HW.a.1.

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{\mu}{2} \| x_2 - x_1 \|_2^2$$

Let $x_2 = x^*$ and $x_1 = x$

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2} \| x^* - x \|_2^2$$

$$f(x^*) - f(x) \leq \underbrace{\nabla f(x)^T (x - x^*) - \frac{\mu}{2} \| x_2 - x \|_2^2}$$

$$= \underbrace{\nabla f(x)^T (x - x^*) - \frac{\mu}{2} \| x^* - x \|_2^2 - \frac{1}{2\mu} \| \nabla f(x) \|_2^2} + \frac{1}{2\mu} \| \nabla f(x) \|_2^2$$

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

$$= -\frac{1}{2} \left\| \sqrt{\mu} (x - x^*) - \frac{1}{\sqrt{\mu}} \nabla f(x) \right\|_2^2 + \frac{1}{2\mu} \| \nabla f(x) \|_2^2$$

$$\leq \frac{1}{2\mu} \| \nabla f(x) \|_2^2$$

HW.a.2.

$$f(x_2) - f(x_1) \geq \nabla f(x_1)^T . (x_2 - x_1) + \frac{\mu}{2} \| x_2 - x_1 \|_2^2$$

$$+ \quad f(x_1) - f(x_2) \geq \nabla f(x_2)^T (x_1 - x_2) + \frac{\mu}{2} \| x_2 - x_1 \|_2^2$$

$$0 \geq (\nabla f(x_1) - \nabla f(x_2))^T (x_2 - x_1) + \mu \| x_2 - x_1 \|_2^2$$

$$= -(\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) + \mu \| x_2 - x_1 \|_2^2$$

$$\langle \nabla f(x_2) - \nabla f(x_1), x_2 - x_1 \rangle \geq \mu \| x_2 - x_1 \|_2^2$$

Applying Cauchy-Schwarz ineq. $\Rightarrow \langle u, v \rangle \leq \|u\| \|v\|$

$$\| \nabla f(x_2) - \nabla f(x_1) \| . \| x_2 - x_1 \| \geq \mu \| x_2 - x_1 \|_2^2$$

$$\| \nabla f(x_2) - \nabla f(x_1) \| \geq \mu \| x_2 - x_1 \|_2^2$$

HW. b. 1.

Using Taylor expansion

$$f(x_2) = f(x_1) + \langle \nabla f(x_1), x_2 - x_1 \rangle + \frac{1}{2} (x_2 - x_1)^T \nabla^2 f(x_1) (x_2 - x_1)$$

Since $\nabla^2 f(x_1) \leq L\, I_d$

$$f(x_2) \leq f(x_1) + \langle \nabla f(x_1), x_2 - x_1 \rangle + \frac{L}{2} \| x_2 - x_1 \|_2^2$$

HW. b. 2

$$f(x_2) - f(x_1) = f(x_2) - f(x_3) + f(x_3) - f(x_1)$$

$$\leq \langle f(x_2), x_2 - x_3 \rangle + \langle \nabla f(x_1), x_3 - x_1 \rangle + \frac{L}{2} \| x_3 - x_1 \|_2^2$$

- $f(x_2) - f(x_3) \leq \langle f(x_2), x_2 - x_3 \rangle$
- $f(x_3) - f(x_1) \leq \langle \nabla f(x_1), x_3 - x_1 \rangle + \frac{L}{2} \| x_3 - x_1 \|_2^2$

To minimize the right hand size of $f(x_2) - f(x_1) \leq \cdots$ ineq.

$$\frac{\partial}{\partial x_3} \left(\langle f(x_2), x_2 - x_3 \rangle + \langle \nabla f(x_1), x_3 - x_1 \rangle + \frac{L}{2} \| x_3 - x_1 \|_2^2 \right) = 0$$

$$\Rightarrow \quad x_3 = x_1 - \frac{1}{L} (\nabla f(x_1) - \nabla f(x_2))$$

$$f(x_2) - f(x_1) \leq \langle \nabla f(x_2), x_2 - x_1 + \frac{1}{L} (\nabla f(x_1) - \nabla f(x_2)) \rangle$$
$$+ \langle \nabla f(x_1), -\frac{1}{L} (\nabla f(x_1) - \nabla f(x_2)) \rangle$$
$$+ \frac{L}{2} \| \frac{1}{L} (\nabla f(x_1) - \nabla f(x_2)) \|_2^2$$

$$= \langle \nabla f(x_2), x_2 - x_1 \rangle - \frac{1}{2L} \| \nabla f(x_1) - \nabla f(x_2) \|_2^2$$

Interchanging x_1 and x_2

$$f(x_1) - f(x_2) \leq \langle \nabla f(x_1), x_1 - x_2 \rangle - \frac{1}{2L} \| \nabla f(x_1) - \nabla f(x_2) \|_2^2$$

$$f(x_2) \geq f(x_1) + \langle \nabla f(x_1), x_2 - x_1 \rangle + \frac{1}{2L} \| \nabla f(x_1) - \nabla f(x_2) \|_2^2$$

HW. 6. 3.

We have

$$f(x_1) - f(x_2) \leq \langle \nabla f(x_1), x_1 - x_2 \rangle - \frac{1}{2L} \| \nabla f(x_1) - \nabla f(x_2) \|_2^2$$

$$+ \quad f(x_2) - f(x_1) \leq \langle \nabla f(x_2), x_2 - x_1 \rangle - \frac{1}{2L} \| \nabla f(x_1) - \nabla f(x_2) \|_2^2$$

$$0 \leq (\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) - \frac{1}{L} \| \nabla f(x_1) - \nabla f(x_2) \|_2^2$$

$$(\nabla f(x_2) - \nabla f(x_1))^T (x_2 - x_1) \geq \frac{1}{L} \| \nabla f(x_1) - \nabla f(x_2) \|_2^2$$

≈ General Info. about Newton's method ≈

In equality constrained Newton's method, we start with x^0 such that $Ax^0 = b$.

Then repeat the updates

$$x^+ = x + \Delta x \quad \text{where} \quad v = \underset{A.z=0}{\operatorname{argmin}} \ \nabla f(x)^T (z-x) + \frac{1}{2}(z-x)^T . \nabla^2 f(x)(z-x)$$

this keeps x^+ in the feasible set, since

$$A.x^+ = A.x + A.\Delta x = b + 0 = b$$

Furthermore, v is the solution to minimize the objective function with equality constraint. From KKT conditions, v satisfies

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}$$

for some w. Thus, Newton criterion v is found by solving a linear system.

$$A.\Delta x = 0 \quad \text{and} \quad \nabla^2 f(x).\Delta x + A^T.w = -\nabla f(x)$$

$$\Rightarrow \Delta x = \left[\nabla^2 f(x)\right]^{-1} . \left[-\nabla f(x) - A^T w\right]$$

Therefore Δx solves the second-order approximation of f at x

$$\begin{aligned} &\text{minimize} && f(x) + \nabla f(x)^T \Delta x + \frac{1}{2}\Delta x^T . \nabla^2 f(x) \Delta x \\ &\text{s.t.} && A(x+\Delta x) = b \end{aligned}$$

⇒ If f is exactly quadratic, Newton method exactly solves the problem. If f is nearly quadratic, Newton method is a very good approximation.

HW. C. 1-2-3

We have $f(x) = \frac{1}{N}\left(f_1(x_1) + f_2(x_2) + \dots + f_N(x_N)\right)$

Then $\nabla f(x) = \frac{1}{N}\left[\nabla f_1(x_1) \quad \nabla f_2(x_2) \quad \cdots \quad \nabla f_N(x_N)\right]^T$

And

$$\nabla^2 f(x) = \begin{bmatrix} \nabla^2 f_1(x_1) & & & 0 \\ & \nabla^2 f_2(x_2) & & \\ & & \ddots & \\ 0 & & & \nabla^2 f_N(x_N) \end{bmatrix}$$

Non-diagonal entries of the Hessian matrix are 0 since when you take the derivative with respect to x_i and x_j where $i \neq j$, the second order derivative of any function $f_i(x_i)$ will be zero.

Since $\nabla^2 f(x)$ is a diagonal matrix, its inverse is trivial. Therefore, using the Newton method for solving the optimization problem is preferable. $N = 1000$ or $N = 10^9$ will not change the calculation of the Hessian matrix.

If we use Newton's method

i) choose an appropriate initial state for x vector such as $A x^0 = b$.

ii) compute Δx by using the KKT conditions

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

and $A . \Delta x = 0$

iii) quit if $\frac{\lambda^2}{2} \leq \epsilon$ where $\lambda^2 \triangleq \nabla f(x)^T . \nabla^2 f(x)^{-1} . \nabla f(x)$

iv) or update $x_{new} = x + \Delta x$

P.S.: I am not sure if $\Delta x = -\nabla^2 f(x)^{-1} . \nabla f(x)$ or

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

HW. C.4

But, if the objective function is

$$f(x) = \frac{1}{N} \sum f_i(x_i) + r(\underline{x})$$

where $r(x)$ is a twice differentiable function, this time the Hessian matrix may not be diagonal since we do not know if

$$\frac{\partial^2 r(\underline{x})}{\partial x_i \, \partial x_j} \overset{?}{=} 0$$

If it is not diagonal, calculating an inverse of a matrix w dimension $N = 10^9$ may take forever.

Now, we can use Quasi-Newton approach. The main idea here is: we can approximate Hessian matrix using only the gradient info.

Consider the following Taylor expansion

$$f(x^t) = f(x^{t+1}) + \left\langle \nabla f(x^{t+1}), x^t - x^{t+1} \right\rangle + \frac{1}{2} (x^t - x^{t+1})^T \cdot H_{t+1}^{-1} \cdot (x^t - x^{t+1})$$

where H_{t+1} is the approximation of the Hessian. H_{t+1} satisfies

$$\nabla f(x^t) = \nabla f(x^{t+1}) + H_{t+1}^{-1} (x^t - x^{t+1})$$

$$\Rightarrow \quad H_{t+1}^{-1} \cdot (x^t - x^{t+1}) = \nabla f(x^t) - \nabla f(x^{t+1})$$