



LINMA2460 Nonlinear Programming

SIMON DESMIDT
ISSAMBRE L'HERMITE DUMONT

Academic year 2024-2025 - Q2



UCLouvain

Contents

1	Definitions, notations and random properties	2
1.1	Properties	3
1.2	Complexity table	3
1.3	GM VS Newton: table	3
2	TODO	4
3	Gradient descent without gradient	6
4	Local rates of convergence	8
4.1	Linear rate of GM	8
4.2	Local quadratic convergence of Newton's method	10
4.3	Quasi Newton methods	11
5	Constrained nonlinear programming problems	17
6	Tips and Tricks	19

Definitions, notations and random properties

- The Taylor expansion of order p of the function f around x_k and evaluated at y is:

$$T_p(y; x_k) = f(x_k) + \sum_{i=1}^p \frac{1}{i!} D^i f(x_k) (y - x_k)^i \quad (1.1)$$

- We can thus define the gradient w.r.t. y of the Taylor expansion of order p of f around x_k and evaluated at x_{k+1} :

$$\nabla_y T_p(x_{k+1}; x_k) = \nabla_y T_p(y; x_k) \big|_{y=x_{k+1}} \quad (1.2)$$

- An oracle is a "black box" that gives information about the derivatives based on x . The general form of an oracle is:

$$\text{p-order oracle: } x \mapsto \{D^i f(x)\}_{i=0}^p \quad (1.3)$$

And so we have the following simple oracles examples:

$$\begin{aligned} \text{Zero}^{th}\text{-order oracle: } x &\mapsto \{f(x)\} \\ \text{First-order oracle: } x &\mapsto \{f(x), \nabla f(x)\} \\ \text{Second-order oracle: } x &\mapsto \{f(x), \nabla f(x), \nabla^2 f(x)\} \end{aligned} \quad (1.4)$$

- $\mathcal{C}_L^p(\mathbb{R}^n)$: Class of functions p -times continuously differentiable with L -Lipschitz continuous p -order derivative, i.e. $\|D^p f(x) - D^p f(y)\| \leq L\|x - y\|, \forall x, y \in \mathbb{R}^n$. And so we have the following simple classes of problems:

- $\mathcal{C}_L^1(\mathbb{R}^n)$: Class of continuously differentiable functions with L -Lipschitz gradient;
- $\mathcal{C}_L^2(\mathbb{R}^n)$: Class of continuously differentiable functions with L -Lipschitz hessian.

- p th-order method (generalization of GM):

$$x_{k+1} = \arg \min_{y \in \mathbb{R}^n} \Omega_{x_k, y, p}(y) \equiv T_{x_k, p}(y) + \frac{M}{(p+1)!} \|y - x_k\|^{p+1} \quad (1.5)$$

- Convergence rate:

– Linear:

$$\|x_{k+1} - x^*\| \leq \alpha \|x_k - x^*\| \quad \forall k \geq 0, \alpha \in (0, 1) \quad (1.6)$$

– Super Linear:

$$\lim_{k \rightarrow +\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0 \quad (1.7)$$

– Quadratic:

$$\|x_{k+1} - x^*\| \leq \beta \|x_k - x^*\|^2 \quad \forall k \geq 0, \beta > 0 \quad (1.8)$$

1.1 Properties

- For a function $f \in \mathcal{C}^1(\Omega)$ and Ω is bounded, the following holds: $\|\nabla f(x)\| \leq L$ for all $x \in \Omega$ for some $L \geq 0$.
- By the mean value theorem, for a continuously differentiable function f , $\forall x, y \in \Omega$, $\exists z \in \Omega : f(y) - f(x) = \langle \nabla f(z), y - x \rangle$.
- For a matrix A and a scalar b , $\|A\| \leq b \implies |\lambda(A)| \leq b \implies |A| \preceq bI_n$, where the absolute value of the matrix is taken component wise.

1.2 Complexity table

Method	Lipschitz	∇f	$\nabla^2 f$...	$\nabla^p f$
Zero order		$O(n\varepsilon^{-2})$			
First order	$p = 1$	$O(\varepsilon^{-2})$			
Second order	$p = 2$	✗	$O(\varepsilon^{-3/2})$		
\vdots		✗	✗	\ddots	
p order		✗	✗	✗	$O(\varepsilon^{-\frac{p+1}{p}})$

1.3 GM VS Newton: table

	cost per iteration	cost of memory	Local rate
GM	$\mathcal{O}(n)$	$\mathcal{O}(n)$	Linear
Quasi-Newton	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	Super Linear
Newton	$\mathcal{O}(n^3)$	$\mathcal{O}(n^2)$	Quadratic

→ For the GM, we assume that we don't need to compute the gradient at each iteration.

TODO

We can generalise the property of a L-Lipschitz function to $f \in \mathcal{C}_L^p(\mathbb{R}^n)$. For $p = 1$, we had

$$f(y) \leq f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2} \|y - x_k\|^2 \quad \forall y \in \mathbb{R}^n \quad (2.1)$$

For a general value of p , it becomes

$$f(y) \leq T_p(y; x_k) + \frac{L}{(p+1)!} \|y - x_k\|^{p+1} \quad \forall y \in \mathbb{R}^n \quad (2.2)$$

Using this, [we need a \$p\$ -th order oracle](#) for the method to work.

To solve $\min_{x \in \mathbb{R}^n} f(x)$, we can use the iteration

$$x_{k+1} = \arg \min_{y \in \mathbb{R}^n} T_p(y; x_k) + \frac{M}{(p+1)!} \|y - x_k\|^{p+1} \quad (2.3)$$

where the constant M is an approximation of the Lipschitz constant L . [Assuming \$f \in \mathcal{C}_L^p\(\mathbb{R}^n\)\$](#) , we have

$$\begin{aligned} f(x_{k+1}) &\leq T_p(x_{k+1}; x_k) + \frac{L}{(p+1)!} \|x_{k+1} - x_k\|^{p+1} \\ &= \underbrace{T_p(x_{k+1}; x_k) + \frac{M}{(p+1)!} \|x_{k+1} - x_k\|^{p+1}}_{\leq f(x_k)} + \frac{(L-M)}{(p+1)!} \|x_{k+1} - x_k\|^{p+1} \end{aligned} \quad (2.4)$$

where the inequality $\leq f(x_k)$ is due to the decrease of f and equation (2.3). [Suppose that \$M > 2L\$](#) . After some algebraic manipulations, we get

$$f(x_k) - f(x_{k+1}) \geq \frac{L}{(p+1)!} \|x_{k+1} - x_k\|^{p+1} \quad (2.5)$$

On the other hand, using the triangular inequality,

$$\begin{aligned} \|\nabla f(x_{k+1})\| &\leq \|\nabla f(x_{k+1}) - \nabla_y T_p(x_{k+1}; x_k)\| \\ &\quad + \underbrace{\left\| \nabla_y T_p(x_{k+1}; x_k) + \nabla \left(\frac{M}{(p+1)!} \|\cdot - x_k\|^{p+1} \right) \right\|_{y=x_{k+1}}}_{=0} \\ &\quad + \left\| \nabla \left(\frac{M}{(p+1)!} \|\cdot - x_k\|^{p+1} \right) \right\|_{y=x_{k+1}} \\ &\leq \frac{L}{p!} \|x_{k+1} - x_k\|^p + \frac{M}{p!} \|x_{k+1} - x_k\|^p \end{aligned} \quad (2.6)$$

$$\implies \|x_{k+1} - x_k\| \geq \left(\frac{p!}{L+M} \right)^{1/p} \|\nabla f(x_{k+1})\|^{1/p} \quad (2.7)$$

Combining equations (2.5) and (2.7),

$$f(x_k) - f(x_{k+1}) \geq \underbrace{\frac{L}{(p+1)!} \left(\frac{p!}{L+M} \right)^{\frac{p+1}{p}}}_{=:C(L)} \|\nabla f(x_{k+1})\|^{\frac{p+1}{p}} \quad (2.8)$$

Let $T(\varepsilon) = \inf\{k \in \mathbb{N} : \|\nabla f(x_k)\| \leq \varepsilon\}$. Assume that $T(\varepsilon) \geq 2$ and $f(x) \geq f_{low} \forall x \in \mathbb{R}^n$. Summing up (2.8) for $k = 0, \dots, T(\varepsilon) - 2$,

$$\begin{aligned} f(x_0) - f_{low} &\geq f(x_0) - f(x_{T(\varepsilon)-1}) = \sum_{k=0}^{T(\varepsilon)-2} f(x_k) - f(x_{k+1}) \\ &\geq (T(\varepsilon) - 1)C(L)\varepsilon^{\frac{p+1}{p}} \\ \implies T(\varepsilon) &\leq 1 + \frac{f(x_0) - f_{low}}{C(L)}\varepsilon^{-\frac{p+1}{p}} \equiv \mathcal{O}\left(\varepsilon^{-\frac{p+1}{p}}\right) \end{aligned} \quad (2.9)$$

Gradient descent without gradient

For this problem consider an adversarial attack on block-based image classifier. We have a machine learning model that given an image $a \in \mathbb{R}^p$ it returns $c(a) \in \mathbb{R}^m$, where $c_j(a) \in [0, 1]$ is the probability of image a to be in class j . The classifier prediction is: $j(a) = \arg \max_{j \in [1, \dots, m]} c_j(a)$.

TODO - Add mise en situation ou pas?

Given x_k let us decide:

$$x_{k+1} = x_k - \frac{1}{\sigma} g_{h_k}(x_k) \quad h_k > 0, \sigma > 0 \quad (3.1)$$

where $g_{h_k}(x_k) \in \mathbb{R}^n$ is given by:

$$[g_{h_k}(x_k)]_j = \frac{f(x_k + h e_j) - f(x_k)}{h_k} \quad \forall j \in [1, \dots, m] \quad (3.2)$$

Suppose that $f \in \mathcal{C}_L^1(\mathbb{R}^n)$. Then,

$$\|\nabla f(x_k) - g_{h_k}(x_k)\| \leq \frac{L\sqrt{n}}{2} h_k \quad (3.3)$$

Thus

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 \\ &= f(x_k) + \langle g_{h_k}(x_k), x_{k+1} - x_k \rangle + \frac{\sigma}{2} \|x_{k+1} - x_k\|^2 \\ &\quad + \langle \nabla f(x_k) - g_{h_k}(x_k), x_{k+1} - x_k \rangle + \frac{(L - \sigma)}{2} \|x_{k+1} - x_k\|^2 \\ &\leq f(x_k) - \frac{1}{\sigma} \|g_{h_k}(x_k)\|^2 + \frac{1}{2\sigma} \|g_{h_k}(x_k)\|^2 \\ &\quad + \|\nabla f(x_k) - g_{h_k}(x_k)\| \frac{1}{\sigma} \|g_{h_k}(x_k)\| + \frac{(L - \sigma)}{2\sigma^2} \|g_{h_k}\|^2 \\ &\leq f(x_k) - \frac{1}{2\sigma} \|g_{h_k}(x_k)\|^2 + \frac{L\sqrt{n}}{2} h_k \frac{1}{\sigma} \|g_{h_k}\| + \frac{(L - \sigma)}{2\sigma^2} \|g_{h_k}\|^2 \\ &\leq f(x_k) - \frac{1}{2\sigma} \|g_{h_k}(x_k)\|^2 + \frac{L}{2} \left(\frac{nh_k^2}{2} + \frac{1}{2\sigma} \|g_{h_k}(x_k)\|^2 \right) + \frac{(L - \sigma)}{2\sigma^2} \|g_{h_k}\|^2 \\ &= f(x_k) - \left(\frac{2\sigma - L - 2(L - \sigma)}{4\sigma^2} \right) \|g_{h_k}(x_k)\|^2 + \frac{Ln}{4} h_k^2 \\ &= f(x_k) - \frac{(4\sigma - 3L)}{4\sigma} \|g_{h_k}(x_k)\|^2 + \frac{Ln}{4} h_k^2 \end{aligned} \quad (3.4)$$

$$\implies \frac{(4\sigma - 3L)}{4\sigma} \|g_{h_k}(x_k)\|^2 \leq f(x_k) - f(x_{k+1}) + \frac{Ln}{4} h_k^2 \quad (3.5)$$

If $\sigma \gg L$, then

$$\frac{1}{4\sigma} \|g_{h_k}(x_k)\|^2 \leq f(x_k) - f(x_{k+1}) + \frac{\sigma n}{4} h_k^2 \quad (3.6)$$

On the other hand, we have

$$\begin{aligned} \|\nabla f(x_k)\| &\leq \|\nabla f(x_k) - g_{h_k}(x_k)\| + \|g_{h_k}(x_k)\| \\ &\leq \frac{L\sqrt{n}}{2} h_k + \|g_{h_k}(x_k)\| \end{aligned} \quad (3.7)$$

Using trick (6.3) in chapter 6,

$$\begin{aligned} \implies \|\nabla f(x_k)\|^2 &\leq 2 \left(\frac{L\sqrt{n}}{2} h_k \right)^2 + 2 \|g_{h_k}(x_k)\|^2 \\ &\leq \frac{L^2 n}{2} h_k^2 + 2 \|g_{h_k}(x_k)\|^2 \end{aligned} \quad (3.8)$$

$$\implies \frac{1}{8\sigma} \|\nabla f(x_k)\|^2 \leq \frac{L^2 n}{16\sigma} h_k^2 + \frac{1}{4\sigma} \|g_{h_k}(x_k)\|^2 \quad (3.9)$$

$$\implies \frac{1}{8\sigma} \|\nabla f(x_k)\|^2 \leq f(x_k) - f(x_{k+1}) + \frac{\sigma n}{4} h_k^2 + \frac{\sigma n}{16} h_k^2 \quad (3.10)$$

Let $T(\varepsilon) = \inf\{k \in \mathbb{N} : \|\nabla f(x_k)\| \leq \varepsilon\}$, with $f(x)$ bounded below by f_{low} , summing up (3.10) for $k = 0, \dots, T(\varepsilon) - 1$:

$$\frac{T(\varepsilon)}{8\sigma} \varepsilon^2 \leq f(x_0) - f_{low} + \frac{5\sigma n}{4} \sum_{k=0}^{T(\varepsilon)-1} h_k^2 \quad (3.11)$$

If $\{h_k^2\}$ is summable

$$\implies T(\varepsilon) \leq 8\sigma \left(f(x_0) - f_{low} + \frac{5\sigma n}{4} \sum_{k=0}^{T(\varepsilon)-1} h_k^2 \right) \varepsilon^2 = \mathcal{O}(\varepsilon^2) \quad (3.12)$$

In terms of call to the oracle, we have a complexity bound of $\mathcal{O}(n\varepsilon^2)$.

Local rates of convergence

4.1 Linear rate of GM

Let $f \in \mathcal{C}_M^{2,2}(\mathbb{R}^n)$. Assume f has a local minimizer x^* such that

$$\mu I_n \preceq \nabla^2 f(x^*) \preceq M I_n \quad (4.1)$$

Let $x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$ for a given $x_0 \in \mathbb{R}^n$.

Notice that

$$\begin{aligned} \nabla f(x_k) &= \nabla f(x_k) - \nabla f(x^*) \\ &= \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) (x_k - x^*) d\tau \\ &= \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau (x_k - x^*) \\ &= G_k(x_k - x^*) \end{aligned} \quad (4.2)$$

Then,

$$\begin{aligned} \|x_{k+1} - x^*\| &= \|x_k - \frac{1}{L} \nabla f(x_k) - x^*\| \\ &= \|(I_n - \frac{1}{L} G_k)(x_k - x^*)\| \\ &\leq \|I_n - \frac{1}{L} G_k\| \|x_k - x^*\| \end{aligned} \quad (4.3)$$

Since $f \in \mathcal{C}_M^{2,2}(\mathbb{R}^n)$, we have $\|\nabla^2 f(x^* + \tau(x_k - x^*)) - \nabla^2 f(x^*)\| \leq \tau M \|x_k - x^*\|$ and using this we get:

$$|\langle \nabla^2 f(x^* + \tau(x_k - x^*)) - \nabla^2 f(x^*) v, v \rangle| \leq \tau M \|x_k - x^*\| \|v\|^2 \quad \forall v \in \mathbb{R}^n \quad (4.4)$$

Using the bound (4.1) and the previous inequality, we get:

$$\begin{aligned} \tau M \|x_k - x^*\| \|v\|^2 &\leq |\langle \nabla^2 f(x^* + \tau(x_k - x^*)) - \nabla^2 f(x^*) v, v \rangle| \leq \tau M \|x_k - x^*\| \|v\|^2 \\ \nabla^2 f(x^*) - \tau M \|x_k - x^*\| I_n &\preceq \nabla^2 f(x^* + \tau(x_k - x^*)) \preceq \nabla^2 f(x^*) + \tau M \|x_k - x^*\| I_n \\ (\mu - \tau M \|x_k - x^*\|) I_n &\preceq \nabla^2 f(x^* + \tau(x_k - x^*)) \preceq (L + \tau M \|x_k - x^*\|) I_n \end{aligned}$$

By the properties of the semi-definite matrices, and the trick (6.4), we have:

$$\begin{aligned} \int_0^1 (\mu - \tau M \|x_k - x^*\|) \|v\|^2 d\tau &\leq \int_0^1 \langle \nabla^2 f(x^* + \tau(x_k - x^*)) v, v \rangle d\tau \\ &\leq \int_0^1 (L + \tau M \|x_k - x^*\|) \|v\|^2 d\tau \quad \forall v \in \mathbb{R}^n \end{aligned} \quad (4.5)$$

By using G_k and some constants, we get:

$$-\frac{1}{L}(L + \frac{M}{2}\|x_k - x^*\|)I_n \preceq -\frac{1}{L}G_k \preceq -\frac{1}{L}(\mu - \frac{M}{2}\|x_k - x^*\|)I_n \quad (4.6)$$

$$\left(1 - \frac{1}{L}(L + \frac{M}{2}\|x_k - x^*\|)\right) I_n \preceq I_n - \frac{1}{L}G_k \preceq \left(1 - \frac{1}{L}(\mu - \frac{M}{2}\|x_k - x^*\|)\right) I_n \quad (4.7)$$

And finally, we get:

$$\begin{aligned} \|I_n - \frac{1}{L}G_k\| &\leq \max \left\{ \left|1 - \frac{1}{L}(L + \frac{M}{2}\|x_k - x^*\|)\right|, \left|1 - \frac{1}{L}(\mu - \frac{M}{2}\|x_k - x^*\|)\right| \right\} \\ &= \max \left\{ \frac{M}{2L}\|x_k - x^*\|, 1 - \frac{\mu}{L} + \frac{M}{2L}\|x_k - x^*\| \right\} \\ &= 1 - \frac{\mu}{L} + \frac{M}{2L}\|x_k - x^*\| \end{aligned} \quad (4.8)$$

Suppose that $\frac{M}{2L}\|x_k - x^*\| \leq \frac{\mu}{2L} \iff \|x_k - x^*\| \leq \frac{\mu}{M}$

Then, in (4.8), we get:

$$\|I_n - \frac{1}{L}G_k\| \leq 1 - \frac{\mu}{2L} < 1 \quad (4.9)$$

And so, by (4.2)

$$\|x_{k+1} - x^*\| \leq \|I_n - \frac{1}{L}G_k\| \|x_k - x^*\| < \|x_k - x^*\| \quad (4.10)$$

If $\|x_0 - x^*\| < \frac{\mu}{M}$, it follows from the previous reasoning that:

$$\|x_2 - x^*\| \leq (1 - \frac{\mu}{2L})\|x_1 - x^*\| \leq (1 - \frac{\mu}{2L})^2\|x_0 - x^*\| \leq \frac{\mu}{M} \quad (4.11)$$

And so by induction, we can conclude that:

$$\|x_k - x^*\| \leq \left(1 - \frac{\mu}{2L}\right)^k \|x_0 - x^*\| \quad \forall k \geq 0 \quad (4.12)$$

\Rightarrow Linear rate of convergence

Given $\varepsilon > 0$, let $T(\varepsilon) = \inf\{k \in \mathbb{N} : \|x_k - x^*\| \leq \varepsilon\}$. Then, if $T(\varepsilon) \geq 1$ and using (4.12), we get:

$$\begin{aligned} \varepsilon &< \|x_{T(\varepsilon)-1} - x^*\| \leq \left(1 - \frac{\mu}{2L}\right)^{T(\varepsilon)-1} \|x_0 - x^*\| \\ \log \left(\frac{\varepsilon}{\|x_0 - x^*\|} \right) &\leq (T(\varepsilon) - 1) \log \left(1 - \frac{\mu}{2L} \right) \\ T(\varepsilon) - 1 &\leq \frac{\log \left(\frac{\varepsilon}{\|x_0 - x^*\|} \right)}{\log \left(1 - \frac{\mu}{2L} \right)} = \frac{\log (\|x_0 - x^*\| \varepsilon^{-1})}{|\log (1 - \frac{\mu}{2L})|} \end{aligned} \quad (4.13)$$

$$\boxed{T(\varepsilon) \leq \mathcal{O}(\log(\varepsilon^{-1}))}$$

\rightarrow Note: convexity was never assumed!

4.2 Local quadratic convergence of Newton's method

Let $f \in \mathcal{C}_M^{2,2}(\mathbb{R}^n)$. Assume f has a local minimizer x^* such that

$$\mu I_n \preceq \nabla^2 f(x^*) \quad \mu > 0 \quad (4.14)$$

Given $x_0 \in \mathbb{R}^n$, let:

$$x_{k+1} = x_k - \nabla^{-2} f(x_k) \nabla f(x_k) \quad (4.15)$$

We have, by the previous equation and the definition of G_k (4.2):

$$\begin{aligned} \|x_{k+1} - x^*\| &= \|x_k - \nabla^{-2} f(x_k) \nabla f(x_k) - x^*\| \\ &= \|(x_k - x^*) - \nabla^{-2} f(x_k) G_k(x_k - x^*)\| \\ &= \|\nabla^{-2} f(x_k) \left(\nabla^2 f(x_k) - \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) (x_k - x^*)\| \\ &= \|\nabla^{-2} f(x_k) \left(\int_0^1 \nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) (x_k - x^*)\| \\ &\leq \|\nabla^{-2} f(x_k)\| \left(\int_0^1 \|\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*))\| d\tau \right) \|x_k - x^*\| \\ &\leq \|\nabla^{-2} f(x_k)\| \left(\int_0^1 M(1 - \tau) \|x_k - x^*\| d\tau \right) \|x_k - x^*\| \\ &\leq \|\nabla^{-2} f(x_k)\| \|x_k - x^*\|^2 \frac{M}{2} \end{aligned} \quad (4.16)$$

Since $f \in \mathcal{C}_M^{2,2}(\mathbb{R}^n)$, we have

$$\nabla^2 f(x^* + \tau(x_k - x^*)) - \nabla^2 f(x^*) \succeq \tau M \|x_k - x^*\| I_n \quad (4.17)$$

$$\begin{aligned} \nabla^2 f(x_k) &\succeq \nabla^2 f(x^*) - M \|x_k - x^*\| I_n \\ &\succeq (\mu - M \|x_k - x^*\|) I_n \end{aligned} \quad (4.18)$$

$$\lambda_{\min}(\nabla^2 f(x_k)) \geq \mu - M \|x_k - x^*\|$$

Suppose that $-M \|x_k - x^*\| \geq -\frac{\mu}{2} \Leftrightarrow \|x_k - x^*\| \leq \frac{\mu}{2M}$

Then,

$$\begin{aligned} \lambda_{\min}(\nabla^2 f(x_k)) &\geq \frac{\mu}{2} \\ \lambda_{\max}(\nabla^{-2} f(x_k)) &\leq \frac{2}{\mu} \\ \Rightarrow \|\nabla^{-2} f(x_k)\| &\leq \frac{2}{\mu} \end{aligned} \quad (4.19)$$

Therefore, by (4.16), we conclude that:

$$\begin{aligned} \|x_{k+1} - x^*\| &\leq \frac{M}{2} \|\nabla^{-2} f(x_k)\| \|x_k - x^*\| \\ &\leq \frac{M}{\mu} \|x_k - x^*\|^2 \end{aligned} \quad (4.20)$$

If $\|x_k - x^*\| \leq \frac{\mu}{2M}$ then,

$$\|x_{k+1} - x^*\| \leq \frac{M}{\mu} \|x_k - x^*\|^2 = \frac{1}{2} \|x_k - x^*\| \quad (4.21)$$

If $\|x_0 - x^*\| \leq \frac{\mu}{2M}$ then $\{x\}_{k \geq 0} \subset B[x^*, \frac{\mu}{2M}]$.

Denote $\delta_k = \frac{M}{\mu} \|x_k - x^*\|$, then we have $\delta_0 = \frac{M}{\mu} \|x_0 - x^*\| \leq \frac{1}{2}$, and if we combine this with (4.21), we get:

$$\delta_{k+1} \leq \delta_k^2 \quad \forall k \geq 0 \quad (4.22)$$

And if we proceed by recurrence, we get:

$$\begin{aligned} \delta_1 &\leq \delta_0^2 \leq \left(\frac{1}{2}\right)^2 \\ \delta_2 &\leq \delta_1^2 \leq \left(\frac{1}{2}\right)^4 \\ &\vdots \\ \delta_k &\leq \left(\frac{1}{2}\right)^{2^k} \quad \forall k \geq 0 \end{aligned} \quad (4.23)$$

$$\Rightarrow \|x_k - x^*\| \leq \frac{\mu}{M} \left(\frac{1}{2}\right)^{2^k} \quad (4.24)$$

Let $T(\varepsilon) = \inf\{k \in \mathbb{N} : \|x_k - x^*\| \leq \varepsilon\}$ and suppose that $T(\varepsilon) \geq 1$. Then using the convergence rate (4.24), we can state the maximal number of iterations:

$$\varepsilon \leq \|x_{T(\varepsilon)-1} - x^*\| \leq \frac{\mu}{M} \left(\frac{1}{2}\right)^{2^{T(\varepsilon)-1}} \quad (4.25)$$

$$2^{2^{T(\varepsilon)-1}} \leq \frac{\mu}{M} \varepsilon^{-1} \quad (4.26)$$

$$\Rightarrow T(\varepsilon) \leq \log_2(\log_2(\frac{\mu}{M} \varepsilon^{-1}))$$

4.3 Quasi Newton methods

4.3.1 SR1 Update

One step of a Quasi-Newton method is given by:

$$x_{k+1} = x_k - B_k \nabla f(x_k) \quad (4.27)$$

With $B_k \in \mathbb{R}^{n \times n}$, symmetric and non-singular

Suppose that $x_k \rightarrow x^*$ when $k \rightarrow \infty$, and that $\nabla^2 f(x_k) \succeq \mu I_n$ with $\mu \geq 0$

We want the condition on B_k to have a Super Linear convergence (1.7) of the Quasi-Newton method. So let's assume that $f \in \mathcal{C}_M^{2,2}(\mathbb{R}^n)$.

Then,

$$\|\nabla^2 f(x_{k+1} - \nabla^2 f(x_k))\| \leq M \|x_{k+1} - x_k\| \quad (4.28)$$

GOOD LABEL ?

$$\|\nabla f(x_{k+1} - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1} - x_k))\| \leq \frac{M}{2} \|x_{k+1} - x_k\|^2 \quad (4.29)$$

Therefore

$$\begin{aligned} \nabla f(x_{k+1}) &= \nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1} - x_k) \\ &\quad + \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k) \end{aligned} \quad (4.30)$$

Using the relation (4.27) we get:

$$\begin{aligned} \nabla f(x_{k+1}) &= \nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1} - x_k) \\ &\quad - B_k^{-1}(x_{k+1} - x_k) \\ &\quad + \nabla^2 f(x_k)(x_{k+1} - x_k) \\ &= \nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1} - x_k) \\ &\quad - \left(B_k^{-1} - \nabla^2 f(x^*) \right) (x_{k+1} - x_k) \\ &\quad + \left(\nabla^2 f(x_k) - \nabla^2 f(x^*) \right) (x_{k+1} - x_k) \\ \|\nabla f(x_{k+1})\| &\leq \|\nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1} - x_k)\| \\ &\quad + \left\| \left(B_k^{-1} - \nabla^2 f(x^*) \right) (x_{k+1} - x_k) \right\| \\ &\quad + \left\| \left(\nabla^2 f(x_k) - \nabla^2 f(x^*) \right) (x_{k+1} - x_k) \right\| \\ &\leq \frac{M}{2} \|x_{k+1} - x_k\|^2 + M \|x_k - x^*\| \|x_{k+1} - x_k\| \\ &\quad + \left\| \left(B_k^{-1} - \nabla^2 f(x_k) \right) (x_{k+1} - x_k) \right\| \end{aligned} \quad (4.31)$$

On the line before we used (4.28) and (4.29). And so we can write:

$$\frac{\|\nabla f(x_{k+1})\|}{\|x_{k+1} - x_k\|} \leq \frac{M}{2} \|x_{k+1} - x_k\| + M \|x_k - x^*\| + \frac{\left\| \left(B_k^{-1} - \nabla^2 f(x_k) \right) (x_{k+1} - x_k) \right\|}{\|x_{k+1} - x_k\|} \quad (4.32)$$

From now on, suppose that this condition (Dimis-Mori condition) is true:

$$\lim_{k \rightarrow \infty} \frac{\left\| \left(B_k^{-1} - \nabla^2 f(x_k) \right) (x_{k+1} - x_k) \right\|}{\|x_{k+1} - x_k\|} = 0 \quad (4.33)$$

Under this condition and by (4.32), we have:

$$\lim_{k \rightarrow \infty} \frac{\|\nabla f(x_{k+1})\|}{\|x_{k+1} - x_k\|} = 0 \quad (4.34)$$

As $\|x_{k+1} - x_k\| \rightarrow 0$, we conclude that $\lim_{x \rightarrow \infty} \|\nabla f(x_{k+1})\| = 0$ and so $\|\nabla f(x^*)\| = 0 \Rightarrow \nabla f(x^*) = 0$. (x^* is a stationary point of f)

We have $\nabla^2 f(x^*) \succeq \mu I_n$ and given $y \in \mathbb{R}^n$, we have:

$$\begin{aligned} \nabla^2 f(y) - \nabla^2 f(x^*) &\succeq -M \|y - x^*\| I_n \\ \nabla^2 f(y) &\succeq (\mu - M \|y - x^*\|) I_n \end{aligned} \quad (4.35)$$

Thus, if $-M \|y - x^*\| \geq -\frac{\mu}{2}$ then $\nabla^2 f(y) \succeq \frac{\mu}{2} I_n$.

Since $x_k \rightarrow x^*$, there exists $k_0 \in \mathbb{N}$ such that $\|x_{k+1} - x^*\| \leq \frac{\mu}{2M} \forall k \geq k_0$. Thus for any $\tau \in [0, 1]$:

$$\|x^* + \tau(x_{k+1} - x^*) - x^*\| \leq \frac{\mu}{2M}, \quad \forall k \geq k_0 \quad (4.36)$$

and so $\nabla^2 f(x^* + \tau(x_{k+1} - x^*)) \succeq \frac{\mu}{2} I_n \forall k \geq k_0$.

$$\begin{aligned}
\|x_{k+1} - x^*\| \|\nabla f(x_{k+1})\| &\geq (x_{k+1} - x^*)^T \nabla f(x_{k+1}) \\
&= (x_{k+1} - x^*)^T (\nabla f(x_{k+1}) - \nabla f(x^*)) \\
&= (x_{k+1} - x^*)^T \int_0^1 \nabla^2 f(x^* + \tau(x_{k+1} - x^*)) (x_{k+1} - x^*) d\tau \\
&\geq \int_0^1 (x_{k+1} - x^*)^T \frac{\mu}{2} I_n (x_{k+1} - x^*) d\tau \\
&= \frac{\mu}{2} \|x_{k+1} - x^*\|^2
\end{aligned} \tag{4.37}$$

$$\|\nabla f(x_{k+1})\| \geq \frac{\mu}{2} \|x_{k+1} - x^*\| \tag{4.38}$$

Let $\rho_k = \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|}$ then, using (6.5), we obtain:

$$\begin{aligned}
\frac{\|\nabla f(x_{k+1})\|}{\|x_{k+1} - x_k\|} &\geq \frac{(\frac{\mu}{2}) \|x_{k+1} - x^*\|}{\|x_{k+1} - x_k\|} \\
&\geq \frac{(\frac{\mu}{2}) \|x_{k+1} - x^*\|}{\|x_{k+1} - x^*\| + \|x_k - x^*\|} \\
&= \frac{(\frac{\mu}{2}) \rho_k}{\rho_k + 1}
\end{aligned} \tag{4.39}$$

Combining (4.39) and (4.32), we get:

$$\frac{\mu}{2} \frac{\rho_k}{\rho_k + 1} \leq \frac{M}{2} \|x_{k+1} - x_k\| + M \|x_k - x^*\| + \frac{\| (B_k^{-1} - \nabla^2 f(x^*)) (x_{k+1} - x_k) \|}{\|x_{k+1} - x_k\|} \tag{4.40}$$

Since the right hand side goes to zero when $k \rightarrow +\infty$, then we have: **IDK how to write that**

$$\begin{aligned}
\lim_{k \rightarrow \infty} \frac{\rho_k}{1 + \rho_k} &= 0 \\
\lim_{k \rightarrow \infty} \frac{1}{\frac{1}{\rho_k} + 1} &= 0 \\
\Rightarrow \lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} &\Rightarrow \lim_{k \rightarrow \infty} \rho_k = 0
\end{aligned} \tag{4.41}$$

Suppose that $n = 1$, then the quasi-newton update is writed:

$$x_{k+1} = x_k - b_k f'(x_k), \quad k \geq 0 \tag{4.42}$$

with $b_k \in \mathbb{R}$. We want $b_k \approx f''(x_k)^{-1}$ and by finite difference we can express it like that $b_k^{-1} \approx \frac{f'(x_{k-1}+h) - f'(x_{k-1})}{h}$. And with $h = x_k - x_{k-1}$, we can define:

$$b_k^{-1} = \frac{f'(x_k) - f'(x_{k-1})}{x_k - x_{k-1}} \tag{4.43}$$

Thus if $x_k \rightarrow x^*$ then:

$$\lim_{k \rightarrow \infty} \frac{|(b_k^{-1} - f''(x^*))(x_k - x_{k-1})|}{|x_k - x_{k-1}|} = 0 \tag{4.44}$$

Because we can notice that:

$$\frac{|(b_k^{-1} - f''(x^*))(x_k - x_{k-1})|}{|x_k - x_{k-1}|} = |b_k^{-1} - f''(x_{k-1})| + |f''(x_{k-1}) - f''(x^*)| \quad (4.45)$$

Since $x_k \rightarrow x^*$, we have $h = x_k - x_{k-1}$ and so:

$$b_k^{-1} = \frac{f'(x_k) - f'(x_{k-1})}{x_k - x_{k-1}} \rightarrow f''(x_{k-1}) \quad (4.46)$$

Thus, $\lim_{k \rightarrow \infty} |b_k^{-1} - f''(x_k)| = 0$.

Assuming that f'' is continuous, we have $\lim_{k \rightarrow \infty} |f''(x_k) - f''(x^*)| = 0$.

If we define $S_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = f'(x_k) - f'(x_{k-1})$ and knowing (4.43), we can write:

$$\begin{aligned} b_k(f'(x_k) - f'(x_{k-1})) &= x_k - x_{k-1} \\ b_k y_{k-1} &= S_{k-1} \end{aligned} \quad (4.47)$$

This suggests that for $n > 1$, we should define the secant condition, B_k such that:

$$B_k y_{k-1} = S_{k-1} \quad (4.48)$$

Lets define $f(x) = \frac{1}{2} \|Ax - b\|^2 = \frac{1}{2} x^T A^T A x - (A^T b)^T x + \frac{1}{2} b^T b$. If A is full rank then f is a strongly convex quadratic function. And we have $\nabla f(x_k) = A^T A x_k - A^T b$. Then,

$$y_{k-1} = \nabla f(x_k) - \nabla f(x_{k-1}) = A^T A (x_k - x_{k-1}) = \nabla^2 f(x_k) S_{k-1} \quad (4.49)$$

And so

$$\nabla^2 f(x_k) y_{k-1} = S_{k-1} \quad (4.50)$$

Therefore, $\nabla^{-2} f$ satisfies the secant condition (4.48), when f is a strongly convex quadratic function. Thus it is reasonable to require the secant for any approximation to $\nabla^{-2} f(x_k)$.

Now, how can we compute B_k such that it satisfies the secant condition (4.48)?

Given a matrix B_{k-1} , our goal is to find a perturbation matrix $P_{k-1} \in \mathbb{R}^{n \times n}$ such that:

$$(B_{k-1} + P_{k-1}) y_{k-1} = S_{k-1} \quad (4.51)$$

If we get such P_{k-1} , we can define $B_k = B_{k-1} + P_{k-1}$, which would satisfy the secant condition (4.48).

For that we need at least n degrees of freedom and a symmetric matrix, so it is natural to try:

$$P_{k-1} = v_{k-1} v_{k-1}^T \quad v_{k-1} \in \mathbb{R}^n \quad (4.52)$$

So we get:

$$(B_{k-1} + v_{k-1} v_{k-1}^T) y_{k-1} = S_{k-1} \quad (4.53)$$

By algebraic manipulations, we get:

$$\begin{aligned} (v_{k-1}^T y_{k-1}) v_{k-1} &= S_{k-1} - B_{k-1} y_{k-1} \\ v_{k-1} &= \frac{S_{k-1} - B_{k-1} y_{k-1}}{\beta} \quad \text{for } \beta = v_{k-1}^T y_{k-1} \end{aligned} \quad (4.54)$$

Combining the two previous equations, we get:

$$\begin{aligned} \left(\frac{1}{\beta} (S_{k-1} - B_{k-1}y_{k-1})^T y_{k-1} \right) \frac{1}{\beta} (S_{k-1} - B_{k-1}y_{k-1}) &= S_{k-1} - B_{k-1}y_{k-1} \\ \frac{1}{\beta^2} (S_{k-1} - B_{k-1}y_{k-1})^T y_{k-1} &= 1 \end{aligned} \quad (4.55)$$

We can isolate β :

$$\beta = \sqrt{(S_{k-1} - B_{k-1}y_{k-1})^T y_{k-1}} \quad (4.56)$$

Combining (4.54) and (4.56), we get:

$$v_{k-1} = \frac{S_{k-1} - B_{k-1}y_{k-1}}{\sqrt{(S_{k-1} - B_{k-1}y_{k-1})^T y_{k-1}}} \quad (4.57)$$

This leads us to the following update for B_k :

$$\begin{aligned} B_k &= B_{k-1} + v_{k-1}v_{k-1}^T \\ &= B_{k-1} + \frac{(S_{k-1} - B_{k-1}y_{k-1})(S_{k-1} - B_{k-1}y_{k-1})^T}{(S_{k-1} - B_{k-1}y_{k-1})^T y_{k-1}} \end{aligned} \quad (4.58)$$

This is called the **SR1 update** (symmetric rank 1 update).

4.3.2 BFGS Update

Lets take back $B_{k+1}y_k = s_k$ and defining $H_{k+1} = B_{k+1}^{-1} \approx \nabla^2 f(x_{k+1})$, we get $H_{k+1}s_k = y_k$.

The idea is to find a rank 2 update that consist of finding $a, b \in \mathbb{R}$ and $v, u \in \mathbb{R}^n$ such that:

$$(H_k + auu^T + bv v^T) s_k = y_k \quad (4.59)$$

Noticing that $u^T s_k$ and $v^T s_k$ are scalars, we can impose that:

$$\begin{cases} a(u^T s_k)u = -H_k s_k \\ b(v^T s_k)v = y_k \end{cases} \quad (4.60)$$

It suggests that we should take $a = \frac{1}{u^T s_k}$ and $b = \frac{1}{v^T s_k}$. Which gives us:

$$\begin{cases} u = -H_k s_k \\ v = y_k \end{cases} \quad (4.61)$$

Combining the two equations, we get:

$$H_{k+1} = H_k - \frac{H_k s_k s_k^T H_k}{s_k^T H_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} \quad (4.62)$$

Using linear algebra, we can compute:

$$B_{k+1} = H_{k+1}^{-1} \\ = \left(I - \rho_k s_k y_k^T \right) B_k \left(I - \rho_k y_k s_k^T \right) + \rho_k s_k s_k^T \text{ with } \rho_k = \frac{1}{y_k^T s_k}$$

Remarks:

- If $B_k \succ 0$ and $s_k^T y_k > 0$ then $B_{k+1} \succ 0$.
- If $B_k \succ 0$ and $d_k = -B_k \nabla f(x_k)$, then

$$\langle \nabla f(x_k), d_k \rangle = -\langle \nabla f(x_k), B_k \nabla f(x_k) \rangle < 0 \quad (4.63)$$

and so d_k is a descent direction for f at x_k .

- The LBFGS is a low memory of BFGS, that does not require the storage of the matrices B_k . Given a vector $v \in \mathbb{R}^n$, it computes $B_k v$, which is all that we need to implement QN method.

Constrained nonlinear programming problems

Consider the constrained problem

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad c_i(x) = 0, \quad i \in \{1, \dots, m\} \quad (5.1)$$

where $f, c_i : \mathbb{R}^n \rightarrow \mathbb{R}$ are \mathcal{C}^1 and there exists a \hat{x} such that $c_i(\hat{x}) = 0$.

A natural approach to solve this problem is to consider the related unconstrained problem in which we try to minimize $f(x)$ plus a term that penalizes the violation of the constraints (quadratic penalty function).

$$\min_{x \in \mathbb{R}^n} Q_\sigma(x) \equiv f(x) + \frac{\sigma}{2} \|c(x)\|_2^2 \quad (5.2)$$

For the problem (5.1), we would like to find a KKT point x^* for which there exists $\lambda^* \in \mathbb{R}^m$ such that:

$$\begin{cases} \nabla f(x^*) - \sum_{i=1}^m \lambda_i^* \nabla c_i(x^*) = 0 & (\text{stationarity}) \\ c(x^*) = 0 & (\text{feasibility}) \end{cases} \quad (5.3)$$

In practice, we are happy if we can find an $(\varepsilon_1, \varepsilon_2)$ -KKT point for (5.1), i.e. a point x^+ such that there exists λ^+ with:

$$\begin{cases} \|\nabla f(x^+) - \sum_{i=1}^m \lambda_i^+ \nabla c_i(x^+)\| \leq \varepsilon_1 \\ \|c(x^+)\| \leq \varepsilon_2 \end{cases} \quad (5.4)$$

Let us relate (5.2) and (5.1). Notice that:

$$\begin{aligned} \|\nabla Q_\sigma(x)\| &= \|\nabla f(x) + \sigma \mathbf{J}_c(x)^T c(x)\| \\ &= \|\nabla f(x) + \sigma \sum_{i=1}^m c_i(x) \nabla c_i(x)\| \\ &= \|\nabla f(x) - \sum_{i=1}^m \lambda_i^+ \nabla c_i(x)\| \quad \text{with } \lambda_i^+ = -\sigma c_i(x^+) \end{aligned} \quad (5.5)$$

Therefore, if $\|\nabla Q_\sigma(x^+)\| \leq \varepsilon_1$, then there exists $\lambda^+ \in \mathbb{R}^m$, $\lambda^+ = -\sigma c(x^+)$ such that $\|\nabla f(x^+) - \sum_{i=1}^m \lambda_i^+ \nabla c_i(x^+)\| \leq \varepsilon_1$.

Given $\bar{x} \in \mathbb{R}^n$, suppose that we compute x^+ such that

$$\begin{aligned}
Q_\sigma(x^+) &\leq Q_\sigma(\bar{x}) \\
f(x^+) + \frac{\sigma}{2} \|c(x^+)\|^2 &\leq f(\bar{x}) + \frac{\sigma}{2} \|c(\bar{x})\|^2 \\
\frac{\sigma}{2} \|c(x^+)\|^2 &\leq f(\bar{x}) - f(x^+) + \frac{\sigma}{2} \|c(\bar{x})\|^2 \\
\|c(x^+)\|^2 &\leq \frac{2}{\sigma} (f(\bar{x}) - f(x^+)) + \|c(\bar{x})\|^2
\end{aligned} \tag{5.6}$$

If $f(x) \geq f_{low} \quad \forall x \in \mathbb{R}^n$, we get $\|c(x^+)\|^2 \leq \frac{2}{\sigma} (f(\bar{x}) - f_{low}) + \|c(\bar{x})\|^2$.

If $\|c(\bar{x})\| \leq \frac{\varepsilon_2}{\sqrt{2}}$ and $\sigma \geq \frac{4}{\varepsilon_2^2} (f(\bar{x}) - f_{low})$, then $\|c(x^+)\|^2 \leq \varepsilon_2^2$ and so $\|c(x^+)\| \leq \varepsilon_2$.

In summary, if we have $\bar{x} \in \mathbb{R}^n$ such that $\|c(\bar{x})\| \leq \frac{\varepsilon_2}{\sqrt{2}}$, and using a method for unconstrained optimization (e.g. GM), we compute x^+ with

$$Q_\sigma(x^+) \leq Q_\sigma(\bar{x}) \quad \text{and} \quad \|\nabla Q_\sigma(x^+)\| \leq \varepsilon_1 \tag{5.7}$$

for $\sigma \geq \frac{4}{\varepsilon_2^2} (f(\bar{x}) - f_{low})$, then x^+ is a $(\varepsilon_1, \varepsilon_2)$ -KKT point for the unconstrained problem (5.1).

Example:

$$\min f(x) \quad \text{such that} \quad Ax = b \quad \text{and} \quad \nabla f(x) \quad L_f\text{-Lipschitz} \tag{5.8}$$

$$\begin{aligned}
Q_\sigma(x) &= f(x) + \frac{\sigma}{2} \|Ax - b\|_2^2 \\
\nabla Q_\sigma(x) &= \nabla f(x) + \sigma A^T (Ax - b) \\
\nabla^2 Q_\sigma(x) &= \nabla^2 f(x) + \sigma A^T A
\end{aligned} \tag{5.9}$$

so ∇Q_σ is L_{Q_σ} -Lipschitz with $L_{Q_\sigma} = L_f + \sigma \|A^T A\|$ so GM takes $\mathcal{O}(L_{Q_\sigma} \varepsilon_1^{-2})$ iterations to find x^+ with $\|\nabla Q_\sigma(x^+)\| \leq \varepsilon_1$, and $Q_\sigma(x^+) \leq Q_\sigma(\bar{x})$, if initialized with \bar{x} . Since $L_{Q_\sigma} = \mathcal{O}(\sigma)$ and $\sigma = \mathcal{O}(\varepsilon_2^{-2})$, we get a complexity bound of $\mathcal{O}(\varepsilon_1^{-2} \varepsilon_2^{-2})$.

Tips and Tricks

1. Approximation of the max:

$$\max\{z, 0\} = \frac{z + |z|}{2} = \frac{z + \sqrt{z^2}}{2} \approx \frac{z + \sqrt{z^2 + \delta}}{2} \quad (6.1)$$

2.

$$ab \leq \frac{a^2 + b^2}{2} \quad (6.2)$$

3.

$$(a + b)^2 \leq 2a^2 + 2b^2 \quad (6.3)$$

4. V-trick:

$$\langle xv, v \rangle \leq \|x\| \|v\|^2 \quad (6.4)$$

5. Triangular inequality by the minimizer:

$$\|x_{k+1} - x_k\| \leq \|x_{k+1} - x^*\| + \|x_k - x^*\| \quad (6.5)$$