

# Option Calcul Scientifique: Optimisation Numérique

Préparation Agrégation de Mathématiques  
Université de Rennes 1  
Isabelle Gruais

7 décembre 2023

## 1 Méthodes de gradient

Soit  $K \subset \mathbb{R}^N$  et soit  $J : K \rightarrow \mathbb{R}$ ,  $\alpha$ -convexe. Soit à résoudre : trouver  $x^* \in K$  solution de :

$$x^* \in K \quad \text{et} \quad J(x^*) = \min_{x \in K} J(x)$$

**Hypothèse 1.0.1.** *On suppose que  $J$  est  $\alpha$ -convexe, différentiable et que  $\nabla J$  est localement lipschitzienne.*

### 1.1 Gradient avec pas optimal

**Définition 1.1.1** (Algorithme du gradient avec pas optimal). On appelle suite engendrée par l'Algorithme du gradient avec pas optimal la suite  $(u_n)_{n \geq 0} \in (\mathbb{R}^N)^{\mathbb{N}}$  définie par la relation de récurrence :

$$u_0 \in \mathbb{R}^N \quad \text{arbitraire,}$$

$$u_{n+1} = u_n - \mu_n \nabla J(u_n)$$

avec

$$J(u_n - \mu_n \nabla J(u_n)) = \min_{\mu \in \mathbb{R}} J(u_n - \mu \nabla J(u_n)).$$

**Définition 1.1.2.** On appelle direction de descente à l'étape  $n$  le vecteur  $r_n := -\nabla J(u_n)$ .

**Proposition 1.1.1.** Dans l'algorithme 1.2.1, deux directions de descente consécutives sont orthogonales.

*Démonstration.* Soit  $n \geq 1$ . On pose :

$$f_n(\mu) = J(u_n + \mu r_n) \quad \text{où} \quad r_n := -\nabla J(u_n). \quad (1)$$

Par définition :

$$f'_n(\mu_n) = 0 = \underbrace{\langle \nabla J(u_n + \mu_n r_n), r_n \rangle}_{=-r_{n+1}} = -\langle r_{n+1}, r_n \rangle.$$

□

**Théorème 1.1.2.** Sous l'Hypothèse 1.0.1, la suite  $(u_n)_{n \geq 0}$  générée par l'algorithme du gradient avec pas optimal converge vers la solution  $u \in \mathbb{R}^N$  du problème de minimisation :

$$J(u) = \min_{v \in \mathbb{R}^N} J(v)$$

et on a l'estimation :

$$\|u_n - u\| \leq \frac{1}{\alpha} \|\nabla J(u_n)\|.$$

*Démonstration.* Par construction, la suite  $(J(u_n))_{n \geq 0}$  est décroissante, minorée par coercivité :

$$J(u_n) \geq \beta \|u_n\|^2 + b \geq b > -\infty$$

donc convergente. On en déduit qu'il existe  $C > 0$  t.q. :  $\forall n \geq 0$ ,

$$C \geq J(u_n) \geq \beta \|u_n\|^2 + b \Rightarrow \|u_n\|^2 \leq \frac{C}{\beta} =: M.$$

i.e. la suite  $(u_n)_{n \geq 0}$  est majorée.

Comme  $J$  est  $\alpha$ -convexe, on aussi :

$$\begin{aligned} J(u_n) &\geq J(u_{n+1}) + \langle \nabla J(u_{n+1}), u_n - u_{n+1} \rangle + \frac{\alpha}{2} \|u_n - u_{n+1}\|^2 = \\ &= J(u_{n+1}) + \underbrace{\mu_n \langle r_n, r_{n+1} \rangle}_{=0} + \frac{\alpha}{2} \|u_n - u_{n+1}\|^2 = J(u_{n+1}) + \frac{\alpha}{2} \|u_n - u_{n+1}\|^2 \end{aligned}$$

$$\Rightarrow \|u_n - u_{n+1}\|^2 \leq \frac{2}{\alpha}(J(u_n) - J(u_{n+1})) \xrightarrow{n \rightarrow +\infty} 0$$

Il en résulte :

$$\begin{aligned} \|\nabla J(u_n)\|^2 &= \langle \nabla J(u_n), \nabla J(u_n) \rangle = \langle \nabla J(u_n), \nabla J(u_n) - \nabla J(u_{n+1}) \rangle \\ &\leq C_M \|\nabla J(u_n)\| \|u_n - u_{n+1}\| \end{aligned}$$

d'où :

$$\|\nabla J(u_n)\| \leq C_M \|u_n - u_{n+1}\| \xrightarrow{n \rightarrow +\infty} 0.$$

Par  $\alpha$ -convexité de  $J$  :

$$\begin{aligned} \alpha \|u_n - u\|^2 &\leq \langle \nabla J(u_n) - \nabla J(u), u_n - u \rangle = \langle \nabla J(u_n), u_n - u \rangle \leq \|\nabla J(u_n)\| \|u_n - u\| \\ \Rightarrow \|u_n - u\| &\leq \frac{1}{\alpha} \|\nabla J(u_n)\| \xrightarrow{n \rightarrow +\infty} 0. \end{aligned}$$

□

**Proposition 1.1.3** (Cas particulier d'une fonctionnelle quadratique). *On suppose que  $J$  est définie par :*

$$J(x) = \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle, \quad \forall x \in \mathbb{R}^N$$

où  $A \in \mathbb{R}^{N \times N}$  est symétrique, définie positive, et qu'il existe  $\alpha > 0$  t.q.

$$\langle Ax, x \rangle \geq \alpha \|x\|^2, \quad \forall x \in \mathbb{R}^N.$$

Alors :

$$\mu_n = \frac{\|r_n\|^2}{\langle Ar_n, r_n \rangle}, \quad \forall n \geq 0. \quad (2)$$

*Démonstration.* Avec la définition (1) :

$$\begin{aligned} f'_n(\mu_n) = 0 &= \langle A(u_n + \mu_n r_n) - b, r_n \rangle = \underbrace{\langle Au_n - b, r_n \rangle}_{=-r_n} + \mu_n \langle Ar_n, r_n \rangle = \\ &= -\|r_n\|^2 + \mu_n \langle Ar_n, r_n \rangle \end{aligned}$$

□

**Théorème 1.1.4.** *Sous les hypothèses de la Proposition 1.1.3, on note  $0 < \lambda_1 \leq \dots \leq \lambda_N$  la suite des valeurs propres de  $A$  comptées avec leur ordre de multiplicité. Alors, l'erreur à l'étape  $n$  définie par*

$$e_n := u_n - u$$

vérifie :

$$\frac{\langle Ae_n, e_n \rangle}{\langle Ae_0, e_0 \rangle} \leq \left( \frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1} \right)^{2n}, \quad \forall n \geq 0$$

*Démonstration.* Le démonstration utilise le résultat suivant qui sera admis :

**Lemme 1.1.5** (Estimation de Kantorovitch).

$$\|x\|^4 \leq \langle Ax, x \rangle \langle A^{-1}x, x \rangle \leq \frac{1}{4} \left( \sqrt{c} + \frac{1}{\sqrt{c}} \right)^2 \|x\|^4 \quad (3)$$

où

$$c := \frac{\lambda_N}{\lambda_1}$$

Soit  $n \geq 0$ . On a :

$$e_{n+1} = u_{n+1} - u = \mu_n r_n + e_n$$

donc

$$\begin{aligned} \langle Ae_{n+1}, e_{n+1} \rangle &= \mu_n \underbrace{\langle Ae_{n+1}, r_n \rangle}_{=-\langle r_{n+1}, r_n \rangle} + \langle Ae_{n+1}, e_n \rangle = \langle Ae_{n+1}, e_n \rangle \\ &= \langle Ae_n, e_n \rangle + \mu_n \langle Ar_n, e_n \rangle = \langle Ae_n, e_n \rangle - \mu_n \|r_n\|^2 \\ &\stackrel{(2)}{=} \langle Ae_n, e_n \rangle - \frac{\|r_n\|^4}{\langle Ar_n, r_n \rangle} \end{aligned}$$

avec :

$$\langle Ae_n, e_n \rangle = \langle r_n, A^{-1}r_n \rangle.$$

On en déduit :

$$\begin{aligned} \frac{\langle Ae_{n+1}, e_{n+1} \rangle}{\langle Ae_n, e_n \rangle} &= 1 - \frac{\|r_n\|^4}{\langle r_n, A^{-1}r_n \rangle \langle A \langle Ar_n, r_n \rangle} \\ &\stackrel{(3)}{\leq} 1 - 4 \left( \sqrt{c} + \frac{1}{\sqrt{c}} \right)^{-2} = \left( \frac{c-1}{c+1} \right)^2 = \left( \frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1} \right)^2. \end{aligned}$$

Il en résulte :

$$\frac{\langle Ae_{n+1}, e_{n+1} \rangle}{\langle Ae_0, e_0 \rangle} = \prod_{k=0}^n \frac{\langle Ae_{k+1}, e_{k+1} \rangle}{\langle Ae_k, e_k \rangle} \leq \left( \frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1} \right)^{2(n+1)}.$$

□

## 1.2 Gradient avec pas fixe

**Définition 1.2.1** (Algorithme du gradient avec pas fixe). On appelle suite engendrée par l'Algorithme du gradient avec pas fixe  $\mu \in \mathbb{R}$  la suite  $(u_n)_{n \geq 0} \in (\mathbb{R}^N)^{\mathbb{N}}$  définie par la relation de récurrence :

$$u_0 \in \mathbb{R}^N \quad \text{arbitraire,}$$

$$u_{n+1} = u_n - \mu \nabla J(u_n)$$

**Théorème 1.2.1.** *Sous l'hypothèse 1.0.1, la suite  $(u_n)_{n \geq 0}$  définie par l'Algorithme 1.2.1 converge vers la solution du problème de minimisation :*

$$J(u) = \min_{v \in \mathbb{R}^N} J(v)$$

pour tout  $\mu > 0$  suffisamment petit. En particulier, elle converge pour tout  $\mu \in ]0, \frac{2\alpha}{C^2}[$  où  $C > 0$  est la constante de Lipschitz de  $\nabla J$  sur la boule fermée  $B := B(u, \|u - u_0\|)$ .

*Démonstration.* Soit  $\mathcal{P}(n)$  la propriété :  $u_n \in B$ . Alors  $\mathcal{P}(0)$  est vraie par définition de  $B$ . On suppose  $\mathcal{P}(n)$  vraie.

De la relation :

$$e_{n+1} = u_{n+1} - u_n + e_n = -\mu \nabla J(u_n) + e_n = -\mu(\nabla J(u_n) - \nabla J(u)) + e_n$$

on déduit que :

$$\begin{aligned} \|e_{n+1}\|^2 &= \|e_n\|^2 - 2\mu \langle \nabla J(u_n) - \nabla J(u), e_n \rangle + \mu^2 \|\nabla J(u_n) - \nabla J(u)\|^2 \\ &\leq \|e_n\|^2 \underbrace{(1 - 2\mu\alpha + C^2\mu^2)}_{=: \theta(\mu)} \end{aligned}$$

avec :  $\forall u, v \in \mathbb{R}^N$ ,

$$\alpha \|u - v\|^2 \leq \langle \nabla J(u) - \nabla J(v), u - v \rangle \leq C\alpha \|u - v\|^2$$

donc  $0 < \alpha < C$ . L'étude des variations de  $\theta : \mu \mapsto 1 - 2\mu\alpha + C^2\mu^2$  montre que

$$0 < \theta(\mu) < 1 \iff \mu \in ]0, \frac{2\alpha}{C^2}[.$$

On en déduit que si  $\mu \in ]0, \frac{2\alpha}{C^2}[$  alors

$$\|e_{n+1}\| < \|e_n\| \underset{\mathcal{P}(n)}{<} \|u - u_0\|$$

i.e.  $u_{n+1} \in B$ . Par récurrence sur  $n \geq 0$  :

$$\|e_n\| \leq \theta(\mu)^n \|e_0\| \xrightarrow{n \rightarrow +\infty} 0.$$

□

## Cas des fonctionnelles quadratiques

Soit  $J(x) = \frac{1}{2}\langle Ax, x \rangle$ ,  $\forall x \in \mathbb{R}^N$ . avec  $A$  symétrique définie positive de valeurs propres :

$$0 < \lambda_1 \leq \dots \leq \lambda_N.$$

Après diagonalisation dans une bon de vecteurs propres,  $J$  se réécrit sous la forme :

$$J(x) = \frac{1}{2} \sum_{i=1}^N \lambda_i x_i^2.$$

Soit  $\mu > 0$ . L'algorithme de gradient avec pas fixe  $\mu$  se réécrit :

$$u^0 \in \mathbb{R}^N, \quad u_i^{n+1} = (1 - \mu\lambda_i)u_i^n, \quad i = 1, \dots, N.$$

La suite  $(u^n)_{n \geq 0}$  converge ssi :  $\forall i \in [[1, N]]$ ,

$$|1 - \mu\lambda_i| < 1 \iff 0 < \mu < \frac{2}{\lambda_i}$$

i.e. ssi  $0 < \mu < \frac{2}{\lambda_N}$ . Le taux de convergence est optimal si  $\mu = \mu_{\text{opt}}$  est solution de

$$\max_{1 \leq i \leq N} |1 - \mu\lambda_i|$$

ce qui est réalisé pour

$$1 - \mu\lambda_1 = -1 + \mu\lambda_N \iff \mu = \frac{2}{\lambda_1 + \lambda_N} =: \mu_{\text{opt}}$$

## 1.3 Gradient conjugué pour une fonctionnelle quadratique

Soit  $J(x) = \frac{1}{2}\langle Ax, x \rangle - \langle b, x \rangle$  où  $A$  est symétrique réelle, définie positive. On construit une suite  $(u_n)_{n \geq 0}$  t.q. :

$$u_{n+1} = u_n + \mu_n d_n, \quad n \geq 0$$

avec  $\mu_n \in \mathbb{R}$  solution du problème

$$J(u_{n+1}) = \min_{\mu \in \mathbb{R}} J(u_n + \mu d_n) \Rightarrow \langle \nabla J(u_n), d_n \rangle = 0$$

et où la suite de directions de descente  $(d_n)_{n \geq 0}$  est choisie t.q. :  $\langle Ad_n, d_{n-1} \rangle = 0$ ,  $n \geq 0$ .

**Définition 1.3.1** (Algorithme du gradient conjugué). On appelle suite engendrée par l'Algorithme du gradient avec pas fixe  $\mu \in \mathbb{R}$  la suite  $(u_n)_{n \geq 0} \in (\mathbb{R}^N)^{\mathbb{N}}$  définie par la relation de récurrence :

$$\begin{aligned} u_0 &\in \mathbb{R}^N \quad \text{arbitraire,} \\ d_0 &= -\nabla J(u_0) = b - Au_0 =: r_0 \\ r_n &= -\nabla J(u_n) = b - Au_n \\ d_n &= r_n + \beta_n d_{n-1}, \quad \text{où } \beta_n = -\frac{\langle r_n, Ad_{n-1} \rangle}{\langle d_{n-1}, Ad_{n-1} \rangle}, \\ u_{n+1} &= u_n + \alpha_n d_n, \quad \alpha_n = \frac{\langle r_n, d_n \rangle}{\langle d_n, Ad_n \rangle}, \end{aligned}$$

**Proposition 1.3.1.** *L'algorithme 1.3.1 vérifie :*

$$\begin{aligned} \langle r_{n+1}, d_n \rangle &= 0, \quad n \geq 0 \\ \langle Ad_n, d_{n-1} \rangle &= 0, \quad n \geq 1. \end{aligned}$$

*Démonstration.* Cela découle de la définition des coefficients  $\alpha_n, \beta_n, n \geq 0$ .  $\square$

**Proposition 1.3.2.** *L'algorithme 1.3.1 vérifie :*

$$\langle r_{n+1}, r_n \rangle = 0, \quad n \geq 0$$

*De plus, on a les expressions :*

$$\alpha_n = \frac{\|r_n\|^2}{\langle Ad_n, d_n \rangle}, \quad \beta_n = -\frac{\|r_n\|^2}{\|r_{n-1}\|^2}$$

*Démonstration.* Par construction :

$$\langle r_n, d_n \rangle = \|r_n\|^2 + \beta_n \underbrace{\langle r_n, d_{n-1} \rangle}_{=0} = \|r_n\|^2$$

et donc :

$$\alpha_n = \frac{\|r_n\|^2}{\langle Ad_n, d_n \rangle}.$$

Alors :

$$\begin{aligned} \langle r_{n+1}, r_n \rangle &= \|r_n\|^2 - \alpha_n \langle Ad_n, r_n \rangle = \\ &= \underbrace{\|r_n\|^2 - \alpha_n \langle Ad_n, d_n \rangle}_{=0} + \alpha_n \beta_n \underbrace{\langle Ad_n, d_{n-1} \rangle}_{=0} = 0 \end{aligned}$$

On a aussi :

$$\begin{aligned}
Ad_{n-1} &= \frac{1}{\alpha_{n-1}} A(u_n - u_{n-1}) = \frac{1}{\alpha_{n-1}} (r_{n-1} - r_n) \\
\Rightarrow \langle r_n, Ad_{n-1} \rangle &= \frac{1}{\alpha_{n-1}} \langle r_n, r_{n-1} - r_n \rangle = -\frac{\|r_n\|^2}{\alpha_{n-1}} \\
\langle d_{n-1}, Ad_{n-1} \rangle &= \frac{1}{\alpha_{n-1}} \langle d_{n-1}, r_{n-1} - r_n \rangle = \frac{1}{\alpha_{n-1}} \langle d_{n-1}, r_{n-1} \rangle \\
&= \frac{\|r_{n-1}\|^2}{\alpha_{n-1}} + \frac{\beta_{n-1}}{\alpha_{n-1}} \underbrace{\langle d_{n-2}, r_{n-1} \rangle}_{=0} = \frac{\|r_{n-1}\|^2}{\alpha_{n-1}}
\end{aligned}$$

d'où on déduit que

$$\beta_n = \frac{\|r_n\|^2}{\|r_{n-1}\|^2}.$$

□

**Théorème 1.3.3.** *On pose :  $e_n = u_n - u$ ,  $\forall n \geq 0$ . Alors :*

$$\frac{\langle Ae_n, e_n \rangle}{\langle Ae_0, e_0 \rangle} \leq \left( \frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1} \right)^{2n}, \quad \forall n \geq 0$$

*Démonstration.* Soit  $n \geq 0$ . De la relation  $e_{n+1} = \alpha_n d_n + e_n$ , on déduit :

$$\begin{aligned}
\langle Ae_{n+1}, e_{n+1} \rangle &= \langle Ae_n, e_n \rangle + 2\alpha_n \langle Ae_n, d_n \rangle + \alpha_n^2 \langle Ad_n, d_n \rangle = \\
&= \langle Ae_n, e_n \rangle - 2\alpha_n \langle r_n, d_n \rangle + \alpha_n^2 \langle Ad_n, d_n \rangle
\end{aligned}$$

avec

$$\alpha_n = \frac{\langle r_n, d_n \rangle}{\langle Ad_n, d_n \rangle}.$$

On en déduit :

$$\langle Ae_{n+1}, e_{n+1} \rangle = \langle Ae_n, e_n \rangle - \frac{\langle r_n, d_n \rangle^2}{\langle Ad_n, d_n \rangle} = \langle Ae_n, e_n \rangle \left( 1 - \frac{\langle r_n, d_n \rangle^2}{\langle Ad_n, d_n \rangle \langle Ae_n, e_n \rangle} \right)$$

avec  $\langle Ae_n, e_n \rangle = \langle r_n, A^{-1}r_n \rangle$ .

De plus, compte tenu de l'expression de  $\beta_n$  :

$$\begin{aligned}
\langle Ad_n, d_n \rangle &= \langle Ar_n, r_n \rangle + 2\beta_n \langle r_n, Ad_{n-1} \rangle + \beta_n^2 \langle d_{n-1}, Ad_{n-1} \rangle = \\
&= \langle Ar_n, r_n \rangle - \frac{\langle r_n, Ad_{n-1} \rangle^2}{\langle d_{n-1}, Ad_{n-1} \rangle} \in ]0, \langle Ar_n, r_n \rangle[.
\end{aligned}$$



Il en résulte, par croissance de  $x \mapsto -\frac{1}{x}$  sur  $]0, +\infty[$  :

$$\langle Ae_{n+1}, e_{n+1} \rangle \leq \langle Ae_n, e_n \rangle \left( 1 - \frac{\langle r_n, d_n \rangle^2}{\langle Ar_n, r_n \rangle \langle r_n, A^{-1}r_n \rangle} \right)$$

et on conclut comme pour le Théorème 1.1.4.  $\square$

**Proposition 1.3.4.** *L'algorithme 1.3.1 vérifie :  $\forall n \geq 1$ ,*

$$\begin{aligned} \langle r_n, d_j \rangle &= 0, & \forall j < n \\ \langle r_n, r_j \rangle &= 0, & \forall j < n \\ \langle Ad_n, d_j \rangle &= 0, & \forall j < n \\ \langle r_{n+1}, Ad_j \rangle &= 0, & \forall j < n \end{aligned} \tag{4}$$

*Démonstration.* Soit  $\mathcal{P}(n)$  la Propriété :

$$\begin{aligned} \langle r_{n+1}, d_j \rangle &= 0, & \forall j \leq n \\ \langle r_{n+1}, r_j \rangle &= 0, & \forall j \leq n \end{aligned} \tag{5}$$

$$\langle Ad_n, d_j \rangle = 0, \quad \forall j < n \tag{6}$$

$$\langle r_{n+1}, Ad_j \rangle = 0, \quad \forall j < n$$

On vérifie directement que  $\mathcal{P}(0)$  est vraie et on suppose  $\mathcal{P}(n-1)$  vraie. Par construction

$$\langle Ad_n, d_{n-1} \rangle = 0.$$

Soit  $j < n-1$ . On a

$$\begin{aligned} \langle Ad_n, d_j \rangle &= \langle d_n, Ad_j \rangle = \langle r_n, Ad_j \rangle + \beta_n \langle d_{n-1}, Ad_j \rangle \\ &\stackrel{\mathcal{P}(n-1)}{=} \beta_n \langle Ad_{n-1}, d_j \rangle \stackrel{\mathcal{P}(n-1)}{=} 0. \end{aligned}$$

i.e. (6) est vraie. Par construction :

$$\langle r_{n+1}, d_n \rangle = 0.$$

Soit  $j < n$ . On a :

$$\langle r_{n+1}, d_j \rangle = \langle r_n, d_j \rangle - \alpha_n \langle Ad_n, d_j \rangle \stackrel{\mathcal{P}(n-1), (6)}{=} 0$$

Par construction :

$$\langle r_{n+1}, r_n \rangle = 0.$$

Soit  $j < n$ . On a :

$$\begin{aligned} \langle r_{n+1}, r_j \rangle &= \langle r_n, r_j \rangle - \alpha_n \langle Ad_n, r_j \rangle \\ &\stackrel{\mathcal{P}(n-1)}{=} -\alpha_n \langle Ad_n, d_j \rangle + \alpha_n \beta_j \langle Ad_n, d_{j-1} \rangle \stackrel{(6)}{=} 0 \end{aligned}$$

i.e. (5). Soit  $j < n$ . On a :

$$\langle r_{n+1}, Ad_j \rangle = \frac{1}{\alpha_j} \langle r_{n+1}, r_j - r_{j+1} \rangle \stackrel{(5)}{=} 0$$

□

**Corollaire 1.3.5.** *L'algorithme 1.3.1 vérifie :*

$$\begin{aligned} d_n &= r_n - \sum_{j < n} \frac{\langle r_n, Ad_j \rangle}{\langle d_j, Ad_j \rangle} d_j \\ r_{n+1} &= r_n - \sum_{j \leq n} \frac{\langle r_n, d_j \rangle}{\langle d_j, Ad_j \rangle} Ad_j \end{aligned}$$

**Corollaire 1.3.6.** *L'algorithme 1.3.1 converge en au plus  $N$  itérations.*

*Démonstration.* On suppose que  $r_{N-1} \neq 0$ . Soit alors  $\mu_0, \dots, \mu_{N-1} \in \mathbb{R}$  t.q.

$$\sum_{k=0}^{N-1} \lambda_k r_k = 0.$$

Alors :

$$0 = \left\langle \sum_{k=0}^{N-1} \lambda_k r_k, r_{N-1} \right\rangle \stackrel{(4)}{=} \lambda_{N-1} \|r_{N-1}\|^2 \Rightarrow \lambda_{N-1} = 0.$$

On suppose que  $\lambda_{N-1} \cdots = \lambda_{N-k+1} = 0$  avec  $1 < k < N$ . Alors :

$$0 = \left\langle \sum_{k=0}^{N-k} \lambda_k r_k, r_{N-k} \right\rangle \stackrel{(4)}{=} \lambda_{N-k} \|r_{N-k}\|^2 \Rightarrow \lambda_{N-k} = 0.$$

Par récurrence sur  $k \in [[1, N-1]]$ , on en déduit que  $\lambda_0 = \dots = \lambda_{N-1} = 0$ , i.e. que  $(r_0, \dots, r_{N-1})$  est une base de  $\mathbb{R}^N$ . En particulier, il résulte de (4) que  $r_N \in \text{Vect}(r_0, \dots, r_{N-1})^\perp = \{0\}$ , i.e.  $r_N = 0$ . □

### Nombre d'opérations :

- (i) : Calcul de  $r_k = b - Au_k$ ,  $k = 0, \dots, N - 1$  : le calcul de  $(Ax)_i = \sum_{j=1}^N a_{ij}x_j$ ,  $i = 1, \dots, N$  comprend  $N - 1$  additions et  $N$  multiplications, soit  $N^2(N - 1)$  additions et  $N^3$  multiplications pour la suite des  $r_k$ ,  $k = 0, \dots, N - 1$ .
- (ii) : Calcul de  $\beta_k = \frac{\|r_k\|^2}{\|r_{k-1}\|^2}$ ,  $k = 1, \dots, N - 1$  : le calcul de  $\|x\|^2 = \sum_{i=1}^N x_i^2$  se décompose en  $N$  multiplications et  $N - 1$  additions, soit  $N(N - 1)$  multiplications et  $(N - 1)^2$  additions pour le calcul des  $\|r_k\|^2$ ,  $k = 1, \dots, N - 1$ . Il reste  $N - 1$  divisions.
- (iii) Calcul de  $d_k = r_k + \beta_k d_{k-1}$ ,  $k = 1, \dots, N - 1$  : se décompose en  $N - 1$  multiplications et  $N - 1$  additions.
- (iv) Calcul de  $\alpha_k = \frac{\|r_k\|^2}{\langle Ad_k, d_k \rangle}$ ,  $k = 0, \dots, N - 1$ . Le calcul de  $\langle Ax, x \rangle = \sum_{i=1}^N (\sum_{j=1}^N a_{ij}x_j)x_i$  se décompose en  $2N$  multiplications et  $2(N - 1)$  additions, soit  $2N^2$  multiplications et  $2N(N - 1)$  additions pour le calcul des  $\langle Ad_k, d_k \rangle$ ,  $k = 0, \dots, N - 1$ . Il reste  $N$  divisions.
- (v) Calcul de  $u_{k+1} = u_k + \alpha_k d_k$ ,  $k = 0, \dots, N - 1$  : se décompose en  $N$  multiplications et  $N$  additions.

A total :

- (i) Additions :

$$\underbrace{N^2(N - 1)}_{r_k} + \underbrace{(N - 1)^2}_{\beta_k} + \underbrace{N - 1}_{d_k} + \underbrace{2N(N - 1)}_{\alpha_k} + \underbrace{N}_{u_{k+1}} \sim N^3 + 2N^2 + 2N \sim N^3$$

- (ii) Multiplications

$$\underbrace{N^3}_{r_k} + \underbrace{(N - 1)N}_{\beta_k} + \underbrace{N - 1}_{d_k} + \underbrace{2N^2}_{\alpha_k} + \underbrace{N}_{u_{k+1}} \sim N^3 + 3N^2 + 2N \sim N^3$$

- (iii) Divisions :

$$\underbrace{0}_{r_k} + \underbrace{(N - 1)}_{\beta_k} + \underbrace{0}_{d_k} + \underbrace{N}_{\alpha_k} + \underbrace{0}_{u_{k+1}} \sim 2N$$

Il y a donc  $2N^3 + O(N^2)$  opérations, à comparer avec  $\frac{2}{3}N^3$  opérations pour la méthode de Gauss avec pivot.

## 2 Optimisation avec contraintes

### 2.1 Gradient à pas fixe avec projection

**Définition 2.1.1.** Soit  $K \subset \mathbb{R}^N$  un convexe fermé et soit  $\mu > 0$ . On appelle suite générée par l'algorithme du gradient à pas fixe  $\mu > 0$  avec projection

sur  $K$  toute suite  $(u_k)_{k \geq 0}$  définie par :

$$u_0 \in \mathbb{R}^N$$

$$u_{k+1} = P_K(u_k - \mu \nabla J(u_k)) \quad (7)$$

où  $P_K$  désigne la projection sur le convexe  $K$ .

**Théorème 2.1.1.** *On suppose que  $J$  est  $\alpha$ -convexe, différentiable sur  $K$ , de gradient  $\nabla J$  lipschitzien sur  $K$ , de constante de Lipschitz  $C > 0$ . Alors, pour tout  $\mu > 0$  vérifiant :*

$$0 < \mu < \frac{2\alpha}{C^2} \quad (8)$$

la suite  $(u_k)_{k \geq 0}$  associée à la Définition 2.1.1 converge vers l'unique solution  $u \in \mathbb{R}^N$  du problème :

$$u \in K \quad \text{et} \quad J(u) = \min_{v \in K} J(v) \quad (9)$$

*Démonstration.* Soit  $\mu \in ]0, \frac{2\alpha}{C^2}[$ . On commence par remarquer que  $v \mapsto v - \mu \nabla J(v)$  est strictement contractante. En effet : soit  $v, w \in \mathbb{R}^N$ . On a :

$$\begin{aligned} \|v - w - \mu(\nabla J(v) - \nabla J(w))\|^2 &= \|v - w\|^2 + \mu^2 \|\nabla J(v) - \nabla J(w)\|^2 + \\ &\quad - 2\mu \langle v - w, \nabla J(v) - \nabla J(w) \rangle \\ &\leq \|v - w\|^2 \underbrace{(1 - 2\mu\alpha + \mu^2 C^2)}_{=: \theta(\mu) \underset{(8)}{< 1}} \end{aligned}$$

Comme  $P_K$  est contractante, il en résulte que  $v \mapsto P_K(-\mu \nabla J(v))$  est strictement contractante. D'après le théorème du point fixe appliqué au convexe fermé  $K$  de l'espace complet  $\mathbb{R}^N$ , on déduit que la suite  $(u_k)_{k \geq 0}$  vérifiant (7) converge vers l'unique solution de (9).  $\square$

## Calcul de $x' = P_K(x)$

Si  $K$  est un pavé de  $\mathbb{R}^N$

Soit  $K = \prod_{i=1}^N [a_i, b_i]$  et soit  $i \in [[1, N]]$ . On a :

$$x'_i = \max(a_i, x_i) = \min(x_i, b_i)$$

Si  $\min(x_i, b_i) = x_i$ , alors  $x_i \leq b_i \Rightarrow x'_i = \max(x_i, a_i)$ . Si  $\min(x_i, b_i) = b_i$ , alors  $x_i \geq b_i \Rightarrow x'_i = b_i$ . Dans tous les cas  $x'_i = \max(a_i, \min(b_i, x_i))$ .

Un raisonnement analogue montre que  $x'_i = \min(b_i, \max(a_i, x_i))$ . Finalement :

$$x'_i = \min(b_i, \max(a_i, x_i)) = \max(a_i, \min(b_i, x_i)).$$

Si  $K$  est une boule de  $\mathbb{R}^N$

Soit  $K = B(a, r) := \{x \in \mathbb{R}^N, \|x - a\| \leq r\}$ . On a immédiatement  $x = x' = a$  si  $x = a$ . Sinon, soit  $x \neq a$ . Alors

$$x' = a + r \frac{x - a}{\|x - a\|}.$$

S

## 2.2 Méthode de relaxation

### Préliminaire : Relaxation sans contrainte

Soit  $J : \mathbb{R}^N \rightarrow \mathbb{R}$  vérifiant l'Hypothèse 1.0.1.

**Définition 2.2.1.** On appelle suite générée par la méthode de relaxation (sans contrainte) toute suite  $(u_k)_{k \geq 0}$  définie par :

$$\begin{aligned} u_0 &\in \mathbb{R}^N \\ J(u_1^{k+1}, u_2^k, \dots, u_N^k) &= \inf_{\mu \in \mathbb{R}} J(\mu, u_2^k, \dots, u_N^k), \\ J(\underbrace{u_1^{k+1}, \dots, u_i^{k+1}}_{=: u_{k,i}}, u_{i+1}^k, \dots, u_N^k) &= \inf_{\mu \in \mathbb{R}} J(u_1^{k+1}, \dots, u_{i-1}^{k+1}, \mu, u_{i+1}^k, \dots, u_N^k), \\ & \qquad \qquad \qquad i = 2, \dots, N-1 \\ J(u_1^{k+1}, \dots, u_N^{k+1}) &= \inf_{\mu \in \mathbb{R}} J(u_1^{k+1}, \dots, u_{N-1}^{k+1}, \mu), \end{aligned}$$

**Théorème 2.2.1.** *Sous les hypothèses de la Définition 2.2.1, la suite  $(u_k)_{k \geq 0}$  converge vers l'unique solution de*

$$J(u) = \inf_{v \in \mathbb{R}^N} J(v).$$

*Démonstration.* Soit  $k \geq 1$  et soit  $i \in [[1, N]]$ . On a :

$$J(u_{k,i}) \leq J(u_{k,i-1}) \leq J(u_{k,0}) = J(u^k) = J(u_{k-1}, N) \leq J(u_{k-1,i}),$$

i.e. la suite  $(J(u_{k,i}))_{k \geq 0}$  est décroissante, minorée par coercivité de  $J$ , donc convergente. La coercivité de  $J$  entraîne qu'il existe  $M > 0$  t.q.

$$\sup_{\substack{k \geq 0 \\ 1 \leq i \leq N}} \|u_{k,i}\| \leq M \Rightarrow \sup_{k \geq 0} \|u^k\| \leq M.$$

On remarque que :

$$J(u_{k,i}) = \inf_{\mu \in \mathbb{R}} J(u_{k,i-1} + \mu e_i) \Rightarrow \frac{\partial J(u_{k,i})}{\partial x_i} = \langle \nabla J(u_{k,i}), e_i \rangle = 0$$

De plus :

$$J(u^{k+1}) = J(u_{k,N}) \leq J(u_{k,0}) = J(u^k)$$

i.e., la suite  $(J(u^k))_{k \geq 0}$  est décroissante, minorée par coercivité de  $J$ , donc convergente. Il en résulte :

$$\begin{aligned} J(u^k) - J(u^{k+1}) &= J(u_{k,0}) - J(u_{k,N}) = \sum_{i=0}^{N-1} (J(u_{k,i}) - J(u_{k,i+1})) \geq \\ &\geq \sum_{i=0}^{N-1} \langle \nabla J(u_{k,i+1}), u_{k,i} - u_{k,i+1} \rangle + \frac{\alpha}{2} \sum_{i=0}^{N-1} \|u_{k,i} - u_{k,i+1}\|^2 \end{aligned}$$

avec :

$$\sum_{i=0}^{N-1} \langle \nabla J(u_{k,i+1}), u_{k,i} - u_{k,i+1} \rangle = \sum_{i=0}^{N-1} \frac{\partial J(u_{k,i+1})}{\partial x_{i+1}} (u_{i+1}^k - u_{i+1}^{k+1}) = 0,$$

$$\sum_{i=0}^{N-1} \|u_{k,i} - u_{k,i+1}\|^2 = \sum_{i=0}^{N-1} \|u_{i+1}^k - u_{i+1}^{k+1}\|^2 = \|u^k - u^{k+1}\|^2.$$

Il en résulte :

$$J(u^k) - J(u^{k+1}) \geq \frac{\alpha}{2} \|u^k - u^{k+1}\|^2,$$

avec  $\lim_{k \rightarrow +\infty} J(u^k) - J(u^{k+1}) = 0$ . Du Théorème d'encadrement des germes, il résulte que  $\lim_{k \rightarrow +\infty} \|u^k - u^{k+1}\| = 0$ . En particulier :

$$\|u^{k+1} - u_{k,i}\|^2 = \sum_{j=i+1}^N |u_j^{k+1} - u_j^k|^2 \leq \|u^{k+1} - u^k\|^2 \xrightarrow{k \rightarrow +\infty} 0, \quad \forall i \in [[1, N]].$$

On a :

$$\begin{aligned} \alpha \|u^{k+1} - u\|^2 &\leq \langle \nabla J(u^{k+1}) - \nabla J(u), u^{k+1} - u \rangle = \langle \nabla J(u^{k+1}), u^{k+1} - u \rangle = \\ &= \sum_{i=1}^N \frac{\partial J(u^{k+1})}{\partial x_i} (u_i^{k+1} - u_i^k) \Rightarrow \alpha \|u^{k+1} - u\| \leq \sum_{i=1}^N \left| \frac{\partial J(u^{k+1})}{\partial x_i} \right| = \sum_{i=1}^N \left| \frac{\partial J(u^{k+1})}{\partial x_i} - \frac{\partial J(u_{k,i})}{\partial x_i} \right| \\ &\leq \sum_{i=1}^N \|\nabla J(u^{k+1}) - \nabla J(u_{k,i})\| \leq C_M \sum_{i=1}^N \underbrace{\|u^{k+1} - u_{k,i}\|}_{\xrightarrow{k \rightarrow +\infty} 0} \xrightarrow{k \rightarrow +\infty} 0 \end{aligned}$$

□

## Cas où $J$ est quadratique

On est ramené à résoudre :  $Au_{k,i} = b$ ,  $i = 1, \dots, N$ ,  $k \geq 0$ , ce qui équivaut à la résolution de

$$(D - E)u^{k+1} = Fu^k + b, \quad k \geq 0$$

où  $A = D - (E + F)$  est la décomposition usuelle de  $A$  pour les méthodes itératives. On retrouve donc le schéma de Gauss-Seidel.

*Remarque 1.* Par extension, la méthode de relaxation est dite de Gauss-Seidel non-linéaire.

## Relaxation avec contrainte

**Définition 2.2.2.** On considère le pavé. de  $\mathbb{R}^N$  :  $U = \prod_{i=1}^N [a_i, b_i]$ . On appelle suite générée par la méthode de relaxation avec contrainte sur  $U$  toute suite  $(u_k)_{k \geq 0}$  définie par :

$$u_0 \in \mathbb{R}^N$$

$$J(u_1^{k+1}, u_2^k, \dots, u_N^k) = \inf_{\mu \in [a_1, b_1]} J(\mu, u_2^k, \dots, u_N^k),$$

$$J(\underbrace{u_1^{k+1}, \dots, u_i^{k+1}}_{=: u_{k,i}}, u_{i+1}^k, \dots, u_N^k) = \inf_{\mu \in [a_i, b_i]} J(u_1^{k+1}, \dots, u_{i-1}^{k+1}, \mu, u_{i+1}^k, \dots, u_N^k),$$

$$i = 2, \dots, N - 1$$

$$J(u_1^{k+1}, \dots, u_N^{k+1}) = \inf_{\mu \in [a_N, b_N]} J(u_1^{k+1}, \dots, u_{N-1}^{k+1}, \mu),$$

**Proposition 2.2.2.** *La suite associée à la Définition 2.2.1 converge vers l'unique solution de*

$$u \in U \quad \text{et} \quad J(u) = \inf_{v \in U} J(v)$$

*Démonstration.* La démonstration est analogue à celle du Théorème 2.2.1. On signale les modifications. Soit  $k \geq 1$  et soit  $i \in [[1, N]]$ . On remarque que :

$$J(u_{k,i}) = \inf_{\mu \in [a_i, b_i]} J(u_{k,i-1} + \mu e_i) \Rightarrow \frac{\partial J(u_{k,i})}{\partial x_i}(v_i - u_i^k) \geq 0, \quad \forall v_i \in [a_i, b_i].$$

De plus :

$$J(u^{k+1}) = J(u_N^k) \leq J(u_0^k) = J(u^k)$$

i.e., la suite  $(J(u^k))_{k \geq 0}$  est décroissante, minorée par coercivité de  $J$ , donc convergente Il en résulte :

$$\begin{aligned} J(u^k) - J(u^{k+1}) &= J(u_{k,0}) - J(u_{k,N}) = \sum_{i=0}^{N-1} (J(u_{k,i}) - J(u_{k,i+1})) \geq \\ &\geq \sum_{i=0}^{N-1} \langle \nabla J(u_{k,i+1}), u_{k,i} - u_{k,i+1} \rangle + \frac{\alpha}{2} \sum_{i=0}^{N-1} \|u_{k,i} - u_{k,i+1}\|^2 \end{aligned}$$

avec :

$$\sum_{i=0}^{N-1} \langle \nabla J(u_{k,i+1}), u_{k,i} - u_{k,i+1} \rangle = \sum_{i=0}^{N-1} \underbrace{\frac{\partial J(u_{k,i+1})}{\partial x_{i+1}} (u_{i+1}^k - u_{i+1}^{k+1})}_{\geq 0} \geq 0,$$

Il en résulte :

$$J(u^k) - J(u^{k+1}) \geq \frac{\alpha}{2} \|u^k - u^{k+1}\|^2,$$

avec  $\lim_{k \rightarrow +\infty} J(u^k) - J(u^{k+1}) = 0$ . Du Théorème d'encadrement des gendarmes, il résulte que  $\lim_{k \rightarrow +\infty} \|u^k - u^{k+1}\| = 0$ .

On a :

$$\begin{aligned} \alpha \|u^{k+1} - u\|^2 &\leq \langle \nabla J(u^{k+1}) - \nabla J(u), u^{k+1} - u \rangle = \langle \nabla J(u^{k+1}), u^{k+1} - u \rangle = \\ &= \sum_{i=1}^N \frac{\partial J(u^{k+1})}{\partial x_i} (u_i^{k+1} - u_i^k) \leq \sum_{i=1}^N \left( \frac{\partial J(u^{k+1})}{\partial x_i} - \frac{\partial J(u_{k,i})}{\partial x_i} \right) (u_i^{k+1} - u_i^k) \\ &\leq \sum_{i=1}^N \left| \frac{\partial J(u^{k+1})}{\partial x_i} - \frac{\partial J(u_{k,i})}{\partial x_i} \right| |u_i^{k+1} - u_i^k| \\ &\leq \sum_{i=1}^N \|\nabla J(u^{k+1}) - \nabla J(u_{k,i})\| \|u^{k+1} - u_{k,i}\| \\ &\Rightarrow \alpha \|u^{k+1} - u\| \leq C_M \sum_{i=1}^N \underbrace{\|u^{k+1} - u_{k,i}\|}_{\xrightarrow[k \rightarrow +\infty]{0}} \xrightarrow[k \rightarrow +\infty]{} 0 \end{aligned}$$

□



## 2.3 Méthode de pénalisation

**Théorème 2.3.1.** Soit  $U \subset \mathbb{R}^N$  un convexe fermé. Soit  $J : \mathbb{R}^N \rightarrow \mathbb{R}$  vérifiant l'Hypothèse 1.0.1 et soit  $\psi : \mathbb{R}^N \rightarrow \mathbb{R}$  convexe vérifiant :

$$\psi(x) \geq 0, \quad \forall x \in \mathbb{R}^N, \quad \text{et} \quad \psi(x) = 0 \iff x \in U$$

Pour tout  $\varepsilon > 0$ , on pose :

$$J_\varepsilon(v) = J(v) + \frac{\psi(v)}{\varepsilon}, \quad \forall v \in \mathbb{R}^N,$$

et on note  $u^\varepsilon \in \mathbb{R}^N$  l'unique solution de :

$$J_\varepsilon(u^\varepsilon) = \min_{v \in \mathbb{R}^N} J_\varepsilon(v).$$

Alors, la suite  $(u^\varepsilon)_{\varepsilon>0}$  converge vers l'unique solution  $u$  de

$$u \in U \quad \text{et} \quad J(u) = \min_{v \in U} J(v).$$

*Démonstration.* Pour tout  $\varepsilon > 0$ , la coercivité de  $J$  entraîne :

$$b + \beta \|u^\varepsilon\|^2 \leq J_\varepsilon(u^\varepsilon) \leq J_\varepsilon(u) = J(u)$$

donc la suite  $(u^\varepsilon)_{\varepsilon>0}$  est bornée dans  $\mathbb{R}^N$ . Par compacité des boules fermées dans  $\mathbb{R}^N$  on déduit qu'il existe une suite extraite  $(u^{\varphi(\varepsilon)})_{\varepsilon>0}$  qui converge vers un  $u' \in \mathbb{R}^N$ . En particulier :

$$J(u') = \lim_{\varepsilon \rightarrow 0} J(u^{\varphi(\varepsilon)}) \leq \liminf_{\varepsilon \rightarrow 0} J_{\varphi(\varepsilon)}(u^{\varphi(\varepsilon)}) \leq J(u) \quad (10)$$

De plus :  $\forall \varepsilon > 0$ ,

$$0 \leq \psi(u^{\varphi(\varepsilon)}) \leq \varepsilon \underbrace{(J(u) - J(u^{\varphi(\varepsilon)}))}_{\xrightarrow{\varepsilon \rightarrow 0} J(u) - J(u')} \xrightarrow{\varepsilon \rightarrow 0} 0$$

Du Théorème d'encadrement des gendarmes et de la continuité de  $\psi$ , on déduit que :

$$\psi(u') = \lim_{\varepsilon \rightarrow 0} \psi(u^{\varphi(\varepsilon)}) = 0$$

i.e.  $u' \in U$ . De (10) et de l'unicité de  $u$  on déduit que  $u = u'$ . La limite  $u$  ne dépendant pas du choix de la suite extraite  $(u^{\varphi(\varepsilon)})_{\varepsilon>0}$ , on en déduit que toute la suite  $(u^\varepsilon)_{\varepsilon>0}$  converge vers  $u$ .  $\square$

## Bibliographie

- [1] P. Ciarlet, Introduction à l'analyse numérique matricielle et à l'optimisation, Masson, Dunod, Paris.
- [2] P. Lascaux, R. Théodor, Analyse numérique matricielle appliquée à l'art de l'ingénieur, Dunod, Paris.