Gradient Descent and Adaptive Optimizers Formulas Summary

Raafat Nagy

Introduction

This document presents key formulas for numerical optimization algorithms commonly used in machine learning (ML) and deep learning (DL), including Gradient Descent, Momentum, Nesterov Accelerated Gradient (NAG), Adagrad, RMSProp, and Adam.

Gradient Descent (GD)

Cost function (MSE)

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

Gradient

$$\nabla J(\theta) = \frac{1}{m} X^T (X\theta - y)$$

Parameter update

$$\theta_{(t+1)} = \theta_{(t)} - \alpha \nabla J(\theta_{(t)})$$

Momentum

Momentum update

$$m_{(t)} = \gamma m_{(t-1)} + \alpha \nabla J(\theta_{(t)})$$

Parameter update

$$\theta_{(t+1)} = \theta_{(t)} - m_{(t)}$$

Nesterov Accelerated Gradient (NAG)

Lookahead position

$$\theta_{\text{lookahead}(t)} = \theta_{(t)} - \gamma m_{(t-1)}$$

Momentum update

$$m_{(t)} = \gamma m_{(t-1)} + \alpha \nabla J(\theta_{\text{lookahead}(t)})$$

Parameter update

$$\theta_{(t+1)} = \theta_{(t)} - m_{(t)}$$

Adagrad

Accumulated squared gradients

$$V_{(t)} = V_{(t-1)} + \left(\nabla J(\theta_{(t)})\right)^2$$

Parameter update

$$\theta_{(t+1)} = \theta_{(t)} - \frac{\alpha}{\sqrt{V_{(t)}} + \epsilon} \nabla J(\theta_{(t)})$$

RMSProp

Exponential moving average of squared gradients

$$V_{(t)} = \beta V_{(t-1)} + (1 - \beta) \left(\nabla J(\theta_{(t)})\right)^2$$

Parameter update

$$\theta_{(t+1)} = \theta_{(t)} - \frac{\alpha}{\sqrt{V_{(t)}} + \epsilon} \nabla J(\theta_{(t)})$$

Adam

First moment (mean of gradients)

$$m_{(t)} = \beta_1 m_{(t-1)} + (1 - \beta_1)\nabla J(\theta_{(t)})$$

Second moment (variance of gradients)

$$v_{(t)} = \beta_2 v_{(t-1)} + (1 - \beta_2)\left(\nabla J(\theta_{(t)})\right)^2$$

Bias-corrected first moment

$$\hat{m}_{(t)} = \frac{m_{(t)}}{1 - \beta_1^t}$$

Bias-corrected second moment

$$\hat{v}_{(t)} = \frac{v_{(t)}}{1 - \beta_2^t}$$

Parameter update

$$\theta_{(t+1)} = \theta_{(t)} - \frac{\alpha}{\sqrt{\hat{v}_{(t)}} + \epsilon}\hat{m}_{(t)}$$

Explanation of Symbols

Symbol	Meaning
$\theta_{(t)}$	Parameter vector at iteration t
α	Learning rate
$J(\theta)$	Cost function (e.g., Mean Squared Error)
$\nabla J(\theta_{(t)})$	Gradient at $\theta_{(t)}$
$m_{(t)}$	Momentum / first moment estimate
$v_{(t)}$	Moving average / second moment estimate
$\hat{m}_{(t)}$	Bias-corrected first moment (Adam)
$\hat{v}_{(t)}$	Bias-corrected second moment (Adam)
$V_{(t)}$	Accumulated / moving avg. of squared gradients
γ	Momentum coefficient (e.g., 0.9)
β	Decay rate (RMSProp, Momentum)
β_1	Decay rate for first moment (Adam)
β_2	Decay rate for second moment (Adam)
ϵ	Small constant for numerical stability
m	Number of training examples
X	Input feature matrix
y	Target vector
$h_\theta(x)$	Model predictions
$\theta_{\text{lookahead}(t)}$	Lookahead position (NAG)