Available online at www.sciencedirect.com
SCIENCE DIRECT®
Operations Research Letters 31 (2003) 167–175
Operations Research Letters
www.elsevier.com/locate/dsw

Mirror descent and nonlinear projected subgradient methods for convex optimization

Amir Beck, Marc Teboulle*

School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel

Received 20 August 2002; received in revised form 28 October 2002; accepted 31 October 2002

Abstract

The mirror descent algorithm (MDA) was introduced by Nemirovsky and Yudin for solving convex optimization problems. This method exhibits an efficiency estimate that is mildly dependent in the decision variables dimension, and thus suitable for solving very large scale optimization problems. We present a new derivation and analysis of this algorithm. We show that the MDA can be viewed as a nonlinear projected-subgradient type method, derived from using a general distance-like function instead of the usual Euclidean squared distance. Within this interpretation, we derive in a simple way convergence and efficiency estimates. We then propose an Entropic mirror descent algorithm for convex minimization over the unit simplex, with a global efficiency estimate proven to be mildly dependent in the dimension of the problem.

Keywords: Nonsmooth convex minimization; Projected subgradient methods; Nonlinear projections; Mirror descent algorithms; Relative entropy; Complexity analysis; Global rate of convergence

1. Introduction

Consider the following nonsmooth convex minimization problem,

(P) minimize $f(x)$ s.t. $x \in X \subset \mathbb{R}^n$.

Throughout the paper we make the following assumptions on problem (P):

Assumption A.

(a) X is a closed convex subset in $\mathbb{R}^n$ with nonempty interior.

(b) The objective function $f : X \to \mathbb{R}$ is a convex Lipschitz continuous function with Lipschitz constant L_f with respect to a fixed given norm $\|\cdot\|$, i.e., $|f(x) - f(y)| \leqslant L_f\|x - y\|$, $\forall x, y \in X$.

(c) The optimal set of (P) denoted by X^* is nonempty.

(d) A subgradient of f at $x \in X$ is computable. An element of the subdifferential $\partial f(x)$ is denoted by $f'(x)$.

We are interested in finding an approximate solution to problem (P), within $\varepsilon > 0$, i.e., to find $x \in X$ such that

$$f(x) - f^* := f(x) - \min_{x \in X} f(x) \leqslant \varepsilon.$$

A standard method to solve (P) is the subgradient projection algorithm, (see e.g. [2] and references

* Corresponding author.
E-mail addresses: becka@post.tau.ac.il (A. Beck), teboulle@post.tau.ac.il (M. Teboulle).

0167-6377/03/$ - see front matter
PII: S0167-6377(02)00231-6

therein), which generates iteratively the sequence $\{x^k\}$ via

$$x_{k+1} = \pi_X(x^k - t_k f'(x^k)),$$
$$t_k > 0 \text{ (a stepsize)}, \tag{1.1}$$

where $\pi_X(x) = \operatorname{argmin}\{\|x - y\| \mid y \in X\}$ is the Euclidean projection onto X.

The key advantage of the subgradient algorithm is its simplicity, provided that projections can be easily computed, which is the case when the constraints set X is described by simple sets, e.g., hyperplanes, balls, bound constraints, etc. Its main drawback is that it has a very slow rate of convergence. Indeed, consider the convex problem (P) with X convex compact, and denote by $\operatorname{Diam}(X)$ the diameter of X, i.e., $\operatorname{Diam}(X) := \max_{x,y\in X} \|x - y\| < \infty$. Then, the optimal efficiency estimate for the subgradient method with stepsizes $t_k = \operatorname{Diam}(X)k^{-1/2}$, $k = 1, \ldots,$ is (see [10]):

$$\min_{1\leqslant s\leqslant k} f(x^s) - \min_{x\in X} f(x)$$
$$\leqslant \mathrm{O}(1)L_f \operatorname{Diam}(X)k^{-1/2}, \tag{1.2}$$

where O(1) stands for a positive absolute constant. Thus, like all gradient based methods, one can obtain in a very small number of iterations a *low accuracy* optimal value, (say one or two digits) but then within further iterations no more progress in accuracy can be achieved and the method is essentially jamming. However, a key feature of gradient methods is also the fact that while their rate of convergence is very slow, the rate is *almost independent* of the dimension of the problem. In contrast to this, more efficient sophisticated algorithms, such as for example interior point-based methods, which require for example at each iteration Newton-type computations, i.e., the solution of a linear system, are often defeated even for problems with a few thousands of variables, and a fortiori for very large-scale nonsmooth problems. Therefore, for constrained problems where low accurate solutions is sufficient and the dimension is huge, gradient type methods appear as natural candidates for developing potential practical algorithms. The recent paper of Ben-Tal et al. [1] on computerized tomography demonstrates very well this situation through an algorithm based on the *mirror descent algorithm* (MDA for short) introduced by Nemirovski and Yudin [10]. It is shown there that it is possible to solve efficiently a convex minimization problem over the unit simplex, with millions of variables.

Motivated by the recent work of Ben-Tal et al. [1], in this paper we concentrate on the analysis of the basic steps of the (MDA) which is recalled in Section 2. We show in Section 3, that the (MDA) can be viewed as a simple *nonlinear subgradient projection* method, where the usual Euclidean projection operator is replaced by a nonlinear/nonorthogonal type projection operator based on a Bregman-like distance function (see e.g. [3,4,14] and references therein). With this new interpretation of the (MDA), we derive in a simple and systematic way convergence proofs and efficiency estimates, see Section 4. In Section 5 we concentrate on optimization problems over the unit simplex and propose a new algorithm called the entropic mirror descent algorithm (EMDA). The EMDA is proven to exhibit an efficiency estimate which is almost independent in the dimension n of the problem and in fact shares the same properties of an algorithm proposed in [1] for the same class of problems, but is given explicitly by a simple formula. Finally, in the last section we outline some potential applications and extensions for further work.

2. The mirror descent algorithm (MDA)

The idea of the algorithm is based on dealing with the structure of the Euclidean norm rather than with local behavior of the objective function in problem (P). Roughly speaking, the method originated from functional analytic arguments arising within the infinite dimensional setting, between primal and dual spaces. The mathematical objects associated with f and x are not vectors from a vector space E, but elements of the dual vector space to E, which consists of linear forms on E. The Euclidean structure is not the only way to identify the primal-dual spaces, and it is possible to identify the primal and dual spaces within a wider family which includes as particular case, the classical Euclidean structures. This idea and approach was introduced by Nemirovsky and Yudin [10], and the reader is refereed to their book for a more detailed motivation and explanations. We will show below, that there is a much simpler and easy way to motivate, explain, and construct the MDA. For now,

let us consider the basic steps involve in the original MDA.

Consider the problem (P) satisfying Assumption A. The (MDA) further assumes the following objects, which can be freely chosen as long as they satisfy the following hypothesis:

- Fix any norm $\|\cdot\|$ in $\mathbb{R}^n$ (which will play a role in the choice of the stepsize).
- Let $\psi : X \to \mathbb{R}$ be a continuously differentiable and strongly convex function on X with strong convexity parameter $\sigma > 0$.
- The conjugate of ψ, defined by $\psi^*(y) = \max_{x\in X}\{\langle x, y\rangle - \psi(x)\}$ is assumed to be easily computable.

The basic steps of the MDA can be described as follows, see [10,1] (for comparison the set Y there is set to be equal to X in [1, p. 84]).

MDA: Start with $y^1 \in \operatorname{dom}\nabla\psi^*$ and generate the sequence $\{x^k\} \in X$ via the iterations

$$x^k = \nabla\psi^*(y^k), \tag{2.3}$$

$$y^{k+1} = \nabla\psi(x^k) - t_k f'(x^k), \tag{2.4}$$

$$\begin{aligned} x_{k+1} &= \nabla\psi^*(y^{k+1}) \\ &= \nabla\psi^*(\nabla\psi(x^k) - t_k f'(x^k)), \end{aligned} \tag{2.5}$$

where $t_k > 0$ are appropriate step sizes.

The method looks at this stage somewhat hard to understand or even to motivate (besides the very rough explanation given above). In the next section we will give a very simple interpretation which will explain and reveal the structure of this algorithm. In the mean time, let us consider a basic example which clearly indicates that the MDA appears to be as a natural generalization of the subgradient algorithm.

Example 1. Let $\|\cdot\|$ be the usual l_2 norm in $\mathbb{R}^n$ and let $\psi(x) := \frac{1}{2}\|x\|^2$ for $x \in X$ and $+\infty$ for $x \notin X$. The function ψ is clearly proper, lsc and strongly convex with parameter $\sigma = 1$, and continuously differentiable on X. A straightforward computation shows that the conjugate of ψ is given by $\psi^* : \mathbb{R}^n \to \mathbb{R}$

$$\psi^*(z) = \tfrac{1}{2}(\|z\|^2 - \|z - \pi_X(z)\|^2)$$

with $\nabla\psi^*(z) = \pi_X(z)$. Indeed, since $\partial\psi(x) = (I + N_X)(x)$, where N_X denotes the normal cone of the closed convex set X, using the well known relations $(I + N_X)^{-1} = \pi_X$ and $(\partial\psi)^{-1} = \partial\psi^*$, (see [11]), one thus has

$$\begin{aligned} z \in \partial\psi(x) \quad &\Leftrightarrow \quad x = (I + N_X)^{-1}(z) \\ &= \pi_X(z) = \nabla\psi^*(z). \end{aligned}$$

Therefore, the (MDA) yields

$$x^k = \pi_X(y^k), \tag{2.6}$$

$$y^{k+1} = x^k - t_k f'(x^k), \tag{2.7}$$

$$x^{k+1} = \pi_X(x^k - t_k f'(x^k)), \tag{2.8}$$

i.e., we have recovered the subgradient projection algorithm.

3. Nonlinear projection methods

It is well known (see e.g. [2]) that the subgradient algorithm can be viewed as *linearization* of the so-called proximal algorithm [12], (or as an explicit scheme of the corresponding subdifferential inclusion). Indeed, it is immediate to verify that the projected subgradient iteration (1.1) can be rewritten equivalently as

$$x^{k+1} \in \operatorname*{argmin}_{x\in X}\left\{\langle x, f'(x^k)\rangle + \frac{1}{2t_k}\|x - x^k\|^2\right\}.$$

In [14], it has been shown that more general proximal maps can be considered by replacing the usual Euclidean quadratic norms with some sort of more general distance-like functions. As explained there, the principal motivation for such kind of distances is to be able to use one which reflects the geometry of the given constraints set X, so that in particular with such an appropriate choice, the constraints can often be *automatically eliminated*. In a similar way, we can thus construct nonlinear projection subgradient methods, by considering iteration schemes of the form

$$x^{k+1} \in \operatorname*{argmin}_{x\in X}\left\{\langle x, f'(x^k)\rangle + \frac{1}{t_k}D(x, x^k)\right\}, \tag{3.9}$$

where $D(u, v)$ replaces $2^{-1}\|u - v\|^2$, and should verify the property $D(u, v)$ is nonnegative, and $D(u, v) = 0$ if and only if $u = v$. We prove below, that the MDA

is nothing else, but the nonlinear subgradient projection method (3.9), with a particular choice of D based on a Bregman-like distance generated by a function ψ. Note, that the hypothesis on D will be somewhat different from the usual Bregman based distances assumed in the literature (see e.g. [8,14], and references therein).

Let $\psi : X \to \mathbb{R}$ be strongly convex and continuously differentiable on $\operatorname{int} X$. The distance-like function is defined by $B_\psi : X \times \operatorname{int}(X) \to \mathbb{R}$ given by

$$B_\psi(x, y) = \psi(x) - \psi(y) - \langle x - y, \nabla\psi(y)\rangle. \tag{3.10}$$

The basic subgradient algorithm based on B_ψ is as follows.

Subgradient algorithm with nonlinear projections (*SANP*): Given B_ψ as defined in (3.10) with ψ as above, start with $x_1 \in \operatorname{int} X$, and generate the sequence $\{x^k\}$ via the iteration

$$x^{k+1} = \operatorname*{argmin}_{x\in X} \left\{ \langle x, f'(x^k)\rangle + \frac{1}{t_k} B_\psi(x, x^k) \right\}, \quad t_k > 0. \tag{3.11}$$

When $\nabla\psi$ can be continuously extended on X, (e.g., as in Example 1), then we can consider the function B_ψ defined on $X \times X$. Note that in this case one needs not to start with $x^1 \in \operatorname{int} X$ and SANP can start with any arbitrary point $x^1 \in \mathbb{R}^n$. With $X = \mathfrak{R}^n$ and $\psi(x) = \frac{1}{2}\|x\|^2$ one obtains $B_\psi(x, y) = \frac{1}{2}\|x - y\|^2$ thus recovering the classical squared Euclidean distance and SANP is just the classical subgradient algorithm.

We now turn to the question of having SANP a well defined algorithm. When ψ is continuously differentiable on X, then the strong convexity assumption immediately implies that the algorithm which starts with $x^1 \in \mathbb{R}^n$ is well defined and produces a sequence $x^k \in X$, $\forall k$. When ψ is only assumed to be differentiable on $\operatorname{int} X$, we clearly need to guarantee that the next iterate stays in the interior of X, so that B_ψ can be defined on $X \times \operatorname{int} X$. For that, it suffices to make the following assumption:

$$\|\nabla\psi(x_t)\| \to +\infty \text{ as } t \to \infty, \ \forall\{x_t\} \in \operatorname{int} X \text{ with } x_t \to x \in \operatorname{bd} X, \tag{3.12}$$

where $\operatorname{bd} X$ denotes the boundary of X. Note that (3.12) is just to say that ψ is essentially smooth, (see [11]). With this additional assumption on ψ together with the strong convexity, it follows that the sequence $\{x^k\}$ is well defined i.e., $x^k \in \operatorname{int} X$, $\forall k$. An interesting choice for ψ satisfying (3.12) will be considered in Section 5.

It is interesting to note the differences between the two classes of algorithms which then emerged from (SANP). The first class with ψ continuously differentiable on X leads to *noninterior* methods with iterates $x^k \in X$. This is exactly the setting of the MDA. Typical examples of ψ in that case will involve power of norms on X, see Example 1 and [1]. On the other hand, the second class, with ψ satisfying (3.12) will be an *interior* type subgradient algorithm producing sequences $x^k \in \operatorname{int} X$. Note that the analysis we develop in the rest of this paper will hold for both classes of algorithms with the additional assumption (3.14) on ψ when needed.

We first recall some useful facts regarding strongly convex functions, and their relations with conjugates and subdifferentials. These results can be found in [13, Section 12H].

Proposition 3.1. *Let $\varphi : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ be a proper convex and lsc function and let $\sigma > 0$. Consider the following statements*:

(a) *φ is strongly convex with parameter σ.*
(b) *$\langle u - v, x - y\rangle \geqslant \sigma\|x - y\|^2$, whenever $u \in \partial\varphi(x)$, $v \in \partial\varphi(y)$; i.e., the map $\partial\varphi$ is strongly monotone.*
(c) *The inverse map $(\partial\varphi)^{-1}$ is everywhere single valued and Lipschitz continuous with constant σ^{-1}.*
(d) *φ^* is finite everywhere and differentiable.*

Then, $(a) \Leftrightarrow (b) \Rightarrow (c) \Leftrightarrow (d)$.

As written above in (3.11), the resemblance between MDA and SANP is still not obvious. However, we first note that the main step of SANP can be written in a more explicit way. Writing down formally the optimality conditions for (3.11), we obtain the following equivalent forms for SANP:

$$\begin{aligned} &0 \in t_k f'(x^k) + \nabla\psi(x^{k+1}) - \nabla\psi(x^k) + N_X(x^{k+1}),\\ &(\nabla\psi + N_X)(x^{k+1}) \in \nabla\psi(x^k) - t_k f'(x^k),\\ &x^{k+1} \in (\nabla\psi + N_X)^{-1}(\nabla\psi(x^k) - t_k f'(x^k)). \end{aligned} \tag{3.13}$$

Proposition 3.2. *The sequence $\{x^k\} \subseteq X$ generated by MDA corresponds exactly to the sequence generated by SANP.*

Proof. By definition of the conjugate function, one has $\psi^*(z) = \max_{x\in X}\{\langle x,z\rangle - \psi(x)\}$. Writing the optimality conditions for the later we obtain $0 \in z - \nabla\psi(x) - N_X(x)$, which is the same as $x \in (\nabla\psi + N_X)^{-1}(z)$. But, since ψ is strongly convex on X, then using Proposition 3.1, ψ^* is finite everywhere and differentiable and one has: $\nabla\psi^* = (\partial\psi)^{-1}$. Thus, the later inclusion is just the equation

$$x = (\nabla\psi + N_X)^{-1}(z) = \nabla\psi^*(z) = (\partial\psi)^{-1}.$$

Using these relations, SANP can be written as follows. Let $y^{k+1} := \nabla\psi(x^k) - t_k f'(x^k)$ and set $x^k = \nabla\psi^*(y^k)$. Then, SANP given by (3.13) reduces to $x^{k+1} = \nabla\psi^*(y^{k+1})$, which are exactly the iterations generated by MDA. □

Note that when ψ satisfies (3.12), then SANP reduces to: $x^{k+1} = (\nabla\psi)^{-1}(\nabla\psi(x^k) - t_k f'(x^k))$.

4. Convergence analysis

With this interpretation of the MDA, viewed as SANP, its convergence analysis can be derived in a simple way. The key of the analysis, relies essentially on the following simple identity which appears to be a natural generalization of the quadratic identity valid for the Euclidean norm.

Lemma 4.1 (Chen and Teboulle [5]). *Let $S \subset \mathbb{R}^n$ be an open set with closure $\bar{S}$ and let $\psi : \bar{S} \to \mathbb{R}$ be continuously differentiable on S. Then for any three points $a, b \in S$ and $c \in \bar{S}$ the following identity holds true*

$$B_\psi(c,a) + B_\psi(a,b) - B_\psi(c,b) = \langle \nabla\psi(b) - \nabla\psi(a), c - a\rangle. \tag{4.14}$$

We will need some further notations. Let $\|z\|_* = \max\{\langle x,z\rangle \mid x \in \mathbb{R}^n, \|x\| \leqslant 1\}$ be the (dual) conjugate norm. The convergence results for the SANP (and hence MDA) are given in the following theorem. We assume that the sequence x^k produced by SANP is well defined (see Section for the appropriate condition on ψ).

Theorem 4.1. *Suppose that assumption A is satisfied for the convex optimization problem (P). Let $\{x^k\}$ be the sequence generated by SANP with starting point $x^1 \in \mathrm{int}(X)$. Then, for every $k \geqslant 1$ one has*

(a)
$$\min_{1\leqslant s\leqslant k} f(x^s) - \min_{x\in X} f(x) \leqslant \frac{B_\psi(x^*,x^1) + 2\sigma^{-1}\sum_{s=1}^k t_s^2 \|f'(x^s)\|_*^2}{\sum_{s=1}^k t_s}. \tag{4.15}$$

(b) *In particular, the method converges, i.e.,* $\min_{1\leqslant s\leqslant k} f(x^s) - \min_{x\in X} f(x) \to 0$ *provided that*

$$\sum_s t_s = \infty, \quad t_k \to 0, \quad k \to \infty.$$

Proof. Let x^* be an optimal solution of (P). Optimality for (3.11) implies:

$$\langle x - x^{k+1}, t_k f'(x^k) + \nabla\psi(x^{k+1}) - \nabla\psi(x^k)\rangle \geqslant 0, \quad \forall x \in X$$

and thus in particular for $x = x^*$ we obtain

$$\langle x^* - x^{k+1}, \nabla\psi(x^k) - \nabla\psi(x^{k+1}) - t_k f'(x^k)\rangle \geqslant 0. \tag{4.16}$$

Using the subgradient inequality for the convex function f one obtains

$$\begin{aligned} 0 &\leqslant t_k(f(x^k) - f(x^*)) \\ &\leqslant t_k\langle x^k - x^*, f'(x^k)\rangle \\ &= \langle x^* - x^{k+1}, \nabla\psi(x^k) - \nabla\psi(x^{k+1}) - t_k f'(x^k)\rangle && (4.17) \\ &\quad + \langle x^* - x^{k+1}, \nabla\psi(x^{k+1}) - \nabla\psi(x^k)\rangle && (4.18) \\ &\quad + \langle x^k - x^{k+1}, t_k f'(x^k)\rangle && (4.19) \\ &:= s_1 + s_2 + s_3, && (4.20) \end{aligned}$$

where s_1, s_2, s_3 denotes the three right-hand side terms (4.17)–(4.19). Now, we have

$$s_1 \leqslant 0, \text{ [by (4.16)]},$$

$$s_2 = B_\psi(x^*,x^k) - B_\psi(x^*,x^{k+1}) - B_\psi(x^{k+1},x^k)$$

(by Lemma 4.1),

$$s_3 \leqslant (2\sigma)^{-1} t_k^2 \|f'(x^k)\|_*^2 + 2^{-1}\sigma\|x^k - x^{k+1}\|^2,$$

the later inequality following from $\langle a,b\rangle \leqslant (2\sigma)^{-1}\|a\|^2 + 2^{-1}\sigma\|b\|_*^2$, $\forall a,b\in\mathbb{R}^n$. Therefore, recalling that $B_\psi(\cdot,\cdot)$ is σ-strongly convex, i.e., $-B_\psi(x^{k+1},x^k) + 2^{-1}\sigma\|x^k - x^{k+1}\|^2 \leqslant 0$, it follows that

$$\begin{aligned} t_k(f(x^k) - f(x^*)) &= s_1 + s_2 + s_3 \\ &\leqslant B_\psi(x^*,x^k) - B_\psi(x^*,x^{k+1}) \\ &\quad + (2\sigma)^{-1} t_k^2 \|f'(x^k)\|_*^2. \end{aligned} \tag{4.21}$$

Summing (4.21) over $k = 1,\ldots,s$ one obtains,

$$\begin{aligned} \sum_{k=1}^{s} t_k(f(x^k) - f(x^*)) &\leqslant B_\psi(x^*,x^1) - B_\psi(x^*,x^{s+1}) \\ &\quad + (2\sigma)^{-1}\sum_{k=1}^{s} t_k^2\|f'(x^k)\|_*^2. \end{aligned}$$

Since $B_\psi(\cdot,\cdot) \geqslant 0$, it follows from the last inequality that

$$\begin{aligned} &\min_{1\leqslant s\leqslant k} f(x^s) - \min_{x\in X} f(x) \\ &\leqslant \frac{B_\psi(x^*,x^1) + (2\sigma)^{-1}\sum_{s=1}^{k} t_s^2\|f'(x^s)\|_*^2}{\sum_{s=1}^{k} t_s}, \end{aligned} \tag{4.22}$$

proving (a). Assuming that $t_k \to 0$ and $\sum t_k = \infty$ as $k\to\infty$, it thus follows from (4.22) that $\min_{1\leqslant s\leqslant k} f(x^s) - \min_{x\in X} f(x) \to 0$ as $k\to\infty$, proving (b). □

The above convergence result allows for deriving the best efficiency estimate of the method, by choosing an appropriate step size. The best stepsize is obviously obtained by minimizing the right-hand side of the inequality (4.22), with respect to $t\in\mathbb{R}^k_{++}$. We need the following technical result.

Proposition 4.1. *Given $c>0$, $b\in\mathbb{R}^d_{++}$ and D a symmetric positive definite matrix one has*

$$\inf_{z\in\mathbb{R}^d_{++}} \frac{c + (2\sigma)^{-1} z^{\mathrm{T}} D z}{b^{\mathrm{T}} z} = \sqrt{\frac{2c}{\sigma b^{\mathrm{T}} D^{-1} b}}$$

with optimal solution $z^* = \sqrt{(2c\sigma/b^{\mathrm{T}}D^{-1}b)}D^{-1}b$.

Proof. Writing the KKT optimality conditions for the (equivalent) convex problem

$$\inf_{z,u>0}\left\{\frac{c + (2\sigma)^{-1} z^{\mathrm{T}} D z}{u} : b^{\mathrm{T}} z = u\right\},$$

yields the desired result. □

We can now derive the efficiency estimate for SANP.

Theorem 4.2. *Suppose that assumption A is satisfied for the convex optimization problem (P). Let $\{x^k\}$ be the sequence generated by SANP with starting point $x^1\in \operatorname{int} X$. Then, with the stepsizes chosen as*

$$t_k := \frac{\sqrt{2\sigma B_\psi(x^*,x^1)}}{L_f}\frac{1}{\sqrt{k}}, \tag{4.23}$$

one has the following efficiency estimate

$$\begin{aligned} &\min_{1\leqslant s\leqslant k} f(x^s) - \min_{x\in X} f(x) \\ &\leqslant L_f\sqrt{\frac{2B_\psi(x^*,x^1)}{\sigma}}\frac{1}{\sqrt{k}}. \end{aligned} \tag{4.24}$$

Proof. The right-hand side of (4.22) is upper bounded by

$$\frac{B_\psi(x^*,x^1) + (2\sigma^{-1})L_f^2\sum_{s=1}^{k} t_s^2}{\sum_{s=1}^{k} t_s}, \tag{4.25}$$

Minimizing (4.25) with respect to $t_1,t_2,\ldots,t_k>0$, and invoking Proposition 4.1 with $c := B_\psi(x^*,x^1)$, $b := e = (1,1,\ldots,1)^{\mathrm{T}}$ and $D = L_f^2\cdot I$ where I is the $k\times k$ identity matrix, one gets the desired step size and efficiency estimate. □

Clearly, in order to make this result practical, one has to be able to upper bound the quantity $B_\psi(x^*,x^1)$, which depends on the (obviously unknown) optimal solution x^*, so that the step size and the efficiency estimate can be computed. This can be done by defining for any $y\in\operatorname{int} X$ $\gamma[\psi,y] := \max_{x\in X} B_\psi(x,y)$. Thus, one can replace in the estimate (4.15), the quantity $B(x^*,x^1)$ by $\gamma[\psi,x^1]$, provided the later quantity is finite. This is particularly true whenever X is assumed compact, as done in [1] for MDA in which case the results of Ben-Tal et al. [1] are recovered through

Theorem 4.2. Also, note that when we replace the differentiability assumption for ψ on $\operatorname{int} X$ and (3.12) holds, then one obtains an interior subgradient algorithm to minimize f over X, where in (4.23) and (4.24) the quantity $B(x^*,x^1)$ is replaced by $\gamma[\psi,x^1]<\infty$ for any $x^1\in\operatorname{int} X$.

5. Application: minimization over the unit simplex

As explained before, the key elements needed to implement the MDA and analyze its efficiency rely on our ability to compute the conjugate function ψ^* of ψ efficiently; to evaluate the strong convexity constant of ψ, and to upperbound the quantity $B_\psi(x^*,x^1)$. In this section, we begin by recalling briefly the results of Ben-Tal et al. [1], where the authors analyze the problem (P) of minimizing a convex function f over the unit simplex $\Delta:=\{x\in\mathbb{R}^n:\ \sum_{j=1}^n x_j=1,\ x\geqslant 0\}$, and we introduce a new method for this class of problems.

The *MDA*$_1$ (Ben-Tal et al. [1]): Let $\psi(x)=\psi_1(x):=2^{-1}\|x\|_p^2$ with $p:=1+(\ln n)^{-1}$. It was proved in [1] that the following efficiency estimate for MDA_1 holds:

$$\min_{1\leqslant s\leqslant k} f(x^s)-\min_{x\in X} f(x) \leqslant O(1)\frac{(\ln n)^{1/2}\max_{1\leqslant s\leqslant k}\|f'(x^s)\|_\infty}{\sqrt{k}} \qquad (5.26)$$

and thus, the MDA_1 with ψ_1 can outperformed the usual gradient method (obtained with ψ_2 on Δ) by a factor of $(n/\ln n)^{1/2}$, which for large n, can make a huge difference. This method was considered as a "nearly optimal" algorithm for the class of problems under consideration. For further details, see [1].

We now propose a different choice for ψ to solve the minimization problem (P) over the unit simplex Δ, which shares the same efficiency estimate. The function appears to be quite "natural" due to the simplex constraints, and is the so-called entropy function defined by

$$\psi_e(x)=\sum_{j=1}^n x_j\ln x_j \text{ if } x\in\Delta,\quad +\infty \text{ otherwise,} \qquad (5.27)$$

where we adopt the convention $0\ln 0\equiv 0$.

The entropy function defined on Δ possesses some remarkable properties collected below.

Proposition 5.1. *Let $\psi_e:\Delta\to\mathbb{R}$ be the entropy function defined in (5.27). Then,*

(a) *ψ_e is 1-strongly convex over $\operatorname{int}\Delta$ with respect to the $\|\cdot\|_1$ norm, i.e.,*

$$\langle\nabla\psi_e(x)-\nabla\psi_e(y),x-y\rangle = \sum_{j=1}^n (x_j-y_j)\ln\frac{x_j}{y_j} \geqslant \|x-y\|_1^2,\quad \forall x,y\in\operatorname{int}\Delta.$$

(b) *The conjugate of ψ_e is the function $\psi_e^*:\mathbb{R}^n\to\mathbb{R}$ with $\psi_e^*\in C^\infty(\mathbb{R}^n)$ given by $\psi_e^*(z)=\ln\sum_{j=1}^n e^{z_j}$, and $\|\nabla\psi_e(x)\|\to\infty$ as $x\to\bar{x}\in\Delta$.*

(c) *For the choice $x^1=n^{-1}e$, and $\psi=\psi_e$ one has $B_\psi(x^*,x^1)\leqslant\ln n$, $\forall x^*\in\Delta$.*

Proof. (a) The strong convexity of ψ_e follows from a fundamental inequality in information Theory [7]. For completeness we give here a different and simple proof. Let $\varphi:\Re_{++}\to\Re$ be defined by

$$\varphi(t)=(t-1)\ln t-2\frac{(t-1)^2}{t+1},\quad \forall t>0.$$

It is easy to verify that $\varphi(1)=\varphi'(1)=0$ and that $\varphi''(t)>0$ $\forall t>0$. Therefore φ is convex on $(0,\infty)$ and it follows that $\varphi(t)\geqslant 0$, $\forall t>0$. Therefore, with $t:=x_j/y_j$ it follows that $\forall x,y\in\operatorname{int}\Delta$:

$$\begin{aligned}\sum_{j=1}^n (x_j-y_j)\ln\frac{x_j}{y_j} &\geqslant \sum_{j=1}^n 2\frac{(x_j-y_j)^2}{x_j+y_j}\\ &= \sum_{j=1}^n \frac{x_j+y_j}{2}\frac{(x_j-y_j)^2}{(\frac{x_j+y_j}{2})^2}\\ &\overset{(*)}{\geqslant} \left(\sum_{j=1}^n \frac{x_j+y_j}{2}\frac{|x_j-y_j|}{\frac{x_j+y_j}{2}}\right)^2\\ &= \|x-y\|_1^2,\end{aligned}$$

where the inequality (*) follows from the convexity of the quadratic function and the fact that $(x+y)/2\in\Delta$.

(b) Using the definition of the conjugate and simple calculus gives the desired results, see also [11].

(c) Substituting $\psi = \psi_e(x) = \sum_{j=1}^n x_j \ln x_j$ in the definition of B_ψ we obtain with $x_j^1 = n^{-1}$, $\forall j$,

$$B_\psi(x^*, x^1) = \sum_{j=1}^n x_j^* \ln\left(\frac{x_j^*}{x_j^1}\right)$$

$$= \sum_{j=1}^n x_j^* \ln x_j^* + \ln n \leqslant \ln n, \quad \forall x^* \in \Delta,$$

the last inequality being true since the entropy function is always nonpositive on Δ. □

Remark 5.1. If for some j one has $y_j = 0, x_j > 0$, the left-hand side of the strong convexity inequality in (a) is $+\infty$ and there is nothing to prove. Likewise, when we reverse x with y. Thus, recalling that $0 \ln 0 \equiv 0$, it follows that the strong convexity inequality given in (a) remains true for all $x, y \in \Delta$.

Using the entropy function ψ_e in (3.12), we thus obtain a very simple algorithm for minimizing the convex function f over Δ, which is given explicitly by

The entropic descent algorithm (EDA)

Start with $x^1 \in \operatorname{int} \Delta$ and generate for $k = 1, \ldots,$ the sequence $\{x_k\}$ via:

$$x_j^{k+1} = \frac{x_j^k e^{-t_k f_j'(x^k)}}{\sum_{j=1}^n x_j^k e^{-t_k f_j'(x^k)}}, \quad t_k = \frac{\sqrt{2 \ln n}}{L_f} \frac{1}{\sqrt{k}},$$

where $f'(x) = (f_1(x)', \ldots, f_n'(x))^T \in \partial f(x)$.

Applying Theorem 4.2 and Proposition 5.1 we immediately obtain the following efficiency estimate for the EMDA.

Theorem 5.1. *Let $\{x^k\}$ be the sequence generated by EMDA with starting point $x^1 = n^{-1}e$. Then, for all $k \geqslant 1$ one has*

$$\min_{1 \leqslant s \leqslant k} f(x^s) - \min_{x \in X} f(x)$$

$$\leqslant \sqrt{2 \ln n} \frac{\|f'(x^s)\|_\infty}{\sqrt{k}}. \tag{5.28}$$

Thus, the EMDA appears as another useful candidate algorithm for solving large scale convex minimization problems over the unit simplex. Indeed, EMDA shares the same efficiency estimate than the (MDA_1) obtained with ψ_1, but has the advantage of being completely explicit, as opposed to the (MDA_1) which still requires the solution of one-dimensional nonlinear equation at each step of the algorithm to compute ψ_1^*.

6. Concluding remarks and further applications

We have presented a new derivation and analysis of mirror descent type algorithms. In its current state, the proposed approach has given rise to new insights on the properties of Mirror descent methods, bringing it in line of subgradient projection algorithms based on Bregman-based distance-like functions. This has led us to provide simple proofs for its convergence analysis and to introduce the new algorithm (EMDA) for solving convex problems over the unit simplex, with efficiency estimate mildly dependent on the problem's dimension. Many issues for potential extensions and further analysis include:

- Extension to the cases where $f(x) = \sum_{l=1}^m f_l(x)$ which can be derived along the analysis of incremental subgradients techniques [9,1] and numerical implementations for the corresponding EMDA.
- The choice of other functions ψ can be considered in SANP, (see for example [14,8]) to produce other interior subgradient (gradient) methods.
- Extension to semidefinite programs, in particular for problems with constraints of the type

  $$Z \in S_n, \quad \operatorname{tr}(Z) = 1, \ Z \succeq 0$$

  and which often arise in relaxations of combinatorial optimization problems. This can be analyzed within the use of a corresponding entropic function defined over the space of positive semidefinite symmetric matrices (see for example [6] and references therein).

References

[1] A. Ben-Tal, T. Margalit, A. Nemirovski, The ordered subsets mirror descent optimization method with applications to tomography, SIAM J. Optim. 12 (2001) 79–108.

[2] D. Bertsekas, Nonlinear Programming, 2nd Edition, Athena Scientific, Belmont, MA, 1999.

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

[4] Y. Censor, A. Lent, An iterative row-action method for interval convex programming, J. Optim. Theory Appl. 34 (3) (1981) 321–353.
[5] G. Chen, M. Teboulle, Convergence analysis of a proximal-like minimization algorithm using Bregman functions, SIAM J. Optim. 3 (1993) 538–543.
[6] M. Doljanski, M. Teboulle, An interior proximal algorithm and the exponential multiplier method for semi-definite programming, SIAM J. Optim. 9 (1998) 1–13.
[7] J.H.B. Kemperman, On the optimum rate of transmitting information, Ann. Math. Statist. 40 (1969) 2156–2177.
[8] K.C. Kiwiel, Proximal minimization methods with generalized Bregman functions, SIAM J. Control Optim. 35 (1997) 1142–1168.
[9] A. Nedic, D. Bertsekas, Incremental subgradient methods for nondifferentiable optimization, SIAM J. Optim. 12 (2001) 109–138.
[10] A. Nemirovski, D. Yudin, Problem complexity and Method Efficiency in Optimization, Wiley, New York, 1983.
[11] R.T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, NJ, 1970.
[12] R.T. Rockafellar, Monotone operators and the proximal point algorithm, SIAM J. Control Optim. 14 (1976) 877–898.
[13] R.T. Rockafellar, R.J.B Wets, Variational Analysis, Springer, New York, 1998.
[14] M. Teboulle, Entropic proximal mappings with application to nonlinear programming, Math. Oper. Res. 17 (1992) 670–690.