Advanced Graph Algorithms and Optimization Spring 2023

Convex Optimization

R. Kyng & M. Probst *Problem Set 2 — Wednesday, March 1st*

These exercises will not count toward your grade, but you are encouraged to solve them all. This exercise sheet contains exercises relating to lectures in Week 2.

To get feedback, you must hand in your solutions by 23.59 pm on March 9th. Both hand-written and LaTeX solutions are acceptable, but we will only attempt to read legible text.

Exercise 1

Prove that if a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is symmetric, then $\|\boldsymbol{A}\| = \max(|\lambda_{\max}(\boldsymbol{A})|, |\lambda_{\min}(\boldsymbol{A})|)$ and give an example of a non-symmetric matrix for which this is not true.

Solution.

By definition, we have

$$\|\boldsymbol{A}\| = \max_{\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\|\boldsymbol{x}\|=\boldsymbol{1}} \|\boldsymbol{A}\boldsymbol{x}\|.$$

Based on the lecture notes we have that $\boldsymbol{A} = \boldsymbol{V}\Lambda\boldsymbol{V}^\top$, where the columns of $\boldsymbol{V}$ form an orthogonal basis and $\Lambda_{ii} = \lambda_i(\boldsymbol{A})$. Thus, we will have

$$\begin{aligned}
\|\boldsymbol{A}\|^2 &= \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\|^2 \\
&= \max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} \\
&= \max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^\top (\boldsymbol{V}\Lambda\boldsymbol{V}^\top \boldsymbol{V}\Lambda\boldsymbol{V}^\top)\boldsymbol{x} \\
&= \max_{\|\boldsymbol{x}\|=1} (\boldsymbol{V}^\top\boldsymbol{x})^\top \Lambda^2 (\boldsymbol{V}^\top\boldsymbol{x}).
\end{aligned}$$

Furthermore, since $\boldsymbol{V}$ is orthogonal

$$\left\| \boldsymbol{V}^\top\boldsymbol{x} \right\| = (\boldsymbol{V}^\top\boldsymbol{x})^\top(\boldsymbol{V}^\top\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{x} = \|\boldsymbol{x}\|^2$$

which implies that $\|\boldsymbol{x}\| = \left\| \boldsymbol{V}^\top\boldsymbol{x} \right\|$.

Overall, we have

$$\|\boldsymbol{A}\|^2 = \max_{\|\boldsymbol{z}\|=1} \boldsymbol{z}^\top \Lambda^2 \boldsymbol{z} = \max_{\|\boldsymbol{z}\|=1} \sum_{i=1}^{n} \lambda_i^2 \boldsymbol{z}(i)^2$$

This implies $\|\boldsymbol{A}\|^2 = \max(\lambda_{\max}^2, \lambda_{\min}^2)$. Thus, we have $\|\boldsymbol{A}\| = \max(\lambda_{\max}, \lambda_{\min})$.

On the other hand, consider the matrix

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

which is not symmetric. We observe that the eigenvalues of $\boldsymbol{A}$ are zero, but $\|\boldsymbol{A}\| = 1$. Therefore, we get $\|\boldsymbol{A}\| \neq \max(\lambda_{\max}, \lambda_{\min})$.

Exercise 2

Consider a twice continuously differentiable function $f : S \to \mathbb{R}$, where $S \subset \mathbb{R}^n$ is a convex open set. Prove that f is β-gradient Lipschitz if and only if for all $\boldsymbol{x} \in S$ we have $\|\boldsymbol{H}_f(\boldsymbol{x})\| \leq \beta$.

Solution.

Assume that $S \subset \mathbb{R}^n$ is open and convex. Let us first prove that if f is β-gradient Lipschitz, then $\|\lambda_{\max}(\boldsymbol{H}_f(\boldsymbol{x}))\| \leq \beta$. For an arbitrary $\boldsymbol{x} \in S$ and any $\boldsymbol{\delta} \neq \boldsymbol{0} \in \mathbb{R}^n$ such that $\boldsymbol{x} + \boldsymbol{\delta} \in S$ we have

$$\boldsymbol{\nabla} f(\boldsymbol{x} + \boldsymbol{\delta}) = \boldsymbol{\nabla} f(\boldsymbol{x}) + \boldsymbol{H}_f(\boldsymbol{x})\boldsymbol{\delta} + q(\boldsymbol{\delta})$$

where $\lim_{\boldsymbol{\delta} \to \boldsymbol{0}} \frac{\|q(\boldsymbol{\delta})\|}{\|\boldsymbol{\delta}\|} = 0$.

Now, we have

$$\|\boldsymbol{\nabla} f(\boldsymbol{x} + \boldsymbol{\delta}) - \boldsymbol{\nabla} f(\boldsymbol{x})\|_2 \leq \beta \|\boldsymbol{\delta}\|_2$$

By combing the aforementioned two inequalities, we get

$$\|\boldsymbol{H}_f(\boldsymbol{x})\boldsymbol{\delta} + q(\boldsymbol{\delta})\|_2 \leq \beta \|\boldsymbol{\delta}\|_2 \Rightarrow \frac{\|\boldsymbol{H}_f(\boldsymbol{x})\boldsymbol{\delta}\|_2}{\|\boldsymbol{\delta}\|_2} \leq \beta + \frac{\|q(\boldsymbol{\delta})\|_2}{\|\boldsymbol{\delta}\|_2}.$$

Since S is open, we can choose $\boldsymbol{\delta}$ such that

$$\frac{\|\boldsymbol{H}_f(\boldsymbol{x})\boldsymbol{\delta}\|_2}{\|\boldsymbol{\delta}\|_2} = \|\boldsymbol{H}_f(\boldsymbol{x})\|_2 .$$

Furthermore, since $\boldsymbol{H}_f(x)$ is symmetric, we have

$$\|\boldsymbol{H}_f(\boldsymbol{x})\| = \max(|\lambda_{\max}(\boldsymbol{H}_f(\boldsymbol{x}))|, |\lambda_{\min}(\boldsymbol{H}_f(\boldsymbol{x}))|).$$

Then, as $\boldsymbol{\delta}$ tends to zero, we have bounded spectral norm, $\|\boldsymbol{H}_f(\boldsymbol{x})\| \leq \beta$.

Now, we prove the other direction. Assume that $\|\boldsymbol{H}_f(\boldsymbol{x})\| \leq \beta$ for all $\boldsymbol{x} \in S$, except a measure zero set. Consider $\boldsymbol{x} \neq \boldsymbol{y} \in S$ and let $\boldsymbol{x}_\theta = \boldsymbol{x} + \theta(\boldsymbol{y} - \boldsymbol{x})$. Since S is convex, $\boldsymbol{x}_\theta \in S$ for any $\theta \in [0, 1]$. By the fundamental theorem of calculus,

$$\boldsymbol{\nabla} f(\boldsymbol{y}) = \boldsymbol{\nabla} f(\boldsymbol{x}) + \int_0^1 \boldsymbol{H}_f(\boldsymbol{x}_\theta)(\boldsymbol{y} - \boldsymbol{x}) d\theta.$$

Thus, we have

$$\begin{aligned}\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|_2 &= \left\| \int_0^1 \boldsymbol{H}_f(\boldsymbol{x}_\theta)(\boldsymbol{y}-\boldsymbol{x}) d\theta \right\|_2 \\ &\leq \int_0^1 \|\boldsymbol{H}_f(\boldsymbol{x}_\theta)(\boldsymbol{y}-\boldsymbol{x})\|_2 \, d\theta \\ &\leq \int_0^1 \|\boldsymbol{H}_f(\boldsymbol{x}_\theta)\| \, \|\boldsymbol{y}-\boldsymbol{x}\|_2 \, d\theta \\ &\leq \int_0^1 \beta \, \|\boldsymbol{y}-\boldsymbol{x}\|_2 \, d\theta \\ &= \beta \, \|\boldsymbol{y}-\boldsymbol{x}\|_2 .\end{aligned}$$

Exercise 3

Prove that when running Gradient Descent, $\|\boldsymbol{x}_i - \boldsymbol{x}^*\|_2 \leq \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2$ for all i.

Solution.

We assume f is β-Lipschitz continuous and convex, our prerequisites for gradient descent. Since f is convex, we have

$$f(\boldsymbol{x}_i) - f(\boldsymbol{x}^*) \leq \nabla f(\boldsymbol{x}_i)^\top (\boldsymbol{x}_i - \boldsymbol{x}^*).$$

By using $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \frac{1}{\beta}\nabla f(\boldsymbol{x}_i)$, we will get

$$f(\boldsymbol{x}_i) - f(\boldsymbol{x}^*) \leq \beta(\boldsymbol{x}_i - \boldsymbol{x}_{i+1})^\top (\boldsymbol{x}_i - \boldsymbol{x}^*).$$

We know that $2\boldsymbol{v}^\top \boldsymbol{u} = \|\boldsymbol{v}\|_2^2 + \|\boldsymbol{u}\|_2^2 - \|\boldsymbol{v}-\boldsymbol{u}\|_2^2$ for two vectors $\boldsymbol{v}, \boldsymbol{u}$. Thus, we have

$$\begin{aligned}f(\boldsymbol{x}_i) - f(\boldsymbol{x}^*) &\leq \frac{\beta}{2}(\|\boldsymbol{x}_i - \boldsymbol{x}_{i+1}\|_2^2 + \|\boldsymbol{x}_i - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}_{i+1} - \boldsymbol{x}^*\|_2^2) \\ &= \frac{1}{2\beta}\|\nabla f(\boldsymbol{x}_i)\|_2^2 + \frac{\beta}{2}(\|\boldsymbol{x}_i - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}_{i+1} - \boldsymbol{x}^*\|_2^2).\end{aligned}$$

Therefore, the above inequality yields

$$\|\boldsymbol{x}_{i+1} - \boldsymbol{x}^*\|_2^2 \leq \|\boldsymbol{x}_i - \boldsymbol{x}^*\|_2^2 - \frac{2}{\beta}(f(\boldsymbol{x}_i) - f(\boldsymbol{x}^*) - \frac{\|\nabla f(\boldsymbol{x}_i)\|_2^2}{2\beta}).$$

Since

$$f(\boldsymbol{x}_i) - f(\boldsymbol{x}^*) \geq f(\boldsymbol{x}_i) - f(\boldsymbol{x}_{i+1}) \geq \frac{\|\nabla f(\boldsymbol{x}_i)\|_2^2}{2\beta},$$

then

$$\|\boldsymbol{x}_{i+1} - \boldsymbol{x}^*\|_2^2 \leq \|\boldsymbol{x}_i - \boldsymbol{x}^*\|_2^2 .$$

Therefore, we can conclude that for any $i \geq 1$,

$$\|\boldsymbol{x}_{i+1} - \boldsymbol{x}^*\|_2 \leq \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2 .$$

Exercise 4

Prove the following theorem.

Theorem. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be an β-gradient Lipschitz, convex function. Let $\boldsymbol{x}_0$ be a given starting point, and let $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ be a minimizer of f. The Gradient Descent algorithm given by $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \frac{1}{\beta}\boldsymbol{\nabla} f(\boldsymbol{x}_i)$ ensures that the kth iterate satisfies*

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{2\beta \left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|_2^2}{k+1}.$$

Hint: do an induction on $1/gap_i$.

Solution.

Let $C = \beta \left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|_2^2$. We want to prove that for any integer $i \geq 0$

$$\text{gap}_i \leq \frac{2C}{i+1}. \tag{1}$$

In the following, we assume $\text{gap}_i > 0$ for all i: if this is ever violated, the algorithm has reached the optimum and will stay there as the gradient is then zero – and in that case Equation (1) holds.

We will prove Equation (1) by proving by induction that $\frac{1}{\text{gap}_i} \geq \frac{i+1}{2C}$. For the base case, using that f is β-gradient Lipschitz, we have

$$\text{gap}_0 = f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*) \leq \boldsymbol{\nabla} f(\boldsymbol{x}^*)^\top (\boldsymbol{x}_0 - \boldsymbol{x}^*) + \frac{\beta}{2}\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|_2^2 = \frac{\beta}{2}\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|_2^2 = C/2 \leq 2C.$$

This rearranges to $\frac{1}{\text{gap}_0} \geq \frac{0+1}{2C}$. From the lecture we know that

$$\text{gap}_{i+1} - \text{gap}_i \leq -\frac{\text{gap}_i^2}{2C}.$$

Since we can assume $\text{gap}_{i+1} > 0$, and we have $\text{gap}_i \geq \text{gap}_{i+1}$, diving through by $\text{gap}_i \cdot \text{gap}_{i+1}$ gives

$$\frac{1}{\text{gap}_i} - \frac{1}{\text{gap}_{i+1}} \leq -\frac{\text{gap}_i^2}{2C\text{gap}_i \cdot \text{gap}_{i+1}} \leq -\frac{1}{2C}.$$

Finally, we have that $\frac{1}{\text{gap}_{i+1}} \geq \frac{1}{2C} + \frac{1}{\text{gap}_i} \geq \frac{(i+1)+1}{2C}$ using the induction hypothesis.

Exercise 5

1. For each of the following functions answer these questions:
   - Is the function convex?
   - Is the function β-gradient Lipschitz for some β?
   - If the function is β-gradient Lipschitz give an upper bound on β – the bound should be within a factor 4 of the true value.

(a) $f(x) = |x|^{1.5}$ on $x \in \mathbb{R}$

(b) $f(x) = \exp(x)$ on $x \in \mathbb{R}$

(c) $f(x) = \exp(x)$ on $x \in (-1, 1)$

(d) $f(x, y) = \sqrt{x + y}$ on $(x, y) \in (0, 1) \times (0, 1)$.

(e) $f(x, y) = \sqrt{x + y}$ on $(x, y) \in (1/2, 1) \times (1/2, 1)$.

(f) $f(x, y) = \sqrt{x^2 + y^2}$ on $(x, y) \in \mathbb{R}^2$.

Solution.

(a) Function $f(x) = |x|^{1.5}$ on $x \in \mathbb{R}$ is convex but it is not β-gradient Lipschitz. We have

$$\boldsymbol{\nabla} f(x) = \frac{3}{2}|x|^{-\frac{1}{2}}x \qquad \text{and} \qquad \boldsymbol{H}_f(x) = \frac{3}{4}|x|^{-\frac{1}{2}}.$$

Since $\boldsymbol{H}_f(x)$ is positive semidefinite, function f is convex. However, it is not β-Lipschitz gradient for any β because $\|\boldsymbol{H}_f(x)\|_2$ tends to infinity when we let x go to zero, which implies that $\|\boldsymbol{H}_f(x)\|_2$ cannot be upper-bounded.

(b) Function $f(x) = \exp(x)$ on $x \in \mathbb{R}$ is convex but it is not β-gradient Lipschitz. We have $\boldsymbol{H}_f(x) = \exp(x)$. Function f is convex because $\exp(x) \geq 0$ for any $x \in \mathbb{R}$, but it is not β-gradient Lipschitz because $\exp(x)$ is not bounded on $\mathbb{R}$.

(c) Function $f(x) = \exp(x)$ on $x \in (-1, 1)$ is convex and β-gradient Lipschitz for $\beta = \exp(1)$. We know that $\boldsymbol{H}_f(x) = \exp(x)$. Since $\exp(x) \geq 0$ for any $x \in (-1, 1)$, then it is convex. Furthermore, since $\exp(-1) < \exp(x) < \exp(1)$ for $x \in (-1, 1)$, then f is β-gradient Lipschitz for $\beta = \exp(1)$. If we set $x = 0$, then the eigenvalue of $\boldsymbol{H}_f(x)$ is equal to 1 which implies that $\beta \geq 1$ by using $\|\lambda_{\max}(\boldsymbol{H}_f(\boldsymbol{x}))\|_2 \leq \beta$ from Exercise 1. Therefore, the upper bound of $\exp(1)$ is within a factor $\exp(1) < 4$ from the true value (actually, $1 + \epsilon$ for any $\epsilon > 0$).

(d) Function $f(x, y) = \sqrt{x + y}$ on $(x, y) \in (0, 1) \times (0, 1)$ is neither convex nor β-gradient Lipschitz for any β. We have

$$\boldsymbol{H}_f(x, y) = \begin{bmatrix} -\frac{1}{4(x+y)^{3/2}} & -\frac{1}{4(x+y)^{3/2}} \\ -\frac{1}{4(x+y)^{3/2}} & -\frac{1}{4(x+y)^{3/2}} \end{bmatrix}.$$

For a non-zero vector $\boldsymbol{z} = [z_1, z_2]^\top$, we have $\boldsymbol{z}^\top \boldsymbol{H}_f(x, y)\boldsymbol{z} = -(z_1 + z_2)^2/4(x + y)^{3/2}$ which is negative for any $(x, y) \in (0, 1) \times (0, 1)$. Thus, f is not convex. Furthermore, it is not β-gradient Lipschitz because $\|\boldsymbol{H}_f(x, y)\|_2 = \frac{1}{2(x+y)^{3/2}}$ which is not bounded on $(0, 1) \times (0, 1)$.

(e) Function $f(x, y) = \sqrt{x + y}$ on $(x, y) \in (1/2, 1) \times (1/2, 1)$ is not convex but it is $\frac{1}{2}$-gradient Lipschitz. We know that

$$\boldsymbol{H}_f(x, y) = \begin{bmatrix} -\frac{1}{4(x+y)^{3/2}} & -\frac{1}{4(x+y)^{3/2}} \\ -\frac{1}{4(x+y)^{3/2}} & -\frac{1}{4(x+y)^{3/2}} \end{bmatrix}.$$

With the same argument as above, f is not convex. However, it is β-gradient Lipschitz for $\beta = 1/2$ because $\frac{1}{2^{5/2}} \leq \|\boldsymbol{H}_f(x, y)\|_2 = \frac{1}{2(x+y)^{3/2}} \leq \frac{1}{2}$ on $(x, y) \in (1/2, 1) \times (1/2, 1)$.

(f) Function $f(x,y) = \sqrt{x^2+y^2}$ on $(x,y) \in \mathbb{R}^2$ is convex but it is not β-gradient Lipschitz for any β. We have

$$\boldsymbol{H}_f(x,y) = (x^2+y^2)^{-\frac{3}{2}} \begin{bmatrix} y^2 & -xy \\ -xy & x^2 \end{bmatrix}$$

for $x, y \neq 0$ and

$$\lim_{(x,y)\to(0,0)} \boldsymbol{H}_f(x,y) = [\infty, \infty]^\top.$$

Furthermore, for any $\boldsymbol{z} = [z_1, z_2]^\top$, we have

$$\boldsymbol{z}^\top \boldsymbol{H}_f(x,y)\boldsymbol{z} = \frac{(z_1 y - z_2 x)^2}{(x^2+y^2)^{\frac{3}{2}}} \geq 0$$

for all $(x,y) \in \mathbb{R}^2$. Therefore, $\boldsymbol{H}_f(x,y)$ is positive semidefinite for any $(x,y) \in \mathbb{R}^2$, which implies that f is convex.

Bonus Exercise 6: Strongly Convex Functions

This longer exercise will teach you about *strongly convex* functions. In it, you will show that, with a small tweak, gradient descent quickly converges to a highly accurate solutions on these functions.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Assume f is twice continuously (Frechét) differentiable and that its first and second (Frechét) derivatives are integrable (basically, don't worry that weird stuff is happening with the derivatives). Assume that for all $\boldsymbol{x}$, we have for some constant $\mu > 0$, that $\lambda_{\min}(H_f(\boldsymbol{x})) \geq \mu$. When this holds, we say that f is *μ-strongly convex*.

Part A. Prove that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{\nabla} f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 .$$

Part B. Prove that there is value $L \in \mathbb{R}$ such that for all $\boldsymbol{x} \in \mathbb{R}^n$, we have $f(\boldsymbol{x}) \geq L$. In other words, the function is not unbounded below.

Part C. Prove that f is *strictly convex* as per Definition 3.2.8 in Chapter 3. Prove also that the minimizer $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ of f is unique.

Part D. Let $\boldsymbol{x}_0$ be a given starting point and $\boldsymbol{x}^*$ be the minimizer of f. Suppose we have an algorithm DecentDescent which takes a starting point $\boldsymbol{x}_0$, and a step count $t \in \mathbb{N}$. DecentDescent$(\boldsymbol{x}_0, t)$ runs for t steps and returns $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$ such that

$$f(\tilde{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \leq \frac{\gamma \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{t+1}$$

where $\gamma > 0$ is a positive number.

Assume that the *cost* of running DecentDescent for t steps is t. Explain how, with a total cost of at most $\frac{8\gamma}{\mu} \log(\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2 / \delta)$, we can produce a point $\widehat{\boldsymbol{x}} \in \mathbb{R}^n$ such that $\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^*\|_2 \leq \delta$ for $\delta > 0$.

Part E. Consider a function $h : \mathbb{R}^n \to \mathbb{R}$ which is both μ-strongly convex and β-gradient Lipschitz. Give an algorithm that returns $\boldsymbol{x}'$ with

$$h(\boldsymbol{x}') - h(\boldsymbol{x}^*) \leq \epsilon$$

by computing the gradient of h at at most $\frac{32\beta}{\mu} \log(2\beta \left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|_2^2 / \epsilon)$ points.

Solution.

Part A. Let $\boldsymbol{z}$ be an arbitrary vector in $\mathbb{R}^n$. Since $\boldsymbol{H}_f(\boldsymbol{z})$ is symmetric, by applying the Courant-Fischer theorem, we have

$$\lambda_{\min}(\boldsymbol{H}_f(\boldsymbol{z})) = \min_{\boldsymbol{u} \in \mathbb{R}^n, \boldsymbol{u} \neq \boldsymbol{0}} \frac{\boldsymbol{u}^\top \boldsymbol{H}_f(\boldsymbol{z}) \boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{u}}.$$

We know that $\lambda_{\min}(\boldsymbol{H}_f(\boldsymbol{z})) \geq \mu$. Therefore, for any $\boldsymbol{u} \in \mathbb{R}^n$

$$\boldsymbol{u}^\top \boldsymbol{H}_f(\boldsymbol{z}) \boldsymbol{u} \geq \mu \left\|\boldsymbol{u}\right\|_2^2. \tag{2}$$

(Note Equation (2) trivially holds for $\boldsymbol{u} = \boldsymbol{0}$.)

Furthermore, based on Taylor's theorem for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, there exits $\boldsymbol{z} \in \mathbb{R}^n$ such that

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{\nabla} f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2} (\boldsymbol{y} - \boldsymbol{x})^\top \boldsymbol{H}_f(\boldsymbol{z}) (\boldsymbol{y} - \boldsymbol{x}). \tag{3}$$

Combining Equations (2) and (3), we conclude that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{\nabla} f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2} \mu \left\|\boldsymbol{y} - \boldsymbol{x}\right\|_2^2.$$

Part B. Consider a fixed vector $\boldsymbol{x} \in \mathbb{R}^n$ with bounded $f(\boldsymbol{x})$ and $\boldsymbol{\nabla} f(\boldsymbol{x})$. From Part A, we know that for any $\boldsymbol{y} \in \mathbb{R}^n$

$$\begin{aligned}
f(\boldsymbol{y}) &\geq f(\boldsymbol{x}) + \boldsymbol{\nabla} f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2} \mu \left\|\boldsymbol{y} - \boldsymbol{x}\right\|_2^2 \\
&= f(\boldsymbol{x}) + \frac{1}{2} \mu \left(\boldsymbol{y} - \boldsymbol{x} + \frac{1}{\mu} \boldsymbol{\nabla} f(\boldsymbol{x}) \right)^2 - \frac{1}{2} \mu \left\| \frac{1}{\mu} \boldsymbol{\nabla} f(\boldsymbol{x}) \right\|_2^2 \\
&\geq f(\boldsymbol{x}) - \frac{1}{2\mu} \left\|\boldsymbol{\nabla} f(\boldsymbol{x})\right\|_2^2.
\end{aligned}$$

Since $\boldsymbol{x}$ is fixed, $L = f(\boldsymbol{x}) - \frac{1}{2\mu} \left\|\boldsymbol{\nabla} f(\boldsymbol{x})\right\|_2^2$ is a constant lower bound on $f(\boldsymbol{y})$.

Part C. Consider two arbitrary vectors $\boldsymbol{x} \neq \boldsymbol{y} \in \mathbb{R}^n$ and $\theta \in (0, 1)$. Let $\boldsymbol{z} = \theta \boldsymbol{x} + (1 - \theta) \boldsymbol{y}$, which implies that $\boldsymbol{x} - \boldsymbol{z} = (1 - \theta)(\boldsymbol{x} - \boldsymbol{y})$ and $\boldsymbol{y} - \boldsymbol{z} = \theta(\boldsymbol{y} - \boldsymbol{x})$. Now, by applying Part A we have

$$f(\boldsymbol{x}) \geq f(\boldsymbol{z}) + \boldsymbol{\nabla} f(\boldsymbol{z})^\top ((1 - \theta)(\boldsymbol{x} - \boldsymbol{y})) + \frac{\mu}{2} (1 - \theta) \left\|\boldsymbol{x} - \boldsymbol{y}\right\|_2^2 \tag{4}$$

and

$$f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + \boldsymbol{\nabla} f(\boldsymbol{z})^\top (\theta(\boldsymbol{y} - \boldsymbol{x})) + \frac{\mu}{2} \theta \left\|\boldsymbol{x} - \boldsymbol{y}\right\|_2^2 \tag{5}$$

By multiplying Equation (4) by θ and Equation (5) by $1-\theta$ and summing them up we get

$$\theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + \mu\theta(1-\theta) \|\boldsymbol{x}-\boldsymbol{y}\|_2^2 .$$

Since $\mu > 0$, $\theta \in (0,1)$, and $\boldsymbol{y} \neq \boldsymbol{x}$, the term $\mu\theta(1-\theta)\|\boldsymbol{y}-\boldsymbol{x}\|_2^2$ is strictly positive. Therefore, we have that

$$\theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{y}) > f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y})$$

which implies that f is strictly convex.

Next, we show that if $\arg\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ is non-empty, then it has exactly one element. Assume that f attains a minimum f^* at both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, where $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$. Since f is strictly convex, we have that

$$f(\frac{1}{2}\boldsymbol{x}_1 + \frac{1}{2}\boldsymbol{x}_2) < \frac{1}{2}f(\boldsymbol{x}_1) + \frac{1}{2}f(\boldsymbol{x}_2) = \frac{1}{2}f^* + \frac{1}{2}f^* = f^*.$$

This is a contradiction.

Part D. Set $t' = 8\gamma/\mu$. We observe that DecentDescent$(\boldsymbol{x}_0, t')$ returns $\boldsymbol{x}_{t'}$ such that

$$f(\boldsymbol{x}_{t'}) - f(\boldsymbol{x}^*) \leq \frac{\gamma\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{t'+1} \leq \frac{\mu\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{8}.$$

Furthermore, from Part A we know that

$$\frac{\mu}{2}\|\boldsymbol{x}_{t'} - \boldsymbol{x}^*\|_2^2 \leq f(\boldsymbol{x}_{t'}) - f(\boldsymbol{x}^*)$$

where we used $\boldsymbol{\nabla} f(\boldsymbol{x}^*) = \boldsymbol{0}$.

Combining the above two equations implies that

$$\frac{\mu}{2}\|\boldsymbol{x}_{t'} - \boldsymbol{x}^*\|_2^2 \leq \frac{\mu\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{8} \Rightarrow \|\boldsymbol{x}_{t'} - \boldsymbol{x}^*\|_2 \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2}{2}.$$

Now, by setting $\boldsymbol{x}_0$ to be $\boldsymbol{x}_{t'}$ and applying the above argument iteratively we can conclude that for any $k \geq 1$

$$\|\boldsymbol{x}_{kt'} - \boldsymbol{x}^*\|_2 \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2}{2^k}.$$

Thus, for $k' = \log(\frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2}{\delta})$, we have

$$\|\boldsymbol{x}_{k't'} - \boldsymbol{x}^*\|_2 \leq \delta.$$

Therefore, with a total cost of at most $t'k' = \frac{8\gamma}{\mu}\log(\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2/\delta)$, we can produce a point $\widehat{\boldsymbol{x}} = \boldsymbol{x}_{t'k'} \in \mathbb{R}^n$ such that $\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^*\|_2 \leq \delta$.

Part E. We know that after t steps of Gradient Descent, each of which requires one computation of the gradient of function h, we obtain a point $\boldsymbol{x}_t$ such that

$$h(\boldsymbol{x}_t) - h(\boldsymbol{x}^*) \leq \frac{2\beta\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{t+1}.$$

Notice that this is the same as DecentDescent from Part D for $\gamma = 2\beta$. Therefore, if we set $\delta = \sqrt{4\epsilon/\gamma}$, after

$$\begin{aligned} t' &= \frac{8\gamma}{\mu} \log\left(\frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2}{\delta}\right) \\ &= \frac{16\beta}{\mu} \log\left(\sqrt{\frac{\beta}{2\epsilon}} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2\right) \\ &= \frac{8\beta}{\mu} \log\left(\frac{\beta \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{2\epsilon}\right) \\ &\leq \frac{32\beta}{\mu} \log\left(\frac{2\beta \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{\epsilon}\right) \end{aligned}$$

steps, we obtain $\boldsymbol{x}'$ such that

$$\left\|\boldsymbol{x}' - \boldsymbol{x}^*\right\|_2 \leq \sqrt{\frac{2\epsilon}{\beta}}. \tag{6}$$

Furthermore, since h is β-gradient Lipschitz, we have

$$h(\boldsymbol{x}') - h(\boldsymbol{x}^*) \leq \boldsymbol{\nabla} h(\boldsymbol{x}^*)^\top (\boldsymbol{x}' - \boldsymbol{x}^*) + \frac{\beta}{2} \left\|\boldsymbol{x}' - \boldsymbol{x}^*\right\|_2^2 = \frac{\beta}{2} \left\|\boldsymbol{x}' - \boldsymbol{x}^*\right\|_2^2 \tag{7}$$

where we used that $\boldsymbol{\nabla} h(x^*) = \mathbf{0}$.

Combining Equation (6) and Equation (7) yields

$$h(\boldsymbol{x}') - h(\boldsymbol{x}^*) \leq \frac{\beta}{2} \left(\sqrt{\frac{2\epsilon}{\beta}}\right)^2 = \epsilon.$$

Thus, by computing the gradient of h for at most $\frac{32\beta}{\mu} \log(2\beta \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 / \epsilon)$ times, we can find $\boldsymbol{x}'$ such that $h(\boldsymbol{x}') - h(\boldsymbol{x}^*) \leq \epsilon$.