Algorithme de la descente de gradient stochastic (SGD)

Intro

- pb avec la descente de gradient ?
- chaque itération requiert de passer par toutes les données.
- matrice $A \in \mathbb{R}^{n \times d}$ (n exemples d dimensions)

Complexité : $k \times n \times d$ (où k = nb d'itérations)

$\Rightarrow$ on va s'intéresser à des algos moins gourmants en données par itération.

Idée : les pb d'optim qui nous intéressent (sur de données), sont souvent de la forme

$$\min_x f(x) \quad \text{où} \quad f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x)$$

ex : régression linéaire.

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x) \quad \text{où} \quad f_i(x) = (a_i^T x - b_i)^2$$

ex : - régression logistique
- réseaux de neurones.

Algo SGD (stochastic gradient Descent)

$x \leftarrow (0, \dots, 0)^T$

$k \leftarrow 0$

répéter

- tirer aléatoirement et uniformément i de l'ensemble $\{1..N\}$
- $x \leftarrow x - \alpha_k \nabla f_i(x)$
- $k \leftarrow k+1$

tant que condition d'arrêt non satisfaite.

- complexité : $k \times d$
- un cas pathologique
  - $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = \frac{1}{2} f_1(x) + \frac{1}{2} f_2(x)$$

$$f_1(x) = 2(x-1)^2 \qquad f_2(x) = -(x-1)^2$$

$$f(x) = \frac{1}{2}(x-1)^2$$

$$\nabla f_1 = 4(x-1) \qquad \nabla f_2 = -2(x-1)$$

  - au début, $x = 0$

$$\nabla f_1(x) = -4 \qquad \nabla f_2(x) = 2$$

  - le SGD produit ici des oscillations sans fin, si le pas α est constant.
  - On aura besoin avec le SGD d'un pas décroissant.

Pourquoi à SGD fonctionne ?

- Rappel sur l'espérance.

  Si une variable aléatoire z prend des valeurs $z_1 .. z_N$ avec probabilité $p_1 .. p_N$, alors on définit l'espérance de z

  $$\mathbb{E}[z] = \sum_{i=1}^{N} p_i z_i$$

  Si z et les z_i sont des vecteurs, la définition est la même.

- A l'étape k, quel est le vecteur ∇f_i espéré ?

$$\mathbb{E}[\nabla f_i] = \sum_{i=1}^{N} \nabla f_i \times p_i$$

$$= \frac{1}{N} \sum_{i=1}^{N} \nabla f_i$$

$$= \nabla f$$