THE PHYSICS OF GRADIENT DESCENT

Aidan Rocke
aidanrocke@gmail.com

September 14, 2021

Nesterov Accelerated Gradient is basically gradient descent with a momentum term, which may be expressed as follows:

$$(x_{k+1} - x_k) - \alpha \cdot (x_k - x_{k-1}) + \eta \cdot \nabla f(x_k + \alpha \cdot (x_k - x_{k-1})) = 0 \tag{1}$$

where α is a damping term and η is the learning rate. In order to perform a continuum-limit approximation of (1), we may define:

$$t = k \cdot h \tag{2}$$

$$X(t) := x_{\lfloor t/h \rfloor} = x_k \tag{3}$$

where we have $x_k = X(t)$ and therefore:

$$X(t+h) = X(t) + \dot{X}(t) \cdot h + \frac{1}{2}\ddot{X}(t) \cdot h^2 + \mathcal{O}(h^3) \tag{4}$$

$$X(t-h) = X(t) - \dot{X}(t) \cdot h + \frac{1}{2}\ddot{X}(t) \cdot h^2 + \mathcal{O}(h^3) \tag{5}$$

This allows us to derive the following continuous-time approximations:

$$x_{k+1} - x_k = \dot{X}(t) \cdot h + \frac{1}{2}\ddot{X}(t) \cdot h^2 \tag{6}$$

$$x_k - x_{k-1} = \dot{X}(t) \cdot h - \frac{1}{2}\ddot{X}(t) \cdot h^2 \tag{7}$$

$$\eta \cdot \nabla f(x_k + \alpha \cdot (x_k - x_{k-1})) = \eta \cdot \nabla f(X(t)) \tag{8}$$

and so in the continuum-limit we have the differential equation for a Damped Harmonic Oscillator:

$$m \cdot \ddot{X}(t) + c \cdot \dot{X}(t) + \nabla f(X(t)) = 0 \tag{9}$$

where $m := \frac{(1+\alpha)\cdot h^2}{2\eta}$ is the particle mass, $c := \frac{(1-\alpha)\cdot h}{\eta}$ is the damping coefficient and $f(\cdot)$ is the potential field.

Therefore, from an optimisation perspective the equilibrium is essentially the minimiser of the potential function.

References

References:

[1] Lin F. Yang et al. The Physical Systems Behind Optimization Algorithms. Neurips. 2016.

[2] Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence o(1/k2). Doklady ANSSSR (translated as Soviet.Math.Docl.), vol. 269, pp. 543– 547.