

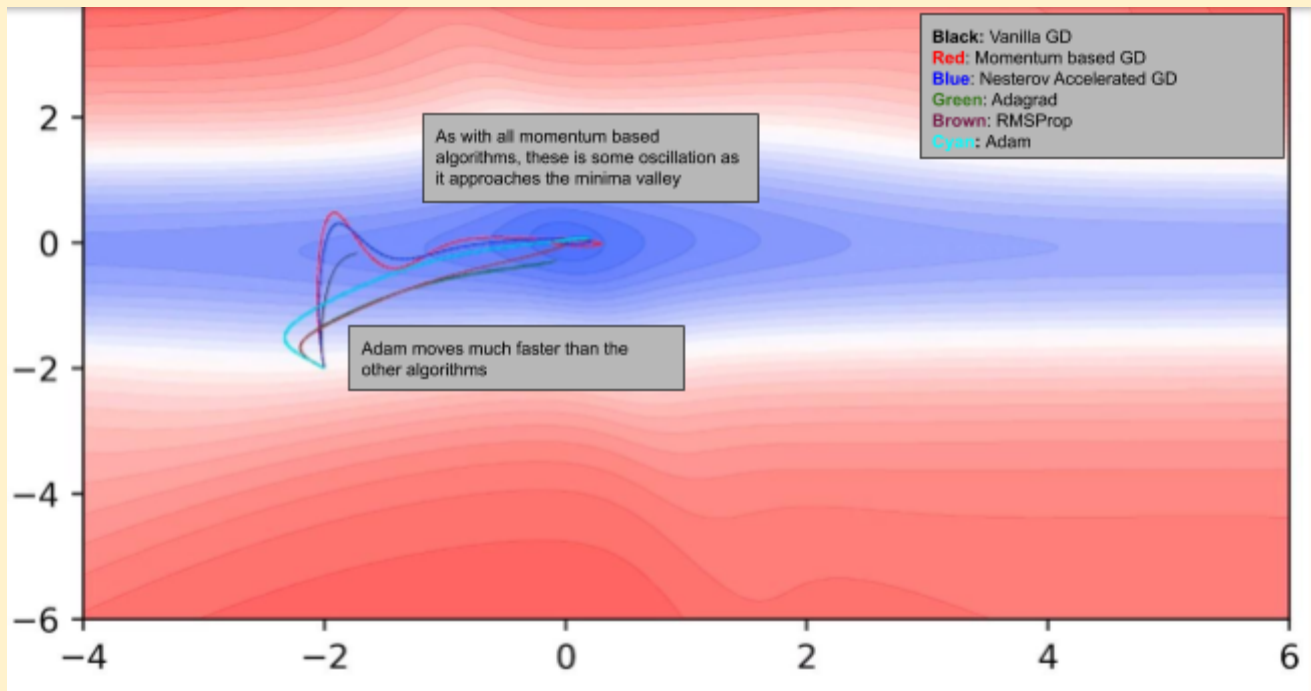
Running and Visualizing Adam

Does it make sense to use a cumulative history of gradients?

1. We have already looked at algorithms that make use of a history term
 - a. Momentum based GD: Makes use of the history of the gradients
 - i. $v_t = \gamma * v_{t-1} + \eta \nabla \omega_t$
 - ii. $\omega_{t+1} = \omega_t - v_t$
 - iii. Here, history is used to calculate the current update
 - b. RMSProp: Makes use of the history of the square of the gradients
 - i. $v_t = \beta * v_{t-1} + (1 - \beta)(\nabla \omega_t)^2$
 - ii. $\omega_{t+1} = \omega_t - \frac{\eta}{\sqrt{(v_t) + \epsilon}} \nabla \omega_t$
 - iii. Here, history is used to adjust the learning-rate
 - c. Can we combine these two ideas?
 - d. Yes, in the form of Adam, which uses both of those history terms
2. Adam
 - a. $m_t = \beta_1 * v_{t-1} + (1 - \beta_1)(\nabla \omega_t)$
 - i. This is very similar to the history that Momentum based GD maintains
 - ii. It's a running sum of all the updates done
 - b. $v_t = \beta_2 * v_{t-1} + (1 - \beta_2)(\nabla \omega_t)^2$
 - i. This is similar to the history that RMSProp maintains
 - ii. It is used to regulate the learning-rate
 - c. $\omega_{t+1} = \omega_t - \frac{\eta}{\sqrt{(v_t) + \epsilon}} m_t$
 - i. Here, the first history m_t is used to make the update, ensuring that the history of derivatives is used to calculate the current update
 - ii. The second derivative v_t is used to regulate the learning rate based on density or sparsity of the feature
 - d. In addition to the above points, Adam performs bias correction by using the following equations
 - i. $m_t = \frac{m_t}{1 - \beta_1^t}$
 - ii. $v_t = \frac{v_t}{1 - \beta_2^t}$
 - iii. It ensures that the training is smoother and also prevents erratic updates in beginning of training.

PadhAI: Variants of Gradient Descent

One Fourth Labs



3.