1D Gradient Descent (GD) :

1D Scalar function $f(x)$

Turning points: $f'(x) = 0$

$w_{\max}$ if $f''(x) < 0$ and $x_{\min}$ if $f''(x) > 0$

GD : $x_{k+1} = x_k - \eta f'(x_k)$

η : learning rate (step size)

1D Gradient Descent with Momentum

Momentum Accelerated GD:**MAGD**

$v_k = \mu v_{k-1} + \eta f'(x_k)$

$x_{k+1} = x_k - v_k$

MAGD : $x_{k+1} = x_k - \eta f'(x_k) - \mu v_{k-1}$

μ is restricted : $0 < \mu < 1$

Current update depends not only on the current gradient but also gradients from previous updates.

1D Gradient Descent with Momentum

Momentum Accelerated GD: **NAGD**

MAGD

$$v_k = \mu v_{k-1} + \eta f'(x_k)$$

$$x_{k+1} = x_k - v_k$$

$$x_{k+1} = x_k - \mu v_{k-1} - \eta f'(x_k)$$

.

1D Gradient Descent with Momentum

Nesterov Accelerated GD: **NAGD**

$$x_{k+1} = x_k - \mu v_{k-1}$$

$$v_k = \mu v_{k-1} + \eta \nabla f(x_k - \mu v_{k-1})$$

$$x_{k+1} = x_k - v_k$$

$$x_{k+1} = x_k - \eta \nabla f(x_k - \mu v_{k-1}) - \mu v_{k-1}$$

μ is restricted : $0 < \mu < 1$

Current update depends not only on the current gradient but also gradients from previous updates.

Gradient Descent in 2D

Scalar function of a vector: $f\left(\mathbf{x}\right)=f\left(x,y\right)\Rightarrow\mathbf{x}=\left(x,y\right)$

Compute partial derivatives: $\frac{\partial f}{\partial x};\frac{\partial f}{\partial y}$

Gradient Descent in 2 Dimensions

$$\left.\begin{aligned} x_{k+1}&=x_k-\eta\frac{\partial f}{\partial x} \\ y_{k+1}&=y_k-\eta\frac{\partial f}{\partial y}\end{aligned}\right\}\Rightarrow\mathbf{x}_{k+1}=\mathbf{x}_k-\eta\nabla f\left(\mathbf{x}\right)$$

Gradient Descent with Momentum in 2D (MAGD)

$$\mathbf{v}_k=\mu\mathbf{v}_{k-1}+\eta\nabla f\left(\mathbf{x}_k\right)$$

$$\mathbf{x}_{k+1}=\mathbf{x}_k-\mathbf{v}_k$$

with μ restricted $0<\mu<1$

$$\mathbf{x}_{k+1}=\mathbf{x}_k-\eta\nabla f\left(\mathbf{x}_k\right)-\mu\mathbf{v}_{k-1}$$

2D Gradient Descent with Nesterov Momentum (NAGD)

$$\mathbf{x}_{k+1}=\mathbf{x}_k-\mu\mathbf{v}_{k-1}$$

$$\mathbf{v}_k=\mu\mathbf{v}_{k-1}+\eta\nabla f\left(\mathbf{x}_k-\mu\mathbf{v}_{k-1}\right)$$

$$\mathbf{x}_{k+1}=\mathbf{x}_k-\mathbf{v}_k$$

$$\mathbf{x}_{k+1}=\mathbf{x}_k-\eta\nabla f\left(\mathbf{x}_k-\mu\mathbf{v}_{k-1}\right)-\mu\mathbf{v}_{k-1}$$

Momentum in Two Dimensions

At any point in 2D: **Infinite number of slopes** (other that the slopes in x and y directions).

Compute **directional derivative** of $f(\mathbf{x})$ at any point i.e.,

Directional Derivative: $\mathbf{u} \bullet \nabla f(x) = \mathbf{u}^T \nabla f(\mathbf{x}) = |\mathbf{u}| |\nabla f(\mathbf{x})| \cos\theta$

where

$|\bullet|$ denotes magnitudes of the vectors and

θ is the angle between them.

$\mathbf{u} \bullet \nabla f(x)$ is maximised if θ is a maximum.

This is the **projection of the gradient to** a unit vector $\mathbf{u}$ through that point

Once that direction has been decided then the momentum
and Nesterov momentum can be applied along that direction.

Adaptive Gradient Methods

RMSProp (Root Mean Square Propagation)

This is a gradient descent which **tracks the value of the gradient as it changes** and
uses that to modify the step.

$$m_{k+1} = \gamma m_k + (1-\gamma)\left(\nabla f(\mathbf{x})\right)^2$$

$$\mathbf{v} = -\frac{\eta}{\sqrt{m}} \nabla f(\mathbf{x})$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{v}$$

m: running average of the squares of the gradient

γ: decay term

RMSProp with Nesterov Momentum

$$m_{k+1} = \gamma m_k + (1-\gamma)\left(\nabla f(\mathbf{x} + \mu\mathbf{v})\right)^2$$

$$\mathbf{v} = \mu\mathbf{v} - \frac{\eta}{\sqrt{m}} \nabla f(\mathbf{x} + \mu\mathbf{v})$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{v}$$

Adaptive Gradient Methods

Adagrad (Adaptive Subgradient Method)

$$\mathbf{v} = \frac{-\eta}{\sqrt{\sum\left[\nabla f\left(\mathbf{x}\right)_i\right]^2}} \nabla f\left(\mathbf{x}\right)_i$$

$$\mathbf{x} = \mathbf{x} + \mathbf{v}$$

Adam (Adaptive Moment Estimation)

$$\mathbf{m}_{k+1} = \beta_1 \mathbf{m}_k + \left(1 - \beta_1\right) \nabla f\left(\mathbf{x}\right)$$

$$\mathbf{v}_{k+1} = \beta_2 \mathbf{v}_k + \left(1 - \beta_2\right) \left[\nabla f\left(\mathbf{x}\right)\right]^2$$

$$\hat{\mathbf{m}}_{k+1} = \frac{\mathbf{m}_k}{1 - \beta_1^t}$$

$$\hat{\mathbf{v}} = -\frac{\mathbf{b}}{1 - \beta_2^t}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{\eta}{\sqrt{\hat{\mathbf{v}}} + \varepsilon} \hat{\mathbf{m}}$$

Bias Correction terms, m and v
Running averages of
1st and 2nd moments
~Mean and ~Variance