



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	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	Accelerated Gradient Method	19
6.1	Derivation of the algorithm	19
6.2	Accelerated Proximal Gradient Method	21
7	Path following Interior Point Method	23
7.1	Self concordant functions	23
7.2	Path-following Interior-point Method	24
7.3	Intermediate Newton method	25
7.4	Path-following Interior point Algorithm	27
8	Tips and Tricks	29

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

	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 (8.3) in chapter 8,

$$\implies \|\nabla f(x_k)\|^2 \leq \frac{L^2 n}{2} h_k^2 + 2\|g_{h_k}(x_k)\|^2 \quad (3.8)$$

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

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

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

$$\frac{T(\varepsilon)}{8\sigma} \varepsilon^2 \leq f(x_0) - f_{low} + \frac{5\sigma n}{16} \sum_{k=0}^{T(\varepsilon)-1} h_k^2 \quad (3.11)$$

If $\{h_k^2\}_{k \geq 0}$ is summable

$$\implies T(\varepsilon) \leq 8\sigma \left(f(x_0) - f_{low} + \frac{5\sigma n}{16} \sum_{k=0}^{T(\varepsilon)-1} h_k^2 \right) \varepsilon^2 = \mathcal{O}(\varepsilon^2) \quad (3.12)$$

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:

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

By the properties of the semi-definite matrices, and the trick (8.4), 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,

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

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:

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

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

\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

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

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

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 $\{x_k\}_{k \geq 0} \subset B[x^*, \frac{\mu}{2M}]$.

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:

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

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

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 us [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:

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

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

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

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

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

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

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 $\|\nabla f(x^*)\| = 0 \Rightarrow \nabla f(x^*) = 0$, meaning that x^* is a stationary point of $f(\cdot)$.

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

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

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

Let $\rho_k = \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|}$ then, using (8.5), we obtain:

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

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

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

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

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

For $n = 1$, the Quasi-Newton update is written:

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

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:

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

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

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

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:

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

Let us 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 consists in finding $a, b \in \mathbb{R}$ and $v, u \in \mathbb{R}^n$ such that:

$$(H_k + a u u^T + b v v^T) s_k = y_k \quad (4.59)$$

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:

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

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 at least 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 that¹:

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

¹ $J_c(\cdot)$ is the Jacobian of $c(\cdot)$.

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

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

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

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

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).

Algorithm 1 Quadratic Penalty Method

- 1: **Input:** $\varepsilon_1, \varepsilon_2 \in (0, 1)$, $x_0 \in \mathbb{R}^n$ such that $\|c(x_0)\|_2 \leq \frac{\varepsilon_2}{\sqrt{2}}$, $\sigma_0 > 0$
- 2: $k = 0$
- 3: **while** $\|c(x_{k+1})\| > \varepsilon_1$ **do**
- 4: Compute $x_{k+1} \in \mathbb{R}^n$ as an approximate solution to

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} Q_{\sigma_k}(x) \\ \text{such that} \quad &Q_{\sigma_k}(x_{k+1}) \leq Q_{\sigma_k}(x_0) \\ \text{and} \quad &\|\nabla Q_{\sigma_k}(x_{k+1})\| \leq \varepsilon_2 \end{aligned} \quad (5.8)$$

- 5: $\sigma_{k+1} \leftarrow 2\sigma_k$
 - 6: $k \leftarrow k + 1$
 - 7: **end while**
-

→ Note: We can compute x_{k+1} satisfying (5.8) by using any monotone optimization method starting from:

$$x_k^* = \arg \min \{Q_{\sigma_k}(x_0), Q_{\sigma_k}(x_k)\} \quad (5.9)$$

- For a constrained problem of the form $\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t. } c_i \leq 0 \quad i = 0, \dots, m$, we can add slack variables to obtain an equivalent equality constrained problem:

$$\begin{aligned} &\min_{x \in \mathbb{R}^n, s \in \mathbb{R}^m} f(x) \\ \text{s.t. } &c_i(x) + s_i^2 = 0 \quad i = 1, \dots, m \end{aligned} \quad (5.10)$$

Accelerated Gradient Method

6.1 Derivation of the algorithm

$$\min_{x \in \mathbb{R}^n} f(x) \quad (6.1)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex, ∇f is L -Lipschitz and has a minimizer x^* . The Accelerated Gradient Method combines present and past information to obtain a point y_k (prediction) and then perform a gradient step using this point as reference point.

$$\begin{cases} y_k = (1 - \gamma_k)x_k + \gamma_k v_k, & \gamma_k \in (0, 1) \\ x_{k+1} = x_k - \frac{1}{L} \nabla f(y_k) \end{cases} \quad (6.2)$$

We will identify ways to define v_k and γ_k based on the following guiding inequalities:

$$\begin{aligned} v_k &= \arg \min_{x \in \mathbb{R}^n} \Psi_k(x) \\ \Psi_k(x) &\leq A_k f(x) + \frac{1}{2} \|x - x_0\|^2 \\ A_k f(x_k) &\leq \min_{x \in \mathbb{R}^n} \Psi_k(x) \equiv \Psi_k^*, \quad A_k \geq 0 \\ A_k &\geq c(k-1)^2 \quad \forall k \geq 2 \end{aligned} \quad (6.3)$$

Assuming the 3 last guiding inequalities (6.3) hold, we have:

$$\begin{aligned} A_k f(x_k) &\leq \min_{x \in \mathbb{R}^n} \Psi_k(x) \\ &\leq \Psi_k(x^*) \\ &\leq A_k f(x^*) + \frac{1}{2} \|x^* - x_0\|^2 \\ (f(x_k) - f(x^*)) &\leq \frac{\|x_k - x^*\|^2}{2A_k} \quad \forall k \geq 2 \\ &\leq \frac{\|x_k - x^*\|^2}{2C(k-1)^2} = \mathcal{O}(k^{-2}) = \mathcal{O}(\varepsilon^{-1/2}) \quad \forall k \geq 2 \end{aligned} \quad (6.4)$$

If we take $A_0 = 0$ and $\Psi_0(x) = \frac{1}{2} \|x - x_0\|^2$, then the second inequality from (6.3) is true for $k = 0$. Let us assume the inequality is true for some $k \geq 0$. Looking at the case $k = 1$, it appears that we can define:

$$\Psi_{k+1}(x) = \Psi_k(x) + b_k (f(y_k) + \langle \nabla f(y_k), x - y_k \rangle) \quad (6.5)$$

with $b_k > 0$ (to be determined).

Suppose that the inequality holds for $k \geq 0$. Then, by the convexity of f and doing an induction assumption:

$$\begin{aligned}\Psi_{k+1}(x) &\leq \Psi_k(x) + b_k f(x) \\ &\leq A_k f(x) + \frac{1}{2} \|x - x_0\|^2 + b_k f(x) \\ &= (A_k + b_k) f(x) + \frac{1}{2} \|x - x_0\|^2\end{aligned}\tag{6.6}$$

Therefore, if we define $A_{k+1} = A_k + b_k$, then the second inequality of (6.3) will also hold for $k + 1$. Regarding of the third inequality of (6.3), notice that:

$$\begin{aligned}A_0 f(x_0) = 0 &= \min_{x \in \mathbb{R}^n} \frac{1}{2} \|x - x_0\|^2 \\ &= \min_{x \in \mathbb{R}^n} \Psi_0(x)\end{aligned}\tag{6.7}$$

It holds for $k = 0$, suppose that it still holds for $k \geq 0$. We want to show that it is also true for $k + 1$. Notice that:

$$\begin{aligned}\Psi_1 &= \frac{1}{2} \|x - x_0\|^2 + b_0 (f(y_0) + \langle \nabla f(y_0), x - y_0 \rangle) \\ \Psi_2 &= \frac{1}{2} \|x - x_0\|^2 + \sum_{i=0}^1 b_0 (f(y_i) + \langle \nabla f(y_i), x - y_i \rangle) \\ &\vdots \\ \Psi_k &= \frac{1}{2} \|x - x_0\|^2 + \sum_{i=0}^{k-1} b_0 (f(y_i) + \langle \nabla f(y_i), x - y_i \rangle)\end{aligned}\tag{6.8}$$

Thus, $\Psi_k(x)$ is a μ -strongly convex function with $\mu = 1$. Therefore:

$$\begin{aligned}\Psi_k(x) &\geq \Psi_k(v_k) + \frac{1}{2} \|v_k - x_0\|^2 \\ &= \min_{x \in \mathbb{R}^n} \Psi_k(x) + \frac{1}{2} \|v_k - x_0\|^2 \\ &\geq A_k f(x_k) + \frac{1}{2} \|v_k - x_0\|^2\end{aligned}\tag{6.9}$$

And so:

$$\begin{aligned}\min_x \Psi_{k+1}(x) &= \min_x \Psi_k + b_k (f(y_k) + \langle \nabla, x - y_k \rangle) \\ &\geq \min_x A_k f(x_k) + \frac{1}{2} \|v_k - x_0\|^2 + b_k (f(y_k) + \langle \nabla, x - y_k \rangle) \\ &\geq \min_x A_k (f(x_k) + \langle \nabla, x_k - y_k \rangle) + b_k (f(y_k) + \langle \nabla, x - y_k \rangle) \\ &\geq (A_k + b_k) f(y_k) + \langle \nabla f(y_k), A_k x_k + b_k x - A_{k+1} y_k \rangle + \frac{1}{2} \|v_k - x_0\|^2 \\ &\geq (A_{k+1}) f(y_k) + \langle \nabla f(y_k), A_k x_k + b_k x - A_{k+1} y_k \rangle + \frac{1}{2} \|v_k - x_0\|^2\end{aligned}\tag{6.10}$$

To make things consistent, let us impose

$$A_k x_k - A_{k+1} y_k + b_k x = b_k (x - v_k) \iff y_k = \frac{A_k}{A_{k+1}} x_k + \frac{b_k}{A_{k+1}} v_k \quad (6.11)$$

And so we can continue equation (6.10):

$$\min_{x \in \mathbb{R}^n} \Psi_{k+1}(x) A_{k+1} \min_{x \in \mathbb{R}^n} \geq f(y_k) + \langle \nabla f(y_k), \gamma_k(x - v_k) \rangle + \frac{1}{2A_{k+1}\gamma_k^2} \|\gamma_k(v_k - x)\|^2 \quad (6.12)$$

To verify the Lipschitz condition, we impose

$$\frac{1}{2A_{k+1}\gamma_k^2} = \frac{L}{2} \iff b_k^2 - \frac{1}{L}b_k - \frac{A_k}{L} = 0 \implies b_k = \frac{1 + \sqrt{1 + 4A_k L}}{2L} \quad (6.13)$$

From all that have been computed previously, we can find a bound in terms of iterations needed. If $x^* = \arg \min f(x)$, we have

$$\begin{aligned} A_k f(x_k) &\leq \min_{x \in \mathbb{R}^n} \Psi_k(x) \leq \Psi_k(x^*) \leq A_k f(x^*) + \frac{1}{2} \|x^* - x_k\|^2 \\ \Rightarrow A_k (f(x_k) - f(x^*)) &\leq \frac{1}{2} \|x^* - x_k\|^2 \\ \Rightarrow f(x_k) - f(x^*) &\leq \frac{1}{2A_k} \|x^* - x_k\|^2 \end{aligned} \quad (6.14)$$

From the relation $A_{k+1} = A_k + b_k$ and the definition of b_k , we can show that $A_k \geq C(k-1)^2$ with $C > 0$ and $k \geq 2$. Thus, we get

$$f(x_k) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2C(k-1)^2} = \mathcal{O}(1/k^2) \quad \forall k \geq 1 \quad (6.15)$$

A recap is given in algorithm 2.

6.2 Accelerated Proximal Gradient Method

In this section, we consider the minimisation of a function over a nonempty, closed and convex set Ω . We decompose the objective function F into a smooth and a possibly non smooth part:

$$\min_{x \in \Omega \subseteq \mathbb{R}^n} F(x) \equiv f(x) + \varphi(x) \quad (6.20)$$

The accelerated proximal gradient method consists in using the proximal operator of the non smooth part φ to define x_{k+1} :

Theorem 6.1. If $\{x_k\}_{k \geq 0}$ is generated by the accelerated proximal gradient method, then

$$F(x_k) - F(x^*) \leq \frac{8L\|x_0 - x^*\|^2}{(k-1)^2} \quad \forall k \geq 2 \quad (6.25)$$

Algorithm 2 Accelerated Gradient Method

1: **Input:** Given $x_0 \in \mathbb{R}^n$, define $\Psi_0(x) = \frac{1}{2}\|x - x_0\|^2$, $A_0 = 0$, $b_0 = 0$, $k = 0$;

2: Compute

$$b_k = \frac{1 + \sqrt{1 + 4A_k L}}{2L} > 0; \quad (6.16)$$

3: Set $\gamma_k = \frac{b_k}{A_{k+1}} \in (0, 1]$ and compute $y_k = (1 - \gamma_k)x_k + \gamma_k v_k$;

4: Set

$$x_{k+1} = \arg \min_{x \in \mathbb{R}^n} f(y_k) + \langle \nabla f(y_k), x - y_k \rangle + \frac{L}{2} \|x - y_k\|^2 \quad (6.17)$$

and $A_{k+1} = A_k + b_k$;

5: Define

$$\Psi_{k+1}(x) = \Psi_k(x) + b_k (f(y_k) + \langle \nabla f(y_k), x - y_k \rangle) \quad \forall x \in \mathbb{R}^n \quad (6.18)$$

and set

$$v_{k+1} = \arg \min_{x \in \mathbb{R}^n} \Psi_{k+1}(x) \quad (6.19)$$

6: $k \leftarrow k + 1$ and go back to Step 1;

Algorithm 3 Accelerated Proximal Gradient Method

1: **Input:** Given $x_0 \in \text{dom}F$, define $\Psi_0(x) = \frac{1}{2}\|x - x_0\|^2$, $A_0 = 0$, $b_0 = 0$, $k = 0$;

2: Compute

$$b_k = \frac{1 + \sqrt{1 + 4A_k L}}{2L} > 0; \quad (6.21)$$

3: Set $\gamma_k = \frac{b_k}{A_{k+1}} \in (0, 1]$ and compute $y_k = (1 - \gamma_k)x_k + \gamma_k v_k$;

4: Set

$$x_{k+1} = \text{Prox}_{\frac{1}{L}\varphi}(y_k - \frac{1}{L}\nabla f(y_k)) \quad (6.22)$$

and $A_{k+1} = A_k + b_k$;

5: Define

$$\Psi_{k+1}(x) = \Psi_k(x) + b_k (f(y_k) + \langle \nabla f(y_k), x - y_k \rangle) \quad \forall x \in \mathbb{R}^n \quad (6.23)$$

and set

$$v_{k+1} = \arg \min_{x \in \mathbb{R}^n} \Psi_{k+1}(x) \quad (6.24)$$

6: $k \leftarrow k + 1$ and go back to Step 1;

Path following Interior Point Method

7.1 Self concordant functions

7.1.1 Definition

Definition 7.1. Given a convex function $f \in \mathcal{C}^3(\text{dom} f)$, with $\text{dom} f \subseteq \mathbb{R}^n$ open and convex, $f(\cdot)$ is said to be self-concordant with constant M_f when

$$\left| D^3 f(x)[u, u, u] \right| \leq 2M_f \|u\|_x^3 \quad \forall x \in \text{dom} f \quad \forall u \in \mathbb{R}^n \quad (7.1)$$

where $\|u\|_x := \sqrt{\langle \nabla^2 f(x) u, u \rangle}$.

From this definition, we can derive two lemmas:

- Let f_1, f_2 be self-concordant functions with constants M_1 and M_2 respectively. Then, given constants $\alpha, \beta > 0$, the function $f = \alpha f_1 + \beta f_2$ is self-concordant with constant $M_f = \max \left\{ \frac{M_1}{\sqrt{\alpha}}, \frac{M_2}{\sqrt{\beta}} \right\}$.
- Let $f(\cdot)$ be a self-concordant function with constant $M_f \geq 0$. Given $x, y \in \text{dom} f$, we have

$$\|y - x\|_y \geq \frac{\|y - x\|_x}{1 + M_f \|y - x\|_x} \quad (7.2)$$

7.1.2 With μ -strongly convex

As a reminder, a function f is said to be μ -strongly convex if

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 \quad \forall x, y \in \text{dom} f \\ &\Rightarrow \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2 \end{aligned} \quad (7.3)$$

Taking $y = x^* = \arg \min f(x)$, we find

$$\|\nabla f(x)\| \geq \mu \|x - x^*\| \quad \forall x \in \text{dom} f \quad (7.4)$$

after using the Cauchy-Schwarz inequality. This implies that the norm of the gradient tends to 0 as x approaches the minimizer x^* .

We can show that, for a self concordant function f with constant M_f , given $x, y \in \text{dom} f$, we have

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{\|y - x\|_x^2}{1 + M_f \|y - x\|_x} \quad (7.5)$$

Theorem 7.2. Let $f(\cdot)$ be a self-concordant function with constant M_f . Consider $x_f^* = \arg \min_{x \in \text{dom} f} f(x)$. Given $x \in \text{dom} f$, with $\nabla^2 f(x)$ is nonsingular, we have

$$\|x - x^*\|_x \leq \frac{\|\nabla f(x)\|_x^*}{1 - M_f \|\nabla f(x)\|_x^*} \quad (7.6)$$

whenever $M_f \|\nabla f(x)\|_x^* < 1$, with $\|\nabla f(x)\|_x^* = \sqrt{\langle h, \nabla^{-2} f(x) h \rangle}$.

→ Note: $|\langle h, u \rangle| \leq \|h\|_x^* \|u\|_x$ if $\nabla^2 f(x)$ is nonsingular.

7.1.3 Self-concordant barrier

Definition 7.3. Let $F(\cdot)$ be a self-concordant function with constant $M_f = 1$. We say that $F(\cdot)$ is a ν -self-concordant barrier for the set $\overline{\text{dom} F}$ when

$$\langle \nabla F(x), u \rangle^2 \leq \nu \langle \nabla^2 F(x) u, u \rangle \quad x \in \text{dom} F \quad \forall u \in \mathbb{R}^n \quad (7.7)$$

The typical example is $F(x) = -\log(x)$.

→ Note: If $F(\cdot)$ is a ν -self-concordant barrier for the set $\overline{\text{dom} F}$, then $\langle \nabla F(x), y - x \rangle < \nu \forall x, y \in \text{dom} F$.

→ If, in addition, $\nabla^2 F(x)$ is nonsingular, then $\|\nabla F(x)\|_x^* \leq \sqrt{\nu}$.

7.2 Path-following Interior-point Method

Consider the optimization problem

$$\min_{x \in \mathbb{R}^n} f_0(x) \equiv \langle c, x \rangle \quad x \in \Omega \quad (7.8)$$

where $\Omega = \overline{\text{dom} F}$ for some ν -self-concordant barrier F and it is bounded. From these assumptions, it follows from the Weierstraß theorem that it has a solution x^* .

The barrier strategy consists in solving the problem iteratively by solving unconstrained optimization problems of the form

$$\min_{x \in \text{dom} F} t f_0(x) + F(x) \quad t > 0 \quad (7.9)$$

Let us denote $f(t; x) \equiv t \langle c, x \rangle + F(x)$, and $x^*(t) = \arg \min_{x \in \text{dom} F} f(t; x)$, which we call the central path function. Then,

$$\nabla_x f(t; x^*(t)) = t c + \nabla F(x^*(t)) = 0 \implies c = -\frac{1}{t} \nabla F(x^*(t)) \quad (7.10)$$

Consequently,

$$f_0(x^*(t)) - f_0(x) = \langle c, x^*(t) - x \rangle = \frac{1}{t} \langle \nabla F(x^*(t)), x^* - x^*(t) \rangle < \frac{\nu}{t} \quad (7.11)$$

The last inequality following equation (7.5). This means that

$$\lim_{t \rightarrow \infty} f_0(x^*(t)) = f_0(x^*) \quad (7.12)$$

And in particular, for $\epsilon > 0$, if $t \geq \nu\epsilon^{-1}$, then

$$f_0(x^*(t)) - f_0(x^*) < \epsilon \quad (7.13)$$

But, since $x^*(t)$ is not computable, one way to get an implementable method is to compute $\bar{x}(t)$ such that

$$\|\nabla_x f(t; \bar{x}(t))\|_x^* \leq \beta \quad \beta \in (0, 1) \quad (7.14)$$

This implies

$$\begin{aligned} f_0(\bar{x}(t)) - f_0(x^*) &= f_0(\bar{x}(t)) - f_0(x^*(t)) - (f_0(x^*) - f_0(x^*(t))) \\ &< \frac{\nu}{t} + f_0(\bar{x}(t)) - f_0(x^*(t)) \\ &= \frac{\nu}{t} + \frac{1}{t} \langle t\mathbf{c}, \bar{x}(t) - x^*(t) \rangle \\ &= \frac{\nu}{t} + \frac{1}{t} \langle \nabla_x f(t; \bar{x}(t)) - \nabla F(\bar{x}(t)), \bar{x}(t) - x^*(t) \rangle \end{aligned} \quad (7.15)$$

To get to the next line, we use the Cauchy-Schwarz and triangular inequalities:

$$\leq \frac{\nu}{t} + \frac{1}{t} [\|\nabla_x f(t; \bar{x}(t))\|_x^* + \|\nabla F(\bar{x}(t))\|_x^*] \|\bar{x}(t) - x^*(t)\|_x \quad (7.16)$$

From equations (7.14) and (7.6), and a property of self-concordant barriers, this means that

$$f_0(\bar{x}(t)) - f_0(x^*) < \frac{\nu}{t} + \frac{1}{t} (\beta + \sqrt{\nu}) \underbrace{\frac{\|\nabla_x f(t; \bar{x}(t))\|_x^*}{1 - \|\nabla_x f(t; \bar{x}(t))\|_x^*}}_{=:\omega(\|\nabla_x f(t; \bar{x}(t))\|_x^*)} \quad (7.17)$$

where $\omega(x) = \frac{x}{1-x}$ is a monotone increasing function, meaning that

$$\omega(\beta) > \omega(\|\nabla_x f(t; \bar{x}(t))\|_x^*) \quad (7.18)$$

and thus

$$f_0(\bar{x}(t)) - f_0(x^*) < \frac{1}{t} \left(\nu + (\beta + \sqrt{\nu}) \frac{\beta}{1 - \beta} \right) \quad (7.19)$$

7.3 Intermediate Newton method

Let us consider the problem (7.8), and let $\hat{f}(\cdot)$ be a self-concordant function with constant $M_{\hat{f}} = 1$. Consider $x \in \text{dom} \hat{f}$ with $\nabla^2 \hat{f}(x)$ nonsingular. Assume that $\|\nabla \hat{f}(x)\|_x^* \leq \tau$ with $\tau + \tau^2 + \tau^3 \leq 1$. The iterate of the intermediate Newton method is given by

$$x^+ = x - \frac{1}{1 + \xi} \nabla^{-2} \hat{f}(x) \nabla \hat{f}(x) \quad \xi = \frac{(\|\nabla \hat{f}(x)\|_x^*)^2}{1 + \|\nabla \hat{f}(x)\|_x^*} \quad (7.20)$$

Then, $x^+ \in \text{dom} \hat{f}$ and

$$\|\nabla \hat{f}(x^+)\|_{x^+}^* \leq \tau^2 \left(1 + \tau + \frac{\tau}{1 + \tau + \tau^2}\right) \quad (7.21)$$

Consider now the function $f(t; x) \equiv t\langle c, x \rangle + F(x)$, a self-concordant function with constant $M_f = 1$. The gradient and hessian are

$$\nabla_x f(t; x) = tc + \nabla F(x) \quad \nabla_x^2 f(t; x) = \nabla^2 F(x) \quad (7.22)$$

Let us define the iterate $t^+ = t + \frac{\gamma}{\|c\|_x^*}$ with $\gamma > 0$. The iterate of the intermediate Newton method becomes

$$x^+ = x - \frac{1}{1 - \xi} \nabla_x^{-2} f(t^+; x) \nabla_x f(t^+; x) = x - \frac{1}{1 + \xi} \nabla^{-2} F(x) (t^+ c + \nabla F(x)) \quad (7.23)$$

As previously, suppose that $\|\nabla_x f(t; x)\|_x^* \leq \beta$. Then,

$$\begin{aligned} \|\nabla_x f(t^+; x)\|_x^* &= \|t^+ c + \nabla F(x)\|_x^* = \|t^+ c - tc + tc + \nabla F(x)\|_x^* \\ &\leq (t^+ - t) \|c\|_x^* + \|\nabla_x f(t; x)\|_x^* = \gamma + \beta \end{aligned} \quad (7.24)$$

This inequality is derived using the hypothesis and the definition of t^+ . This means that, choosing $\gamma \leq \tau - \beta$ for $\tau + \tau^2 + \tau^3 \leq 1$, we get

$$\|\nabla_x f(t^+; x)\|_x^* \leq \tau \quad (7.25)$$

By equation (7.21), we have

$$\|\nabla_x f(t^+; x^+)\|_{x^+}^* \leq \tau^2 \left(1 + \tau + \frac{\tau}{1 + \tau + \tau^2}\right) = \frac{\tau^2(1 + \tau)}{1 - \tau^3} \quad (7.26)$$

And so taking $\beta = \tau^2 \left(1 + \tau + \frac{\tau}{1 + \tau + \tau^2}\right)$ seems reasonable.

→ Note: notice that $\tau > \beta$ for every $\tau \in (0, 1/2]$ and verifies $\tau + \tau^2 + \tau^3 \leq 1$.

From all those inequalities and properties, we can derive an algorithm.

7.4 Path-following Interior point Algorithm

7.4.1 Algorithm

Algorithm 4 Path-following Interior Point Algorithm

- 1: **Input:** Given $\tau \in (0, 1/2]$, define $\beta = \tau^2 \left(1 + \tau + \frac{\tau}{1+\tau+\tau^2}\right)$. Choose $0 < \gamma \leq \tau - \beta$. Find $x_0 \in \text{dom}F$ such that $\|\nabla F(x_0)\|_{x_0}^* \leq \beta$ and set $t_0 = 0$ and $k := 0$;
- 2: **Step 1:** Compute

$$\begin{aligned}
 t_{k+1} &= t_k + \frac{\gamma}{\|c\|_x^*} \\
 x_{k+1} &= x_k - \frac{1}{1 + \tilde{\xi}_k} \nabla^{-2} F(x_k) (t_{k+1} c + \nabla F(x_k)) \\
 \tilde{\xi}_k &= \frac{(\|\nabla f(t_k; x_k)\|_{x_k}^*)^2}{1 + \|\nabla f(t_k; x_k)\|_{x_k}^*}
 \end{aligned} \tag{7.27}$$

- 3: **Step 2:** $k \leftarrow k + 1$ and go back to Step 1.
-

7.4.2 Complexity bound

Notice that, by construction, $\|\nabla_x f(t_k; x_k)\|_{x_k}^* \leq \beta$, $\forall k \geq 0$, and so

$$t_k \|c\|_{x_k}^* = \|\nabla_x f(t_k; x_k) - \nabla F(x_k)\|_{x_k}^* \leq \beta + \sqrt{\nu} \tag{7.28}$$

This can be used to bound t_{k+1} :

$$t_{k+1} - t_k = \frac{\gamma}{\|c\|_{x_k}^*} \geq \frac{\gamma t_k}{\beta + \sqrt{\nu}} \iff \left(1 + \frac{\gamma}{\beta + \sqrt{\nu}}\right) t_k \quad \forall k \geq 0 \tag{7.29}$$

Thus,

$$t_k \geq \left(1 + \frac{\gamma}{\beta + \sqrt{\nu}}\right)^{k-1} t_1 = \left(1 + \frac{\gamma}{\beta + \sqrt{\nu}}\right)^{k-1} \frac{\gamma}{\|c\|_{x_0}^*} \tag{7.30}$$

Combining this to (7.19), it follows that

$$\begin{aligned}
 f_0(x_k) - f_0^* &\leq \frac{1}{t_k} \left(\nu + \frac{(\beta + \sqrt{\nu})\beta}{1 - \beta} \right) \\
 &\leq \frac{\|c\|_{x_0}^* \left(\nu + \frac{(\beta + \sqrt{\nu})\beta}{1 - \beta} \right)}{\gamma \left(1 + \frac{\gamma}{\beta + \sqrt{\nu}}\right)^{k-1}}
 \end{aligned} \tag{7.31}$$

Thus, to obtain a point x_k with $f_0(x_k) - f_0^* \leq \epsilon$, it is sufficient to have

$$\begin{aligned} \frac{\|c\|_{x_0}^* \left(\nu + \frac{(\beta + \sqrt{\nu})\beta}{1 - \beta} \right)}{\gamma \left(1 + \frac{\gamma}{\beta + \sqrt{\nu}} \right)^{k-1}} &\leq \epsilon \\ (k-1) \ln \left(1 + \frac{\gamma}{\beta + \sqrt{\nu}} \right) &\geq \ln \left(\frac{\|c\|_{x_0}^* \left(\nu + \frac{(\beta + \sqrt{\nu})\beta}{1 - \beta} \right)}{\gamma} \epsilon^{-1} \right) \\ &\Rightarrow k \geq \mathcal{O}(\epsilon^{-1}) \end{aligned} \quad (7.32)$$

Notice that $\ln(1+x) \geq cx$ for $x > 0$ and c a constant **TO BE CHECKED**. We can apply it to $x = \frac{\gamma}{\beta + \sqrt{\nu}}$ to find a bound on the number of iterations: we will have $f_0(x_k) - f_0^* \leq \epsilon$ whenever

$$(k-1)c \left(\frac{\gamma}{\beta + \sqrt{\nu}} \right) \geq \ln \left(\|c\|_{x_0}^* \left(\nu + \frac{(\beta + \sqrt{\nu})\beta}{1 - \beta} \right) \gamma^{-1} \epsilon^{-1} \right) \quad (7.33)$$

Therefore, to find a ϵ -approximate solution of problem (7.8), the algorithm 4 takes no more than $\mathcal{O}(\sqrt{\nu} \ln(\epsilon^{-1}))$ iterations.

7.4.3 Example

Consider the following problem:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} q_0(x) &\equiv c_0 + \langle b_0, x \rangle + \frac{1}{2} \langle A_0 x, x \rangle \\ \text{s.t. } q_i(x) &\equiv c_i + \langle b_i, x \rangle + \frac{1}{2} \langle A_i x, x \rangle \leq \beta_i \quad i = 1, \dots, m \end{aligned} \quad (7.34)$$

where $A_i = A_i^T \succeq 0$ for $i = 0, \dots, m$. To be able to use the algorithm derived previously, we need to change the objective function:

$$\min_{(x, \beta) \in \mathbb{R}^n \times \mathbb{R}} \beta_0 \equiv f_0(x, \beta_0) \quad \text{s.t. } q_i(x) \leq \beta_i \quad i = 0, \dots, m \quad (7.35)$$

The feasible set of this problem is the closure of the domain of the following self-concordant barrier, with constant $\nu = m + 1$:

$$F(x, \beta_0) = - \sum_{i=0}^m \ln(\beta_i - q_i(x)) \quad (7.36)$$

From the complexity of algorithm 4, it takes at most $\mathcal{O}(\sqrt{m+1} \ln(\epsilon^{-1}))$ iterations to find x_k such that

$$f_0(x_k, \beta_{0,k}) - f_0^* \leq \epsilon \quad (7.37)$$

and the operation complexity multiplies it by $\mathcal{O}(m^3)$ because it solves a linear system at each iteration.

Tips and Tricks

1. Approximation of the max:

$$\max\{z, 0\} = \frac{z + |z|}{2} = \frac{z + \sqrt{z^2}}{2} \approx \frac{z + \sqrt{z^2 + \delta}}{2} \quad (8.1)$$

2.

$$ab \leq \frac{a^2 + b^2}{2} \quad (8.2)$$

3.

$$(a + b)^2 \leq 2a^2 + 2b^2 \quad (8.3)$$

4. V-trick:

$$\langle xv, v \rangle \leq \|x\| \|v\|^2 \quad (8.4)$$

5. Triangular inequality by the minimizer:

$$\|x_{k+1} - x_k\| \leq \|x_{k+1} - x^*\| + \|x_k - x^*\| \quad (8.5)$$

6. Mean Value Theorem $\forall x, y \in \Omega, \exists z \in \Omega$ s.t.:

$$f(y) - f(x) = \langle \nabla f(z), y - x \rangle \quad z \in [x, y] \quad (8.6)$$

7. By definition if a function is C_M^p , then

$$|f(y) - T_p(y; x)| \leq \frac{M}{(p+1)!} \|y - x\|^{p+1} \quad (8.7)$$