

Acceleration of GD via Momentum

Axel Böhm

November 9, 2021

- 1 Optimal methods
- 2 Nesterov momentum

Smooth convex functions: less than $\mathcal{O}(\epsilon^{-1})$ steps?

Given L and $D = \|x_0 - x^*\|$ we know that

- ◇ GD converges with $\mathcal{O}(1/k)$
- ◇ cannot go faster (“lower bound”)

Maybe gradient descent is not the best possible algorithm?

After all it is arguably the simplest possible method using the gradient.

Smooth convex functions: less than $\mathcal{O}(\epsilon^{-1})$ steps?

So let's look at the following classes of methods:

First-order method:

- ◇ Access to the data only via an oracle which returns f and ∇f at given points.
- ◇ Clearly GD is a first order method.

Q: What is the **best** first-order method for smooth convex functions. *best* means: smallest upper bound on the number of oracle calls *in the worst case*.

Nemirovski and Yudin 1979 proved that every first-order method needs at least $\Omega(1/\sqrt{\epsilon})$ iterations (no method can be faster than $\mathcal{O}(1/k^2)$).

Acceleration to $\mathcal{O}(1/\sqrt{\epsilon})$ steps

- ◇ Nesterov 1983 came up with a method that needs only $\mathcal{O}(1/\sqrt{\epsilon})$ iterations (and is therefore the *best one*).
- ◇ Known as **Nesterov's accelerated gradient** method.
- ◇ By now multiple similar algorithms with same complexity exist.
- ◇ Proofs are generally not really instructive (some are computer assisted).

Nesterov's accelerated gradient method

Algorithm Nesterov's accelerated gradient method (NAG)

```
1: for  $k = 0, 1, \dots$  do
2:    $x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k)$ 
3:    $z_{k+1} = z_k - \frac{t+1}{2L} \nabla f(y_k)$ 
4:    $y_{k+1} = \frac{t+1}{t+3} x_{k+1} + \frac{2}{t+3} z_{k+1}$ 
```

- ◇ perform “**smooth step**” from y_k to x_{k+1}
- ◇ perform **aggressive step** from z_k to z_{k+1}
- ◇ form **weighted average** of the two
compensate for the aggressive step by giving less weight

Nesterov's algorithm as a momentum method

A different way to write the method is via **momentum**

$$\begin{aligned}y_k &= x_k + \beta_k(x_k - x_{k-1}) \\x_{k+1} &= y_k - \frac{1}{L}\nabla f(y_k).\end{aligned}$$

- ◇ differs from GD on in momentum/inertia term $\beta_k(x_k - x_{k-1})$
- ◇ has to be chosen carefully $\beta_k = \frac{k-1}{k+2}$
- ◇ coefficient approaches $\frac{k-1}{k+2} \approx 1 - \frac{3}{k}$

Nesterov's accelerated gradient method: convergence

Theorem

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex and L -smooth with minimum x^* , then NAG yields

$$f(x_k) - f(x^*) \leq \frac{2L\|x_0 - x^*\|^2}{k(k+1)}$$

Recall that the gradient descent bound was

$$f(x_k) - f(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2k}.$$

Proof idea

Potential function Φ that decreases along trajectory (standard technique). Out of the blue: Use

$$\Phi(k) := k(k+1)(f(x_k) - f^*) + 2L\|z_k - x^*\|^2.$$

Then show that

$$\Phi(k+1) \leq \Phi(k).$$

Results in

$$\Phi(k+1) \leq \Phi(k) \leq \dots \leq \Phi(0)$$

and therefore

$$k(k+1)(f(x_k) - f^*) \leq 2L\|z_0 - x^*\|^2.$$

Why momentum?

- ◇ GD has problems with **ravines**, i.e. areas where the surface curves much more steeply in one dimension than in another.
- ◇ Results in zig-zagging.



Figure: no momentum



Figure: with momentum

Momentum in terms of velocity

Consider a ball rolling down a slope. Its **velocity** is

$$v_k = \beta v_{k-1} + \alpha \nabla f(x_k)$$

$$x_{k+1} = x_k - v_k$$

- ◇ a fraction β of the **previous velocity** (friction)
- ◇ plus, steepness of the **slope**

In terms of iterates:

$$x_{k+1} = x_k - v_k$$

$$= x_k - \alpha \nabla f(x_k) - \beta v_{k-1}$$

$$= x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$$

Heavy ball: Polyak 1964

We derived

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}),$$

while Nesterov's method was

$$\begin{aligned} y_k &= x_k + \beta_k(x_k - x_{k-1}) \\ x_{k+1} &= y_k - \frac{1}{L} \nabla f(y_k). \end{aligned}$$

However, **Polyak's** momentum provides no speedup over $\mathcal{O}(1/k)$ for smooth convex function.

What's the difference?

- ◇ Both types of momentum seem so similar.
- ◇ Heavy ball does not care if do momentum or gradient first.
- ◇ Nesterov momentum applies inertia first then gradient.

$$\begin{aligned}v_k &= \beta v_{k-1} + \alpha \nabla f(x_k - \beta v_{k-1}) \\x_{k+1} &= x_k - v_k\end{aligned}$$

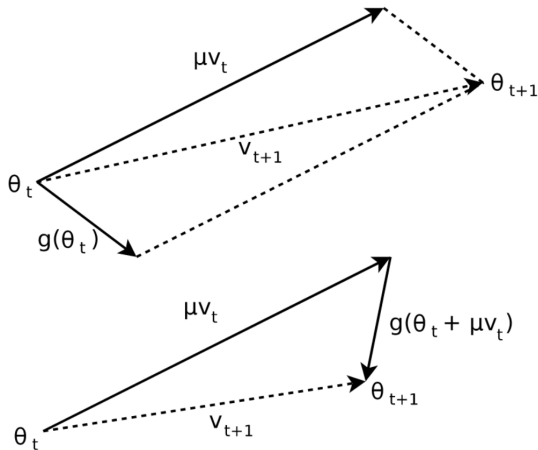


Figure: Nesterov vs Polyak momentum.

Momentum for strongly convex functions

For smooth strongly convex we know that GD obtains

$$\|x_{k+1} - x^*\|^2 \leq \left(1 - \frac{\mu}{L}\right) \|x_k - x^*\|^2$$

and

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k \frac{L \|x_0 - x^*\|^2}{2}.$$

Performance depends heavily on the **condition number** $\kappa := L/\mu$:

Contraction coefficient is $(1 - 1/\kappa)$.

Nesterov and Polyak momentum improve this to $(1 - 1/\sqrt{\kappa})$