Estimate sequence methods

Michel Baes*

September 24, 2008

Abstract

The approach of estimate sequence offers an interesting rereading of a number of accelerating schemes proposed by Nesterov [Nes03], [Nes05b], and [Nes06]. It seems to us that this framework is the most appropriate descriptive framework to develop and analyze several improving strategies, such as restarting procedures, aggressive choice of parameters, and adaptive parameters updates.

We develop in this work a simple, self-consistent, and unified framework for the study of estimate sequences, with which we can recover each 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.

1 Introduction

The concept of *estimate sequences*, introduced by Nesterov in 1983 [Nes83] to define the provably fastest gradient-type schemes for convex optimization. Since then, this concept, in spite of its conceptual simplicity, has not attracted a lot of attention. Except in some papers of Nesterov [Nes05b, Nes06, Nes05a], and in a paper of Auslender and Teboulle [AT06], this idea has practically not been exploited yet in optimization.

The purpose of this paper is to provide a simple framework for the study of estimate sequences schemes, and to demonstrate its power by generalizing 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. Indeed, these schemes require not more than $\mathcal{O}((1/\epsilon)^{m+1})$ iterations to find an ϵ-approximation of the solution. In turn, our results allows us to reinterpret the accelerated cubic regularization schemes of Nesterov, and to improve some hidden constants in his complexity analysis.

The essence of our approach is crystallized in Lemma 2.3, which provides essentially the sole condition to check for proving the convergence of the scheme and determine its speed. This lemma suggests at least two different strategies for designing an accelerated scheme, which we illustrate successively in Section 3, where we offer a new reading of accelerated gradient schemes, and in Sections 4, 5, and 6, where we reinterpret the accelerated cubic scheme of Nesterov with simplified proofs, and we extend it to m-th regularization scheme.

*M. Baes is with the Department of Electrical Engineering (ESAT), Research Group SCD-SISTA, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium. E-mail: Michel.Baes@esat.kuleuven.be.

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/or differentiability.

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:

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

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 X^* before any computation started, 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 $f(x_k) - f^* \leq \lambda_k(\phi_0(x^*) - f^*)$.*

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.

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 $\left((\phi_k)_{k\geq 0}; (\lambda_k)_{k\geq 0}\right)$ 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,

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

which proves that we have built an estimate sequence. The second equality in (1) comes from an easy 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.
- ◇ a way of constructing, preferably very cheaply, points x_k that satisfy the inequality prescribed at Proposition 2.1.

In view of Lemma 7.1, 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 $\mathcal{O}(1/(k^n\beta))$.

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 $(x_k)_{k\geq 0}$.

Lemma 2.3 *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*:*

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

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:*

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

where v_k *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. The above inequality 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 the left-hand side on Q. ∎

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

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 cases, constructing a point x_{k+1} can be done very cheaply, by an appropriate gradient-like step.

Remark 2.1 *Let us assume that we do not have access to an exact subgradient of the objective function, but rather to an ϵ-subgradient for a given $\epsilon > 0$. In other words, we have an oracle that outputs, for an input point $y \in Q$, a vector $g \in \mathbb{R}^n$ such that for every $x \in Q$, the inequality:*

$$f(x) \geq f(y) + \langle g, y - x \rangle - \epsilon$$

holds. We denote the set of such vectors g as $\partial_\epsilon f(y)$.

Given $g(y_k) \in \partial_\epsilon f(y_k)$, we choose $f_k(x) := f(y_k) + \langle g(y_k), y_k - x\rangle - \epsilon$ as underestimate. The functions of the estimate sequence become:

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

We can reformulate the inequality (3) as follows.

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

The influence of the subgradient inaccuracy vanishes as fast as α_k goes to zero. Suppose that the sequence x_{k+1} satisfies:

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

that is, that this sequence is constructed exactly as for a scheme where the subgradient is available with an absolute precision. Then, repeating the same reasoning as for Proposition 2.1, we get from $\phi_k(v_k) \geq f(x_k) - \alpha_{k-1}\epsilon$ that:

$$f(x_k) - f^* \leq \lambda_k(\phi_0(x^*) - f(x^*)) + \epsilon\alpha_{k-1} + \epsilon,$$

for every $k > 0$. Therefore, we must let α_k go to zero as well. ■

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 $x, y \in Q$. 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 $x_0 \in \operatorname{int} Q$, 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:

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

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. 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 [Nes05b]).

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 σ-strongly convex for the norm $\|\cdot\|$. Also $\min_{x \in Q} f(x) = f(x_0) \geq f(x^*)$. 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 Norm lower bound on ϕ_k

Let us consider here the case where $\chi_k(x,y) := L\|x-y\|^2/2$, where $\|\cdot\|$ is a norm on $\mathbb{R}^n$. 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

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

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

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

A important quantity in the subsequent analysis is the *gradient mapping*:

$$g_Q(y;h) \in -\frac{1}{2h}\partial_x\|x-y\|^2\Big|_{x=x_Q(y;h)}.$$

For the unconstrained gradient method with the Euclidean norm, this dual vector $g_{\mathbb{R}^n}(y;h)$ is nothing but the usual gradient of the objective function at y. The following lemma shows the elementary properties of the gradient mapping. We give its proof for the sake of completeness.

Lemma 3.1 *For every $y \in \mathbb{R}^n$, and $h > 0$, we have*

$$\|y - x_Q(y;h)\| = h\|g_Q(y;h)\|_*, \quad \text{and} \quad \langle g_Q(y;h), y - x_Q(y;h)\rangle = \|g_Q(y;h)\|_*\|y - x_Q(y;h)\|$$

Proof

Let $\phi(x) := \|x-y\|^2/2$. The conjugate of ϕ is $\phi^*(s) = \|s\|_*^2/2 + \langle s, y\rangle$, and, as

$$-hg_Q(y;h) \in \partial_y\phi(x_Q(y;h)),$$

we have $\phi^*(-hg_Q(y;h)) + \phi(x_Q(y;h)) = -\langle hg_Q(y;h), x_Q(y;h)\rangle$. Thus

$$h^2\|g_Q(y;h)\|_*^2/2 + \|x_Q(y;h)-y\|^2/2 = -\langle hg_Q(y;h), x_Q(y;h)-y\rangle \leq h\|g_Q(y;h)\|_*\|x_Q(y;h)-y\|.$$

Therefore $\|y - x_Q(y;h)\| = h\|g_Q(y;h)\|_*$, and the above inequality is tight. ∎

The following lemma gives a lower estimate for f that uses $x_Q(y;h)$ and $g_Q(y;h)$.

Lemma 3.2 *Let f be a convex function with a Lipschitz continuous gradient with constant $L > 0$ for the norm $\|\cdot\|$. For every $y \in \mathbb{R}^n$, $x \in Q$, and every $0 < h \leq 1/L$, we have:*

$$f(x) \geq f(x_Q(y;h)) + \langle g_Q(y;h), x - x_Q(y;h)\rangle - \frac{h}{2}\|g_Q(y;h)\|_*^2.$$

Proof
Let $\phi(x) := f(y) + \langle f'(y), x - y\rangle + ||x-y||^2/2h$, so that $\phi(x) \geq f(y)$ for every $x \in Q$ in view of the Lipschitz continuity of f's gradient. By definition, the minimizer of $\phi(x)$ over Q is $x_Q(y;h)$, thus

$$\langle f'(y) - g_Q(y;h), x - x_Q(y;h)\rangle \geq 0 \quad \text{for every } x \in Q. \tag{6}$$

Therefore, we have for every $x \in Q$:

$$\begin{aligned} f(x) &\geq f(y) + \langle f'(y), x - y\rangle = f(y) + \langle f'(y), x - x_Q(y;h)\rangle + \langle f'(y), x_Q(y;h) - y\rangle \\ &\geq f(y) + \langle g_Q(y;h), x - x_Q(y;h)\rangle + \langle f'(y), x_Q(y;h) - y\rangle \\ &\geq f(x_Q(y;h)) - \frac{L}{2}||y - x_Q(y;h)||^2 + \langle g_Q(y;h), x - x_Q(y;h)\rangle \\ &= f(x_Q(y;h)) - \frac{Lh^2}{2}||g_Q(y;h)||_*^2 + \langle g_Q(y;h), x - x_Q(y;h)\rangle, \end{aligned}$$

in view of the previous lemma; it remains to use the hypothesis on the step-size $0 < hL \leq 1$. ■

We are ready now to deal properly with the right-hand side of (5). Observe that:

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

The minimizer x^* of the above problem satisfies:

$$x_k^* = \frac{\alpha_k}{\lambda_{k+1}}(x_Q(y_k; 1/L) - y_k) + v_k, \tag{7}$$

so that:

$$m_k^* = \frac{\alpha_k^2}{\lambda_{k+1}} \left\{ \langle f'(y_k), x_Q(y_k; 1/L) - y_k\rangle + \frac{L}{2}||x_Q(y_k; 1/L) - y_k||^2 \right\}.$$

Note that m_k^* is not positive. Indeed, using the inequality (6) and Lemma 3.2, we can write:

$$\begin{aligned} \frac{\lambda_{k+1}}{\alpha_k^2} m_k^* &= \langle f'(y_k), x_Q(y_k; 1/L) - y_k\rangle + \frac{L}{2}||x_Q(y_k; 1/L) - y_k||^2 \\ &\leq \langle g_Q(y_k; 1/L), x_Q(y_k; 1/L) - y_k\rangle + \frac{L}{2}||x_Q(y_k; 1/L) - y_k||^2 \\ &= -||g_Q(y_k; 1/L)||_* ||x_Q(y_k; 1/L) - y_k|| + \frac{L}{2}||x_Q(y_k; 1/L) - y_k||^2 \\ &= \left(-\frac{1}{h} + \frac{L}{2}\right)||x_Q(y_k; 1/L) - y_k||^2 \leq -\frac{L}{2}||x_Q(y_k; 1/L) - y_k||^2 \leq 0 \end{aligned}$$

Therefore, if $\alpha_k^2/\lambda_{k+1} \leq 1$, 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
Assumption: *f is strongly convex with constant σ for the norm $||\cdot||$, and has a Lipschitz continuous gradient with constant L for the same norm; the set Q is closed, convex, and has a nonempty interior*
Choose $x_0 \in \mathbb{R}^n$, 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 7.1 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 and inexpensive shortcut for the first two steps of the recursive loop of the above algorithm. The requirement $\lambda_{k+1} \geq \alpha_k^2$ is satisfied for every $k/geq0$ when $\lambda_k = 4/(k+2)^2$ for $k \geq 2$. With this choice, we obtain $\alpha_k = (2k + 5)/(k + 3)^2$ and $\alpha_k^2/\lambda_{k+1} \leq (7/8)^2$.

3.2 Bregman distance lower bound on ϕ_k

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. 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 more, we have for every $x, y \in Q$ that :

$$\chi_k(x, y) \geq \frac{\gamma_k \sigma}{2}||x - y||^2. \tag{8}$$

If the coefficients γ_k are bounded from above by L/σ, we can apply Lemma 2.3. Therefore, with

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

we have:

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

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

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

This point plays the role of x_k^* in the previous section. Equation (7) hints a possible candidate for the next approximation x_{k+1} for the minimizer x^* of f on Q. Let us check that:

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

yields a sufficient decrease of the objective, that is, that

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

Observe that:

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

Now, if $\alpha_k^2 = \lambda_{k+1} = (1 - \alpha_k)\lambda_k$ as in the previous algorithm, and $\gamma_k := L/\sigma$, 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
Assumption: *f is strongly convex with constant σ for the norm $\|\cdot\|$, and has a Lipschitz continuous gradient with constant L for the same norm; the set Q is closed, convex, and has a nonempty interior*
Choose $x_0 \in \mathbb{R}^n$, 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 for constrained problems

Cubic regularization has been developed by Nesterov and Polyak in [NP06], and further extended by Nesterov in [Nes06, Nes05a] and by Cartis, Gould, and Toint [CGT07].

We consider here convex objective functions f that are twice continuously differentiable and that have a Lipschitz continuous Hessian with respect to an *operator* norm $||z||$, 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, $||A|| = \max_{||x||=1} ||Ax||_*$. An easy consequence of the Lipschitz continuity of the Hessian (see Lemma 1 in [NP06] for a proof) reads:

$$||f'(y) - f'(x) - f''(x)(y - x)||_* \leq \frac{M}{2}||x - y||^2 \quad \forall x, y \in \operatorname{dom} f. \tag{9}$$

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 construct 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 7.4, 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:

$$\begin{aligned} f(y_k) \; + \; & \langle f'(y_k), (1 - \alpha_k)x_k + \alpha_k v_k - y_k\rangle \\ & + \min_{x \in Q} \left\{ \alpha_k \langle f'(y_k), x - v_k\rangle + \lambda_{k+1}\frac{M}{12}||x - v_k||^3 \right\} \geq f(x_{k+1}). \end{aligned} \tag{10}$$

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 two last terms of the above 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:

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

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 problems of the above type can be solved efficiently because their 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:

$$\langle f'(x)+f''(x)(x_N(x)-x), y-x_N(x)\rangle+\frac{N}{4}||x-x_N(x)||\langle G(x_N(x)-x), y-x_N(x)\rangle \geq 0 \quad \forall y \in Q. \quad (11)$$

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

$$\min_{y\in Q} \alpha\langle g, y-v\rangle + \lambda\chi_k(y,v) \geq \min_{y\in Q}\langle g, y-z\rangle + \frac{\lambda_{k+1}}{\alpha^3}\chi_k(y,z).$$

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:

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

■

The next lemma plays the crucial role in the validation of the desired inequality. 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 (11), we have:

$$0 \leq \langle f'(x) + f''(x)(x_N(x)-x), y - x_N(x)\rangle + \frac{Nr_N(x)}{4}\langle G(x_N(x)-x), y-x_N(x)\rangle. \quad (12)$$

Observe that:

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

Moreover, in view of the Hessian Lipschitz continuity (9), we have:

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

Summing up these two inequalities with appropriate multiplicative coefficients, we get form (12):

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

■

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 norm $\|\cdot\|$.*
Choose $x_0 \in Q$, set $v_0 := x_0$ and $\lambda_0 := 1$.
Set $\phi_0(x) := f(x_0) + M\|x - x_0\|^3/6$.
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 (10) 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 Lemmas 4.1 and Lemma 4.2, the second term of the above inequality can be bounded from below as follow:

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

Therefore, the inequality we need to prove becomes:

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

Relaxing the constraint of the above minimization problem, we get:

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

Thus, the inequality is satisfied as soon as:

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

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 with respect to N. Its maximizer turns out to be attained for $N := 5M$, in which case its value is $1/12$. Note that these constant can be seen in the Algorithm 4.1.

It suffices now to apply Lemma 7.1 to obtain the complexity of the algorithm. ■

5 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

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

which can be highly nontrivial. Nevertheless, let us analyze this scheme, as an illustration of the power of the estimate sequence framework.

Given a norm $\|\cdot\|$, we define the norm of an m-tensor A as:

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

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:

$$\left\| f^{(m-j)}(y) - f^{(m-j)}(x) - f^{(m-j+1)}(x)[y-x] - \cdots - \frac{1}{j!} f^{(m)}(x)[y-x, \ldots, y-x] \right\|$$
$$\leq \frac{M}{(j+1)!} \|y-x\|^{j+1}. \tag{13}$$

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 scheme proposed by 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 an operator norm as in the previous section. Lemma 7.4 provides us with a constant c_{m+1} such that we can use 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:

$$\langle f'(y_k), (1-\alpha_k)x_k + \alpha_k v_k - y_k \rangle + \min_{y \in Q} \{ \alpha_k \langle f'(y_k), y - v_k \rangle + \lambda_{k+1} \chi_k(y, v_k) \} \geq 0. \tag{14}$$

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

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

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

$$\left\langle f'(x) + f''(x)[x_N(x) - x] + \cdots + \frac{1}{(m-1)!} f^{(m)}(x)[x_N(x) - x, \cdots, x_N(x) - x], y - x_N(x) \right\rangle$$
$$+ \frac{N r_N(x)^{m-1}}{m!} \left\langle G(x_N(x) - x), y - x_N(x) \right\rangle \geq 0 \quad \forall y \in Q. \tag{15}$$

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

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

Lemma 5.2 *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 (13) to get:

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

Using the latter inequality in (15), we get:

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

It remains to use

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

to get the desired inequality. ■

The m-regularization algorithm looks as follows.

Algorithm 5.1
Assumption: *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 $\phi_0(x) := f(x_0) + M\|x - x_0\|^{m+1}/m!$.
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 $y_k := \arg\min_{y \in Q} \langle f'(z_k), y - z_k\rangle + \cdots + \frac{1}{m!}\langle f^{(m)}(z_k)(y-z_k), y - z_k\rangle + \frac{(2m+1)M}{(m+1)!}\|y - z_k\|^{m+1}$;
  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 5.1 *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 (14) 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 5.1 and Lemma 5.2, 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 7.1. ∎

6 Cubic and m-th regularization for unconstrained problems

It is possible to improve slightly 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. In the viewpoint of our complexity analysis, the case $m = 2$ does not bear anything special.

There are essentially two things that change with respect to the constrained situation. Firstly, the inequality (14) that we need to check can be simplified because we can compute the exact value of the minimum. It can be rewritten as:

$$\begin{aligned}
\langle f'(y_k), (1-\alpha_k)x_k + \alpha_k v_k - y_k\rangle &\geq \min_{y\in\mathbb{R}^n}\{\alpha_k\langle f'(y_k), y - v_k\rangle + \lambda_{k+1}\chi_k(y, v_k)\} \\
&= \frac{m}{m+1}\left(\frac{m!}{c_{m+1}}\right)^{1/m}\left(\frac{\alpha_k^{m+1}}{\lambda_{k+1}}\frac{\|f'(y_k)\|_*^{m+1}}{M}\right)^{1/m}. \qquad (16)
\end{aligned}$$

Secondly, the form of the optimality condition for $x_N(x)$ changes. When $Q = \mathbb{R}^n$ in (15), we have for all $x \in \mathbb{R}^n$:

$$f'(x) + \cdots + \frac{1}{(m-1)!}f^{(m)}(x)[x_N(x) - x, \cdots, x_N(x) - x] + \frac{Nr_N(x)^{m-1}}{m!}G(x_N(x) - x) = 0. \qquad (17)$$

This relation get a kind of counterpart to Lemma 5.2.

Lemma 6.1 *For every $x \in \mathbb{R}^n$, we have*

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

Proof
We can transform the Lipschitz continuity of $f^{(m)}$ (see (13) with $j := m-1$) using the optimality condition (17):

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

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

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

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

Now, according to the algorithm, $y_k = x_N(z_k)$, and the above inequality resembles strikingly to the desired inequality (16). Identifying the coefficients, we get:

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

or:

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

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

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

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 Section 4.

The algorithm now reads as follows:

Algorithm 6.1
Assumption: *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 $y_k := \arg\min_{y\in Q} \langle f'(z_k), y - z_k\rangle + \cdots + \frac{1}{m!}\langle f^{(m)}(z_k)(y-z_k), y-z_k\rangle + \frac{\sqrt{m}M}{(m+1)!}\|y - z_k\|^{m+1}$;
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.

7 Appendix: useful technical results

Lemma 7.1 *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 $n > 0$ for which $\alpha_k^n/\lambda_{k+1} \geq \beta$ for every $k \geq 0$, then, for all $N \geq 1$,*

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

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

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

Using this inequality, we can write:

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

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

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

which is equivalent to the desired result. ■

The following theorem quantifies the objective decrease induced by a standard gradient method step:

$$x_+ = x - hf'(x).$$

Lemma 7.2 *Let f be a proper convex function that has a Lipschitz continuous gradient with constant L with respect to the Euclidean norm. Considering a non-stationary point x in the domain of f, we have, for every h such that $x - f'(x) \in \operatorname{dom} f$:*

$$f\left(x - hf'(x)\right) \leq f(x) - h||f'(x)||^2 + \frac{Lh^2}{2}||f'(x)||^2.$$

The maximal guaranteed decrease is attained for $h := 1/L$, provided that the corresponding point is in $\operatorname{dom} f$.

Proof

This is an immediate application of the definition of Lipschitz continuous gradient. ■

When the norm used is not Euclidean, the standard gradient step does something quite unnatural. Indeed, it adds an object x that belongs to the primal space $(\mathbb{R}^n, ||\cdot||)$ to an object $-hf'(x)$ that belongs to the dual space $(\mathbb{R}^n, ||\cdot||_*)$. In order to circumvent this problem, we can define:

$$x_+ := \arg\min_{y \in \mathbb{R}^n} \left\{ f(x) + \langle f'(x), y - x \rangle + \frac{1}{2h}||y - x||^2 \right\}.$$

Note that the usual gradient method corresponds to the case where the norm is Euclidean. A statement similar to Lemma 7.2 can be shown for this extension.

Lemma 7.3 *Let f be a proper convex function that has a Lipschitz continuous gradient with constant L with respect to the norm $||\cdot||$. Considering a non-stationary point x in the domain of f, we have, for every $0 < h \leq 1/L$:*

$$f(x_+) \leq f(x) - \frac{h}{2}||f'(x)||_*^2.$$

The maximal guaranteed decrease is attained for $h := 1/L$, provided that the corresponding point is in $\operatorname{dom} f$.

Proof

Since f has a Lipschitz continuous gradient, we have

$$f(y) \leq f(x) + \langle f'(x), y - x \rangle + \frac{1}{2h}||y - x||^2 \quad \text{for all } y \in \mathbb{R}^n$$

since $1/h \geq L$. Minimizing over $y \in \mathbb{R}^n$, we have:

$$f(x_+) \leq \min_{y \in \mathbb{R}^n} \left\{ f(x) + \langle f'(x), y - x \rangle + \frac{1}{2h}||y - x||^2 \right\} = f(x) - \frac{h}{2}||f'(x)||_*^2,$$

using $||a||_* = \sup_{||b|| \leq 1} \langle a, b \rangle$. ■

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

Lemma 7.4 *Consider an operator norm* $||\cdot|| := \langle G\cdot,\cdot\rangle^{1/2}$. *Let* $x_0 \in \mathbb{R}^n$, *and define* $\phi(x) = ||x - x_0||^p$ *for every natural* $p \geq 2$. *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* 1 *when* $p = 2$.

Proof

Let $C_p(x,y) := [\phi(y) - \phi(x) - \langle \phi'(x), y - x\rangle]/[||y - x||^p]$. We have successively, noting that $\min C_p(x,y) = \min C_p(x + x_0, y + y_0)$:

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

where $t = ||y||/||x||$, and $\alpha = \langle Gx, y\rangle/||x||||y||$. 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

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

on $t > 0$. If $p = 2$, this function reduces to 1. Assume that $p > 2$. It can be easily checked that ψ has only one minimizer τ, for which $g(\tau) := \tau^{p-1} - (p-1)\tau - (p-2) = 0$ (because the above polynomial equation, obtained from the derivative of f, has only one positive root in view of a sign analysis if g'). 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,

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

■

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 $t \mapsto \sqrt[p-2]{t}$ when $p \geq 2$, we have

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

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 every norm with a unit ball containing an exposed face that is not reduced to a singleton, as for instance the 1-norm and the ∞-norm.

Acknowledgments

I would like to thank Ion Necoara for pointing him the reference ??. This research was 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.

[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.

[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.

[Nes05a] ———, *Cubic regularization of Newton's method for convex problems with constraints*, CORE Discussion Paper 68, Center for Operation Research and Econometrics, Université catholique de Louvain, 2005.

[Nes05b] ———, *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.

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