Master thesis in non-linear optimization

Global Complexity Analysis For The Second-Order Methods

by

Scieur Damien

Advisor: Y. Nesterov

Readers: P.-A. Absil
F. Glineur

June 2015
Louvain-la-Neuve

Acknowledgements

I want sincerely to thank:

My supervisor, Yurii Nesterov,
for his support and advices,

My friends Mathieu Dath and Léopold Cambier,
for reading my master thesis,

And of course all people who support me.

Contents

Introduction

In many fields of sciences and engineering, we often need to minimize (or maximize) a function f. There exists a lot of different kinds of optimization programs, for example combinatorial optimization. In this master thesis, we will be interested in unconstrained optimization of a multivariate function

$$\min_{x \in \mathbb{R}^n} f(x)$$

where $f(x)$ is twice differentiable. There exists a lot of algorithms which try to minimize such functions, for example the gradient scheme:

$$x_{k+1} = x_k - \alpha_k f'(x_k)$$

for some step size $\alpha_k > 0$. It is one of the most common and known scheme. This algorithm, quite simple, is quite natural: we just follow the direction of the steepest descend. However, We do not use the second order information, i.e. the Hessian of the function. One other very famous scheme is the Newton method:

$$x_{k+1} = x_k - [f''(x)]^{-1} f'(x)$$

The main property of this algorithm is its capacity to converge faster (under some conditions) when we are close to the optimum. However, the scheme has several drawbacks: the inverse of the Hessian is not always well-defined, and we do not have any guarantee on the global convergence of the algorithm.

In paper [5] was proposed a cubic regularisation of the Newton's method. The procedure is to minimize a cubic global upper-estimation of function f at each iteration. With this trick, the step $x_{k+1} - x_k$ will always be well defined, and we can estimate the global rate convergence.

The goal of this master thesis will be firstly to analyze with precision the behavior of this new algorithm on strongly convex functions. During this analysis we will find that the regularisation does not work as well as expected on *smooth functions*, unlike the gradient method. We will thus propose some variants of the original algorithm in order to have better global performances.

Chapter 1
Definitions and notations

1.1 General

We suppose that we work in the space $\mathbb{R}^n$. The inner product $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ between two vectors $v, w \in \mathbb{R}^n$ is

$$\langle v, w \rangle = \sum_{i=1}^{n} v_i w_i.$$

We can thus define the norm $\| \cdot \|_M$ for a vector v where M is a symmetric positive definite matrix:

$$\|v\|_M = \sqrt{\langle Mv, v \rangle} = \sqrt{\langle v, Mv \rangle}$$

We will also use the Euclidean norm $\| \cdot \|_2 = \| \cdot \|_I$, where I is the identity matrix. For more convenience, the notation $\| \cdot \|$ will be used for this norm.

The norm of a square symmetric matrix M can be defined as

$$\|M\| = \|M\|_2 = \max_v \frac{|\langle Mv, v \rangle|}{\|v\|^2} = \sigma_1$$

where σ_i are the singular values of matrix M, indexed by decreasing order.

A square matrix M is called positive semi-definite, or $M \succeq 0$, if and only if all eigenvalues λ_i are non-negatives. For two matrices A and B, the notation $A \succeq B$ means that the matrix $(A - B)$ is positive semi-definite.

In this master thesis, we work with twice-differentiable functions $f : \mathbb{R}^n \mapsto \mathbb{R}, x \mapsto f(x)$. The gradient $f'(x)$ of these functions belongs to $\mathbb{R}^n$ and the Hessian belongs to $\mathbb{R}^{n \times n}$. If a function f is twice differentiable and defined all over $\mathbb{R}^n$, we say that $f \in \mathcal{S}$:

$$f \in \mathcal{S} \Leftrightarrow f : \mathbb{R}^n \mapsto \mathbb{R},\ x \mapsto f(x) \text{ and } f'(x),\ f''(x) \text{ exist on } \mathbb{R}^n.$$

Since the main topic of this master thesis is optimization, we are interested in the minimal value of function f, called f^*. This value is reached at some point x^*. We have thus the relations

$$f^* = \min_x f = f(x^*)$$

and

$$x^* \in \arg\min_x f$$

If we suppose that f has one and only one global minimum x^* then the notation $x^* = \arg\min_x f$ will be used. Moreover, if the function f is differentiable, we have $\|f'(x^*)\| = 0$.

1.2 Convex and strongly convex functions

A twice differentiable function f is called convex (resp. strongly convex with constant $\mu \in \mathbb{R}_0^+$) if and only if for any x in $\operatorname{dom} f$ we have $f''(x) \succeq 0$ (resp. $f''(x) \succeq \mu I$). For such functions, we say $f \in \mathcal{S}^0$ (resp. $f \in \mathcal{S}^\mu$).

1.3 Lipschitz-continuous functions

A function f is called Lipschitz-continuous of constant l if and only if for all x, y we have

$$|f(x) - f(y)| \leq l\|x - y\|_2$$

where l is a positive constant.

If f is twice-differentiable and has a Lipschitz continuous gradient of constant σ then we say $f \in \mathcal{S}_\sigma$. If f is twice-differentiable and has a Lipschitz continuous Hessian of constant L then we say $f \in \mathcal{S}_{\infty,L}$. Indeed, if the function has Lipschitz-continuous gradient and Hessian then $f \in \mathcal{S}_{\sigma,L}$.

1.4 Performance of a scheme

In this paper, we will build some algorithms in order to find an approximation of the minimum of a function, i.e. we want to find $f(x)$ s.t.

$$f(x) - f^* \leq \varepsilon$$

where ε, the accuracy of the approximation, is a real positive value. Any x satisfying this condition is called ε-solution of the problem. The way used to find this approximation (in our case) consists in building an sequence x_k and $f(x_k)$ s.t. $f(x_{k+1}) \leq f(x_k)$ and when $k \to \infty$ then $f(x_k) \to f^*$. For our purpose, it is more convenient to use the notation $\delta_k = f(x_k) - f^*$. When the minimum is unique, then the notation $\Delta_k = x_k - x^*$ can be also used.

Indeed, the number of intermediate points x_k depend also of the quality of the initial point x_0. We will use two different measures of quality of the initial value. The first one is very intuitive and is δ_0. The second one is about the distance $\|x_0 - x^*\|$. Let us define the level set

$\mathcal{D} = \{x : f(x) \leq f(x_0)\}$ and suppose that $\mathcal{D}$ is bounded. Then we can define the diameter of this set D :

$$D = \max_{x,y \in \mathcal{D}} \|x - y\|$$

This value is very important because it is often used in the analysis of the performances of our algorithms.

We will also have to analyse the performance of a scheme as a function of accuracy ε, δ_0 and D. There exists some other measures of performance, but the only one that we will use is very common. It is the worst-case bound in terms of the number of calls of oracle (i.e. computing $f(x)$, $f'(x)$ and $f''(x)$). The main advantage of this point of view is that we can have a guarantee about the maximal number of iterations. But the main drawback is that we cannot deduce the average number of iterations. For example, the Simplex algorithm requires at most an exponential number of iterations to reach the desired solution, but in practice this algorithm converges very quickly.

Sometimes we will also compare different schemes and try to determine which one is the best. We will say that one scheme is better than another when the maximal number of iterations $k_{\max}$ of the first one is better than the maximal number of iterations $k_{\max}$ of the second for any value of ε, δ_0 and D.

1.5 Rate of convergence

In this report, we will describe the rate of convergence of different schemes. For example, we will say that a scheme has a linear rate of convergence when

$$\delta_{k+1} \leq \frac{1}{1+c}\, \delta_k$$

for a positive constant c (often not very large). This constant c must be independent of δ_0 and D. In this case the maximal number of iterations of the scheme in function of D and ε can be computed. Suppose we want $\delta_k \leq \varepsilon$; then the condition

$$\left(\frac{1}{1+c}\right)^k \delta_0 \leq \varepsilon$$

is sufficient. However, since c is not very large, we can use the following relation

$$\frac{1}{1+c} \leq e^{-c}$$

to have a better interpretation of the final result. We have now the stronger condition $e^{-ck}\delta_0 \leq \varepsilon$ and $k_{\max}$ can be easily deduced:

$$k_{\max} = \frac{1}{c} \log\left(\frac{\delta_0}{\varepsilon}\right). \tag{1.1}$$

This rate of convergence is very fast : for example if we want to have a one hundred time more accurate solution then we need $\log(100)/c$ more iterations. Any multiplication in the accuracy or the quality in the initial point intervene in a additive way in the number of iterations (because of the logarithm).

There also exists some faster rates of convergence. For example, the well-known Newton method has a quadratic rate of convergence when the initial point is close enough to the optimal solution (under some assumptions). We will say that a scheme has a quadratic rate of convergence when

$$\delta_{k+1} \leq \frac{1}{c}\delta_k^2.$$

for c a real positive value. Let us write an intermediate value $\alpha_k = \frac{1}{c}\delta_k$. Indeed, we have $\alpha_{k+1} \leq \alpha_k^2$ and a sufficient condition to have an ε solution is

$$\alpha_0^{2^k} \leq \frac{\varepsilon}{c}.$$

We see here that we need $\alpha_0 < 1$, or $\delta_k \leq c$, to have a guarantee of convergence. Now we can easily deduce the maximum number of iterations $k_{\max}$:

$$k_{\max} = \log_2 \left[\frac{\log(\varepsilon/c)}{\log(\delta_0/c)} \right].$$

Between these two rates of convergence there exists so-called super-linear rate of convergence:

$$k_{\max} = \log_\kappa \left[\frac{\log(\varepsilon/c_1)}{\log(\delta_0/c_2)} \right], \tag{1.2}$$

where κ is between one and two for some constants c_1 and c_2. Last but not least, there also exists a rate of convergence called sub-linear. This rate is slower than the linear one, and takes sometimes the form

$$\delta_k \leq \frac{\delta_0}{p(k)} \tag{1.3}$$

where $p(k)$ is a polynomial function in k. To compute the maximum number of iterations we need to solve

$$p(k_{\max}) = \frac{\delta_0}{\varepsilon}.$$

Chapter 2
Main inequalities

In this chapter we introduce all necessary results related to strongly convex functions, functions with Lipschitz-continuous gradient and Lipschitz-continuous Hessian. The majority of the results of this chapter can be found in the book [4].

2.1 Strongly convex functions

There exists a well-known property coming from convex function : when we found a local minimum in a convex function, then we are sure that this minimum is *global*. However, we cannot be sure that a minimum exists and is unique. But for all strongly convex functions there always exists one minimum, which is indeed unique.

Moreover, this class ensures that the Hessian of the function is non-degenerate because if $f \in \mathcal{S}^\mu$ then

$$f''(x) \succeq \mu I, \quad \mu > 0,$$

which means that the Hessian is positive definite everywhere. With this information we will be able to build some algorithms which have a good rate of convergence to the global minimum. For example, there exists a constant step size gradient method which has a linear rate of convergence to the global minimum of such functions.

Note that the definition of strongly convex functions can be extended to differentiable functions. If the following inequality is satisfied for any x, y

$$f(y) - f(x) - \langle f'(x), y - x \rangle \geq \frac{\mu}{2}\|y - x\|^2 \tag{2.1}$$

then f is strongly convex. There exists a geometric interpretation of this result: for any point x there exists a quadratic function which supports the function f. This property is very useful when evaluated at the optimum. Since $f'(x^*) = 0$,

$$f(x) - f(x^*) \geq \frac{\mu}{2}\|x - x^*\|^2 \tag{2.2}$$

which gives us a relation between the accuracy of the value of the function and the proximity of x to the solution.

There exists a lot of very interesting inequalities for strongly convex functions but we will very often use the following theorem.

Theorem 2.1.1. *Suppose $f \in \mathcal{S}^\mu$. For any $x, y \in \mathbb{R}^n$ we have*

$$f(y) - f(x) - \langle f'(x), y - x \rangle \leq \frac{1}{2\mu} \|f'(x) - f'(y)\|^2 \tag{2.3}$$

Proof: At first, suppose x is fixed. Then we can build a function $\phi(y)$:

$$\phi(y) = f(y) - \langle f'(x), y \rangle .$$

By construction, $\phi \in \mathcal{S}^\mu$. Because $\phi'(x) = 0$, we deduce that x minimize $\phi(y)$. If we use also (2.1) we get the following relation

$$\phi(x) = \min_z \phi(z) \geq \min_z \left[\phi(y) + \langle \phi'(y), z - y \rangle + \frac{\mu}{2} \|z - y\|^2 \right] \quad \forall y.$$

We can solve this minimization problem using the first optimality condition

$$\phi'(y) + \mu(z - y) = 0 \qquad \Rightarrow \quad z = y - \frac{\phi'(y)}{\mu}.$$

By consequence,

$$\phi(x) \geq \phi(y) - \frac{1}{2\mu} \|\phi'(y)\|^2.$$

Since this development is valid for any x, it is exactly (2.3). □

A very useful result using this theorem is when we use (2.3) at the optimum x^* :

$$f(x) - f(x^*) \leq \frac{1}{2\mu} \|f'(x)\|^2 \tag{2.4}$$

because it gives us a very useful relation between the gradient of the function at x and the error at point x.

Last but not least, there exists one other property which comes directly from convexity. For such functions it is well-known that the tangent of the function at point x is a global lower-bound for the whole function f:

$$f(y) \geq f(x) + \langle f'(x), y - x \rangle .$$

We can thus write

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

This property will be more useful for us when applied at the optimum:

$$f(x) - f(x^*) \leq \|f'(x)\| \, \|x - x^*\|. \tag{2.5}$$

2.2 Functions with Lipschitz-continuous gradient

This assumption is like the "dual" of strongly convex functions. The eigenvalues of the Hessian have a positive *lower bound*, while the ones of functions with Lipschitz-continuous gradient are bounded above by a positive constant σ. It means that for any $f \in \mathcal{S}_\sigma$ we have

$$\|f'(y) - f'(x)\| \leq \sigma\|y - x\| \quad \text{and} \quad f''(x) \preceq \sigma I \tag{2.6}$$

Knowing that a function has a Lipschitz-continuous gradient (also called *smooth functions*) is very helpful: the graph of such function is between two quadratic functions[1].

Theorem 2.2.1. *Suppose $f \in \mathcal{S}_\sigma$. For any $x, y \in \mathbb{R}^n$ we have*

$$|f(y) - f(x) - \langle f'(x), y - x\rangle| \leq \frac{\sigma}{2}\|y - x\|^2. \tag{2.7}$$

Proof: For any x, y we have

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \langle f'(x + \tau(y - x)), y - x\rangle \, \mathrm{d}\tau \\ &= f(x) + \langle f'(x), y - x\rangle + \int_0^1 \langle f'(x + \tau(y - x)) - f'(x), y - x\rangle \, \mathrm{d}\tau. \end{aligned}$$

Therefore, using the definition of Lipschitz-continuous gradient,

$$\begin{aligned} |f(y) - f(x) + \langle f'(x), y - x\rangle| &= \left| \int_0^1 \langle f'(x + \tau(y - x)) - f'(x), y - x\rangle \, \mathrm{d}\tau \right| \\ &\leq \int_0^1 \|f'(x + \tau(y - x)) - f'(x)\|\|y - x\| \mathrm{d}\tau \\ &\leq \sigma\|y - x\|^2 \int_0^1 \tau \mathrm{d}\tau \\ &= \frac{\sigma}{2}\|y - x\|^2. \end{aligned}$$

□

With (2.7) we can prove a result which is very similar to (2.4).

Theorem 2.2.2. *Suppose $f \in \mathcal{S}_\sigma$. For any $x, y \in \mathbb{R}^n$ we have*

$$f(x) + \langle f'(x), y - x\rangle + \frac{1}{2\sigma}\|f'(x) - f'(y)\|^2 \leq f(y). \tag{2.8}$$

We can use this inequality at the point x^*:

$$f(x) - f(x^*) \geq \frac{1}{2\sigma}\|f'(x)\|^2. \tag{2.9}$$

[1] In this case the quadratic lower bound is concave, unlike strongly convex functions.

Proof: The proof is very similar to (2.3). Let us fix some $x \in \mathbb{R}^n$ and consider the function

$$\phi(y) = f(y) - \langle f'(x), y \rangle .$$

Using (2.7) we have

$$\phi(y^*) \leq \phi\left(y - \frac{1}{\sigma}\phi'(y)\right) \leq \phi(y) - \frac{1}{2\sigma}\|\phi'(y)\|^2,$$

and it is exactly (2.8) □

Smooth functions are very good for first-order methods. For example, there exists a fixed-step gradient method for such functions which converges to a point which is not a maximum s.t. the gradient at this point is zero. The intuition behind this condition is the following: if we ask for the values $f(x)$ and $f'(x)$, then these values will be very close to the ones at $x + \epsilon$ where $\|\epsilon\|$ is small:

$$f(x) - \langle f'(x), \epsilon \rangle - \mathcal{O}(\|\epsilon\|^2) \leq f(x+\epsilon) \leq f(x) + \langle f'(x), \epsilon \rangle + \mathcal{O}(\|\epsilon\|^2)$$

$$\|f'(x) - f'(x+\epsilon)\| \leq \sigma\|\epsilon\|.$$

In other words, we want a function which is robust over the impact of the argument x.

2.3 Functions with Lipschitz-continuous Hessian

Like above, we will make another assumption over the robustness of the function. Since we are interested in second order methods, we want functions for which the variation of the Hessian is bounded:

$$\|f''(y) - f''(x)\| \leq L\|y - x\|. \tag{2.10}$$

We will thus assume that $f \in \mathcal{S}_{\infty,L}$. We can integrate two times the above condition, leading to two very interesting inequalities.

Theorem 2.3.1. *Suppose $f \in \mathcal{S}_{\infty,L}$. Then for any x, y we have*

$$\|f'(y) - f'(x) - f''(x)(y-x)\| \leq \frac{1}{2}L\|y-x\|^2. \tag{2.11}$$

$$\left|f(y) - f(x) - \langle f'(x), y-x \rangle - \frac{1}{2}\langle f''(x)(y-x), y-x \rangle\right| \leq \frac{1}{6}L\|y-x\|^3. \tag{2.12}$$

Proof: Indeed,

$$\begin{aligned}
\|f'(y) - f'(x) - f''(x)(y-x)\| &= \left\| \int_0^1 [f''(x+\tau(y-x)) - f''(x)](y-x)\mathrm{d}\tau \right\| \\
&\leq \|y-x\| \int_0^1 \|[f''(x+\tau(y-x)) - f''(x)]\| \,\mathrm{d}\tau \\
&\leq L\|y-x\|^2 \int_0^1 \tau \mathrm{d}\tau \\
&= \frac{L}{2}\|y-x\|^2.
\end{aligned}$$

It is exactly (2.11). With a similar idea,

$$\begin{aligned}
&\left| f(y) - f(x) - \langle f'(x), y - x \rangle - \frac{1}{2} \langle f''(x)(y-x), y-x \rangle \right| \\
&= \left| \int_0^1 \langle f'(x + \tau(y-x)) - f'(x) - \tau f''(x)(y-x), y-x \rangle \, \mathrm{d}\tau \right| \\
&\leq \|y - x\| \int_0^1 \|f'(x + \tau(y-x)) - f'(x) - \tau f''(x)(y-x)\| \mathrm{d}\tau.
\end{aligned}$$

We can now use (2.11), leading to (2.12).

□

Chapter 3

Cubic regularization of the Newton's Method (CNM)

We present now an algorithm (presented in [5]) used for minimizing unconstrained functions with Lipschitz-continuous Hessian. This algorithm is an improvement of the Newton's method, because this scheme converges everywhere in the domain and is always well-defined.

This new second-order scheme, the *cubic regularisation of the Newton's method (CNM)*, minimizes a cubic model at each step. This cubic model is a global upper estimation of the objective function. By analyzing the decreasing of this cubic model, we will be able to deduce the rate of convergence of the algorithm.

3.1 Regular algorithm

The idea of the CNM is to minimize the cubic model (2.12). By minimizing this expression, we are sure that the next iterate x_{k+1} satisfies

$$f(x_{k+1}) \leq f(x_k).$$

Let us introduce the following mapping :

$$T_M(x) \in \arg\min_{y \in E} f(x) + \langle f'(x), y - x \rangle + \frac{1}{2} \langle f''(x)(y-x), y-x \rangle + \frac{M}{6} \|y - x\|^3. \tag{3.1}$$

Where $M \geq L$ is a positive parameter.

It is important to note that $T_M(x)$ is the solution of the following system of equations

$$f'(x) + f''(x)(T_M(x) - x) + \frac{1}{2} M \|T_M(x) - x\| \cdot (T_M(x) - x) = 0. \tag{3.2}$$

The basic algorithm is to choose $x_{k+1} = T_M(x_k)$. In the rest of this report we assume that $M = L$ and $T_L(x) = T$ for more simplicity.

From the first optimality condition (3.2) we can derive the rate of convergence (more detail in [5]):

$$\min_k \|f'(x_k)\| \leq \mathcal{O}\left(\frac{\delta_0}{k}\right)^{2/3}$$

If we suppose f convex, we can have a better result.

Theorem 3.1.1. *Suppose that f is convex and has a non-empty set of global minimum X^*. Also, suppose that the value D (recall: the radius of the level set $\{x : f(x) \leq f(x_0)\}$) is finite. Then,*

- *If $\delta_0 > \frac{3}{2}LD^3$ we have*
$$f(x_1) - f(x^*) \leq \frac{1}{2}LD^3. \tag{3.3}$$
- *If $\delta_0 \leq \frac{3}{2}LD^3$ we have*
$$f(x_k) - f(x^*) \leq \frac{9LD^3}{(k+4)^2}. \tag{3.4}$$

We will admit this theorem without proof. The complete development can be found in [2].

This theorem means that the method convergences to a global minimum. We see also that the rate of convergence is polynomial. Let us estimate the number of iterations to reach a precision ε. For that we need to solve the following sufficient condition,

$$\frac{9LD^3}{(k+4)^2} \leq \varepsilon,$$

and we can thus find an estimation on the maximum number of iterations:

$$k \leq \sqrt{\frac{9LD^3}{\epsilon}} - 4 \leq \sqrt{\frac{9LD^3}{\epsilon}}. \tag{3.5}$$

Moreover, if we assume that the set of optimal points is globally non-degenerate with parameter $\mu > 0$ (i.e. (2.1) holds for $x = x^*$ and for any y) then we can prove a better local result.

Theorem 3.1.2. *Suppose that $f(x)$ is convex and admits a globally non-degenerate optimal set. Then*

1. *If $f(x_0) - f(x^*) \geq \left(\frac{2}{3L}\right)^2 \left(\frac{\mu}{2}\right)^3 = \bar{\omega}$ the process converge at the following rate.*
$$\delta_k^{1/4} \leq \delta_0^{1/4} - \frac{k}{6}\bar{\omega}^{1/4} \tag{3.6}$$
2. *If $f(x_0) - f(x^*) \leq \bar{\omega}$ then the convergence becomes super-linear.*
$$\delta_{k+1} \leq \sqrt{\frac{1}{9\bar{\omega}}}\delta_k^{3/2} \tag{3.7}$$

The complete development can be found in [5].

We will try to deduce using expression (3.6) the maximum number of iterations needed from δ_0 to $\bar{\omega}$. Indeed, we need to satisfy this condition to ensure $\delta_k \leq \bar{\omega}$:

$$\delta_0^{1/4} - \frac{k}{6}\bar{\omega}^{1/4} \leq \bar{\omega}^{1/4}.$$

Thus, we find an estimation of the maximal number of iterations:

$$k \leq 6\left(\left[\frac{\delta_0}{\bar{\omega}}\right]^{1/4} - 1\right) \leq 6\left[\frac{\delta_0}{\bar{\omega}}\right]^{1/4}. \tag{3.8}$$

We will now try to find the switching value ζ on δ_k which minimizes the total number of iterations using the combination of (3.5) and (3.8):

$$\zeta = \arg\min_x \sqrt{\frac{9LD^3}{x}} + 6\left[\frac{x}{\bar{\omega}}\right]^{1/4}.$$

The first optimality condition is

$$-\frac{\sqrt{9LD^3}}{2}\zeta^{-3/2} + \frac{6}{4\bar{\omega}^{1/4}}\zeta^{-3/4} = 0.$$

Assuming $\zeta \neq 0$ we find

$$\zeta = \left(\bar{\omega}^{1/4}\sqrt{LD^3}\right)^{4/3} = \frac{\mu D^2}{18^{1/3}}.$$

The number of iterations needed for going to $\varepsilon = \bar{\omega}$ is thus bounded by

$$9 \cdot 18^{1/6}\sqrt{\frac{LD}{\mu}} \leq 14.6\sqrt{\frac{LD}{\mu}}. \tag{3.9}$$

After this phase, the scheme converge with a super-linear rate of convergence. We will wait a little bit before making the precise upper bound, because we will see later better results when we add the strong convexity assumption.

3.2 Minimal decreasing

Let us will analyse the minimal decrease between two iterations of CNM. Using the first order optimality condition (3.2):

$$\|f'(T)\| = \left\|f'(T) - f'(x) - f''(x)(T-x) - \frac{L}{2}\|T-x\|(T-x)\right\|.$$

If we use now (2.11) we get (using $M = L$):

$$\|f'(T)\| \leq L\|T-x\|^2. \tag{3.10}$$

Now consider again (3.2) and multiply it by $(T-x)$. It becomes (recall that we supposed $M = L$)

$$\langle f'(x), T-x\rangle = -\langle f''(x)(T-x), T-x\rangle - \frac{1}{2}L\|T-x\|^3. \tag{3.11}$$

Using (2.12) with $y = T$ in combination with (3.11) we get

$$\begin{aligned} f(T) &\leq f(x) + \langle f'(x), T - x\rangle + \frac{1}{2}\langle f''(x)(T-x), T-x\rangle + \frac{L}{6}\|T-x\|^3 \\ &= f(x) - \frac{1}{2}\langle f''(x)(T-x), T-x\rangle - \frac{L}{3}\|T-x\|^3, \end{aligned}$$

wich leads us to an expression of the minimal decreasing $f(x) - f(T)$:

$$f(x) - f(T) \geq \frac{1}{2}\langle f''(x)(T-x), T-x\rangle + \frac{L}{3}\|T-x\|^3. \tag{3.12}$$

We can also use (3.10) in this equation to get a relation between the decreasing and the gradient at the point T.

3.3 Accelerated algorithm

By using some appropriate estimate sequences, the algorithm can be improved for convex functions. All details can be found in [2].

Initialization : Choose $x_0 \in E$. Set $M = 2L$ and $N = 12L$. Compute $x_1 = T_L(x)$ and define $\psi_1(x) = f(x_1) + \frac{N}{6}\|x - x_0\|^3$.
Iteration $k \geq 1$ **:**

1. Compute $v_k = \arg\min_x \psi(x)$ and choose $y_k = \frac{k}{k+3}x_k + \frac{3}{k+3}v_k$

2. Compute $x_{k+1} = T_M(y_k)$ and update

$$\psi_{k+1} = \psi_k + \frac{(k+1)(k+2)}{2}[f(x_{k+1}) + \langle f'(x_{k+1}, x - x_{k+1}\rangle]$$

We can show that the convergence of this algorithm is

$$f(x_k) - f(x^*) \leq \mathcal{O}\left(\frac{LD^3}{k(k+1)(k+2)}\right). \tag{3.13}$$

The big $\mathcal{O}$ is there because the complexity change in function of the update of $\psi_k(x)$ up to a factor. This function can, for example, be updated with linear of quadratic functions (in the above case we showed the linear update). Indeed this result is much better than the regular algorithm.

Chapter 4

Properties of the intersection of functional classes

In general, we analyse the performances of a scheme over one "simple" and specific functional class. However, sometimes an algorithm can have very different behaviour when we add additional information over the function that we want to minimize. For example, under some assumptions we can prove that the Newton's method converges quadratically when we are close enough to the optimum. But when we add the fact that the function is quadratic then the Newton's method converges in only *one* iteration.

The goal of this chapter is to derive some properties that we can deduce using the information of the *intersection* of some functional classes. It will help us to analyse the behaviour of some algorithms for functions which belong to several functional classes. Also, some classes will give us a global upper bound, which can be useful for building a better algorithm based on the minimization of the upper-estimations.

4.1 Functions with Lipschitz-continuous gradient and Hessian

We start from analysing this very important class: if some function f belongs to $\mathcal{S}_{\sigma,L}$ then we can apply either the gradient method or the CNM with some guarantees.

All functions $f \in \mathcal{S}_{\sigma,L}$ have the following two properties:

$$f''(y) \preceq \sigma I, \quad \sigma > 0,$$
$$\|f''(y) - f''(x)\| \leq L\|y - x\|.$$

However, we can put these two conditions together and try to have an equivalent definition of the class $\mathcal{S}_{\sigma,L}$.

Theorem 4.1.1. *A function f belongs to $\mathcal{S}_{\sigma,L}$ if and only if for all $x, y \in \mathbb{R}^n$ and for all vectors $u \in \mathbb{R}^n$ we have*

$$\langle [f''(y) - f''(x)]u, u\rangle \leq \min\left\{L\|y - x\|\|u\|^2 \; ; \; \langle [\sigma I - f''(x)]u, u\rangle\right\}. \tag{4.1}$$

Proof: First suppose that $f \in \mathcal{S}_{\sigma,L}$. Since $f''(x) \preceq \sigma I$ we have for any $u \in \mathbb{R}^n$

$$\begin{aligned} &\langle f''(y)u, u\rangle \leq \sigma\|u\|^2 = \langle[\sigma I]u, u\rangle \\ \Leftrightarrow &\langle[f''(y) - f''(x)]u, u\rangle \leq \langle[\sigma I - f''(x)]u, u\rangle . \end{aligned}$$

Moreover,

$$\langle[f''(y) - f''(x)]u, u\rangle \leq \|f''(y) - f''(x)\|\|u\|^2 \leq L\|y - x\|\|u\|^2.$$

By combining these two inequalities we get the desired result. Now suppose that f satisfies (4.1) and let us show that $f \in \mathcal{S}_{\sigma,L}$. First, we will prove that $f(y) \preceq \sigma I$. Indeed,

$$\begin{aligned} &\langle[f''(y) - f''(x)]u, u\rangle \leq \langle[\sigma I - f''(x)]u, u\rangle \\ \Leftrightarrow &\langle[f''(y)]u, u\rangle \leq \sigma\|u\|^2 \\ \Rightarrow &\frac{\langle[f''(y)]u, u\rangle}{\|u\|^2} \leq \sigma, \quad u \neq 0 \\ \Leftrightarrow &\max_{u\neq 0} \frac{\langle[f''(y)]u, u\rangle}{\|u\|^2} \leq \sigma. \end{aligned}$$

Meaning that $\lambda_{\max}(f''(y)) \leq \sigma$. This condition is equivalent to

$$f''(y) \preceq \sigma I$$

for all y. Now we will prove the second condition. Indeed,

$$\langle[f''(y) - f''(x)]u, u\rangle \leq L\|y - x\|\|u\|^2.$$

We can use this inequality with x and y interchanged to conclude that

$$|\langle[f''(y) - f''(x)]u, u\rangle| \leq L\|y - x\|\|u\|^2.$$

Finally,

$$\frac{|\langle[f''(y) - f''(x)]u, u\rangle|}{\|u\|^2} \leq L\|y - x\|.$$

and we get the result by taking the maximum over u. □

Now suppose that the direction of u is fixed. One interesting thing to know is the norm of $y - x$ such that the minimum will switch between two values. We need to compute $\|y - x\|$ which solve the following equation:

$$L\|y - x\|\|u\|^2 = \langle[\sigma I - f''(x)]u, u\rangle .$$

We find easily that

$$\|y - x\| = \frac{\langle[\sigma I - f''(x)]u, u\rangle}{L\|u\|^2}.$$

We can see that in fact the right hand side is homogeneous of degree zero in u. Let us introduce a parameter $\gamma_x(y)$:

$$\gamma_x(y) = \min\left\{1; \frac{\langle[\sigma I - f''(x)](y-x), y-x\rangle}{L\|y-x\|^3}\right\}.$$

We can recognize on the right the "relative" switching value when $u = y - x$. We can now derive a weaker condition (but more useful for later) of inequality (4.1).

Corollary 4.1.1. *If a function f belongs to $\mathcal{S}_{\sigma,L}$ then*

$$\langle[f''(y) - f''(x)](y-x), y-x\rangle \leq \gamma_x(y) L\|y-x\|^3 \tag{4.2}$$

Proof: First, suppose that $\gamma_x(y) = 1$. In this case, by definition of $\gamma_x(y)$,

$$L\|y-x\|^3 \leq \langle[\sigma I - f''(x)](y-x), y-x\rangle$$

Thus we get the right expression using (4.1) with $u = y - x$. Suppose now $\gamma_x(y) < 1$. Using the same argument (4.2) is proved. □

We have now a simpler expression for the integration. We will now see two very interesting inequalities, very similar to (2.11) and (2.12).

Theorem 4.1.2. *If $f \in S_{\sigma,L}$, then for any x, y:*

$$\langle f'(y) - f'(x) - f''(x)(y-x), y-x\rangle \leq \left(\gamma_x(y) - \frac{\gamma_x(y)^2}{2}\right) L\|y-x\|^3 \tag{4.3}$$

$$f(y) - f(x) - \langle f'(x), y-x\rangle - \frac{1}{2}\langle f''(x)(y-x), y-x\rangle \leq \left(\frac{\gamma_x(y)(1-\gamma_x(y))}{2} + \frac{\gamma_x(y)^3}{6}\right) L\|y-x\|^3 \tag{4.4}$$

Proof: Using the definition of $\gamma_x(y)$,

$$\begin{aligned}
\langle f'(y) - f'(x) - f''(x)(y-x), y-x\rangle &= \int_0^1 \langle(f''(x+\tau(y-x)) - f''(x))(y-x), y-x\rangle \,\mathrm{d}\tau \\
&\leq \int_0^1 \min\left\{\tau L\|y-x\|^3;\, \langle(\sigma I - f''(x))(y-x), y-x\rangle\right\} \mathrm{d}\tau \\
&= \int_0^{\gamma_x(y)} \tau L\|y-x\|^3 \mathrm{d}\tau + \int_{\gamma_x}^1 \langle(\sigma I - f''(x))(y-x), y-x\rangle \,\mathrm{d}\tau \\
&= \frac{\gamma_x^2(y) L}{2}\|y-x\|^3 + (1-\gamma_x(y))\langle(\sigma I - f''(x))(y-x), y-x\rangle.
\end{aligned}$$

If $\gamma_x(y) = 1$, then we get (4.3). If $\gamma_x(y) < 1$, we just replace $\langle(\sigma I - f''(x))(y-x), y-x\rangle$ by

$L\|y-x\|^3\gamma_x(y)$ to have also (4.3). Now we can use this result for the proof of (4.4):

$$\begin{aligned}
&f(y)-f(x)-\langle f'(x),y-x\rangle-\frac{1}{2}\langle f''(x)(y-x),y-x\rangle\\
=&\int_0^1\langle f'(x+\tau(y-x))-f'(x)-f''(x)\tau(y-x),y-x\rangle\,\mathrm{d}\tau\\
=&\int_0^1\langle f'(x+\tau(y-x))-f'(x)-f''(x)\tau(y-x),\tau(y-x)\rangle\,\frac{1}{\tau}\mathrm{d}\tau\\
\leq&\int_0^{\gamma_x(y)}\frac{1}{2}L\|y-x\|^3\tau^2\mathrm{d}\tau+\int_{\gamma_x(y)}^1\left(\frac{\gamma_x(y)}{\tau}-\frac{\gamma_x(y)^2}{2\tau^2}\right)L\|y-x\|^3\tau^2\mathrm{d}\tau\\
=&L\|y-x\|^3\left(\frac{\gamma_x(y)-\gamma_x(y)^2}{2}+\frac{\gamma_x(y)^3}{6}\right).
\end{aligned}$$

□

We can remark that we have now a new upper bound (4.4) for functions $f\in\mathcal{S}_{\mu,L}$. This upper-bound combines two interesting aspects of the functional class:

- The first one is the local cubic approximation. When y is close to x, then the model is cubic and thus gives us a very accurate approximation of f.
- However, when y is far from x then the model becomes quadratic. This quadratic model will grow very slowly in comparison with the cubic one.

We will see later that this "switched model" (locally cubic, globally quadratic) will be very useful for the interpretation of the behaviour of the CNM on some functional classes.

4.2 Strongly convex functions with Lipschitz-continuous Hessian

In this section, all proofs are very similar. That is why all results will be admitted without any proof.

Adding the strongly convex property will ensure us that the function has one and only one global minimum. Also, the quadratic lower bound (2.2) due to the strong convexity assumption will ensure us a minimal rate of growth [1]. The consequence of this property is that the global rate of convergence of the CNM will be improved on such functions.

Let us write down the formal definition of this class. A function $f\in\mathcal{S}^{\mu}_{\infty,L}$ if and only if the function f follows these two properties:

$$\begin{aligned}
f''(y)&\succeq\mu I,\quad\mu>0\\
\|f''(y)-f''(x)\|&\leq L\|y-x\|
\end{aligned}$$

[1] In general we do not use (2.2). We prefer to use (2.3) which comes also from the strong convexity assumption.

Like in the section above, we will use a parameter $\beta_x(y)$:

$$\beta_x(y) = \min\left\{1\ ;\ \frac{\langle [f''(x) - \mu I](y-x), y-x\rangle}{L\|y-x\|^2}\right\}.$$

We can derive with this parameter the following theorem, which bounds below the function f by a switched model (locally cubic, globally quadratic).

Theorem 4.2.1. *Suppose $f \in \mathcal{S}^{\mu}_{\infty,L}$. We have*

$$\langle f'(y) - f'(x) - f''(x)(y-x), y-x\rangle \geq -\left(\beta_x(y) - \frac{\beta_x(y)^2}{2}\right) L\|y-x\|^3 \tag{4.5}$$

$$f(y) - f(x) - \langle f'(x), y-x\rangle - \frac{1}{2}\langle f''(x)(y-x), y-x\rangle \geq -\left(\frac{\beta_x(y)(1-\beta_x(y))}{2} + \frac{\beta_x(y)^3}{6}\right) L\|y-x\|^3 \tag{4.6}$$

Unfortunately, we will see later that having a very precise lower bound will not give us more information about the minimal decrease of the function f. However it does not mean that having a strongly convex function does not impact the global complexity.

4.3 Strongly convex functions with Lipschitz-continuous gradient and Hessian

The last class which remains to be analyzed is $\mathcal{S}^{\mu}_{\sigma,L}$. Any function $f \in \mathcal{S}^{\mu}_{\sigma,L}$ satisfies

$$\mu I \preceq f''(y) \preceq \sigma I, \quad 0 < \mu \leq \sigma,$$
$$\|f''(y) - f''(x)\| \leq L\|y-x\|.$$

Unfortunately, we cannot derive better bounds than (4.4) and (4.6). But functions which belong to this class are very interesting for this reason: in [4] it is shown that strongly convex functions with Lipschitz-continuous gradient are *easy* for first order methods. Moreover, adding the Lipschitz-continuous Hessian condition will allow us to apply and compare the CNM or its variants to the first order methods.

In [2] it is shown that the CNM enters polynomially to the quadratic region of convergence, while some first order methods (say, for example, the gradient method) converge globally with a linear rate of convergence, which is indeed much better. In the same paper it was introduced the following open question to which we will try to find an answer:

"For the problem class $\mathcal{S}^{\mu}_{\sigma,L}$, can we get any advantages from the second order schemes being used at the initial stage of the minimization process?".

4.4 Relaxation of the bounds for $\mathcal{S}^{\mu}_{\sigma,L}$

In this section we will show weaker results of (4.4) and (4.6) by relaxing parameters $\gamma_x(y)$ and $\beta_x(y)$. The aim of this relaxation is to have simpler bounds, which is useful for example when we want to minimize the model.

First of all we will introduce a new parameter $\theta_x(y)$:

$$\theta_x(y) = \min\left\{1\ ;\ \frac{\sigma-\mu}{L\|y-x\|}\right\}.$$

The main advantage of $\theta_x(y)$ is that this parameter does not depend on a scalar product. Also, for any $x,\ y$ we have

$$\gamma_x(y) \le \theta_x(y),$$
$$\beta_x(y) \le \theta_x(y).$$

Which allow us to replace $\gamma_x(y)$ and $\beta_x(y)$ by $\theta_x(y)$ in (4.4) and (4.6). Finally, if $f \in \mathcal{S}^{\mu}_{\sigma,L}$ we can put the two bounds together:

Corollary 4.4.1. *If $f \in S^{\mu}_{\sigma,L}$, then for any x, y in* dom f *:*

$$\left|\langle f'(y) - f'(x) - f''(x)(y-x), y-x\rangle\right| \le \left(\theta_x(y) - \frac{\theta_x(y)^2}{2}\right) L\|y-x\|^3 \tag{4.7}$$

$$\left|f(y) - f(x) - \langle f'(x), y-x\rangle - \frac{1}{2}\langle f''(x)(y-x), y-x\rangle\right| \le \left(\frac{\theta_x(y)(1-\theta_x(y))}{2} + \frac{\theta_x(y)^3}{6}\right) L\|y-x\|^3 \tag{4.8}$$

Chapter 5
CNM applied to strongly convex functions

Let us analyse the global behaviour of CNM on strongly convex functions. A local result is given in [2]: the CNM converges quadratically when we are close to the optimum, like the Newton's method. But the convergence of the algorithm in the first stage is still proportional to $\sqrt{D}$ (where D is the diameter of the set $\{x : f(x) \leq f(x_0)\}$) when we add the smooth assumption. This is quite bad because the gradient method achieves a much better global complexity.

The goal of this section is first to describe with precision the behaviour of CNM on strongly convex functions with Lipschitz-continuous Hessian. Then we will show some specific difficult functions for CNM.

5.1 Impact of the strong convexity assumption

We have seen before (equation (4.6)) a new expression of the lower bound of a function $f \in \mathcal{S}^{\mu}_{\infty,L}$. This new expression can now be used to characterize with more precision the decrease of the function between two iterations:

Lemma 5.1.1. *Suppose $f \in \mathcal{S}^{\mu}_{\infty,L}$. Then*

$$f(x) - f(T) \leq \left(\frac{\beta_x(T)(1-\beta_x(T))+1}{2} - \frac{\beta_x(T)^3}{6} \right) L\|T-x\|^3 + \frac{1}{2}\langle f''(x)(T-x), T-x\rangle \quad (5.1)$$

Proof: By using (4.6) in combination (3.11) with we get:

$$f(T)-f(x)+\frac{1}{2}\langle f''(x)(T-x), T-x\rangle+\frac{L}{2}\|T-x\|^3 \geq -\left(\frac{\beta_x(T)(1-\beta_x(T))}{2} - \frac{\beta_x(T)^3}{6} \right) L\|T-x\|^3$$

which leads us to the desired result. □

This is for us a bad news: improving the lower bound (2.12) with the strongly convex assumption does not gives us additional information on the *minimal* decrease. However, we can use this assumption in (3.12):

$$\begin{aligned} f(x) - f(T) &\geq \frac{1}{2}\langle f''(x)(T-x), T-x\rangle + \frac{L}{3}\|T-x\|^3 \\ &\geq \frac{\mu}{2}\|T-x\|^2 + \frac{L}{3}\|T-x\|^3. \end{aligned}$$

However this relation is difficult to use. This is why we will decompose the inequality into two simpler relations:

$$f(x) - f(T) \geq \frac{\mu}{2}\|T - x\|^2, \tag{5.2}$$

$$f(x) - f(T) \geq \frac{L}{3}\|T - x\|^3. \tag{5.3}$$

The first inequality will be used when we are close to the optimum because $\|T - x\|^2$ will be larger than $\|T - x\|^3$. We will thus use the second one to describe the first stage of the algorithm. Sometimes we prefer to use (5.2) and (5.3) with (3.10):

$$f(x) - f(T) \geq \frac{\mu}{2L}\|f'(T)\|, \tag{5.4}$$

$$f(x) - f(T) \geq \frac{1}{3\sqrt{L}}\|f'(T)\|^{3/2}. \tag{5.5}$$

5.2 Stopping criterion

The main goal is to find $x : f(x) - f(x^*) \leq \varepsilon$. We will try here to have similar stopping criterion. For example, a more interesting condition can be on $\|x - T\|$ or $\|f'(T)\|$, much easier to compute.

Lemma 5.2.1. *We have*

$$f(T) - f(x^*) \leq \frac{1}{2\mu}\|f'(T)\|^2 \tag{5.6}$$

$$f(T) - f(x^*) \leq \frac{L^2}{2\mu}\|T - x\|^4 \tag{5.7}$$

By consequence, if one of these conditions is satisfied,

$$\|f'(T)\| \leq \sqrt{2\mu\varepsilon},$$
$$\|x - T\| \leq \sqrt[4]{\frac{2\mu}{L^2}\varepsilon},$$

then $f(T) - f(x^*) < \varepsilon$.

Proof: First we will prove (5.6). Using (2.4) at $x = T$:

$$f(T) - f(x^*) \leq \frac{1}{2\mu}\|f'(T)\|^2. \tag{5.8}$$

Having $\frac{1}{2\mu}\|f'(T)\|^2 \leq \varepsilon$ ensures us the desired accuracy. Now we will prove (5.7). If we use (3.10) on (5.8) at $x = T$:

$$f(T) - f(x^*) \leq \frac{L^2}{2\mu}\|T - x\|^4.$$

Asking $\frac{L^2}{2\mu}\|T - x\|^4$ to be smaller than ε gives us (5.7). □

For theoretical purposes we may be interested in a stopping criterion on $\|x - x^*\|$. We will see later that we can characterize the quadratic region of convergence with this value.

Lemma 5.2.2. *We have*

$$f(T) - f(x^*) \leq \frac{L}{3}\|x - x^*\|^3 \tag{5.9}$$

Proof: Start with the upper bound of (2.12) at $y = T$:

$$f(T) \leq f(x) + \langle f'(x), T - x\rangle + \frac{1}{2}\langle f''(x)(T - x), T - x\rangle + \frac{L}{6}\|T - x\|^3.$$

By definition of T:

$$f(T) \leq \min_y \left[f(x) + \langle f'(x), y - x\rangle + \frac{1}{2}\langle f''(x)(y - x), y - x\rangle + \frac{L}{6}\|y - x\|^3\right].$$

We can now use the lower bound of (2.12) in the previous inequality:

$$f(x) + \langle f'(x), y - x\rangle + \frac{1}{2}\langle f''(x)(y - x), y - x\rangle + \frac{L}{6}\|y - x\|^3 \leq f(y) + \frac{L}{3}\|y - x\|^3.$$

leading to

$$f(T) \leq \min_y \left[f(y) + \frac{L}{3}\|y - x\|^3\right].$$

If we choose y to be equal to x^*, we get exactly (5.9). □

One can see that we could also use the lower bound (4.6). Using this inequality leads us to a little bit stronger (but very similar) result, but the expression is too complex to be useful.

5.3 Global complexity

5.3.1 First stage of the minimization process

We will now try to describe with precision the global complexity of the CNM when applied on strongly convex functions with Lipschitz-continuous Hessian. We will derive first the asymptotic rate of convergence, i.e. the bound on the maximum number of iterations when we are very far from the optimum.

Theorem 5.3.1. *When applied on functions which belongs to $\mathcal{S}^{\mu}_{\infty,L}$, the rate of convergence of the CNM is bounded as follow:*

$$\delta_k \geq \left(1 + \frac{K}{\sqrt{D}}\right)\delta_{k+1} \tag{5.10}$$

where $K = \frac{1}{3}\sqrt{2\frac{\mu}{L}}$. Using the relation (1.1) the maximum number of iterations is bounded by

$$k^{(1)}_{\max} \leq \frac{\sqrt{D}}{K}\log\left(\frac{\delta_0}{\epsilon}\right) = \frac{3}{\sqrt{2}}\sqrt{\frac{LD}{\mu}}\log\left(\frac{\delta_0}{\epsilon}\right) \tag{5.11}$$

Proof: Let us start from (5.5) and decompose $\|f'(T)\|^{3/2}$ into $\|f'(T)\|\|f'(T)\|^{1/2}$. We can use the expression (2.4) to bound $\|f'(T)\|$ and (2.5) to bound $\|f'(T)\|^{1/2}$:

$$\begin{aligned}\|f'(T)\| &\geq \sqrt{2\mu(f(T)-f(x^*))} \\ \|f'(T)\|^{1/2} = \sqrt{\|f'(T)\|\frac{\|T-x^*\|}{\|T-x^*\|}} &\geq \sqrt{\frac{f(T)-f(x^*)}{D}} \qquad \text{because } D \geq \|T-x^*\|.\end{aligned}$$

Finally, using these two inequalities on (5.5) (where $x_k = x$ and $x_{k+1} = T$),

$$\begin{aligned}f(x_k) - f(x_{k+1}) = \delta_k - \delta_{k+1} &\geq \frac{1}{3\sqrt{L}}\|f'(k+1)\|\|f'(k+1)\|^{1/2} \\ &\geq \underbrace{\frac{1}{3}\sqrt{2\frac{\mu}{L}}}_{=K}\frac{1}{\sqrt{D}}\delta_{k+1},\end{aligned}$$

we get the desired result. □

Despite how the expression looks, the rate of convergence *is not* linear because $k_{\max}$ grows in $\sqrt{D}$ and not in $\log(D)$. However, the result is *global*, meaning that the CNM will converge for any starting point x_0. Also, the CNM will reach the optimum at any accuracy with a finite number of operations.

We can also see that the expression is quite strange: the decreasing is leaded by a coefficient which depends of D. Before going further, we will analyse this expression. Assume D very large, then we can use the first-order approximation:

$$\delta_k \leq \left(\frac{1}{1+\frac{K}{\sqrt{D}}}\right)\delta_{k-1} \approx \left(1-\frac{K}{\sqrt{D}}\right)\delta_{k-1}.$$

With (2.2) we can deduce that

$$D \leq \left[\frac{2}{\mu}\delta_0\right]^{1/2}. \tag{5.12}$$

Then,

$$\delta_k \leq \left(1-\frac{K}{\left[\frac{2}{\mu}\delta_0\right]^{1/4}}\right)\delta_{k-1} \leq \left(1-\frac{K}{\left[\frac{2}{\mu}\delta_0\right]^{1/4}}\right)^k\delta_0.$$

Now we can use one more time a first-order approximation to have a lower bound on the rate of convergence in the worst case:

$$\delta_0 - k\left[\frac{\mu K^4}{2}\right]^{1/4}\delta_0^{3/4}.$$

We see here clearly a polynomial expression of the decrease of the function. Note that we can

also replace the value of D using (5.12) in (5.11):

$$k_{\max}^{(1)} \leq \frac{\left[\frac{2}{\mu}\delta_0\right]^{1/4}}{K} \log\left(\frac{\delta_0}{\epsilon}\right).$$

Surprising enough, this expression is worse than (3.8) despite the "linear-like" expression (5.10). However, it gives us a good estimation of the relation between the number of iterations and D: (5.10) grows is $\mathcal{O}\left(\sqrt{D}\log(D)\right)$ while (5.10) grows in $\mathcal{O}\left(D^{3/4}\right)$ (if we use (3.3), assuming that we have done already one iteration). If we compare it with (3.9) (assuming that $\varepsilon \geq \bar{\omega}$), then we remark that the estimation is not so bad : indeed asymptotically this precise bound is much better, but the constant is also higher, which means that for D not too big, the two bounds are quite equivalent.

Despite the fact that (5.10) is not the best bound, the expression of its maximum number of iterations and its rate of convergence summarize well the behaviour of the CNM on strongly convex function with Lipschitz-continuous Hessian. We will see also later that the expression (5.10) is very similar to the complexity of the fast gradient method applied on $\mathcal{S}^{\mu}_{\infty,L}$

5.3.2 Super-linear and quadratic rate of convergence

Like the Newton's method the CNM is able to converge faster when we get closer to the optimum. Let us introduce two switching values ω_1 and ω_2:

$$\omega_1 = \left(\frac{3}{L}\right)^2 \left(\frac{\mu}{2}\right)^3 \quad , \quad \omega_2 = \frac{\mu^3}{2L^2} = \frac{4}{9}\omega_1. \tag{5.13}$$

Those two switching values are conditions on δ_k. Let us first prove the super-linear then the quadratic rate of convergence.

Lemma 5.3.1. *Suppose $f(x) - f(x^*) \leq \omega_1$. Then the CNM converges superlineary:*

$$\delta_{k+1} \leq \sqrt{\frac{1}{\omega_1}}\delta_k^{3/2} \tag{5.14}$$

and the number of iterations needed to reach ε is

$$k_{\max}^{(3)} \leq \log_{3/2}\left[\frac{\log\left(\frac{\omega_1}{\epsilon}\right)}{\log\left(\frac{\omega_1}{\delta_0}\right)}\right]. \tag{5.15}$$

Proof: Use (5.9) then (2.2):

$$f(T) - f(x^*) \leq \frac{L}{3}\|x - x^*\|^3 \leq \frac{L}{3}\left(\frac{2}{\mu}(f(x) - f(x^*))\right)^{3/2}.$$

It is exactly (5.14). Denote

$$\alpha_k = \frac{\delta_k}{\omega_1}.$$

Then,

$$\alpha_k \leq \alpha_{k-1}^{3/2} \leq \alpha_0^{(3/2)^k}.$$

Since we want a precision ε, the condition

$$\alpha_0^{(3/2)^k} \leq \frac{\varepsilon}{\omega_1}$$

is sufficient, leading us to (5.14). □

We can compare this super-linear rate with (3.7). First of all it is obvious that the region of super linear convergence of (5.14) is bigger and the coefficient which multiply $\delta_k^{3/2}$ is also smaller. We can thus conclude that the rate (5.14) is a better result that (3.7) in any cases. Now we will see the quadratic rate of convergence of δ_k (presented in [2]) and $\|x_{k+1} - x_k\|$.

Lemma 5.3.2. *Suppose $f(x) - f(x^*) \leq \omega_2$ or $\|x_0 - x_1\| \leq \frac{3\mu}{2L}$. Then the CNM converges quadratically:*

$$\|x_{k+1} - x_k\| \leq \frac{L}{\mu}\|x_k - x_{k-1}\|^2, \tag{5.16}$$

$$\delta_{k+1} \leq \frac{1}{\omega_2}\delta_k^2. \tag{5.17}$$

Therefore, the number of iteration needed to reach ε is

$$k_{\max}^{(4)} \leq \log_2 \left[\frac{\log\left(\frac{\omega_2}{\epsilon}\right)}{\log\left(\frac{\omega_2}{\delta_0}\right)} \right]. \tag{5.18}$$

Proof: Let us use firstly (3.11). Indeed,

$$\|f'(x)\|\|T - x\| \geq \langle f''(x)(T - x), T - x\rangle + \frac{1}{2}L\|T - x\|^3.$$

Since f is strongly convex (meaning that $f''(x) \succeq \mu I$) and if we forget the last term,

$$\|f'(x_k)\| \geq \mu\|x_{k+1} - x_k\|.$$

Now, use (3.10). We get

$$\|x_{k+1} - x_k\| \leq \frac{L}{\mu}\|x_{k-1} - x_k\|^2.$$

If $\|x_0 - x_1\| \leq \frac{\mu}{L}$ is satisfied, then the above sequence converges to zero.

We can now build a stronger condition if we use (5.2):

$$f(x) - f(T) \leq \frac{\mu^3}{2L^2} = \omega_1 \qquad \Leftarrow \qquad \delta_0 \leq \frac{\mu^3}{2L^2} = \omega_1.$$

We can also describe the rate of convergence of δ_k. If we use (5.4) then (2.4):

$$\delta_k \geq f(x_k) - f(x_{k+1}) \geq \frac{\mu}{2L}\|f'(x_{k+1})\| \geq \frac{\mu}{2L}\sqrt{2\mu\delta_{k+1}}.$$

Leading us to the expression (5.17), with the same condition of convergence. □

5.3.3 Bound on the total number of iterations

We will try here to put together all previous results in order to have a precise bound on the total number of iterations. We will assume that we begin very far from the solution (D is big) and that we want a very accurate solution (ε is very small).

We have already computed the number of iterations in the first phase (see (3.9)). We will now try to compute the minimal number of iterations in the second phase, i.e. when $\delta_0 \leq \bar{\omega}$ (where $\bar{\omega} = \frac{\mu^3}{18L^2}$). Let us first estimate a condition on δ_k for which the super-linear rate converges faster than the quadratic rate. We need to find ζ such that

$$\sqrt{\frac{1}{\omega_1}}\zeta^{3/2} = \frac{1}{\omega_2}\zeta^2.$$

We find

$$\zeta = \frac{\omega_2^2}{\omega_1} = \frac{2\mu^3}{9L^2}. \tag{5.19}$$

We see here that the value of ζ is larger than $\bar{\omega}$. It means that at the end of the first phase, we have a point which is already in the region of quadratic convergence. Let us estimate the number of iterations in this phase using (5.18):

$$k \leq \log_2\left(\frac{\log\left(\frac{\omega_2}{\varepsilon}\right)}{\log\left(\frac{\omega_2}{\bar{\omega}}\right)}\right) = \log_2\left(\log_4\left(\frac{\omega_2}{\varepsilon}\right)\right).$$

We have now a good idea of the total complexity of the CNM :

$$k_{\max} \leq 14.6\sqrt{\frac{LD}{\mu}} + \log_2\left(\log_4\left(\frac{\omega_2}{\varepsilon}\right)\right). \tag{5.20}$$

We see that the number of iterations is proportional to the square-root of D. Paper [5] compared the performances of the CNM with the performances of the optimal first-order methods for smooth strongly convex functions (see [4] for more information). Let us call $\hat{L}$ the largest eigenvalue of $f''(x)$. Since we work with functions with Lipschitz-continuous Hessian, we can estimate $\sigma = \hat{L} + LD$. The complexity of the optimal first-order method is of the order of

$$\mathcal{O}\left(\sqrt{\frac{\hat{L}+LD}{\mu}}\log\left(\frac{(\hat{L}+LD)D^2}{\varepsilon}\right)\right).$$

We conclude that for strongly convex functions the performance of the CNM is much better than the performance of the first order method on our class of problem. Note that with the complexity bound (5.11) we have the same conclusion.

5.4 Examples where CNM works bad

We have seen before a bound on the total number of iteration for the CNM. We have also seen that the number of iterations grows in $\mathcal{O}\left(\sqrt{\frac{LD}{\mu}}\right)$. We will try here to analyse several functions which are difficult to minimize with the CNM.

5.4.1 Intuitive example : smooth approximation of absolute value

First of all we will analyse the performance of the CNM on a smooth approximation of absolute value. Let[1] $f(x) = \log\left(e^x + e^{-x}\right)$. Indeed,

- $f'(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \approx \text{sign}(x)$.
- $f''(x) = 1 - f'(x)^2, \quad f'(0) = 0 \Rightarrow x^* = 0$.
- $|f'''(x)| = \left|\frac{8(e^x-1)(e^x+1)e^{2x}}{(e^{2x}+1)^3}\right| < 1 \quad \rightarrow L = 1$.

Let us call $D = |x_0 - x^*| = |x_0|$. For more simplicity, we will assume that x_0 is positive and very big and that ε is not too small. We can now deduce μ

$$\mu = f''(x_0) - f'(x_0)^2 > 0$$

meaning that the function is strongly convex over the domain $\{x : |x| \leq D\}$. Indeed, the parameter μ decreases a lot in function of D and tends to zero, meaning that f is "only" strictly convex on $\mathbb{R}$. The goal of this example is not to show the worst case function but to understand intuitively the characteristics of a function which is difficult for the CNM. Since we assumed that ϵ is not very small and x positive, we can write an approximation of the mapping $T_L(x)$:

$$T_L(x) = \arg\min_y (y - x) + \frac{1}{6}|y - x|^3.$$

because $f'(x) \approx 1$ and $f''(x) \approx 0$. We have thus

$$T_L(x) = x - \sqrt{2}.$$

We can easily compute an approximation of the number of iterations:

$$k \approx \frac{D - \varepsilon}{\sqrt{2}}.$$

[1] There exists other kinds of smooth approximation of $|x|$, like $\sqrt{\epsilon + x^2}$.

We see here that the number of iteration is much worse than (5.20). It is due to the fact that μ is function of D.

The main characteristic of this function is the (almost) constant gradient $f'(x)$. The norm of the gradient does not increases when D is larger. So one characteristic of difficult functions is having a gradient which does not increase a lot when we are far from the optimum, while staying strongly convex. If we use (2.2) and (2.4) we get

$$\mu\|x - x^*\| \leq \|f'(x)\|.$$

For more simplicity we will work in one dimension. If we take the lower-bound and integrate one time, we get

$$f(x) - f(x^*) = \frac{\mu}{2}(x - x^*)^2.$$

This is indeed a quadratic function. We can thus expect that kind of functions is difficult to minimize with the CNM.

5.4.2 One-dimensional quadratic function

We have seen before an intuition about a difficult class of functions: the quadratic functions. However, for such functions, $L = 0$ and the CNM converges in only one iteration. For now, assume that we work with $f(x) = \frac{x^2}{2}$ and we want a precision not too small. Indeed, $\mu = 1$. However, the estimation of the Lipschitz constant is pessimistic: $L = 1$. Let us now write the mapping $T_L(x)$.

$$T_L(x) = \arg\min_y \left[x(y - x) + \frac{1}{2}(y - x)^2 + \frac{1}{6}|y - x|^3 \right].$$

The first order optimality condition is

$$T - \frac{1}{2}(T - x)^2 = 0.$$

We can now compute the explicit expression of the solution T

$$T = x + 1 - \sqrt{2x + 1} \geq x - \sqrt{2x}.$$

To avoid negative numbers, we suppose $x \geq 2$. Suppose we want to reach a precision ε. This condition is strictly equivalent to

$$\|x_k - x^*\| = \|x_k\| \leq \sqrt{\varepsilon}.$$

We can now compute the minimal number of iterations to reach this accuracy:

$$\sqrt{\varepsilon} \geq x_k = T_L(x_{k-1}) \geq x_{k-1} - \sqrt{2x_{k-1}} \geq x_0 - k\sqrt{x_0}.$$

Leading us to

$$k_{\min} \geq \sqrt{\frac{D}{2}} - \sqrt{\frac{\varepsilon}{2D}} \geq \sqrt{\frac{D}{2}}.$$

We can compare this lower bound to (3.9):

$$k_{\max} \leq 14.6\sqrt{\frac{LD}{\mu}} = 14.6\sqrt{D}.$$

Both bounds grow in $\sqrt{D}$, meaning that we have a good upper bound, describing nicely the complexity of the CNM applied on the class $\mathcal{S}^{\mu}_{\infty,L}$. One can say that assuming an error on the parameter L is kind of "cheating". In practice, we can run the algorithm with bad estimation of the parameter L, but let us suppose that we have a subroutine which estimates perfectly a local value of L. We will now build a very similar function. Call

$$h_1 = 2\frac{\sigma - \mu}{L} \quad , \quad h_2 = \frac{\sigma - \mu}{2}.$$

Now we will build $g(x) \in \mathcal{S}^{\mu}_{\sigma,L}$:

$$g(x) = -\left(\frac{h_2}{h_1^2}\right)\sin(h_1 x) + (h_2 + \mu)\frac{x^2}{2} + \frac{h_2}{h_1}x \tag{5.21}$$

$$g'(x) = -\left(\frac{h_2}{h_1}\right)\cos(h_1 x) + (h_2 + \mu)x + \frac{h_2}{h_1} \tag{5.22}$$

$$g''(x) = h_2 \sin(h_1 x) + h_2 + \mu. \tag{5.23}$$

The behaviour of the CNM on $g(x)$ is very similar to $(1/2)x^2$. For the illustration (see figure (5.1)), let us see the graph of $g(x)$, $(1/2)x^2$ and x with $\mu = L = 1$ and $\sigma = 2$. We see on the figure that $g(x)$ are between $\frac{\mu}{2}x^2$ and $\frac{\sigma}{2}x^2$. We have also that $g'(x)$ and $g''(x)$ is between the gradient/Hessian of the two square functions. We can thus assume that when we apply the CNM algorithm on $g(x)$ the number of iterations will also grow in $\mathcal{O}(\sqrt{D})$, while having a local value of the parameter L equal to one.

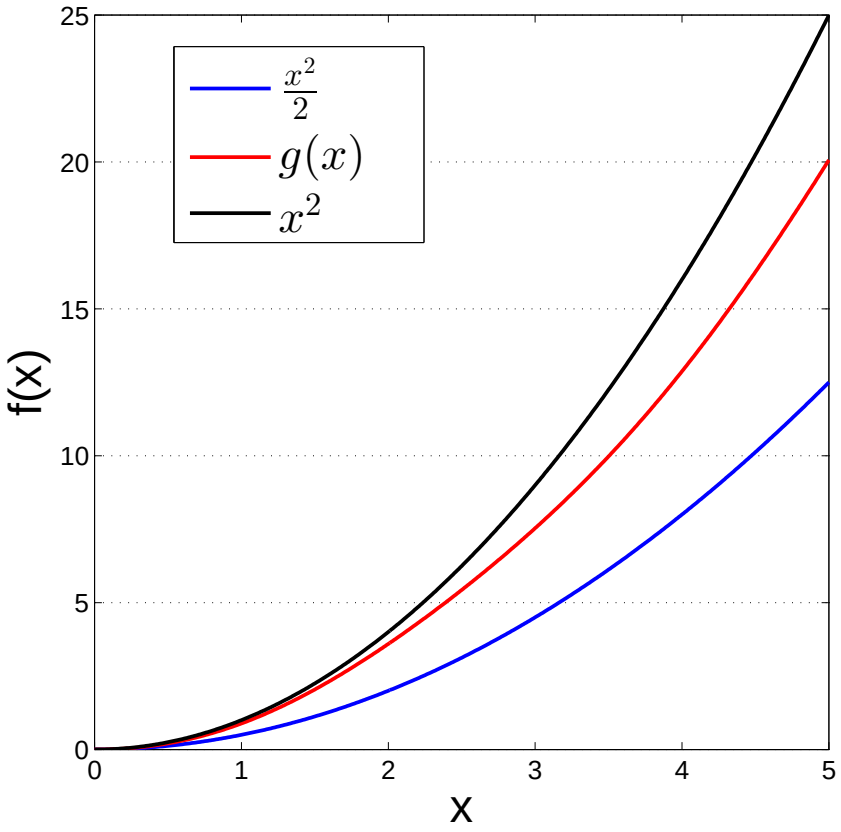

Figure 5.1: Comparison of the graph of some functions in $\mathcal{S}^{\mu}_{\sigma,L}$.

Chapter 6

Combining gradient method and CNM: the hybrid scheme

6.1 Differences between the gradient method and CNM

In the previous section, we analysed the complexity of the CNM on strongly convex functions, and we found that a bad function is the quadratic one. The bad news is the following: the example x^2 belongs also to $\mathcal{S}^{\mu}_{\sigma,L}$. We know that for this class there exists a gradient method which converges linearly, and the optimal method for this class has the following complexity:

$$k_{\max}^{\text{Optimal method}} \leq \sqrt{\frac{\sigma}{\mu}} \log\left(\frac{\delta_0}{\varepsilon}\right).$$

The number of iterations grows in $\log(\delta_0)$(and, by consequence, also in $\log(D)$), which is *much* better than the CNM. It is very weird, because adding one more information (in this case: the function is *smooth*), does not affect the behaviour of the CNM.

The main reason is the following: we have seen before that the CNM minimizes an auxiliary function $m_3(x)$ (model of order 3) to find the next iterate. This secondary function is a global upper-bound, defined at (3.1). Let us compare this cubic model with the bound computed at (2.7):

$$m_2(y) = f(x) + \langle f'(x), y - x\rangle + \frac{\sigma}{2}\|y - x\|^2$$

The minimiser of this function is $x_{k+1} = x_k - \frac{1}{\sigma}f'(x_k)$. We can easily deduce that $f(y) \leq m_2(y)$. We have in fact here a quadratic model, which is also a global upper-approximation of function f.

We will now compare the two bounds. Suppose that we are looking at y which is very far from x (for example when we perform large steps). Then the error between the two models is of the order of $\frac{L}{6}\|y - x\|^3$, meaning that for large steps the cubic model $m_3(y)$ is not useful in comparison with $m_2(y)$. However, when we are very close to x, then the conclusion is not the same: the approximation is much better with the cubic model. Because the two aspects are very important for the approximation of the function, we will thus analyse the hybrid method, which consists to take x_{k+1} :

$$x_{k+1} \in \arg\min_y \left[m_2(y),\, m_3(y)\right].$$

The general behaviour of this new algorithm will be the following : we can expect that the hybrid method will take gradient steps at the first stage of the minimization process and CNM steps for

the last phase.

We should think at the fact that for this algorithm we need one more parameter σ. Having a good estimation for such parameter can be sometimes difficult to obtain. Also, an adaptive algorithm which implements this method has to handle two different parameters, which can be hard.

In this section we will work firstly with function with Lipschitz-continuous gradient and Hessian (i.e. functional class $\mathcal{S}_{\sigma,L}$). If needed, the convexity or strong convexity assumption will be added.

6.2 Complexity analysis

We will now try to derive the rate of convergence of the hybrid algorithm. We have seen before that we expect the gradient method better than the CNM when we are far from the optimum. If the algorithm takes the gradient steps at the first stage of the process, it means that

$$-\frac{1}{2\sigma}\|f'(x)\|^2 \leq \langle f'(x), T-x\rangle + \frac{1}{2}\langle f''(x)(T-x), T-x\rangle + \frac{L}{6}\|T-x\|^3.$$

We can thus expect a condition over the gradient : when the norm of $f'(x)$ is large enough then the gradient step will be taken.

Lemma 6.2.1. *Suppose $f \in \mathcal{S}_{\sigma,L}$. If we have*

$$\|f'(x)\| \geq 8\frac{\sigma^2}{L} \tag{6.1}$$

then $\min_y m_2(y) \leq \min_y m_3(y)$.

Proof: For more convenience in this proof, call $Q = \sigma\sqrt{\frac{2}{L}}$. Consider the following inequality in variable z:

$$\frac{1}{2}z^2 - \frac{2}{3}Qz - \frac{Q^2}{2} \geq 0.$$

We can show that this inequality holds for $z \geq 2Q$. Let $\sqrt{\|f'(x)\|} \geq 2Q$. We can thus write

$$\frac{\|f'(x)\|}{2} - \frac{2\|f'(x)\|^{1/2}}{3}Q - \frac{Q^2}{2} \geq 0.$$

After a rearrangement of the terms, and by multiplying both sides by $\frac{\|f'(x)\|}{\sigma}$:

$$\frac{\|f'(x)\|^2}{2\sigma} \geq \frac{2\|f'(x)\|^{3/2}}{3\sigma}Q + \frac{Q^2}{2\sigma}\|f'(x)\|.$$

We can now replace Q by its expression:

$$\frac{\|f'(x)\|^2}{2\sigma} \geq \frac{L}{3}\left(\frac{2\|f'(x)\|}{L}\right)^{3/2} + \frac{\sigma}{L}\|f'(x)\|.$$

The left-hand side is exactly $f(x) - \min_y m_2(y)$. So, focus now on the right hand side of the inequality. We can use (3.11) to deduce that

$$\|f'(x)\| \geq \frac{L}{2}\|T - x\|^2.$$

If we use this intermediate result on our right hand side, we get

$$\frac{\|f'(x)\|^2}{2\sigma} \geq \frac{L}{3}\|T - x\|^3 + \frac{\sigma}{2}\|T - x\|^2.$$

Since f has a Lipschitz-continuous gradient, $f''(x) \preceq \sigma I$:

$$\frac{\|f'(x)\|^2}{2\sigma} \geq \frac{L}{3}\|T - x\|^3 + \frac{1}{2}\left\langle f''(x)(T - x), T - x\right\rangle.$$

We can now use (3.11) to add "zero" to our expression, leading us to

$$\frac{\|f'(x)\|^2}{2\sigma} \geq -\left\langle f'(x), T - x\right\rangle - \frac{1}{2}\left\langle f''(x)(T - x), T - x\right\rangle - \frac{L}{6}\|T - x\|^3.$$

We can recognize the right hand side of the inequality to be $f(x) - \min_y m_3(y)$, which proves the desired result. □

We have seen here that the gradient method will be very efficient during a long time. Whatever the initial point will be, the gradient step will be chosen while the norm of the gradient is bigger than a *constant* value. Thus, for D big, we can suppose that the total complexity is comparable to the complexity of the gradient method only. However, if very-high accuracy is needed, then the CNM step can be also useful: for example the convergence is quadratic for strongly convex functions.

6.2.1 Global complexity for convex functions

Now we will add the convexity hypothesis. Recall that for such problem there exists (not always) a convex set of global minimum X^* of the function f. We will suppose in this section that this set exists and is bounded. We will also suppose that the value D is also bounded and very big, and the accuracy $\varepsilon > 0$ is very small.

First stage: gradient method

We have seen before that the gradient method will be used when $\|f'(x)\| \geq 8\frac{\sigma^2}{L}$. We can prove that the rate of this method is polynomial. Indeed, by definition of the gradient step,

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \frac{1}{2\sigma}\|f'(x)\|^2 \\ &\geq 32\frac{\sigma^3}{L^2}. \end{aligned}$$

Since $f(x_0) - f(x^*) \geq f(x_0) - f(x_k) \geq 4\frac{\sigma}{L}$, we have

$$k \leq \frac{L^2}{32\sigma^3}\delta_0.$$

By (2.7) we have that δ_0 is bounded by $\frac{\sigma}{2}D^2$, leading us to a bound on the number of iterations k_1 in the first phase:

$$k_1 \leq \frac{L^2}{64\sigma^2}D^2. \tag{6.2}$$

Last stage: Hybrid method

We have computed before the number of iterations for the gradient method. We can derive here the general expression of the rate of convergence:

$$f(x_k) - f(x^*) \leq \frac{2\sigma D^2}{k+4}.$$

We can now use this expression with (3.4) to deduce the rate of convergence of δ_k. Since we take the best step between the gradient method and the CNM, we have

$$f(x_k) - f(x^*) \leq \min\left\{\frac{9LD^3}{(k+4)^2}\ ;\ \frac{2\sigma D^2}{k+4}\right\}.$$

Let us have a switching value $\tilde{k}$ in function of D, meaning that $\tilde{k}$ solves

$$\frac{9LD^3}{(\tilde{k}+4)^2} = \frac{2\sigma D^2}{\tilde{k}+4}.$$

We find

$$\tilde{k} = \frac{9L}{2\sigma}D - 4.$$

Note that we cannot say that we already reached this switching value after the first phase because (6.2) is an upper-bound on the number of iterations. Also, having more than (6.2) iterations does not means that we will always use CNM because. This switching value just tells us that the rate of convergence of the algorithm is characterized by the rate of the CNM.

Once this switching value is attained, we have to finish the minimization process by k_2 iterations, with k_2 s.t.

$$\frac{9LD^3}{(k_2 + \max\{\tilde{k}; k_1\} + 4)^2} \leq \varepsilon.$$

Note that we cannot bound D because it can be as large as we want (if we want to bound D we need a stronger assumption than convexity). We can thus obtain a sufficient condition for k_2:

$$k_2 \geq \sqrt{\frac{9LD^3}{\varepsilon}} - \max\{\tilde{k}; k_1\} - 4.$$

Total complexity

We have seen that for the first phase we bounded the k_1 first iterations by a fixed value (defined in (6.2)). Then, we needed to reach the switching value $\tilde{k}$. In this case the required number of iterations is $\max\{0; \tilde{k} - k_1\}$. During this phase, the rate of convergence is characterized by the decrease of the gradient method. Finally, once this switching value is attained, the rate of convergence is bounded by the expression of the rate of the CNM. The total number of iterations $k_{\max}$ is thus bounded by

$$k_{\max} \leq k_1 + \max\{0; \tilde{k} - k_1\} + \sqrt{\frac{9LD^3}{\varepsilon}} - \max\{\tilde{k}; k_1\} - 4 = \sqrt{\frac{9LD^3}{\varepsilon}} - 4.$$

Surprising enough the total number of iterations is in fact the number of iterations needed when using CNM. We can thus deduce that the hybrid method is not really useful on this class of function. However, we need to keep in mind that at the first stage of the minimization process we use only gradient steps, which are much more easier to compute than a CNM step.

6.2.2 Global complexity for strongly convex functions

In chapter 5 we noticed that the number of iterations of the CNM is quite big in comparison with the gradient method when we begin far from the optimum. However when the accuracy ε is very small, then the CNM is much more useful. Combining the two methods to minimize a function f which belongs to the class $\mathcal{S}^{\mu}_{\sigma,L}$ seems intuitive and efficient.

Before entering in the complexity analysis, we have to notice that for this class the optimal step for the fixed-step gradient method is in fact (see [4])

$$x_{k+1} = x_k - \frac{2}{\sigma + \mu} f'(x).$$

However, this step length is too big when $\mu \to 0$ (i.e. when a strongly convex function is close to be a strictly convex function). Because we were interested in convex functions, we will let the coefficient $\frac{1}{\sigma}$ instead of $\frac{2}{\sigma+\mu}$. This will not affect the final conclusion, but the bound on the total number of iterations will indeed change up to a scalar factor.

First stage: gradient method

In this case, we can show that the rate of convergence is linear. Indeed, since we use the gradient method, we can deduce the minimal decrease by replacing $y - x$ by $\frac{-1}{\sigma} f'(x)$ in (2.7):

$$\delta_{k+1} \leq \delta_k - \frac{1}{2\sigma} \|f'(x_k)\|^2.$$

We can now use (2.4) to bound $\|f'(x)\|$:

$$\delta_{k+1} \leq \left(1 - \frac{\mu}{\sigma}\right) \delta_k \leq \delta_0 e^{-\frac{\mu}{\sigma}(k+1)}. \tag{6.3}$$

We will now derive the number of iterations in the first phase. If we use (2.9), then

$$\frac{1}{2\sigma}\|f'(x_k)\|^2 \leq \delta_k.$$

We can deduce with this result and (6.1) that the number of iterations k_1 in the first phase is bounded by

$$32\frac{\sigma^3}{L^2} \leq \delta_{k_1} \leq \delta_0 e^{-\frac{\mu}{\sigma}k_1} \quad \Rightarrow \quad k_1 \leq \frac{\sigma}{\mu}\log\left(\frac{L^2}{32\sigma^3}\delta_0\right). \tag{6.4}$$

At the end of this phase we have the following accuracy

$$\delta_{k_1} \leq \frac{1}{2\mu}\|f'(x_{k_1})\|^2 \leq 32\frac{\sigma^4}{\mu L^2}. \tag{6.5}$$

Second stage: Hybrid method

When the first phase is finished, we need to reach the region of super-linear convergence of the CNM with the hybrid method. We have seen that after the gradient method phase, we have that the accuracy δ_{k_1} is bounded (see the expression (6.5)). The region of super-linear convergence is $\{x : f(x) \leq \omega_1\}$, where ω_1 is defined in (5.13).

In this phase we do not have any guarantee that we will go faster than the gradient method. Therefore, the rate of convergence is still (6.3). The number of iterations k_2 of the hybrid method in this phase must ensure that $\delta_{k\in[k_1,k_2]}$ goes from δ_{k_1} to the region of super-linear convergence. At first, let us see for which value of δ the rate of convergence of the gradient method is equal to the super-linear rate:

$$\left(1 - \frac{\mu}{\sigma}\right)\delta = \sqrt{\frac{1}{\omega_1}}\delta^{3/2} \qquad \Rightarrow \qquad \delta = \omega_1\left(1 - \frac{\mu}{\sigma}\right)^2.$$

The number of iterations k_2 must therefore satisfy $\delta_{k_2} \leq \omega_1\left(1 - \frac{\mu}{\sigma}\right)^2$. The following condition

$$\delta_{k_1} e^{-\frac{\mu}{\sigma}k_2} \leq \omega_1$$

is thus sufficient. We can thus bound k_2 by

$$k_2 \leq \frac{\sigma}{\mu}\log\left(\frac{\delta_{k_1}}{\omega_1\left(1 - \frac{\mu}{\sigma}\right)^2}\right) \leq \frac{\sigma}{\mu}\left[4\log\left(\frac{\sigma}{\mu}\right) + 2\log\left(\frac{16}{3\left(1 - \frac{\mu}{\sigma}\right)}\right)\right]. \tag{6.6}$$

The accuracy at the end of this phase is as follow:

$$\delta_{k_2} \leq \omega_1 \left(1 - \frac{\mu}{\sigma}\right)^2. \tag{6.7}$$

Last stage: CNM

Now we are in the region of super-linear convergence of the CNM. The gradient method becomes now less useful because we do not have any guarantee that the gradient method will converge faster when we are close to the optimum.

We need now to reach the region of quadratic convergence. We already computed the switching value ζ in (5.19) (i.e. the condition on δ when the super-linear rate is slower than the quadratic rate). We need thus to go from δ_{k_2} to ζ with a super-linear rate, leading to the following sufficient condition on the number of iterations k_3 in this phase (where $\alpha_k = \delta_k/\omega_1$):

$$\alpha_0^{(3/2)^k} \leq \frac{\zeta}{\omega_1}.$$

By consequence, the bound on k_3 is

$$k_3 \leq \log_{3/2}\left(\frac{\log\left(\frac{\omega_1}{\zeta}\right)}{\log\left(\frac{\omega_1}{\delta_{k_2}}\right)}\right) \leq \log_{3/2}\left(\frac{\log\left(\frac{4}{9}\right)}{-\log\left(1-\frac{\mu}{\sigma}\right)}\right).$$

At the end of this phase, we have $\delta_{k_3} \leq \zeta$. Now we need to compute one last time the number of iterations k_4 for going from ζ to ε with a quadratic rate of convergence. Let α_k be now $\frac{\delta_k}{\omega_2}$:

$$\alpha_0^{2^k} \leq \varepsilon \qquad \Rightarrow \qquad k_4 \leq \log_2 \log_{\frac{4}{9}} \frac{1}{\varepsilon}. \tag{6.8}$$

Total complexity

Since we have computed the complexity of all phases, we can sum all k_i in order to have an idea of the maximal number of iterations $k_{\max}$ of the hybrid method:

$$k_{\max} \leq \frac{\sigma}{\mu}\left[\log\left(\frac{L^2}{32\sigma^3}\delta_0\right) + 4\log\left(\frac{\sigma}{\mu}\right) + 2\log\left(\frac{16}{3\left(1-\frac{\mu}{\sigma}\right)}\right)\right] + \log_{3/2}\left(\frac{\log\left(\frac{4}{9}\right)}{-\log\left(1-\frac{\mu}{\sigma}\right)}\right) + \log_2 \log_{\frac{4}{9}} \frac{1}{\varepsilon}.$$

We can summarize this result with the $\mathcal{O}$ notation, assuming δ_0 very big and ε very small:

$$k_{\max} = \mathcal{O}\left(\frac{\sigma}{\mu}\log\left(\frac{L^2}{\sigma^3}\delta_0\right) + \log_2 \log_{\frac{4}{9}} \frac{1}{\varepsilon}\right). \tag{6.9}$$

This expression tells us that the number of iterations is, as expected, of the order of the number of iterations of the gradient method needed to reach the quadratic region of convergence combined

with the number of iterations of the CNM once this region is attained.

Note that for this scheme the most of the work (the k_1 first iterations) is achieved by the gradient method. Also, the CNM step is at least as complicated to be computed as the gradient step, meaning that the total number of iterations k_{max} itself is not really relevant alone, but must be presented as $k_{\text{gradient}} + k_{\text{hybrid}}$.

6.3 Conclusion

The hybrid method is very intuitive and gives us acceptable results: for strongly convex functions we ensure at least the linear rate of convergence for any starting point, and at the end we will converge very quickly to the optimum.

For convex function the conclusion is mitigated: we have seen that the number of iterations of the hybrid method is not better than the CNM. However we need to keep in mind that this analysis only take care of the worst case: In the average case, it is obvious that the hybrid method is at least better than the CNM or the gradient.

The main drawback of this method is σ, a new parameter needed to run the hybrid algorithm. The estimation of one parameter is much easier than two. Despite this fact, the gradient step is free to compute in comparison of the CNM step. If we can have a good *a priori* estimation of σ and L then the hybrid method is a good choice for minimizing strongly convex functions with Lipschitz-continuous gradient and Hessian.

Chapter 7

Minimizing a more accurate model: the γ-method

7.1 Motivations

In the previous section we analysed the hybrid method: assuming that we know the two parameters σ and L, we perform simultaneously a gradient and a CNM step, and we take the best one between the two. However, the model that we minimize is not convex. We can imagine taking the "convex hull" of the two models, but this task can be complex[1].

We can us make another suggestion: instead of taking the minimum of the two models $m_2(y)$ and $m_3(y)$, we can minimize directly the model (4.4) (we will call it the γ-model or $m_\gamma(y)$):

$$\min_y \underbrace{f(x) + \langle f'(x), y-x\rangle + \frac{1}{2}\langle f''(x)(y-x), y-x\rangle + \frac{L}{6}\|y-x\|^3\left(\frac{\gamma_x(y)(1-\gamma_x(y))}{2} + \frac{\gamma_x(y)}{6}\right)}_{=m_\gamma(y)}.$$

By construction, the γ-model belongs to $\mathcal{S}_{\sigma,L}$. Also, we have

$$m_\gamma(x) = f(x), \qquad m'_\gamma(x) = f'(x), \qquad m''_\gamma(x) = f''(x)$$

(the derivatives are easy to compute since we are around $y = x$, meaning that $\gamma_x(x) = 1$). The second derivative of this model is

$$m''_\gamma(y) = \begin{cases} f''(x) + \frac{L}{\|y-x\|}(y-x)(y-x)^T & \text{if } \gamma_x(y) = 1 \\ \sigma I & \text{if } \gamma_x(y) < 1 \end{cases}$$

If we add the fact that $m_\gamma(x) \in \mathcal{S}_{\sigma,L}$, this expression leads us to the relation $m''_\gamma(y) \succeq f''(x)$, meaning that if f is strongly convex, then the model will be also strongly convex (of parameter $\lambda_{\min}[f''(x)]$).

[1] In fact, taking the minimum of the two models is more or less equivalent of minimizing the convex hull of the two models.

7.2 Complexity analysis

This model looks very promising: we bound a function $f \in \mathcal{S}^{\mu}_{\sigma,L}$ by another function in $\mathcal{S}^{\mu}_{\sigma,L}$. We can thus conclude that $m_\gamma(y)$ is the best global upper-bound for this class. We can thus expect from this model a performance which is at least as good as the hybrid method.

We can also remark that m_γ is similar to the hybrid method, because the model is both quadratic and cubic. When the step size is small, then $\gamma_x(y)$ will be equal to one. Therefore, we have in this case $m_\gamma(y) = m_3(y)$. However, suppose the step size is very large, i.e. $\gamma_x(y) \to 0$. In this case $m_\gamma(y) = m_2(y)$. So for the extreme cases the two methods are equivalent. The advantage of this method is when $m_3(y)$ is comparable to $m_2(y)$, i.e when $\gamma_x(y)$ is not too small, but below one. In this case we can see with the γ-model a modified gradient step which is more aggressive. Suppose $\gamma_x(y) < 1$, then $m_\gamma(y)$ can be written as

$$m_\gamma(y) = f(x) + \langle f'(x), y - x\rangle + \frac{\sigma}{2}\|y-x\|^2 + L\|y-x\|^3 \underbrace{\left(\frac{\gamma_x(y)^3}{6} - \frac{\gamma_x(y)^2}{2}\right)}_{<0} < m_2(y). \quad (7.1)$$

We will see that this expression leads to longer steps than the usual gradient method.

With the previous results, we can thus deduce that

$$m_\gamma(y) \leq \min\left\{m_2(y)\ ;\ m_3(y)\right\}.$$

The complexity of the γ-method is thus at least as good as the hybrid method.

Unfortunately, because $\gamma_x(y)$ can be as close to zero as possible we cannot improve the complexity bound (6.9). But in practice the algorithm converge a little bit faster when $\gamma_x(y)$ becomes not too small. The usefulness of this scheme is finally quite mitigated: we have a method which is a least better than the hybrid method, but the global performances are not really improved.

We can confirm this fact by a numerical example. Let us take the function $g(x)$ defined in (5.21) with parameters $\mu = 1$, $\sigma = 20$ and $L = 0.01$. Suppose that we want a very large ε such that we have always $\gamma_x(x_{k+1}) < 1$. Suppose also that we minimize this function with the gradient method and the γ-method with an error on the parameters ($\hat{\sigma} = 1000\sigma$ and $\hat{L} = 100L$).

In this case, we can see in figure (7.2) that at the beginning of the process, $\gamma_x(y)$ is very close to zero. This is why the global complexity is not better than the gradient method (this fact is confirmed with figure (7.1): at the beginning, both method converge at the same rate). However, when we get closer to the minimum of this function, then we see that the convergence becomes incredibly fast. We can explain this fact with the figure (7.2) and formula (7.1). We see that because of the right parenthesis we use a much larger step than the gradient method,

leading to a better local rate of convergence, which spares more or less the half of the number of iterations of the gradient method. Despite the fact that the theoretical conclusion are not really optimistic, the performance of the process is much better when applied in practice. Also, we applied this scheme on a one-dimensional example. The algorithm can thus be more efficient on a more complex function because unlike the gradient method, we also take care of the Hessian.

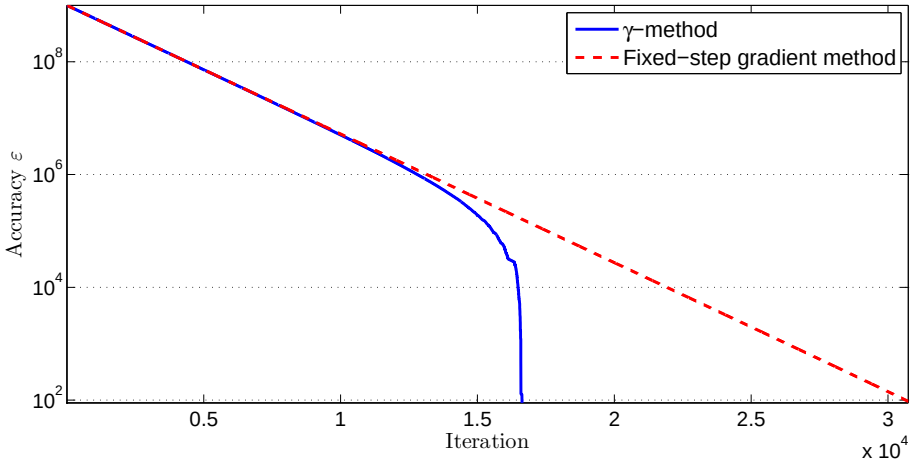

Figure 7.1: Comparison of the convergence between the gradient method and the γ-method

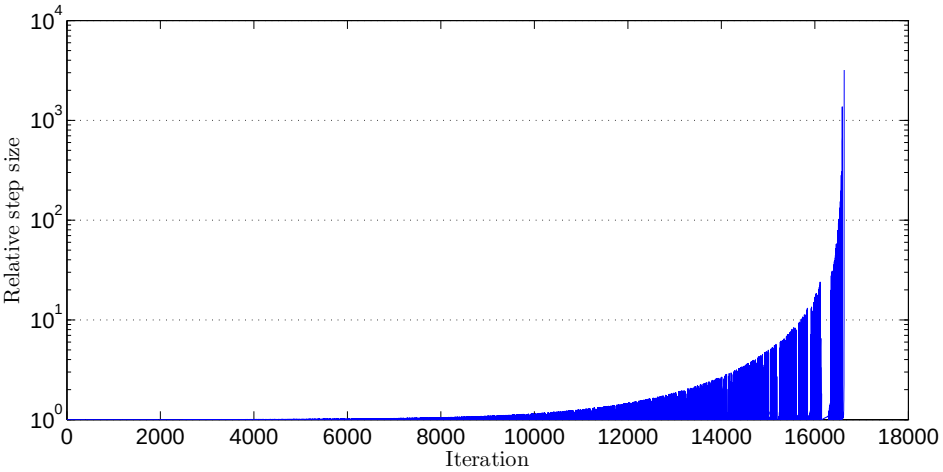

Figure 7.2: Relative step size of the γ-method compared to the fixed step gradient method.

The main drawback of this scheme is minimizing $m_\gamma(y)$: computing the next iterate x_{k+1} is harder than a simple gradient step, but as we have seen before, the gradient step can be very close to the γ-method step: we can thus use it as a good initial point for an iterative method which minimizes $m_\gamma(y)$. We will thus see two variants of this method where the next iterate is simpler to compute.

7.3 The γ-method with non-optimal direction

We have seen before that minimizing $m_\gamma(y)$ can be difficult. The real difficulty is in fact the computation of the optimal direction. But when the direction is fixed the problem is much easier, and we can have an explicit formula for the optimal norm associated to a direction.

Let us write the step $s = \alpha u$, where $\alpha = \pm\|s\|$ and $u = \pm\frac{s}{\|s\|}$. In this case both s and αu are equivalent. Suppose that α is very big such that $\gamma_x(x + \alpha u) < 1$. Indeed,

$$f(x+s) \leq f(x) + \alpha \langle f'(x), u\rangle + \frac{\sigma}{2}\alpha^2 + \alpha^3 L \left(\frac{\langle f''(x)u, u\rangle^3}{6L^3\alpha^3} - \frac{\langle f''(x)u, u\rangle^2}{2L^2\alpha^2} \right).$$

Suppose now that u is fixed. The first optimality condition is

$$\langle f'(x), u\rangle + \sigma\alpha - \frac{\langle f''(x)u, u\rangle^2}{2L} = 0.$$

We can thus find the optimal norm of a given direction:

$$\alpha^* = \frac{1}{\sigma}\left(\frac{\langle f''(x)u, u\rangle^2}{2L} - \langle f'(x), u\rangle \right). \tag{7.2}$$

We thus see that the optimal norm is given by something proportional to the norm of the gradient plus a flat amount which depends of a value defined by the matrix $f''(x)$.

We can apply this formula with the gradient direction. The direction is $u = -\frac{f'(x)}{\|f'(x)\|}$ while the optimal norm is

$$\alpha^* = \frac{1}{\sigma}\left(\frac{\langle f''(x)f'(x), f'(x)\rangle^2}{2L\|f'(x)\|^4} + \|f'(x)\| \right).$$

If write the full step,

$$\alpha^* u = -\frac{1}{\sigma}\left(\frac{\langle f''(x)f'(x), f'(x)\rangle^2}{2L\|f'(x)\|^5} + 1 \right) f'(x),$$

we can easily see that the improved step is in fact the old step $-\frac{1}{\sigma}f'(x)$ plus a "constant" which depends only of the direction of the steepest descend. Therefore, when we are far from the optimum, $\|f'(x)\|$ will be very big, and the optimal norm will tend to the norm of the gradient step.

7.4 Minimizing the relaxation of the γ-model: the θ-method

We have seen before that minimizing the γ-model is hard, and we have also seen that when we fixed a direction then we can compute easily the associated optimal norm. We will thus see here a new method which consists in minimizing a model which is very similar to the γ-model, the

θ-model (see (4.8)):

$$m_\theta(y) = f(x)+\langle f'(x), y-x\rangle+\frac{1}{2}\langle f''(x)(y-x), y-x\rangle+L\|y-x\|^3\left(\frac{\theta_x(y)(1-\theta_x(y))}{2}+\frac{\theta_x(y)^3}{6}\right).$$

Suppose now that $\theta_x(y) < 1$. The new iterate x_{k+1} is thus equal to

$$x_{k+1} = \arg\min_y \left[\langle f'(x), y-x\rangle + \frac{1}{2}\langle f''(x)(y-x), y-x\rangle + L\|y-x\|^3\left(\frac{\theta_x(y)(1-\theta_x(y))}{2}\right)\right].$$

because $\theta^3(y)\|y-x\|^3$ is in fact constant, like $f(x)$. Also, the model need 3 parameters in theory, but in practice we just need to estimate L and $(\sigma-\mu)$, so we do not have more parameters than before. Note that the θ-model is convex by construction.

7.4.1 Minimizing the model

We will now use another model which implies $\theta_x(y)$ instead of $\gamma_x(y)$:

$$\theta_x(y) = \min\left\{1\ ;\ \frac{\sigma-\mu}{L\|y-x\|}\right\} \qquad ; \qquad \gamma_x(y) = \min\left\{1; \frac{\langle[\sigma I - f''(x)](y-x), y-x\rangle}{L\|y-x\|^3}\right\}.$$

The difference between the two is the lower-approximation of $f''(x)$ by μI. The main advantage is to avoid a dependence with the direction $y-x$ and the value of $\theta_x(y)$. By consequence the first-order condition when $\theta_x(y) < 1$ is much simpler than for the γ-method:

$$f'(x) + f''(x)(y-x) + (\sigma-\mu)(y-x) - \frac{(\sigma-\mu)^2}{2L}\frac{(y-x)}{\|y-x\|} = 0. \tag{7.3}$$

Now, denote $y-x$ by αu where $\|u\| = 1$. We have thus

$$f'(x) + \alpha f''(x)u + \alpha(\sigma-\mu)u - \frac{(\sigma-\mu)^2}{2L}u = 0. \tag{7.4}$$

Let us write u in function of α:

$$u = -\left(\alpha\left[f''(x) + (\sigma-\mu)I\right] - \frac{(\sigma-\mu)^2}{2L}I\right)^{-1} f'(x), \quad \|u\| = 1, \quad \alpha > \frac{\sigma-\mu}{L}\ (\Leftrightarrow \theta_x(x+\alpha u) < 1).$$

We thus need to find the right α such that the norm of u is equal to one. We can find the solution to this equation with a binary search algorithm (where at each step we solve a linear system of n variables). We will prove that the norm of the right-hand-side of the equation decreases when α becomes larger. Let

$$A = f''(x) + (\sigma-\mu)I, \qquad cI = \frac{(\sigma-\mu)^2}{2L}I.$$

Let us use the SVD algorithm on A. Since A is square and symmetric, the SVD of A is $U\Sigma U^T$,

where $U^TU = UU^T = I$ and Σ is diagonal. Indeed,

$$u = -\left(\alpha U\Sigma U^T - cUU^T\right)^{-1} f'(x) = -U\left(\alpha\Sigma - c\right)^{-1} U^T f'(x).$$

Let us call $v = U^T f'(x)$. We have thus

$$\|u\|^2 = \|(\alpha\Sigma - c)^{-1}v\|^2 = \sum_{i=1}^{n}\left(\frac{v_i}{\alpha\sigma_i - c}\right)^2.$$

It is now clear that $\|u\|$ decreases when we increase α. We can thus use the binary search algorithm to solve the intermediate minimization problem (or any variant, like the secant method, which is faster): If the norm of u is below one, then we need to increase α, and if the norm of u is too large then we need to decrease α. Let us now prove that, for all i, we have

$$\alpha\sigma_i - c > 0. \tag{7.5}$$

Let us first have a lower bound on σ_i, the singular values of A. Indeed,

$$A = f''(x) + (\sigma - \mu)I \succeq \sigma I \Rightarrow \sigma_i \geq \sigma.$$

Since we forced α to be larger than $\frac{\sigma-\mu}{L}$, the product $\alpha\sigma_i$ is bounded as follow:

$$\alpha\sigma_i \geq \sigma\frac{\sigma-\mu}{L} \quad \Rightarrow \quad \frac{\alpha\sigma_i}{c} \geq \frac{2\sigma}{\sigma-\mu} > 2,$$

leading to the fact that (7.5) is true.

We have already a lower-bound for α, but we need also an upper bound α_{ub} in order to run the binary-search. We will now compute an upper-bound using (7.4). Indeed,

$$\alpha\|f''(x)u + (\sigma-\mu)u\| \leq \|f'(x)\| + \frac{(\sigma-\mu)^2}{2L}.$$

Since $f''(x) \succeq \mu I$, we have

$$\alpha \leq \frac{1}{\sigma}\left(\|f'(x)\| + \frac{(\sigma-\mu)^2}{2L}\right).$$

We can thus now write explicitly our specific binary-search algorithm to find the minimum of the $\theta-$model.

Initialization : Let $\alpha_{lb} = \frac{\sigma-\mu}{L}$ and $\alpha_{ub} = \frac{1}{\sigma}\left(\|f'(x)\| + \frac{(\sigma-\mu)^2}{2L}\right)$.
While $\alpha_{ub} - \alpha_{lb} >$**tol:**

1. Compute $\alpha_{new} = \frac{\alpha_{ub}-\alpha_{lb}}{2}$, and $u = -\left(\alpha\left[f''(x) + (\sigma-\mu)I\right] - \frac{(\sigma-\mu)^2}{2L}I\right)^{-1} f'(x)$.
2. If $\|u\| > 1$, then $\alpha_{lb} = \alpha_{new}$. Else, $\alpha_{ub} = \alpha_{new}$.

Note that in this section we have considered the binary search algorithm, but the secant method (or the Newton's method) can be also used for better performances.

7.4.2 Complexity analysis

We will now analyze the complexity of the θ-method. There is no *a priori* reason of having a method which is better than the gradient method, so we will compare at the end the maximum number of iterations of the two algorithms. We will suppose that we are far from the solution, so that $\theta(y) < 1$. Let us take (7.3) and call y_θ the solution of this equation. Indeed, if we use (4.7),

$$\begin{aligned}
\|f'(y_\theta)\| &= \|f'(y_\theta) - f'(x) - \langle f''(x), y_\theta - x\rangle - \left(1 - \frac{\theta_x(y_\theta)}{2}\right)(\sigma-\mu)(y_\theta - x)\| \\
&\leq \|f'(y_\theta) - f'(x) - \langle f''(x), y_\theta - x\rangle\| + \left(1 - \frac{\theta_x(y_\theta)}{2}\right)(\sigma-\mu)\|y_\theta - x\| \\
&\leq L\|y_\theta - x\|^2 \left(\theta_x(y_\theta) - \frac{\theta_x^2(y_\theta)}{2}\right) + \left(1 - \frac{\theta_x(y_\theta)}{2}\right)(\sigma-\mu)\|y_\theta - x\| \\
&= L\|y_\theta - x\|^2 \left(2\theta_x(y_\theta) - \theta_x^2(y_\theta)\right).
\end{aligned}$$

If we replace $\theta_x(y_\theta)$ in the last inequality, we can deduce

$$\frac{\|f'(y_\theta)\|}{2(\sigma-\mu)} + \frac{\sigma-\mu}{2L} \leq \|y_\theta - x\|. \tag{7.6}$$

Also, we can use one more time (7.3) multiplied by $y_\theta - x$:

$$0 = \langle f'(x), y_\theta - x\rangle + \langle f''(x)(y_\theta - x), y_\theta - x\rangle + \left(\theta_x(y) - \frac{\theta_x^2(y)}{2}\right) L\|y - x\|^3. \tag{7.7}$$

This result can be injected in (4.8) by replacing $\langle f'(x), y_\theta - x\rangle$. Therefore,

$$f(x) - f(y_\theta) \geq \frac{1}{2}\langle f''(x)(y_\theta - x), y_\theta - x\rangle + L\|y_\theta - x\|^3 \left(\frac{\theta_x(y_\theta)}{2} - \frac{\theta_x^3(y_\theta)}{6}\right). \tag{7.8}$$

Since for all $z \in [0,1]$ we have $\frac{z}{2} - \frac{z^3}{6} \geq \frac{1}{3}z$, and because $f''(x)$ is strongly convex, we can transform the inequality (7.8) into something very similar to (3.12).

$$f(x) - f(y_\theta) \geq \frac{\mu}{2}\|y_\theta - x\|^2 + \frac{\theta_x(y_\theta)L}{3}\|y_\theta - x\|^3 = \left(\frac{\sigma}{3} + \frac{\mu}{6}\right)\|y_\theta - x\|^2.$$

We can now use (7.6) on this inequality to get

$$f(x) - f(y_\theta) \geq \left(\frac{\sigma}{3} + \frac{\mu}{6}\right)\left(\frac{\|f'(y_\theta)\|}{2(\sigma-\mu)} + \frac{\sigma-\mu}{2L}\right)^2. \tag{7.9}$$

We can see in this relation two very interesting things. The first one is the constant term: at

each iteration, the model will decrease of at least

$$\left(\frac{\sigma}{3}+\frac{\mu}{6}\right)\left(\frac{\sigma-\mu}{2L}\right)^2.$$

It is due to the fact that we need to be far enough from the optimal point. The second thing is that the decrease is proportional to the square of $\|f'(x)\|$, like in the gradient method. We can thus expect a linear rate of convergence for this method.

Now, we will drop the constant term in (7.9), and we will replace $\|f'(x)\|$ using (2.4). Therefore,

$$\delta_k - \delta_{k+1} \geq \frac{\mu}{2(\sigma-\mu)^2}\left(\frac{\sigma}{3}+\frac{\mu}{6}\right)\delta_{k+1}.$$

For more convenience we will call G the constant in the right hand side of the above inequality. We can now deduce the rate of convergence of the method:

$$\delta_k \leq \frac{1}{1+G}\delta_{k-1} \leq (1+G)^{-k}\delta_0.$$

Suppose we want to reach $\varepsilon > 0$ not too small, the following condition is thus sufficient

$$e^{-kG}\delta_0 \leq \varepsilon.$$

The number of iterations $k_{\max}$ is finally bounded by

$$k_{\max} \leq G \log \frac{\delta_0}{\varepsilon}.$$

We have here, as expected, a linear rate of convergence. This method can thus be used in the first phase of the minimization process.

7.4.3 Comparison with the quadratic model

In this section we will compare the θ-model with the quadratic model used by the gradient method. We are interested in a condition on y or the norm of $y-x$ such that $m_\theta(y) \leq m_2(y)$ when $\theta_x(y) < 1$. After the expansion of $\theta_x(y)$ we find

$$\frac{1}{2}\langle [f''(x)-\mu](y-x), y-x\rangle - \frac{(\sigma-\mu)^2}{2L}\|y-x\| + \frac{(\sigma-\mu)^3}{6L^2} \leq 0.$$

Suppose that $y-x=\alpha u$ and the direction u is fixed. Let us call $\mathcal{E} = \frac{\langle [f''(x)-\mu]u,u\rangle}{\sigma-\mu}$ the relative error of the approximation of $f''(x)$ by μI and $\Delta = \frac{\sigma-\mu}{L}$. Then we can find a condition on α :

$$\alpha \leq \frac{\Delta}{\mathcal{E}}\left(1+\sqrt{1-\frac{4}{3}\mathcal{E}}\right).$$

If the error $\mathcal{E}$ is too big, then we do not have a solution to the second order equation, meaning

that the θ-model will be in this case strictly above the quadratic model. To avoid this fact, we should have μ as big as possible and σ close to μ.

This condition also tells us that the θ-method is worse that the gradient method when we are very far from the optimum. The method becomes efficient more or less at the same time when the γ-method becomes much faster than the gradient. However, the θ-method can be used to compute another direction, different from the gradient, which can be used in combination with the $\gamma-$method with fixed direction.

7.5 Conclusion

In this chapter, we have seen several methods, with their own benefits and drawbacks. The common and main problem of all previous methods is the presence of two parameters (L and σ or L and $(\sigma - \mu)$) in the model needed to compute the next iterate.

The $\gamma-$method is the best algorithm in term of maximum number of iterations. However, the subproblem is hard to minimize because its structure is quite complex. However, when the direction is fixed, everything becomes much easier and we can have the expression of the optimal norm (i.e. the value of the norm which minimizes the γ-model) associated to any direction.

We have also analysed the $\theta-$method, much easier to compute. This scheme is finally not as good as expected, because of its only local efficiency (when we are far, the gradient method is better, and when we are too close we have to switch to the CNM method). The scheme is more useful when used to compute a direction, because we can adapt the norm with the γ-model. However in this case we will need 3 parameters, because σ appears alone in the γ-model.

Finally, assuming that all parameters are known, the best method which makes a compromise between the difficulty of the resolution of the sub-problem and the total complexity should be an hybrid method using the expression of the optimal norm, which choose the best step between the direction given by the gradient or by the θ-method. This new hybrid scheme is by construction more efficient than the old one but slower than the γ-method in term of number of iterations. For all previous methods the number of parameters can be a real difficulty when used in an adaptive algorithm. We will thus see in the next section a new algorithm which will use only one parameter.

Chapter 8

Using line-search on the parameter L: adaptive CNM

8.1 Motivations

In the previous sections we have seen methods which need more than one parameter, like the hybrid method of the γ-method. However, the two algorithms works at least as good as the gradient. They are all based on a local cubic and global quadratic upper-approximation of the function $f \in \mathcal{S}_{\sigma,L}$.

We have seen that the CNM works bad on $f \in \mathcal{S}^{\mu}_{\sigma,L}$ because the method is too conservative. We will thus now change the strategy for this class: we will use only the cubic model (2.12) used by the CNM but this time we will try to find a smaller L at each iteration. The consequence of this choice is that the model will not be a global upper-approximation anymore, but we will try to keep the inequality at x_{k+1} to ensure a certain decrease.

8.2 Intuition: A smaller L for a larger step size

The main reason of using a line-search on the parameter L instead of the norm of the step $(y-x)$ is because the direction will also change, leading to a better theoretical decreasing. However by changing L, the norm of the step will also be modified.

We will see here the link between the parameter L and the norm of the step $(y-x)$. Indeed, by using the cubic model (3.2), we can deduce the step $(y-x)$ by an implicit way. Let us denote $s = T - x$ and $r = \|T - x\|$. We have thus

$$s = -\left(f''(x) + \frac{Lr}{2} I\right)^{-1} f'(x).$$

We *cannot* say that Lr is a constant, because we have to follow $\|s\| = r$. We will thus one more time use the SVD of $f''(x) = U\Sigma U^T$ because $f''(x)$ is symmetric. We can thus rewrite the step

$$s = -\left(U\Sigma U^T + UU^T \frac{Lr}{2}\right)^{-1} f'(x) = -U\left(\Sigma + \frac{Lr}{2} I\right)^{-1} U^T f'(x).$$

Let us call $U^T f'(x) = v$. We have thus

$$\|s\|^2 = \sum_{i=1}^{n} \left(\frac{v_i}{\sigma_i + \frac{L}{2} r} \right)^2 .$$

Finally, using the fact that $\|s\|^2 = r^2$,

$$1 = \sum_{i=1}^{n} \left(\frac{v_i}{\sigma_i r + \frac{L}{2} r^2} \right)^2 .$$

We can see with this equation that a decrease of L must be compensated by an increase of the norm of the step s. We can also see that the impact of the decrease of L is more important when the singular values of $f''(x)$ are small.

8.3 The line-search algorithm

In this section we will present the line-search algorithm itself. Our goal is to find one good value $\tilde{L}$ under the condition

$$f(T_{\tilde{L}}(x)) \leq m_3(x; \tilde{L})$$

where $m_3(x; \tilde{L})$ is the cubic model used with another value of L. The subroutine is described below.

Subroutine: Line search on parameter L
Goal: Find a good value for parameter L, called $\tilde{L}$.
Initialization : Choose $L_0 = L$ and $i = 0$.
Iteration :

1. Set $i = i + 1$ and $L_i = 2^{-i} L$.
2. Compute $T_i = T_{L_i}(x)$.
3. *If* $[f(T_i) \leq m_3(x; L_i)]$ go to step 1.
   Else, go to step 4.
4. Return $\tilde{L} = L_{i-1}$.

The goal of this algorithm is to divide at each step the old value of the parameter by two, until $f(T_{\tilde{L}/2}) \geq m_3(x; \tilde{L}/2)$. We are now sure that $\tilde{L}$ is quite small and we can thus deduce the rate of convergence of the algorithm by analyzing the decrease of the modified cubic model.

We need to be careful when we use this algorithm: such a value of L does not always exists. For example, suppose that for some point x we have $f''(x) = \sigma I$, and for any other points y we have $f''(y) \prec f''(x)$. In this case, according to inequality (4.4), we will thus have

$$f(y) \leq f(x) + \langle f'(x), y - x \rangle + \frac{1}{2} \langle f''(x)(y - x), y - x \rangle .$$

One way to avoid this case is to test the "cubic" model with $L = 0$ (indeed, in this case it becomes quadratic) before running the line-search subroutine.

However, such a value of L can be also as close as possible to zero. In this case the line search will be very slow. The solution is to set a lower bound on L. For example, we can stop the algorithm when for some i the decrease of the cubic model with L_i is half a time the decrease of the classical Newton model:

$$m_3(x; L_i) \leq f(x) - \frac{1}{2}\left\langle f'(x), f''(x)^{-1} f'(x)\right\rangle$$

We can also fix an arbitrary L_{lb} in advance, and stop when $L_i \leq L_{lb}$.

8.4 Complexity analysis

We will now analyze the complexity of the line-search algorithm. Suppose that we have fixed L_{lb} in advance, and we start with the value L. Then in this case the complexity of the line-search subroutine is

$$\mathcal{O}\left(\log_2 \frac{L}{L_{lb}}\right). \tag{8.1}$$

Let us now estimate the number of iterations of the main algorithm. We will compute the biggest possible value for $\tilde{L}$, denoted by L_{ub}. Let us fix some x. Indeed, since for any y we have $f(y) \leq m_\gamma(y)$, L_{ub} must follow

$$\max L_{ub} \; s.t. \qquad m_3(y, L_{ub}) \geq m_\gamma(y) \qquad \text{and} \qquad m_3(y, L_{ub}/2) \leq m_\gamma(y).$$

By consequence, we see that L_{ub} satisfies $m_3(y, L_{ub}/2) = m_\gamma(y)$. If we develop the two expressions we find:

$$\frac{L_{ub}}{2}\|y - x\|^3 = \frac{L}{6}\|y - x\|^3 \left(\frac{\gamma_x(y)(1 - \gamma_x(y))}{2} + \frac{\gamma_x(y)}{6}\right).$$

Leading to

$$L_{ub} = \frac{2L}{6}\left(\frac{\gamma_x(y)(1 - \gamma_x(y))}{2} + \frac{\gamma_x^3(y)}{6}\right).$$

We can thus conclude that

$$f(T_{\tilde{L}}) \leq m_3(T_{\tilde{L}}, \tilde{L}) \leq m_3(T_{L_{ub}}, L_{ub}) \leq m_\gamma(T_{2L}; 2L).$$

This result is very surprising: with a line-search on the parameter L we are able to reach an accuracy comparable to the γ-model with a small error on the parameter L. It means that the total complexity of our algorithm, when applied on $\mathcal{S}^{\mu}_{\sigma,L}$ is simply (8.1) times (6.9) with L

multiplied by two:

$$\mathcal{O}\left(\left(\log_2 \frac{L_{lb}}{L}\right)\left[\frac{\sigma}{\mu}\log\left(\frac{4L^2}{\sigma^3}\delta_0\right)+\log_2\left(\log\frac{1}{\varepsilon}\right)\right]\right).$$

8.5 Discussion

We have seen here an adaptive algorithm which is an improvement of the CNM: with the line-search subroutine, we are able to achieve an accuracy almost as good as the γ-model, at the price of an additional factor $\log_2 \frac{L_{lb}}{L}$. With this little trick we have a linear rate of convergence at the first phase of the minimization process instead of $\mathcal{O}\left(\sqrt{\frac{LD}{\mu}}\right)$ iterations.

We have proposed here the strategy of dividing by two the value of L at each iteration of the line-search. However, one can try another type of algorithm like a secant method, in order to try to minimize the complexity of the line-search. In the same idea, we can decrease a little bit the number of iterations of the sub-problem by knowing a better upper-bound for L_0 or L_{lb} if we have an information about σ. Suppose that σ is know, then $\tilde{L}$ should be

$$\tilde{L} \approx \frac{2L}{6}\left(\frac{\gamma_x(x+s)(1-\gamma_x(x+s))}{2}+\frac{\gamma_x^3(x+s)}{6}\right)$$

for s the gradient step $-\frac{1}{\sigma}f'(x)$ (or the improved gradient step using the γ-model). It comes from the fact that when we are far from the optimum the CNM step is similar to the gradient step. The subroutine can by consequence be improved with some various techniques but the complexity in $\mathcal{O}\left(\log_2(L_{lb}/L)\right)$ is already quite small.

Conclusion

The second order methods were often used at the termination stage of the minimization process because of their quadratic convergence. In this master thesis we tried to analyse the global complexity of such methods and answer to the following question, presented in [2]: *"For the problem class $\mathcal{S}^{\mu}_{\sigma,L}$, can we get any advantages from the second order schemes being used at the initial stage of minimization process?"*.

We first noticed that the CNM converges very slowly on this class of functions because the steps were too conservatives, caused by the global cubic model. Once this problem was identified, we tried to improve the global performance of the scheme with a switching strategy: the upper-estimation became locally cubic but globally quadratic. With this trick the other methods became much more efficient and had a complexity comparable to the one of the gradient method when used at the first stage of the minimization process.

However, the answer to the above question is quite mitigated. Assuming that we know all constants, at the very beginning of the algorithm we have seen that our methods are comparable to a simple gradient step but the computation of a second-order step is at least as hard as computing the gradient step. So there is *a priori* no reason to believe that second-order schemes are much more efficient than first-order methods on the class $\mathcal{S}^{\mu}_{\sigma,L}$.

Nevertheless in this work we have only analysed simple steps methods. But it is well known that with a multi-step strategy, like the fast gradient method presented in [4], we are able to improve the complexity bound. If we want to know if the above methods are very comparable to the first-order algorithms, then we need to derive their accelerated version and compare it with the complexity of the fast-gradient.

Also, all of the algorithms presented in this thesis work only on unconstrained domains. For many practical problems we have some constrains to handle (for example, let us say that $x \in Q$ for Q a convex set). The way to avoid this problem is to choose the minimum of the sub-problem on Q. It is shown in [3] that we can find in this case $T_L(x)$ efficiently. One improvement of all schemes presented here can be to make a variant which works on constrained domains.

Finally, we also took care of the number of parameters present in all the methods presented here. Note that for the CNM an adaptive algorithm is presented in [1], where the parameter L is estimated at each iteration, and the Hessian $f''(x)$ is approximated. A natural improvement can be an adaptive algorithm which tries to estimate both L and σ.

Bibliography

[1] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. "Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results". In: *Mathematical Programming* 127.2 (2011), pp. 245–295.

[2] Yurii Nesterov. "Accelerating the cubic regularization of Newton's method on convex problems". In: *Mathematical Programming* 112.1 (2008), pp. 159–181.

[3] Yurii Nesterov. "Cubic regularization of Newton's method for convex problems with constraints". In: *Available at SSRN 921825* (2006).

[4] Yurii Nesterov. *Introductory lectures on convex optimization*. Vol. 87. Springer Science & Business Media, 2004.

[5] Yurii Nesterov, Boris Polyak, et al. *Cubic regularization of a Newton scheme and its global performance*. Universite catholique de Louvain, 2003.