Estimate sequence methods: extensions and approximations

Michel Baes*

August 11, 2009

Abstract

The approach of estimate sequence offers an interesting rereading of a number of accelerating schemes proposed by Nesterov [Nes03], [Nes05], and [Nes06]. It seems to us that this framework is the most appropriate descriptive framework to develop an analysis of the sensitivity of the schemes to approximations.

We develop in this work a simple, self-contained, and unified framework for the study of estimate sequences, with which we can recover some accelerating scheme proposed by Nesterov, notably the acceleration procedure for constrained cubic regularization in convex optimization, and obtain easily generalizations to regularization schemes of any order. We analyze carefully the sensitivity of these algorithms to various types of approximations: partial resolution of subproblems, use of approximate subgradients, or both, and draw some guidelines on the design of further estimate sequence schemes.

1 Introduction

The concept of *estimate sequences* was introduced by Nesterov in 1983 [Nes83] to define the provably fastest gradient-type schemes for convex optimization. This concept, in spite of its conceptual simplicity, has not attracted a lot of attention during the 20 first years of its existence. Some interest for this concept resurrected in 2003, when Nesterov wrote his seminal paper on smoothing techniques [Nes05]. Indeed, the optimization method Nesterov uses on a smoothed approximation of the convex non-smooth objective function can be seen as an estimate sequence method. These estimate sequence methods play a crucial role in further papers of Nesterov [Nes06, Nes07]. Auslender and Teboulle [AT06] managed to extend the estimate sequence method, stated in Section 2.2 of [Nes03] for squared Euclidean norms as prox-functions, to general Bregman distances at the cost of a supplementary technical assumption on the domain of these Bregman distances.

Several other papers propose generalizations of Nesterov's smoothing algorithm, and can be interpreted in the light of the estimate sequence concept or sight generalizations of it. For instance,

*M. Baes is with the Institute for Operations Research, ETH, Rämistrasse 101, CH-8092 Zürich, Switzerland. Part of this work has been done while the author was at the Department of Electrical Engineering (ESAT), Research Group SCD-SISTA and the Optimization in Engineering Center OPTEC, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium. E-mail: Michel.Baes@ifor.math.ethz.ch.

Lan, Lu, and Monteiro [LLM] propose an accelerating strategy for Nesterov's smoothing, which can be interpreted as a simple restarting procedure of an estimate sequence scheme — albeit their algorithm is not an estimate sequence scheme as we define it in the present paper. D'Aspremont investigates another interesting aspect of Nesterov's algorithm: its robustness with respect to incorrect data [d'A08], more specifically to incorrect computation of the objective's gradient. In this paper, we show how we can benefit from the estimate sequence framework to carry out an analysis of the effect not only of approximate gradient on general estimate sequence schemes, but also of approximate resolution of subproblems one has to solve at every iteration. This result can support a strategy for reducing the iteration cost by solving only coarsely these subproblems.

The purpose of this paper is to provide a simple framework for the study of estimate sequences schemes, and to demonstrate its power in various ways. First, we generalize the accelerated cubic regularization scheme of Nesterov to "m-th regularization" schemes (cubic regularization represents the case where $m = 2$), with the fastest global convergence properties seen so far: these schemes require not more than $\mathcal{O}((1/\epsilon)^{m+1})$ iterations to find an ϵ-approximation of the solution. Also, our results allows us to reinterpret the accelerated cubic regularization schemes of Nesterov, and to improve some hidden constants in his complexity analysis.

Second, we show how accurately subgradients and the solution of intermediate problems have to be computed in order to guarantee that no error propagation occurs from possible approximations during the course of iterations.

The essence of our approach is crystallized in Lemma 2.3, which provides the sole condition to check for proving the convergence of the scheme, determine its speed, and investigating every extension we study in this paper. Interestingly, we can interpret Lan, Lu, and Monteiro's restarting scheme as an application of Lemma 2.3, albeit their scheme does not fall into the estimate sequence paradigm.

The paper is organized as follows. In Section 2, we recall the concept of estimate sequence, and we establish Lemma 2.3, which plays a central role in our paper. As in [Nes03], we particularize in the next section the general estimate sequence method to smooth problems. However, we use a slightly more general setting, allowing ourselves non-Euclidean norms. We show in Section 4 a simplified presentation of the fast cubic regularization for convex constrained problems developed by Nesterov in [Nes07]. This presentation allows us to extend the idea of cubic regularization to m-th regularization, and to obtain, at least theoretically the fastest black-box methods obtained so far, provided that the subproblems one must solve at every iteration are simple enough. Interestingly, we can improve some constants in the complexity results when we focus on unconstrained problem, due to the much more tractable optimality condition. Section 5 displays a careful sensitivity analysis of the various algorithms developed in the paper with respect to different kind of approximations. Section 6 shows briefly how Lan, Lu, and Monteiro's restarting scheme can be analyzed easily with the sole Lemma 2.3.

Finally, the Appendix contains some useful technical results.

2 Foundation of estimate sequence methods

Consider a general nonlinear optimization problem $\inf_{x\in Q} f(x)$, where $Q \subseteq \mathbb{R}^n$ is a closed set, and f is a proper convex function on Q. We assume that f attains its infimum f^* in Q, and we denote

by X^* the set of its minimizers. Later on, we will introduce more assumptions on f and Q, such as convexity, Lipschitz regularity, and differentiability a.o.

An *estimate sequence* (see Chapter 2 in [Nes03]) is a sequence of convex functions $(\phi_k)_{k\geq 0}$ and a sequence of positive number $(\lambda_k)_{k\geq 0}$ satisfying:

$$\lim_{k\to\infty} \lambda_k = 0, \quad \text{and} \quad \phi_k(x) \leq (1-\lambda_k)f(x) + \lambda_k\phi_0(x) \quad \text{for all } x \in Q,\ k \geq 1.$$

We also need to guarantee the inequality $\phi_0(x^*) \geq f^*$ for an element x^* of X^*. Since we obviously do not have access to any point of the set X^* before any computation starts, the latter condition has to be relaxed, e.g. to $\min_{x\in Q} \phi_0(x) \geq f(y)$ for a point $y \in Q$.

The first proposition indicates how estimate sequences can be used for solving an optimization problem, and, if implementable, how fast the resulting procedure would converge.

Proposition 2.1 *Suppose that the sequence $x_0, x_1, x_2, \ldots$ of Q satisfies $f(x_k) \leq \min_{x\in Q} \phi_k(x)$. Then $f(x_k) - f^* \leq \lambda_k(\phi_0(x^*) - f^*)$ for every $k \geq 1$.*

Proof

It suffices to write:

$$f(x_k) \leq \min_{x\in Q} \phi_k(x) \leq \min_{x\in Q} f(x) + \lambda_k(\phi_0(x) - f(x)) \leq f(x^*) + \lambda_k(\phi_0(x^*) - f(x^*)).$$

■

The next proposition describes a possible way of constructing an estimate sequence. It is a slight extension of Lemma 2.2.2 in [Nes03].

Proposition 2.2 *Let $\phi_0 : Q \to \mathbb{R}$ be a convex function such that $\min_{x\in Q} \phi_0(x) \geq f^*$. Let $(\alpha_k)_{k\geq 0} \subset (0,1)$ be a sequence whose sum diverges. Suppose also that we have a sequence $(f_k)_{k\geq 0}$ of functions from Q to $\mathbb{R}$ that underestimate f:*

$$f_k(x) \leq f(x) \text{ for all } x \in Q \text{ and all } k \geq 0.$$

We define recursively $\lambda_0 := 1$, $\lambda_{k+1} := \lambda_k(1-\alpha_k)$, and

$$\phi_{k+1}(x) := (1-\alpha_k)\phi_k(x) + \alpha_k f_k(x) = \lambda_{k+1}\phi_0(x) + \sum_{i=0}^{k} \frac{\lambda_{k+1}\alpha_i}{\lambda_{i+1}} f_i(x), \tag{1}$$

for all $k \geq 0$. Then $\Big((\phi_k)_{k\geq 0}; (\lambda_k)_{k\geq 0}\Big)$ is an estimate sequence.

Proof

Since

$$\ln(\lambda_{k+1}) = \sum_{j=0}^{k} \ln(1-\alpha_j) \leq -\sum_{j=0}^{k} \alpha_j$$

for each $k \geq 0$, the sequence $(\lambda_k)_{k\geq 0}$ converges to zero, as the sum of α_j's diverges. Let us now check that $\phi_k(x) \leq (1-\lambda_k)f(x) + \lambda_k\phi_0(x)$ for $k \geq 1$. For $k = 1$, this condition is immediately verified. Using now a recursive argument,

$$\begin{aligned}
\phi_{k+1}(x) = (1-\alpha_k)\phi_k(x) + \alpha_k f_k(x) &\leq (1-\alpha_k)\phi_k(x) + \alpha_k f(x) \\
&\leq (1-\alpha_k)(1-\lambda_k)f(x) + (1-\alpha_k)\lambda_k\phi_0(x) + \alpha_k f(x) \\
&= (1-\lambda_{k+1})f(x) + \lambda_{k+1}\phi_0(x),
\end{aligned}$$

which proves that we have built an estimate sequence. The second equality in (1) is obtained by an elementary recurrence on k. ■

All the estimate sequence methods we describe in this text are constructed on the basis of this fundamental proposition. In a nutshell, each of these methods can be defined by the specification of four elements:

- ◇ a function ϕ_0 that is easy to minimize on Q and that is bounded from below by $f(y)$ for a $y \in Q$;
- ◇ a sequence of weights α_k in $]0, 1[$;
- ◇ a strategy for constructing the successive lower estimates f_k of f. For convex objective functions, affine underestimates constitute the most natural choice as they are cheap to build. For strongly convex functions, we can also think of quadratic lower estimates (see Section 2.2.4 in [Nes03]).
- ◇ A way of constructing, preferably very cheaply, points x_k that satisfy the inequality prescribed in Proposition 2.1, namely $f(x_k) \leq \min_{x \in Q} \phi_k(x)$.

In view of Lemma 8.1, the existence of a positive constant β such that $\alpha_k^n / \lambda_{k+1} \geq \beta$ proves that the sequence $(\lambda_k)_{k \geq 0}$ decreases to zero as fast as $\mathcal{O}(1/(k^n \beta))$, i.e. the resulting algorithm requires $\mathcal{O}((1/\epsilon)^{1/k})$ iterations, where $\epsilon > 0$ is the desired accuracy.

The following lemma concentrates on the case where the feasible set Q as well as the objective function f are convex, and where the underestimates f_k of f are affine. It provides an intermediate inequality that we will use in the construction of the sequence $(x_k)_{k \geq 0}$ and for exploring various extensions and adaptations of the estimate sequence scheme. We denote the subdifferential of f at x as $\partial f(x)$ (see e.g. [Roc70], Section 23).

Lemma 2.3 *We are given an estimate sequence* $\left((\phi_k)_{k \geq 0}; (\lambda_k)_{k \geq 0}\right)$ *for the convex problem*

$$\min_{x \in Q} f(x),$$

constructed according to Proposition 2.2 using affine underestimates of f*:*

$$f_k(x) := f(y_k) + \langle g(y_k), x - y_k \rangle \quad \text{for some } y_k \in Q,$$

where $g(y_k) \in \partial f(y_k)$*. We also assume that* ϕ_0 *is continuously differentiable. Suppose that for every* $k \geq 0$*, we have a function* $\chi_k : Q \times Q \to \mathbb{R}_+$ *such that* $\chi_k(x, y) = 0$ *implies* $x = y$*, and such that:*

$$\phi_0(x) \geq \phi_0(v_k) + \langle \phi_0'(v_k), x - v_k \rangle + \chi_k(x, v_k) \quad \text{for all } x \in Q, k \geq 0, \tag{2}$$

where v_k *is the minimizer of* ϕ_k *on* Q*. We denote by* $(x_k)_{k \geq 0}$ *a sequence satisfying* $f(x_k) \leq \phi_k(v_k)$*. Then*

$$\phi_{k+1}(v_{k+1}) \geq f(y_k) + \langle g(y_k), (1 - \alpha_k) x_k + \alpha_k v_k - y_k \rangle + \min_{x \in Q} \{\alpha_k \langle g(y_k), x - v_k \rangle + \lambda_{k+1} \chi_k(x, v_k)\} \tag{3}$$

for every $k \geq 0$*.*

Proof
Observe first that the condition (2) can be rewritten as

$$\phi_k(x) \geq \phi_k(v_k) + \langle \phi_k'(v_k), x - v_k \rangle + \lambda_k \chi_k(x, v_k) \quad \text{for all } x \in Q, k \geq 0. \tag{4}$$

Indeed, in view of Proposition 2.2, we have $\phi_k(x) = \lambda_k \phi_0(x) + \sum_{i=0}^{k-1} \frac{\lambda_k \alpha_i}{\lambda_{i+1}} f_i(x) = \lambda_k \phi_0(x) + l_k(x)$, where $l_k(x)$ is an affine function. Inequality (4) can be rewritten as:

$$\lambda_k \phi_0(x) + l_k(x) \geq \lambda_k \phi_0(v_k) + \lambda_k \langle \phi_0'(v_k), x - v_k \rangle + l_k(x) + \lambda_k \chi_k(x, v_k),$$

and results immediately from (2). Now, fixing $k \geq 0$ and $x \in Q$, we can write successively:

$$\begin{aligned}
\phi_{k+1}(x) &= (1-\alpha_k)\phi_k(x) + \alpha_k f_k(x) \\
&\geq (1-\alpha_k)\left(\phi_k(v_k) + \langle \phi_k'(v_k), x - v_k \rangle + \lambda_k \chi_k(x, v_k)\right) + \alpha_k (f(y_k) + \langle g(y_k), x - y_k \rangle) \\
&\geq (1-\alpha_k)\left(f(x_k) + \lambda_k \chi_k(x, v_k)\right) + \alpha_k (f(y_k) + \langle g(y_k), x - y_k \rangle) \\
&\geq (1-\alpha_k)\left(f(y_k) + \langle g(y_k), x_k - y_k \rangle + \lambda_k \chi_k(x, v_k)\right) + \alpha_k (f(y_k) + \langle g(y_k), x - y_k \rangle) \\
&= f(y_k) + \langle g(y_k), (1-\alpha_k)x_k + \alpha_k v_k - y_k \rangle + \alpha_k \langle g(y_k), x - v_k \rangle + \lambda_{k+1} \chi_k(x, v_k).
\end{aligned}$$

The first inequality comes from (4). The second one uses the fact that v_k is a minimizer of ϕ_k on Q, so that $\langle \phi_k'(v_k), x - v_k \rangle \geq 0$ for each $x \in Q$. The third one comes form $g(y_k) \in \partial f(y_k)$.

It remains to minimize both sides on Q. ■

The inequality (3) suggests at least two different lines of attack to construct the next approximation x_{k+1} of an optimum x^*.

First, if we can ensure that the sequence $(y_k)_{k \geq 0}$ satisfies at every $k \geq 0$ and at every $x \in Q$ the inequality:

$$\langle g(y_k), (1-\alpha_k)x_k + \alpha_k v_k - y_k \rangle + \{\alpha_k \langle g(y_k), x - v_k \rangle + \lambda_{k+1} \chi_k(x, v_k)\} \geq 0,$$

it suffices to set $x_{k+1} := y_k$.

Another possibility is to build a sequence $(y_k)_{k \geq 0}$ for which the inequality

$$\langle g(y_k), (1-\alpha_k)x_k + \alpha_k v_k - y_k \rangle \geq 0$$

holds for every $k \geq 0$ — for instance by letting $y_k := (1-\alpha_k)x_k + \alpha_k v_k$. Then, the inequality of the above lemma reduces to

$$\phi_{k+1}(v_{k+1}) \geq f(y_k) + \min_{x \in Q} \{\alpha_k \langle g(y_k), x - v_k \rangle + \lambda_{k+1} \chi_k(x, v_k)\}.$$

In some situations, constructing a point x_{k+1} for which $f(x_{k+1})$ is lower than the above right-hand side can be done very cheaply by an appropriate subgradient-like step. More details are given in the subsequent sections.

Finally, the above lemma suggests a new type of scheme, where one only ensures that the inequality (3) is maintained at every iteration regardless of the fact that it originates from the construction of an estimate sequence. Under some conditions on χ_k, the convergence speed of the resulting scheme also relies on how fast we can drive the sequence $(\lambda_k)_{k \geq 0}$ to 0. More details are given in Section 6.

3 Strongly convex estimates for convex constrained optimization

In this setting, the function ϕ_0 we choose is a strongly convex function, not necessarily quadratic. Let us fix a norm $||\cdot||$ of $\mathbb{R}^n$. We assume that the objective function f is differentiable and has a Lipschitz continuous gradient with constant L for the norm $||\cdot||$:

$$\forall x, y \in Q, \quad |f(y) - f(x) - \langle f'(x), y - x\rangle| \leq \frac{L}{2}||y - x||^2.$$

Equivalently, denoting by $||\cdot||_*$ the dual norm of $||\cdot||$, the inequality

$$||f'(y) - f'(x)||_* \leq L||y - x||$$

holds for every $x, y \in Q$. Observe that we do not assume the strong convexity of f. The function ϕ_0 is constructed from a *prox-function* d for Q. A prox-function d is a nonnegative convex function minimized at a point $x_0 \in \operatorname{relint} Q$, and for which $d(x_0) = 0$. Also, a prox-function is supposed to be strongly convex: there exists a constant $\sigma > 0$ for which every $x, y \in Q$ and $\lambda \in [0, 1]$, we have:

$$\lambda d(x) + (1 - \lambda)d(y) \geq d(\lambda x + (1 - \lambda)y) + \frac{\sigma}{2}\lambda(1 - \lambda)||x - y||^2.$$

If the function d is differentiable, this condition can be rewritten as (see e.g Theorem IV.4.1.1 in [HUL93a]):

$$\forall x, y \in Q, \quad \langle d'(x) - d'(y), x - y\rangle \geq \sigma||x - y||^2.$$

The prox-function d is a crucial building tool for an estimate sequence, and it should be chosen carefully. Indeed, at each step, we will have to solve one (or two) problem(s) of the form $\min_{x\in Q} d(x) + l(x)$, where l is a linear mapping. Sometimes (in cubic and m-th regularization schemes, see in Section 4, we even need to solve $\min_{x\in Q} d(x) + p(x)$, where p is a polynomial function. Having an easy access to its minimizer is a *sine qua non* requirement for the subsequent algorithm to work efficiently. The best instances for this scheme are of course those for which this minimizer can be computed analytically. Interestingly enough, the set of these instances is not reduced to a few trivial ones (see [Nes05, NP06]).

These ingredients allow us to define the first function of our estimate sequence:

$$\phi_0(x) := f(x_0) + \frac{L}{\sigma}d(x),$$

which is L-strongly convex for the norm $||\cdot||$. Also $\min_{x\in Q} \phi_0(x) = f(x_0) \geq f(x^*)$. Moreover, we have for every $x \in Q$:

$$\phi_0(x) - \phi_0(x_0) - \langle \phi_0'(x_0), x - x_0\rangle \geq \frac{L}{2}||x - x_0||^2 \geq f(x) - f(x_0) - \langle f'(x_0), x - x_0\rangle.$$

Therefore, $\phi_0(x) \geq f(x) - \langle f'(x_0), x - x_0\rangle$.

The underestimates f_k of f are chosen to be linear underestimates:

$$f_k(x) := f(y_k) + \langle f'(y_k), x - y_k\rangle.$$

Following the construction scheme of Proposition 2.2, we have

$$\phi_{k+1}(x) = \lambda_{k+1}\left(f(x_0) + \frac{L}{\sigma}d(x)\right) + \sum_{i=0}^{k}\frac{\lambda_{k+1}\alpha_i}{\lambda_{i+1}}(f(y_i) + \langle f'(y_i), x - y_i\rangle)$$

for all $k \geq 0$. As the function ϕ_k is strongly convex — with constant $\lambda_k L$ for the norm $||\cdot||$ — it has a unique minimizer. We denote this minimizer by v_k.

At iteration $k+1$, we need to construct a point $x_{k+1} \in \mathbb{R}^n$ such that $f(x_{k+1}) \leq \phi_{k+1}(v_{k+1})$, given that $f(x_k) \leq \phi_k(v_k)$. There are many ways to achieve this goal; we consider here two possibilities hinted by Lemma 2.3. Each of them are defined by a nonnegative function χ_k for which $\chi_k(x, x) = 0$ and

$$d(x) \geq d(v_k) + \langle d'(v_k), x - v_k\rangle + \frac{\sigma}{L}\chi_k(x, v_k) \quad \text{for all } x \in Q, k \geq 0.$$

A first possibility is to choose $\chi_k(x, y) := L||x-y||^2/2$; the above inequality reduces to the σ-strong convexity of d. We can also consider a suitable multiple of the Bregman distance induced by d, that is,

$$\chi_k(x, y) := \gamma_k(d(x) - d(y) - \langle d'(y), y - x\rangle).$$

The required inequality is ensured as soon as $\sigma/L \geq \gamma_k > 0$. The potential advantage of this second approach resides in the fact that, in the resulting algorithm, the computation of the next point x_{k+1} requires to solve a problem of the same type than for computing the minimizer of ϕ_k, that is, minimizing a d plus a linear function over Q.

3.1 When the lower bound on ϕ_k is a norm

Let us consider here the case where $\chi_k(x, y) := L||x - y||^2/2$, where $||\cdot||$ is a norm on $\mathbb{R}^n$, not necessarily Euclidean like in [Nes03]. According to Lemma 2.3, we can set

$$y_k := (1 - \alpha_k)x_k + \alpha_k v_k.$$

With this choice, we obtain:

$$\phi_{k+1}(v_{k+1}) \geq f(y_k) + \min_{x \in Q}\left\{\alpha_k\langle f'(y_k), x - v_k\rangle + \frac{\lambda_{k+1}L}{2}||x - v_k||^2\right\} \tag{5}$$

for every $k \geq 0$.

The minimization problem on the right-hand side is closely related to the standard gradient method. We denote by $x_Q(y; h)$ the minimizer of

$$\langle f'(y), x - y\rangle + \frac{1}{2h}||x - y||^2$$

over Q. If the considered norm is Euclidean, this minimizer is simply the Euclidean projection of a gradient step over Q:

$$x_Q(y; h) = \arg\min_{x \in Q}||x - (y - hf'(y))||_2.$$

Observe that:

$$\begin{aligned} m_k^* &:= \min_{x\in Q}\left\{\alpha_k\langle f'(y_k), x - v_k\rangle + \frac{\lambda_{k+1}L}{2}||x - v_k||^2\right\} \\ &= \min_{x\in Q}\left\{\langle f'(y_k), \alpha_k x + (1-\alpha_k)x_k - y_k\rangle + \frac{\lambda_{k+1}L}{2\alpha_k^2}||\alpha_k x + (1-\alpha_k)x_k - y_k||^2\right\} \\ &\geq \min_{x\in Q}\left\{\langle f'(y_k), x - y_k\rangle + \frac{\lambda_{k+1}L}{2\alpha_k^2}||x - y_k||^2\right\}. \end{aligned}$$

because $\alpha_k Q + (1-\alpha_k)x_k \subseteq Q$ in view of the convexity of Q. Thus:

$$m_k^* \geq \min_{x\in Q}\left\{\langle f'(y_k), x - y_k\rangle + \frac{L}{2}||x - y_k||^2\right\},$$

provided that $\lambda_{k+1}/\alpha_k^2 \geq 1$. Hence, we can bound $\phi_{k+1}(v_{k+1})$ from below by:

$$f(y_k) + \langle f'(y_k), x_Q(y_k; 1/L) - y_k\rangle + \frac{L}{2}||x_Q(y_k; 1/L) - y_k||^2.$$

By Lipschitz continuity of the gradient of f, this quantity is larger than $f(x_Q(y_k; 1/L))$. Therefore, setting $x_{k+1} := x_Q(y_k; 1/L)$ is sufficient to ensure the required decrease of the objective. However, this choice assumes that optimization problems of the form $\min_{x\in Q} ||x - z||^2 + l(x)$, where l is a linear function, are easy to solve as well. An alternative is presented in the next subsection, where only optimization problems of the form $\min_{x\in Q} d(x) + l(x)$ should be solved at every iteration.

Algorithm 3.1
Assumptions: *f has a Lipschitz continuous gradient with constant L for the norm $||\cdot||$; the set Q is closed, convex, and has a nonempty interior.*
Choose $x_0 \in Q$, set $v_0 := x_0$ and $\lambda_0 := 1$.
Choose a strongly convex function d with strong convexity constant $\sigma > 0$ for the norm $||\cdot||$, minimized in $x_0 \in Q$, and that vanishes in x_0 and set $\phi_0 := f(x_0) + Ld(x)/\sigma$.
For $k \geq 0$,
Find α_k such that $\alpha_k^2 = (1-\alpha_k)\lambda_k$.
Set $\lambda_{k+1} := (1-\alpha_k)\lambda_k$.
Set $y_k := \alpha_k v_k + (1-\alpha_k)x_k$.
Set $x_{k+1} := x_Q(y_k; 1/L) = \arg\min_{x\in Q}\{\langle f'(y_k), x - y_k\rangle + \frac{L}{2}||x - y_k||^2\}$.
Set $\phi_{k+1}(x) := (1-\alpha_k)\phi_k(x) + \alpha_k(f(y_k) + \langle f'(y_k), x - y_k\rangle)$.
Set $v_{k+1} := \arg\min_{x\in Q}\phi_{k+1}(x)$.
End ■

Assuming that $\alpha_k^2 = \lambda_{k+1} = (1-\alpha_k)\lambda_k$, we obtain in view of Lemma 8.1 and Proposition 2.1 a complexity of

$$2\sqrt{\frac{1}{\epsilon}\left(f(x_0) - f^* + \frac{L}{\sigma}d(x^*)\right)} \text{ iterations.}$$

There is a simple variant for the computation of the sequences $(\lambda_k)_{k\geq 0}$ and $(\alpha_k)_{k\geq 0}$. The requirement $\lambda_k \geq \alpha_k^2/(1-\alpha_k)$ is satisfied for every $k \geq 0$ when $\lambda_k = 4/(k+2)^2$ for $k \geq 2$. With this choice, we obtain $\alpha_k = 1 - \frac{\lambda_{k+1}}{\lambda_k} = (2k+5)/(k+3)^2$ and $\alpha_k^2/\lambda_{k+1} = (1 - 1/(2k+6))^2 \in [25/36, 1[$.

3.2 When the lower bound on ϕ_k is a Bregman distance

In this setting, we define $\chi_k(x,z) := \gamma_k(d(x) - d(z) - \langle d'(z), x - z\rangle)$, where γ_k is a positive coefficient that will be determined in the course of our analysis. Observe that, in contrast with [AT06], we do not assume anything on the domain of d, except that it contains Q. For a fixed $z \in Q$, the function $x \mapsto \chi_k(x,z)$ is strongly convex with constant $\sigma\gamma_k$ for the norm $||\cdot||$. Even better, we have for every $x, y \in Q$ that:

$$\chi_k(x,y) \geq \frac{\gamma_k \sigma}{2}||x-y||^2. \tag{6}$$

If the coefficients γ_k are bounded from above by L/σ, we can apply Lemma 2.3 because the inequality (2) is satisfied in view of the Lipschitz continuity of f'. Therefore, with:

$$y_k := (1-\alpha_k)x_k + \alpha_k v_k,$$

we have:

$$\phi_{k+1}(v_{k+1}) \geq f(y_k) + \min_{x \in Q}\left\{\alpha_k\langle f'(y_k), x - v_k\rangle + \lambda_{k+1}\chi_k(x, v_k)\right\}$$

for every $k \geq 0$. Let us denote:

$$w_k := \arg\min_{x \in Q}\left\{\alpha_k\langle f'(y_k), x - v_k\rangle + \lambda_{k+1}\chi_k(x, v_k)\right\},$$

and check that:

$$x_{k+1} := \alpha_k(w_k - v_k) + y_k = \alpha_k w_k + (1-\alpha_k)x_k$$

yields a sufficient decrease of the objective. We can write:

$$\begin{aligned}
& f(y_k) + \min_{x \in Q}\left\{\alpha_k\langle f'(y_k), x - v_k\rangle + \lambda_{k+1}\chi_k(x, v_k)\right\} \\
= \; & f(y_k) + \alpha_k\langle f'(y_k), w_k - v_k\rangle + \lambda_{k+1}\chi_k(w_k, v_k) \\
\geq \; & f(y_k) + \alpha_k\langle f'(y_k), w_k - v_k\rangle + \frac{\lambda_{k+1}\gamma_k\sigma}{2}||w_k - v_k||^2 \\
= \; & f(y_k) + \langle f'(y_k), x_{k+1} - y_k\rangle + \frac{\lambda_{k+1}\gamma_k\sigma}{2\alpha_k^2}||x_{k+1} - y_k||^2.
\end{aligned}$$

Now, if $\lambda_{k+1}/\alpha_k^2 \geq 1$ and $\gamma_k \geq L/\sigma$, the right-hand side is clearly larger than $f(x_{k+1})$ in view of the Lipschitz continuity of the gradient of f. The corresponding algorithm can be written as follows.

Algorithm 3.2
Assumptions: *f has a Lipschitz continuous gradient with constant L for the norm $||\cdot||$; the set Q is closed, convex, and has a nonempty interior.*
Choose $x_0 \in Q$, set $v_0 := x_0$ and $\lambda_0 := 1$.
Choose a strongly convex function d with strong convexity constant $\sigma > 0$ for the norm $||\cdot||$, minimized in $x_0 \in Q$, and that vanishes in x_0 and set $\phi_0 := f(x_0) + Ld(x)/\sigma$.
For $k \geq 0$,
Find α_k such that $\alpha_k^2 = (1-\alpha_k)\lambda_k$.
Set $\lambda_{k+1} := (1-\alpha_k)\lambda_k$.
Set $y_k := \alpha_k v_k + (1-\alpha_k)x_k$.
Set $w_k := \arg\min_{x \in Q}\{\alpha_k\langle f'(y_k), x - v_k\rangle + \lambda_{k+1}\chi_k(x, v_k)\}$.

Set $x_{k+1} := \alpha_k w_k + (1-\alpha_k)x_k$.
Set $\phi_{k+1}(x) := (1-\alpha_k)\phi_k(x) + \alpha_k(f(y_k) + \langle f'(y_k), x - y_k\rangle)$.
Set $v_{k+1} := \arg\min_{x\in Q} \phi_{k+1}(x)$.
End ■

4 Cubic regularization and beyond

4.1 Cubic regularization for constrained problems

Cubic regularization has been developed by Nesterov and Polyak in [NP06], and further extended by Nesterov in [Nes06, Nes07] and by Cartis, Gould, and Toint [CGT07]. We derive in this section a slight modification of Nesterov's algorithm for constrained convex problems and establish its convergence speed using an alternative proof to his, which allows us to further extend the accelerated algorithm to other types of regularization.

We consider here a convex objective function f that is twice continuously differentiable and that has a Lipschitz continuous Hessian with respect to a *matrix induced norm* $||\cdot||$, that is, a norm of the form $||a|| = \langle Ga, a\rangle^{1/2}$, where G is a positive definite Gram matrix. In other words, we assume that there exists $M > 0$ such that for every $x, y \in \operatorname{dom} f$, we can write:

$$|||f''(y) - f''(x)||| \leq M||y - x||.$$

The matrix norm used above is the one induced by the norm we have chosen, that is, $|||A||| := \max_{||x||=1} ||Ax||_*$. A consequence of the Lipschitz continuity of the Hessian (see Lemma 1 in [NP06] for a proof) reads:

$$||f'(y) - f'(x) - f''(x)(y-x)||_* \leq \frac{M}{2}||y-x||^2 \quad \forall x, y \in \operatorname{dom} f. \tag{7}$$

The optimization problem of interest here consists in minimizing f on a closed convex set $Q \subseteq \mathbb{R}^n$ with a nonempty interior.

Let us fix a starting point $x_0 \in Q$. We initialize the construction of our estimate sequence by:

$$\phi_0(x) := f(x_0) + \frac{M}{6}||x - x_0||^3.$$

As in Proposition 2.2, we consider linear underestimates of f:

$$f_k(x) = f(y_k) + \langle f'(y_k), x - y_k\rangle$$

to build the estimate sequence. Let us define $\chi_k(x, z) := M||x - z||^3/12$ for every $k \geq 0$. Since the inequality (2) is satisfied in view of Lemma 8.2, we can use Lemma 2.3 to define our goal: at iteration k, we need to find a point $x_{k+1} \in Q$ and suitable coefficients α_k for which:

$$\begin{aligned} f(y_k) \;+\; &\langle f'(y_k), (1-\alpha_k)x_k + \alpha_k v_k - y_k\rangle \\ &+ \min_{x\in Q}\left\{\alpha_k\langle f'(y_k), x - v_k\rangle + \lambda_{k+1}\frac{M}{12}||x - v_k||^3\right\} \geq f(x_{k+1}). \end{aligned} \tag{8}$$

Recall that v_k is the minimizer of ϕ_k on Q. Our strategy here is to take $x_{k+1} := y_k$, so we are left with the problem of determining a point y_k for which the sum of the two last terms on the left-hand side is nonnegative.

In parallel with what we have done in the previous sections where the objective function had a Lipschitz continuous gradient, we define:

$$x_N(x) := \arg\min_{y \in Q} \left\{ f(x) + \langle f'(x), y - x \rangle + \frac{1}{2}\langle f''(x)(y-x), y-x \rangle + \frac{N}{6}||y-x||^3 \right\}$$

for every $N \geq M$. For the subsequent method to have a practical interest at all, the above optimization problem has to be easy to solve. As noticed by Nesterov and Polyak in Section 5 of [NP06], *unconstrained* nonconvex — although this paper does not leave the convex realm — problems of the above type can be solved efficiently because their *strong* dual boils down to a convex optimization problem with only one variable. Moreover, the optimal solution of the original problem can be easily reconstructed from the dual optimum. For constrained problems, Nesterov observed in Section 6 of [Nes06] that, as long as convex quadratic functions can be minimized easily on Q, we can guarantee an easy access to $x_N(x)$.

The optimality condition for $x_N(x)$ reads as follows:

$$\big\langle f'(x)+f''(x)(x_N(x)-x), y-x_N(x)\big\rangle + \frac{N}{2}||x-x_N(x)||\big\langle G(x_N(x)-x), y-x_N(x)\big\rangle \geq 0 \quad \forall y \in Q. \quad (9)$$

We start our analysis with an easy lemma, an immediate generalization of which will be exploited in the next section as well.

Lemma 4.1 *Let $g \in \mathbb{R}^n$, $\lambda > 0$, $x, v \in Q$ and $z := (1-\alpha)x + \alpha v$ for an $\alpha \in [0,1]$. We have:*

$$\min_{y \in Q} \{\alpha\langle g, y - v\rangle + \lambda\chi_k(y,v)\} \geq \min_{y \in Q} \left\{ \langle g, y - z\rangle + \frac{\lambda}{\alpha^3}\chi_k(y,z) \right\}.$$

Proof
By convexity of Q, we have $Q = \alpha Q + (1-\alpha)Q$. Therefore:

$$Q - z = (1-\alpha)(Q-x) + \alpha(Q-v) \supseteq \alpha(Q-v)$$

because x belongs to Q. Now, we can write:

$$\begin{aligned}
\min\{\alpha\langle g, y-v\rangle + \lambda\chi_k(y,v) : y \in Q\} &= \min\{\langle g, u\rangle + \lambda M||u||^3/(12\alpha^3) : u \in \alpha(Q-v)\} \\
&\geq \min\{\langle g, u\rangle + \lambda M||u||^3/(12\alpha^3) : u \in Q - z\} \\
&= \min\{\langle g, y-z\rangle + \lambda\chi_k(y,z)/\alpha^3 : y \in Q\}.
\end{aligned}$$

■

The next lemma plays the crucial role in the validation of the desired inequality. (Compare with item 1 of Theorem 1 in [Nes07]). We write $r_N(x)$ for $||x - x_N(x)||$.

Lemma 4.2 *For every $x, y \in Q$, we have*

$$\langle f'(x_N(x)), y - x_N(x)\rangle \geq -\frac{M+N}{2}r_N(x)^2||y-x|| + \frac{N-M}{2}r_N(x)^3.$$

Proof
By the optimality condition (9), we have for every $x, y \in Q$:

$$0 \leq \big\langle f'(x) + f''(x)(x_N(x) - x), y - x_N(x) \big\rangle + \frac{N r_N(x)}{2} \big\langle G(x_N(x) - x), y - x_N(x) \big\rangle. \qquad (10)$$

Observe that:

$$\big\langle G(x_N(x) - x), y - x_N(x) \big\rangle \leq r_N(x)||y - x|| - r_N(x)^2.$$

Moreover, in view of the Hessian Lipschitz continuity (7), we have:

$$\begin{aligned} & \langle f'(x) + f''(x)(x_N(x) - x), y - x_N(x) \rangle \\ \leq\ & ||f'(x) + f''(x)(x_N(x) - x) - f'(x_N(x))||_* ||y - x_N(x)|| + \langle f'(x_N(x)), y - x_N(x) \rangle \\ \leq\ & \frac{M}{2} r_N(x)^2 \big(||y - x|| + r_N(x) \big) + \langle f'(x_N(x)), y - x_N(x) \rangle. \end{aligned}$$

Summing up these two inequalities with appropriate multiplicative coefficients, we get form (10):

$$\begin{aligned} 0 \leq\ & \frac{N r_N(x)}{2} \big(r_N(x)||y - x|| - r_N(x)^2 \big) + \frac{M}{2} r_N(x)^2 \big(||y - x|| + r_N(x) \big) + \langle f'(x_N(x)), y - x_N(x) \rangle \\ =\ & \frac{N + M}{2} r_N(x)^2 ||y - x|| + \frac{M - N}{2} r_N(x)^3 + \langle f'(x_N(x)), y - x_N(x) \rangle. \end{aligned}$$

■

We are now ready to design an estimate sequence scheme for the constrained optimization problem we are interested in.

Algorithm 4.1
Assumption: *f is convex and has a Lipschitz continuous Hessian with constant M for the matrix norm induced by $||\cdot||$.*
Choose $x_0 \in Q$, set $v_0 := x_0$ and $\lambda_0 := 1$.
Set $\phi_0(x) := f(x_0) + M||x - x_0||^3/6$.
For *$k \geq 0$,*
Find α_k such that $12\alpha_k^3 = (1 - \alpha_k)\lambda_k$.
Set $\lambda_{k+1} := (1 - \alpha_k)\lambda_k$.
Set $z_k := \alpha_k v_k + (1 - \alpha_k) x_k$.
Set $y_k := \arg\min_{y \in Q} \{ \langle f'(z_k), y - z_k \rangle + \frac{1}{2} \langle f''(z_k)(y - z_k), y - z_k \rangle + \frac{5M}{6} ||x - z_k||^3 \}$.
Set $x_{k+1} := y_k$.
Set $\phi_{k+1}(x) := (1 - \alpha_k)\phi_k(x) + \alpha_k (f(y_k) + \langle f'(y_k), x - y_k \rangle)$.
Set $v_{k+1} := \arg\min_{x \in Q} \phi_{k+1}(x)$.
End ■

Theorem 4.1 *The above algorithm takes not more than*

$$\left\lceil 12 \left(\frac{1}{\epsilon} \right)^{1/3} \left(f(x_0) - f(x^*) + \frac{M}{6} ||x_0 - x^*||^3 \right)^{1/3} \right\rceil$$

iterations to find a point x_k for which $f(x_k) - f^* \leq \epsilon$.

Proof
Let us show first that the inequality (8) is satisfied, that is, that:

$$\langle f'(y_k), (1-\alpha_k)x_k + \alpha_k v_k - y_k\rangle + \min_{x\in Q}\left\{\alpha_k\langle f'(y_k), x - v_k\rangle + \lambda_{k+1}\frac{M}{12}||x-v_k||^3\right\} \geq 0.$$

In view of the algorithm, we have set $z_k := \alpha_k v_k + (1-\alpha_k)x_k$ and $y_k := x_N(z_k)$. Using Lemma 4.1 and Lemma 4.2, the second term of the above inequality can be bounded from below as follows:

$$\begin{aligned}
&\min_{y\in Q}\left\{\alpha_k\langle f'(y_k), y - v_k\rangle + \lambda_{k+1}\frac{M}{12}||y-v_k||^3\right\}\\
\geq\ &\min_{y\in Q}\left\{\langle f'(y_k), y - z_k\rangle + \frac{\lambda_{k+1}}{\alpha_k^3}\frac{M}{12}||y-z_k||^3\right\}\\
=\ &\langle f'(y_k), y_k - z_k\rangle + \min_{y\in Q}\left\{\langle f'(y_k), y - y_k\rangle + \frac{\lambda_{k+1}}{\alpha_k^3}\frac{M}{12}||y-z_k||^3\right\}\\
\geq\ &\langle f'(y_k), y_k - z_k\rangle + \min_{y\in Q}\left\{-\frac{M+N}{2}r_N(z_k)^2||y-z_k|| + \frac{N-M}{2}r_N(z_k)^3 + \frac{\lambda_{k+1}}{\alpha_k^3}\frac{M}{12}||y-z_k||^3\right\}.
\end{aligned}$$

Therefore, the inequality we need to prove becomes:

$$\frac{N-M}{2}r_N(z_k)^3 + \min_{y\in Q}\left\{-\frac{M+N}{2}r_N(z_k)^2||y-z_k|| + \frac{\lambda_{k+1}}{\alpha_k^3}\frac{M}{12}||y-z_k||^3\right\} \geq 0.$$

Suppressing the constraint of the above minimization problem, we get the following lower bound:

$$\begin{aligned}
&\min_{y\in\mathbb{R}^n}\left\{-\frac{M+N}{2}r_N(z_k)^2||y-z_k|| + \frac{\lambda_{k+1}}{\alpha_k^3}\frac{M}{12}||y-z_k||^3\right\}\\
=\ &\min_{t\geq 0}\left\{-\frac{M+N}{2}r_N(z_k)^2 t + \frac{\lambda_{k+1}}{\alpha_k^3}\frac{M}{12}t^3\right\} = -\frac{\sqrt{2}}{3}r_N(z_k)^3\sqrt{\frac{(M+N)^3}{M}}\sqrt{\frac{\alpha_k^3}{\lambda_{k+1}}}.
\end{aligned}$$

Thus, the inequality is satisfied as soon as:

$$\frac{N-M}{2} - \frac{\sqrt{2}}{3}\sqrt{\frac{(M+N)^3}{M}}\sqrt{\frac{\alpha_k^3}{\lambda_{k+1}}} \geq 0,$$

or

$$\frac{9}{8}\frac{M(N-M)^2}{(M+N)^3} \geq \frac{\alpha_k^3}{\lambda_{k+1}}.$$

The left-hand side can be maximized in N. Its maximizer turns out to be attained for $N := 5M$, in which case its value is $1/12$. Note the constants prescribed here have been integrated in Algorithm 4.1.

It suffices now to apply Lemma 8.1 to obtain the complexity of the algorithm. ■

4.2 Beyond cubic regularization

In principle, the above reasoning can be applied in the study of an optimization scheme for constrained convex optimization with higher regularity. However, the obtained scheme would imply to solve at every iteration a problem of the form

$$\min_{y \in Q} f(x) + f'(x)[y-x] + \cdots + \frac{1}{m!} f^{(m)}(x)[y-x, \ldots, y-x] + \frac{M}{(m+1)!} ||y-x||^{m+1},$$

which can be highly nontrivial. A discussion of the cases where this problem is reasonably easy — e.g. where the above objective function is convex and/or has an easy dual — can be the topic of a further paper. In fact, this problem does not need to be solved extremely accurately. We show in this paper that, in the case where D_Q is bounded, a solution with accuracy $\mathcal{O}(\epsilon^{1.5})$ is amply sufficient to guarantee the reliability of the algorithm — see Subsection 5.1 for more details. Nevertheless, let us analyze this scheme, as an illustration of the power of the estimate sequence framework.

Given a norm $||\cdot||$ on $\mathbb{R}^n$, we define the norm of a tensor A of rand d as:

$$||A|| := \sup_{||x_1||=1} \cdots \sup_{||x_d||=1} \langle A[x_1, \ldots, x_{d-1}], x_d \rangle.$$

Let us assume that the m-th derivative of f is Lipschitz continuous:

$$||f^{(m)}(y) - f^{(m)}(x)|| \leq M||y-x||$$

for every $x, y \in Q$. By integrating several times the above inequality, we can easily deduce that for every j between 0 and m, and for every $x, y \in \mathbb{R}^n$, we have:

$$\begin{aligned} &\left\| f^{(m-j)}(y) - f^{(m-j)}(x) - f^{(m-j+1)}(x)[y-x] - \cdots - \frac{1}{j!} f^{(m)}(x)[y-x, \ldots, y-x] \right\| \\ &\leq \frac{M}{(j+1)!} ||y-x||^{j+1}. \end{aligned} \tag{11}$$

Actually, we only need the inequality for $j := m-1$ in our reasoning.

For constructing the first function of our estimate sequence, we choose a starting point $x_0 \in Q$, and set:

$$\phi_0(x) := f(x_0) + \frac{M}{(m+1)!} ||x - x_0||^{m+1}.$$

Then, following the construction outlined in Proposition 2.2, we define

$$\phi_{k+1}(x) = (1-\alpha_k)\phi_k(x) + \alpha_k (f(y_k) + \langle f'(y_k), x - y_k \rangle)$$

for an appropriate choice of α_k and y_k.

Let us restrict our analysis to the case where the norm $||\cdot||$ is a matrix induced norm as in the previous section. Lemma 8.2 provides us with a constant c_{m+1} such that the function

$$\chi_k(x, y) := \frac{M c_{m+1}}{(m+1)!} ||y-x||^{m+1} \quad \forall k \geq 0$$

can be used in Lemma 2.3. As for cubic regularization, our strategy for exploiting this lemma consists in trying to find at every iteration k a point y_k that satisfies the inequality:

$$\langle f'(y_k), (1-\alpha_k)x_k + \alpha_k v_k - y_k\rangle + \min_{y\in Q}\{\alpha_k\langle f'(y_k), y - v_k\rangle + \lambda_{k+1}\chi_k(y, v_k)\} \geq 0. \qquad (12)$$

The structure of our construction parallels the one for cubic regularization. Our main tool is the minimizer:

$$x_N(x) := \arg\min_{y\in Q}\left\{f(x) + f'(x)[y-x] + \cdots + \frac{1}{m!}f^{(m)}(x)[y-x,\ldots,y-x] + \frac{N}{(m+1)!}||y-x||^{m+1}\right\},$$

where $N \geq M$. A necessary optimality condition reads, with $r_N(x) := ||x - x_N(x)||$:

$$\begin{aligned}&\left\langle f'(x) + f''(x)[x_N(x)-x] + \cdots + \frac{1}{(m-1)!}f^{(m)}(x)[x_N(x)-x,\cdots,x_N(x)-x], y - x_N(x)\right\rangle \\ &+\frac{Nr_N(x)^{m-1}}{m!}\left\langle G(x_N(x)-x), y - x_N(x)\right\rangle \geq 0 \quad \forall y \in Q. \qquad (13)\end{aligned}$$

Let us extend the two lemmas of the previous section. We omit the proof of the first one, as it is a trivial extension of the one of Lemma 4.1.

Lemma 4.3 *Let $g \in \mathbb{R}^n$, $\lambda > 0$, $x, v \in Q$ and $z := (1-\alpha)x + \alpha v$ for an $\alpha \in [0,1]$. We have:*

$$\min_{y\in Q}\{\alpha\langle g, y - v\rangle + \lambda\chi_k(y,v)\} \geq \min_{y\in Q}\left\{\langle g, y - z\rangle + \frac{\lambda}{\alpha^{m+1}}\chi_k(y,z)\right\}.$$

Lemma 4.4 *For every $x, y \in Q$, we have*

$$\langle f'(x_N(x)), y - x_N(x)\rangle \geq -\frac{M+N}{m!}r_N(x)^m||y-x|| + \frac{N-M}{m!}r_N(x)^{m+1}.$$

Proof
First, we can use (11) to get:

$$\begin{aligned}&\left\langle f'(x) + \cdots + \frac{1}{(m-1)!}f^{(m)}(x)[x_N(x)-x,\cdots,x_N(x)-x] - f'(x_N(x)), y - x_N(x)\right\rangle \\ &\leq \left\|f'(x) + \cdots + \frac{1}{(m-1)!}f^{(m)}(x)[x_N(x)-x,\cdots,x_N(x)-x] - f'(x_N(x))\right\|_* ||y - x_N(x)|| \\ &\leq \frac{M}{m!}r_N(x)^m||y-x_N(x)|| \leq \frac{M}{m!}r_N(x)^m\big(||y-x|| + r_N(x)\big).\end{aligned}$$

Using the latter inequality in (13), we get:

$$\left\langle f'(x_N(x)), y - x_N(x)\right\rangle + \frac{M}{m!}r_N(x)^m(||y-x|| + r_N(x)) + \frac{Nr_N(x)^{m-1}}{m!}\left\langle G(x_N(x)-x), y - x_N(x)\right\rangle \geq 0.$$

It remains to use

$$\left\langle G(x_N(x)-x), y - x_N(x)\right\rangle \leq r_N(x)||y-x|| - r_N(x)^2$$

to get the desired inequality. ■

The m-regularization algorithm looks as follows.

Algorithm 4.2
Assumptions: *f is convex and has a Lipschitz continuous m-th differential with constant M for the norm $||\cdot||$.*
Choose $x_0 \in Q$, set $v_0 := x_0$ and $\lambda_0 := 1$.
Set $\phi_0(x) := f(x_0) + M||x - x_0||^{m+1}/(m+1)!$.
For $k \geq 0$,
Find α_k such that $(2m+2)\alpha_k^{m+1} = c_{m+1}(1-\alpha_k)\lambda_k$.
Set $\lambda_{k+1} := (1-\alpha_k)\lambda_k$.
Set $z_k := \alpha_k v_k + (1-\alpha_k)x_k$.
Set $y_k := \arg\min_{y\in Q}\{\langle f'(z_k), y - z_k\rangle + \cdots + \frac{1}{m!}\langle f^{(m)}(z_k)[\cdots], y - z_k\rangle + \frac{(2m+1)M}{(m+1)!}||y - z_k||^{m+1}\}$.
Set $x_{k+1} := y_k$.
Set $\phi_{k+1}(x) := (1-\alpha_k)\phi_k(x) + \alpha_k(f(y_k) + \langle f'(y_k), x - y_k\rangle)$.
Set $v_{k+1} := \arg\min_{x\in Q}\phi_{k+1}(x)$.
End ■

Theorem 4.2 *The above algorithm takes not more than*

$$\left\lceil \frac{2m+2}{c_{m+1}} \left(\frac{1}{\epsilon}\right)^{1/(m+1)} \left(f(x_0) - f(x^*) + \frac{M}{(m+1)!}||x_0 - x^*||^{m+1}\right)^{1/(m+1)} \right\rceil$$

iterations to find a point x_k for which $f(x_k) - f^* \leq \epsilon$.

Proof
The proof is nothing more than an adaptation of the demonstration of Theorem 4.1. With $z_k := \alpha_k v_k + (1-\alpha_k)x_k$ and $y_k := x_N(z_k)$, the inequality (12) becomes

$$\langle f'(y_k), z_k - y_k\rangle + \min_{y\in Q}\{\alpha_k\langle f'(y_k), y - v_k\rangle + \lambda_{k+1}\chi_k(y, v_k)\} \geq 0.$$

Applying successively Lemma 4.3 and Lemma 4.4, we can transform this inequality into:

$$\begin{aligned}
&\langle f'(y_k), z_k - y_k\rangle + \min_{y\in Q}\{\alpha_k\langle f'(y_k), y - v_k\rangle + \lambda_{k+1}\chi_k(y, v_k)\} \\
\geq\ &\langle f'(y_k), z_k - y_k\rangle + \min_{y\in Q}\left\{\langle f'(y_k), y - z_k\rangle + \frac{\lambda_{k+1}}{\alpha_k^{m+1}}\chi_k(y, z_k)\right\} \\
\geq\ &\min_{y\in Q}\left\{-\frac{M+N}{m!}r_N(x)^m||y - x|| + \frac{N-M}{m!}r_N(z_k)^{m+1} + \frac{\lambda_{k+1}}{\alpha_k^{m+1}}\chi_k(y, z_k)\right\} \\
\geq\ &\frac{N-M}{m!}r_N(z_k)^{m+1} + \min_{t\geq 0}\left\{-\frac{M+N}{m!}r_N(z_k)^m t + \frac{\lambda_{k+1}}{\alpha_k^{m+1}}\frac{Mc_{m+1}}{(m+1)!}t^{m+1}\right\} \\
=\ &\frac{r_N(z_k)^{m+1}}{m!}\left((N-M) - \frac{m}{m+1}\left(\frac{(M+N)^{m+1}\alpha_k^{m+1}}{Mc_{m+1}\lambda_{k+1}}\right)^{1/m}\right).
\end{aligned}$$

This quantity is nonnegative as soon as:

$$c_{m+1}\frac{M(N-M)^m}{(N+M)^{m+1}}\left(\frac{m+1}{m}\right)^m \geq \frac{\alpha_k^{m+1}}{\lambda_{k+1}}.$$

Maximizing the left-hand side with respect to N, we get a value of $c_{m+1}/(2m+2)$, attained for $N = (2m+1)M$. It remains to apply Lemma 8.1. ■

Interestingly, the case $m := 1$, we get a new algorithm for minimizing a convex function with a Lipschitz continuous gradient. However, we need in this algorithm to evaluate the gradient of the function f in two points at every iteration, instead of just one as in Algorithms 3.1 and 3.2.

We conclude this section with a short note on solving the equation:

$$(2m+2)\alpha_k^{m+1} = c_{m+1}(1-\alpha_k)\lambda_k.$$

With $\gamma := c_{m+1}\lambda_k/(2m+2) > 0$, the equation to solve has the form $p(t) = t^{m+1} + \gamma t - \gamma$. As $p(0) < 0 < p(1)$ and $p'(t) > 0$ on $t \in [0,1]$, this equation has a unique solution and can be solved in a few steps of Newton's algorithm initialized at 0.

4.3 Cubic and m-th regularization for unconstrained problems

It is possible to improve some constants in the complexity analysis for cubic regularization and m-th regularization when problems are unconstrained. We consider here a function f with a M-Lipschitz continuous m-th differential. From the viewpoint of our complexity analysis, the case $m = 2$ does not bear anything special.

There are essentially two elements in the proof that change with respect to the constrained situation. Firstly, the inequality (12) that we need to check can be simplified because we can compute the exact value of the minimum. It can be rewritten as:

$$\begin{aligned}\langle f'(y_k), (1-\alpha_k)x_k + \alpha_k v_k - y_k\rangle &\geq -\min_{y\in\mathbb{R}^n}\{\alpha_k\langle f'(y_k), y - v_k\rangle + \lambda_{k+1}\chi_k(y, v_k)\} \\ &= \frac{m}{m+1}\left(\frac{m!}{c_{m+1}}\right)^{1/m}\left(\frac{\alpha_k^{m+1}}{\lambda_{k+1}}\frac{||f'(y_k)||_*^{m+1}}{M}\right)^{1/m}. \qquad (14)\end{aligned}$$

Secondly, the form of the optimality condition for $x_N(x)$ changes. When $Q = \mathbb{R}^n$ in (13), we have for all $x \in \mathbb{R}^n$:

$$f'(x) + \cdots + \frac{1}{(m-1)!}f^{(m)}(x)[x_N(x) - x, \cdots, x_N(x) - x] + \frac{Nr_N(x)^{m-1}}{m!}G(x_N(x) - x) = 0. \quad (15)$$

This relation allows us to get a kind of counterpart to Lemma 4.4.

Lemma 4.5 *For every $x \in \mathbb{R}^n$ and $N \geq M$, we have*

$$||f'(x_N(x))||_*^2 \leq \frac{(m+1)N}{m\cdot m!}\left(\frac{m-1}{m}\frac{Nm!}{N^2 - M^2}\right)^{\frac{m-1}{m+1}}\langle f'(x_N(x)), x - x_N(x)\rangle^{\frac{2m}{m+1}}.$$

Proof
We can transform the Lipschitz continuity of $f^{(m)}$ (see (11) with $j := m-1$) using the optimality

condition (15):

$$\begin{aligned}
0 &\leq \left(\frac{M}{m!}\right)^2 r_N(x)^{2m} - \left\| f'(x_N(x)) - f'(x) - \cdots - \frac{1}{(m-1)!} f^{(m)}(x)[x_N(x)-x, \ldots, x_N(x)-x] \right\|_*^2 \\
&= \left(\frac{M}{m!}\right)^2 r_N(x)^{2m} - \left\| f'(x_N(x)) + \frac{N r_N(x)^{m-1}}{m!} G(x_N(x)-x) \right\|_*^2 \\
&= \frac{M^2 - N^2}{(m!)^2} r_N(x)^{2m} - \|f'(x_N(x))\|_*^2 - \frac{2N r_N(x)^{m-1}}{m!} \langle f'(x_N(x)), x_N(x) - x \rangle.
\end{aligned}$$

Since $N \geq M$, we can see that $\langle f'(x_N(x)), x_N(x) - x \rangle$ is negative. Also:

$$\begin{aligned}
||f'(x_N(x))||_*^2 &\leq \frac{M^2 - N^2}{(m!)^2} r_N(x)^{2m} - \frac{2N r_N(x)^{m-1}}{m!} \langle f'(x_N(x)), x_N(x) - x \rangle \\
&\leq \frac{(m+1)N}{m \cdot m!} \left(\frac{m-1}{m} \frac{N m!}{N^2 - M^2} \right)^{\frac{m-1}{m+1}} \langle f'(x_N(x)), x - x_N(x) \rangle^{\frac{2m}{m+1}}.
\end{aligned}$$

The last bound comes from the maximization of its left-hand side with respect to $r_N(x)$. ■

Now, if we take $x := (1-\alpha_k)x_k + \alpha_k v_k$ in the previous lemma, the inequality resembles strikingly to the desired inequality (14), provided that we choose $y_k := x_N(x)$. In light of the previous lemma, the following relation ensures that (14) is satisfied:

$$\frac{m}{m+1} \left(\frac{m!}{c_{m+1}}\right)^{\frac{1}{m}} \left(\frac{\alpha_k^{m+1}}{\lambda_{k+1}}\right)^{\frac{1}{m}} \left(\frac{1}{M}\right)^{\frac{1}{m}} \leq \left(\frac{m \cdot m!}{(m+1)N}\right)^{\frac{m+1}{2m}} \left(\frac{m}{m-1} \frac{N^2 - M^2}{N m!}\right)^{\frac{m-1}{2m}}.$$

We can reformulate this inequality as:

$$\frac{\alpha_k^{m+1}}{\lambda_{k+1}} \leq c_{m+1} \left(\frac{m+1}{m-1}\right)^{\frac{m-1}{2}} \frac{M (N^2 - M^2)^{\frac{m-1}{2}}}{N^m}.$$

Maximizing the right-hand side with respect to N, we get:

$$\frac{\alpha_k^{m+1}}{\lambda_{k+1}} \leq \frac{c_{m+1}}{\sqrt{m}} \left(\frac{m+1}{m}\right)^{\frac{m-1}{2}}.$$

and this optimum is attained for $N := \sqrt{m}M$. Comparing this value with the one obtained in the previous section, we see that the improvement is rather significant: their ratio is as large as:

$$2\sqrt{\frac{(m+1)^{m+1}}{m^m}},$$

that is, of order $\mathcal{O}(\sqrt{m})$ for large values of m. In particular, for cubic regularization ($m = 2$) our constant equals $\sqrt{3}/4$, while we obtained only $1/12$ in the constrained case.

The algorithm now reads as follows:

Algorithm 4.3
Assumptions: *f is convex and has a Lipschitz continuous m-th differential with constant M for the norm $||\cdot|| = \langle G\cdot,\cdot\rangle^{1/2}$; $Q \equiv \mathbb{R}^n$.*
Choose $x_0 \in \mathbb{R}^n$, set $v_0 := x_0$ and $\lambda_0 := 1$.
Set $\phi_0(x) := f(x_0) + M||x - x_0||^{m+1}/m!$.
For *$k \geq 0$,*
Find α_k such that $\alpha_k^{m+1} = \frac{c_{m+1}}{\sqrt{m}} \left(\frac{m+1}{m}\right)^{\frac{m-1}{2}} (1-\alpha_k)\lambda_k$.
Set $\lambda_{k+1} := (1-\alpha_k)\lambda_k$.
Set $z_k := \alpha_k v_k + (1-\alpha_k)x_k$.
Set $y_k := \arg\min_{y\in Q}\{\langle f'(z_k), y - z_k\rangle + \cdots + \frac{1}{m!}\langle f^{(m)}(z_k)[\cdots], y - z_k\rangle + \frac{\sqrt{m}M}{(m+1)!}||y - z_k||^{m+1}\}$.
Set $x_{k+1} := y_k$.
Set $\phi_{k+1}(x) := (1-\alpha_k)\phi_k(x) + \alpha_k(f(y_k) + \langle f'(y_k), x - y_k\rangle)$.
Set $v_{k+1} := \arg\min_{x\in Q} \phi_{k+1}(x)$.
End ■

The above algorithm does not take more than:

$$\left\lceil \frac{\sqrt{m}}{c_{m+1}} \left(\frac{m}{m+1}\right)^{(m-1)/2} \left(\frac{1}{\epsilon}\right)^{1/(m+1)} \left(f(x_0) - f(x^*) + \frac{M}{(m+1)!}||x_0 - x^*||^{m+1}\right)^{1/(m+1)} \right\rceil$$

iterations to find an ϵ-approximate solution.

5 Estimate sequences and approximations

In the course of an algorithm based on estimate sequences, we must compute at every iteration the minimizer of usually two optimization problems. In order to accelerate the scheme, we can consider solving these subproblems only approximately. A natural question arises immediately: how does the accuracy of the resolution of these subproblems relate to the precision of the final answer given by the algorithm? In particular, how do the successive errors accumulate in the course of the algorithm? For some optimization problems, computing the needed differentials can be quite hard to do accurately (e.g. in Stochastic Optimization, see [SN05], or in large-scaled Semidefinite Optimization, see [d'A08]). How precisely must we compute these differentials in order to avoid any accumulation of error? How can we combine these approximations with an accelerated computation of subproblems? We answer these questions in this section for Algorithms 3.1, 3.1, and 4.2.

5.1 Inexact resolution of intermediate problems

In this subsection, we assume that we have access to accurate gradients of the objective function f, but that we do not have the time or the patience of computing v_k and/or x_{k+1} (in Algorithm 3.1) or y_k (in the other algorithms).

In order to carry out our analysis, we need to formulate a few assumptions on the original optimization problem. First, we assume that the feasible set Q is compact. Fixing a norm $||\cdot||$ on $\mathbb{R}^n$, we denote the finite diameter of Q by:

$$D_Q := \sup\{||y - x|| : x, y \in Q\}.$$

Second, we must formulate some regularity assumptions on the function ϕ_0. In view of the examples studied in the previous sections, we can consider bounds of the form:

$$L_0||y-x||^2 \geq \langle \phi_0'(y) - \phi_0'(x), y-x\rangle \geq \sigma_0||y-x||^p \quad \text{for every } x,y \in Q, \tag{16}$$

where $p \geq 1$ is an appropriately chosen number, and $L_0, \sigma_0 \geq 0$. We easily deduce that for every x and y, the following inequality holds:

$$\frac{L_0}{2}||y-x||^2 \geq \phi_0(y) - \phi_0(x) - \langle \phi_0'(x), y-x\rangle \geq \frac{\sigma_0}{p}||y-x||^p. \tag{17}$$

Also, we can write in view of Theorem 2.1.5 in [Nes03]:

$$L_0||y-x|| \geq ||\phi_0'(y) - \phi_0'(x)||_*. \tag{18}$$

Before studying the effect of solving subproblems inexactly, let us check that the above condition is satisfied in two most typical settings. The following lemma deals with the situation we have considered in Section 4.

Lemma 5.1 *Consider a matrix induced norm $||\cdot|| = \langle G\cdot, \cdot\rangle^{1/2}$ and a number $m \geq 1$. The inequality (16) is satisfied when*

$$\phi_0(x) = f(x_0) + \frac{M}{(m+1)!}||x-x_0||^{m+1},$$

with $p := m+1$, $L_0 := MD_Q^{p-2}/(p-2)!$, and $\sigma_0 := 2Mc_p/p!$, where c_p is given in Lemma 8.2. When $m = 1$, one can take $L_0 = \sigma_0 = M$.

Proof

Let $p := m+1$. In view of Lemma 8.2, we have:

$$\phi_0(y) - \phi_0(x) - \langle \phi_0'(x), y-x\rangle \geq \frac{M}{p!}c_p||y-x||^p$$

for every $x, y \in Q$. Adding this inequality to the one obtained by inverting x and y, we obtained the desired value of σ. For proving the upper bound, let $F_p(x,y) := \langle \phi_0'(y) - \phi_0'(x), y-x\rangle(p-1)!/M$, and bound $F_p(x,y)/||y-x||^2$ from above. Without loss of generality, we can assume that $x_0 = 0$. First,

$$\begin{aligned}\max_{x\neq y\in Q} \frac{F_p(x,y)}{||y-x||^2} &= \max_{x\neq y\in Q} \frac{||y||^p - (||y||^{p-2} + ||x||^{p-2})\langle Gy, x\rangle + ||x||^p}{||y||^2 - 2\langle Gy, x\rangle + ||x||^2} \\ &= \max_{\substack{x\neq y\in Q \\ -1\leq\alpha\leq 1}} \frac{||y||^p - (||y||^{p-1}||x|| + ||y||\cdot||x||^{p-1})\alpha + ||x||^p}{||y||^2 - 2||y||\cdot||x||\alpha + ||x||^2}\end{aligned}$$

Fixing two distinct points x and y in Q, we denote by $\psi(\alpha)$ the above right-hand side. After some trivial rearrangements, the numerator of its derivative is:

$$\begin{aligned}&2(||y||^p + ||x||^p)||y||\cdot||x|| - (||y||^{p-1}||x|| + ||y||\cdot||x||^{p-1})(||y||^2 + ||x||^2) \\ &= ||y||\cdot||x||(||y||^{p-2} - ||x||^{p-2})(||y||^2 - ||x||^2),\end{aligned}$$

which is nonnegative, thus the maximum of $\psi(\alpha)$ on $[-1,1]$ is attained when $\alpha = 1$. The maximum of $F_p(x,y)/||y-x||^2$ reduces to:

$$\begin{aligned}
&\max_{\substack{x\neq y\in Q\\ y=tx}} \frac{||y||^p - (||y||^{p-1}||x|| + ||y||\cdot||x||^{p-1}) + ||x||^p}{(||y|| - ||x||)^2} \\
&\leq \max_{x\neq y\in Q} \frac{(||y|| - ||x||)(||y||^{p-1} - ||x||^{p-1})}{(||y|| - ||x||)^2} \\
&= \max_{x\neq y\in Q} (||y||^{p-2} + ||y||^{p-3}||x|| + \cdots + ||x||^{p-2}) \\
&\leq (p-1)D_Q^{p-2}.
\end{aligned}$$

When $p = 2$, $\langle \phi_0'(y) - \phi_0'(x), y - x\rangle = M||y-x||^2$, and $L_0 = \sigma_0 = M$ works. ■

The entropy function is a common choice for constructing ϕ_0 in Algorithm 3.1 or 3.2. In this setting, used when the feasible set Q is a simplex, that is, $Q := \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}$, we choose the prox-function d as follows:

$$d(x) := \sum_{i=1}^n x_i \ln(x_i) + \ln(n).$$

Its minimizer on Q is the all-$1/n$ vector, and the second inequality in (16) holds with $\sigma := 1$ when we use the 1-norm (see Lemma 3 in [Nes05] for a proof that $d(x)$ is 1-strongly convex on Q or that norm; we show below an extension of this result). However, the first inequality does not hold[1], and our result cannot be applied. Nevertheless, Ben-Tal and Nemirovski [BTN05] suggest a slight modification of d to regularize this function. Let $\delta > 0$ and

$$d_\delta(x) := \sum_{i=1}^n \left(x_i + \frac{\delta}{n}\right) \ln\left(x_i + \frac{\delta}{n}\right) - (1+\delta)\ln\left(\frac{1+\delta}{n}\right).$$

Lemma 5.2 *Using the 1-norm for $||\cdot||$, we have:*

$$L_0||y-x||_1^2 \geq \langle d_\delta'(y) - d_\delta'(x), y - x\rangle \geq \sigma_0||y-x||_1^2 \quad \text{for every } x, y \in Q, \tag{19}$$

where $L_0 := n/\delta$, $\sigma_0 := 1/(1+\delta)$.

Proof

We have to show that $\langle d_\delta''(x)h, h\rangle/||h||_1^2$ is bounded from above by n/δ and from below by $1/(1+\delta)$. Using Cauchy-Schwartz's Inequality, we have for every $x \in Q$:

$$\begin{aligned}
\langle d_\delta''(x)h, h\rangle = \sum_{i=1}^n \frac{h_i^2}{x_i + \delta/n} &= \left(\sum_{i=1}^n \frac{h_i^2}{x_i + \delta/n}\right)\left(\sum_{i=1}^n x_i + \frac{\delta}{n}\right) - \delta\left(\sum_{i=1}^n \frac{h_i^2}{x_i + \delta/n}\right) \\
&\geq \left(\sum_{i=1}^n |h_i|\right)^2 - \delta\left(\sum_{i=1}^n \frac{h_i^2}{x_i + \delta/n}\right) = ||h||_1^2 - \delta\langle d_\delta''(x)h, h\rangle.
\end{aligned}$$

[1]For a simple check of this assertion, consider $y_\epsilon := (1-(n-1)\epsilon, \epsilon, \ldots, \epsilon)^T$ and $x := (1/n, \ldots, 1/n)^T$, for $1/n > \epsilon > 0$. As $||y_\epsilon - x||_1 = 2(n-1)(1/n - \epsilon) < 2$, and $\langle d'(y_\epsilon) - d'(x), y_\epsilon - x\rangle = (n-1)(1/n-\epsilon)\ln(1/\epsilon - (n-1))$ is unbounded when $\epsilon \to 0$, the upper bound in (16) cannot be guaranteed, whatever L_0 is.

From the other side, we have:

$$\sum_{i=1}^{n} \frac{h_i^2}{x_i + \delta/n} \leq \sum_{i=1}^{n} \frac{h_i^2}{\delta/n} \leq \frac{n}{\delta}||h||_1^2.$$

■

The following proposition indicates the effect of constructing an approximate minimizer $\hat{v}_{k+1}$ to ϕ_{k+1} and an approximate point $\hat{x}_{k+1}$ on the fundamental inequality $\phi_k(v_k) \geq f(x_k)$. At the end of this subsection, we particularize this proposition to the three algorithms under consideration. The notation comes from Lemma 2.3.

Proposition 5.3 *Assume that inequality (16) holds, and that the following slight extension of Inequality (2) in Lemma 2.3 is satisfied with functions $\chi_k(x, y) \geq \sigma_0||y - x||^p/p$:*

$$\phi_0(y) \geq \phi_0(x) + \langle \phi_0'(x), y - x\rangle + \chi_k(y, x) \quad \textit{for all } x, y \in Q. \tag{20}$$

Let $\epsilon \geq 0$, let $\gamma \in [0, 1]$ and fix $k \geq 0$. Assume that

$$\hat{x}_k, \hat{v}_k \in Q, \quad \textit{and} \quad \min_{x \in Q} \phi_k(x) \geq f(\hat{x}_k) - \epsilon.$$

Suppose that the accuracy $\hat{\epsilon}_k$ by which we have computed $\hat{v}_k$, that is a constant verifying $\hat{\epsilon}_k \geq \phi_k(\hat{v}_k) - \phi_k(v_k)$, satisfies the following bound:

$$0 \leq \hat{\epsilon}_k \leq \min\left\{1, \left(\frac{\alpha_k\gamma}{1-\alpha_k}\right)^p \left(\frac{\epsilon}{1 + D_Q L_0 \sqrt[p]{p\lambda_k^{p-1}/\sigma_0}}\right)^p\right\}, \tag{21}$$

and suppose that the accuracy by which we compute $\hat{x}_{k+1}$ guarantees:

$$f(y_k)+\langle g(y_k), (1-\alpha_k)\hat{x}_k+\alpha_k\hat{v}_k-y_k\rangle+\min_{x\in Q}\{\alpha_k\langle g(y_k), x-\hat{v}_k\rangle + \lambda_{k+1}\chi(x,\hat{v}_k)\} \geq f(\hat{x}_{k+1})-\alpha_k(1-\gamma)\epsilon, \tag{22}$$

where $g(y_k) \in \partial f(y_k)$. Then:

$$\min_{x \in Q} \phi_{k+1}(x) \geq f(\hat{x}_{k+1}) - \epsilon.$$

Moreover, if

$$\epsilon \leq \frac{\lambda_{k+1}}{\alpha_k\gamma}\left(\frac{D_Q^p L_0^p}{\sigma_0}\right)^{\frac{1}{p-1}}, \tag{23}$$

the bound on $\hat{\epsilon}_k$ can be improved to:

$$0 \leq \hat{\epsilon}_k \leq \left(\frac{\alpha_k\gamma}{1-\alpha_k}\right)^p \left(\frac{\epsilon}{D_Q L_0}\right)^p \left(\frac{\sigma_0}{p\lambda_k^{p-1}}\right). \tag{24}$$

Proof
Let us fix $k \geq 0$. Observe that the condition (20) implies:

$$\phi_k(y) \geq \phi_k(x) + \langle \phi_k'(x), y - x\rangle + \lambda_k\chi_k(y, x) \quad \text{for all } x, y \in Q, k \geq 0.$$

First, we bound $\min_{x\in Q}\langle\phi_k'(\hat{v}_k), x-\hat{v}_k\rangle$ from below. Obviously, the function ϕ_k has a Lipschitz continuous gradient with constant $\lambda_k L_0$. Observe that, in view of (18):

$$L_0\lambda_k||\hat{v}_k - v_k|| \geq ||\phi_k'(\hat{v}_k) - \phi_k'(v_k)||_*,$$

and

$$\hat{\epsilon}_k \geq \phi_k(\hat{v}_k) - \phi_k(v_k) \geq \frac{\sigma_0\lambda_k}{p}||\hat{v}_k - v_k||^p. \tag{25}$$

$$\begin{aligned}
\min_{x\in Q}\langle\phi_k'(\hat{v}_k), x-\hat{v}_k\rangle &\geq \min_{x\in Q}\{\langle\phi_k'(v_k), x-\hat{v}_k\rangle - ||\phi_k'(\hat{v}_k) - \phi_k'(v_k)||_*||x-\hat{v}_k||\} \\
&\geq \min_{x\in Q}\{\langle\phi_k'(v_k), x-v_k\rangle + \langle\phi_k'(v_k), v_k-\hat{v}_k\rangle - L_0\lambda_k||\hat{v}_k-v_k||\cdot||x-\hat{v}_k||\} \\
&\geq \langle\phi_k'(v_k), v_k-\hat{v}_k\rangle - L_0\lambda_k||\hat{v}_k-v_k||D_Q \\
&\geq \phi_k(v_k) - \phi_k(\hat{v}_k) + \lambda_k\chi_k(v_k,\hat{v}_k) - L_0\lambda_k||\hat{v}_k-v_k||D_Q \\
&\geq -\hat{\epsilon}_k + \frac{\sigma_0\lambda_k}{p}||\hat{v}_k-v_k||^p - L_0\lambda_k||\hat{v}_k-v_k||D_Q.
\end{aligned}$$

Now, the function $t\mapsto\sigma_0 t^p/p - L_0 D_Q t$ is decreasing in $[0,t^*]$, where $t^* := (L_0 D_Q/\sigma_0)^{1/(p-1)}$. We know from (25) that we can estimate $||\hat{v}_k - v_k||$ by $\hat{t} := \sqrt[p]{p\hat{\epsilon}_k/\sigma_0\lambda_k}$. If $\hat{t}\leq t^*$, we can write the following bound:

$$\min_{x\in Q}\langle\phi_k'(\hat{v}_k), x-\hat{v}_k\rangle \geq -L_0\lambda_k\sqrt[p]{\frac{p\hat{\epsilon}_k}{\sigma_0\lambda_k}}D_Q.$$

Observe that $\hat{t}\leq t^*$ is ensured when (23) and (24) hold. Also, the bound (24) implies:

$$\min_{x\in Q}\langle\phi_k'(\hat{v}_k), x-\hat{v}_k\rangle \geq -\frac{\alpha_k\gamma\epsilon}{1-\alpha_k}.$$

If we cannot guarantee that $\hat{t}\leq t^*$, we can use the following slightly less favorable estimation, provided that $\hat{\epsilon}_k\leq 1$:

$$\begin{aligned}
\min_{x\in Q}\langle\phi_k'(\hat{v}_k), x-\hat{v}_k\rangle &\geq -\hat{\epsilon}_k + \frac{\sigma_0\lambda_k}{p}||\hat{v}_k-v_k||^p - L_0\lambda_k||\hat{v}_k-v_k||D_Q. \\
&\geq -\hat{\epsilon}_k - L_0\lambda_k||\hat{v}_k-v_k||D_Q \geq -\hat{\epsilon}_k - L_0\lambda_k\sqrt[p]{\frac{p\hat{\epsilon}_k}{\sigma_0\lambda_k}}D_Q. \\
&\geq -\sqrt[p]{\hat{\epsilon}_k} - L_0\lambda_k\sqrt[p]{\frac{p\hat{\epsilon}_k}{\sigma_0\lambda_k}}D_Q = -\sqrt[p]{\hat{\epsilon}_k}\left(1 + L_0\lambda_k\sqrt[p]{\frac{p}{\sigma_0\lambda_k}}D_Q\right).
\end{aligned}$$

The bound (21) on $\hat{\epsilon}_k$ ensures that:

$$\min_{x\in Q}\langle\phi_k'(\hat{v}_k), x-\hat{v}_k\rangle \geq -\frac{\alpha_k\gamma\epsilon}{1-\alpha_k}.$$

We conclude our proof by following essentially the same steps as in the argument of Lemma 2.3:

$$\begin{aligned}
&\min_{x\in Q}\phi_{k+1}(x) = \min_{x\in Q}\Big\{(1-\alpha_k)\phi_k(x) + \alpha_k\Big[f(y_k) + \langle g(y_k), x-y_k\rangle\Big]\Big\}\\
\geq\ &\min_{x\in Q}\Big\{(1-\alpha_k)\Big[\phi_k(\hat{v}_k) + \langle \phi_k'(\hat{v}_k), x-\hat{v}_k\rangle\Big] + \lambda_{k+1}\chi_k(x,\hat{v}_k) + \alpha_k\Big[f(y_k) + \langle g(y_k), x-y_k\rangle\Big]\Big\}\\
\geq\ &\min_{x\in Q}\Big\{(1-\alpha_k)\phi_k(\hat{v}_k) - \alpha_k\gamma\epsilon + \lambda_{k+1}\chi_k(x,\hat{v}_k) + \alpha_k\Big[f(y_k) + \langle g(y_k), x-y_k\rangle\Big]\Big\}\\
\geq\ &(1-\alpha_k)f(\hat{x}_k) - (1-\alpha_k+\alpha_k\gamma)\epsilon + \alpha_k\Big[f(y_k) + \langle g(y_k), \hat{v}_k - y_k\rangle\Big]\\
&+\min_{x\in Q}\{\alpha_k\langle g(y_k), x-\hat{v}_k\rangle + \lambda_{k+1}\chi_k(x,\hat{v}_k)\}\\
\geq\ &f(y_k) + \langle g(y_k), (1-\alpha_k)\hat{x}_k + \alpha_k\hat{v}_k - y_k\rangle - (1-\alpha_k+\alpha_k\gamma)\epsilon\\
&+\min_{x\in Q}\{\alpha_k\langle g(y_k), x-\hat{v}_k\rangle + \lambda_{k+1}\chi_k(x,\hat{v}_k)\}\\
\geq\ &f(\hat{x}_{k+1}) - \epsilon.
\end{aligned}$$

■

The previous proposition has a clear meaning with respect to the computation of $\hat{v}_k$. The designer of an estimate sequence scheme must take a particular care in the choice of ϕ_0 because the computation of v_k, that is, of a minimizer of ϕ_0 plus a linear term over Q, must be done quite accurately: for instance, for Algorithms 3.1 and 3.2, we have $\alpha_k = \Theta(1/k) \geq \Omega(\sqrt{L/\epsilon})$, and we get the lower bound $\epsilon_k \geq \Omega(\gamma^2\epsilon^3/L)$. It would be desirable that this computation remains relatively cheap, or even that this minimizer can be computed analytically.

As far as the criterion (22) on the accuracy of $\hat{x}_{k+1}$ is concerned, it is not difficult to relate it with a condition on how precisely the corresponding intermediate optimization problem has to be solved. Roughly speaking, this intermediate problem must be solved within an accuracy of $(1-\gamma)\alpha_k\epsilon$ or of $(1-\gamma)\alpha_k\epsilon/D_Q$.

For instance, in Algorithms 3.1 and 3.2, the inequality (22) is guaranteed as soon as $h(\hat{x}_{k+1}) - \min_{x\in Q} h_k(x) \leq (1-\gamma)\alpha_k\epsilon$, where $\alpha_k\langle f'(y_k), x-v_k\rangle + \lambda_{k+1}\chi_k(x,v_k)$.

For Algorithm 4.2, we can replace the condition (22) by the following one, provided that the feasible set Q has a finite diameter D_Q:

$$||h_k'(\hat{x}_{k+1})||_* \leq \frac{(1-\gamma)\alpha_k\epsilon}{D_Q},$$

where

$$h_k(x) := \langle f'(z_k), x-z_k\rangle + \cdots + \frac{1}{m!}\langle f^{(m)}(z_k)[x-z_k,\cdots,x-z_k], x-z_k\rangle + \frac{(2m+1)M}{(m+1)!}||x-z_k||^{m+1}.$$

Obviously, this condition implies the following approximate optimality criterion:

$$\langle h_k'(\hat{x}_{k+1}), y-\hat{x}_{k+1}\rangle \geq -(1-\gamma)\alpha_k\epsilon \quad \forall y\in Q.$$

In order to show how the above criterion implies (22), one can easily adapt the technical Lemma 4.4 into:

$$\forall x\in Q \quad \langle f'(\hat{x}_{k+1}), y-\hat{x}_{k+1}\rangle - (1-\gamma)\alpha_k\epsilon \geq -\frac{M+N}{m!}\hat{r}_k^m||y-z_k|| + \frac{N-M}{m!}\hat{r}_k^{m+1},$$

where $\hat{z}_k = (1-\alpha_k)\hat{x}_k + \alpha_k \hat{v}_k$, and $\hat{r}_k := ||\hat{x}_{k+1} - \hat{z}_k||$. Using this inequality and the same argument as in the proof of Theorem 4.2, we can immediately show that the desired inequality (22) holds.

5.2 Approximate subgradients and higher-order differentials

In some circumstances, e.g. in the framework of stochastic optimization where a prior Monte-Carlo sampling is used to approximate the actual objective function (see [SN05]), we do not have access to an exact subgradient of the objective function f.

Specifically, we assume that, for a given accuracy $\epsilon > 0$ and a point x of $\operatorname{dom} f$, we can only determine in a reasonable time a subgradient g that satisfies the two following properties:

$$\forall\, y \in \operatorname{dom} f, \quad f(y) \geq f(x) + \langle g, y - x\rangle - \epsilon, \tag{26}$$

and

$$\forall\, y \in \operatorname{dom} f, \quad f(y) \geq f(x) + \langle g, y - x\rangle - \epsilon||y - x||, \tag{27}$$

where $||\cdot||$ is an appropriate norm. The first inequality is used in Chapter XI of [HUL93b] in the definition of ϵ-subgradients. The second one is defined in Section 1.3 of [Mor05] as *analytic ϵ-subgradients*.

We shall denote the set of the approximate subgradients that satisfy (26) and (27) by $\partial_\epsilon f(x)$. The interest of mixing these two notions of subgradient lies in the fact that we can use affine underestimates for constructing our estimate sequence, and, at the same time, employ the following lemma on the error of the approximate subgradient over the actual one. In a more careful analysis, we could make a distinction between the required accuracy in (26) and (27), defining "(ϵ_1, ϵ_2)-subgradients". It not difficult to incorporate this extra feature in our argument.

The following lemma shows a useful consequence of the inequality (27).

Lemma 5.4 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a closed convex function, let $x \in \operatorname{dom} f$, $\epsilon \geq 0$, and $g \in \mathbb{R}^n$ satisfying (27). Then*

$$||g(x) - g||_* \leq \epsilon$$

for a subgradient $g(x) \in \partial f(x)$, where the norm $||\cdot||_*$ is dual to the norm used in (27).

Proof

The subgradient of the function $h(y) := ||y - x||$ in $y = x$ is $B_*[0,1] := \{s \in \mathbb{R}^n : ||s||_* \leq 1\}$. Indeed, for every $\hat{s} \in B_*[0,1]$ and every $y \in \mathbb{R}^n$ we have:

$$h(y) = ||y - x|| = \max_{s \in B_*[0,1]} \langle s, y - x\rangle \geq \langle \hat{s}, y - x\rangle = h(x) + \langle \hat{s}, y - x\rangle.$$

On the other hand, if some $\hat{s}$ verifies $h(y) \geq h(x) + \langle \hat{s}, y - x\rangle$ for every $z \in \mathbb{R}^n$, then $1 \geq \max\{\langle \hat{s}, u\rangle : ||u|| \leq 1\} = ||\hat{s}||_*$.

Let g be a vector satisfying (27), or equivalently $g \in \partial(f + \epsilon h)(x)$. According to Theorem 23.8 in [Roc70], we have:

$$\partial(f + \epsilon h)(x) = \partial f(x) + \epsilon \partial h(x) = \partial f(x) + \epsilon B_*[0,1].$$

Therefore, there exist a $g(x) \in \partial f(x)$ and a $\xi \in B_*[0,1]$ such that $g = g(x) + \epsilon\xi$, which implies $||g - g(x)||_* \leq \epsilon$. ∎

Given an approximate subgradient $g_k \in \partial_{\bar{\epsilon}_k} f(y_k)$ for an $\bar{\epsilon}_k > 0$, a natural candidate for the underestimate f_k is:

$$f_k(x) = f(y_k) + \langle g_k, x - y_k \rangle - \bar{\epsilon}_k. \tag{28}$$

The functions of the estimate sequence become:

$$\phi_k(x) = \lambda_k \phi_0(x) + \sum_{i=0}^{k-1} \frac{\lambda_k \alpha_i}{\lambda_{i+1}} (f(y_i) + \langle g_i, x - y_i \rangle) - \sum_{i=0}^{k-1} \frac{\lambda_k \alpha_i}{\lambda_{i+1}} \bar{\epsilon}_i.$$

The instrumental inequality in Lemma 2.3 can be easily extended to approximate subgradients. Its demonstration follows closely the proof of Lemma 2.3, and we just sketch the small variations between the two proofs.

Lemma 5.5 *We are given an estimate sequence* $\Big((\phi_k)_{k\geq 0}; (\lambda_k)_{k\geq 0}\Big)$ *for the convex problem*

$$\min_{x \in Q} f(x),$$

constructed according to Proposition 2.2 using affine underestimates of f*:*

$$f_k(x) := f(y_k) + \langle g_k, x - y_k \rangle - \bar{\epsilon}_k \quad \text{for some } y_k \in Q, \bar{\epsilon}_k > 0 \text{ and } g_k \in \partial_{\bar{\epsilon}_k} f(y_k).$$

We also assume that ϕ_0 *is continuously differentiable and that we have functions* $\chi_k : Q \times Q \to \mathbb{R}_+$ *for* $k \geq 0$ *such that* $\chi_k(x, y) = 0$ *implies* $x = y$*, and for which:*

$$\phi_0(y) \geq \phi_0(x) + \langle \phi_0'(x), y - x \rangle + \chi_k(y, x) \quad \text{for all } y, x \in Q.$$

If x_k *and* v_k *satisfy* $\phi_k(v_k) \geq f(x_k) - \epsilon$ *for an* $\epsilon > 0$*, then:*

$$\min_{x \in Q} \{\phi_{k+1}(x)\} - f(y_k) - \langle g_k, (1-\alpha_k)x_k + \alpha_k v_k - y_k \rangle - \min_{x \in Q} \{\alpha_k \langle g_k, x - v_k \rangle + \lambda_{k+1} \chi_k(x, v_k)\} \geq \bar{\epsilon}_k + (1-\alpha_k)\epsilon$$

for every $k \geq 0$*.*

Moreover, if $y_k := (1 - \alpha_k)x_k + \alpha_k v_k$*, the right-hand side can be replaced by*

$$\alpha_k (1 + (1 - \alpha_k) D_Q) \bar{\epsilon}_k + (1 - \alpha_k)\epsilon.$$

Proof
We have for every $x \in Q$:

$$\begin{aligned}
\phi_{k+1}(x) &\geq (1-\alpha_k)(f(x_k) - \epsilon) + \lambda_{k+1}\chi_k(x, v_k) + \alpha_k(f(y_k) + \langle g_k, x - y_k \rangle - \bar{\epsilon}_k) \\
&\geq (1-\alpha_k)(f(y_k) + \langle g_k, x_k - y_k \rangle - \bar{\epsilon}_k - \epsilon) + \lambda_{k+1}\chi_k(x, v_k) + \alpha_k(f(y_k) + \langle g_k, x - y_k \rangle - \bar{\epsilon}_k),
\end{aligned}$$

which is exactly the desired inequality. For the second bound, we can proceed as follows. Here $g(y_k)$ is a subgradient of f at y_k such that $||g_k - g(y_k)||_* \leq \bar{\epsilon}_k$, and $y_k := (1 - \alpha_k)x_k + \alpha_k v_k$, like in Algorithm 3.1 and Algorithm 3.2.

$$\begin{aligned}
\phi_{k+1}(x) &\geq (1-\alpha_k)(f(x_k) - \epsilon) + \lambda_{k+1}\chi_k(x, v_k) + \alpha_k(f(y_k) + \langle g_k, x - y_k \rangle - \bar{\epsilon}_k) \\
&\geq (1-\alpha_k)(f(y_k) + \langle g(y_k), x_k - y_k \rangle - \epsilon) + \lambda_{k+1}\chi_k(x, v_k) + \alpha_k(f(y_k) + \langle g_k, x - y_k \rangle - \bar{\epsilon}_k) \\
&= f(y_k) + \alpha_k \langle g(y_k) - g_k, y_k - v_k \rangle - (1-\alpha_k)\epsilon + \lambda_{k+1}\chi_k(x, v_k) + \alpha_k(\langle g_k, x - v_k \rangle - \bar{\epsilon}_k).
\end{aligned}$$

Observe that:

$$\langle g(y_k) - g_k, y_k - v_k \rangle = (1-\alpha_k)\langle g(y_k) - g_k, x_k - v_k \rangle \geq -(1-\alpha_k)\bar{\epsilon}_k D_Q,$$

which yields the desired inequality. ■

Remark 5.1 *For the strongly convex estimate sequences constructed in Section 3, the above lemma implies that no error propagation occur from an iteration to the other if, in regard to the first bound, $\bar{\epsilon}_k \leq \alpha_k \epsilon$. We have seen that $\alpha_k = \Theta(1/k)$, and, given the complexity of these algorithm, is bounded from below by $\Omega(\sqrt{L/\epsilon})$. Thus, the first bound tells us that the subgradient have to be computed with an accuracy with respect to ϵ of $\mathcal{O}(\epsilon^{1.5})$. The second bound is more favorable with respect to ϵ, as it an accuracy of $\mathcal{O}(\epsilon/(1+D_Q))$ is enough to tame the subgradient errors. This latter conclusion corresponds to the bound obtained by d'Aspremont in [d'A08].* ■

It remains to answer two questions in this section. First, how does the imperfect knowledge of a subgradient affects the computation of the point x_{k+1}? Unfortunately, this analysis has to be conducted case by case, because the strategy and the information required for constructing x_{k+1} varies a lot from an algorithm to another. Moreover, for m-th regularization schemes, we also need higher order information on the objective function. Hence, we need an extension of the concept of estimate subgradient to Hessians and higher-order differentials.

The second and seemingly trickier question addresses the effects of approximate subgradient combined with inexact resolution of intermediate problems, and will be addressed in the next subsection.

The answer to the first question for Algorithms 3.1 and 3.2 can be determined reasonably easily, following essentially the line of the derivations done in Section 3, and using Lemma 5.5. However, we need to assume the boundedness of the feasible set Q.

Here, the function f must be differentiable, convex, and has a L-Lipschitz continuous gradient with respect to a norm $||\cdot||$.

Proposition 5.6 *Let $\epsilon > 0$ and assume that for a fixed k the inequality $\phi_k(v_k) \geq f(x_k) - \epsilon$ holds, where ϕ_k is the k-th estimate function generated by Algorithm 3.1 or by Algorithm 3.2. If $0 \leq \bar{\epsilon}_k \leq \epsilon/(2D_Q + 1)$, then*

$$\phi_{k+1}(v_{k+1}) \geq f(x_{k+1}) - \epsilon,$$

with ϕ_{k+1}, x_{k+1}, and v_{k+1} as generated by Algorithm 3.1 or by Algorithm 3.2 where $f'(y_k)$ has been replaced by $g_k \in \partial_{\bar{\epsilon}_k} f(y_k)$.

Proof

The only change with respect to the derivations in Section 3 is the use of the inequality:

$$\langle g_k, x - v_k \rangle \geq \langle g(y_k), x - v_k \rangle - ||g_k - g(y_k)||_* ||x - v_k|| \geq \langle g(y_k), x - v_k \rangle - \bar{\epsilon}_k D_Q \quad \forall x \in Q.$$

for $g(y_k) \in \partial f(y_k) = \{f'(y_k)\}$, as stated in Lemma 5.4. ■

For m-th regularization algorithms, we use the following extension of approximate subgradients. Let f be a m times differentiable real-valued function, and let x be a point in the domain of f. For every $d \geq 2$, we call a d-tensor A_d *an ϵ-approximation of the d-th differential in x* if:

$$||A_d - f^{(d)}(x)|| \leq \epsilon,$$

where we use here the appropriate tensor norm induced by a matrix induced norm $||\cdot||$. We can extend Lemma 4.4 for approximate subgradients. Here, the differential $f^{(m)}$ is assumed to be M-Lipschitz continuous, and f is supposed to be convex.

Lemma 5.7 *Let $\bar{\epsilon}', \epsilon', \epsilon'', \ldots, \epsilon^{(m)} \geq 0$, and A_d be an $\epsilon^{(d)}$-approximation of $f^{(d)}(x)$ for $1 \leq d \leq m$. We write for every $x \in Q$:*

$$\hat{x}_N(x) := \arg\min_{y \in Q} \left\{ \langle A_1, y - x \rangle + \cdots + \frac{1}{m!} A_m[y - x, \ldots, y - x] + \frac{N}{(m+1)!} ||y - x||^{m+1} \right\}.$$

Then, for every $x, y \in Q$ and $g \in \partial_{\bar{\epsilon}'} f(\hat{x}_N(x))$, we have:

$$\langle g, y - \hat{x}_N(x) \rangle \geq -\frac{M+N}{m!} \hat{r}_N(x)^m ||y - x|| + \frac{N-M}{m!} \hat{r}_N(x)^{m+1} - \bar{\epsilon}' D_Q - \sum_{d=1}^{m} \frac{\epsilon^{(d)} D_Q^d}{d!},$$

where $\hat{r}_N(x) := ||\hat{x}_N(x) - x||$.

Proof
In view of the optimality conditions of the problem defining $\hat{x}_N(x)$, we can write, for every $y \in Q$:

$$\begin{aligned}
0 \leq & \left\langle A_1 + A_2[\hat{x}_N(x) - x] + \cdots + \frac{1}{(m-1)!} A_m[\hat{x}_N(x) - x, \cdots, \hat{x}_N(x) - x], y - \hat{x}_N(x) \right\rangle \\
& + \frac{N \hat{r}_N(x)^{m-1}}{m!} \left\langle G(\hat{x}_N(x) - x), y - \hat{x}_N(x) \right\rangle \\
\leq & \left\langle f'(x) + f''(x)[\hat{x}_N(x) - x] + \cdots + \frac{1}{(m-1)!} f^{(m)}(x)[\hat{x}_N(x) - x, \cdots, \hat{x}_N(x) - x], y - \hat{x}_N(x) \right\rangle \\
& + \frac{N \hat{r}_N(x)^{m-1}}{m!} \left\langle G(\hat{x}_N(x) - x), y - \hat{x}_N(x) \right\rangle + \sum_{d=1}^{m} \frac{\epsilon^{(d)} D_Q^d}{d!}.
\end{aligned}$$

Following the same reasoning as in the proof of Lemma 4.4, we get to:

$$0 \leq \langle f'(\hat{x}_N(x)), y - \hat{x}_N(x) \rangle + \frac{M+N}{m!} \hat{r}_N(x)^m ||y - x|| + \frac{M-N}{m!} \hat{r}_N(x)^{m+1} + \sum_{d=1}^{m} \frac{\epsilon^{(d)} D_Q^d}{d!}.$$

It remains to use:

$$\langle f'(\hat{x}_N(x)), y - \hat{x}_N(x) \rangle \leq \langle g, y - \hat{x}_N(x) \rangle + \bar{\epsilon}' D_Q$$

to get the desired inequality. ■

This lemma allows us to understand how the propagation error due to approximate knowledge of differentials can be controlled in the m-th regularization algorithm for constrained problems. Essentially, we need to merge the previous lemma with Lemma 5.5.

Proposition 5.8 *Let $\epsilon > 0$ and assume that for a fixed k the inequality $\phi_k(v_k) \geq f(x_k) - \epsilon$ holds, where ϕ_k is the k-th estimate function generated by Algorithm 4.2. Let $z_k := \alpha_k v_k + (1 - \alpha_k) x_k$. Suppose that we perform one iteration of Algorithm 4.2 where $f'(y_k)$ has been replaced by $g_k \in$*

$\partial_{\bar\epsilon'} f(y_k)$ and $f^{(d)}(z_k)$ by one of its $\epsilon^{(d)}$-approximation A_d, generating the function ϕ_{k+1} and the points x_{k+1}, and v_{k+1}. If

$$\bar\epsilon' + D_Q\epsilon' + \sum_{d=1}^{m} \frac{\epsilon^{(d)} D_Q^d}{d!} \leq \alpha_k \epsilon,$$

then

$$\phi_{k+1}(v_{k+1}) \geq f(x_{k+1}) - \epsilon.$$

Proof
According to Lemma 5.5, we need to check that:

$$\langle g_k, z_k - y_k\rangle + \min_{y\in Q}\left\{\alpha_k\langle g_k, y - v_k\rangle + \lambda_{k+1}\frac{Mc_{m+1}}{(m+1)!}||y-v_k||^{m+1}\right\} - \bar\epsilon' - (1-\alpha_k\epsilon) \geq -\epsilon.$$

Using Lemma 4.3 and Lemma 5.7, we can convert this inequality into the following stronger version:

$$\min_{y\in Q}\left\{-\frac{M+N}{m!}||z_k-y_k||^m||y-z_k|| + \frac{N-M}{m!}||z_k-y_k||^{m+1} + \frac{\lambda_{k+1}}{\alpha_k^{m+1}}\frac{Mc_{m+1}}{(m+1)!}||y-z_k||^{m+1}\right\}$$
$$\geq \bar\epsilon' + D_Q\epsilon' + \sum_{d=1}^{m}\frac{\epsilon^{(d)}D_Q^d}{d!} - \alpha_k\epsilon.$$

This inequality can be guaranteed with the same choice of constants α_k, L, and M as for Algorithm 4.2 provided that $\bar\epsilon' + D_Q\epsilon' + \sum_{d=1}^{m}\epsilon^{(d)}D_Q^d/d! - \alpha_k\epsilon$ is not positive. ■

5.3 Inexact resolutions and approximate subgradients

It remains to resolve the question of the combined effect of solving problems approximately and using approximate subgradient. The reasoning leading to the following result blends arguments from the various propositions and lemmas of this section in a pretty straightforward way. As in the previous section, we must treat minimization algorithms for convex function with a Lipschitz continuous gradient and m-th regularization separately.

Theorem 5.1 *Let f be a differentiable convex function with an L-Lipschitz gradient with respect to a norm $||\cdot||$. Assume that inequality (16) holds for $p := 2$, an $L_0 > 0$, and $\sigma_0 := L$, so that inequality (20) in Proposition 5.3, namely:*

$$\phi_0(y) \geq \phi_0(x) + \langle \phi_0'(x), y - x\rangle + \chi_k(y,x) \quad \textit{for all } x, y \in Q$$

is satisfied for a sequence $(\chi_k)_{k\geq 0}$ of nonnegative functions with $\chi_k(x,y) = 0 \Rightarrow x = y$ and $\chi_k(x,y) \geq L||y-x||^2/2$ for all k.

Let $\epsilon \geq 0$, $\gamma \in [0,1]$, D_Q be the diameter of Q, and fix $k \geq 0$. Assume that:

1. *$\hat x_k \in Q$, and $\min_{x\in Q}\phi_k(x) \geq f(\hat x_k) - \epsilon$;*
2. *$\hat\epsilon_k \geq \phi_k(\hat v_k) - \phi_k(v_k)$, where*

$$0 \leq \hat\epsilon_k \leq \min\left\{1, \left(\frac{\alpha_k\gamma}{1-\alpha_k}\right)^2\left(\frac{\epsilon}{1 + D_Q L_0\sqrt{2\lambda_k/L}}\right)^2\right\};$$

3. $g_k \in \partial_{\bar{\epsilon}'_k} f(y_k)$;

4. $h_k(\hat{x}_{k+1}) \leq \min_{x \in Q} h_k(x) + \alpha_k \epsilon_k^{(x)}$, *where* $h_k(x) := \alpha_k \langle g_k, x - \hat{v}_k \rangle + \lambda_{k+1} \chi_k(x, \hat{v}_k)$;

5. $\epsilon(1-\gamma) \geq \epsilon^{(x)} + \bar{\epsilon}'_k \max\{1 + (2-\alpha_k)D_Q, 1/\alpha_k + D_Q\}$.

Then:

$$\min_{x \in Q} \phi_{k+1}(x) \geq f(\hat{x}_{k+1}) - \epsilon.$$

Proof
Observe first that:

$$\begin{aligned} f(y_k) + \min_{x \in Q} h_k(x) &\geq f(y_k) + h_k(\hat{x}_{k+1}) - \alpha_k \epsilon^{(x)} \\ &\geq f(y_k) + \langle f'(y_k), \hat{x}_{k+1} - y_k \rangle + \frac{L}{2} ||\hat{x}_{k+1} - y_k||^2 - \alpha_k D_Q \bar{\epsilon}'_k - \alpha_k \epsilon^{(x)} \\ &\geq f(\hat{x}_{k+1}) - \alpha_k D_Q \bar{\epsilon}'_k - \alpha_k \epsilon^{(x)}. \end{aligned} \tag{29}$$

It suffices to write:

$$\begin{aligned} \min_{x \in Q} \phi_{k+1}(x) &= \min_{x \in Q} (1-\alpha_k)\phi_k(x) + \alpha_k (f(y_k) + \langle g_k, x - y_k \rangle - \bar{\epsilon}'_k) \\ &\geq (1-\alpha_k)\phi_k(\hat{v}_k) - \alpha_k \gamma \epsilon + \min_{x \in Q} \{\lambda_{k+1} \chi_k(x, \hat{v}_k) + \alpha_k (f(y_k) + \langle g_k, x - y_k \rangle - \bar{\epsilon}'_k)\} \\ &\geq (1-\alpha_k) f(\hat{x}_k) - (1 - \alpha_k(1-\gamma))\epsilon + \min_{x \in Q} \{\lambda_{k+1} \chi_k(x, \hat{v}_k) + \alpha_k (f(y_k) + \langle g_k, x - y_k \rangle - \bar{\epsilon}'_k)\}. \end{aligned}$$

The first inequality comes from a lower bound of $\langle \phi_k(\hat{v}_k), x - \hat{v}_k \rangle$ established in the proof of Proposition 5.3. We now have two possibilities for continuing our reasoning. Either we use:

$$f(\hat{x}_k) \geq f(y_k) + \langle f'(y_k), \hat{x}_k - y_k \rangle,$$

and, with some rearrangements:

$$\begin{aligned} \min_{x \in Q} \phi_{k+1}(x) &\geq f(y_k) + \min_{x \in Q} h_k(x) + (1-\alpha_k)\alpha_k \langle f'(y_k) - g_k, \hat{x}_k - \hat{v}_k \rangle - (1 - \alpha_k(1-\gamma))\epsilon - \alpha_k \bar{\epsilon}'_k \\ &\geq f(\hat{x}_{k+1}) - \alpha_k D_Q \bar{\epsilon}'_k - \alpha_k \epsilon^{(x)} - (1-\alpha_k)\alpha_k \bar{\epsilon}'_k D_Q - (1 - \alpha_k(1-\gamma))\epsilon - \alpha_k \bar{\epsilon}'_k, \end{aligned}$$

by (29). Or we write:

$$f(\hat{x}_k) \geq f(y_k) + \langle g_k, \hat{x}_k - y_k \rangle - \bar{\epsilon}'_k,$$

and we get, given that $y_k = (1-\alpha_k)x_k + \alpha_k v_k$:

$$\begin{aligned} \min_{x \in Q} \phi_{k+1}(x) &\geq f(y_k) + \min_{x \in Q} h_k(x) - \bar{\epsilon}'_k - (1 - \alpha_k(1-\gamma))\epsilon \\ &\geq f(\hat{x}_{k+1}) - \alpha_k D_Q \bar{\epsilon}'_k - \alpha_k \epsilon^{(x)} - \bar{\epsilon}'_k - (1 - \alpha_k(1-\gamma))\epsilon. \end{aligned}$$

The condition on ϵ suffices to prove the desired inequality, be it with the first or the second variant. ■

Similar derivations allow us to get the following result. Its proof follows the line of the last theorem's demonstration, and we omit it.

Theorem 5.2 *Let f be an m times differentiable convex function with an M-Lipschitz continuous m-th differential with respect to a norm $||\cdot||$. Assume that inequality (16) holds for $p := m+1$, a $L_0 > 0$, and a $\sigma_0 \geq 0$, so that inequality (20) in Proposition 5.3, namely:*

$$\phi_0(y) \geq \phi_0(x) + \langle \phi_0'(x), y - x\rangle + \chi_k(y,x) \quad \textit{for all } x, y \in Q$$

is satisfied with $\chi_k(x,y) \geq \sigma_0 ||y-x||^{m+1}/(m+1)$ for all k.

Let $\epsilon \geq 0$, $\gamma \in [0,1]$, D_Q be the diameter of Q, and fix $k \geq 0$. Assume that:

1. *$\hat{x}_k \in Q$, and $\min_{x\in Q} \phi_k(x) \geq f(\hat{x}_k) - \epsilon$;*
2. *$\hat{\epsilon}_k \geq \phi_k(\hat{v}_k) - \phi_k(v_k)$, where*

$$0 \leq \hat{\epsilon}_k \leq \min\left\{1, \left(\frac{\alpha_k \gamma}{1-\alpha_k}\right)^p \left(\frac{\epsilon}{1 + D_Q L_0 \sqrt[p]{p\lambda_k^{p-1}/\sigma_0}}\right)^p\right\},$$

   and $p := m+1$;

3. *$g_k \in \partial_{\bar{\epsilon}_k'} f(y_k)$, and A_d is an $\epsilon_k^{(d)}$-approximation of $f^{(d)}(\hat{z}_k)$ for $1 \leq d \leq m$; here $\hat{z}_k := (1-\alpha_k)\hat{x}_k + \alpha_k \hat{v}_k$;*
4. *$||h_k(\hat{x})||_* \leq \alpha_k \epsilon^{(x)}$, where*

$$h_k(x) := \langle f'(z_k), x - z_k\rangle + \cdots + \frac{1}{m!}\langle f^{(m)}(z_k)[x-z_k, \cdots, x-z_k], x-z_k\rangle + \frac{(2m+1)M}{(m+1)!}||x-z_k||^{m+1};$$

5. *$\alpha_k \epsilon(1-\gamma) \geq \alpha_k \epsilon^{(x)} + \bar{\epsilon}_k'(D_Q + 1) + \sum_{d=1}^m \frac{\epsilon^{(d)} D_Q^d}{d!}$.*

Then:

$$\min_{x\in Q} \phi_{k+1}(x) \geq f(\hat{x}_{k+1}) - \epsilon.$$

■

6 A variant of the estimation scheme

We explore in this section the possibility of ignoring some of the knowledge previously accumulated to construct the estimate sequence, in the hope that this cleaning of old data could accelerate the method. Actually, it is possible to devise such a strategy and to establish its efficiency theoretically, but we need to put ourselves in a slightly different framework than estimate sequences. The results of this subsection are strongly inspired by the approach of Lan, Lu, and Monteiro [LLM].

Actually, in the convergence analysis of the various methods proposed in this paper, the only link with the estimate sequence construction comes precisely from Lemma 2.3: the central question of Sections 3 and 4 was to find a point x_{k+1} compatible with the requirements of the fundamental Proposition 2.1. Let us step a little bit backwards and contemplate the following question.

Suppose that for every $k \geq 1$, we can guarantee that the sequence $(x_j)_{j\geq 0}$ satisfies:

$$f(y_k) + \langle g(y_k), (1-\alpha_k)x_k + \alpha_k v_k - y_k\rangle + \min_{x\in Q}\{\alpha_k\langle g(y_k), x - v_k\rangle + \lambda_{k+1}\chi_k(x, v_k)\} \geq f(x_{k+1}) \quad (30)$$

for well-chosen sequences $(v_j)_{j\geq 0}$, $(y_j)_{j\geq 0}$ of the feasible set Q and coefficients $(\alpha_j)_{j\geq 0}$ in $[0, 1]$. Also, $g(y_k) \in \partial f(y_k)$. Moreover, we assume that $\lambda_0 := 1$ and $\lambda_{k+1} := (1-\alpha_k)\lambda_k$. Finally, $\chi_k(x, y)$ is a nonnegative differentiable function which vanishes only when $x = y$.

Assume that

$$y_k := (1-\alpha_k)x_k + \alpha_k v_k.$$

Then, if the above inequality is true, we can successively write:

$$\begin{aligned}
f(x_{k+1}) &\leq f(y_k) + \min_{x\in Q}\{\alpha_k\langle g(y_k), x - v_k\rangle + \lambda_{k+1}\chi_k(x, v_k)\} \\
&= f(y_k) + \min_{x\in Q}\{\langle g(y_k), \alpha_k x + (1-\alpha_k)x_k - y_k\rangle + \lambda_{k+1}\chi_k(x, v_k)\} \\
&= (1-\alpha_k)\left(f(y_k) + \langle g(y_k), x_k - y_k\rangle\right) + \min_{x\in Q}\{\alpha_k\left(f(y_k) + \langle g(y_k), x - y_k\rangle\right) + \lambda_{k+1}\chi_k(x, v_k)\} \\
&\leq (1-\alpha_k)f(x_k) + \min_{x\in Q}\{\alpha_k f_k(x) + \lambda_{k+1}\chi_k(x, v_k)\},
\end{aligned}$$

where $f_k(x) := f(y_k) + \langle g(y_k), x - y_k\rangle$. Let

$$u_k := \arg\min_{x\in Q}\{\alpha_k f_k(x) + \lambda_{k+1}\chi_k(x, v_k)\},$$

so that:

$$\langle \alpha_k f_k'(u_k) + \lambda_{k+1}\chi_k'(u_k, v_k), x - u_k\rangle \geq 0 \quad \forall x \in Q,$$

where, for notational convenience, $\chi_k'(x, y) := \partial\chi_k(x, y)/\partial x$. Since f_k is affine, this condition is equivalent to:

$$\alpha_k f_k(x) + \lambda_{k+1}\langle \chi'(u_k, v_k), x - u_k\rangle \geq \alpha_k f_k(u_k) \quad \forall x \in Q.$$

Therefore, taking $x := x^*$ above, we can continue as follows:

$$\begin{aligned}
f(x_{k+1}) &\leq (1-\alpha_k)f(x_k) + \alpha_k f_k(u_k) + \lambda_{k+1}\chi_k(u_k, v_k) \\
&\leq (1-\alpha_k)f(x_k) + \alpha_k f_k(x^*) + \lambda_{k+1}\left(\chi_k(u_k, v_k) + \langle \chi_k'(u_k, v_k), x^* - u_k\rangle\right),
\end{aligned}$$

and:

$$\frac{f(x_{k+1}) - f(x^*)}{\lambda_{k+1}} \leq \frac{f(x_k) - f(x^*)}{\lambda_k} + \chi_k(u_k, v_k) + \langle \chi_k'(u_k, v_k), x^* - u_k\rangle.$$

It suffices to construct the sequence $(v_k)_{k\geq 0}$ so that

$$\sum_{k\geq 0}^{N} \chi_k(u_k, v_k) + \langle \chi_k'(u_k, v_k), x^* - u_k\rangle$$

remains smaller than a constant C for each N to prove that the sequence $(f(x_k))_{k\geq 0}$ converges to f^* as fast as $(\lambda_k)_{k\geq 0}$. Since the choice of $(\lambda_k)_{k\geq 0}$ is solely determined by the inequality (30), we can

retrieve the fastness of estimate sequence methods for this new class of schemes. We can interpret these schemes as restated estimate sequence schemes, setting:

$$\phi_k(x) := f(y_k) + \alpha_k \langle g(y_k), x - v_k \rangle + \lambda_{k+1} \chi_k(x, v_k).$$

The following proposition gives an example where the obtained scheme converges indeed asymptotically as fast as the corresponding estimate sequence scheme.

Proposition 6.1 *Suppose that $v_k := u_{k-1}$ and that $\chi_k(a,b) \equiv \chi(a,b) = d(a) - d(b) - \langle d'(b), a - b \rangle$ is a Bregman distance generated by some strongly convex function d. Then C can be chosen as $\chi(x^*, v_0)$.*

Proof
It is a simple application of the well-know triangular equality for the Bregman distance:

$$\chi(u_k, v_k) + \langle \chi'(u_k, v_k), x^* - u_k \rangle = \chi(x^*, v_k) - \chi(x^*, u_k).$$

With $v_k := u_{k-1}$, the sum of interest equals $\chi(x^*, v_0) - \chi(x^*, u_N) \leq \chi(x^*, v_0)$. ■

We recover the algorithm proposed in [LLM].

7 Conclusions and outlook

We have shown in this paper that estimate sequence methods can be used as a unifying paradigm for analyzing a number of accelerated algorithms. We have demonstrated their sensibility with respect to approximations, suggesting some accelerating possibilities, by allowing oneself to cut off some intermediate computations.

We speculate that estimate sequences constitute an appropriate descriptive framework to develop and analyze several improving strategies, such as restarting procedures, aggressive choice of parameters, and adaptive parameters updates.

8 Appendix: useful technical results

Lemma 8.1 *Consider a sequence $\{\alpha_k : k \geq 0\}$ of $]0,1[$, and define $\lambda_0 := 1$, $\lambda_{k+1} = (1 - \alpha_k)\lambda_k$ for every $k \geq 0$. If there exists a constant $\beta > 0$ and an integer $p > 0$ for which $\alpha_k^p / \lambda_{k+1} \geq \beta$ for every $k \geq 0$, then, for all $N \geq 1$,*

$$\lambda_N \leq \left(\frac{p}{p + N \sqrt[p]{\beta}} \right)^p \leq \frac{1}{\beta} \left(\frac{p}{N} \right)^p$$

Proof
Obviously, the sequence $\{\lambda_k : k \geq 0\}$ is positive and decreasing. Let us fix $k \geq 0$. We have:

$$\begin{aligned} \lambda_k - \lambda_{k+1} &= \left(\sqrt[p]{\lambda_k^{p-1}} + \sqrt[p]{\lambda_k^{p-2} \lambda_{k+1}} + \cdots + \sqrt[p]{\lambda_k \lambda_{k+1}^{p-2}} + \sqrt[p]{\lambda_{k+1}^{p-1}} \right) \left(\sqrt[p]{\lambda_k} - \sqrt[p]{\lambda_{k+1}} \right) \\ &\leq p \sqrt[p]{\lambda_k^{p-1}} \left(\sqrt[p]{\lambda_k} - \sqrt[p]{\lambda_{k+1}} \right). \end{aligned}$$

This inequality implies:

$$\frac{1}{\sqrt[p]{\lambda_{k+1}}} - \frac{1}{\sqrt[p]{\lambda_k}} = \frac{\sqrt[p]{\lambda_k} - \sqrt[p]{\lambda_{k+1}}}{\sqrt[p]{\lambda_{k+1}\lambda_k}} \geq \frac{\lambda_k - \lambda_{k+1}}{p\lambda_k \sqrt[p]{\lambda_{k+1}}} = \frac{\alpha_k \lambda_k}{p\lambda_k \sqrt[p]{\lambda_{k+1}}} \geq \frac{\sqrt[p]{\beta}}{p}.$$

Summing up these inequalities for $k = 0$ and $k = N - 1$, we obtain:

$$\frac{1}{\sqrt[p]{\lambda_N}} - \frac{1}{\sqrt[p]{\lambda_0}} = \frac{1}{\sqrt[p]{\lambda_N}} - 1 \geq N\frac{\sqrt[p]{\beta}}{p},$$

which is equivalent to the desired result. ■

The next lemma establishes an inequality on powers of matrix induced norms and is useful for determining how cubic methods (and others) can be accelerated.

Lemma 8.2 *Consider a matrix induced norm $||\cdot|| := \langle G\,\cdot, \cdot\rangle^{1/2}$. Let $x_0 \in \mathbb{R}^n$, and, given a natural number $p \geq 2$, define $\phi(x) := ||x - x_0||^p$ for every $x \in \mathbb{R}^n$. We have for every $x, y \in \mathbb{R}^n$:*

$$\phi(y) - \phi(x) - \langle \phi'(x), y - x\rangle \geq c_p||y - x||^p,$$

where the constant c_p can be chosen as large as $(p-1)/(\sqrt[p-2]{2p-3}+1)^{p-2}$ when $p > 2$, or as 1 when $p = 2$.

Proof
Let $C_p(x, y) := [\phi(y) - \phi(x) - \langle \phi'(x), y - x\rangle]/[||y - x||^p]$ for $x \neq y$. We have successively, noting that $\min\{C_p(x, y) : x \neq y\} = \min\{C_p(x + x_0, y + x_0) : x \neq y\}$:

$$\begin{aligned}\min_{x\neq y} C_p(x, y) &= \min_{x\neq y} \frac{||y||^p + (p-1)||x||^p - p||x||^{p-2}\langle Gx, y\rangle}{(||x||^2 - 2\langle Gx, y\rangle + ||y||^2)^{p/2}} \\ &= \min_{\alpha\in[-1,1], t>0} \frac{t^p + p - 1 - pt\alpha}{(t^2 - 2t\alpha + 1)^{p/2}},\end{aligned}$$

where t and α represent respectively $||y||/||x||$ and $\langle Gx, y\rangle/(||x|| \cdot ||y||)$ (note that $C_p(0, y) = C_p(x, 0) = 1$ for all x, y, while $C_p(x, -x) = 2p/2^p \leq 1$). Let us denote by $\psi(\alpha, t)$ the objective function in the above minimization problem. For every fixed $t > 0$, the function $\alpha \mapsto \psi(\alpha, t)$ is increasing. Indeed, its derivative is nonnegative iff $t^p - t^2 + (p-2)(1 - t\alpha) \geq 0$ for every feasible α. As $p \geq 2$, this condition is implied by the inequality

$$t^p - t^2 + (p-2)(1-t) = (1-t)^2(t^{p-2} + 2t^{p-3} + \cdots + (p-2)t + p - 2) \geq 0,$$

which is obviously true when $t > 0$. Now, we must minimize the function

$$f(t) := \psi(-1, t) = \frac{t^p + pt + p - 1}{(t+1)^p}$$

on $t > 0$. If $p = 2$, this function reduces to 1. Assume that $p > 2$. It can be easily checked that f has only one minimizer τ, for which $g(\tau) := \tau^{p-1} - (p-1)\tau - (p-2) = 0$. Indeed, the above polynomial equation, obtained from the derivative of f, has only one positive root: $g(0) < 0$,

$g'(t) = (p-1)(t^{p-2}-1) \leq 0$ for $t \in [0,1]$, $g'(t) > 0$ for $t > 1$, and $g(t) \to +\infty$ as $t \to +\infty$. Let us check that $\tau \leq \sqrt[p-2]{2p-3}$. We have indeed $g(0) < 0$ and:

$$g(\sqrt[p-2]{2p-3}) = \sqrt[p-2]{2p-3}\,(2p-3-(p-1)) - (p-2) \geq (2p-3-(p-1)) - (p-2) = 0.$$

Therefore, using $\tau^{p-1} = (p-1)\tau + (p-2)$, we get:

$$\begin{aligned}\min_{t>0} f(t) &= \frac{\tau^p + p\tau + p - 1}{(\tau+1)^p} = \frac{(p-1)\tau^2 + (p-2)\tau + p\tau + p - 1}{(\tau+1)^p} \\ &= \frac{p-1}{(\tau+1)^{p-2}} \geq \frac{p-1}{(\sqrt[p-2]{2p-3}+1)^{p-2}}.\end{aligned}$$

■

The above lemma improves slightly a previous result of Nesterov (see Lemma 4 in [Nes06]), where he proved that c_p could be as large as $1/2^{p-2}$. Our result is stronger, because, from concavity of the function $h(t) := \sqrt[p-2]{t}$ when $p \geq 2$, we have

$$h(p-1) \geq \frac{h(2p-3)+h(1)}{2} \;\Leftrightarrow\; \sqrt[p-2]{p-1} \geq \frac{\sqrt[p-2]{2p-3}+1}{2} \;\Leftrightarrow\; \frac{p-1}{(\sqrt[p-2]{2p-3}+1)^{p-2}} \geq \frac{1}{2^{p-2}}.$$

It should be noted that an extension of this lemma to more general norms would entail the applicability of m-th regularization to these norms. However, one can prove that the function $C(x,y)$ is not bounded away from 0 for all those norm whose unit ball contains a non-trivial exposed face, i.e. an exposed face that is not reduced to a singleton, as for instance the 1-norm and the ∞-norm.

Acknowledgments

This research was partially supported by: ⋄ Research Council KUL: GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering(OPTEC), IOF-SCORES4CHEM; ⋄ Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011).

References

[AT06] A. Auslender and M. Teboulle, *Interior gradient and proximal methods for convex and conic optimization*, SIAM Journal on Optimization **16** (2006), no. 3, 697–725.

[BTN05] A. Ben-Tal and A. Nemirovski, *Non-Euclidean restricted memory level method for large-scale convex optimization*, Mathematical Programming **102** (2005), no. 3, 407–456.

[CGT07] C. Cartis, N. Gould, and P. Toint, *Adaptive cubic overestimation methods for unconstrained optimization*, Technical report ral-tr-2007-016, Rutherford Appleton Laboratory, Chilton, England, 2007.

[d'A08] A. d'Aspremont, *Smooth optimization with approximate gradient*, SIAM Journal on Optimization **19** (2008), no. 3, 1171–1183.

[HUL93a] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms I*, Grundlehren der mathematischen Wissenschaften, vol. 305, Springer, 1993.

[HUL93b] _____, *Convex Analysis and Minimization Algorithms II*, Grundlehren der mathematischen Wissenschaften, vol. 306, Springer, 1993.

[LLM] G. Lan, Z. Lu, and R. Monteiro, *Primal-dual First-order Methods with $\mathcal{O}(1/\epsilon)$ Iteration-complexity for Cone Programming*, To appear in Mathematical Programming.

[Mor05] B. Mordukhovich, *Variational Analysis And Generalized Differentiation I, Theory And Examples*, Grundlehren der mathematischen Wissenschaften, no. 330, Springer, 2005.

[Nes83] Y. Nesterov, *A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$*, Doklady AN SSSR (translated as Soviet Math. Dokladi) **3** (1983), 543–547.

[Nes03] _____, *Introductory lectures on convex optimization: a basic course*, Applied Optimization, vol. 87, Kluwer Academic Publishers, 2003.

[Nes05] _____, *Smooth minimization of non-smooth functions*, Mathematical Programming **103** (2005), no. 1, 127–152.

[Nes06] _____, *Accelerating the cubic regularization of Newtons method on convex problems*, CORE Discussion Paper 39, Center for Operation Research and Econometrics, Université catholique de Louvain, 2006.

[Nes07] _____, *Gradient methods for minimizing composite objective function*, CORE Discussion Paper 76, Center for Operation Research and Econometrics, Université catholique de Louvain, 2007.

[NP06] Y. Nesterov and B. Polyak, *Cubic regularization of Newton's method and its global performance*, Mathematical Programming **108** (2006), 177–205.

[Roc70] R. T. Rockafellar, *Convex Analysis*, Princeton Mathematics Series, vol. 28, Princeton University Press, 1970.

[SN05] A. Shapiro and A. Nemirovski, *On complexity of stochastic programming problems*, Continuous Optimization: Current Trends and Applications (V. Jeyakumar and A. Rubinov, eds.), Springer, 2005, pp. 111–144.