Optimization Methods		Date:	*23.03.2018*
Instructor: *Sujit Prakash Gujar*		Scribes:	Adithya Avvaru (20162116) Sahithi Jonnalagadda (201401233)

Lecture 19: Line Search Methods

1 Recap

In last class, we have discussed about application of optimization methods to Neural networks where we formulate the error function as $f(x) = \frac{1}{2}\|g(x)\|^2$ where $g(x) = [g_1(x)\ g_2(x)\ g_3(x)\ ...\ g_m(x)]^T$ and each $g_i(x) = \phi(x, d_i) - y_i$. Here, each d_i represent i^{th} input vector, y_i represent i^{th} output and x represent weight vector of neural network.

Linear approximation for $g(x)$ at $x = x_k$ is

$$\tilde{g}(x, x_k) = g(x_k) + \nabla g(x_k)^T(x - x_k) \tag{1}$$

After solving for $min\ \ f(x)$, we get

$$x_{k+1} = x_k - \alpha_k(\nabla g(x_k)\nabla g(x_k)^T)^{-1}\nabla g(x_k)g(x_k) \tag{2}$$

where α_k is step size.

Considering

$$\begin{aligned} D_k &= (\nabla g(x_k)\nabla g(x_k)^T)^{-1} \\ \nabla f(x) &= \nabla g(x_k)g(x_k) \\ d_k &= -D_k\nabla f(x) \end{aligned}$$

Then,

$$x_{k+1} = x_k + \alpha_k d_k \tag{3}$$

Getting best α_k using $\nabla f(x) = 0$ is difficult as $\nabla f(x) = 0$ has n equations and n unknowns. Let $g : \mathbb{R} \to \mathbb{R}$ such that $g(\alpha_k) = f(x_k + \alpha_k d_k)$. So, it is sufficient to find α_k such that $g'(\alpha_k) = 0$. However, it is difficult to solve $g'(\alpha_k) = 0$ for complex functions $f(x)$.
So, we use line search methods, which we will be discussed in further sections, for solving $g'(\alpha_k) = 0$.

2 Line Search Methods

2.1 Cubic Interpolation

In cubic interpolation, we find an interval $[a, b]$ such that there is at least one local minima. Then, the algorithm tries to fit a cubic polynomial to the four values $g(a)$, $g'(a)$, $g(b)$, $g'(b)$. The following are the steps of the algorithm:

1. $\Big[$**Determination of Initial Interval** $[a, b]\Big]$: To get the points a, b for the initial iteration we evaluate the function $g(\alpha)$ and its derivative $g'(\alpha)$ at points 0, s, 2s, 4s, 8s, ... (s is taken as some fixed scalar), until we get 2 consecutive points a, b such that a, b satisfy either of cases

(1) and (2) below.

$\exists$ at least one $\bar{\alpha} \in (a, b]$ such that $g'(\bar{\alpha}) = 0$ if

$$g'(a) < 0 \; and \; g'(b) \geq 0 \qquad ... \, case \, (1)$$
$$(or)$$
$$g'(a) < 0 \; and \; g(b) \geq g(a) \qquad ... \, case \, (2)$$

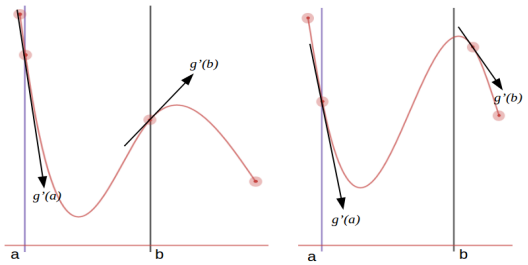

Figure 1: Left to Right : Cases (1) and (2)

2. $\Big[$**Finding $\bar{\alpha}$ in the interval** $[a, b]\Big]$: In this step, we try to fit a cubic polynomial $\bar{g}(\alpha) = A\alpha^3 + B\alpha^2 + C\alpha + D$ between a & b.

   To solve 4 unknowns A, B, C, D in $\bar{g}(\alpha)$, we need 4 equations, which can be obtained by the following equations:

$$\begin{aligned} \bar{g}(a) &= g(a) \\ \bar{g}(b) &= g(b) \\ \bar{g}'(a) &= g'(a) \\ \bar{g}'(b) &= g'(b) \end{aligned}$$

   These 4 equations are solved to uniquely determine $\bar{g}(\alpha)$ and solving for α which minimizes $\bar{g}(\alpha)$ gives a closed form solution of

$$\bar{\alpha} = b - \left\{ \frac{g'(b) - w - z}{g'(b) - g'(a) + 2w} \right\} (b - a) \tag{4}$$

   where,

$$\begin{aligned} z &= \frac{3\big(g(a) - g(b)\big)}{b - a} + g'(a) + g'(b) \\ w &= \sqrt{z^2 - g'(a)g'(b)} \end{aligned}$$

   This $\bar{\alpha}$ is our next guess i.e. a value between the interval $[a, b]$.

3. **[Updating the Current Interval]** :

   (a) If $g'(\bar{\alpha}) = 0$, then we found the required local minimum of $g(\alpha)$ and the value is $\bar{\alpha}$. **Exit**

   (b) If $g'(\bar{\alpha}) > 0$, then $a \leftarrow a$, $b \leftarrow \bar{\alpha}$.

   (c) If $g'(\bar{\alpha}) < 0$ and $g(\bar{\alpha}) < g(a)$, then $a \leftarrow \bar{\alpha}$, $b \leftarrow b$.

   (d) If $g'(\bar{\alpha}) < 0$ and $g(\bar{\alpha}) \geq g(a)$, then $a \leftarrow a$, $b \leftarrow \bar{\alpha}$.

   Geometrical illustration of these cases is shown in Figure 2 except case (a) which is trivial.

4. **[Repeat]** :

   After the interval is updated in step 3, we repeat the steps 2 and 3 till we find $\bar{\alpha}$ such that $g'(\bar{\alpha}) = 0$.

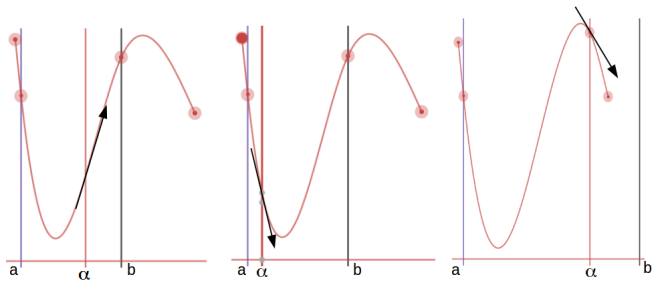

Figure 2: From Left to Right : Cases (b), (c) and (d) of Updating the Current Interval; Black arrow represent $g'(\alpha)$ at $\alpha = \bar{\alpha}$ in all the 3 cases

2.2 Quadratic Interpolation

Though Cubic Interpolation fits well compared to Quadratic Interpolation, finding derivatives in Cubic is costly compared to Quadratic. Due to this drawback, we, sometimes, use Quadratic Interpolation method.

This method requires 3 points $a < b < c$ s.t. $g(a) > g(b)$ and $g(c) > g(b)$. It is very easy to see that a local minima of function $g(\alpha)$ lies between the points a and c.

At each step we try to fit a quadratic polynomial, which is uniquely determined by any three points, to these 3 points. After fitting a quadratic polynomial, we find a minima of this polynomial. Such a minima can be easily found out analytically. Using this minima, we then update our a, b, c values. We keep continuing this process of polynomial fitting, finding a minima and updating the points till the length of the interval (a, c) becomes smaller than some predefined value.

The following are the steps involved in Quadratic Interpolation method:

1. We start with three points a, b, c. To get the points a, b, c for the initial iteration we evaluate the function $g(\alpha)$ at the points $0, s, 2s, 4s, 8s, ...$(s is taken as some fixed constant) until we get three consecutive points a, b, c s.t. $a < b < c$, $g(a) > g(b)$ and $g(c) > g(b)$.

2. Now we fit a quadratic polynomial to these three points.

$$g(\alpha) = A\alpha^2 + B\alpha + C \tag{5}$$

   We need to find values for constants A, B, C using our three points a, b, c. Once these constants are found out, we can easily find the minima of the interpolated polynomial:

$$\bar{\alpha} = -\frac{B}{2A} \tag{6}$$

3. Put a, b, c in $g(\alpha)$ so as to get:

$$g(a) = Aa^2 + Ba + C \tag{7}$$
$$g(b) = Ab^2 + Bb + C \tag{8}$$
$$g(c) = Ac^2 + Bc + C \tag{9}$$

   Eq (8) - Eq (7) $\Rightarrow g(b) - g(a) = A(b^2 - a^2) + B(b - a)$

$$\Rightarrow B = \frac{g(b) - g(a)}{b - a} - A(b + a) \tag{10}$$

   Eq (9) - Eq (8) $\Rightarrow g(c) - g(b) = A(c^2 - b^2) + B(c - b)$

$$\Rightarrow B = \frac{g(c) - g(b)}{c - b} - A(c + b) \tag{11}$$

Solving for A using Eqs (10) and (11) gives:

$$A = \frac{(c-b)g(a) + (a-c)g(b) + (b-a)g(c)}{(c-b)(a-c)(b-a)} \tag{12}$$

Putting this A in any of the Eqs (10) or (11) gives us $-B$ as

$$-B = \frac{(c^2-b^2)g(a) + (a^2-c^2)g(b) + (b^2-a^2)g(c)}{(c-b)(a-c)(b-a)} \tag{13}$$

From Eq (6), the minimum of $g(\alpha)$, $\bar{\alpha}$ is calculated as

$$\bar{\alpha} = \frac{(c^2-b^2)g(a) + (a^2-c^2)g(b) + (b^2-a^2)g(c)}{2((c-b)g(a) + (a-c)g(b) + (b-a)g(c))} \tag{14}$$

4. Now the position of $\bar{\alpha}$ matters. $\bar{\alpha}$ should be minimum of $g(a), g(b), g(c)$. According to this, we have to update a, b, c.
   (a) If $\bar{\alpha} > b$ and $g(\bar{\alpha}) > g(b)$, then $a \leftarrow a, b \leftarrow b, c \leftarrow \bar{\alpha}$
   (b) If $\bar{\alpha} > b$ and $g(\bar{\alpha}) < g(b)$, then $a \leftarrow b, b \leftarrow \bar{\alpha}, c \leftarrow c$
   (c) If $\bar{\alpha} < b$ and $g(\bar{\alpha}) > g(b)$, then $a \leftarrow \bar{\alpha}, b \leftarrow b, c \leftarrow c$
   (d) If $\bar{\alpha} < b$ and $g(\bar{\alpha}) < g(b)$, then $a \leftarrow a, c \leftarrow b, b \leftarrow \bar{\alpha}$
5. Repeat steps 2, 3 and 4 until convergence.

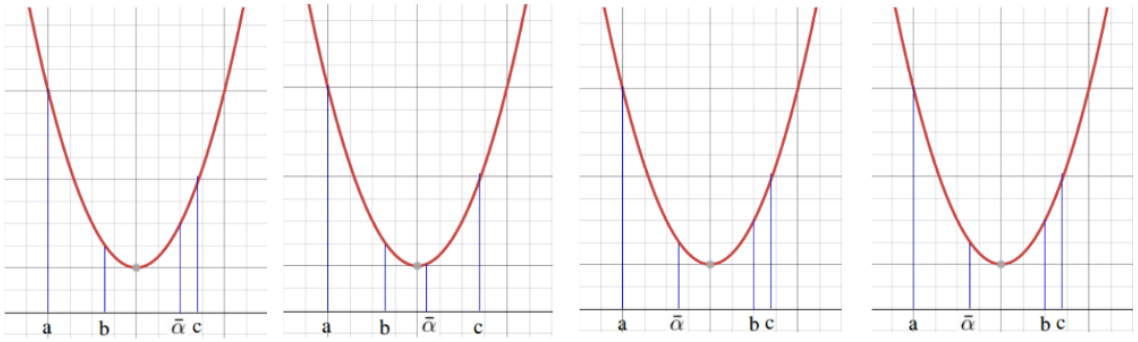

Figure 3: From Left to Right : Cases (a), (b), (c) and (d) of step 4 above

3 Other Methods

Line Search methods, described above, are computationally expensive due to expensive calculation of derivatives of higher order terms. So, we use other methods, based on successive step size α reduction, which follow the rules mentioned in subsections below.

3.1 Armijo's Rule

Armijo's Rule ensures sufficient decrease in value of the function, by finding acceptable step size α at each iteration.
We choose a step size α such that it satisfy

$$g(\alpha) = f(x_k + \alpha d_k) \leq f(x_k) + c_1 \alpha d_k^T \nabla f(x_k) \tag{15}$$

i.e. we ensure the reduction of function value is at least $c_1 \alpha d_k^T \nabla f(x_k)$, which is lower bound on the magnitude of reduction in function $f(x)$ value. Eq. 15 explains that we select a value for α which ensure the graph $g(\alpha)$ lies below the line $f(x_k) + c_1 \alpha d_k^T \nabla f(x_k)$.

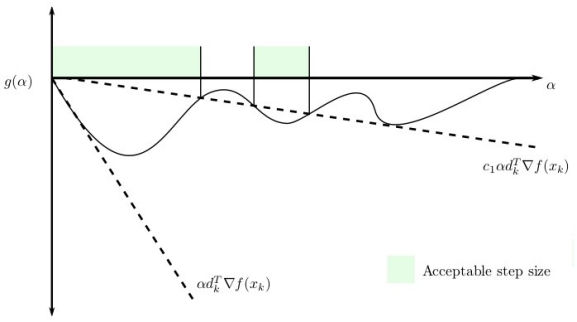

Figure 4: Graphical interpretation of Arjimo's rule

3.1.1 Convergence Using Armijo's Rule

1. Take initial step size $\alpha_0 = $ s, where s is some large number, $k = 0$ and some $\beta \in (0, 1)$.
2. Check if it satisfies Eq. 15.
3. If yes, then return the step size α_k
4. Else,

$$k = k + 1$$
$$\alpha_k = \beta^k s$$

5. Repeat steps 2, 3 and 4.

When the above algorithm converges, $f(x_k + \beta^k s d_k) \leq f(x_k) + c_1 \beta^k s d_k^T \nabla f(x_k)$ as $\alpha_k = \beta^k s$.

3.2 Wolfe or Curvature Conditions

We say that the step size α_k at iteration k is said to satisfy *Wolfe or Curvature Conditions*, if

$$d_k \nabla f(x_k + \alpha_k d_k) \geq c_2 d_k^T \nabla f(x_k) \tag{16}$$

This ensures a sufficient reduction in $\nabla f(x)$. In general, one should ensure that step size satisfies the Armijo's rule, and if possible, the Wolfe conditions. It is useful for problems where it is difficult to find an α_k that minimizes g(α).

3.3 Golden Section Method

In Golden Section Method, we minimize the function $g(\alpha)$ over an interval $[0, s]$ by determining at the k^{th} iteration an interval $[\alpha_k, \bar{\alpha}_k]$ containing the minimum value α_k. The reduction in interval is determined by the number $\mathcal{T}$, which is related to the Fibonacci number sequence, where $\mathcal{T} = \dfrac{3 - \sqrt{5}}{2}$.
For this method, we assume the function $g(\alpha)$ to be **strictly unimodal**[1] in the interval $[0, s]$.

The idea behind Golden Section method is to determine the interval $[\alpha_{k+1}, \bar{\alpha}_{k+1}]$ such that $\alpha^* \in$

[1] A strictly unimodal function g over an interval $[0, s]$ is defined as a function that has a unique global minimum α^* in $[0, s]$ and if α_1, α_2 are two points in $[0, s]$ such that $\alpha_1 < \alpha_2 < \alpha^*$ or $\alpha_1 > \alpha_2 > \alpha^*$, then $g(\alpha_1) > g(\alpha_2) > g(\alpha^*)$ or $g(\alpha^*) < g(\alpha_1) < g(\alpha_2)$, respectively.

$[\alpha_{k+1}, \bar{\alpha}_{k+1}]$ from interval $[\alpha_k, \bar{\alpha}_k]$, starting initially from $[\alpha_0, \bar{\alpha}_0] = [0, s]$.

The steps to follow are as follows:

1. Determine initial interval as $[\alpha_0, \bar{\alpha}_0] = [0, s]$ and repeat the following steps till the value $(\bar{\alpha}_k - \alpha_k)$ become smaller than the tolerance level.
2. Calculate b_k, $\bar{b}_k$, $g(b_k)$ and $g(\bar{b}_k)$, where

$$b_k \leftarrow \alpha_k + \mathcal{T}(\bar{\alpha}_k - \alpha_k)$$
$$\bar{b}_k \leftarrow \bar{\alpha}_k - \mathcal{T}(\bar{\alpha}_k - \alpha_k)$$

   Later, we find the interval based on the update rules explained in below steps.
3. If $g(b_k) < g(\bar{b}_k)$, then
   (a) If $g(\alpha_k) \leq g(b_k)$, then $\alpha_{k+1} \leftarrow \alpha_k, \bar{\alpha}_{k+1} \leftarrow b_k$
   (b) If $g(\alpha_k) > g(b_k)$, then $\alpha_{k+1} \leftarrow \alpha_k, \bar{\alpha}_{k+1} \leftarrow \bar{b}_k$
4. If $g(b_k) > g(\bar{b}_k)$, then
   (a) If $g(\bar{b}_k) \geq g(\bar{\alpha}_k)$, then $\alpha_{k+1} \leftarrow \bar{b}_k, \bar{\alpha}_{k+1} \leftarrow \bar{\alpha}_k$
   (b) If $g(\bar{b}_k) < g(\alpha_k)$, then $\alpha_{k+1} \leftarrow b_k, \bar{\alpha}_{k+1} \leftarrow \bar{\alpha}_k$
5. If $g(b_k) = g(\bar{b}_k)$, then $\alpha_{k+1} \leftarrow b_k, \bar{\alpha}_{k+1} \leftarrow \bar{b}_k$

4 Example

Given $f(x, y) = 2x^2 + y^2$, $X_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $s = 0.5$, $\beta = 0.7$ and $c_1 = \frac{1}{4}$, Find α using Armijo's Rule.

Solution:

$$d_k = \nabla f(X) = \begin{bmatrix} 4x \\ 2y \end{bmatrix} \tag{17}$$
$$\alpha_k = \beta^k s \tag{18}$$
$$X_{k+1} = X_k - \alpha_k d_k \tag{19}$$

$$d_{k(X=X_0)} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$
$$d_k^T \nabla f(X_0) = 20$$
$$f(X_0) = 3$$

At $k = 0$**:**
$\alpha_0 = \beta^0 s = 0.5$ $\quad X_1 = X_0 - \alpha_0 d_{k(X=X_0)} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$ (using Eq. 19) $\quad f(X_1) = 2$
Using Eq 15, consider

$$f(x_1) \leq f(x_0) + c_1 \alpha_0 (-d_k \nabla f(X_0)) \quad [Negative\ sign\ to\ indicate\ slope\ as\ negative]$$
$$2 \leq 3 + \frac{1}{4} * 0.5 * (-20)$$
$$2 \leq 0.5 \quad \textbf{[False]}$$

So, we can't choose α as $\alpha_0 = 0.5$. Increment k.

At $k = 1$:
$\alpha_1 = \beta^1 s = 0.35 \quad X_1 = X_0 - \alpha_1 d_{k(X=X_0)} = \begin{bmatrix} -0.4 \\ 0.3 \end{bmatrix}$ (using Eq. 19) $\quad f(X_1) = 0.41$
Using Eq 15, consider

$$f(x_1) \leq f(x_0) + c_1\alpha_1(-d_k \nabla f(X_0)) \quad [Negative\ sign\ to\ indicate\ slope\ as\ negative]$$
$$0.41 \leq 3 + \frac{1}{4} * 0.35 * (-20)$$
$$0.41 \leq 1.25 \quad \textbf{[True]}$$

So, we stop here and choose α as $\alpha_1 = 0.35$.

5 Introduction to Conjugate Gradient Method

Consider the following optimization problem:

$$min\ f(x) = \frac{1}{2}x^T Q x - x^T b$$

where Q is positive definite and symmetric.

Analytically Solving: $\nabla f(x) = Qx - b = 0 \Rightarrow x = Q^{-1}b$, where x is the optimal solution.

We do not prefer the analytical solution, as computing inverse of Q is an unstable operation and due to floating point operations, precision is often lost. Such problems can be solved efficiently using conjugate[2] gradient methods, which will be discussed in next class.

[2]The vectors u and v are conjugate to each other, with respect to a real, symmetric matrix A, if $u^T Av = 0$.