Optimisation convexe – Méthodes itératives

Descentes de gradient

Bashar Dudin

June 8, 2018

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 sur $\mathbb{R}$ 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 les problèmes quadratiques.

[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 sur $\mathbb{R}$ 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 les problèmes quadratiques. 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 sur $\mathbb{R}$ 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 les problèmes quadratiques. 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.

Quand on vous dit la vérité.

Principe des méthodes de descente

- Prinicpe général
- Calcule du pas de la descente
- Convexité forte et conditionnement de la hessienne

La classique

Les descentes de plus fortes pentes

La méthode de Newton

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 différentielle seconde continue.

Cadre

Hypothèse

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

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

Le problème (P)

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

où 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: $x \leftarrow x_0$
3: **while** `not` stopping condition **do**
4: compute a direction Δx to update x
5: compute step $t > 0$ of descent
6: $x \leftarrow x + t\Delta x$
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'est plus significative.

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: $t*$: an 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 condtion 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 $\mathscr{C}^2$, sous l'une des deux conditions suivantes:

1. La hessienne de f est majorée sur S.

Analyse de convergence

On désigne 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 $\mathscr{C}^2$, 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 $x \in S$, $\nabla^2 f(x) \geq m$.

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 $\mathscr{C}^2$, 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 $x \in S$, $\nabla^2 f(x) \geq m$.

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

Nombre de conditionnement de la hessienne

Définition

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

- La hessienne de f étant symétrique positive et à valeur réelle, 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 $\mathscr{C}^2$, le nombre de conditionnement de $\nabla^2 f(x)$ est le rapport de sa plus grande valeur propre à sa plus petite.

- La hessienne de f étant symétrique positive et à valeur réelle, 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 $\mathscr{C}^2$, le nombre de conditionnement de $\nabla^2 f(x)$ est le rapport de sa plus grande valeur propre à sa plus petite.

- La hessienne de f étant symétrique positive et à valeur réelle, 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 $\mathscr{C}^2$, le nombre de conditionnement de $\nabla^2 f(x)$ est le rapport de sa plus grande valeur propre à sa plus petite.

- La hessienne de f étant symétrique positive et à valeur réelle, 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**