Name: Hongda Li
AMATH 516 FALL 2021 HW1

Exercise 1.1

Given a collection of real $m \times n$ matrices, $A_1, A_2, \ldots, A_l$, define the linear mapping $\mathcal{A} : \mathbb{R}^{m\times n} \mapsto \mathbb{R}^l$ by setting:

$$\mathcal{A} := \begin{bmatrix} \langle A_1, X\rangle \\ \langle A_2, X\rangle \\ \vdots \\ \langle A_l, X\rangle \end{bmatrix}$$

And show that the mapping $A^*y = \sum_{i=1}^{l} y_i A_i$ is the adjoint mapping.
The operator and its adjoint should satisfy that $\langle AX, y\rangle = \langle X, A^*y\rangle$. In which we can start considering the left hand side:

$$\begin{aligned} \langle \mathcal{A}X, y\rangle &= \begin{bmatrix} \langle A_1, X\rangle \\ \langle A_2, X\rangle \\ \vdots \\ \langle A_l, X\rangle \end{bmatrix} \\ &= \sum_{i=1}^{l} y_i \langle A_i, X\rangle \\ &= \sum_{i=1}^{l} y_i \mathrm{Tr}(A_i^T X) \end{aligned} \tag{1.1.1}$$

And starting from the right hand side of the equation we have:

$$\begin{aligned} &\langle X, \mathcal{A}^* y\rangle \\ &= \langle X, \mathcal{A}^* y\rangle \\ &= \sum_{i=1}^{l} \langle X, y_i A_i\rangle \\ &= \sum_{i=1}^{l} y_i \langle X, A_i\rangle \end{aligned} \tag{1.1.2}$$

The last line of 1.1.2 and 1.1.1 is the same. Therefore the defined operator $\mathcal{A}^*$ is the adjoint operator.

Exercise 1.2

The inner product induced by a matrix A where A is SPD is an inner product and look for dual norm induces by this inner product.

From the property of Positive Definite Matrices we know that: $A \in \mathbf{S}^n_{++}$ where $\forall x \in \mathbf{E} - \{\mathbf{0}\} : \langle Ax, x\rangle > 0$. We wish to prove that $\langle v, w\rangle_A := \langle Av, w\rangle$ is an inner product. To do that we need to test on three properties of an inner product on $\mathbf{E}$ which would be:

1. The inner product is symmetric:

$$\langle v, w\rangle_A = \langle Av, w\rangle = \langle v, A^T w\rangle = \langle v, Aw\rangle = \langle v, w\rangle_A$$

   And it's symmetric.

2. It's a binlinear operator:

$$\begin{aligned} &\langle v, u+w\rangle_A \\ &= \langle Av, u+w\rangle \\ &= \langle Av, u\rangle + \langle Av, w\rangle \\ &= \langle v, u\rangle_A + \langle v, w\rangle_A \end{aligned} \tag{1.2.1}$$

By the fact that it's symmetric, we can prove that it's also linear for its first argument.

3. It's positive definite:

$$\begin{aligned} &\langle x, x\rangle_A \, \forall x \neq \mathbf{0} \\ &\langle x, x\rangle_A = \langle Ax, x\rangle \\ &= x^T Ax > 0 \quad \forall\, x \neq \mathbf{0} \end{aligned} \tag{1.2.2}$$

The last line is by the defintion of positive definite matrix applied on A.

In our case, the definition of the dual norm is:

$$\|x\|_A^* = \sup_z \{\langle x, z\rangle : \|z\|_A \leq 1\} \tag{1.2.3}$$

Here we invoke the property that A is Positive definite and hence there exists the factorization that $A = LL$ where $L = A^{1/2}$. And $L^T = L$.

$$\begin{aligned} &\|z\|_A \leq 1 \implies \langle Az, z\rangle \leq 1 \\ \implies &\langle LLz, z\rangle \leq 1 \\ \implies &\left\langle Lz, L^T z\right\rangle \leq 1 \\ \implies &\langle Lz, Lz\rangle \leq 1 \end{aligned} \tag{1.2.4}$$

Let $y = LZ$, then $Z = L^{-1}Y$, $\|z\|_A = \|y\|_2$, and this is possible because L is PSD. And hence the dual norm can be written in the form of:

$$\|x\|_A^* = \sup_{\|y\|_2 \leq 1} \left\langle x, L^{-1}y\right\rangle = \sup_{\|y\|_2 \leq 1} \left\langle L^{-T}x, y\right\rangle$$

To maximaize the quantity, we choose unit vector Y that points towards the direction of $L^{-T}x$, this is possible for all x because the matrix L is PSD and hence, invertible and it's full-ranked. Then we can say that:

$$\begin{aligned} y &= \frac{L^{-T}}{\|L^{-T}x\|_2} \\ \implies \|x\|_A^* &= \frac{\left\langle L^{-T}x, L^{-T}x\right\rangle}{\|L^{-T}x\|_2} \\ &= \sqrt{\langle L^{-T}x, L^{-T}x\rangle} \\ &= \sqrt{\langle x, L^{-1}L^{-T}x\rangle} \end{aligned} \tag{1.2.5}$$

Here, take notice that $L^{-T} = Q\sqrt{\Lambda^{-1}}Q^T = L^{-1}$, and that would mean $L^{-1}L^{-T} = L^{-1}L^{-1} = A^{-1}$. And we can do that without worrying because SPD matrix gives real positive λ on the diagonal of Λ, and hence $\|x\|_A^* = \sqrt{\langle x, A^{-1}x\rangle} = \|x\|_{A^{-1}}$

Exercise 1.6

Consider a closed function $f : \mathbf{E} \mapsto \bar{\mathbb{R}}$ and a nonempty compact set $Q \subset \mathbf{E}$. Then the infimum value of $\inf_{x \in Q} f(x)$ is attained at some point in Q.

The function is closed, which means that its lower-semicontinuity asserts:

$$\exists\, x^* \in \mathbf{E} : f(x^*) = \inf_{x \in Q} f(x)$$

The set Q is closed and compact using the Bozano-Weistrass theorem, all sequence converge to some point, and that limit point will be in the set. Therefore, x^* is attained inside of the set Q. And therefore:

$$\exists\, x^* \in Q : f(x^*) = \inf_{x \in Q} f(x)$$

Exercise 1.8

Any coercive closed function $f : \mathbf{E} \mapsto \bar{\mathbb{R}}$ has a minimizer.

By the definition that the function is coercive, we know that:

$$\forall\, r\ \exists\, \delta : \|x\| > \delta \implies f(x) > r \tag{1.8.1}$$

By the fact that the function is closed, we know that the set $Q = \{x : f(x) \leq r\}$ for a fixed r is a closed set, and it's compact too because $\|x\| > \delta \implies f(x) > r$, which is then not in the set, so it's inside of the $\|x\| \leq \delta$.
Using this new closed and compact set Q and ues what we proved in 1.6, we know that the function attains a minimum and the solution is in the domain Q.

Exercise 1.9

The function is:

$$f(x) = \frac{1}{2} \langle x, \mathcal{A}x \rangle + \langle \mathcal{A}, x \rangle + c$$

1.9.1

Replacing $\mathcal{A}$ by a Adjoint operator $\frac{1}{2}(\mathcal{A} + \mathcal{A}^*)$, then it won't change the value of the function. This is true because:

$$\begin{aligned} \frac{1}{2} \langle x, \mathcal{A}x \rangle &= \frac{1}{2} \left\langle x, \frac{1}{2}(\mathcal{A} + \mathcal{A}^*) \right\rangle \\ &= \frac{1}{4} \langle x, \mathcal{A}x + \mathcal{A}^* x \rangle \\ &= \frac{1}{4} \langle x, \mathcal{A}x \rangle + \frac{1}{4} + \langle x, \mathcal{A}^* x \rangle \\ &= \frac{1}{4} \langle x, \mathcal{A}x \rangle + \frac{1}{4} + \langle \mathcal{A}x, x \rangle \\ &= \frac{1}{2} \langle x, \mathcal{A}x \rangle \end{aligned} \tag{1.9.1.1}$$

And this would imply that:

$$f(x) = \frac{1}{2} \langle x, \mathcal{A}x \rangle + \langle v, x \rangle + c = \frac{1}{2} \langle x, \mathcal{A}x \rangle + \langle v, x \rangle + c \tag{1.9.1.2}$$

1.9.2

We assume that the $\mathcal{A} = \mathcal{A}^*$, then we want to figure out the gradient of the function.

In general, the rule for $\langle f, g \rangle$ where f, g maps from $\mathbb{R}^m$ to $\mathbb{R}^n$, then:

$$\nabla \langle f, g \rangle = \nabla f^T g + \nabla g^T f$$

For the convenience of notation and type setting, $\mathcal{A}$ is A and $\mathcal{A}^*$ is A^T Where, $\nabla f, \nabla g$ are the Jacobian of the function f, g. Therefore, the Gradient is going to be like:

$$\begin{aligned} \nabla f(x) &= \nabla \left[\frac{1}{2} x^T A x + v^T x + c \right] \\ &= \frac{1}{2} \nabla \left[x^T A x \right] + v \\ &= \frac{1}{2} \nabla [x]^T A x + \frac{1}{2} \nabla [Ax]^T x + v \\ &= \frac{1}{2} (A + A^T) x + v \\ &= \frac{1}{2} (2A) x + v \\ &= Ax + v \end{aligned} \tag{1.9.2.1}$$

The the gradient on the function $Ax + v$ is just A. The Gradient of a linear operator is just the operator itself, because it's a linear operator.

1.9.3

If $\mathcal{A}$ is PD(positive Definite) if and only if f is Coercive.

To prove this, we need to prove it both directions.

First assuming that the matirx A is PD, then we wish to show that the function $f(x)$ is coercive. Start by considering the $\langle x, Ax \rangle$ part of the function which we have:

$$\begin{aligned} &= \frac{1}{2} \langle x, Q \Lambda Q^T x \rangle \\ &= \frac{1}{2} \langle Q^T x, \Lambda Q^T x \rangle \\ \text{Let: } y = Q^T x, x = Qy \text{ Then:} \\ &= \frac{1}{2} \langle y, \Lambda y \rangle \\ &= \frac{1}{2} \sum_{i=1}^{n} \lambda_i y^2 \end{aligned} \tag{1.9.3.1}$$

Using the fact that the matrix A is Positive Definite, we know that all of the eigenvalues of the matrix A is positive, and hence, the function $f(x)$ is a quadratic function in y, and all of the coefficient of the quadratic terms are positive. The remaining component $v^T x = v^T Q_y$, which is a linear function wrt to variable y. It's obvious that a function that is a quadratic with positive coefficients on the quardratic term is coercive (Do a completing square thing).

Nest we wish to show that if $f(x)$ is Coercive, then A is PD, which by contradictions, we want to show contradiction on the statement that if $f(x0)$ is coercive, then A is PD. Assuming that $f(x)$ is coercive but the matrix A is not PD. Not PD implies that:

$$\exists y : \frac{1}{2} \langle y, Ay \rangle \leq 0$$

Along the directioin of y consider $y = \alpha\hat{y}$ then:

$$\begin{aligned} f(x) &= \frac{\alpha^2}{2}\langle \hat{y}, A\hat{y}\rangle + v^T\hat{y} \\ &= \frac{\alpha^2}{2}\langle \hat{y}, A\hat{y}\rangle + \alpha v^T\hat{y} + c \end{aligned} \tag{1.9.3.2}$$

Then, by the fact that the matrix A is not positive definite, we have a quadratic function that has negative quadratic term, hence the function decreases to infinity, making it not coercive at all.

Exercise 1.12

Define 2 sets:

$$\begin{aligned} &\mathbb{R}^n_{++} = \{x \in \mathbb{R}^n : x_i > 0 \; \forall \; 1 \leq i \leq n\} \\ &\mathbf{S}^n_{++} := \{X \in \mathbf{S}_n : X \prec 0\} \end{aligned} \tag{1.12.1.0.1}$$

To denote the positive quadrant and all the PD matrices. Then define function f mapping from $\mathbb{R}^n_{++}$ to $\mathbb{R}$ and function $F(X)$ mapping from $\mathbf{S}^n_{++}$ to $\mathbb{R}$ which is:

$$\begin{aligned} f : \mathbb{R}^n_{++} &\mapsto \mathbb{R} = \sum_{i=1}^{n} -\log(x_i) \\ F : \mathbf{S}^n_{++} &\mapsto \mathbb{R} = -\log\det(X) \end{aligned} \tag{1.12.0.2}$$

1.12.1

Figure out the Gradient and Hessian of the function $f(x)$. We use e_i to denote the standard basis vector for $\mathbb{R}^n$

$$\begin{aligned} \nabla f(x) &= \nabla\left[\sum_{i=1}^{n} -\log(x_i)\right] \\ &= \sum_{i=1}^{n} -\nabla\left[\log(x_i)\right] \\ &= \sum_{i=1}^{n} -\left(\frac{1}{x_i}\right)e_i \end{aligned} \tag{1.12.2.1}$$

And the Hessian is just the gradient of the gradient giving us:

$$\begin{aligned} \nabla f(x) &= \nabla\left[\sum_{i=1}^{n} -\left(\frac{1}{x_i}\right)e_i\right] \\ &= \sum_{i=1}^{n} \nabla\left[\frac{-1}{x_i}e_i\right] \\ &= \sum_{i=1}^{n} \frac{1}{x_i^2}e_i e_i^T \\ &= \operatorname{diag}\left(\frac{1}{x_1^2}, \frac{1}{x_2^2}, \ldots, \frac{1}{x_n^2}\right) \end{aligned} \tag{1.12.2.2}$$

1.12.2

We wish to prove that the gradient of the function $F(X)$ is:

$$\nabla F(X) = -X^{-1}$$

The key is to justify, firstly that the below equalty is true:

$$F(X+tV) - F(X) + t * \left\langle X^{-1}, V \right\rangle = -\log\det(I + tX^{-1/2}VX^{-1/2}) + t * \mathrm{Tr}(X^{-1/2}VX^{-1/2})$$

We here we are going to use the sevral identities about the matrix determinant and the trace of the matrices for the derivation:

$$\begin{aligned} &\det(AB) = \det(A)\det(B) \\ &\det(A^{-1}) = \det(A)^{-1} \\ &\langle A, B \rangle = \mathrm{Tr}(A^T B) \\ &\mathrm{Tr}(AB) = \mathrm{Tr}(BA) \end{aligned} \tag{1.12.3.1}$$

The justification for the equality goes like this:

$$\begin{aligned} F(X+tV) - F(x) &= -\log(\det(X+tV)) - \log\det(X) \\ &= \log\left(\frac{\det(X)}{\det(X+tV)}\right) \\ &= -\log\left(\frac{\det(X+tV)}{\det(X)}\right) \\ &= -\log\left(\det(X^{-1/2})\det(X+tV)\det(X^{-1/2})\right) \\ &= -\log\det(I + tX^{-1/2}VX^{-1/2}) \end{aligned} \tag{1.12.3.2}$$

Here we use the fact that the matrix X is Symmetric Definite, and hence it has the factorization of $X = X^{-1/2}X^{-1/2}$, where $X^{-1/2}$ is also a Positive Definite Matrix.
Where, the first 2 terms of the left hand side of the equation is the same as the first term on the right hand side of the equation, next we have:

$$\begin{aligned} t\left\langle X^{-1}, V \right\rangle &= t * \mathrm{Tr}(X^{-T}V) \\ &= t * \mathrm{Tr}(X^{-1/2}VX^{-1/2}) \end{aligned} \tag{1.12.3.3}$$

Therefore, the above equality is true, Next, we wish to prove that the RHS of the equation is $o(|t|)$.
We consider the substitution with $A = xX^{-1/2}VX^{-1/2}$ then the RHS of the equation becomes:

$$\begin{aligned} &= -\log\det(I + tX^{-1/2}VX^{-1/2}) + t * \mathrm{Tr}(X^{-1/2}VX^{-1/2}) \\ &= -\log\det(I + tA) + t * \mathrm{Tr}(A) \end{aligned} \tag{12.3.3.4}$$

For convenience, we use $\lambda_i[M]$ to denote the i Eigenvalues of the matrix, where $\lambda_n = \lambda_{\max}$ and λ_1 is $\lambda_{\min}$. Then:

$$\begin{aligned} &= -\log\left(\prod_{i=1}^{n}\lambda_i[I+tA]\right) + t * \mathrm{Tr}(A) \\ &\underset{[1]}{=} -\log\left(\prod_{i=1}^{n}\lambda_i[I+tA]\right) + t\sum_{i=1}^{n}\lambda_i[A] \\ &= -\sum_{i=1}^{n}\log\left(1 + t\lambda_i[A]\right) + t\sum_{i=1}^{n}\lambda_i[A] \end{aligned} \tag{12.3.3.5}$$

[1]: Here we use the fact the the sume of the eigenvalues are the Trace of the matrix which is easy to deduce using the Canonical Jordan Form decomposition of any Matrices. [1]

[1] $\mathrm{Tr}(XJX^{-1}) = \mathrm{Tr}(X^{-1}XJ) = \prod_{i=1}^{n}\lambda_i[J]$

We can take the limit of the last part of the expression 12.3.3.5 using the Lopital's Rule and have:

$$\begin{aligned}
\lim_{t\to 0}\frac{-\sum_{i=1}^{n}\log(1+t\lambda_i[A])}{t} &= \lim_{t\to 0} -\sum_{i=1}^{n}\frac{\lambda_i[A]}{1+t\lambda_i[A]} \\
&= -\sum_{i=1}^{n}\lambda_i[A] \\
\lim_{t\to 0}\frac{t\sum_{i=1}^{n}\lambda_i[A]}{t} &= \sum_{i=1}^{n}\lambda_i[A] \\
\implies \lim_{t\to 0}\frac{\sum_{i=1}^{n}\log\left(1+t\lambda_i[A]\right)+t\sum_{i=1}^{n}\lambda_i[A]}{t} &= 0
\end{aligned} \tag{12.3.3.6}$$

Therefore, we have shown that the RHS of the equation is zero if we take the Limit, and taking the limit on the Left hand side we will obetain the direction derivative on the direction V for the function $F(X)$, which is:

$$\begin{aligned}
\lim_{t\to 0}\frac{F(X+tV)-F(X)+t*\left\langle X^{-1},V\right\rangle}{t} &= 0 \\
\lim_{t\to 0}\frac{F(X+tV)-F(X)}{t} &= -\left\langle X^{-1},V\right\rangle \\
\nabla F(X)[V] &= -\left\langle X^{-1},V\right\rangle \\
\underset{[1]}{\implies} \nabla F(X) &= -X^{-1}
\end{aligned} \tag{12.3.3.7}$$

[1]: The direction derivative is analogus to derivative derivative where V is just a vector.
We have shown that the Gradient of $F(X)$ is $-X^{-1}$. $\square$
Next we wish to figure out the directional derivative on the gradient which is:

$$\begin{aligned}
\nabla^2 F(X)[V] &= \lim_{t\to 0}\frac{\nabla F(X+tV)-\nabla F(X)}{t} \\
&= \lim_{t\to 0}\frac{-(X+tV)^{-1}+X^{-1}}{t}
\end{aligned} \tag{12.3.3.8}$$

Take notice that:

$$\begin{aligned}
(X+tV)^{-1} &= X^{-1/2}(I+tX^{-1/2}VX^{-1/2})^{-1}X^{-1/2} \\
\text{Let: } A &= X^{-1/2}VX^{-1/2} \\
(I+tA)^{-1} &= \sum_{k=0}^{\infty}(-t)^k A^k \\
&= I-tA+\mathcal{O}(|t^2|\|A\|_{op}^2)
\end{aligned} \tag{12.3.3.9}$$

Take note that this inverse only equals to the series when the Matrix $I-A$ is invertible and the spectral radius of the matrix $I-A$ is strictly less than one.
And in that sense, we can make the susbtitution from equation above to the equation above above, giving us:

$$\begin{aligned}
&-(X+tV)^{-1}+X^{-1} \\
&= -X^{-1/2}(I+tX^{-1/2}VX^{-1/2})^{-1}X^{-1/2}+X^{-1} \\
&= X^{-1/2}(-I-tX^{-1/2}VX^{-1/2})^{-1}X^{-1/2}+X^{-1/2}IX^{-1/2} \\
&= X^{-1/2}(-(I+tA)^{-1}+I)X^{-1/2} \\
\implies X^{-1/2}\left(\lim_{t\to 0}\frac{(-I+tA)^{-1}+I}{t}\right)X^{-1/2} &= X^{-1/2}\left(\lim_{t\to 0}\frac{-I+tA+\mathcal{O}(|t|^2\|A\|_{op}^2)+I}{t}\right)X^{-1/2} \\
&= X^{-1/2}AX^{-1/2} = X^{-1/2}(X^{-1/2}VX^{-1/2})X^{-1/2} = X^{-1}AX^{-1}
\end{aligned} \tag{12.3.3.10}$$

Therefore when the operator norm of A is bouned, then the limit in 12.3.3.8 is approaching zero, proving the the direction derivative along the matrix V is $X^{-1}VX^{-1}$. $\square$

1.12.3

In this section we are going to show that the Direction Derivative of the Gradient in the direction of matrix V is a Positive Definite Operator. We want to show that:

$$\left\langle \nabla^2 F(X)[V], V \right\rangle = \|X^{-1/2}VX^{-1/2}\|_F^2$$

$$\begin{aligned} \left\langle X^{-1}VX^{-1}, V \right\rangle &= \left\langle X^{-1/2}X^{-1/2}VX^{-1}, V \right\rangle \\ &= \left\langle X^{-1/2}VX^{-1}, X^{-1/2}V \right\rangle \\ &= \left\langle X^{-1/2}VX^{-1/2}, X^{-1/2}VX^{-1/2} \right\rangle \\ &= \|X^{-1/2}VX^{-1/2}\|_F^2 \end{aligned} \tag{1.12.3.1}$$

The frobenius norm is induced by the Inner product defined for matrices. Therefore the last line follows directly from the second last line. Using the fact that the right hand side of the equation is always positive, we know that operator is positive definite.

1.13

Define $f : U \mapsto \mathbb{R}$ and 2 points $x, y \in U$. Define the univariate function $\varphi : [0,1] \mapsto \mathbb{R} := f(x + t(y - x))$ and let $x_t := x + t(y - x)$.

1.13.1

If the function f is C^1 smooth, then show the equality:

$$\varphi'(t) = \langle \nabla f(x_t), y - x \rangle \text{ holds for any } t \in (0,1)$$

The function $f(x)$ is C^1 smooth implies that the function's derivative is Lipschitz Continuous, and the second derivative is discontinuous. Consider:

$$\begin{aligned} \partial_t \varphi(t) &= \partial_t f(x + t(y - x)) \\ &= \nabla f(x + t(y - x))^T \partial_t (x + t(y - x)) \\ &= \nabla f(x + t(y - x))^T (y - x) \\ &= \langle \nabla f(x_t), y - x \rangle \end{aligned} \tag{1.12.1.1}$$

Choosing $t \in (0,1)$ will means that all these points interpolated by t are valid and they are in the domain of the function f.

1.13.2

Show that:

$$\varphi''(t) = \left\langle \nabla^2 f(x_t)(y - x), (y - x) \right\rangle$$

Firstly, taking the directly derivative through a multi-variabel function is:

$$\begin{aligned} \partial_t [\nabla f(x_t)] &= \partial_t [\nabla f(x + t(y - x))] \\ &= \nabla^2 f(x_t) \partial_t [x + t(y - x)] \\ &= \nabla^2 f(x_t)(y - x) \end{aligned} \tag{1.13.2.1}$$

Taking the scalar derivative through a multivariable function requires the Jacobian and the direction derivative (The Jacobian of the Gradient is the Hessian).
Straight from 1.12.1.1, we have:

$$\begin{aligned}\partial_t[\varphi'(t)] &= \langle \partial_t[\nabla f(x + t(y-x))], y - x\rangle \\ &= \left\langle \nabla^2 f(x_t)(y-x)), y - x\right\rangle\end{aligned} \tag{1.13.2.2}$$

The differential operator goes into the inner product because $y - x$ is a fixed constant wrt to t, and then we just make a substitution using the results from 1.13.2.1

1.16

The function $f : U \mapsto \bar{\mathbb{R}}$ where U is a convex set, if f is C^2 smooth, then f is β-smooth if and only if $\|\nabla^2 f(x)\|_{\text{op}} \leq \beta$.

To show that the statement is true, we need to show it in 2 ways. Firstly we wish to assume that the oiperator norm of the Hessian is bounded by a factor β, which means thta, for all x in U, the domain of the function, $\left\|\nabla^2 f(x)\right\|_{\text{op}}$ is bounded by the factor β. Let's start with a line with 2 points $x, y \in U$, and interpolated by a parameter $t \in (0, 1)$ where $x_t = x + t(y - x)$, then we make use of Taylor series on the gradient:

$$\nabla f(x_t) - \nabla f(x) = \int_0^t \nabla^2 f(x_s)(x_s - x)ds \tag{1.16.1}$$

$$\begin{aligned}\|\nabla f(x_t) - \nabla f(x)\| &= \left\| \int_0^t \nabla^2 f(x_s)(x_s - x)ds \right\| \\ &\leq \int_0^t \left\|\nabla^2 f(x_s)(x_s - x)\right\| ds \\ &\leq \int_0^t \left\|\nabla^2 f(x_s)\right\|_{\text{op}} \|x_s - x\| ds \\ &= \int_0^t \left\|\nabla^2 f(x_s)\right\|_{\text{op}} \|s(y - x)\| ds \\ &\leq t \max_{t\in[0,1]} \|\nabla^2 f(x_t)\|_{\text{op}} \|y - x\| \\ \|\nabla f(y) - \nabla f(x)\| &\leq \max_{t\in[0,1]} \left\|\nabla^2 f(x_t)\right\|_{\text{op}} \|y - x\|\end{aligned}$$

And in fact, we know that the operator norm is going to be bounded by β, and therefore, we arrive at the fact that the gradient is L-Continuous, which is one of the characterizations of beta smoothness of the function. □

The converse of the argument is arguing that if a function's gradient is L-Continuous, then the opperator norm is bounded by the same constant from L-continuity. Let's start with Taylor Expansion for the gradient of the function. We start of by choosing to points $x, y \in U$, where the direction of $y - x$ will be chosen later, and then we set $x_t = x + t(y - x)$, then consider this:

$$\|\nabla f(x_t) - \nabla f(x)\| \leq \beta \|x_t - x\| \tag{1.16.2}$$

At this point, I want to bring out the Taylor series:

$$\nabla f(x_t) = \nabla f(x) + \int_0^t \nabla^2 f(x_s)(y - x)ds \tag{1.16.3}$$

$$\nabla f(x_t) - \nabla f(x) = \int_0^t \nabla^2 f(x_s)(y - x)ds$$

Substituting 1.16.3 into 1.16.2, giving us:

$$\left\| \int_0^t \nabla^2 f(x_s)(y-x)ds \right\| \leq \beta \|x_t - x\| \tag{1.16.4}$$

$$\left\| \frac{\int_0^t \nabla^2 f(x_s)(y-x)ds}{x_t - x} \right\| \leq \beta$$

$$\left\| \lim_{t\to 0} \frac{\int_0^t \nabla^2 f(x_s)(y-x)ds}{x_t - x} \right\| \leq \beta$$

$$\left\| \lim_{t\to 0} \frac{\nabla^2 f(x_t)(y-x)}{y-x} \right\| \leq \beta \quad \text{Lopital's Rule}$$

Now we make the choice $(y-x)$ to be the direction of the maximal Eigenvalues for the Hessian matrix, then the limit on the LHS becomes the operation norm. Therefore, we arrive at the conclusion that:

$$\left\| \lim_{t\to 0} \frac{\nabla^2 f(x+tv)v}{tv} \right\| \leq \beta$$

Using the fact that the function f is C^2 smooth, then taking the limit will give us the Operator Norm of the matrix $\|\nabla^2 f(x)\|_{\text{op}} \leq \beta$ (the pertubation in the hessian will settle at the end.). Therefore, if the gradient of the function is L-Continuous, we know thta the operator is less than Beta, for every point $x \in U$.