Descente de gradient
Le 15/10/2020

$f : \mathbb{R}^n \to \mathbb{R}$ $\quad (n \geq 1)$

Problème : trouver $x_* \in \mathbb{R}^n$ tq $f(x_*) = \min_{x \in \mathbb{R}^n} f(x)$.

On suppose qu'un minimiseur existe. On le note x_*.

Déf : $\forall x \in \mathbb{R}^n$, $\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x) \right)$ (gradient)

si f est différentiable

$$\text{Hess } f(x) = \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right)_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}.$$ (hessienne)

si f est 2 fois différentiable

Prop. informelle : si f est différentiable, $\forall x \in \mathbb{R}^n$,

pour tout y proche de x,

$$f(y) \approx f(x) + \langle \nabla f(x), y - x \rangle.$$

$-\nabla f(x)$: direction dans laquelle f décroît le plus vite près de x.

Déf (descente de gradient) :

- on part de $x_0 \in \mathbb{R}^n$;
- on définit, $\forall t \in \mathbb{N}$,

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

(α_t : pas (« stepsize »))

Idéalement, on espère $x_t \xrightarrow{t\to+\infty} x_*$.

ou au moins $f(x_t) \xrightarrow{t\to+\infty} f(x_*) = \min f$.

But aujourd'hui : sous quelles hypothèses de f ces propriétés sont-elles vraies ?
Vitesse de convergence ?

1.2) Choix des pas $(\alpha_t)_{t\in\mathbb{N}}$.
α_t trop petit $\longrightarrow$ algorithme trop lent
α_t trop grand $\longrightarrow$ l'algorithme diverge.

1. Pas pré-définis
Par exemple, $\alpha_t = \eta \quad \forall t \in \mathbb{N}$ (pas constant)
(pour un $\eta > 0$)
$\alpha_t = \frac{1}{t+1} \quad \forall t \in \mathbb{N}$ (pas décroissant)

2. Exact line search :
$\forall t, \quad \alpha_t = \underset{\alpha \geq 0}{\operatorname{argmin}} f(x_t - \alpha \nabla f(x_t))$

Difficile à calculer.

3. Backtracking line search

Principe : $\forall t$, $f(x_t - \alpha_t \nabla f(x_t))$
$\approx f(x_t) + \langle \nabla f(x_t), (x_t - \alpha \nabla f(x_t)) - x_t \rangle$
$\approx f(x_t) - \alpha_t \|\nabla f(x_t)\|^2$
(si α_t est assez petit)

Algo: 1) on pose $\alpha_t = \alpha_{max}$ (α_{max} paramètre choisi à l'avance)

2) tant que $f(x_t - \alpha_t \nabla f(x_t)) > f(x_t) - c\,\alpha_t \|\nabla f(x_t)\|^2$,

↳ paramètre choisi à l'avance dans $]0;1[$

$\alpha_t = \tau\,\alpha_t$

↳ paramètre choisi à l'avance dans $]0;1[$.

3) on renvoie α_t.

1.3) Rappel : fonctions quadratiques

Soit $f : x \longrightarrow \frac{1}{2}\langle x, Mx\rangle + \langle b, x\rangle$

où $M \in \mathbb{R}^{n\times n}$ est symétrique et $b \in \mathbb{R}^n$.

Prop : $\forall x \in \mathbb{R}^n$, $\nabla f(x) = Mx + b$,

$\text{Hess}\, f(x) = M$.

f convexe $\iff$ M semi-définie positive ($M \geq 0$)

$\iff$ les valeurs propres de M sont ≥ 0.

Thm : Supposons $M \geq 0$

Considérons la descente de gradient avec pas constant $\alpha < \frac{2}{\lambda_{max}(M)}$.

(i) Si $\lambda_{min}(M) > 0$,

$\forall t \in \mathbb{N}$, $\|x_t - x_*\| \leq \rho^t \|x_0 - x_*\|$

pour un $\rho \in]0;1[$.

(ii) Même si $\lambda_{min}(M) = 0$,

$\forall t \in \mathbb{N}$, $f(x_t) - f(x_*) \leq \frac{\|x_0 - x_*\|}{\tau\, t}$

pour un $\tau > 0$.

II) Convergence

2.1) Fonctions lisses

Déf: soit $M>0$. On dit que f est M-lisse (M-smooth) si ∇f est M-lipschitzien, càd $\forall x,y\in\mathbb{R}^n$, $\|\nabla f(x)-\nabla f(y)\|\leq M\|x-y\|$

Rq: si f est deux fois différentiable,
f est L-lisse $\iff$ $|||\text{Hess } f(x)||| \leq L \ \forall x$
(norme spectrale : plus grande valeur propre en valeur absolue)

Ex: pour la fonction quadratique $f: x\to \frac{1}{2}\langle x, Mx\rangle + \langle x, p\rangle$, f est L-lisse ssi $-L\leq \lambda_{min}(M)\leq \lambda_{max}(M)\leq L$.

Lemme: Soit $L>0$ fixé. Si f est L-lisse, alors $\forall x,y\in\mathbb{R}^n$,
$$f(y)\leq f(x)+\langle \nabla f(x), y-x\rangle + \frac{L}{2}\|x-y\|^2$$

Dém: Soit $x,y\in\mathbb{R}^n$.

$$f(y) = f(x) + \int_0^1 \underbrace{\langle \nabla f(x+t(y-x)), y-x\rangle}_{=g'(t) \text{ où } g(t)=f(x+t(y-x))}\, dt$$

$g(1)=g(0)+\int_0^1 g'(t)\,dt$

$$= f(x) + \int_0^1 \big[\langle \nabla f(x+t(y-x))-\nabla f(x), y-x\rangle + \langle \nabla f(x), y-x\rangle\big]\, dt$$

$$\leq f(x) + \int_0^1 \langle \nabla f(x), y-x\rangle\, dt + \int_0^1 \underbrace{\|\nabla f(x+t(y-x))-\nabla f(x)\|}_{\leq L\|t(y-x)\| = Lt\|y-x\|} \times \|y-x\|\, dt$$

$$= f(x) + \int_0^1 \langle \nabla f(x), y-x\rangle\, dt + \int_0^1 Lt\|y-x\|^2\, dt$$

$$= f(x) + \langle \nabla f(x), y-x \rangle + \frac{L}{2}\|y-x\|^2. \qquad \square$$

Corollaire : pour tout t,

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1}-x_t \rangle + \frac{L}{2}\|x_{t+1}-x_t\|^2$$

Or $x_{t+1} = x_t - \alpha_t \nabla f(x_t)$
donc $x_{t+1} - x_t = -\alpha_t \nabla f(x_t)$

$$= f(x_t) - \alpha_t \|\nabla f(x_t)\|^2 + \frac{L}{2}\alpha_t^2 \|\nabla f(x_t)\|^2$$

$$= f(x_t) - \alpha_t \left(1 - \frac{L\alpha_t}{2}\right) \|\nabla f(x_t)\|^2.$$

Si $\alpha_t = \frac{1}{L}$, $\forall t$, $f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2$.

Corollaire : Soit $L>0$. Supposons f L-lisse et bornée inférieurement. Alors :

(i) $(f(x_t))_{t\in\mathbb{N}}$ converge vers une valeur finie.
(ii) $\|\nabla f(x_t)\| \xrightarrow{t\to+\infty} 0$:

→ pour la descente de gradient à pas constant $\alpha_t = \frac{1}{L}$ $\forall t$

Pour des résultats de convergence plus forts, il faut plus d'hypothèses.

2.2) Fonctions convexes lisses

Déf : f est convexe si $\forall x, y \in \mathbb{R}^n$, $\forall t \in [0;1]$,

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y).$$

Prop : si f est différentiable, f est convexe ssi
$\forall x, y \in \mathbb{R}^n$,
$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

Rq : si f est deux fois différentiable, f est convexe ssi $\text{Hess } f(x) \geq 0$ pour tout $x \in \mathbb{R}^n$.

Thm : Soit $L > 0$. On suppose f L-lisse et convexe.
On considère la descente de gradient à pas constant égal à $1/L$.
$$\forall t \in \mathbb{N}, \quad f(x_t) - f(x_*) \leq \underbrace{\frac{2L \|x_0 - x_*\|^2}{t+4}}_{O(1/t)}.$$

Dém : première étape :
$$\forall t \in \mathbb{N}, \quad \|x_{t+1} - x_*\| \leq \|x_t - x_*\|. \quad \text{(Admis).}$$

Deuxième étape :
on trouve une inégalité qui relie $f(x_{t+1}) - f(x_*)$ et $f(x_t) - f(x_*)$.
$$f(x_{t+1}) - f(x_*) \leq f(x_t) - f(x_*) - \frac{1}{2L} \|\nabla f(x_t)\|^2 \quad (*)$$
(propriété)
$$f(x_*) \geq f(x_t) + \langle \nabla f(x_t), x_* - x_t \rangle$$
(par convexité de f)
donc $f(x_t) - f(x_*) \leq \langle \nabla f(x_t), x_* - x_t \rangle \leq \|\nabla f(x_t)\| \, \|x_* - x_t\|$
$\leq \|\nabla f(x_t)\| \, \|x_* - x_0\|$
donc $$\|\nabla f(x_t)\| \geq \frac{f(x_t) - f(x_*)}{\|x_* - x_0\|}.$$
Avec $(*)$: $$f(x_{t+1}) - f(x_*) \leq f(x_t) - f(x_*) - \frac{1}{2L} \left(\frac{f(x_t) - f(x_*)}{\|x_* - x_0\|} \right)^2$$

$$\forall t,\ \frac{1}{f(x_{t+1})-f(x_*)} \geq \frac{1}{(f(x_t)-f(x_*))\left(1-\frac{f(x_t)-f(x_*)}{2L\|x_0-x_*\|^2}\right)}$$

$$\geq \frac{1}{f(x_t)-f(x_*)}\left(1+\frac{f(x_t)-f(x_*)}{2L\|x_0-x_*\|^2}\right)$$

$$\left(\forall c\in[0;1],\ \frac{1}{1-c}\geq 1+c\right)$$

$$= \frac{1}{f(x_t)-f(x_*)} + \frac{1}{2L\|x_0-x_*\|^2}$$

Donc $\forall t$,

$$\frac{1}{f(x_t)-f(x_*)} \geq \frac{1}{f(x_0)-f(x_*)} + \frac{t}{2L\|x_0-x_*\|^2}$$

$$\geq \frac{2}{L\|x_0-x_*\|^2} + \frac{t}{2L\|x_0-x_*\|^2}$$

(admis)

$$= \frac{t+4}{2L\|x_0-x_*\|^2}.$$

Donc $\forall t$, $f(x_t)-f(x_*) \leq \frac{2L\|x_0-x_*\|^2}{t+4}$. □

2.3) Fonctions lisses et fortement convexes

Déf: si f est différentiable, on dit qu'elle est μ-fortement convexe (pour un $\mu>0$) si $\forall x, y$,

$$f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle + \frac{\mu}{2}\|x-y\|^2.$$

Rq: f μ-fortement convexe $\Leftrightarrow$ $f_\mu : x \to f(x) - \frac{\mu}{2}\|x\|^2$ est convexe.

Rq: si f est deux fois dérivable,
f μ-fortement convexe
$\Leftrightarrow$ $\forall x \in \mathbb{R}^n$, $Hess f(x) - \mu Id \geq 0$.
$\Leftrightarrow$ Les valeurs propres de $Hess f(x)$ sont $\geq \mu$.

Ex: pour $f : x \to \frac{1}{2}\langle x, Mx\rangle + \langle x, b\rangle$,
f est fortement convexe ssi $\lambda_{min}(M) > 0$.
Si $\lambda_{min}(M) > 0$, f est $(\lambda_{min}(M))$-fortement convexe.

Thm: Soient $0 < \mu < L$ fixés. Soit f L-lisse et μ-fortement convexe.
On considère la descente de gradient à pas constant $\alpha_t = \frac{1}{L}$ $\forall t$.
Pour tout $t \in \mathbb{N}$,

$$f(x_t) - f(x_*) \leq \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^t \|x_* - x_0\|^2.$$

$\frac{L-\mu}{L+\mu} \in]0;1[$ → vitesse géométrique

Rq: $\frac{L-\mu}{L+\mu} = \frac{L/\mu - 1}{L/\mu + 1}$

$\frac{L}{\mu}$ est le « conditionnement »
Plus $\frac{L}{\mu}$ est proche de 1, plus la vitesse de convergence est rapide.

III) Accélération

Revenons au cas où f est L-lisse et convexe mais pas fortement convexe.
On a vu que la descente de gradient convergeait à vitesse $O(1/t)$. Résultat optimal : pour certaines fonctions f, c'est réellement la vitesse observée.

Algorithme plus rapide que la descente de gradient?

Oui, descente accélérée, algorithme de Nesterov

On se limite aux algos de 1er ordre (qui n'ont accès qu'aux valeurs de f et ∇f).

Deux "erreurs" faites par la descente de gradient:
- oubli de l'info des étapes précédentes
- regarde $\nabla f(x_t)$ alors que la valeur du gradient serait peut-être plus informative en un autre point.

Algo de Nesterov:

2 suites définies en parallèle,

$(x_t)_{t\in\mathbb{N}}$ (minimiseurs approchés)

$(y_t)_{t\in\mathbb{N}}$ (points où évaluer le gradient)

Définition de x_{t+1} et y_{t+1} $\forall t$:

$$x_{t+1} = y_t - \frac{1}{L}\nabla f(y_t)$$

$$y_{t+1} = x_{t+1} + \underbrace{\gamma_t (x_{t+1} - x_t)}_{\text{<< moment >>}}$$

où $(\gamma_t)_{t\in\mathbb{N}}$ est définie par

$$\begin{cases} \lambda_{-1} = 0 \\ \forall t\in\mathbb{N},\ \lambda_t = \dfrac{1+\sqrt{1+4\lambda_{t-1}^2}}{2} \\ \forall t,\ \gamma_t = \dfrac{\lambda_t - 1}{\lambda_{t+1}}. \end{cases}$$

Thm: soit $L>0$. On suppose que f est convexe et L-lisse.
Pour tout $t\in\mathbb{N}$,

$$f(x_t) - f(x_*) \leq \frac{2L}{(t+1)^2}\|x_0 - x_*\|^2. \quad) \ O(1/t^2)$$