MATH-329 Nonlinear optimization
Exercise session 4: Hessians and Newton's method

Instructor: Nicolas Boumal
TAs: Andrew McRae, Andreea Musat

Document compiled on September 25, 2024

Exercises marked with $(*)$ will be used in later exercises or in the homeworks: you might want to prioritize those.

1. Good and bad global behavior of Newton's method.

1. Consider the function $f(x) = \frac{1}{4}x^4 - x^2 + 2x + 1$. What is the behavior of Newton's method on f if the initial point is $x_0 = 0$? (Observe numerically first.)

2. Argue that $f(x) = \log(e^x + e^{-x})$ has a Lipschitz continuous gradient, a Lipschitz continuous Hessian, and is strictly convex (it helps to plot the function). What is the behavior of Newton's method on this function with $x_0 = 1$? And with $x_0 = 1.5$?

3. Consider the Rosenbrock function
$$f(x, y) = (a - x)^2 + b(y - x^2)^2$$
with $a = 1$ and $b = 100$. Run Newton's method with $x_0 = (-1.2, 1)$. Compare the performance with the one of gradient descent for this initial point.

Answer.

1. The function f is 1-dimensional (see Figure 1). In this case the iterates are
$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}.$$
Therefore we have
$$x_1 = 0 - \frac{0^3 - 2 \cdot 0 + 2}{3 \cdot 0^2 - 2} = 1 \qquad \text{and} \qquad x_2 = 1 - \frac{1^3 - 2 \cdot 1 + 2}{3 \cdot 1^2 - 2} = 0.$$
This pattern repeats indefinitely: for all k we have
$$x_k = \begin{cases} 0 & \text{if } k \text{ is even} \\ 1 & \text{if } k \text{ is odd.} \end{cases}$$
In this situation Newton's method exhibits a limit cycle. It is also an attracting limit cycle in the sense that starting from any point in some neighborhood of 0 and 1 will produce iterates with this limit cycle. This phenomenon is not specific to this very simple optimization instance and can also occur in more realistic situations.

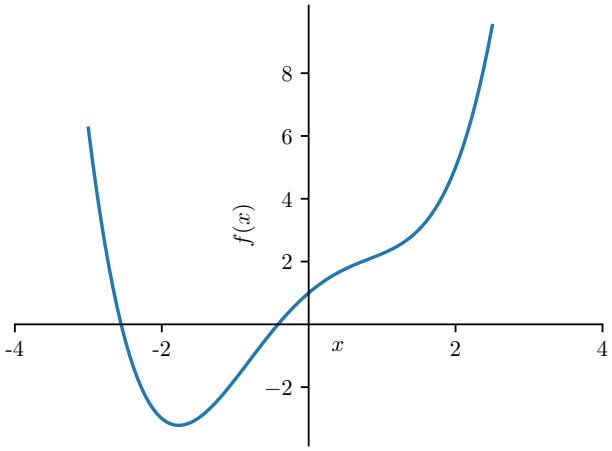

Figure 1: $f(x) = \frac{1}{4}x^4 - x^2 + 2x + 1$.

2. The function f is 1-dimensional (see Figure 2). We find that for all $x \in \mathbb{R}$ the first and

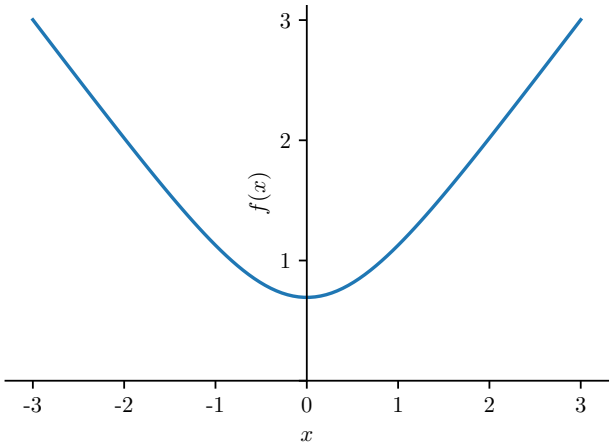

Figure 2: $f(x) = \log(e^x + e^{-x})$.

second derivatives are

$$f'(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = 1 - 2\frac{e^{-x}}{e^x + e^{-x}} \qquad \text{and} \qquad f''(x) = \frac{4}{(e^x + e^{-x})^2}.$$

We bound the second derivative as $0 < f''(x) \leq 1$ for all $x \in \mathbb{R}$. This shows that f is strictly convex and has 1-Lipschitz continuous gradients. (Notice that the bound $|f'(x)| < 1$ gives that f is also 1-Lipschitz.) For the Lipschitz continuity of f'' we can either find some constant L' to bound $|f''(x) - f''(y)| \leq L'|x - y|$. Or an alternative method consists in bounding the third derivative

$$f'''(x) = -\frac{8(e^x - e^{-x})}{(e^x + e^{-x})^3}$$

(see Figure 3). Clearly f''' is an odd function, non-negative on $\mathbb{R}_-$, and non-positive on

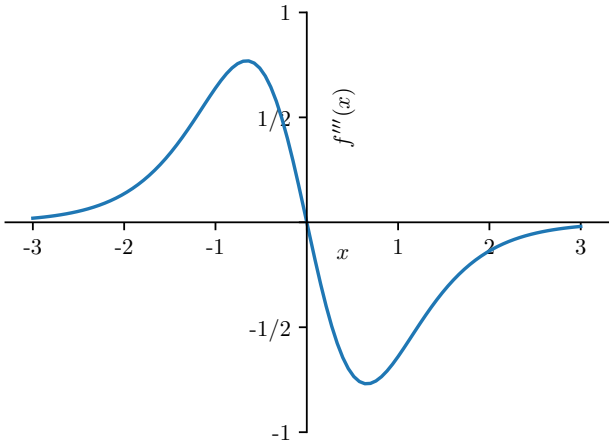

Figure 3: $f'''(x) = -\frac{8(e^x - e^{-x})}{(e^x + e^{-x})^3}$.

$\mathbb{R}_+$. It is also easy to see that $\lim_{x \to -\infty} f'''(x) = 0$ so f''' attains its maximum on $\mathbb{R}_-$. The maximum must be a critical point; we find that the derivative of f''' is given by

$$f''''(x) = \frac{16e^{2x}(1 - 4e^{2x} + e^{4x})}{(1 + e^{2x})^4}$$

for all x. This quantity is zero if and only if $1 - 4e^{2x} + e^{4x} = 0$. This equation has a unique solution in $\mathbb{R}_-$ given by

$$x^\star = \frac{1}{2}\log(2 - \sqrt{3}).$$

So $|f'''|$ is bounded above by $f'''(x^\star) = \frac{4}{3\sqrt{3}}$. We conclude that f'' is L'-Lipschitz continuous with $L' = \frac{4}{3\sqrt{3}}$.

With $x_0 = 1$ Newton's method achieves convergence with machine precision in a few iterations. However, despite the favorable properties of f, with $x_0 = 1.5$ the iterations diverge. This is another example of what can go wrong with Newton's method.

3. Gradient descent with constant step-size needs several dozens of thousands of iterations to find a point such that the gradient norm is less than 10^{-6}. Gradient descent with backtracking linesearch is also very slow: the number of iterations that we need is within the same order of magnitude. In contrast Newton's method converges with machine precision in less than 10 iterations.

■

2.$(*)$ **Regression loss function.** We let $y \in \mathbb{R}^m$ and $F \colon \mathbb{R}^n \to \mathbb{R}^m$ be a C^2 map. Consider the C^2 regression function $f \colon \mathbb{R}^n \to \mathbb{R}$ defined by

$$f(x) = \frac{1}{2}\|F(x) - y\|^2.$$

Find the gradient and the Hessian of f as a function of the derivatives of F.
Hint: There are two ways to see this. You can see F as m scalar functions that have a gradient and a Hessian and express everything as a function of this. Or you can define the Jacobian of F as $J(x) = \mathrm{D}F(x) \in \mathbb{R}^{m\times n}$. Then J is a function $\mathbb{R}^n \to \mathbb{R}^{m\times n}$ and for all $x, u \in \mathbb{R}^n$ we have $\mathrm{D}J(x)[u] \in \mathbb{R}^{m\times n}$.

Answer. Remember that if $h\colon \mathcal{E} \to \mathcal{F}$, $g\colon \mathcal{F} \to \mathcal{G}$ and $f = g \circ h\colon \mathcal{E} \to \mathcal{G}$ then for all $x, u \in \mathcal{E}$ we have

$$\mathrm{D}f(x)[u] = \mathrm{D}g(h(x))[\mathrm{D}h(x)[u]].$$

This is known as the chain rule.

Using the chain rule (or the product rule for the inner product) we find that for all $x, u \in \mathbb{R}^n$ we have

$$\begin{aligned}\mathrm{D}f(x)[u] &= \langle F(x) - y, \mathrm{D}F(x)[u]\rangle \\ &= \langle \mathrm{D}F(x)^*[F(x) - y], u\rangle\,,\end{aligned}$$

where $\mathrm{D}F(x)^*\colon \mathbb{R}^m \to \mathbb{R}^n$ is the adjoint of $\mathrm{D}F(x)\colon \mathbb{R}^n \to \mathbb{R}^m$. We deduce that for all $x \in \mathbb{R}^n$ we have

$$\nabla f(x) = \mathrm{D}F(x)^*[F(x) - y].$$

For all $x \in \mathbb{R}^n$ we let $J(x) = \mathrm{D}F(x)$ denote the Jacobian of F at x so we can rewrite

$$\nabla f(x) = J(x)^*[F(x) - y].$$

Using the product rule we find that for all $x, u \in \mathbb{R}^n$ we have

$$\nabla^2 f(x)[u] = (\mathrm{D}J(x)[u]^*)[F(x) - y] + J(x)^*[J(x)[u]]$$

To convince yourself you can also work from the definition:

$$\nabla^2 f(x)[u] = \lim_{t\to 0} \frac{\nabla f(x + tu) - \nabla f(x)}{t}.$$

Use the Taylor expansion $J(x + tu) = J(x) + t\mathrm{D}J(x)[u] + O(t^2)$ to compute the limit above.

If you are worried about what happens with the adjoint, check that taking the adjoint of a linear map is a linear operation and think about what happen when we differentiate linear operations. ■

3. Computing Hessians. For the following functions $f\colon \mathcal{E} \to \mathbb{R}$ give an expression for the gradient and the Hessian. Specifically, for the Hessian compute $\nabla^2 f(x)[u]$ for all $x, u \in \mathcal{E}$.

1. Given $a \in \mathbb{R}^n$, consider $f(x) = \frac{1}{2}(x^\top x + (a^\top x)^2)$ for all $x \in \mathbb{R}^n$.
2. $(*)$ Given $A \in \mathbb{R}^{m\times n}$, $M \in \mathbb{R}^{n\times n}$, consider the function $f\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ defined by

$$f(X) = \frac{1}{2}\|A^\top X - M\|_{\mathrm{F}}^2,$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm.

Answer.

1. We have already encountered this type of function, but not in this form. In fact, note that we may rewrite f as

$$\begin{aligned} f(x) &= \frac{1}{2}(x^\top x + x^\top a a^\top x) \\ &= \frac{1}{2} x^\top (I + aa^\top) x. \end{aligned}$$

From this we deduce that for all $x \in \mathbb{R}^n$

$$\nabla f(x) = (I + aa^\top)x \qquad \text{and} \qquad \nabla^2 f(x) = (I + aa^\top). \tag{1}$$

In particular for all $x, u \in \mathbb{R}^n$ we have

$$\nabla^2 f(x)\,[u] = u + aa^\top u.$$

2. In exercise 3 of exercise sheet 1 we found that for all X we have

$$\nabla f(X) = AA^\top X - AM.$$

Therefore, it follows that for all $X, U \in \mathbb{R}^{m \times n}$

$$\begin{aligned} \nabla^2 f(X)[U] &= \lim_{t \to 0} \frac{\nabla f(X + tU) - \nabla f(X)}{t} \\ &= \lim_{t \to 0} \frac{AA^\top (X + tU) - AA^\top X}{t} \\ &= AA^\top U. \end{aligned}$$

■