Inequalities Via Lagrange Multipliers

Many (classical) inequalities can be proven by setting up and solving certain optimization problems. In turn, such optimization problems can be handled using the method of *Lagrange Multipliers* (see the Theorem 2 below).

A. Compactness (in $\mathbb{R}^N$)

When solving optimization problems, the following notions are extremely important.

Definitions. Suppose N is a positive integer, and A is a non-empty subset in $\mathbb{R}^N$.

A. We say that A is *bounded*, if there exists some constant m, such that: $\|a\| \leq m$, for all a in A.

B. We say that A is *closed*, if whenever $(a_n)_{n=1}^{\infty}$ is a sequence of points, which converges to some x, it follows that x itself belongs to A.

C. We say that A is *compact*, if it is both closed and bounded.

Comments. The norm $\| \, . \, \|$ mentioned in Definition A can be any norm[1]. Two frequently used choices are

$$\begin{aligned} \|x\|_2 &= \sqrt{x_1^2 + \cdots + x_N^2}, \\ \|x\|_\infty &= \max\{|x_1|, \ldots, |x_N|\}, \\ x &= (x_1, \ldots, x_N) \in \mathbb{R}^N \end{aligned}$$

For a sequence $(a_n)_{n=1}^{\infty}$ of vectors in $\mathbb{R}^N$, the condition that $(a_n)_{n=1}^{\infty}$ converges to $x \in \mathbb{R}^N$ means that $\lim_{n\to\infty} \|a_n - x\| = 0$, were $\| \, . \, \|$ is any norm on $\mathbb{R}^N$.

The above definition of compactness is ad-hoc, and specific to $\mathbb{R}^N$. Another important characterization, specific to *metric spaces*, is the following: *every sequence in A has a sub-sequence which converges to a point in A.* (The fact that this characterization is equivalent to the condition stated in Definition C follows from the well-known property of bounded sequences: *every bounded sequence in $\mathbb{R}$ has a convergent sub-sequence.* Using this property, one can easily show that the same statement is true with $\mathbb{R}$ replaced with $\mathbb{R}^N$.)

The most general characterization of compactness (for so-called *topological spaces*) is beyond the scope of these notes.

Compactness is an essential property that is sought when attacking optimization problems, as suggested by the following results.

Theorem 1. *If $A \subset \mathbb{R}^N$ is compact, and $f : A \to \mathbb{R}^M$ is continuous, then $f(A)$ is compact.*

[1] On $\mathbb{R}^N$ any two norms $\| \, . \, \|$ and $\| \, . \, \|'$ are *equivalent*, in the sense that there exists positive constants C and D, such that $C\|x\| \leq \|x\|' \leq D\|x\|$, for every x.

Corollary. *If $A \subset \mathbb{R}^N$ is compact, and $f : A \to \mathbb{R}$ is continuous, then there exist points $a_0, a_1 \in A$, such that: $f(a_0) \le f(a) \le f(a_1)$, $\forall\, a \in A$.* (That is, f attains its maximum and minimum values on A.)

B. Lagrange Multipliers

Theorem 2 (Lagrange). *Assume g is some continuously differentiable real-valued function, defined on some (open) domain D in $\mathbb{R}^N$, and K is some real number, such that the level set $A = \{x \in D \,|\, g(x) = K\}$ has the following property:*

($*$) *for every $a \in A$ the gradient $(\nabla g)(a)$ is non-zero, i.e.*

$$\Big(\frac{\partial g}{\partial x_1}(a), \dots, \frac{\partial g}{\partial x_N}(a)\Big) \neq (0, \dots, 0).$$

Suppose f is another real-valued continuously differentiable function, defined on D. If $x \in A$ is local maximum point for f on A, i.e.

($**$) *there exists some neighborhood V of x in $\mathbb{R}^N$, such that*

$$f(x) \ge f(a), \ \ \forall\, a \in A \cap V,$$

then there exists some real number λ, such that $(\nabla f)(x) = \lambda (\nabla g)(x)$, that is,

$$\frac{\partial f}{\partial x_j}(x) = \lambda \frac{\partial g}{\partial x_j}(x), \ \ \forall\, j = 1, \dots, N.$$

Comments. A level set A satisfying condition ($*$) is called a *hyper-surface in* $\mathbb{R}^N$. In practice, "the candidates" for x satisfying ($**$) are found by solving (or just narrowing down the solution set of) the system of equations

$$\begin{cases} \dfrac{\partial f}{\partial x_1}(x) &= \lambda \dfrac{\partial g}{\partial x_1}(x) \\ \quad \vdots & \\ \dfrac{\partial f}{\partial x_N}(x) &= \lambda \dfrac{\partial g}{\partial x_N}(x) \\ \\ \quad g(x) &= K \end{cases} \tag{1}$$

The system (1) has $N + 1$ equations and $N + 1$ unknowns: λ and the coordinates of x.

In our examples, the hyper-surface A will be *compact*, so the function f will have an *absolute* maximum value on A.

C. Applications

Arithmetic Vs. Geometric Means Inequality. *Given $a_1, a_2, \dots, a_N \ge 0$, one has the inequality:*

$$\sqrt[N]{a_1 a_2 \cdots a_N} \le \frac{a_1 + a_2 + \cdots + a_N}{N} \tag{2}$$

Moreover, one has equality in (2), if and only if $a_1 = a_2 = \cdots = a_N$.

Proof. It is clear that the statement does not change if all a's are multiplied by a positive constant. Therefore[2] the above statement can be rephrased as follows.

($*$) *Whenever $a_1, a_2, \dots, a_N \geq 0$ are such that $a_1 + a_2 + \cdots + a_N = N$, it follows that*

$$a_1 a_2 \cdots a_N \leq 1 \tag{3}$$

Moreover, one has equality in (3), if and only if $a_1 = a_2 = \cdots = a_N = 1$.

We now restate ($*$) as an optimization problem, using the substitution $a_j = x_j^2$, so we consider the sphere of radius $\sqrt{N}$ in $\mathbb{R}^N$, that is, the set

$$A = \{x = (x_1, x_2, \dots, x_N) \in \mathbb{R}^N \mid x_1^2 + \cdots + x_N^2 = N\},$$

which is a level set for the function $g(x_1, \dots, x_N) = x_1^2 + \cdots + x_N^2$. Of course, the gradient of g is: $(\nabla g)(x_1, \dots, x_N) = (2x_1, \dots, 2x_N)$, which never vanishes on A, so A is indeed a hyper-surface. Note also that A is *compact.* The objective function is $f : A \to \mathbb{R}$ defined by

$$f(x) = x_1 x_2 \cdots x_N, \quad x = (x_1, x_2, \dots, x_N) \in A,$$

for which we must find the absolute maximum on A. Our "guess" is the follwoing.

(**) *The maximum value* $\max_{x \in A} f(x)$ *is equal to* 1*, and all points $x = (x_1, x_2, \dots, x_N) \in A$, with $f(x) = 1$, satisfy the condition*

$$|x_1| = |x_2| = \cdots = |x_N| = 1. \tag{4}$$

The system of equations (1) reads (after replacing λ with $\lambda/2$)

$$\begin{cases} x_2 x_3 \cdots x_N = \lambda x_1 \\ x_1 x_3 \cdots x_N = \lambda x_2 \\ \vdots \\ x_1 \cdots \widehat{x_j} \cdots x_N = \lambda x_j \\ \vdots \\ x_1 x_2 \cdots x_{N-1} = \lambda x_N \\ x_1^2 + \cdots + x_N^2 = N \end{cases} \tag{5}$$

(The notation $\widehat{x_j}$ indicates that the factor x_j is missing.)

Multiplying each one of the top N equations by its missing factor from the left-hand side, yields:

$$\begin{cases} \lambda x_1^2 = \lambda x_2^2 = \cdots = \lambda x_N^2 = x_1 x_2 \cdots x_N \\ x_1^2 + \cdots + x_N^2 = N \end{cases} \tag{6}$$

Now we see that every solution (λ, x) of (3) satisfies one of the following conditions:

(A) $\lambda = 0$, in which case $f(x) = 0$;

[2] The trivial case when $a_1 = a_2 = \cdots = a_N = 0$ is omitted.

(B) $\lambda \neq 0$, in which case we have (using the last equation) $x_1^2 = x_2^2 = \cdots = x_N^2 = 1$, which forces $f(x) = \pm 1$.

Clearly (A) does not yield a maximum. Case (B), however proves statement (∗∗). □

Hölder's Inequality. *Let $p, q > 1$ be two real numbers satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Given $a_1, a_2, \dots, a_N, b_1, b_2, \dots, b_N \geq 0$, one has the inequality:*

$$a_1 b_1 + a_2 b_2 + \dots a_N b_N \leq [a_1^p + a_2^p + \cdots + a_N^p]^{1/p} \cdot [b_1^q + a_2^q + \cdots + b_N^q]^{1/q} \tag{7}$$

Moreover, one has equality in (7), if and only the N-tuples $(a_1^p, a_2^p, \dots, a_N^p)$ and $(b_1^q, b_2^q, \dots, b_N^q)$ are proportional.

Proof. It is clear that the statement does not change if all a's are multiplied by a positive constant and all the b's are multiplied by another positive constant. Therefore[3] the above statement can be rephrased as follows.

(∗) *Whenever $a_1, a_2, \dots, a_N, b_1, b_2, \dots, b_N \geq 0$ are such that*

$$a_1^p + a_2^p + \cdots + a_N^p = b_1^q + b_2^q + \cdots + b_n^q = 1,$$

it follows that

$$a_1 b_1 + a_2 b_2 + \dots a_N b_n \leq 1 \tag{8}$$

Moreover, one has equality in (8), if and only if $a_j^p = b_j^q$, for all $j = 1, 2, \dots, N$.

We now restate (∗) as an optimization problem, as follows. We first fix $a_1, a_2, \dots, a_N \geq 0$, such that $a_1^p + a_2^p + \cdots + a_N^p = 1$. Second (with the substitutions $b_j = x_j^2$ in mind) we consider the function $g : \mathbb{R}^N \to \mathbb{R}$ defined by

$$g(x) = (x_1^2)^q + (x_2^2)^q + \cdots + (x_N^2)^q, \;\; x = (x_1, x_2, \dots, x_N) \in \mathbb{R}^N.$$

Note that the definition of g involves the "funny" function[4] $\phi(t) = (t^2)^q$, whose derivative is

$$\phi'(t) = \begin{cases} \dfrac{(2q)(t^2)^q}{t} & \text{if } t \neq 0 \\ 0 & \text{if } t = 0 \end{cases} \tag{9}$$

In particular, ϕ is continuously differentiable, and so is g, whose gradient is

$$(\nabla g)(x) = \big(\phi'(x_1), \phi'(x_2), \dots, \phi'(x_N)\big), \;\; x = (x_1, x_2, \dots, x_N) \in \mathbb{R}^N. \tag{10}$$

Consider now the set $A = \{x \in \mathbb{R}^N \mid g(x) = 1\}$. Since for every $x = (x_1, x_2, \dots, x_N) \in A$, at least one coordinate is non-zero, using (10) and (9) we see that $(\nabla g)(x)$ is non-zero, for each $x \in A$, so A is a hyper-surface. Obviously A is also compact. Our objective function is $f : A \to \mathbb{R}$ defined by

$$f(x) = a_1 x_1^2 + a_2 x_2^2 + \dots a_n x_N^2, \;\; x = (x_1, x_2, \dots, x_N) \in A,$$

for which we must find the absolute maximum on A. Our "guess" is the following.

[3] The trivial case when $a_1 = a_2 = \cdots = a_N = b_1 = b_2 = \cdots = b_N = 0$ is omitted.

[4] One might be tempted to write $\phi_(t) = t^{2q}$, but this would be incorrect, since we allow t to be *negative*.

$(**)$ *The maximum value* $\max_{x \in A} f(x)$ *is equal to* 1*, and all points* $x = (x_1, x_2, \ldots, x_N) \in A$*, with* $f(x) = 1$*, satisfy the condition*

$$(x_j^2)^q = a_j^p, \ \ \forall\, j = 1, 2, \ldots, N. \tag{11}$$

If we write down the system of equations (1) for this particular setting, it reads:

$$\begin{cases} 2a_1x_1 = \lambda\phi'(x_1) \\ 2a_2x_2 = \lambda\phi'(x_2) \\ \vdots \\ 2a_Nx_N = \lambda\phi'(x_N) \\ (x_1^2)^q + \cdots + (x_N^2)^q = 1 \end{cases} \tag{12}$$

Fix a solution (λ, x) of (12). If we multiply the j^{th} equation by x_j, then using (9) we obtain the system

$$\begin{cases} a_1x_1^2 = (\lambda q)(x_1^2)^q \\ a_2x_2^2 = (\lambda q)(x_2^2)^q \\ \vdots \\ a_Nx_N^2 = (\lambda q)(x_N^2)^q \\ (x_1^2)^q + \cdots + (x_N^2)^q = 1 \end{cases} \tag{13}$$

so upon adding the top N equations, and using the bottom we one, we obtain

$$f(x) = a_1x_1^2 + a_2x_2^2 + \cdots + a_Nx_N^2 = \lambda q[(x_1^2)^q + \cdots + (x_N^2)^q] = \lambda q. \tag{14}$$

Going back to the system (13), let $j_1, \ldots, j_m$ be the indices for which $x_j \neq 0$, so that for each $k = 1, \ldots, m$, the $j_k{}^{\text{th}}$ equation yields

$$a_{j_k} = (\lambda q)(x_{j_k}^2)^{q-1}.$$

Taking then the p^{th} power (and using the identity $\dfrac{1}{p} + \dfrac{1}{q} = 1$, which gives $p(q-1) = q$) we now obtain

$$a_{j_k}^p = (\lambda q)^p (x_{j_k}^2)^q, \ \ \forall\, k = 1, \ldots, m. \tag{15}$$

If we sum up the above identities (and use the fact that $x_j = 0$, for all $j \notin \{j_1, \ldots, j_m\}$), then by the last equation in (12) we obtain

$$a_{j_1}^p + \cdots + a_{j_m}^p = (\lambda q)^p[(x_{j_1}^2)^q + \cdots + (x_{j_m}^2)^q] = (\lambda q)^p[(x_1^2)^q + (x_2^2)^q + \cdots + (x_N^2)^q] = (\lambda q)^p,$$

so using the assumption on the a's we will get:

$$(\lambda q)^p = a_{j_1}^p + \cdots + a_{j_m}^p \leq a_1^p + a_2^p + \cdots + a_N^p = 1. \tag{16}$$

This of course yields $\lambda q \leq 1$, so going back to (14) we obtain $f(x) \leq 1$. Up to this point we have shown that $\max_{x \in A} f(x) \leq 1$, which, by the way, proves (8).

To finish the proof of $(**)$ we need to:

(A) show that $\max_{x\in A} f(x) = 1$, and

(B) show that any $x \in A$ with $f(x) = 1$ satisfies (11).

We begin with assertion (B). If $f(x) = 1$, then there is some λ such that (λ, x) is a solution of (12). As we have seen in (14), this forces $\lambda q = 1$, so going back to (16) this also forces the inequality there to be an equality. This would mean, of course, that $a_j = 0 (= x_j)$, $\forall\, j \notin \{j_1, \dots, j_m\}$, so by (15) the desired conclusion (11) follows. To prove the above assertion (A), all we have to do is to produce one $x \in A$, with $f(x) = 1$. Such an x, however, can be easily defined by $x_j = a_j^{\frac{p}{2q}}$. □

Hölder's Inequality for Complex Numbers. *Let $p, q > 1$ be two real numbers satisfying $\dfrac{1}{p} + \dfrac{1}{q} = 1$. Given $x_1, x_2, \dots, x_N, y_1, y_2, \dots, y_N \in \mathbb{C}$, one has the inequality:*

$$|x_1y_1 + x_2y_2 + \dots x_Ny_N \le \left[|x_1|^p + |x_2|^p + \dots + |x_N|^p\right]^{1/p} \cdot \left[|y_1|^q + |y_2|^q + \dots + |y_N|^q\right]^{1/q} \quad (17)$$

Moreover, equality holds in (17)*, if and only the N-tuples $(x_1|x_1|^{p-1}, x_2|x_2|^{p-1}, \dots, x_N|x_N|^{p-1})$ and $(\overline{y}_1|y_1|^{q-1}, \overline{y}_2|y_2|^{q-1}, \dots, \overline{y}_N|y_N|^{q-1})$ are proportional.*

Proof. Without loss of generality (multiply, if necessary all the x's by a positive constant, and multiply the all the y's by another positive constant), we can assume that

$$|x_1|^p + |x_2|^p + \dots + |x_N|^p = |y_1|^q + |y_2|^q + \dots + |y_N|^q. \quad (18)$$

The inequality (17) follows immediately from (7), applied with $a_j = |x_j|$ and $b_j = |y_j|$:

$$|x_1y_1 + x_2y_2 + \dots + x_Ny_N| \le |x_1| \cdot |y_1| + |x_2| \cdot |y_2| + \dots + |x_N| \cdot |y_N| \le \quad (19)$$

$$\le \left[|x_1|^p + |x_2|^p + \dots + |x_N|^p\right]^{1/p} \cdot \left[|y_1|^q + |y_2|^q + \dots + |y_N|^q\right]^{1/q} \quad (20)$$

Assume now we have equality in (17), which forces equality in both inequalities from (19) and (20). First of all, by the preceding result, the equality in (20) forces the sequences $(a_1^p, a_2^p, \dots, a_N^p)$ and $(b_1^q, b_2^q, \dots, b_N^q)$, so by (18) we have in fact the equalities $|y_j|^q = |x_j|^p$, $\forall\, j = 1, \dots, N$, so if we denote this common value by c_j, we have

$$|x_j| = c_j^{1/p} \text{ and } |y_j| = c_j^{1/q}, \ \ \forall\, j = 1, \dots, N. \quad (21)$$

. Secondly, let us observe that the fact that we have equality in the first inequality in (19) is an instance of equality in the "polygon" inequality

$$|z_1 + z_2 + \dots + z_N| \le |z_1| + |z_2| + \dots + |z_N|, \quad (22)$$

and we know that equality holds in (22) if and only if all z's are on a *ray*, i.e. a set of the form $R_w = \{\rho w \,|\, \rho \ge 0\}$, for some $w \in \mathbb{C}$, with $|w| = 1$, which means that

$$z_j = w|z_j|, \ \ \forall\, j = 1, \dots, N. \quad (23)$$

If we apply this to $z_j = x_j y_j$, using (21) it follows that we have the equalities $x_j y_j = wc_j$, so upon after multiplying by $\overline{y}_j$ we now have $x_j |y_j|^2 = w\overline{y}_j c_j$, i.e.

$$x_j c_j^{2/q} = w\overline{y}_j c_j, \ \ \forall\, j = 1, \dots, N. \tag{24}$$

Dividing by $c_j^{1/q}$ (in the case when $c_j \neq 0$) yields $w\overline{y}_j c_j^{1-1/q} = x_j c_j^{1/q} = x_j c_j^{1-1/p}$, which by (21) reads:

$$w\overline{y}_j |y_j|^{q-1} = x_j |x_j|^{p-1}, \ \ \forall\, j = 1, \dots, N. \tag{25}$$

Strictly speaking, (25) was obtained only when $c_j \neq 0$, but is is trivially satisfied if $c_j = 0$, in which case we have $x_j = y_j = 0$. $\square$

Comment. The special case of Hölder's Inequality that corresponds to $p = q = 2$ is known as the *Cauchy-Bunyakovsky-Schwartz Inequality.* In this special case the inequality reads

$$|x_1 y_1 + x_2 y_2 + \dots x_N y_N| \leq \left[|x_1|^2 + |x_2|^2 + \dots + |x_N|^2 \right]^{1/2} \cdot \left[|y_1|^2 + |y_2|^2 + \dots + |y_N|^2 \right]^{1/2}, \tag{26}$$

and equality holds in (26), if and only if *the N-tuples* $(x_1, x_2, \dots, x_N)$ *and* $(\overline{y}_1, \overline{y}_2, \dots, \overline{y}_N)$ *are proportional.*

Indeed, in the case $p = q = 2$, if we argue as above (with same notations), under the assumption (18), the equality in (26) will yield the equality (24), which in this case reads $x_j c_j = w\overline{y}_j c_j$, thus forcing directly: $x_j = w\overline{y}_j$, $\forall\, j = 1, \dots, N$.