Regret bounds

and their applications to online learning, optimization and games

Joon Kwon

January 25, 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.

Acknowledgements

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

Introduction

Chapter 1

Convexity tools

1.1 Preliminairies

Let $d \geqslant 1$. Throughout the chapter, we consider Euclidean space $\mathbb{R}^d$. 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

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

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

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

Proposition 1.1.3. *Let $\|\cdot\|$ be a norm in $\mathbb{R}^d$. Then, $\|\cdot\|_{**} = \|\cdot\|$.*

Example 1.1.4 (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 1.1.5. 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 (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.

Definition 1.2.3 (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 1.2.4. The convexity of a function f is closely related to the convexity sets, as the former can be equivalently defined as the epigraph

$$\left\{ (a, x) \in \mathbb{R}^d \times \mathbb{R},\ a \geqslant f(x) \right\}$$

being convex.

Example 1.2.5. 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 pointwise supremum of convex functions.

Proposition 1.2.6 (Jensen's inequality)**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a measurable convex function, and X a random variable with values in* $\operatorname{dom} f$ *so that $\mathbb{E}[X]$ exists. Then,*

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

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 $\lambda_1 + \cdots + \lambda_n > 0$,

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

Example 1.2.7 (Norms)**.** Let $a \geqslant 1$. Any norm to the power a ($x \mapsto \|x\|^a$) is a convex function. In particular, convex functions $h_2 : x \mapsto \frac{1}{2} \|\cdot\|_2^2$ and more generally $h_p : x \mapsto \frac{1}{2} \|\cdot\|_p^2$ ($p \geqslant 1$), as well as squared Mahalanobis norms $h_A : x \mapsto \frac{1}{2} \langle x, Ax \rangle$ (where A is a positive definite $d \times d$ matrix) will be used in future chapters.

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 convex 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

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

If $\mathcal{X}$ is closed, then $I_{\mathcal{X}}$ is also lower-semicontinuous.

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 *Euclidean regularizer.*

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 $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. Then $x_* \in \mathbb{R}^d$ 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 $x \in \operatorname{int}\operatorname{dom} f$. Then, f is differentiable in x if, and only if, $\partial f(x) = \{\nabla f(x)\}$.*

Remark 1.3.5. Even in the case of a point in the domain of a convex function, the 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. *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 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 (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 diffrentiable and strictly convex), and is therefore 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:

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

As the primitive of an increasing function, f^* is also convex. It is easy to see that the operation is involutional: $f^{**} = f$. 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 would correspond to the usual duality between points and hyperplanes. The Lengendre–Fenchel transform in fact generalizes the above for higher dimension, and for nondifferentiable functions. The neatest properties of e.g. Propositions 1.4.3 and 1.4.4 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 *correspondance*.

Proposition 1.4.3. *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper function.*

(i) f^* is proper, lower-semicontinuous and convex.

(ii) If f is proper, lower-semicontinuous and convex, then $f^{**} = f$.

Proposition 1.4.4. *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 1.4.5 (Dual 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.

Example 1.4.6 (Squared norms)**.** Let $\|\cdot\|$ be a norm in $\mathbb{R}^d$ and consider $f : x \mapsto \frac{1}{2}\|x\|^2$. 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.

Definition 1.5.1. Let $\mathcal{X} \subset \mathbb{R}^d$, $f : \mathcal{X} \to \mathbb{R}$ and $x, x' \in \mathcal{X}$ 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 second 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,

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 subgradient may be empty. In the case of a convex function, it is nonnegative as soon as it exists. In the case of a convex function f that is 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-semicontinous 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.

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

where for the second equality, we applied Fenchel's identity from property (iii) in Proposition 1.4.4. □

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. Then for all $x, x' \in \mathbb{R}^d$, $D_h(x', x) = \frac{1}{2} \|x' - x\|_2^2$.

1.6 Smoothness and strong convexity

We now introduce strongly convexity and smoothness which intuitively correspond to the curvature of a function being respectively bounded from below (be 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 $\lambda \in [0,1]$,

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

Definition 1.6.2 (Smoothness)**.** Let $\Omega \subset \mathbb{R}^d$ be an open set and $f : \Omega \to \mathbb{R}$ 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 $x, x' \in \Omega$, $|D_f(x', x)| \leqslant \frac{K}{2} \|x' - x\|$, in other words,

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

Proposition 1.6.3 (Duality between strong convexity and smoothness)**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a 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 $y \in \partial f(x)$, $D_f(x', x;\ y) \geqslant \frac{1}{2} \|x' - x\|^2$, on other words*

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

(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*

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

(vi) *f^* is differentiable and $1/K$-smooth with respect to $\|\cdot\|_*$.*

Proposition 1.6.4 (Second order characterization of strong convexity)**.** *Let $\Omega \subset \mathbb{R}^d$ be an open set, $f : \Omega \to \mathbb{R}$ 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,*

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

Proposition 1.6.5 (First and second order characterizations of smoothness)**.** *Let $\Omega \subset \mathbb{R}^d$ be an open set, $f : \Omega \to \mathbb{R}$ a differentiable function 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 if, and only if, for all $u \in \mathbb{R}^d$,

$$\left| \left\langle u, \nabla^2 f(x) u \right\rangle \right| \leqslant L \left\| u \right\|^2 .$$

Corollary 1.6.6. *The squared Euclidean norm $h_2 : x \mapsto \frac{1}{2} \left\| x \right\|_2^2$ is 1-strongly convex with respect to $\|\cdot\|_2$.*

Proposition 1.6.7. *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 1.6.8. *The negative entropy h_{ent} is 1-strongly convex with respect to ℓ_1.*