Neural Networks and Deep Learning: Optimization Issues

Nicolas Thome

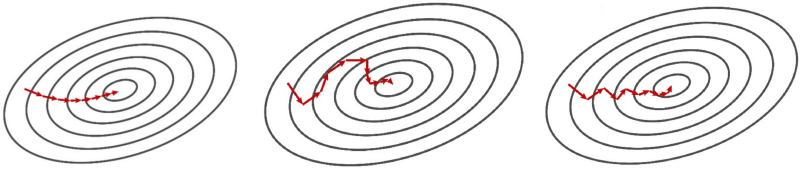
Conservatoire National des Arts et Métiers (Cnam)
Département Informatique

Beyond Stochastic Gradient Descent (SGD)

- Gradient descent optimization:

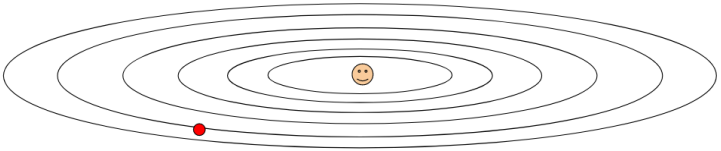
$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta \frac{\partial f}{\partial \mathbf{w}} (\mathbf{w}^t) = \mathbf{w}^t - \eta \nabla f (\mathbf{w}^t)$$

- **1st issue:** objective f changes quickly in one direction and slowly in another
- **2nd issue:** Stochastic Gradient Descent (SGD)

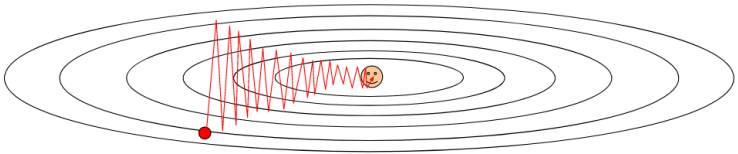


Full gradient | SGD (online) | SGD (mini-batch)

Beyond Stochastic Gradient Descent (SGD)



- Poor conditioning on Hessian matrix, *i.e.* large condition number (largest/smallest singular value)

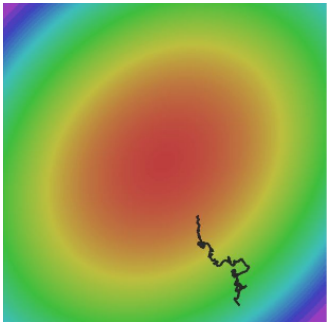


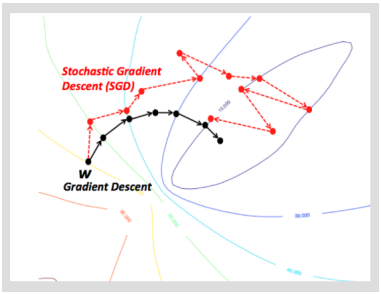
- **Gradient descent: very slow progress along shallow dimension, jitter along steep direction**

Beyond Stochastic Gradient Descent (SGD)

- Fonction $f(\mathbf{w}) = \sum_{i=1}^{N} f_i(\mathbf{w})$, *e.g.* $f_i(\mathbf{w}) = \ell_{CE}(\hat{\mathbf{y}}_i, \mathbf{y}_i^*)$
- SGD: approximation of the true gradient, *e.g.* for mini-batch:

$$\nabla f\left(\mathbf{w}^t\right) \approx \frac{1}{B} \sum_{i=1}^{B} \frac{\partial f_i(\mathbf{w})}{\partial \mathbf{w}} \left(\mathbf{w}^{(t)}\right)$$



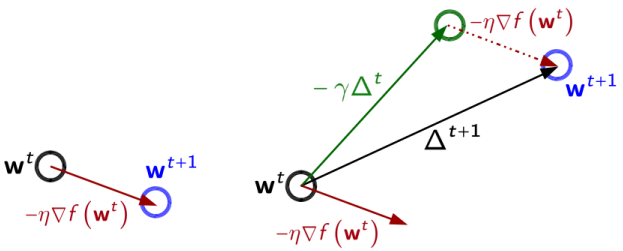


⇒ **Noisy gradient and poor convergence**

Momentum

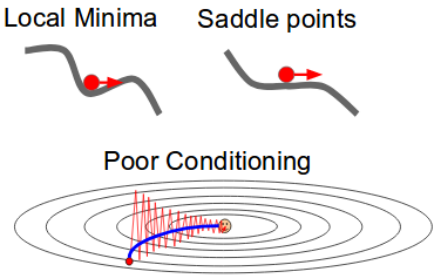
- In all SGD variants/improvement: $\boxed{\mathbf{w}^{t+1} = \mathbf{w}^t - \Delta^{t+1}}$
  - Δ^{t+1}: update vector $\mathbf{w}^t \rightarrow \mathbf{w}^{t+1}$
  - Ex: Gradient descent: $\Delta^{t+1} = \eta \nabla f(\mathbf{w}^t)$
- **Momentum:** use previous gradient memory, *e.g.* running average:

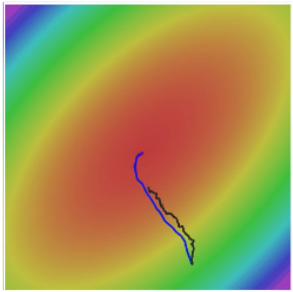
$$\Delta^{t+1} = \eta \nabla f(\mathbf{w}^t) + \gamma \Delta^t \qquad \gamma \in [0;1[\ (0.5, 0.9)$$



Momentum

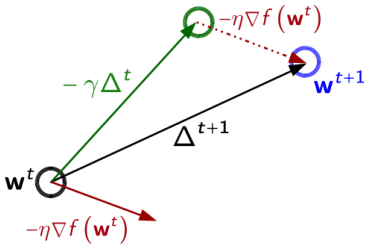
- Δ^t: $\mathbf{v}^t \sim$ velocity, inertia or memory: $\mathbf{w}^{t+1} = \mathbf{w}^t - \mathbf{v}^{t+1}$
  - Dimensions with oscillating gradient directions $\Rightarrow \mathbf{v}^t$ damped
  - Dimensions with small but consistent gradient direction $\Rightarrow \mathbf{v}^t$ increased
- More robust to local minima/saddle points, poor conditioning and noisy gradients



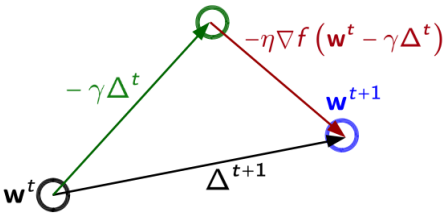


Nesterov Accelerate Gradient (NAG)

- NAG [Sutskever et al., 2013]
- ~ Momentum, but compute gradient at position predicted by Δ^t
  - $\Delta^{t+1} = \eta \nabla f\left(\mathbf{w}^t - \gamma \Delta^t\right) + \gamma \Delta^t \qquad \gamma \sim 0.9$
  - $\mathbf{w}^{t+1} = \mathbf{w}^t - \Delta^{t+1}$



Momentum



NAG

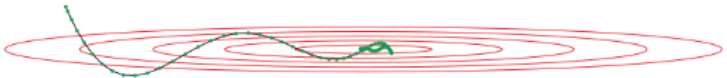
Nesterov Accelerate Gradient (NAG) [Sutskever et al., 2013]

With $\mathbf{x}^t = \mathbf{w}^t - \gamma\Delta^t$, more convenient update rule:

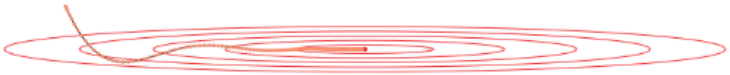
- $\Delta^{t+1} = \eta\nabla f\left(x^t\right) + \gamma\Delta^t$
- $\mathbf{x}^{t+1} = \mathbf{x}^t + \gamma\Delta^t - (\gamma + 1)\Delta^{t+1}$

- $\oplus$: anticipatory update $\Rightarrow$ ~~too large updates and overshooting~~
- $\oplus$: Increased responsiveness to the landscape of loss function f

Momentum

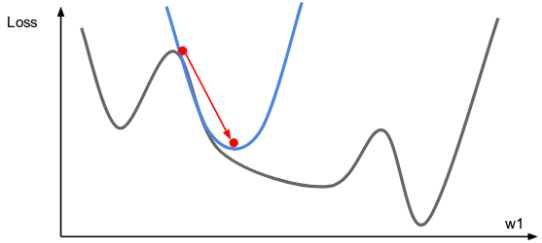


NAG



Optimization Schemes: Conclusion

- First-order methods, *e.g.* Momentum, NAG: better convergence
- Second-order methods: little adapted to stochastic training



- **Advanced optimization methods, per-dimension learning rate adaptation $\Rightarrow$ following !**

References I

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013).
On the importance of initialization and momentum in deep learning.
In Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, ICML'13, pages III–1139–III–1147. JMLR.org.