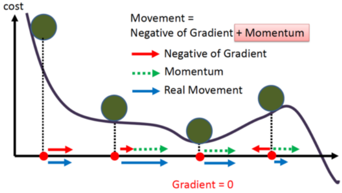
Optimization in Machine Learning

First order methods: GD with Momentum



Learning goals

- Recap of GD problems
- Momentum definition
- Unrolling formula
- Examples
- Nesterov

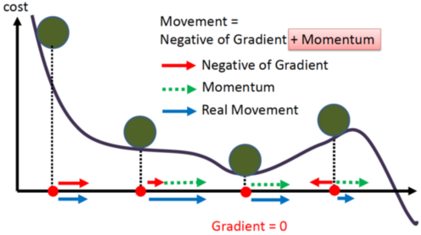
RECAP: WEAKNESSES OF GRADIENT DESCENT

- **Zig-zagging behavior:** For ill-conditioned problems, GD moves with a zig-zag course to the optimum, since the gradient points approximately orthogonal in the shortest direction to the minimum.
- **Slow crawling:** may vanish rapidly close to stationary points (e.g. saddle points) and hence also slows down progress.
- **Trapped in stationary points:** In some functions GD converges to stationary points (e.g. saddle points) since gradient on all sides is fairly flat and the step size is too small to pass this flat part.

Aim: More efficient algorithms which quickly reach the minimum.

GD WITH MOMENTUM

- **Idea:** "Velocity" $\boldsymbol{\nu}$. Velocity increases if successive gradients point in the same direction, but decreases if are opposite / different.



Source: H. Khandewal, *Gradient Descent with Momentum, RMSprop And Adam Optimizer*, Medium 2020.

- $\boldsymbol{\nu}$ is the (weighted) moving average of the previous gradients:

$$\begin{aligned} \boldsymbol{\nu}^{[t+1]} &\leftarrow \varphi \boldsymbol{\nu}^{[t]} - \alpha \nabla f(\mathbf{x}^{[t]}) \\ \mathbf{x}^{[t+1]} &\leftarrow \mathbf{x}^{[t]} + \boldsymbol{\nu}^{[t+1]} \end{aligned}$$

- α is the step size and $\varphi \in [0, 1]$ is an additional hyperparameter.

GD WITH MOMENTUM

- In GD: the step size is simply the gradient multiplied by the learning rate α
- Now, the step size depends on how large and how aligned a sequence of gradients is. The step size grows when many successive gradients point in the same direction.
- φ determines how strongly previous gradients are included in $\boldsymbol{\nu}$.
- Common values for φ are 0.5, 0.9 and even 0.99
- In general, the larger φ is relative to the learning rate α, the more previous gradients affect the current direction.
- $\varphi = 0$ equals gradient descent.
- Can be seen as GD with “short term memory” for the direction of motion.

MOMENTUM: EXAMPLE

$$
\begin{aligned}
\boldsymbol{\nu}^{[1]} &\leftarrow \varphi \boldsymbol{\nu}^{[0]} - \alpha \nabla f(\mathbf{x}^{[0]}) \\
\mathbf{x}^{[1]} &\leftarrow \mathbf{x}^{[0]} + \varphi \boldsymbol{\nu}^{[0]} - \alpha \nabla f(\mathbf{x}^{[0]})
\end{aligned}
$$

MOMENTUM: EXAMPLE

$$
\begin{aligned}
\boldsymbol{\nu}^{[1]} &\leftarrow \varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\mathbf{x}^{[0]}) \\
\mathbf{x}^{[1]} &\leftarrow \mathbf{x}^{[0]} + \varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\mathbf{x}^{[0]}) \\
\boldsymbol{\nu}^{[2]} &\leftarrow \varphi\boldsymbol{\nu}^{[1]} - \alpha\nabla f(\mathbf{x}^{[1]}) \\
&= \varphi(\varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\mathbf{x}^{[0]})) - \alpha\nabla f(\mathbf{x}^{[1]}) \\
\mathbf{x}^{[2]} &\leftarrow \mathbf{x}^{[1]} + \varphi(\varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\mathbf{x}^{[0]})) - \alpha\nabla f(\mathbf{x}^{[1]})
\end{aligned}
$$

MOMENTUM: EXAMPLE

$$
\begin{aligned}
\boldsymbol{\nu}^{[1]} &\leftarrow \varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\boldsymbol{x}^{[0]}) \\
\boldsymbol{x}^{[1]} &\leftarrow \boldsymbol{x}^{[0]} + \varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\boldsymbol{x}^{[0]}) \\
\boldsymbol{\nu}^{[2]} &\leftarrow \varphi\boldsymbol{\nu}^{[1]} - \alpha\nabla f(\boldsymbol{x}^{[1]}) \\
&= \varphi(\varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\boldsymbol{x}^{[0]})) - \alpha\nabla f(\boldsymbol{x}^{[1]}) \\
\boldsymbol{x}^{[2]} &\leftarrow \boldsymbol{x}^{[1]} + \varphi(\varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\boldsymbol{x}^{[0]})) - \alpha\nabla f(\boldsymbol{x}^{[1]}) \\
\boldsymbol{\nu}^{[3]} &\leftarrow \varphi\boldsymbol{\nu}^{[2]} - \alpha\nabla f(\boldsymbol{x}^{[2]}) \\
&= \varphi(\varphi(\varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\boldsymbol{x}^{[0]})) - \alpha\nabla f(\boldsymbol{x}^{[1]})) - \alpha\nabla f(\boldsymbol{x}^{[2]}) \\
\boldsymbol{x}^{[3]} &\leftarrow \boldsymbol{x}^{[2]} + \varphi(\varphi(\varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\boldsymbol{x}^{[0]})) - \alpha\nabla f(\boldsymbol{x}^{[1]})) - \alpha\nabla f(\boldsymbol{x}^{[2]}) \\
&= \boldsymbol{x}^{[2]} + \varphi^{3}\boldsymbol{\nu}^{[0]} - \varphi^{2}\alpha\nabla f(\boldsymbol{x}^{[0]}) - \varphi\alpha\nabla f(\boldsymbol{x}^{[1]}) - \alpha\nabla f(\boldsymbol{x}^{[2]}) \\
&= \boldsymbol{x}^{[2]} - \alpha(\varphi^{2}\nabla f(\boldsymbol{x}^{[0]}) + \varphi^{1}\nabla f(\boldsymbol{x}^{[1]}) + \varphi^{0}\nabla f(\boldsymbol{x}^{[2]})) + \varphi^{3}\boldsymbol{\nu}^{[0]}
\end{aligned}
$$

MOMENTUM: EXAMPLE

$$\begin{aligned}
\boldsymbol{\nu}^{[1]} &\leftarrow \varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\boldsymbol{x}^{[0]}) \\
\boldsymbol{x}^{[1]} &\leftarrow \boldsymbol{x}^{[0]} + \varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\boldsymbol{x}^{[0]}) \\
\boldsymbol{\nu}^{[2]} &\leftarrow \varphi\boldsymbol{\nu}^{[1]} - \alpha\nabla f(\boldsymbol{x}^{[1]}) \\
&= \varphi(\varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\boldsymbol{x}^{[0]})) - \alpha\nabla f(\boldsymbol{x}^{[1]}) \\
\boldsymbol{x}^{[2]} &\leftarrow \boldsymbol{x}^{[1]} + \varphi(\varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\boldsymbol{x}^{[0]})) - \alpha\nabla f(\boldsymbol{x}^{[1]}) \\
\boldsymbol{\nu}^{[3]} &\leftarrow \varphi\boldsymbol{\nu}^{[2]} - \alpha\nabla f(\boldsymbol{x}^{[2]}) \\
&= \varphi(\varphi(\varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\boldsymbol{x}^{[0]})) - \alpha\nabla f(\boldsymbol{x}^{[1]})) - \alpha\nabla f(\boldsymbol{x}^{[2]}) \\
\boldsymbol{x}^{[3]} &\leftarrow \boldsymbol{x}^{[2]} + \varphi(\varphi(\varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\boldsymbol{x}^{[0]})) - \alpha\nabla f(\boldsymbol{x}^{[1]})) - \alpha\nabla f(\boldsymbol{x}^{[2]}) \\
&= \boldsymbol{x}^{[2]} + \varphi^3\boldsymbol{\nu}^{[0]} - \varphi^2\alpha\nabla f(\boldsymbol{x}^{[0]}) - \varphi\alpha\nabla f(\boldsymbol{x}^{[1]}) - \alpha\nabla f(\boldsymbol{x}^{[2]}) \\
&= \boldsymbol{x}^{[2]} - \alpha(\varphi^2\nabla f(\boldsymbol{x}^{[0]}) + \varphi^1\nabla f(\boldsymbol{x}^{[1]}) + \varphi^0\nabla f(\boldsymbol{x}^{[2]})) + \varphi^3\boldsymbol{\nu}^{[0]} \\
\boldsymbol{x}^{[t+1]} &= \boldsymbol{x}^{[t]} - \alpha\sum_{j=0}^{t}\varphi^j\nabla f(\boldsymbol{x}^{[t-j]}) + \varphi^{t+1}\boldsymbol{\nu}^{[0]}
\end{aligned}$$

MOMENTUM: EXAMPLE

Suppose momentum always observes the same gradient $\nabla f(\boldsymbol{x})$:

$$\begin{aligned}
\boldsymbol{x}^{[t+1]} &= \boldsymbol{x}^{[t]} - \alpha \sum_{j=0}^{t} \varphi^{j} \nabla f(\boldsymbol{x}^{[j]}) + \varphi^{t+1} \boldsymbol{\nu}^{[0]} \\
&= \boldsymbol{x}^{[t]} - \alpha \nabla f(\boldsymbol{x}) \sum_{j=0}^{t} \varphi^{j} + \varphi^{t+1} \boldsymbol{\nu}^{[0]} \\
&= \boldsymbol{x}^{[t]} - \alpha \nabla f(\boldsymbol{x}) \frac{1 - \varphi^{t+1}}{1 - \varphi} + \varphi^{t+1} \boldsymbol{\nu}^{[0]} \\
&\to \boldsymbol{x}^{[t]} - \alpha \nabla f(\boldsymbol{x}) \frac{1}{1 - \varphi} \qquad \text{for } t \to \infty.
\end{aligned}$$

Thus, momentum will accelerate in the direction of $-\nabla f(\boldsymbol{x})$ until reaching terminal velocity with step size:

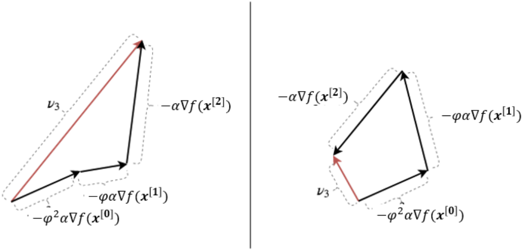
$$-\alpha \nabla f(\boldsymbol{x})(1 + \varphi + \varphi^2 + \varphi^3 + ...) = -\alpha \nabla f(\boldsymbol{x}) \frac{1}{1 - \varphi}$$

E.g. a momentum with $\varphi = 0.9$ corresponds to multiplying the maximum speed by 10 relative to the gradient descent algorithm.

MOMENTUM: EXAMPLE

The vector $\boldsymbol{\nu}^{[3]}$ (for $\boldsymbol{\nu}^{[0]} = 0$):

$$\begin{aligned}\boldsymbol{\nu}^{[3]} &= \varphi(\varphi(\varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\boldsymbol{x}^{[0]})) - \alpha\nabla f(\boldsymbol{x}^{[1]})) - \alpha\nabla f(\boldsymbol{x}^{[2]}) \\ &= -\varphi^2\alpha\nabla f(\boldsymbol{x}^{[0]}) - \varphi\alpha\nabla f(\boldsymbol{x}^{[1]}) - \alpha\nabla f(\boldsymbol{x}^{[2]})\end{aligned}$$


ν_3
$-\alpha\nabla f(\boldsymbol{x}^{[2]})$
$-\varphi\alpha\nabla f(\boldsymbol{x}^{[1]})$
$-\varphi^2\alpha\nabla f(\boldsymbol{x}^{[0]})$
$-\alpha\nabla f(\boldsymbol{x}^{[2]})$
$-\varphi\alpha\nabla f(\boldsymbol{x}^{[1]})$
ν_3
$-\varphi^2\alpha\nabla f(\boldsymbol{x}^{[0]})$

If consecutive (negative) gradients point mostly in the same direction, the velocity "builds up". On the other hand, if consecutive (negative) gradients point in very different directions, the velocity "dies down".

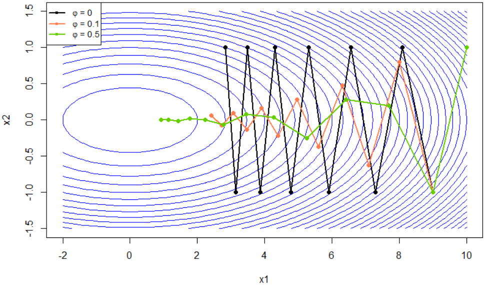
Further geometric intuitions as well as a detailed explanation can be found on the following website: `https://distill.pub/2017/momentum/`

GD WITH MOMENTUM: ZIG-ZAGGING BEHAVIOR

Consider a 2D quadratic form

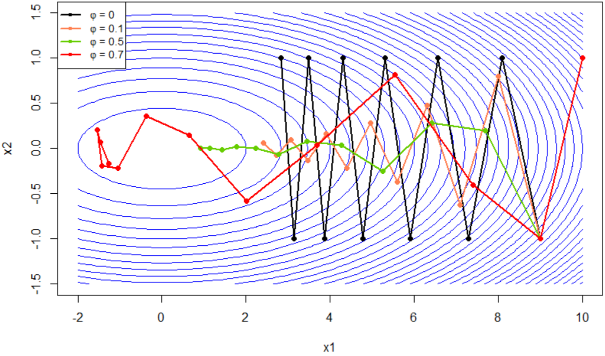
$$f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{C}\mathbf{x} = \mathbf{x}^\top \begin{pmatrix} 0.5 & 0 \\ 0 & 10 \end{pmatrix} \mathbf{x}, \quad \mathbf{x}^* = (0, 0)^\top.$$

Let $\mathbf{x}^{[0]} = (10, 1)^\top$, and $\alpha = 0.1$. GD shows stronger zig-zagging behavior than GD with momentum.



GD WITH MOMENTUM: ZIG-ZAGGING BEHAVIOR

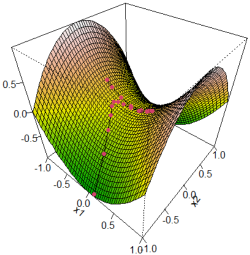
Be cautios: If the momentum is too high, the optimum will probably be missed, rolling past it and back. We might swing back and forth between local optima.

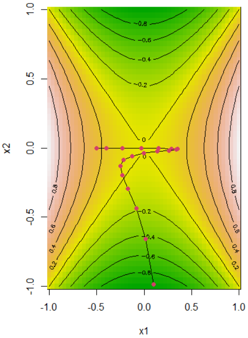


GD WITH MOMENTUM: MAGNITUDE VANISHING

Consider the 2D quadratic function $f(\boldsymbol{x}) = x_1^2 - x_2^2$with a saddle point at $(0,0)^\top$. Let $\boldsymbol{x}^{[0]} = (-0.5, 0.001)^\intercal$, and $\alpha = 0.1$.

The GD is slowing down at the saddle point (vanishing magnitude of the gradient), while GD with momentum “breaks out” of the saddle point and converges towards the minimum.



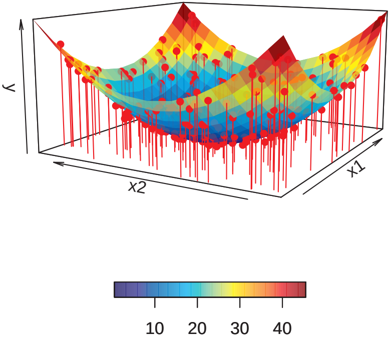


ERM FOR NN WITH GD

Let $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)\right)$, with $y = x_1^2 + x_2^2$ and minimize

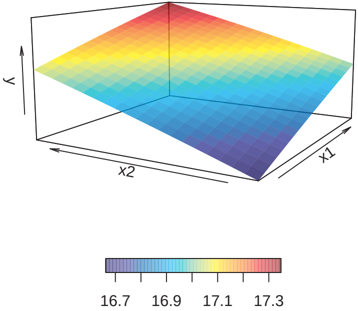
$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(f(\mathbf{x} \mid \boldsymbol{\theta}) - y^{(i)}\right)^2$$

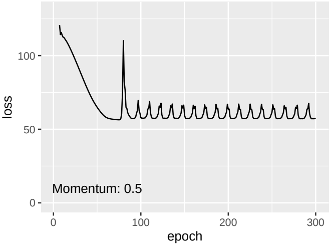
where $f(\mathbf{x} \mid \boldsymbol{\theta})$ is a neural network with 2 hidden layers (2 units each).



ERM FOR NN WITH GD

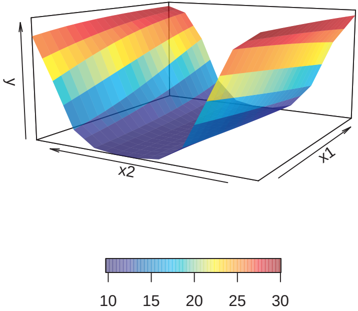
After 10 iters of GD:

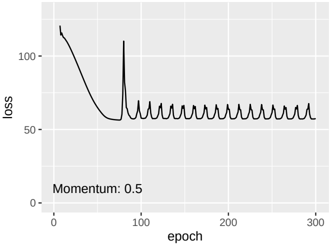




ERM FOR NN WITH GD

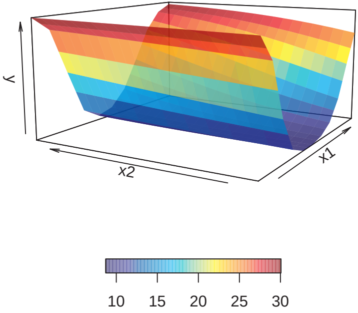
After 100 iters of GD:

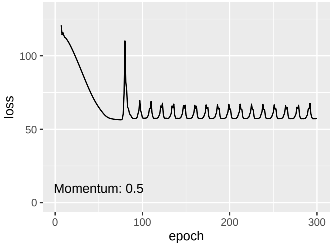




ERM FOR NN WITH GD

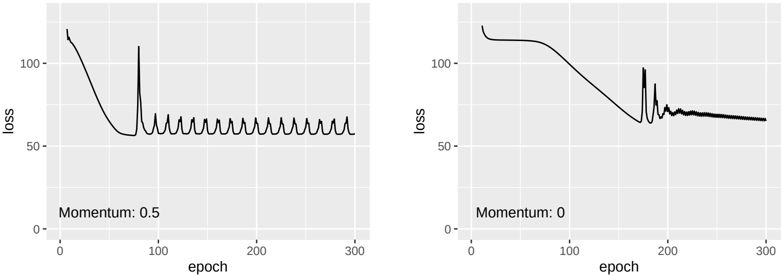
After 300 iters of GD:





ERM FOR NN WITH GD

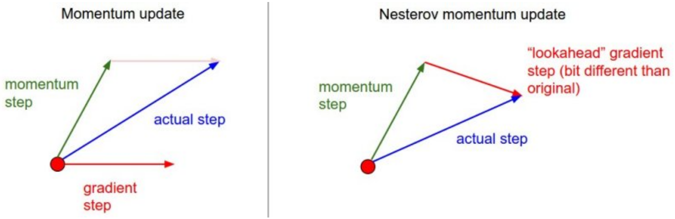
Gradient Descent with and without momentum



NESTEROV'S ACCELERATED GRADIENT

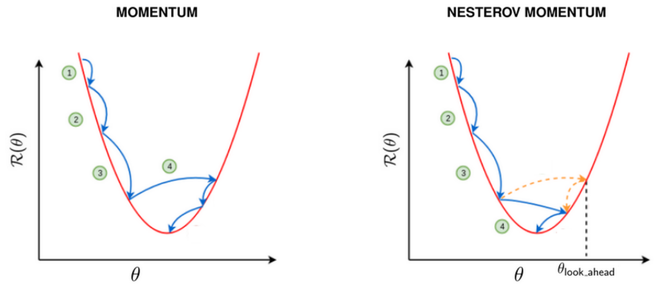
A slightly modified version is Nesterov momentum with stronger theoretical convergence guarantees for convex functions. Here, the gradient is computed at updated position

$$\tilde{\boldsymbol{x}} = \boldsymbol{x}^{[t]} + \varphi \boldsymbol{\nu}^{[t]}$$



Instead of evaluating the gradient at the current position, with Nesterov momentum we evaluate the gradient at the "looked-ahead" position. Source:https://cs231n.github.io/neural-networks-3/

MOMENTUM VS. NESTEROV MOMENTUM



Source: Chandra (2015)

Comparison GD with momentum (left) and GD with Nesterov momentum (right) for one parameter θ. The first three updates of θ are very similar in both cases and the updates become larger due to momentum (accumulation of previous negative gradients). Update 4 is different. In case of momentum, the update overshoots as it makes an even bigger step due to the gradient history. In contrast, Nesterov momentum first evaluates a "look-ahead" point $\theta_{\text{look\_ahead}}$, detects that it overshoots, and slightly reduces the overall magnitude of the fourth update. Thus, Nesterov momentum reduces overshooting and leads to smaller oscillations than momentum.