Optimization Methods		Date:	*23/03/2018*
Instructor: *Sujit Prakash Gujar*		Scribes:	Amit Kumar Gupta Shubham Kumar Vishnu Vardhan

Lecture 19: Line Search Methods

1 Recap

- Till now we have seen various kinds of methods to minimize a function such as steepest descent, Newton's methods.
- In Neural Network, to minimize the error, our objective function become

  $$\min_{x \epsilon R^n} f(x) = \min_{x \epsilon R^n} \frac{||g(x)||^2}{2}$$

  Where,

  $$||g(x)||^2 = \sum_{i=1}^{m} g^2(x)$$

- To find the best parameter x of a neural network which reduces the resultant error we must keep updating our x in the direction of local minima of the objective function. This is called linear approximation. Using first order Taylor series expansion we approximate g to $\bar{g}$ at x_k i.e. at k^{th} iteration as :

  $$\bar{g}(x, x_k) = g(x_k) + \nabla g(x_k)^T (x - x_k)$$

  Solving further, we get

  $$x_{k+1} = x_k - D_k^{-1} \nabla g(x_k) g(x_k)$$

  Where,

  $$D_k = [\nabla g(x_k) \nabla g(x_k)^T]$$

  If D_k is positive definite, it is of the same form as the gradient descent algorithm. If it is not positive definite, we can make it positive definite by adding some positive values to the diagonal elements.

- If we use D_k of below form, it is called Gauss Newton Method.

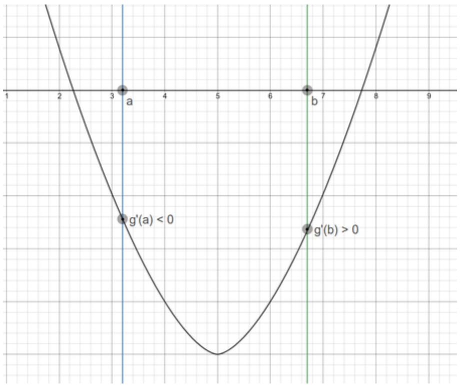
  $$D_k = [\nabla g(x_k) \nabla g(x_k)^T + \Delta_k]$$

2 Line Search Methods

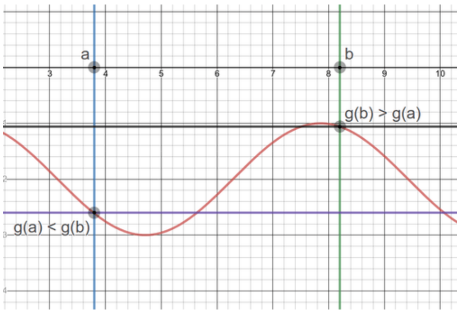
2.1 Cubic Interpolation

- Higher the order, better is the approximation. So, cubic approximation is one of the best possible options for higher order functions. In this method, we try to fit a cubic curve on the points from our given function.
- As a cubic polynomial has 4 unknowns, we will need 4 equations but instead of choosing 4 points we choose 2 consecutive points a, b from the sequence $[0, s, 2s, 3s, ...]$ where $a < b$ such that

(i). $g'(a) \leqslant 0$ and $g'(b) \geqslant 0$



or (ii). $g(b) \geqslant g(a)$



- Cubic interpolation consists of several steps and is iterative in nature.
  - Find an approximate cubic function to fit the given points.
  - Compute the point where the function minima occurs.
  - Update the values of a and b.

We repeat these steps again and again until convergence of actual local minima of the original function $g(x)$.

$$g(\alpha) = A(\alpha)^3 + B(\alpha)^2 + C(\alpha) + D$$

$$g'(\alpha) = 3A(\alpha)^2 + 2B(\alpha) + C$$

So, we know the values of g(a), g(b), g'(a), g'(b). Thus, 4 unknowns and 4 equations, we need to solve these 4 equations.

$$g(a) = A(a)^3 + B(a)^2 + C(a) + D \tag{1}$$

$$g(b) = A(b)^3 + B(b)^2 + C(b) + D \tag{2}$$

$$g'(a) = 3A(a)^2 + 2B(a) + C \tag{3}$$

$$g'(b) = 3A(b)^2 + 2B(b) + C \tag{4}$$

Now after solving equations $1, 2, 3, 4$ and minimizing $g(\alpha)$, we get

$$\bar{\alpha} = b - \frac{g'(b) + W - Z}{g'(b) - g'(a) + ZW} \times (b - a)$$

Where,

$$Z = \frac{3(g(a) - g(b))}{b - a} + g'(a) - g'(b)$$

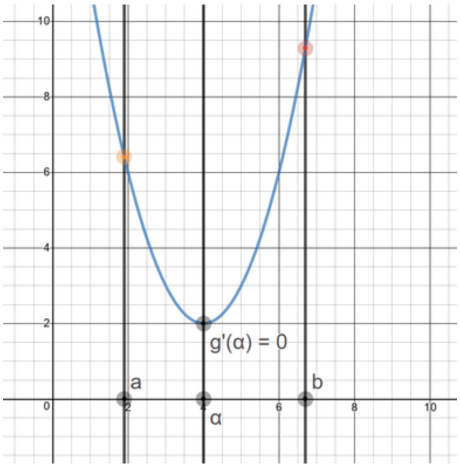
and

$$W = \sqrt{Z^2 - g'(b)g'(a)}$$

Now, we need to update a and b

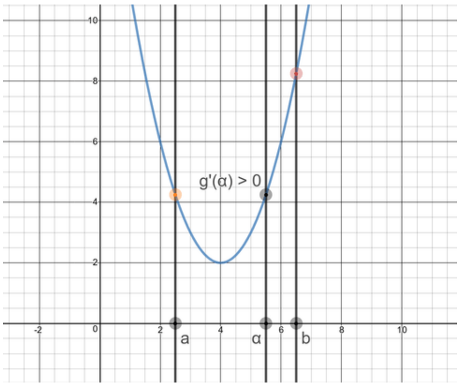
- Case 1 : If $g'(\bar{\alpha}) = 0$, Stop. $(\bar{\alpha})$ is our required (α).
- Case 2 : If $g'(\bar{\alpha}) > 0$, $a = a$ & $b = \bar{\alpha}$.
- Case 3 : If $g'(\bar{\alpha}) < 0$, then
  - If $g(\bar{\alpha}) \leqslant g(a)$, $a = \bar{\alpha}$ & $b = b$.
  - If $g(\bar{\alpha}) > g(a)$, $a = a$ & $b = \bar{\alpha}$.

Case 1 : $g'(\bar{\alpha}) = 0$.



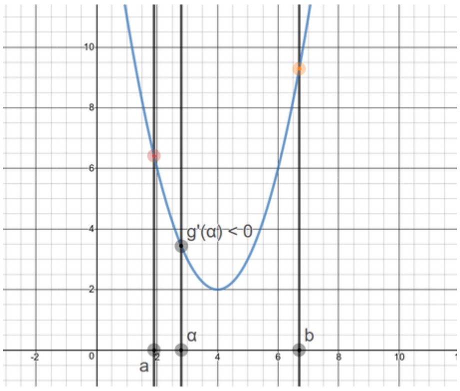
We can see that we got the minimum value, so no need to move further and we stop.

Case 2 : $g'(\bar{\alpha}) > 0$.



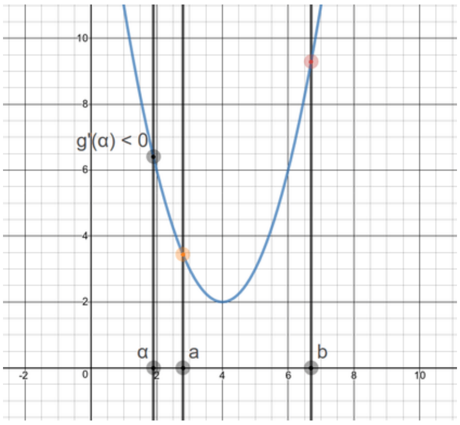
We can see that setting our new values, $a = a$ and $b = \bar{\alpha}$ would satisfy all our conditions.

Case 3.1 : $g'(\bar{\alpha}) < 0$ & $g(\bar{\alpha}) \leqslant g(a)$.



We can see that setting our new values, $a = \bar{\alpha}$ and $b = b$ would satisfy all our conditions.

Case 3.2 : $g'(\bar{\alpha}) < 0$ & $g(\bar{\alpha}) > g(a)$.



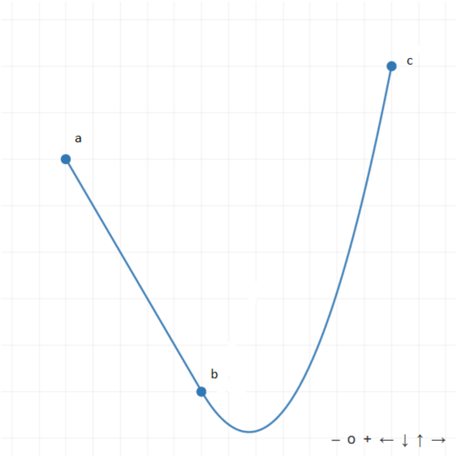
We can see that setting our new values, $a = a$ and $b = \bar{\alpha}$ would satisfy all our conditions.

2.2 Quadratic Interpolation

- A cubic interpolation requires computing derivatives at different points and is expensive in computation. Therefore a quadratic curve can rather be used instead of a cubic polynomial.
- In this method, quadratic curves are fitted to function data to output a sequence of approximations to the actual minimum. Let the quadratic interpolation function be represented by :

$$g(\alpha) = A\alpha^2 + B\alpha + C$$

- Three points can always be interpolated to fit a quadratic curve. Consider three points $a < b < c$ such that $g(a) > g(b)$ and $g(b) < g(c)$, then we are sure that there exists at least one local minima in the interval (a, c).



- Since a quadratic equation has three unknowns, three equations are required to find these unknown parameters namely A, B and C.
- Substitute a, b and c in the quadratic to get the following three equations :

$$g(a) = Aa^2 + Ba + C \tag{1}$$

$$g(b) = Ab^2 + Bb + C \tag{2}$$

$$g(c) = Ac^2 + Bc + C \tag{3}$$

- (2) - (1) gives

$$g(b) - g(a) = A(b^2 - a^2) + B(b - a)$$

$$B = \frac{g(b) - g(a)}{b - a} - A(b + a) \tag{4}$$

  (3) - (2) gives

$$g(c) - g(b) = A(c^2 - b^2) + B(c - b)$$

$$B = \frac{g(c) - g(b)}{c - b} - A(c + b) \tag{5}$$

- On equating (4) and (5) and solving for A

$$A = \frac{(c - b)g(a) + (a - c)g(b) + (b - a)g(c)}{(c - b)(a - c)(b - a)} \tag{6}$$

  Putting this A in (4) and solving for B

$$B = \frac{(c^2 - b^2)g(a) + (a^2 - c^2)g(b) + (b^2 - a^2)g(c)}{(b - c)(a - c)(b - a)} \tag{7}$$

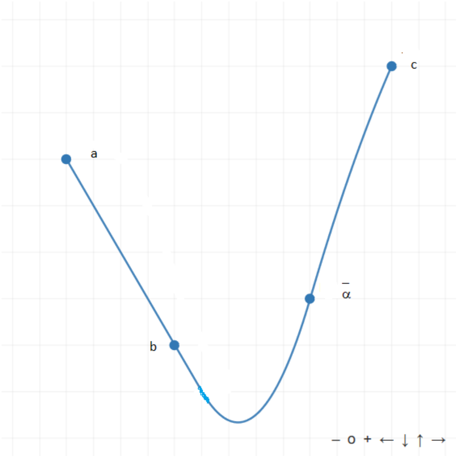
- Let the minimum value of the function $g(\alpha)$ occur at some point $\bar{\alpha}$. We can easily find $\bar{\alpha}$ by differentiating the function and equating it to 0 .

$$\bar{\alpha} = \frac{-B}{2A} \tag{8}$$

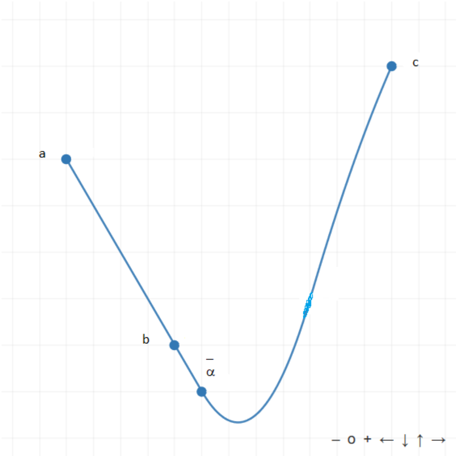
$$\bar{\alpha} = \frac{(c^2 - b^2)g(a) + (a^2 - c^2)g(b) + (b^2 - a^2)g(c)}{2((c - b)g(a) + (a - c)g(b) + (b - a)g(c))} \tag{9}$$

- If this $\bar{\alpha}$ is also a minima for the actual function, then we found a solution and hence output this $\bar{\alpha}$.
- Otherwise we need to update our a, b and c for the next iteration in the following manner until desired precision of interval (a, c) is reached.

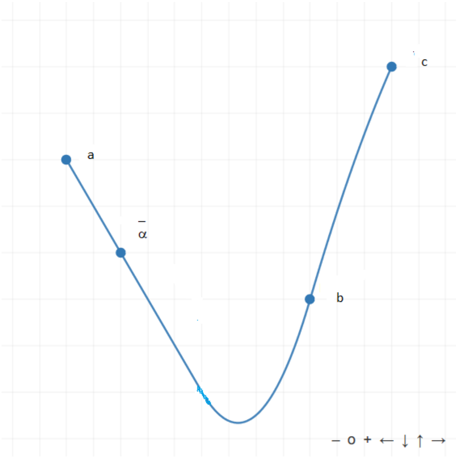
  (a) If $\bar{\alpha} > b$ and $g(\bar{\alpha}) > g(b)$: $a = a, b = b, c = \bar{\alpha}$



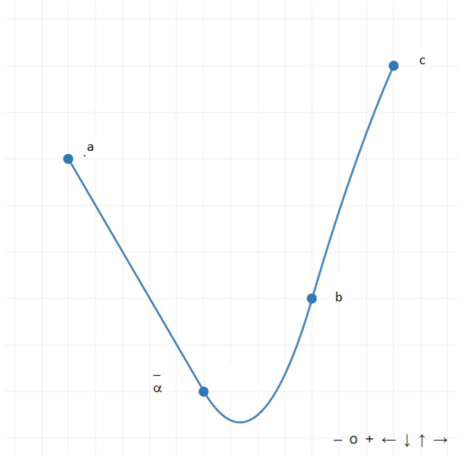
  (b) If $\bar{\alpha} > b$ and $g(\bar{\alpha}) < g(b)$: $a = b, b = \bar{\alpha},\ c = c$



  (c) If $\bar{\alpha} < b$ and $g(\bar{\alpha}) > g(b)$: $a = \bar{\alpha},\ b = b, c = c$



(d) If $\bar{\alpha} < b$ and $g(\bar{\alpha}) < g(b)$: $a = a, b = \bar{\alpha}, \ c = b$



1. Note : In the first iteration, we evaluate the function $g(\alpha)$ at the points $0, s, 2s, 4s, 8s, ...$ where s is some constant value until we get three points a, b, c such that $a < b < c$, $g(a) > g(b)$ and $g(c) > g(b)$.

3 Golden section Method

- Golden Section is a technique to find out the extremum of the strictly unimodal function by successively narrowing the range of values.
- This method maintains the function values for triples of points whose distances form a golden ratio.

To find the extremum for a unimodal function in the range $[a, b]$. In every iteration we select the interval $[a, b - \alpha(b - a)]$ or $[a + \alpha(b - a), b]$ to reduce the search range.

4 Step Size (α_k)

Whatever method we use, we must ensure that certain reduction in function value at α_k.

- *Armijo's rule*: At iteration k, we choose α_k such that $f(x_{k+1}) \leq f(x_k) + c_1\alpha_k\nabla f(x_k)^T d_k$ for some $c_1 > 0$. i.e. reduction is at least $c_1\alpha_k\nabla f(x_k)^T d_k$.
- *Wolfe condition*: Slope should decrease sufficiently.
  $d_k^T\nabla f(x_k + \alpha_k d_k) \geq c_2 d_k^T\nabla f(x_k)$ for some $c_2 > 0$, $0 < c_1 < c_2 < 1$.

Without using interpolation technique, we can use standard technique.
Starting from some initial guess, $s, \beta < 1$, m_k be the 1^{st} non-negative integer such that

$$f(x_k + \beta^{m_k})s) < f(x_k) + c_1\beta^{m_k} s d_k^T\nabla f(x_k)$$

$$\alpha_k = \beta^{m_k} s$$

Consider the following optimization problem:

$$\min_{x\in\mathbb{R}^n} f(x)$$

with $f(x) = \frac{1}{2}x^\intercal Qx - b^\intercal x$, where Q is a real, symmetric, positive definite matrix.

We can solve it analytically as follows:

$$\nabla f(x) = 0$$

$$Qx - b = 0$$

$$x = Q^{-1}b$$

Here, we can find the critical point analytically. However, matrix inversion is computationally unstable and inefficient operation and hence, if possible we prefer to avoid matrix inversion. In this particular case, as Q is symmetric and positive definite, we can use *Cholesky* decomposition.

Cholesky decomposition is unsuitable for larger problem where results needs to obtained very quickly. So, for such problems, we can use *Conjugate Gradient* method which is computationally fast. This method will be covered in the coming class.