and $(u_t)_{t \geq 0}$ a sequence in \mathbb{R}^d . A sequence $((x_t, y_t))_{t \geq 0}$ in $\mathbb{R}^d \times \mathbb{R}^d$ is a sequence of *strict UMD iterates* associated with h and $(u_t)_{t \geq 0}$ if for all $t \geq 0$,

$$(I) \quad y_t \in \partial h(x_t),$$

$$(II) \quad \forall x \in \mathcal{X}, \langle y_t + u_t - y_{t+1} | x - x_{t+1} \rangle \leq 0.$$

Proposition 2.3.5. Let $((x_t, y_t))_{t \geq 0}$ be a sequence of strict UMD iterates defined as above. Then for all $t \geq 0$, $x_{t+1} = \nabla h^*(y_t + u_t)$ and thus $((x_t, y_t))_{t \geq 0}$ are UMD iterates.

Proof. Let $t \geq 0$ and let us equivalently prove that $y_t + u_t \in \partial h(x_{t+1})$. Let $x \in \mathbb{R}^d$. If $x \notin \text{dom } h$,

$$+\infty = h(x) - h(x_{t+1}) \geq \langle y_t + u_t, x - x_{t+1} \rangle.$$

If $x \in \text{dom } h$, using assumption (II) and the fact that $y_{t+1} \in h(x_{t+1})$,

$$h(x) - h(x_{t+1}) \geq \langle y_{t+1}, x - x_{t+1} \rangle \geq \langle y_t + u_t, x - x_{t+1} \rangle.$$

Therefore, $y_t + u_t \in \partial h(x_{t+1})$, in other words $x_{t+1} = \nabla h^*(y_t + u_t)$. \square

Example 2.3.6 (Euclidean regularizer). Denote $\Pi_{\mathcal{X}}$ the Euclidean projection onto \mathcal{X} and consider the Euclidean regularizer on \mathcal{X} : $h = \frac{1}{2} \|\cdot\|_2^2 + I_{\mathcal{X}}$ and $x_0 \in \mathbb{R}^d$.

- If $x_{t+1} = \Pi_{\mathcal{X}}(x_t + u_t)$ for all $t \geq 0$, then $((x_t, x_t))_{t \geq 0}$ can be proved to be a sequence of strict UMD iterates.
- If $y_{t+1} = y_t + u_t$ and $x_{t+1} = \Pi_{\mathcal{X}}(y_{t+1})$ for all $t \geq 0$, then $((x_t, x_t))_{t \geq 0}$ can also be proved to be a sequence of strict UMD iterates.

Remark 2.3.7 (Non-unicity of strict UMD iterates). As already seen in the above example, an interesting character of (strict) UMD iterates is that for a given sequence $(u_t)_{t \geq 1}$ of dual increments and initial points (x_0, y_0) such that $y_0 \in \partial h(x_0)$, there may be several possible strict UMD iterates. Here is a simple and explicit example. Consider $d = 1$, $\mathcal{X} = [0, 1]$, $h(x) = \frac{1}{2}x^2 + I_{\mathcal{X}}(x)$, $(x_0, y_0) = (1, 1)$ and $u_t = (-1)^t$ for $t \geq 0$. Then, one can verify that $((1, \frac{3+(-1)^t}{2}))_{t \geq 0}$ is a strict UMD sequence, and so is $((x_t, y_t))_{t \geq 0}$ where $x_t = y_t = \frac{1+(-1)^{t+1}}{2}$ for $t \geq 1$.

Remark 2.3.8 (Existence of strict UMD iterates). As soon as regularizer h and sequence of dual increments $(u_t)_{t \geq 0}$ are given, we can see that associated strict UMD iterates always exist. Indeed, from the definition of a regularizer, $\text{dom } h^* = \mathbb{R}^d$ and thus one can choose any $y_0 \in \mathbb{R}^d$ and consider $x_0 := \nabla h^*(y_0)$, which satisfies $y_0 \in \partial h(x_0)$. Then, for $t \geq 0$, one can consider $y_{t+1} := y_t + u_t$ which indeed satisfies variational condition (II), and then define $x_{t+1} := \nabla h^*(y_{t+1})$, which ensures $y_{t+1} \in \partial h(x_{t+1})$, as required by (i).

Remark 2.3.9 (Alternative notation for strict UMD iterates). For a given regularizer h , let $\Pi_h : \mathbb{R}^d \rightrightarrows \mathcal{X} \times \mathbb{R}^d$ be a set-valued mapping defined as follows. For $y_0 \in \mathbb{R}^d$, $\Pi_h(y_0)$ is the set of couples (x, y) satisfying

$$x = \nabla h^*(y_0), \quad y \in \partial h(x), \quad \text{and} \quad \forall x' \in \mathcal{X}, \quad \langle y_0 - y, x' - x \rangle \leq 0.$$

Then, one can verify that $((x_t, y_t))_{t \geq 0}$ is a strict UMD sequence associated with h and given sequence $(u_t)_{t \geq 0}$ if, and only if, $y_0 \in \partial h(x_0)$ and

$$(x_{t+1}, y_{t+1}) \in \Pi_h(y_t + u_t), \quad t \geq 0.$$

2.4 Regret bounds

Lemma 2.4.1 (UMD lemma). *Let h be a regularizer on \mathcal{X} , $(u_t)_{t \geq 0}$ a sequence in \mathbb{R}^d , and $((x_t, y_t))_{t \geq 0}$ a sequence of UMD iterates associated with regularizer h and dual increments $(u_t)_{t \geq 0}$ and $x \in \text{dom } h$. Consider notation*

$$D_t = D_h(x, x_t; y_t), \quad D'_t = D_h(x_{t+1}, x_t; y_t), \quad D_t^* = D_{h^*}(y_t + u_t; y_t), \quad t \geq 0.$$

(i) *Then for all $t \geq 1$,*

$$\langle u_t, x - x_{t+1} \rangle = D_t - D_{t+1} - D'_t + \langle y_t + u_t - y_{t+1}, x - x_{t+1} \rangle,$$

and

$$\langle u_t, x - x_t \rangle = D_t - D_{t+1} + D_t^* + \langle y_t + u_t - y_{t+1}, x - x_{t+1} \rangle.$$

(ii) *Moreover, $((x_t, y_t))_{t \geq 0}$ are strict UMD iterates, then for all $t \geq 0$,*

$$\langle u_t, x - x_{t+1} \rangle \leq D_t - D_{t+1} - D'_t,$$

and

$$\langle u_t, x - x_t \rangle \leq D_t - D_{t+1} + D_t^*.$$

(iii) *Besides, if h is K -strongly convex with respect to some norm $\|\cdot\|$ and some $K > 0$, then for all $t \geq 0$,*

$$D_t^* \leq \frac{1}{2K} \|u_t\|_*^2.$$

Proof. The first identity from (i) can be verified by writing explicitly the difference between both sides and simplifying. The second identity follows from noticing that for all $t \geq 0$,

$$\langle u_t, x_{t+1} - x_t \rangle = D'_t + D_h(x_t, x_{t+1}; y_t + u_t) = D'_t + D_t^*,$$

where the second equality comes from Proposition 1.5.5; and adding to the first equality. Equalities in (ii) are an immediate consequence of (i). iii follows from Proposition (1.6.4). \square

2.5 Time-dependent regularizers

Definition 2.5.1. Let $(h_t)_{t \in \frac{1}{2}\mathbb{N}}$ be a sequence of regularizers and $(u_t)_{t \geq 0}$ a sequence in \mathbb{R}^d . An associated sequence $((x_t, y_t))_{t \geq 0}$ in $\mathbb{R}^d \times \mathbb{R}^d$ of UMD iterates satisfy for all $t \in \mathbb{N}$,

- (i) $y_t \in \partial h_t(x_t)$,
- (ii) $x_{t+1} = \nabla h_{t+1/2}^*(y_t + u_t)$.

Lemma 2.5.2 (UMD lemma with time-dependent regularizers). *Let $(h_t)_{t \in \frac{1}{2}\mathbb{N}}$ be a sequence of regularizers, $(u_t)_{t \geq 0}$ a sequence in \mathbb{R}^d , $((x_t, y_t))_{t \geq 1}$ associated UMD iterates, and $x \in \bigcap_{t \in \frac{1}{2}\mathbb{N}} \text{dom } h_t$. For each $t \in \mathbb{N}$, consider notation*

- $D_t = D_{h_t}(x, x_t; y_t)$,
- $D'_t = h_{t+1/2}(x_{t+1}) - h_t(x_t) - \langle y_t, x_{t+1} - x_t \rangle$,
- $D_t^* = h_{t+1/2}^*(y_t + u_t) - h_t^*(y_t) - \langle u_t, x_t \rangle$,
- For $x' \in \text{dom } h_t$ (resp. $\text{dom } h_{t+1/2}$), $\Delta h_t(x') = h_{t+1/2}(x') - h_t(x')$ (resp. $\Delta h_{t+1/2}(x') = h_{t+1}(x') - h_{t+1/2}(x')$),
- $D_{t+1/2}^\Delta = \Delta h_{t+1/2}(x) - \Delta h_{t+1/2}(x_{t+1}) - \langle y_{t+1} - y_t - u_t, x - x_{t+1} \rangle$.

(i) Then for all $t \in \mathbb{N}$,

$$\langle u_t, x - x_{t+1} \rangle = D_t - D_{t+1} - D'_t + D_{t+1/2}^\Delta + \Delta h_t(x),$$

and

$$\langle u_t, x - x_t \rangle = D_t - D_{t+1} + D_t^* + D_{t+1/2}^\Delta + \Delta h_t(x).$$

(ii) If $\Delta h_t = 0$ for a given $t \in \mathbb{N}$, then

$$\langle u_t, x - x_{t+1} \rangle = D_t - D_{t+1} - D'_t + D_{t+1/2}^\Delta,$$

and

$$\langle u_t, x - x_t \rangle = D_t - D_{t+1} + D_t^* + D_{t+1/2}^\Delta.$$

(iii) If $\Delta h_{t+1/2} = 0$ and $\langle y_{t+1} - y_t - u_t, x - x_{t+1} \rangle \geq 0$ for a given $t \in \mathbb{N}$, then

$$\langle u_t, x - x_{t+1} \rangle \leq D_t - D_{t+1} - D'_t + h_{t+1}(x) - h_t(x),$$

and

$$\langle u_t, x - x_t \rangle \leq D_t - D_{t+1} + D_t^* + h_{t+1}(x) - h_t(x).$$

(iv) If for a given $t \in \mathbb{N}$, $h_{t+1/2} \geq h_t$ and h_t is K_t -strongly convex with respect to some norm $\|\cdot\|$ and some $K_t > 0$, then

$$D_t^* \leq \frac{1}{2K_t} \|u_t\|_*^2.$$

Proof. The first equality in (i) can be proved by merely simplifying. For the second inequality, we notice that

$$\begin{aligned} \langle u_t, x_t - x_{t+1} \rangle &= D'_t + h_t(x_t) - h_{t+1/2}(x_{t+1}) - \langle y_t + u_t, x_t - x_{t+1} \rangle \\ &= D'_t + \langle y_t, x_t \rangle - h_t^*(y_t) - \langle y_t + u_t, x_{t+1} \rangle + h_{t+1/2}^*(y_t + u_t) \\ &\quad - \langle y_t + u_t, x_t - x_{t+1} \rangle \\ &= D'_t + D_t^*, \end{aligned}$$

where for the second equality, we used Fenchel's identity from Proposition 1.4.6. Adding the above to the first equality give the second equality in (i). Then, (ii) and (iii) are easy consequences.

Let us prove (iv). The assumption $h_{t+1/2} \geq h_t$ and the definition of the Legendre–Fenchel transform immediately implies $h_{t+1/2}^* \leq h_t^*$. Therefore,

$$D_t^* = h_{t+1/2}^*(y_t + u_t) - h_t^*(y_t) - \langle u_t, x_t \rangle \leq D_{h^*}(y_t + u_t, y_t) \leq \frac{1}{2K_t} \|u_t\|_*^2.$$

□

Chapter 3

Online optimization

This chapter first presents the topic of regret minimization in sequential decision problems and then makes the connection with online learning and optimization through the frameworks of online linear optimization and online convex optimization. Several important families of algorithms (dual averaging, mirror descent, follow the regularized leader) are introduced and analyzed using UMD theory from Chapter 2.

Throughout the chapter, \mathcal{X} is a nonempty closed convex set, and $\Pi_{\mathcal{X}}$ denotes the Euclidean projection onto \mathcal{X} . For a function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$, we denote $\mathcal{D}_f = \text{int dom } f$.

3.1 Introduction to regret minimization

Let us first consider a simple sequential decision problem where the Decision Maker chooses its actions in the finite set $\{1, \dots, d\}$, possibly at random. At step $t \geq 0$,

- the Decision Maker chooses $x_t \in \Delta_d$,
- Nature chooses and reveals *payoff vector* $u_t \in [0, 1]^d$,
- i_t is a random element in $\{1, \dots, d\}$ drawn according to distribution x_t and revealed to the Decision Maker,
- the Decision Maker obtains payoff u_{t,i_t} .

The choice of x_t by the Decision Maker may depend on all past information known to him, meaning $(x_0, u_0, i_0, \dots, x_{t-1}, u_{t-1}, i_{t-1})$. Similarly, the choice of u_t by Nature may depend on all past information including x_t : $(x_0, u_0, i_0, \dots, x_{t-1}, u_{t-1}, i_{t-1}, x_t)$.

In a restrictive variant of this problem, called *multi-armed bandit* and which will be considered later, the decision maker only observes the actually obtained payoff u_{t,i_t} , and not the whole payoff vector u_t .

The Decision Maker wishes to maximize its cumulative payoff $\sum_{t=0}^T u_{t,i_t}$. More specifically, we aim at constructing decision rules for the Decision Maker (which we will simply call *algorithms*) that offer some *worst-case guarantee* on the cumulative payoff which holds *for all possible sequence* $(u_t)_{t \geq 0}$ chosen by Nature. Therefore, the guarantee must be relative to the sequence of payoff vectors $(u_t)_{t \geq 0}$. One possible type of guarantee is an upper bound on the *regret*, defined as

$$\max_{1 \leq i \leq d} \sum_{t=0}^T u_{t,i} - \sum_{t=0}^T u_{t,i_t},$$

which compares the actual cumulative payoff with the best cumulative payoff that would have been obtained¹ by constantly choosing a given element $i \in \{1, \dots, d\}$, meaning $i_t = i$ for all $t \geq 0$.

For a given sequence of payoff vectors $(u_t)_{t \geq 0}$, the above regret is a random variable (because for each $t \geq 0$, i_t is a random variable), and we are interested in analyzing its expectation. Using the law of total expectation, we can write

$$\begin{aligned} \mathbb{E} \left[\max_{1 \leq i \leq d} \sum_{t=0}^T u_{t,i} - \sum_{t=0}^T u_{t,i_t} \right] &= \mathbb{E} \left[\max_{1 \leq i \leq d} \sum_{t=0}^T u_{t,i} - \sum_{t=0}^T \mathbb{E}[u_{t,i_t} | x_t] \right] \\ &= \mathbb{E} \left[\max_{1 \leq i \leq d} \sum_{t=0}^T u_{t,i} - \sum_{t=0}^T \langle u_t, x_t \rangle \right]. \end{aligned} \quad (3.1)$$

Besides, because the maximum of linear function on a convex compact set is attained at its edges,

$$\max_{1 \leq i \leq d} \sum_{t=0}^T u_{t,i} = \max_{1 \leq i \leq d} \left\langle \sum_{t=0}^T u_t, e_i \right\rangle = \max_{x \in \Delta_d} \sum_{t=0}^T \langle u_t, x_t \rangle.$$

Therefore, upper bounds on the quantity

$$\max_{x \in \Delta_d} \sum_{t=0}^T \langle u_t, x \rangle - \sum_{t=0}^T \langle u_t, x_t \rangle = \max_{x \in \Delta_d} \sum_{t=0}^T \langle u_t, x - x_t \rangle, \quad (3.2)$$

which we will also call the *regret*, will yield the same bound on the expected regret from (3.1).

A common assumption is that the sequence of payoff vectors $(u_t)_{t \geq 0}$ is bounded. In that case, we will see that there exists algorithms that

¹This interpretation holds only if we assume that the choice of the payoff vectors $(u_t)_{t \geq 0}$ by Nature does not depend on the choices of the Decision Maker. Nevertheless, all guarantees on the regret will still hold, even in the case where Nature does react to the choices of the Decision Maker; only the interpretation will not stand.

guarantee that above (cumulative) regret grows at most in \sqrt{T} , in other words, the *average* regret is bounded from above by a quantity that vanishes as $1/\sqrt{T}$. This has the following interpretation called *prediction with expert advice*: there are d experts, and at each step $t \geq 0$, the Decision Maker has to choose one of the experts and follow his advice, and obtains the corresponding payoff. Then, as will be proved below, there exists algorithms such that the Decision Maker is guaranteed to perform *as well as the best expert* (asymptotically and in average).

Another important question will be the optimal dependence of the regret bound in d , in the case e.g. where the payoff vectors are assumed to be in $[0, 1]^d$.

Example 3.1.1 (Follow the leader). The simplest algorithm we may think of is called *follow the leader*, and picks the decision $i \in \{1, \dots, d\}$ which would have yielded the highest cumulative payoff on the previous steps:

$$i_{t+1} = \arg \max_{1 \leq i \leq d} \sum_{s=1}^t u_{s,i}, \quad t \geq 0,$$

which can be equivalently written

$$x_{t+1} = \arg \max_{x \in \Delta_d} \left\langle \sum_{s=1}^t u_s, x \right\rangle, \quad t \geq 0. \quad (3.3)$$

Unfortunately, this algorithm is too simple and it is easy to find a bounded sequence for which the regret grows linearly, e.g. for $d = 2$,

$$u_0 = \begin{pmatrix} 1/2 \\ 0 \end{pmatrix}, \quad u_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad u_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad u_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \dots$$

yields for all $T \geq 1$,

$$\max_{1 \leq i \leq d} \sum_{t=0}^T u_{t,i} - \sum_{t=0}^T u_{t,i_t} \geq \frac{T-1}{2} - \frac{1}{2}.$$

Intuitively, the issue with the follow the leader algorithm is that it follows previous data too closely, so that there exists a payoff vector for the next step for which the decision of the algorithm is the worst. This is a kind of *overfitting*. To address the issue, one possible approach is to *regularize* the quantity that is maximized in (3.3), which will lead to the *dual averaging* and *follow the regularized leader* algorithms below.

We now make the connection with online learning and optimization by presenting two successive extensions of the above framework.

Online linear optimization The quantity (3.2) inspires the following natural extension. Let $\mathcal{X} \subset \mathbb{R}^d$ be a nonempty closed convex set, and $\mathcal{U} \subset \mathbb{R}^d$ any nonempty set. At each step $t \geq 0$,

- the Decision Maker chooses $x_t \in \mathcal{X}$,
- Nature chooses and reveals $u_t \in \mathcal{U}$,
- the Decision Maker obtains payoff $\langle u_t, x_t \rangle$.

In the case where \mathcal{X} is bounded, the natural definition of the regret is

$$\max_{x \in \mathcal{X}} \sum_{t=0}^T \langle u_t, x - x_t \rangle.$$

When \mathcal{X} is unbounded, it will be possible to guarantee upper bounds on

$$\sum_{t=0}^T \langle u_t, x - x_t \rangle,$$

that depend on the *comparison point* $x \in \mathcal{X}$. In the case where the set \mathcal{U} of payoff vectors is bounded, typical regret bounds also grow as \sqrt{T} , as will be established below in Sections 3.2 and 3.3.

Online convex optimization We now further generalize by considering convex loss functions, which is motivated by e.g. online learning. Let $\mathcal{X} \subset \mathbb{R}^d$ be a nonempty closed convex set and \mathcal{L} a nonempty set of convex function $\ell : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ such that $\mathcal{X} \subset \text{dom } \ell$. At each step $t \geq 0$,

- the Decision Maker chooses $x_t \in \mathcal{X}$,
- Nature chooses and reveals $\ell_t \in \mathcal{L}$,
- the Decision Maker incurs loss $\ell_t(x_t)$.

The corresponding definition for the regret is

$$\max_{x \in \mathcal{X}} \sum_{t=0}^T (\ell_t(x_t) - \ell_t(x)).$$

In the case where the loss functions have some *curvature*, the latter can be leveraged to achieve regret bounds that grow strictly slower than \sqrt{T} .

The case where the loss function is constant boils down to convex optimization.

Example 3.1.2 (Online linear regression). By slightly modifying the above problem, the Decision Maker can observe some contextual information z_t before choosing $x_t \in \mathcal{X}$. Online versions of supervised learning problem can then be considered with Nature choosing loss function of the form e.g.

$$\ell(x) = (\langle z, x \rangle - y)^2, \quad z \in \mathbb{R}^d, y \in \mathbb{R}.$$

3.2 Dual averaging

We define and analyze the dual averaging family of algorithm which can be seen as a *regularized* version of the follow the leader algorithm from Example (3.3). It is a special case of UMD iterates where for all $t \geq 0$, the regularizers satisfy $h_{t+1/2} = h_{t+1}$ and where the next dual point y_{t+1} is uniquely defined as $y_{t+1} = y_t + u_t$.

We obtain in Proposition 3.2.5 below a regret bound in the context of online linear optimization, which will be applied and transposed to numerous problems.

Definition 3.2.1. Let $(h_t)_{t \geq 0}$ be a sequence of regularizers on \mathcal{X} , $(u_t)_{t \geq 0}$ a sequence in \mathbb{R}^d , and $y_0 \in \mathbb{R}^d$. The associated sequence $(x_t)_{t \geq 0}$ of *dual averaging* (DA) iterates is defined for $t \geq 0$ as

$$x_t = \nabla h_t^*(y_t) \quad \text{and} \quad y_{t+1} = y_t + u_t.$$

Remark 3.2.2. The above definition can be equivalently written

$$\begin{aligned} x_{t+1} &= \arg \max_{x \in \mathbb{R}^d} \left\{ \left\langle y_0 + \sum_{s=0}^t u_s, x \right\rangle - h_t(x) \right\} \\ &= \arg \max_{x \in \mathcal{X}} \{ \langle u_t, x \rangle - D_h(x, x_t; y_t) \}. \end{aligned}$$

In the case $y_0 = 0$, the above second expression gives the following interpretation: x_{t+1} is the maximizer not of the past cumulative payoff function $x \mapsto \langle \sum_{s=0}^t u_s, x \rangle$ as in the *follow the leader* algorithm (see Section 3.1), but of a *regularized* version. For this reason, this algorithm is sometimes called *follow the regularized leader*. We will use this name below to designate a more general algorithm in the context of regret minimization with convex losses.

Example 3.2.3 (Lazy Mirror Descent). Let $y_0 \in \mathbb{R}^d$ and denote $\Pi_{\mathcal{X}}$ the Euclidean projection onto \mathcal{X} . For a given sequence $(u_t)_{t \geq 0}$, $x_0 = \Pi_{\mathcal{X}}(y_0)$ and

$$x_{t+1} = \Pi_{\mathcal{X}}(x_t + u_t), \quad t \geq 0$$

corresponds to dual averaging iterates with constant Euclidean regularizer $h = \frac{1}{2} \|\cdot\|_2^2 + I_{\mathcal{X}}$ (see Proposition 2.2.5 and Example 2.3.6).

Example 3.2.4 (Exponential weights algorithm). For a given sequence $(u_t)_{t \geq 0}$, consider

$$x_{t+1} = \left(\frac{\exp(\sum_{s=0}^t u_{s,i})}{\sum_{j=1}^d \exp(\sum_{s=0}^t u_{s,j})} \right)_{1 \leq j \leq d}.$$

These correspond to dual averaging iterates with constant entropic regularizer h_{ent} (see Proposition 2.2.6)

The following statement gives a regret bound for online linear optimization, which will also be applied and transposed to numerous problems.

Proposition 3.2.5. *Let $(x_t)_{t \geq 0}$ and $(y_t)_{t \geq 0}$ be defined as in Definition 3.2.1, and $x \in \bigcap_{t \geq 0} \text{dom } h_t$. Then,*

(i) *$((x_t, y_t))_{t \geq 0}$ is a sequence of UMD iterates associated with regularizers $(h_t)_{t \in \frac{1}{2}\mathbb{N}}$ and dual increments $(u_t)_{t \geq 0}$, where $h_{t+1/2} := h_{t+1}$ for all $t \geq 0$;*

(ii) *for all $T \geq 0$,*

$$\sum_{t=0}^T \langle u_t, x - x_t \rangle \leq D_0 + (h_{T+1}(x) - h_0(x)) + \sum_{t=0}^T D_t^*,$$

where

$$\begin{aligned} D_0 &= h_0(x) - h_0(x_0) - \langle y_0, x - x_0 \rangle \\ D_t^* &= h_{t+1}^*(y_t + u_t) - h_t^*(y_t) - \langle u_t, x_t \rangle, \quad t \geq 0; \end{aligned}$$

(iii) *if for $t \geq 0$, $h_{t+1} \geq h_t$ and h_t is K_t -strongly convex for some norm $\|\cdot\|$, then D_t^* is bounded as*

$$D_t^* \leq \frac{1}{2K_t} \|u_t\|_*^2.$$

Proof. (i) holds because the conditions from Definition 2.3.1 are trivially satisfied. (ii) and (iii) easily follow from Lemma 2.5.2. \square

Dual averaging with nonincreasing parameter Let h be a regularizer on \mathcal{X} , $(u_t)_{t \geq 0}$ a sequence in \mathbb{R}^d , $y_0 = 0$ and $(\eta_t)_{t \geq 0}$ a positive nonincreasing sequence. Consider iterates $(x_t)_{t \geq 0}$ defined for $t \geq 0$ as

$$x_t = \nabla h^*(\eta_t y_t) \quad \text{and} \quad y_{t+1} = y_t + u_t. \quad (3.4)$$

Then, $((x_t, y_t))_{t \geq 0}$ are UMD iterates associated with dual increments $(u_t)_{t \geq 0}$ and regularizers $(h_t)_{t \in \frac{1}{2}\mathbb{N}}$ where

$$h_t(x) = \frac{h(x) - \min h}{\eta_t}, \quad t \geq 0, \quad (3.5)$$

and $h_{t+1/2} = h_{t+1}$ for $t \geq 1$. The definition of UMD iterates is invariant when constants are added to regularizers, so it would be equivalent to simply consider $h_t = h/\eta_t$ but the above regularizers have the advantage of ensuring $h_{t+1} \geq h_t$ and therefore makes the analysis simpler.

Proposition 3.2.6. *Consider the iterates defined in (3.4).*

(i) For all $T \geq 0$ and $x \in \text{dom } h$,

$$\sum_{t=0}^T \langle u_t, x - x_t \rangle \leq \frac{h(x) - \min h}{\eta_T} + \sum_{t=0}^T \eta_t D_{h^*}(y_t + u_t, y_t).$$

(ii) Moreover, if h is K -strongly convex for some norm $\|\cdot\|$, for all $T \geq 0$,

$$\sum_{t=0}^T \langle u_t, x - x_t \rangle \leq \frac{h(x) - \min h}{\eta_T} + \frac{1}{2K} \sum_{t=0}^T \eta_t \|u_t\|_*^2.$$

(iii) Moreover, if there exists $L > 0$ such that $\|u_t\|_* \leq L$, then the choice $\eta_t = \eta \sqrt{2K}/(L\sqrt{t+1})$ for some $\eta > 0$ yields

$$\sum_{t=0}^T \langle u_t, x - x_t \rangle \leq \left(\frac{h(x) - \min h}{\eta} + \eta \right) L \sqrt{\frac{T+1}{2K}}.$$

(iv) Moreover, if $\sup_{x \in \mathcal{X}} h(x) < +\infty$, then $\eta = \sqrt{\max_{\mathcal{X}} h - \min h}$ yields

$$\max_{x \in \mathcal{X}} \sum_{t=0}^T \langle u_t, x - x_t \rangle \leq L \sqrt{\frac{2(\max_{\mathcal{X}} h - \min h)(T+1)}{K}}.$$

Proof. (i) Applying Proposition 3.2.5, using $y_0 = 0$, and simplifying gives

$$\sum_{t=0}^T \langle u_t, x - x_t \rangle \leq \frac{h(x) - \min h}{\eta_{T+1}} + \sum_{t=0}^T (h_{t+1}^*(y_t + u_t) - h_t(y_t) - \langle u_t, x_t \rangle).$$

For a given $T \geq 0$, iterates x_0, \dots, x_T do not depend on η_{T+1} , therefore, Proposition 3.2.5 can be applied for $\eta_{T+1} = \eta_T$. Besides, because $h_{t+1} \geq h_t$ by definition in (3.5), $h_{t+1}^*(y_t + u_t) \leq h_t^*(y_t + u_t)$, hence the result.

(ii) follows from Proposition 1.6.4, (iii) and (iv) are immediate consequences. \square

Remark 3.2.7. According to the above statements (iii) and (iv) give that the average regret is minimized at speed $1/\sqrt{T}$, for all $T \geq 0$, without prior knowledge of the latter: the algorithm and the guarantee are said to be *horizon-free*. In the case of a bounded domain such that $\sup_{\mathcal{X}} h < +\infty$, the bound (iv) does not depend on the comparison point x . When \mathcal{X} is unbounded, or when $\sup_{\mathcal{X}} h = +\infty$, the bound (iii) depend on the comparison point x , but still guarantees a horizon-free average regret bound of order $1/\sqrt{T}$, which is an important feature of dual averaging.

3.3 Online mirror descent

We define and analyze the the online mirror descent family of algorithms, which is an extension of the online projected gradient descent:

$$x_{t+1} = \Pi_{\mathcal{X}}(x_t + u_t) = \arg \max_{x \in \mathcal{X}} \left\{ \langle u_t, x \rangle - \frac{1}{2} \|x - x_t\|_2^2 \right\},$$

where the above Euclidean distance is replaced by the Bregman divergence associated with a differentiable function H . This yields a special case of UMD iterates where for all $t \geq 0$, $h_t = h_{t+1/2}$ and where dual point y_t is uniquely defined as $y_t = \nabla H_t(x_t)$.

Proposition 3.3.13 below gives a regret bound in the context of online linear optimization, and will also be applied and transposed to various problems.

Definition 3.3.1. Let $H : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$. Denote $\mathcal{D}_H := \text{int dom } H$. H is a *mirror map* compatible with \mathcal{X} if

- (i) H is lower semi-continuous, strictly convex, and differentiable on \mathcal{D}_H ,
- (ii) the gradient of H takes all possible values, i.e. $\nabla H(\mathcal{D}_H) = \mathbb{R}^d$,
- (iii) $\mathcal{X} \subset \text{cl } \mathcal{D}_H$,
- (iv) $\mathcal{X} \cap \mathcal{D}_H \neq \emptyset$.

Example 3.3.2 (ℓ^p norms). For $p > 1$, $x \mapsto \frac{1}{2} \|x\|_p^2$ is a mirror map compatible with all nonempty closed convex sets.

Example 3.3.3 (Generalized negative entropy). The generalized negative entropy on the closed positive orthant, defined as

$$H(x) = \begin{cases} \sum_{i=1}^d x_i \log x_i & \text{if } x \in \mathbb{R}_+^d \\ +\infty & \text{otherwise,} \end{cases}$$

with convention $0 \cdot \log 0 = 0$, is a mirror map compatible with all nonempty closed convex subsets of \mathbb{R}_+^d (e.g. the simplex).

Example 3.3.4 (Log barrier). The log barrier on the open positive orthant, defined as

$$H(x) = \begin{cases} -\sum_{i=1}^d \log x_i & \text{if } x \in (\mathbb{R}_+^*)^d, \\ +\infty & \text{otherwise,} \end{cases}$$

is a mirror map compatible with all nonempty closed convex subsets of \mathbb{R}_+^d (e.g. the simplex), although its domain is the open positive orthant $(\mathbb{R}_+^*)^d$.

Proposition 3.3.5. Let H be a mirror map compatible with \mathcal{X} , H^* its Legendre–Fenchel transform. Then,

- (i) $\text{dom } H^* = \mathbb{R}^d$,
- (ii) H^* is differentiable on \mathbb{R}^d ,
- (iii) $\nabla H^*(\mathbb{R}^d) = \mathcal{D}_H$,
- (iv) For $x \in \mathcal{D}_H$ and $y \in \mathbb{R}^d$, $\nabla F^*(\nabla F(x)) = x$ and $\nabla F(\nabla F^*(y)) = y$.

Proof. Let $y_t \in \mathbb{R}^n$. By property (ii) from Definition 3.3.1, there exists $x_0 \in \mathcal{D}_H$ such that $\nabla H(x_0) = y$. Therefore, function $\varphi_y : x \mapsto \langle y|x \rangle - H(x)$ is differentiable at x_0 and $\nabla \varphi_y(x_0) = 0$. Moreover, φ_y is strictly concave as a consequence of property (i) from Definition 3.3.1. Therefore, x_0 is the unique maximizer of φ_y and:

$$H^*(y) = \max_{x \in \mathbb{R}^n} \{\langle y|x \rangle - H(x)\} < +\infty,$$

which proves property (i).

Besides, we have

$$x_0 \in \partial H^*(y) \iff y = \nabla H(x_0) \iff x_0 \text{ minimizer of } \phi_y, \quad (3.6)$$

where the first equivalence comes from Proposition 1.4.6. Point x_0 being the unique maximizer of φ_y , we have that $\partial H^*(y)$ is a singleton. In other words, H^* is differentiable in y and

$$\nabla H^*(y) = x_0 \in \mathcal{D}_H. \quad (3.7)$$

First, the above (3.7) proves property (ii). Second, this equality combined with the equality from (3.6) gives the second identity from property (iv). Third, this proves that $\nabla H^*(\mathbb{R}^n) \subset \mathcal{D}_H$.

It remains to prove the reverse inclusion to get property (iii). Let $x \in \mathcal{D}_H$. By property (i) from Definition 3.3.1, H is differentiable in x . Consider

$$y := \nabla H(x), \quad (3.8)$$

and all the above holds with this special point y . In particular, $x_0 = x$ by uniqueness of x_0 . Therefore (3.7) gives

$$\nabla H^*(y) = x, \quad (3.9)$$

and this proves $\nabla H^*(\mathbb{R}^n) \supset \mathcal{D}_H$ and thus property (iii). Combining (3.8) and (3.9) gives the first identity from property (iv). \square

Proposition 3.3.6 (OMD iteration). *Let H be a mirror map compatible with \mathcal{X} , $x \in \mathcal{D}_H$, $u \in \mathbb{R}^d$ and consider*

$$x' = \arg \max_{x'' \in \mathcal{X}} \{\langle u, x'' \rangle - D_H(x'', x)\}.$$

Then, $x' \in \mathcal{X} \cap \mathcal{D}_H$ and can also be written

$$\begin{aligned} x' &= \arg \max_{x'' \in \mathcal{X}} \{ \langle \nabla H(x) + u, x'' \rangle - H(x'') \} \\ &= \arg \min_{x'' \in \mathcal{X}} D_H(x'', \nabla H^*(\nabla H(x) + u)). \end{aligned}$$

Proof. The proof that x' is well-defined and belongs to \mathcal{D}_H is given in [JKM23]. Besides,

$$\begin{aligned} x' &= \arg \max_{x'' \in \mathcal{X}} \{ \langle u, x'' \rangle - D_H(x'', x) \} \\ &= \arg \max_{x'' \in \mathcal{X}} \{ \langle u, x'' \rangle - H(x'') + \langle \nabla H(x), x'' - x \rangle \} \\ &= \arg \max_{x'' \in \mathcal{X}} \{ \langle \nabla H(x) + u, x'' \rangle - H(x'') \} \\ &= \arg \max_{x'' \in \mathcal{X}} \{ -H(x'') + H(\nabla H^*(\nabla H(x) + u)) + \langle \nabla H(x) + u, x'' - x \rangle \} \\ &= \arg \min_{x'' \in \mathcal{X}} D_H(x'', \nabla H^*(\nabla H(x) + u)). \end{aligned}$$

□

Definition 3.3.7. A sequence $(H_t)_{t \geq 0}$ of mirror maps has *nondecreasing domains* if $\text{dom } H_t \subset \text{dom } H_{t+1}$ for all $t \geq 0$.

Definition 3.3.8. Let $(H_t)_{t \geq 0}$ be a sequence of mirror maps compatible with \mathcal{X} with nondecreasing domains and $(u_t)_{t \geq 0}$ a sequence in \mathbb{R}^d . A sequence $(x_t)_{t \geq 0}$ in \mathbb{R}^d is an sequence of *online mirror descent (OMD)* iterates on \mathcal{X} associated with mirror maps $(H_t)_{t \geq 0}$ and *dual increments* $(u_t)_{t \geq 0}$ if $x_0 \in \mathcal{X} \cap \text{int dom } H_0$ and

$$x_{t+1} = \arg \max_{x \in \mathcal{X}} \{ \langle u_t, x \rangle - D_{H_t}(x, x_t) \}, \quad t \geq 0.$$

Remark 3.3.9. The above iterates are well-defined. Indeed, we first note that $x \in \mathcal{D}_{H_0}$, and then by induction, as soon as $x_t \in \mathcal{D}_{H_t}$, Proposition 3.3.6 ensures that x_{t+1} is well-defined and unique, and that it belongs to \mathcal{D}_{H_t} . Because the mirror maps have nondecreasing domains, x_{t+1} also belongs to $\mathcal{D}_{H_{t+1}}$.

Example 3.3.10 (Online (projected) gradient descent). Denote $\Pi_{\mathcal{X}}$ the Euclidean projection onto \mathcal{X} . Let $y_0 \in \mathbb{R}^d$ and a sequence $(u_t)_{t \geq 0}$ in \mathbb{R}^d . Then, $x_0 = \Pi_{\mathcal{X}}(y_0)$ and

$$x_{t+1} = \Pi_{\mathcal{X}}(x_t + u_t), \quad t \geq 0,$$

are OMD iterates on \mathcal{X} associated with constant Euclidean mirror map $H = \frac{1}{2} \|\cdot\|_2^2$.

Example 3.3.11 (Exponential weights algorithm). The exponential weights algorithm from Example 3.2.4 gives OMD iterates on the simplex Δ_d associated with generalized entropic mirror map from Example 3.3.3.

The following proposition indicates the regularizers to consider for making the connection with the definition of UMD iterates.

Proposition 3.3.12. *Let H be a mirror map compatible with \mathcal{X} . Then $h = H + I_{\mathcal{X}}$ is a regularizer on \mathcal{X} and for all $x \in \mathcal{D}_H$, $\nabla H(x) \in \partial h(x)$.*

Proof. See [JKM23]. □

The following statement gives a regret bound for online linear optimization, and will also be applied and transposed to numerous problems.

Proposition 3.3.13. *Let $(H_t)_{t \geq 0}$ be a sequence of mirror maps compatible with \mathcal{X} with nondecreasing domains, $(u_t)_{t \geq 0}$ a sequence in \mathbb{R}^d , $(x_t)_{t \geq 0}$ associated OMD iterates in \mathcal{X} and $x \in \mathcal{X} \cap \text{dom } H_0$. Then,*

- (i) *$((x_t, \nabla H_t(x_t)))_{t \geq 0}$ is a sequence of UMD iterates associated with regularizers $(h_t)_{t \geq 0}$ defined as*

$$h_t = h_{t+1/2} = H_t + I_{\mathcal{X}}, \quad t \in \mathbb{N},$$

and dual increments $(u_t)_{t \geq 0}$;

- (ii) *for all $T \geq 0$,*

$$\sum_{t=0}^T \langle u_t, x - x_t \rangle \leq D_{H_0}(x, x_0) - D_{H_{T+1}}(x, x_{T+1}) + \sum_{t=0}^T (\tilde{D}_{t+1/2}^{\Delta} + \tilde{D}_t^*),$$

where

$$\begin{aligned} \tilde{D}_{t+1/2}^{\Delta} &= D_{H_{t+1}-H_t}(x, x_{t+1}), \\ \tilde{D}_t^* &= D_{H_t^*}(\nabla H_t(x_t) + u_t, \nabla H_t(x_t)); \end{aligned}$$

- (iii) *if for $t \geq 0$, H_t is K_t -strongly convex for some norm $\|\cdot\|$, then \tilde{D}_t^* is bounded as*

$$\tilde{D}_t^* \leq \frac{1}{2K_t} \|u_t\|_*^2.$$

Proof. (i) $(h_t)_{t \in \frac{1}{2}\mathbb{N}}$ are indeed regularizers thanks to Proposition 3.3.12. Let us prove that $((x_t, \nabla H_t(x_t)))_{t \geq 0}$ satisfy the definition of UMD iterates. Let $t \geq 0$ and $x' \in \mathbb{R}^d$. By definition of h_t , $h_t \geq H_t$. Moreover, $h_t(x_t) = H(x_t)$ because $x_t \in \mathcal{D}_{H_t}$ by Remark 3.3.9. Then,

$$h_t(x') - h_t(x_t) \geq H_t(x') - H_t(x_t) \geq \langle \nabla H_t(x_t), x - x_t \rangle,$$

which means $\nabla F_t(x_t) \in \partial h_t(x_t)$, and the first condition from Definition 2.3.1 is satisfied. Besides, using Proposition 3.3.6 and the definition of $h_{t+1/2}$,

$$\begin{aligned} x_{t+1} &= \arg \max_{x \in \mathcal{X}} \{ \langle \nabla H_t(x_t) + u_t, x \rangle - H_t(x) \} \\ &= \arg \max_{x \in \mathbb{R}^d} \{ \langle \nabla H_t(x_t) + u_t, x \rangle - h_{t+1/2}(x) \} = \nabla h_{t+1/2}^*(\nabla H_t(x_t) + u_t), \end{aligned}$$

which proves the second condition, and $((x_t, \nabla H_t(x_t)))_{t \geq 0}$ is indeed a sequence of UMD iterates associated with $(H_t)_{t \geq 0}$ and $(u_t)_{t \geq 0}$.

(ii) For all $t \in \frac{1}{2}\mathbb{N}$, the definition of h_t imply that $\text{dom } h_t = \mathcal{X} \cap \text{dom } H_t$. Because $(H_t)_{t \geq 0}$ has nondecreasing domain,

$$\begin{aligned} x \in \mathcal{X} \cap \text{dom } H_0 &= \mathcal{X} \cap \bigcap_{t \in \mathbb{N}} \text{dom } H_t = \bigcap_{t \in \mathbb{N}} (\mathcal{X} \cap \text{dom } H_t) \\ &= \bigcap_{t \in \mathbb{N}} \text{dom } h_t = \bigcap_{t \in \frac{1}{2}\mathbb{N}} \text{dom } h_t. \end{aligned}$$

Therefore, Lemma 2.5.2 can be applied with x and with notation therein, we get for all $T \geq 0$,

$$\sum_{t=0}^T \langle u_t, x - x_t \rangle = D_0 - D_{T+1} + \sum_{t=0}^T (D_{t+1/2}^\Delta + D_t^*).$$

First, because $x, x_0 \in \mathcal{X} \cap \text{dom } H_0$,

$$\begin{aligned} D_0 &= h_0(x) - h_0(x_0) - \langle \nabla H_0(x_0), x - x_0 \rangle \\ &= H_0(x) - H_0(x_0) - \langle \nabla H_0(x_0), x - x_0 \rangle \\ &= D_{H_0}(x, x_0). \end{aligned}$$

Similarly, x belongs to $\mathcal{X} \cap \text{dom } H_{T+1}$ and so does x_{T+1} as a consequence of $(x_t)_{t \geq 0}$ being OMD iterates (see Remark 3.3.9) and $D_{T+1} = D_{H_{T+1}}(x, x_{T+1})$.

Let us now bound the two remaining Bregman divergences. For $t \geq 0$, denote $y_t = \nabla H_t(x_t)$, then Proposition 3.3.6 ensures that

$$x_{t+1} = \arg \max_{x \in \mathcal{X}} \{ \langle y_t + u, x \rangle - H_t(x) \},$$

in other words x_{t+1} is a minimizer of a differentiable convex function over \mathcal{X} . Therefore, Proposition 1.2.10 gives the following variational characterization:

$$\langle \nabla H_t(x_{t+1}) - y_t - u_t, x' - x_{t+1} \rangle \geq 0, \quad x' \in \mathcal{X}.$$

We use it for $x' = x$ to bound $D_{t+1/2}^\Delta$ as follows.

$$\begin{aligned}
D_{t+1/2}^\Delta &= \Delta h_{t+1/2}(x) - \Delta h_{t+1/2}(x_{t+1}) \\
&\quad - \langle y_{t+1} - y_t - u_t, x - x_{t+1} \rangle \\
&\leq H_{t+1}(x) - H_t(x) - H_{t+1}(x_{t+1}) + H_t(x_{t+1}) \\
&\quad - \langle \nabla H_{t+1}(x_{t+1}) - \nabla H_t(x_{t+1}), x - x_{t+1} \rangle \\
&= D_{H_{t+1}-H_t}(x, x_{t+1}) \\
&= \tilde{D}_{t+1/2}^\Delta.
\end{aligned}$$

We now turn to D_t^* . Because $h_t \geq H_t$ by definition of h_t , the reverse inequality holds for the Legendre–Fenchel transform, and in particular, $h_t^*(y_t + u_t) \leq H_t^*(y_t + u_t)$. Besides, it holds that

$$x_t = \nabla H_t^*(y_t) \quad \text{and} \quad x_t = \nabla h_t^*(y_t),$$

where the first equality comes from Proposition 3.3.5, whereas the second is obtained combining the fact that $\nabla H_t(x_t) = y_t \in \partial h_t(x_t)$ (proved above) with Proposition 1.4.6. The latter proposition also gives a characterization of both above gradients as maximizers of concave functions, which yields

$$\begin{aligned}
h_t^*(y_t) &= \max_{x \in \mathbb{R}^d} \{ \langle y_t, x \rangle - h_t(x) \} = \langle y_t, x_t \rangle - h_t(x_t) \\
&= \langle y_t, x_t \rangle - H_t(x_t) = \max_{x \in \mathbb{R}^d} \{ \langle y_t, x \rangle - H_t(x) \} \\
&= H_t^*(y_t).
\end{aligned}$$

Hence

$$\begin{aligned}
D_t^* &= h_{t+1/2}^*(y_t + u_t) - h_t^*(y_t) - \langle u_t, x_t \rangle \leq H_t^*(y_t + u_t) - H_t^*(y_t) - \langle u_t, x_t \rangle \\
&= D_{H_t^*}(y_t + u_t, y_t).
\end{aligned}$$

□

OMD with nonincreasing step-sizes An important special case of OMD iterates is the following. Let H be a mirror map compatible with \mathcal{X} , $(\gamma_t)_{t \geq 0}$ a positive and nonincreasing sequence, and $(u_t)_{t \geq 0}$ a sequence in \mathbb{R}^d . Then, associated OMD iterates on \mathcal{X} with time-dependent *step-sizes* (or *learning rate*) are defined as

$$x_{t+1} = \arg \max_{x \in \mathcal{X}} \{ \langle \gamma_t u_t, x \rangle - D_H(x, x_t) \} \quad (3.10)$$

$$= \arg \max_{x \in \mathcal{X}} \{ \langle \nabla H(x_t) + \gamma_t u_t, x \rangle - H(x) \}. \quad (3.11)$$

This corresponds to OMD iterates from Definition 3.3.8 with time-dependent mirror maps $(H_t)_{t \geq 0}$ defined as $H_t = \gamma_t^{-1} H$ for all $t \geq 0$, which is indeed a sequence of mirror maps compatible with \mathcal{X} with nondecreasing domain.

Corollary 3.3.14. *Consider the OMD iterates with nonincreasing step-sizes defined in (3.10).*

(i) *For all $T \geq 0$ and $x \in \mathcal{X} \cap \text{dom } H$,*

$$\sum_{t=0}^T \langle u_t, x - x_t \rangle \leq \frac{\max_{0 \leq t \leq T} D_H(x, x_t)}{\gamma_T} + \sum_{t=0}^T \tilde{D}_t^*,$$

where $\tilde{D}_t^ = \gamma_t^{-1} D_{H^*}(\nabla H(x_t) + \gamma_t u_t, \nabla H_t(x_t))$.*

(ii) *Moreover, if H is K -strongly convex with respect to some norm $\|\cdot\|$, for all $T \geq 0$,*

$$\sum_{t=0}^T \langle u_t, x - x_t \rangle \leq \frac{\max_{0 \leq t \leq T} D_H(x, x_t)}{\gamma_T} + \frac{1}{2K} \sum_{t=0}^T \gamma_t \|u_t\|_*^2.$$

(iii) *Moreover, if there exists known constants $R > 0$ and $L > 0$, such that for all $t \geq 0$, $D_H(x, x_t) \leq R^2$ and $\|u_t\|_* \leq L$, then step-sizes $\gamma_t = R\sqrt{K}/(L\sqrt{t+1})$ yield for all $T \geq 0$,*

$$\max_{x \in \mathcal{X}} \sum_{t=0}^T \langle u_t, x - x_t \rangle \leq RL \sqrt{\frac{2(T+1)}{K}}.$$

Proof. (i) Applying Proposition 3.3.13,

$$\begin{aligned} \sum_{t=0}^T \langle u_t, x - x_t \rangle &= \frac{D_H(x, x_0)}{\gamma_0} - \frac{D_H(x, x_{T+1})}{\gamma_{T+1}} \\ &\quad + \sum_{t=0}^T \left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t} \right) D_H(x, x_{t+1}) + \sum_{t=0}^T \tilde{D}_t^* \\ &= \frac{D_H(x, x_0)}{\gamma_0} + \sum_{t=0}^{T-1} \left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t} \right) D_H(x, x_{t+1}) \\ &\quad - \frac{D_H(x, x_{T+1})}{\gamma_T} + \sum_{t=0}^T \tilde{D}_t^* \\ &\leq \left(\max_{1 \leq t \leq T} D_H(x, x_t) \right) \left(\frac{1}{\gamma_0} + \sum_{t=0}^{T-1} \left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t} \right) \right) + \sum_{t=0}^T \tilde{D}_t^* \\ &= \frac{\max_{1 \leq t \leq T} D_H(x, x_t)}{\gamma_T} + \sum_{t=0}^T \tilde{D}_t^*. \end{aligned}$$

Besides, for all $t \geq 0$ and $y \in \mathbb{R}^d$,

$$H_t^*(y) = \max_{x \in \mathbb{R}^d} \left\{ \langle y, x \rangle - \frac{H(x)}{\gamma_t} \right\} = \frac{1}{\gamma_t} \max_{x \in \mathbb{R}^d} \{ \langle \gamma_t y, x \rangle - H(x) \},$$

which yields

$$\tilde{D}_t^* = D_{H_t^*}(\nabla H_t(x_t) + u_t, \nabla H_t(x_t)) = \frac{D_{H^*}(\nabla H(x_t) + \gamma_t u_t, \nabla H(x_t))}{\gamma_t}.$$

(ii) and (iii) follow immediately. \square

Remark 3.3.15. Even for a fixed point $x \in \mathcal{X} \cap \text{dom } H$, unless there is a known bound on $\max_{t \geq 0} D_H(x, x_t)$, for instance in the case where \mathcal{X} is bounded, there is no known sublinear regret bound for OMD. This is contrast with DA which, with the right nonincreasing parameters from Proposition 3.2.5, guarantees a regret bound that grows as \sqrt{T} with time, as soon as a bound on the norms of vectors $(u_t)_{t \geq 0}$ is known.

Corollary 3.3.16. *Consider the special case of online gradient descent iterations*

$$x_{t+1} = \Pi_{\mathcal{X}}(x_t + u_t), \quad t \geq 0.$$

Then, for all $T \geq 0$ and $x \in \mathcal{X}$,

$$\sum_{t=0}^T \langle u_t, x - x_t \rangle \leq \frac{\max_{0 \leq t \leq T} \|x - x_t\|_2^2}{2\gamma_T} + \sum_{t=0}^T \frac{\gamma_t \|u_t\|_2^2}{2}.$$

Proof. Combine Corollaries 3.3.16 and 1.6.8. \square

Bibliography

- [Bub11] Sébastien Bubeck. *Introduction to Online Optimization: Lecture Notes*. Princeton University, 2011.
- [CBL06] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- [Haz16] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.
- [JKM23] Anatoli Juditsky, Joon Kwon, and Éric Moulines. Unifying mirror descent and dual averaging. *Mathematical Programming*, (199):793–830, 2023.
- [McM17] H. Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *The Journal of Machine Learning Research*, 18(1):3117–3166, 2017.
- [Ora23] Francesco Orabona. A modern introduction to online learning. *arXiv:1912.13213*, 2023.
- [Roc70] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
- [SS11] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.