Lagrange Multipliers, Constrained Maximization, and the Maximum Entropy Principle

John Wyatt
4/6/04

I. The Geometric Idea Behind Lagrange Multipliers

Example 1

Suppose a bead is allowed to slide along a smooth curved wire under the influence of gravity until it comes to rest as is shown in fig. 1. The only possible rest points are points where the wire is parallel to the ground, i.e., the interval *a* and the points *b, c, d,* and *e*. These are precisely the points where the wire (i.e., the constraint set *C*) is perpendicular to the downward force *f*, the gradient of the gravitational potential $\phi = mgh$. Thus the maximum and minimum of ϕ , points *d* and *e*, are both points at which $f \perp C$. The converse, however, does not hold, since *f* is also perpendicular to *C* on the interval *a* at the points *b* and *c*, which are not global maxima or minima for ϕ .

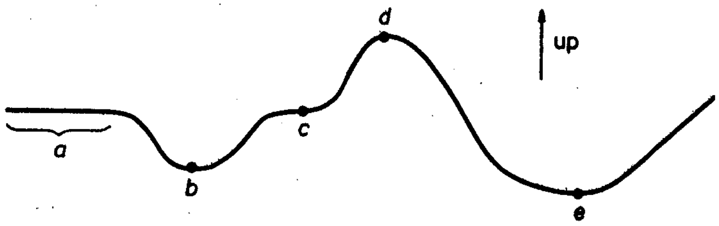

FIGURE 1

Figure 1

The Gradient of a Scalar Function

Let

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

represent position in n-dimensional space, and $\phi(\boldsymbol{x})$ be a scalar function of **x**. For example

$$\phi(\boldsymbol{x}) = h(x_1, x_2)$$

could be the altitude h at a point $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ on a map of an area of terrain, or

$$\phi(\boldsymbol{x}) = T(x_1, x_2, x_3)$$

could be the temperature at a point $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ in the classroom.

Similarly

$$\boldsymbol{x} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix} = \boldsymbol{p}$$

could be a vector of probabilities, and

$$H(\boldsymbol{p}) = \sum_{k=1}^{n} p_k \log\left(\frac{1}{p_k}\right)$$

could be the entropy of that set of probabilities.

At any point $\hat{\boldsymbol{x}}$, *the gradient of ϕ at $\hat{\boldsymbol{x}}$*, is the vector

$$\nabla\phi(\hat{\boldsymbol{x}}) = \begin{pmatrix} \frac{\delta\phi}{\delta x_1}(\hat{\boldsymbol{x}}) \\ \frac{\delta\phi}{\delta x_2}(\hat{\boldsymbol{x}}) \\ \vdots \\ \frac{\delta\phi}{\delta x_n}(\hat{\boldsymbol{x}}) \end{pmatrix}$$

This can be visualized as a vector with its tail at $\hat{x}$ pointing in the direction of steepest increase of ϕ.

For example, if you are skiing on a mountain with elevation $h(x_1, x_2)$, at any point $\hat{x}$ the negative gradient of *h*, $(-\nabla h(\hat{x}))$ points in the direction of steepest descent (the “fall line”) down the mountain from $\hat{x}$.

Example 2

Find the point on the unit circle which maximizes $\phi(x,y)=x+y$. The result may be seen from inspection of fig. 2; $x^*=\left(\sqrt{2/2}\sqrt{2/2}\right)$. The unit circle is the constraint set *C*, and the vector field $\nabla\phi$ is sketched in fig. 2. Notice that $\nabla\phi$ is perpendicular to *C*, at **x*** and also at the point -**x***, which minimizes ϕ.

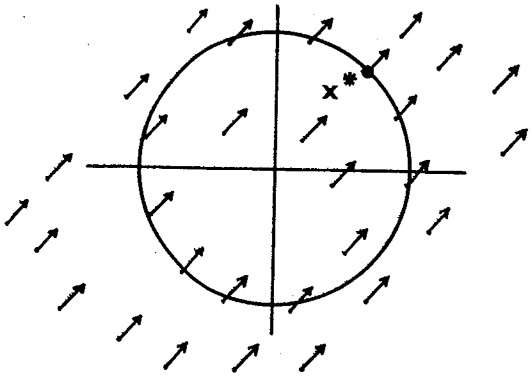

FIGURE . 2

Figure 2

Geometric Principle of Lagrange Multipliers

Suppose we wish to find the maximum of a scalar function $\phi(\boldsymbol{x})$ subject to the constraint that the point **x** must lie in a given smooth curve or surface, *C*, in the space. If **x*** is the point in *C* at which the maximum occurs, then it can be seen that $\nabla\phi$, the gradient of ϕ, must be perpendicular to *C* at **x***. For if any component of $\nabla\phi$ were parallel to *C* at **x***, we could move away from **x*** and up the gradient of ϕ while remaining on *C*, contrary to the assumption that this point was a maximum. Note that $\nabla\phi\perp C$ is a necessary but not sufficient condition for a maximum.

Since maximizing ϕ is the same as minimizing $-\phi$, and since $\nabla\phi \perp C$ is equivalent to $\nabla(-\phi) \perp C$, it follows that the minimum of ϕ must also occur at a point where $\nabla\phi \perp C$.

Example 3

Let C be a surface in 3-dimensional space. We want to find the maximum value attained by ϕ on C. A possible vector field $\nabla\phi$ is sketched in fig. 3, and this allows us to look for possible maxima.

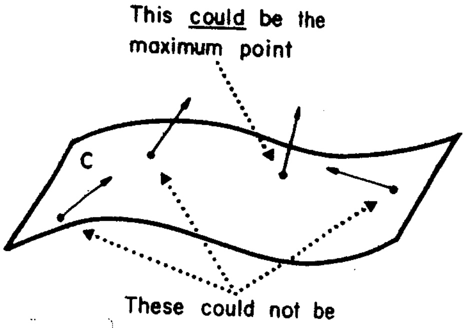

FIGURE 3

Figure 3

II. Turning the Geometric Principle into a Calculation – Surfaces Defined by Constraints

For a surface defined by a constraint given by a scalar function g,

$g(x) = G,$

a *normal vector to the surface* at any point $\hat{x}$ in the surface (i.e., $\hat{x}$ satisfying $g\ (\hat{x}) = G$) is given by

$$\nabla g(\hat{x})$$

Example 4

The unit sphere centered at the origin in 3-space is the set of points satisfying the constraint

$$g_1(x,y,z) = x^2 + y^2 + z^2 - 1 = 0$$

See fig. 4.

The normal to the sphere at the point $(x,y,z) = (1,0,0)$ is just the set of all vectors of the form $\lambda \cdot \nabla g_1(1,0,0) = \lambda \cdot (2,0,0)$, where the quantity λ is any scalar constant. It is the set of all vectors with tails at (1,0,0) which are perpendicular to the sphere. In this example it is the x axis itself.

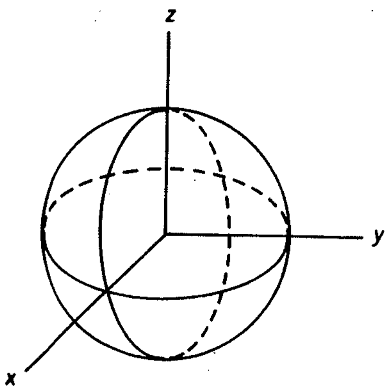

FIGURE 4

Figure 4

Definition

In n-dimentional space let C be the *(n-k)* dimensional surface consisting of all $\mathbf{x}$ that satisfy all k constraint equations

$$\begin{aligned} g_1(\boldsymbol{x}) &= G_1 \\ g_2(\boldsymbol{x}) &= G_2 \\ &\vdots \\ g_k(\boldsymbol{x}) &= G_k \end{aligned}$$

At any point $\hat{\boldsymbol{x}} \ \varepsilon \ C$ (i.e., $g_1(\hat{\boldsymbol{x}}) = G_1,\ g_2(\hat{\boldsymbol{x}}) = G_2, \cdots,\ g_k(\hat{\boldsymbol{x}}) = G_k$) the *normal space to C at* $\hat{\boldsymbol{x}}$ consists of all vectors (with their tails at $\hat{\boldsymbol{x}}$) that are linear combinations of $\nabla g_1(\hat{\boldsymbol{x}}), \nabla g_2(\hat{\boldsymbol{x}}), \cdots,\ \nabla g_k(\hat{\boldsymbol{x}})$.

In other words, a vector $\mathbf{v}$ with its tail at $\hat{\boldsymbol{x}}$ is normal to $C \Leftrightarrow$

$$\boldsymbol{v} = \lambda_1 \nabla g_1(\hat{\boldsymbol{x}}) + \lambda_2 \nabla g_2(\hat{\boldsymbol{x}}) + \cdots + \lambda_k \nabla g_k(\hat{\boldsymbol{x}}).$$

Example 5

The unit circle lying in the *x*-*y* plane in 3-space is the set of points satisfying

$$g_1(x,y,z) = x^2 + y^2 + z^2 - 1 = 0$$
$$g_2(x,y,z) = z = 0$$

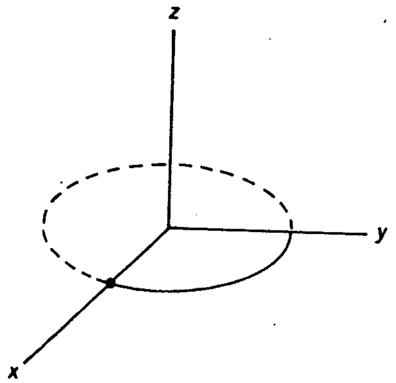

FIGURE 5

Figure 5

Exercise

Find the normal space to the unit circle defined above at the point

$$\hat{\boldsymbol{x}} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

Describe the normal space geometrically. Find the gradients of g_1 and g_2 at $\hat{x}$ and give a formula for all points in the normal space in terms of ∇g_1 and ∇g_2.

III. Lagrange Multiplier Theorem[1]

The maximum value of $\phi(\boldsymbol{x})$ subject to the constraints that

$$\begin{gathered} g_1(\boldsymbol{x}) = G_1 \\ g_2(\boldsymbol{x}) = G_2 \\ \vdots \\ g_k(\boldsymbol{x}) = G_k \end{gathered}$$

if such a value exists, must occur at a point **x*** in C such that

$$\nabla\phi(\boldsymbol{x}^*) = \sum_{j=1}^{k} \lambda_j \nabla g_j(\boldsymbol{x}^*)$$

is satisfied for *some value* of $\lambda_1, \ldots, \lambda_n$. The λ's are called *Lagrange multipliers.*

[1] The technical assumptions are that $\phi(\boldsymbol{x})$ and $g_j(\boldsymbol{x}), j = 1, 2, \ldots, k < n,$ are defined and continuously differentiable on an open set in n-dimensional Euclidean space. We further assume that at each point where $g_1 = G_1, \cdots, g_k = G_k$, $\nabla g_1, \nabla g_2, \cdots, \nabla g_k$ are linearly independent.

IV. Maximizing the Entropy Subject to a Linear Constraint

We will often use Lagrange multipliers to solve the following type of problem:

$$\text{Maximize} \qquad H(p)=\sum_{j=1}^{n} p_j \log\left(1/p_j\right)=\frac{\sum_{j=1}^{n} p_j \ln\left(1/p_j\right)}{\ln 2}$$

subject to the constraints[2]

$$g_1(\boldsymbol{p})=\sum_{j=1}^{n} p_j = 1$$

$$g_2(\boldsymbol{p})=\sum_{j=1}^{n} p_j g_j = G$$

We can simplify the algebra slightly by maximizing instead

$$\hat{H}(p)=\sum_{j=1}^{n} p_j \ln\left(1/p_j\right)$$

To find the various gradients, note that

[2] The third constraint, each $p_j \geq 0$, turns out to be automatically satisfied in this class of problems and does not need to be handled separately. (We are very lucky!)

$$\frac{\delta \hat{H}}{\delta p_j} = \ln\left(1/p_j\right) + 1$$

$$\frac{\delta g_1}{\delta p_j} = 1$$

$$\frac{\delta g_2}{\delta p_j} = g_j$$

The Lagrange multiplier equations tell us that the vector **p*** of the probabilities that maximizes H subject to the two linear constraints must satisfy

$$\nabla \hat{H}\left(p^*\right) = \alpha \nabla g_1\left(\boldsymbol{p}^*\right) + \beta \nabla g_2\left(p^*\right) \tag{1}$$

$$g_1\left(\boldsymbol{p}^*\right) = 1 \tag{2}$$

$$g_2\left(\boldsymbol{p}^*\right) = G, \tag{3}$$

or, more concretely

$$\ln\left(1/p_j^*\right) + 1 = \alpha + \beta g_j, \; j = 1, \cdots, n \tag{1'}$$

$$\sum_{j=1}^{n} p_j^* = 1 \tag{2'}$$

$$\sum_{m=1}^{n} g_m p_m^* = G \tag{3'}$$

The first equation tells us

$$p_j^* = e^{1-\alpha} e^{-\beta g_j}, \quad j = 1, 2, \cdots, n$$

We can guarantee the second equation (2') is satisfied by defining

$$p_j^* = \frac{e^{1-\alpha}e^{-\beta g_j}}{\sum_{j=1}^{n} e^{1-\alpha}e^{-\beta g_j}} = \frac{e^{-\beta g_j}}{\sum_{j=1}^{n} e^{-\beta g_j}},$$

which eliminates α. The remaining unknown is β, which must be chosen to satisfy the third equation (3')

$$\frac{\sum_{m=1}^{n} g_m e^{-\beta g_m}}{\sum_{j=1}^{n} e^{-\beta g_j}} = G,$$

i.e.,

$$\sum_{m=1}^{n} (g_m - G) e^{-\beta g_m} = 0.$$

This equation must be solved numerically in most cases.

Additional Justification for Using Maximum Entropy

We use the maximum entropy method to estimate probabilities when we have insufficient data to determine them accurately. The justification has been that the maximum entropy distribution, subject to whatever constraints are known, introduces no unjustified bias or constraint. A second justification comes from this important example due to Boltzmann.

Dice Example

Suppose n independent fair 6-sided dice are thrown in sequence. Given that the total number of spots showing is na $(1 \leq a \leq 6)$, find a rational basis for estimating the proportion of dice showing face i, $i = 1, 2, \cdots 6$, for large n.

Approach 1: Maximum Entropy

Suppose n_1 tosses yield 1 spot, ..., n_6 tosses yield 6 spots, with

$$n_1 + n_2 + \cdots + n_6 = n.$$

Then if a toss is chosen at random, the probability it will have i spots showing is

$$p_i = \frac{n_i}{n}.$$

The constraint on the set of outcomes is

$$\sum_{i=1}^{6} in_i = na,$$

i.e.,

$$\sum_{i=1}^{6} i\frac{n_i}{n} = a,$$

i.e.,

$$\sum_{i=1}^{6} i\, p_i = a.$$

According to the maximum entropy approach we maximize

$$H = \sum_{i=1}^{6} p_i \log \frac{1}{p_i}$$

subject to the constraints

$$\sum_{i=1}^{6} i\, p_i = a$$

$$\sum_{i=1}^{6} p_i = 1$$

np_i an integer, $i = 1, \cdots, 6.$

If we ignore the last constraint, this becomes our most standard Lagrange multiplier problem, with solution

$$p_i = \frac{e^{-i\beta}}{\sum_{i=1}^{6} e^{-i\beta}},$$

with β chosen so that

$$\sum_{i=1}^{6} i\, p_i = a.$$

Approach 2: Finding Most Likely Values of $n_1, \ldots, n_6$

Note that for independent fair dice, all sequences of outcomes are equally probable, with probability $\left(\frac{1}{6}\right)^n$. For any specific numbers of tosses $n_1, \ldots, n_6$ showing 1,..., 6 spots, the number of possible sequences is

$$\binom{n}{n_1, n_2, \cdots n_6} = \frac{n!}{n_1!n_2!n_3!n_4!n_5!n_6}.$$

Therefore $p(n_1, n_2, \ldots, n_6)$ is proportional to this quantity, and the value of $n_1, \ldots, n_6$ that maximizes it, subject to the constraints

$$\sum_{i=1}^{n} n_i = n$$

$$\sum_{i=1}^{n} i\ n_i = na$$

is the most probable. This quantity is hard to manage, but using the crude Stirling's approximation for large n,

$$n! \approx \left(\frac{n}{e}\right)^n,$$

we have

$$\binom{n}{n_1, n_2, \cdots, n_6} \approx \frac{\left(\frac{n}{e}\right)^n}{\prod_{i=1}^{6}\left(\frac{n_i}{e}\right)^{n_i}} =$$

$$\prod_{i=1}^{6}\left(\frac{n}{n_i}\right)^{n_i}$$

To put this into a more familiar form, as before let

$$p_i = \frac{n_i}{n}$$

and note that

$$e^{\frac{nH}{\ln 2}(p_1, \cdots, p_6)} = e^{\frac{n}{\ln 2}\sum_{i=1}^{6} p_i \ln(1/p_i)} =$$

$$\left(\prod_{i=1}^{6}\left(\frac{1}{p_i}\right)^{np_i}\right)^{\frac{1}{\ln 2}} = \left(\prod_{i=1}^{6}\left(\frac{n}{n_i}\right)^{n_i}\right)^{\frac{1}{\ln 2}},$$

so

$$\binom{n}{n_1, n_2, \cdots, n_6} \approx e^{\frac{nH(p_1, \cdots, p_6)}{\ln 2}},$$

which has a maximum wherever the entropy has a maximum.

Conclusion

For large n, maximizing the entropy of a set of probabilities $(p_1,\cdots,p_m)$ subject to any set of constraints is approximately equivalent, for large n, to maximizing the probability of

$$n_i \triangleq n\,p_i, \quad 1 \le i \le m$$

outcomes of type i in a sequence of n independent trials.