ENGR 4350: Applied Deep Learning

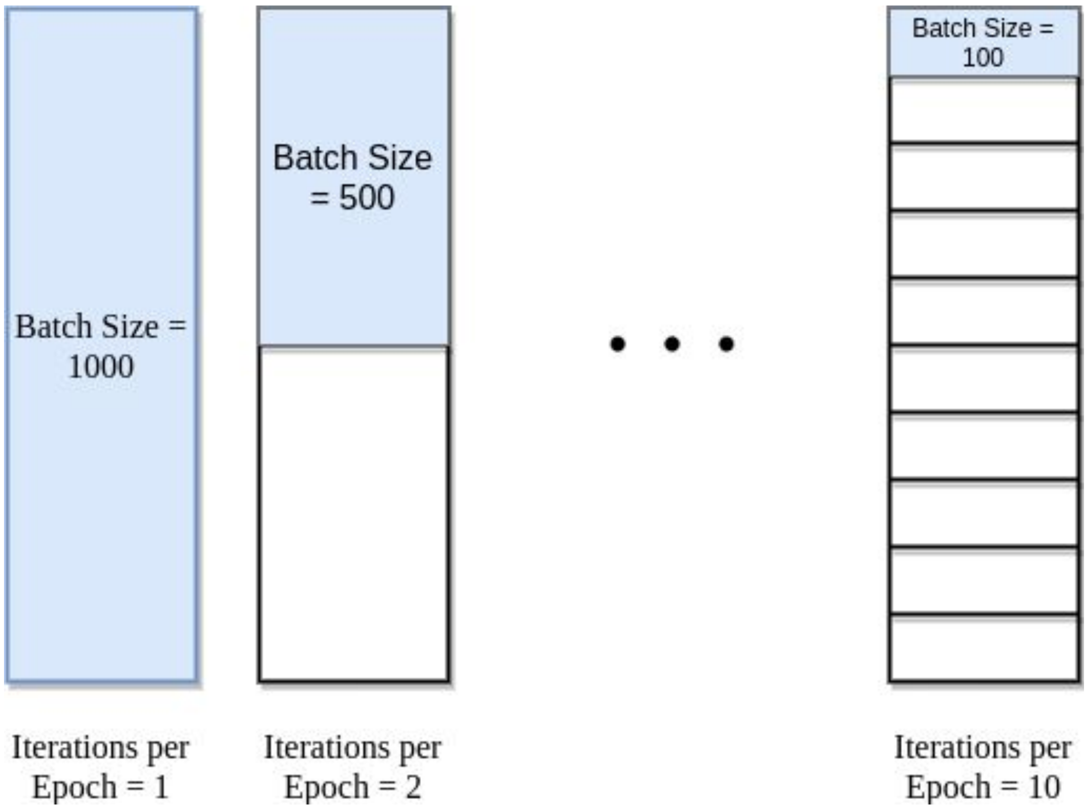
Optimization

10/05/2022

Outline

- Mini-Batch Gradient Descent
- Gradient Descent with Momentum
- RMSProp Optimization
- Adam Optimization

Mini-Batch Gradient Descent



Mini-Batch Gradient Descent

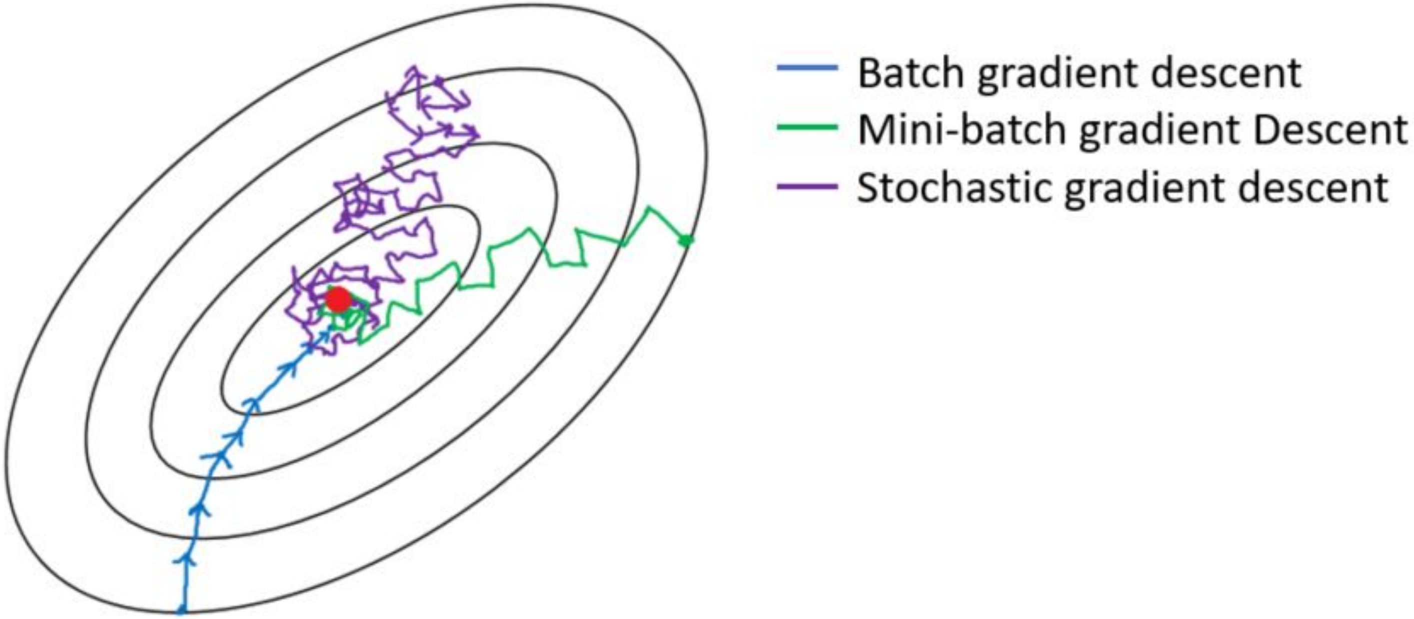
- Batch gradient descent uses all (M) examples in each iteration.
- Stochastic gradient descent (SGD) uses only 1 example in each iteration.
- Batch GD is the slowest, but the most stable. It eats more memory.
- SGD is the fastest, but quite unstable.
- Mini-batch gradient descent is a compromise.

For mini-batch $= 1$ to number of batches

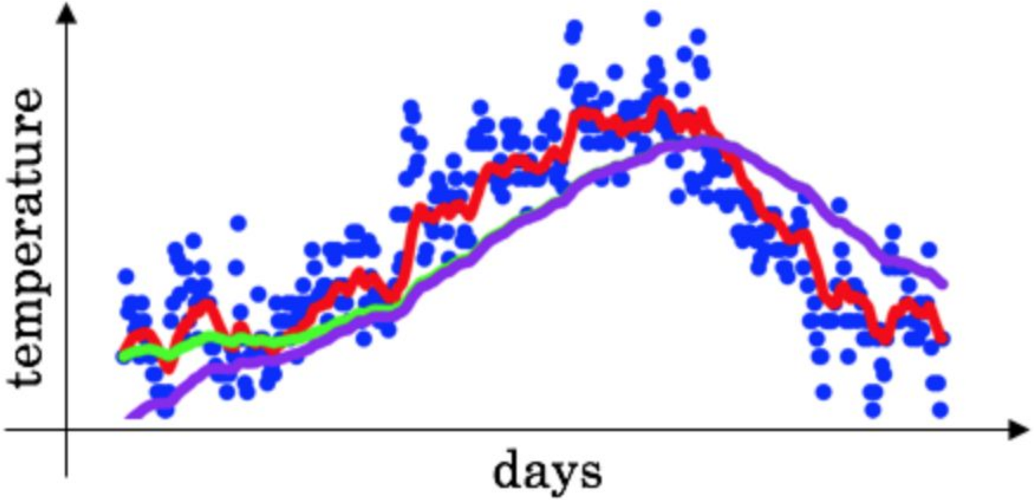
$$\mathbf{W} := \mathbf{W} - \gamma \frac{\partial J}{\partial \mathbf{W}}$$

$$\mathbf{b} := \mathbf{b} - \gamma \frac{\partial J}{\partial \mathbf{b}}$$

Mini-Batch Gradient Descent



Exponentially Weighted Average



$$y_0 = x_0$$

$$y_t = \beta y_{t-1} + (1-\beta) x_t$$

Gradient Descent with Momentum

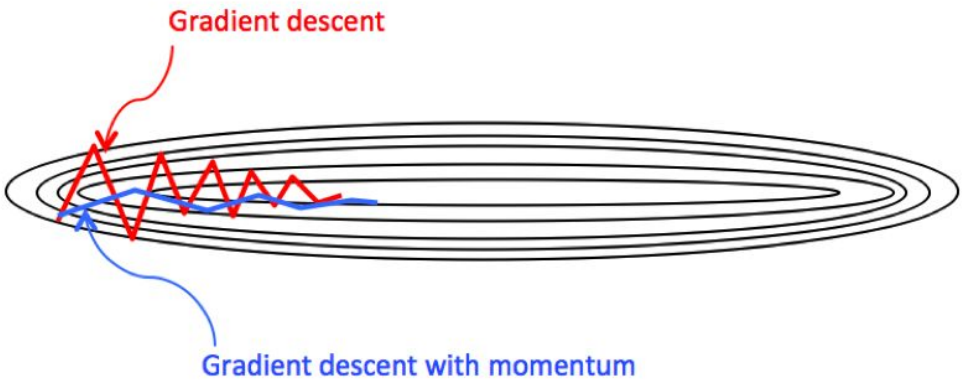
For iteration t = 1 to T

$$m_{t,\mathbf{W}} = \beta m_{t-1,\mathbf{W}} + (1-\beta)\frac{\partial J}{\partial \mathbf{W}}$$

$$m_{t,\mathbf{b}} = \beta m_{t-1,\mathbf{b}} + (1-\beta)\frac{\partial J}{\partial \mathbf{b}}$$

$$\mathbf{W} := \mathbf{W} - \gamma m_{t,\mathbf{W}}$$

$$\mathbf{b} := \mathbf{b} - \gamma m_{t,\mathbf{b}}$$



Root Mean Square Propagation (RMSProp)

For iteration t = 1 to T

$$v_{t,\mathbf{W}} = \beta v_{t-1,\,\mathbf{W}} + (1-\beta)\left(\frac{\partial J}{\partial \mathbf{W}}\right)^2$$

$$v_{t,\mathbf{b}} = \beta v_{t-1,\,\mathbf{b}} + (1-\beta)\left(\frac{\partial J}{\partial \mathbf{b}}\right)^2$$

$$\mathbf{W} := \mathbf{W} - \gamma \frac{\partial J}{\partial \mathbf{W}} \cdot \frac{1}{\sqrt{v_{t,\mathbf{W}}} + \epsilon}$$

$$\mathbf{b} := \mathbf{b} - \gamma \frac{\partial J}{\partial \mathbf{b}} \frac{1}{\sqrt{v_{t,\mathbf{b}}} + \epsilon}$$

Adam Optimization

Initialize: $m_0 = 0,\ v_0 = 0,\ \theta_0$

$\theta = (\mathbf{W}, \mathbf{b})$

For iteration t = 1 to T

$$g_t = \nabla_\theta \mathcal{L}(\theta_{\mathrm{t}-1})$$

$$m_t = \beta_1 \cdot m_{\mathrm{t}-1} + (1 - \beta_1) \cdot g_{\mathrm{t}}$$

Momentum

$$v_t = \beta_2 \cdot v_{\mathrm{t}-1} + (1 - \beta_2) \cdot g_{\mathrm{t}}^2$$

RMSProp

$$\hat{m_t} = \frac{m_t}{(1 - \beta_1^t)}$$

Bias correction

$$\hat{v_t} = \frac{v_t}{(1 - \beta_2^t)}$$

Bias correction

$$\theta_{\mathrm{t}} := \theta_{\mathrm{t}-1} - \gamma \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

Mini-batch gradient descent

Optimizations