Advanced Graph Algorithms and Optimization
Graded Homework 1

Jonas Hübotter

April 24th, 2022

1 Strongly Convex Functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a μ-strongly convex and β-smooth[1] function that is twice continuously differentiable[2] and whose first and second order derivatives are integrable.

[1] I say *β-smooth* to mean *β-gradient Lipschitz* as I am more used to this wording.

[2] We use the notion of Frechét differentiability.

1.1 Part A

Lemma 1. *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a convex function that is continuously differentiable. Then,*

$$(\nabla h(x) - \nabla h(y))^\top (x - y) \geq 0. \tag{1}$$

Proof. We have,

$$\begin{aligned}
&(\nabla h(x) - \nabla h(y))^\top (x - y) \\
&= \nabla h(x)^\top x - \nabla h(x)^\top y - \nabla h(y)^\top x + \nabla h(y)^\top y \\
&= -\nabla h(x)^\top (y - x) - \nabla h(y)^\top (x - y).
\end{aligned}$$

Thus, it suffices to show,

$$\nabla h(x)^\top (y - x) + \nabla h(y)^\top (x - y) \leq 0.$$

By the first-order characterization of convexity, we have,[3]

[3] theorem 2.3.7

$$\begin{aligned}
h(y) &\geq h(x) + \nabla h(x)^\top (y - x) \quad \text{and} \\
h(x) &\geq h(y) + \nabla h(y)^\top (x - y).
\end{aligned}$$

Rearranging terms, we obtain,

$$\nabla h(x)^\top (y - x) + \nabla h(y)^\top (x - y) \leq h(y) - h(x) + h(x) - h(y) = 0.$$

□

1.2 Part B

We first prove the following lemma.[4]

[4] This is analogous to lemma 3.5 in [1], however, they use a different strategy in their proof.

Lemma 2. *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a convex function that is continuously differentiable and β-smooth. Then,*

$$h(y) \geq h(x) + \nabla h(x)^\top (y - x) + \frac{1}{2\beta} \|\nabla h(y) - \nabla h(x)\|_2^2. \tag{2}$$

Proof. Let $\phi_x(z) \doteq h(z) - \nabla h(x)^\top z$. Note $\nabla \phi_x(z) = \nabla h(z) - \nabla h(x)$. We have that ϕ_x is convex,[5]

[5] We show the first-order characterization of convexity.

$$\begin{aligned}
&\phi_x(z_1) + \nabla \phi_x(z_1)^\top (z_2 - z_1) \\
&= h(z_1) - \nabla h(x)^\top z_1 + \nabla h(z_1)^\top (z_2 - z_1) + \nabla h(x)^\top (z_1 - z_2) \\
&\leq h(z_2) - \nabla h(x)^\top z_2 && \text{using the first-order characterization of convexity for } h \\
&= \phi_x(z_2).
\end{aligned}$$

We also have that ϕ_x is β-smooth,

$$\begin{aligned}
\|\nabla \phi_x(z_1) - \nabla \phi_x(z_2)\|_2 &= \|\nabla h(z_1) - \nabla h(x) - \nabla h(z_2) + \nabla h(x)\|_2 \\
&= \|\nabla h(z_1) - \nabla h(z_2)\|_2 \\
&\leq \beta \|z_1 - z_2\|_2 . && \text{using that } h \text{ is } \beta\text{-smooth}
\end{aligned}$$

Thus,[6]

[6] proposition 3.3.3

$$\phi_x(z) \leq \phi_x(y) + \nabla \phi_x(y)^\top (z - y) + \frac{\beta}{2} \|z - y\|_2^2$$

and therefore,

$$\min_{z \in \mathbb{R}^n} \phi_x(z) \leq \min_{z \in \mathbb{R}^n} \phi_x(y) + \nabla \phi_x(y)^\top (z - y) + \frac{\beta}{2} \|z - y\|_2^2 .$$

We have $\min_{z \in \mathbb{R}^n} \phi_x(z) = \phi_x(x)$ as $\nabla \phi_x(x) = 0$ and ϕ_x is convex. In the lecture,[7] we have seen in an analogous argument that the right-hand side is minimized for $z = y - 1/\beta \nabla \phi_x(y)$. The inequality simplifies to,

[7] section 3.3.2

$$\begin{aligned}
h(x) - \nabla h(x)^\top x &= \min_{z \in \mathbb{R}^n} \phi_x(z) \\
&\leq \min_{z \in \mathbb{R}^n} \phi_x(y) + \nabla \phi_x(y)^\top (z - y) + \frac{\beta}{2} \|z - y\|_2^2 \\
&= h(y) - \nabla h(x)^\top y - \frac{1}{2\beta} \|\nabla \phi_x(y)\|_2^2 .
\end{aligned}$$

By reordering the terms, we obtain,

$$\begin{aligned}
h(y) &\geq h(x) + \nabla h(x)^\top (y - x) + \frac{1}{2\beta} \|\nabla \phi_x(y)\|_2^2 \\
&= h(x) + \nabla h(x)^\top (y - x) + \frac{1}{2\beta} \|\nabla h(y) - \nabla h(x)\|_2^2 ,
\end{aligned}$$

as desired. □

Lemma 3. *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a convex function that is continuously differentiable and β-smooth. Then,*

$$(\nabla h(x) - \nabla h(y))^\top (x - y) \geq \frac{1}{\beta} \|\nabla h(x) - \nabla h(y)\|_2^2 . \tag{3}$$

Proof. Recall from section 1.1 that

$$(\nabla h(x) - \nabla h(y))^\top (x - y) = -\nabla h(x)^\top (y - x) - \nabla h(y)^\top (x - y).$$

Using eq. (2), we obtain,

$$\begin{aligned}
&- \nabla h(x)^\top (y - x) - \nabla h(y)^\top (x - y) \\
&\geq h(x) - h(y) + \frac{1}{2\beta} \|\nabla h(y) - \nabla h(x)\|_2^2 \\
&\qquad + h(y) - h(x) + \frac{1}{2\beta} \|\nabla h(y) - \nabla h(x)\|_2^2 \\
&= \frac{1}{\beta} \|\nabla h(y) - \nabla h(x)\|_2^2 .
\end{aligned}$$

□

1.3 Part C

Lemma 4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable, μ-strongly convex, β-smooth function. Then,*

(1) $h(x) \doteq f(x) - \mu/2 \|x\|_2^2$ is a convex and, if $\beta \neq \mu$, $(\beta - \mu)$-smooth function; and

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

Proof of (1). Let us compute the Hessian H_h of h.

$$\begin{aligned}
H_h(x)(i,j) &= \frac{\partial^2}{\partial x(i)\, \partial x(j)} h(x) \\
&= \frac{\partial^2}{\partial x(i)\, \partial x(j)} \left(f(x) - \frac{\mu}{2} \|x\|_2^2 \right) \\
&= H_f(x)(i,j) - \frac{\mu}{2} \underbrace{\frac{\partial^2}{\partial x(i)\, \partial x(j)} \|x\|_2^2}_{=2} \\
&= H_f(x)(i,j) - \mu.
\end{aligned}$$

Thus, $H_h(x) = H_f(x) - \mu I$ for all $x \in \mathbb{R}^n$. In particular, if $\{\lambda_i\}_i$ are the eigenvalues of H_f, then $\{\lambda_i - \mu\}_i$ are the eigenvalues of H_h.[8]

To show that h is convex, it suffices to show that H_h is positive semi-definite and therefore that $\lambda_{\min}(H_h(x)) \geq 0$ for all $x \in \mathbb{R}^n$.[9] Using that f is μ-strongly convex, we have for all $x \in \mathbb{R}^n$,

$$\lambda_{\min}(H_h(x)) = \underbrace{\lambda_{\min}(H_f(x))}_{\geq \mu} - \mu \geq \mu - \mu = 0.$$

To show that h is $(\beta - \mu)$-smooth, it suffices to show that $\lambda_{\max}(H_h(x)) \leq \beta - \mu$ for all $x \in \mathbb{R}^n$.[10] Using that f is β-smooth, we

[8] Let $A \in \mathbb{R}^{n \times n}$ and $c \in \mathbb{R}$. Then, for any eigenvalue $\lambda \in \mathbb{R}$ of A and corresponding eigenvector $x \in \mathbb{R}^n$,

$$\begin{aligned}(A + cI)x &= Ax + cIx \\ &= \lambda x + cx = (\lambda + c)x.\end{aligned}$$

Hence, $\lambda + c$ is the eigenvalue of $A + cI$ corresponding to the eigenvector x.

[9] using theorem 3.2.9 and theorem 3.1.2

[10] using proposition 3.3.2

have for all $\boldsymbol{x} \in \mathbb{R}^n$,

$$\lambda_{\max}(\boldsymbol{H}_h(\boldsymbol{x})) = \underbrace{\lambda_{\max}(\boldsymbol{H}_f(\boldsymbol{x}))}_{\leq \beta} - \mu \leq \beta - \mu. \qquad \square$$

Proof of (2). We consider two cases. First, suppose $\beta = \mu$. We have,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{\beta}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \qquad \text{using exercise 19 (A) from the first problem set, where } f \text{ is } \beta\text{-strongly convex}$$

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2\beta} \|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|_2^2. \qquad \text{as we have shown for } \beta\text{-smooth } f \text{ in eq. (2)}$$

We obtain,

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

which is what we wanted to show.

Now, suppose $\beta \neq \mu$. Let $h(\boldsymbol{x}) \doteq f(\boldsymbol{x}) - \mu/2 \|\boldsymbol{x}\|_2^2$ be defined as in (1). Using our results from (1), h is convex and $(\beta - \mu)$-smooth. By eq. (3), we have

$$(\nabla h(\boldsymbol{x}) - \nabla h(\boldsymbol{y}))^\top (\boldsymbol{x} - \boldsymbol{y}) \geq \frac{1}{\beta - \mu} \|\nabla h(\boldsymbol{x}) - \nabla h(\boldsymbol{y})\|_2^2.$$

Note that $\nabla h(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) - \mu \boldsymbol{x}$. This gives us,

$$\begin{aligned}
&(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^\top (\boldsymbol{x} - \boldsymbol{y}) \\
&= (\nabla h(\boldsymbol{x}) - \nabla h(\boldsymbol{y}))^\top (\boldsymbol{x} - \boldsymbol{y}) + \underbrace{(\mu \boldsymbol{x} - \mu \boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y})}_{= \mu \|\boldsymbol{x} - \boldsymbol{y}\|_2^2} \\
&\geq \frac{1}{\beta - \mu} \|\nabla h(\boldsymbol{x}) - \nabla h(\boldsymbol{y})\|_2^2 + \mu \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \\
&= \frac{1}{\beta - \mu} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}) + \mu(\boldsymbol{y} - \boldsymbol{x})\|_2^2 + \mu \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \\
&= \frac{1}{\beta - \mu} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 - \frac{2\mu}{\beta - \mu} (\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^\top (\boldsymbol{x} - \boldsymbol{y}) \qquad \text{expanding the squared norm} \\
&\qquad + \frac{\beta \mu}{\beta - \mu} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2.
\end{aligned}$$

Rearranging the terms, we get,

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

Finally, multiplying both sides by $\frac{\beta-\mu}{\beta+\mu} > 0$, we obtain,

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{1}{\beta+\mu} \|\nabla f(x) - \nabla f(y)\|_2^2 + \frac{\beta\mu}{\beta+\mu} \|x - y\|_2^2. \quad \square$$

Lemma 5. *When f is β-smooth and μ-strongly convex, we always have $\mu \leq \beta$.*

Proof. Suppose for a contradiction that $\mu > \beta$. Recall that for any $x \in \mathbb{R}^n$, $\lambda_{\min}(H_f(x)) \geq \mu$ and $\lambda_{\max}(H_f(x)) \leq \beta$ as f is μ-strongly convex and β-smooth. But this yields,

$$\lambda_{\min}(H_f(x)) \geq \mu > \beta \geq \lambda_{\max}(H_f(x)). \quad \square$$

1.4 Part D

Lemma 6. *Let f be defined as in the beginning of this section. When using a version of gradient descent with $x_{i+1} \doteq x_i - \alpha \nabla f(x_i)$ for some $\alpha \in \mathbb{R}$, then*

$$\|x_{i+1} - x^*\|_2^2 \leq \left(1 - \frac{2\alpha\mu\beta}{\mu+\beta}\right) \|x_i - x^*\|_2^2 + \alpha\left(\alpha - \frac{2}{\mu+\beta}\right) \|\nabla f(x_i)\|_2^2, \tag{4}$$

where $x^* \in \arg\min_{x \in \mathbb{R}^n} f(x)$.

Proof. We have,

$$\begin{aligned} \|x_{i+1} - x^*\|_2^2 &= \|x_i - x^* - \alpha \nabla f(x_i)\|_2^2 \\ &= \|x_i - x^*\|_2^2 - 2\alpha \nabla f(x_i)^\top (x_i - x^*) + \alpha^2 \|\nabla f(x_i)\|_2^2. \end{aligned}$$

It suffices to show,

$$2\alpha \nabla f(x_i)^\top (x_i - x^*) \geq \frac{2\alpha}{\mu+\beta} \|\nabla f(x_i)\|_2^2 + \frac{2\alpha\mu\beta}{\mu+\beta} \|x_i - x^*\|_2^2.$$

Dividing by 2α, observe that this is precisely what we have proven in part (2) of lemma 4 where $x \doteq x_i$ and $y \doteq x^*$.[11] $\square$

[11] Note that $\nabla f(x^*) = 0$.

1.5 Part E

Lemma 7. *In the setting of lemma 6, we have for $\alpha \doteq 1/\beta$,*

$$\|x_k - x^*\|_2^2 \leq \exp\left(-\frac{\mu}{\beta} k\right) \|x_0 - x^*\|_2^2. \tag{5}$$

Proof. Unraveling the recurrence from eq. (4), we get,

$$\|x_k - x^*\|_2^2 \leq \left(1 - \frac{2\alpha\mu\beta}{\mu+\beta}\right)^k \|x_0 - x^*\|_2^2 + \alpha\left(\alpha - \frac{2}{\mu+\beta}\right) \|\nabla f(x_i)\|_2^2.$$

Plugging in $\alpha \doteq 1/\beta$, yields,

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

Using $\mu \leq \beta$, we have,

$$\frac{2\mu}{\mu+\beta} \geq \frac{\mu}{\beta} \quad \text{and} \quad \frac{2}{\mu+\beta} \geq \frac{1}{\beta}.$$

We obtain,

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

using that $1 + x \leq \exp(x)$ for all $x \in \mathbb{R}$

1.6 Part F

We will first show a result for the version of gradient descent we have seen in parts D and E. We will then improve on this result using acceleration. The proof of this statement is not required for our improved version.

Theorem 8. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a μ-strongly convex and β-smooth function that is twice continuously differentiable. Then, gradient descent with $\boldsymbol{x}_{i+1} \doteq \boldsymbol{x}_i - 1/\beta \nabla f(\boldsymbol{x}_i)$ yields an approximate solution $\boldsymbol{x}_k$ such that for any $\epsilon > 0$,*

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \epsilon$$

where $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x})$ and the gradient of f is computed at at most $\kappa \log(\beta \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 / 2\epsilon)$ points.[12]

[12] $\kappa \doteq \beta/\mu$ is the *condition number* of f.

Proof. First, note that during each iteration of the given scheme, the gradient of f is evaluated at exactly one point. It therefore suffices to bound the number of iterations until we get an ϵ-optimal solution.

As f is β-smooth, we have,

$$f(\boldsymbol{x}_k) \leq f(\boldsymbol{x}^*) + \nabla f(\boldsymbol{x}^*)^\top (\boldsymbol{x}_k - \boldsymbol{x}^*) + \frac{\beta}{2} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 .$$

Noting that $\nabla f(\boldsymbol{x}^*) = 0$ and rearranging the terms, we obtain,

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{\beta}{2} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 .$$

Using eq. (5), we get,

$$\leq \frac{\beta}{2} \exp\left(-\frac{k}{\kappa}\right) \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \overset{!}{\leq} \epsilon.$$

Solving the inequality for k, yields,

$$k \geq \kappa \log\left(\frac{\beta \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{2\epsilon}\right)$$

as desired. $\square$

We now show that we can improve the previous result using acceleration to only require order $\sqrt{\kappa}$ rather than order of κ iterations to converge to an ϵ-optimal solution.

Theorem 9. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a μ-strongly convex and β-smooth function that is twice continuously differentiable. Let $\boldsymbol{x}_0 \in \mathbb{R}$ be any initial guess. Then, the iterative scheme,*

$$\boldsymbol{y}_0 \doteq \boldsymbol{x}_0 \tag{6}$$

$$\boldsymbol{y}_{i+1} \doteq \boldsymbol{x}_i - \frac{1}{\beta}\nabla f(\boldsymbol{x}_i) \tag{7}$$

$$\boldsymbol{x}_{i+1} \doteq (1+\theta)\boldsymbol{y}_{i+1} - \theta \boldsymbol{y}_i \quad \text{for } \theta \doteq \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, \tag{8}$$

yields an approximate solution $\boldsymbol{y}_k$ such that for any $\epsilon > 0$,

$$f(\boldsymbol{y}_k) - f(\boldsymbol{x}^*) \leq \epsilon$$

where $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ and the gradient of f is computed at at most $\sqrt{\kappa}\log(\beta\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2/\epsilon)$ points.[13]

[13] The proof of this theorem is inspired by the lecture on accelerated gradient descent and section 3.7.1 of [1].

Note that the sequence $\{\boldsymbol{y}_i\}_i$ is similar to the gradient descent scheme that we have examined previously. We choose x_i as a convex combination of the previous and current best guess. Our approach will be to (1) upper bound $f(\boldsymbol{y}_i)$ by a function $\phi_i : \mathbb{R}^n \to \mathbb{R}$, of which we (2) show that $\phi_i(\boldsymbol{x})$ converges to $f(\boldsymbol{x})$ quickly.

We define ϕ_i iteratively,

$$\phi_0(\boldsymbol{x}) \doteq f(\boldsymbol{x}_0) + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2 \tag{9}$$

$$\phi_{i+1}(\boldsymbol{x}) \doteq (1-\gamma)\phi_i(\boldsymbol{x}) + \gamma\Big(f(\boldsymbol{x}_i) + \nabla f(\boldsymbol{x}_i)^\top(\boldsymbol{x} - \boldsymbol{x}_i) + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}_i\|_2^2\Big), \tag{10}$$

as the convex combination of itself and a second-order Taylor approximation of f at $\boldsymbol{x}_i$ where we write $\gamma \doteq 1/\sqrt{\kappa} = \sqrt{\mu/\beta}$ to simplify notation. It is easy to see that ϕ_i is convex.[14] Our analysis rests on the following two claims, which we will prove later.

[14] We will later show that ϕ_i is μ-strongly convex.

Claim 10 (Upper bound)**.** $f(\boldsymbol{y}_i) \leq \min_{\boldsymbol{x}\in\mathbb{R}^n} \phi_i(\boldsymbol{x})$.

Claim 11 (Fast convergence)**.** $\phi_i(\boldsymbol{x}) \leq f(\boldsymbol{x}) + (1-\gamma)^i(\phi_0(\boldsymbol{x}) - f(\boldsymbol{x}))$.

Proof of theorem 9. By claim 10, $f(\boldsymbol{y}_i) \leq \phi_i(\boldsymbol{x}^*)$ during all iterations i. Therefore,

$$\begin{aligned} f(\boldsymbol{y}_k) - f(\boldsymbol{x}^*) &\leq \phi_k(\boldsymbol{x}^*) - f(\boldsymbol{x}^*) \\ &\leq (1-\gamma)^k(\phi_0(\boldsymbol{x}^*) - f(\boldsymbol{x}^*)) && \text{using claim 11} \\ &= (1-\gamma)^k(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*) + \frac{\mu}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2). && \text{using the definition of } \phi_0\text{, eq. (9)} \end{aligned}$$

$$\leq (1-\gamma)^k \frac{\mu+\beta}{2} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \qquad \text{using } f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \leq \tfrac{\beta}{2}\|\boldsymbol{x}-\boldsymbol{x}^*\|_2^2 \text{ as } f \text{ is } \beta\text{-smooth, see the proof of theorem 24}$$

$$\leq (1-\gamma)^k \beta \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \qquad \text{using } \mu \leq \beta$$

$$\leq \exp\left(-\frac{k}{\sqrt{\kappa}}\right)\beta \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \overset{!}{\leq} \epsilon \qquad \text{using that } 1 + x \leq \exp(x) \text{ for all } x \in \mathbb{R}$$

Solving the inequality for k, yields,

$$k \geq \sqrt{\kappa} \log\left(\frac{\beta \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{\epsilon}\right)$$

as desired. □

It remains to prove the two claims.

Proof of claim 11. We prove the claim by induction on i. In the base case, $i = 0$, we immediately have,

$$f(\boldsymbol{x}) + (1-\gamma)^0(\phi_0(\boldsymbol{x}) - f(\boldsymbol{x})) = \phi_0(\boldsymbol{x}).$$

Let us now consider any fixed $i \in \mathbb{N}_0$ and suppose that the statement holds for i. We have,

$$\phi_{i+1}(\boldsymbol{x}) = (1-\gamma)\phi_i(\boldsymbol{x}) + \gamma\Big(f(\boldsymbol{x}_i) + \nabla f(\boldsymbol{x}_i)^\top(\boldsymbol{x}-\boldsymbol{x}_i) + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{x}_i\|_2^2\Big) \qquad \text{using the definition of } \phi_{i+1}\text{, eq. (10)}$$

$$\leq (1-\gamma)^{i+1}(\phi_0(\boldsymbol{x}) - f(\boldsymbol{x})) + (1-\gamma)f(\boldsymbol{x}) \qquad \text{using the induction hypothesis}$$

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

Finally, observe that

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

as f is μ-strongly convex. Noting that $(1-\gamma)f(\boldsymbol{x}) + \gamma f(\boldsymbol{x}) = f(\boldsymbol{x})$, completes the proof. □

To prove the final claim, we define $\boldsymbol{v}_i \doteq \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \phi_i(\boldsymbol{x})$ and $\phi_i^* \doteq \min_{\boldsymbol{x}\in\mathbb{R}^n} \phi_i(\boldsymbol{x})$.

Claim 12. $\phi_{i+1}^* \geq (1-\gamma)\phi_i^* + (1-\gamma)\nabla f(\boldsymbol{x}_i)^\top(\boldsymbol{x}_i - \boldsymbol{y}_i) + \gamma f(\boldsymbol{x}_i) - \frac{1}{2\beta}\|\nabla f(\boldsymbol{x}_i)\|_2^2.$

Proof of claim 10. We prove the claim by induction on i. In the base case, $i = 0$, we have,

$$f(\boldsymbol{y}_0) = f(\boldsymbol{x}_0) \leq \min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}_0) + \underbrace{\frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{x}_0\|_2^2}_{\geq 0} = \min_{\boldsymbol{x}\in\mathbb{R}^n} \phi_0(\boldsymbol{x}),$$

using that $\boldsymbol{y}_0 = \boldsymbol{x}_0$.

Let us now consider any fixed $i \in \mathbb{N}_0$ and suppose that the statement holds for i. By the β-smoothness of f, we have,

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

By the definition of $\boldsymbol{y}_{i+1}$, we have $\boldsymbol{y}_{i+1} - \boldsymbol{x}_i = -\nabla f(\boldsymbol{x}_i)/\beta$ and the inequality simplifies to,

$$\begin{aligned}
f(\boldsymbol{y}_{i+1}) &\leq f(\boldsymbol{x}_i) - \frac{1}{2\beta} \|\nabla f(\boldsymbol{x}_i)\|_2^2 \\
&= (1-\gamma) f(\boldsymbol{y}_i) - (1-\gamma) f(\boldsymbol{y}_i) + f(\boldsymbol{x}_i) - \frac{1}{2\beta} \|\nabla f(\boldsymbol{x}_i)\|_2^2 \\
&\leq (1-\gamma)\phi_i^* - (1-\gamma) f(\boldsymbol{y}_i) + f(\boldsymbol{x}_i) - \frac{1}{2\beta} \|\nabla f(\boldsymbol{x}_i)\|_2^2 && \text{using the induction hypothesis} \\
&= (1-\gamma)\phi_i^* + (1-\gamma)(f(\boldsymbol{x}_i) - f(\boldsymbol{y}_i)) + \gamma f(\boldsymbol{x}_i) - \frac{1}{2\beta} \|\nabla f(\boldsymbol{x}_i)\|_2^2
\end{aligned}$$

Using the first-order characterization of convexity, we have,

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

Combining the previous two inequalities, yields,

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

$f(\boldsymbol{y}_{i+1}) \leq \phi_{i+1}^*$ follows by claim 12. □

Claim 13. *We use the following simple observations for our proof of claim 12.*

(1) $\phi_i(\boldsymbol{x}) = \phi_i^* + \frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{v}_i\|_2^2$.

(2) $\boldsymbol{v}_{i+1} = (1-\gamma)\boldsymbol{v}_i + \gamma\Big(\boldsymbol{x}_i - \frac{1}{\mu}\nabla f(\boldsymbol{x}_i)\Big)$.

(3) $\boldsymbol{v}_i - \boldsymbol{x}_i = \frac{\boldsymbol{x}_i - \boldsymbol{y}_i}{\gamma}$.

Proof of claim 12. We have,

$$\begin{aligned}
\phi_{i+1}^* + \frac{\mu}{2} \|\boldsymbol{x}_i - \boldsymbol{v}_{i+1}\|_2^2 &= \phi_{i+1}(\boldsymbol{x}_i) && \text{using claim 13(1)} \\
&= (1-\gamma)\phi_i(\boldsymbol{x}_i) + \gamma f(\boldsymbol{x}_i) && \text{using the definition of } \phi_{i+1}\text{, eq. (10)} \\
&= (1-\gamma)\phi_i^* + (1-\gamma)\frac{\mu}{2} \|\boldsymbol{x}_i - \boldsymbol{v}_i\|_2^2 + \gamma f(\boldsymbol{x}_i). && \text{using claim 13(1)}
\end{aligned}$$

By rearranging the terms, we get,

$$\phi_{i+1}^* = (1-\gamma)\phi_i^* + (1-\gamma)\frac{\mu}{2} \|\boldsymbol{x}_i - \boldsymbol{v}_i\|_2^2 + \gamma f(\boldsymbol{x}_i) - \frac{\mu}{2} \|\boldsymbol{x}_i - \boldsymbol{v}_{i+1}\|_2^2 .$$

Using claim 13(2), we have,

$$\|\boldsymbol{x}_i - \boldsymbol{v}_{i+1}\|_2^2 = \left\| (1-\gamma)(\boldsymbol{x}_i - \boldsymbol{v}_i) + \frac{\gamma}{\mu}\nabla f(\boldsymbol{x}_i) \right\|_2^2$$

$$= (1-\gamma)^2 \|x_i - v_i\|_2^2 - 2(1-\gamma)\frac{\gamma}{\mu}\nabla f(x_i)^\top (v_i - x_i) + \frac{\gamma^2}{\mu^2}\|\nabla f(x_i)\|_2^2.$$

Combining the two equalities, we obtain,

$$\begin{aligned}\phi_{i+1}^* &= (1-\gamma)\phi_i^* + \underbrace{\gamma(1-\gamma)\frac{\mu}{2}\|x_i - v_i\|_2^2}_{\geq 0} \\ &\quad + \gamma f(x_i) + \gamma(1-\gamma)\nabla f(x_i)^\top (v_i - x_i) - \frac{1}{2\beta}\|\nabla f(x_i)\|_2^2 \\ &\geq (1-\gamma)\phi_i^* + \gamma f(x_i) + (1-\gamma)\nabla f(x_i)^\top (x_i - y_i) - \frac{1}{2\beta}\|\nabla f(x_i)\|_2^2, \qquad \text{using claim 13(3)}\end{aligned}$$

as desired. □

We finish by giving formal proofs of the statements in claim 13 even though they are similar to proofs we have seen in class and the weekly problem sets.

Proof of claim 13(1). We first show by induction on i that $H_{\phi_i}(x) = \mu I$ for all $x \in \mathbb{R}^n$ and $i \geq 0$. By a simple calculation, we have,

$$\nabla\phi_0(x) = \mu(x - x_0) \quad \text{and} \quad H_{\phi_0}(x) = \mu I.$$

Let us consider any fixed $i \in \mathbb{N}_0$ and suppose that the statement holds for i. Following from the definition of ϕ_{i+1}, we have,

$$\begin{aligned}\nabla\phi_{i+1}(x) &= (1-\gamma)\nabla\phi_i(x) + \gamma(\nabla f(x_i) + \mu(x - x_i)) \quad \text{and} \\ H_{\phi_{i+1}}(x) &= (1-\gamma)H_{\phi_i}(x) + \gamma\mu I.\end{aligned}$$

Using the induction hypothesis, we conclude,

$$H_{\phi_{i+1}}(x) = (1-\gamma)\mu I + \gamma\mu I = \mu I.$$

In particular, this shows that ϕ_i is μ-strongly convex.

Note that the highest-order term in ϕ_i must therefore be of order two. It is easy to see that any quadratic function that satisfies $H_{\phi_{i+1}}(x) = \mu I$ and $\phi_i^* = \min_{x\in\mathbb{R}^n} \phi_i(x)$, can be written as[15]

$$\phi_i(x) = \frac{\mu}{2}\|x - z\|_2^2 + \phi_i^*$$

for some $z \in \mathbb{R}^n$. We immediately see that ϕ_i is minimized by z, and hence, $z = v_i$. □

[15] Consider an arbitrary quadratic function $g(x) = x^\top Ax + x^\top b + c$ with minimum m and $H_g(x) = A + A^\top = \mu I$. Thus, $A = \mu/2 I$. Now, consider the function

$$\begin{aligned}h(x) &= \frac{\mu}{2}\|x - z\|_2^2 + m \\ &= \frac{\mu}{2}\|x\|_2^2 - \mu x^\top z + \frac{\mu}{2}\|z\|_2^2 + m.\end{aligned}$$

To get $h \equiv g$, we simply need to set $z = -b/\mu$. As z is the minimizer of both h and g, $c = \mu/2\|z\|_2^2 - m$ is uniquely determined using that the minimum of h and g is m.

Proof of claim 13(2). Recall,

$$\nabla\phi_{i+1}(x) = (1-\gamma)\nabla\phi_i(x) + \gamma(\nabla f(x_i) + \mu(x - x_i)).$$

Using claim 13(1), we get,

$$\begin{aligned}
&= (1-\gamma)\nabla\Big(\frac{\mu}{2}\|x - v_i\|_2^2 + \phi_i^*\Big) + \gamma(\nabla f(x_i) + \mu(x - x_i)) \\
&= (1-\gamma)\mu(x - v_i) + \gamma(\nabla f(x_i) + \mu(x - x_i)) \\
&= \mu x - \mu(1-\gamma)v_i - \mu\gamma x_i + \gamma\nabla f(x_i) \stackrel{!}{=} 0.
\end{aligned}$$

Solving the equation for x, yields,

$$x = (1-\gamma)v_i + \gamma x_i - \frac{\gamma}{\mu}\nabla f(x_i).$$

As ϕ_{i+1} is convex, x minimizes ϕ_{i+1}, and hence, $v_{i+1} = x$.[16] □

[16] As ϕ_{i+1} is μ-strongly convex, it is strictly convex, and therefore x is its unique minimizer.

Proof of claim 13(3). We prove the statement by induction on i. For $i = 0$, note that the minimizer v_0 of ϕ_0 is x and hence,

$$v_0 - x_0 = 0 = x_0 - y_0.$$ using that $x_0 = y_0$

Let us consider any fixed $i \in \mathbb{N}_0$ and suppose that the statement holds for i. We have,

$$\begin{aligned}
v_{i+1} - x_{i+1} &= (1-\gamma)v_i + \gamma x_i - \frac{1}{\gamma\beta}\nabla f(x_i) - x_{i+1} && \text{using claim 13(2) and the identity } \gamma/\mu = 1/\gamma\beta \\
&= \frac{1}{\gamma}x_i - \Big(\frac{1}{\gamma} - 1\Big)y_i - \frac{1}{\gamma\beta}\nabla f(x_i) - x_{i+1} && \text{using the induction hypothesis} \\
&= \frac{1}{\gamma}y_{i+1} - \Big(\frac{1}{\gamma} - 1\Big)y_i - x_{i+1} && \text{using the definition of } y_{i+1}, \; x_i = y_{i+1} + 1/\beta\nabla f(x_i) \\
&\stackrel{!}{=} \frac{x_{i+1} - y_{i+1}}{\gamma}.
\end{aligned}$$

Solving the equation for x_{i+1}, we obtain,

$$x_{i+1} = (1+\theta)y_{i+1} - \theta y_i \quad \text{for } \theta = \frac{1-\gamma}{\gamma+1} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1},$$

which coincides precisely with our choice of x_{i+1}. □

2 *A different kind of smoothness*

Definition 14. A *norm* on $\mathbb{R}^n$ is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ such that
(1) for every $a \in \mathbb{R}$ and $x \in \mathbb{R}^n$, $\|ax\| = |a|\,\|x\|$;
(2) for every $x, y \in \mathbb{R}^n$, $\|x + y\| \le \|x\| + \|y\|$; and
(3) for every $x \in \mathbb{R}^n$, $\|x\| = 0$ implies $x = \mathbf{0}$.

Definition 15. Given the norm $\|\cdot\|$ on $\mathbb{R}^n$ its *dual norm* $\|\cdot\|_*$ is defined as,

$$\|x\|_* \doteq \sup\{z^\top x \mid z \in \mathbb{R}^n, \|z\| = 1\}. \tag{11}$$

2.1 Part A

Lemma 16.
(1) *The supremum in the definition of the dual norm is obtained.*
(2) $\|\cdot\|_*$ *is a norm on* $\mathbb{R}^n$.
(3) $x^\top y \leq \|x\| \|y\|_*$.
(4) $(\|x\|_*)_* \leq \|x\|$.[17]

[17] The other direction holds too, but is not shown here.

Proof of (1). Let $B \doteq \{z \in \mathbb{R}^n \mid \|z\| = 1\} \subseteq \mathbb{R}^n$ be the unit ball and consider the linear functional,

$$f_x : B \to \mathbb{R}, z \mapsto z^\top x.$$

We want to show that $\operatorname{im} f_x$ has a supremum. By the completeness axiom, it is sufficient to show that $\operatorname{im} f_x$ is nonempty and bounded (as $\operatorname{im} f_x \subseteq \mathbb{R}$).

Note that $\operatorname{im} f_x \neq \emptyset$ follows from the simple observation that $B \neq \emptyset$.[18]

[18] We have that B is nonempty, as we have for any $x \in \mathbb{R}^n \setminus \{0\}$ that the unit vector $x/\|x\| \in B$.

To show that $\operatorname{im} f_x$ is bounded, recall that the unit ball B is bounded. Thus, it suffices to show that f_x is a bounded operator. For any $z \in \mathbb{R}^n$, we have,

$$|f_x(z)| = \left|z^\top x\right| \leq \|z\|_2 \|x\|_2 ,$$

using the Cauchy-Schwartz inequality. Now, recall that all norms on $\mathbb{R}^n$ are equivalent.[19] Using this fact, we obtain,

[19] In particular, there exists $C \in \mathbb{R}$ such that for all $x \in \mathbb{R}^n$, $\|x\|_2 \leq C \|x\|$.

$$\leq \underbrace{C \|x\|_2}_{\text{const.}} \|z\| ,$$

proving that f_x is a bounded operator and $\operatorname{im} f_x$ is bounded. □

Proof of (2). We check the three properties of a norm. We fix arbitrary $a \in \mathbb{R}$ and $x, y \in \mathbb{R}^n$.

(1) $\|ax\|_* = \sup_{\|z\|=1} az^\top x = \sup_{\|z\|=1} |a| z^\top x = |a| \sup_{\|z\|=1} z^\top x = |a| \|x\|_*$. using that $\sup z^\top x = \sup z^\top (-x)$ as $\|z\| = 1$ implies $\|-z\| = 1$

(2) $\|x+y\|_* = \sup_{\|z\|=1} z^\top (x+y) = \sup_{\|z\|=1} z^\top x + z^\top y$
$\leq \sup_{\|z\|=1} z^\top x + \sup_{\|z\|=1} z^\top y = \|x\|_* + \|y\|_*$. using that $\sup a + b \leq \sup a + \sup b$

(3) We prove the contrapositive of positive definiteness. Suppose $x \neq 0$. Then, using the unit vector $z \doteq x/\|x\|$, $\|x\|_* = \sup_{\|z\|=1} z^\top x \geq \|x\|_2^2/\|x\| > 0$. □

Proof of (3). Taking the unit vector $z \doteq x/\|x\|$, we get,

$$\|y\|_* = \sup_{\|z\|=1} z^\top y \geq \frac{x^\top y}{\|x\|}.$$

Rearranging the terms, yields the desired result. □

Proof of (4). First, note that the dual norm can be characterized equivalently as,

$$\|x\|_* = \sup_{\|z\|=1} z^\top x = \sup_{y \neq 0} \frac{y^\top x}{\|y\|}, \tag{12}$$

by taking the unit vector $z \doteq y/\|y\|$. Using this characterization, we obtain,

$$\begin{aligned}
(\|x\|_*)_* &= \sup_{z \neq 0} \frac{z^\top x}{\|z\|_*} \\
&= \sup_{z \neq 0} \frac{z^\top x}{\sup_{y \neq 0} \frac{y^\top z}{\|y\|}} \\
&= \sup_{z \neq 0} z^\top x \inf_{y \neq 0} \frac{\|y\|}{y^\top z} \\
&= \sup_{z \neq 0} \inf_{y \neq 0} \|y\| \frac{z^\top x}{y^\top z} \\
&\leq \inf_{y \neq 0} \|y\| \sup_{z \neq 0} \frac{z^\top x}{y^\top z}. && \text{using the max-min inequality}
\end{aligned}$$

Observe that when x and y are not linearly dependent, their fraction can be made arbitrarily large, and hence, in this case the supremum is ∞. If, on the other hand $y = \alpha x$ for some $\alpha \in \mathbb{R}^n$, then the fraction evaluates to $1/\alpha$. This observation yields,

$$= \alpha \|x\| \cdot \frac{1}{\alpha} = \|x\|,$$

where we used absolute homogeneity. □

2.2 *Part B*

Definition 17.

(1) Given any positive definite matrix M, the *Mahalanobis norm* is defined as $\|x\|_M \doteq \sqrt{x^\top M x}$.
(2) The *uniform norm* is defined as $\|x\|_\infty \doteq \max_i |x(i)|$.
(3) The *Manhattan norm* is defined as $\|x\|_1 \doteq \sum_i |x(i)|$.

Lemma 18.

(1) $(\|\cdot\|_M)_* = \|\cdot\|_{M^{-1}}$.
(2) $(\|\cdot\|_\infty)_* = \|\cdot\|_1$.

Proof of (1). As M is positive definite, it can be factorized uniquely[20] into $M = LL^\top$ where L is lower triangular with positive entries on the diagonal. We write $M^{1/2} \doteq L$. Also note that as M is positive definite, it is symmetric. We have for any $x \in \mathbb{R}^n$,

[20] by the Cholesky decomposition

$$(\|x\|_M)_* = \sup_{\|z\|_M=1} z^\top x.$$

We substitute $y \doteq M^{1/2}z$,[21]

$$\begin{aligned}
&= \sup_{\|y\|_2=1} x^\top M^{-1/2} y \\
&= \sup_{\|y\|_2=1} \left(M^{-1/2}x\right)^\top y \\
&\le \sup_{\|y\|_2=1} \left\|M^{-1/2}x\right\|_2 \underbrace{\|y\|_2}_{=1} = \left\|M^{-1/2}x\right\|_2 \\
&= \sqrt{\left(M^{-1/2}x\right)^\top M^{-1/2}x} = \sqrt{x^\top M^{-1}x} = \|x\|_{M^{-1}}.
\end{aligned}$$

[21] We have $z = M^{-1/2}y$ and

$$\begin{aligned}
\|z\|_M &= \sqrt{z^\top M z} \\
&= \sqrt{\left(M^{-1/2}y\right)^\top M M^{-1/2}y} \\
&= \sqrt{y^\top y} = \|y\|_2 .
\end{aligned}$$

using the Cauchy-Schwartz inequality

Moreover, for $y \doteq {}^{M^{-1/2}x}/\left\|M^{-1/2}x\right\|_2$, we have,[22]

$$\begin{aligned}
(\|x\|_M)_* &\ge \frac{\left\|M^{-1/2}x\right\|_2^2}{\left\|M^{-1/2}x\right\|_2} \\
&= \left\|M^{-1/2}x\right\|_2 = \|x\|_{M^{-1}}.
\end{aligned}$$

[22] Note that y is normalized to unit length, i.e., $\|y\|_2 = 1$.

Hence, $(\|x\|_M)_* = \|x\|_{M^{-1}}$. □

Corollary 19. *In particular, the euclidean norm $\|\cdot\|_2$ is self-dual.*

Proof. $(\|\cdot\|_2)_* = (\|\cdot\|_I)_* = \|\cdot\|_I = \|\cdot\|_2$. □

Proof of (2). We have,

$$\begin{aligned}
(\|\cdot\|_\infty)_* &= \sup_{\|z\|_\infty=1} z^\top x \\
&= \sup_{\max_i |z(i)|=1} z^\top x.
\end{aligned}$$

Clearly,

$$z(i) \doteq \begin{cases} 1 & x(i) \ge 0 \\ -1 & x(i) < 0 \end{cases}$$

is a least upper bound for $z^\top x$. To see this, suppose for a contradiction that there exists a $y \in \mathbb{R}^n$ such that $y^\top x > z^\top x$. But then, we must have for at least one coordinate i that $|y(i)| > 1$, contradicting $\|y\|_\infty = 1$. We obtain,

$$(\|\cdot\|_\infty)_* = z^\top x = \sum_i |x(i)| = \|x\|_1 .$$ □

2.3 Part C

Definition 20. Given a norm $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$, the *dual vector map* is a function $(\cdot)^\# : \mathbb{R}^n \to \mathbb{R}^n$ such that $\boldsymbol{x}^\top (\boldsymbol{x})^\# = \|\boldsymbol{x}\|$ and $\left\|(\boldsymbol{x})^\#\right\|_* = 1$.

We will often work with the dual vector map with respect to the dual norm of a given norm $\|\cdot\|$. We denote this dual vector map by $(\cdot)^\#_*$. Using the aforementioned properties, we have,

(1) $\boldsymbol{x}^\top (\boldsymbol{x})^\#_* = \|\boldsymbol{x}\|_*$ and

(2) $\left(\left\|(\boldsymbol{x})^\#_*\right\|_*\right)_* = \left\|(\boldsymbol{x})^\#_*\right\| = 1.$ using that $(\|\cdot\|_*)_* = \|\cdot\|$

Lemma 21.

(1) The dual vector map for $\|\cdot\|_{\boldsymbol{M}}$ is $(\boldsymbol{x})^\# \doteq \boldsymbol{M}\boldsymbol{x}/\sqrt{\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}}$ and unique.

(2) A (non-unique) dual vector map for $\|\cdot\|_1$ is given by,

$$(\boldsymbol{x})^\# (i) \doteq \begin{cases} 1 & \boldsymbol{x}(i) \geq 0 \\ -1 & \boldsymbol{x}(i) < 0. \end{cases} \tag{13}$$

(3) A (non-unique) dual vector map for $\|\cdot\|_\infty$ is given by,

$$(\boldsymbol{x})^\# (i) \doteq \begin{cases} 1 & i = j \text{ and } \boldsymbol{x}(i) \geq 0 \\ -1 & i = j \text{ and } \boldsymbol{x}(i) < 0 \\ 0 & \text{otherwise}, \end{cases} \tag{14}$$

where $j \in \arg\max_j |\boldsymbol{x}(j)|$ is arbitrary but fixed.

Note that in our analysis of the dual vector map, we exclude the case $\boldsymbol{x} = \boldsymbol{0}$, as any unit vector can be chosen as the dual vector to $\boldsymbol{0}$.[23]

[23] This would immediately imply that there are infinitely many dual vector maps.

Proof of (1). We have,

$$\boldsymbol{x}^\top (\boldsymbol{x})^\# = \frac{\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}}{\sqrt{\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}}} = \sqrt{\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}} = \|\boldsymbol{x}\|_{\boldsymbol{M}} \quad \text{and}$$

$$\left\|(\boldsymbol{x})^\#\right\|_{\boldsymbol{M}^{-1}} = \frac{\|\boldsymbol{M}\boldsymbol{x}\|_{\boldsymbol{M}^{-1}}}{\sqrt{\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}}} = \frac{\sqrt{(\boldsymbol{M}\boldsymbol{x})^\top \boldsymbol{M}^{-1}\boldsymbol{M}\boldsymbol{x}}}{\sqrt{\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}}} = \frac{\sqrt{\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}}}{\sqrt{\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}}} = 1.$$

as $\boldsymbol{M}$ is positive definite, it is also symmetric

It remains to show that this choice of $(\boldsymbol{x})^\#$ is unique. Consider the special case where $\boldsymbol{M} \doteq \boldsymbol{I}$.[24] As $\|\cdot\|_2$ is self-dual, we need that

[24] As we have seen, $\|\cdot\|_{\boldsymbol{I}} = \|\cdot\|_2$.

$$\left\|(\boldsymbol{x})^\#\right\|_2 = (\boldsymbol{x})^{\#^\top} (\boldsymbol{x})^\# \stackrel{!}{=} 1,$$

implying that $(\boldsymbol{x})^\#$ must have unit length. Then, to satisfy $\boldsymbol{x}^\top (\boldsymbol{x})^\# \stackrel{!}{=} \|\boldsymbol{x}\|_2$, we must have $(\boldsymbol{x})^\# = \boldsymbol{x}/\|\boldsymbol{x}\|_2$, which corresponds uniquely to our choice of the dual vector map for $\|\cdot\|_{\boldsymbol{I}}$. □

Proof of (2). We have,

$$\boldsymbol{x}^\top (\boldsymbol{x})^\# = \sum_i \boldsymbol{x}(i) \cdot (\boldsymbol{x})^\# (i) = \sum_i |\boldsymbol{x}(i)| = \|\boldsymbol{x}\|_1 \quad \text{and}$$

$$\left\|(x)^{\#}\right\|_{\infty} = \max_i \left|(x)^{\#}(i)\right| = 1.$$

Clearly, our choice of $(\cdot)^{\#}$ is not unique, as if x contains zeros, the coordinates of the dual vector map may be either positive or negative. □

Proof of (3). Observe that, by definition, $(x)^{\#}$ has only one non-zero coordinate. This coordinate corresponds precisely to the coordinate of x with the largest absolute value. We therefore have,

$$x^{\top}(x)^{\#} = \max_i |x(i)| = \|x\|_{\infty} \quad \text{and}$$
$$\left\|(x)^{\#}\right\|_1 = 1.$$

Again, our choice of $(x)^{\#}$ is not unique, as when x has multiple coordinates with maximal absolute value, any one of them can be selected by the dual vector map. □

2.4 Part D

Definition 22. A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is *β-smooth with respect to a norm $\|\cdot\|$* if for all $x, y \in \mathbb{R}^n$,

$$\|\nabla f(x) - \nabla f(y)\|_* \leq \beta \|x - y\|. \tag{15}$$

Lemma 23. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and β-smooth with respect to the norm $\|\cdot\|$. Then,*

$$f(y) \leq f(x) + \nabla f(x)^{\top}(y - x) + \frac{\beta}{2}\|y - x\|^2. \tag{16}$$

Proof. We fix any $x, y \in \mathbb{R}^n$. We define $g(\theta) \doteq f(x_\theta)$ where we let $x_\theta \doteq x + \theta(y - x)$. Note that $g(1) - g(0) = f(y) - f(x)$. We have,

$$\begin{aligned}
f(y) &= f(x) + g(1) - g(0) \\
&= f(x) + \int_0^1 \frac{dg(\theta)}{d\theta}\, d\theta && \text{by the fundamental theorem of calculus} \\
&= f(x) + \int_0^1 \nabla f(x_\theta)^{\top}(y - x)\, d\theta && \text{by the chain rule} \\
&= f(x) + \int_0^1 \nabla f(x)^{\top}(y - x)\, d\theta \\
&\quad + \int_0^1 (\nabla f(x_\theta) - \nabla f(x))^{\top}(y - x)\, d\theta \\
&= f(x) + \nabla f(x)^{\top}(y - x) + \int_0^1 (\nabla f(x_\theta) - \nabla f(x))^{\top}(y - x)\, d\theta
\end{aligned}$$

For the integrand, we obtain,

$$(\nabla f(x_\theta) - \nabla f(x))^{\top}(y - x) \leq \|\nabla f(x_\theta) - \nabla f(x)\|_* \|y - x\| \qquad \text{using } x^{\top}y \leq \|x\| \|y\|_*$$

$$\leq \beta \|x_\theta - x\| \|y - x\| \qquad \text{using that } f \text{ is } \beta\text{-smooth, eq. (15)}$$
$$= \theta\beta \|y - x\|^2 .$$

We get,

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \int_0^1 \theta\beta \|y - x\|^2 \, d\theta$$
$$= f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2} \|y - x\|^2 . \qquad \square$$

2.5 Part E

Theorem 24. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, convex, and β-smooth with respect to the norm $\|\cdot\|$. Then, gradient descent with*

$$x_{i+1} \doteq x_i - \frac{1}{\beta} \|\nabla f(x_i)\|_* (\nabla f(x_i))_*^\# \tag{17}$$

yields an approximate solution x_k such that for any $\epsilon > 0$,

$$f(x_k) - f(x^*) \leq \epsilon$$

where $x^* \in \arg\min_{x \in \mathbb{R}^n} f(x)$, ∇f and $(\cdot)_*^\#$ are evaluated at most $\mathcal{O}\left(\beta R^2/\epsilon\right)$ times and at most $\mathcal{O}\left(n\beta R^2/\epsilon\right)$ additional arithmetic operations are used. Here,

$$R \doteq \max_{\substack{x \in \mathbb{R}^n \\ f(x) \leq f(x_0)}} \|x - x^*\| . \tag{18}$$

Proof. We will show that $k = \mathcal{O}\left(\beta R^2/\epsilon\right)$ is sufficient. Clearly, by the choice of the update rule, each iteration computes the gradient and dual vector only once. As we work with vectors in n dimensions, the addition and scalar multiplications take $\mathcal{O}(n)$ time per iteration.

By β-smoothness of f, we have,

$$f(x_{i+1}) \leq f(x_i) + \nabla f(x_i)^\top (x_{i+1} - x) + \frac{\beta}{2} \|x_{i+1} - x\|$$
$$= f(x_i) - \frac{1}{\beta} \|\nabla f(x_i)\|_* \underbrace{\nabla f(x_i)^\top (\nabla f(x_i))_*^\#}_{= \|\nabla f(x_i)\|_*}$$
$$+ \frac{1}{2\beta} \|\nabla f(x_i)\|_*^2 \underbrace{\left\| (\nabla f(x_i))_*^\# \right\|^2}_{=1}$$
$$= f(x_i) - \frac{1}{2\beta} \|\nabla f(x_i)\|_*^2 . \tag{19}$$

The remainder of the proof is analogous to the proof of gradient descent in $\|\cdot\|_2$ we have seen in the lecture and the exercises. We define $\text{gap}_i \doteq f(x_i) - f(x^*)$. We have,

$$\text{gap}_i = f(x_i) - f(x^*) \leq \nabla f(x_i)^\top (x_i - x^*) \qquad \text{using the first-order characterization of convexity, } f(x^*) \geq f(x_i) + \nabla f(x_i)^\top (x^* - x_i)$$

$$\begin{aligned} &\leq \|\nabla f(x_i)\|_* \|x_i - x^*\| \\ &\leq R \|\nabla f(x_i)\|_*, \end{aligned} \tag{20}$$

where we note that for all i, $f(x_i) \leq f(x_0)$. We obtain,

$$\text{gap}_{i+1} - \text{gap}_i = f(x_{i+1} - x_i) \leq -\frac{1}{2\beta} \|\nabla f(x_i)\|_*^2 \leq -\frac{1}{2\beta} \left(\frac{\text{gap}_i}{R}\right)^2. \tag{21}$$

using eq. (19) and then rearranging eq. (20)

Claim 25. $f(x_k) - f(x^*) \leq \frac{2\beta R^2}{k+1}$.

Using this claim, solving

$$f(x_k) - f(x^*) \leq \frac{2\beta R^2}{k+1} \overset{!}{\leq} \epsilon$$

for k, yields $k = \Omega\left(\beta R^2/\epsilon\right)$. Thus, choosing $k = \mathcal{O}\left(\beta R^2/\epsilon\right)$ is sufficient. □

Proof of claim 25. We prove $1/\text{gap}_i \geq i+1/2\beta R^2$ analogously to the proof of exercise 15 on the first problem set by an induction on $1/\text{gap}_i$.[25] In the base case,

[25] We assume that $\text{gap}_i > 0$ for all i, as otherwise our algorithm has already converged to the optimal solution.

$$\begin{aligned} \text{gap}_0 = f(x_0 - f(x^*) &\leq \nabla f(x^*)^\top (x_0 - x^*) + \frac{\beta}{2} \|x_0 - x^*\|^2 \\ &= \frac{\beta}{2} \|x_0 - x^*\|^2 \leq 2\beta R^2, \end{aligned}$$

using that f is β-smooth and $\nabla f(x^*) = 0$

from which we obtain $1/\text{gap}_0 \geq 1/2\beta R^2$. Let us now consider an arbitrary but fixed $i \in \mathbb{N}_0$ and suppose the statement holds for i. Dividing eq. (21) by $\text{gap}_i \cdot \text{gap}_{i+1}$, yields,

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

using $\text{gap}_i \geq \text{gap}_{i+1}$

Rearranging and using the induction hypothesis, yields,

$$\frac{1}{\text{gap}_{i+1}} \geq \frac{1}{2\beta R^2} + \frac{1}{\text{gap}_i} \geq \frac{i+2}{2\beta R^2}.$$ □

2.6 Part F

Lemma 26. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and convex such that for all $x, y \in \mathbb{R}^n$,*

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2} \|y - x\|^2.$$

Then, f is β-smooth with respect to the norm $\|\cdot\|$, i.e.,

$$\|\nabla f(x) - \nabla f(y)\|_* \leq \beta \|x - y\|.$$

Proof. We adopt a similar approach to our proof of lemma 2. Let $\phi_x(z) \doteq f(z) - \left(f(x) + \nabla f(x)(z-x)^\top\right)$. Note that this yields, $\nabla\phi_x(z) = \nabla f(z) - \nabla f(x)$. We have that ϕ_x is convex,

$$\begin{aligned}
&\phi_x(z_1) + \nabla\phi_x(z_1)^\top(z_2 - z_1) \\
&= f(z_1) - f(x) - \nabla f(x)^\top(z_1 - x) \\
&\qquad + \nabla f(z_1)^\top(z_2 - z_1) - \nabla f(x)^\top(z_2 - z_1) \\
&\leq f(z_2) - f(x) - \nabla f(x)^\top(z_2 - x) && \text{using the first-order characterization of convexity for } f \\
&= \phi_x(z).
\end{aligned}$$

Using the β-smoothness of f, we have for any $y \in \mathbb{R}^n$,

$$\begin{aligned}
\phi_x(z) &= f(x) - \left(f(x) + \nabla f(x)(z-x)^\top\right) \\
&\leq f(y) + \nabla f(y)^\top(z-y) + \frac{\beta}{2}\|z-y\|^2 \\
&\qquad - \left(f(x) + \nabla f(x)(z-x)^\top\right).
\end{aligned}$$

Rearranging to group terms that depend on z, we obtain,

$$\begin{aligned}
&= f(y) - \left(f(x) + \nabla f(x)^\top(y-x)\right) \\
&\qquad + (\nabla f(y) - \nabla f(x))^\top(z-y) + \frac{\beta}{2}\|z-y\|^2.
\end{aligned}$$

As $\nabla\phi_x(x) = 0$ and ϕ_x is convex, $\min_{z\in\mathbb{R}^n}\phi_x(z) = \phi_x(x) = 0$. Taking the minimum of both sides of the previous inequality, we get,

$$\begin{aligned}
0 &= \min_{z\in\mathbb{R}^n}\phi_x(z) \\
&\leq f(y) - \left(f(x) + \nabla f(x)^\top(y-x)\right) \\
&\qquad + \min_{z\in\mathbb{R}^n}(\nabla f(y) - \nabla f(x))^\top(z-y) + \frac{\beta}{2}\|z-y\|^2 \\
&= f(y) - \left(f(x) + \nabla f(x)^\top(y-x)\right) \\
&\qquad + \min_{\delta\in\mathbb{R}^n}(\nabla f(y) - \nabla f(x))^\top\delta + \frac{\beta}{2}\|\delta\|^2.
\end{aligned}$$

Claim 27. *For any $z \in \mathbb{R}^n$, we have* $\min_{\delta\in\mathbb{R}^n} z^\top\delta + \frac{\beta}{2}\|\delta\|^2 = -\frac{1}{2\beta}\|z\|_*^2$.

Using this claim, rearranging the terms of the previous inequality gives,

$$f(y) \geq f(x) + \nabla f(x)^\top(y-x) + \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|_*^2. \tag{22}$$

Now, recall from section 1.1 that

$$(\nabla f(x) - \nabla f(y))^\top(x-y) = -\nabla f(x)^\top(y-x) - \nabla f(y)^\top(x-y).$$

Using eq. (22), we obtain,

$$\begin{aligned}
&\|\nabla f(x) - \nabla f(y)\|_* \|x - y\| \\
&\geq (\nabla f(x) - \nabla f(y))^\top (x - y) && \text{using } x^\top y \leq \|x\| \|y\|_* \\
&= -\nabla f(x)^\top (y - x) - \nabla f(y)^\top (x - y) \\
&\geq f(x) - f(y) + \frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|_*^2 \\
&\qquad + f(y) - f(x) + \frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|_*^2 \\
&= \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\|_*^2 .
\end{aligned}$$

Rearranging gives the desired inequality. □

Proof of claim 27. We will prove both directions separately. To see that $\min_{\delta \in \mathbb{R}^n} z^\top \delta + \frac{\beta}{2} \|\delta\|^2 \leq -\frac{1}{2\beta} \|z\|_*^2$, we choose $\delta \doteq -\frac{1}{\beta} \|z\|_* (z)_*^\#$, and obtain,[26]

[26] This is similar to our choice of the update rule of gradient descent from the previous section.

$$\begin{aligned}
z^\top \delta + \frac{\beta}{2} \|\delta\|^2 &= -\frac{1}{\beta} \|z\|_* \underbrace{z^\top (z)_*^\#}_{=\|z\|_*} + \frac{1}{2\beta} \|z\|_*^2 \underbrace{\left\| (z)_*^\# \right\|^2}_{=1} \\
&= -\frac{1}{2\beta} \|z\|_*^2 .
\end{aligned}$$

To see that $\min_{\delta \in \mathbb{R}^n} z^\top \delta + \frac{\beta}{2} \|\delta\|^2 \geq -\frac{1}{2\beta} \|z\|_*^2$, we bound,

$$\begin{aligned}
z^\top \delta + \frac{\beta}{2} \|\delta\|^2 &= -(-z)^\top \delta + \frac{\beta}{2} \|\delta\|^2 \\
&\geq -\|z\|_* \|\delta\| + \frac{\beta}{2} \|\delta\|^2 && \text{using } x^\top y \leq \|x\| \|y\|_* \\
&\geq \min_{\Delta \in \mathbb{R}} \underbrace{-\|z\|_* \Delta + \frac{\beta}{2} \Delta^2}_{\doteq \Phi_z(\Delta)} . && \text{choosing } \Delta \doteq \|\delta\| \text{ and minimizing}
\end{aligned}$$

Clearly, Φ_z is a quadratic with positive curvature, and hence, convex. We have that

$$\frac{d\Phi_z(\Delta)}{d\Delta} = -\|z\|_* + \beta\Delta \stackrel{!}{=} 0,$$

is solved for $\Delta = 1/\beta \|z\|_*$, which therefore is a minimizer of Φ_z. Substituting for this minimizer in our previous inequality, we obtain,

$$z^\top \delta + \frac{\beta}{2} \|\delta\|^2 \geq -\frac{1}{2\beta} \|z\|_*^2 .$$ □

2.7 Part G

Lemma 28. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function such that for all $x, y \in \mathbb{R}^n$,*

$$0 \leq y^\top H_f(x) y \leq \beta \|y\|^2 . \tag{23}$$

Then, f is convex and β-smooth with respect to the norm $\|\cdot\|$.

Proof. To show that f is convex, it suffices that $\boldsymbol{H}_f$ is positive semi-definite.[27] This corresponds precisely to the condition that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{y}^\top \boldsymbol{H}_f(\boldsymbol{x})\boldsymbol{y} \geq 0$. Thus, it only remains to show that f is also β-smooth.

[27] using theorem 3.2.9

Similarly to our proof of lemma 23, we employ the fundamental theorem of calculus. We fix arbitrary $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and let $g(\theta) \doteq f(\boldsymbol{x}_\theta)$ for $\boldsymbol{x}_\theta \doteq \boldsymbol{x} + \theta(\boldsymbol{y} - \boldsymbol{x})$. Analogously to the mentioned proof, we have,

$$
\begin{aligned}
f(\boldsymbol{y}) &= f(\boldsymbol{x}) + g(1) - g(0) \\
&= f(\boldsymbol{x}) + \int_0^1 \frac{dg(\theta)}{d\theta}\, d\theta && \text{by the fundamental theorem of calculus} \\
&= f(\boldsymbol{x}) + \int_0^1 \nabla f(\boldsymbol{x}_\theta)^\top (\boldsymbol{y} - \boldsymbol{x})\, d\theta && \text{by the chain rule} \\
&= f(\boldsymbol{x}) + \int_0^1 \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x})\, d\theta \\
&\qquad + \int_0^1 (\nabla f(\boldsymbol{x}_\theta) - \nabla f(\boldsymbol{x}))^\top (\boldsymbol{y} - \boldsymbol{x})\, d\theta \\
&= f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \int_0^1 (\nabla f(\boldsymbol{x}_\theta) - \nabla f(\boldsymbol{x}))^\top (\boldsymbol{y} - \boldsymbol{x})\, d\theta.
\end{aligned}
$$

Now, let us shift our attention to bounding the integrand. We define $h(\tau) \doteq \nabla f(\boldsymbol{x}_\tau)^\top (\boldsymbol{y} - \boldsymbol{x})$ where we let $\boldsymbol{x}_\tau \doteq \boldsymbol{x} + \tau(\boldsymbol{x}_\theta - \boldsymbol{x})$. Note that

$$
(\nabla f(\boldsymbol{x}_\theta) - \nabla f(\boldsymbol{x}))^\top (\boldsymbol{y} - \boldsymbol{x}) = h(1) - h(0).
$$

By the chain rule,

$$
\begin{aligned}
\frac{dh(\tau)}{d\tau} &= (\boldsymbol{x}_\theta - \boldsymbol{x})^\top \boldsymbol{H}_f(\boldsymbol{x}_\tau)(\boldsymbol{y} - \boldsymbol{x}) \\
&= \theta(\boldsymbol{y} - \boldsymbol{x})^\top \boldsymbol{H}_f(\boldsymbol{x}_\tau)(\boldsymbol{y} - \boldsymbol{x}).
\end{aligned}
$$

We obtain the bound,

$$
\begin{aligned}
h(1) - h(0) &= \int_0^1 \frac{dh(\tau)}{d\tau}\, d\tau && \text{by the fundamental theorem of calculus} \\
&= \int_0^1 \theta \underbrace{(\boldsymbol{y} - \boldsymbol{x})^\top \boldsymbol{H}_f(\boldsymbol{x}_\tau)(\boldsymbol{y} - \boldsymbol{x})}_{\leq \beta \|\boldsymbol{y} - \boldsymbol{x}\|^2}\, d\tau && \text{using the assumption} \\
&\leq \int_0^1 \theta\beta \|\boldsymbol{y} - \boldsymbol{x}\|^2\, d\tau \\
&= \theta\beta \|\boldsymbol{y} - \boldsymbol{x}\|^2 .
\end{aligned}
$$

Substituting this bound for the integrand, we obtain,

$$
\begin{aligned}
f(\boldsymbol{y}) &\leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \int_0^1 \theta\beta \|\boldsymbol{y} - \boldsymbol{x}\|^2\, d\theta \\
&= f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{\beta}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2 .
\end{aligned}
$$

Using lemma 26, we conclude that f is indeed β-smooth. □

2.8 Part H

We will consider the function,

$$m : \mathbb{R}^n \to R, \quad \boldsymbol{x} \mapsto \frac{1}{\lambda} \log\left(\sum_i \exp(\lambda \boldsymbol{x}(i))\right), \tag{24}$$

for some $\lambda > 0$. We will see that m is a well-behaved approximation to a slight variation of the uniform norm.

Lemma 29.
(1) $\max_i \boldsymbol{x}(i) \leq m(\boldsymbol{x}) \leq \frac{\log n}{\lambda} + \max_i \boldsymbol{x}(i)$.
(2) *m is convex and λ-smooth with respect to $\|\cdot\|_\infty$.*

Proof of (1). Fix any $\boldsymbol{x} \in \mathbb{R}^n$. We have,

$$\begin{aligned} m(\boldsymbol{x}) &= \frac{1}{\lambda} \log\left(\sum_i \exp(\lambda \boldsymbol{x}(i))\right) \\ &\leq \frac{1}{\lambda} \log\left(n \exp(\lambda \max_i \boldsymbol{x}(i))\right) \\ &= \frac{1}{\lambda}\left(\log n + \lambda \max_i \boldsymbol{x}(i)\right) \\ &= \frac{\log n}{\lambda} + \max_i \boldsymbol{x}(i). \end{aligned}$$

For the other direction,

$$\begin{aligned} m(\boldsymbol{x}) &= \frac{1}{\lambda} \log\left(\sum_i \exp(\lambda \boldsymbol{x}(i))\right) \\ &\geq \frac{1}{\lambda} \log\left(\exp(\lambda \max_i \boldsymbol{x}(i))\right) \\ &= \max_i \boldsymbol{x}(i). \end{aligned}$$

□

Proof of (2). First, we show that m is convex. To begin with, recall Hölder's inequality,

$$\sum_i |\boldsymbol{x}(i)\boldsymbol{y}(i)| \leq \left(\sum_i |\boldsymbol{x}(i)|^p\right)^{\frac{1}{p}} \left(\sum_i |\boldsymbol{y}(i)|^q\right)^{\frac{1}{q}}, \tag{25}$$

for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\frac{1}{p} + \frac{1}{q} = 1$. Fix any $\theta \in [0, 1]$. Then,

$$\begin{aligned} m(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) &= \frac{1}{\lambda} \log\left(\sum_i e^{\lambda(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y})}\right) \\ &= \frac{1}{\lambda} \log\left(\sum_i e^{\theta\lambda\boldsymbol{x}} e^{(1-\theta)\lambda\boldsymbol{y}(i)}\right) \\ &\leq \frac{1}{\lambda} \log\left(\left(\sum_i e^{\lambda\boldsymbol{x}(i)}\right)^{\theta} \left(\sum_i e^{\lambda\boldsymbol{y}(i)}\right)^{1-\theta}\right) \end{aligned}$$

using Hölder's inequality with $1/p \doteq \theta$ and $1/q = 1 - \theta$

$$= \theta \frac{1}{\lambda} \log\left(\sum_i e^{\lambda \boldsymbol{x}(i)}\right) + (1-\theta)\frac{1}{\lambda}\log\left(\sum_i e^{\lambda \boldsymbol{y}(i)}\right)$$
$$= \theta m(\boldsymbol{x}) + (1-\theta) m(\boldsymbol{y}).$$

To prove smoothness of m, we first compute its Hessian and then apply lemma 28. For the Hessian of m, we have for any fixed $\boldsymbol{x} \in \mathbb{R}^n$,

$$\begin{aligned}\boldsymbol{H}_m(\boldsymbol{x})(i,j) &= \frac{\partial^2}{\partial \boldsymbol{x}(i)\,\partial \boldsymbol{x}(j)} m(\boldsymbol{x}) \\ &= \frac{\partial}{\partial \boldsymbol{x}(i)}\left(\frac{1}{\lambda}\frac{\partial}{\partial \boldsymbol{x}(j)}\log\left(\sum_k e^{\lambda \boldsymbol{x}(k)}\right)\right).\end{aligned}$$

We compute,

$$\frac{\partial}{\partial \boldsymbol{x}(j)}\log\left(\sum_k e^{\lambda \boldsymbol{x}(k)}\right) = \frac{\frac{\partial}{\partial \boldsymbol{x}(j)}\sum_k e^{\lambda \boldsymbol{x}(k)}}{\sum_k e^{\lambda \boldsymbol{x}(k)}} = \frac{\lambda e^{\lambda \boldsymbol{x}(j)}}{\sum_k e^{\lambda \boldsymbol{x}(k)}}.$$ using the chain rule in each step

We write $\Phi \doteq \sum_k e^{\lambda \boldsymbol{x}(k)}$ and $\Phi_{-i} \doteq \Phi - e^{\lambda \boldsymbol{x}(i)}$. Then,

$$\begin{aligned}\boldsymbol{H}_m(\boldsymbol{x})(i,j) &= \frac{\partial}{\partial \boldsymbol{x}(i)}\frac{e^{\lambda \boldsymbol{x}(j)}}{\Phi} \\ &= \frac{\left(\frac{\partial}{\partial \boldsymbol{x}(i)} e^{\lambda \boldsymbol{x}(j)}\right)\cdot \Phi - e^{\lambda \boldsymbol{x}(j)} \cdot \frac{\partial}{\partial \boldsymbol{x}(i)}\Phi}{\Phi^2} && \text{using the quotient rule} \\ &= \frac{1}{\Phi^2}\begin{cases}\lambda e^{\lambda \boldsymbol{x}(i)}\Phi - \lambda e^{2\lambda \boldsymbol{x}(i)} & i = j \\ -\lambda e^{\lambda(\boldsymbol{x}(i)+\boldsymbol{x}(j))} & i \neq j\end{cases} \\ &= \frac{1}{\Phi^2}\begin{cases}\lambda e^{\lambda \boldsymbol{x}(i)}\Phi_{-i} & i = j \\ -\lambda e^{\lambda(\boldsymbol{x}(i)+\boldsymbol{x}(j))} & i \neq j.\end{cases}\end{aligned}$$

Fixing any $\boldsymbol{y} \in \mathbb{R}^n$, we have,

$$\begin{aligned}\boldsymbol{y}^\top \boldsymbol{H}_m(\boldsymbol{x})\boldsymbol{y} &= \sum_{i,j} \boldsymbol{H}_m(\boldsymbol{x})(i,j)\cdot \boldsymbol{y}(i)\cdot \boldsymbol{y}(j) \\ &\leq \lambda \sum_i \boldsymbol{y}(i)^2 e^{\lambda \boldsymbol{x}(i)} \cdot \frac{\Phi_{-i}}{\Phi^2} && \text{using that the off-diagonal entries of the Hessian are negative} \\ &\leq \lambda \left\|\boldsymbol{y}\right\|_\infty^2 \frac{1}{\Phi}\underbrace{\sum_i e^{\lambda \boldsymbol{x}(i)}}_{=\Phi} && \text{using } \tfrac{\Phi_{-1}}{\Phi} < 1 \text{ and } \boldsymbol{y}(i) \leq \|\boldsymbol{y}\|_\infty \\ &= \lambda \left\|\boldsymbol{y}\right\|_\infty^2. && \square\end{aligned}$$

2.9 *Part I*

We consider the flow problem on a weighted and undirected graph $G = (V, E, \boldsymbol{c})$ with incidence matrix $\boldsymbol{B}$ and $\boldsymbol{U} \doteq \operatorname{diag}_{e\in E}\boldsymbol{c}(e)$ for capacities $\boldsymbol{c} \in \mathbb{R}_{\geq 0}^{|E|}$:

$$\min_{\substack{\boldsymbol{f}\in\mathbb{R}^{|E|} \\ \boldsymbol{B}^\top \boldsymbol{f} = \boldsymbol{b}}} \left\|\boldsymbol{U}^{-1}\boldsymbol{f}\right\|_\infty$$

for demands $\boldsymbol{b} \in \mathbb{R}^{|V|}$. The flow problem can be characterized equivalently as,

$$\min_{\substack{\boldsymbol{d}\in\mathbb{R}^{|E|}\\ \boldsymbol{B}^\top \boldsymbol{d}=\boldsymbol{0}}} \left\| \boldsymbol{U}^{-1}(\tilde{\boldsymbol{f}}_0 + \boldsymbol{d}) \right\|_\infty$$

for any feasible flow $\tilde{\boldsymbol{f}}_0$, i.e., $\boldsymbol{B}^\top \tilde{\boldsymbol{f}}_0 = \boldsymbol{b}$. We can also characterize the problem as,

$$\min_{\boldsymbol{x}\in\mathbb{R}^{|E|}} \|\boldsymbol{f}_0 + \boldsymbol{P}\boldsymbol{x}\|_\infty$$

where $\hat{\boldsymbol{P}} \in \mathbb{R}^{|E|\times|E|}$ is a projection matrix such that for all $\boldsymbol{x} \in \mathbb{R}^{|E|}$ we have $\boldsymbol{B}^\top \hat{\boldsymbol{P}}\boldsymbol{x} = \boldsymbol{0}$ and for every circulation[28] $\boldsymbol{d}$ there exists an $\boldsymbol{x} \in \mathbb{R}^{|E|}$ so that $\hat{\boldsymbol{P}}\boldsymbol{x} = \boldsymbol{d}$. We let $\boldsymbol{P} \doteq \boldsymbol{U}^{-1}\hat{\boldsymbol{P}}\boldsymbol{U}$ and $\boldsymbol{f}_0 \doteq (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{U}^{-1}\tilde{\boldsymbol{f}}_0$. We write,

$$OPT \doteq \min_{\boldsymbol{x}\in\mathbb{R}^{|E|}} \|\boldsymbol{f}_0 + \boldsymbol{P}\boldsymbol{x}\|_\infty .$$

[28] A *circulation* is a vector $\boldsymbol{d} \in \mathbb{R}^{|E|}$ such that $\boldsymbol{B}^\top \boldsymbol{d} = \boldsymbol{0}$.

Because the uniform norm is not smooth, we will use a smooth approximation similar to the function m we have seen in the previous section to approximately solve the optimization problem using gradient descent. As a smooth approximation, we use,

$$s : \mathbb{R}^n \to R, \quad \boldsymbol{x} \mapsto \frac{1}{\lambda}\log\left(\frac{\sum_{e\in E}\exp(\lambda \boldsymbol{x}(e)) + \exp(-\lambda \boldsymbol{x}(e))}{2|E|}\right),$$

for some $\lambda > 0$, which is convex, $\mathcal{O}(\lambda)$-smooth with respect to $\|\cdot\|_\infty$, and satisfies,

$$\|\boldsymbol{x}\|_\infty \le s(\boldsymbol{x}) \le 2\frac{\log|E|}{\lambda} + \|\boldsymbol{x}\|_\infty . \tag{26}$$

We will therefore optimize the function $g(\boldsymbol{x}) \doteq s(\boldsymbol{f}_0 + \boldsymbol{P}\boldsymbol{x})$. Note that, as g is the composition of two convex functions, it is also convex.

Lemma 30. *g is $\mathcal{O}\left(\lambda \|\boldsymbol{P}\|^2_{\infty\to\infty}\right)$-smooth with respect to $\|\cdot\|_\infty$.*[29]

Proof. By $\mathcal{O}(\lambda)$-smoothness of s, we have for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$s(\boldsymbol{y}) \le s(\boldsymbol{x}) + \boldsymbol{\nabla} s(\boldsymbol{x})^\top(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mathcal{O}(\lambda)}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2_\infty .$$

Let us now fix any $\boldsymbol{x}', \boldsymbol{y}' \in \mathbb{R}^n$. We substitute $\boldsymbol{x} \doteq \boldsymbol{f}_0 + \boldsymbol{P}\boldsymbol{x}'$ and $\boldsymbol{y} \doteq \boldsymbol{f}_0 + \boldsymbol{P}\boldsymbol{y}'$. Note that by the chain rule, $\boldsymbol{\nabla} g(\boldsymbol{x}') = \boldsymbol{P}^\top \boldsymbol{\nabla} s(\boldsymbol{f}_0 + \boldsymbol{P}\boldsymbol{x}')$.[30] We obtain,

$$\begin{aligned} g(\boldsymbol{y}') &= s(\boldsymbol{f}_0 + \boldsymbol{P}\boldsymbol{y}') \\ &\le s(\boldsymbol{f}_0 + \boldsymbol{P}\boldsymbol{x}') + \boldsymbol{\nabla} s(\boldsymbol{f}_0 + \boldsymbol{P}\boldsymbol{x}')^\top \boldsymbol{P}(\boldsymbol{y}' - \boldsymbol{x}') + \frac{\mathcal{O}(\lambda)}{2}\left\|\boldsymbol{P}(\boldsymbol{y}' - \boldsymbol{x}')\right\|^2_\infty \end{aligned}$$

[29] Here, $\|\boldsymbol{A}\|_{\alpha\to\beta} \doteq \max_{\|\boldsymbol{x}\|_\alpha = 1}\|\boldsymbol{A}\boldsymbol{x}\|_\beta$ is the *operator norm* of $\boldsymbol{A}$ induced by norms $\|\cdot\|_\alpha$ on the input space and $\|\cdot\|_\beta$ on the output space. We have,

$$\|\boldsymbol{A}\boldsymbol{x}\|_\beta \le \|\boldsymbol{A}\|_{\alpha\to\beta}\|\boldsymbol{x}\|_\alpha . \tag{27}$$

[30] The chain rule says that for a function $g(\boldsymbol{x}) \doteq s(h(\boldsymbol{x}))$,

$$Dg(\boldsymbol{x}) = Df(h(\boldsymbol{x}))Dh(\boldsymbol{x}).$$

Moreover, $\boldsymbol{\nabla} g(\boldsymbol{x}) = (Dg(\boldsymbol{x}))^\top$. In this case, $h(\boldsymbol{x}) = \boldsymbol{f}_0 + \boldsymbol{P}\boldsymbol{x}$, so $Dh(\boldsymbol{x}) = \boldsymbol{P}$.

$$= g(\boldsymbol{x}') + (\underbrace{\boldsymbol{P}^\top \boldsymbol{\nabla} s(\boldsymbol{f}_0 + \boldsymbol{P}\boldsymbol{x}')}_{=\boldsymbol{\nabla} g(\boldsymbol{x}')})^\top (\boldsymbol{y}' - \boldsymbol{x}') + \frac{\mathcal{O}(\lambda)}{2} \left\|\boldsymbol{P}(\boldsymbol{y}' - \boldsymbol{x}')\right\|_\infty^2$$

$$\leq g(\boldsymbol{x}') + \boldsymbol{\nabla} g(\boldsymbol{x}')^\top (\boldsymbol{y}' - \boldsymbol{x}') + \frac{\mathcal{O}(\lambda)}{2} \|\boldsymbol{P}\|_{\infty\to\infty}^2 \left\|\boldsymbol{y}' - \boldsymbol{x}'\right\|_\infty^2. \qquad \text{using eq. (27)}$$

Applying lemma 26, completes the proof. □

We denote by $X^\star$ the set of vectors $\boldsymbol{x}^\star$ that are minimizers of g, i.e., for which we have $g(\boldsymbol{x}^\star) = \min_{\boldsymbol{x}\in\mathbb{R}^n} g(\boldsymbol{x})$.

Lemma 31. *For any $\epsilon > 0$ and $\hat{\boldsymbol{x}} \in \mathbb{R}^n$ such that $g(\hat{\boldsymbol{x}}) \leq g(\boldsymbol{x}^\star) + \frac{\epsilon}{2} OPT$, we have,*

$$\|\boldsymbol{f}_0 + \boldsymbol{P}\hat{\boldsymbol{x}}\|_\infty \leq (1+\epsilon) OPT$$

for some $\lambda = \Theta(\log|E|/\epsilon OPT)$.

Proof. We know,

$$\|\boldsymbol{f}_0 + \boldsymbol{P}\hat{\boldsymbol{x}}\|_\infty \leq s(\boldsymbol{f}_0 + \boldsymbol{P}\hat{\boldsymbol{x}}) = g(\hat{\boldsymbol{x}}) \leq g(\boldsymbol{x}^\star) + \frac{\epsilon}{2} OPT.$$

Thus, it suffices to show, $g(\boldsymbol{x}^\star) \leq (1 + \epsilon/2) OPT$. We have,

$$\begin{aligned} g(\boldsymbol{x}^\star) &= \min_{\boldsymbol{x}\in\mathbb{R}^n} g(\boldsymbol{x}) \\ &= \min_{\boldsymbol{x}\in\mathbb{R}^n} s(\boldsymbol{f}_0 + \boldsymbol{P}\boldsymbol{x}) \\ &\leq 2\frac{\log|E|}{\lambda} + \min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{f}_0 + \boldsymbol{P}\boldsymbol{x}\|_\infty \\ &= 2\frac{\log|E|}{\lambda} + OPT \\ &\overset{!}{\leq} \left(1 + \frac{\epsilon}{2}\right) OPT. \end{aligned}$$

Solving for λ, we obtain,

$$\lambda \geq \frac{4\log|E|}{\epsilon OPT}.$$

Hence, choosing $\lambda = \Theta(\log|E|/\epsilon OPT)$ is sufficient. □

2.10 Part J

It can be shown that,

$$R \doteq \max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ g(\boldsymbol{x})\leq g(\boldsymbol{x}_0)}} \|\boldsymbol{x} - \boldsymbol{x}^*\|_\infty = \max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ g(\boldsymbol{x})\leq g(\boldsymbol{x}_0)}} \min_{\boldsymbol{x}^\star\in X^\star} \|\boldsymbol{x} - \boldsymbol{x}^\star\|_\infty. \qquad (28)$$

Lemma 32.
(1) $\|\boldsymbol{f}_0\|_\infty \leq (1 + \|\boldsymbol{P}\|_{\infty\to\infty}) OPT.$

(2) For any $\boldsymbol{y}$ such that $g(\boldsymbol{y}) \leq g(\mathbf{0})$, we have,

$$g(\boldsymbol{y}) \leq (1 + \|\boldsymbol{P}\|_{\infty\to\infty})OPT + 2\frac{\log|E|}{\lambda}.$$

(3) $R = \mathcal{O}((1 + \|\boldsymbol{P}\|_{\infty\to\infty})OPT + \log|E|/\lambda)$ when $\boldsymbol{x}_0 \doteq \mathbf{0}$.

Proof of (1). Consider an optimal circulation $\boldsymbol{d}^* \in \mathbb{R}^{|E|}$. As the discussed optimization problems are all equivalent, we have,

$$\left\|\boldsymbol{U}^{-1}(\boldsymbol{f}_0 + \boldsymbol{d}^*)\right\|_\infty = OPT.$$

We have,

$$\begin{aligned}
\|\boldsymbol{f}_0\|_\infty &= \left\|(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{U}^{-1}\tilde{\boldsymbol{f}}_0\right\|_\infty \\
&= \left\|\boldsymbol{U}^{-1}\tilde{\boldsymbol{f}}_0 + \boldsymbol{U}^{-1}\hat{\boldsymbol{P}}\boldsymbol{d}^* - \boldsymbol{P}\boldsymbol{U}^{-1}\tilde{\boldsymbol{f}}_0 - \boldsymbol{U}^{-1}\hat{\boldsymbol{P}}\boldsymbol{d}^*\right\|_\infty \\
&\leq \left\|\boldsymbol{U}^{-1}\tilde{\boldsymbol{f}}_0 + \boldsymbol{U}^{-1}\hat{\boldsymbol{P}}\boldsymbol{d}^*\right\|_\infty + \left\|\boldsymbol{P}\boldsymbol{U}^{-1}\tilde{\boldsymbol{f}}_0 + \boldsymbol{U}^{-1}\hat{\boldsymbol{P}}\boldsymbol{d}^*\right\|_\infty && \text{using the triangle inequality} \\
&= \left\|\boldsymbol{U}^{-1}(\tilde{\boldsymbol{f}}_0 + \hat{\boldsymbol{P}}\boldsymbol{d}^*)\right\|_\infty + \left\|\boldsymbol{P}\boldsymbol{U}^{-1}(\tilde{\boldsymbol{f}}_0 + \boldsymbol{d}^*)\right\|_\infty. && \text{using } \hat{\boldsymbol{P}} = \boldsymbol{U}\boldsymbol{P}\boldsymbol{U}^{-1}
\end{aligned}$$

Recall that $\boldsymbol{d}^*$ was the result of the projection $\hat{\boldsymbol{P}}\boldsymbol{x}^*$ for some $\boldsymbol{x}^*$. Therefore, due to the idempotency of projections, $\hat{\boldsymbol{P}}\boldsymbol{d}^* = \hat{\boldsymbol{P}}^2\boldsymbol{x}^* = \hat{\boldsymbol{P}}\boldsymbol{x}^* = \boldsymbol{d}^*$, and we obtain,

$$\begin{aligned}
&= \left\|\boldsymbol{U}^{-1}(\tilde{\boldsymbol{f}}_0 + \boldsymbol{d}^*)\right\|_\infty + \left\|\boldsymbol{P}\boldsymbol{U}^{-1}(\tilde{\boldsymbol{f}}_0 + \boldsymbol{d}^*)\right\|_\infty && \text{using } \hat{\boldsymbol{P}} = \boldsymbol{U}\boldsymbol{P}\boldsymbol{U}^{-1} \\
&\leq \left\|\boldsymbol{U}^{-1}(\tilde{\boldsymbol{f}}_0 + \boldsymbol{d}^*)\right\|_\infty + \|\boldsymbol{P}\|_{\infty\to\infty}\left\|\boldsymbol{U}^{-1}(\tilde{\boldsymbol{f}}_0 + \boldsymbol{d}^*)\right\|_\infty && \text{using eq. (27)} \\
&= (1 + \|\boldsymbol{P}\|_{\infty\to\infty})\left\|\boldsymbol{U}^{-1}(\tilde{\boldsymbol{f}}_0 + \boldsymbol{d}^*)\right\|_\infty \\
&= (1 + \|\boldsymbol{P}\|_{\infty\to\infty})OPT. && \square
\end{aligned}$$

Proof of (2). We have,

$$\begin{aligned}
g(\boldsymbol{y}) \leq g(\mathbf{0}) = s(\boldsymbol{f}_0) &\leq 2\frac{\log|E|}{\lambda} + \|\boldsymbol{f}_0\|_\infty \\
&\leq 2\frac{\log|E|}{\lambda} + (1 + \|\boldsymbol{P}\|_{\infty\to\infty})OPT. && \square
\end{aligned}$$

Proof of (3). We have,

$$R = \max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ g(\boldsymbol{x})\leq g(\mathbf{0})}} \min_{\boldsymbol{x}^\star\in X^\star} \|\boldsymbol{x} - \boldsymbol{x}^\star\|_\infty.$$

Observe that given any $\boldsymbol{x}^\star \in X^\star$, we have for $\boldsymbol{y} \doteq \boldsymbol{P}\boldsymbol{x}^\star + \boldsymbol{x} - \boldsymbol{P}\boldsymbol{x}$ that $\boldsymbol{P}\boldsymbol{y} = \boldsymbol{P}\boldsymbol{x}^\star$ and therefore $\boldsymbol{y} \in X^\star$. This gives a feasible solution for the minimum,

$$\leq \max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ g(\boldsymbol{x})\leq g(\mathbf{0})}} \|\boldsymbol{x} - \boldsymbol{y}\|_\infty$$

$$
\begin{aligned}
&= \max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ g(\boldsymbol{x})\le g(\boldsymbol{0})}} \|\boldsymbol{P}\boldsymbol{x}-\boldsymbol{P}\boldsymbol{x}^\star\|_\infty \\
&= \max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ g(\boldsymbol{x})\le g(\boldsymbol{0})}} \|\boldsymbol{f}_0+\boldsymbol{P}\boldsymbol{x}-\boldsymbol{f}_0-\boldsymbol{P}\boldsymbol{x}^\star\|_\infty \\
&\le \max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ g(\boldsymbol{x})\le g(\boldsymbol{0})}} \|\boldsymbol{f}_0+\boldsymbol{P}\boldsymbol{x}\|_\infty + \|\boldsymbol{f}_0+\boldsymbol{P}\boldsymbol{x}^\star\|_\infty && \text{using the triangle inequality} \\
&\le \max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ g(\boldsymbol{x})\le g(\boldsymbol{0})}} g(\boldsymbol{x}) + \underbrace{g(\boldsymbol{x}^\star)}_{\le g(\boldsymbol{x})} \\
&\le 2 \max_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ g(\boldsymbol{x})\le g(\boldsymbol{0})}} g(\boldsymbol{x}) \\
&\le 2(1+\|\boldsymbol{P}\|_{\infty\to\infty})OPT + 4\frac{\log|E|}{\lambda}. && \square
\end{aligned}
$$

2.11 Part K

We choose $\hat{\boldsymbol{P}} \doteq \boldsymbol{I} - \boldsymbol{U}\boldsymbol{B}\boldsymbol{L}^{+}\boldsymbol{B}^\top$ with the Laplacian $\boldsymbol{L} \doteq \boldsymbol{B}^\top\boldsymbol{U}\boldsymbol{B}$.[31] Then,

$$\left\|\boldsymbol{U}^{-1}\hat{\boldsymbol{P}}\boldsymbol{U}\right\|_{\infty\to\infty} = \|\boldsymbol{P}\|_{\infty\to\infty} \le 1 + 8\frac{\log|E|}{\Phi^2},$$

where Φ is the expansion of the graph.

[31] Note that the definition of the incidence matrix used here is the transpose of the incidence matrix as defined in the lecture notes.

Theorem 33. *Gradient descent with respect to g and $\|\cdot\|_\infty$ yields a $(1+\epsilon)$ approximate solution to OPT in time $\tilde{\mathcal{O}}(|E|/\epsilon^2\Phi^8)$ under the assumption that solving a Laplacian linear system exactly is as expensive as finding a $1/|V|^{100}$-approximate solution.*

Proof. First, recall that g is continuously differentiable, convex, and $\mathcal{O}\left(\lambda\|\boldsymbol{P}\|^2_{\infty\to\infty}\right)$-smooth with respect to $\|\cdot\|_\infty$. Using our analysis from theorem 24, gradient descent with respect to g and $\|\cdot\|_\infty$ will evaluate $\boldsymbol{\nabla} g$ and $(\cdot)^{\#}_{*}$ at most $\mathcal{O}\left(\beta R^2/\epsilon OPT\right)$ times and use at most $\mathcal{O}\left(|E|\beta R^2/\epsilon OPT\right)$ additional arithmetic operations to find an $\hat{\boldsymbol{x}}$ such that $g(\hat{\boldsymbol{x}}) - g(\boldsymbol{x}^*) \le \frac{\epsilon}{2}OPT$. By lemma 31, we know that this yields a $(1+\epsilon)$ approximation of OPT for some $\lambda = \Theta(\log|E|/\epsilon OPT)$.

We have for β,

$$
\begin{aligned}
\beta = \mathcal{O}\left(\lambda\|\boldsymbol{P}\|^2_{\infty\to\infty}\right) &= \mathcal{O}\left(\frac{\log|E|}{\epsilon OPT}\left(1+8\frac{\log|E|}{\Phi^2}\right)^2\right) \\
&= \mathcal{O}\left(\frac{\log|E|}{\epsilon OPT}\cdot\frac{\log^2|E|}{\Phi^4}\right) \\
&= \mathcal{O}\left(\frac{\log^3|E|}{\epsilon OPT\Phi^4}\right)
\end{aligned}
$$

and for R,

$$R = \mathcal{O}\left((1+\|\boldsymbol{P}\|_{\infty\to\infty})OPT + \frac{\log|E|}{\lambda}\right)$$

$$= \mathcal{O}\left(\left(2 + 8\frac{\log|E|}{\Phi^2} + \epsilon\right)OPT\right)$$
$$= \mathcal{O}\left(\frac{\log|E|}{\Phi^2}OPT\right).$$

Combining these bounds, we get,

$$\mathcal{O}\left(\frac{\beta R^2}{\epsilon OPT}\right) = \mathcal{O}\left(\frac{\log^5|E|}{\epsilon^2\Phi^8}\right) = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2\Phi^8}\right).$$

It therefore remains to show that each iteration of gradient descent takes $\tilde{\mathcal{O}}(|E|)$ time.

For a matrix $\boldsymbol{A}$, let $T(\boldsymbol{A})$ be the maximum time to compute $\boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{A}^\top\boldsymbol{x}$ for any vector $\boldsymbol{x}$. We use the following claim:

Claim 34. $T(\boldsymbol{P}) = \tilde{\mathcal{O}}(|E|)$.

By the chain rule, we have $\nabla g(\boldsymbol{x}) = \boldsymbol{P}^\top \nabla s(\boldsymbol{f}_0 + \boldsymbol{P}\boldsymbol{x})$. Thus, ∇g can be computed in time $T(\boldsymbol{P})$ plus the time to compute ∇s. It is not hard to show that ∇s can be computed in time $\mathcal{O}(|E|)$.[32] Finally, observe that $(\cdot)^{\#}_*$ corresponds to the dual vector map of the Manhattan norm, which we stated in eq. (13). Clearly, this mapping can be computed in $\mathcal{O}(|E|)$ time. Therefore, each iteration of gradient descent can be computed in $\tilde{\mathcal{O}}(|E|)$ time. □

[32] The argument is similar to our computation of the first-order derivative of the function m in the proof of lemma 29(2).

Proof of claim 34. We have,

$$\boldsymbol{P} = \boldsymbol{U}^{-1}\hat{\boldsymbol{P}}\boldsymbol{U} = \boldsymbol{I} - \boldsymbol{B}\boldsymbol{L}^+\boldsymbol{B}^\top\boldsymbol{U} \quad \text{and}$$
$$\boldsymbol{P}^\top = \boldsymbol{I} - \boldsymbol{U}^\top\boldsymbol{B}^\top\boldsymbol{L}^+\boldsymbol{B}.$$

using $(\boldsymbol{L}^+)^\top = \boldsymbol{L}^+$

Trivially, $\boldsymbol{I}\boldsymbol{x}$, $\boldsymbol{U}\boldsymbol{x}$, and $\boldsymbol{U}^\top\boldsymbol{x}$ can be computed in $\mathcal{O}(|E|)$ time. As by definition of the incidence matrix $\boldsymbol{B}$, $\operatorname{nnz}(\boldsymbol{B}) = \mathcal{O}(|E|)$, $\boldsymbol{B}\boldsymbol{x}$ and $\boldsymbol{B}^\top\boldsymbol{x}$ can also be computed in $\mathcal{O}(|E|)$ time.

It follows from the definition of the incidence matrix that $\mathbf{1} \in \ker \boldsymbol{B}$ and $\mathbf{1} \in \ker \boldsymbol{B}^\top$. Therefore, we have for any $\boldsymbol{y} \in \mathbb{R}^{|V|}$ that $\boldsymbol{B}\boldsymbol{y} \perp \mathbf{1}$ and for any $\boldsymbol{x} \in \mathbb{R}^{|E|}$ that $\boldsymbol{B}^\top\boldsymbol{x} \perp \mathbf{1}$.[33] Therefore, $\boldsymbol{B}^\top\boldsymbol{U}\boldsymbol{x} \perp \mathbf{1}$ and $\boldsymbol{B}\boldsymbol{y} \perp \mathbf{1}$ for any $\boldsymbol{x}$ and $\boldsymbol{y}$.[34]

[33] $(\boldsymbol{B}\boldsymbol{y})^\top\mathbf{1} = \boldsymbol{y}^\top\boldsymbol{B}^\top\mathbf{1} = \boldsymbol{y}^\top\mathbf{0} = \mathbf{0}$; the other case is symmetric

[34] $\boldsymbol{B}^\top\boldsymbol{U}\boldsymbol{x}$ and $\boldsymbol{B}\boldsymbol{y}$ can be interpreted as demand vectors.

Using the result of Kyng and Sachdeva,[35] we can find an ϵ-approximate solution $\tilde{\boldsymbol{z}}$ to $\boldsymbol{L}\boldsymbol{z} = \boldsymbol{d}$ in time $\mathcal{O}\left(|E|\log^3|V|\log(1/\epsilon)\right)$, where $\boldsymbol{d} \perp \mathbf{1}$. Using our assumption that finding $\tilde{\boldsymbol{z}}$ for $\epsilon = 1/|V|^{100}$ is as expensive as finding $\boldsymbol{z}$ exactly, we conclude that $\boldsymbol{L}^+\boldsymbol{B}^\top\boldsymbol{U}\boldsymbol{x}$ and $\boldsymbol{L}^+\boldsymbol{B}\boldsymbol{y}$ can be computed in

[35] Corollary 10.2.5 in the lecture notes

$$\mathcal{O}\left(|E|\log^3|V|\log|V|^{100}\right) = \mathcal{O}\left(|E|\log^4|V|\right) = \tilde{\mathcal{O}}(|E|)$$

time. □

References

[1] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.