MATH-329 Continuous optimization
Exercise session 1: getting started

Instructor: Nicolas Boumal
TAs: Andreea Musat and Andrew McRae

Document compiled on September 10, 2024

The course textbook is *Numerical Optimization* by Nocedal and Wright, second edition. You can get a PDF of it here: https://link.springer.com/book/10.1007/978-0-387-40065-5. (Download is free if you are on EPFL network.)

The official programming language for the course is Matlab. This being said, you are free to use another language to complete assignments (for example, Python or Julia) **provided** (a) you get approval from the TAs (who will have to run your code), and (b) you are comfortable enough to translate Matlab instructions to your chosen language yourself.

You can get Matlab here: https://ch.mathworks.com/academia/tah-portal/ecole-polytechnique-federale-de-lausanne-epfl-303238.html.

Matlab Tutorials: http://ubcmatlabguide.github.io/ and https://learnxinyminutes.com/docs/matlab/. See also the Matlab course linked on the Moodle page.

Matlab cheat sheet: http://web.mit.edu/18.06/www/MATLAB/matlab-cheatsheet.pdf.

Absolute best Matlab command ever: `help`. For example, type `help fminunc` (or `help ... fmin` + hit TAB for auto-completion) for an explanation of how `fminunc` works. Want more info? Type `doc fminunc` instead. At the end of a help section, there are links to related functions: explore.

1. Model the following situations as optimization problems. Remember, an optimization problem has two ingredients: a set S and a function $f\colon S \to \mathbb{R}$. Write them in minimization form, as: $\min_{x \in S} f(x)$. (**No need to solve the resulting optimization problems**; just write them down mathematically.)

1. A probability distribution over n objects is a vector $p \in \mathbb{R}^n$ whose entries sum to one and are nonnegative. The entropy of p is

$$H(p) = -\sum_{i=1}^{n} p_i \log(p_i),$$

   where we continuously extend the function $x \mapsto x\log(x)$ at zero (so $0\log(0) \triangleq 0$). Which distribution has maximal entropy?

2. Given k points $x_1, \ldots, x_k \in \mathbb{R}^n$, find the smallest ball in $\mathbb{R}^n$ that contains them.

3. We have n power plants that need to provide electricity to m cities. Each power plant has a maximum energy production capacity of w_i megawatt-hours for $i = 1, \ldots, n$ and each city has an energy demand of d_j megawatt-hours for $j = 1, \ldots, m$. The cost of sending energy from plant i to city j is c_{ij} CHF per megawatt-hour. How can we minimize the cost of energy distribution to meet the demand?

2. Gradients. We will often need to compute gradients (and later, also Hessians) of multivariate functions. Recall that on a Euclidean space $\mathcal{E}$ with inner product $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \|$ the gradient $\nabla f(x)$ of a function $f \colon \mathcal{E} \to \mathbb{R}$ at x is defined by the property:

$$\forall v \in \mathcal{E}, \qquad \langle \nabla f(x), v \rangle = \mathrm{D}f(x)[v], \qquad \text{where} \qquad \mathrm{D}f(x)[v] = \lim_{t \to 0} \frac{f(x+tv) - f(x)}{t}.$$

Let's make sure we are fully comfortable with that definition.

1. Show that if f is differentiable at x then $\nabla f(x)$ as characterized above exists and is unique. Hint: as often when working in linear spaces, it is convenient to introduce a basis.

2. For the special case $\mathcal{E} = \mathbb{R}^n$ with the usual inner product $\langle u, v \rangle = u^\top v = u_1 v_1 + \cdots + u_n v_n$, show that the gradient is nothing but the vector of partial derivatives of f,

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix},$$

   thus recovering (in this special case) what is often presented as the definition of the gradient.

3. For the special case just described, compute the gradient of the function

$$f(x) = \frac{1}{2} x^\top A x + b^\top x + c,$$

   where $A \in \mathbb{R}^{n \times n}$ is a symmetric matrix, $b \in \mathbb{R}^n$ is a vector and $c \in \mathbb{R}$ is a scalar. Try to do it using the definition of the gradient, that is: obtain an expression for $\mathrm{D}f(x)[v]$ first, then try to express the latter in the form $\langle \ldots, v \rangle$. Use a uniqueness argument to deduce a formula for $\nabla f(x)$. We believe that you will, with experience, find this approach much faster and more comfortable than using partial derivatives.

4. We now let $S \in \mathbb{R}^{n \times n}$ be an arbitrary symmetric positive definite matrix, and we define an inner product $\langle u, v \rangle = u^\top S v$. Using this inner product, compute the gradient of the quadratic function f from the previous part. As a sanity check, note that we recover the usual Euclidean inner product when S is the identity matrix.

3. A Euclidean space of matrices. We get another interesting Euclidean space by considering matrices. Let $\mathcal{E} = \mathbb{R}^{m \times n}$ and define the inner product

$$\langle U, V \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} U_{ij} V_{ij}.$$

Convince yourself that this is the same as the Euclidean space of vectors above, only with n replaced by mn.

1. Verify that $\langle U, V \rangle = \mathrm{Tr}(U^\top V)$ where $\mathrm{Tr}(M) = \sum_i M_{ii}$.

2. What is the norm $\|U\| = \sqrt{\langle U, U \rangle}$ associated to this inner product? (It has a name.)

3. Show that for all $U, V \in \mathcal{E}$ we have $\operatorname{Tr}(U^\top V) = \operatorname{Tr}(VU^\top)$. This implies that when we take the trace of a product of matrices of appropriate shapes we can cycle the order:

$$\operatorname{Tr}(ABCD) = \operatorname{Tr}(BCDA) = \operatorname{Tr}(CDAB) = \operatorname{Tr}(DABC).$$

This property is known as the *cyclic invariance* of the trace. What is the equivalent property in terms of inner products?

4. Compute the gradient of the function $f\colon \mathcal{E} \to \mathbb{R}$ defined by

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

where $A \in \mathbb{R}^{m\times n}$ and $M \in \mathbb{R}^{n\times n}$ are fixed matrices. Here too, the most convenient way is to expand the norm using inner products, obtain an expression for $\mathrm{D}f(X)[V]$, and massage that expression until it has the form $\langle \ldots, V\rangle$. You can then identify the gradient by uniqueness.

4. Strict vs isolated local minima (optional). Read page 13 of the N&W textbook.

1. Plot the function $f(x) = x^4 \cos(1/x) + 2x^4$ near $x = 0$ (we define $f(0) = 0$).

2. Convince yourself that it has a strict local minimum at $x^* = 0$ yet that this point is *not* an isolated local minimum.

3. The other way around, argue briefly but conclusively that isolated local minima (of any function f) are strict.