E0 230 CMO - Assignment 2 (20th September, 2021)

Instructions:

- This is an assignment, and **all work submitted must be your own**!
- Attempt all questions
- Last date to submit your answers is **30th September 11:59 PM**.
- 1. Question 1 needs to be answered in the pdf on at most ONE SIDE of an A4 sheet.
  2. Question 2 needs to be answered in the pdf on at most ONE SIDE of an A4 sheet.
  3. Question 3 needs to be answered in the form.
  4. Questions 4 (a) - 4 (c) need to be answered in the pdf on at most ONE SIDE of an A4 sheet. Question 4 (d) needs to be answered in the form.
  5. Questions 5 (a) - 5 (b) need to be answered in the pdf on at most ONE SIDE of an A4 sheet. Question 5 (c) needs to be answered in the form.
  6. Question 6 needs to be answered in the pdf on at most TWO SIDES of an A4 sheet.
- **NOTE:** Everything written in the pdf needs brief justifications.
- A Teams form with slots for the answers to be written in the form will be uploaded 24 hours before the submission deadline. If you are asked to provide a number, enter it into the teams form. You answer should be correct to 3 decimal places unless stated otherwise.
- Both your code and your PDF must be submitted in a single zip file, which should be called **student_name_cmo21assn2.zip**.
- Choose the files required for your setup from the concerned directory in the zip file
- For the numericals, you need to call the executables we have created from your script.
- The oracles provided this time are slightly different. In the command line, type './Q3_oracle'. Then give your input in the usual form (SRNo, point) (e.g., 1,[1,2,3,4]). The oracle will return the function output and stay running, waiting for the next input. You can give the oracle as many inputs as you please.
- In case using linux or mac ensure that you add chmod permissions for executing the file code.
- For mac use the following instructions:
  - Run `chmod 777 <executable name>`
  - You may need to do: `Preferences > Security and Privacy > General` and allow the executable (app) to run (`click "allow"`).

1. (5 points) Let $f : \mathbb{R}^4 \to \mathbb{R}$ be a twice continuously differentiable quadratic function.

   (a) (1 point) What is the second-order Taylor expansion of $f(x)$ at x_0?

   Solution: Since f is a quadratic function, the second-order Taylor expansion is exact,

   $$f(x) = f(x_0) + \nabla f(x_0)^T(x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0).$$

   (b) Let $0 \prec \nabla^2 f(x) \preceq 25I$ for any $x \in \mathbb{R}^4$. At $x_0 = [4, 0, -2, 1]^T$ we are given that $f(x_0) = 6, \nabla f(x_0) = [8, 4, 4, 2]^T$. Let $g(x)$ be a convex quadratic function of the form $\frac{1}{2}x^T Q x + b^T x + c$ such that $f(x) \leq g(x)$ for all $x \in \mathbb{R}^4$ and $f(x_0) = g(x_0)$.

   i. (1 point) What is Q?

   ii. (1 point) What is b?

   iii. (1 point) What is c?

   Solution: Since $\nabla^2 f(x) \preceq 25I$ at every x, we have that $(x')^T \nabla^2 f(x)(x') \leq 25\|x'\|^2$ for any $x, x' \in \mathbb{R}^4$. Therefore,

   $$f(x) \leq f(x_0) + \nabla f(x_0)^T(x - x_0) + \frac{25}{2}\|x - x_0\|^2.$$

   Therefore,

   $$f(x) \leq 6 + [8, 4, 4, 2](x - [4, 0, -2, 1]^T) + \frac{25}{2}\|x - [4, 0, -2, 1]^T\|^2$$

   Using $g(x)$ as the R.H.S. of the inequality above we get that $f(x_0) = g(x_0)$ and $f(x) \leq g(x)$ for all $x \in \mathbb{R}^4$. By appropriate comparison of terms we get that $Q = 25I_4, b = [-92, 4, 54, -23]^T$ and $c = \frac{485}{2}$, where I_4 is the identity matrix in 4 dimensions.

   (c) (1 point) What is the minimum value of g?

   Solution: Since g is also a convex quadratic function and Q is symmetric, the minimum occurs at $x^* = -Q^{-1}b$. The minimum value of g is

   $$\frac{1}{2}(-Q^{-1}b)^T Q(-Q^{-1}b) + b^T(-Q^{-1}b) + c = \frac{1}{2}b^T Q^{-1}b - b^T Q^{-1}b + c = -\frac{1}{2}b^T Q^{-1}b + c = 4.$$

2. (5 points) Let $f(x) = x^T Q x + b^T x + c$ be a quadratic function with $Q \succ 0$. Let $g_k = \nabla f(x^{(k)})$. Suppose we apply the following fixed step-size update step to gradient descent:

   $$x^{(k+1)} = x^{(k)} - \alpha g_k.$$

   In the first part of this question, we will derive constraints on the step size α that guarantee global convergence, i.e. gradient descent with step size α converges for any choice of starting point. Consider how the function value changes as we move from $x^{(k)}$ to $x^{(k+1)}$. We have that

   $$f(x^{(k+1)}) = f(x^{(k)}) - \alpha g_k^T g_k + \frac{\alpha^2}{2} g_k^T \nabla^2 f(x^{(k)}) g_k.$$

   (a) (3 points) Determine the values a, b such that for any $\alpha \in (a, b)$ we have that $f(x^{(k+1)})$ is always smaller than $f(x^{(k)})$. Ensure that a and b are independent of the iteration number k.

Solution: To ensure this we need to ensure that

$$\frac{\alpha^2}{2} g_k^T \nabla^2 f(x^{(k)}) g_k - \alpha g_k^T g_k < 0.$$

This is a quadratic equation in α with a positive coefficient of α^2 (since $\nabla^2 f(x)$ is positive definite), which means that it is negative in the open interval $(0, \frac{2\|g_k\|^2}{g_k^T \nabla^2 f(x)} g_k)$. We then have,

$$\frac{2\|g_k\|^2}{g_k^T \nabla^2 f(x) g_k} \geq \frac{2}{\lambda_{max}(\nabla^2 f(x))} = \frac{1}{\lambda_{max}(Q)}.$$

Thus, any $\alpha \in (0, \frac{1}{\lambda_{max}(Q)})$ will ensure that $f(x^{(k+1)}) < f(x^{(k)})$.

(b) (2 points) We can show that gradient descent with constant step size will always converge for any α satisfying the constraints derived above. You need not prove this in this question, but it would be a good exercise to prove this for yourself.
Now, consider the quadratic function

$$f(x) = 3(x_1^2 + x_2^2) + 4x_1x_2 + 5x_1 + 6x_2 + 7$$

Use 2(a) to find an interval (a, b) such that gradient descent with constant step size $\alpha \in (a, b)$ will ensure global convergence for f.

Solution: $f(x) = X^T \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix} X + [5, 6]^T X + 7.$

Here $Q = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}$ and its eigenvalues are 5 and 1. Therefore, $\alpha \in (0, \frac{1}{5})$ and $a = 0, b = \frac{1}{5}$ is the required solution.

3. (10 points) *In-exact line search.* Backtracking is a form of inexact line search in which a step size is determined at each step which satisfies the Armio-Goldstein condition. Given constants $\alpha, \beta \in (0, 1)$, at each step of the algorithm, if the current point is $x \in \mathbb{R}^d$, the direction of line search is chosen as $u = -\nabla f(x)$, and for determining the step size, an initial step size $t = 1$ is chosen and is repeatedly updated as $t \leftarrow \beta t$ until $f(x + tu) \leq f(x) + \alpha t \nabla f(x)^T u$ and then x is updated as $x \leftarrow x + tu$. Once the update distance $||tu||_2$ for the point x becomes less than ϵ during any epoch, the algorithm is stopped. The following is the pseudocode for implementing inexact line search with backtracking:

```
i ← 0
u ← −∇f(x)
t ← 1
while i < maxIterations do
    if f(x + tu) ≤ f(x) + αt∇f(x)^T u then
        return x + tu
    end if
    if ‖tu‖ < ϵ then
        return x + tu
    end if
    t ← βt
    i ← i + 1
end while
```

You are given oracles for $f : \mathbb{R}^4 \to \mathbb{R}$, ∇f and $\nabla^2 f$.

(a) Implement inexact line search with backtracking (described above) using $\alpha = 0.5, \beta = 0.5, \epsilon = 10^{-7}$ and $maxIterations = 30$. To test the correctness of your implementation, apply backtracking line search to f starting from $[-1, -1, -1, -1]$. Report two things:

i. (1 point) The value of t at the end of the search. Report a where $t = 2^{-a}$.

ii. (1 point) The value of $f(x)$ at the end of the search.

(b) Now use the provided oracles to implement the following two algorithms:

1. Gradient descent with inexact line search using backtracking with the same parameters as above, i.e. $\alpha = 0.5, \beta = 0.5, \epsilon = 10^{-7}$ and $maxIterations = 30$.
2. Newton's Method.

Apply each algorithm to f using the starting point $[5, -3, -5, 3]$. Run each algorithm for 50 iterations or until $\|\nabla f(x)\| < 10^{-10}$, whichever occurs first.

i. (5 points) Write down the value of $f(x)$ after 1, 5, 10 and 20 iterations of gradient descent. Also report the final value of $f(x)$ after gradient descent terminates.

ii. (2+1 points) Write down the solution obtained for Newton's method, and the number of iterations in which termination occured.

4. (10 points) Consider the one dimensional function $f(x) = \frac{1}{2}\log(x^2 + 1)$.

(a) (1 point) Prove or disprove: f is convex.

Solution: The gradient of f is given by,

$$f'(x) = \frac{x}{x^2 + 1}.$$

The double derivative of f is given by,

$$f''(x) = \frac{1 - x^2}{(1 + x^2)^2}.$$

The double derivative of f is negative at $|x| > 1$. Therefore the function is not convex.

(b) (4 points) Is the function L-smooth for some $L > 0$? If yes, find the smallest positive integer value of L for which this function is L-smooth. If not, disprove.

Solution: We know that $|f''(x)| \leq 1$ for all $x \in \mathbb{R}$ (check). Now using the fundamental theorem of calculus we have,

$$\begin{aligned}
f'(y) - f'(x) &= \int_0^1 f''(x + t(y - x))(y - x)dt \\
\implies |f'(y) - f'(x)| &= \left| \int_0^1 f''(x + t(y - x))(y - x)dt \right| \\
\implies |f'(y) - f'(x)| &\leq \left| \int_0^1 f''(x + t(y - x))dt \right| \cdot |y - x| \\
\implies |f'(y) - f'(x)| &\leq \left(\int_0^1 |f''(x + t(y - x))dt| \right) \cdot |y - x| \\
\implies |f'(y) - f'(x)| &\leq \left(\int_0^1 |1dt| \right) \cdot |y - x| \\
\implies |f'(y) - f'(x)| &\leq 1 \cdot |y - x|.
\end{aligned}$$

Therefore $L = 1$.

(c) (1 point) Can we use gradient descent to find a local minimum? Why or why not?

Solution: The given function is in C_1^L for $L = 1$ and is bounded from below, that is, the smallest value of $f(x)$ is 0. Hence we can apply gradient descent to converge to a point where gradient is 0.

(d) (4 points) Apply gradient descent by choosing the stepsize using in-exact line search with backtracking. Use $\alpha = 0.5, \beta = 0.1, \epsilon = 0.01$ and $maxIterations = 30$ for backtracking. For each of the following initial points report the values of k and $x^{(k)}$ (rounded to 3 decimal places) with stopping condition: $|f'(x^{(k)})| \leq \delta$, where $\delta = 0.05$,

i. (1 point) $x^{(0)} = 0.1$.
ii. (1 point) $x^{(0)} = 0.6$.
iii. (1 point) $x^{(0)} = -0.5$.
iv. (1 point) $x^{(0)} = 1.2$.

5. (10 points) *Newton's Method.* Consider the same function $f(x) = \frac{1}{2}\log(x^2 + 1)$.

(a) (1 point) Prove or disprove if $x^* = 0$ is a local minimum of f.

Solution: The gradient of f is given by,

$$f'(x) = \frac{x}{x^2 + 1}.$$

This is 0 at $x = 0$. Therefore $x^* = 0$ is a local minimum.

(b) (5 points) For $x^* = 0$, derive the largest a such that Newton's Method is effective for any $x^{(0)} \in (x^* - a, x^* + a)$.

Solution: Using the Newton's Method, the update rule for x is,

$$x = x - \frac{f'(x)}{f''(x)} = x - \frac{x(x^2+1)}{1 - x^2} = \frac{x - x^3 - x^3 - x}{1 - x^2} = \frac{-2x^3}{1 - x^2}.$$

For the Newton's method to be effective for $x^* = 0$ we either should start at $x = 0$ or we want a starting point x such that,

$$\left|\frac{-2x^3}{1 - x^2}\right| < |x|. \tag{1}$$

That is, the value of x should move closer to 0 after every iteration. Let $f_1(x) = \frac{-2x^3}{1-x^2}$ and $f_2(x) = x$. Both f_1 and f_2 are odd functions. Note that modulus of an odd function is symmetric about $x = 0$,

$$|f_1(-x)| = |-f_1(x)| = |f_1(x)|.$$

Therefore, both $|f_1(x)|$ and $|f_2(x)|$ are symmetric about $x = 0$. Hence, we only need to check the cases where $x > 0$. Then by symmetry about $x = 0$, if $0 < x < a$ is a valid starting point, so is $-a < x < 0$.

Case A: Let $x \in (0, 1)$. Then $f_1(x) < 0$ and $f_2(x) > 0$. We want all the values of x such that $-f_1(x) < f_2(x)$. That is,

$$\frac{2x^3}{1 - x^2} < x.$$

Which is true for $x < \frac{1}{\sqrt{3}}$. Therefore $0 < x < \frac{1}{\sqrt{3}}$ is a valid starting point. By symmetry about $x = 0$, $-\frac{1}{\sqrt{3}} < x < 0$ is also a valid starting point.

Case B: Let $x \in (1, \infty)$. Then $f_1(x) > 0$ and $f_2(x) > 0$. We want all the values of x such that $f_1(x) < f_2(x)$. That is $\frac{-2x^3}{1-x^2} < x$ This is true when $\frac{2x^3}{x^2-1} < x$. But for $x > 1$, this is not possible. Therefore in this case, there are no valid values of x.
Therefore, Equation (1) is true for $|x| < \frac{1}{\sqrt{3}}$. Note that for x such that $|x| = \frac{1}{\sqrt{3}}$ the value of x does not get updated with the update rule. Therefore, $a = \frac{1}{\sqrt{3}}$.

(c) (4 points) Run Newton's method for each of the following initial points report the values of k and $x^{(k)}$ (rounded to 3 decimal places) with stopping condition: $|f'(x^{(k)})| \leq \delta$, where $\delta = 0.05$,

i. (1 point) $x^{(0)} = 0.1$.

ii. (1 point) $x^{(0)} = 0.6$.

iii. (1 point) $x^{(0)} = -0.5$.

iv. (1 point) $x^{(0)} = 1.2$.

6. (20 points) As we have seen in the class, the condition number ($\rho = \frac{\lambda_{\max}}{\lambda_{\min}}$) of the Hessian of a function determines the convergence rate of the gradient descent algorithm as $\left(\frac{\rho-1}{\rho+1}\right)^2$. Therefore, if the Hessian has very large condition number, the algorithm requires large number of iterations to converge. One way to avoid this is via appropriate linear transformations of the decision variables.

Consider the unconstrained minimization problem,

$$\min_{x \in \mathbb{R}^d} \quad f(x),$$

where f is a strictly-convex function, i.e., Hessian of f is positive definite everywhere. Now for any non-singular matrix $S \in \mathbb{R}^{d \times d}$, consider the linear transformation $x = Sy$. Then define $g(y) := f(Sy)$.

(a) (2 points) Show that finding an x that minimizes $f(x)$ is the same as finding a y that minimizes $g(y)$. That is, show that $x^* = Sy^*$, where $x^* = \arg\min_x f(x)$ and $y^* = \arg\min_y g(y)$.

Solution: Note that since S is invertible, it forms a bijective map from $\mathbb{R}^d$ to $\mathbb{R}^d$. Therefore,

$$x^* = \arg\min_x f(x) = \arg\min_{Sy} f(Sy) = \arg\min_{Sy} g(y) = Sy^*.$$

(b) (3 points) By chain rule we have $\nabla g(y) = \nabla f(Sy) = S^T \nabla f(x)$. Let α_k be the stepsize of the gradient descent algorithm on the transformed problem, at iteration k. Then the update to y is,

$$y^{(k+1)} = y^{(k)} - \alpha_k \nabla g(y^{(k)}).$$

Substituting $\nabla g(y)$ and left multiplying by S on both sides gives,

$$x^{(k+1)} = x^{(k)} - \alpha_k S S^T \nabla f(x^{(k)}) \tag{2}$$

Show that $-SS^T \nabla f(x^{(k)})$ is a descent direction of f at $x^{(k)}$, when $\nabla f(x^{(k)}) \neq \mathbf{0}$.

Solution: Any direction is a descent direction if it makes an obtuse angle with the gradient. We have that,

$$\begin{aligned}
(-SS^T \nabla f(x^{(k)}))^T \nabla f(x^{(k)}) &= -\nabla f(x^{(k)})^T S^T S \nabla f(x^{(k)}) \\
&= -\left(S \nabla f(x^{(k)})\right)^T \left(S \nabla f(x^{(k)})\right) \\
&= -\|S \nabla f(x^{(k)})\|^2 \leq 0.
\end{aligned}$$

Note that since S is non-singular the null-space of S consists only of the vector $\mathbf{0}$. Therefore, $S\nabla f(x^{(k)}) = 0$ if and only if $\nabla f(x^{(k)}) = \mathbf{0}$. But given that $\nabla f(x^{(k)}) \neq \mathbf{0}$. Therefore, $-\|S\nabla f(x^{(k)})\|^2 < 0$. Therefore, $(-SS^T\nabla f(x^{(k)}))$ is a descent direction.

(c) (1 point) What is Hessian of g in terms of S and $\nabla^2 f$.

Solution: $\nabla^2 g(y) = S^T\left(\nabla^2 f(x)\right)S$.

(d) (1 point) When $SS^T = \left(\nabla^2 f(x^{(k)})\right)^{-1}$ and $\alpha_k = 1, \forall k$, which other algorithm we have seen in class does the update rule in Equation (2) correspond to?

Solution: Newton's method. The update rule becomes

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

(e) (3 points) If condition number of $\nabla^2 f(x)$ is ρ, what is the condition number of $\nabla^2 g(y)$ when $SS^T = \left(\nabla^2 f(x)\right)^{-1}$?

Solution: From part (b)

$$\begin{aligned}
&\nabla^2 g(y) = S^T\left(\nabla^2 f(x)\right)S \\
\implies &S\nabla^2 g(y) = SS^T\left(\nabla^2 f(x)\right)S && \text{(left multiply both sides by } S\text{)} \\
\implies &S\nabla^2 g(y) = \left(\nabla^2 f(x)\right)^{-1}\left(\nabla^2 f(x)\right)S && \text{(substitute } SS^T\text{)} \\
\implies &S\nabla^2 g(y) = S && \text{(left multiply both sides by } S^{-1}\text{)} \\
\implies &\nabla^2 g(y) = I
\end{aligned}$$

Therefore, condition number of the Hessian of $g(y) = 1$, no matter what the value of ρ is.

(f) We have provided an oracle to a convex quadratic function $f : \mathbb{R}^{10} \to \mathbb{R}$ where $f(x) = \frac{1}{2}x^TQx + b^Tx + c$ and Q has a large condition number. You can query at any point x, the value and gradient of x and also Qx, the action of Q on x. This exercise is to check that a scaled gradient helps decrease the number of iterations until convergence, by well-conditioning the problem.

i. (1 point) With descent direction u at a point x, what is the optimal stepsize for gradient descent with exact line search?

Solution: For a convex quadratic function with Hessian at any point given by $\nabla^2 f(x)$, and for any descent direction u, the optimal stepsize at x is,

$$\alpha^* = -\frac{\nabla f(x)^T u}{u^T\nabla^2 f(x)u}.$$

ii. (3 points) What is the stepsize chosen by exact line search for the update rule in Equation (2)?

Solution: In Equation (2) the descent direction is $u = -SS^T\nabla(f(x^{(k)})$ as derived above.

Therefore,

$$\alpha_k = -\frac{\nabla f(x^{(k)})^T(-SS^T\nabla(f(x^{(k)}))}{(-SS^T\nabla(f(x^{(k)}))^T\nabla^2 f(x)(-SS^T\nabla(f(x^{(k)}))} = \frac{\|S^T\nabla f(x^{(k)})\|^2}{\nabla(f(x^{(k)})^T SS^T\nabla^2 f(x)SS^T\nabla(f(x^{(k)})}.$$

iii. (2 points) Implement gradient descent with exact line search. In how many iterations does it converge starting at $x^{(0)} = [1, 50, \cdots, 50]^T$, with stopping condition: $\|x^{(k)} - x^{(k+1)}\| \leq \delta$, where $\delta = 0.01$.

iv. (3+1 points) Implement gradient descent with the update rule in Equation (2). Note that the stepsize also changes as in subpart (ii). Use

$$SS^T = \begin{pmatrix} \beta & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

SS^T is the 10×10 identity matrix with its $(1,1)^{th}$ entry replaced by β. You need to choose the best value of β for your matrix amongst $\frac{1}{200}, \frac{1}{700}$ and $\frac{1}{2000}$. Report in how many iterations scaled gradient descent converges for each β starting at $x^{(0)} = [1, 50, \cdots 50]^T$, with stopping condition: $\|x^{(k)} - x^{(k+1)}\| \leq \delta$, where $\delta = 0.01$. Use the number of iterations to deduce the best value of β for your function. Report the point to which scaled gradient descent converges when the SS^T defined above is initialized with this best value of β.

Solution: One can easily check Q is symmetric by querying the oracle with the standard basis vectors.