Apprentissage en grande domension

January 25, 2017

$$\min_{\beta \in \mathbb{R}} f(\beta) \tag{1}$$

Conditions: f convexe:

$$f(y) >= f(x) + \nabla f(x)^T (y - x) \tag{2}$$

Definition 1:

$$\forall \theta \in [0, 1] \tag{3}$$

Def 3 M
Def 4 Lipschizsienne

$$\forall x, y ||f(x) - f(y)||_2 <= L||x - y||_2 \tag{4}$$

Def 5 contractant

$$LLipschitzavec 0 <= L < 1 \tag{5}$$

Them 1 Thm point fixe: f est $\alpha-$contractant,

$$\exists x^* telque f^* = f(x^*) \tag{6}$$

La suite definie par $x_{n+1} = f(x_n)$ converge vers x^* et vérifie

$$||x_n - x^*||_2 <= \frac{\alpha^n}{1 - \alpha} ||x_0 - x_1||_2 \tag{7}$$

Gradient Algo
Prop 5 Gradient monotone f diff est convexe, si et seulement si

$$\begin{aligned} &(\nabla f(x) - \nabla f(y))^T (x - y) >= 0 \\ &= \nabla f(x) f - \text{consistante} \end{aligned} \tag{8}$$

PREUVE 1.$\Rightarrow$:

$$f(y) >= f(x) + \nabla f(x)^T (y - x) \tag{9}$$

$$f(x) >= f(y) + \nabla f(y)^T (x - y) \tag{10}$$

$$-f(x) - f(y) < -f(x) - f(y) + \nabla f(x)^T (x - y) - \nabla f(y)^T (x - y) \tag{11}$$

$$(\nabla f(x) - \nabla f(y))^T (x - y) >= 0 \tag{12}$$

2. $\Leftarrow$: On introduit une fonction Φ:

$$\Phi(t) = f(x + t(y - x)) \tag{13}$$

$$\Phi'(t) = \nabla f(x + t(y - x))^T (y - x) \tag{14}$$

Comme ∇f est monotone

$$\Phi'(t) >= \Phi'(0), t >= 0 \tag{15}$$

$$f(y) - \Phi(1) = \Phi(0) + \int_0^1 \Phi'(t)dt \tag{16}$$

$$f(y) >= \Phi(0) + \Phi'(0) = f(x) + \nabla f(x)^T (y - x) \tag{17}$$

Theorème Boîte quadratique supérieure

$$f \sim L^1, \nabla f est L - lipschitz \tag{18}$$

Alors

$$g(x) = \frac{L}{2} x^T x - f(x) est convexe \tag{19}$$

$$f(y) <= \nabla <= \nabla f(x)^T (y - x) + \frac{L}{2} ||x - y||_2^2 \tag{20}$$

1. ∇f Lipschitz

$$||\nabla f(y) - \nabla f(x)||_2 <= L||y - x||_2 \tag{21}$$

2.

$$\begin{aligned} (\nabla f(y) - \nabla f(x))^T (y - x) &<= ||\nabla f(y) - \nabla f(x)||_2 ||y - x||_2 \\ &<= L||y - x||_2^2 \end{aligned} \tag{22}$$

$$\nabla g(x) = Lx - \nabla f \tag{23}$$

$$\begin{aligned} &(\nabla g(x) - \nabla g(y))^T (x - y) \\ =&(Lx - \nabla f(x) - Ly + \nabla f(y))^T (x - y) \\ =& - (\nabla f(y) - \nabla f(x))^T (y - x) + L||x - y||_2^2 \\ >=&0 \end{aligned} \tag{24}$$

$$y = x - t\nabla f(x) \tag{25}$$

$$f(x - t\nabla f(x)) <= f(x) + t(1 - \frac{Lt}{2})||\nabla f(x)||_2^2 \tag{26}$$

choix de t tel que $0 <= t < \frac{1}{2}$

$$x^+ = x - t\nabla f(x) \tag{27}$$

$$\begin{aligned} f(x^+) <=&f(x) + f(1 - \frac{Lt}{2})||\nabla f(x)||_2^2 \\ <=&f(x) - \frac{t}{2}||\nabla f(x)||_2^2 \end{aligned}$$

$$\begin{aligned} <= f^* + \nabla f(x)^T (x - x^*) - \frac{t}{2}||\nabla f(x)||^2 \\ =&f^* + \frac{1}{2t}(||x - x^*||_2^2 - ||x - x^* - t\nabla f(x)||_2^2) \\ =&f^* + \frac{1}{2t}(||x - x^*||_2^2 - ||x^+ - x^*||_2^2) \end{aligned} \tag{28}$$

$$\begin{aligned}\sum_{k=1}^{N}(f(x_k)-k^*) <=& \frac{1}{2t}\sum_{k=1}^{N}(||x_{k-1}-x^*||_2^2-||x_k-x^*||_2^2)\\ =&\frac{1}{2t}(||x_0-x^*||_2^2-||x_N-x^*||_2^2)\\ <= \frac{1}{2t}||x_0-x^*||_2^2\end{aligned} \tag{29}$$

Prop: Quand f est differenciable

$$f(y) >= f(x) + \nabla f(x)^T(y-x) \tag{30}$$

Definition: sous gradient g est un sous gradient de f en x, ssi

$$\forall y, f(y) >= f(x) + g^T(y-x) \tag{31}$$

Definition: sous differentielle f convexe, on definit la sous differentielle de f en x comme

$$\partial f(x) = \{g|\forall y, f(y) >= f(x) + g^T(y-x)\} \tag{32}$$

Theoreme 3:

$$x^* = argminf \Leftrightarrow 0 \in \partial f(x^*) \tag{33}$$

Si $0 \in \partial f(x^*)$, alors

$$\forall y, f(y) >= f(x^*) + 0^T(y-x^*) \Leftrightarrow x^= argminf \tag{34}$$

Prop 7: linéarité non négative f_1 et f_2 convexes, $\alpha_1, \alpha_2 >= 0$

$$f >= \partial(\alpha_1 f_1 + \alpha_2 f_2)(x) = \alpha_1 \partial f_1(x) + \alpha_2 \partial f_x(x) \tag{35}$$

+ addition d'ensemble

$$E + F = \{e + favece \in E, f \in F\} \tag{36}$$

Prop 8: combinaison affine: Si $h(x) = f(Ax+b)$, alors

$$\partial h(x) = A^T \partial f(Ax+b) \tag{37}$$

f est une fonction G-Lipschitzienne
ALGO: Méthode du "sous-gradient"

$$x_k \leftarrow x_{k-1} - t_k g_{k-1} \tag{38}$$

ou

$$g_{k-1} \in \partial f(x_k - 1) \tag{39}$$

Trois possibilité pour t_k

1. $t_k = t$

2. "Longueur constante" $t_k||g_{k-1}||_2 estconstante$

3.

$$t_k \to_{k\to+\infty} 0 \tag{40}$$

$$\sum_{k=1}^{+\infty} = +\infty \tag{41}$$

$$\sum_{k=1}^{+\infty} t_k^2 = \text{limite finie} \tag{42}$$

Theoreme: f convexe et non differentielle f est G-Lipschitzienne $\Leftrightarrow ||g||_2 <= G, \forall g \in \partial f(x)$

Preuve: $\Leftarrow$

On suppose $\forall x, \forall g \in \partial f(x)$

$$||g||_2 <= G \tag{43}$$

Soit $x(g_x)$ et $y(g_y)$

$$g_x^T(x-y) >= f(x) - f(y) >= g_y^T(x-y) \tag{44}$$

$$G||x-y||_2 >= f(x) - f(y) >= -G||x-y||_2 \tag{45}$$

$$\forall x, y, ||f(x) - f(y)|| <= G||x-y||_2 \tag{46}$$

$\Rightarrow \exists g$ tel que $||g||_2 > G$

$$y = x + \frac{g}{||g||_2} \tag{47}$$

$$f(y) >= f(x) + g^T(y-x) = f(x) + ||g||_2 > f(x) + G \tag{48}$$

Pas possible car f est G-Lipschitzienne

Attention: La méthode du sous-gradient n'est pas une méthode de descente.

$$x^+ = tg \tag{49}$$

g sous-gradient de f en x.

$$\begin{aligned} ||x^+ - x^*||_2^2 = ||x - tg - x^*||_2^2 \\ =&||x-x^*||_2^2 + t^2||g||_2^2 - 2tg^T(x-x^*) \\ <=&||x-x^*||_2^2 + t^2||g||_2^2 - 2t(f(x) - f^*) \end{aligned} \tag{50}$$

Pour une iteration k:

$$2t_k(f(x_{k-1}) - f^*) < ||x_{k-1} - x^*||_2^2 - ||x_k - x^*||_2^2 + t_k^2||g_{k-1}||_2^2 \tag{51}$$

en sommant les inégalités

$$\begin{aligned} 2(\sum_{k=1}^{N} t_k)(f_{best}^{(}N) - f^*) &<= ||x_0 - x^*||_2^2 - ||x_N - x^*||_2^2 + \sum_{k=1}^{N} t_k^2 ||g_{k-1}||_2^2 \\ &<= ||x_0 - x^*||_2^2 + \sum_{k=1}^{N} t_k^2 ||g_{k-1}||_2^2 \end{aligned} \tag{52}$$

1. $t_k = t$

$$f_{best}^{(N)} - f^* <= \frac{||x_0 - x^*||_2^2}{2Nt} + \frac{G^2 t}{2} \tag{53}$$

2. $t_k ||g_{k-1}||_2 = s$

$$f_{best}^{(N)} - f^* <= \frac{G||x_0 - x^*||_2^2}{2Ns} + \frac{Gs}{2} \tag{54}$$

3. $t_k \to 0, \sum t_k \to +\infty, \sum t_k^2$ converge

$$f_{best}^{(N)} - f^* <= \frac{||x_0 - x^*||_2^2 + \sigma^2 \sum t_k^2}{2 \sum t_k} \tag{55}$$

Conclusion: La méthode du sous gradient n'est pas facile à paramétrer pour obtenir sa convergence.

Exercise:

$$f(\beta) = ||X\beta - y||_2^2 + \lambda ||\beta||_1 \tag{56}$$

$$\partial f(\beta) = X^T (X\beta - y) + \lambda \partial_{||\dot{|}|_1}(\beta) \tag{57}$$

$$[\partial_{||\dot{|}|_1}(\beta)] = \left\{ \begin{array}{ll} sign(\beta_i) & \text{si}\beta_i \neq 0 \\ [-1, 1] & \text{si}\beta_i = 0 \end{array} \right. \tag{58}$$

Definition Operateur proximal

$$prox_f(x) = argmin_u \{f(u) + \frac{1}{2}||u - x||_2^2\} \tag{59}$$

f convexe "semi-continue inférieurement"(sci). alors, $prox_f(x)$ existe et est unique.

Theoreme Caractérisation par le sous-gradient

$$u = prox_f(x) \Leftrightarrow x - u \in \partial f(u) \tag{60}$$

Preuve:

$$\begin{aligned} u = prox_f(x) &\Leftrightarrow u = argmin\{f(u) + \frac{1}{2}||u - x||_2^2\} \\ &\Leftrightarrow 0 \in \partial g(u) \\ &\Leftrightarrow 0 \in \partial g_1(u) + \partial g_2(u) \\ &\Leftrightarrow 0 \in \partial f(u) + (u - x) \Leftrightarrow x - u \in \partial f(u) \end{aligned} \tag{61}$$

$$g(y) = g_1(y) + g_2(y) = f(y) + \frac{1}{2}||y-x||_2^2 \tag{62}$$

Algorithme du gradient proximal

$$0 \in \partial f(x^*) \Leftrightarrow x^* = argmin_x f(x) \tag{63}$$

$$\partial(f_1 + f_2) = \partial f_1 + \partial f_2 \tag{64}$$

Si f est différentielle en x, alors

$$\partial f(x) = \nabla f(x) \tag{65}$$

Norme euclidienne

$$f(x) = ||x||_2 \tag{66}$$

$$prox_{tf}(x) = \begin{cases} (1 - \frac{t}{||x||_2})x & , ||x||_2 >= t \\ 0 & , sinon \end{cases} \tag{67}$$

Multiplication par un scalaire ¿0

$$f(x) = \lambda g(x/\lambda) \tag{68}$$

$$prox_f(x) = \lambda prox_{\frac{1}{\lambda}g}(\frac{x}{\lambda}) \tag{69}$$

Somme séparable (Group LASSO)

$$f([x, y] = g(x) + h(y) \tag{70}$$

$$prox_f([x, y]) = [prox_g(x), prox_h(y)] \tag{71}$$

Norme l_1

$$f(x) = ||x||_1 \tag{72}$$

$$[prox_f(x)]_i \begin{cases} x_i - 1 & \text{si} x_i >= 1 \\ 0 & \text{si} |x_i| < 1 \\ x_i + 1 & \text{si} x_i <= -1 \end{cases} \tag{73}$$

Numériquement

$$proxl_1(x) = sign(x) \times pmax(abs(x) - 1, x) \tag{74}$$

$$min_\beta f(\beta) = min_\beta \{g(\beta) + h(\beta)\} \tag{75}$$

Algorithme du gradient proximal g convexe et differentiable, ∇g est L-Lipschitzienne

h convexe et non-differentiable (sci pour avoir $prox_{l_2}(x)$)

Exercise

$$f(\beta) = ||X\beta - y||_2^2 + \lambda||\beta||_1 \tag{76}$$

Algorithme:

$$x_k \leftarrow prox_{t_k h}(x_{k-1-t_k \nabla g(x_{k-1})}) \tag{77}$$

$$f^* = f(x^*) \text{ fini} \tag{78}$$

$$t_k = \frac{1}{L}, (0 <= t_k < \frac{1}{L}) \tag{79}$$

Gradient Map

$$G_t(x) = \frac{1}{t}(x - prox_{tl_2}(x - t\nabla g(x))) \tag{80}$$

Pourquoi?

$$x^+ = x - tG_t(x) \tag{81}$$

Attention:

- $G_t(x)$ n'est pas un gradient pour g, n'est pas un sous-gradient pour h ou pour f
- $G_t(x^*) = 0$ ssi $x^* = argminf$

Borne Quadratique Supérieure (BQS)

$$g(y) <= g(x) + \nabla g(x)^T (y - x) + \frac{L}{2}||y - x||_2^2 \tag{82}$$

Pour

$$y(= x^+) = x - tG_t(x) \tag{83}$$

$$\begin{aligned} g(x - tG_t(x)) <=& g(x) - t\nabla g(x)^T G_t(x) + \frac{L}{2}t^2||G_t(x)||_2^2 \\ <= g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2}||G_t(x)||_2^2 \end{aligned} \tag{84}$$

Théorème: L'inégalité précédente nous permet de montrer

$$f(x - tG_t(x)) <= f(z) + G_t(x)^T (x - z) - \frac{t}{2}||G_t(x)||_2^2 \tag{85}$$

$$\begin{aligned} f(x - tG_t(x)) <=& g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2}||G_t(x)||_2^2 + h(x - tG_t(x)) \\ <=& g(z) + \nabla g(z)^T (x - z) - t\nabla g(x)^T G_t(x) + \frac{t}{2}||G_t(x)||_2^2 + h(z) + v^T (x - z - tG_t(x)) \\ =& f(z) + G_t(x)^T (x - z) - \frac{t}{2}||G_t(x)||_2^2 \end{aligned} \tag{86}$$

Pour

$$z = x \tag{87}$$

on a

$$f(x^+) <= f(x) - \frac{t}{2}||G_t(x)||_2^2 \tag{88}$$

$$f(x^+) \to f(x_k) \tag{89}$$

Donc, on a une méthode de descente !

Pour $z = x^*$

$$\begin{aligned} f(x^*) - f^* <= & G_t(x)^T(x - x^*) - \frac{t}{2}||G_t(x)||_2^2 \\ = & \frac{1}{2t}(||x - x^*||_2^2 - ||x - x^* - tG_t(x)||_2^2) \\ = & \frac{1}{2t}(||x - x^*||_2^2 - ||x^+ - x^*||_2^2) \end{aligned} \tag{90}$$

$$f(x_N) - f^* <= \frac{1}{2Nt}||x_0 - x^*||_2^2 \tag{91}$$

$$[prox_{t||||_1}](x) = \begin{cases} x_i - t & \text{si} x_i >= t \\ 0 & \text{si} |x_i| < t \\ x_i + t & \text{si} x_i <= t \end{cases} \tag{92}$$

Fast Proximal gradient algorithm

Convexe & differentielle

$$f(y) >= f(x) + \nabla f(x)^T(y - x) \tag{93}$$

Sous-gradient — sous differentielle

$$\partial f(x) = \{g | g^T(x - y) <= f(y) - f(x)\} \tag{94}$$

Prox.

$$prox_f(x) = argmin_\mu\{f(\mu) + \frac{1}{2}||x - \mu||_2^2\} \tag{95}$$

$$x - u \in \partial f(u) \Leftrightarrow u = prox_f(x) \tag{96}$$

$$\min f(\beta) = g(\beta) + h(\beta) \tag{97}$$

∇g L-Lipschitzienne $prox_{th}$ convexe

FISTA: (n'est pas une méthode de descente)

$$y = x_{k-1} + \frac{k-2}{k+1}(x_{k-1} - x_{k-2}) \tag{98}$$

$$x_k = prox_{t_k h}(y - t_k \nabla g(y)) \tag{99}$$

$$t_k = \frac{1}{L}\text{constant} \tag{100}$$

Reformulation

$$\theta_k = \frac{2}{k+1} \tag{101}$$

v_k tel que $v_0 = x_0$ et $\forall k >= 1$

$$\begin{cases} y = (1-\theta_k)x_{k-1} + \theta_k v_{k-1} \\ x_k = prox_{th}(y - t_k \nabla g(y)) \\ v_k = x_{k-1} + \frac{1}{\theta_k}(x_k - x_{k-1}) \end{cases} \tag{102}$$

Inégalité

$$\forall k >= 2, \frac{1-\theta_k}{\theta_k} <= \frac{1}{\theta_{k-1}^2} \tag{103}$$

BQS(g)

$$g(u) <= g(z) + \nabla g^T(z)(u-z) + \frac{L}{2}||u-z||_2^2 \tag{104}$$

$BQS(h)$

$$u = prox_{th}(w) \tag{105}$$

alors

$$\forall z, h(u) <= h(z) + \frac{1}{t}(v-u)^T(u-z) \tag{106}$$

1.

$$g(x^+) <= g(y) + \nabla g^T(y)(x^+ - y) + \frac{1}{2t}||x^+ - y||_2^2 \tag{107}$$

2.

$$\begin{aligned} h(x^+) &<= h(z) + \frac{1}{t}(y - t\nabla g(y)x^+)^T(x^+ - z) \\ &= h(z) + \nabla g(y)^T(z - x^+) + \frac{1}{t}(x^+ - y)^T(z - x^+) \end{aligned} \tag{108}$$

1+2:

$$\begin{aligned} f(x^+) &= g(x^+) + h(x^+) \\ &<= g(y) + h(z) + \nabla g(y)^T(x^+ - y + z - x^+) + \frac{1}{2t}||x^+ - y||_2^2 + \frac{1}{t}(x^+ - y)^T(z - x^+) \\ &<= f(z) + \frac{1}{2t}||x^+ - y||_2^2 + \frac{1}{t}(x^+ - y)^T(z - x^+) \end{aligned} \tag{109}$$

$$\begin{aligned} & f(x^+) - f^* - (1-\theta)(f(x) - f^*) \\ & <= \frac{\theta^2}{2t}(||v - x^*||_2^2) - ||v^+ - x^*||_2^2 \\ \Leftrightarrow \frac{t}{\theta^2}(f(x) - f^* + \frac{1}{2}||v_1 - x^*||_2^2 & <= \frac{1-\theta_1^2}{\theta_1^2}(f(z) - f^*) + \frac{1}{2}||v - x^*||_2^2 \end{aligned} \tag{110}$$

Comme

$$\frac{1-\theta_1}{\theta_1^2} <= \frac{1}{\theta_{i-1}^2} \tag{111}$$

Conclusion

$$\frac{t}{\theta_k^2}(f(x_k) - f^*) - \frac{1}{2}||v_1 - x^*||_2^2 <= \frac{(1-\theta_1)^t}{\theta_1^2}(f(x_0) - f^*) + \frac{1}{2}||v_0 - x^+||_2^2 \quad (112)$$

Ainsi

$$\frac{t}{\theta_k^2}f(x_k) - f^* <= \frac{(1-\theta_1)^t}{\theta_1^2}(f(x_0 - f^*)) + \frac{1}{2}||v_k - x^*||_2^2 - \frac{1}{2}||v_0 - x^*||_2^2 \quad (113)$$

$$f(x_k) - f^* <= \frac{2L}{(k+1)^2}||x_0 - x^*||_2^2 \quad (114)$$