UNIFYING MIRROR DESCENT AND DUAL AVERAGING

Anatoli Juditsky, Joon Kwon, Éric Moulines

UNIFYING MIRROR DESCENT AND DUAL AVERAGING

ANATOLI JUDITSKY, JOON KWON, AND ÉRIC MOULINES

Abstract. We introduce and analyze a new family of first-order optimization algorithms which generalizes and unifies both mirror descent and dual averaging. Within the framework of this family, we define new algorithms for constrained optimization that combines the advantages of mirror descent and dual averaging. Our preliminary simulation study shows that these new algorithms significantly outperform available methods in some situations.

Contents

1. Introduction

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 mirror descent algorithms.

Let us start with the most basic setting, in which the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is convex on $\mathbb{R}^n$, 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 is allowed to query the values of the objective $f(x_t)$ and of its gradient $\nabla f(x_t)$ at a search points $x_t \in \mathbb{R}^n$) and which outputs points where the value of the objective function f is provably close to the minimum $f_* = f(x_*)$. The most basic of such algorithm is the (Euclidean) gradient descent, which starts at some 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:

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

Date: June 13, 2022.

where x_{t+1} appears as the solution of a simplified minimization problem where the objective function f has been replaced by its linearization at x_t *plus* a Euclidean *proximal term* $\frac{1}{2\gamma}\left\|x-x_t\right\|_2^2$ which prevents the next iterate x_{t+1} from being too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, introduced in [29, 32], can be seen as an extension of the above gradient descent, in which Euclidean proximal term is replaced with a *Bregman divergence* [6]:

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

where, for any $x, x' \in \mathbb{R}^n$ the Bregman divergence $D_F(\cdot,\cdot)$ associated with a differentiable strictly convex function $F:\mathbb{R}^n\to\mathbb{R}$ is defined as

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

Such 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 smooth with respect to a non-Euclidean norm. One can see that the above mirror descent iteration can be 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 Fenchel–Legendre transform of F,

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

This formulation makes explicit the distinction between the primal space of iterates $(x_t)_{t\geqslant 1}$ and the dual space where the gradients $(\nabla f(x_t))_{t\geqslant 1}$ belong: search point x_t is mapped from the primal into the dual space 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. To force the trajectory of the method to stay in $\mathcal{X}$, the mirror descent should be properly adapted. Such modification can be implemented 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 [32, Chapter 3] and *dual averaging* (DA) introduced in [23, 35], sometimes called *lazy* mirror descent. To illustrate the similarities and differences between MD and DA, we here describe their implementation in the simple Euclidean case. The MD algorithm in this case corresponds to the *projected gradient descent*, in which, given an initial point $x_1\in\mathcal{X}$, for $t\geqslant 1$,

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

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 x_{t+1}, and so on.

For a given initial point $\vartheta_1\in\mathbb{R}^n$, the corresponding algorithm in the DA family writes, for all $t\geqslant 1$:

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

The difference with the projected gradient descent is that the gradient increment is performed 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 led to the two families of algorithms being used and studied in different situations. For instance, DA algorithms seem to be advantageous in distributed problems [16, 17], and manifold identification [18, 25]. They are also believed to possess better averaging properties in the presence of noise [19]. On the other hand, MD is shown to provide better convergence rates in some cases (e.g., in the case of smooth objective f, cf. [19, Section 4.2]). MD also achieves the optimal rate in the adversarial multi-armed bandit problem [1, 2] and the online combinatorial optimization problem with semi-bandit [3] and bandit feedback [10, 13].

Algorithms of the mirror descent type were also transposed to provide solutions for other problems with convex structure. We already mentioned bandit problems. More generally, the mirror descent algorithms have been successful in online learning, see e.g. [7, 11, 20, 37, 39–41], when solving saddle-point problems and variational inequalities [21, 30, 34], similar procedures were used for estimator aggregation in statistical learning [22, 23], etc.[1]

Main contribution. In this paper, we introduce and study a new family of algorithms, which we refer to as *unified mirror descent* (UMD) which unifies and extends both mirror descent and dual averaging. The general algorithm has the property of offering at each step a set of possible iterations. We also construct, in the context of this new family, two algorithms for constrained optimization we refer to as *alternating primal-dual descent* (APDD) and *interpolating primal-dual descent* (IPDD) capable of outperforming mirror descent and dual averaging algorithms in some situations.

It should be mentioned that MD and DA algorithms were studied in a common framework in [26, 27]. Those works are, however, are very different in spirit—they deal with unconstrained problems, the difference between mirror descent and dual averaging appears as a result of utilizing the regularizers/mirror maps which vary over time, and the unification is then achieved by tweaking the way the time-varying regularizers/mirror maps are defined.

Paper outline. In Section 2.1 (resp. 2.2) we recall the definitions of the mirror descent (resp. dual averaging) algorithms. Then, we introduce in Section 3 a new family of algorithms called UMD and show that MD and DA are special cases. We then establish some basic properties of UMD, to be used to derive complexity estimates for UMD-type algorithms in various contexts. Next, in Section 4, we study applications of UMD to smooth and nonsmooth convex optimization. In Section 5 we introduce two new algorithms—APDD (alternating primal-dual descent) and IPDD (interpolating primal-dual descent)—and present results of some preliminary numerical experiments which show that they compare favorably to MD and DA.

Proofs which are longer than few lines are postponed to Appendix B.

1.1. **Preliminaries and notation.**

For $x, \vartheta \in \mathbb{R}^n$, $\langle \vartheta | x \rangle$ stands for 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 $\|\cdot\|$ on $\mathbb{R}^n$, we denote $\|\cdot\|_*$ the conjugate norm,

$$\|\vartheta\|_* = \max_{\|x\| \leqslant 1} \langle \vartheta | x \rangle , \qquad \vartheta \in \mathbb{R}^n.$$

[1] We also refer to [27, Appendix C] for a discussion comparing MD and DA.

The characteristic function $I_{\mathcal{C}} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ of a convex set $\mathcal{C} \subset \mathbb{R}^n$ is zero on $\mathcal{C}$ and equal to $+\infty$ elsewhere. Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be 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 $x \in \mathbb{R}^n$ is the set of $\vartheta \in \mathbb{R}^n$ such that

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

we refer to $\vartheta \in \partial g(x)$ as subgradients of g at x. The Fenchel–Legendre 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 $x \in \mathbb{R}^n$, its Bregman divergence between $x, x' \in \mathbb{R}^n$ is defined as

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

Some convexity definitions and results are recalled in Section A.

Throughout the paper, we consider algorithms associated with an arbitrary sequence $(\xi_t)_{t \geqslant 1}$ in $\mathbb{R}^n$, and the problem domain $\mathcal{X} \subset \mathbb{R}^n$ is a closed and nonempty convex set.

2. Mirror descent and dual averaging algorithms

2.1. **Mirror descent.**

The 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 $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. Denote $\mathcal{D}_F := \operatorname{int} \operatorname{dom} F$. We say that F is an $\mathcal{X}$-compatible *mirror map* if

(i) F is lower-semicontinuous 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) $\mathcal{X} \cap \mathcal{D}_F \neq \emptyset$.

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

Proposition 2.2. *Let $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be an $\mathcal{X}$-compatible mirror map, F^* the Fenchel–Legendre 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 $y \in \mathbb{R}^n$, $\nabla F^*(\nabla F(x)) = x$ and $\nabla F(\nabla F^*(y)) = y$.

We can now define the mirror descent algorithm [29].

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

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

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

The above is well-defined thanks to the following recursive argument. 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.

Theorem 2.4 (Bregman projection onto $\mathcal{X}$)**.** [2] *Let $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be an $\mathcal{X}$-compatible mirror map, and let $\mathcal{D}_F := \operatorname{int} \operatorname{dom} F$. Then for any point $x_0 \in \mathcal{D}_F$, the minimizer of $D_F(x, x_0)$ over $x \in \mathcal{X}$ exists, is unique, and 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 $\nabla F(\tilde{x}_t) = \nabla F(x_t) + \xi_t$. Then, using the definition of the Bregman divergence:

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

The last expression is called *primal formulation*, and is usually taken as the definition of mirror descent, cf. [4, Section 3]. Introducing the *prox-mapping*:

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

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

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

Some examples of widely used mirror maps are as follows.

Example 2.5 (Gradient descent)**.** The simplest example is provided by $\mathcal{X} = \mathbb{R}^n$ and $F(x) = \frac{1}{2} \|x\|_2^2$. One can easily see that F is indeed an $\mathbb{R}^n$-compatible mirror map. In this case, $\nabla F(x) = \nabla F^*(x) = x$ are identity mappings, and the update rule (MD) boils down to the gradient descent iteration if we consider dual increments $\xi_t := -\gamma \nabla f(x_t)$ where $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable objective function.

Example 2.6 (Projected gradient descent)**.** A common variant of the previous example is the case where $\mathcal{X}$ is some closed proper subset of $\mathbb{R}^n$. One can easily see that $F(x) = \frac{1}{2} \|x\|_2^2$ is an $\mathcal{X}$-compatible mirror map. If f is an objective function which is differentiable on $\mathcal{X}$, then (MD) with $\xi_t = -\gamma \nabla f(x_t)$ corresponds to the standard projected gradient descent algorithm.

[2] Similar statements can be found in the literature (cf. e.g., [14, Lemma A.1]) but we could not find one that exactly matches assumptions of Theorem 2.4 on F and $\mathcal{X}$. We provide a detailed proof in Appendix B for completeness.

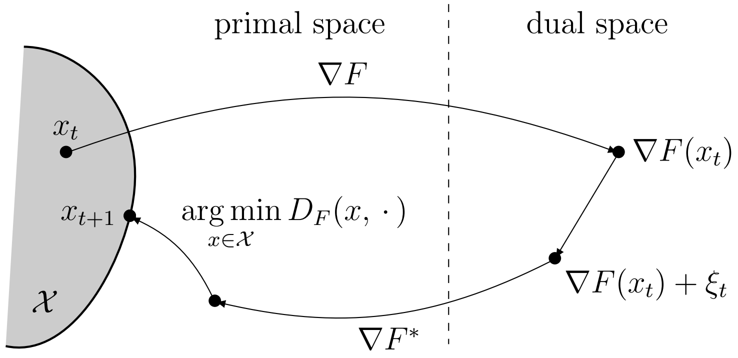


FIGURE 1. Mirror descent

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

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

and F given 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 $x \notin \mathbb{R}_+^n$. Then F is an $\mathcal{X}$-compatible mirror map.

2.2. **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 2.11 that for a given mirror map, there always exists a corresponding regularizer but the converse is not true.

Definition 2.8 (Regularizers)**.** A function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is an $\mathcal{X}$-*pre-regularizer* if it is strictly convex, lower-semicontinuous, and if $\operatorname{cl} \operatorname{dom} h = \mathcal{X}$. Moreover, if $\operatorname{dom} h^* = \mathbb{R}^n$, then h is said to be an $\mathcal{X}$-*regularizer*.

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

Proposition 2.9. *Let h be an $\mathcal{X}$-pre-regularizer.*

(i) If $\mathcal{X}$ is compact, then h is an $\mathcal{X}$-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 an $\mathcal{X}$-regularizer.
(iii) If h is strongly convex with respect to some norm $\|\cdot\|$, then h is an $\mathcal{X}$-regularizer.

Proposition 2.10 (Differentiability of h^*)**.** *Let h be an $\mathcal{X}$-regularizer. Then h^* is differentiable on $\mathbb{R}^n$.*

Proposition 2.11. *Let F be an $\mathcal{X}$-compatible mirror map. Then, $h := F + I_{\mathcal{X}}$ is an $\mathcal{X}$-regularizer, and, moreover, $\nabla F(x) \in \partial h(x)$ for all $x \in \mathcal{D}_F$.*

Proofs of Propositions 2.9–2.11 are postponed to Appendix B.1.

Corollary 2.12. *(i) $h(x) := \frac{1}{2}\|x\|_2^2 + I_{\mathcal{X}}(x)$ is an $\mathcal{X}$-regularizer.*
(ii) The entropy defined as:

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

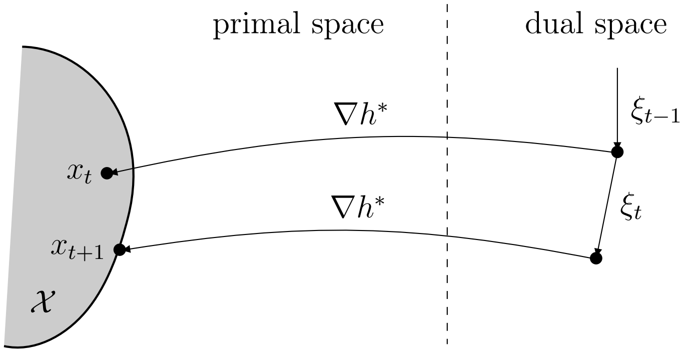


FIGURE 2. Dual averaging

where $\Delta_n := \left\{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\right\}$ *(where* $0 \log 0 = 0$ *by convention) is a* Δ_n*-regularizer.*

Example 2.13 (Elastic-net regularization)**.** An example of a 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 due to the strong convexity (cf. Proposition 2.9).

We now recall the definition of the dual averaging (DA) iterates in the case of the constant regularizer.[3]

Definition 2.14 (Dual averaging [34, 35])**.** Let h be an $\mathcal{X}$-regularizer and let $\xi := (\xi_t)_{t\geqslant 1}$ 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.

3. The unified mirror descent algorithm

In this section, we introduce a general family of algorithms which we refer to as *unified mirror descent (UMD)* and show that MD and DA are special cases.

[3] In its general form [35], the DA algorithm allows for a time-variable regularizer. For the sake of clarity, we consider here the simple case of time-invariant regularizers which already captures some essential differences between MD and DA.

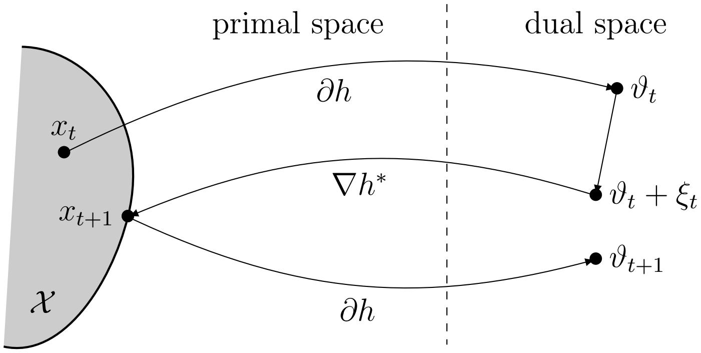


FIGURE 3. Unified mirror descent

3.1. **Definition, properties and special cases.**

Definition 3.1. Let h be an $\mathcal{X}$-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 sequence of UMD iterates associated with h and ξ (or a UMD(h,ξ) sequence for short) if for all $t \geqslant 1$:

$$x_t = \nabla h^*(\vartheta_t), \tag{2a}$$

$$\forall x \in \mathcal{X},\ \langle \vartheta_{t+1} - \vartheta_t - \xi_t | x - x_{t+1}\rangle \geqslant 0. \tag{2b}$$

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 3.2. *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, $x_t = \nabla h^*(\vartheta_t)$, which combined with Proposition A.3 of Appendix A implies (i). Furthermore, 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 the second inequality results from (2b) in the definition of UMD iterates. On the other hand, $h(x) = +\infty$ for $x \notin \mathcal{X}$ implying the inequality $h(x) - h(x_{t+1}) \geqslant \langle \vartheta_t + \xi_t | x - x_{t+1}\rangle$ in this case. We conclude that $\vartheta_t + \xi_t$ also belongs to $\partial h(x_{t+1})$, i.e., (ii) holds true. □ □

Remark 3.3 (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) iterates always exist. Indeed, from the definition of regularizers, it follows that there exists a primal point $x_1 \in \mathcal{X}$ such that $\partial h(x_1) \neq \emptyset$; 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 (2b), and then define $x_{t+1} := \nabla h^*(\vartheta_{t+1})$. This choice of ϑ_{t+1} actually corresponds to the iteration of the DA algorithm.

Proposition 3.4 (DA is a special case of UMD)**.** *Let h be an $\mathcal{X}$-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 (2a) is true by definition of (DA). Besides, the relation $\vartheta_{t+1} = \vartheta_t + \xi_t$ makes condition (2b) trivially satisfied because one of the arguments of the scalar product is zero. □ □

Proposition 3.5 (MD is a special case of UMD)**.** *Let F be an $\mathcal{X}$-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 (2a) and (2b) are satisfied with $h := F + I_{\mathcal{X}}$.

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.3 of the appendix. We can then rewrite the (MD) iteration as follows:

$$\begin{aligned} x_{t+1} &= \arg\min_{x \in \mathcal{X}} D_F(x,\ \tilde{x}_t) \\ &= \arg\min_{x \in \mathcal{X}} \left\{ F(x) - F(\tilde{x}_t) - \langle \nabla F(\tilde{x}_t) | x - \tilde{x}_t \rangle \right\} \\ &= \arg\min_{x \in \mathcal{X}} \left\{ F(x) - \langle \vartheta_t + \xi_t | x \rangle \right\}. \end{aligned} \tag{3}$$

In other words, x_{t+1} is the minimizer on $\mathcal{X}$ of the convex function $F(x) - \langle \vartheta_t + \xi_t | x \rangle$. 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 $\mathcal{D}_F$, so that the optimality conditions for (3) imply that

$$\forall x \in \mathcal{X}, \quad \langle \nabla F(x_{t+1}) - \vartheta_t - \xi_t | x - x_{t+1} \rangle \geqslant 0, \tag{4}$$

which is exactly condition (2b) due to $\vartheta_{t+1} = \nabla F(x_{t+1})$.

Thanks to Proposition 2.11, we know that $\vartheta_t = \nabla F(x_t) \in \partial h(x_t)$ which is equivalent to $x_t \in \nabla h^*(\vartheta_t)$ (see Proposition A.3); thus, condition (2a) is satisfied. □ □

One may 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:

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

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:

$$\begin{gathered} \vartheta_1 \in \partial h(x_1), \\ (x_{t+1}, \vartheta_{t+1}) \in \Pi_h(\vartheta_t + \xi_t), \quad t \geqslant 1. \end{gathered}$$

Remark 3.6 (On the non-unicity of UMD iterates). An interesting characteristic of UMD iterates 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 particular, in such case, DA and MD coincide. This is the case, for instance, if $\mathcal{X} = \mathbb{R}^n$ and the regularizer h is differentiable on $\mathbb{R}^n$.

3.2. **Simple examples.** As an illustration, let us describe the iterates of MD, DA and UMD in the Euclidean setting corresponding to the $\mathcal{X}$-regularizer $h(x) = \frac{1}{2}\left\|x\right\|_2^2 + I_{\mathcal{X}}$ and the mirror map $F(x) = \frac{1}{2}\left\|x\right\|_2^2$. It is easy to check that the map ∇h^* is the Euclidean projection onto $\mathcal{X}$. We consider below two simple cases of the set $\mathcal{X}$. We denote $(x_t, \vartheta_t)_{t\geqslant 1}$ a sequence of UMD(h, ξ) iterates.

Euclidean ball. Here we consider the case of $\mathcal{X} = \overline{B}(0,1)$, the closed unit Euclidean ball of $\mathbb{R}^2$. Let $t \geqslant 1$ and assume that $\vartheta_t + \xi_t$ is outside $\overline{B}(0,1)$, so that x_{t+1} which is the Euclidean projection of $\vartheta_t + \xi_t$ belongs to the boundary of $\overline{B}(0,1)$; thus, $\left\|x_{t+1}\right\|_2 = 1$. From this point x_{t+1}, an MD iteration corresponds to choosing $\vartheta_{t+1}^{\mathrm{MD}} := x_{t+1}$, and a DA iteration corresponds to choosing $\vartheta_{t+1}^{\mathrm{DA}} := \vartheta_t + \xi_t$. Besides this, we can see that the set of points ϑ_{t+1} which have x_{t+1} as Euclidean projection is $[1, +\infty)\, x_{t+1}$ and that the set of points ϑ_{t+1} satisfying condition (2b) is $(-\infty, 1]\,(\vartheta_t + \xi_t)$. Therefore, the set of vectors ϑ_{t+1} satisfying both conditions (2b) and (2a) is the convex hull of x_{t+1} and $\vartheta_t + \xi_t$, which is represented by a thick segment in Figure 4.

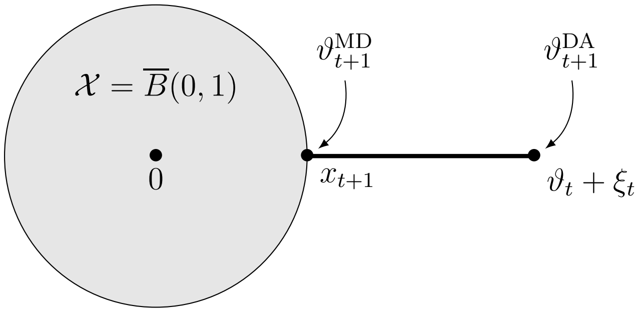

$\mathcal{X} = \overline{B}(0,1)$
0
$\vartheta_{t+1}^{\mathrm{MD}}$
$\vartheta_{t+1}^{\mathrm{DA}}$
x_{t+1}
$\vartheta_t + \xi_t$

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

Simplex in $\mathbb{R}^2$. Suppose that the ambient space is $\mathbb{R}^2$, and

$$\mathcal{X} = \{(x^{(1)}, x^{(2)}) \in \mathbb{R}_+^2 : x^{(1)} + x^{(2)} \leqslant 1\}$$

is the “full simplex” of $\mathbb{R}^2$. Let $t \geqslant 1$ and assume that $\eta = \vartheta_t + \xi_t$ belongs to the normal cone $\mathcal{N}_{\mathcal{X}}(\overline{x})$ of $\mathcal{X}$ at $\overline{x} = (0,1)$, i.e.,

$$\mathcal{N}_{\mathcal{X}}(\overline{x}) = \{\vartheta \in \mathbb{R}^2 : \vartheta^{(2)} \geqslant 1,\, \vartheta^{(2)} - \vartheta^{(1)} \geqslant 1\},$$

cf. Figure 5. In this case, $x_{t+1} = \overline{x}$ and the set of points ϑ_{t+1} which have $\overline{x}$ as Euclidean projection is $\mathcal{N}_{\mathcal{X}}(\overline{x})$, while the set of ϑ_{t+1} satisfying (2b) is

$$\mathcal{S} = \{\vartheta \in \mathbb{R}^2 : \vartheta^{(2)} \leqslant \eta_t^{(2)},\, \vartheta^{(2)} - \vartheta^{(1)} \leqslant \eta^{(2)} - \eta^{(1)}\}.$$

The set of vectors satisfying both (2b) and (2a) is represented by the dashed area in Figure 5.

3.3. **A class of iterates interpolating MD and DA.** As an example which goes beyond MD and DA, we consider a class of UMD iterates which, at each step, interpolates between a MD and a DA iteration; the IPDD algorithm considered in Section 5.2 is from this class.

Let F be an $\mathcal{X}$-compatible mirror map, and $h = F + I_{\mathcal{X}}$ the associated regularizer. Let $\xi := (\xi_t)_{t\geqslant 1}$ be a sequence of dual increments in $\mathbb{R}^n$, and $(\alpha_t)_{t\geqslant 1}$ be a sequence of coefficients

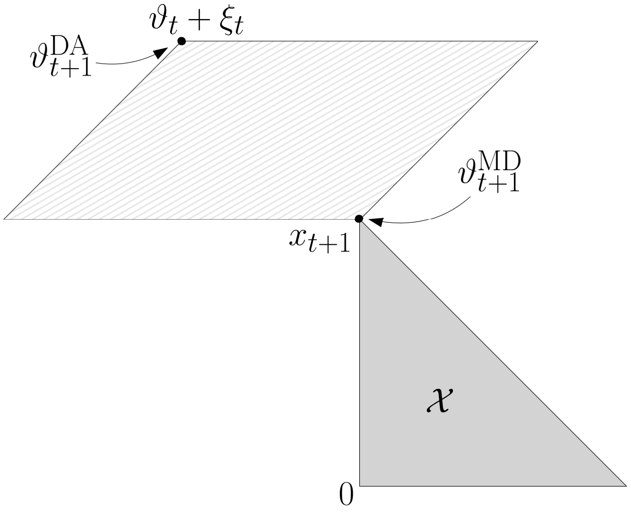

$\vartheta_t + \xi_t$
$\vartheta_{t+1}^{\mathrm{DA}}$
$\vartheta_{t+1}^{\mathrm{MD}}$
x_{t+1}
$\mathcal{X}$
0

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

in $[0, 1]$. Consider a sequence of iterates $(x_t, \vartheta_t)_{t\geqslant 1}$ as follows: $\vartheta_1 \in \mathbb{R}^n$, $x_1 = \nabla h^*(\vartheta_1)$, and for $t \geqslant 2$

$$\begin{aligned} x_{t+1} &= \nabla h^*(\vartheta_t + \xi_t), \\ \vartheta_{t+1} &= \alpha_t \nabla F(x_t) + (1 - \alpha_t)(\vartheta_t + \xi_t). \end{aligned} \tag{5}$$

Note that selecting constant weights $(\alpha_t)_{t\geqslant 1}$ equal to 0 (resp. equal to 1) corresponds to DA (resp. MD).

Proposition 3.7. *Sequence $(x_t, \vartheta_t)_{t\geqslant 1}$ as defined in (5) is a sequence of UMD(h, ξ) iterates.*

Proof. $x_1 = \nabla h^*(\vartheta_1)$ by definition. For $t \geqslant 2$, establishing $x_t = \nabla h^*(\vartheta_t)$ is equivalent (by Proposition A.3 of Appendix A) to proving that $\vartheta_t \in \partial h(x_t)$. By definition, ϑ_t is a convex combination of $\nabla F(x_t)$ and $\vartheta_t + \xi_t$ which are both elements of a convex set $\partial h(x_t)$. Therefore, relationship (2a) from Definition 3.1 holds true. On the other hand, for $x \in \mathcal{X}$ and $t \geqslant 1$,

$$\begin{aligned} \langle \vartheta_{t+1} - \vartheta_t - \xi_t | x - x_{t+1} \rangle &= \langle \alpha_t \nabla F(x_t) + (1 - \alpha_t)(\vartheta_t + \xi_t) - \vartheta_t - \xi_t | x - x_{t+1} \rangle \\ &= \alpha_t \langle \nabla F(x_t) - \vartheta_t - \xi_t | x - x_{t+1} \rangle \geqslant 0 \end{aligned}$$

(the last inequality holds for the same reasons as the similar inequality (4) in the proof of Proposition 3.5), implying variational condition (2b). □ □

3.4. **Analysis.** We introduce a natural extension of the Bregman divergence which will be central when analyzing the properties of UMD iterates.

Definition 3.8 (Bregman divergence)**.** Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function. For $x \in \mathbb{R}^n$ such that $\partial g(x) \neq \emptyset$, $x' \in \mathbb{R}^n$, and $\vartheta \in \partial g(x)$, we define the Bregman divergence from x to x' with subgradient ϑ as

$$D_g(x', x;\ \vartheta) := g(x') - g(x) - \langle \vartheta | x' - x \rangle .$$

Remark 3.9. If g is differentiable at x, the *traditional* Bregman divergence from x to x' is well-defined and is equal to the Bregman divergence (as defined above) from x to x' with (the only) subgradient $\nabla g(x)$; in other words: $D_g(x', x;\ \nabla g(x)) = D_g(x', x)$.

Note that a different generalization of the Bregman divergence using directional derivatives was proposed in [15, 27]; it does not lead however to a unifying view of mirror descent and dual averaging algorithms, due to the uniqueness of directional derivatives.

Proposition 3.10. *Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower-semicontinuous convex function. Let $x, x', \vartheta, \vartheta' \in \mathbb{R}^n$ be 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^*} is the Bregman divergence associated with the Fenchel–Legendre transform g^* of g.

(ii) Moreover, if g is strongly convex (with modulus 1)[4] *with respect to a given norm $\|\cdot\|$, g^* is differentiable on $\mathbb{R}^n$ and*

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

Proof. (i) The nonnegativity of D_g is an immediate consequence of the convexity of g. Using the Fenchel identity (property (iii) from Proposition A.3), we write

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

(ii) The differentiability of g^* and the second inequality is given by [39, Lemma 15]. For the first inequality, we refer to [39, Lemma 13]. □

□

We now establish the following fundamental result, which is an extension of classical statements [12, Lemma 3] and [34, Lemma 4]. It underlies the analysis of the algorithms of the mirror descent type and is operational when deriving accuracy guarantees in various applications of the UMD presented below.

Lemma 3.11. *Let h be an $\mathcal{X}$-regularizer, $\xi := (\xi_t)_{t\geqslant 1}$ be 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 $t \geqslant 1$,*

$$\langle \xi_t | x - x_{t+1}\rangle \leqslant D_h(x, x_t;\ \vartheta_t) - D_h(x, x_{t+1};\ \vartheta_{t+1}) - D_h(x_{t+1}, x_t;\ \vartheta_t). \tag{6}$$

As a consequence,

$$\langle \xi_t | x - x_t\rangle \leqslant D_h(x, x_t;\ \vartheta_t) - D_h(x, x_{t+1};\ \vartheta_{t+1}) + D_{h^*}(\vartheta_t + \xi_t, \vartheta_t). \tag{7}$$

[4] With some terminological abuse, we say that g is strongly convex when it is strongly convex with modulus 1.

Proof. Let $x \in \operatorname{dom} h$ and $t \geqslant 1$. Using variational inequality (2b) from the definition of UMD iterates we write

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

The above 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 3.2).

To prove (7), 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 3.2. Moreover,

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

where the first equality is due to Proposition 3.10–(i) and the second equality—to the differentiability of h^*. Combining the two previous displays and adding to (6) gives the result.
□ □

Remark 3.12. The "high level idea" underlying the construction of the unified mirror descent is to combine elements of DA and MD algorithms. MD makes use of a mirror map, which is differentiable on the interior of its domain. As shown in Figure 1, its gradient ∇F is used to go from the primal space to the dual space, where the *dual iteration* is performed. Then, ∇F^* is used to come back to the primal space, where the Bregman projection is done.

Our first observation is that this *back-and-forth* between the primal and dual spaces can be extended to regularizers which may be non-differentiable. Indeed, the corresponding regularizer $h = F + I_{\mathcal{X}}$ can be seen as a "limit" of mirror maps whose values outside of $\mathcal{X}$ are sent up to $+\infty$; when the regularizer is not differentiable, the corresponding "limit" of the gradients corresponds to the subdifferential. Therefore, a natural idea is to define dual UMD iterates using the subdifferential whenever the gradient does not exist, as in the diagram from Figure 3, meaning that $\vartheta_t \in \partial h(x_t)$, with $x_{t+1} = \nabla h^*(\vartheta_t + \xi_t)$, $t \geqslant 1$. Note that both MD and DA satisfy these relationships.

However, the above recursion "as is" may not obey accuracy bounds which hold for MD and DA recursions. To make this recursion "interesting" we need to constrain the choice of ϑ_t in the subdifferential of $h(x_t)$. Our second observation is that in the proof of accuracy bounds for MD and DA, results similar to Lemma 3.11 are operational (cf. e.g. [12, Lemma 3] and [34, Lemma 4]), with MD and DA satisfying inequality (6) of Lemma 3.11 "by construction". One easily check that the variational condition (2b) is exactly the constraint needed for inequality (6) to hold.

An immediate consequence of Lemma 3.11 and property (ii) of Proposition 3.10 is the following inequality (sometimes called *regret bound*), which extends and unifies classical guarantees on MD and DA—cf., e.g., [31, Lemma 2.1], [34, Lemma 4], [9, Theorems 5.2 & 5.4], etc.

Corollary 3.13. *Let h be an $\mathcal{X}$-regularizer which is assumed to be strongly convex with respect to some norm $\|\cdot\|$, and $\xi := (\xi_t)_{t\geqslant 1}$ be 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 $x \in \operatorname{dom} h$,*

$$\sum_{t=1}^{T} \langle \xi_t | x - x_t \rangle \leqslant D_h(x, x_1;\ \vartheta_1) - D_h(x, x_{T+1};\ \vartheta_{T+1}) + \frac{1}{2}\sum_{t=1}^{T} \|\xi_t\|_*^2 \tag{8}$$

where $\|\cdot\|_*$ is the norm conjugate to $\|\cdot\|$.

4. Accuracy bounds for convex optimization

Throughout this section, let $\mathcal{X}$ be a convex and closed domain of $\mathbb{R}^n$, h an $\mathcal{X}$-regularizer, $\|\cdot\|$ some norm in $\mathbb{R}^n$, and we assume that h is strongly convex w.r.t. $\|\cdot\|$.

4.1. UMD for nonsmooth convex optimization.

Let $M > 0$, and let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function with domain $\operatorname{dom} f \supset \mathcal{X}$, which is sub-differentiable 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

$$f_* = \min_{x\in\mathcal{X}} f(x) \tag{9}$$

is solvable, and denote $x_* \in \mathcal{X}$ a minimizer. Let $(\gamma_t)_{t\geqslant 1}$ be a sequence of positive *step-sizes*. We consider a sequence $(x_t, \vartheta_t)_{t\geqslant 1}$ of UMD(h,ξ) iterates associated with dual increments $\xi := (-\gamma_t f'(x_t))_{t\geqslant 1}$, where $f'(x_t) \in \partial f(x_t)$. Namely, $\vartheta_1 \in \partial h(x_1)$, 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 $\overline{x}_T$ by UMD after T iterations, computed either as

$$\overline{x}_T = \left(\sum_{t=1}^{T} \gamma_t\right)^{-1} \sum_{t=1}^{T} \gamma_t x_t, \qquad \text{or} \qquad \overline{x}_T \in \operatorname*{Arg\,min}_{t=1,\dots,T} f(x_t).$$

In particular, it recovers known guarantees for the mirror descent [32, Theorem 3.3.5], [4, Theorem 4.1] and dual averaging [35, Theorem 1] algorithms. The following result is a straightforward consequence of Corollary 3.13.

Proposition 4.1. *Suppose that $x_* \in \operatorname{dom} h$. Then for all $T \geqslant 1$,*

$$f(\overline{x}_T) - f_* \leqslant \left(\sum_{t=1}^{T} \gamma_t\right)^{-1} \left[D_h(x_*, x_1;\ \vartheta_1) + \frac{M^2}{2}\sum_{t=1}^{T} \gamma_t^2\right].$$

Let $\Omega_{\mathcal{X}}$ be an upper estimate of $\sqrt{2D_h(x_*, x_1;\ \vartheta_1)}$.[5] UMD algorithm with constant step-sizes

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

[5] 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 $\Omega_{\mathcal{X}} \geqslant \max_{x\in\mathcal{X}} \|x - x_1\|$.

satisfies:

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

Next, following [28, 36], let us consider an alternative algorithm whose iterates $(x_t, y_t, \vartheta_t)_{t\geqslant 1}$ are defined as $x_1 = y_1$, $\vartheta_1 \in \partial h(x_1)$, and for $t \geqslant 1$,

$$\begin{aligned} (x_{t+1}, \vartheta_{t+1}) &\in \Pi(\vartheta_t - \gamma_t f'(y_t)), \\ y_{t+1} &= (1-\nu_t) y_t + \nu_t x_{t+1}, \end{aligned} \tag{10}$$

where $\nu_t \in (0,1)$ is given by

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

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

The following statement provides accuracy bounds for the last iterate y_T of the recursion (10) and generalizes respective accuracy bounds of [28, 36].

Proposition 4.2. *Suppose that $x_* \in \operatorname{dom} h$. Then for $T \geqslant 1$,*

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

In particular, for constant step-sizes

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

where $\Omega_{\mathcal{X}}$ is an upper bound for of $\sqrt{2D_h(x_*, x_1;\ \vartheta_1)}$, one has

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

4.2. **UMD for smooth convex optimization.** In this section, in the context of smooth convex optimization, we first present a class of algorithms which generalizes gradient descent and enjoys a $1/T$ convergence rate. The new algorithms defined below in Section 5 belong to this class. Moreover, we also define a generalization of Nesterov's accelerated gradient method which guarantees a $1/T^2$ convergence rate.

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function which is continuously differentiable on $\mathcal{X}$ with Lipschitz-continuous gradient, i.e.,

$$\left\| \nabla f(x) - \nabla f(x') \right\|_* \leqslant L \left\| x - x' \right\|, \qquad x, x' \in \mathcal{X}. \tag{11}$$

We assume that the problem

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

is solvable and denote $x_* \in \mathcal{X}$ a minimizer. Let $(\gamma_t)_{t\geqslant 1}$ be a sequence of positive step-sizes, and $(x_t, \vartheta_t)_{t\geqslant 1}$ be UMD(h, ξ) iterates with $\xi := (-\gamma_t \nabla f(x_t))_{t\geqslant 1}$. In other words, $\vartheta_1 \in \partial h(x_1)$, and

$$x_{t+1} = \Pi_h(\vartheta_t - \gamma \nabla f(x_t)), \qquad t \geqslant 1.$$

Theorem 4.3. *Suppose that $x_* \in \operatorname{dom} h$. Assume that step-sizes $(\gamma_t)_{t\geqslant 1}$ are chosen in such a way that condition*

$$\gamma_t D_f(x_{t+1}, x_t) \leqslant D_h(x_{t+1}, x_t;\ \vartheta_t) \tag{12}$$

is satisfied for all $t \geqslant 1$, which is always the case for $\gamma_t \leqslant 1/L$. Then for $T \geqslant 1$, one has

$$f(x_{T+1}) - f_* \leqslant \left(\sum_{t=1}^{T} \gamma_t\right)^{-1} D_h(x_*, x_1;\ \vartheta_1). \tag{13}$$

In particular, for constant step-sizes $\gamma_t \equiv \gamma = 1/L$, one has

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

We now aim at presenting an "UMD analogue" of Nesterov's accelerated gradient descent algorithm, which unifies and generalizes classic algorithmic schemes for smooth optimization, like e.g. optimal scheme for smooth minimization from [33], optimal method from [24], etc.

Let points $(x_t, y_t, z_t, \vartheta_t)_{t\geqslant 1}$ satisfy $x_1 = z_1 = \nabla h^*(\vartheta_1)$, and for $t \geqslant 1$,

$$y_t = (1-\nu_t)z_t + \nu_t x_t \tag{14a}$$

$$(x_{t+1}, \vartheta_{t+1}) \in \Pi_h(\vartheta_t - \gamma_t \nabla f(y_t)) \tag{14b}$$

$$z_{t+1} = (1-\nu_t)z_t + \nu_t x_{t+1}, \tag{14c}$$

with positive step-sizes $(\gamma_t)_{t\geqslant 1}$ and $(\nu_t)_{t\geqslant 1}$ a sequence in $(0,1)$.

We refer to $(x_t, y_t, z_t, \vartheta_t)_{t\geqslant 1}$ as *accelerated unified mirror descent (AUMD)* iterates. Note that AUMD iterates are well defined. Indeed, it follows from the above definition that $(x_t, \vartheta_t)_{t\geqslant 1}$ is a sequence of UMD(h, ξ) iterates (associated with dual increments $\xi := (-\gamma_t \nabla f(y_t))_{t\geqslant 1}$). Therefore, f being differentiable on $\mathcal{X}$, x_{t+1} do exist whenever $y_t \in \mathcal{X}$. To show the latter, y_t being a convex combination of x_t and z_t, it suffices to check that x_t, z_t are well-defined and belong to $\mathcal{X}$, which can be done recursively. Indeed, $x_1, z_1 \in \mathcal{X}$ by construction. Then, assuming that $x_t, z_t \in \mathcal{X}$, we get that $y_t \in \mathcal{X}$. As a result, x_{t+1} is well-defined and belongs to $\mathcal{X}$, and so does z_{t+1}, being a convex combination of z_t and x_{t+1}.

The following result states the accuracy guarantees for the AUMD algorithm (14a)–(14c) and generalizes corresponding statements from [24, Theorem 2] and [5, Theorem 4.4].

Theorem 4.4. *Suppose that $x_* \in \operatorname{dom} h$. Let $\gamma_1 = 1/L$ and for $t \geqslant 1$,*

$$\gamma_{t+1} = (2L)^{-1}\left(1 + \sqrt{1 + (2L\gamma_t)^2}\right), \qquad \nu_t = (L\gamma_t)^{-1}. \tag{15}$$

Then for $T \geqslant 1$, it holds:

$$f(z_{T+1}) - f_* \leqslant \frac{4LD_h(x_*, x_1;\ \vartheta_1)}{(T+1)^2}. \tag{16}$$

5. APDD and IPDD: new algorithms for constrained optimization

In constrained convex optimization problems with minimizer lying at the boundary of the feasible domain, (primal) mirror descent algorithm often shows fast convergence, but is sensitive to the selected step-size, whereas dual averaging achieves slower convergence, but is robust to the step-size choice (cf. the discussion below).

We introduce two new algorithms from the UMD family which we refer to as *Alternating Primal-Dual Descent (APDD)* and *Interpolating Primal-Dual Descent (IPDD)*, and which belong to the class described in Section 3.3. They, therefore, benefit from the theoretical guarantees from Section 4. The idea behind their construction is to alternate between different iterations types in order to combine the advantages of mirror descent and dual averaging so that they perform better than (or as well as) the best of both algorithms with moderate computational overhead.[6] In the reminder of this section, $f : \mathbb{R}^d \to \mathbb{R}$ is a differentiable function.

5.1. **The APDD algorithm.** Every k steps (where $k \geqslant 1$ is a parameter) the APDD algorithm computes both a mirror descent update and a dual averaging update, and then compares the value of the objective function at the two primal iterates thus obtained. The algorithm then records the so-determined best update and performs for the remaining $k-1$ steps dual averaging updates. In a variant of the algorithm, the best update is determined over not one, but over several steps ahead.

Definition 5.1 (k-APDD algorithm)**.** Let $k \geqslant 1$ be an integer, $\gamma > 0$, F be an $\mathcal{X}$-compatible mirror map, $h = F + I_{\mathcal{X}}$, and $(x_1, \vartheta_1) \in \mathcal{X} \times \mathbb{R}^d$ such that $x_1 = \nabla h^*(\vartheta_1)$.

The primal and dual iterates $(x_t, \vartheta_t)_{t \geqslant 1}$ of the k-APDD algorithm with step-size γ and objective function f are defined as $x_2 = \nabla h^*(\vartheta_1 - \gamma \nabla f(x_1))$ and for $t \geqslant 2$,

- if $t \equiv 2 \mod k$, $\vartheta_t \in \operatorname{Arg\,min}_{\vartheta \in \left\{\vartheta_t^{\mathrm{MD}}, \vartheta_t^{\mathrm{DA}}\right\}} f\left(\nabla h^*(\vartheta - \gamma \nabla f(x_t))\right)$,
  where $\vartheta_t^{\mathrm{MD}} = \nabla F(x_t)$ and $\vartheta_t^{\mathrm{DA}} = \vartheta_{t-1} - \gamma \nabla f(x_{t-1})$;
- if $t \not\equiv 2 \mod k$, $\vartheta_t = \vartheta_{t-1} - \gamma \nabla f(x_{t-1})$.
- $x_{t+1} = \nabla h^*(\vartheta_t - \gamma \nabla f(x_t))$.

Observe that numerical complexity of the iterate of the APDD does not exceed twice the complexity of the iterate of the MD/DA-algorithm.

5.2. **The IPDD algorithm.** The IPDD algorithm performs at time $t \geqslant 1$ an iteration which interpolates DA and MD (with coefficient α) if the condition

$$\gamma D_f(x_{t+1}, x_t) \leqslant D_h(x_{t+1}, x_t;\ \vartheta_t) \tag{17}$$

is satisfied at this iteration.[7] When this condition is not satisfied, the IPDD algorithm performs a DA iteration instead.

Definition 5.2 (The α-IPDD algorithm)**.** Let $\alpha \in (0, 1]$, $\gamma > 0$; let also F be a $\mathcal{X}$-compatible mirror map, $h = F + I_{\mathcal{X}}$, and $(x_1, \vartheta_1) \in \mathcal{X} \times \mathbb{R}^d$ such that $x_1 = \nabla h^*(\vartheta_1)$.

[6] APDD and IPDD algorithms should be seen as mere examples having nothing special which sets them apart from other possible UMD implementations.

[7] Recall, that satisfaction of such condition (cf. (12)) at each iteration of the method, ensures that the bound (13) of Theorem 4.3 holds true.

The primal and dual iterates $(x_t, \vartheta_t)_{t\geqslant 1}$ of the α-IPDD algorithm with step-size γ and objective function f are defined as $x_2 = \nabla h^*(\vartheta_1 - \gamma\nabla f(x_1))$ and for $t \geqslant 2$,

$$\begin{aligned}
\vartheta_t^{(0)} &= \alpha\nabla F(x_t) + (1-\alpha)(\vartheta_{t-1} - \gamma\nabla f(x_{t-1}))\\
x_{t+1}^{(0)} &= \nabla h^*(\vartheta_t^{(0)} - \gamma\nabla f(x_t))\\
\vartheta_t &= \begin{cases}\vartheta_t^{(0)}, & \text{if } \gamma D_f(x_{t+1}^{(0)}, x_t) \leqslant D_h(x_{t+1}^{(0)}, x_t; \vartheta_t^{(0)})\\ \vartheta_{t-1} - \gamma\nabla f(x_{t-1}), & \text{otherwise,}\end{cases}\\
x_{t+1} &= \nabla h^*(\vartheta_t - \gamma\nabla f(x_t)).
\end{aligned}$$

Note that at a given step, the IPDD algorithm first computes an iteration which interpolates DA and MD if condition (17) is satisfied; additionally, a DA iteration is computed if condition (17) is not satisfied. Thus, in the worst case, the algorithm computes two UMD iterations per step.

Remark 5.3. Clearly, the use of the APDD and IPDD algorithms makes no sense in situations where mirror descent and dual averaging produce identical iterates, which is the case, in particular, when the problem is unconstrained. Even in a constrained case, when the algorithm iterates belong to the interior of $\mathcal{X}$ (e.g., when the problem solution is an interior point of $\mathcal{X}$), the mirror descent and dual averaging update will end up coinciding and the APDD and IPDD algorithms will no longer provide any improvement. However, the situation changes dramatically when the iterates of the method lie at the boundary of the feasible set.

From now on, let us consider a convex optimization problem with minimizer at the boundary of the feasible set $\mathcal{X}$. For simplicity, we discuss the Euclidean instances of the algorithms, meaning that the associated mirror map (resp. regularizer) is $F = \frac{1}{2}\|\cdot\|_2^2$ (resp. $h = F + I_{\mathcal{X}}$).

As the algorithm iterates reach the boundary of the feasible set, possible UMD iterates differ. Informally, with dual increments $\xi_t = -\gamma\nabla f(x_t)$ in the normal cone of $\mathcal{X}$ at x_t, the dual iterate ϑ_t of DA (the iterate "before the projection") gets farther and farther away from $\mathcal{X}$. By contrast, the dual iterate ϑ_t of MD always belongs to the feasible set, by definition. Consequently, successive primal DA iterates becomes closer to each other and more conservative, compared to the MD iterates which vary more aggressively. As a result, MD with large step-sizes produces iterates which vary too much (thus compromising convergence), whereas variations of DA iterates with large step-sizes are quickly attenuated exactly because the step-sizes are large (and dual iterates are "far from $\mathcal{X}$"), implying that dual iterates ϑ_t are getting away from $\mathcal{X}$ even faster. This difference is clearly observed in the numerical experiments below.

An alternative intuition on the robustness of DA to the choice of step-sizes is as follows. As stated in Theorem 4.3, in the case of the smooth objective function, accuracy bounds for the UMD still hold if condition (17) is satisfied for all $t \geqslant 1$. Observe that

$$D_h(x_{t+1}, x_t;\ \theta_t) = D_F(x_{t+1}, x_t) + \langle \nabla F(x_t) - \vartheta_t | x_{t+1} - x_t\rangle\,.$$

In the Euclidean case, when $\nabla F(x_t) = x_t$ and x_{t+1} are in $\mathcal{X}$, the farther ϑ_t from $\mathcal{X}$, the larger the above quantity. As ϑ_t gets farther and farther away from $\mathcal{X}$, the ratio $D_h(x_{t+1}, x_t;\ \vartheta_t)/D_f(x_{t+1}, x_t)$ becomes large, and condition (17) for DA iterates is satisfied with a large step-size γ.

The k-APDD algorithm mostly performs DA iterations, and MD iterations once in a while. Each MD iteration brings the dual point ϑ_t back to the feasible set $\mathcal{X}$, which has the effect of making subsequent iterations more aggressive. When the step-size is too large, these iterations may lead to a temporary increase of the objective value, but the following DA iterations push the dual point ϑ_t away from $\mathcal{X}$, recovering proper convergence in most cases.

The α-IPDD algorithm performs a "moderately aggressive iteration" most of the time which is an interpolation between DA and MD iterations, with coefficient $\alpha \in (0, 1]$. When such interpolation is too aggressive for maintaining proper convergence, a DA iteration is performed instead, pushing the dual point ϑ_t further away from $\mathcal{X}$, thus making subsequent iterations more conservative. Occasionally, aggressive iterations may lead to a temporary increase of the objective value, which is always less significant than in the case of APDD. In the numerical experiments we report on below, we choose $\alpha = .1$ which seems to be a good trade-off between aggressiveness (for achieving fast convergence) and conservativeness (for maintaining the dual point ϑ_t in a region were convergence is sustainable).

5.3. **Numerical experiments.** To illustrate the numerical performance of the proposed algorithms, we present here results of a preliminary computational experiment involving comparison of APDD and IPDD with existing algorithms in two optimization problems arising in the statistical treatment of large datasets.

5.3.1. *Least-squares regression.* We consider a least-squares regression problem using the training sample of the BlogFeedback dataset from the UCI Machine Learning Repository[8] with rescaled—multiplied by 0.005—features (regressors). The corresponding least-squares problem with dimensions $n = 52397$ and $d = 280$ writes

$$\min_{x \in \mathcal{X}} \left\{ f(x) = \frac{1}{n} \sum_{i=1}^{n} (b_i - \langle x | a_i \rangle)^2 \right\}$$

where

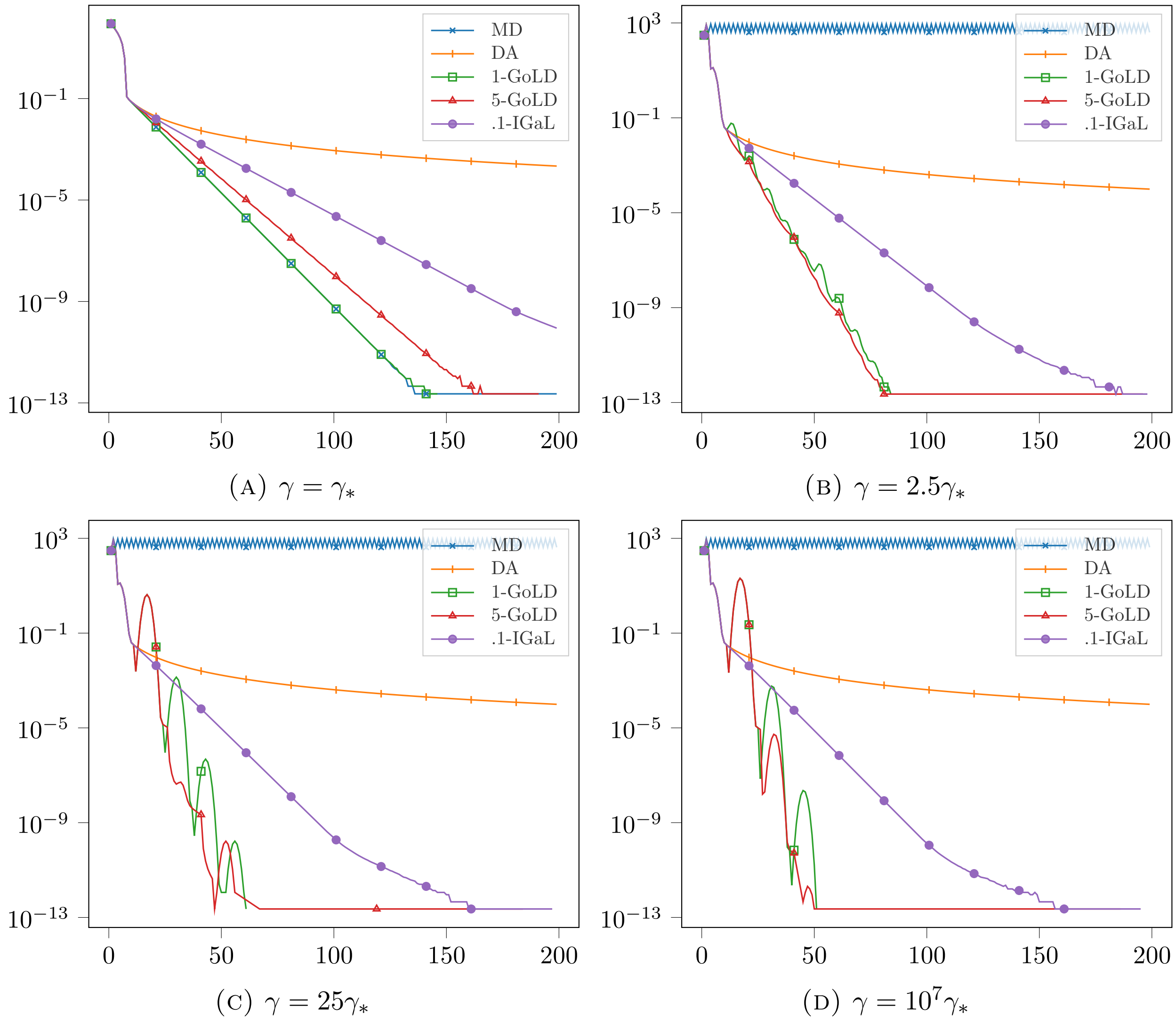
$$\mathcal{X} = \left\{ x \in \mathbb{R}^d : \|x\|_2 \leqslant 1 \right\}$$

The spectral norm of the matrix $Q = \frac{1}{n} \sum_{i=1}^{n} a_i a_i^T$ (the Lipschitz constant of the gradient of the objective function) is $\lambda_{\max}(Q) = 571.34$; the matrix is ill conditioned with only 20 eigenvalues exceeding $10^{-6} \lambda_{\max}(Q)$. A high accuracy approximation of the optimal value f_* (with optimal solution at the boundary of the feasible set) is first computed with a long run of projected gradient descent with manually tuned step-size.

We compute approximate solutions by the following algorithms: mirror descent (MD), dual averaging (DA), 1-APDD, 5-APDD, and .1-IPDD. We consider the Euclidean setting which corresponds to the mirror map $F = \frac{1}{2} \| \cdot \|_2^2$ and regularizer $h = F + I_{\mathcal{X}}$. Iterations are initialized with $\vartheta_1 = 0$ and $x_1 = \nabla h^*(\vartheta_1) = 0$, we run $T = 200$ steps of each methods with the constant step-size. The results of the experiment are presented in Figure 6 where we plot the evolution of the suboptimality gap $f(x_t) - f_*$ for several values of the step-sizes. When the step-size is less than its "theoretically justified" value $\gamma_* = (\lambda_{\max}(Q))^{-1} = 0.0018$ all algorithms converge slowly. For $\gamma = \gamma_*$, all algorithms except DA converge linearly, with MD and 1-APDD achieving the fastest convergence. MD does not converge for $\gamma \geqslant 10\gamma_*$, DA converges sublinearly, and the APDD and the IPDD algorithms exhibit linear convergence (with APDD showing temporal increase in the objective value); convergence of the IPDD algorithm seems to be unaffected by the large value of the step-size.

[8] https://archive.ics.uci.edu/ml/datasets/BlogFeedback

FIGURE 6. Least-squares regression: evolution of suboptimality of approximate solution by mirror descent, dual averaging, 1-APDD, 5-APDD, and IPDD algorithms for different values of γ.



(A) $\gamma = \gamma_*$

(B) $\gamma = 2.5\gamma_*$

(C) $\gamma = 25\gamma_*$

(D) $\gamma = 10^7\gamma_*$

5.3.2. *Logistic regression.* We consider a logistic regression problem using the training sample of the Madelon dataset from the UCI Machine Learning Repository[9] with all features multiplied by 10^{-3}. The corresponding minimization problem of dimensions $n = 2000$ and $d = 500$ writes

$$\min_{x\in\mathcal{X}} \left\{ f(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-b_i\langle x|a_i\rangle}\right) \right\}$$

with the same as in the preceding example feasible set

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d : \|x\|_2 \leqslant 1 \right\}.$$

[9] https://archive.ics.uci.edu/ml/datasets/Madelon

In the present case, the spectral norm of the matrix $Q = \frac{1}{n}\sum_{i=1}^{n} a_i a_i^T$ (the Lipschitz constant of the gradient of the objective function) is $\lambda_{\max}(Q) = 119.16$. An estimate of the optimal value f_* (with optimal solution at the boundary of the feasible set) is first computed with a long run of projected gradient descent with manually tuned step-size.

We consider the following algorithms: mirror descent (MD), dual averaging (DA), 20-APDD, 20-7-APDD,[10] and .1-IPDD. As in Section 5.3.1, we consider the Euclidean setup of the problem, with the mirror map $F = \frac{1}{2}\|\cdot\|_2^2$ and regularizer $h = F + I_{\mathcal{X}}$. Iterations are initialized with $\vartheta_1 = 0$ and $x_1 = \nabla h^*(\vartheta_1) = 0$. We run $T = 5000$ steps of each algorithm for several values of constant step-sizes; the results are presented in Figure 8.

In view of the problem parameters, the "theoretically justified" choice of the step-size is $\gamma_* = (\lambda_{\max}(Q))^{-1} = 0.0084$. In our experiments, we observe linear convergence of all algorithms except for DA when the algorithm step-size $\gamma < 0.07 (\approx 8\gamma_*)$. MD does not converge for $\gamma = 0.07$, DA converges sublinearly, the remaining algorithms achieving the same linear convergence. When $\gamma \geqslant .1$, 20-APDD and 20-7-APDD algorithms start to oscillate without converging, while DA converges sublinearly, and IPDD continues exhibiting linear convergence. We observe yet another regime for larger step-sizes, e.g. $\gamma = 1$ or $\gamma = 200$: MD does not converge, IPDD converging linearly, 20-APDD still oscillating, but, surprisingly, 20-7-APDD enjoying the same sublinear convergence as DA.

The observed results may be summarized as follows: in our experiments, MD converges linearly only when run with a properly selected step-size, while DA always converges sublinearly. The k-ℓ-APDD algorithm is robust w.r.t. the step-size choice, it converges linearly in a wide range of step-sizes, and, same as DA, enjoys sublinear convergence for large step-sizes. The IPDD algorithm converges linearly and seems to be insensitive to the choice of step-size.

Acknowledgements

The authors are grateful to Roberto Cominetti, Cristóbal Guzmán, Nicolas Flammarion and Sylvain Sorin for inspiring discussions and suggestions. A. Juditsky was supported by MIAI @ Grenoble Alpes (`ANR-19-P3IA-0003`). J. Kwon was supported by a public grant as part of the "Investissement d'avenir" project (`ANR-11-LABX-0056-LMH`), LabEx LMH.

References

[1] 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.

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

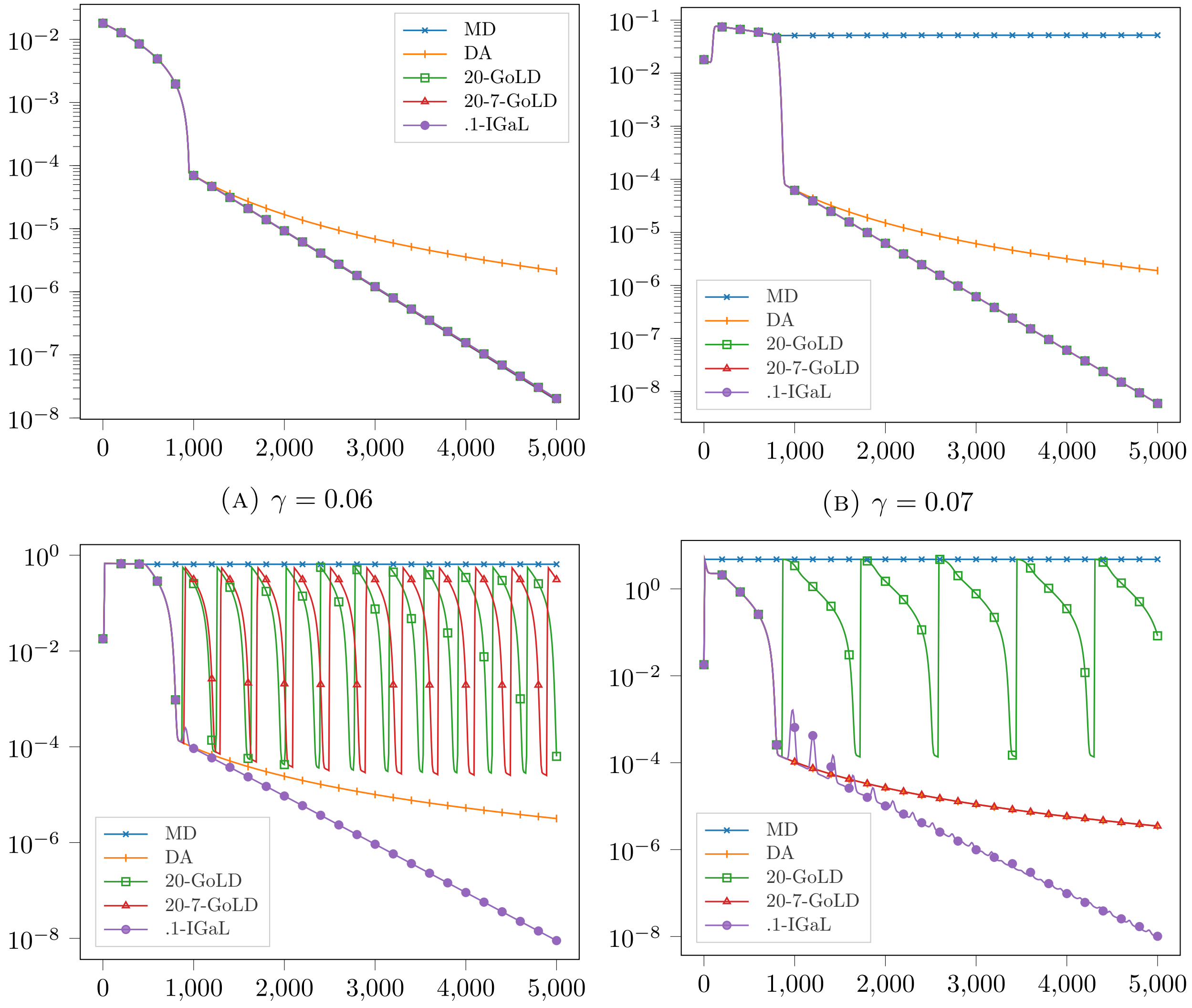
[3] Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2013.

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

[5] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[10] The second parameter in the definition of the k-ℓ-ADPP corresponds to the ℓ-step ahead computation of the objective when determining the choice of update every k steps of the algorithm.

FIGURE 8. Logistic regression: evolution of suboptimality of approximate solution by mirror descent, dual averaging, 20-APDD, 20-7-APDD and .1-IPDD algorithms for different values of γ.



(A) $\gamma = 0.06$

(B) $\gamma = 0.07$

(C) $\gamma = 0.1$

(D) $\gamma = 200$

[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] 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.

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

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

[15] Sanjoy Dasgupta and Matus J. Telgarsky. Agglomerative Bregman clustering. In *Proceedings of the 29th International Conference on Machine Learning (ICML 12)*, pages 1527–1534, 2012.

[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 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.

[18] John C. Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, to appear.

[19] 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.

[20] 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.

[21] Anatoli Juditsky and Arkadi Nemirovski. First order methods for nonsmooth convex large-scale optimization, II: utilizing problems structure. *Optimization for Machine Learning*, pages 149–183, 2011.

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

[23] 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.

[24] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

[25] 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.

[26] 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.

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

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

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

[30] 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.

[31] 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.

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

[33] Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[34] Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

[35] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

[36] Yurii Nesterov and Vladimir Shikhman. Quasi-monotone subgradient methods for nonsmooth convex minimization. *Journal of Optimization Theory and Applications*, 165(3):917–940, 2015.

[37] Alexander Rakhlin and Ambuj Tewari. Lecture notes on online learning. 2009.

[38] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[39] Shai Shalev-Shwartz. *Online learning: Theory, algorithms, and applications*. PhD thesis, The Hebrew University of Jerusalem, 2007.

[40] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

[41] 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.

Definition A.2 (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 strongly convex with modulus κ with respect to norm $\|\cdot\|$ if for all $x, x' \in \mathbb{R}^n$ and $\lambda \in [0,1]$,

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

Proposition A.3 (Theorem 23.5 in [38]). *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) $x \in \operatorname{Arg\,max}_{x' \in \mathbb{R}^n} \{\langle y|x'\rangle - g(x')\}$;
(v) $y \in \operatorname{Arg\,max}_{y' \in \mathbb{R}^n} \{\langle y'|x\rangle - g^*(y')\}$.

Appendix B. Postponed proofs

B.1. Proofs for Section 2.

B.1.1. *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

$$x_1 \in \partial F^*(\vartheta) \quad \Longleftrightarrow \quad \vartheta = \nabla F(x_1) \quad \Longleftrightarrow \quad x_1 \text{ minimizer of } \phi_\vartheta, \tag{18}$$

where the first equivalence comes from Proposition A.3. 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

$$\nabla F^*(\vartheta) = x_1 \in \mathcal{D}_F. \tag{19}$$

First, the above (19) proves property (ii). Second, this equality combined with the equality from (18) 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

$$\vartheta := \nabla F(x), \tag{20}$$

and all the above holds with this special point ϑ. In particular, $x_1 = x$ by uniqueness of x_1. Therefore (19) gives

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

and this proves $\nabla F^*(\mathbb{R}^n) \supset \mathcal{D}_F$ and thus property (iii). Combining (20) and (21) gives the first identity from property (iv).

B.1.2. *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: consequently, 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 $x_0 + \frac{x_{\phi(k)} - x_0}{\left\|x_{\phi(k)} - x_0\right\|}$ 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

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

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. Consequently, $D_F(\,\cdot\,, x_0)$ 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 continuous 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

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

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[11] $\mathcal{D}_F$ as the convex combination of two points in $\mathcal{D}_F$. Therefore, $(x_*, x_1]$ is indeed a subset of $\mathcal{D}_F$.

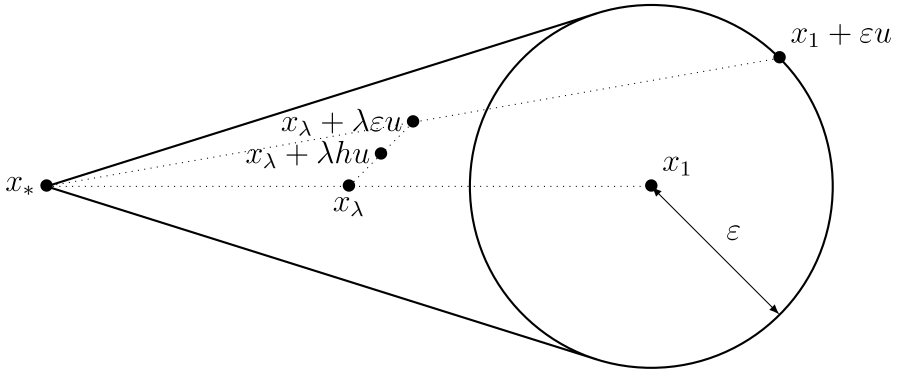
G being differentiable on $\mathcal{D}_F$ by definition of the mirror map, the gradient of G exists at each point of $(x_*, x_1]$. Let us prove that ∇G is bounded on $(x_*, x_1]$. Let $x_\lambda \in (x_*, x_1]$, where $\lambda \in (0, 1]$ is such that

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

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 $\mathcal{C}$:

$$x_\lambda + \lambda \varepsilon u = (1 - \lambda)x_* + \lambda(x_1 + \varepsilon u) \in \mathcal{C}. \tag{22}$$

[11] 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.



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

$$x_\lambda + \lambda h u = \left(1 - \frac{h}{\varepsilon}\right) x_\lambda + \frac{h}{\varepsilon}\left(x_\lambda + \lambda\varepsilon u\right) \in \mathcal{C}. \tag{23}$$

Now using for G the convexity inequality associated with the convex combination from (23), we write:

$$\begin{aligned} G(x_\lambda + h\lambda u) - G(x_\lambda) &\leqslant \frac{h}{\varepsilon}\left(G(x_\lambda + \lambda\varepsilon u) - G(x_\lambda)\right) \\ &= \frac{h}{\varepsilon}\left(G(x_\lambda + \lambda\varepsilon u) - G(x_*) + G(x_*) - G(x_\lambda)\right) \\ &\leqslant \frac{h}{\varepsilon}\left(G(x_\lambda + \lambda\varepsilon u) - G(x_*)\right), \end{aligned} \tag{24}$$

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 (22), we also write

$$\begin{aligned} G(x_\lambda + \lambda\varepsilon u) - G(x_*) &\leqslant \lambda\left(G(x_1 + \varepsilon u) - G(x_*)\right) \\ &\leqslant \lambda\left(G_{\max} - G_{\min}\right). \end{aligned} \tag{25}$$

Combining (24) and (25) 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:

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

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 $x_* \in \mathcal{D}_F$.

B.1.3. *Proof of Proposition 2.9.* 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.3, $h^*(\vartheta) = \langle \vartheta | x \rangle - h(x) < +\infty$.

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

B.1.4. *Proof of Proposition 2.10.* Let $\vartheta \in \mathbb{R}^n$. Because $\operatorname{dom} h^* = \mathbb{R}^n$, the subdifferential $\partial h^*(\vartheta)$ is nonempty—see e.g. [38, Theorem 23.4]. By Proposition A.3, $\partial h^*(\vartheta)$ 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.

Let $x \in \mathcal{D}_F$ and let us prove that $\nabla F(x) \in \partial h(x)$. By convexity of F, the following is true

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

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

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

In other words, $\nabla F(x) \in \partial f(x)$.

B.1.5. *Proof of Proposition 2.11.* 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 $\mathcal{X} \cap \mathcal{D}_F \neq \emptyset$. 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 neighborhood $U \subset \mathcal{D}_F$ 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 neighborhood 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 an $\mathcal{X}$-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 an $\mathcal{X}$-regularizer.

B.2. **Proofs for Section 4.**

B.2.1. *Proof of Proposition 4.2.* 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, utilizing the convexity of f, we get

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

Besides this, for $t = 1$, we have $\gamma_1 \langle f'(y_1) | x_1 - x_* \rangle \geqslant \gamma_1 (f(y_1) - f_*)$ because $x_1 = y_1$ by definition. Then, summing over $t = 1, \dots, T$, we obtain after simplifications:

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

Using the definition of coefficients ν_t, the above left-hand side simplifies to result in the inequality

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

Finally, because $(x_t, \vartheta_t)_{t \geqslant 1}$ is a sequence of $\mathrm{UMD}(h, \xi)$ iterates with dual increments $\xi := (-\gamma_t f'(y_t))_{t \geqslant 1}$, the result then follows by applying inequality (8) from Corollary 3.13 and dividing by $\sum_{t=1}^{T} \gamma_t$. □

B.2.2. *Proof of Theorem 4.3.* First, observe that whenever $\gamma_t \leqslant 1/L$, due to (11),

$$\begin{aligned} f(x_{t+1}) - f(x_t) - \langle \nabla f(x_t) | x_{t+1} - x_t \rangle &\leqslant \frac{L}{2} \|x_{t+1} - x_t\|^2 \\ &\leqslant (2\gamma_t)^{-1} \|x_{t+1} - x_t\|^2. \end{aligned} \tag{26}$$

Thus,

$$\begin{aligned} \gamma_t D_f(x_{t+1}, x_t) &= \gamma_t [f(x_{t+1}) - f(x_t) - \langle f'(x_t) | x_{t+1} - x_t \rangle] \\ &\leqslant \tfrac{1}{2} \|x_{t+1} - x_t\|^2 \\ &\leqslant D_h(x_{t+1}, x_t;\ \vartheta_t). \end{aligned}$$

by the strong convexity of D_h. On the other hand, by (6) of Lemma 3.11, for any $x \in \mathcal{X} \cap \operatorname{dom} h$,

$$\begin{aligned} D_h(x, x_{t+1}; \vartheta_{t+1}) &\leqslant D_h(x, x_t; \vartheta_t) + \gamma_t \langle \nabla f(x_t) | x - x_{t+1} \rangle - D_h(x_{t+1}, x_t; \vartheta_t) \\ [\text{by } (12)] &\leqslant D_h(x, x_t; \vartheta_t) + \gamma_t \langle \nabla f(x_t) | x - x_t \rangle \\ &\quad - \gamma_t \langle \nabla f(x_t) | x_{t+1} - x_t \rangle - D_f(x_{t+1}, x_t) \\ [\text{due to } (26)] &\leqslant D_h(x, x_t; \vartheta_t) + \gamma_t \langle \nabla f(x_t) | x - x_t \rangle - \gamma_t [f(x_{t+1} - f(x_t)] \\ [\text{by convexity of } f] &\leqslant D_h(x, x_t; \vartheta_t) - \gamma_t (f(x_{t+1}) - f(x)). \end{aligned}$$

Consequently, $\forall x \in \mathcal{X} \cap \operatorname{dom} h$,

$$\gamma_t (f(x_{t+1}) - f(x_t)) \leqslant D_h(x, x_t; \vartheta_t) - D_h(x, x_{t+1}; \vartheta_{t+1}).$$

When applying the above inequality to $x = x_t$ we conclude that

$$\gamma_t (f(x_{t+1}) - f(x_t)) \leqslant -D_h(x_t, x_{t+1}; \vartheta_{t+1}) \leqslant 0.$$

Finally, when setting $x = x_*$, we obtain

$$\left(\sum_{t=1}^{T} \gamma_t \right) (f(x_{T+1}) - f_*) \leqslant \sum_{t=1}^{T} \gamma_t (f(x_{t+1}) - f(x_*)) \leqslant D_h(x_*, x_1; \vartheta_1)$$

which implies (13). □

B.2.3. *Proof of Theorem 4.4.* We start with the following technical result.

Lemma B.1. *Assume that positive step-sizes $\nu_t \in (0,1]$ and $\gamma_t > 0$ are such that the relationship*

$$f(z_{t+1}) \leqslant f(y_t) + \nu_t \langle \nabla f(y_t) | x_{t+1} - x_t \rangle + \frac{\nu_t}{\gamma_t} D_h(x_{t+1}, x_t; \vartheta_t), \tag{27}$$

holds for all t which is certainly the case if $\nu_t \gamma_t \leqslant L^{-1}$. Denote $s_t = f(z_t) - f_*$; then

$$\gamma_t \nu_t^{-1}(s_{t+1} - s_t) + \gamma_t s_t \leqslant D_h(x_*, x_t; \vartheta_t) - D_h(x_*, x_{t+1}; \vartheta_{t+1}). \tag{28}$$

Proof of the lemma. Observe first that by construction,

$$z_{t+1} - y_t = (1-\nu_t) z_t + \nu_t x_{t+1} - [(1-\nu_t) z_t + \nu_t x_t] = \nu_t (x_{t+1} - x_t)$$

By strong convexity of h, for $\nu_t \gamma_t \leqslant L^{-1}$ we have

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

what is (27).

Next, observe that by (14a),

$$\nu_t (x_* - x_t) = (\nu_t x_* + (1-\nu_t) z_t) - y_t,$$

whence, by convexity of f,

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

When substituting the latter bound into (27) we get

$$\begin{aligned} f(z_{t+1}) \leqslant f(y_t) &+ \nu_t \langle \nabla f(y_t) | x_{t+1} - x_* \rangle + \nu_t (f(x_*) - f(y_t)) \\ &+ (1-\nu_t)(f(z_t) - f(y_t)) + \frac{\nu_t}{\gamma_t} D_h(x_{t+1}, x_t; \vartheta_t), \end{aligned}$$

or

$$f(z_{t+1}) - f(z_t) \leqslant \nu_t \langle \nabla f(y_t) | x_{t+1} - x_* \rangle + \nu_t (f_* - f(z_t)) + \frac{\nu_t}{\gamma_t} D_h(x_{t+1}, x_t; \vartheta_t).$$

Now, because $(x_t, \vartheta_t)_{t \geqslant 1}$ is a sequence of UMD iterates, by (6) of Lemma 3.11,

$$\gamma_t \langle \nabla f(y_t) | x_{t+1} - x_* \rangle \leqslant D_h(x_*, x_t; \vartheta_t) - D_h(x_*, x_{t+1}; \vartheta_{t+1}) - D_h(x_{t+1}, x_t; \vartheta_t),$$

and we arrive at

$$\gamma_t \nu_t^{-1} (f(z_{t+1}) - f(z_t)) \leqslant D_h(x_*, x_t; \vartheta_t) - D_h(x_*, x_{t+1}; \vartheta_{t+1}) + \gamma_t (f_* - f(z_t)),$$

what is (28). □

Proof of the theorem. Assume that ν_t and γ_t satisfy

$$\nu_1 = 1, \quad \nu_t \in (0, 1], \quad \gamma_{t+1}(\nu_{t+1}^{-1} - 1) \leqslant \gamma_t \nu_t^{-1}. \tag{29}$$

When summing (28) up from 1 to T we get

$$\begin{aligned} D_h(x_*, x_t; \vartheta_1) &\geqslant \sum_{t=1}^{T} [\gamma_t \nu_t^{-1}(s_{t+1} - s_t) + \gamma_t s_t] \\ &= \gamma_T \nu_T^{-1} s_{T+1} + \sum_{t=2}^{T} s_t \left(\gamma_{t-1}\nu_{t-1}^{-1} - \gamma_t(\nu_t^{-1} - 1)\right) - \gamma_1(\nu_1^{-1} - 1)s_1 \\ &\underbrace{\geqslant}_{[\text{by } (29)]} \gamma_T \nu_T^{-1} s_{T+1} = \gamma_T \nu_T^{-1}(f(z_{T+1}) - f_*). \end{aligned}$$

It is clear that the choice of $\gamma_1 = L^{-1}$, $\nu_1 = 1$ and $\nu_t = (\gamma_t L)^{-1}$ satisfies the relationship $\gamma_t \nu_t \leqslant L^{-1}$. In this case, when choosing step-sizes $(\gamma_t)_{t \geqslant 1}$ to saturate recursively the last relation in (29), specifically,

$$\gamma_{t+1}^2 L - \gamma_{t+1} = \gamma_t^2 L$$

we come to celebrated Nesterov step-sizes (15) which satisfy $\gamma_t \nu_t^{-1} \geqslant \frac{(t+1)^2}{4L}$, and we arrive at (16). □