Momentum Methods

Zachary Ross

February 10, 2023

In the following writeup, we explore accelerated gradient descent (GD) methods. These are adaptations to the general method

$$x_{t+1} = x_t - \eta \nabla f(x_t), \tag{1}$$

where η is the learning rate and $f : \mathbb{R}^n \to \mathbb{R}$ is a *convex* objective function with a unique minimum, which use previous descent directions as part of current iterates computation. The intuition behind using accelerated methods is that it introduces a physics-inspired component to GD which mimics a dampened oscillator ordinary differential equation (ODE). This effect minimizes the oscillations that tend to occur in gradient descent by using previous gradient computations to penalize rapid changes in direction and reward movement towards an optimal point.

We reconstruct this via a method similar to as was done by Gadat et al.: the ODE limit for gradient descent is given by the equation

$$\frac{dx}{dt} = -\nabla f(x). \tag{2}$$

This can be modified into a dampened oscillator ODE by including acceleration and dampening term $\gamma \geq 0$

$$\frac{d^2x}{dt^2} + \gamma \frac{dx}{dt} = -\nabla f(x) \tag{3}$$

and then rearranged to interpret in terms of acceleration, along with a simplification to velocity v of x, i.e. $v = \frac{dx}{dt}$,

$$\frac{dv}{dt} = -\gamma v - \nabla f(x). \tag{4}$$

Conversion to a discrete time step $\sqrt{\eta}$ with a right-derivative yields the equations

$$\frac{dv_t}{dt} \approx \frac{v_{t+1} - v_t}{\sqrt{\eta}}, \quad \text{and} \quad v_t = \frac{dx_t}{dt} \approx \frac{x_{t+1} - x_t}{\sqrt{\eta}} \tag{5}$$

which can be used to derive a formula for the dampened oscillator GD. Let $\beta = 1 - \gamma\sqrt{\eta}$. Then

$$\begin{aligned} \frac{v_t - v_{t-1}}{\sqrt{\eta}} &= -\gamma v_{t-1} - \nabla f(x_{t-1}) \\ v_t &= \beta v_{t-1} - \sqrt{\eta}\nabla f(x_{t-1}) \\ \frac{x_{t+1} - x_t}{\sqrt{\eta}} &= \beta \frac{x_t - x_{t-1}}{\sqrt{\eta}} - \sqrt{\eta}\nabla f(x_{t-1}) \\ x_{t+1} &= x_t + \beta(x_t - x_{t-1}) - \eta \nabla f(x_{t-1}) \end{aligned} \tag{6}$$

By modifying this equation to evaluate the gradient at x_t rather than x_{t-1}, this yields the *heavyball* method as was discovered by Polyak

$$x_{t+1} = x_t + \beta(x_t - x_{t-1}) - \eta \nabla f(x_t). \tag{7}$$

The update to this method

Despite the effectiveness of this method, Polyak's still has the potential to oscillate infinitely under specific conditions, as observed by Lessard et al.. This is due to the gradient being evaluated *before* momentum is

applied and x_t sometimes serving as a poor approximate for x_{t-1}. Nesterov altered this method slightly by evaluating the gradient *after* momentum is applied, i.e.

$$x_{t+1} = x_t + \beta(x_t - x_{t-1}) - \eta\nabla f(x_t + \beta(x_t - x_{t-1})). \tag{8}$$

This derivation can be made by reinterpreting Equation 9 as a left derivative with timestep $\sqrt{\delta}$

$$\frac{dv_t}{dt} \to \frac{v_t - v_{t-1}}{\sqrt{\delta}}, \quad \text{and} \quad v_t = \frac{dx_t}{dt} \to \frac{x_t - x_{t-1}}{\sqrt{\delta}} \tag{9}$$

and re-deriving Equation 10 under this condition for v_{t+1}

$$\begin{aligned}
\frac{v_{t+1} - v_t}{\sqrt{\delta}} &= -\gamma v_{t+1} - \nabla f(x_{t+1}) \\
v_{t+1} - v_t &= -\gamma\sqrt{\delta} v_{t+1} - \sqrt{\delta}\nabla f(x_{t+1}) \\
\frac{x_{t+1} - x_t - x_t + x_{t-1}}{\sqrt{\delta}} &= -\gamma\sqrt{\delta}\frac{x_{t+1} - x_t}{\sqrt{\delta}} - \sqrt{\delta}\nabla f(x_{t+1}) \\
x_{t+1} - x_t - x_t + x_{t-1} &= -\gamma\sqrt{\delta}(x_{t+1} - x_t) - \delta\nabla f(x_{t+1}) \\
x_{t+1} &= x_t + \beta(x_t - x_{t-1}) - \beta\delta\nabla f(x_{t+1})
\end{aligned} \tag{10}$$

$$\beta = \frac{1}{1 + \gamma\sqrt{\delta}} \tag{11}$$

1 Questions

1. The derivation in Equation 9 defines a right-derivative to be used in the computation, which

References

[1] Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12(1):461 – 529, 2018. doi: 10.1214/18-EJS1395. URL `https://doi.org/10.1214/18-EJS1395`.

[2] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[3] Yurii Evgen'evich Nesterov. A method of solving a convex programming problem with convergence rate o\bigl(k^2\bigr). In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences, 1983.

[4] Boris Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr Computational Mathematics and Mathematical Physics*, 4:1–17, 12 1964. doi: 10.1016/0041-5553(64)90137-5.

2 Closed Form

Conversion to a discrete time step $\sqrt{\eta}$ yields the equations

$$\frac{dv_t}{dt} \to \frac{v_t - v_{t-1}}{\sqrt{\eta}}, \quad \text{and} \quad v_t = \frac{dx_t}{dt} \to \frac{x_t - x_{t-1}}{\sqrt{\eta}} \tag{12}$$

which can be used to derive a formula for a dampened oscillator for GD. Let $\beta = 1 - \gamma\sqrt{\eta}$. Then

$$\begin{aligned}
\frac{v_t - v_{t-1}}{\sqrt{\eta}} &= -\gamma v_{t-1} - \nabla f(x_t) \\
v_t &= \beta v_{t-1} - \sqrt{\eta}\nabla f(x_t) \\
\frac{x_{t+1} - x_t}{\sqrt{\eta}} &= \beta\frac{x_t - x_{t-1}}{\sqrt{\eta}} - \sqrt{\eta}\nabla f(x_t).
\end{aligned} \tag{13}$$

This yields the *heavyball* method as was discovered by Polyak.

$$x_{t+1} = x_t + \beta(x_t - x_{t-1}) - \eta \nabla f(x_t) \tag{14}$$