Advanced Graph Algorithms and Optimization Spring 2020

Karush-Kuhn-Tucker Conditions, Fenchel Conjugates, Newton's Method

Rasmus Kyng, Scribe: Hongjie Chen *Lecture 13 — Wednesday, May 20th*

1 Langrange Multipliers and Convex Duality Recap

Recall the convex optimization problem we studied last lecture,

$$\begin{aligned} \min \quad & \mathcal{E}(\boldsymbol{y}) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{y} = \boldsymbol{b} \\ & \boldsymbol{c}(\boldsymbol{y}) \leq \boldsymbol{0}, \end{aligned} \tag{1}$$

where $\mathcal{E}(\boldsymbol{y}) : S \to \mathbb{R}$ is defined on a subset $S \subseteq \mathbb{R}^n$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{c}(\boldsymbol{y})$ is a vector of constraints $\boldsymbol{c}(\boldsymbol{y}) = (c_i(\boldsymbol{y}))_{i \in [k]}$. For every $i \in [k]$ the function $c_i : S \to \mathbb{R}$ is convex. We call (1) the primal (problem) and denote its optimal value by α^*.

The associated Lagrangian is defined by

$$L(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{s}) = \mathcal{E}(\boldsymbol{y}) + \boldsymbol{x}^T(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{y}) + \boldsymbol{s}^T\boldsymbol{c}(\boldsymbol{y}).$$

where $\boldsymbol{x} \in \mathbb{R}^m$, $\boldsymbol{s} \in \mathbb{R}^k$ are dual variables. The dual (problem) is given by

$$\max_{\substack{\boldsymbol{x}, \boldsymbol{s} \\ \boldsymbol{s} \geq 0}} \quad L(\boldsymbol{x}, \boldsymbol{s}) \tag{2}$$

whose optimal value is denoted by β^*. The dual is always a convex optimization problem even though the primal is non-convex. The optimal value of the primal (1) can also be written as

$$\alpha^* = \inf_{\boldsymbol{y}} \sup_{\boldsymbol{x}; \boldsymbol{s} \geq \boldsymbol{0}} L(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{s}), \tag{3}$$

where no constraint is imposed on the primal variable $\boldsymbol{y}$. The optimal value of the dual (2) is

$$\beta^* = \sup_{\boldsymbol{x}; \boldsymbol{s} \geq \boldsymbol{0}} \inf_{\boldsymbol{y}} L(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{s}). \tag{4}$$

Note the only difference between (3) and (4) is that the positions of "inf" and "sup" are swapped. The weak duality theorem states that the dual optimal value is a lower bound of the primal optimal value, i.e. $\beta^* \leq \alpha^*$.

The Slater's condition for (1) requires the existence of a *strictly feasible* point, i.e. there exists $\tilde{\boldsymbol{y}} \in S$ s.t. $\boldsymbol{A}\tilde{\boldsymbol{y}} = \boldsymbol{b}$ and $\boldsymbol{c}(\tilde{\boldsymbol{y}}) < \boldsymbol{0}$. This means that the strictly feasible point $\tilde{\boldsymbol{y}}$ lies inside the interior of the set $\{\boldsymbol{y} : \boldsymbol{c}(\boldsymbol{y}) \leq \boldsymbol{0}\}$ defined by the inequality constraints. The strong duality theorem says, the Slater's condition implies strong duality, $\beta^* = \alpha^*$.

Example. In lecture 10, we gave a combinatorial proof of the min-cut max-flow theorem, and showed that the min-cut problem can be expressed as a linear program. Now, we will use the

strong duality theorem to give an alternative proof, and directly find the min-cut linear program is the dual program to our maximum flow linear program.

We will assume that Slater's condition holds for our primal program. Since scaling the flow down enough will always ensure that capacity constraints are strictly satisfied i.e. $\boldsymbol{f} < \boldsymbol{c}$, the only concern is to make sure that non-negativity constraints are satisfied. This means that there is an s-t flow that sends a non-zero flow on every edge. In fact, this may not always be possible, but it is easy to detect such edges and remove them without changing the value of the program: an edge (u, v) should be removed if there is no path s to u or no path v to t. We can identify all such edges using a BFS from s along the directed edges and a BFS along reversed directed edges from t.

Slater's condition holds whenever there is a directed path from s to t with non-zero capacity (and if there is not, the maximum flow and minimum cut are both zero).

$$\begin{aligned}
\min_{\substack{F\in\mathbb{R}\\ \boldsymbol{B}\boldsymbol{f}=F\boldsymbol{b}_{s,t}\\ \boldsymbol{0}\le\boldsymbol{f}\le\boldsymbol{c}}} -F &= \min_{F;\boldsymbol{f}\ge\boldsymbol{0}} \max_{\boldsymbol{x};\boldsymbol{s}\ge\boldsymbol{0}} -F + \boldsymbol{x}^\top(F\boldsymbol{b}_{s,t} - \boldsymbol{B}\boldsymbol{f}) + (\boldsymbol{f}-\boldsymbol{c})^\top\boldsymbol{s} \\
&\qquad \text{(Slater's condition } \implies \text{ strong duality)} \\
-\max_{\substack{F\in\mathbb{R}\\ \boldsymbol{B}\boldsymbol{f}=F\boldsymbol{b}_{s,t}\\ \boldsymbol{0}\le\boldsymbol{f}\le\boldsymbol{c}}} F &= \max_{\boldsymbol{x};\boldsymbol{s}\ge\boldsymbol{0}} \min_{F;\boldsymbol{f}\ge\boldsymbol{0}} F(\boldsymbol{b}_{s,t}^\top\boldsymbol{x} - 1) + \boldsymbol{f}^\top(\boldsymbol{s} - \boldsymbol{B}^\top\boldsymbol{x}) - \boldsymbol{c}^\top\boldsymbol{s} \\
&= \max_{\substack{\boldsymbol{x};\,\boldsymbol{s}\ge\boldsymbol{0}\\ \boldsymbol{b}_{s,t}^\top\boldsymbol{x}=1\\ \boldsymbol{s}\ge\boldsymbol{B}^\top\boldsymbol{x}}} -\boldsymbol{c}^\top\boldsymbol{s} \\
\max_{\substack{F\in\mathbb{R}\\ \boldsymbol{B}\boldsymbol{f}=F\boldsymbol{b}_{s,t}\\ \boldsymbol{0}\le\boldsymbol{f}\le\boldsymbol{c}}} F &= \min_{\substack{\boldsymbol{x};\,\boldsymbol{s}\ge\boldsymbol{0}\\ \boldsymbol{b}_{s,t}^\top\boldsymbol{x}=1\\ \boldsymbol{s}\ge\boldsymbol{B}^\top\boldsymbol{x}}} \boldsymbol{c}^\top\boldsymbol{s} \qquad (5)
\end{aligned}$$

The LHS of (5) is exactly the LP formulation of max-flow, while the RHS is exactly the LP formulation of min-cut.

Let $\boldsymbol{y}^*$ be an optimizer of the primal problem and $\boldsymbol{x}^*, \boldsymbol{s}^*$ for the dual. Since $\boldsymbol{y}^*, \boldsymbol{x}^*, \boldsymbol{s}^*$ are also primal/dual feasible, then

$$\alpha^* = \mathcal{E}(\boldsymbol{y}^*) \ge L(\boldsymbol{y}^*, \boldsymbol{x}^*, \boldsymbol{s}^*) \ge L(\boldsymbol{x}^*, \boldsymbol{s}^*) = \beta^*.$$

Now, suppose strong duality holds, i.e. $\alpha^* = \beta^*$. Then,

$$\mathcal{E}(\boldsymbol{y}^*) = L(\boldsymbol{y}^*, \boldsymbol{x}^*, \boldsymbol{s}^*) = L(\boldsymbol{x}^*, \boldsymbol{s}^*).$$

If further assume $\mathcal{E} : S \to \mathbb{R}$ and $\boldsymbol{c}$ are differentiable and the minimizer $\boldsymbol{y}^*$ is not on the boundary of S, then

$$\begin{aligned}
\boldsymbol{\nabla}_{\boldsymbol{y}} L(\boldsymbol{y}, \boldsymbol{x}^*, \boldsymbol{s}^*)\big|_{\boldsymbol{y}=\boldsymbol{y}^*} &= \boldsymbol{0} \\
\boldsymbol{\nabla}_{\boldsymbol{y}} \Big(\mathcal{E}(\boldsymbol{y}) + (\boldsymbol{x}^*)^\top(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{y}) + (\boldsymbol{s}^*)^\top\boldsymbol{c}(\boldsymbol{y})\Big)\Big|_{\boldsymbol{y}=\boldsymbol{y}^*} &= \boldsymbol{0} \\
\boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y}^*) - \boldsymbol{A}^\top\boldsymbol{x}^* - \boldsymbol{s}^\top\boldsymbol{\nabla}\boldsymbol{c}(\boldsymbol{y}^*) &= \boldsymbol{0}.
\end{aligned}$$

And this connects to our point of the parallel gradients.

Furthermore, we also have *complementary slackness*,

$$\boldsymbol{s}^*(i) \cdot \boldsymbol{c}_i(\boldsymbol{y}^*) = 0 \text{ for all } i,$$

which means the i-th optimal dual varible $\boldsymbol{s}^*(i)$ is zero unless the i-th constraint is active at the optimum, i.e. $\boldsymbol{c}_i(\boldsymbol{y}^*) = 0$.

2 Karush-Kuhn-Tucker Theorem

Let us continue to consider the convex optimization problem (1) with $\mathcal{E}$ and $\boldsymbol{c}$ differentiable and assume that the domain S of $\mathcal{E}$ is an *open* convex set. We assume S is open to rule out having an optimal $\boldsymbol{y}$ on the boundary of S. Suppose $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{s}}$ satisfy the following conditions:

- $\boldsymbol{A}\tilde{\boldsymbol{y}} = \boldsymbol{b}$ and $\boldsymbol{c}(\tilde{\boldsymbol{y}}) \leq 0$ (primal feasible)
- $\tilde{\boldsymbol{s}} \geq 0$ (dual feasible)
- $\boldsymbol{\nabla}_y \mathcal{E}(\tilde{\boldsymbol{y}}) - \boldsymbol{A}^\top \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{s}}^\top \boldsymbol{\nabla} \boldsymbol{c}(\boldsymbol{y}^*) = \boldsymbol{0}$ $(\boldsymbol{\nabla}_y L(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{s}}) = \boldsymbol{0})$
- $\tilde{\boldsymbol{s}}(i) \cdot \boldsymbol{c}_i(\tilde{\boldsymbol{y}}) = 0$ for all i (complementary slackness)

These conditions are called the *Karush-Kuhn-Tucker* (KKT) conditions.

Theorem 2.1. *$\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{s}}$ are primal/dual optimal.*

Proof. $\tilde{\boldsymbol{y}}$ is global minimizer of $\boldsymbol{y} \mapsto L(\boldsymbol{y}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{s}})$, since this function is convex with vanishing gradient at $\tilde{\boldsymbol{y}}$. Hence,

$$L(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{s}}) = \inf_{\boldsymbol{y}} L(\boldsymbol{y}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{s}}) = L(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{s}}) \leq \beta^*.$$

On the other hand, due to primal feasibility and complementary slackness,

$$L(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{s}}) = \mathcal{E}(\tilde{\boldsymbol{y}}) + \tilde{\boldsymbol{x}}^\top (\boldsymbol{b} - \boldsymbol{A}^\top \tilde{\boldsymbol{y}}) + \tilde{\boldsymbol{s}}^\top \boldsymbol{c}(\tilde{\boldsymbol{y}}) = \mathcal{E}(\tilde{\boldsymbol{y}}) \geq \alpha^*.$$

Thus, $\beta^* \geq \alpha^*$. But also $\beta^* \leq \alpha^*$ by weak duality. Therefore, $\beta^* = \alpha^*$ and $\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{s}}$ are primal/dual optimal. □

3 Fenchel Conjugate

Definition 3.1 (Fenchel conjugate)**.** Given a (convex) function $\mathcal{E} : S \subseteq \mathbb{R}^n \to \mathbb{R}$, its *Fenchel conjugate* is a function $\mathcal{E}^* : \mathbb{R}^n \to \mathbb{R}$ defined as

$$\mathcal{E}^*(\boldsymbol{z}) = \sup_{\boldsymbol{y} \in S} \langle \boldsymbol{z}, \boldsymbol{y} \rangle - \mathcal{E}(\boldsymbol{y}).$$

Remark 3.2. $\mathcal{E}^*$ is a convex function whether $\mathcal{E}$ is convex or not, since $\mathcal{E}^*(z)$ is pointwise supremum of a family of convex (here, affine) functions of z.

In this course, we have only considered convex function that are real-valued and continuous and defined on a convex domain. For any such $\mathcal{E}$, we have $\mathcal{E}^{**} = \mathcal{E}$, i.e. the Fenchel conjugate of the Fenchel conjugate is the original function. This is a consquence of the Fenchel-Moreau theorem, which establishes this under slightly more general conditions. We will not prove this generally, but as part of Theorem 3.3 below, we prove it under more restrictive assumptions.

Example. Let $\mathcal{E}(\boldsymbol{y}) = \frac{1}{p}\|\boldsymbol{y}\|_p^p$ $(p > 1)$. We want to evaluate its Fenchel conjugate $\mathcal{E}^*$ at any given point $\boldsymbol{z} \in \mathbb{R}^n$. Since $\mathcal{E}$ is convex and differentiable, the supremum must be achieved at some $\boldsymbol{y}$ with vanishing gradient

$$\boldsymbol{\nabla}_{\boldsymbol{y}}\langle \boldsymbol{z}, \boldsymbol{y}^*\rangle - \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}^*) = \boldsymbol{z} - \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}^*) = \boldsymbol{0} \iff \boldsymbol{z} = \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}^*).$$

It's not difficult to see, for all i,

$$\boldsymbol{z}(i) = \operatorname{sgn}(\boldsymbol{y}(i))|\boldsymbol{y}(i)|^{p-1}.$$

Then,

$$\begin{aligned}
\mathcal{E}^*(\boldsymbol{z}) &= \langle \boldsymbol{z}, \boldsymbol{y}\rangle - \mathcal{E}(\boldsymbol{y}) \\
&= \sum_i |\boldsymbol{z}(i)|^{\frac{1}{p-1}+1} - \frac{1}{p}|\boldsymbol{z}(i)|^{\frac{p}{p-1}} \\
&\text{(define } q \text{ s.t. } \frac{1}{q} + \frac{1}{p} = 1) \\
&= \frac{1}{q}\|\boldsymbol{z}\|_q^q
\end{aligned}$$

More generally, given a convex and differentiable function $\mathcal{E} : S \to \mathbb{R}$, if there exists $\boldsymbol{y} \in S$ s.t. $\boldsymbol{z} = \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})$, then $\mathcal{E}^*(\boldsymbol{z}) = (\boldsymbol{y}^*)^\top \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}^*) - \mathcal{E}(\boldsymbol{y}^*)$. The Fenchel conjugate and Lagrange duality are closely related, which is demonstrated in the following example.

Example. Consider a convex optimization problem with only linear constraints,

$$\begin{aligned}
\min_{\boldsymbol{y}\in\mathbb{R}^n} \quad & \mathcal{E}(\boldsymbol{y}) \\
\text{s.t.} \quad & \boldsymbol{A}\boldsymbol{y} = \boldsymbol{b}
\end{aligned}$$

where $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ is a convex function and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Then the corresponding dual problem is

$$\begin{aligned}
\sup_{\boldsymbol{x}\in\mathbb{R}^n} \inf_{\boldsymbol{y}\in\mathbb{R}^n} \mathcal{E}(\boldsymbol{y}) + \boldsymbol{x}^\top(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{y}) &= \sup_{\boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{b}^\top\boldsymbol{x} - \sup_{\boldsymbol{y}\in\mathbb{R}^n} \Big(\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{y} - \mathcal{E}(\boldsymbol{y})\Big) \\
&= \sup_{\boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{b}^\top\boldsymbol{x} - \mathcal{E}^*(\boldsymbol{A}^\top\boldsymbol{x})
\end{aligned}$$

Theorem 3.3 (properties of Fenchel conjugate)**.** *When $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ is bounded below, strictly convex, differentiable, with a Hessian that is positive definite everywhere, we have the following three properties:*

1. $\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{\nabla}\mathcal{E}^*(\boldsymbol{z})) = \boldsymbol{z}$ *and* $\boldsymbol{\nabla}\mathcal{E}^*(\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})) = \boldsymbol{y}$
2. $(\mathcal{E}^*)^* = \mathcal{E}$*, i.e. the Fenchel conjugate of the Fenchel conjugate is the original function.*

3. $\boldsymbol{H}_{\mathcal{E}^*}(\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})) = \boldsymbol{H}_{\mathcal{E}}^{-1}(\boldsymbol{y})$

primal point $\boldsymbol{y}$	$\xrightarrow{\boldsymbol{\nabla}_y\mathcal{E}}$ $\xleftarrow{\boldsymbol{\nabla}_z\mathcal{E}^*}$	dual point $\boldsymbol{z}$
gradient $\boldsymbol{\nabla}_y\mathcal{E}(\boldsymbol{y}) = \boldsymbol{z}$		gradient $\boldsymbol{\nabla}_z\mathcal{E}^*(\boldsymbol{z}) = \boldsymbol{y}$
Hessian $\boldsymbol{H}_{\mathcal{E}}(\boldsymbol{y}) = \boldsymbol{H}_{\mathcal{E}^*}^{-1}(\boldsymbol{z})$		Hessian $\boldsymbol{H}_{\mathcal{E}^*}(\boldsymbol{y}) = \boldsymbol{H}_{\mathcal{E}}^{-1}(\boldsymbol{y})$

Figure 1: Properties of Fenchel conjugate

Proof sketch. Part 1. Given $\boldsymbol{z}$, let $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{z})$ be the $\boldsymbol{y}$ achieving the supremum in the Fenchel conjugate program such that $\mathcal{E}^*(\boldsymbol{z}) = \langle \boldsymbol{z}, \boldsymbol{y}(\boldsymbol{z})\rangle - \mathcal{E}(\boldsymbol{y}(\boldsymbol{z}))$. One can show that because $\mathcal{E}$ is bounded below and strictly convex, $\boldsymbol{y} \mapsto \langle \boldsymbol{z}, \boldsymbol{y}\rangle - \mathcal{E}(\boldsymbol{y})$ is bounded above and is strictly concave, and is hence maximized at some $\boldsymbol{y}$ such that $\boldsymbol{z} = \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}(\boldsymbol{z}))$. Since $\mathcal{E}$ is strictly convex, $\boldsymbol{y}(\boldsymbol{z})$ is unique. Then, using the product rule and composition rule of derivatives,

$$\begin{aligned}
\boldsymbol{\nabla}\mathcal{E}^*(\boldsymbol{z}) &= \boldsymbol{\nabla}_{\boldsymbol{z}}\left(\langle \boldsymbol{z}, \boldsymbol{y}(\boldsymbol{z})\rangle - \mathcal{E}(\boldsymbol{y}(\boldsymbol{z}))\right) \\
&= \boldsymbol{y}(\boldsymbol{z}) + \operatorname{diag}(\boldsymbol{z})\boldsymbol{\nabla}_{\boldsymbol{z}}\boldsymbol{y}(\boldsymbol{z}) - \operatorname{diag}(\underbrace{\boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y}(\boldsymbol{z}))}_{=\boldsymbol{z}})\boldsymbol{\nabla}_{\boldsymbol{z}}\boldsymbol{y}(\boldsymbol{z}) \\
&= \boldsymbol{y}(\boldsymbol{z})
\end{aligned}$$

Thus we have $\boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y}(\boldsymbol{z})) = \boldsymbol{z}$ and $\boldsymbol{\nabla}_{\boldsymbol{z}}\mathcal{E}^*(\boldsymbol{z}) = \boldsymbol{y}(\boldsymbol{z})$. Combining the two, we have $\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{\nabla}\mathcal{E}^*(\boldsymbol{z})) = \boldsymbol{z}$.

We can also see that for any $\boldsymbol{y}$, there exists a $\boldsymbol{z}$ such that $\boldsymbol{\nabla}_{\boldsymbol{z}}\mathcal{E}^*(\boldsymbol{z}) = \boldsymbol{y}$, namely, this is attained by $\boldsymbol{z} = \boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y})$. Thus, $\boldsymbol{\nabla}\mathcal{E}^*(\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})) = \boldsymbol{y}$.

Part 2. Observe that

$$\mathcal{E}^{**}(\boldsymbol{u}) = \sup_{\boldsymbol{z}\in\mathbb{R}^n} \langle \boldsymbol{u}, \boldsymbol{z}\rangle - \mathcal{E}^*(\boldsymbol{z})$$

and let $\boldsymbol{z}(\boldsymbol{u})$ denote the $\boldsymbol{z}$ obtaining the supremum, in the above program. We then have $\boldsymbol{u} = \boldsymbol{\nabla}\mathcal{E}^*(\boldsymbol{z}(\boldsymbol{u}))$. Letting $\boldsymbol{y}(\boldsymbol{z})$ be defined as in Part 1, we get $\boldsymbol{y}(\boldsymbol{z}(\boldsymbol{u})) = \boldsymbol{\nabla}_{\boldsymbol{z}}\mathcal{E}^*(\boldsymbol{z}(\boldsymbol{u})) = \boldsymbol{u}$

$$\mathcal{E}^{**}(\boldsymbol{u}) = \langle \boldsymbol{u}, \boldsymbol{z}(\boldsymbol{u})\rangle - (\langle \boldsymbol{z}(\boldsymbol{u}), \boldsymbol{y}(\boldsymbol{z}(\boldsymbol{u}))\rangle - \mathcal{E}(\boldsymbol{y}(\boldsymbol{z}(\boldsymbol{u})))) = \mathcal{E}(\boldsymbol{u}).$$

Part 3. Now we add two infinitesimals $\boldsymbol{\tau}$ and $\boldsymbol{\delta}$ to $\boldsymbol{z}$ and $\boldsymbol{y}$ respectively s.t.

$$\boldsymbol{\nabla}_{\boldsymbol{z}}\mathcal{E}^*(\boldsymbol{z}+\boldsymbol{\tau}) = \boldsymbol{y} + \boldsymbol{\delta}, \quad \boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y}+\boldsymbol{\delta}) = \boldsymbol{z} + \boldsymbol{\tau}.$$

Then,

$$\boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y}+\boldsymbol{\delta}) - \boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y}) = \boldsymbol{\tau}, \quad \boldsymbol{\nabla}_{\boldsymbol{z}}\mathcal{E}^*(\boldsymbol{z}+\boldsymbol{\tau}) - \boldsymbol{\nabla}_{\boldsymbol{z}}\mathcal{E}^*(\boldsymbol{z}) = \boldsymbol{\delta}.$$

Since $\boldsymbol{H}_{\mathcal{E}}(\boldsymbol{y})$ measures the change of $\boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y})$ when $\boldsymbol{y}$ changes by an infinitesimal $\boldsymbol{\delta}$, then

$$\begin{aligned}
&\boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y}+\boldsymbol{\delta}) - \boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y}) \approx \boldsymbol{H}_{\mathcal{E}}(\boldsymbol{y})\boldsymbol{\delta} \\
\iff &\boldsymbol{H}_{\mathcal{E}}^{-1}(\boldsymbol{y})\left(\boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y}+\boldsymbol{\delta}) - \boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y})\right) \approx \boldsymbol{\delta} \\
\iff &\boldsymbol{H}_{\mathcal{E}}^{-1}(\boldsymbol{y})\boldsymbol{\tau} \approx \boldsymbol{\delta} = \boldsymbol{\nabla}\mathcal{E}^*(\boldsymbol{z}+\boldsymbol{\tau}) - \boldsymbol{\nabla}\mathcal{E}^*(\boldsymbol{z}) \\
\iff &\boldsymbol{H}_{\mathcal{E}}^{-1}(\boldsymbol{y})\boldsymbol{\tau} \approx \boldsymbol{\nabla}\mathcal{E}^*(\boldsymbol{z}+\boldsymbol{\tau}) - \boldsymbol{\nabla}\mathcal{E}^*(\boldsymbol{z}) \qquad (6)
\end{aligned}$$

Similarly,

$$\boldsymbol{H}_{\mathcal{E}^*}(\boldsymbol{z})\boldsymbol{\tau} \approx \boldsymbol{\nabla}_{\boldsymbol{z}}\mathcal{E}^*(\boldsymbol{z}+\boldsymbol{\tau}) - \boldsymbol{\nabla}_{\boldsymbol{z}}\mathcal{E}^*(\boldsymbol{z}) \tag{7}$$

Comparing (6) and (7), it is easy to see

$$\boldsymbol{H}_{\mathcal{E}^*}(\boldsymbol{z}) = \boldsymbol{H}_{\mathcal{E}}^{-1}(\boldsymbol{y}) \iff \boldsymbol{H}_{\mathcal{E}^*}(\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})) = \boldsymbol{H}_{\mathcal{E}}^{-1}(\boldsymbol{y}).$$

□

4 Newton's Method

4.1 Warm-up: Quadratic Optimization

First, let us play with a toy example, minimizing a quadratic function

$$\mathcal{E}(\boldsymbol{y}) = \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}^\top \boldsymbol{y} + \boldsymbol{c}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is positive definite. By setting the gradient w.r.t. $\boldsymbol{y}$ to zero,

$$\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b} = \boldsymbol{0},$$

we obtain the global minimizer

$$\boldsymbol{y}^* = -\boldsymbol{A}^{-1}\boldsymbol{b}.$$

To make it more like gradient descent, let us start at some "guess" point $\boldsymbol{y}$ and take a step $\boldsymbol{\delta}$ to move to the new point $\boldsymbol{y}+\boldsymbol{\delta}$. Then we try to minimize $\mathcal{E}(\boldsymbol{y}+\boldsymbol{\delta})$ by setting the gradient w.r.t. $\boldsymbol{\delta}$ to zero,

$$\begin{aligned}
\boldsymbol{\nabla}_{\boldsymbol{\delta}}\mathcal{E}(\boldsymbol{y}+\boldsymbol{\delta}) = \boldsymbol{A}(\boldsymbol{y}+\boldsymbol{\delta}) &= \boldsymbol{0} \\
\boldsymbol{\delta} &= -\boldsymbol{y} - \boldsymbol{A}^{-1}\boldsymbol{b} \\
\boldsymbol{y}+\boldsymbol{\delta} &= -\boldsymbol{A}^{-1}\boldsymbol{b}
\end{aligned}$$

This gives us exactly global minimizer in just one step. However, the situation changes when the function is not quadratic anymore and thus we do not have constant Hessian. But taking a step which try to set the gradient to zero might still be a good idea.

4.2 K-stable Hessian

Next, consider a convex function $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ whose Hessian is "nearly constant". Recall the Hessian $\boldsymbol{H}_{\mathcal{E}}(\boldsymbol{y})$ aka $\boldsymbol{\nabla}^2\mathcal{E}(\boldsymbol{y})$ at a point $\boldsymbol{y}$ is just a matrix of pairwise 2nd order partial derivatives $\frac{\partial^2\mathcal{E}(\boldsymbol{y})}{\partial \boldsymbol{y}_i \partial \boldsymbol{y}_j}$. We say $\mathcal{E}$ has k-stable Hessian if there exists a constant matrix $\boldsymbol{A}$ s.t. for all $\boldsymbol{y}$

$$\boldsymbol{H}_{\mathcal{E}}(\boldsymbol{y}) \approx_K \boldsymbol{A} \iff \frac{1}{1+K}\boldsymbol{A} \preceq \boldsymbol{H}_{\mathcal{E}}(\boldsymbol{y}) \preceq (1+K)\boldsymbol{A}.$$

Note that we just require the existence of $\boldsymbol{A}$ and do not assume we know $\boldsymbol{A}$. Then a natural question is to ask what convergence rate can be achieved if we take a gradient step "guided" by the

Hessian, which is called a "Newton step". Such method is also known as the 2nd order method. Note that this is very similar to preconditioning.

Now, let us make our setting precise. We want minimize a convex function $\mathcal{E}$ with k-stable Hessian $\boldsymbol{A} \succ \boldsymbol{0}$. And $\boldsymbol{y}^*$ is a global minimizer of $\mathcal{E}$. Start from some initial point $\boldsymbol{y}_0$. The update rule is

$$\boldsymbol{y}_{i+1} = \boldsymbol{y}_i - \alpha \cdot \boldsymbol{H}_{\mathcal{E}}^{-1}(\boldsymbol{y}_i)\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}_i),$$

where α is the step size and it will be decided later.

Theorem 4.1. *$\mathcal{E}(\boldsymbol{y}_k) - \mathcal{E}(\boldsymbol{y}^*) \le \epsilon\,(\mathcal{E}(\boldsymbol{y}_0) - \mathcal{E}(\boldsymbol{y}^*))$ when $k > (K+1)^6 \log(1/\epsilon)$.*

Proof. By Taylor's theorem, there exists $\tilde{\boldsymbol{y}} \in [\boldsymbol{y}, \boldsymbol{y}+\boldsymbol{\delta}]$ s.t.

$$\begin{aligned}\mathcal{E}(\boldsymbol{y}+\boldsymbol{\delta}) &= \mathcal{E}(\boldsymbol{y}) + \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})^\top\boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}^\top\boldsymbol{H}_{\mathcal{E}}(\tilde{\boldsymbol{y}})\boldsymbol{\delta} \\ &\le \underbrace{\mathcal{E}(\boldsymbol{y}) + \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})^\top\boldsymbol{\delta} + \frac{(K+1)^2}{2}\boldsymbol{\delta}^\top\boldsymbol{H}_{\mathcal{E}}(\boldsymbol{y})\boldsymbol{\delta}}_{=:f(\boldsymbol{\delta})} \qquad (8)\end{aligned}$$

where the inequality comes from K-stability of Hessian,

$$\boldsymbol{H}_{\mathcal{E}}(\tilde{\boldsymbol{y}}) \preceq (1+K)\boldsymbol{A} \preceq (1+K)^2\boldsymbol{H}_{\mathcal{E}}(\boldsymbol{y})$$

Observe that $f(\boldsymbol{\delta})$ is a convex quadratic function in $\boldsymbol{\delta}$. By setting minizing it, or equivalently setting $\boldsymbol{\nabla}_{\boldsymbol{\delta}} f(\boldsymbol{\delta}^*) = \boldsymbol{0}$, we get

$$\boldsymbol{\delta}^* = -\frac{1}{(K+1)^2}\boldsymbol{H}_{\mathcal{E}}^{-1}(\boldsymbol{y})\boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y}) \qquad (9)$$

Here, the step size α is equal to $(K+1)^{-2}$. Then, plugging (9) into (8),

$$\begin{aligned}\mathcal{E}(\boldsymbol{y}+\boldsymbol{\delta}^*) &\le \mathcal{E}(\boldsymbol{y}) - \frac{1}{2(K+1)^2}\boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y})^\top\boldsymbol{H}_{\mathcal{E}}^{-1}(\boldsymbol{y})\boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y}) \\ &\le \mathcal{E}(\boldsymbol{y}) - \frac{1}{2(K+1)^3}\boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y})^\top\boldsymbol{A}^{-1}\boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y}) \\ &\text{(subtract } \mathcal{E}(\boldsymbol{y}^*) \text{ on both sides)} \\ \mathcal{E}(\boldsymbol{y}+\boldsymbol{\delta}^*) - \mathcal{E}(\boldsymbol{y}^*) &\le \mathcal{E}(\boldsymbol{y}) - \mathcal{E}(\boldsymbol{y}^*) - \frac{1}{2(K+1)^3}\underbrace{\boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y})^\top\boldsymbol{A}^{-1}\boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y})}_{=:\sigma}\end{aligned}$$

where the second inequality is due to K-stability of the inverse Hessian,

$$\frac{1}{1+K}\boldsymbol{A}^{-1} \preceq \boldsymbol{H}_{\mathcal{E}}(\boldsymbol{y})^{-1} \preceq (1+K)\boldsymbol{A}^{-1}.$$

Meanwhile, using Taylor's theorem and K-stability, for some $\hat{\boldsymbol{y}}$ between $\boldsymbol{y}$ and $\boldsymbol{y}^*$, and noting $\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}^*) = \boldsymbol{0}$, we have

$$\begin{aligned}\mathcal{E}(\boldsymbol{y}) &= \mathcal{E}(\boldsymbol{y}^*) + \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}^*)^\top(\boldsymbol{y}-\boldsymbol{y}^*) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{y}^*)^\top\boldsymbol{H}_{\mathcal{E}}(\hat{\boldsymbol{y}})(\boldsymbol{y}-\boldsymbol{y}^*) \\ \mathcal{E}(\boldsymbol{y}) - \mathcal{E}(\boldsymbol{y}^*) &\le \frac{(K+1)}{2}\underbrace{(\boldsymbol{y}-\boldsymbol{y}^*)^\top\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{y}^*)}_{=:\gamma}\end{aligned}$$

Next, our task is reduced to comparing σ and γ. $\boldsymbol{y}_t := \boldsymbol{y}^* + t(\boldsymbol{y} - \boldsymbol{y}^*)$ ($t \in [0,1]$) is a point on the segment connecting $\boldsymbol{y}^*$ and $\boldsymbol{y}$. Since

$$\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}) = \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}) - \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}^*) = \int_0^1 H(\boldsymbol{y}_t)(\boldsymbol{y} - \boldsymbol{y}^*)\mathrm{d}t,$$

then

$$\begin{aligned}(\boldsymbol{y} - \boldsymbol{y}^*)^\top \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}) &= \int_0^1 (\boldsymbol{y} - \boldsymbol{y}^*)^\top H(\boldsymbol{y}_t)(\boldsymbol{y} - \boldsymbol{y}^*)\mathrm{d}t \\ &\geq \frac{1}{K+1}\int_0^1 (\boldsymbol{y} - \boldsymbol{y}^*)^\top \boldsymbol{A}(\boldsymbol{y} - \boldsymbol{y}^*)\mathrm{d}t \\ &= \frac{\gamma}{K+1} \end{aligned} \tag{10}$$

On the other hand, define $\boldsymbol{z}_s = \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}^*) + s(\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}) - \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}^*))$ and then $\mathrm{d}\boldsymbol{z}_s = \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})\mathrm{d}s$. Using *Theorem 3.3*, we have

$$\boldsymbol{y} - \boldsymbol{y}^* = \int_0^1 \boldsymbol{H}_{\mathcal{E}^*}(\boldsymbol{z}_s)\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})\mathrm{d}s.$$

Then,

$$\begin{aligned}\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})^\top(\boldsymbol{y} - \boldsymbol{y}^*) &= \int_0^1 \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})^\top \boldsymbol{H}_{\mathcal{E}^*}(\boldsymbol{z}_s)\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})\mathrm{d}s \\ &\leq (K+1)\int_0^1 \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})^\top \boldsymbol{A}^{-1}\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})\mathrm{d}s \\ &\leq (K+1)\sigma \end{aligned} \tag{11}$$

Combining (10) and (11) yields

$$\gamma \leq (K+1)^2\sigma.$$

Therefore,

$$\mathcal{E}(\boldsymbol{y} + \boldsymbol{\delta}^*) - \mathcal{E}(\boldsymbol{y}^*) \leq (\mathcal{E}(\boldsymbol{y}) - \mathcal{E}(\boldsymbol{y}^*))\left(1 - \frac{1}{(K+1)^6}\right).$$

□

Remark 4.2. The basic idea of relating σ and γ in the above proof is writing the same quantity, $\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})^\top(\boldsymbol{y} - \boldsymbol{y}^*)$, as two integrations along different lines. $(K+1)^6$ can be reduced to $(K+1)^2$ and even to $(K+1)$ with more care. In some settings, Newton's method converges in $\log\log(1/\epsilon)$ steps.

4.3 Linearly Constrained Newton's Method

Let us apply Newton's method to convex optimization problems with only linear constraints,

$$\begin{aligned}\min_{\boldsymbol{f}\in\mathbb{R}^m} \quad & \mathcal{E}(\boldsymbol{f}) \\ \text{s.t.} \quad & \boldsymbol{B}\boldsymbol{f} = \boldsymbol{d}\end{aligned}$$

where $\mathcal{E} : S \subseteq \mathbb{R}^m \to \mathbb{R}$ is a convex function and $\boldsymbol{B} \in \mathbb{R}^{n\times m}$. Wlog, let $\boldsymbol{d} = \boldsymbol{0}$, since otherwise we can equivalently deal the following problem with $\boldsymbol{B}\boldsymbol{f}_0 = \boldsymbol{d}$,

$$\begin{aligned} \min_{\boldsymbol{\rho}\in\mathbb{R}^m} \quad & \mathcal{E}(\boldsymbol{f}_0 + \boldsymbol{\rho}) \\ \text{s.t.} \quad & \boldsymbol{B}\boldsymbol{\rho} = \boldsymbol{0} \end{aligned}$$

It is useful to think the variable $\boldsymbol{f} \in \mathbb{R}^m$ as a flow in a graph. Define $C := \{\boldsymbol{f} : \boldsymbol{B}\boldsymbol{f} = \boldsymbol{0}\}$ which is essentially the kernel space of $\boldsymbol{B}$. C is also called the "cycle space" as it is the set of cycle flows when treating $\boldsymbol{f}$ as flows. Restricting the domain of $\mathcal{E}$ to the cycle space yields a new function $\hat{\mathcal{E}} : S \cap C \to \mathbb{R}$ s.t. $\hat{\mathcal{E}}(\boldsymbol{f}) = \mathcal{E}(\boldsymbol{f})$ for any $\boldsymbol{f} \in C$. How does $\boldsymbol{\nabla}\hat{\mathcal{E}}$ look like compared to $\boldsymbol{\nabla}\mathcal{E}$? Let Π_C be the orthogonal projection matrix onto $\boldsymbol{C}$, meaning Π_C is symmetric and $\Pi_C\boldsymbol{\delta} = \boldsymbol{\delta}$ for any $\boldsymbol{\delta} \in C$. Given any $\boldsymbol{f} \in \mathbb{R}^m$, add to it an infinitesimal $\boldsymbol{\delta} \in C$, then

$$\begin{aligned} \mathcal{E}(\boldsymbol{f} + \boldsymbol{\delta}) &\approx \mathcal{E}(\boldsymbol{f}) + \langle \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{f}), \boldsymbol{\delta}\rangle \\ &= \mathcal{E}(\boldsymbol{f}) + \langle \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{f}), \Pi_C\boldsymbol{\delta}\rangle \\ &= \mathcal{E}(\boldsymbol{f}) + \langle \Pi_C\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{f}), \boldsymbol{\delta}\rangle \end{aligned}$$

This tells us the gradient of $\hat{\mathcal{E}}$ at a point $\boldsymbol{f} \in C$ is equal to the projection of gradient of $\boldsymbol{\nabla}\mathcal{E}$ at $\boldsymbol{f}$ onto the subspace C. Similarly,

$$\mathcal{E}(\boldsymbol{f} + \boldsymbol{\delta}) = \mathcal{E}(\boldsymbol{f}) + \langle \Pi_C\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{f}), \boldsymbol{\delta}\rangle + \frac{1}{2}\langle \boldsymbol{\delta}, \Pi_C\boldsymbol{H}_{\mathcal{E}}(\boldsymbol{f})\Pi_C\boldsymbol{\delta}\rangle$$

Note that if $\mathcal{E}$ has K-stable Hessian, then $\hat{\mathcal{E}}$ also has K-stable Hessian.

What is a Newton step $\boldsymbol{\delta}^*$ in a linearly constrained optimization problem? $\boldsymbol{\delta}^*$ should be a minimizer of

$$\begin{aligned} &\min_{\substack{\boldsymbol{\delta}\in\mathbb{R}^m \\ \boldsymbol{B}\boldsymbol{\delta}=\boldsymbol{0}}} \langle \underbrace{\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{f})}_{=:\boldsymbol{g}}, \boldsymbol{\delta}\rangle + \frac{1}{2}\langle \boldsymbol{\delta}, \underbrace{\boldsymbol{H}_{\mathcal{E}}(\boldsymbol{f})}_{=:\boldsymbol{H}}\boldsymbol{\delta}\rangle \\ &\quad \text{(Lagrange duality)} \\ \iff &\max_{\boldsymbol{x}\in\mathbb{R}^n}\min_{\boldsymbol{\delta}\in\mathbb{R}^m} \underbrace{\langle \boldsymbol{g}, \boldsymbol{\delta}\rangle + \frac{1}{2}\langle \boldsymbol{\delta}, \boldsymbol{H}\boldsymbol{\delta}\rangle - \boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{\delta}}_{\text{Lagrangian } L(\boldsymbol{\delta},\boldsymbol{x})} \end{aligned} \tag{12}$$

Applying the KKT optimality conditions, one has

$$\begin{aligned} \boldsymbol{B}\boldsymbol{\delta} &= \boldsymbol{0}, \\ \boldsymbol{\nabla}_{\boldsymbol{\delta}} L(\boldsymbol{\delta}, \boldsymbol{x}) = \boldsymbol{g} + \boldsymbol{H}\boldsymbol{\delta} - \boldsymbol{B}^\top\boldsymbol{x} &= \boldsymbol{0}, \end{aligned}$$

from which we get

$$\begin{aligned} \boldsymbol{\delta} + \boldsymbol{H}^{-1}\boldsymbol{g} &= \boldsymbol{H}^{-1}\boldsymbol{B}^\top\boldsymbol{x} \\ \underbrace{\boldsymbol{B}\boldsymbol{\delta}}_{=\boldsymbol{0}} + \boldsymbol{B}\boldsymbol{H}^{-1}\boldsymbol{g} &= \boldsymbol{B}\boldsymbol{H}^{-1}\boldsymbol{B}^\top\boldsymbol{x} \\ \boldsymbol{B}\boldsymbol{H}^{-1}\boldsymbol{g} &= \underbrace{\boldsymbol{B}\boldsymbol{H}^{-1}\boldsymbol{B}^\top}_{=:\boldsymbol{L}}\boldsymbol{x} \end{aligned}$$

Finally, the solutions to (12) are

$$\begin{cases} \boldsymbol{x}^* &= \boldsymbol{L}^{-1}\boldsymbol{B}\boldsymbol{H}^{-1}\boldsymbol{g} \\ \boldsymbol{\delta}^* &= -\boldsymbol{H}^{-1}\boldsymbol{g} + \boldsymbol{H}^{-1}\boldsymbol{B}^\top \boldsymbol{x}^* \end{cases}$$

It is easy to verify that $\boldsymbol{B}\boldsymbol{\delta}^* = \boldsymbol{0}$. Thus, our update rule is $\boldsymbol{f}_{i+1} = \boldsymbol{f}_i + \boldsymbol{\delta}^*$. And we have the following convergence result.

Theorem 4.3. $\hat{\mathcal{E}}(\boldsymbol{f}_k) - \hat{\mathcal{E}}(\boldsymbol{f}^*) \leq \epsilon \cdot \left(\hat{\mathcal{E}}(\boldsymbol{f}_0) - \hat{\mathcal{E}}(\boldsymbol{f}^*)\right)$ *when* $k > (K+1)^6 \log(1/\epsilon)$.

Remark 4.4. Note if $\mathcal{E}(\boldsymbol{f}) = \sum_{i=1}^m \mathcal{E}_i(\boldsymbol{f}(i))$, then $\boldsymbol{H}_{\mathcal{E}}(\boldsymbol{f})$ is diagonal. Thus, $\boldsymbol{L} = \boldsymbol{B}\boldsymbol{H}^{-1}\boldsymbol{B}^\top$ is indeed a Laplacian provided that $\boldsymbol{B}$ is an incidence matrix. Therefore, the linear equations we need to solve to apply Newton's method in a network flow setting are Laplacians, which means we can solve them very quickly.