PadhAI: Variants of Gradient Descent

One Fourth Labs

Intuition behind nesterov accelerated gradient descent

Can we do something to reduce the oscillation in Momentum based GD

1. Let us consider the Momentum based Gradient Descent Update Rule
   a. $v_t = \gamma * v_{t-1} + \eta \nabla \omega_t$
   b. $\omega_{t+1} = \omega_t - v_t$
   c. $\omega_{t+1} = \omega_t - \gamma * v_{t-1} - \eta \nabla \omega_t$
   d. Here, we can see that the movement occurs in two steps
      i. The first is with the history-term $\gamma * v_{t-1}$
      ii. The second is with the weight term $\eta \nabla \omega_t$
      iii. When moving both steps each time, it is possible to overshoot the minima between the two steps
      iv. So we can consider first moving with the history term, then calculate the second step from where we were located after the first step (ω_{temp}).
2. Using the above intuition, the Nesterov Accelerated Gradient Descent solves the problem of overshooting and multiple oscillations
   a. $\omega_{temp} = \omega_t - \gamma * v_{t-1}$ compute ω_{temp} based on movement with history
   b. $\omega_{t+1} = \omega_{temp} - \eta \nabla \omega_{temp}$ move further in the direction of the derivative of ω_{temp}
   c. $v_t = \gamma * v_{t-1} + \eta \nabla \omega_{temp}$ update history with movement due to derivative of ω_{temp}