Medium Article 8

Kunj Mehta

20 October 2021

1 Gradient Descent

Randomly initialize parameters $\overline{\theta}$

Iterate over t from $t = 1$ to $t = t_{max}$

$$X_{t+1} = X_t - \alpha_t \nabla f(\theta)$$

where $\nabla f(\theta) = \begin{bmatrix} \frac{\partial f(\theta)}{\partial w_1} & \frac{\partial f(\theta)}{\partial w_2} & \dots \frac{\partial f(\theta)}{\partial w_n} \end{bmatrix}$

2 Minimization

Iterative minimization over t steps

$\theta_{t+1} = \theta_t + d_t$

where d_t is the direction of a step towards minimization

3 n-dim Taylor Series

Using n-dimensional Taylor Series for $f(\theta_{t+1})$ centered at $f(\theta_t)$

$$f(\theta_{t+1}) = f(\theta_t) + [\nabla f(\theta_t)]^T(\theta_{t+1} - \theta_t)$$
$$+ \tfrac{1}{2}(\theta_{t+1} - \theta_t)^T H(\theta_t)(\theta_{t+1} - \theta_t)$$
$$+ O(||(\theta_{t+1} - \theta_t)||^3)$$

where
$\nabla f(\theta_t)$ corresponds to the first order derivative

$(\theta_{t+1} - \theta_t)^T H(\theta_t)(\theta_{t+1} - \theta_t)$ the general matrix representation of the quadratic form corresponds to the second derivative

$O(||(\theta_{t+1} - \theta_t)||^3)$ corresponds to all higher order terms

$$H(\theta_t) = \begin{bmatrix} \frac{\partial^2 f(\theta_t)}{\partial^2 \theta_1} & \frac{\partial^2 f(\theta_t)}{\partial \theta_1 \theta_2} & \cdots \frac{\partial^2 f(\theta_t)}{\partial \theta_1 \theta_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial^2 f(\theta_t)}{\partial \theta_n \theta_1} & \frac{\partial^2 f(\theta_t)}{\partial \theta_n \theta_2} & \cdots \frac{\partial^2 f(\theta_t)}{\partial^2 \theta_n} \end{bmatrix}$$

4 Solving Taylor Series

Substituting $\theta_{t+1} = \theta_t + d_t$

$$f(\theta_{t+1}) = f(\theta_t) + [\nabla f(\theta_t)]^T d_t$$
$$+ \tfrac{1}{2} d_t^T H(\theta_t) d_t$$
$$+ O(||d_t||^3)$$

$$f(\theta_{t+1}) = f(\theta_t) + [\nabla f(\theta_t)]^T d_t + O(||d_t||^2)$$

Now let,
$d_t = -\nabla f(\theta_t)$

$$f(\theta_{t+1}) - f(\theta_t) = -||\nabla f(\theta_t)||^2 + O(||f(\theta_t)||^2)$$

Now let,
$d_t = -\alpha_t \nabla f(\theta_t)$

$$f(\theta_{t+1}) - f(\theta_t) = -\alpha_t ||\nabla f(\theta_t)||^2 + O(\alpha_t^2 ||f(\theta_t)||^2)$$