TP4 : OPTIMISATION CONVEXE

COURS D'APPRENTISSAGE, ECOLE NORMALE SUPÉRIEURE

Raphaël Berthier
raphael.berthier@inria.fr

1. Choix des pas dans la descente de gradient pour la minimisation de quadratiques

1) De $\nabla f(x^*) = 0$ il découle que $Hx^* = b$. Par conséquent,

$$x_{t+1} - x^* = x_t - \gamma(Hx_t - b) - x^* = x_t - x^* - \gamma H(x_t - x^*) = (I_d - \gamma H)(x_t - x^*)\,.$$

2) La matrice H étant la hessienne de f, on cherche donc μ et L tels que $\mu I_d \preccurlyeq H \preccurlyeq L I_d$. $L = \lambda_1$ et $\mu = \lambda_n$ sont les constantes optimales.

3)

$$\begin{aligned}\langle x_{t+1} - x^*, u_i\rangle &= \langle (I_d - \gamma H)(x_t - x^*), u_i\rangle \\ &= \langle x_t - x^*, (I_d - \gamma H)u_i\rangle \qquad \text{car } H \text{ est symétrique} \\ &= \langle x_t - x^*, (1 - \gamma\lambda_i)u_i\rangle \\ &= (1 - \gamma\lambda_i)\langle x_t - x^*, u_i\rangle\,,\end{aligned}$$

et le résultat suit par récurrence.

4) On a donc $\|x_t - x^*\|_2^2 = \sum_{i=1}^d (1-\gamma\lambda_i)^{2t}\langle x_0 - x^*, u_i\rangle^2$. Pour minimiser la vitesse de convergence, il faut minimiser en γ la quantité

$$\max_{i=1,\dots,d} |1 - \gamma\lambda_i| = \max(1 - \gamma\mu, \gamma L - 1)\,.$$

Ceci est minimal en $\gamma = 2/(L+\mu)$. La vitesse de convergence associée est

$$\|x_t - x^*\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^t \|x_0 - x^*\|_2\,.$$

5) Si on note $a_t = \langle x_t - x^*, u_i\rangle$, alors $a_{t+1} = (1-\gamma\lambda_i)a_t + \beta(a_t - a_{t-1})$. On applique ensuite la méthode classique de résolution des équations linéaires d'ordre 2. Après quelques calculs, le discriminant du polynôme caractéristique vaut $\delta = 16\lambda_i(\lambda_i - L - \mu)/(\sqrt{L} + \sqrt{\mu})^4$, qui est négatif. Le polynôme caractéristique a deux racines complexes conjuguées z_+, z_-. La vitesse de décroissance de a_t est donnée par $a_t = O(m^t)$, où $m = |z_+| = |z_-|$. Le calcul donne $m = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$.

2. Descente de gradient pour la régression ridge

6) On vérifie $H = \frac{1}{n}X^TX + \lambda I_d$, $b = \frac{1}{n}X^Ty$. Puisque $p > n$, on a que X^TX ne peut pas être de rang plein. Par conséquence, $\mu = \lambda$. Par ailleurs, $L = \lambda + \lambda_{\max}(X^TX)/n$.

7)

$$w^* = \left(\frac{1}{n}X^TX + \lambda I\right)^{-1}\frac{1}{n}X^Ty\,.$$

8) **9**) **10**) Voir Jupyter notebook.

11) Lorsque $\lambda \to 0$, le problème devient de moins en moins bien conditionné, de sorte que (a) la vitesse asymptotique de la méthode heavy ball devient infiniment meileure que la descente de gradient simple, et (b) l'optimisation, pour les deux méthodes, devient de plus en plus difficile.