LOVELY PROFESSIONAL UNIVERSITY

INT247
Machine Learning Foundations

Lecture #1.1

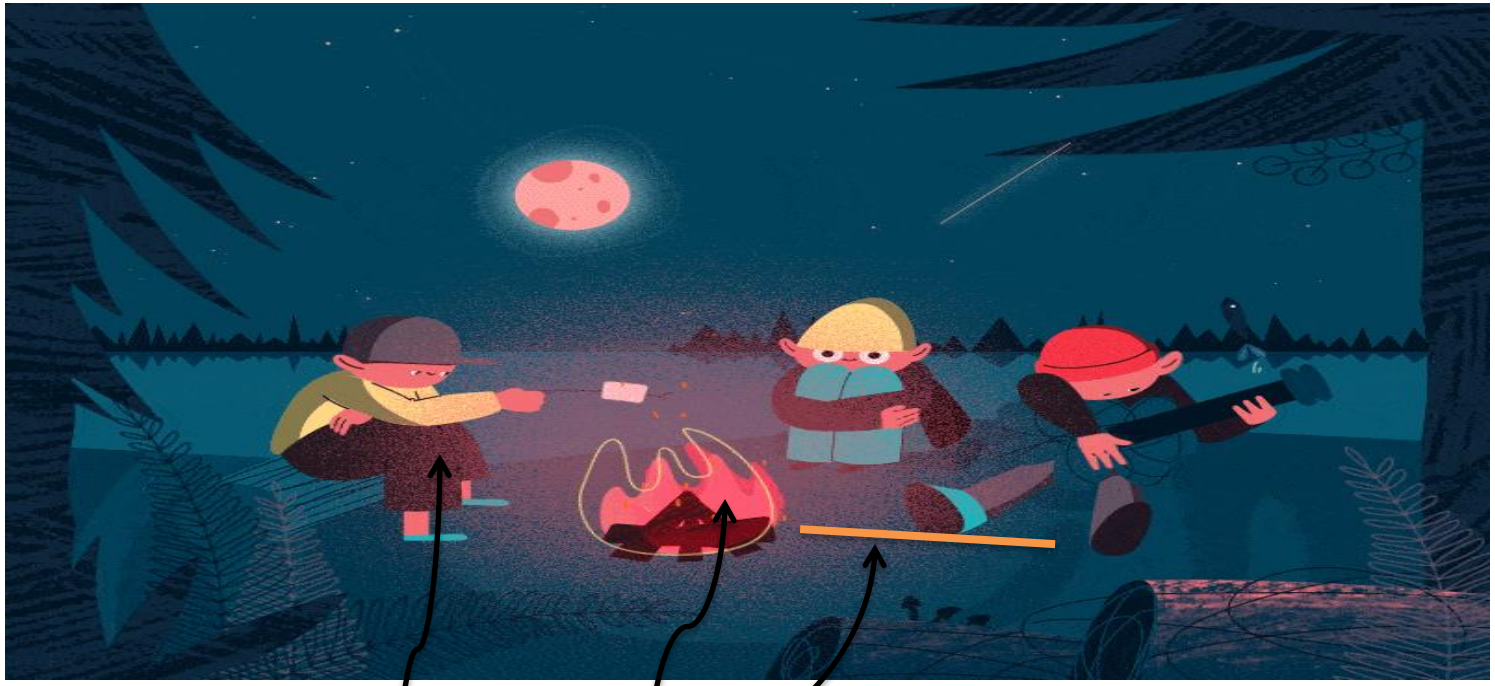
Optimization Techniques

© LPU :: INT247 Machine Learning Foundations :: Usha Mittal & Sanjay Kumar Singh

Cost Function

4 year-old sitting by a fire to keep warm without knowing the danger of fire, puts her finger into it and gets burned.
Next time she sits by the fire, doesn't get burned, but sits too close, gets too hot and has to move away.
The third time she sits by the fire and finds the distance that keeps her warm without exposing her to any danger.

Cost Function



Cost Function ????

Learner ???

Optmization ??

© LPU :: INT247 Machine Learning Foundations :: Usha Mittal & Sanjay Kumar Singh

Cost Function

- **Measure of how wrong the model is in terms of its ability to estimate the relationship between *X* and *Y*.**
- **The objective of machine learning model is**
  - **to find parameters**
  - **weights**
  - **a structure that minimises the cost function.**

Different Types of Optimization Algorithms

- **Gradient Descent**
- **Stochastic Gradient Descent**
- **Mini Batch Gradient Descent**
- **Newton's Method**

Gradient Descent

- **Finds the local or global minima.**
- **Control the variance.**
- **Update the model's parameters.**

Gradient Descent

- **Condition for optimality:**

$\nabla e(w^*)=0$.....(1)

Where ∇ is gradient operator.

- **Gradient vector of the cost function**

$$\nabla e(w)=\left[\frac{\partial e}{\partial w1}, \frac{\partial e}{\partial w2}, \dots\dots, \frac{\partial e}{\partial wm}\right] \text{.....(2)}$$

$$g = \nabla e(w) \dots\dots\dots\dots\dots\dots\dots\dots (3)$$

$$w(n+1) = w(n) - \eta g(n) \text{....(4)}$$

Where η is a positive constant and g(n) is the gradient vector.

$$\Delta w(n) = w(n+1) - w(n) \text{----(5)}$$

$$\Delta w(n) = -\eta g(n) \text{--------(6)}$$

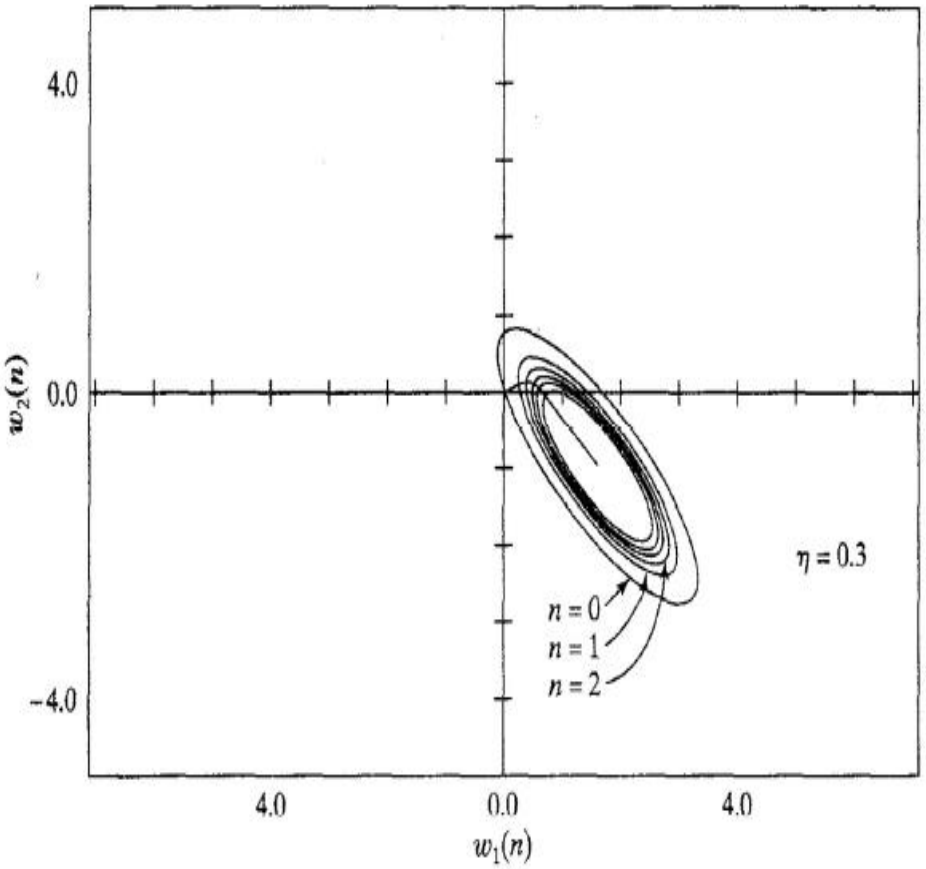
$$e(w(n+1)) \cong e(w(n)) + gT(n)\Delta w(n) \text{-----(7)}$$

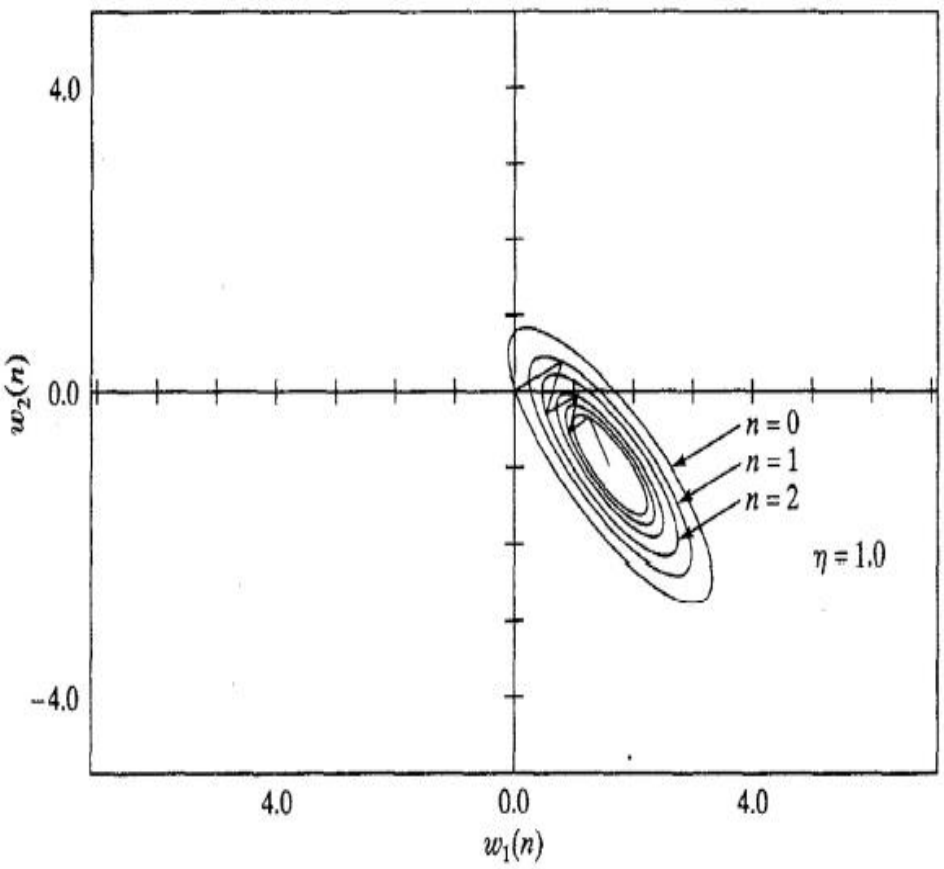
Substitute the value of $\Delta w(n)$ in eq. 7

$$e(w(n+1)) \cong e(w(n)) - \eta g^T(n) g(n) \text{-----(8)}$$

$$e(w(n+1)) \cong e(w(n)) - \eta ||g(n)||^2 \text{--------(9)}$$

Influence of η on convergence

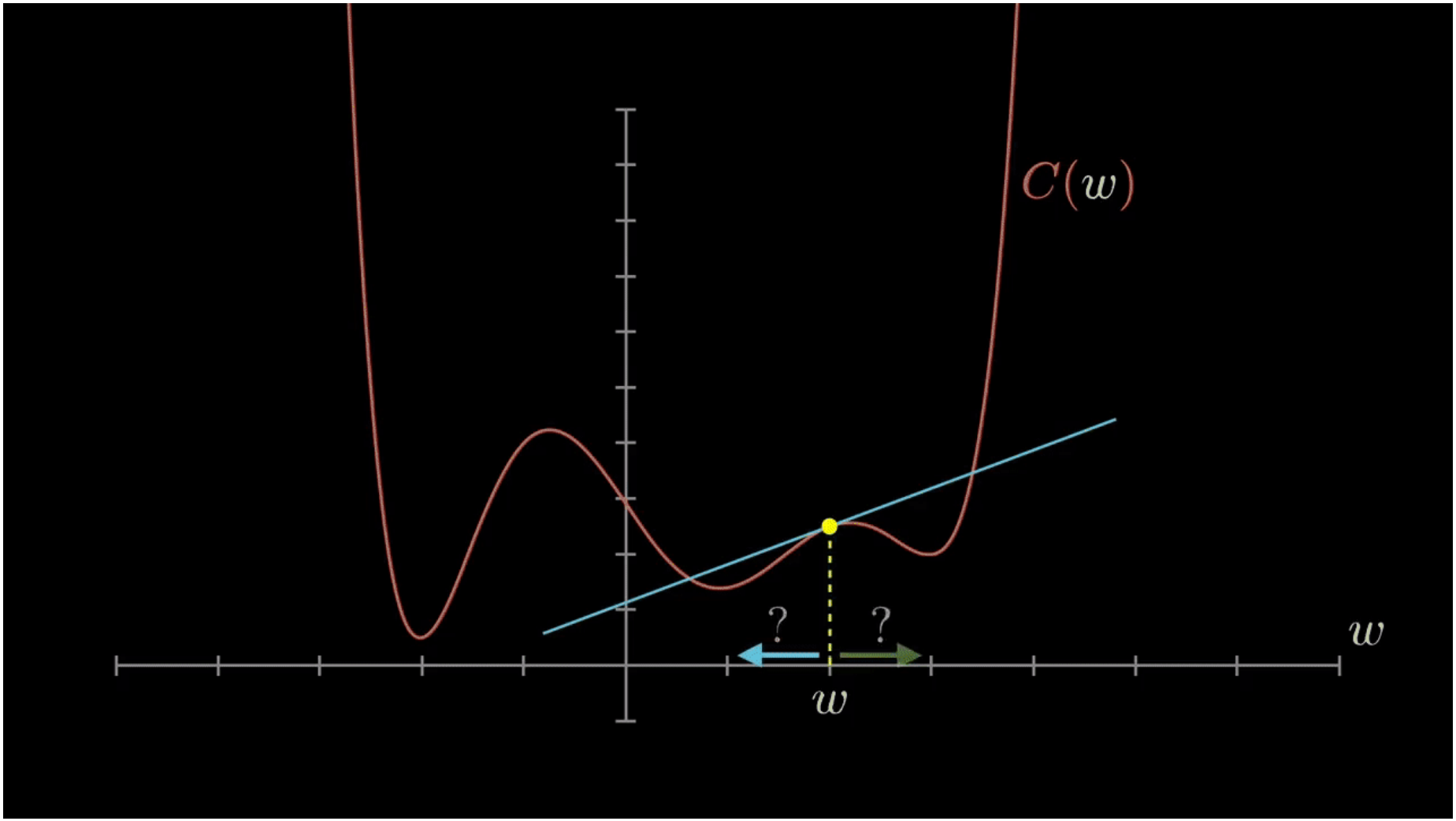




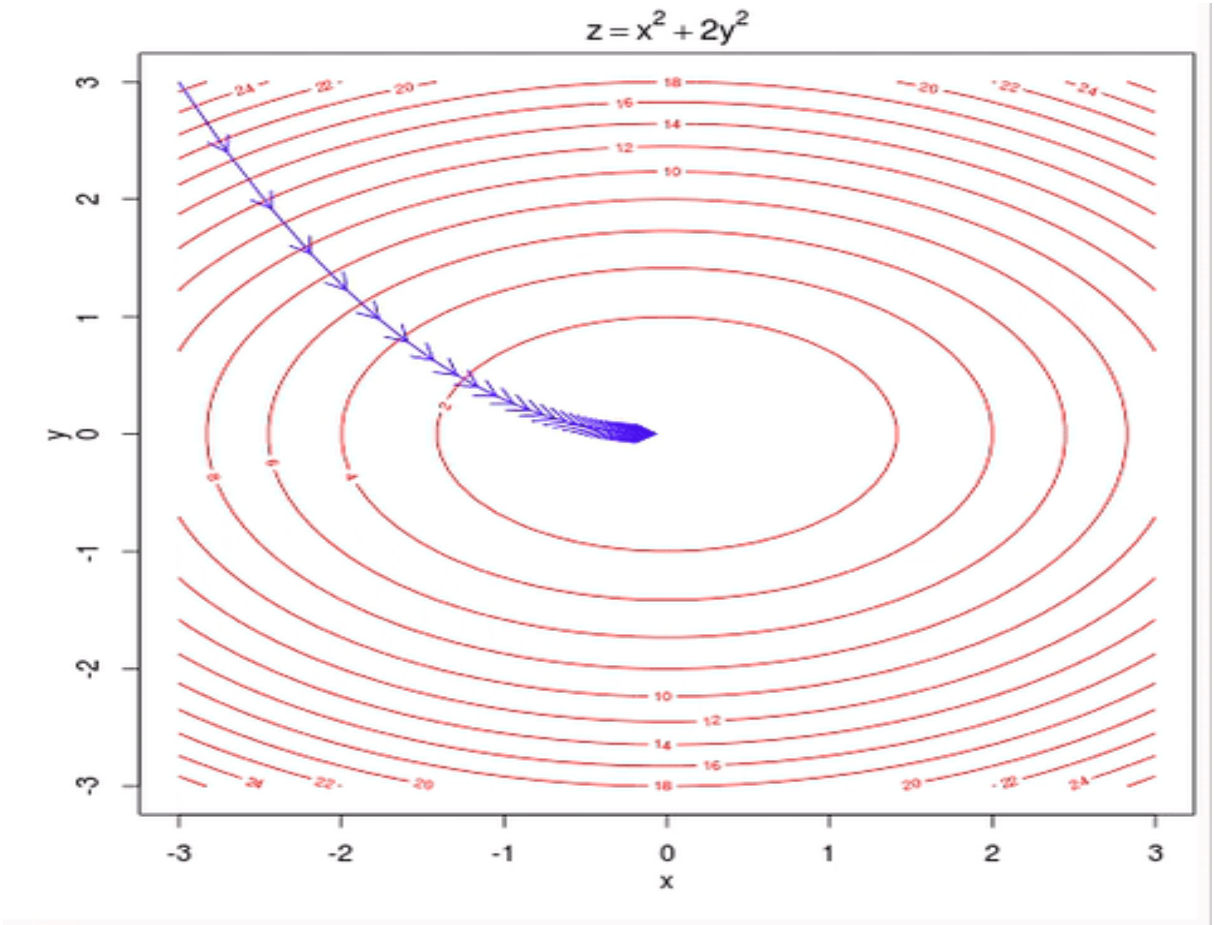
η is small i.e. η=0.3

η is large i.e. η=1.0

Gradient Descent

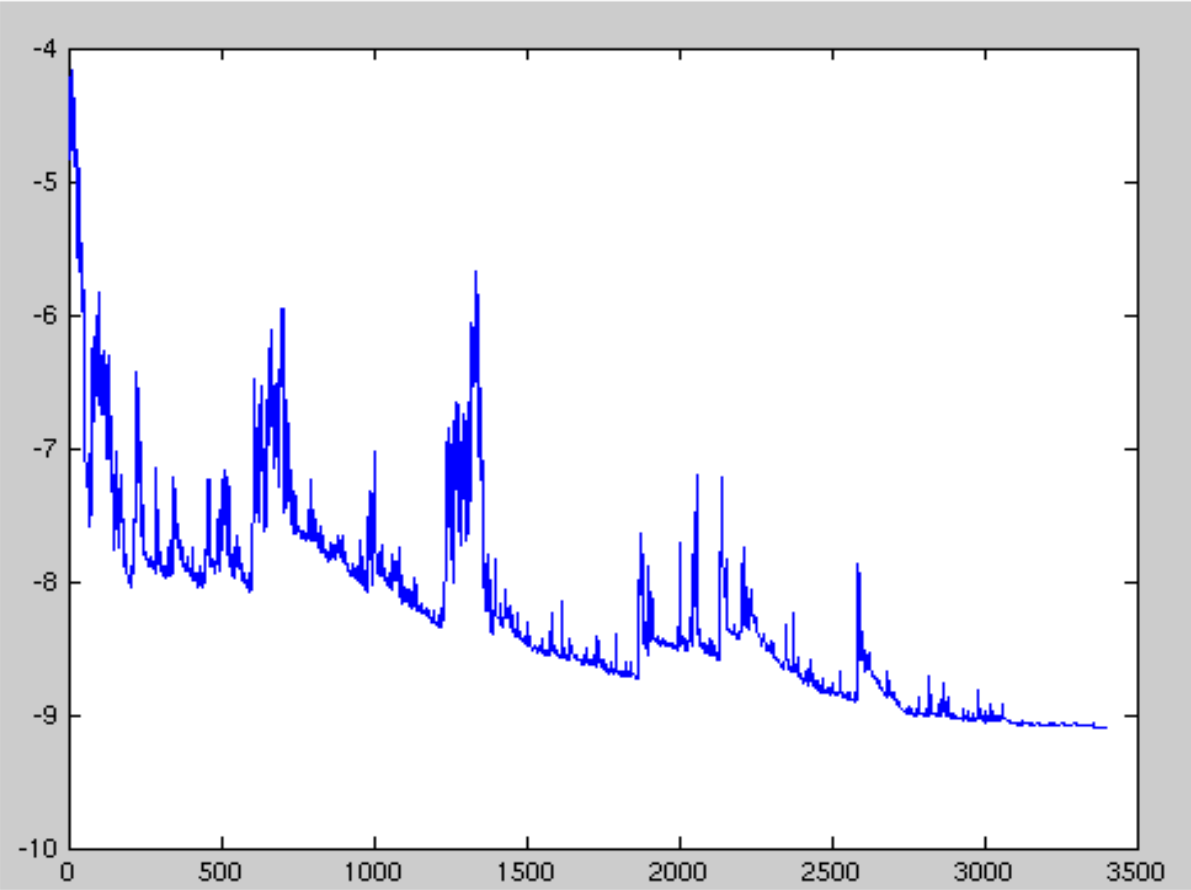


Gradient Descent



Stochastic Gradient Descent

- **Update parameters for each training example.**



© LPU :: INT247 Machine Learning Foundations :: Usha Mittal & Sanjay Kumar Singh

Mini Batch Gradient Descent

- **Update parameters for each mini batch range from 50 to 256.**
- **Reduces the variance in the parameter updates.**

Newton's Method

- **Using second order Taylor series expansion of the cost function.**

$$\Delta w(n) = w(n+1) - w(n) \text{.....(1)}$$

$$\cong gT(n)\Delta w(n) + \frac{1}{2}\Delta wT(n)H(n)\Delta w(n) \text{...(2)}$$

$$H = \nabla^2 e(w) \text{.....(3)}$$

Differentiate eq. 2 w.r.t. Δw

$$g(n) + H(n)\Delta w(n) = 0 \text{.....(4)}$$

$$\Delta w(n) = -H^{-1}(n)g(n) \dots\dots\dots (5)$$

$$w(n+1) = w(n) + \Delta w(n) \text{.....(6)}$$

$$= w(n) - H^{-1}(n)g(n) \text{.....(7)}$$

© LPU :: INT247 Machine Learning Foundations :: Usha Mittal & Sanjay Kumar Singh