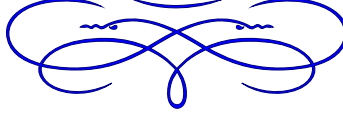


## Notes optimisation avancée



**Théorème 0.1.** Soit  $f$  une fonction réelle  $\alpha$ -fortement convexe et  $L$ -lipschitzienne sur  $E$ . Soit  $x^\star := \arg \min_{x \in E} f(x)$ . On a le résultat suivant:

$$f(\hat{x}) - f(x^\star) \leq \frac{2L^2}{\alpha T} \quad (1)$$

*Démonstration.* Pour  $t = 1$  à  $T$ , introduisons le pas:

$$\mu_t := \frac{2}{\alpha t} \quad (2)$$

En utilisant l'identité classique  $2\langle x, y \rangle = \|x + y\|^2 - \|x\|^2 - \|y\|^2$ , on a pour  $t = 0$  à  $T - 1$ :

$$\begin{aligned} f(\hat{x}) - f(x^\star) &\leq f(x_t) - \left( f(x_t) + u_t^T (x^\star - x_t) + \frac{\alpha}{2} \|x^\star - x_t\|^2 \right) \\ &= -u_t^T (x^\star - x_t) - \frac{\alpha}{2} \|x^\star - x_t\|^2 \\ &= -\frac{\mu_{t+1}}{\mu_{t+1}} \cdot u_t^T (x^\star - x_t) - \frac{\alpha}{2} \|x^\star - x_t\|^2 \\ &= \frac{1}{2\mu_{t+1}} \left( \|\mu_{t+1} \cdot u_t\|^2 + \|x^\star - x_t\|^2 - \|\mu_{t+1} \cdot u_t + x^\star - x_t\|^2 \right) - \frac{\alpha}{2} \|x^\star - x_t\|^2 \\ &= \frac{1}{2\mu_{t+1}} \left( \|\mu_{t+1} \cdot u_t\|^2 + \|x^\star - x_t\|^2 - \|x^\star - y_t\|^2 \right) - \frac{\alpha}{2} \|x^\star - x_t\|^2 \\ &\leq \frac{1}{2\mu_{t+1}} \left( \mu_{t+1}^2 \cdot L^2 + \|x^\star - x_t\|^2 - \|x^\star - x_{t+1}\|^2 \right) - \frac{\alpha}{2} \|x^\star - x_t\|^2 \\ &= \frac{\mu_{t+1} L^2}{2} + \underbrace{\left( \frac{1}{2\mu_{t+1}} - \frac{\alpha}{2} \right)}_{\frac{\alpha(t-1)}{4}} \|x^\star - x_t\|^2 - \underbrace{\frac{1}{2\mu_{t+1}}}_{\frac{\alpha(t+1)}{4}} \|x^\star - x_{t+1}\|^2 \end{aligned}$$

En utilisant la définition de  $\hat{x}$ , la convexité de  $f$  et l'inégalité d'au-dessous pour tout  $t$ , on a successivement:

$$\begin{aligned} f(\hat{x}) - f(x^\star) &= f\left( \sum_{t=0}^T \frac{2t}{T(T+1)} x_t \right) - f(x^\star) \\ &\leq \sum_{t=0}^T \frac{2t}{T(T+1)} (f(x_t) - f(x^\star)) \\ &\leq \sum_{t=0}^T \frac{2t}{T(T+1)} \left( \frac{\mu_{t+1} L^2}{2} + \frac{\alpha(t-1)}{4} \cdot \|x^\star - x_t\|^2 - \frac{\alpha(t+1)}{4} \cdot \|x^\star - x_{t+1}\|^2 \right) \\ &\leq \underbrace{\sum_{t=0}^T \frac{t}{T(T+1)} \mu_{t+1} L^2}_{:=A} + \underbrace{\sum_{t=0}^T \frac{\alpha t}{2T(T+1)} \left( (t-1) \cdot \|x^\star - x_t\|^2 - (t+1) \cdot \|x^\star - x_{t+1}\|^2 \right)}_{:=B} \end{aligned}$$

En posant  $\delta_t := t(t-1) \cdot \|x^\star - x_t\|^2$ , on peut réécrire la somme de gauche:

$$B = \frac{\alpha}{2T(T+1)} \sum_{t=0}^T (\delta_t - \delta_{t+1}) = \frac{\alpha}{2T(T+1)} (\delta_0 - \delta_{T+1}) \leq 0$$

Montrons maintenant que  $A \leq \frac{2L^2}{\alpha T}$ . En utilisant la définition du pas  $\mu_t$ , on a:

$$A = \frac{L^2}{T(T+1)} \sum_{t=0}^T \frac{2t}{\alpha(t+1)} = \frac{2L^2}{\alpha T} \cdot \frac{1}{(T+1)} \sum_{t=0}^T \underbrace{\frac{t}{t+1}}_{\leq 1} \leq \frac{2L^2}{\alpha T}$$

d'où le résultat. □

Un exemple très important est celui des fonctions  $\alpha$ -fortement convexe et  $\beta$  régulière. Remarquons que nécessairement  $\alpha \leq \beta$ .

## 1 Cas de fonctions $\alpha$ -fortement convexe et $\beta$ régulière

Soit  $f : \mathbb{R}^d \mapsto \mathbb{R}$  une fonction  $\alpha$ -fortement convexe et  $\beta$  régulière. On a donc pour tout  $x, x_0$  appartenant à  $\mathbb{R}^d$ :

**$\beta$ -régularité**  $f(x) \leq f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{\beta}{2} \|x - x_0\|^2$

**$\alpha$ -fortement convexe**  $f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{\alpha}{2} \|x - x_0\|^2 \leq f(x)$

Plusieurs cas :

- Si  $\alpha = \beta$ , cela signifie que  $f$  est quadratique;
- si  $\alpha \neq \beta$ , alors nécessairement  $\alpha \leq \beta$ .
- Si  $\alpha \approx \beta$ , alors on peut confondre la courbe de  $f$  et la courbe de  $\beta$  et on peut alors trouver le minimum très facilement.

**Théorème 1.1.** En reprenant les hypothèses sur  $f$  introduites au-dessus, on a le résultat suivant

$$\|\hat{x} - x^\star\|^2 \leq e^{-\frac{T}{K}} \|x_0 - x^\star\|^2 \tag{3}$$

où  $K := \frac{\beta}{\alpha}$ .

Remarquons tout d'abord que l'on a l'égalité suivante:

$$f(\hat{x}) - f(x^\star) = \int_0^1 \nabla f(x^\star + t(\hat{x} - x^\star))^T \cdot (\hat{x} - x^\star) dt \tag{4}$$

En introduisant la fonction auxiliaire  $\phi(t) := f(x^\star + t(\hat{x} - x^\star))$  définie sur  $[0; 1]$ , on a :

$$\phi(1) - \phi(0) = \int_0^1 \phi'(t) dt = \int_0^1 \nabla f(x^\star + t(\hat{x} - x^\star))^T \cdot (\hat{x} - x^\star) dt$$

ce qui montre bien (1).

En utilisant la forme intégrale pour  $f(\hat{x}) - f(x^\star)$  et en appliquant le théorème ci dessous, on obtient donc successivement:

$$\begin{aligned}
 f(\hat{x}) - f(x^\star) &= \int_0^1 \nabla f(x^\star + t(\hat{x} - x^\star)) - \nabla f(x^\star)^T \cdot (\hat{x} - x^\star) dt \\
 &\leq \int_0^1 \left\| \nabla f(x^\star + t(\hat{x} - x^\star)) - \nabla f(x^\star) \right\| \cdot \|\hat{x} - x^\star\| dt \\
 &\leq \int_0^1 \beta \|t(\hat{x} - x^\star)\| \cdot \|\hat{x} - x^\star\| dt \\
 &= \beta \|\hat{x} - x^\star\|^2 \cdot \int_0^1 t dt \\
 &= \frac{\beta \|\hat{x} - x^\star\|^2}{2} \\
 &\leq \frac{\beta}{2} e^{-T/K} \cdot \|x_0 - x^\star\|
 \end{aligned}$$

*Remarque à propos de la complexité de cet algorithme:*

- il faut que l'on ne parte pas trop loin de  $x^\star$  à cause de la dépendance au terme  $\|x_0 - x^\star\|$ .
- La complexité est en  $\log(\frac{1}{\varepsilon})$  ce qui est très rapide. Si  $\varepsilon \sim 10^{-9}$ ,  $\log(\frac{1}{\varepsilon}) \sim 9$  on l'on a besoin de seulement d'une dizaine d'étages pour avoir une précision à  $\varepsilon$ -près.

## 2 Exemple de fonction à la fois $\alpha$ -fortement convexe et $\beta$ régulière

Il suffit de prendre n'importe quelle fonction quadratique (non nulle).

Si  $A \in S_d^{++}$ ,  $b \in \mathbb{R}^d$  et  $f: \mathbb{R}^d \rightarrow \mathbb{R}$ ,  $x \mapsto \frac{1}{2}x^T A x - b^T x$ . Alors on a  $f$  qui est :

- $\lambda_{\min}(A)$ -fortement convexe,
- $\lambda_{\max}(A)$ -régulière.

En effet, on a  $f \in \mathcal{L}^2$  et  $\forall x \in \mathbb{R}^d$ ,  $\nabla^2 f(x) = A$ , d'où l'on a bien

$$\forall x \in \mathbb{R}^d, \quad \lambda_{\min} I_d \leq \nabla^2 f(x) \leq \lambda_{\max} I_d$$

En 10 étages, on peut trouver une solution à l'équation  $Ax = b$  à  $\varepsilon$  près, tandis que résoudre cette équation directement donne une complexité très importante.

En effet,  $f$  est minimale en  $\nabla f(x) = Ax - b = 0 \Leftrightarrow x = A^{-1}b$ . En appliquant l'algorithme de descente de gradient, on trouve  $\hat{x} \in \mathbb{R}^d$  tel que

$$\|\hat{x} - x^\star\|^2 \leq e^{-T/K} \cdot \|x_0 - x^\star\|^2$$

Si  $T \gg \underbrace{K}_{=\frac{\lambda_{\max}}{\lambda_{\min}}} \log\left(\frac{1}{\varepsilon}\right)$ , egsi la matrice est bien conditionné ( $K$  faible), alors on approche  $x^\star$  très rapidement.

*Démonstration.* On démontre ici le théorème énoncé plus haut.

On aimerait avoir pour tout  $t = 1$  à  $T$  :

$$\|x_t - x^\star\|^2 \leq \left(1 - \frac{1}{K}\right) \cdot \|x_{t-1} - x^\star\|^2 \quad (5)$$

ce qui impliquerait par récurrence immédiate

$$\|x_T - x^\star\|^2 \leq \underbrace{\left(1 - \frac{1}{K}\right)^T}_{e^{-T/K}} \cdot \|x_0 - x^\star\|^2 \quad (6)$$

On a :

$$\begin{aligned}\|x_t - x^\star\|^2 &= \left\| x_{t-1} - \frac{1}{\beta} \nabla f(x_{t-1}) - x^\star \right\|^2 \\ &= \|x_{t-1} - x^\star\|^2 + \underbrace{\frac{1}{\beta^2} \|\nabla f(x_{t-1})\|^2 - \frac{2}{\beta} \nabla f(x_{t-1})^T \cdot (x_{t-1} - x^\star)}_{\leq -\frac{1}{K} \|x_{t-1} - x^\star\|^2?}\end{aligned}$$

□