Optimisation convexe – Méthodes itératives

Descentes de gradient

Bashar Dudin

June 10, 2020

EPITA

Quand on vous dit la vérité.

Les limites de la démarche analytique

On s'intéresse dans ces slides à la résolution par des techniques itératives des problèmes d'optimisation[1].

[1] En premier lieu convexes.

Les limites de la démarche analytique

On s'intéresse dans ces slides à la résolution par des techniques itératives des problèmes d'optimisation[1].

La résolution analytique d'équations réelles est rarement aisée. Dans le cas où de telles expressions analytiques existent, elles peuvent être coûteuses à calculer, comme c'est déjà le cas pour la résolution de l'équation normale dans le cas de la régression linéaire.

[1] En premier lieu convexes.

Les limites de la démarche analytique

On s'intéresse dans ces slides à la résolution par des techniques itératives des problèmes d'optimisation[1].

La résolution analytique d'équations réelles est rarement aisée. Dans le cas où de telles expressions analytiques existent, elles peuvent être coûteuses à calculer, comme c'est déjà le cas pour la résolution de l'équation normale dans le cas de la régression linéaire. Les solutions sont parfois même impossibles à expliciter à l'aide des fonctions usuelles et des données du problème.

[1] En premier lieu convexes.

Les limites de la démarche analytique

On s'intéresse dans ces slides à la résolution par des techniques itératives des problèmes d'optimisation[1].

La résolution analytique d'équations réelles est rarement aisée. Dans le cas où de telles expressions analytiques existent, elles peuvent être coûteuses à calculer, comme c'est déjà le cas pour la résolution de l'équation normale dans le cas de la régression linéaire. Les solutions sont parfois même impossibles à expliciter à l'aide des fonctions usuelles et des données du problème.

Question

On se place dans le cadre de la régression logistique. Pourriez-vous calculer analytiquement une expression d'un point optimal?

[1] En premier lieu convexes.

Prinicpe général

Cadre

Hypothèse

Dans la suite, et à moins de faire explicitement mention du contraire, toutes nos fonctions sont supposées 2-fois différentiables de hessienne continue.

Cadre

Hypothèse

Dans la suite, et à moins de faire explicitement mention du contraire, toutes nos fonctions sont supposées 2-fois différentiables de hessienne continue.

On s'intéresse à un problème d'optimisation de la forme

Le problème (P)

$$\min_{x \in \mathbb{R}^n} f(x) \qquad \text{(P)}$$

où f sera supposée convexe et vérifiant l'hypothèse de régularité ci-dessus.

Principe des descentes de gradient

Algorithm 1 Principe des descentes de gradient

Input: f : a function, x_0 : an initial point in the domain of f
Output: x^* : an optimal solution of (P) if bounded from below

```
1: function GRADIENT_DESCENT(f, x_0)
2:     x ← x_0
3:     while not stopping condition do
4:         compute a direction Δx to update x
5:         compute step t > 0 of descent
6:         x ← x + tΔx
7:     end while
8:     return x
9: end function
```

Principe des descentes de gradient

Le principe général précédent nécessite quelques remarques, qu'ils nous faudra garder en mémoire par la suite.

Principe des descentes de gradient

Le principe général précédent nécessite quelques remarques, qu'ils nous faudra garder en mémoire par la suite.

- Chaque choix dans un algorithme de descente impacte la convergence de celui-ci ; l'initialisation, le choix de la direction ou celle du pas.

Principe des descentes de gradient

Le principe général précédent nécessite quelques remarques, qu'ils nous faudra garder en mémoire par la suite.

- Chaque choix dans un algorithme de descente impacte la convergence de celui-ci ; l'initialisation, le choix de la direction ou celle du pas.
- En pratique la boucle `while` apparaît à l'intérieur d'une boucle finie.

Principe des descentes de gradient

Le principe général précédent nécessite quelques remarques, qu'ils nous faudra garder en mémoire par la suite.

- Chaque choix dans un algorithme de descente impacte la convergence de celui-ci ; l'initialisation, le choix de la direction ou celle du pas.
- En pratique la boucle `while` apparaît à l'intérieur d'une boucle finie.
- Le pas est souvent (du moins dans un premier temps) pris constant. Cela peut poser des problèmes de convergence.

Principe des descentes de gradient

Le principe général précédent nécessite quelques remarques, qu'ils nous faudra garder en mémoire par la suite.

- Chaque choix dans un algorithme de descente impacte la convergence de celui-ci ; l'initialisation, le choix de la direction ou celle du pas.
- En pratique la boucle `while` apparaît à l'intérieur d'une boucle finie.
- Le pas est souvent (du moins dans un premier temps) pris constant. Cela peut poser des problèmes de convergence.
- La condition d'arrêt s'exprime souvent par le fait que la mise à jour n'a plus d'impact significatif.

Calcule du pas de la descente

Calculer le pas

Il arrive parfois qu'on ne se préoccupe que peu du pas de la descente. Notamment quand on lui garde une valeur constante. C'est un choix qui comporte des dangers, un pas constant peut donner lieu à une descente divergente.

Calculer le pas

Il arrive parfois qu'on ne se préoccupe que peu du pas de la descente. Notamment quand on lui garde une valeur constante. C'est un choix qui comporte des dangers, un pas constant peut donner lieu à une descente divergente.

Pour éviter cet écueil les matheux s'arment de deux approches:

Calculer le pas

Il arrive parfois qu'on ne se préoccupe que peu du pas de la descente. Notamment quand on lui garde une valeur constante. C'est un choix qui comporte des dangers, un pas constant peut donner lieu à une descente divergente.

Pour éviter cet écueil les matheux s'arment de deux approches:

- Le calcul du pas optimal : pour une direction choisie on calcule t minimisant $f(x + t\Delta x)$.

Calculer le pas

Il arrive parfois qu'on ne se préoccupe que peu du pas de la descente. Notamment quand on lui garde une valeur constante. C'est un choix qui comporte des dangers, un pas constant peut donner lieu à une descente divergente.

Pour éviter cet écueil les matheux s'arment de deux approches:

- Le calcul du pas optimal : pour une direction choisie on calcule t minimisant $f(x + t\Delta x)$.
- Le ***backtracking*** : une heuristique qui mime le précédent point tout en étant moins coûteuse.

Algorithm 2 Backtracking

Input: f : a function, x : a point in the domain of f
Input: Δx a descent direction
Input: $\alpha \in]0, 0.5[$, $\beta \in]0, 1[$
Output: t^* : a sub-optimal point minimizimg $f(x + t\Delta x)$ if bounded from below

1: **function** BACKTRACKING(f, x, $\alpha = 0.1$, $\beta = 0.8$)
2: $t \leftarrow 1$
3: **while** $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$ **do**
4: $t \leftarrow \beta t$
5: **end while**
6: **return** t
7: **end function**

Calculer le pas | *Backtracking*

À la fin de l'execution de `backtracking` on se retrouve dans l'une des deux situations suivantes:

Calculer le pas | *Backtracking*

À la fin de l'execution de `backtracking` on se retrouve dans l'une des deux situations suivantes:

- $t = 1$

Calculer le pas | *Backtracking*

À la fin de l'execution de `backtracking` on se retrouve dans l'une des deux situations suivantes:

- $t = 1$
- $t \in]\beta t_0, t_0]$ où t_0 est le plus grand réel satisfaisant la condition de boucle.

Convexité forte et conditionnement de la hessienne

Analyse de convergence

On désgine par S l'ensemble $\{x \in \mathrm{Dom}(f) \mid f(x) \leq f(x_0)\}$ où x_0 est sous-entendu être un point initial de descente de gradient. C'est un fermé de $\mathbb{R}^n$; toute suite de S convergente dans $\mathbb{R}^n$ a une limite dans S.

Analyse de convergence

On désgine par S l'ensemble $\{x \in \mathrm{Dom}(f) \mid f(x) \leq f(x_0)\}$ où x_0 est sous-entendu être un point initial de descente de gradient. C'est un fermé de $\mathbb{R}^n$; toute suite de S convergente dans $\mathbb{R}^n$ a une limite dans S.

L'analyse de convergence, sous-entendu de la vitesse de convergence, des méthodes de descentes s'effectue, outre l'hypothèse de régularité $\mathcal{C}^2$, sous l'une des deux conditions suivantes:

1. La hessienne de f est majorée sur S.

Analyse de convergence

On désgine par S l'ensemble $\{x \in \mathrm{Dom}(f) \mid f(x) \leq f(x_0)\}$ où x_0 est sous-entendu être un point initial de descente de gradient. C'est un fermé de $\mathbb{R}^n$; toute suite de S convergente dans $\mathbb{R}^n$ a une limite dans S.

L'analyse de convergence, sous-entendu de la vitesse de convergence, des méthodes de descentes s'effectue, outre l'hypothèse de régularité $\mathcal{C}^2$, sous l'une des deux conditions suivantes:

1. La hessienne de f est majorée sur S.
2. f est ***strictement convexe*** sur S ; c'est-à-dire qu'il existe $m > 0$ tel que pour tout $x \in S$, $H_f(x) \geq mI$. C'est une inégalité fonctionnelle entre formes quadratiques!

Analyse de convergence

On désgine par S l'ensemble $\{x \in \mathrm{Dom}(f) \mid f(x) \leq f(x_0)\}$ où x_0 est sous-entendu être un point initial de descente de gradient. C'est un fermé de $\mathbb{R}^n$; toute suite de S convergente dans $\mathbb{R}^n$ a une limite dans S.

L'analyse de convergence, sous-entendu de la vitesse de convergence, des méthodes de descentes s'effectue, outre l'hypothèse de régularité $\mathcal{C}^2$, sous l'une des deux conditions suivantes:

1. La hessienne de f est majorée sur S.
2. f est ***strictement convexe*** sur S ; c'est-à-dire qu'il existe $m > 0$ tel que pour tout $x \in S$, $H_f(x) \geq mI$. C'est une inégalité fonctionnelle entre formes quadratiques!

La seconde condition est la plus forte des deux ; elle implique la première.

Nombre de conditionnement de la hessienne

Définition

Si $f : \mathbb{R}^n \to \mathbb{R}$ est une fonction convexe $\mathcal{C}^2$, le nombre de conditionnement de $H_f(x)$ est le rapport de sa plus grande valeur propre à sa plus petite.

Nombre de conditionnement de la hessienne

Définition

Si $f : \mathbb{R}^n \to \mathbb{R}$ est une fonction convexe $\mathcal{C}^2$, le nombre de conditionnement de $H_f(x)$ est le rapport de sa plus grande valeur propre à sa plus petite.

- La hessienne de f étant symétrique positive et à valeurs réelles, elle est diagonalisable sur $\mathbb{R}$ à valeurs propres positives.

Nombre de conditionnement de la hessienne

Définition

Si $f : \mathbb{R}^n \to \mathbb{R}$ est une fonction convexe $\mathcal{C}^2$, le nombre de conditionnement de $H_f(x)$ est le rapport de sa plus grande valeur propre à sa plus petite.

- La hessienne de f étant symétrique positive et à valeurs réelles, elle est diagonalisable sur $\mathbb{R}$ à valeurs propres positives.
- Le nombre de conditionnement est ≥ 1.

Nombre de conditionnement de la hessienne

Définition

Si $f : \mathbb{R}^n \to \mathbb{R}$ est une fonction convexe $\mathcal{C}^2$, le nombre de conditionnement de $H_f(x)$ est le rapport de sa plus grande valeur propre à sa plus petite.

- La hessienne de f étant symétrique positive et à valeurs réelles, elle est diagonalisable sur $\mathbb{R}$ à valeurs propres positives.
- Le nombre de conditionnement est ≥ 1.
- Le rapport des bornes qui encadrent f dans le cas strictement convexe est un majorant des nombres de conditionnement.

Nombre de conditionnement de la hessienne

Définition

Si $f : \mathbb{R}^n \to \mathbb{R}$ est une fonction convexe $\mathcal{C}^2$, le nombre de conditionnement de $H_f(x)$ est le rapport de sa plus grande valeur propre à sa plus petite.

- La hessienne de f étant symétrique positive et à valeurs réelles, elle est diagonalisable sur $\mathbb{R}$ à valeurs propres positives.
- Le nombre de conditionnement est ≥ 1.
- Le rapport des bornes qui encadrent f dans le cas strictement convexe est un majorant des nombres de conditionnement.

Convergence

Les nombres de conditionnements de f sont corrélés à la vitesse de convergence de la descente de gradient.

La descente de gradient à l'ancienne

Algorithm 3 Descente de gradient

Input: f : a function, x_0 : an initial point in the domain of f, ε : tolerance
Output: x^* : an optimal solution of (P) if bounded from below

1: **function** GRADIENT_DESCENT(f, x_0, ε)
2: $x \leftarrow x_0$
3: **while** $\|\nabla f(x)\| > \varepsilon$ **do**
4: $\Delta x \leftarrow -\nabla f(x)$
5: compute step $t > 0$ of descent
6: $x \leftarrow x + t\Delta x$
7: **end while**
8: **return** x
9: **end function**

Proposition

Supposons f $\mathcal{C}^2$ ayant une hessienne majorée ; il existe $M \in \mathbb{R}_+$ tel que pour tout $x \in S$, $H_f(x)$. La descente de gradient avec un pas constant $t \leq \frac{1}{M}$ ou via *backtracking* garantit

$$|f(x_k) - f(x^*)| \leq \frac{\|x_0 - x^*\|_2}{2ck}$$

où

- x_0 est le point initial de la descente
- x_k le k-ème itéré de la descente
- c est égal à t dans le cas du pas constant et à $\min\{1, \frac{\beta}{M}\}$ dans le cas du *backtracking*.

$$|f(x_k) - f(x^*)| \leq \frac{\|x_0 - x^*\|_2}{2ck}$$

- La valeur absolue du membre de gauche est superflue.

$$|f(x_k) - f(x^*)| \leq \frac{\|x_0 - x^*\|_2}{2ck}$$

- La valeur absolue du membre de gauche est superflue.
- La vitesse de convergence dépend du point initial (étonnant ...).

Analyse de convergence | Hessienne majorée

$$|f(x_k) - f(x^*)| \leq \frac{\|x_0 - x^*\|_2}{2ck}$$

- La valeur absolue du membre de gauche est superflue.
- La vitesse de convergence dépend du point initial (étonnant ...).
- Pour un point sous-optimal à ε près on est sur une complexité en $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$.

$$|f(x_k) - f(x^*)| \leq \frac{\|x_0 - x^*\|_2}{2ck}$$

- La valeur absolue du membre de gauche est superflue.
- La vitesse de convergence dépend du point initial (étonnant ...).
- Pour un point sous-optimal à ε près on est sur une complexité en $\mathscr{O}\left(\frac{1}{\varepsilon}\right)$.
- L'intérêt du *backtracking* réside dans le fait qu'on n'a pas à calculer le pas à la main. Si β est proche de 1 on ne perd pas grand chose par rapport au pas constant.

Proposition

Supposons f $\mathcal{C}^2$ fortement convexe ; il existe m, $M \in \mathbb{R}_+$ encadrant la hessienne en tout point de S. La descente de gradient à pas constant $t \leq \frac{2}{M}$ ou via *backtracking* garantit

$$|f(x_k) - f(x^*)| \leq c^k \|x_0 - x^*\|_2$$

avec $c \in]0, 1[$.

- x_0 est le point initial de la descente
- x_k le k-ème itéré de la descente
- c est égal à $(1 - \frac{m}{M})$ dans le cas du pas constant et à $1 - \min\{2m\alpha, 2\beta\alpha\frac{m}{M}\}$ dans le cas du *backtracking*.

$$|f(x_k) - f(x^*)| \leq c^k \|x_0 - x^*\|_2$$

- La vitesse de convergence qu'on obtient ainsi est dite linéaire ; elle l'est si graphée contre une échelle logarithmique.

Analyse de convergence | Cas fortement convexe

$$|f(x_k) - f(x^*)| \leq c^k \|x_0 - x^*\|_2$$

- La vitesse de convergence qu'on obtient ainsi est dite linéaire ; elle l'est si graphée contre une échelle logarithmique.
- Pour un point sous-optimal à ε près on est sur une complexité en $\mathscr{O}\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$.

$$|f(x_k) - f(x^*)| \leq c^k \|x_0 - x^*\|_2$$

- La vitesse de convergence qu'on obtient ainsi est dite linéaire ; elle l'est si graphée contre une échelle logarithmique.
- Pour un point sous-optimal à ε près on est sur une complexité en $\mathscr{O}\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$.
- La constante c dépend très fortement de $\frac{M}{m}$.

Généraliser la descente classique

L'intuition qui mène à la descente de gradient classique répond en réalité à un problème de minimisation.

Généraliser la descente classique

L'intuition qui mène à la descente de gradient classique répond en réalité à un problème de minimisation. Étant donné un point $x \in S$ et un vecteur v de norme assez petite, on peut écrire

$$f(x+v) = f(x) + \nabla f(x)^T v + o(v).$$

La *direction* de descente qui minimise au plus la valeur objective est donnée par

$$\Delta x_{nsd} = \operatorname{argmin} \left\{ \nabla f(x)^T v \mid \|v\|_2 = 1 \right\}.$$

C'est une conséquence de l'inégalité de Cauchy-Shwarz.

La descente de plus forte pente

La démarche précédente nous permet de définir différentes variantes des descentes de gradients ; en réalité l'essentielle des descentes de gradients.

La descente de plus forte pente

La démarche précédente nous permet de définir différentes variantes des descentes de gradients ; en réalité l'essentielle des descentes de gradients.

Définition

Soient $x \in S$ et $\|\cdot\|$ une norme sur $\mathbb{R}^n$. On désigne par Δx_{nsd} (nsd pour *nomralized steepest descent*) la quantité

$$\Delta x_{nsd} = \operatorname{argmin} \left\{ \nabla f(x)^T v \mid \|v\| = 1 \right\}$$

La direction de descente de plus forte pente (sous-entendu pour la norme $\|\cdot\|$) est donnée par

$$\Delta_{sd} = \|\nabla f(x)\|_* \Delta x_{nsd}$$

où $\|\cdot\|_*$ désigne la norme d'opérateur associée à $\|\cdot\|$.

La descente de plus forte pente pour une norme $\|\cdot\|$ s'interprète comme

le vecteur de la sphère unité pour $\|\cdot\|$ de plus grande projection sur $-\nabla f(x)$.

La descente de plus forte pente pour une norme $\|\cdot\|$ s'interprète comme

le vecteur de la sphère unité pour $\|\cdot\|$ de plus grande projection sur $-\nabla f(x)$.

Pour tout vecteur v dans la sphère unité pour la norme $\|\cdot\|$, $\nabla f(x)^T v$ a pour valeur absolue la norme de la projection orthogonale de v sur $\nabla f(x)$. Dans la mesure où l'on cherche un minimisant c'est la plus grande projection contre $-\nabla f(x)$.

Algorithm 4 Descente de gradient de plus forte pente

Input: f : a function, x_0 : an initial point in the domain of f, ε : tolerance, $\|\cdot\|$ une norme sur $\mathbb{R}^n$.

Output: x^* : an optimal solution of (P) if bounded from below

1: **function** STEEPEST_GRADIENT_DESCENT(f, x_0, ε, $\|\cdot\|$)
2: $x \leftarrow x_0$
3: **while** $\|\nabla f(x)\| > \varepsilon$ **do**
4: Compute steepest descent direction Δx_{sd} for $\|\cdot\|$.
5: compute step $t > 0$ of descent
6: $x \leftarrow x + t\Delta x_{sd}$
7: **end while**
8: **return** x
9: **end function**

- La stratégie de descente de plus forte pente permet de varier les types de descentes de gradients. Les géométries sous-jacentes sont parfois plus adaptées à certains problèmes qu'à d'autres.

- La stratégie de descente de plus forte pente permet de varier les types de descentes de gradients. Les géométries sous-jacentes sont parfois plus adaptées à certains problèmes qu'à d'autres.
- L'analyse de la convergence dans le cas de descente classique s'étend au cas de plus forte pente. La raison en est l'équivalence des différentes normes sur $\mathbb{R}^n$. L'essentiel des propriétés utilisés lors des encadrements étant partagées par celles-ci.

C’est tout pour cette séance!
Un TP sera mis à votre disposition en début de semaine prochaine pour la prochaine séance.