Cours Apprentissage - ENS Math/Info
Optimisation Convexe

Francis Bach

16 Octobre 2015

Ce cours s'appuie sur le livre "Convex Optimization" de Stephen Boyd et Lieven Vandenberghe (disponible gratuitement : http://www.stanford.edu/~boyd/cvxbook/) et les livres "Non-linear programming" de Dimitri Bertsekas (Athena Scientific), et "Introductory lectures on convex optimization : A basic course" de Yurii Nesterov (Kluwer Academic Publishers).

Dans ce cours, on se limitera principalement aux problèmes d'optimisation non-contraints, i.e., minimiser $f(x)$ pour $x \in \mathbb{R}^d$. Les minimas locaux sont tels que $f'(x) = 0$, et sont globaux lorsque f est convexe.

Afin de trouver les minimiseurs de x, deux grandes stratégies sont possibles : l'utilisation de boites à outils génériques ou des algorithmes itératifs simples.

1 Boîte à outil générique

- Utile pour la programmation linéaire, i.e., $\min_{Ax=b, x \geqslant 0} c^\top x = \max_{A^\top y \leqslant c} b^\top y$ et pour ses extensions.
- Optimisation à moyenne échelle (pas plus de dizaines de milliers de contraintes et/ou variables) avec une forte précision.
- Voir http://cvxr.com/cvx/

2 Méthode de l'ellipsoide

- Minimisation d'une fonction convexe dérivable sur $\mathbb{R}^n$ (extensions possibles aux cas non-différentiables et contraints).
- Extension de la méthode de bisection utilisée pour les fonctions à une seule variable.
- Complexité théorique polynômiale (résultat très général), peu utilisé en pratique (à cause des inversions de matrices et la convergence lente)
- **Algorithme** :
  - Initialisation : ellipsoide $\mathcal{E}_0 = \{(x - x_0)^\top P_0^{-1}(x - x_0) \leqslant 1\}$ qui contient x_* (avec typiquement $P_0 = R^2 I$)
  - Pour $t \geqslant 0$, contruire l'ellipsoide de volume minimum qui contient $\{(x - x_t)^\top P_t^{-1}(x - x_t) \leqslant 1\}$ et le demi-plan $\{f'(x_t)^\top (x - x_t) \leqslant 0\}$.

  - Ceci correspond à $x_{t+1} = x_t - \frac{1}{n+1} P_t g_t$ et $P_{t+1} = \frac{n^2}{n^2-1}(P_t - \frac{2}{n+1} P_t g_t g_t^\top P_t)$ avec $g_t = \frac{1}{\sqrt{f'(x_t)^\top P_t f'(x_t)}} f'(x_t)$.
    Voir `http://www-math.mit.edu/~goemans/18433S09/ellipsoid.pdf` et
    `http://sma.epfl.ch/~eisenbra/OptInFinance/Slides/ellipsoid.pdf`

- **Propriété** : le volume de P_{t+1} est plus petit que $e^{-1/2n}$ fois le volume de P_t.

- **Conséquence** : si f est B-Lipschitz, alors la précision ε est atteinte après au plus $2n^2 \log(RG/\varepsilon)$.
  Voir `http://www.stanford.edu/class/ee392o/elp.pdf`.
  Preuve (hypothèse, f G-Lipschitz et $B(x_*, \varepsilon/G) \subset \mathcal{E}_0$) : on cherche x tel que $f(x) \leqslant f(x_*) + \varepsilon = \inf_{x \in \mathbb{R}^d} f(x) + \varepsilon = f^* + \varepsilon$. Si on suppose par l'absurde que $\forall k \leqslant t$, $f(x_k) > f^* + \varepsilon$, alors la boule $B(x_*, \varepsilon/G)$ est incluse dans $\mathcal{E}_t$, ce qui donne, en comparant les volumes $(\varepsilon/G)^n \leqslant R^n e^{-t/2n}$ et donc $t \leqslant 2n^2 \log \frac{RG}{\varepsilon}$.

3 Descente de gradient

- Minimisation d'une fonction convexe dérivable sur $\mathbb{R}^n$

- **Algorithme** :
  - Initialisation : $x_0 \in \mathbb{R}^n$,
  - Pour $t \geqslant 0$, $x_{t+1} = x_t - \gamma_t f'(x_t)$.

- **Valeurs de pas** γ_t :
  - Pas constant : $\gamma_t = 1/L$ où L est une borne supérieure uniforme de la plus grande valeur propre de la Hessienne de f
  - "Line search" : optimisation totale ou partielle par rapport à γ. Voir par exemple `http://www-personal.umich.edu/~mepelman/teaching/IOE511/Handouts/511notes07-5.pdf` pour la règle dite d'Armijo.

- **Analyse théorique**
  - Hypothèses pour résultat théorique simple : f deux fois dérivable *convexe*, $LI \succcurlyeq f''(x) \succcurlyeq \mu I$ (forte convexité).
  - $x_{t+1} - x_* = x_t - x_* - \gamma_t f''(y_t)(x_t - x_*)$ pour un $y_t \in [x_t, x_*]$
  - Pour $\gamma_t = 1/L$, on peut montrer que $\|x_{t+1} - x_*\|^2 \leqslant (1 - \frac{\mu}{L})\|x_t - x_*\|^2 \leqslant (1 - \frac{\mu}{L})^{(t+1)} \|x_0 - x_*\|^2$.
    On se limitera à la preuve dans la cas quadratique (par simplicité).
  - La convergence est alors dite linéaire.
  - Si on ne suppose que la convexité, alors on obtient le résultat $f(x_t) - f(x_*) \leqslant \frac{L}{2t} \|x_0 - x_*\|^2$,

- Cadre convexe : convergence vers un minimum global

- Cadre non convexe : convergence vers un point stationnaire

- Critère d'arrêt : gradient de norme inférieure à ε

4 Méthode de Newton

- Minimisation d'une fonction convexe deux fois dérivable sur $\mathbb{R}^n$

- **Principe** : minimiser l'approximation quadratique autour de x_t

- **Algorithme** :
  - Initialisation : $x_0 \in \mathbb{R}^n$,
  - Pour $t \geqslant 0$, $x_{t+1} = x_t - f''(x_t)^{-1} f'(x_t)$.

– **Convergence quadratrique** (cas convexe) : $\|x_{t+1} - x_*\|^2 \leqslant C(\|x_t - x_*\|^2)^2$