UNIFYING MIRROR DESCENT AND DUAL AVERAGING

ANATOLI JUDITSKY, JOON KWON, AND ÉRIC MOULINES

Abstract. We introduce and analyse a new family of algorithms which generalizes and unifies both the mirror descent and the dual averaging algorithms. The unified analysis of the algorithms involves the introduction of a generalized Bregman divergence which utilizes subgradients instead of gradients. Our approach is general enough to encompass classical settings in convex optimization, online learning, and variational inequalities such as saddle-point problems.

1. Introduction

The family of mirror descent algorithms were initially introduced as first-order convex optimization algorithms, and were then extended to a variety of (online) optimization problems. Let us quickly recall the succession of ideas which have led to the the mirror descent algorithms.

Let us first consider the most basic setting, where the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and finite on the whole space $\mathbb{R}^n$, differentiable, and admits a unique minimizer $x_* \in \mathbb{R}^n$. We focus on the construction of algorithms based on first-order oracles (in other words, the algorithm may obtain the values of the function $f(x)$ and of its gradient $\nabla f(x)$ at querried points $x \in \mathbb{R}^n$) and which outputs points where the value of the objective function f is provably close to the minimum $f(x_*)$. The most basic such algorithm is the gradient descent, which starts at some initial point $x_1 \in \mathbb{R}^n$ and iterates:

$$x_{t+1} = x_t - \gamma \nabla f(x_t), \quad t \geqslant 1,$$

where $\gamma > 0$ is the step-size. An equivalent way of writing the above is the so-called *proximal* formulation:

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

where x_{t+1} appears as the solution of a simplified minimization problem where the objective function f has been replaced by its linearization at x_t *plus* a Euclidean *proximal term* $\frac{1}{2\gamma} \|x - x_t\|_2^2$ which prevents the next iterate x_{t+1} from being too far from x_t. This algorithm is well-suited to assumptions regarding the objective function f which involve the Euclidean norm (e.g. if ∇f is bounded (or Lipschitz-continuous) with respect to the Euclidean norm).

The mirror descent algorithm, first introduced in [38, 41] and further studied in [5], can be presented as an extension of the above gradient descent, by replacing the Euclidean proximal term by a *Bregman divergence* [6] associated with a differentiable convex function $F : \mathbb{R}^n \to \mathbb{R}$:

$$x_{t+1} = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \left\{ f(x_t) + \langle \nabla f(x_t) | x - x_t \rangle + \frac{1}{\gamma} D_F(x, x_t) \right\},$$

where, for any $(x, x') \in \mathbb{R}^n$ the Bregman divergence is defined as:

$$D_F(x', x) := F(x') - F(x) - \left\langle \nabla F(x) | x' - x \right\rangle .$$

These mirror descent algorithms, with a carefully chosen function F, are used to better suit the geometry of the problem, for instance when the objective function f is Lipschitz-continuous or

Date: October 31, 2019.

smooth with respect to a non-Euclidean norm. One can see that the above mirror descent iteration can we equivalently written (under appropriate assumptions on F):

$$x_{t+1} = \nabla F^*(\nabla F(x_t) - \gamma \nabla f(x_t)), \tag{1}$$

where F^* is the Legendre–Fenchel transform of F given by:

$$F^*(\vartheta) = \max_{x \in \mathbb{R}^n} \left\{ \langle \vartheta | x \rangle - F(x) \right\}, \qquad \vartheta \in \mathbb{R}^n.$$

This formulation makes explicit the distinction between the primal space (where the iterates $(x_t)_{t \geqslant 1}$ live) and the dual space (where the gradients $(\nabla f(x_t))_{t \geqslant 1}$ belong): point x_t is mapped from the primal into the dual using ∇F, the gradient step is then performed in the dual space $(\nabla F(x_t) - \gamma \nabla f(x_t))$, and the point thus obtained is finally mapped back into the primal space using ∇F^*.

We now move on to *constrained* problems. Let $\mathcal{X} \subset \mathbb{R}^n$ be a closed convex set. Even if an iterate x_t belongs to the set $\mathcal{X}$, the above gradient/mirror descent iteration may well propose an iterate x_{t+1} which lays outside of $\mathcal{X}$. Therefore, to force the next iterate x_{t+1} to belong to $\mathcal{X}$, the mirror descent can be adapted with the help of an additional *projection step*. But this projection can be introduced in at least two ways, which give rise to two families of algorithms: *mirror descent* (MD) which can be traced back to the pioneering work of [41, Chapter 3] and *dual averaging* (DA) introduced in [45, 52], sometimes called *lazy* mirror descent or *follow the regularizer leader*. To illustrate the similarities and differences between MD and DA, we here describe the simple Euclidean case for each of the two families. The MD algorithm in the Euclidean case corresponds to the *projected gradient descent* [21, 33]. For a given initial point $x_1 \in \mathcal{X}$, it writes, for all $t \geqslant 1$,

$$y_{t+1} = x_t - \gamma \nabla f(x_t) \quad \text{and} \quad x_{t+1} = \operatorname{proj}_{\mathcal{X}}(y_{t+1}),$$

where $\operatorname{proj}_{\mathcal{X}}$ denotes the Euclidean projection onto $\mathcal{X}$. In other words, it first performs a gradient step, then projects the point thus obtained onto the set $\mathcal{X}$. Then the next gradient step is performed starting from point x_{t+1}.

For a given initial point $\vartheta_0 \in \mathbb{R}^n$, the corresponding algorithm in the DA family writes, for all $t \geqslant 0$:

$$\vartheta_{t+1} = \vartheta_t - \gamma \nabla f(x_t) \quad \text{and} \quad x_{t+1} = \operatorname{proj}_{\mathcal{X}}(\vartheta_{t+1}).$$

The difference with the projected gradient descent is that the gradient steps are performed starting from the *unprojected* point ϑ_t.

The MD and DA algorithms share similarities in their analysis and in the guarantees they provide. However, their differences lead the two families to be well-suited for different situations. DA were advantageous in distributed problems [16, 19], and manifold identification [18, 32], for instance. They also show better averaging properties in the presence of noise [20]. On the other hand, MD is known to provide better convergence rates in some cases (e.g. when the objective function f is assumed to be smooth, see also [20, Section 4.2] for another simple case). MD also achieves (unlike DA) the optimal rate in the adversarial multi-armed bandit problem [2, 3] and the online combinatorial optimization problem with semi-bandit feedback [4] and bandit feedback [10, 14].

The mirror descent algorithms were also tranposed to provide solutions for other problems than convex optimization. We already mentionned bandit problems. More generally, the mirror descent algorithms have been successful in online learning—see e.g. [7, 11, 22, 48, 50, 51, 53]. They also provided solutions for saddle-point problems [40], and similar procedures were used for estimator aggregation in statistical learning [28, 29]. See also [35, Appendix C] for a discussion comparing MD and DA.

A closely related family of algorithms, which has attracted much attention, is obtained by allowing the regularizers or mirror maps—see below Sections 2 and 3—to vary over time (possibly as a function of previous data) [12, 17, 24–26, 36, 47]—see [35] for a recent survey. This will not be the case in this work.

1.1. **Main contributions.** We introduce and study a new family of algorithms which unifies and extends both the mirror descent and the dual averaging. The general algorithm has the property of offering at each step several possible iterations. Our main result is an analysis of the algorithm which is general enough to be then applied to various settings such as convex optimization, regret minimization, saddle-point problems, etc.

1.2. **Other works.** To the best of our knowledge, the only other works which present in a common unifying framework the mirror descent and the dual averaging algorithms [34, 35], and is quite different from the present work. In our approach, the difference between the two methods appears as a result of the problem being constrained in a set $\mathcal{X} \subsetneq \mathbb{R}^n$. In [34, 35], although the problems are unconstrained, the difference between mirror descent and dual averaging appears as a result of having regularizers/mirror maps which vary over time. The unification is then achieved by tweaking the way the time-varying regularizers/mirror map are defined. Our unified algorithm is based on quite different ideas and has the property of offering several possible iterations at some steps, which is not the case in [34, 35].

1.3. **Paper outline.** In Section 2 (resp. 3) we recall the definition of the mirror descent (resp. dual averaging) algorithms. In Section 4 we define our new family of algorithms which we call *unified mirror descent*, and establish that the mirror descent and dual averaging families are special cases. We then establish our main result, which is the guarantee offered by the unified mirror descent algorithms. In Section 5, we present various classical problems in which the unifed mirror descent algorithms can be applied.

1.4. **Preliminaries and notation.** Throughout the paper, we consider algorithms associated with an arbitrary sequence $(\xi_t)_{t\geqslant 1}$ in $\mathbb{R}^n$. The special case of gradient/mirror descent for the minimization of a differentiable objective function f is recovered by considering sequences of the form $\xi_t = -\gamma\nabla f(x_t)$.

For $x, \vartheta \in \mathbb{R}^n$, $\langle \vartheta | x\rangle$ denotes the canonical scalar product. For a given set $\mathcal{A} \subset \mathbb{R}^n$, $\operatorname{int}\mathcal{A}$ and $\operatorname{cl}\mathcal{A}$ denote its interior and closure respectively. For a given norm $\|\cdot\|$ in $\mathbb{R}^n$, we denote $\|\cdot\|_*$ its dual norm is defined as

$$\|\vartheta\|_* := \max_{\|x\|\leqslant 1} \langle \vartheta | x\rangle \ .$$

The convex characteristic $I_{\mathcal{C}} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ of a convex set $\mathcal{C} \subset \mathbb{R}^n$ is zero on $\mathcal{C}$ and equal to $+\infty$ elsewhere. Denote $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ a convex function. Its domain $\operatorname{dom} g$ is the set $\{x \in \mathbb{R}^n : g(x) < \infty\}$. Its subdifferential $\partial g(x)$ at point $x \in \mathbb{R}^n$ is the set of vectors $\vartheta \in \mathbb{R}^n$ such that

$$\forall x' \in \mathbb{R}^n, \quad g(x') - g(x) \geqslant \left\langle \vartheta | x' - x\right\rangle \ .$$

The Legendre–Fenchel transform of g is defined by:

$$g^*(\vartheta) = \max_{x\in\mathbb{R}^n} \left\{\langle \vartheta | x\rangle - g(x)\right\}, \qquad \vartheta \in \mathbb{R}^n.$$

If g is differentiable at a given point $x \in \mathbb{R}^n$, its Bregman divergence between x and any point $x' \in \mathbb{R}^n$ is defined as

$$D_g(x', x) = g(x') - g(x) - \left\langle \nabla g(x) | x' - x\right\rangle \ .$$

Further convexity definitions and results are recalled in Section A.

Throughout the paper, $\mathcal{X} \subset \mathbb{R}^n$ will be a closed and nonempty convex set.

2. Mirror descent

The greedy mirror descent algorithms rely on the notion of *mirror maps* that we now recall. Our presentation draws inspiration from [8], with a few differences in definitions and conventions.

Definition 2.1. Let $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a function. Denote $\mathcal{D}_F := \operatorname{int}\operatorname{dom} F$. F is a $\mathcal{X}$-compatible *mirror map* if

(i) F is lower-semicontinous and strictly convex,
(ii) F is differentiable on $\mathcal{D}_F$,
(iii) the gradient of F takes all possible values, i.e. $\nabla F(\mathcal{D}_F) = \mathbb{R}^n$.
(iv) $\mathcal{X} \subset \operatorname{cl}\mathcal{D}_F$,
(v) $\mathcal{X} \cap \mathcal{D}_F \neq \varnothing$,

The following proposition gathers a few properties about mirror maps. For the sake of completeness, the proofs given in Appendix B.

Proposition 2.2. *Let $F : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ be a $\mathcal{X}$-compatible mirror map, F^* the Legendre–Fenchel transform of F, and $\mathcal{D}_F := \operatorname{int}\operatorname{dom} F$. Then,*

(i) $\operatorname{dom} F^* = \mathbb{R}^n$,
(ii) F^* *is differentiable on* $\mathbb{R}^n$,
(iii) $\nabla F^*(\mathbb{R}^n) = \mathcal{D}_F$,
(iv) For all $x \in \mathcal{D}_F$ *and* $y \in \mathbb{R}^n$, *we have* $\nabla F^*(\nabla F(x)) = x$ *and* $\nabla F(\nabla F^*(y)) = y$.

We can now define the mirror descent algorithm [38], [8, Section 4.2].

Definition 2.3. Let F be a $\mathcal{X}$-compatible mirror map, $\xi := (\xi_t)_{t\geqslant 1}$ be a sequence in $\mathbb{R}^n$ and $x_1 \in \mathcal{X} \cap \mathcal{D}_F$ an initial point. We define the associated MD iterates as follows:

$$x_{t+1} = \operatorname*{arg\,min}_{x\in\mathcal{X}} D_F(x,\ \nabla F^*(\nabla F(x_t) + \xi_t)), \qquad t \geqslant 1. \tag{MD}$$

$(x_t)_{t\geqslant 1}$ is then said to be a $\mathrm{MD}(\mathcal{X}, F, \xi)$ sequence and $\xi = (\xi_t)_{t\geqslant 1}$ is called the sequence of *dual increments.*

The above is well-defined thanks to the following induction. As soon as x_t ($t \geqslant 1$) belongs to $\mathcal{X}\cap\mathcal{D}_F$, $\nabla F(x_t)$ exists because F is differentiable on $\mathcal{D}_F$ by Definition 2.1. Then, $\nabla F^*(\nabla F(x_t)+\xi_t)$ exists because F^* is differentiable on $\mathbb{R}^n$ by Proposition 2.2–(ii). Then, the next iterate x_{t+1} is obtained using the Bregman projection onto $\mathcal{X}$, which is well-defined and belongs to $\mathcal{X}\cap\mathcal{D}_F$ thanks to Theorem 2.4 below.

Statements similar to the following can be found in the literature (see e.g. [15, Lemma A.1]) but we could not find one which matches our exact assumptions on F and $\mathcal{X}$. Therefore, we give a detailed proof in Appendix B for completeness.

Theorem 2.4 (Bregman projection onto $\mathcal{X}$)**.** *Let $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a $\mathcal{X}$-compatible mirror map and denote $\mathcal{D}_F := \operatorname{int}\operatorname{dom} F$. Then for any $x_0 \in \mathcal{D}_F$, $\operatorname{arg\,min}_{x\in\mathcal{X}} D_F(x, x_0)$ exists and is unique. Moreover, it belongs to $\mathcal{X} \cap \mathcal{D}_F$; in other words:*

$$\operatorname*{arg\,min}_{x\in\mathcal{X}} D_F(x, x_0) = \operatorname*{arg\,min}_{x\in\mathcal{X}\cap\mathcal{D}_F} D_F(x, x_0).$$

The above (MD) iteration can be rewritten as follows. Denote $\tilde{x}_t := \nabla F^*(\nabla F(x_t) + \xi_t)$. Using Proposition 2.2–(iv), we get $\nabla F(\tilde{x}_t) = F(x_t) + \xi_t$. Then, using the definition of the Bregman

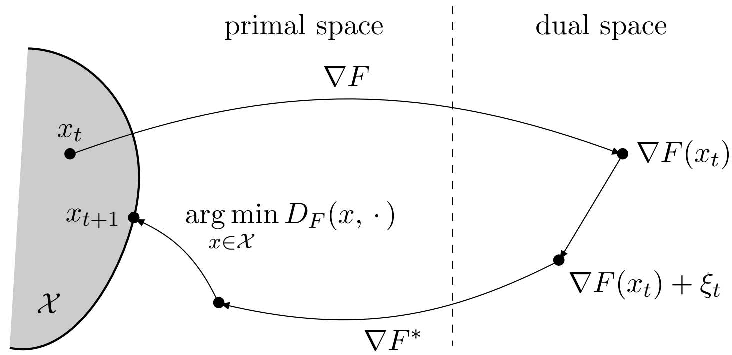


FIGURE 1. Mirror descent

divergence:

$$\begin{aligned} x_{t+1} &= \arg\min_{x\in\mathcal{X}} \{F(x) - F(\tilde{x}_t) - \langle \nabla F(\tilde{x}_t) | x - \tilde{x}_t \rangle\} \\ &= \arg\min_{x\in\mathcal{X}} \{F(x) - \langle \nabla f(x_t) + \xi_t | x \rangle\} \\ &= \arg\min_{x\in\mathcal{X}} \{F(x) - F(x_t) - \langle \nabla f(x_t) | x - x_t \rangle - \langle \xi_t | x \rangle\} \\ &= \arg\min_{x\in\mathcal{X}} \{- \langle \xi_t | x \rangle + D_F(x, x_t)\} . \end{aligned}$$

The above last expression is called *primal formulation*, and is taken as the definition of mirror descent algorithms in some works—see e.g. [5, Section 3]. Introducing the so-called *prox-mapping*:

$$T_{\mathcal{X},F}(u,x) := \arg\min_{x'\in\mathcal{X}} \left\{- \left\langle u \middle| x' \right\rangle + D_F(x', x)\right\}, \quad u\in\mathbb{R}^n,\ x\in\mathcal{X}\cap\mathcal{D}_F,$$

the MD iterates starting from some point $x_1 \in \mathcal{X}\cap\mathcal{D}_F$ can then be alternatively written as:

(MD-prox) $$x_{t+1} = T_{\mathcal{X},F}(\xi_t, x_t), \quad t \geqslant 1.$$

We now give a few examples of mirror maps.

Example 2.5 (Gradient descent)**.** The most simple example is $\mathcal{X} = \mathbb{R}^n$ and F defined on $\mathbb{R}^n$ by $F(x) = \frac{1}{2}\|x\|_2^2$. One can easily see that F is indeed a $\mathbb{R}^n$-compatible mirror map. We then have $\nabla F = \nabla F^* = \mathrm{I}_n$. Besides, if $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable objective function, and if we consider dual increments $\xi_t := -\gamma\nabla f(x_t)$, then update rule (MD) boils down to the gradient descent algorithm.

Example 2.6 (Projected gradient descent)**.** A common variant of the above is the case where $\mathcal{X}$ is some closed proper subset of $\mathbb{R}^n$. The same function $F(x) = \frac{1}{2}\|x\|_2^2$ is a $\mathcal{X}$-compatible mirror map. Again, if f is a objective function which is differentiable on the interior of $\mathcal{X}$, then considering $\xi_t = -\gamma\nabla f(x_t)$ makes (MD) correspond to the projected gradient algorithm.

Example 2.7 (Exponential weights)**.** A frequent special case corresponds to $\mathcal{X}$ being the n-simplex:

$$\mathcal{X} = \left\{ x\in\mathbb{R}_+^n \,\middle|\, \sum_{i=1}^n x_i = 1 \right\},$$

and F defined by $F(x) = \sum_{i=1}^n x_i \log x_i$ for $x\in\mathbb{R}_+^n$ (using convention $0\log 0 = 0$) and $F(x) = +\infty$ for $x \notin \mathbb{R}_+^n$. F is then a $\mathcal{X}$-compatible mirror map.

3. Dual averaging

The dual averaging algorithms rely on the notion of regularizers which we now recall. These are less restrictive than mirror maps: we see below in Proposition 3.4 that for a given mirror map, there exists a corresponding regularizer but the converse is not true. A regularizer may have a domain with empty interior (and therefore be nowhere differentiable), whereas mirror maps must be differentiable on the interior of their domain.

Definition 3.1 (Regularizers)**.** Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a function. h is a *$\mathcal{X}$-pre-regularizer* if it is strictly convex, lower-semicontinous, and if $\operatorname{cl} \operatorname{dom} h = \mathcal{X}$. Moreover, if $\operatorname{dom} h^* = \mathbb{R}^n$, then h is said to be a *$\mathcal{X}$-regularizer*.

The following proposition gives several sufficient conditions for the above condition $\operatorname{dom} h^* = \mathbb{R}^n$ to be satisfied. The proof is postponed to Appendix B.

Proposition 3.2. *Let h be a $\mathcal{X}$-pre-regularizer.*

(i) If $\mathcal{X}$ is compact, then h is a $\mathcal{X}$-regularizer.
(ii) If h is differentiable on $\mathcal{D}_h := \operatorname{int} \operatorname{dom} h$ and $\nabla h(\mathcal{D}_h) = \mathbb{R}^n$, then h is a $\mathcal{X}$-regularizer.
(iii) If h is strongly convex with respect to some norm $\|\cdot\|$, then h is a $\mathcal{X}$ regularizer.

Proposition 3.3 (Differentiability of h^*)**.** *Let h be a $\mathcal{X}$-regularizer. Then h^* is differentiable on $\mathbb{R}^n$.*

Proposition 3.4. *Let F be a $\mathcal{X}$-compatible mirror map. Then, $h := F + I_{\mathcal{X}}$ is a $\mathcal{X}$-regularizer.*

Corollary 3.5. *(i) $h(x) := \frac{1}{2}\|x\|_2^2 + I_{\mathcal{X}}(x)$ is a $\mathcal{X}$-regularizer.*
(ii) The entropy defined as:

$$h(x) := \begin{cases} \sum_{i=1}^n x_i \log x_i & \text{if } x \in \Delta_n \\ +\infty & \text{otherwise}, \end{cases}$$

where $\Delta_n := \left\{ x \in \mathbb{R}_+^n \,\middle|\, \sum_{i=1}^n x_i = 1 \right\}$ and where we use convention $0 \log 0 = 1$, is a Δ_n-regularizer.

Example 3.6 (Elastic net regularization)**.** An example of regularizer which does not have a mirror map counterpart, because it fails to be differentiable, is the so-called *elastic net* regularizer:

$$h(x) := \|x\|_1 + \|x\|_2^2,$$

which is indeed a $\mathbb{R}^n$-regularizer by strong convexity (Proposition 3.2).

We now recall the definition of the dual averaging (DA) iterates.

Definition 3.7 (Dual averaging [45, 52])**.** Let h be a $\mathcal{X}$-regularizer and $\xi := (\xi_t)_{t \geqslant 1}$ be a sequence in $\mathbb{R}^n$. A sequence $(x_t, \vartheta_t)_{t \geqslant 1}$ is said to be a sequence of DA iterates associated with h and ξ (DA(h, ξ) for short) if for $t \geqslant 1$:

$$\begin{aligned} x_t &= \nabla h^*(\vartheta_t) \\ \vartheta_{t+1} &= \vartheta_t + \xi_t. \end{aligned} \tag{DA}$$

Points $(x_t)_{t \geqslant 1}$ (resp. $(\vartheta_t)_{t \geqslant 1}$) are then called *primal iterates* (resp. *dual iterates*), and vectors $(\xi_t)_{t \geqslant 1}$ are called *dual increments*.

We can see that for a given couple (x_1, ϑ_1) of initial points satisfying $x_1 = \nabla h^*(\vartheta_1)$, and a sequence $(\xi_t)_{t \geqslant 1}$ of dual increments, the subsequent iterates $(x_t, \vartheta_t)_{t \geqslant 2}$ are well-defined and unique.

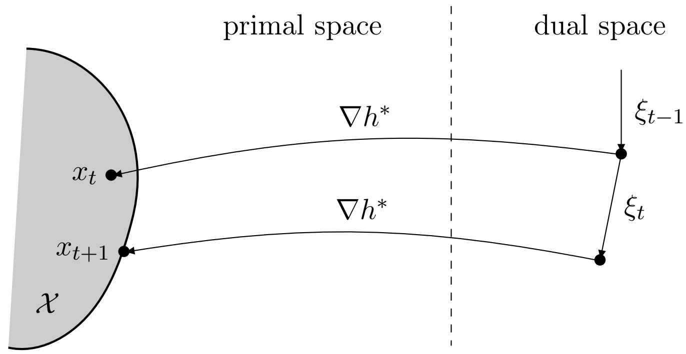


FIGURE 2. Dual averaging

4. The unified mirror descent algorithm

We introduce in this section our general family of algorithms which we call unified mirror descent (UMD) and prove that MD and DA are special cases. We then establish a guarantee provided by the UMD algorithm.

4.1. Definition, properties and special cases.

Definition 4.1. Let h be a $\mathcal{X}$-regularizer and $\xi := (\xi_t)_{t\geqslant 1}$ be a sequence in $\mathbb{R}^n$. We say that $(x_t, \vartheta_t)_{t\geqslant 1}$ is a UMD sequence associated with h and ξ (or a UMD(h,ξ) sequence for short) if for all $t \geqslant 1$:

$$x_t = \nabla h^*(\vartheta_t), \tag{I}$$

$$\forall x \in \mathcal{X},\ \langle \vartheta_{t+1} - \vartheta_t - \xi_t | x - x_{t+1}\rangle \geqslant 0. \tag{II}$$

Points $(x_t)_{t\geqslant 1}$ (resp. $(\vartheta_t)_{t\geqslant 1}$) are called *primal iterates* (resp. *dual iterates*), and vectors $(\xi_t)_{t\geqslant 1}$ are called *dual increments.*

Proposition 4.2. *Let $(x_t, \vartheta_t)_{t\geqslant 1}$ be an UMD(h, ξ) sequence defined as above. Then for all $t \geqslant 1$,*

(i) $\vartheta_t \in \partial h(x_t)$;
(ii) $\vartheta_t + \xi_t \in \partial h(x_{t+1})$ and $x_{t+1} = \nabla h^*(\vartheta_t + \xi_t)$.

Proof. Let $t \geqslant 1$. By definition of UMD iterates, we have $x_t = \nabla h^*(\vartheta_t)$, which combined with Propositions A.2 and A.3 gives property (i).

For all $x \in \mathcal{X}$ we deduce from $\vartheta_{t+1} \in \partial h(x_{t+1})$ that:

$$h(x) - h(x_{t+1}) \geqslant \langle \vartheta_{t+1} | x - x_{t+1}\rangle \geqslant \langle \vartheta_t + \xi_t | x - x_{t+1}\rangle .$$

where we used variational condition (II) from the definition of UMD iterates to get the second inequality. Then, inequality $h(x) - h(x_{t+1}) \geqslant \langle \vartheta_t + \xi_t | x - x_{t+1}\rangle$ also holds for $x \notin \mathcal{X}$ because then $h(x) = +\infty$. This proves that $\vartheta_t + \xi_t$ also belongs to $\partial h(x_{t+1})$, i.e. property (ii). □

Remark 4.3 (Existence of UMD iterates). As soon as $\mathcal{X}$-regularizer h and sequence of dual increments $(\xi_t)_{t\geqslant 1}$ are given, we can see that UMD(h, ξ) always exist. Indeed, from the definition of regularizers, it follows that there exists a primal point $x_1 \in \mathcal{X}$ such that $\partial h(x_1) \neq \varnothing$; in other words, there exists (x_1, ϑ_1) such that $x_1 = \nabla h^*(\vartheta_1)$. Then, for $t \geqslant 1$, one can consider $\vartheta_{t+1} := \vartheta_t + \xi_t$ which indeed satisfies variational condition (II), and then define $x_{t+1} := \nabla h^*(\vartheta_{t+1})$. This choice of ϑ_t actually corresponds to the iteration of the DA algorithm, as will be detailed in the proof of Proposition 4.5.

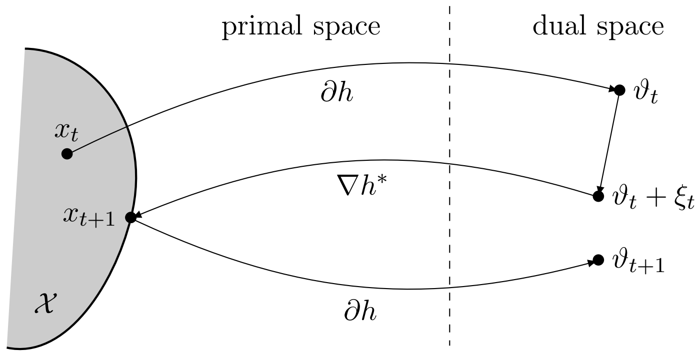


FIGURE 3. Unified mirror descent

We can consider the following alternative definition of UMD iterates. Let $\Pi_h : \mathbb{R}^n \rightrightarrows \mathcal{X} \times \mathbb{R}^n$ be a multi-valued *prox-mapping* defined as follows. $\Pi_h(\zeta)$ is the set of couples (x, ϑ) satisfying:

$$\begin{aligned}
&x = \nabla h^*(\zeta) \\
&\vartheta \in \partial h(x) \\
&\forall x' \in \mathcal{X},\ \left\langle \vartheta - \zeta | x' - x \right\rangle \geqslant 0.
\end{aligned}$$

Then, it can be easily checked that $(x_t, \vartheta_t)_{t\geqslant 1}$ is a sequence of UMD(h, ξ) iterates if and only if:

$$\text{(UMD-prox)} \qquad \begin{aligned}
&\vartheta_1 \in \partial h(x_1) \\
&(x_{t+1}, \vartheta_{t+1}) \in \Pi_h(\vartheta_t + \xi_t), \quad t \geqslant 1.
\end{aligned}$$

Remark 4.4 (On the non-unicity of UMD iterates). An interesting character of the UMD algorithm is that for a given sequence $(\xi_t)_{t\geqslant 1}$ of dual increments and initial points (x_1, ϑ_1), there may be several possible UMD sequences because the prox-mapping Π_h is multi-valued. However, as soon as the subdifferential $\partial h(x)$ is at most a singleton at each point $x \in \mathcal{X}$, the prox-mapping Π_h is single-valued and the UMD sequence is thus unique; in particuliar, DA and MD then coincide. This is the case for instance if $\mathcal{X} = \mathbb{R}^n$ and if the regularizer h is differentiable on $\mathbb{R}^n$.

Proposition 4.5 (DA is a special case of UMD)**.** *Let h be a $\mathcal{X}$-regularizer, $\xi := (\xi_t)_{t\geqslant 1}$ be a sequence in $\mathbb{R}^n$. Let $(x_t, \vartheta_t)_{t\geqslant 1}$ be DA(h, ξ) iterates. Then, $(x_t, \vartheta_t)_{t\geqslant 1}$ are UMD(h, ξ) iterates.*

Proof. First, condition (I) is true by definition of (DA). Besides, the relation $\vartheta_{t+1} = \vartheta_t + \xi_t$ makes condition (II) trivially satisfied because one of the arguments of the scalar product is zero. □

Proposition 4.6 (MD is a special case of UMD)**.** *Let F be a $\mathcal{X}$-compatible mirror map and $\xi := (\xi_t)_{t\geqslant 1}$ be a sequence in $\mathbb{R}^n$. Let $(x_t)_{t\geqslant 1}$ be a sequence of MD$(F, \mathcal{X}, \xi)$ iterates. Then, $(x_t, \nabla F(x_t))_{t\geqslant 1}$ is a sequence of UMD$(F + I_{\mathcal{X}}, \xi)$ iterates.*

Proof. For $t \geqslant 1$, we consider $\vartheta_t := \nabla F(x_t)$. Let us prove that conditions (I) and (II) are satisfied with regularizer $h := F + I_{\mathcal{X}}$.

For $t \geqslant 1$, denote $\tilde{x}_t := \nabla F^*(\nabla F(x_t) + \xi_t)$, which implies $\nabla F(\tilde{x}_t) = \nabla F(x_t) + \xi_t = \vartheta_t + \xi_t$ thanks to Proposition A.2. We can then rewrite the (MD) iteration as follows:

$$\begin{aligned}
x_{t+1} &= \operatorname*{arg\,min}_{x\in\mathcal{X}} D_F(x,\ \tilde{x}_t) \\
&= \operatorname*{arg\,min}_{x\in\mathcal{X}} \left\{ F(x) - F(\tilde{x}_t) - \langle \nabla F(\tilde{x}_t) | x - \tilde{x}_t \rangle \right\} \\
&= \operatorname*{arg\,min}_{x\in\mathcal{X}} \left\{ F(x) - \langle \vartheta_t + \xi_t | x \rangle \right\}.
\end{aligned}$$

In other words, x_{t+1} is the minimiser on $\mathcal{X}$ of the convex function $x \mapsto F(x) - \langle \vartheta_t + \xi_t | x \rangle$. This function is differentiable at x_{t+1} because we know by Theorem 2.4 that $x_{t+1} \in \mathcal{D}_F := \operatorname{int} \operatorname{dom} F$ and F is differentiable on $\mathcal{D}_F$. Applying the variational characterization from Proposition A.4, we get

$$\forall x \in \mathcal{X}, \quad \langle \nabla F(x_{t+1}) - \vartheta_t - \xi_t | x - x_{t+1} \rangle \geqslant 0,$$

which is exactly condition (II) because we just set $\vartheta_{t+1} = \nabla F(x_{t+1})$.

By convexity of F, the following is true

$$\forall x \in \mathbb{R}^n, \quad F(x) - F(x_t) \geqslant \langle \nabla F(x_t) | x - x_t \rangle \, .$$

By definition of h, we obviously have $h(x) \geqslant F(x)$ for all $x \in \mathbb{R}^n$, and $h(x_t) = F(x_t) + I_{\mathcal{X}}(x_t) = F(x_t)$ because $x_t \in \mathcal{X}$. Therefore, the following is also true

$$\forall x \in \mathbb{R}^n, \quad h(x) - h(x_t) \geqslant \langle \nabla F(x_t) | x - x_t \rangle \, .$$

In other words, $\vartheta_t = \nabla F(x_t) \in \partial h(x_t)$, which is equivalent to $x_t \in \nabla h^*(\vartheta_t)$ (see Propositions A.2 and (A.3)), and condition (I) is satisfied. □

4.2. **Simple examples.** Let us describe and compare the iterates of MD, DA and UMD in a Euclidean setting: in this section only we consider $\mathcal{X}$-regularizer $h := \frac{1}{2} \|x\|_2^2 + I_{\mathcal{X}}$ as well as mirror map $F := \frac{1}{2} \|x\|_2^2$. It is then easy to check that the map ∇h^* is the Euclidean projection onto $\mathcal{X}$. We denote $(x_t, \vartheta_t)_{t \geqslant 1}$ a sequence of UMD(h, ξ) iterates. We consider below two simple cases for the set $\mathcal{X}$.

4.2.1. *Euclidean ball.* We here consider the case where the set $\mathcal{X} = \overline{B}(0,1)$ is the closed unit Euclidean ball. Let $t \geqslant 1$ and assume that $\vartheta_t + \xi_t$ is outside of $\overline{B}(0,1)$ so that x_{t+1}, which is the Euclidean projection of $\vartheta_t + \xi_t$, belongs to the boundary of $\overline{B}(0,1)$, in other words, $\|x_{t+1}\|_2 = 1$. From this point x_{t+1}, a MD iteration corresponds to choosing $\vartheta_{t+1}^{\mathrm{MD}} := x_{t+1}$, and a DA interation corresponds to choosing $\vartheta_{t+1}^{\mathrm{DA}} := \vartheta_t + \xi_t$. Besides, we can see that the set of points ϑ_{t+1} with have x_{t+1} as Euclidean projection is is $[1, +\infty)\, x_{t+1}$ and that the set of points ϑ_{t+1} satisfying condition (II) is $(-\infty, 1]\, (\vartheta_t + \xi_t)$. Therefore, the set of vectors ϑ_{t+1} satisfying both conditions (II) and (I) is the convex hull of x_{t+1} and $\vartheta_t + \xi_t$, which is represented by a thick segment in Figure 4.

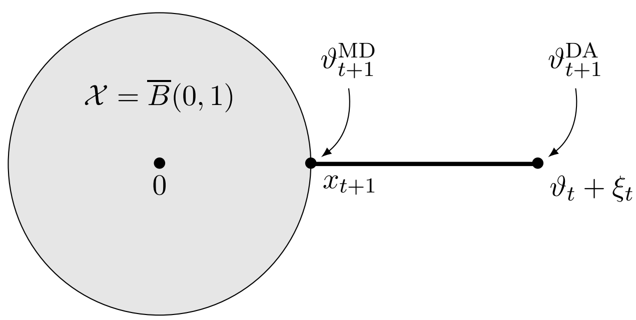


FIGURE 4. Comparison of MD, DA and UMD iterations when $\mathcal{X}$ is the Euclidean ball.

4.2.2. *Segment in* $\mathbb{R}^2$. We here assume the ambiant space to be $\mathbb{R}^2$ and consider $\mathcal{X} = [0,1] \times \{0\}$. Let $t \geqslant 1$ and assume that $\vartheta_t + \xi_t$ belongs to $\mathbb{R}_-^* \times \mathbb{R}$ so that x_{t+1}, the Euclidean projection of $\vartheta_t + \xi_t$, is equal to $(0,0)$. Then, we can see that the set of points ϑ_{t+1} which have $x_{t+1} = (0,0)$ as Euclidean projection is $\mathbb{R}_- \times \mathbb{R}$ and that the set of points ϑ_{t+1} satisfying condition (II) are those with first coordinate greater or equal than the first coordinatcoordinate of $\vartheta_t + \xi_t$. The set of vectors satisfying both conditions (II) and (I) is represented by the dashed area in Figure 5.

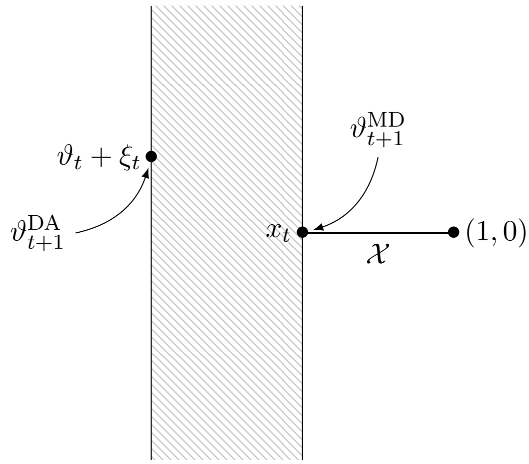


FIGURE 5. Comparison of MD, DA and UMD iterations when $\mathcal{X}$ is a segment in $\mathbb{R}^2$.

4.3. **Analysis.** We introduce a natural extension of the Bregman divergence, which utilizes subgradients instead of gradients. This notion will be central in the statement and the analysis of the guarantee provided by UMD iterates.

Definition 4.7 (Bregman divergence)**.** Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function. For $x \in \mathbb{R}^n$ such that $\partial g(x) \neq \varnothing$, $x' \in \mathbb{R}^n$, and $\vartheta \in \partial g(x)$, we define the Bregman divergence from x to x' with subgradient ϑ as

$$D_g(x', x;\ \vartheta) := h(x') - h(x) - \left\langle \vartheta \middle| x' - x \right\rangle .$$

Remark 4.8. If g is differentiable at point x, the *traditional* Bregman divergence from x to x' is well-defined and is equal to the the only the Bregman divergence (as defined above) from x to x' with (only) subgradient $\nabla g(x)$, in other words: $D_g(x', x;\ \nabla g(x)) = D_g(x', x)$.

Proposition 4.9. *Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower-semicontinuous convex function. Let $x, x', \vartheta, \vartheta' \in \mathbb{R}^n$ such that $\vartheta \in \partial g(x)$ and $\vartheta' \in \partial g(x')$.*

(i) Then,

$$0 \leqslant D_g(x', x;\ \vartheta) = D_{g^*}(\vartheta, \vartheta';\ x'),$$

where $D_{g^*}(\,\cdot\,, \cdot;\ \cdot)$ is the Bregman divergence associated with the Legendre–Fenchel transform g^* of g.

(ii) Moreover, if g is K-strongly convex with respect to a given norm $\|\cdot\|$, g^* is differentiable on $\mathbb{R}^n$ and:

$$\frac{K}{2}\left\|x' - x\right\|^2 \leqslant D_g(x', x;\ \vartheta) = D_{g^*}(\vartheta, \vartheta') \leqslant \frac{1}{2K}\left\|\vartheta - \vartheta'\right\|_*^2 .$$

Proof. (i) The nonnegativity is satisfied because, by simply using Definition 4.7, it can be seen to be equivalent to the convexity of g. Using the Fenchel identity (property (iii) form Proposition A.2), we write

$$\begin{aligned} D_g(x', x;\ \vartheta) &= g(x') - g(x) - \left\langle \vartheta \middle| x' - x \right\rangle \\ &= \left\langle \vartheta' \middle| x' \right\rangle - g^*(\vartheta') - \langle \vartheta | x \rangle + g^*(\vartheta) - \left\langle \vartheta \middle| x' - x \right\rangle \\ &= g^*(\vartheta) - g^*(\vartheta') - \left\langle \vartheta - \vartheta' \middle| x' \right\rangle = D_{g^*}(\vartheta, \vartheta';\ x'). \end{aligned}$$

(ii) The differentiability of g^* and the second inequality is given by [50, Lemma 15]. For the first inequality, we refer to [50, Lemma 13].

□

The idea of analysing mirror descent and dual averaging with the help a generalized version of the Bregman divergence appears in [26] but differs from our work in several ways: the Bregman divergence is defined with directional derivatives (which are unique), and does not lead to an unification of both families of algorithms.

We now establish the following fundamental inequalities, which are the extention to UMD of a classical tool due to [13]. They will be operational in the analysis of the various applications of UMD presented in Section 5.

Lemma 4.10. *Let h be a $\mathcal{X}$-regularizer, $\xi := (\xi_t)_{t\geqslant 1}$ a sequence in $\mathbb{R}^n$, and $(x_t, \vartheta_t)_{t\geqslant 1}$ a sequence of UMD(h, ξ) iterates. Then, for all $x \in \operatorname{dom} h$ and $t \geqslant 1$,*

$$\langle \xi_t | x - x_{t+1} \rangle \leqslant D_h(x, x_t;\ \vartheta_t) - D_h(x, x_{t+1};\ \vartheta_{t+1}) - D_h(x_{t+1}, x_t;\ \vartheta_t), \tag{3}$$

$$\langle \xi_t | x - x_t \rangle \leqslant D_h(x, x_t;\ \vartheta_t) - D_h(x, x_{t+1};\ \vartheta_t) + D_{h^*}(\vartheta_t + \xi_t, \vartheta_t). \tag{4}$$

Proof. Let $x \in \operatorname{dom} h$ and $t \geqslant 1$. Using variational inequality (II) from the definition of the UMD iterates, we write:

$$\begin{aligned}
\langle \xi_t | x - x_{t+1} \rangle &\leqslant \langle \vartheta_{t+1} - \vartheta_t | x - x_{t+1} \rangle \\
&= \langle \vartheta_{t+1} | x - x_{t+1} \rangle - \langle \vartheta_t | x - x_t \rangle + \langle \vartheta_t | x_{t+1} - x_t \rangle \\
&= (h(x) - h(x_t) - \langle \vartheta_t | x - x_t \rangle) + (h(x) - h(x_{t+1}) - \langle \vartheta_{t+1} | x - x_{t+1} \rangle) \\
&\qquad - (h(x_{t+1}) - h(x_t) - \langle \vartheta_t | x_{t+1} - x_t \rangle) \\
&= D_h(x, x_t;\ \vartheta_t) - D_h(x, x_{t+1};\ \vartheta_{t+1}) - D_h(x_{t+1}, x_t;\ \vartheta_t).
\end{aligned}$$

The above Bregman divergences are indeed well-defined because $\vartheta_t \in \partial h(x_t)$ and $\vartheta_{t+1} \in \partial h(x_{t+1})$ as a consequence of the definition of UMD iterates (property (i) from Proposition 4.2).

To prove (4), we note that

$$\langle \xi_t | x_{t+1} - x_t \rangle = D_h(x_{t+1}, x_t;\ \vartheta_t) + D_h(x_t, x_{t+1};\ \vartheta_t + \xi_t),$$

where the second Bregman divergence is well-defined because $\vartheta_t + \xi_t \in \partial h(x_{t+1})$ according to property (ii) from Proposition 4.2. Moreover, we have

$$D_h(x_t, x_{t+1};\ \vartheta_t + \xi_t) = D_{h^*}(\vartheta_t + \xi_t, \vartheta_t;\ x_t) = D_{h^*}(\vartheta_t + \xi_t, \vartheta_t),$$

where the first equality comes from Proposition 4.9–(i) and the second equality from the differentiability of h^*. Combining the two previous displays and adding to (3) gives the result. □

An immediate consequence of Lemma 4.10 and property (ii) of Proposition 4.9 is the following celebrated inequality (sometimes called *regret bound*), which extends and unifies classical guarantees on MD and DA—see e.g. [48, Proposition 11], [51, Lemma 2.20], [9, Theorems 5.2 & 5.4].

Corollary 4.11. *Let h be a $\mathcal{X}$-regularizer which we assume to be K-strongly convex with respect to some norm $\|\cdot\|$, and $\xi := (\xi_t)_{t\geqslant 1}$ a sequence in $\mathbb{R}^n$. Let $(x_t, \vartheta_t)_{t\geqslant 1}$ be a sequence of UMD(h, ξ) iterates. Then for $T \geqslant 1$ and $x \in \operatorname{dom} h$,*

$$\sum_{t=1}^{T} \langle \xi_t | x - x_t \rangle \leqslant D_h(x, x_1;\ \vartheta_1) - D_h(x, x_{T+1};\ \vartheta_{T+1}) + \frac{1}{2K} \sum_{t=1}^{T} \|\xi_t\|_*^2. \tag{5}$$

5. Applications

This section illustrates how the UMD algorithm can be used as a building block for various (online) optimization problems. In the examples presented below, the analysis of the algorithms will always make use of Lemma 4.10.

5.1. **UMD for nonsmooth convex optimization.** Let $M > 0$, $\mathcal{X} \subset \mathbb{R}^n$ be closed and convex, $\|\cdot\|$ be a norm on $\mathbb{R}^n$, and $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ a convex function such that $\mathcal{X} \subset \operatorname{dom} f$. We assume that f has nonempty subdifferential on $\mathcal{X}$ with bounded subgradients:

$$\forall x \in \mathcal{X},\ \forall \xi \in \partial f(x), \quad \|\xi\|_* \leqslant M.$$

We suppose that the optimization problem

$$f_* = \min_{x \in \mathcal{X}} f(x) \tag{6}$$

admits a minimizer $x_* \in \mathcal{X}$.

We now define UMD iterates for solving (6). Let h be an $\mathcal{X}$-regularizer and $(\gamma_t)_{t \geqslant 1}$ be a sequence of positive *step-sizes*. We consider $(x_t, \vartheta_t)_{t \geqslant 1}$ a sequence of UMD(h, ξ) with associated dual increments $\xi := (-\gamma_t f'(x_t))_{t \geqslant 1}$ where $f'(x_t)$ is a subgradient in $\partial f(x_t)$. In other words, $\vartheta_1 \in \partial h(x_1)$ and for $t \geqslant 1$:

$$(x_{t+1}, \vartheta_{t+1}) \in \Pi_h(\vartheta_t - \gamma_t f'(x_t)).$$

The following result provides accuracy estimates for approximate solutions x'_T by UMD after T iterations, computed either as

$$x'_T = \frac{\sum_{t=1}^T \gamma_t x_t}{\sum_{t=1}^T \gamma_t} \qquad \text{or as} \qquad x'_T \in \operatorname*{arg\,min}_{t=1,\dots,T} f(x_t).$$

In particular, it recovers known guarantees for mirror descent [5, Theorem 4.1] and dual averaging [45, Theorem 1] algorithms.

Proposition 5.1. *We assume h to be K-strongly convex with respect to $\|\cdot\|$ and that $x_* \in \operatorname{dom} h$. For $T \geqslant 1$, one has*

$$f(x'_T) - f_* \leqslant \frac{D_h(x_*, x_1;\ \vartheta_1) + \frac{M^2}{2K} \sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t}.$$

For instance, with constant step-sizes

$$\gamma_t = \gamma := \frac{\Omega_{\mathcal{X}}}{M\sqrt{T}}, \qquad t \geqslant 1,$$

where $\Omega_{\mathcal{X}}$ is an upper estimate of $\sqrt{2D_h(x_*, x_1;\ \vartheta_1)}$[1]*, the following holds:*

$$f(x'_T) - f_* \leqslant \frac{\Omega_{\mathcal{X}} M}{\sqrt{T}}.$$

Proof. Applying inequality (5) from Corollary 4.11 with $x = x_*$ gives:

$$\sum_{t=1}^T \gamma_t \left\langle f'(x_t) \middle| x_t - x_* \right\rangle \leqslant D_h(x_*, x_1;\ \vartheta_1) + \frac{M^2}{2K} \sum_{t=1}^T \gamma_t^2.$$

On the other hand, by convexity of f,

$$\sum_{t=1}^T \gamma_t \left\langle f'(x_t) | x_t - x_* \right\rangle \geqslant \sum_{t=1}^T \gamma_t (f(x_t) - f(x_*)) \geqslant \left(\sum_{t=1}^T \gamma_t \right) (f(x'_T) - f(x_*)),$$

and the result follows. $\square$

[1]In the case of compact $\mathcal{X}$ one can take $\Omega_X = [\max_{x \in \mathcal{X}} 2D_h(x, x_1;\ \vartheta_1)]^{1/2}$. Note that in this case due to strong convexity of $D_h(\cdot, x_1, \vartheta_1)$ one has $\Omega_{\mathcal{X}} \geqslant \max_{x \in \mathcal{X}} \|x - x_1\|$.

5.2. **A quasi-monotone UMD for nonsmooth convex optimization.** We consider the same problem as in Section 5.1. Similarly to [37, 46], we construct an algorithm which guarantees the same convergence as in Proposition 5.1 but for the last iterate.

Let function f and regularizer h satisfy the same assumptions as in Section 5.1 and $(\gamma_t)_{t\geqslant 1}$ be a sequence of positive step-sizes. We define iterates $(x_t, y_t, \vartheta_t)_{t\geqslant 1}$ according to $x_1 = y_1$, $\vartheta_1 \in \partial h(x_1)$ and for $t \geqslant 1$:

$$\begin{aligned}(x_{t+1}, \vartheta_{t+1}) &\in \Pi(\vartheta_t - \gamma_t f'(y_t))\\ y_{t+1} &= (1-\nu_t)y_t + \nu_t x_{t+1}.\end{aligned}$$

where $f'(y_t)$ is some subgradient in $\partial f(y_t)$ and where coefficient $\nu_t \in (0,1)$ is given for $t \geqslant 1$ by:

$$\nu_t = \gamma_{t+1}\left(\sum_{s=1}^{t+1}\gamma_s\right)^{-1}.$$

Note that $(x_t, \vartheta_t)_{t\geqslant 1}$ are UMD(h,ξ) iterates, where $\xi := (-\gamma_t f'(y_t))_{t\geqslant 1}$.

Proposition 5.2. *We assume h to be K-strongly convex with respect to $\|\cdot\|$ and that $x_* \in \operatorname{dom} h$. For $T \geqslant 1$,*

$$f(y_T) - f_* \leqslant \frac{D_h(x_*, x_1;\ \vartheta_1) + \frac{M^2}{2K}\sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t}.$$

Like in Proposition 5.1, choosing constant step-sizes

$$\gamma_t = \gamma := \frac{\Omega_{\mathcal{X}}}{M\sqrt{T}}, \qquad t \geqslant 1$$

where $\Omega_{\mathcal{X}}$ is an upper estimate of $\sqrt{2D_h(x_*, x_1;\ \vartheta_1)}$, we obtain:

$$f(y_T) - f_* \leqslant \frac{\Omega_{\mathcal{X}} M}{\sqrt{T}}.$$

Proof. Let $t \geqslant 2$. It follows from the definition of the iterates that $x_t - y_t = (\nu_{t-1}^{-1} - 1)(y_t - y_{t-1})$. Therefore, using the convexity of f, we can write:

$$\begin{aligned}\left\langle \gamma_t f'(y_t)\middle| x_t - x_*\right\rangle &= \gamma_t \left\langle f'(y_t)\middle| y_t - x_*\right\rangle + \gamma_t \left\langle f'(y_t)\middle| x_t - y_t\right\rangle\\ &= \gamma_t \left\langle f'(y_t)\middle| y_t - x_*\right\rangle + \gamma_t(\nu_{t-1}^{-1} - 1)\left\langle f'(y_t)\middle| y_t - y_{t-1}\right\rangle\\ &\geqslant \gamma_t\left(f(y_t) - f_*\right) + \gamma_t(\nu_{t-1}^{-1}-1)\left(f(y_t) - f(y_{t-1})\right)\\ &= \gamma_t\nu_{t-1}^{-1} f(y_t) - \gamma_t(\nu_{t-1}^{-1} - 1) f(y_{t-1}) - \gamma_t f_*.\end{aligned}$$

Besides, for $t=1$, we simply have $\gamma_1\left\langle f'(y_1)|x_1 - x_*\right\rangle \geqslant \gamma_1(f(y_1) - f_*)$ because $x_1 = y_1$ by definition. Then, summing over $t = 1, \ldots, T$ and simplifying, we get:

$$(\gamma_1 - \gamma_2(\nu_1^{-1} - 1))f(y_1) + \sum_{t=2}^{T-1}(\gamma_t\nu_{t-1}^{-1} - \gamma_{t+1}(\nu_t^{-1} - 1))f(y_t) + \gamma_T\nu_{t-1}^{-1} f(y_T)$$
$$- \left(\sum_{t=1}^T \gamma_t\right) f_* \leqslant \sum_{t=1}^T \left\langle \gamma_t f'(x_t)\middle| x_t - x_*\right\rangle.$$

Using the definition of coefficients ν_t, the above left-hand side simplifies as follows:

$$\left(\sum_{t=1}^T \gamma_t\right)(f(y_T) - f_*) \leqslant \sum_{t=1}^T \left\langle \gamma_t f'(x_t)\middle| x_t - x_*\right\rangle.$$

Finally, because $(x_t, \vartheta_t)_{t\geqslant 1}$ is a sequence of $\mathrm{UMD}(h,\xi)$ iterates where $\xi = (-\gamma_t f'(y_t))_{t\geqslant 1}$, the result then follows by applying inequality (5) from Corollary 4.11 and dividing by $\sum_{t=1}^T \gamma_t$. $\square$

5.3. **Accelerated UMD for smooth convex optimization.** We present a first-order algorithm for the constrained minimization of a smooth convex function (with respect to an arbitrary norm). Like the Nesterov's accelerated method [42], of which it is an extension, the following algorithm achieves a $1/T^2$ convergence rate. The construction of the involves a sequence of UMD iterates, and its analysis makes use of inequality (3) from Lemma 4.10.

Let $\mathcal{X}$ be a closed and convex subset of $\mathbb{R}^n$, $\mathcal{U}$ an open and convex neighborhood of $\mathcal{X}$ and $\|\cdot\|$ a norm on $\mathbb{R}^n$. Let $f : \mathcal{U} \to \mathbb{R}$ be a convex function which we assume to be L-smooth with respect to $\|\cdot\|$ (and therefore differentiable) on $\mathcal{U}$:

$$\left\|\nabla f(x) - \nabla f(x')\right\|_* \leqslant L \left\|x - x'\right\|, \qquad x, x' \in \mathcal{U}.$$

We assume that the optimization problem

$$f_* = \min_{x\in\mathcal{X}} f(x)$$

admits an optimal solution $x_* \in \mathcal{X}$.

Let $K > 0$ and let h be a $\mathcal{X}$-regularizer. Points $(x_t, y_t, z_t, \vartheta_t)_{t\geqslant 1}$ are AUMD iterates if they satisfy $x_1 = y_1 = \nabla h^*(\vartheta_1)$ and for $t \geqslant 1$:

$$(x_{t+1}, \vartheta_{t+1}) \in \Pi_h(\vartheta_t - \gamma_t \nabla f(y_t)) \tag{7a}$$

$$z_{t+1} = y_t + \nu_t(x_{t+1} - x_t) \tag{7b}$$

$$y_{t+1} = (1 - \nu_{t+1}) z_{t+1} + \nu_{t+1} x_{t+1}, \tag{7c}$$

where the coefficients $(\gamma_t)_{t\geqslant 1}$ and $(\nu_t)_{t\geqslant 1}$ are given for $t \geqslant 1$ by:

$$\gamma_1 = K/L, \qquad \gamma_{t+1} = \frac{K}{2L}\left(1 + \sqrt{1 + (2L\gamma_t/K)^2}\right), \qquad \nu_t = \frac{K}{L\gamma_t}.$$

Remark 5.3 (AUMD iterates always exist). It follows from the above definition that $(x_t, \vartheta_t)_{t\geqslant 1}$ is a sequence of $\mathrm{UMD}(h,\xi)$ iterates (associated with dual increments $\xi := (-\gamma_t \nabla f(y_t))_{t\geqslant 1}$). Therefore, such points $(x_t)_{t\geqslant 1}$ do exist as long as gradients $(\nabla f(y_t))_{t\geqslant 1}$ exists (see below). Applying Lemma 4.10 to the UMD iterates $(x_t, \vartheta_t)_{t\geqslant 1}$ immediately gives Lemma 5.7 below. Besides, one can easily check that for $t \geqslant 1$, point z_{t+1} can we written as a convex combination of z_t and x_{t+1}, specifically: $z_{t+1} = (1-\nu_t) z_t + \nu_t x_{t+1}$. Since x_{t+1} belongs to $\mathcal{X}$ by construction, an immediate induction proves that z_t belongs to $\mathcal{X}$ for all $t \geqslant 1$. Then, y_t being by construction a convex combination of z_t and x_t, it also belongs to $\mathcal{X}$. Since f is differentiable on $\mathcal{U} \supset \mathcal{X}$, $\nabla f(y_t)$ indeed exists. Therefore, AUMD iterates always exist.

Remark 5.4 (Bibliographic remarks). Accelerated versions of mirror descent [30] and dual averaging [43] are both special cases of our AUMD algorithm.

The accelerated version of DA from [43, Section 3] is slightly different from our version, as the points $(z_t)_{t\geqslant 1}$ are updated as:

$$z_{t+1} = \operatorname*{arg\,min}_{z\in\mathcal{X}} \left\{ f(y_t) + \langle \nabla f(y_t) | z - y_t \rangle + \frac{L}{2}\left\|z - y_t\right\|^2 \right\}, \tag{8}$$

Therefore, our version (while providing the same convergence guarantee) has a computational advantage since update (7b) is much less demanding computationnaly than the above MD-style update (8).

The idea of emphasizing—both in the presentation and the analysis of the algorithm—the presence of a sequence of mirror descent iterates is borrowed from [1, 43].

Theorem 5.5. *We assume h to be K-strongly convex with respect to $\|\cdot\|$ and that $x_* \in \operatorname{dom} h$. For $T \geqslant 1$, AUMD iterates defined as above guarantee:*

$$f(z_{T+1}) - f_* \leqslant \frac{4LD_h(x_*, x_1;\ \vartheta_1)}{KT^2}.$$

We give without proof the following lemma which gathers a few immediate consequences of the above definitions.

Lemma 5.6. *For $t \geqslant 1$,*
(i) $\nu_t \in (0, 1)$,
(ii) $y_t - x_t = (\nu_t^{-1} - 1)(z_t - y_t)$,
(iii) $\gamma_t \nu_t^{-1} - \gamma_{t+1}(\nu_{t+1}^{-1} - 1) = 0$,
(iv) For $T \geqslant 1$, $\gamma_T \nu_T^{-1} = \sum_{t=1}^{T} \gamma_t \geqslant \frac{KT^2}{4L}$.

Sequence $(x_t, \vartheta_t)_{t \geqslant 1}$ being UMD iterates, the following lemma follows from summing inequality (3) from Lemma 4.10.

Lemma 5.7. *For $T \geqslant 1$,*

$$\sum_{t=1}^{T} \gamma_t \langle \nabla f(y_t) | x_{t+1} - x_* \rangle \leqslant D_h(x_*, x_1;\ \vartheta_1) - \sum_{t=1}^{T} D_h(x_{t+1}, x_t;\ \vartheta_t).$$

Proof of Theorem 5.5. Let $t \geqslant 1$. We begin by expressing the smoothness of f between points y_t and z_{t+1} (see e.g. [31, Proposition 3]):

$$\begin{aligned} f(z_{t+1}) - f(y_t) &\leqslant \langle \nabla f(y_t) | z_{t+1} - y_t \rangle + \frac{L}{2} \|z_{t+1} - y_t\|^2 \\ &= \nu_t \langle \nabla f(y_t) | x_{t+1} - x_t \rangle + \frac{L\nu_t^2}{2} \|x_{t+1} - x_t\|^2 \\ &\leqslant \nu_t \langle \nabla f(y_t) | x_{t+1} - x_t \rangle + \frac{L\nu_t^2}{K} D_h(x_{t+1}, x_t;\ \vartheta_t), \end{aligned}$$

where we used relation (7b) from the definition of the algorithm to get the second line, and the K-strong convexity of h (Proposition 4.9) to get the third line. Multiplying by $L\gamma_t^2/K$ and simplifying gives:

$$\gamma_t \nu_t^{-1}(f(z_{t+1}) - f(y_t)) \leqslant \gamma_t \langle \nabla f(y_t) | x_{t+1} - x_t \rangle + D_h(x_{t+1}, x_t;\ \vartheta_t). \tag{9}$$

Besides, we can write

$$\begin{aligned} x_{t+1} - x_t &= (x_{t+1} - x_*) + (x_* - y_t) + (y_t - x_t) \\ &= (x_{t+1} - x_*) + (x_* - y_t) + (\nu_t^{-1} - 1)(z_t - y_t), \end{aligned}$$

where the second line uses relation (ii) from Lemma 5.6. Injecting the above into $\langle \nabla f(y_t) | x_{t+1} - x_t \rangle$ gives:

$$\begin{aligned} \langle \nabla f(y_t) | x_{t+1} - x_t \rangle &= \langle \nabla f(y_t) | x_{t+1} - x_* \rangle + \langle \nabla f(y_t) | x_* - y_t \rangle + (\nu_t^{-1} - 1) \langle \nabla f(y_t) | z_t - y_t \rangle \\ &\leqslant \langle \nabla f(y_t) | x_{t+1} - x_* \rangle + f(x_*) - f(y_t) + (\nu_t^{-1} - 1)\left(f(z_t) - f(y_t) \right). \end{aligned} \tag{10}$$

Combining inequalities (9) and (10), and summing over $t = 1, \dots, T$ gives:

$$\begin{aligned} \sum_{t=1}^{T} \gamma_t \nu_t^{-1} \left(f(z_{t+1}) - f(y_t) \right) \leqslant D_h(x_*, x_1;\ \vartheta_1) &- \sum_{t=1}^{T} D_h(x_{t+1}, x_t;\ \vartheta_t) + \sum_{t=1}^{T} \gamma_t \left(f(x_*) - f(y_t) \right) \\ &+ \sum_{t=1}^{T} \gamma_t (\nu_t^{-1} - 1) \left(f(z_t) - f(y_t) \right) + \sum_{t=1}^{T} D_h(x_{t+1}, x_t;\ \vartheta_t), \end{aligned}$$

where we used Lemma 5.7 to get the first two terms of the right-hand side. Then, simplifying and moving all values of f (except for $f(x_*)$) to the left-hand side, we get:

$$\sum_{t=1}^{T}\left(\gamma_t+\gamma_t(\nu_t^{-1}-1)-\gamma_t\nu_t^{-1}\right)f(y_t)+\sum_{t=2}^{T}\left(\gamma_{t-1}\nu_{t-1}^{-1}-\gamma_t(\nu_t^{-1}-1)\right)f(z_t)$$
$$+\gamma_1(\nu_1^{-1}-1)f(z_1)+\gamma_T\nu_T^{-1}f(z_{T+1})\leqslant D_h(x_*,x_1;\ \vartheta_1)+\left(\sum_{t=1}^{T}\gamma_t\right)f(x_*).$$

The factor in front of $f(y_t)$ is clearly zero, as well as $\gamma_1(\nu_1^{-1}-1)$. The result then follows by applying properties (iii) and (iv) from Lemma 5.6. □

5.4. **Solving variational inequalities.** Let $\mathcal{X}\subset\mathbb{R}^n$ be closed and convex and $\Phi:\mathcal{X}\to\mathbb{R}^n$ a *monotone* operator:

$$\forall x,x'\in\mathcal{X},\quad \left\langle\Phi(x')-\Phi(x)\middle|x'-x\right\rangle\geqslant 0.$$

A point $x_*\in\mathcal{X}$ is a (weak) *solution* of the variational inequality associated with Φ if it satisfies:

$$\forall x\in\mathcal{X},\quad \langle\Phi(x)|x_*-x\rangle\leqslant 0. \tag{11}$$

The is goal is to construct algorithms which outputs approximate solutions. This framework is a generalization of convex optimization and contains various problems such as convex-concave saddle-point problems and convex Nash equilibrium problems—see e.g. [27, 39, 44].

Remark 5.8 (Weak and strong solutions). Since Φ is monotone, condition (11) is implied by $\langle\Phi(x_*)|x-x_*\rangle\geqslant 0$ for all $x\in\mathcal{X}$, which is the standard definition of a (strong) solution associated with Φ. The converse—a weak solution as defined by (11) is a strong solution as well—is also true, provided, e.g., that Φ is continuous. An advantage of the concept of weak solution is that such a solution always exists as soon as $\mathcal{X}$ is assumed to be compact.

Let h be a $\mathcal{X}$-regularizer and $\gamma>0$ a step-size. We say that points $(x_t,y_t,\vartheta_t,\zeta_t)_{t\geqslant 1}$ are associated *unified mirror prox* (UMP) iterates if they satisfy $x_1=\nabla h^*(\vartheta_1)$ and for $t\geqslant 1$:

$$\zeta_t\in\partial h(x_t) \tag{12a}$$
$$\forall x\in\mathcal{X},\ \langle\zeta_t-\vartheta_t|x-x_t\rangle\geqslant 0 \tag{12b}$$
$$y_t=\nabla h^*(\zeta_t-\gamma\Phi(x_t)) \tag{12c}$$
$$(x_{t+1},\vartheta_{t+1})\in\Pi_h(\vartheta_t-\gamma\Phi(y_t)). \tag{12d}$$

Note that the above definition implies that $(x_t,\vartheta_t)_{t\geqslant 1}$ is a sequence of UMD(h,ξ) iterates where $\xi=(-\gamma\Phi(y_t))_{t\geqslant 1}$.

Remark 5.9 (Mirror prox [39] is special case of UMP). The original mirror prox algorithm [39] associated with a $\mathcal{X}$-compatible mirror map F is defined as follows. Let $x_1\in\mathcal{X}\cap\mathcal{D}_F$ be an initial point and for $t\geqslant 1$:

$$y_t=\operatorname*{arg\,min}_{x\in\mathcal{X}}D_F(x,\nabla F^*(\nabla F(x_t)-\gamma\Phi(x_t))) \tag{13a}$$
$$x_{t+1}=\operatorname*{arg\,min}_{x\in\mathcal{X}}D_F(x,\nabla F^*(\nabla F(x_t)-\gamma\Phi(y_t))). \tag{13b}$$

By setting $\vartheta_t:=\zeta_t:=\nabla F(x_t)$, we can easily check that points $(x_t,y_t,\vartheta_t,\zeta_t)_{t\geqslant 1}$ satisfy above conditions (12a)–(12d) with regularizer $h:=F+I_{\mathcal{X}}$ and are therefore UMP iterates.

Remark 5.10 (Nesterov's dual extrapolation [44] is a special case of UMP). Let F be an $\mathcal{X}$-compatible mirror map and $h := F + I_{\mathcal{X}}$ the associated $\mathcal{X}$-regularizer. Then, the dual extrapolation algorithm is defined for $t \geqslant 1$ as [44, Section 3]:

$$x_t = \nabla h^*(\vartheta_t) \tag{14a}$$
$$y_t = \operatorname*{arg\,min}_{x \in \mathcal{X}} D_F(x, \nabla F^*(\nabla F(x_t) - \gamma \Phi(x_t))) \tag{14b}$$
$$\vartheta_{t+1} = \vartheta_t - \gamma \Phi(y_t). \tag{14c}$$

Then, one can check that considering $\zeta_t := \nabla F(x_t)$ makes the above satisfy conditions (12a)–(12d) from the definition of UMP. Moreover, Theorem 5.11 below generalizes the original guarantee [44, Theorem 2].

Theorem 5.11. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. We assume monotone operator $\Phi : \mathcal{X} \to \mathbb{R}^n$ to be L-Lipschitz continuous with respect to the following norms:*

$$\left\|\Phi(x) - \Phi(x')\right\|_* \leqslant L \left\|x - x'\right\|, \qquad x, x' \in \mathcal{X},$$

regularizer h to be K-strongly convex with respect to $\|\cdot\|$. Denote $\overline{y}_t = \frac{1}{T}\sum_{t=1}^T y_t$ the average of points y_t from some unified mirror prox iterates associated with Φ and parameter $\gamma \leqslant K/L$. Then, for $T \geqslant 1$, the following guarantee holds:

$$\forall x \in \operatorname{dom} h, \quad \langle \Phi(x) | \overline{y}_T - x \rangle \leqslant \frac{D_h(x, x_1;\ \vartheta_1)}{\gamma T}.$$

Proof. Let $x \in \operatorname{dom} h$ and $T \geqslant 1$. Using the definition of $\overline{y}_T$ and the monotonicity of Φ, we write:

$$\begin{aligned}
\langle \gamma\Phi(x) | \overline{y}_T - x \rangle &= \frac{1}{T}\sum_{t=1}^T \langle \gamma\Phi(x) | y_t - x \rangle \leqslant \sum_{t=1}^T \langle \gamma\Phi(y_t) | y_t - x \rangle \\
&= \sum_{t=1}^T \left(\langle \gamma\Phi(y_t) | x_{t+1} - x \rangle + \langle \gamma\Phi(y_t) | y_t - x_{t+1} \rangle\right) \\
&\leqslant D_h(x, x_1;\ \vartheta_1) + \sum_{t=1}^T \left(-D_h(x_{t+1}, x_t;\ \vartheta_t) + \langle \gamma\Phi(y_t) | y_t - x_{t+1} \rangle\right),
\end{aligned}$$

where the second inequality comes from summing inequality (3) from Lemma 4.10 (because $(x_t, \vartheta_t)_{t \geqslant 1}$ is a sequence of UMD iterates as we noticed). We bound the above last two terms as follows. Let $t \geqslant 1$ and denote δ_t the content of the above last sum and let us bound it as follows.

$$\begin{aligned}
\delta_t &:= -D_h(x_{t+1}, x_t;\ \vartheta_t) - \langle \gamma\Phi(y_t) | x_{t+1} - y_t \rangle \\
&= -h(x_{t+1}) + h(x_t) + \langle \vartheta_t | x_{t+1} - x_t \rangle + \langle \gamma(\Phi(x_t) - \Phi(y_t)) | x_{t+1} - y_t \rangle - \langle \gamma\Phi(x_t) | x_{t+1} - y_t \rangle.
\end{aligned} \tag{15}$$

Condition (12b) from the definition of UMP iterates gives:

$$\langle \vartheta_t | x_{t+1} - x_t \rangle \leqslant \langle \zeta_t | x_{t+1} - x_t \rangle. \tag{16}$$

Besides, using basic inequality $\langle y | x \rangle \leqslant \frac{1}{2}\|y\|_*^2 + \frac{1}{2}\|x\|^2$, we can write:

$$\begin{aligned}
\langle \gamma(\Phi(x_t) - \Phi(y_t)) | x_{t+1} - y_t \rangle &\leqslant \frac{\gamma^2}{2K}\|\Phi(x_t) - \Phi(y_t)\|_*^2 + \frac{K}{2}\|x_{t+1} - y_t\|^2 \\
&\leqslant \frac{\gamma^2 L^2}{2K}\|x_t - y_t\|^2 + \frac{K}{2}\|x_{t+1} - y_t\|^2,
\end{aligned} \tag{17}$$

where we used the Lipschitz continuity of operator Φ. Injecting (16) and (17) into (15) and simplifying gives

$$\delta_t \leqslant -h(x_{t+1}) + h(x_t) + \langle \zeta_t | y_t - x_t \rangle + \langle \zeta_t - \gamma\Phi(x_t) | x_{t+1} - y_t \rangle + \frac{\gamma^2 L^2}{2K} \|y_t - x_t\|_*^2 + \frac{K}{2} \|x_{t+1} - y_t\|^2 . \tag{18}$$

First, we have $\zeta_t \in \partial h(x_t)$ thanks to condition (12a) and $D_h(y_t, x_t;\ \zeta_t)$ is well-defined. Besides, thanks to Proposition A.2, condition (12c) is equivalent to $\zeta_t - \gamma\Phi(x_t) \in \partial h(y_t)$, and $D_h(x_{t+1}, y_t;\ \zeta_t - \gamma\Phi(x_t))$ is thus well-defined. We can make those two generalized Bregman divergences appear in the above right-hand side, which is consequently equal to:

$$\delta_t \leqslant \frac{\gamma^2 L^2}{2K} \|y_t - x_t\|^2 - D_h(y_t, x_t;\ \zeta_t) + \frac{K}{2} \|x_{t+1} - y_t\|^2 - D_h(x_{t+1}, y_t;\ \zeta_t - \gamma\Phi(x_t)).$$

Using the K-strong convexity of h (Proposition 4.9) and the fact that $\gamma \leqslant K/L$ by assumption, the above simplifies to $\delta_t \leqslant 0$. The result follows. □

5.5. **Regret minimization.** Like the MD and DA algorithms, the UMD algorithm can also be used for online optimization. Let us recall the classical problem called *online linear optimization* [7, 9, 11, 22, 23, 50, 51]. We assume $\mathcal{X}$ to be a convex compact subset of $\mathbb{R}^n$. The problem is a repeated game between a *decision maker* and *Nature* and is played as follows. For each stage $t \geqslant 1$:

- the decision maker chooses an *action* $x_t \in \mathcal{X}$,
- Nature chooses an action $\zeta_t \in \mathbb{R}^n$,
- decision maker gets *payoff* $\langle \zeta_t | x_t \rangle$.

The goal of the decision maker is to minimize its *regret* R_T (where $T \geqslant 1$ is the number of stages), which is defined as the difference between its cumulative payoff and the cumulative of the best constant choice of action in hindsight:

$$R_T := \max_{x \in \mathcal{X}} \sum_{t=1}^{T} \langle \zeta_t | x \rangle - \sum_{t=1}^{T} \langle \zeta_t | x_t \rangle .$$

The decision maker can choose its actions according to the UMD algorithm to minimize the regret. The following result is an immediate consequence of Corollary 4.11 and extends classical guarantees—see e.g. [48, Proposition 11], [Lemma 2.20 51] and [9, Theorem 5.4].

Theorem 5.12. *Let $M > 0$ and $\eta > 0$. Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$, h a $\mathcal{X}$-regularizer which we assume to be K-strongly convex with respect to $\|\cdot\|$ and such that $\operatorname{dom} h = \mathcal{X}$, and assume (x_t, ϑ_t) to be a sequence of UMD(h, ξ) iterates where $\xi := (\eta\zeta_t)_{t \geqslant 1}$. We assume that $\|\zeta_t\|_* \leqslant M$ for all $t \geqslant 1$. Then, for all $T \geqslant 1$,*

$$R_T \leqslant \frac{\Omega_{\mathcal{X}}}{\eta} + \frac{\eta M T}{2K},$$

where $\Omega_{\mathcal{X}} := \max_{x \in \mathcal{X}} D_h(x, x_1;\ \vartheta_1)$. In particular, the choice $\eta = \sqrt{2K\Omega_{\mathcal{X}}/(MT)}$ gives

$$R_T \leqslant \sqrt{\frac{2MT\Omega_{\mathcal{X}}}{K}}.$$

Acknowledgements

The authors are grateful to Roberto Cominetti, Cristóbal Guzmán, Nicolas Flammarion and Sylvain Sorin for inspiring discussions and suggestions. The second author was supported by a public grant as part of the Investissement d'avenir project, reference `ANR-11-LABX-0056-LMH`, LabEx LMH.

References

[1] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *ITCS*, 2017.

[2] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, pages 217–226, 2009.

[3] Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research*, 11:2785–2836, 2010.

[4] Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2013.

[5] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[6] Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[7] Sébastien Bubeck. *Introduction to Online Optimization: Lecture Notes*. Princeton University, 2011.

[8] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[9] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, 5(1):1–122, 2012.

[10] Sébastien Bubeck, Nicolò Cesa-Bianchi, and Sham M. Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *JMLR: Workshop and Conference Proceedings (COLT)*, volume 23, pages 41.1–41.14, 2012.

[11] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[12] Nicolò Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352, 2007.

[13] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[14] Alon Cohen, Tamir Hazan, and Tomer Koren. Tight bounds for bandit combinatorial optimization. *Proceedings of Machine Learning Research (COLT 2017)*, 65:1–14, 2017.

[15] Bruce Cox, Anatoli Juditsky, and Arkadi Nemirovski. Dual subgradient algorithms for large-scale nonsmooth learning problems. *Mathematical Programming*, 148(1-2):143–180, 2014.

[16] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(Jan):165–202, 2012.

[17] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[18] John Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *arXiv preprint arXiv:1612.05612*, 2016.

[19] John C. Duchi, Alekh Agarwal, and Martin J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2012.

[20] Nicolas Flammarion and Francis Bach. Stochastic composite least-squares regression with convergence rate O(1/n). *Proceedings of Machine Learning Research (COLT 2017)*, 65:1–44, 2017.

[21] Alan A. Goldstein. Convex programming in Hilbert space. *Bulletin of the American Mathematical Society*, 70(5):709–710, 1964.

[22] Elad Hazan. The convex optimization approach to regret minimization. In S. Nowozin S. Sra and S. Wrigh, editors, *Optimization for Machine Learning*, pages 287–303. MIT press, 2012.

[23] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[24] Elad Hazan and Satyen Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80(2-3):165–188, 2010.

[25] Elad Hazan, Alexander Rakhlin, and Peter L Bartlett. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems*, pages 65–72, 2008.

[26] Pooria Joulani, András György, and Csaba Szepesvári. A modular analysis of adaptive (non-) convex optimization: Optimism, composite objectives, and variational bounds. *Proceedings of Machine Learning Research (ALT 2017)*, 1:40, 2017.

[27] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[28] Anatoli Juditsky, Philippe Rigollet, Alexandre B Tsybakov, et al. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.

[29] Anatoli B. Juditsky, Alexander V. Nazin, Alexandre B. Tsybakov, and Nicolas Vayatis. Recursive aggregation of estimators by the mirror descent algorithm with averaging. *Problems of Information Transmission*, 41(4):368–384, 2005.

[30] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in neural information processing systems*, pages 2845–2853, 2015.

[31] Joon Kwon and Panayotis Mertikopoulos. A continuous-time approach to online optimization. *Journal of Dynamics and Games*, 4(2):125–148, 2017.

[32] Sangkyun Lee and Stephen J. Wright. Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research*, 13(Jun):1705–1744, 2012.

[33] Evgeny S. Levitin and Boris T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1–50, 1966.

[34] Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 525–533, 2011.

[35] H. Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *The Journal of Machine Learning Research*, 18(1):3117–3166, 2017.

[36] H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Conference on Learning Theory (COLT)*, page 244, 2010.

[37] Alexander V. Nazin. Algorithms of inertial mirror descent in convex problems of stochastic optimization. *Autom. Remote Control*, 79(1):78–88, 2018.

[38] Arkadi Nemirovski. Efficient methods for large-scale convex optimization problems. *Ekonomika i Matematicheskie Metody*, 15, 1979.

[39] Arkadi Nemirovski. Prox-method with rate of convergence O(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[40] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[41] Arkadi Nemirovski and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.

[42] Yurii Nesterov. A method of solving a convex programming problem with convergence rate O(k^2). In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences,

1983.
[43] Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.
[44] Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.
[45] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.
[46] Yurii Nesterov and Vladimir Shikhman. Quasi-monotone subgradient methods for nonsmooth convex minimization. *Journal of Optimization Theory and Applications*, 165(3):917–940, 2015.
[47] Francesco Orabona, Koby Crammer, and Nicolò Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3):411–435, 2015.
[48] Alexander Rakhlin and Ambuj Tewari. Lecture notes on online learning. 2009.
[49] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
[50] Shai Shalev-Shwartz. *Online learning: Theory, algorithms, and applications*. PhD thesis, The Hebrew University of Jerusalem, 2007.
[51] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.
[52] Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.
[53] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, 2003.

Appendix A. Convex analysis tools

Definition A.1 (Lower-semicontinuity)**.** A function $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is lower-semicontinuous if for all $c \in \mathbb{R}$, the sublevel set $\{x \in \mathbb{R}^n \,|\, f(x) \leqslant c\}$ is closed.

One can easily check that the sum of two lower-semicontinuous functions is lower-semicontinuous. Continuous functions and characteristic functions $I_{\mathcal{X}}$ of closed sets $\mathcal{X} \subset \mathbb{R}^n$ are examples of lower-semicontinuous functions.

Proposition A.2 (Theorem 23.5 in [49])**.** *Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower-semicontinuous convex function with nonempty domain. Then for all $x, y \in \mathbb{R}^n$, the following statements are equivalent:*

(i) $x \in \partial g^*(y)$;
(ii) $y \in \partial g(x)$;
(iii) $\langle y|x\rangle = g(x) + g^*(y)$;
(iv) $x \in \arg\max_{x' \in \mathbb{R}^n} \{\langle y|x'\rangle - g(x')\}$;
(v) $y \in \arg\max_{y' \in \mathbb{R}^n} \{\langle y'|x\rangle - g^*(y')\}$.

Proposition A.3 (Theorem 25.1 in [49])**.** *Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function and $x \in \mathbb{R}^n$. Then $\partial g(x)$ is a singleton if and only if g is differentiable at x. Then, $\partial g(x) = \{\nabla g(x)\}$.*

Proposition A.4. *Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function, $\mathcal{X} \subset \mathbb{R}^n$ a convex set and $x_0 \in \mathcal{X}$. We assume that g is differentiable in x_0. Then*

$$x_0 \in \arg\min_{x \in \mathcal{X}} g(x) \quad \iff \quad \forall x \in \mathcal{X},\ \langle \nabla g(x_0)|x - x_0\rangle \geqslant 0.$$

Definition A.5 (Strong-convexity)**.** Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, $\|\cdot\|$ be a norm in $\mathbb{R}^n$ and $K > 0$. Function g is said to be K-strongly convex with respect to norm $\|\cdot\|$ if for all $x, x' \in \mathbb{R}^n$ and $\lambda \in [0, 1]$,

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

Appendix B. Postponed proofs

Proof of Proposition 2.2. Let $\vartheta \in \mathbb{R}^n$. By property (iii) from Definition 2.1, there exists $x_1 \in \mathcal{D}_F$ such that $\nabla F(x_1) = \vartheta$. Therefore, function $\varphi_\vartheta : x \mapsto \langle \vartheta | x \rangle - F(x)$ is differentiable at x_1 and $\nabla \varphi_\vartheta(x_1) = 0$. Moreover, φ_ϑ is strictly concave as a consequence of property (i) from Definition 2.1. Therefore, x_1 is the unique maximizer of φ_ϑ and:

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

which proves property (i).

Besides, we have

$$x_1 \in \partial F^*(\vartheta) \iff \vartheta = \nabla F(x_1) \iff x_1 \text{ minimizer of } \phi_\vartheta, \tag{19}$$

where the first equivalence comes from Proposition A.2. Point x_1 being the unique maximizer of φ_ϑ, we have that $\partial F^*(\vartheta)$ is a singleton. In other words, F^* is differentiable in ϑ and

$$\nabla F^*(\vartheta) = x_1 \in \mathcal{D}_F. \tag{20}$$

First, the above (20) proves property (ii). Second, this equality combined with the equality from (19) gives the second identity from property (iv). Third, this proves that $\nabla F^*(\mathbb{R}^n) \subset \mathcal{D}_F$.

It remains to prove the reverse inclusion to get property (iii). Let $x \in \mathcal{D}_F$. By property (ii) from Definition 2.1, F is differentiable in x. Consider

$$\vartheta := \nabla F(x), \tag{21}$$

and all the above holds with this special point ϑ. In particular, $x_1 = x$ by uniqueness of x_1. Therefore (20) gives

$$\nabla F^*(\vartheta) = x, \tag{22}$$

and this proves $\nabla F^*(\mathbb{R}^n) \supset \mathcal{D}_F$ and thus property (iii). Combining (21) and (22) gives the first identity from property (iv). □

Proof of Theorem 2.4. Let $x_0 \in \mathcal{D}_F$. By definition of the mirror map, F is differentiable at x_0. Therefore, $D_F(x, x_0)$ is well-defined for all $x \in \mathbb{R}^n$.

For all real value $\alpha \in \mathbb{R}$, consider the sublevel set $S_\mathcal{X}(\alpha)$ of function $x \mapsto D_F(x, x_0)$ associated with value α and restricted to $\mathcal{X}$:

$$S_\mathcal{X}(\alpha) := \{x \in \mathcal{X} \,|\, D_F(x, x_0) \leqslant \alpha\}.$$

Inheriting properties from F, function $D_F(\,\cdot\,, x_0)$ is lower-semicontinuous and strictly convex: consequentely, the sublevel sets $S_\mathcal{X}(\alpha)$ are closed and convex.

Let us also prove that the sublevel sets $S_\mathcal{X}(\alpha)$ are bounded. For each value $\alpha \in \mathbb{R}$, we write

$$S_\mathcal{X}(\alpha) \subset S_{\mathbb{R}^n}(\alpha) := \{x \in \mathbb{R}^n \,|\, D_F(x, x_0) \leqslant \alpha\}$$

and aim at proving that the latter set is bounded. By contradiction, let us suppose that there exists an unbounded sequence in $S_{\mathbb{R}^n}(\alpha)$: let $(x_k)_{k \geqslant 1}$ be such that $0 < \|x_k - x_0\| \xrightarrow[k \to +\infty]{} +\infty$ and $D_F(x_k, x_0) \leqslant \alpha$ for all $k \geqslant 1$. Using the Bolzano–Weierstrass theorem, there exists $v \neq 0$ and a subsequence $(x_{\phi(k)})_{k \geqslant 1}$ such that

$$\frac{x_{\phi(k)} - x_0}{\left\|x_{\phi(k)} - x_0\right\|} \xrightarrow[k \to +\infty]{} v.$$

The point $x_0+\frac{x_{\phi(k)}-x_0}{\|x_{\phi(k)}-x_0\|}$ being a convex combination of x_0 and $x_{\phi(k)}$, we can write the corresponding convexity inequality for function $D_F(\,\cdot\,, x_0)$:

$$\begin{aligned} D_F\left(x_0 + \lambda_k(x_{\phi(k)} - x_0), x_0\right) &\leqslant (1-\lambda_k) D_F(x_0, x_0) + \lambda_k D_F(x_{\phi(k)}, x_0) \\ &\leqslant \lambda_k \alpha \xrightarrow[k\to+\infty]{} 0, \end{aligned}$$

where we used shorthand $\lambda_k := \left\|x_{\phi(k)} - x_0\right\|^{-1}$. For the first above inequality, we used $D_F(x_0, x_0) = 0$ and that $D_F(x_{\phi(k)}, x_0) \leqslant \alpha$ by definition of $(x_k)_{k\geqslant 1}$. Then, using the lower-semicontinuity of $D_F(\,\cdot\,, x_0)$ and the fact that $x_0 + \lambda_k(x_{\phi(k)} - x_0) \xrightarrow[k\to+\infty]{} x_0 + v$, we have

$$D_F(x_0 + v, x_0) \leqslant \liminf_{k\to+\infty} D_F(x_0 + \lambda_k(x_{\phi(k)} - x_0), x_0) \leqslant \liminf_{k\to+\infty} \lambda_k \alpha = 0.$$

The Bregman divergence of a convex function being nonnegative, the above implies $D_F(x_0+v, x_0) = 0$. Thus, function $D_F(\,\cdot\,, x_0)$ attains its minimum (0) at two different points (at x_0 and at $x_0 + v$): this contradicts its strong convexity. Therefore, sublevel sets $S_{\mathcal{X}}(\alpha)$ are bounded and thus compact.

We now consider the value α_{inf} defined as

$$\alpha_{\text{inf}} := \inf \left\{\alpha \,|\, S_{\mathcal{X}}(\alpha) \neq \varnothing\right\}.$$

In other words, α_{inf} is the infimum value of $D_F(\,\cdot\,, x_0)$ on $\mathcal{X}$, and thus the only possible value for the minimum (if it exists). We know that $\alpha_{\text{inf}} \geqslant 0$ because the Bregman divergence is always nonnegative. From the definition of the sets $S_{\mathcal{X}}(\alpha)$, it easily follows that:

$$S_{\mathcal{X}}(\alpha_{\text{inf}}) = \bigcap_{\alpha > \alpha_{\text{inf}}} S_{\mathcal{X}}(\alpha).$$

Naturally, the sets $S_{\mathcal{X}}(\alpha)$ are increasing in α with respect to the inclusion order. Therefore, $S_{\mathcal{X}}(\alpha_{\text{inf}})$ is the intersection of a nested sequence of nonempty compact sets. It is thus nonempty as well by Cantor's intersection theorem. Consequentely, $D_F(\,\cdot\,, x_0)$ does admit a minimum on $\mathcal{X}$, and the minimizer is unique because of the strong convexity.

Let us now prove that the minimizer $x_* := \arg\min_{x\in\mathcal{X}} D_F(x,\ x_0)$ also belongs to $\mathcal{D}_F$. Let us assume by contradiction that $x_* \in \mathcal{X} \setminus \mathcal{D}_F$. By definition of the mirror map, $\mathcal{X} \cap \mathcal{D}_F$ is nonempty; let $x_1 \in \mathcal{X} \cap \mathcal{D}_F$. The set $\mathcal{D}_F$ being open by definition, there exists $\varepsilon > 0$ such that the closed Euclidean ball $\overline{B}(x_1, \varepsilon)$ centered in x_1 and of radius ε is a subset of $\mathcal{D}_F$. We consider the convex hull

$$\mathcal{C} := \operatorname{co}\left(\{x_*\} \cup \overline{B}(x_1, \varepsilon)\right),$$

which is clearly is a compact set.

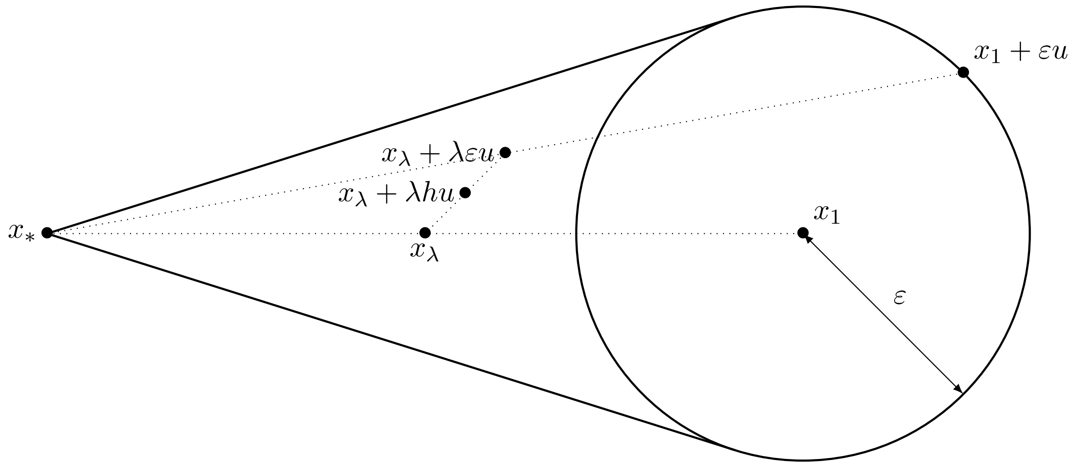
Consider function G defined by:

$$G(x) := D_F(x, x_0) = F(x) - F(x_0) - \langle \nabla F(x_0) | x - x_0 \rangle,$$

so that x_* is the minimizer of G on $\mathcal{X}$. In particular, G is finite in x_*. G inherits strict convexity, lower-semicontinuity, and differentiability on $\mathcal{D}_F$ from function F. G is continous on the compact set $\overline{B}(x_1, \varepsilon)$ because G is convex on the open set $\mathcal{D}_F \supset \overline{B}(x_1, \varepsilon)$. Therefore, G is bounded on $\overline{B}(x_1, \varepsilon)$. Let us prove that G is also bounded on $\mathcal{C}$. Let $x \in \mathcal{C}$. By definition of $\mathcal{C}$, there exists $\lambda \in [0, 1]$ and $x' \in \overline{B}(x_1, \varepsilon)$ such that $x = \lambda x_* + (1-\lambda)x'$. By convexity of G, we have:

$$G(x) \leqslant \lambda G(x_*) + (1-\lambda) G(x') \leqslant G(x_*) + G(x').$$

We know that $G(x_*)$ is finite and that $G(x')$ is bounded for $x' \in \overline{B}(x_1, \varepsilon)$. Therefore G is bounded on $\mathcal{C}$: let us denote $G_{\max}$ and $G_{\min}$ some upper and lower bounds for the value of G on $\mathcal{C}$.



Because $\mathcal{X}$ is a convex set, the segment $[x_*, x_1]$ (in other words the convex hull of $\{x_*, x_1\}$) is a subset of $\mathcal{X}$. Besides, let us prove that the set

$$(x_*, x_1] := \{(1-\lambda)x_* + \lambda x_1 \mid \lambda \in (0,1]\}$$

is a subset of $\mathcal{D}_F$. Let $x_\lambda := (1-\lambda)x_* + \lambda x_1$ (with $\lambda \in (0,1]$) a point in the above set, and let us prove that it belongs to $\mathcal{D}_F$. By definition of the mirror map, we have $\mathcal{X} \subset \operatorname{cl} \mathcal{D}_F$, and besides $x_* \in \mathcal{X}$ by definition. Therefore, there exists a sequence $(x_k)_{k \geqslant 1}$ in $\mathcal{D}_F$ such that $x_k \to x_*$ as $k \to +\infty$. Then, we can write

$$\begin{aligned} x_\lambda &= (1-\lambda)x_* + \lambda x_1 \\ &= (1-\lambda)x_k + (1-\lambda)(x_* - x_k) + \lambda x_1 \\ &= (1-\lambda)x_k + \lambda \left(x_1 + \frac{1-\lambda}{\lambda}(x_* - x_k) \right). \end{aligned}$$

Since $x_k \to x_*$, for high enough k, the point $x_1 + (1-\lambda)\lambda^{-1}(x_* - x_k)$ belongs to $\overline{B}(x_1, \varepsilon)$ and therefore to $\mathcal{D}_F$. Then, the point x_λ belongs to the convex set[2] $\mathcal{D}_F$ as the convex combination of two points in $\mathcal{D}_F$. Therefore, $(x_*, x_1]$ is indeed a subset of $\mathcal{D}_F$.

G being differentiable on $\mathcal{D}_F$ by definition of the mirror map, the gradient of G exists at each point of $(x_*, x_1]$. Let us prove that ∇G is bounded on $(x_*, x_1]$. Let $x_\lambda \in (x_*, x_1]$, where $\lambda \in (0,1]$ is such that

$$x_\lambda = (1-\lambda)x_* + \lambda x_1,$$

and let $u \in \mathbb{R}^n$ such that $\|u\|_2 = 1$. The point $x_1 + \varepsilon u$ belongs to $\mathcal{C}$ because it belongs to $\overline{B}(x_1, \varepsilon)$. The following point also belongs to convex set $\mathcal{C}$ as the convex combination of x_* and $x_1 + \varepsilon u$ which both belong to $\mathcal{C}$:

$$x_\lambda + \lambda \varepsilon u = (1-\lambda)x_* + \lambda(x_1 + \varepsilon u) \in \mathcal{C}. \tag{23}$$

Let $h \in (0, \varepsilon]$. The following point also belongs to $\mathcal{C}$ as a convex combination of x_λ and the above point $x_\lambda + \lambda \varepsilon u$:

$$x_\lambda + \lambda h u = \left(1 - \frac{h}{\varepsilon}\right) x_\lambda + \frac{h}{\varepsilon}\left(x_\lambda + \lambda \varepsilon u\right) \in \mathcal{C}. \tag{24}$$

[2]The domain of a convex function is convex, and therefore $\mathcal{D}_F = \operatorname{int} \operatorname{dom} F$ is convex as the interior of a convex set.

Now using for G the convexity inequality associated with the convex combination from (24), we write:

$$
\begin{aligned}
G(x_\lambda + h\lambda u) - G(x_\lambda) &\leqslant \frac{h}{\varepsilon}\left(G(x_\lambda + \lambda\varepsilon u) - G(x_\lambda)\right) \\
&= \frac{h}{\varepsilon}\left(G(x_\lambda + \lambda\varepsilon u) - G(x_*) + G(x_*) - G(x_\lambda)\right) \\
&\leqslant \frac{h}{\varepsilon}\left(G(x_\lambda + \lambda\varepsilon u) - G(x_*)\right),
\end{aligned}
\tag{25}
$$

where for the last line we used $G(x_*) \leqslant G(x_\lambda)$ which is true because x_λ belongs to $\mathcal{X}$ and x_* is by definition the minimizer of G on $\mathcal{X}$. Using the convexity inequality associated with the convex combination from (23), we also write

$$
\begin{aligned}
G(x_\lambda + \lambda\varepsilon u) - G(x_*) &\leqslant \lambda\left(G(x_1 + \varepsilon u) - G(x_*)\right) \\
&\leqslant \lambda\left(G_{\max} - G_{\min}\right).
\end{aligned}
\tag{26}
$$

Combining (25) and (26) and dividing by $h\lambda$, we get

$$\frac{G(x_\lambda + h\lambda u) - G(x_\lambda)}{h\lambda} \leqslant \frac{G_{\max} - G_{\min}}{\varepsilon}.$$

Taking the limit as $h \to 0^+$, we get that $\langle \nabla G(x_\lambda) | u \rangle \leqslant (G_{\max} - G_{\min})/\varepsilon$. This being true for all vector u such that $\|u\|_2 = 1$, we have

$$\|\nabla G(x_\lambda)\|_2 = \max_{\|u\|_2 = 1} \langle \nabla G(x_\lambda) | u \rangle \leqslant \frac{G_{\max} - G_{\min}}{\varepsilon}.$$

As a result, ∇G is bounded on $(x_*, x_1]$.

Let us deduce that $\partial G(x_*)$ is nonempty. The sequence $(\nabla G(x_{1/k}))_{k\geqslant 1}$ is bounded. Using the Bolzano–Weierstrass theorem, there exists a subsequence $(\nabla G(x_{1/\phi(k)}))_{k\geqslant 1}$ which converges to some vector $\vartheta_* \in \mathbb{R}^n$. For each $k \geqslant 1$, the following is satisfied by convexity of G:

$$\left\langle \nabla G(x_{1/\phi(k)}) \middle| x - x_{1/\phi(k)} \right\rangle \leqslant G(x) - G(x_{1/\phi(k)}), \quad x \in \mathbb{R}^n.$$

Taking the limsup on both sides for each $x \in \mathbb{R}^n$ as $k \to +\infty$, we get (because obviously $x_{1/\phi(k)} \to x_*$):

$$\langle \vartheta_* | x - x_* \rangle \leqslant G(x) - \liminf_{k\to+\infty} G(x_{1/\phi(k)}) \leqslant G(x) - G(x_*), \quad x \in \mathbb{R}^n,$$

where the second inequality follows from the lower-semicontinuity of G. Consequently, ϑ_* belongs to $\partial G(x_*)$.

But by definition of the mirror map ∇F takes all possible values and so does ∇G, because it follows from the definition of G that $\nabla G = \nabla F - \nabla F(x_0)$. Therefore, there exists a point $\tilde{x} \in \mathcal{D}_F$ (thus $\tilde{x} \neq x_*$) such that $\nabla G(\tilde{x}) = \vartheta_*$. Considering the point $x_{\text{mid}} = \frac{1}{2}(x_* + \tilde{x})$, we can write the following convexity inequalities:

$$
\begin{aligned}
\langle \vartheta_* | x_{\text{mid}} - x_* \rangle &\leqslant G(x_{\text{mid}}) - G(x_*) \\
\langle \vartheta_* | x_{\text{mid}} - \tilde{x} \rangle &\leqslant G(x_{\text{mid}}) - G(\tilde{x}).
\end{aligned}
$$

We now add both inequalities and use the fact that $x_{\text{mid}} - \tilde{x} = x_* - x_{\text{mid}}$ by definition of x_{mid} to get $0 \leqslant 2G(x_{\text{mid}}) - G(x_*) - G(\tilde{x})$, which can also be written

$$G\left(\frac{x_* + \tilde{x}}{2}\right) \geqslant \frac{G(x_*) + G(\tilde{x})}{2},$$

which contradicts the strong convexity of G. We conclude that $x_* \in \mathcal{D}_F$. □

Proof of Proposition 3.2. Let $\vartheta \in \mathbb{R}^n$. For each of the three assumptions, let us prove that $h^*(\vartheta)$ is finite. This will prove that $\operatorname{dom} h^* = \mathbb{R}^n$.

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

$$h^*(\vartheta) = \max_{x\in\mathbb{R}^n} \{\langle\vartheta|x\rangle - h(x)\} = \max_{x\in\mathcal{X}} \{\langle\vartheta|x\rangle - h(x)\}.$$

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

(ii) Because $\nabla h(\mathcal{D}_h) = \mathbb{R}^n$ by assumption, there exists $x \in \mathcal{D}_h$ such that $\nabla h(x) = \vartheta$. Then, by Proposition A.2, $h^*(\vartheta) = \langle\vartheta|x\rangle - h(x) < +\infty$.

(iii) The function $x \mapsto \langle\vartheta|x\rangle - h(x)$ is strongly concave on $\mathbb{R}^n$ and therefore admits a maximum. Therefore, $h^*(\vartheta) < +\infty$.

□

Proof of Proposition 3.3. Let $\vartheta \in \mathbb{R}^n$. Because $\operatorname{dom} h^* = \mathbb{R}^n$, the subdifferential $\partial h^*(\vartheta)$ is nonempty—see e.g. [49, Theorem 23.4]. By Proposition A.2, $\partial h^*(\vartheta)$ is the set of maximizers of function $x \mapsto \langle\vartheta|x\rangle - h(x)$, which is strictly concave. Therefore, the maximizer is unique and h^* is differentiable at ϑ by Proposition A.3. □

Proof of Proposition 3.4. h is strictly convex as the sum of two convex functions, one of which (F) is strictly convex. h is lower-semicontinuous as the sum of two lower-continuous functions.

Let us now prove that $\operatorname{cl}\operatorname{dom} h = \mathcal{X}$. First, we write

$$\operatorname{dom} h = \operatorname{dom}(F + I_{\mathcal{X}}) = \operatorname{dom} F \cap \operatorname{dom} I_{\mathcal{X}} = \operatorname{dom} F \cap \mathcal{X}.$$

Let $x \in \operatorname{cl}\operatorname{dom} h = \operatorname{cl}(\operatorname{dom} F \cap \mathcal{X})$. There exists a sequence $(x_k)_{k\geqslant 1}$ in $\operatorname{dom} F \cap \mathcal{X}$ such that $x_k \to x$. In particular, each x_k belongs to closed set $\mathcal{X}$, and so does the limit: $x \in \mathcal{X}$.

Conversely, let $x \in \mathcal{X}$ and let us prove that $x \in \operatorname{cl}(\operatorname{dom} F \cap \mathcal{X})$ by constructing a sequence $(x_k)_{k\geqslant 1}$ in $\operatorname{dom} F \cap \mathcal{X}$ which converges to x. By definition of the mirror map, we have $\mathcal{X} \subset \operatorname{cl}\mathcal{D}_F$, where $\mathcal{D}_F := \operatorname{int}\operatorname{dom} F$. Therefore, there exists a sequence $(x'_l)_{l\geqslant 1}$ in $\mathcal{D}_F$ such that $x'_l \to x$ as $l \to +\infty$. From the definition of the mirror map, we also have that $\mathcal{X}\cap\mathcal{D}_F \neq \varnothing$. Let $x_0 \in \mathcal{X}\cap\mathcal{D}_F$. In particular, x_0 belongs $\mathcal{D}_F$ which is an open set by definition. Therefore, there exists a neighbourhood $U \subset \mathcal{D}_F$ of point x_0. We now construct the sequence $(x_k)_{k\geqslant 1}$ as follows:

$$x_k := \left(1 - \frac{1}{k}\right) x + \frac{1}{k} x_0, \quad k \geqslant 1.$$

x_k belongs to $\mathcal{X}$ as the convex combination of two points in the convex set $\mathcal{X}$, and obviously converges to x. Besides, x_k can also be written, for any $k, l \geqslant 1$,

$$\begin{aligned} x_k &= \left(1 - \frac{1}{k}\right) x'_l + \left(1 - \frac{1}{k}\right)(x - x'_l) + \frac{1}{k} x_0 \\ &= \left(1 - \frac{1}{k}\right) x'_l + \frac{1}{k}\left(x_0 + (k-1)(x - x'_l)\right) \\ &= \left(1 - \frac{1}{k}\right) x'_l + \frac{1}{k} x_0^{(kl)}, \end{aligned}$$

where we set $x_0^{(kl)} := x_0 + (k-1)(x - x'_l)$. For a given $k \geqslant 1$, we see that $x_0^{(kl)} \to x_0$ as $l \to +\infty$ because $x'_l \to x$ by definition of $(x'_l)_{l\geqslant 1}$. Therefore, for large enough l, $x_0^{(kl)}$ belongs to the neighbourhood U and therefore to $\mathcal{D}_F$. x_k then appears as the convex combination of x'_l and $x_0^{(kl)}$ which both belong to the convex set $\mathcal{D}_F \subset \operatorname{dom} F$. (x_k) is thus a sequence in $\operatorname{dom} F \cap \mathcal{X}$ which converges to x. Therefore, $x \in \operatorname{cl}(\operatorname{dom} F \cap \mathcal{X})$ and h is a $\mathcal{X}$-pre-regularizer.

Finally, we have $F \leqslant h$ by definition of h. One can easily check that this implies $h^* \leqslant F^*$ and we know from Proposition 2.2 that $\operatorname{dom} F^* = \mathbb{R}^n$, in other words that F^* only takes finite values. Therefore, so does h^* and h is a $\mathcal{X}$-regularizer. □