Descente stochastique du gradient SGD

Introduction à l'apprentissage automatique
Master Sciences Cognitives
Aix Marseille Université

Alexis Nasr

Plan

Recherche des valeurs extremales d'une fonction d'une variable

Recherche des valeurs extremales d'une fonction de plusieurs variables

Descente stochastique du gradient

Recherche d'extremums d'une fonction

- La recherche des extremums de la fonction $f(x)$ suppose de résoudre l'équation :

$$f'(x) = 0$$

- Dans certains cas, on peut trouver une solution **exacte** de l'équation.
- Mais la plupart du temps (dans notre cas) on a recours à une méthode itérative qui donnera une solution **approchée** de l'équation.

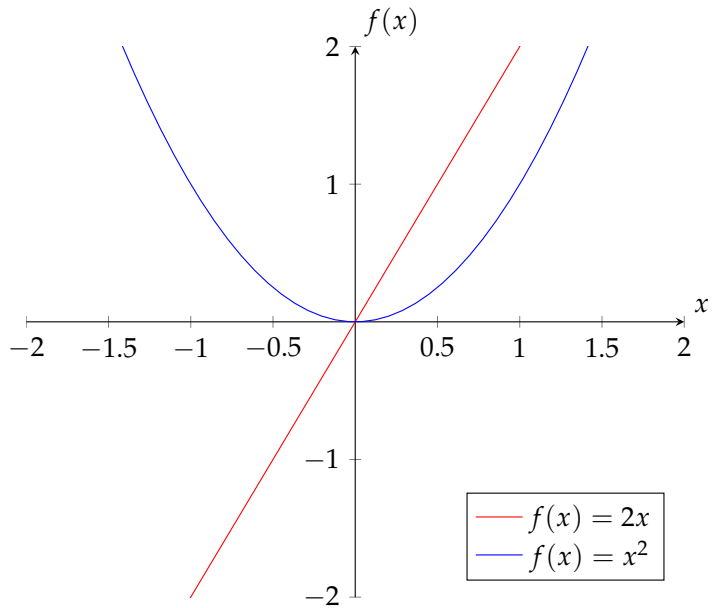
Recherche itérative de minimums d'une fonction

- On utilise le signe de la dérivée en un point a .
  - Si $\left.\frac{df}{dx}\right|_{x=a} > 0$, alors f est croissante en a.
    - Dans ce cas, si on **diminue** la valeur de x ($x = a - \varepsilon$), la valeur de $f(x)$ diminue.
  - Si, $\left.\frac{df}{dx}\right|_{x=a} < 0$, alors f est décroissant en a.
    - Dans ce cas, si on **augmente** la valeur de x ($x = a + \varepsilon$), la valeur de $f(x)$ diminue.
- Méthode **itérative** : on procède par étapes
- à chaque étape on modifie la valeur de x :

$$x \leftarrow x - \eta \frac{d}{dx} f(x)$$

- jusqu'à atteindre un critère d'arrêt.

Recherche du minimum de la fonction $f(x) = x^2$

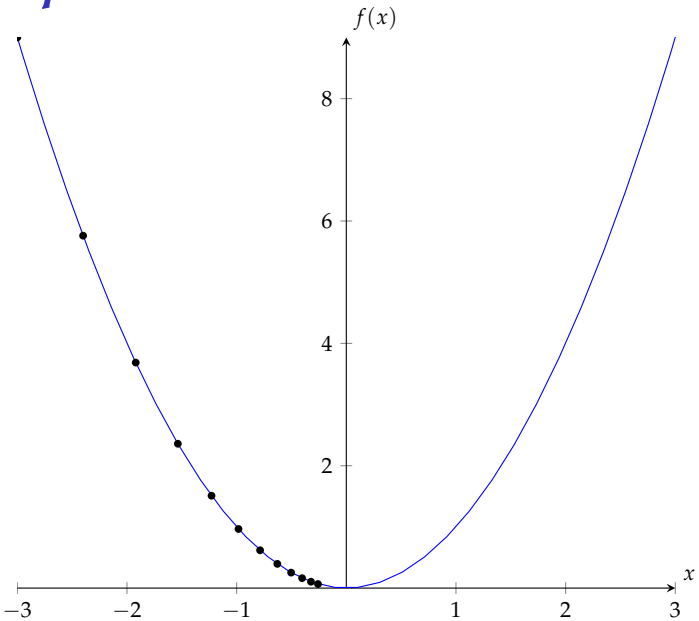


- Etape de mise à jour :

$$x \leftarrow x - 2\eta x$$

Exemple : $f(x) = x^2$, $\frac{df}{dx} = 2x$, $\eta = 0.1$

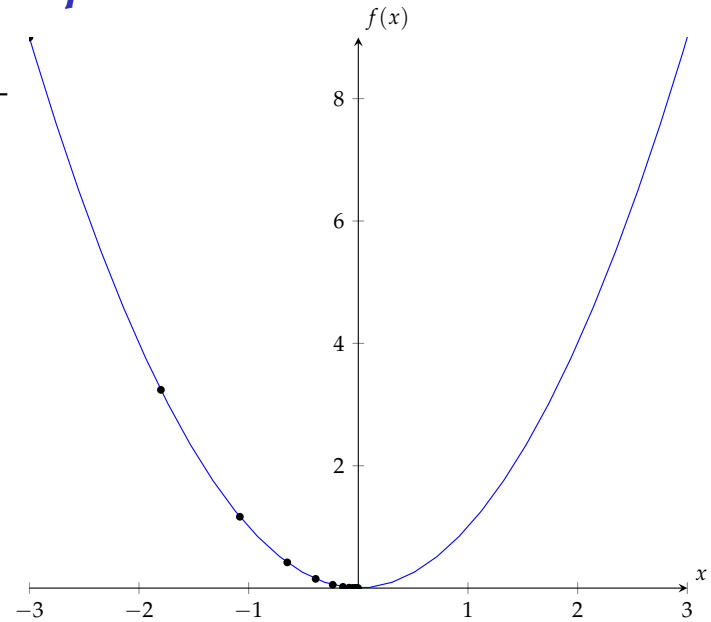
i	x	$2x$	x^2	$x - \eta \times 2x$
0	-3.000	-6.000	9.000	-2.400
1	-2.400	-4.800	5.760	-1.920
2	-1.920	-3.840	3.686	-1.536
3	-1.536	-3.072	2.359	-1.229
4	-1.229	-2.458	1.510	-0.983
5	-0.983	-1.966	0.966	-0.786
6	-0.786	-1.573	0.618	-0.629
7	-0.629	-1.258	0.396	-0.503
8	-0.503	-1.007	0.253	-0.403
9	-0.403	-0.805	0.162	-0.322
10	-0.322	-0.644	0.104	-0.258
11	-0.258	-0.515	0.066	-0.206
12	-0.206	-0.412	0.043	-0.165
13	-0.165	-0.330	0.027	-0.132
14	-0.132	-0.264	0.017	-0.106
15	-0.106	-0.211	0.011	-0.084
16	-0.084	-0.169	0.007	-0.068
17	-0.068	-0.135	0.005	-0.054
18	-0.054	-0.108	0.003	-0.043
19	-0.043	-0.086	0.002	-0.035
20	-0.035	-0.069	0.001	-0.028



plus vite !

Exemple : $f(x) = x^2$, $\frac{df}{dx} = 2x$, $\eta = 0.2$

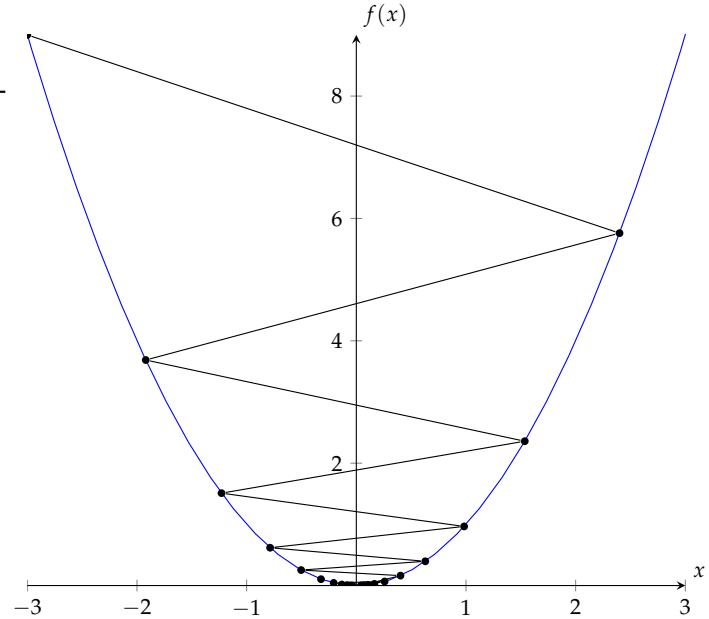
i	x	$2x$	x^2	$x - \eta \times 2x$
0	-3.000	-6.000	9.000	-1.800
1	-1.800	-3.600	3.240	-1.080
2	-1.080	-2.160	1.166	-0.648
3	-0.648	-1.296	0.420	-0.389
4	-0.389	-0.778	0.151	-0.233
5	-0.233	-0.467	0.054	-0.140
6	-0.140	-0.280	0.020	-0.084
7	-0.084	-0.168	0.007	-0.050
8	-0.050	-0.101	0.003	-0.030
9	-0.030	-0.060	0.001	-0.018
10	-0.018	-0.036	0.000	-0.011
11	-0.011	-0.022	0.000	-0.007
12	-0.007	-0.013	0.000	-0.004
13	-0.004	-0.008	0.000	-0.002
14	-0.002	-0.005	0.000	-0.001
15	-0.001	-0.003	0.000	-0.001
16	-0.001	-0.002	0.000	-0.001
17	-0.001	-0.001	0.000	-0.000
18	-0.000	-0.001	0.000	-0.000
19	-0.000	-0.000	0.000	-0.000
20	-0.000	-0.000	0.000	-0.000



encore plus vite !

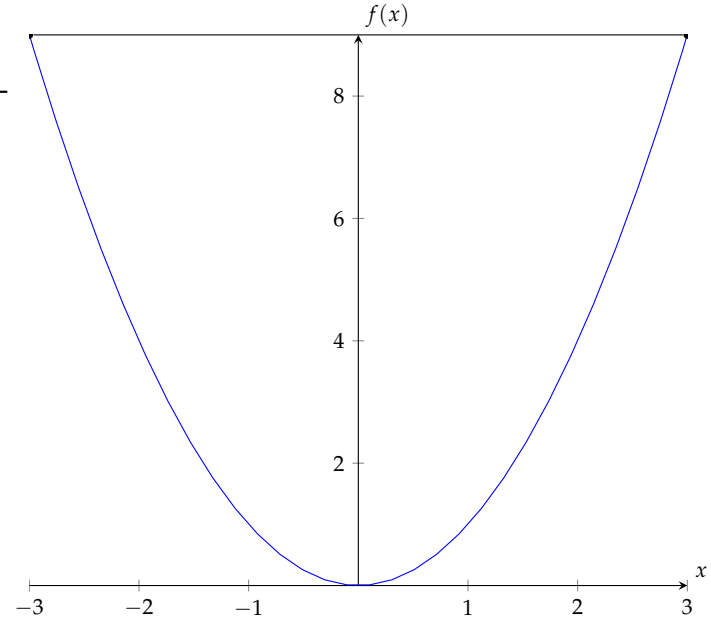
Exemple : $f(x) = x^2$, $\frac{df}{dx} = 2x$, $\eta = 0.9$

i	x	$2x$	x^2	$x - \eta \times 2x$
0	-3.000	-6.000	9.000	2.400
1	2.400	4.800	5.760	-1.920
2	-1.920	-3.840	3.686	1.536
3	1.536	3.072	2.359	-1.229
4	-1.229	-2.458	1.510	0.983
5	0.983	1.966	0.966	-0.786
6	-0.786	-1.573	0.618	0.629
7	0.629	1.258	0.396	-0.503
8	-0.503	-1.007	0.253	0.403
9	0.403	0.805	0.162	-0.322
10	-0.322	-0.644	0.104	0.258
11	0.258	0.515	0.066	-0.206
12	-0.206	-0.412	0.043	0.165
13	0.165	0.330	0.027	-0.132
14	-0.132	-0.264	0.017	0.106
15	0.106	0.211	0.011	-0.084
16	-0.084	-0.169	0.007	0.068
17	0.068	0.135	0.005	-0.054
18	-0.054	-0.108	0.003	0.043
19	0.043	0.086	0.002	-0.035
20	-0.035	-0.069	0.001	0.028



Exemple : $f(x) = x^2$, $\frac{df}{dx} = 2x$, $\eta = 1$

i	x	$2x$	x^2	$x - \eta \times 2x$
0	-3.000	-6.000	9.000	3.000
1	3.000	6.000	9.000	-3.000
2	-3.000	-6.000	9.000	3.000
3	3.000	6.000	9.000	-3.000
4	-3.000	-6.000	9.000	3.000
5	3.000	6.000	9.000	-3.000
6	-3.000	-6.000	9.000	3.000
7	3.000	6.000	9.000	-3.000
8	-3.000	-6.000	9.000	3.000
9	3.000	6.000	9.000	-3.000
10	-3.000	-6.000	9.000	3.000
11	3.000	6.000	9.000	-3.000
12	-3.000	-6.000	9.000	3.000
13	3.000	6.000	9.000	-3.000
14	-3.000	-6.000	9.000	3.000
15	3.000	6.000	9.000	-3.000
16	-3.000	-6.000	9.000	3.000
17	3.000	6.000	9.000	-3.000
18	-3.000	-6.000	9.000	3.000
19	3.000	6.000	9.000	-3.000
20	-3.000	-6.000	9.000	3.000



Points critiques et extremums d'une fonction de plusieurs variables

- La recherche des extremums de f suppose de résoudre l'équation :

$$\nabla f(\mathbf{x}) = 0$$

- Dans certains cas, on peut trouver une solution **exacte** de l'équation.
- Mais la plupart du temps (dans notre cas) on a recours à une méthode itérative qui donnera une solution **approchée** de l'équation.

Descente du gradient : idée générale

- Le gradient de la fonction f, calculé au point $\mathbf{x}$: $\nabla f(\mathbf{x})$, indique comment faire varier le vecteur $\mathbf{x}$ pour aboutir à l'augmentation maximale de f.
- Conséquences :

$$\mathbf{x}' = \mathbf{x} + \eta \nabla f(\mathbf{x}) \quad \Rightarrow f(\mathbf{x}') \geq f(\mathbf{x})$$

$$\mathbf{x}' = \mathbf{x} - \eta \nabla f(\mathbf{x}) \quad \Rightarrow f(\mathbf{x}') \leq f(\mathbf{x})$$

- η est appelé **pas d'apprentissage**, il permet de contrôler la variation de $\mathbf{x}$.

Exemple

- fonction

$$f(x,y) = (x+y) \times (y+1) = xy + x + y^2 + y$$

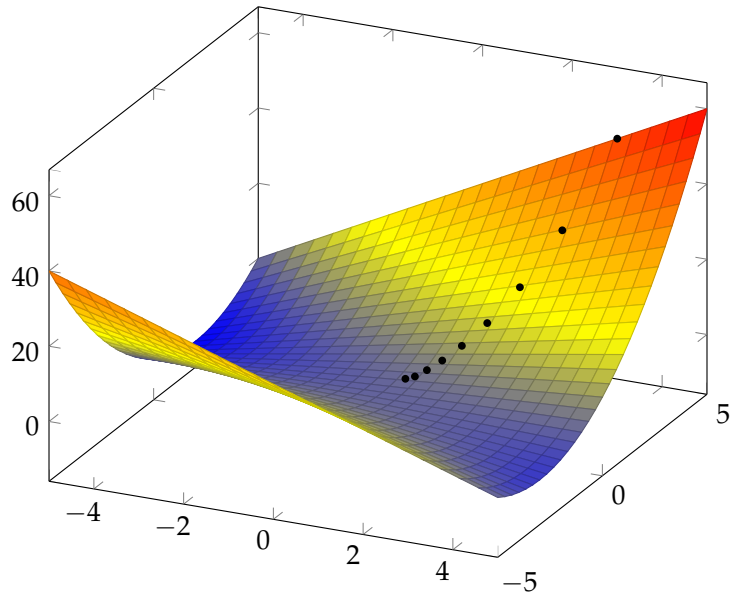
- dérivées partielles

$$\frac{\partial}{\partial x} f(x,y) = y + 1$$
$$\frac{\partial}{\partial y} f(x,y) = x + 2y + 1$$

- gradient : $\nabla f(x,y) = [y+1, x+2y+1]^T$
- gradient au point $(3,5)$:$\nabla f(3,5) = [6,14]^T$
- mise à jour ($\eta = 0.1$)

$$x = 3 - 0.1 \times 6$$
$$y = 5 - 0.1 \times 14$$

Exemple : $f(x,y) = (x+y) \times (y+1), \eta = 0.1$



i	x	y	$f(x,y)$
0	3	5	48
1	2.4	3.66	28.24
2	1.934	2.635	16.605
3	1.571	1.851	9.75
4	1.285	1.252	5.714
5	1.060	0.796	3.332
6	0.881	0.448	1.925
7	0.736	0.185	1.091
8	0.617	-0.014	1.091

Descente stochastique du gradient

- Etant donné :
  - des données d'apprentissage $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$
  - un modèle de prédiction $f(\mathbf{x}; \mathbf{w}) = \mathbf{y}$
  - une fonction d'erreur $E(\hat{\mathbf{y}}, \mathbf{y})$
- On souhaite déterminer le jeu de paramètres $\mathbf{w}^*$ qui minimise la fonction d'erreur sur les données $\mathcal{D}$.

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathcal{W}} E(\mathbf{w}; \mathcal{D})$$

- La fonction d'erreur sur $\mathcal{D}$ est calculée à partir de l'erreur sur chaque exemple $(\mathbf{x}_i, \mathbf{y}_i)$:

$$E(\mathbf{w}; \mathcal{D}) = \sum_{i=1}^{N} E(f(\mathbf{x}_i; \mathbf{w}), \mathbf{y}_i)$$

Descente stochastique du gradient

- Méthode itérative
- On met à jour $\mathbf{w}$ de manière à diminuer la fonction d'erreur :

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla_{\mathbf{w}} E(\mathbf{w}; \mathcal{D})$$

- Pour chaque mise à jour, le gradient $\nabla_{\mathbf{w}} E(\mathbf{w}; \mathcal{D})$ doit être calculé.
- Problème : il peut y avoir beaucoup de données d'apprentissage, le coût d'une mise à jour peut être prohibitif.
- Extrême inverse : on peut mettre à jour $\mathbf{w}$ après chaque exemple lu.
- Problème : l'erreur sur un exemple n'est qu'une estimation grossière de l'erreur sur l'ensemble $\mathcal{D}$.
- Le gradient calculé et la mise à jour qui en découle peut emmener l'algorithme vers une mauvaise solution.

Descente stochastique du gradient

- Compromis : on calcule l'erreur (et le gradient correspondant) sur un sous ensemble des données de taille m, appelé **minibatch** (mini lot).
- On calcule la moyenne du gradient sur les exemples du minibatch

$$\mathbf{g} = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\mathbf{w}} E(f(\mathbf{x}_i; \mathbf{w}), \mathbf{y}_i)$$

- Puis on fait la mise à jour :

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \mathbf{g}$$

Sources

- Ian Goodfellow, Yoshua Bengio, Aaron Courville, *Deep Learning*, MIT Press, 2016.