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?

Answer.

1. We define the simplex as

$$S = \{x \in \mathbb{R}^n \mid x_1, \ldots, x_n \geq 0 \text{ and } x_1 + \cdots + x_n = 1\}.$$

This is the set of discrete distributions on n elements. Maximizing the entropy is equivalent to minimizing the negative of the entropy. So we define $f(x) = \sum_{i=1}^{n} x_i \log(x_i)$ and we formulate the problem as

$$\min_{p \in S} f(p).$$

2. This problem is known as the minimum bounding sphere problem. It is a particular instance of a more general class of 1-center problems, where we wish to locate a center point such that it minimizes the maximum distance to some target points. Variations include constraints on the locations of the center point and various notions of distance.

Balls in $\mathbb{R}^n$ are defined by a center and a radius so we can see it as an element of $\mathbb{R}^n \times \mathbb{R}_+$. A ball $(c, r) \in \mathbb{R}^n \times \mathbb{R}_+$ is feasible if for all i we have $\|x_i - c\| \leq r$. So we can define the feasible set as

$$S = \{(c, r) \in \mathbb{R}^n \times \mathbb{R}_+ \mid \max_{1 \leq i \leq k} \|x_i - c\| \leq r\}.$$

Finally we solve the minimization problem

$$\min_{(c,r) \in S} f(c, r)$$

where $f(c, r) = r$.

Notice however that, given the center of the ball, the optimal radius can be computed as the maximal distance to the points. Using this we can optimize only over the center of the ball. We let the feasible set $S = \mathbb{R}^n$ (all possible centers for the ball), and define

$$f(c) = \max_{1 \leq i \leq k} \{\|c - x_i\|\}.$$

Finally we solve the optimization problem

$$\min_{c \in S} f(c).$$

In this second formulation we simplified the feasible set S at the cost of making the function f more complicated.

3. This is a classical instance of a linear transportation problem. We must first define the decision variables to optimize. Let $x_{ij} \geq 0$ denote the energy in megawatt-hours that power plant i sends to city j for all $(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, m\}$. We collect these in a matrix $X = (x_{ij}) \in \mathbb{R}^{n \times m}$. Each power plant i will have the capacity to produce enough energy provided that

$$\sum_{j=1}^{m} x_{ij} \leq w_i.$$

Each city j will have its energy demand satisfied if

$$\sum_{i=1}^{n} x_{ij} \geq d_j.$$

The cost associated to a distribution plan can be computed as

$$\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}c_{ij}.$$

Putting all this together gives an optimization problem of the form

$$\min_{X\in S} \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}c_{ij},$$

with the admissible set S being

$$S = \left\{X \in \mathbb{R}^{n\times m} : X \geq 0,\, X\mathbb{1}_m \leq w,\, X^\top \mathbb{1}_n \geq d\right\}$$

where inequalities should be intended pointwise and we have defined $w = (w_i) \in \mathbb{R}^n$, $d = (d_j) \in \mathbb{R}^m$, $\mathbb{1}_p = (1, \ldots, 1)^\top \in \mathbb{R}^p$.

■

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 $\langle u, v\rangle = u^\top v = u_1v_1 + \cdots + u_nv_n$, 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

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

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 $\langle u, v\rangle = u^\top Sv$. 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.

Answer.

1. Let $\mathcal{E}$ be of dimension n. We can find a basis $\{e_i\}_{i=1}^n$ of $\mathcal{E}$ that is orthonormal with respect to the inner product $\langle \cdot, \cdot \rangle$. It satisfies $\langle e_i, e_j \rangle = \delta_{ij}$ (Kronecker delta) and is such that any $v \in \mathcal{E}$ can be uniquely expanded in the basis as $v = \sum_{i=1}^n \langle v, e_i \rangle e_i$. Therefore the directional derivative of f at x along the direction v can be written as

$$\begin{aligned} \mathrm{D}f(x)[v] &= \mathrm{D}f(x)\left[\sum_{i=1}^n \langle v, e_i \rangle e_i\right] \\ &= \sum_{i=1}^n \langle v, e_i \rangle \mathrm{D}f(x)[e_i] \\ &= \left\langle v, \sum_{i=1}^n \mathrm{D}f(x)[e_i] e_i \right\rangle =: \langle v, \nabla f(x) \rangle , \end{aligned}$$

   where the second and third equality follow from the linearity of the differential and of the inner product respectively. This proves existence and uniqueness by construction.
   An alternative non-constructive proof can be given by noting that the directional derivative of f at x along v is a functional on the vector space $\mathcal{E}$ (a linear mapping associating a scalar to each vector). By the Riesz representation theorem, the exists a unique vector we denote as $\nabla f(x)$ such that

$$\mathrm{D}f(x)[v] = \langle \nabla f(x), v \rangle .$$

2. If $\mathcal{E} = \mathbb{R}^n$ is endowed with the usual inner product, a possible choice for the orthonormal basis $\{e_i\}$ defined in question 1 is the canonical basis: e_i is the vector of all zeros except for entry i that is set to 1. Since by definition of partial derivative we have

$$\frac{\partial f}{\partial x_i} = \mathrm{D}f(x)[e_i] ,$$

   then following the constructive proof of 1 we retrieve

$$\nabla f(x) = \sum_{i=1}^d \mathrm{D}f(x)[e_i] e_i = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix} .$$

3. This objective function is an example where the definition of gradient using directional derivative is more convenient than the partial derivative one. In fact

$$\begin{aligned} f(x+tv) &= \frac{1}{2}(x+tv)^\top A(x+tv) + b^\top(x+tv) + c \\ &= \frac{1}{2}x^\top Ax + b^\top x + c + \frac{t^2}{2}v^\top Av + tb^\top v + \frac{t}{2}x^\top Av + \frac{t}{2}v^\top Ax \\ &= f(x) + \frac{t^2}{2}v^\top Av + t(v^\top b + v^\top Ax), \end{aligned}$$

   since A is symmetric. Therefore

$$\begin{aligned} \mathrm{D}f(x)[v] &= \lim_{t\to 0} \frac{f(x+tv) - f(x)}{t} \\ &= \lim_{t\to 0} \frac{\frac{t^2}{2}v^\top Av + t(v^\top b + v^\top Ax)}{t} \\ &= v^\top(Ax + b) \\ &= \langle v, Ax + b \rangle \\ &= \langle v, \nabla f(x) \rangle . \end{aligned}$$

So we have shown $\nabla f(x) = Ax + b$.

4. With the same computations as in Question 3 we find

$$\begin{aligned} f(x + tv) &= f(x) + t(v^\top b + v^\top Ax) + O(t^2) \\ &= f(x) + t\left\langle v, S^{-1}(b + Ax)\right\rangle + O(t^2). \end{aligned}$$

We conclude that the gradient is given by

$$\nabla f(x) = S^{-1}(b + Ax).$$

We recover the result from the previous question when $S = I$. Changing the inner product defines another geometry where the angles and distances are different. Notice that even though the formula is different this gradient is still the steepest ascent direction for f, only for a different notion of a norm. Can you see why?

■

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 $\mathrm{Tr}(U^\top V) = \mathrm{Tr}(VU^\top)$. This implies that when we take the trace of a product of matrices of appropriate shapes we can cycle the order:

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

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.

Answer.

1. For all $U, V \in \mathcal{E}$ we have

$$\begin{aligned}\operatorname{Tr}(U^\top V) &= \sum_{j=1}^{n} (U^\top V)_{jj} \\ &= \sum_{j=1}^{n} \sum_{i=1}^{m} U_{ij} V_{ij} \\ &= \langle U, V \rangle .\end{aligned}$$

2. From the definition of $\langle \cdot, \cdot \rangle$ it follows immediately that for all $U \in \mathcal{E}$ we have

$$\|U\| = \sqrt{\langle U, U \rangle} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} U_{ij}^2}.$$

   This is the *Frobenius norm*, often denoted by $\| \cdot \|_{\mathrm{F}}$.

3. Given $U, V \in \mathcal{E}$ we have

$$\begin{aligned}\operatorname{Tr}(U^\top V) &= \langle U, V \rangle \\ &= \langle V, U \rangle \\ &= \operatorname{Tr}(V^\top U),\end{aligned}$$

   where we used the symmetry of the inner product. In terms of inner products, this means that given matrices A, B, C of appropriate shapes we have

$$\langle AB, C \rangle = \left\langle B, A^\top C \right\rangle \quad \text{and} \quad \langle AB, C \rangle = \left\langle A, C B^\top \right\rangle .$$

   In words, we can swap the first or the last matrix from a side of the inner product to the other if we transpose it. This property is very useful when computing gradients.

4. Let $X, V \in \mathcal{E}$ and $t \in \mathbb{R}$. Then

$$\begin{aligned}f(X + tV) &= \frac{1}{2} \left\langle A^\top X + tA^\top V - M, A^\top X + tA^\top V - M \right\rangle \\ &= \frac{1}{2} \|A^\top X - M\|^2 + t \left\langle A^\top V, A^\top X - M \right\rangle + O(t^2) \\ &= f(X) + t \left\langle A^\top V, A^\top X - M \right\rangle + O(t^2).\end{aligned}$$

   From this we deduce that

$$\begin{aligned}\mathrm{D}f(X)[V] &= \lim_{t \to 0} \frac{f(X + tV) - f(X)}{t} \\ &= \left\langle A^\top V, A^\top X - M \right\rangle \\ &= \left\langle V, AA^\top X - AM \right\rangle ,\end{aligned}$$

   where we used the trace cyclic invariance for the last equality (that is, the adjoint of A is $A^\top$ for the canonical inner product). The quantity $AA^\top X - AM$ satisfies the definition of the gradient. As the gradient is unique we conclude that $\nabla f(X) = AA^\top X - AM$.

■

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.

Answer.

1. Figure 1 shows the function f continuously extended with $f(0) = 0$.

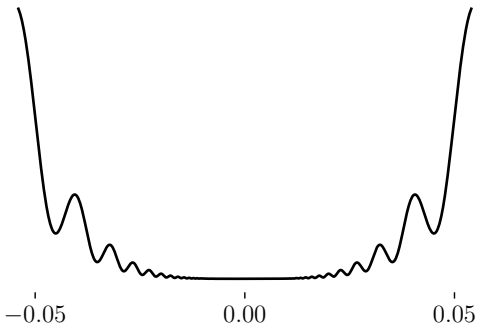

Figure 1: Function $x \mapsto x^4 \cos(1/x) + 2x^4$ around $x = 0$.

2. For all $x \in \mathbb{R}$ we have $f(x) \geq x^4$. From this we deduce that $x^* = 0$ is a strict global (hence local) minimum of f. Yet, it is possible to find points x arbitrarily close to x^* such that the derivative $f'(x) = x^2(8x + \sin(1/x) + 4x\cos(1/x))$ is zero and the second derivative $f''(x) = (12x^2 - 1)\cos(1/x) + 6x(4x + \sin(1/x))$ is positive. This means that these points are local minima and that x^* is *not* isolated.
3. Let x^* be an isolated local minimum of f. There exists a neighborhood $\mathcal{U}$ of x^* such that x^* is the only local minimum of f in $\mathcal{U}$. Since x^* is a local minimum there also exists a neighborhood $\mathcal{V}$ of x^* such that for all $x \in \mathcal{V}$ we have $f(x) \geq f(x^*)$. Now consider the set $\mathcal{N} = \mathcal{U} \cap \mathcal{V}$, which is also a neighborhood of x^*. For all $x \in \mathcal{N}$ we have $f(x) \geq f(x^*)$. Moreover, suppose for contradiction that there exists a point $x \in \mathcal{N} \setminus \{x^*\}$ such that $f(x) = f(x^*)$. Then x would also be a local minimum of f, and it is in $\mathcal{U}$ because it is in $\mathcal{N}$: that is a contradiction. So we conclude that for all $x \in \mathcal{N} \setminus \{x^*\}$ we have $f(x) > f(x^*)$, which means that x^* is a strict local minimum.

■