Advanced Graph Algorithms and Optimization Spring 2023

Fenchel Conjugates and Newton Steps

R. Kyng & M. Probst Gutenberg *Problem Set 9 — Wednesday, May 10th*

The exercises for this week will not count toward your grade, but you are highly encouraged to solve them all. This exercise sheet has exercises related to week 9. We encourage you to start early so you have time to go through everything.

To get feedback, you must hand in your solutions by 23:59 on May 18. Both hand-written and LaTeX solutions are acceptable, but we will only attempt to read legible text.

Exercise 1: Fenchel Conjugate Examples

- Consider the function $\mathcal{E}(y) = y^4/4$, where $\mathcal{E} : \mathbb{R} \to \mathbb{R}$. Compute the Fenchel conjugate $\mathcal{E}^*$ of $\mathcal{E}$. Verify the gradient and Hessian relationships between these that we discussed in class.

  Solution We have,

$$\mathcal{E}^*(z) = \sup_{y \in \mathbb{R}} \left(yz - \frac{y^4}{4} \right)$$

The expression is maximized when its derivative with respect to y is 0 (Since it is a concave function in y over an open domain). Therefore,

$$\begin{aligned} \frac{d\left(yz - \frac{y^4}{4}\right)}{dy} &= 0 \\ \implies z - y^3 &= 0 \\ \implies y &= z^{1/3} \end{aligned}$$

This implies,

$$\begin{aligned} \mathcal{E}^*(z) &= z \cdot z^{1/3} - \frac{(z^{1/3})^4}{4} \\ &= \frac{z^{4/3}}{4/3} \end{aligned}$$

We see that $\boldsymbol{\nabla}\mathcal{E}(y) = y^3$, $\boldsymbol{\nabla}\mathcal{E}^*(z) = z^{1/3}$, $H_{\mathcal{E}}(y) = 3y^2$ and $H_{\mathcal{E}^*}(z) = \frac{z^{-2/3}}{3}$. We can then indeed verify that,

$$\begin{aligned}
\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{\nabla}\mathcal{E}^*(z)) &= \boldsymbol{\nabla}\mathcal{E}(z^{1/3}) = z^{3/3} = z \\
\boldsymbol{\nabla}\mathcal{E}^\star(\boldsymbol{\nabla}\mathcal{E}(y)) &= \boldsymbol{\nabla}\mathcal{E}^*(y^3) = y^{3/3} = y \\
H_{\mathcal{E}^*}(\boldsymbol{\nabla}\mathcal{E}(y)) &= H_{\mathcal{E}^*}(y^3) \\
&= \frac{y^{-6/3}}{3} = \frac{1}{3y^2} \\
&= \frac{1}{H_{\mathcal{E}}(y)} = H_{\mathcal{E}}^{-1}(y)
\end{aligned}$$

- Consider the function $\mathcal{E}(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}\|^2$, where $\|\cdot\|$ is an arbitrary norm. Compute the Fenchel conjugate $\mathcal{E}^*$ the function.
  Hint: express your answer using the dual norm

$$\|\boldsymbol{z}\|_* = \max_{\substack{\boldsymbol{y} \text{ s.t.} \\ \|\boldsymbol{y}\| \le 1}} \boldsymbol{z}^\top \boldsymbol{y}.$$

Solution We have,

$$\begin{aligned}
\mathcal{E}^*(\boldsymbol{z}) &= \sup_{\boldsymbol{y}\in\mathbb{R}^n} \left(\boldsymbol{z}^\top\boldsymbol{y} - \frac{1}{2}\|\boldsymbol{y}\|^2\right) \\
&= \sup_{c\in\mathbb{R}_{\ge 0}} \sup_{\substack{\boldsymbol{y} \text{ s.t.} \\ \|\boldsymbol{y}\|=c}} \left(\boldsymbol{z}^\top\boldsymbol{y} - \frac{1}{2}c^2\right) \\
&= \sup_{c\in\mathbb{R}_{\ge 0}} -\frac{1}{2}c^2 + \sup_{\substack{\boldsymbol{y} \text{ s.t.} \\ \|\boldsymbol{y}\|=c}} \boldsymbol{z}^\top\boldsymbol{y} \\
&= \sup_{c\in\mathbb{R}_{\ge 0}} -\frac{1}{2}c^2 + c\cdot \sup_{\substack{\boldsymbol{y} \text{ s.t.} \\ \|\boldsymbol{y}\|=1}} \boldsymbol{z}^\top\boldsymbol{y} \\
&= \sup_{c\in\mathbb{R}_{\ge 0}} -\frac{1}{2}c^2 + c\|\boldsymbol{z}\|_*
\end{aligned}$$

The expression is maximized when its derivative with respect to c is 0 or potentially when $c = 0$ (boundry case since $\mathbb{R}_{\ge 0}$ is not an open domain). We an ignore the boundry case because we will see that it will only be optimal when $\|\boldsymbol{z}\|_* = 0$. Therefore,

$$\begin{aligned}
\frac{d\left(-\frac{1}{2}c^2 + c\|\boldsymbol{z}\|_*\right)}{dc} &= 0 \\
\implies -c + \|\boldsymbol{z}\|_* &= 0 \\
\implies c &= \frac{\|\boldsymbol{z}\|_*}{2}
\end{aligned}$$

This implies,

$$\begin{aligned}\mathcal{E}^*(\boldsymbol{z}) &= -\frac{1}{2}\|\boldsymbol{z}\|_*^2 + \|\boldsymbol{z}\|_*^2 \\ &= \|\boldsymbol{z}\|_*^2\end{aligned}$$

- Consider the function $\mathcal{E}(\boldsymbol{y}) = \|\boldsymbol{y}\|$, where $\|\cdot\|$ is an arbitrary norm. Compute the Fenchel conjugate $\mathcal{E}^*$ the function. *Warning: you may get ∞ somewhere....*

  Solution We have,

$$\begin{aligned}\mathcal{E}^*(\boldsymbol{z}) &= \sup_{\boldsymbol{y}\in\mathbb{R}^n} \left(\boldsymbol{z}^\top \boldsymbol{y} - \|\boldsymbol{y}\|\right) \\ &= \sup_{c\in\mathbb{R}_{\geq 0}} \sup_{\substack{\boldsymbol{y} \text{ s.t.} \\ \|\boldsymbol{y}\|=c}} \left(\boldsymbol{z}^\top \boldsymbol{y} - c\right) \\ &= \sup_{c\in\mathbb{R}_{\geq 0}} -c + \sup_{\substack{\boldsymbol{y} \text{ s.t.} \\ \|\boldsymbol{y}\|=c}} \boldsymbol{z}^\top \boldsymbol{y} \\ &= \sup_{c\in\mathbb{R}_{\geq 0}} -c + c \cdot \sup_{\substack{\boldsymbol{y} \text{ s.t.} \\ \|\boldsymbol{y}\|=1}} \boldsymbol{z}^\top \boldsymbol{y} \\ &= \sup_{c\in\mathbb{R}_{\geq 0}} -c + c\|\boldsymbol{z}\|_* \\ &= \sup_{c\in\mathbb{R}_{\geq 0}} c \cdot (\|\boldsymbol{z}\|_* - 1)\end{aligned}$$

  In this situation, we have 2 cases. If $(\|\boldsymbol{z}\|_* - 1) \leq 0$ then $c = 0$ is the optimal. Otherwise, we can push c as high as we want and get ∞. Therefore,

$$\mathcal{E}^*(\boldsymbol{z}) = \begin{cases} 0 & \text{if } \|\boldsymbol{z}\|_* \leq 1 \\ \infty & \text{otherwise} \end{cases}$$

- *UPDATED.* Consider the function $\mathcal{E} : \mathbb{R} \to \mathbb{R}$ given by

$$\mathcal{E}(y) = \begin{cases} |y|^{2.2}/(2.2) \text{ if } |y| \leq 1 \\ |y|^{1.1} - (1 - 1/2.2) \text{ if } |y| > 1 \end{cases}$$

  Compute its Fenchel conjugate $\mathcal{E}^*$.

  Solution We have,

$$\mathcal{E}^*(z) = \sup_{y\in\mathbb{R}} (yz - \mathcal{E}(y))$$

  We notice that since $\mathcal{E}(y)$ is symmetric along $y = 0$ (i.e $\mathcal{E}(y) = \mathcal{E}(-y)$), we can argue that $\mathcal{E}^*(z)$ is too. This is because for any $y \in \mathbb{R}$, $yz - \mathcal{E}(y) = (-y)(-z) - \mathcal{E}(-y)$ and we take a supremum over the domain. So $\mathcal{E}^*(z) = \mathcal{E}^*(-z) = \mathcal{E}^*(|z|)$. Therefore, we can just focus on $z \geq 0$ for now. $(yz - \mathcal{E}(y))$ is a continuous concave function, but it is not differentiable when $|y| = 1$. Thus, we need to consider 3 cases, $|y| < 1$, $|y| > 1$ and $|y| = 1$. Setting derivative to 0 for $|y| < 1$ we get,

$$\begin{aligned}
|y|^{1.2} &= z \\
\implies y &= z^{1/1.2} \\
\implies (yz - \mathcal{E}(y)) &= z \cdot z^{1/1.2} - \frac{z^{2.2/1.2}}{2.2} \\
&= z^{2.2/1.2} (1 - 1/2.2) \\
&= \frac{z^{2.2/1.2}}{2.2/1.2}
\end{aligned}$$

But the above is only valid for when $|y| < 1$ which implies that it only applies when $z < 1$. Setting derivative to 0 for $|y| > 1$ we get,

$$\begin{aligned}
1.1 \, |y|^{0.1} &= z \\
\implies y &= (z/1.1)^{10} \\
\implies (yz - \mathcal{E}(y)) &= z \cdot (z/1.1)^{10} - (z/1.1)^{11} + (1 - 1/2.2) \\
&= (z/1.1)^{10} \cdot (z - z/1.1) + (1 - 1/2.2) \\
&= \frac{z \cdot (z/1.1)^{10}}{11} + (1 - 1/2.2)
\end{aligned}$$

But the above is only valid for when $|y| > 1$ which implies that it only applies when $z > 1.1$. In the case when $|y| = 1$ we get,

$$(yz - \mathcal{E}(y)) = z - 1/2.2$$

Notice that the function is convex and thus either will have one maximizer (or will approach infinity, which is not relevant here). Therefore, we are already done for when $0 \geq z < 1$ and $z > 1.1$ according to setting derivative to zero. In the case of $1 \leq z \leq 1.1$ the only maximizing point can be $y = 1$. Thus,

$$\mathcal{E}^*(z) = \begin{cases} \frac{|z|^{2.2/1.2}}{2.2/1.2} & \text{if } |z| < 1 \\ \frac{|z| \cdot (|z|/1.1)^{10}}{11} + (1 - 1/2.2) & \text{if } |z| > 1.1 \\ |z| - 1/2.2 & \text{o.w.} \end{cases}$$

- *BONUS (NEW).* Consider the function $\mathcal{E} : \mathbb{R} \to \mathbb{R}$ given by

$$\mathcal{E}(y) = \begin{cases} |y|^{2.2}/(2.2) \text{ if } |y| \leq 1 \\ |y|^{1.1}/1.1 - (1/1.1 - 1/2.2) \text{ if } |y| > 1 \end{cases}$$

Compute its Fenchel conjugate $\mathcal{E}^*$.

We have,

$$\mathcal{E}^*(z) = \sup_{y \in \mathbb{R}} (yz - \mathcal{E}(y))$$

Similar to the previous problem, we have $\mathcal{E}^*(z) = \mathcal{E}^*(-z) = \mathcal{E}^*(|z|)$ implying we can just focus on $z \geq 0$ for now. However, in this problem, $(yz - \mathcal{E}(y))$ is not just convex and continuous, but also differentiable everywhere. Thus, we can directly consider only the cases $|y| \leq 1$ and $|y| > 1$.

Similar to the previous problem, for the case $|y| \leq 1$, setting derivative to 0 gives,

$$\begin{aligned}
|y|^{1.2} &= z \\
\implies y &= z^{1/1.2} \\
\implies (yz - \mathcal{E}(y)) &= \frac{z^{2.2/1.2}}{2.2/1.2}
\end{aligned}$$

The above case is valid when $|y| \leq 1$, implying it applies to $z \leq 1$. Now, setting derivative to 0 for $|y| > 1$,

$$\begin{aligned}
|y|^{0.1} &= z \\
\implies y &= z^{10} \\
\implies (yz - \mathcal{E}(y)) &= z^{11} - z^{11}/1.1 + (1/1.1 - 1/2.2) \\
&= \frac{z^{11}}{11} + (1/1.1 - 1/2.2)
\end{aligned}$$

We have all cases covered for z, therefore,

$$\mathcal{E}^*(z) = \begin{cases} \frac{|z|^{2.2/1.2}}{2.2/1.2} & \text{if } |z| \leq 1 \\ \frac{z^{11}}{11} + (1/1.1 - 1/2.2) & \text{o.w.} \end{cases}$$

Exercise 2: Fenchel Twice

- Consider a convex function $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ with Fenchel conjugate $\mathcal{E}^*$. Suppose $\boldsymbol{\nabla}\mathcal{E}$ is a bijection. Explain why $\boldsymbol{\nabla}\mathcal{E}^*$ is the inverse function of $\boldsymbol{\nabla}\mathcal{E}$.

  Solution Because the gradient $\boldsymbol{\nabla}_{\boldsymbol{y}}\mathcal{E}$ is a bijection onto $\mathbb{R}^n$, given any $\boldsymbol{z} \in \mathbb{R}^n$, there exists a unique $\boldsymbol{y}$ such that $\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}) = \boldsymbol{z}$. Let $\boldsymbol{y}(\boldsymbol{z})$ be a $\boldsymbol{y}$ s.t. $\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}) = \boldsymbol{z}$.

  The function $\boldsymbol{y} \mapsto \langle \boldsymbol{z}, \boldsymbol{y} \rangle - \mathcal{E}(\boldsymbol{y})$ is concave in $\boldsymbol{y}$ and has gradient $\boldsymbol{z} - \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})$ and is hence maximized at $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{z})$. This follows because we know a differentiable convex function is minimized when its gradient is zero and so a differentiable concave function is maximized when its gradient is zero.

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 $\boldsymbol{\nabla}_{\boldsymbol{z}}\mathcal{E}^*(\boldsymbol{z}) = \boldsymbol{y}(\boldsymbol{z})$. We 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, $\boldsymbol{\nabla}\mathcal{E}^*(\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y})) = \boldsymbol{y}$.

- Sketch a proof that given a convex function $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ with Fenchel conjugate $\mathcal{E}^*$, we have $(\mathcal{E}^*)^* = \mathcal{E}$. I.e. the Fenchel conjugate transformation applied twice returns the original function. You may impose any technical conditions on $\mathcal{E}$ that you like.

Solution 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 (assume that $\boldsymbol{z}$ is unique), in the above program. We then have $\boldsymbol{u} = \boldsymbol{\nabla}\mathcal{E}^*(\boldsymbol{z}(\boldsymbol{u}))$. Let us also assume that $\boldsymbol{\nabla}\mathcal{E}$ is a bijection. Letting $\boldsymbol{y}(\boldsymbol{z})$ be defined as in the previous part, we get $\boldsymbol{y}(\boldsymbol{z}(\boldsymbol{u})) = \boldsymbol{\nabla}_{\boldsymbol{z}}\mathcal{E}^*(\boldsymbol{z}(\boldsymbol{u})) = \boldsymbol{u}$

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

Exercise 3: Fenchel Transformation

Consider a convex function $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ with Fenchel conjugate $\mathcal{E}^*$.

- Give a convenient expression for the Fenchel conjugate of $\widehat{\mathcal{E}}(\boldsymbol{y}) = \mathcal{E}(\boldsymbol{y} + \boldsymbol{t})$, where $\boldsymbol{t}$ is a fixed vector.

Solution We write the Fenchel conjugate as:

$$\begin{aligned}
\widehat{\mathcal{E}}^*(\boldsymbol{y}) &= \sup_{\boldsymbol{y}\in\mathbb{R}^n} \boldsymbol{z}^\top\boldsymbol{y} - \widehat{\mathcal{E}}(\boldsymbol{y}) \\
&= \sup_{\boldsymbol{y}\in\mathbb{R}^n} \boldsymbol{z}^\top\boldsymbol{y} + \boldsymbol{z}^\top\boldsymbol{t} - \boldsymbol{z}^\top\boldsymbol{t} - \mathcal{E}(\boldsymbol{y} + \boldsymbol{t}) \\
&= \boldsymbol{z}^\top\boldsymbol{t} + \sup_{\boldsymbol{y}\in\mathbb{R}^n} \boldsymbol{z}^\top(\boldsymbol{y} - \boldsymbol{t}) - \mathcal{E}(\boldsymbol{y} + \boldsymbol{t}) \\
&= \boldsymbol{z}^\top\boldsymbol{t} + \sup_{(\boldsymbol{y}+\boldsymbol{t})\in\mathbb{R}^n} \boldsymbol{z}^\top(\boldsymbol{y} - \boldsymbol{t}) - \mathcal{E}(\boldsymbol{y} + \boldsymbol{t}) \\
&= \boldsymbol{z}^\top\boldsymbol{t} + \mathcal{E}^*(\boldsymbol{z})
\end{aligned}$$

- Give a convenient expression for the Fenchel conjugate of $\widehat{\mathcal{E}}(\boldsymbol{y}) = \mathcal{E}(\boldsymbol{M}\boldsymbol{y})$, where $\boldsymbol{M}$ is an invertible matrix.

Solution We write the Fenchel conjugate as:

$$
\begin{aligned}
\widehat{\mathcal{E}}^*(\boldsymbol{y}) &= \sup_{\boldsymbol{y}\in\mathbb{R}^n} \boldsymbol{z}^\top \boldsymbol{y} - \widehat{\mathcal{E}}(\boldsymbol{y}) \\
&= \sup_{\boldsymbol{y}\in\mathbb{R}^n} \boldsymbol{z}^\top \boldsymbol{M}^{-1}\boldsymbol{M}\boldsymbol{y} - \mathcal{E}(\boldsymbol{M}\boldsymbol{y}) \\
&= \sup_{\boldsymbol{y}\in\mathbb{R}^n} ((\boldsymbol{M}^{-1})^\top \boldsymbol{z})^\top \boldsymbol{M}\boldsymbol{y} - \mathcal{E}(\boldsymbol{M}\boldsymbol{y}) \\
&= \sup_{\boldsymbol{M}\boldsymbol{y}\in\mathbb{R}^n} ((\boldsymbol{M}^{-1})^\top \boldsymbol{z})^\top \boldsymbol{M}\boldsymbol{y} - \mathcal{E}(\boldsymbol{M}\boldsymbol{y}) \\
&= \sup_{\boldsymbol{y}'\in\mathbb{R}^n} ((\boldsymbol{M}^{-1})^\top \boldsymbol{z})^\top \boldsymbol{y}' - \mathcal{E}(\boldsymbol{y}') \quad \text{(Transformation valid because } \boldsymbol{M} \text{ is invertible)} \\
&= \mathcal{E}^*((\boldsymbol{M}^{-1})^\top \boldsymbol{z})
\end{aligned}
$$

- Give a convenient expression for the Fenchel conjugate of $\widehat{\mathcal{E}}(\boldsymbol{y}) = \mathcal{E}(\boldsymbol{y}) + \boldsymbol{g}^\top \boldsymbol{y}$, where $\boldsymbol{g}$ is a fixed vector.

Solution We write the Fenchel conjugate as:

$$
\begin{aligned}
\widehat{\mathcal{E}}^*(\boldsymbol{y}) &= \sup_{\boldsymbol{y}\in\mathbb{R}^n} \boldsymbol{z}^\top \boldsymbol{y} - \widehat{\mathcal{E}}(\boldsymbol{y}) \\
&= \sup_{\boldsymbol{y}\in\mathbb{R}^n} \boldsymbol{z}^\top \boldsymbol{y} - \mathcal{E}(\boldsymbol{y}) - \boldsymbol{g}^\top \boldsymbol{y} \\
&= \sup_{\boldsymbol{y}\in\mathbb{R}^n} (\boldsymbol{z} - \boldsymbol{g})^\top \boldsymbol{y} - \mathcal{E}(\boldsymbol{y}) \\
&= \mathcal{E}^*(\boldsymbol{z} - \boldsymbol{g})
\end{aligned}
$$

Exercise 4: Newton Steps are affine invariant

Consider a convex function $\mathcal{E}(\boldsymbol{y}) : \mathbb{R}^n \to \mathbb{R}$.

Suppose that starting from $\boldsymbol{y}_0$ we take a Newton step given by $\boldsymbol{\delta} = -H_{\mathcal{E}}(\boldsymbol{y}_0)^{-1}\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}_0)$.

Next, consider the reparameterized function $\widehat{\mathcal{E}}(\boldsymbol{x}) = \mathcal{E}(\boldsymbol{M}\boldsymbol{x})$, where $\boldsymbol{M}$ is invertible. Suppose that starting from $\boldsymbol{x}_0 = \boldsymbol{M}^{-1}\boldsymbol{y}_0$ we take a Newton step given by $\widehat{\boldsymbol{\delta}} = -H_{\widehat{\mathcal{E}}}(\boldsymbol{x}_0)^{-1}\boldsymbol{\nabla}\widehat{\mathcal{E}}(\boldsymbol{x}_0)$.

- Show that $\boldsymbol{y}_0 + \boldsymbol{\delta} = \boldsymbol{M}(\boldsymbol{x}_0 + \widehat{\boldsymbol{\delta}})$.

Solution For $\widehat{\mathcal{E}}$, we have $\boldsymbol{\nabla}\widehat{\mathcal{E}}(\boldsymbol{x}) = \boldsymbol{M}^\top \boldsymbol{\nabla}\mathcal{E}(\boldsymbol{M}\boldsymbol{x})$ and $H_{\widehat{\mathcal{E}}}(\boldsymbol{x}) = \boldsymbol{M}^\top H_{\mathcal{E}}(\boldsymbol{M}\boldsymbol{x})\boldsymbol{M}$. We see that,

$$
\begin{aligned}
\boldsymbol{M}(\boldsymbol{x}_0 + \widehat{\boldsymbol{\delta}}) &= \boldsymbol{y}_0 + \boldsymbol{M}\left(-H_{\widehat{\mathcal{E}}}(\boldsymbol{x}_0)^{-1}\boldsymbol{\nabla}\widehat{\mathcal{E}}(\boldsymbol{x}_0)\right) \\
&= \boldsymbol{y}_0 + \boldsymbol{M}\left(-H_{\widehat{\mathcal{E}}}(\boldsymbol{x}_0)^{-1}\boldsymbol{\nabla}\widehat{\mathcal{E}}(\boldsymbol{x}_0)\right) \\
&= \boldsymbol{y}_0 - \boldsymbol{M}\left(\boldsymbol{M}^{-1}H_{\mathcal{E}}(\boldsymbol{M}\boldsymbol{x})^{-1}(\boldsymbol{M}^\top)^{-1}\boldsymbol{M}^\top\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{M}\boldsymbol{x}_0)\right) \\
&= \boldsymbol{y}_0 - H_{\mathcal{E}}(\boldsymbol{M}\boldsymbol{x})^{-1}\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{M}\boldsymbol{x}_0) \\
&= \boldsymbol{y}_0 + \boldsymbol{\delta}
\end{aligned}
$$

- Suppose our updates $\boldsymbol{\delta}$ came from gradient descent instead of Newton steps – would a similar relationship still hold? What if $\boldsymbol{M}$ is orthonormal?

Solution We have $\boldsymbol{\delta} = -\frac{1}{\beta}\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}_0)$ and $\widehat{\boldsymbol{\delta}} = -\frac{1}{\beta}\boldsymbol{\nabla}\widehat{\mathcal{E}}(\boldsymbol{x}_0)$. We again need to show that $\boldsymbol{y}_0 + \boldsymbol{\delta} = \boldsymbol{M}(\boldsymbol{x}_0 + \widehat{\boldsymbol{\delta}})$. We see that,

$$
\begin{aligned}
\boldsymbol{M}(\boldsymbol{x}_0 + \widehat{\boldsymbol{\delta}}) &= \boldsymbol{y}_0 + \boldsymbol{M}\left(-\frac{1}{\beta}\boldsymbol{\nabla}\widehat{\mathcal{E}}(\boldsymbol{x}_0)\right) \\
&= \boldsymbol{y}_0 + \boldsymbol{M}\left(-\frac{1}{\beta}\cdot\boldsymbol{M}^\top\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{M}\boldsymbol{x}_0)\right) \\
&= \boldsymbol{y}_0 + \boldsymbol{M}\boldsymbol{M}^\top\left(-\frac{1}{\beta}\boldsymbol{\nabla}\mathcal{E}(\boldsymbol{y}_0)\right) \\
&= \boldsymbol{y}_0 + \boldsymbol{M}\boldsymbol{M}^\top\boldsymbol{\delta}
\end{aligned}
$$

Therefore $\boldsymbol{M}(\boldsymbol{x}_0+\widehat{\boldsymbol{\delta}}) = \boldsymbol{y}_0+\boldsymbol{\delta}$ if we have $\boldsymbol{M}\boldsymbol{M}^\top = \boldsymbol{I}$ i.e. $\boldsymbol{M}$ is orthogonal but not in general.