

The Method of Lagrange Multipliers

S. Sawyer — October 25, 2002

1. Lagrange's Theorem. Suppose that we want to maximize (or minimize) a function of n variables

$$f(x) = f(x_1, x_2, \dots, x_n) \quad \text{for} \quad x = (x_1, x_2, \dots, x_n) \quad (1.1a)$$

subject to p constraints

$$g_1(x) = c_1, \quad g_2(x) = c_2, \quad \dots, \quad \text{and} \quad g_p(x) = c_p \quad (1.1b)$$

As an example for $p = 1$, find

$$\min_{x_1, \dots, x_n} \left\{ \sum_{i=1}^n x_i^2 : \sum_{i=1}^n x_i = 1 \right\} \quad (1.2a)$$

or

$$\min_{x_1, \dots, x_5} \sum_{i=1}^5 x_i^2 \quad \text{subject to} \quad \begin{cases} x_1 + 2x_2 + x_3 = 1 & \text{and} \\ x_3 - 2x_4 + x_5 = 6 \end{cases} \quad (1.2b)$$

A first guess for (1.1) (with $f(x) = \sum_{i=1}^n x_i^2$ in (1.2)) might be to look for solutions of the n equations

$$\frac{\partial}{\partial x_i} f(x) = 0, \quad 1 \leq i \leq n \quad (1.3)$$

However, this leads to $x_i = 0$ in (1.2), which does not satisfy the constraint.

Lagrange's solution is to introduce p new parameters (called *Lagrange Multipliers*) and then solve a more complicated problem:

Theorem (Lagrange) *Assuming appropriate smoothness conditions, minimum or maximum of $f(x)$ subject to the constraints (1.1b) that is not on the boundary of the region where $f(x)$ and $g_j(x)$ are defined can be found by introducing p new parameters $\lambda_1, \lambda_2, \dots, \lambda_p$ and solving the system*

$$\frac{\partial}{\partial x_i} \left(f(x) + \sum_{j=1}^p \lambda_j g_j(x) \right) = 0, \quad 1 \leq i \leq n \quad (1.4a)$$

$$g_j(x) = c_j, \quad 1 \leq j \leq p \quad (1.4b)$$

This amounts to solving $n+p$ equations for the $n+p$ real variables in x and λ . In contrast, (1.3) has n equations for the n unknowns in x . Fortunately, the system (1.4) is often easy to solve, and is usually much easier than using the constraints to substitute for some of the x_i .

2. Examples. (1) There are $p = 1$ constraints in (1.2a), so that (1.4a) becomes

$$\frac{\partial}{\partial x_i} \left(\sum_{k=1}^n x_k^2 + \lambda \sum_{k=1}^n x_k \right) = 2x_i + \lambda = 0, \quad 1 \leq i \leq n$$

Thus $x_i = -\lambda/2$ for $1 \leq i \leq n$. The constraint $\sum_{i=1}^n x_i = 1$ in (1.4b) implies $\sum_{i=1}^n x_i = -n\lambda/2 = 1$ or $\lambda = -2/n$, from which $x_i = 1/n$ for $1 \leq i \leq n$.

For $x_i = 1/n$, $f(x) = n/n^2 = 1/n$. One can check that this is a minimum as opposed to a maximum or saddle point by noting that $f(x) = 1$ if $x_1 = 1$, $x_i = 0$ for $2 \leq i \leq n$.

(2) *A System with Two Constraints:* There are $p = 2$ constraints in (1.2b), which is to find

$$\min \sum_{i=1}^5 x_i^2 \quad \text{subject to} \quad \begin{cases} x_1 + 2x_2 + x_3 = 1 & \text{and} \\ x_3 - 2x_4 + x_5 = 6 \end{cases} \quad (2.1)$$

The method of Lagrange multipliers says to look for solutions of

$$\frac{\partial}{\partial x_i} \left(\sum_{k=1}^5 x_k^2 + \lambda(x_1 + 2x_2 + x_3) + \mu(x_3 - 2x_4 + x_5) \right) = 0 \quad (2.2)$$

where we write λ, μ for the two Lagrange multipliers λ_1, λ_2 .

The equations (2.2) imply $2x_1 + \lambda = 0$, $2x_2 + 2\lambda = 0$, $2x_3 + \lambda + \mu = 0$, $2x_4 - 2\mu = 0$, and $2x_5 + \mu = 0$. Combining the first three equations with the first constraint in (2.1) implies $2 + 6\lambda + \mu = 0$. Combining the last three equations in (2.2) with the second constraint in (2.1) implies $12 + \lambda + 6\mu = 0$. Thus

$$6\lambda + \mu = -2$$

$$\lambda + 6\mu = -12$$

Adding these two equations implies $7(\lambda + \mu) = -14$ or $\lambda + \mu = -2$. Subtracting the equations implies $5(\lambda - \mu) = 10$ or $\lambda - \mu = 2$. Thus $(\lambda + \mu) + (\lambda - \mu) = 2\lambda = 0$ and $\lambda = 0, \mu = -2$. This implies $x_1 = x_2 = 0$, $x_3 = x_5 = 1$, and $x_4 = -2$. The minimum value in (2.1) is 6.

(3) *A BLUE problem:* Let X_1, \dots, X_n be independent random variables with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma_i^2$. Find the coefficients a_i that minimize

$$\text{Var} \left(\sum_{i=1}^n a_i X_i \right) \quad \text{subject to} \quad E \left(\sum_{i=1}^n a_i X_i \right) = \mu \quad (2.3)$$

This asks us to find the Best Linear Unbiased Estimator $\sum_{i=1}^n a_i X_i$ (abbreviated BLUE) for μ for given values of σ_i^2 .

Since $\text{Var}(aX) = a^2 \text{Var}(X)$ and $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ for independent random variables X and Y , we have $\text{Var}(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i^2 \text{Var}(X_i) = \sum_{i=1}^n a_i^2 \sigma_i^2$. Thus (2.3) is equivalent to finding

$$\min \sum_{i=1}^n a_i^2 \sigma_i^2 \quad \text{subject to} \quad \sum_{i=1}^n a_i = 1$$

Using one Lagrange multiplier λ for the constraint leads to the equations $2a_i \sigma_i^2 + \lambda = 0$ or $a_i = -\lambda/(2\sigma_i^2)$. The constraint $\sum_{i=1}^n a_i = 1$ then implies that the BLUE for μ is

$$\sum_{i=1}^n a_i X_i \quad \text{where} \quad a_i = c/\sigma_i^2 \quad \text{for} \quad c = 1 / \sum_{k=1}^n (1/\sigma_k^2) \quad (2.4)$$

If $\sigma_i^2 = \sigma^2$ for all i , then $a_i = 1/n$ and $\sum_{i=1}^n a_i X_i = (1/n) \sum_{i=1}^n X_i = \bar{X}$ is the BLUE for μ .

Conversely, if $\text{Var}(X_i) = \sigma_i^2$ is variable, then the BLUE $\sum_{i=1}^n a_i X_i$ for μ puts relatively less weight on the noisier (higher-variance) observations (that is, the weight a_i is smaller), but still uses the information in the noisier observations. Formulas like (2.4) are often used in survey sampling.

3. A Short Proof of Lagrange’s Theorem. The extremal condition (1.3) (without any constraints) can be written in vector form as

$$\nabla f(x) = \left(\frac{\partial}{\partial x_1} f(x), \frac{\partial}{\partial x_2} f(x), \dots, \frac{\partial}{\partial x_n} f(x) \right) = 0 \quad (3.1)$$

Now by Taylor’s Theorem

$$f(x + hy) = f(x) + hy \cdot \nabla f(x) + O(h^2) \quad (3.2)$$

where h is a scalar, $O(h^2)$ denotes terms that are bounded by h^2 , and $x \cdot y$ is the dot product. Thus (3.1) gives the vector direction in which $f(x)$ changes the most per unit change in x , where unit change is measured in terms of the length of the vector x .

In particular, if $y = \nabla f(x) \neq 0$, then

$$f(x - hy) < f(x) < f(x + hy)$$

for sufficiently small values of h , and the only way that x can be a local minimum or maximum would be if x were on the boundary of the set of points where $f(x)$ is defined. This implies that $\nabla f(x) = 0$ at non-boundary minimum and maximum values of $f(x)$.

Now consider the problem of finding

$$\max f(x) \quad \text{subject to} \quad g(x) = c \tag{3.3}$$

for one constraint. If $x = x_1(t)$ is a path in the surface defined by $g(x) = c$, then by the chain rule

$$\frac{d}{dt}g(x_1(t)) = \frac{d}{dt}x_1(t) \cdot \nabla g(x_1(t)) = 0 \tag{3.4}$$

This implies that $\nabla g(x_1(t))$ is orthogonal to the tangent vector $(d/dt)x_1(t)$ for any path $x_1(t)$ in the surface defined by $g(x) = c$.

Conversely, if x is any point in the surface $g(x) = c$ and y is any vector such that $y \cdot \nabla g(x) = 0$, then it follows from the Implicit Function Theorem there exists a path $x_1(t)$ in the surface $g(x) = c$ such that $x_1(0) = x$ and $(d/dt)x_1(0) = y$. This result and (3.4) imply that the gradient vector $\nabla g(x)$ is always orthogonal to the surface defined by $g(x) = c$ at x .

Now let x be a solution of (3.3). I claim that $\nabla f(x) = \lambda \nabla g(x)$ for some scalar λ . First, we can always write $\nabla f(x) = c \nabla g(x) + y$ where $y \cdot \nabla g(x) = 0$. If $x(t)$ is a path in the surface with $x(0) = x$ and $(d/dt)x(0) \cdot \nabla f(x) \neq 0$, it follows from (3.2) with $y = (d/dt)x(0)$ that there are values for $f(x)$ for $x = x(t)$ in the surface that both larger and smaller than $f(x)$.

Thus, if x is a maximum or minimum of $f(x)$ in the surface and $\nabla f(x) = c \nabla g(x) + y$ for $y \cdot \nabla g(x) = 0$, then $y \cdot \nabla f(x) = y \cdot \nabla g(x) + y \cdot y = y \cdot y = 0$ and $y = 0$. This means that $\nabla f(x) = c \nabla g(x)$, which completes the proof of Lagrange's Theorem for one constraint ($p = 1$).

Next, suppose that we want to solve

$$\max f(x) \quad \text{subject to} \quad g_1(x) = c_1, \dots, g_p(x) = c_p \tag{3.5}$$

for p constraints. Let x be a solution of (3.5). Recall that the each vector $\nabla g_j(x)$ is orthogonal to the surface $g_j(x) = c_j$ at x . Let \mathcal{L} be the linear space

$$\mathcal{L} = \text{span}\{ \nabla g_j(x) : 1 \leq j \leq p \}$$

I claim that $\nabla f(x) \in \mathcal{L}$. This would imply

$$\nabla f(x) = \sum_{j=1}^p \lambda_j \nabla g_j(x)$$

for some choice of scalar values λ_j , which would prove Lagrange's Theorem.

To prove that $\nabla f(x) \in \mathcal{L}$, first note that, in general, we can write $\nabla f(x) = w + y$ where $w \in \mathcal{L}$ and y is perpendicular to \mathcal{L} , which means that $y \cdot z = 0$ for any $z \in \mathcal{L}$. In particular, $y \cdot \nabla g_j(x) = 0$ for $1 \leq j \leq p$. Now find a path $x_1(t)$ through x in the intersection of the surfaces $g_j(x) = c_j$ such that $x_1(0) = x$ and $(d/dt)x_1(0) = y$. (The existence of such a path for sufficiently small t follows from a stronger form of the Implicit Function Theorem.) It then follows from (3.2) and (3.5) that $y \cdot \nabla f(x) = 0$. Since $\nabla f(x) = w + y$ where $y \cdot w = 0$, it follows that $y \cdot \nabla f(x) = y \cdot w + y \cdot y = y \cdot y = 0$ and $y = 0$. This implies that $\nabla f(x) = w \in \mathcal{L}$, which completes the proof of Lagrange's Theorem.

4. Warnings. The same warnings apply here as for most methods for finding a maximum or minimum:

The system (1.4) does not look for a maximum (or minimum) of $f(x)$ subject to constraints $g_j(x) = c_j$, but only a point x on the set of values determined by $g_j(x) = c_j$ whose first-order changes in x are zero. This is satisfied by a value $x = x_0$ that provides a minimum or maximum typical for $f(x)$ in a neighborhood of x_0 , but may only be a local minimum or maximum. There may be several local minima or maxima, each yielding a solution of (1.4). The criterion (1.4) also holds for "saddle points" of $f(x)$ that are local maxima in some directions or coordinates and local minima in others. In these cases, the different values $f(x)$ at the solutions of (1.4) have to be evaluated individually to find the global maximum.

A particular situation to avoid is to look for a maximum value of $f(x)$ by solving (1.4) or (1.3) when $f(x)$ takes arbitrarily large values when any of the components of x are large (as is the case for $f(x)$ in (1.2)) and (1.4) has a unique solution x_0 . In that case, x_0 is probably the global minimum of $f(x)$ subject to the constraints, and not a maximum. In that case, rather than find the best possible value of $f(x)$, one may end up with the worst possible value. After solving (1.3) or (1.4), one often has to look at the problem more carefully to see if it is a global maximum, a global minimum, or neither.

Another situation to avoid is when the maximum or minimum is on the boundary of the values for which $f(x)$ is defined. In that case, the maximum or minimum is not an interior value, and the first-order changes in $f(x)$ (that is, the partial derivatives of $f(x)$) may not be zero at that point. An example is $f(x) = x$ on the unit interval $0 \leq x \leq 1$. The minimum value of $f(x) = x$ on the interval is $x = 0$ and the maximum is $x = 1$, but neither are solutions of $f'(x) = 0$.