Regret bounds

and their applications to online learning, optimization and games

Joon Kwon

February 1, 2024

Contents

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 (Extragradient, 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 $d \geqslant 1$. Throughout the chapter, we consider Euclidean space $\mathbb{R}^d$ equipped with its canonical inner product. For a set $A \subset \mathbb{R}^d$, $\operatorname{int} A$ and $\operatorname{cl} A$ denote its interior and closure, respectively.

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

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

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

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

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 $p, q \geqslant 1$ 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 $x, y \in \mathbb{R}^d$, $\langle y, x \rangle \leqslant \|y\|_* \|x\|$, which, together with the above examples recovers Cauchy-Schwarz and Hölder's inequalities.

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.

Definition 1.2.4 (Convex functions)**.** A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is *convex* if for all $x, x' \in \mathbb{R}^d$ and $\lambda \in [0,1]$,

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

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

Remark 1.2.5. The convexity of a function f is closely related to the convexity of sets, as the former can be equivalently defined as the epigraph

$$\operatorname{epi} f = \Big\{ (a, x) \in \mathbb{R}^d \times \mathbb{R},\ a \geqslant f(x) \Big\}$$

being convex.

Example 1.2.6. 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 $a \geqslant 1$. Function $x \mapsto \|x\|^a$ is convex.

Proposition 1.2.7 (Jensen's inequality)**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function, and X a random variable with values in* $\operatorname{dom} f$ *so that* $\mathbb{E}[X]$ *exists. Then, f is measurable,* $\mathbb{E}[f(X)]$ *exists in* $\mathbb{R} \cup \{+\infty\}$ *and*

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

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

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

Proposition 1.2.8 (First and second order characterizations of convexity)**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ a function with open domain.*

(i) If f is differentiable on its domain, f is convex if, and only if, for all $x, x' \in \operatorname{dom} f$,
$$f(x') \geqslant f(x) + \left\langle \nabla f(x), x' - x \right\rangle .$$

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

Proposition 1.2.9. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a nonempty convex set, and $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ 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,*
$$\forall x \in \mathcal{X},\ \langle \nabla f(x_*), x - x_* \rangle \geqslant 0.$$

Definition 1.2.10 (Lower semi-continuity)**.** A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous if for all $a \in \mathbb{R}^d$, the set $\left\{x \in \mathbb{R}^d,\ f(x) \leqslant a\right\}$ is closed.

Example 1.2.11. 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 1.2.12 (Negative entropy on the simplex)**.** The function $h_{\text{ent}} : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ 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 1.2.13. *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function and $\mathcal{X}_0 \subset \mathbb{R}^d$ a compact set such that $\operatorname{dom} f \cap \mathcal{X}_0 \neq \varnothing$. Then, f attains a minimum on $\mathcal{X}_0$.*

1.3 Subgradients

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

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

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

Proposition 1.3.4. *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a convex function and $x \in \operatorname{int} \operatorname{dom} f$. 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 $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ defined as:

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

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

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

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

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

1.4 Legendre–Fenchel transform

Definition 1.4.1. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper function. The *Legendre–Fenchel transform* (or *convex conjugate*) of f is a function $f^* : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ 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 $x, y \in \mathbb{R}^d$, $\langle y, x\rangle \leqslant f(x) + f^*(y)$.

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

Theorem 1.4.5 (Fenchel–Moreau). *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ 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 $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ 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) $x \in \operatorname{Arg\,max}_{x' \in \mathbb{R}^d} \{\langle y, x \rangle - f(x)\}$,

(v) $y \in \operatorname{Arg\,max}_{y' \in \mathbb{R}^d} \{\langle y', x \rangle - f^*(y')\}$.

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 involutional property of the Legendre-Fenchel transform is an extension of the involutional 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 $\mathcal{X} \subset \mathbb{R}^d$, $f : \mathcal{X} \to \mathbb{R}$, $x \in \operatorname{int} \mathcal{X}$ and $x' \in \mathcal{X}$ such that f is differentiable in x. Then, the *Bregman divergence* from x to x' is defined as

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

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

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

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 $x \in \operatorname{dom} f$, 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 $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ 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. ☐

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 $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, $\|\cdot\|$ 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 $\lambda \in [0, 1]$,

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

Definition 1.6.2 (Smoothness)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ 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 $x, x' \in \mathbb{R}^d$, $|D_f(x', x)| \leqslant \frac{L}{2} \|x' - x\|^2$, in other words,

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

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

Proposition 1.6.4 (Duality between strong convexity and smoothness)**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ 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 $y \in \partial f(x)$, $D_f(x', x;\ y) \geqslant \frac{K}{2}\|x'-x\|^2$, in other words*

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

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

$$\left\langle y'-y, x'-x\right\rangle \geqslant K\left\|x'-x\right\|^2.$$

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

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

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

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

(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 $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ a proper lower semicontinuous function which we assume K-strongly convex with respect to $\|\cdot\|$. Then, for all $y, y' \in \mathbb{R}^d$,*

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

Proposition 1.6.6 (Second order characterization of strong convexity)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ 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,*

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

Proposition 1.6.7 (First and second order characterizations of smoothness)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ 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,*

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

Corollary 1.6.8. *The squared Euclidean norm* $h_2 : x \mapsto \frac{1}{2}\left\|x\right\|_2^2$ *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* $h_p : x \mapsto \frac{1}{2}\left\|x\right\|_p^2$ *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* $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ *attains a minimum 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 $(x_t)_{t\geqslant 0}$ in $\mathcal{X}$ based on another sequence $(u_t)_{t\geqslant 0}$ in $\mathbb{R}^d$, where for each $t \geqslant 0$, 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. 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

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

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. Such a sequence satisfies the following elementary result.

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

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

Consequently, for all $T \geqslant 0$,

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

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

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

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: if $\frac{d\tilde{x}_t}{dt} = \tilde{u}_t$, then

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

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

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

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

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

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

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

We now turn to a constrained setting were 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 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 $t \geqslant 0$, $x_{t+1} = \operatorname{arg\,min}_{x \in \mathcal{X}} \left\|x_t + u_t - x\right\|_2^2$ if, and only if,*

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

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

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

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 $(x_t)_{t \geqslant 0}$ and $(y_t)_{t \geqslant 0}$ satisfy*

$$y_{t+1} = y_t + u_t \quad \text{and} \quad x_{t+1} = \operatorname*{arg\,min}_{x \in \mathcal{X}} \left\|y_{t+1} - x\right\|_2^2 , \qquad t \geqslant 0,$$

then for all $t \geqslant 0$,

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

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 $t \geqslant 0$, $x_{t+1} = x_t + A^{-1}u_t$ if, and only if*

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

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

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

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

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 $t \geqslant 0$,*

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

then for all $x \in \Delta_d$,

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

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

2.2 Regularizers

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

(ii) If h is differentiable on $\mathcal{D}_h := \operatorname{int}\operatorname{dom} h$ and $\nabla h(\mathcal{D}_h) = \mathbb{R}^d$, then h is an regularizer on $\mathcal{X}$.

(iii) If h is strongly convex, then h is an regularizer $\mathcal{X}$.

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 $\operatorname{dom} h^* = \mathbb{R}^d$.

(i) Because $\operatorname{cl}\operatorname{dom} h = \mathcal{X}$ 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 1.2.13, 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 strongly concave 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 an regularizer on $\mathcal{X}$. Then, h^* is differentiable on $\mathbb{R}^d$, ∇h^* takes values in* $\operatorname{dom} h \subset \mathcal{X}$*, and*

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

Proof. Let $y \in \mathbb{R}^d$. Because $\operatorname{dom} h^* = \mathbb{R}^d$, 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 $\operatorname{dom} h$, is unique, and thus h^* is differentiable at y by Proposition 1.3.4. □

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

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

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

$$\nabla h^*(y) = \operatorname*{arg\,min}_{x\in\mathcal{X}} \|y - x\|.$$

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

Proof. □

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

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

is a regularizer on Δ_d *and*

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

Proof. □

2.3 UMD iterates

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

(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 $y_t + u_t \in \partial h(x_t)$.

Definition 2.3.3. Let h be an regularizer on $\mathcal{X}$ and $(u_t)_{t \geqslant 0}$ a sequence in $\mathbb{R}^d$. A sequence $((x_t, y_t))_{t \geqslant 0}$ in $\mathbb{R}^d \times \mathbb{R}^d$ is a sequence of *strict UMD iterates* associated with h and $(u_t)_{t \geqslant 0}$ if for all $t \geqslant 0$,

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

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

Proposition 2.3.4. *Let* $((x_t, y_t))_{t \geqslant 0}$ *be a sequence of strict UMD iterates defined as above. Then for all* $t \geqslant 0$, $x_{t+1} = \nabla h^*(y_t + u_t)$ *and thus* $((x_t, y_t))_{t \geqslant 0}$ *are UMD iterates.*

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

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

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

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

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

Example 2.3.5 (Euclidean regularizer)**.** Denote $\Pi_{\mathcal{X}}$ the Euclidean projection onto $\mathcal{X}$ and consider the Euclidean regularizer on $\mathcal{X}$: $h = \frac{1}{2}\left\|\cdot\right\|_2^2 + I_{\mathcal{X}}$ and $x_0 \in \mathbb{R}^d$.

- If $x_{t+1} = \Pi_{\mathcal{X}}(x_t + u_t)$ for all $t \geqslant 0$, then $((x_t, x_t))_{t\geqslant 0}$ can be proved to be a sequence of strict UMD iterates.
- If $x_0 = y_0$, and $y_{t+1} = y_t + u_t$ and $x_{t+1} = \Pi_{\mathcal{X}}(y_{t+1})$ for all $t \geqslant 0$, then $((x_t, x_t))_{t\geqslant 0}$ can be proved to be a sequence of strict UMD iterates.

Remark 2.3.6 (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 $(u_t)_{t\geqslant 1}$ 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 $t \geqslant 0$. Then, one can verify that $((1, \frac{3+(-1)^t}{2}))_{t\geqslant 0}$ is a strict UMD sequence, and so is $((x_t, y_t))_{t\geqslant 0}$ where $x_t = y_t = \frac{1+(-1)^{t+1}}{2}$ for $t \geqslant 1$.

Remark 2.3.7 (Existence of strict UMD iterates)*.* As soon as regularizer h and sequence of dual increments $(u)_{t\geqslant 0}$ are given, we can see that associated strict UMD iterates always exist. Indeed, from the definition of regularizers, it follows that there exists $x_0 \in \mathcal{X}$ such that $\partial h(x_0) \neq \emptyset$; in other words, there exists (x_0, y_0) such that $x_0 = \nabla h^*(y_0)$. Then, for $t \geqslant 0$, 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 $y_{t+1} = \partial h(x_{t+1})$, as requires by (i).

Remark 2.3.8 (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

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

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

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

2.4 Regret bounds

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

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

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

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

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

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

and

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

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

$$D_t^* \leqslant \frac{1}{2K} \left\| u_t \right\|_*^2 .$$

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 $t \geqslant 0$,

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

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. ☐

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 $(u_t)_{t \geqslant 0}$ a sequence in $\mathbb{R}^d$. An associated sequence $((x_t, y_t))_{t \geqslant 0}$ 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,* $(u_t)_{t \geqslant 0}$ *a sequence in* $\mathbb{R}^d$*,* $((x_t, y_t))_{t \geqslant 1}$ *associated UMD iterates, and* $x \in \bigcap_{t \in \frac{1}{2}\mathbb{N}} \operatorname{dom} h_t$*. 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* $x' \in \operatorname{dom} h_t$ *(resp.* $\operatorname{dom} h_{t+1/2}$*),* $\Delta h_t(x') = h_{t+1/2}(x') - h_t(x')$ *(resp.* $\Delta h_{t+1/2}(x') = h_{t+1}(x) - h_{t+1/2}(x')$*),*
- $D^{\Delta}_{t+1/2} = \Delta h_{t+1/2}(x) - \Delta h_{t+1/2}(x_{t+1}) - \langle y_{t+1} - y_t - u_t, x - x_{t+1} \rangle$.

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

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

and

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

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

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

and

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

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

$$\langle u_t, x - x_{t+1} \rangle \leqslant \Phi_t - \Phi_{t+1} - D'_t,$$

and

$$\langle u_t, x - x_t \rangle \leqslant \Phi_t - \Phi_{t+1} + D^*_t,$$

where $\Phi_t = \langle y_t, x - x_t \rangle - h_t(x_t)$ *(and similarly for* Φ_{t+1}*).*

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

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

Proof. (i) can be proved by merely simplifying. Then, (ii) and (iii) are easy consequences.

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

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

□

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.