Numerical Optimization with Python

Dry HW 01

Part I: Modeling optimization problems:

For each of the following descriptions, formulate a mathematical optimization problem (constrained or unconstrained) that models the given definitions (not solving the problem at all).

1. An ice cream manufacturer produces three flavors: vanilla, chocolate and strawberry. The profit per Kg from manufacturing vanilla is 7NIS, for chocolate is 6NIS and for strawberry is 5NIS. There's one machine, which is capable of producing no more than 100Kg of ice cream per day. Each Kg of Vanilla requires 0.5Kg of milk, 0.4Kg of cream and 0.1Kg of vanilla extract. Each Kg of chocolate requires 0.2Kg of milk, 0.7Kg of cream and 0.1Kg of chocolate. Each Kg of strawberry requires 0.4Kg of milk, 0.4Kg of cream and 0.2Kg of strawberries. The manufacturer has the following limits on daily supplies: milk 45Kg, cream 60Kg, vanilla 10Kg, chocolate 10Kg, strawberries 15Kg. Formulate an optimization problem that finds the quantities to manufacture that maximize the manufacturer profit.

   Solution

   Denote the unknowns as follows:

   x_1: Vanilla production in Kg per day in the optimal design

   x_2: Chocolate production in Kg per day in the optimal design

   x_3: Strawberry production in Kg per day in the optimal design

   Therefore, accounting for the given profits per ice cream flavor, the profit per day is:

$$7x_1 + 6x_2 + 5x_3$$

Which is the objective function we attempt to maximize. Capacity constraint: a single constraint, following from the fact that a single machine exists and is capable of producing no more than 100Kg per day, hence:

$$x_1 + x_2 + x_3 \leq 100$$

Supply constraints: there are five involved ingredients (milk, cream, vanilla extract, chocolate and strawberries). The first two are involved in production of all flavors, and the last three constrain only the production of one flavor each:

Milk: $0.5x_1 + 0.2x_2 + 0.4x_3 \leq 45$

Cream: $0.4x_1 + 0.7x_2 + 0.4x_3 \leq 60$

Vanilla extract: $0.1x_1 \leq 10$

Chocolate: $0.1x_2 \leq 10$

Strawberries: $0.2x_3 \leq 15$

Finally, non-negative production constraints (restrict the amounts to be non-negative:

$$x_i \geq 0, i = 1,2,3$$

Summarizing, we have formulated the following maximization problem (a Linear Programming problem):

$$\max[7x_1 + 6x_2 + 5x_3]$$

Subject to:

$$x_1 + x_2 + x_3 \leq 100$$

$$0.5x_1 + 0.2x_2 + 0.4x_3 \leq 45$$

$$0.7x_1 + 0.4x_2 + 0.4x_3 \leq 60$$

$$0.1x_1 \leq 10$$

$$0.1x_2 \leq 10$$

$$0.2x_3 \leq 15$$

$$x_i \geq 0, i = 1,2,3$$

---

2. Given a vector x of n real values $x_1, \dots, x_n$, a vector $p \in \mathbb{R}^n$ is called a ***probability distribution*** on x, if $p_i \geq 0$ and $p_1 + \cdots + p_n = 1$. In other words, the p_i's are non-negative and sum to 1. Given a probability vector p for x, the ***expected value*** (or expectation) of x is the weighted average $\mathrm{E}x = \sum_{i=1}^{n} p_i x_i$. Given a probability distribution p on n values, the ***entropy*** of p is defined to be: $\mathrm{H} = -\sum_{i=1}^{n} p_i \log p_i$.

For the values $x_1 = -10.2, x_2 = 0.4, x_3 = 16.6, x_4 = 10.3$, formulate an optimization problem that finds a probability distribution for x with expected value zero and has maximal entropy.

Solution:

In the problem description $n = 4$ hence our unknown vector is $[p_1, p_2, p_3, p_4]^T$. Maximizing entropy is equivalent to minimizing the negated objective, hence we can write our minimization objective:

$$\sum_{i=1}^{4} p_i \log p_i$$

The expectation constraint on the specific given values: $-10.2p_1 + 0.4p_2 + 16.6p_3 + 10.3p_4 = 0$.

Finally, we constrain p to be a probability vector: $p_i \geq 0, i = 1, \ldots, 4$ and $p_1 + \cdots + p_4 = 1$.

Summarizing, we have formulated the following problem:

$$\min \left[\sum_{i=1}^{4} p_i \log p_i \right]$$

Subject to:

$$-10.2p_1 + 0.4p_2 + 16.6p_3 + 10.3p_4 = 0$$

$$p_1 + p_2 + p_3 + p_4 = 1$$

$$p_i \geq 0, i = 1, \ldots, 4$$

---

3. For some dimension $n > 1$, consider upper ("northern") half of the closed unit ball (boundary as well as interior). Formulate an optimization problem that seeks the point in the above described domain, that is closest to a given point $p \in \mathbb{R}^n$ in the Euclidean norm (L_2).
   a. If the problem you formulated is not smooth, adjust it so it is equivalent (same solution) but smooth.
   b. Formulate the same problem but now seeking a point in the domain that is closest to the given point $p \in \mathbb{R}^n$ in L_1 norm. Is your problem smooth?
   c. Transform the problem in (b) to a smooth problem that seeks the same solution (hint: introduce helper variables that will increase the dimension and add constraints, but provide a smooth problem with the same interpretation, regarding the original variables)

Solution:

Part (a):

Denote the unknown location by $x = [x_1, \dots, x_n]^T \in \mathbb{R}^n$. The objective function measures the distance from the given, constant point p and therefore, in our first formulation, the objective we minimize is $\|x - p\|$ which, explicitly, is the Euclidean distance: $[(x_1 - p_1)^2 + \cdots + (x_n - p_n)]^{\frac{1}{2}}$.

The (Euclidean) unit ball in n dimensions is simply given by the constraint $\|x\| \leq 1$, which, written explicitly is: $[x_1^2 + \cdots + x_n^2]^{\frac{1}{2}} \leq 1$. By "upper" we mean that the last coordinate being non-negative, which is expressed by $x_n \geq 0$.

Now, the above involved functions have square roots, which are non-differentiable at the origin, thus we make the following arguments, to formulate a smooth, but equivalent problem. First, we minimize $\|x - p\|^2$ instead of $\|x - p\|$, and this will yield the same solution, since the function $t \mapsto t^2$ is monotone increasing on $t \geq 0$, this does not change ordering and the same location will be obtained. Similarly, the unit ball can be equivalently written by the constraint $\|x\|^2 \leq 1$, and we arrive at the following problem formulation, in which all the involved functions are smooth:

$$\min[(x_1 - p_1)^2 + \cdots + (x_n - p_n)^2]$$

Subject to:

$$x_1^2 + \cdots + x_n^2 \leq 1$$

$$x_n \geq 0$$

Part (b):

In this part, only the objective changes. Recall that for a given vector $v \in \mathbb{R}^n$, the L_1 norm (or 1-norm) is defined by: $\|v\|_1 := \sum_{i=1}^{n} |v_i|$. Note that we did not ask for the domain to change, the unit ball still refers to the Euclidean norm (2-norm, or L_2). Thus, the revised problem for this section is:

$$\min[|x_1 - p_1| + \cdots + |x_n - p_n|]$$

Subject to:

$$x_1^2 + \cdots + x_n^2 \leq 1$$

$$x_n \geq 0$$

Part (c):

The formulation in part (b) is not smooth, since the absolute value function is not differentiable at the origin. However, we now use a standard method to convert L_1 minimization problems to smooth problems. We introduce n additional variables to the problem: $t_1, \dots, t_n$.

Replace the objective in (b) with the following new objective:

$$\min[t_1 + \cdots + t_n],$$

and add $2n$ new constraints. Two constraints per each t_i that will capture what we require from i'th term in the original objective:

$$t_i \geq x_i - p_i$$

$$t_i \geq p_i - x_i.$$

Note: this means the dimension of the problem is now $2n$, and a solution is a vector of the form:

$$[x_1, \dots, x_n, t_1, \dots, t_n]^T,$$

despite the fact that x_i's do not appear in the objective.

Summarizing, the smooth version that we have formulated and is equivalent to the problem in (b), in the sense that it yields the same optimal $x_1, \dots, x_n$:

$$\min[t_1 + \cdots + t_n]$$

Subject to:

$$x_1^2 + \cdots + x_n^2 \leq 1$$

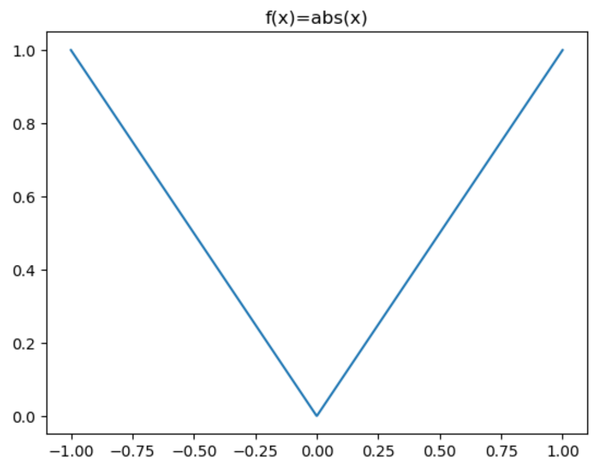
$$x_n \geq 0$$

$$t_i \geq x_i - p_i, i = 1, \dots, n$$

$$t_i \geq p_i - x_i, i = 1, \dots, n$$

To understand the claimed equivalence in the alternative formulation, consider the following simple problem (unconstrained but non-smooth, single variable):

$$\min|x|$$



For which the unconstrained minimizer is $x = 0$.

Now, replace it with the following (constrained, smooth, 2 variables)

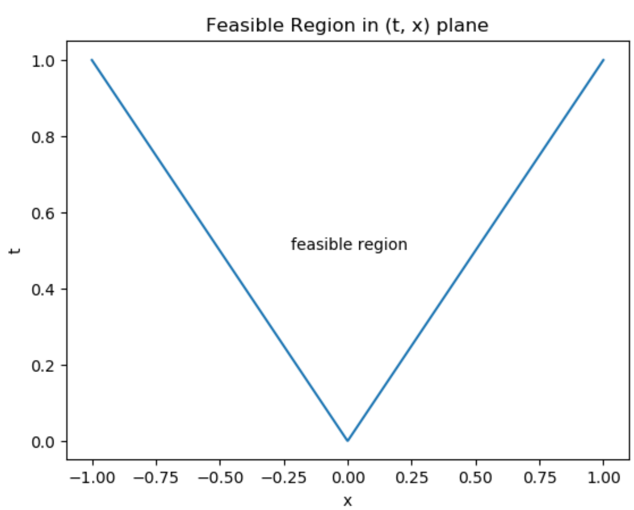
$$\min t$$

Subject to:

$$t \geq x$$

$$t \geq -x$$

Indeed, this is simply a different point of view on the same problem:



it is easy to see that in the $[x, t]$ plane, the feasible region is the wedge above the lines $t = x$ and $t = -x$, and the optimal $[x, t]^T$ location in this feasible region, is $x = 0, t = 0$ because it minimizes t

from all other possible locations. This shows we can replace occurrences of absolute values in minimization problems with a helper variable and two constraints. This argument follows easily to L_1 norm, as it is a separable sum of absolute values, in the sense that term i depends only on x_i, hence each term is minimized separately by the above method, replacing the absolute value by t_i and 2 constraints.

Part II: General preview material – multivariate differentiation:

1. Let $f: \mathbb{R}^n \to \mathbb{R}$ be defined by: $f(x) = x_1 + x_2 + \cdots + x_n$.
    a. What is the partial derivative: $\frac{\partial f(x)}{\partial x_i}$?
    b. What is $\nabla f(x)$? (remember – the gradient is a column vector)
    c. What is the directional derivative $\frac{\partial f}{\partial u}$ w.r.t the unit vector: $u = \left[\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right]^T$? Does it depend on x? Why?
    d. For a constant vector $a \in \mathbb{R}^n$, write the gradient of the linear function: $f(x) = a^T x$

Solution:

Part (a)

$\frac{\partial f}{\partial x_i} = 1$, by direct differentiation.

Part (b)

Since the gradient is the vector of partial derivatives as in (a), we have: $\nabla f(x) = [1, \dots, 1]^T$

Part (c)

We have shown in class that $\frac{\partial f}{\partial u}(x) = u^T \nabla f(x)$ and therefore: $\frac{\partial f}{\partial u}(x) = \left[\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right] [1, \dots, 1]^T = \sqrt{n}$. It does not depend on x since f is linear, and hence its derivatives are constant.

Part (d)

$$f(x) = a^T x = a_1 x_1 + \cdots + a_n x_n \Rightarrow \frac{\partial f}{\partial x_i} = a_i \Rightarrow \nabla f(x) = a$$

---

2. Let $f: \mathbb{R}^2 \to \mathbb{R}$ be defined by: $f(x) = x_1^2 + 2x_2^2 + x_1 x_2$.
    a. Write f as a quadratic function in matrix form, namely find a symmetric matrix Q such that f is given by: $f(x) = x^T Q x$.
    b. Find $\frac{\partial f}{\partial x_1}$ and $\frac{\partial f}{\partial x_2}$, put them together and write $\nabla f(x)$.
    c. Express $\nabla f(x)$ found in (b) using the matrix Q from (a). Hint: it should be a generalization of the derivative of a quadratic function in one variable.
    d. Find $\frac{\partial^2 f}{\partial x_1^2}, \frac{\partial^2 f}{\partial x_1 \partial x_2}$ and $\frac{\partial^2 f}{\partial x_2^2}$. Arrange them together and write the Hessian $\nabla^2 f(x)$.
    e. Express $\nabla^2 f(x)$ found in (d) in terms of the matrix Q from (a). Hint: it should again be a generalization of the second derivative of a quadratic function in one variable.
    f. The general case: for a constant matrix $Q \in \mathbb{R}^{n \times n}$, write the gradient $\nabla f(x)$ and the Hessian $\nabla^2 f(x)$ of the quadratic function $f(x) = x^T Q x$.

Solution:

Part (a)

From looking at the quadratic form we can check that $Q = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}$. However, if you are not used to quadratic forms, we now find the coefficients systematically:

For the 2x2 case, a symmetric matrix operates as a quadratic form as follows:

$$Q = \begin{bmatrix} q_1 & q_2 \\ q_2 & q_3 \end{bmatrix} \Rightarrow x^T Q x = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} q_1 & q_2 \\ q_2 & q_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} q_1 x_1 + q_2 x_2 \\ q_2 x_1 + q_3 x_2 \end{bmatrix}$$
$$= q_1 x_1^2 + q_2 x_1 x_2 + q_2 x_1 x_2 + q_3 x_2^2 = q_1 x_1^2 + 2 q_2 x_1 x_2 + q_3 x_2^2$$

Now comparing the coefficients with $f(x) = x_1^2 + 2x_2^2 + x_1 x_2$, we see that indeed $q_1 = 1, q_2 = 0.5$ and $q_3 = 2$.

NOTE:

it is VERY useful to know the same for the $n \times n$ case: given a matrix $Q \in \mathbb{R}^{n \times n}$, the scalar expression for the quadratic form $x^T Q x$ is $\sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij} x_i x_j$. To verify that, simply do the matrix vector product.

Now, when Q is symmetric, we have that $q_{ij} = q_{ji}$, and it is convenient to write the above sum as follows:

$$x^T Q x = \sum_{i=1}^{n} \sum_{j=1}^{n} q_{ij} x_i x_j = \sum_{i=1}^{n} q_{ii} x_i^2 + 2 \sum_{i>j} q_{ij} x_i x_j$$

The first term ($i = j$) corresponds to all the diagonal elements, that yield the square terms, the second term corresponds to the off-diagonal elements, and accounts for them twice, traversing only the lower triangle ($i > j$ indices), yielding the mixed products ($x_i x_j$ where $i \neq j$). Make sure you understand the above and how the exercise we did is the 2x2 special case of that.

Part (b)

In this part we find the gradient by differentiating element-wise, namely:

$f(x) = x_1^2 + 2x_2^2 + x_1 x_2 \Rightarrow \frac{\partial f}{\partial x_1}(x) = 2x_1 + x_2$ and $\frac{\partial f}{\partial x_2}(x) = 4x_2 + x_1$, Hence put together:

$$\nabla f(x) = \begin{bmatrix} 2x_1 + x_2 \\ 4x_2 + x_1 \end{bmatrix}$$

Part (c)

For $x^T Q x$ we expect that $\nabla f(x)$ will be $2Qx$, and we now check that:

$$2Qx = 2 \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2 \begin{bmatrix} x_1 + 0.5x_2 \\ 0.5x_1 + 2x_2 \end{bmatrix} = \begin{bmatrix} 2x_1 + x_2 \\ x_1 + 4x_2 \end{bmatrix}$$

As found directly in part (a).

Note:

The fact that we could guess this expression $2Qx$, can either be done by differentiating each element in the general case and showing that indeed $\nabla[x^T Q x] = 2Qx$, or just guessing from the knowledge in the 1-dimensional case.

Part (d)

$$\frac{\partial^2 f}{\partial x_1^2} = \frac{\partial}{\partial x_1}[2x_1 + x_2] = 2$$

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial}{\partial x_2}[2x_1 + x_2] = 1$$

$$\frac{\partial^2 f}{\partial x_2^2} = \frac{\partial}{\partial x_2}[4x_2 + x_1] = 4$$

Therefore the Hessian matrix is:

$$\nabla^2 f(x) = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

Part (e):

The Hessian in terms of Q is $2Q$, which can be seen directly from Q or obtained by differentiating the general case of a quadratic f.

Part (f):

In the general case of a quadratic form $x \mapsto x^T Qx$, (symmetric Q) the gradient is $2Qx$ and the Hessian is $2Q$.

---

3. Let $f: \mathbb{R}^n \to \mathbb{R}$ and $u \in \mathbb{R}^n$, $\|u\| = 1$. Recall from class: the directional derivative $\frac{\partial f(x)}{\partial u}$ is the slope of the graph of f when marching in the direction u at the point x, we have shown this is given by $u^T \nabla f(x)$. In this exercise we develop the matrix form expression for the directional <u>second</u> derivative.
    a. Write a parametrized expression for a line $l(t)$ in the direction u that through x.

<u>Solution</u>:

The line is parameterized by $l(t) = x + tu, t \in \mathbb{R}$. The location x and direction u are constant, the single real parameter t runs over all $\mathbb{R}$, and the point x is realized by $t = 0$, namely $x = l(0)$.

---

b. Write a function of a single variable $g(t)$, defined to be the restriction of f to the line $l(t)$

Solution:

The required $g: \mathbb{R} \to \mathbb{R}$ is the restriction of f to the line, namely $g(t) = f\big(l(t)\big) = f(x + tu)$.

---

c. Write an expression for the first derivative $g'(t)$. Hint: g is a composition $f \circ l: \mathbb{R} \to \mathbb{R}$, use the chain rule (repeat as done in class).

Solution:

Differentiating using the chain rule gives:

$$g'(t) = \nabla f\big(l(t)\big)^T l'(t) = \nabla f\big(l(t)\big)^T u = u^T \nabla f\big(l(t)\big)$$

Note that we have used the fact that the inner product is symmetric and that $l'(t)$ is the constant direction u.

---

d. Using $g'(t)$ from (c), write an expression for $g''(t)$ (Hint: use the chain rule again, and recall we noted in class what the differential matrix of the vector function: $x \mapsto \nabla f$ is.)

Solution:

Differentiating a second time using the chain rule:

$$\frac{d}{dt}\left[u^T \nabla f\big(l(t)\big)\right] = u^T \nabla^2 f\big(l(t)\big) l'(t) = u^T \nabla^2 f\big(l(t)\big) u$$

We have again used the fact that l' is the constant direction u, and that the derivative (Jacobi matrix) of the vector valued function $x \mapsto \nabla f(x)$ is the Hessian matrix $\nabla^2 f(x)$.

---

e. Finally, use g'' from (d) to obtain $\frac{\partial^2 f(x)}{\partial u^2}$. Hint: which t value gives x along the line l?

Solution:

To locate the value of the directional second derivative to x, we recall that $x = l(0)$ thus we are interested in $g''(0)$ which is $u^T \nabla^2 f\big(l(0)\big)u = u^T \nabla^2 f(x) u$.

This concludes showing that the directional second derivative is expressed by the quadratic form of the Hessian, operating on the direction.

---

f. For the quadratic function in question 2, what is $\frac{\partial^2 f(x)}{\partial u^2}$ in the unit direction $u = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^T$? Does it depend on x? Why?

Solution:

The Hessian matrix of the quadratic form is $2Q = 2\begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$ and hence the second directional derivative is the quadratic form:

$$\left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right] \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = 2\left(\frac{1}{\sqrt{2}}\right)^2 + 4\left(\frac{1}{\sqrt{2}}\right)^2 + 2\left[1\left(\frac{1}{\sqrt{2}}\right)\left(\frac{1}{\sqrt{2}}\right)\right] = 1 + 2 + 1 = 4$$

g. (5 point bonus) we know that $\frac{\partial f(x)}{\partial u}$ has a geometric interpretation, as the slope at x in the direction u. What is the geometric interpretation of $\frac{\partial^2 f(x)}{\partial u^2}$? (what does it measure? No need to be precise regarding units/scale)

Solution:

The geometric quantity that the directional second derivative measure is (up to unit scaling) the ***curvature*** of the path traveling in the u direction on the graph of f.

Curvature is defined differently according to context (of a curve, of a surface, of higher dimensional manifolds, and also depends on the representation). You are encouraged to read further!

For our course, we remember that positive curvature in every direction is equivalent to convexity, and is what characterizes minimum points of unconstrained optimization problems, or of constrained problems in the situation that no constraints are active (away from the boundary of the feasible set).