Fondamentaux Mathématiques pour les Data Sciences

M2 Data Science

Alexandre Aussem

LIRIS UMR 5205 CNRS
Data Mining & Machine Learning Group (DM2L)
University of Lyon 1
Web: perso.univ-lyon1.fr/alexandre.aussem

October 24, 2017

Université Claude Bernard Lyon 1

Outline

La descente de gradient : intuition

- Enjeu : minimiser f (dans $\mathbb{R}^d$) en trouvant un nouveau point pour lequel f diminue le plus.
- Approximation du premier ordre :

$$f(x) \approx f(x^0) + \langle \nabla f(x^0), x - x^0 \rangle$$

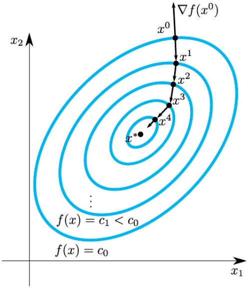
- Solution : il faut "s'aligner" avec la direction opposée au gradient $x - x_0 = -\alpha \nabla f(x^0)$
  $\alpha > 0$ contrôle la "vitesse" avec laquelle on progresse dans la direction. Ce paramètre est appelé le **pas** de la méthode.

La descente de gradient : algorithme

Data: initialisation x^0, nb max. d'itérations T, critère d'arrêt ε, pas α
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T$ **do**
  $x^{t+1} \leftarrow x^t - \alpha \nabla f(x^t)$
  STOP si critère d'arrêt inférieur à ε
end

Critères d'arrêts possibles :

- $\|\nabla f(x^t)\| \leq \varepsilon$
- $f(x^{t+1}) - f(x^t) \leq \varepsilon$
- $\|x^{t+1} - x^t\| \leq \varepsilon$ ou $\frac{\|x^{t+1}-x^t\|}{\|x^t\|} \leq \varepsilon$



Choix du pas : Recherche linéaire

Parfois, il faut choisir le pas à chaque itération : α^t évolue avec les itérations. On note $d^t = -\nabla f(x^t)$ une direction de descente

Règle de la minimisation

Minimisation sur l'amplitude : il faut résoudre le problème 1D :

$$f(x^t + \alpha^t d^t) = \min_{\alpha \geq 0} f(x^t + \alpha d^t)$$

Rem: Pour cela il faut que le problème 1D soit simple à résoudre

Méthode de Newton

Objectif : la méthode de Newton (ou Newton-Raphson) sert à trouver les zéros d'une fonction, *i.e.*, résoudre $f(x) = 0$
L'idée : approximation locale par une fonction affine

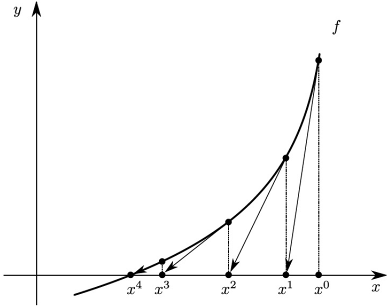
$$f(x) \approx f(x^0) + f'(x^0)(x - x^0)$$

La règle de mise à jour est donc :

$$\boxed{x^{t+1} \leftarrow x^t - \frac{f(x^t)}{f'(x^t)}}$$

Méthode de Newton

Data: point initial x^0, nombre max. d'itérations T, critère d'arrêt ε
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T-1$ **do**
  $x^{t+1} \leftarrow x^t - \frac{f(x^t)}{f'(x^t)}$
  STOP si critère d'arrêt inférieur à ε
end



Méthode de Newton : cas multidimensionnel

Localement, en un point x^0 une fonction deux fois différentiable ressemble à :

$$f(x) \approx f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \frac{1}{2}(x - x^*)^\top \nabla^2 f(x^*)(x - x^*)$$

- Enjeu : minimiser en x l'approximation (quadratique) précédente
- Solution : CNO

$$\nabla f(x^*) + \nabla^2 f(x^*)(x - x^*) = 0$$

- Nouvelle règle de mise à jour :

$$\boxed{x^{t+1} \leftarrow x^t - (\nabla^2 f(x^t))^{-1} \nabla f(x^t)}$$

Rem: C'est donc la méthode de Newton appliquée à la recherche de zéros d'une approximation du gradient de f

Exemples de problèmes avec contraintes

En pratique : on optimise souvent avec contraintes (physiques)

- Contrainte de **positivité** :
  $K = \{x \in \mathbb{R}^d : \forall i \in [\![1, d]\!], x_i \geq 0\}$
- Contrainte de type **simplexe** (pour des probabilités) :
  $K = \Delta_d = \{x \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1 \text{ et } \forall i \in [\![1, d]\!], x_i \geq 0\}$
- **Moindres carrés contraints** : on cherche x tel que $Ax = b$ avec une contrainte linéaire sur x, *e.g.*, $Bx = 0$ pour une matrice $B \in \mathbb{R}^{m \times d}$

On cherche alors à résoudre

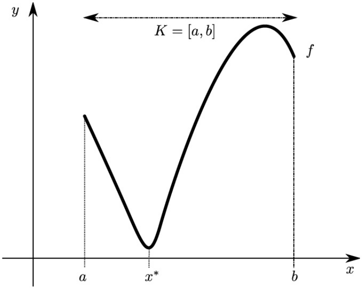
$$\boxed{x^* \in \arg\min_{x \in K} f(x)}$$

où $K \subset \mathbb{R}^d$ est un ensemble qui encode les contraintes

Condition d'existence d'un minimum

Théorème de Weierstrass

Si une fonction $f : \mathbb{R}^d \mapsto \mathbb{R}$ est continue sur un ensemble fermé et borné K (*i.e.*, un ensemble **compact**) alors il existe un point x^* qui atteint le minimum : $x^* \in \underset{x \in K}{\arg\min} f(x)$

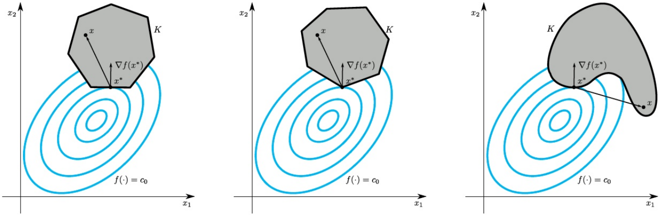


Condition du premier ordre

Théorème : CNO cas contraint

Si f a un minimum local en x^* sur un convexe K, alors

$$\forall x \in K, \langle \nabla f(x^*), x - x* \rangle \geq 0$$



Projection sur les convexes fermés

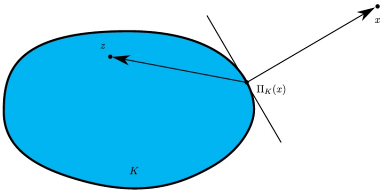
Théorème de projection

Si $K \subset \mathbb{R}^d$ est un convexe fermé non-vide, alors pour tout point $x \in \mathbb{R}^d$ il y a un unique point noté $\Pi_K(x)$ qui satisfait :

$$\Pi_K(x) = \arg\min_{z \in K} \frac{1}{2}\|x - z\|^2$$

De plus un point x^* est solution de ce problème ssi

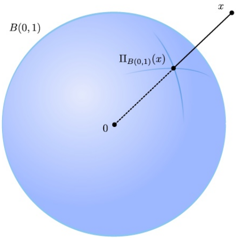
$$\forall z \in K, \langle z - x^*, x - x^* \rangle \leq 0$$



Illustration

- Le projecteur sur $B(0,1)$ (la boule centrée en 0 et de rayon unité) est la fonction

$$\Pi_{B(0,1)}(x) = \begin{cases} x & \text{si } \|x\| \leq 1 \\ \frac{x}{\|x\|} & \text{si } \|x\| > 1 \end{cases}$$



Contraintes et Lagrangien

En pratique : forme explicite pour les contraintes, avec m contraintes d'égalité, et r contraintes d'inégalité

$$(\mathcal{P}) \qquad \begin{array}{ll} \min & f(x) \\ \text{s. c.} & h_1(x) = 0, \ldots, h_m(x) = 0, \\ & g_1(x) \leq 0, \ldots, g_r(x) \leq 0, \end{array}$$

Définition : Lagrangien

On appelle **Lagrangien** du problème $(\mathcal{P})$ la fonction

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \mu_j g_j(x)$$

Conditions de Karush-Khunn-Tucker

Théorème : KKT

Si x^* est un minimum local du problème $(\mathcal{P})$, que f, h_i, g_j sont dérivables avec des gradients continus, sous des conditions de qualification sur x^*, il existe $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*)$ et $\mu^* = (\mu_1^*, \ldots, \mu_r^*)$ tel que :

$$
\begin{aligned}
\forall j \in [\![1, r]\!], \quad \mu_j^* &\geq 0, \\
\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) &= 0, && \text{(CNO)} \\
h_1(x^*) = 0, \ldots, h_m(x^*) &= 0, && \text{(satisfiabilité)} \\
g_1(x^*) \leq 0, \ldots, g_r(x^*) &\leq 0, && \text{(satisfiabilité)} \\
\forall j \in [\![1, r]\!], \quad \mu_j^* g_j(x^*) &= 0. && \text{(complémentarité)}
\end{aligned}
$$

Exemple de résolution

Objectif quadratique et contrainte affine

$$(\mathcal{P}) \qquad \min_{x_1,x_2} \quad \frac{1}{2}(x_1^2 + x_2^2) \qquad \text{s. c.} \quad x_1 + x_2 \leq -2,$$

$$\mathcal{L}(x,\mu) = \frac{1}{2}(x_1^2 + x_2^2) + \mu(x_1 + x_2 + 2)$$

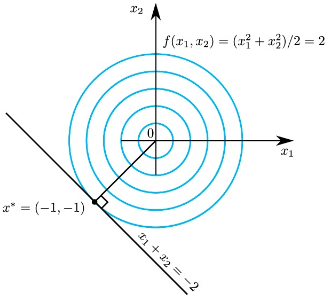
La CNO donne $x_1^* + \mu^* = x_2^* + \mu^* = 0$. Par complémentarité, on peut traiter deux cas exclusifs

1. $x_1^* + x_2^* < -2$ et $\mu^* = 0$ (absurde !)
2. $x_1^* + x_2^* = -2$ et $\mu^* = 1$, puis $x_1^* = x_2^* = -1$

Vérification visuelle

Objectif quadratique et contrainte affine

$$(\mathcal{P}) \qquad \begin{array}{ll} \displaystyle\min_{x_1,x_2} & \dfrac{1}{2}(x_1^2 + x_2^2) \\ \text{s. c.} & x_1 + x_2 \leq -2, \end{array}$$



References I

- Christopher M. Bishop.
  Pattern Recognition and Machine Learning.
  Springer, 2006.
- Anne Sabourin et Joseph Salmon.
  Fondamentaux pour le Big Data, Télécom ParisTech.