Convergence and Complexity in nonlinear Optimization

Ana Friedlander and José Mario Martínez

www.ime.unicamp.br/∼martinez

Universidade Estadual de Campinas, Brasil

2017

Outline

The Optimization Problem

$$\text{Minimize } f(x) \text{ subject to } x \in S,$$

where

$$f : \mathbb{R}^n \to \mathbb{R} \text{ and } S \subset \mathbb{R}^n.$$

S is called *feasible set* and f is the *objective function*

Minimizers

$x^* \in S$ is a *global minimizer* if $f(x^*) \leq f(x)$ for all $x \in S$. If $f(x^*) < f(x)$ for all $x \in S$ such that $x \neq x^*$ we say that x^* is a *strict* global minimizer.
$x^* \in S$ is a *local minimizer* if there exists $\varepsilon > 0$ such that $f(x^*) \leq f(x)$ for all $x \in S$ such that $\|x - x^{\|} \leq \varepsilon$. If $f(x^*) < f(x)$ for all $x \in S$ such that $0 < \|x - x^*\| \leq \varepsilon$ we say that x^* is a *strict* local minimizer.

Theorem of Weierstrass

Every continuous function defined on a compact (closed and bounded) set $S \subset \mathbb{R}^n$ admits a global minimizer and a global maximizer on S.

Unconstrained Optimization

If the feasible set is $\mathbb{R}^n$ we talk about *Unconstrained Optimization*. Maximizing f and Minimizing $-f$ are equivalent problems, so we always talk about one of them : *Minimization*.

First-order stationarity condition

In Unconstrained Minimization, if x^* is a local minimizer and the first-order partial derivatives exist at x^*, then : the stationarity condition

$$\nabla f(x^*) = 0.$$

Recall that $\nabla f(x)$ is the gradient of f at x, that is, the vector of first-order derivatives.
Our vectors will be always “columns" in the present text.

Second-order stationarity condition

In Unconstrained Minimization, if x^* is a local minimizer and both first and second partial derivatives are continuous at x^* we have that $\nabla f(x^*) = 0$ and

$$\nabla^2 f(x^*) \text{ is positive semidefinite .}$$

$\nabla^2 f(x)$ will always denote the Hessian matrix of second partial derivatives, which is symmetric, so its eigenvalues are all real. Positive semidefinitess of A means that $v^T A v \geq 0$ for all $v \in \mathbb{R}^n$, or, equivalently that all the eigenvalues of A are non-negative. Positive semidefinitess is also denoted $A \geq 0$.

Sufficient Second-Order conditions for local minimizers

The conditions $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \geq 0$ are *necessary*.
Sufficient conditions for local minimizer :
If $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) > 0$ (positive definite, positive eigenvalues), then x^* is a *strict local minimizer*.

Descent directions

If $x \in \mathbb{R}^n$ and $\nabla f(x) \neq 0$ we know, by the necessary first-order optimality condition, that x *is not* a local minimizer. Therefore, in every neighborhood of x there is a point z such that $f(z) < f(x)$. It is interesting to find directions d along which such points z can be computed.

Descent Direction

Assume that $f : \mathbb{R}^n \to \mathbb{R}$ admits continuous first derivatives ($f \in C^1$), $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$ such that

$$\nabla f(x)^T d < 0.$$

Then, there exists $\bar{\alpha} > 0$ such that

$$f(x + \alpha d) < f(x)$$

for all $x \in (0, \bar{\alpha}]$.

In this case we say that d is a descent direction .

Algorithmic Model based on descent directions

This is an iterative algorithm that generates points $x^k \in \mathbb{R}^n$ such that, whenever $\nabla f(x^k) \neq 0$, we will have that $f(x^{k+1}) < f(x^k)$.
So, at a typical iteration we have that $\nabla f(x^k) \neq 0$ and we proceed as follows :
Step 1 : Find $d_k \in \mathbb{R}^n$ such that $\nabla f(x^k)^T d_k < 0$.
Step 2 : (Determination of the "steplength")
Compute $\lambda_k > 0$ such that $f(x^k + \lambda_k d_k) < f(x^k)$. (This is called the "line search subproblem".)
Step 3 : Define $x^{k+1} = x^k + \lambda_k d_k$.

If, at some iteration, we compute x^k such that $\nabla f(x^k) = 0$, then x^k is a stationary point and the algorithm stops.
Most probably, the process continuous indefinitely without stopping. In this case one generates an infinite sequence $\{x^k\}$ and the following questions are relevant :

- Does the limit $\lim_{k\to\infty} x^k$ exist ?
- In the positive case, is it possible to guarantee that such limit is a solution of the problem, or at least, that it is a stationary point ?

Let us try to answer those questions.
Clearly, the algorithm generates a sequence such that $f(x^{k+1}) < f(x^k)$ for all k.
Consider the function of one variable given by $f(x) = x^2$, whose only minimizer is $x^* = 0$.
The sequence defined by

$$x^k = 1 + 1/k$$

may be generated by the algorithm because

$$f(x^{k+1}) = [1 + 1/(k+1)]^2 < [1 + 1/k]^2 = f(x^k).$$

However,

$$\lim_{k \to \infty} x^k = 1.$$

This example shows that the limit of a sequence generated by the algorithm may exist but could not be a stationary point of the problem.
The decrease between two consecutive iterations could also be very small with a big distance between x^{k+1} and $x\ k$, as shown in the following picture.

Picture with big distance and small decrease

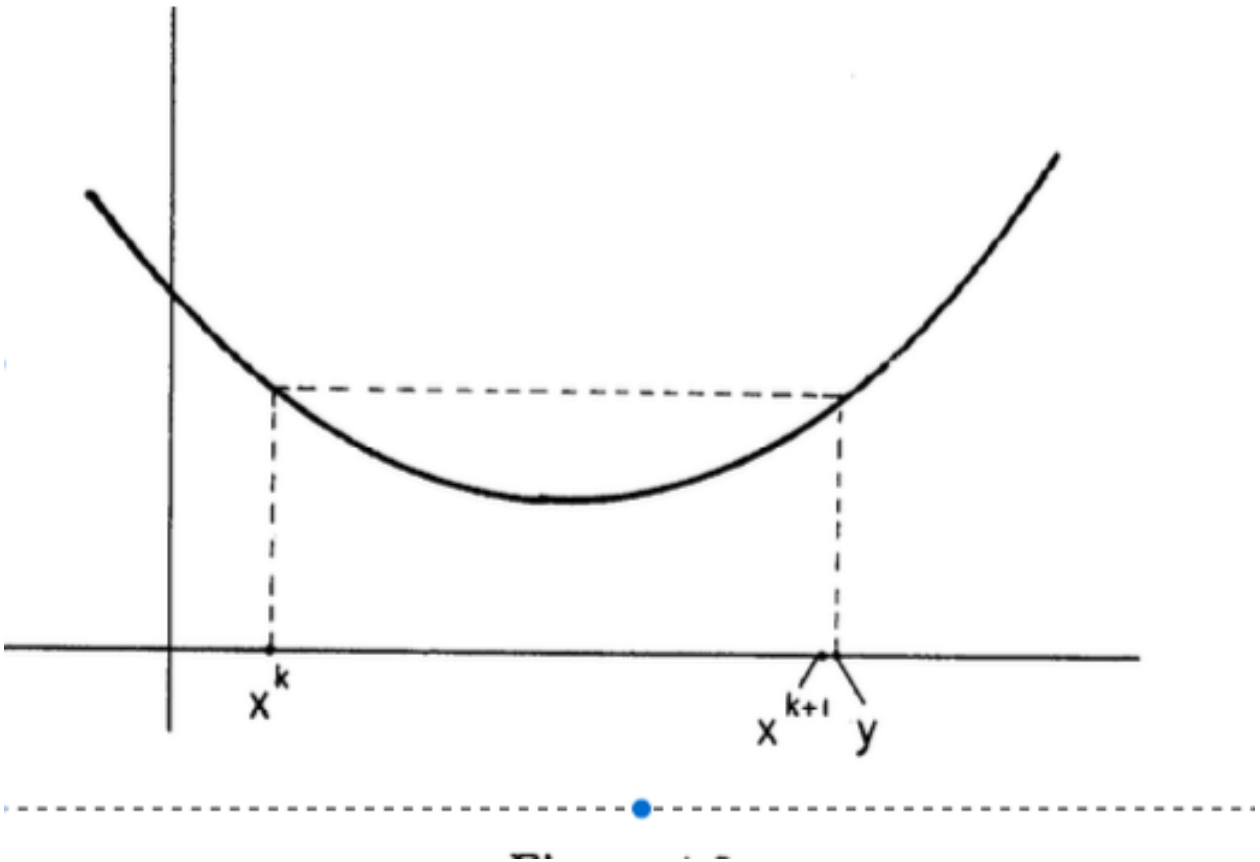


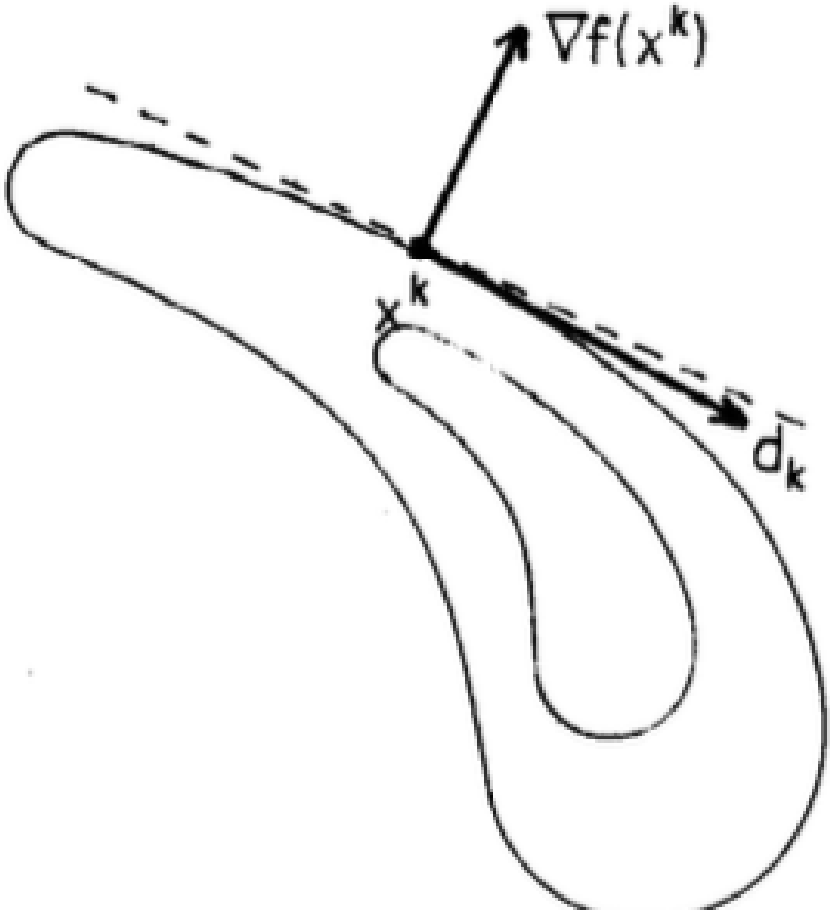
Figure – Big distance with almost null decrease

There exists a third situation in which descent directions could lead us to very small functional decrease.
In fact, consider the *level set* defined by

$$\Gamma = \{x \in \mathbb{R}^n \mid f(x) = f(x^k)\}.$$

Clearly, the decrease is null along Γ. Therefore, if the descent direction d_k is almost orthogonal to Γ, such direction would be almost tangent to Γ. In this case, the decrease along that direction could also be very small, as shown in the following picture.

Picture with small decrease along an almost tangent direction



Avoiding short directions far from the solution

Short steps far from the solution inhibit convergence to the solution, as shown in picture before.
In order to avoid such anomaly we will impose that the direction would be big enough in comparison to the gradient at the current point.

$$\|d_k\| \geq \sigma \|\nabla f(x^k)\|$$

for some $\sigma > 0$.
This requirement is easy to satisfy since we may simply multiply a trial direction by a suitable number for its fulfillment.

Avoiding almost orthogonal directions to the gradient

We have seen that, if a direction is almost tangent, that is, almost orthogonal to the gradient, the space for decrease could be very small and, so, very small decrease should be obtained at the current iteration.
In order to avoid this anomaly, we impose that the angle between d_k and $-\nabla f(x^k)$ should be bounded away from 90 degrees.
Namely, the cosine of the angle should be bounded away from zero :

$$\nabla f(x^k)^T d_k \leq -\theta \|d_k\| \|\nabla f(x^k)\|$$

with $\theta \in (0, 1)$.

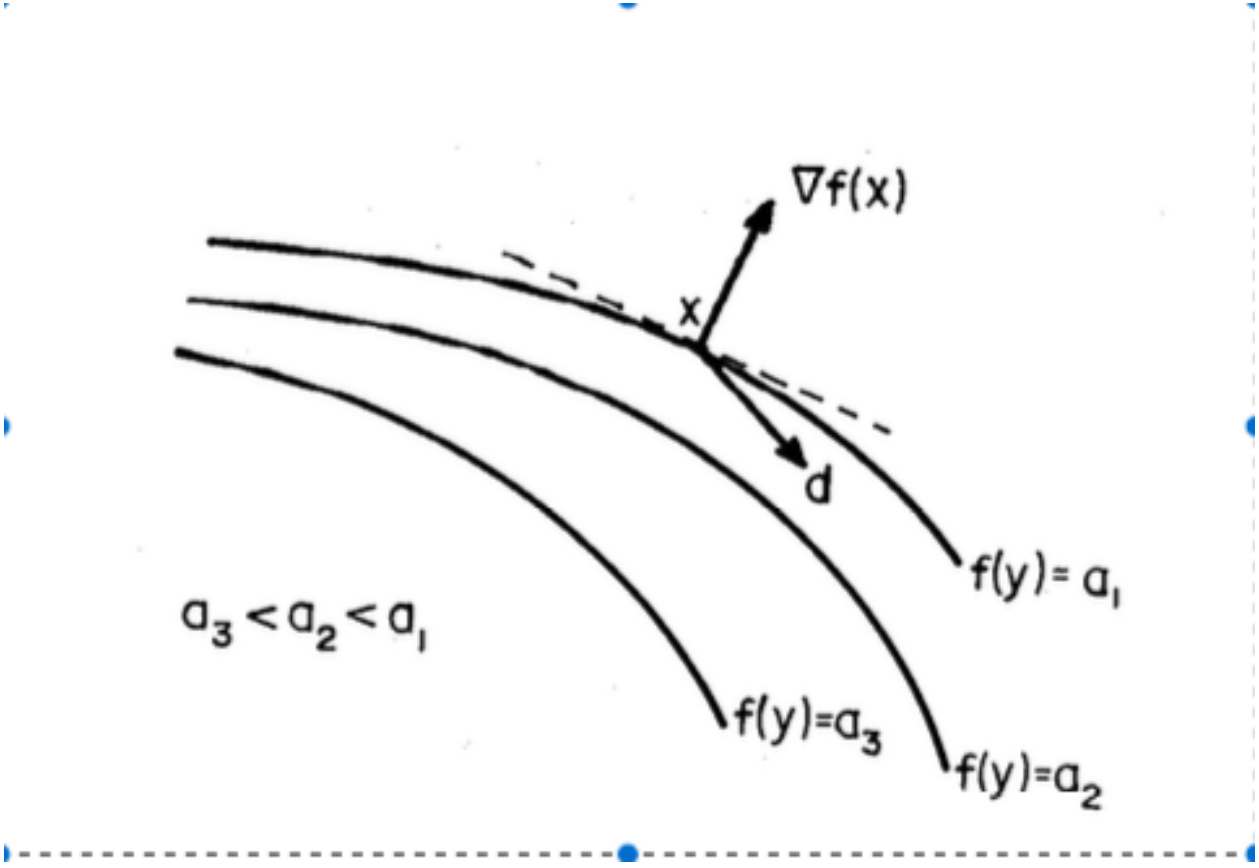


Figure – Angle Condition

Avoiding steps with negligible descent

The condition $f(x^{k+1}) < f(x^k)$ could be extremely weak. Therefore, we impose a *sufficient descent* condition (or Armijo condition) :

$$f(x^k + \lambda d_k) \leq f(x^k) + \lambda \alpha \nabla f(x^k)^T d_k$$

with $\alpha \in (0, 1/2)$.

Picture explaining Armijo

This picture shows that, for $\lambda > 0$ small enough, Armijo's condition holds.

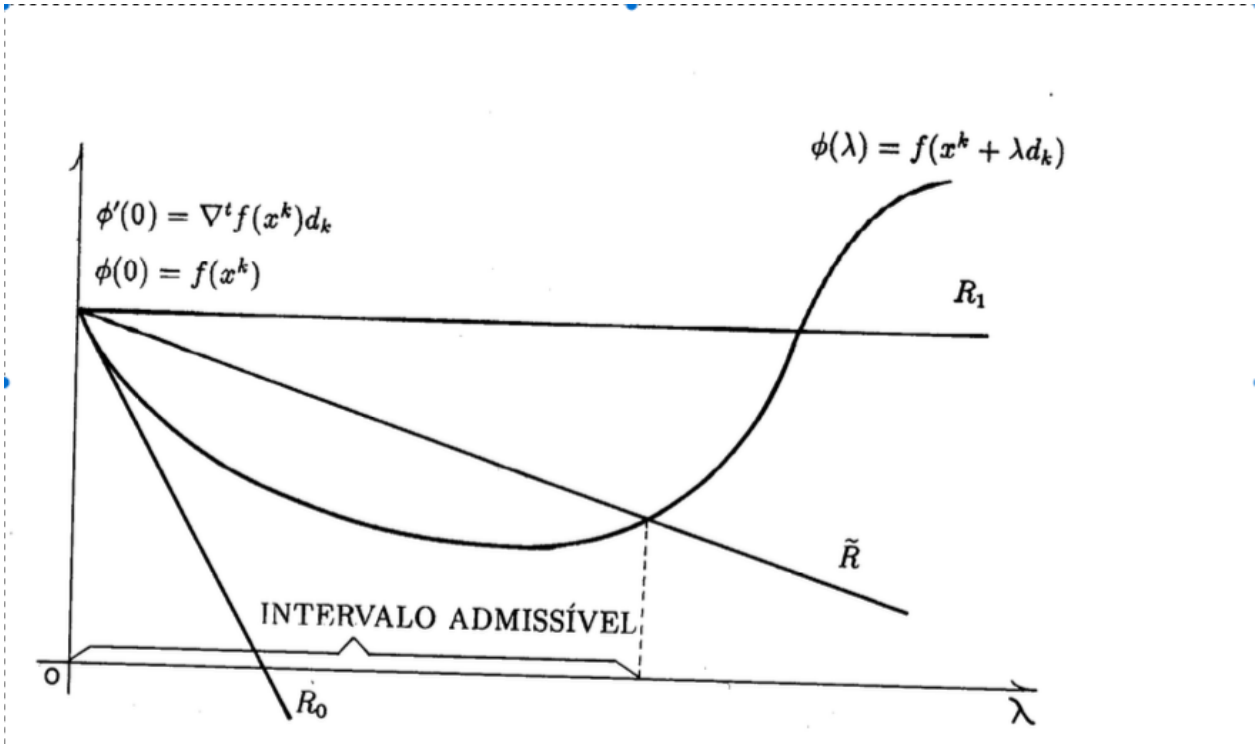


Figure – Armijo

Formal definition of the first-order descent algorithm

Let $\sigma > 0$, $\alpha \in (0, 1/2)$, and $\theta \in (0, 1)$ be given. Assume that $\nabla f(x^k) \neq 0$. Then, the steps for obtaining x^{k+1} are :

Step 1 Choose $d_k \in \mathbb{R}^n$ such that

$$\|d_k\| \geq \|\nabla f(x^k)\|$$

and

$$\nabla f(x^k)^T d_k \leq -\theta \|\nabla f(x^k)\| \|d_k\|.$$

Set $\lambda \leftarrow 1$.

Step 2 Test the Armijo condition

$$f(x^k + \lambda d^k) \leq f(x^k) + \lambda \alpha \nabla f(x^k)^T d_k.$$

Step 3 If Armijo holds, define $x^{k+1} = x^k + \lambda d_k$. Otherwise, choose $\tilde{\lambda} \in (0.1\lambda, 0.5\lambda)$, set $\lambda \leftarrow \tilde{\lambda}$ and to go Step 2.

The algorithm is well defined

This means that, given an iterate at which the gradient does not vanishes, we can find x^{k+1} in finite time.
This theorem is a consequence of the fact that d_k is a descent direction.

Global convergence theorem

The so called “global convergence theorems" in this area are theorems that say that, given an arbitrary initial point, the sequence converges, in some sense, to a stationary point (point where the gradient vanishes).
The word “global" is not related, in this case, to global minimization.
The assumption for the fulfillment of this theorem is that $\nabla f(x)$ is continuous and the theorem says that at every limit point of the sequence generated by the algorithm, the gradient is null.
Note that the theorem does not guarantee that the sequence converges. In fact, many different accumulation points (even infinitely many) may exist, but all of then will be stationary.
Note that only continuity of the gradient is assumed for the fulfillment of this theorem. This fact will be relevant when we talk about complexity later.

Classical descent methods : Steepest Descent

Gradient methods, or steepest descent methods, are methods of the type defined above in which $d_k = -\nabla f(x_k)$. Obviously, the condition that says that the norm of the direction is proportional to the gradient trivially holds and the angle condition too.
Sometimes, in steepest descent methods one employs stricter descent devices than the one described in the General Method. One possibility is to minimize along the direction d_k.

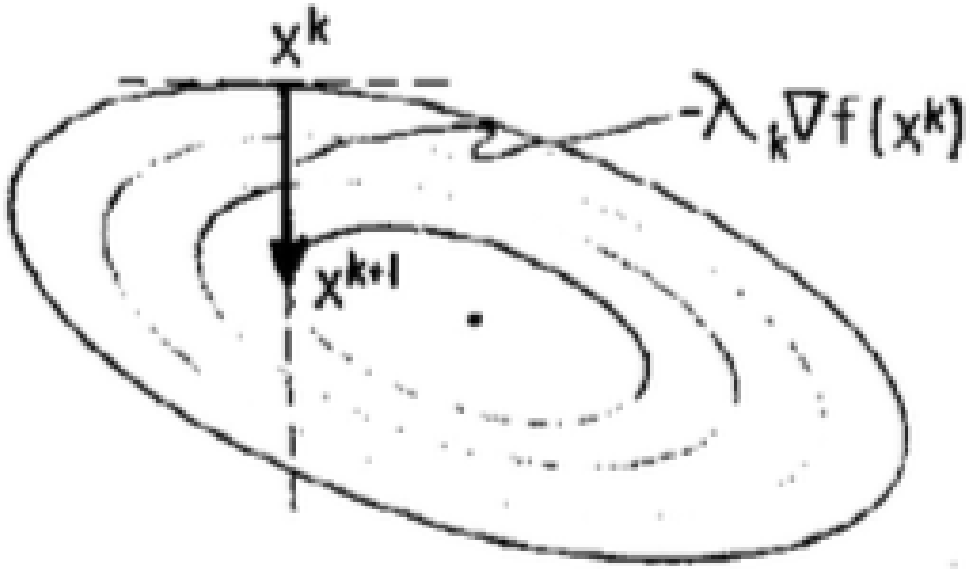


Figure – Steepest Descent

Gradient method with one-dimensional minimization

If $x^k \in \mathbb{R}^n$ is such that $\nabla f(x^k) \neq 0$, the steps to determine x^{k+1} are the following :

Step 1 Compute $d_k = -\nabla f(x^k)$

Step 2 Exact line search.

Compute λ_k, minimizer of $f(x^k + \lambda d_k)$ subject to $\lambda \geq 0$.

Step 3 Define $x^{k+1} = x^k + \lambda_k d_k$.

Note that the one-dimensional problem at Step 2 not always is well defined, since the global minimizer along the direction could not exist at all.

Steepest descent method for minimizing quadratics

If $f(x)$ is a strictly convex quadratic we have :

$$f(x) = a + b^T x + \frac{1}{2} x^T G x$$

with $G = \nabla^2 f(x)$ positive definite. In this case, $x^* = -G^{-1} b$ is the unique global minimizer of f. The application of the steepest descent method to minimize a quadratic is justified in some (large-scale) problems, although, in general, is not better than the Hestenes-Stiefel conjugate gradient method.
It can be proved that, in this case, the generated sequence converges to x^* with the following "speed on f" :

$$f(x^{k+1}) - f(x^*) \leq \left(\frac{A - a}{A + a} \right)^2 [f(x^k) - f(x^*)],$$

where $A \geq a > 0$ are the extreme eigenvalues of G.

Gradient method for locally strict convex functions

Assume that $f \in C^2$ (continuous second derivatives) and let x^* be a local minimizer of f at which the Hessian is positive definite with extreme eigenvalues $A \geq a > 0$. If the Gradient method is well defined for all k and x^k converges to x^*, then the sequence $\{f(x^k)\}$ converges to $f(x^*)$ with the speed given in the previous slide. (Linear convergence of $f(x^k)$ towards $f(x^*)$ with rate

$$r = \left(\frac{A-a}{A+a}\right)^2.$$

Newton's Method

It is easy to see that, if f is quadratic with positive definite Hessian G and x^0 is arbitrary, then the global minimizer x^* is given by

$$x^* = x^0 + d$$

where

$$d = -G^{-1}[Gx^0 + b].$$

This results from

$$\nabla f(x) = Gx + b$$

after some manipulation.

Picture of Newton for Quadratic

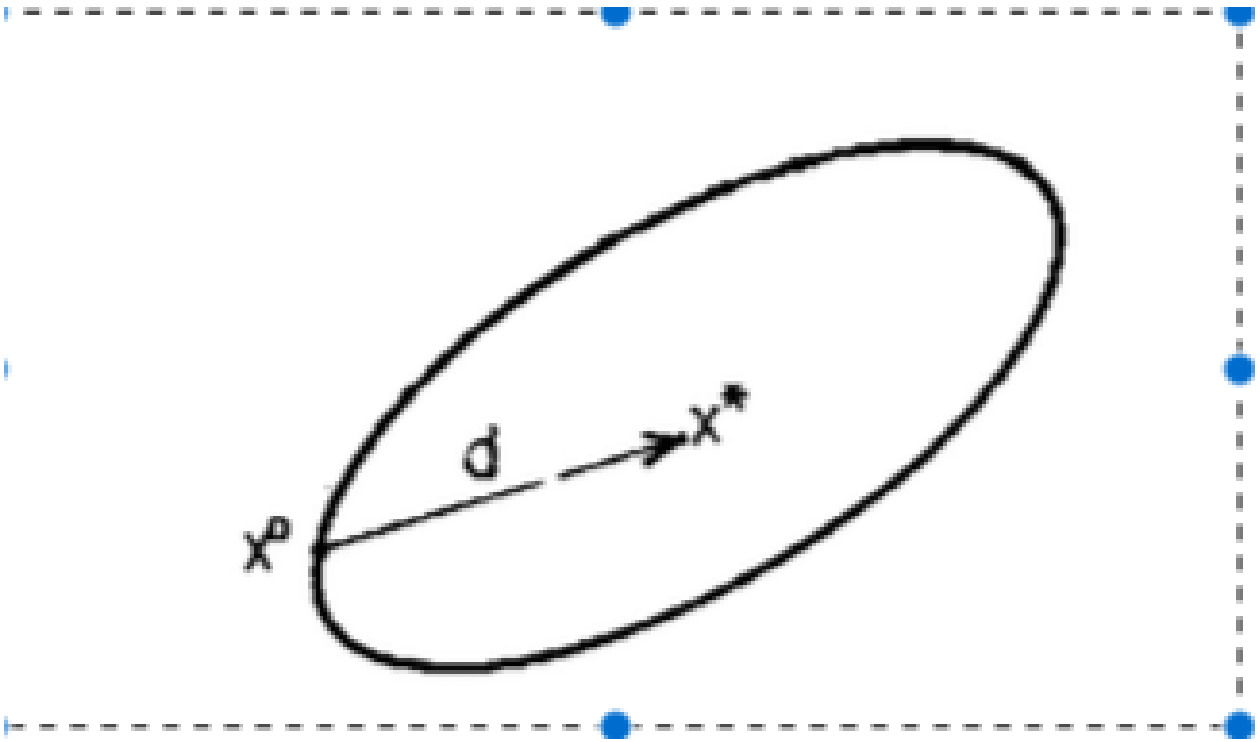


Figure – Newton for quadratics

For defining Newton's method for general (not necessarily quadratic minimization) we consider the Taylor expansion up to second-order of $f(x)$ around x^k. Namely,

$$q(d) = f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T \nabla^2 f(x^k) d.$$

Calling $c = q(0) = f(x^k)$, $b = \nabla q(0) = \nabla f(x^k)$, and $G = \nabla^2 q(0) = \nabla^2 f(x^k)$, if $\nabla^2 f(x^k)$ is positive definite, we may compute the minimizer of this quadratic starting from $d_0 = 0$, and we obtain :

$$d^* = -G^{-1}(Gd_0 + b) = -G^{-1}b = -\nabla^2 f(x^k)^{-1} \nabla f(x^k).$$

This suggests the choice of the direction

$$d_k = -\nabla^2 f(x^k)^{-1} \nabla f(x^k)$$

in the context of general unconstrained minimization.

The following questions are relevant :

- Is d_k always a descent direction ?
- In the answer is positive, are the "not-small d_k" and the angle conditions fulfilled ?

Unfortunately d_k could not be a descent direction if the Hessian is not positive definite. Take, for example, $f(x, y) = x^2 - y^2$ and $x^0 = (0, 1)$.
In this case, the Newtonian direction so far generated is not a descent direction, in fact it is an ascent direction. It could be argued that it is enough to take the opposite direction to obtain a descent direction. However, it may be possible that neither the Newton direction, nor its opposite one are descent directions. This occurs when the direction is orthogonal to the gradient.
An example of this situation is

$$f(x, y) = x^4 + xy + (y + 1)^2$$

with $x^0 = (0, 0)$. In this case, $\nabla f(x^0) = -(0, 2)$, the Newton direction is $-(2, 0)$ and $\nabla^2 f(x^0)^T d_0 = 0$.

Newton's method for general function minimization

If $\nabla f(x^k) \neq 0$, the steps for computing x^{k+1} are :
Step 1 Determine $d_k \in \mathbb{R}^n$ such that

$$\nabla^2 f(x^k) d_k = -\nabla f(x^k).$$

This means to solve this linear system. Note that this step may not be well defined if the Hessian is singular.
Step 2 Define $x^{k+1} = x^k + \lambda_k d^k$ where λ_k is determined as in the first-order methods.

Local Quadratic Convergence of Newton

Assume that f has bounded third derivatives (alternative, second derivatives satisfy a Lipschitz condition). Let x^* be a local minimizer at which the Hessian is positive definite. Then, there exists a neighborhood of x^* such that, if x^0 belongs to that neighborhood, the sequence generated by Newton is well-defined, the Hessian is positive definite at all the iterates and *quadratic convergence* towards x^* takes place. Namely, there exists $c > 0$ such that

$$\|x^{k+1} - x^*\| \leq c\|x^k - x^*\|^2$$

for all k.

Local quadratic convergence is the best property of Newton.
However, Newton needs to be modified in order to obtain global convergence.
One way for modification consists of forcing the fulfillment the requirements of the first-order global convergence theorem.
Other possibility is to employ the so called trust-region technique.
The third possibility is to use regularization, which is the main topic of the second part of this course.

Unconstrained Optimization

Problem :

$$\text{Minimize } f(x), x \in \mathbb{R}^n.$$

Notation : $g = \nabla f$, $H = \nabla^2 f$ (Hessian).
Necessary first-order condition : $g(x) = 0$. (Stationary points.)

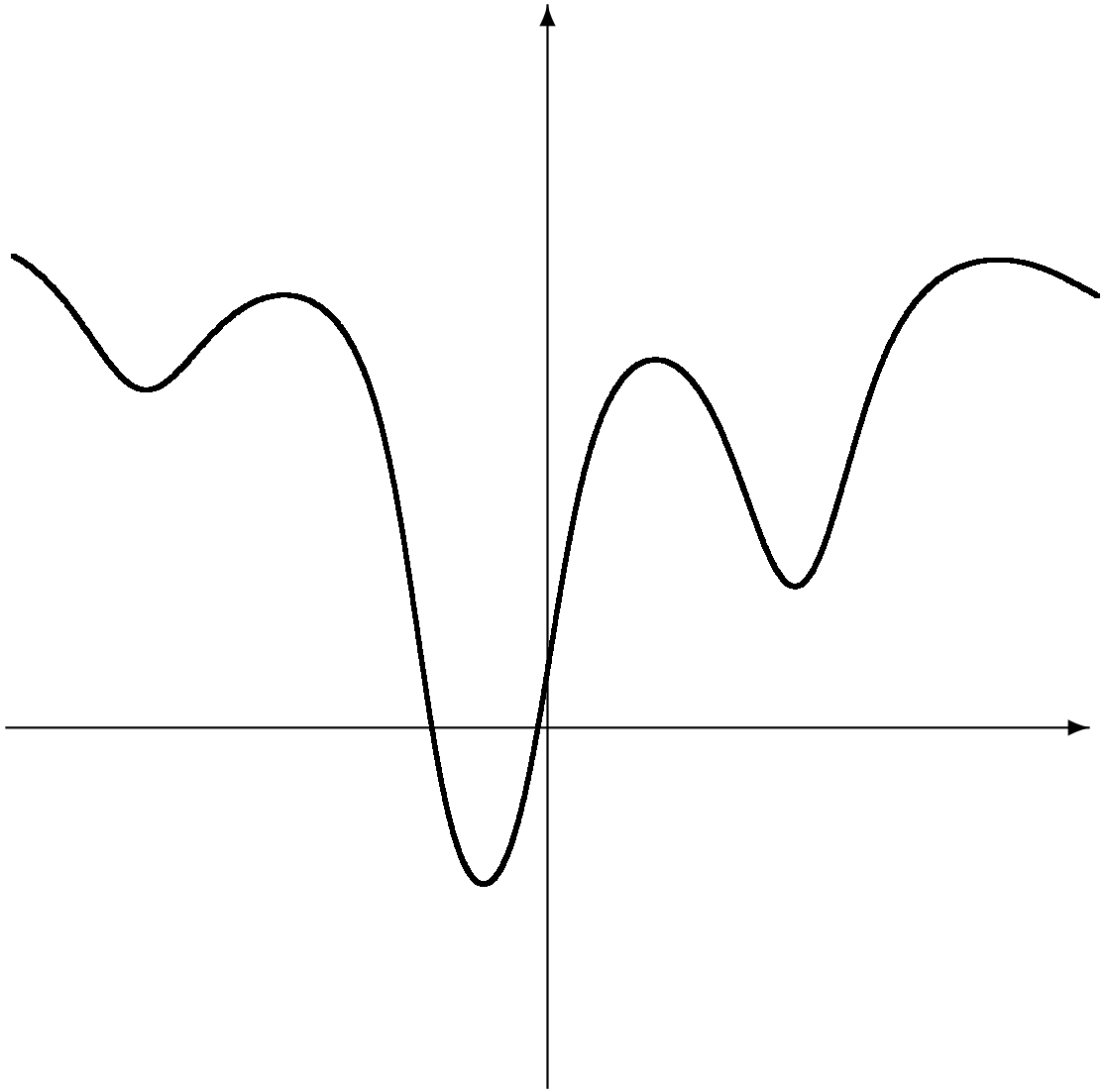
Iterative methods

Traditional "global" convergence results :

1. Every limit point is stationary ;
2. $\lim \|g(x^k)\| = 0$.

Complexity question

Given $\varepsilon > 0$ and $x^0 \in \mathbb{R}^n$, How many iterations-evaluations are needed to guarantee that $\|g(x^k)\| \leq \varepsilon$?
Or :
Given $\varepsilon > 0$, $x^0 \in \mathbb{R}^n$, and $f_{target} < f(x^0)$, which is the maximal number of iterations-evaluations that may occur satisfying $f(x^k) > f_{target}$ and $\|g(x^k)\| \geq \varepsilon$? ?



Cauchy-like method

Cauchy is based on the approximation in the linear model

$$f(x^k + s) \approx f(x^k) + \langle g(x^k), s \rangle.$$

Therefore, the Cauchy idea should be based on the minimization of the linear model at each iteration.
But the linear model does not have a minimizer.
Therefore, one minimizes the linear model plus a quadratic regularization :

$$f(x^k) + \langle g(x^k), s \rangle + \rho \|s\|^2.$$

Algorithm : Cauchy with regularization

Given $x^0 \in \mathbb{R}^n$, $\alpha > 0$, $k \leftarrow 0$.
Step 1 : Set $\rho \leftarrow 1$.
Step 2 : Solve "the subproblem"

$$\text{Minimize } \langle g(x^k), s \rangle + \rho \|s\|^2,$$

obtaining the solution s^{trial}. Note that $s^{trial} = -\frac{1}{2\rho} g(x^k)$.
Step 3 : Test the sufficient descent condition
If

$$f(x^k + s^{trial}) \leq f(x^k) - \alpha \|s^{trial}\|^2,$$

set $s^k = s^{trial}$, $x^{k+1} = x^k + s^k$, $k \leftarrow k + 1$, and go to Step 1.
Otherwise, set $\rho \leftarrow 2\rho$ and go to Step 2.
(Note that this is the steepest descent method with elementary backtracking line search.)

Assumption : $f(x+s) \leq f(x) + \langle g(x), s \rangle + L\|s\|^2$

Property 1 : If $\rho \geq L + \alpha$, sufficient descent is fulfilled.
Proof : $f(x^k + s) \leq f(x^k) + \langle g(x^k), s \rangle + L\|s\|^2$

$$\leq f(x^k) + \langle g(x^k), s \rangle + (L+\alpha)\|s\|^2 - \alpha\|s\|^2$$
$$\leq f(x^k) + \langle g(x^k), s \rangle + \rho\|s\|^2 - \alpha\|s\|^2$$
$$\leq f(x^k) - \alpha\|s\|^2.$$

Property 2 : At each iteration, after a maximum of $\log_2(L+\alpha)$ tests (backtrackings, functional evaluations) we necessarily obtain the descent condition with

$$\rho < 2(L+\alpha).$$

Complexity of Cauchy $= O(\varepsilon^{-2})$

From $s^{trial} = -\frac{1}{2\rho} g(x^k)$ we get $\|s^k\| \geq \frac{1}{4(L+\alpha)} \|g(x^k)\|$. Then, by the sufficient descent condition,

$$f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{16(L+\alpha)^2} \|g(x^k)\|^2.$$

Then, if $\|g(x^k)\| \geq \varepsilon$,

$$f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{16(L+\alpha)^2} \varepsilon^2.$$

This implies that the maximal number of iterations at which $\|g(x^k)\| \geq \varepsilon$ and $f(x^k) > f_{target}$ is :

$$[f(x^0) - f_{target}] \frac{16(L+\alpha)^2}{\alpha} \varepsilon^{-2}.$$

Maximum of evaluations :

$$[f(x^0) - f_{target}] \log_2(L+\alpha) \frac{16(L+\alpha)^2}{\alpha} \varepsilon^{-2}.$$

Other methods with complexity $O(\varepsilon^{-2})$

1. Cauchy-related methods with different line searches.
2. Cauchy-related methods with different line searches and non-monotone strategies.
3. Quasi-Newton methods with line searches and safeguards. (Not without safeguards !)
4. Newton's method with safeguards and line searches.
5. Newton's method with trust regions whose complexity is not better than $O(\varepsilon^{-2})$, which is very unfair ! !

Is the complexity bound $O(\varepsilon^{-2})$ sharp for gradient-like methods ?

Yes.
An example due to Cartis, Gould, and Toint for which the worst complexity holds "almost" exactly.
But :
In this example third derivatives are unbounded.
Open question : Existence of a sharpness example with bounded third derivatives.
We will return to this later.

Newton method and cubic regularization

Imitating Cauchy, Newton is based on the approximation by the quadratic model

$$f(x^k + s) \approx f(x^k) + \langle g(x^k), s\rangle + \frac{1}{2} s^T H(x^k) s.$$

Therefore, the Newtonian idea should be based on the minimization of the quadratic model at each iteration.
But the quadratic model not always has a minimizer.
Therefore, one minimizes the linear model plus a cubic regularization :

$$f(x^k) + \langle g(x^k), s\rangle + \frac{1}{2} s^T H(x^k) s + \rho \|s\|^3.$$

Algorithm : Newton with cubic regularization

Given $x^0 \in \mathbb{R}^n$, $\alpha > 0$, $k \leftarrow 0$.
Step 1 : Set $\rho \leftarrow 0$.
Step 2 : Solve "the subproblem"

$$\text{Minimize } \langle g(x^k), s\rangle + \frac{1}{2} s^T H(x^k) s + \rho \|s\|^3,$$

obtaining the solution s^{trial}.
Step 3 : Test the sufficient descent condition
If

$$f(x^k + s^{trial}) \leq f(x^k) - \alpha \|s^{trial}\|^3,$$

set $s^k = s^{trial}$, $x^{k+1} = x^k + s^k$, $k \leftarrow k + 1$, and go to Step 1.
Otherwise, set $\rho \leftarrow \max\{1, 2\rho\}$ and go to Step 2.

Assumption : The Hessian is Lipschitz with constant L_3

Property 1 : If $\rho \geq L_3 + \alpha$, sufficient descent is fulfilled.
Proof : $f(x^k + s) \leq f(x^k) + \langle g(x^k), s\rangle + \frac{1}{2} s^T H(x^k) s + L_3 \|s\|^3$

$$\leq f(x^k) + \langle g(x^k), s\rangle + \frac{1}{2} s^T H(x^k) s + (L_3 + \alpha)\|s\|^3 - \alpha\|s\|^3$$

$$\leq f(x^k) + \langle g(x^k), s\rangle + \frac{1}{2} s^T H(x^k) s + \rho\|s\|^3 - \alpha\|s\|^3$$

$$\leq f(x^k) - \alpha\|s\|^3.$$

Property 2 : At each iteration, after a maximum of $\log_2(L_3 + \alpha)$ tests (backtrackings, functional evaluations) we necessarily obtain the descent condition with

$$\rho < 2(L_3 + \alpha).$$

Complexity of Newton-Cubic $= O(\varepsilon^{-3/2})$

$$\|g(x^{k+1})\| = \|g(x^k + s^k)\| \leq \|g(x^k) + H_k s^k\| + L_3\|s\|^2$$

and

$$\nabla_s[\langle g(x^k), s^k\rangle + \frac{1}{2}(s^k)^T H(x^k)s^k + \rho\|s^k\|^3] = 0.$$

$$\Rightarrow g(x^k) + H_k s^k + 3\rho s^k\|s^k\| = 0 \Rightarrow \|g(x^k) + H_k s^k\| \leq 3\rho\|s^k\|^2$$

$$\Rightarrow \|g(x^{k+1})\| \leq (L_3 + 3\rho)\|s^k\|^2 \leq (L_3 + 6(L_3 + \alpha))\|s^k\|^2$$

$$\Rightarrow \|s^k\| \geq \left(\|g(x^{k+1})\|/(7L_3 + \alpha)\right)^{1/2}.$$

Then, by the sufficient descent condition,

$$f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{[7L_3 + \alpha]^3}\|g(x^{k+1})\|^{3/2}.$$

Then, if $\|g(x^{k+1})\| \geq \varepsilon$,

$$f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{[7L_3 + \alpha]^3}\varepsilon^{3/2}.$$

This implies that the maximal number of iterations at which $\|g(x^{k+1})\| \geq \varepsilon$ and $f(x^k) > f_{target}$ is :

$$[f(x^0) - f_{target}]\frac{[7L_3 + \alpha]^3}{\alpha}\varepsilon^{-3/2}.$$

Maximum of evaluations :

$$[f(x^0) - f_{target}]\log_2(L_3 + \alpha)\frac{[7L_3 + \alpha]^3}{\alpha}\varepsilon^{-3/2}.$$

What about naive quadratic regularization with cubic descent ?

Given $x^0 \in \mathbb{R}^n$, $\alpha > 0$, $k \leftarrow 0$.
Step 1 : Set $\rho \leftarrow 0$.
Step 2 : Solve "the subproblem"

$$\text{Minimize } \langle g(x^k), s\rangle + \frac{1}{2} s^T H(x^k) s + \frac{\rho}{2}\|s\|^2,$$

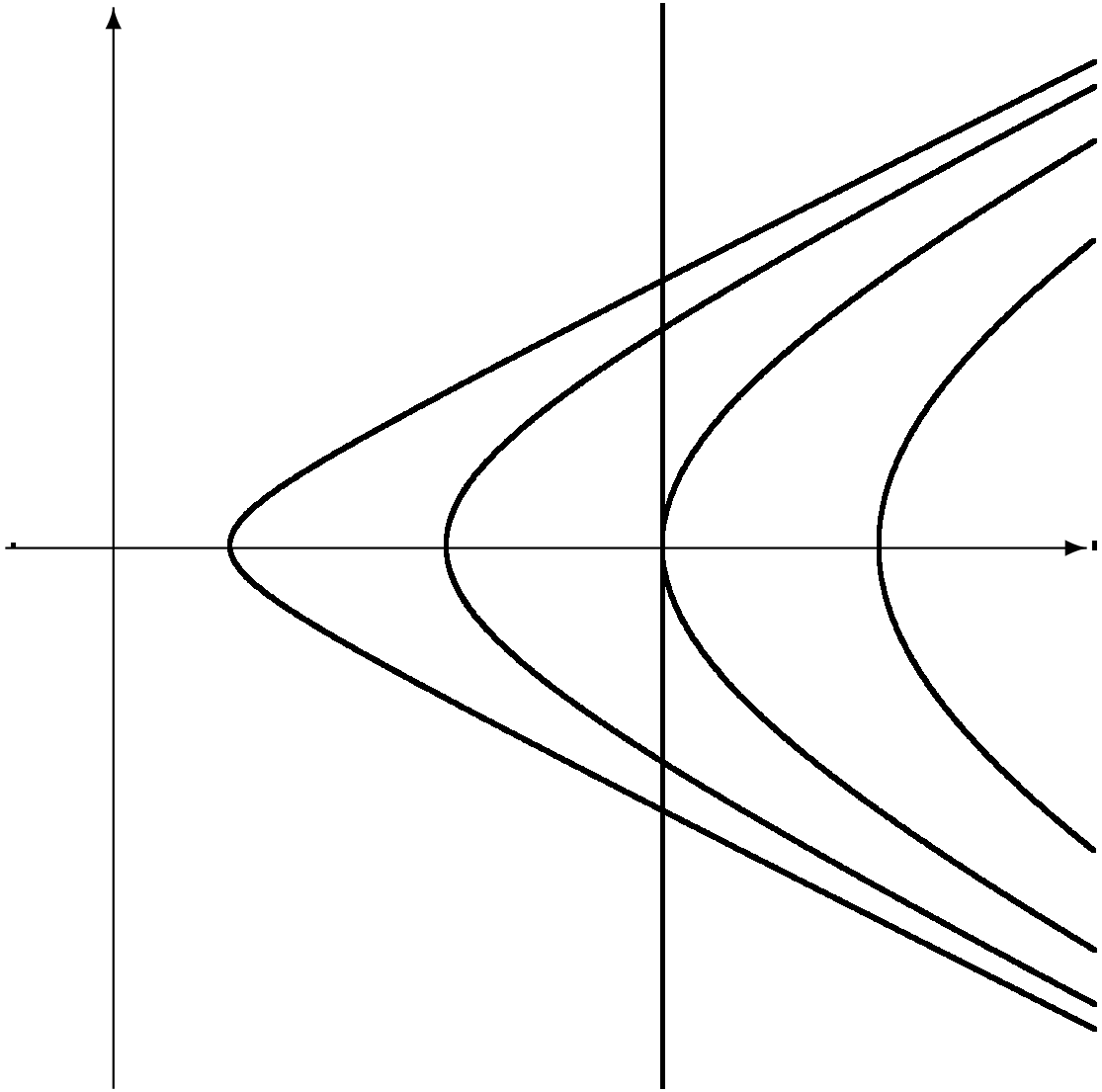
obtaining the solution s^{trial}. (Jump to "Otherwise,..." if there is no solution ($\rho < -\lambda_1$ or, perhaps, $\rho = -\lambda_1$))
Step 3 : Test the sufficient descent condition
If

$$f(x^k + s^{trial}) \leq f(x^k) - \alpha\|s^{trial}\|^3,$$

set $s^k = s^{trial}$, $x^{k+1} = x^k + s^k$, $k \leftarrow k+1$, and go to Step 1.
Otherwise, set $\rho \leftarrow \max\{1, 2\rho\}$ and go to Step 2.

Naive quadratic regularization with cubic descent has complexity $O(\varepsilon^{-2})$ and does not have complexity $O(\varepsilon^{-3/2})$

Existence of an example in which the complexity is sharply $O(\varepsilon^{-2})$:
Such example exists : Take an one-dimensional example of minimizing $f_1(x_1)$ in which Cauchy is sharply $O(\varepsilon^{-2})$ and define $f(x_1, x_2) = f_1(x_1) - x_2^2$. Minimizing $f(x_1, x_2)$ with Naive Quadratic Regularization results sharply $O(\varepsilon^{-2})$.
The problem is that known examples in which Cauchy is sharply $O(\varepsilon^{-2})$ do not have Lipschitz Hessians.
Picture :



Religious belief

Every algorithmic modification that improves theoretical properties will eventually improve practical performance.

Not-so-naive quadratic regularization with cubic descent

Given $x^0 \in \mathbb{R}^n$, $\alpha > 0$, $k \leftarrow 0$.
Step 1 : Choose "carefully" the regularization parameter ρ, in particular, considering separately the "hard case" (in which $\rho = -\lambda_1$).
Step 2 : Solve the subproblem :

$$\text{Minimize } \langle g(x^k), s\rangle + \frac{1}{2} s^T H(x^k) s + \frac{\rho}{2}\|s\|^2,$$

obtaining the solution s^{trial}.
Step 3 : Test the sufficient descent condition
If

$$f(x^k + s^{trial}) \leq f(x^k) - \alpha \|s^{trial}\|^3,$$

set $s^k = s^{trial}$, $x^{k+1} = x^k + s^k$, $k \leftarrow k+1$, and go to Step 1.
Otherwise, go to Step 1.

Main feature of the not-so-naive quadratic regularization method

Consider the Hard Case !
This involves different trial points with the same regularization parameter $-\lambda_1$.

Theoretical improvement provided by the not-so-naive quadratic regularization with respect to naive quadratic regularization

Worst-case Complexity $O(\varepsilon^{-3/2})$.

Generalization of Newton with cubic regularization

Instead of minimizing the quadratic Taylor approximation plus cubic regularization

Minimize the p-th Taylor approximation plus $p+1$-th regularization

Complexity result : $O(\varepsilon^{-1-\frac{1}{p}})$. (Tends to -1 if $p \to \infty$.)

Generalization of Newton with quadratic regularization and cubic descent

Instead of minimizing the quadratic Taylor approximation plus quadratic regularization with $p+1$-th descent
Minimize the p-th Taylor approximation plus p-th regularization and $p+1$-th descent.
Unknown theoretical properties and probably without practical relevance.

Some words about Complexity in Constrained Optimization

1. What is a solution ?
2. Depending of the definition of solution you obtain different complexity results
3. The method with best complexity results (very close to the unconstrained case) is a method with short steps, analogous to GRG

Conclusions

1. WCC (Worst-Case Complexity) analysis provides (very pessimistic) estimates of number of evaluations and computer time.
2. However, one believes that better worst-case implies better average case.
3. Cauchy-related are $O(\varepsilon^{-2})$.
4. Standard Newton are $O(\varepsilon^{-2})$.
5. Newton with Cubic Regularization is $O(\varepsilon^{-3/2})$.
6. Naive Newton with quadratic regularization is $O(\varepsilon^{-2})$.
7. Not-so-naive Newton with quadratic regularization is $O(\varepsilon^{-3/2})$.

Constrained Minimization Problem

Assume that $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^n$, $h_E : \mathbb{R}^n \to \mathbb{R}^m$, $h_I : \mathbb{R}^n \to \mathbb{R}^q$, $h'_E : \mathbb{R}^n \to \mathbb{R}^{m \times n}$, and $h'_I : \mathbb{R}^n \to \mathbb{R}^{q \times n}$. Our problem will be

$$\text{Minimize } f(x) \text{ subject to } h_E(x) = 0 \text{ and } h_I(x) \leq 0. \qquad (1)$$

Local Models and Examples

For all $\bar{x} \in \mathbb{R}^n$ we define $M_{\bar{x}} : \mathbb{R}^n \to \mathbb{R}$ (intended to be a "model" of $f(x)$ around $\bar{x}$) and $\nabla M_{\bar{x}} : \mathbb{R}^n \to \mathbb{R}^n$.

$$\|g(x) - \nabla M_{\bar{x}}(x)\| \leq L\|x - \bar{x}\|^{p+\beta-1} \tag{2}$$

and

$$M_{\bar{x}}(\bar{x}) = f(\bar{x}) \text{ and } f(x) \leq M_{\bar{x}}(x) + L\|x - \bar{x}\|^{p+\beta}. \tag{3}$$

Local Model approximately minimizes

$$M_{\bar{x}}(x) + \sigma \|x - \bar{x}\|^{p+1} \leq f(\bar{x}) \tag{4}$$

and

$$\|\nabla[M_{\bar{x}}(x)+\sigma\|x-\bar{x}\|^{p+1}]+h_E'(x)^T\lambda+h_I'(x)^T\mu\| \leq \theta\|x-\bar{x}\|^p, \tag{5}$$

where

$$\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^q_+, \| \min\{\mu, -h_I(x)\}\| \leq \delta \tag{6}$$

$$\|h_E(x)\| \leq \delta, \text{ and } \|h_I(x)_+\| \leq \delta. \tag{7}$$

$$\text{Minimize } M_{\bar{x}}(x)+\sigma\|x-\bar{x}\|^{p+1} \text{ subject to } h_E(x)=0 \text{ and } h_I(x)\leq 0. \tag{8}$$

Algorithm

Assume that $x^0 \in \mathbb{R}^n$, $\alpha \in (0,1)$, $\varepsilon \in (0,1)$,$\delta > 0$, $f_{target} \in \mathbb{R}$, $\theta > 0$, and $\sigma_{min} > 0$,
Initialize $k \leftarrow 0$ and $\sigma_0 = \sigma_{min}$.
Step 1. Set $\sigma \leftarrow \sigma_k$.
Step 2. Find $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^m$, and $\mu \in \mathbb{R}^q_+$ such that (4) and (5) hold with $\bar{x} = x^k$.
Step 3. If $\|g(x) + h'_E(x)^T\lambda + h'_I(x)^T\mu\| \leq \varepsilon$ or $f(x) \leq f_{target}$, stop.
Step 4. Test the sufficient descent condition

$$f(x) \leq f(x^k) - \frac{\alpha}{(2p+4)^{\frac{p+1}{p}}} \frac{\varepsilon^{\frac{p+1}{p}}}{\sigma^{\frac{1}{p}}}. \tag{9}$$

If (9) does not hold, set $\sigma \leftarrow 2\sigma$ and go to Step 2. Else, continue at Step 5.
Step 5. Define $x^{k+1} = x$, $k \leftarrow k+1$, $\sigma_k = \sigma$, and go to Step 1.

Lemma

Assume that $\alpha \in (0,1)$ and (2), (3), (4), (5), (6), and (7) are satisfied by $\bar{x}$, x, λ, and μ. Then,

$$\|g(x)+h_E'(x)^T\lambda+h_I'(x)^T\mu\| \leq (\theta+(p+1)\sigma)\|x-\bar{x}\|^p+L\|x-\bar{x}\|^{p+\beta-1}. \tag{10}$$

Moreover, if

$$\|g(x) + h'_E(x)^T\lambda + h'_I(x)^T\mu\| \geq \varepsilon > 0 \tag{11}$$

and

$$\sigma \geq \max\left\{\theta, c_\sigma \varepsilon^{\frac{\beta-1}{p+\beta-1}}\right\} \tag{12}$$

we have that

$$\sigma\|x - \bar{x}\|^p \geq \frac{\varepsilon}{2p+4}, \tag{13}$$

$$f(x) \leq f(\bar{x}) - \alpha\sigma\|x - \bar{x}\|^{p+1}, \tag{14}$$

and

$$f(x) \leq f(\bar{x}) - \frac{\alpha}{(2p+4)^{\frac{p+1}{p}}} \frac{\varepsilon^{\frac{p+1}{p}}}{\sigma^{\frac{1}{p}}}. \tag{15}$$

$$c_\sigma = \max\left\{\frac{2^{\frac{-\beta+1}{p+\beta-1}}}{p+2} L^{\frac{p}{p+\beta-1}}, \frac{L}{(1-\alpha)((2p+4)^{(\beta-1)/(p+\beta-1)}}\right\}$$

Theorem 1

Assume that x^{k+1} is computed by Algorithm and the assumptions (2)–(7) hold for $\bar{x} = x^k$ at all the trial points x computed at Step 2 of iteration k.
Then,

$$f(x^{k+1}) \leq f(x^k) - \alpha c_p \varepsilon^{\frac{p+\beta}{p+\beta-1}} \tag{16}$$

$$c_p = \min\left\{ \frac{1}{(2p+4)^{\frac{p+1}{p}}} \frac{1}{(2\theta)^{\frac{1}{p}}}, \frac{1}{(2p+4)^{\frac{p+1}{p}} \left\{ 2\max\left\{ \frac{2^{\frac{-\beta+1}{p+\beta-1}}}{p+2} L^{\frac{p}{p+\beta-1}}, \frac{L}{(1-\alpha)((2p+4)^{(\beta-1)/(p+\beta-1)}} \right\} \right\}^{\frac{1}{p}}} \right\}.$$

Theorem 2

Assume that (2)–(7) hold for $\bar{x} = x^k$ and all the trial points x computed at every iteration k performed by Algorithm. Then, after, at most,

$$(f(x^0) - f_{target}) \frac{\varepsilon^{-\frac{p+\beta}{p+\beta-1}}}{\alpha c_p}. \tag{17}$$

iterations, Algorithm computes $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^m$, and $\mu \in \mathbb{R}^q_+$ verifying $f(x) \leq f_{target}$ or

$$\|g(x) + h'_E(x)^T \lambda + h'_I(x)^T \mu\| \leq \varepsilon, \tag{18}$$

and

$$\|h_E(x)\| \leq \delta, \|h_I(x)_+\| \leq \delta, \text{ and } \|\min\{\mu, -h_I(x)\}\| \leq \delta. \tag{19}$$

Theorem 3

Assume that the hypotheses of Theorem hold. Then, the number of evaluations of f employed by Algorithm bounded above by

$$(f(x^0) - f_{target}) \frac{\varepsilon^{-\frac{p+\beta}{p+\beta-1}}}{\alpha c_p} +$$

$$+ \left[\max \left\{ \log_2(\theta), \left(\frac{1-\beta}{p+\beta-1} \log_2(\varepsilon^{-1}) + c_\ell \right) \right\} \right] - \log_2(\sigma_{min}) + 1.$$

$$c_\ell = \log_2(\max\left\{\frac{2^{\frac{-\beta+1}{p+\beta-1}}}{p+2} L^{\frac{p}{p+\beta-1}}, \frac{L}{(1-\alpha)((2p+4)^{(\beta-1)/(p+\beta-1)}}\right\}).$$

Computational effort when $\varepsilon = 10^{-4}$

$$p = 1, \beta = 1 : 10^8;$$

$$p = 1, \beta = 0.01 : 10^{400};$$

$$p = 2, \beta = 0 : 10^8;$$

$$p = 2, \beta = 1 : 10^6;$$

$$p = 3, \beta = 0 : 10^6;$$

$$p = 3, \beta = 1 : 10^{16/3} \approx 10^5;$$

$$p \approx \infty : 10^4.$$