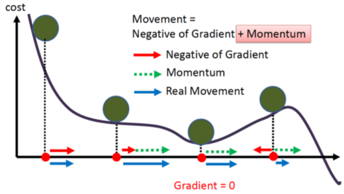
Optimization in Machine Learning

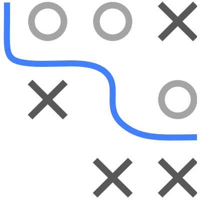
First order methods
GD with Momentum



Learning goals

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

RECAP: WEAKNESSES OF GRADIENT DESCENT

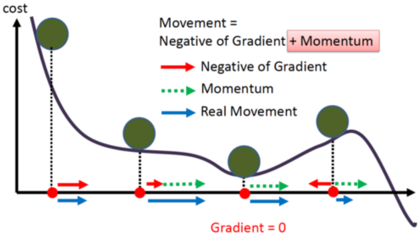


- **Zig-zagging behavior:** For ill-conditioned problems, GD moves with a zig-zag course to the optimum, since the gradient points approximately orthogonal in the shortest direction to the minimum.
- **Slow crawling:** may vanish rapidly close to stationary points (e.g. saddle points) and hence also 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” $\boldsymbol{\nu}$: Increasing if successive gradients point in the same direction but decreasing if they point in opposite directions



Source: Khandewal, *GD with Momentum, RMSprop and Adam Optimizer*, 2020.

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

$$\boldsymbol{\nu}^{[t+1]} = \varphi \boldsymbol{\nu}^{[t]} - \alpha \nabla f(\mathbf{x}^{[t]})$$
$$\mathbf{x}^{[t+1]} = \mathbf{x}^{[t]} + \boldsymbol{\nu}^{[t+1]}$$

- $\varphi \in [0, 1)$ is additional hyperparameter

GD WITH MOMENTUM / 2

- Length of a single step depends on how large and aligned a sequence of gradients is
- Length of a single step grows if 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 in relation to α, the more strongly previous gradients influence the current direction
- **Special case** $\varphi = 0$: "vanilla" gradient descent
- **Intuition:** GD with "short term memory" for the direction of motion

MOMENTUM: ANALYSIS

$$\boldsymbol{\nu}^{[1]} = \varphi \boldsymbol{\nu}^{[0]} - \alpha \nabla f(\mathbf{x}^{[0]})$$

$$\mathbf{x}^{[1]} = \mathbf{x}^{[0]} + \varphi \boldsymbol{\nu}^{[0]} - \alpha \nabla f(\mathbf{x}^{[0]})$$

MOMENTUM: ANALYSIS

$$\boldsymbol{\nu}^{[1]} = \varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\mathbf{x}^{[0]})$$

$$\mathbf{x}^{[1]} = \mathbf{x}^{[0]} + \varphi\boldsymbol{\nu}^{[0]} - \alpha\nabla f(\mathbf{x}^{[0]})$$

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

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

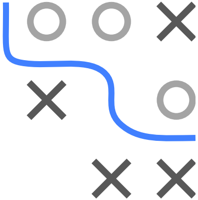
MOMENTUM: ANALYSIS

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

MOMENTUM: ANALYSIS

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

MOMENTUM: INTUITION



Suppose momentum always observes the same gradient $\nabla f(\mathbf{x}^{[t]})$:

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

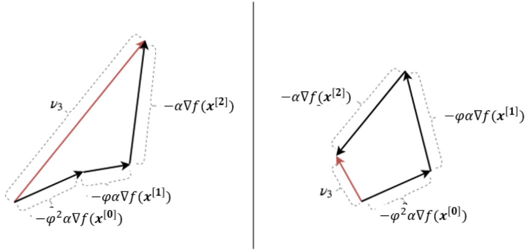
Momentum accelerates along $-\nabla f(\mathbf{x}^{[t]})$ to terminal velocity yielding step size $\alpha/(1 - \varphi)$.

Example: Momentum with $\varphi = 0.9$ corresponds to a tenfold increase in original step size α compared to vanilla gradient descent

MOMENTUM: INTUITION / 2

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

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

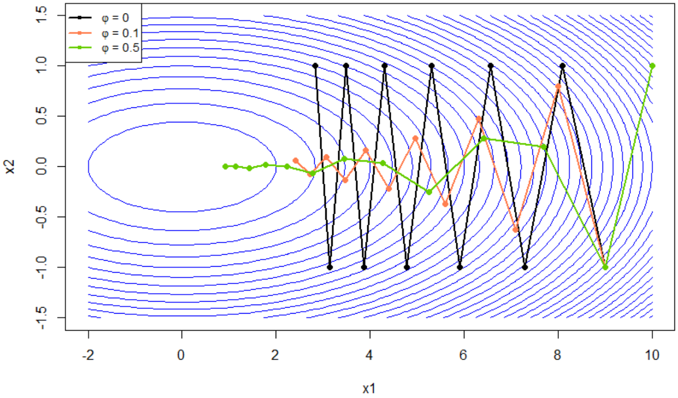


Successive gradients pointing in same/different directions increase/decrease velocity.
Further geometric intuitions and detailed explanations:
`https://distill.pub/2017/momentum/`

GD WITH MOMENTUM: ZIG-ZAG BEHAVIOUR

Consider a two-dimensional quadratic form $f(\mathbf{x}) = x_1^2/2 + 10x_2$.

Let $\mathbf{x}^{[0]} = (10, 1)^\top$ and $\alpha = 0.1$.

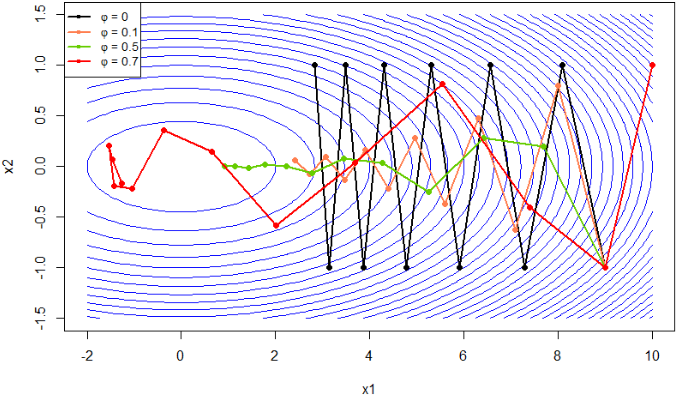


GD shows stronger zig-zag behaviour than GD with momentum.

GD WITH MOMENTUM: ZIG-ZAG BEHAVIOUR / 2

Caution:

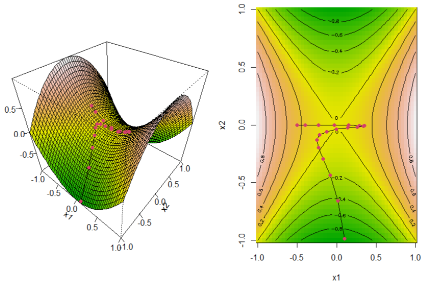
- If momentum is too high, minimum is possibly missed
- We might go back and forth around or between local minima



GD WITH MOMENTUM: SADDLE POINTS

Consider the two-dimensional quadratic form $f(\mathbf{x}) = x_1^2 - x_2^2$ with a saddle point at $(0, 0)^\top$.

Let $\mathbf{x}^{[0]} = (-1/2, 10^{-3})^\top$ and $\alpha = 0.1$.



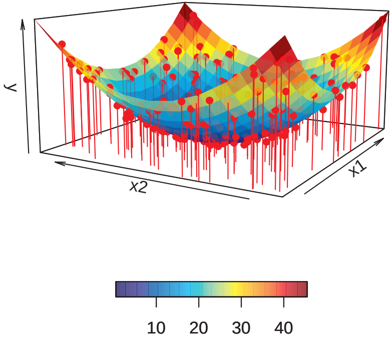
GD was slowing down at the saddle point (vanishing gradient).
GD with momentum “breaks out” of the saddle point and moves on.

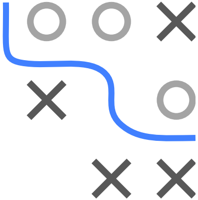
ERM FOR NN WITH GD

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

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

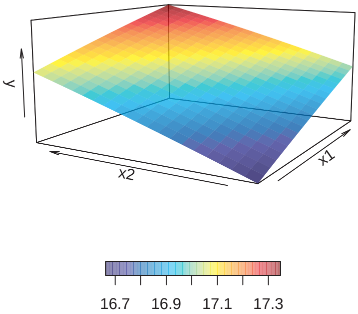
where $f(\mathbf{x} \mid \boldsymbol{\theta})$ is a neural network with 2 hidden layers (2 units each).

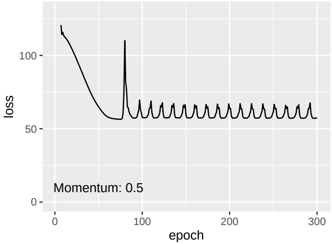


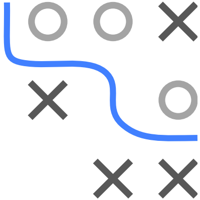


ERM FOR NN WITH GD / 2

After 10 iters of GD:

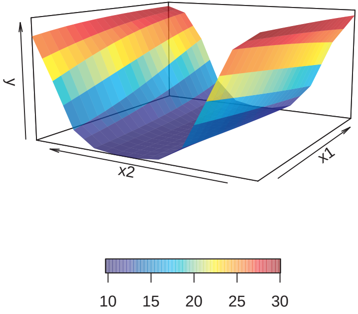


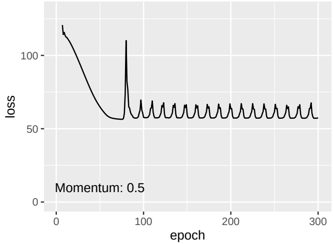


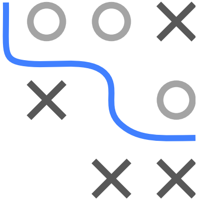


ERM FOR NN WITH GD / 3

After 100 iters of GD:

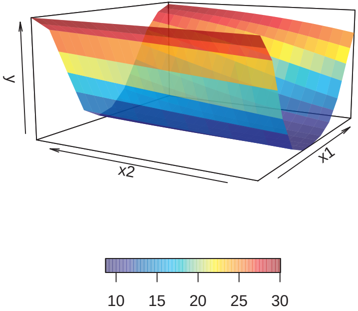


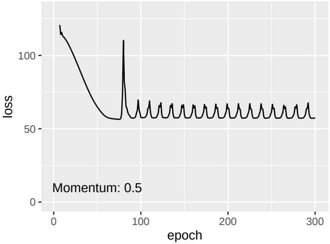


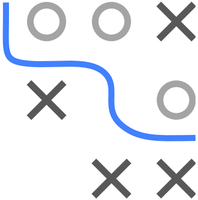


ERM FOR NN WITH GD / 4

After 300 iters of GD:

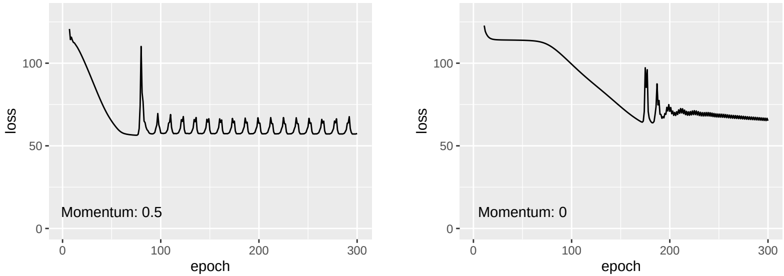


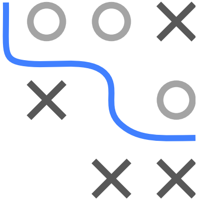




ERM FOR NN WITH GD / 5

Gradient Descent with and without momentum

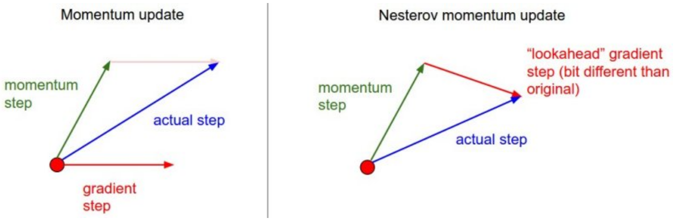




NESTEROV ACCELERATED GRADIENT

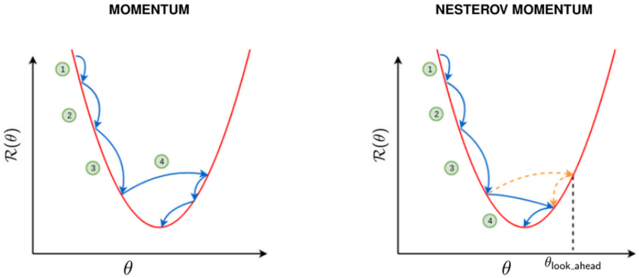
- Slightly modified version: **Nesterov accelerated gradient**
- Stronger theoretical convergence guarantees for convex functions
- Avoid moving back and forth near optima

$$\boldsymbol{\nu}^{[t+1]} = \varphi \boldsymbol{\nu}^{[t]} - \alpha \nabla f(\mathbf{x}^{[t]} + \varphi \boldsymbol{\nu}^{[t]})$$
$$\mathbf{x}^{[t+1]} = \mathbf{x}^{[t]} + \boldsymbol{\nu}^{[t+1]}$$



Nesterov momentum update evaluates gradient at the "look-ahead" position. (Source: `https://cs231n.github.io/neural-networks-3/`)

MOMENTUM VS. NESTEROV



GD with momentum (**left**) vs. GD with Nesterov momentum (**right**).
Near minima, momentum makes a large step due to gradient history.
Nesterov momentum "looks ahead" and reduces effect of gradient history.
(Source: Chandra, 2015)