OM Homework set - 5

Akshay Bankar
2019201011

PAGE NO.
DATE / /

1.1 Jensen's inequality: Let $f(x)$ be convex in interval I,
Then for any $x_1, \ldots, x_t \in I$

f[avg. of $\{x_i\}$) $\leq$ avg. of $\{f(x_i)\}$

and if f were concave, the inequality is reversed i.e

f(avg. of $\{x_i\}$) $\geq$ avg. of $\{f(x_i)\}$

Given inequality to prove,

$$\frac{1}{x-1} + \frac{1}{x} + \frac{1}{x+1} > \frac{3}{x}$$

Let $f(x) = \frac{1}{x}$ which is convex in interval $(0, \infty)$

Let the set $X = \{(x-1), x, (x+1)\}$

$\therefore$ Avg. of $\{f(x_i)\} = \frac{1}{3}\left(f(x-1) + f(x) + f(x+1)\right)$

$$= \frac{1}{3}\left[\frac{1}{x-1} + \frac{1}{x} + \frac{1}{x+1}\right] - (1)$$

and f(Avg. of $\{x_i\}$) $= f\left(\frac{1}{3}(x-1 + x + x+1)\right)$

$$= f\left(\frac{3x}{3}\right)$$

$$= f(x)$$

$$= \frac{1}{x} - (2)$$

From Jensen's inequality, (1), (2)

$$\frac{1}{3}\left(\frac{1}{x-1} + \frac{1}{x} + \frac{1}{x+1}\right) > \frac{1}{x}$$

or $$\frac{1}{x-1} + \frac{1}{x} + \frac{1}{x+1} > \frac{3}{x}$$

→ Given series, $S = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} + \ldots$

S can be written as,

$$S = \left(\frac{1}{2-1} + \frac{1}{2} + \frac{1}{2+1}\right) + \frac{1}{4} + \left(\frac{1}{6-1} + \frac{1}{6} + \frac{1}{6+1}\right) + \frac{1}{8} + \left(\frac{1}{10-1} + \frac{1}{10} + \frac{1}{10+1}\right) + \frac{1}{12} + \cdots$$

Using Jensen's inequality for numbers in brackets we get.

$$S > \left(\frac{3}{2} + \frac{3}{6} + \frac{3}{10} + \cdots\right) + \left(\frac{1}{4} + \frac{1}{8} + \frac{1}{12} + \cdots\right)$$

or $$S > \frac{3}{2}\left(1 + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \cdots\right) + \frac{1}{4}\underbrace{\left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots\right)}_{S}$$

$$\Rightarrow \quad S > \frac{3}{2}\left(1 + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \cdots\right) + \frac{S}{4}$$

$$\Rightarrow \quad \frac{3S}{4} > \frac{3}{2}\left(1 + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \cdots\right) \quad - (1)$$

Now, let $S_1 = 1 + \left(\frac{1}{3}\right) + \left(\frac{1}{5} + \frac{1}{7}\right) + \left(\frac{1}{9} + \frac{1}{11} + \frac{1}{13} + \frac{1}{15}\right) + \cdots$

$$S_1 > 1 + \frac{1}{4} + \left(\frac{1}{8} + \frac{1}{8}\right) + \left(\frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16}\right) + \cdots$$

$$= 1 + \frac{1}{4} + \frac{1}{4} + \cdots$$

$$\text{i.e. } \sum_{n=1}^{\infty} \frac{n}{4}$$

$\therefore$ as $n \to \infty \quad \sum \frac{n}{4} \to \infty$

$\therefore \quad S_1 \to \infty$ and hence S_1 is divergent.

Now from (1),

$$S > \frac{4}{2} S_1$$

$$\text{or} \quad S > 2S_1$$

Since S_1 is divergent which implies S is also divergent.

Therefore the given series does not converge to real number.

DATE / /

1.2 Three Chords Lemma: Let $f: I \to \mathbb{R}$. Then f is convex on I if and only if, for any points $a, b, c \in I$ with $a < b < c$ we have

$$\frac{f(b)-f(a)}{b-a} \leq \frac{f(c)-f(a)}{c-a} \leq \frac{f(c)-f(b)}{c-b}$$

Proof: Given that f is convex on I and $a < b < c$.

We can write $b = \lambda a + (1-\lambda)c$

where $\lambda = \frac{c-b}{c-a} \in [0,1]$

Since f is convex,

$$f(b) \leq \left(\frac{c-b}{c-a}\right) f(a) + \left(\frac{b-a}{c-a}\right) f(c) \quad - (1)$$

$$\Rightarrow f(b) - f(a) \leq \left(\frac{a-b}{c-a}\right) f(a) + \left(\frac{b-a}{c-a}\right) f(c)$$

$$\Rightarrow f(b) - f(a) \leq \frac{b-a}{c-a}\left(f(c) - f(a)\right)$$

$$\Rightarrow \frac{f(b)-f(a)}{b-a} \leq \frac{f(c)-f(a)}{c-a} \quad - (2)$$

This proves the 1st inequality.

From eq (1) we have

$$f(b) \leq \lambda f(a) + (1-\lambda) f(c)$$

$$\Rightarrow \lambda [f(a) - f(c)] + [f(c) - f(b)] \geq 0$$

$$\Rightarrow \frac{c-b}{c-a}[f(c) - f(a)] \leq f(c) - f(b)$$

$$\Rightarrow \frac{f(c)-f(a)}{c-a} \leq \frac{f(c)-f(b)}{c-b} \quad - (3)$$

Hence from eqn (2) and (3) we prove,

$$\frac{f(b)-f(a)}{b-a} \leq \frac{f(c)-f(a)}{c-a} \leq \frac{f(c)-f(b)}{c-b}$$

DATE / /

1.3 Given: x, y, z are positive real numbers with $x+y+z=1$

To show: $\left(1+\frac{1}{x}\right)\left(1+\frac{1}{y}\right)\left(1+\frac{1}{z}\right) \geq 64$

Soln: Let $f(x) = \log\left(1+\frac{1}{x}\right)$ which is convex in $(0, \infty)$

Consider average of functions, $f(x)$, $f(y)$, $f(z)$

$$\Rightarrow \frac{1}{3}\left(f(x)+f(y)+f(z)\right)$$

$$= \frac{1}{3}\left(\log\left(1+\frac{1}{x}\right)+\log\left(1+\frac{1}{y}\right)+\log\left(1+\frac{1}{z}\right)\right)$$

$$= \frac{1}{3}\log\left[\left(1+\frac{1}{x}\right)\left(1+\frac{1}{y}\right)\left(1+\frac{1}{z}\right)\right] \quad - (1)$$

Now, consider function f over average of x, y, z

i.e $f\left(\frac{x+y+z}{3}\right) = \log\left(1+\frac{3}{x+y+z}\right)$

$$= \log\left(1+\frac{3}{1}\right) \quad [\because x+y+z=1]$$

$$= \log(4) \quad - (2)$$

Now, from Jensen's inequality,

$$\frac{1}{3}\left(f(x)+f(y)+f(z)\right) \geq f\left(\frac{x+y+z}{3}\right)$$

From (1) & (2)

$$\frac{1}{3}\log\left[\left(1+\frac{1}{x}\right)\left(1+\frac{1}{y}\right)\left(1+\frac{1}{z}\right)\right] \geq \log 4$$

Taking exponentiation on both sides,

$$\left(1+\frac{1}{x}\right)\left(1+\frac{1}{y}\right)\left(1+\frac{1}{z}\right) \geq 64.$$

1.4 Given: $x^3 - 2x - 5 = 0$, $x_0 = 2$, $x_1 = 3$

Soln: The secant method is a root-finding algorithm that uses a succession of roots of secant lines to better approximate root of a function f.

$$x_2 = x_0 - f(x_0)\frac{(x_1 - x_0)}{f(x_1) - f(x_0)}$$

1st iteration: $x_0 = 2$, $x_1 = 3$

$f(x_0) = f(2) = (2)^3 - 2(2) - 5 = -1$

$f(x_1) = f(3) = 16$

$$\therefore \quad x_2 = x_0 - f(x_0) \cdot \frac{x_1 - x_0}{f(x_1) - f(x_0)}$$

$$= 2 - (-1) \cdot \frac{(3-2)}{16-(-1)}$$

$x_2 = 2.0588$

$\therefore \quad f(x_0) = f(2.0588) = -0.3908$

2nd iteration: $x_1 = 3$, $x_2 = 2.0588$

$$\therefore \quad x_3 = x_1 - f(x_1)\frac{(x_2 - x_1)}{f(x_2) - f(x_1)}$$

$$= 3 - 16 \cdot \frac{(2.0588 - 3)}{-0.3908 - 16}$$

$\Rightarrow \quad x_3 = 2.0813$

$\therefore \quad f(x_3) = f(2.0813) = -0.1472$

DATE / /

1.5 Given: $xe^x = 1$, $x \in [0, 1]$

Soln: Bisection method: It is a root-finding method that repeatedly bisects an interval and then selects a subinterval in which a root must lie for further processing.

Here $xe^x - 1 = 0$

Let $f(x) = xe^x - 1$

1st iteration: ~~f to)~~

$f(0) = -1 < 0$ and $f(1) = 1.7183 > 0$

∴ Root lies between 0 and 1

$x_0 = \frac{0+1}{2} = 0.5$

$f(x_0) = f(0.5) = 0.5e^{0.5} - 1 = -0.1756 < 0$

2nd iteration:

$f(0.5) = -0.1756 < 0$ & $f(1) = 1.7183 > 0$

∴ Root lies between 0.5 and 1

$x_1 = \frac{0.5+1}{2} = 0.75$

$f(x_1) = f(0.75) = 0.75e^{0.75} - 1 = 0.5878 > 0$

DATE / /

1.6 Given - $x^2 - y^2 = 3$, $x^2 + y^2 = 13$, $x_0 = y_0 = \sqrt{6.5}$

To find - Real roots of given equations by Newton's Method (2 iter)

Soln: Let $f_1 = x^2 - y^2 - 3 = 0$ - (1)

$f_2 = x^2 + y^2 - 13 = 0$ - (2)

Determine functional form of the partial derivatives,

$$J_{1,1} = \frac{\partial f_1}{\partial x} = 2x, \quad J_{1,2} = \frac{\partial f_1}{\partial y} = -2y$$

$$J_{2,1} = \frac{\partial f_2}{\partial x} = 2x, \quad J_{2,2} = \frac{\partial f_2}{\partial y} = 2y$$

$$\therefore J = \begin{bmatrix} 2x & -2y \\ 2x & 2y \end{bmatrix}$$

with $x_0 = \sqrt{6.5}$ $y_0 = \sqrt{6.5} = 2.5495$., we get

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} x_n \\ y_n \end{pmatrix} - J^{-1} f\left(\begin{bmatrix} x_n \\ y_n \end{bmatrix}\right)$$

$\therefore$ n = 1st iteration,

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 2.5495 \\ 2.5495 \end{bmatrix} - \begin{bmatrix} 5.099 & -5.099 \\ 5.099 & 5.099 \end{bmatrix}^{-1} \times f\left(\begin{bmatrix} 2.5495 \\ 2.5495 \end{bmatrix}\right) \rightarrow \begin{bmatrix} -3 \\ 0 \end{bmatrix}$$

$$\begin{matrix} x_1 \\ y_1 \end{matrix} = \begin{bmatrix} 2.8437 \\ 2.2553 \end{bmatrix}$$

n = 2nd iteration, $x_1 = 2.8437$, $y_1 = 2.2553$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2.8437 \\ 2.2553 \end{bmatrix} - \begin{bmatrix} 5.6874 & -4.5107 \\ 5.6874 & 4.5107 \end{bmatrix}^{-1} \times \begin{bmatrix} 0 \\ 0.1731 \end{bmatrix}$$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2.82847 \\ 2.23615 \end{bmatrix}$$

$\therefore$ After 2 iterations, approx roots are $x = 2.82847$, $y = 2.23615$

1.7 Fixed point iteration method: Given transcendental equation $f(x)=0$ can be converted into the form $x=g(x)$ & then using the iterative scheme with the recursive relation,

$$x_{i+1} = g(x_i), \quad i = 0, 1, 2, \ldots$$

with some initial guess x_0 is called fixed point method.

Given: $f(x) = 2x - \cos x - 3 = 0$

$$g(x) = \frac{\cos x + 3}{2}, \quad x_0 = \pi/2$$

From $f(x)$ we have, $x = \frac{\cos x + 3}{2} = g(x)$

Iteration-1 :- $x_0 = 1.57\pi/2$, $f(x_0) = \pi - 3 = 0.14159$

$$x_1 = g(x_0) = \frac{\cos(\pi/2) + 3}{2}$$

$x_1 = 1.5$, $f(x_1) = 0.3472$

Iteration-2 : $x_1 = 1.5$

$$x_2 = g(1.5) = \frac{\cos(1.5) + 3}{2} = \frac{0.0707 + 3}{2}$$

~~$x_2 = 1.$~~ ~~0007~~

$x_2 = 1.53537$, $f(x_2) = -0.0173$

DATE / /

1.8 Pseudocode for gradient ascent:

Let X → Data of m samples and n features

y → Output

alpha → learning rate.

weights → To learn.

Update step of gradient ascent is given by

$$w = w + \alpha$$

```
Gradient_descent (X, y, t, num_iters)
    m, n = shape (X)
    alpha = t
    weights = zeros (n)
    for i = 1: num_iters :
        ŷ = X * weights
    error = y - ŷ
    weights = weights + alpha * error * X.
```

Maximum of $6 - (x_1^2 (x_1^2 - 16) + x_2^2 (x_2^2 - 9))$ #

→ Starting from (0,0), the gradient does not update and the maximum value of function stays at 6.

- If the starting point is changed to (0.1, 0.1), the maximum value the function reaches is 90.25

DATE / /

1-9 Batch gradient descent: In this, entire training set is used to perform one iteration of gradient descent. The average of the gradients of all the training example is taken and this mean gradient is used to update parameters.
Mini-batch GD - This uses a small subset of training data to compute the gradient.
Stochastic GD - A single training sample is used to find gradient and update parameters.

Batch GD is used because it is good for convex or relatively smooth error manifolds. as it moves somewhat directly towards an optimum solution.
But using SGD or mini-batch GD, the optimization path taken is erratic and gives inaccurate gradients but with advantage that computing the gradient is lot faster. Hence there is trade-off between accuracy in optimization steps versus the speed.
~~Batch GD takes faster~~
Batch GD gives more accurate weights but is slow if the data size is large in which case requires high, slow computation.

DATE / /

1.10 Convergence proof of gradient descent (with fixed step size)

Theorem: Suppose the function $f: \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and that it gradient is Lipschitz continuous with constant $L > 0$ i.e $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x-y\|_2$ for any x, y. Then if grad. descent is run for k iterations with a fixed step size $t \leq 1/L$, it will yield a solution which satisfies,

$$f(x^{(k)}) - f(x^*) \leq \frac{\|x^{(0)} - x^*\|_2^2}{2tk} \quad - (1)$$

where $f(x^*)$ is optimal value. This means GD is guaranteed to converge with rate $1/k$

∎

Proof: Since ∇f is Lipschitz continuous with const. L

$$\Rightarrow \quad \nabla^2 f(x) \preceq LI,$$

or $\nabla^2 f - LI$ is negative semidefinite matrix.

Using this, quadratic expansion of f around $f(x)$ gives following inequality

$$\begin{aligned} f(y) &\leq f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}\nabla^2 f(x)\|y-x\|_2^2 \\ &\leq f(x) + \nabla f(x)^T (y-x) + \frac{1}{2} L\|y-x\|_2^2 \end{aligned}$$

By putting grad. desc. update in this, let $y = x^+ = x - t\nabla f(x)$,

$$\begin{aligned} f(x^+) &\leq f(x) + \nabla f(x)^T (x^+ - x) + \frac{1}{2} L\|x^+ - x\|_2^2 \\ &= f(x) + \nabla f(x)^T (x - t\nabla f(x) - x) + \frac{1}{2} L\|x - t\nabla f(x) - x\|_2^2 \\ &= f(x) + \nabla f(x)^T t\nabla f(x) + \frac{1}{2} L\|t\nabla f(x)\|_2^2 \\ &= f(x) - t\|\nabla f(x)\|_2^2 + \frac{1}{2} L t^2 \|\nabla f(x)\|_2^2 \\ &= f(x) - (1 - \frac{1}{2}Lt) t \|\nabla f(x)\|_2^2 \quad - (2) \end{aligned}$$

Using $t \leq 1/L$, we know that,

$$-(1 - \frac{1}{2}Lt) = \frac{1}{2}Lt - 1 \leq \frac{1}{2} L\left(\frac{1}{L}\right) - 1 = \frac{1}{2} - 1 = \frac{-1}{2}$$

Putting this in (2)

$$f(x^+) \leq f(x) - \frac{1}{2} t \|\nabla f(x)\|_2^2 \quad - (3)$$

DATE / /

Since $\frac{1}{2}t\|\nabla f(x)\|$ will always be +ve unless $\nabla f(x)=0$, this inequality implies that objective function value strictly decreases with each iteration of grad. desc. until it reaches the optimal value $f(x) = f(x^*)$. This holds only if t is chosen small enough i.e $t \leq 1/L$

We can bound $f(x^*)$, the objective value at next iteration in terms of $f(x^*)$, the optimal objective value, since f is convex,

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x)$$
$$f(x) \leq f(x^*) + \nabla f(x)^T (x - x^*)$$

Putting this in (3)

$$f(x^+) \leq f(x^*) + \nabla f(x)^T (x - x^*) - \frac{t}{2}\|\nabla f(x)\|_2^2$$
$$f(x^+) - f(x^*) \leq \frac{1}{2t}\left(2t\nabla f(x)^T (x - x^*) - t^2\|\nabla f(x)\|_2^2\right)$$

$$f(x^+) - f(x^*) \leq \frac{1}{2t}\left(2t\nabla f(x)^T (x - x^*) - t^2\|\nabla f(x)\|_2^2 - \|x - x^*\|_2^2 + \|x - x^*\|_2^2\right)$$
$$f(x^+) - f(x) \leq \frac{1}{2t}\left(\|x - x^*\|_2^2 - \|x - t\nabla f(x) - x^*\|_2^2\right) \quad - (4)$$

By definition, $x^+ = x - t\nabla f(x)$, putting this in (4)

$$f(x^+) - f(x^*) \leq \frac{1}{2t}\left(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2\right) \quad - (5)$$

This inequality holds for x^+ in every iteration of GD, Summing over iterations

$$\sum_{i=1}^{k} f(x^{(i)} - f(x^+)) \leq \sum_{i=1}^{k} \frac{1}{2t}\left(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2\right)$$

$$= \frac{1}{2t}\left(\|x^{(0)} - x^*\|_2^2 - \|x^{(k)} - x^*\|_2^2\right)$$
$$\leq \frac{1}{2t}\left(\|x^{(0)} - x^*\|_2^2\right) \quad - (6)$$

Now, f is decreasing on every iteration, we can conclude that.

$$f(x^{(k)} - f(x^*)) \leq \frac{1}{k}\sum_{i=1}^{k} f(x^{(i)}) - f(x^*)$$

$$\leq \frac{\|x^{(0)} - x^*\|_2^2}{2tk} \quad [\text{substituting from (6)}]$$

Hence proves (1), the convergence of gradient descent.