OCVX Case Study

Exam II - Juillet 2019

Daoud Karim, Misk Nicolas, Andersen Simon

Les problèmes étudiés

Descente de gradient

- Comparaison différents pas

- Comparaison norme L1-L2

Méthodes d’optimisation

- Learning rate scheduling

- Nesterov

- Adam

Méthode de Newton

- Descente de gradient avec Newton

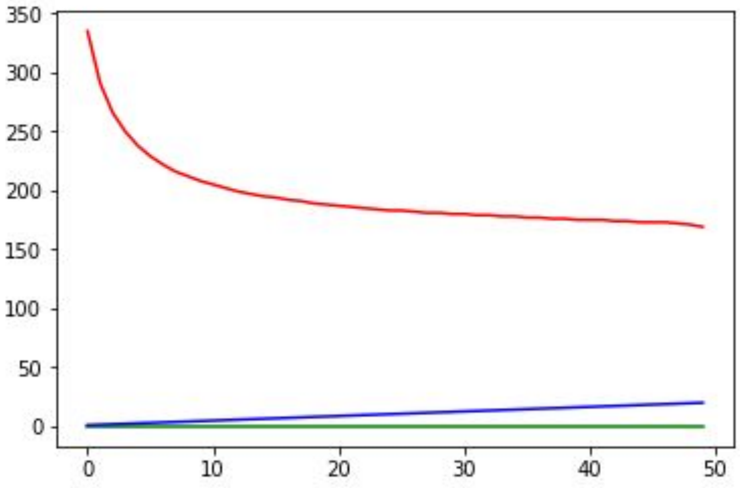
- Quasi-newton

Descente de gradient

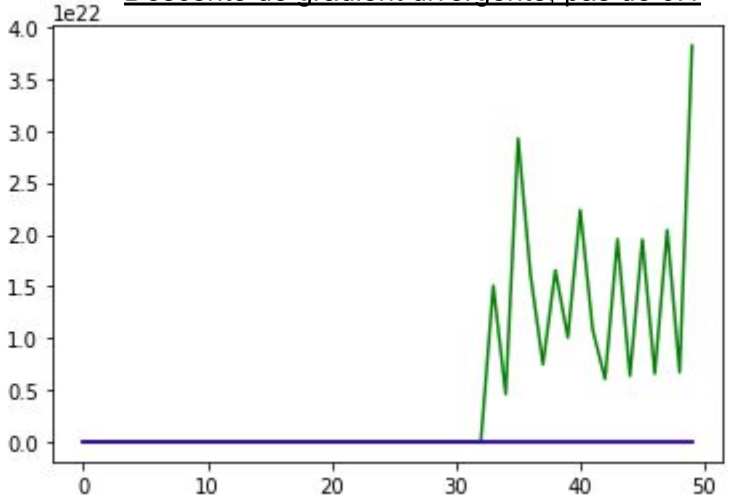
- Permet de minimiser une fonction
- Algorithme très utilisé dans le machine learning
- De nombreuses variantes et optimisations possible

<u>Légende:</u>
Nombre d’itérations
Valeur objective trouvée
Nombre de conditionnement

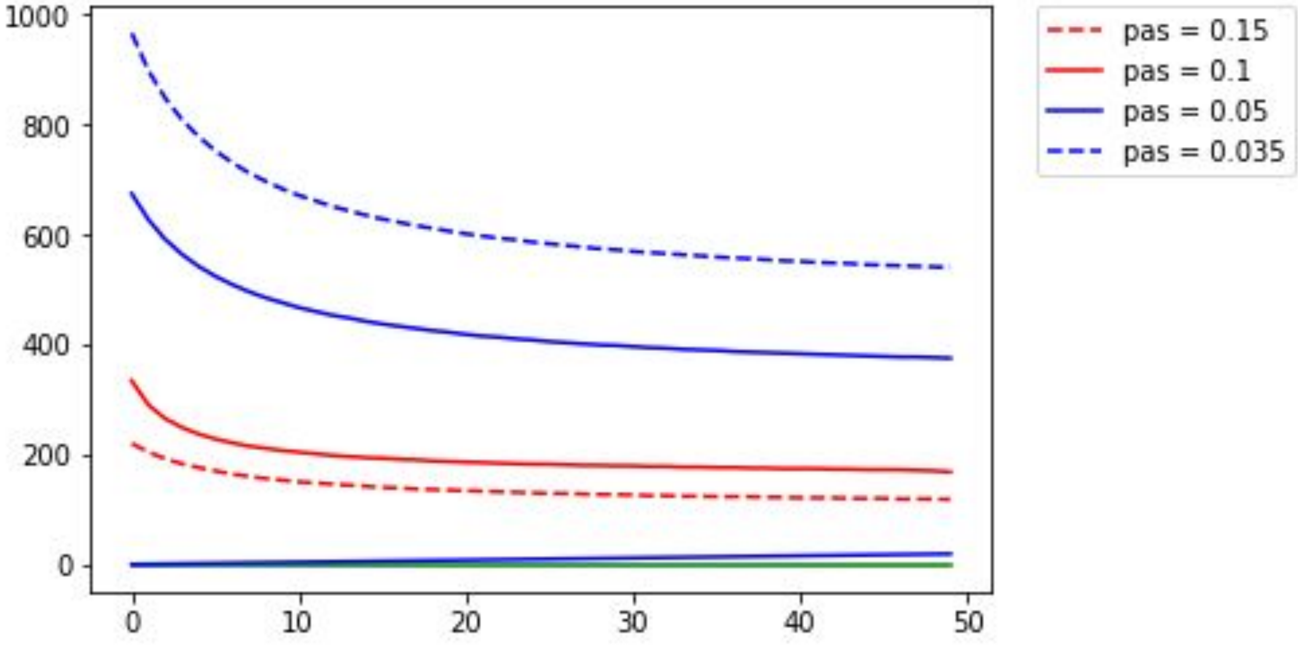
<u>Descente de gradient, pas de 0.1</u>



<u>Descente de gradient divergente, pas de 0.1</u>

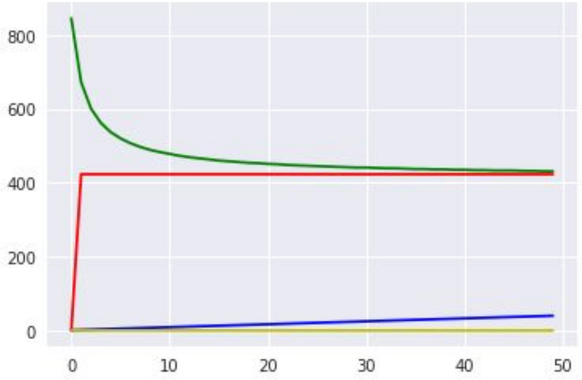


Descente de gradient : comparatif

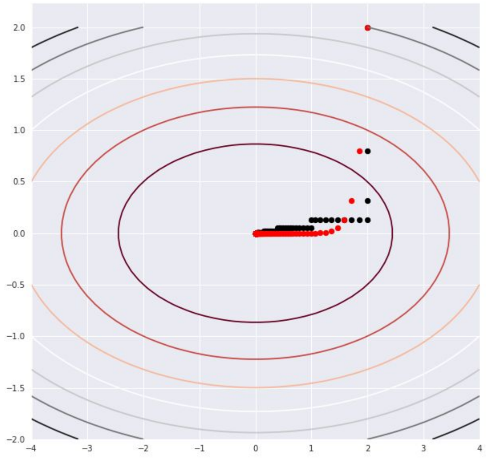


Comparaison norme L1 - L2

- Le nombre d’itération converge vers la même valeur quand le nombre de conditionnement augmente.
- En norme L1 les itérations suivent la direction du gradient.
- En norme L2 les itérations suivent les directions de l’axe orthogonal.

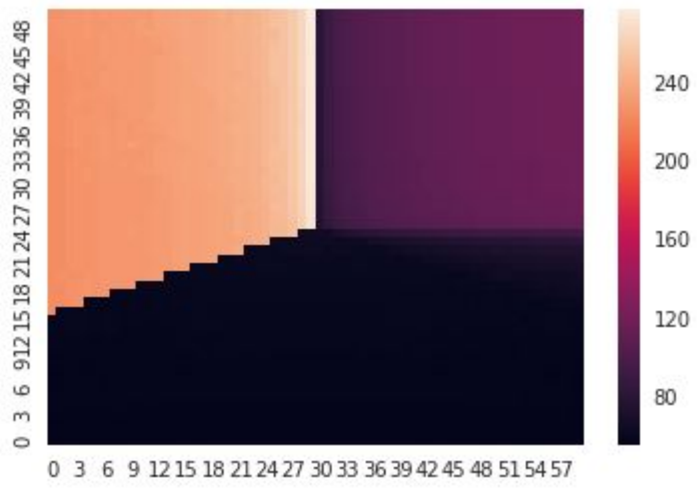


Nombre d’itération en fonction du conditionnement



Visualisation des itérations

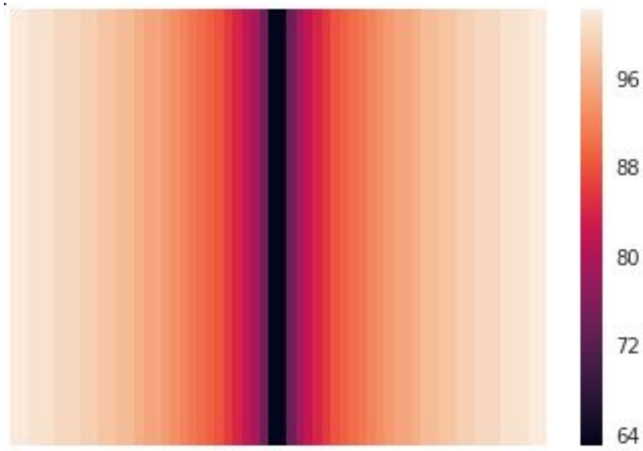
Etude du point de départ de la descente



Nombre d’itérations en fonction du point de départ en norme L1

- Le nombre d’itérations ne dépend pas de la distance par rapport au point optimal.
- Les coordonnées négatives en Y semblent donner de meilleurs résultats.
- Les pires résultats sont pour les coordonnées en X négatifs et Y positifs.

Etude du point de départ de la descente

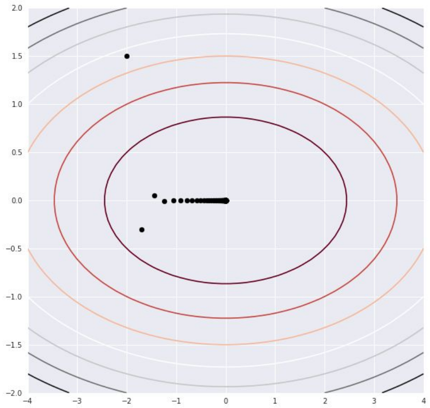


Nombre d’itérations en fonction du point de départ
en norme L2

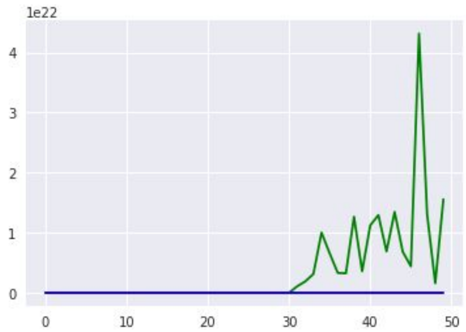
- Dépend fortement de la distance en X.

Méthodes d’optimisations

Learning rate scheduling



Learning rate scheduling, pas de 0.3



Descente classique, pas de 0.3

- Variation du step en fonction du nombre d'itérations.

- Implémentation d'un learning rate décroissant, méthode exponentielle

$$S_n = S_0.e^{-k.t}$$

Avec S_i le step à l'itération i, et k un hyperparamètre introduit.

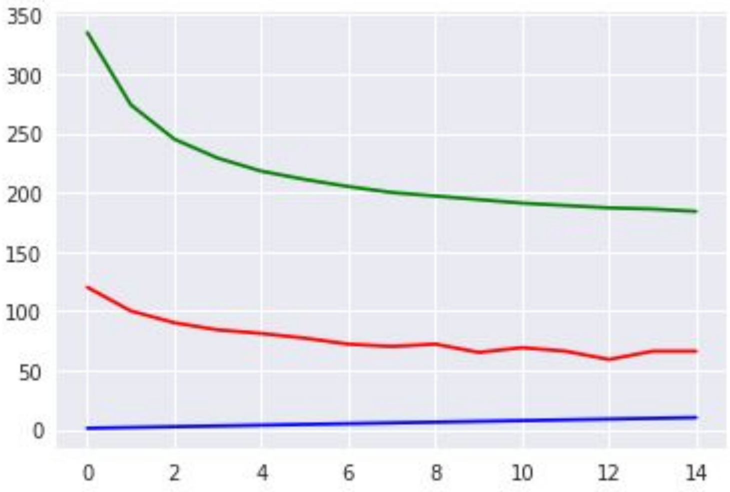
- On obtient une convergence pour les nombres de conditionnement élevés avec un step de 0.3

Nesterov

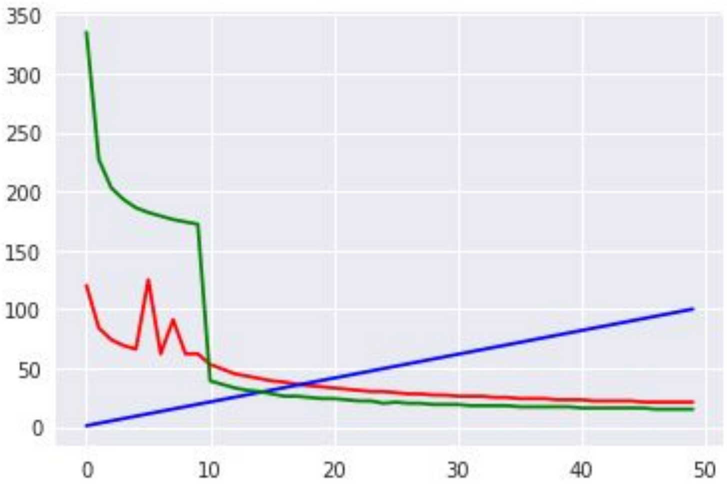
<u>Légende:</u>
Nombre d'itérations (Nesterov)
Nombre d'itérations (Descente classique)
Nombre de conditionnement

- Accélération de la descente via une nouvelle méthode de calcul des itérations

<u>Pas de 0.1, nombre de conditionnement de 0 à 10</u>



<u>Pas de 0.1, nombre de conditionnement de 0 à 100</u>

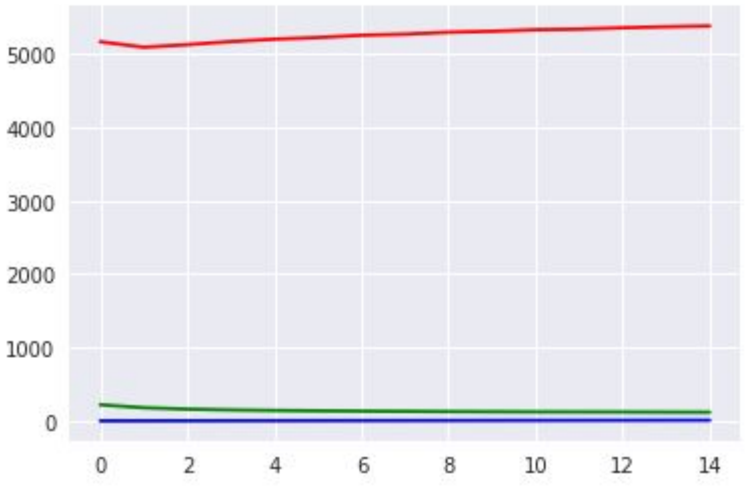


Adam optimization

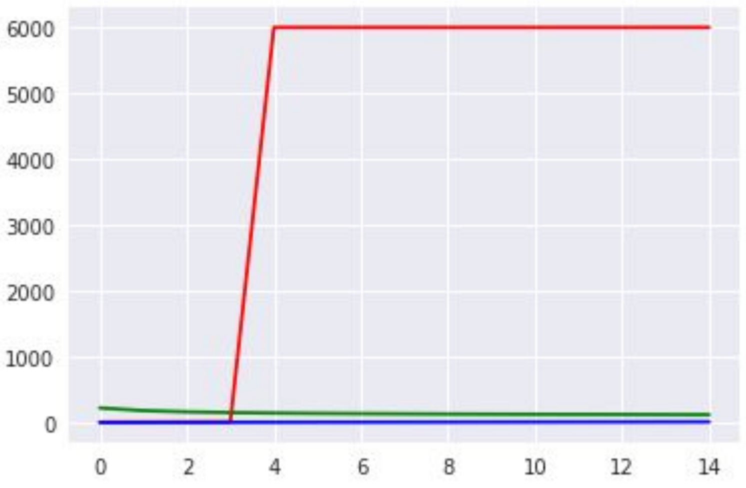
<u>Légende:</u>
Nombre d’itérations (Adam)
Nombre d’itérations (Descente classique)
Nombre de conditionnement

- Adaptive Moment Estimation
- Moyennes mobiles du moment d’ordre 1 et 2

<u>Hyperparamètres par défault</u>

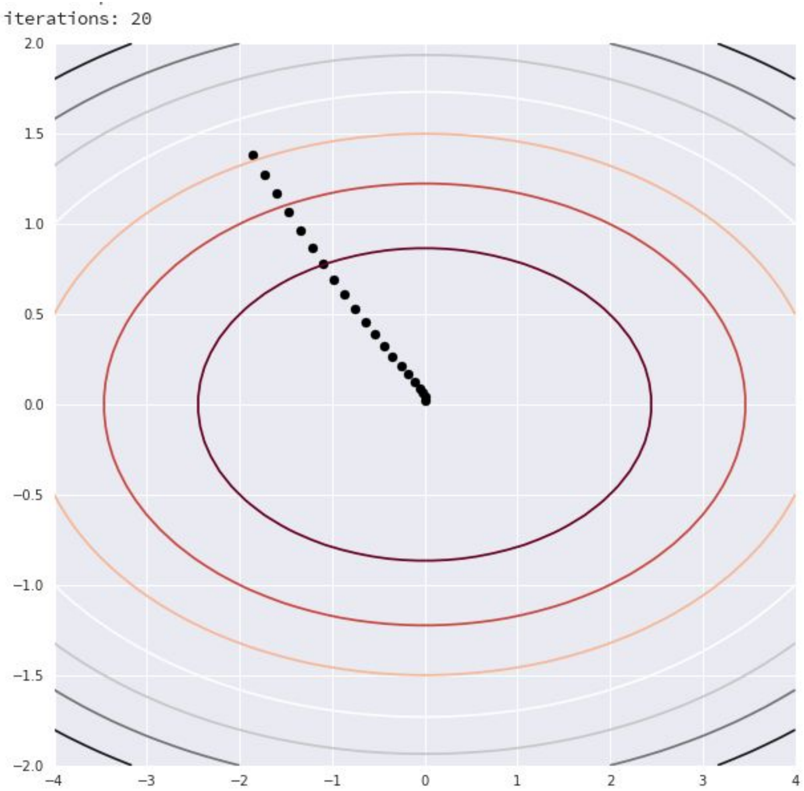


<u>Hyperparamètres adaptés</u>



Adam Optimization: direction de la descente

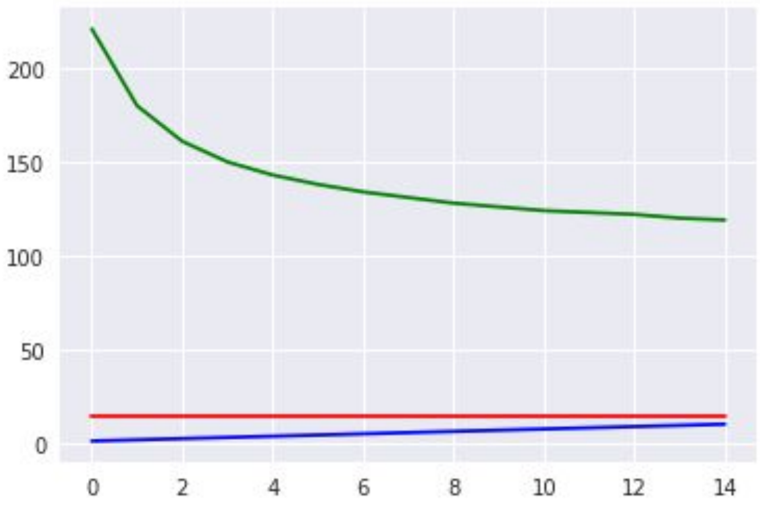
- Converge très rapidement sur les problèmes convexes.
- 20 itérations dans notre exemple

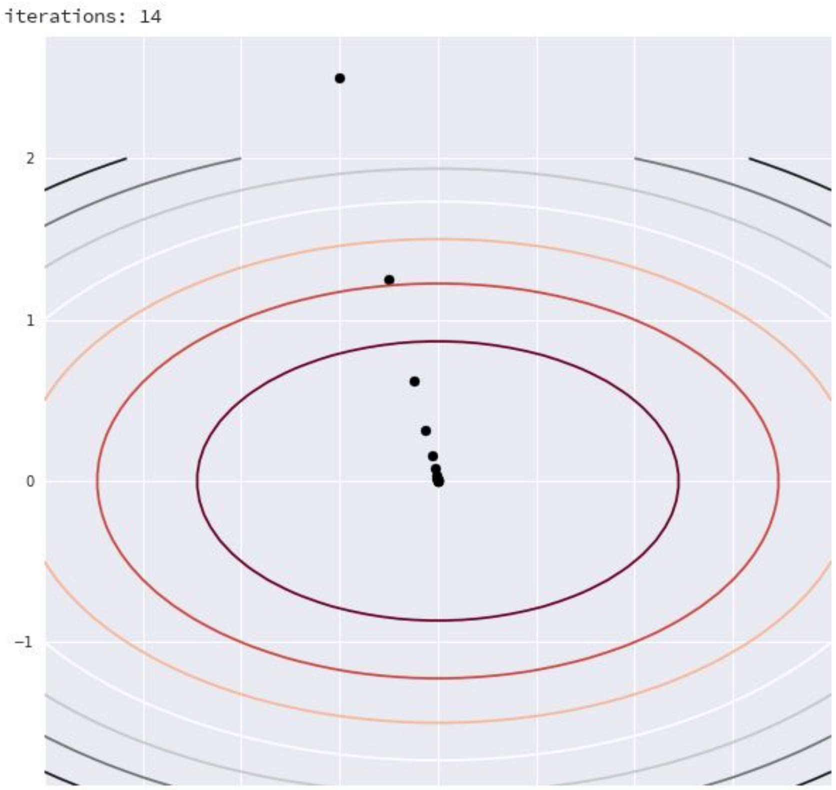


Méthode de Newton

Descente de gradient avec Newton

- La méthode de Newton permet d’obtenir les zéros d’une fonction
- Dans le cas d’un descente de gradient, un minimum x est disponible lorsque grad(f(x)) = 0





Quasi Newton

- Dérive de la méthode de Newton
- Utilise une approximation du de la Jacobienne
- Permet d’optimiser les calculs lorsque les matrices deviennent importantes

Conclusion

- Il existe différentes méthodes pour optimiser une descente de gradient
- Selon le besoin, il convient d’utiliser une méthode plutôt qu’une autre
- Il est important de chercher les paramètres les mieux adaptés à notre problème (learning rate, hyperparamètres, …)

Des questions?