# OPTIMIZATION (SI 416) – LECTURE 7

**Harsha Hutridurga**

IIT Bombay

## RECAP (SUFFICIENCY)

- Take a twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$
- If $x_* \in \mathbb{R}^n$ is such that

$$\nabla f(x_*) = 0 \qquad \text{and} \qquad \nabla^2 f(x_*) \quad \text{is positive definite}$$

  then

  - $x_*$ is a strict local minimizer of $f$, i.e.

$$f(x_*) < f(x) \qquad \text{for all } x \in B_r(x_*)$$

    for some $r > 0$

## RECAP (NEWTON'S ALGORITHM)

- The iterates in Newton's algorithm are given by

$$x^{(n+1)} = x^{(n)} - \left(\nabla^2 f(x^{(n)})\right)^{-1} \nabla f(x^{(n)})$$

- For the algorithm to be well-defined, Hessian matrix needs to be invertible at $x^{(n)}$
- Hessian matrix $\nabla^2 f$ is said to be Lipschitz continuous if

$$\left\|\nabla^2 f(x)v - \nabla^2 f(y)v\right\| \leq \beta \left\|x - y\right\| \left\|v\right\| \quad \forall x, y, v \in \mathbb{R}^n$$

for some $\beta > 0$

## Theorem

*Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function. Let $x_* \in \mathbb{R}^n$ be such that*

$$\nabla f(x_*) = 0 \qquad \text{and} \qquad \nabla^2 f(x_*) \quad \text{is positive definite.}$$

*Suppose that $\nabla^2 f$ is Lipschitz continuous in a neighbourhood of $x_*$. Then, the iterates built by the Newton's algorithm satisfy:*

- ♣ *if the starting point $x^{(0)}$ is sufficiently close to $x_*$, then the sequence of iterates converges to $x_*$*
- ♣ *the rate of convergence of $\{x^{(n)}\}$ is quadratic*
- ♣ *the sequence of gradient norms $\{\|\nabla f(x^{(n)})\|\}$ converges quadratically to zero*

♣ Note that we have

$$\begin{aligned} x^{(n+1)} - x_* = x^{(n)} - x_* - \left(\nabla^2 f(x^{(n)})\right)^{-1} \nabla f(x^{(n)}) \\ = \left(\nabla^2 f(x^{(n)})\right)^{-1} \nabla^2 f(x^{(n)}) \left(x^{(n)} - x_*\right) \\ - \left(\nabla^2 f(x^{(n)})\right)^{-1} \left(\nabla f(x^{(n)}) - \nabla f(x_*)\right) \end{aligned}$$

thanks to $\nabla f(x_*) = 0$

♣ Observe that (thanks to fundamental theorem of calculus)

$$\nabla f(x_*) - \nabla f(x^{(n)}) = \int_0^1 \nabla^2 f(x^{(n)} + \alpha(x_* - x^{(n)}))(x_* - x^{(n)})\, \mathrm{d}\alpha$$

♣ Hence we have

$$\nabla^2 f(x^{(n)}) \left(x^{(n)} - x_*\right) - \left(\nabla f(x^{(n)}) - \nabla f(x_*)\right)$$
$$= \int_0^1 \left(\nabla^2 f(x^{(n)}) - \nabla^2 f(x^{(n)} + \alpha(x_* - x^{(n)}))\right) \left(x^{(n)} - x_*\right) \,\mathrm{d}\alpha$$

♣ Thus we deduce

$$\left\|\nabla^2 f(x^{(n)}) \left(x^{(n)} - x_*\right) - \left(\nabla f(x^{(n)}) - \nabla f(x_*)\right)\right\|$$
$$\leq \int_0^1 \left\|\nabla^2 f(x^{(n)}) - \nabla^2 f(x^{(n)} + \alpha(x_* - x^{(n)}))\right\| \left\|x^{(n)} - x_*\right\| \,\mathrm{d}\alpha$$
$$\leq \left\|x^{(n)} - x_*\right\|^2 \beta \int_0^1 \alpha \,\mathrm{d}\alpha = \frac{\beta}{2} \left\|x^{(n)} - x_*\right\|^2$$

- Recall that we had

$$x^{(n+1)} - x_* = \left(\nabla^2 f(x^{(n)})\right)^{-1} \left(\nabla^2 f(x^{(n)}) \left(x^{(n)} - x_*\right) - \left(\nabla f(x^{(n)}) - \nabla f(x_*)\right)\right)$$

- Therefore it follows that

$$\left\|x^{(n+1)} - x_*\right\| \leq \left\|\left(\nabla^2 f(x^{(n)})\right)^{-1}\right\| \left\|\nabla^2 f(x^{(n)}) \left(x^{(n)} - x_*\right) - \left(\nabla f(x^{(n)}) - \nabla f(x_*)\right)\right\| \leq \frac{\beta}{2} \left\|\left(\nabla^2 f(x^{(n)})\right)^{-1}\right\| \left\|x^{(n)} - x_*\right\|^2$$

- It is given that $\nabla^2 f(x_*)$ is positive definite
- Hence is invertible
- As $f$ is twice continuously differentiable, $\exists\, r > 0$ such that

$$\left\|\left(\nabla^2 f(x)\right)^{-1}\right\| \leq 2\left\|\left(\nabla^2 f(x_*)\right)^{-1}\right\| \qquad \text{for all } x \in B_r(x_*)$$

- Therefore it follows that

$$\left\|x^{(n+1)} - x_*\right\| \leq \beta\left\|\left(\nabla^2 f(x_*)\right)^{-1}\right\| \left\|x^{(n)} - x_*\right\|^2$$

- From the above recursive relation, we deduce that

$$\left\|x^{(n)} - x_*\right\| \leq \beta^{2^n-1}\left\|\left(\nabla^2 f(x_*)\right)^{-1}\right\|^{2^n-1} \left\|x^{(0)} - x_*\right\|^{2^n}$$

♣ Suppose we choose the starting point $x^{(0)}$ such that

$$\left\| x^{(0)} - x_* \right\| \leq \min \left\{ r, \frac{1}{2\beta \left\| \left(\nabla^2 f(x_*)\right)^{-1} \right\|} \right\}$$

♣ Hence we deduce that

$$\left\| x^{(n)} - x_* \right\| \leq \frac{2^{-2^n}}{\beta \left\| \left(\nabla^2 f(x_*)\right)^{-1} \right\|} \qquad \text{for } n = 1, 2, \ldots$$

♣ This helps us conclude that

$$\lim_{n \to \infty} x^{(n)} = x_*$$

and that the convergence is quadratic.

♣ Recall that we have

$$x^{(n+1)} = x^{(n)} + p^{(n)} \quad \text{and} \quad p^{(n)} = -\left(\nabla^2 f(x^{(n)})\right)^{-1} \nabla f(x^{(n)})$$

i.e.

$$\nabla^2 f(x^{(n)}) p^{(n)} + \nabla f(x^{(n)}) = 0$$

♣ Note that

$$\begin{aligned}
\nabla f(x^{(n+1)}) &= \nabla f(x^{(n+1)}) - \nabla f(x^{(n)}) - \nabla^2 f(x^{(n)}) p^{(n)} \\
&= \int_0^1 \nabla^2 f(x^{(n)} + \alpha p^{(n)}) \left(x^{(n+1)} - x^{(n)}\right) \mathrm{d}\alpha - \nabla^2 f(x^{(n)}) p^{(n)} \\
&= \int_0^1 \left(\nabla^2 f(x^{(n)} + \alpha p^{(n)}) - \nabla^2 f(x^{(n)})\right) p^{(n)} \, \mathrm{d}\alpha
\end{aligned}$$

♣ From the earlier equality, we deduce

$$\left\|\nabla f(x^{(n+1)})\right\| \le \int_0^1 \left\|\nabla^2 f(x^{(n)} + \alpha p^{(n)}) - \nabla^2 f(x^{(n)})\right\| \left\|p^{(n)}\right\| \,\mathrm{d}\alpha$$

$$\le \beta \left\|p^{(n)}\right\|^2 \int_0^1 \alpha \,\mathrm{d}\alpha = \frac{\beta}{2}\left\|p^{(n)}\right\|^2$$

♣ Using the definition of $p^{(n)}$, we deduce

$$\left\|\nabla f(x^{(n+1)})\right\| \le \frac{\beta}{2}\left\|\left(\nabla^2 f(x^{(n)})\right)^{-1}\right\|^2 \left\|\nabla f(x^{(n)})\right\|^2$$

$$\le \beta \left\|\left(\nabla^2 f(x_*)\right)^{-1}\right\|^2 \left\|\nabla f(x^{(n)})\right\|^2$$

thanks to $\nabla^2 f(x_*)$ being invertible.

♣ From the above recursive relation, we deduce

$$\left\|\nabla f(x^{(n)})\right\| \le \beta^{2^n-1}\left\|\left(\nabla^2 f(x_*)\right)^{-1}\right\|^{2^{n+1}-2} \left\|\nabla f(x^{(0)})\right\|^{2^n}$$

- Suppose we choose the starting point $x^{(0)}$ such that

$$\left\|\nabla f(x^{(0)})\right\| \leq \frac{1}{2\beta \left\|\left(\nabla^2 f(x_*)\right)^{-1}\right\|^2}$$

- Then we deduce that

$$\left\|\nabla f(x^{(n)})\right\| \leq \frac{2^{-2^n}}{\beta \left\|\left(\nabla^2 f(x_*)\right)^{-1}\right\|^2} \qquad \text{for } n = 1, 2, \ldots$$

- This helps us conclude that

$$\lim_{n\to\infty} \left\|\nabla f(x^{(n)})\right\| = 0$$

  and that the convergence is quadratic.
- Note that the above assumption on $x^{(0)}$ is not impractical as $\nabla f(x_*) = 0$ and $f$ is smooth. So, for points close to $x_*$, the above condition is satisfied

- Take a twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$
- Begin with a starting point $x^{(0)}$
- Pick a matrix $B^{(0)}$ and build the next iterate $x^{(1)}$ as follows

$$x^{(1)} = x^{(0)} - \alpha_0 \left(B^{(0)}\right)^{-1} \nabla f(x^{(0)})$$

  where $\alpha_0$ is the step length
- General recipe in building the iterates is as follows:

$$x^{(k+1)} = x^{(k)} - \alpha_k \left(B^{(k)}\right)^{-1} \nabla f(x^{(k)}) \qquad \text{for } k = 0, 1, 2, \ldots$$

- In the classical Newton's method, one takes

$$B^{(k)} = \nabla^2 f(x^{(k)})$$

- Rather than computing the hessian, the idea is to build $B^{(k)}$ which approximate the hessian $\nabla^2 f(x^{(k)})$

- What properties should $B^{(k)}$ have?
- Take a quadratic approximation of $f$ around $x^{(k)}$

$$m_k(p) = f(x^{(k)}) + \left\langle \nabla f(x^{(k)}), p \right\rangle + \frac{1}{2} \left\langle B^{(k)} p, p \right\rangle$$

- Here $B^{(k)}$ is a symmetric positive definite matrix to be found
- Observe that

$$m_k(0) = f(x^{(k)}) \quad \text{and} \quad \nabla m_k(0) = \nabla f(x^{(k)})$$

- The minimizer $p^{(k)}$ of the above convex quadratic model is

$$p^{(k)} = -\left(B^{(k)}\right)^{-1} \nabla f(x^{(k)})$$

- One can use this direction to build the next iterate

$$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$$

- Now take a quadratic approximation of $f$ around $x^{(k+1)}$

$$m_{k+1}(p) = f(x^{(k+1)}) + \left\langle \nabla f(x^{(k+1)}), p \right\rangle + \frac{1}{2} \left\langle B^{(k+1)} p, p \right\rangle$$

- Here again, $B^{(k+1)}$ is a symmetric positive definite matrix to be found
- Observe that

$$m_{k+1}(0) = f(x^{(k+1)}) \quad \text{and} \quad \nabla m_{k+1}(0) = \nabla f(x^{(k+1)})$$

- Suppose we demand that

$$\nabla m_{k+1}(-\alpha_k p^{(k)}) = \nabla f(x^{(k)})$$

i.e. the gradient of the above quadratic matches with that of $f$ at $x^{(k)}$ as well

- Observe that

$$\nabla m_{k+1}(-\alpha_k p^{(k)}) = \nabla f(x^{(k+1)}) - \alpha_k B^{(k+1)} p^{(k)}$$

- Our earlier demand results in the relation

$$\begin{aligned} \nabla f(x^{(k+1)}) - \alpha_k B^{(k+1)} p^{(k)} &= \nabla f(x^{(k)}) \\ \Longrightarrow \qquad B^{(k+1)} \alpha_k p^{(k)} &= \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) \end{aligned}$$

- Let us use the notations

$$\begin{aligned} s^{(k)} &:= \alpha_k p^{(k)} = x^{(k+1)} - x^{(k)} \\ y^{(k)} &:= \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) \end{aligned}$$

- Then the above relation involving the hessian approximation reads

$$B^{(k+1)} s^{(k)} = y^{(k)}$$

- Given $s^{(k)}$ and $y^{(k)}$, we need to find a symmetric positive definite matrix $B^{(k+1)}$ such that

$$B^{(k+1)}s^{(k)} = y^{(k)}$$

- This is referred to as the secant equation
- Observe that by taking the inner product of the secant equation with $s^{(k)}$, we obtain

$$\left\langle B^{(k+1)}s^{(k)}, s^{(k)} \right\rangle = \left\langle y^{(k)}, s^{(k)} \right\rangle$$

- As we are looking for a positive definite $B^{(k+1)}$, the input $s^{(k)}$ and $y^{(k)}$ should necessarily satisfy

$$\left\langle y^{(k)}, s^{(k)} \right\rangle > 0$$

- Recall that for a strongly convex function $f$, we have strict monotonicty of the gradient, i.e.

$$(\nabla f(x) - \nabla f(y)) \cdot (x - y) > 0 \qquad \text{for all distinct } x, y \in \mathbb{R}^n.$$

- Hence the necessary condition for the secant equation is satisfied
- For a general objective function, this may not be true
- Suppose we are performing a line search algorithm where the step lengths satisfy the Wolfe conditions
- The second Wolfe condition says that for some $c_2 \in (0,1)$,

$$\left\langle \nabla f(x^{(k)} + \alpha_k p^{(k)}), p^{(k)} \right\rangle \geq c_2 \left\langle \nabla f(x^{(k)}), p^{(k)} \right\rangle$$

- Hence we deduce that

$$\left\langle y^{(k)}, s^{(k)} \right\rangle \geq (c_2 - 1)\alpha_k \left\langle \nabla f(x^{(k)}), p^{(k)} \right\rangle > 0,$$

thanks to $p^{(k)}$ being a descent direction

♣ Given an objective function $f : \mathbb{R}^n \to \mathbb{R}$

- Begin with a starting point $x^{(0)}$
- Begin with a symmetric positive definite matrix $B^{(0)}$
- Take the decent direction to be

$$p^{(0)} = -\left(B^{(0)}\right)^{-1} \nabla f(x^{(0)})$$

- Picking a step length $\alpha_0$ satisfying the Wolfe conditions, define

$$x^{(1)} = x^{(0)} + \alpha_0 p^{(0)}$$

- Build a symmetric positive definite matrix $B^{(1)}$ satisfying

$$B^{(1)} s^{(0)} = y^{(0)}$$

where $\qquad s^{(0)} = \alpha_0 p^{(0)} \qquad$ and $\qquad y^{(0)} = \nabla f(x^{(1)}) - \nabla f(x^{(0)})$

- Take the next descent direction to be

$$p^{(1)} = -\left(B^{(1)}\right)^{-1} \nabla f(x^{(1)})$$

## FINDING A SOLUTION TO THE SECANT EQUATION

- Goal is to find a symmetric $n \times n$ positive definite matrix satisfying the secant equation

$$Bs = y$$

  where the input $s, y$ satisfy $\langle s, y \rangle > 0$
- Symmetric condition implies that we need to find only $\frac{n(n+1)}{2}$ entries in $B$
- Secant equation is a collection of $n$ equations
- Recall that a matrix is positive definite if and only if all its leading principal minors are positive
- Demanding $B$ to be positive definite thus gives $n$ more equations

END OF LECTURE 7

THANK YOU FOR YOUR ATTENTION