

The Method of Lagrange Multipliers

Phil Lucht

Rimrock Digital Technology, Salt Lake City, Utah 84103

last update: June 8, 2015

rimrock@xmission.com

Maple code is available upon request. Comments and errata are welcome.

The material in this document is copyrighted by the author.

The graphics look ratty in Windows Adobe PDF viewers when not scaled up, but look just fine in this excellent freeware viewer: <http://www.tracker-software.com/pdf-xchange-products-comparison-chart>.

The table of contents has live links. Most PDF viewers provide these links as bookmarks on the left.

Overview and Summary	2
1. The R matrix and its Rank.....	3
2. The Method of Lagrange Multipliers.....	6
3. Proof of Theorem 1	9
(a) Proof for $N = 6$ and $S = 4$	9
(b) Proof for general $S \leq N$	13
4. The Gradient Interpretation: Example 1.....	17
5. Example 2: Extremal distances between an ellipse and a circle	22
6. Example 3: The Boltzmann factor in Statistical Mechanics	29
6.1 Statement of The Boltzmann Extremum Problem	29
6.2 Solution of The Boltzmann Extremum Problem.....	31
6.3 More details of the solution	34
6.4 A numerical example with plots	37
Appendix A : Proof that matrix rank is the number of independent columns and rows	43
References	46

Overview and Summary

The intent of the following presentation is to provide a reasonable derivation of why the Method of Lagrange Multipliers works. Although we do discuss the relevant "gradient approach", it seems more convincing to use a "matrix approach" where the solution points of a constrained extremum problem are associated with the points at which a certain matrix drops below full rank (an approach used by Buck, see References). Then the gradient statement of "why it works" follows at once. It certainly is a mind-boggling exercise to visualize the solution sets as intersections of multiple surfaces in a space of more than three dimensions.

Every document on this subject needs a "Theorem 1" and ours is no exception. Fortunately for the reader, there are no other theorems (ignoring Appendix A), and Theorem 1 is fairly easy to prove. The reader should be at least dimly familiar with basic linear algebra facts.

We first state our Theorem 1 in Section 1, then show in Section 2 how it provides an immediate derivation of the Method of Lagrange Multipliers. Section 3 then proves Theorem 1. Section 4 gives the gradient interpretation and provides an example which at least can be visualized with simple surfaces. Section 5 presents another example which, though simple, requires numeric work to solve. Finally, Section 6 provides an extended standard example involving statistical mechanics. Appendix A proves for a general matrix that column rank = row rank = rank.

What the reader will not find below is a treatment of questions of continuity and whether solutions do or do not exist for special situations. These topics seems well treated in other papers.

The general tone is more that of an engineer or physicist, not that of an abstract mathematician. Maple code is used where it seems useful to implement calculations or display graphs.

When an earlier equation is quoted, its equation number is put in italics.

A very short list of References is provided on the last page.

1. The R matrix and its Rank

We shall operate in N dimensional space E^N with coordinates $\mathbf{r} = (x_1, x_2 \dots x_N)$.

We seek an extremum of a real function $f(\mathbf{r})$ subject to S-1 constraints $a(\mathbf{r}) = 0, b(\mathbf{r}) = 0, \dots q(\mathbf{r}) = 0$ where we use q generically to represent the last constraint function. For example, if there are two constraints, then $S = 3$ and the two constraint equations are $a(\mathbf{r}) = 0$ and $b(\mathbf{r}) = 0$. So here is the problem:

$$\text{find extremum of } f(\mathbf{r}) \text{ subject to constraints } \begin{matrix} a(\mathbf{r})=0, & b(\mathbf{r})=0, & c(\mathbf{r})=0, & \dots & q(\mathbf{r})=0 \\ 1 & 2 & 3 & & S-1 \end{matrix} \quad (1.1)$$

Construct the following matrix of partial derivatives, where $f_i(\mathbf{r}) \equiv \partial_i f(\mathbf{r}) = \partial f(\mathbf{r}) / \partial x_i$. That is, the subscript on f indicates which of the arguments $x_1, x_2, x_3, \dots, x_N$ the partial is with respect to. So here is that matrix, which we shall call R :

$$R = \begin{pmatrix} f_1 & f_2 & f_3 & \dots & f_N \\ a_1 & a_2 & a_3 & \dots & a_N \\ b_1 & b_2 & b_3 & \dots & b_N \\ \dots & & & & \\ q_1 & q_2 & q_3 & \dots & q_N \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{a} \\ \mathbf{b} \\ \dots \\ \mathbf{q} \end{pmatrix} \quad (1.2)$$

For example, $f_2 \equiv \partial f(x_1, x_2, x_3, \dots, x_N) / \partial x_2$.

On the far right we abbreviate each row of functions in **bold**, indicating a row vector like $\mathbf{f}$.

The matrix has S rows and N columns. We shall only be interested in $S \leq N$ (see comment below). If $S = N$, the matrix is square, otherwise it has more columns than rows, so has a horizontal band shape.

The maximum possible **rank** of the above matrix is S since this is the smaller of the number of rows and columns (see below).

Since all the functions are functions of $\mathbf{r}$, we really have $R(\mathbf{r})$. As the point $\mathbf{r}$ is varied, the elements of the matrix $R(\mathbf{r})$ generally change.

If $S = N$, the matrix is square and we can talk about the determinant $\det(R)$. In this case there is only one $S \times S$ "submatrix" and it is the entire matrix R.

If $S < N$, the matrix is wider than it is tall. In this case, we can talk about square submatrices within R which are of shape $S \times S$. For example, for $N = 6$ and $S = 4$ we have three constraints and the R matrix is this,

$$R = \begin{array}{|c|c|c|c|c|c|} \hline f_1 & f_2 & f_3 & f_4 & f_5 & f_6 \\ \hline a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \\ \hline b_1 & b_2 & b_3 & b_4 & b_5 & b_6 \\ \hline c_1 & c_2 & c_3 & c_4 & c_5 & c_6 \\ \hline \end{array} \quad (1.3)$$

Here we have outlined two 4x4 submatrices in red and blue. The columns of a submatrix need not be contiguous, so the three green boxes indicate another 4x4 submatrix. In this example, the number of submatrices is given by the binomial factor (6,4) -- pick a committee of 4 columns from 6 candidates. So we have shown only 3 out of $(6,4) = 6!/(4!2!) = 15$ possible 4x4 submatrices. We can indicate a submatrix using this notation:

$$\text{red} = (f_1, f_2, f_3, f_4) \quad \text{blue} = (f_2, f_3, f_4, f_5) \quad \text{green} = (f_1, f_2, f_4, f_6) \quad (1.4)$$

That is to say, we abbreviate the submatrix by stating only the values in the first row of the submatrix. Since our main interest will be whether or not the determinants of these 4x4 submatrices vanish, the ordering of the columns within a submatrix is not important.

For the general R matrix shown above with S columns and N rows, there are (N,S) possible submatrices of size SxS. A particular submatrix is then indicated by a sequence of S subscripted f values.

Why not $S > N$? One can regard $u = f(x_1, x_2, x_3, \dots, x_N)$ as an N-dimensional surface in E^{N+1} where the extra dimension is u. For example, for $N = 2$, $u = f(x, y)$ is a 2-dimensional surface in E^3 .

The constraint equation $a(x_1, x_2, x_3, \dots, x_N) = 0$ is an N-1 dimensional surface in E^N . Again for $N = 2$, $a(x, y) = 0$ is a 1-dimensional surface (a curve) in E^2 . If $a(x, y) = x^2 - y^2 - 1$, that curve is a circle.

A constraint surface can be "extruded" in the u direction of E^N to give an N dimensional surface in E^{N+1} . For $N=2$, if the constraint is a circle, that extruded surface is a cylinder in E^3 .

The intersection of the N-dimensional surface $u = f$ with the N-dimensional surface $a = 0$ is a surface of dimension N-1. In our $N = 2$ example, the 2D surface $u = f$ intersects the cylinder in a curve.

Each extra constraint lowers the dimensionality of the final intersection surface. If $S = 2$ (one constraint), the intersection surface was just seen to be of dimension N-1. For $S = 3$ (two constraints), that surface is down to dimension N-2. The dimension of the final intersecting surface is thus $N-S+1$. When $S = N$, the final surface is a 1D curve. If we try to set $S > N$, the intersection has no degree of freedom at all and one cannot have a continuous extremum situation.

An $N = 2$ example of this paragraph's discussion is presented in Section 4.

We now quote a few facts from linear algebra concerning the rank of a matrix:

Definition: The **rank** r of an $m \times n$ matrix A is the dimension of the largest non-vanishing subdeterminant (minor) within A. [Shilov 1.92] Thus, $r \leq \min(m, n)$. (1.5)

Definitions: [Shilov 2.21] Consider an $m \times n$ matrix A. Let $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ be an arbitrary selection of N rows (row vectors) from the matrix A. This set of N rows is **linearly dependent** if one can find a set of constants k_i at least one of which is non-zero such that

$$k_1 \mathbf{r}_1 + k_2 \mathbf{r}_2 + \dots + k_N \mathbf{r}_N = 0 \quad (1.6)$$

There could be several such equations, or there could be only one. This generally means that at least one row can be expressed as a linear combination of the other rows. A legal such equation would be $\mathbf{r}_2 = 0$ in which case $k_2 = 1$. One could then say that $\mathbf{r}_2$ is a linear combination of the other rows with coefficients all zero. In any event, if one or more rows are all zeros, a set of N rows are linearly dependent. If no such

equation (1.6) exists, then the set of N rows is **linearly independent**. For example, the set of vectors $\mathbf{r}_1 = (1,0)$ and $\mathbf{r}_2 = (0,1)$ are linearly independent.

The above discussion also applies to columns. Take "row" $\rightarrow$ "column" and $\mathbf{r}_i \rightarrow \mathbf{c}_i$.

Facts: The rank r as defined in (1.5) has these properties for any $m \times n$ matrix A : (1.7)

- (a) the rank r equals the number of linearly independent columns of A [Shilov 3.12]
- (b) the rank r equals the number of linearly independent rows of A [Shilov 3.13]

See Appendix A below for a brief proof of the claims of (1.7).

With the above as background, the following theorem is fundamental to our Lagrange Multiplier presentation:

Theorem 1. If the point $\mathbf{r}$ is a solution to the extremal problem (1.1), then $\text{rank}[R(\mathbf{r})] < S$. (1.8)

According to the rank definition (1.5),

$$\text{rank}(R) < S \quad \Leftrightarrow \quad \text{all } S \times S \text{ submatrices must have zero determinant} \quad (1.9)$$

If, for example, the red submatrix in Fig 1.3 had a non-zero determinant, then $\text{rank}(R) = 4 = S$.

The point of Theorem 1 is that the solution point $\mathbf{r}$ of the constrained extremum problem must be a point where the $R(\mathbf{r})$ matrix drops below full rank. We shall prove this theorem in Section 1.3 by showing that, if $\mathbf{r}$ is a solution to the extremum problem, then all those $S \times S$ submatrices discussed above have zero determinant and thus $\text{rank}(R) < S$. If one is searching for candidate solutions $\mathbf{r}$ to the extremum problem, one can restrict one's search to points $\mathbf{r}$ for which $\text{rank}[R(\mathbf{r})] < S$.

Comment on the R matrix. Imagine a general transformation from x -space $\mathbf{x} = (x_1, x_2, \dots, x_N)$ to u -space with coordinates $\mathbf{u} = (u_1, u_2, \dots, u_S)$. One might write such a transformation as $\mathbf{u} = \mathbf{F}(\mathbf{x})$ where $\mathbf{F}: E^N \rightarrow E^S$. Although this transformation in general is non-linear, in a tiny neighborhood of point $\mathbf{x}$ in x -space (and the corresponding point $\mathbf{u}$ in u -space), the transformation will be linear if certain conditions are met. That local linear relation is then described by $d\mathbf{u} = R d\mathbf{x}$ where R is a matrix (N columns, S rows) whose matrix elements are $R_{ij} = \partial u_i / \partial x_j$. If we think of

$$\begin{array}{llll} f = u_1(x_1, x_2, \dots, x_N) & R_{11} = \partial u_1 / \partial x_1 = \partial f / \partial x_1 = f_1 & R_{12} = f_2 & \text{etc.} \\ a = u_2(x_1, x_2, \dots, x_N) & R_{21} = \partial u_2 / \partial x_1 = \partial a / \partial x_1 = a_1 & R_{22} = a_2 & \text{etc} \\ b = u_3(x_1, x_2, \dots, x_N) & & & \\ \dots & & & \\ q = u_S(x_1, x_2, \dots, x_N) & & & \end{array} \quad (1.10)$$

then the matrix shown in (1.2) is exactly the R matrix for this general transformation $\mathbf{u} = \mathbf{F}(\mathbf{x})$. If $S < N$, then the R matrix is not square, the relation $d\mathbf{u} = R d\mathbf{x}$ is a "projection" of a larger space into a smaller one, and the equation therefore cannot be inverted. This notation " R " is used extensively in our Tensor Analysis document, see in particular Chapter 2, though only invertible transformations $\mathbf{F}: E^N \rightarrow E^N$ are considered there. Sometimes R is referred to as "the differential" of a transformation.

Before proving Theorem 1, we want to cut to the chase and state the Method of Lagrange Multipliers to create a framework in which candidate solutions for the extremum problem can be found.

2. The Method of Lagrange Multipliers

The extremum problem stated above in (1.1),

$$\text{find extremum of } f(\mathbf{r}) \text{ subject to constraints } \begin{matrix} a(\mathbf{r})=0, & b(\mathbf{r})=0, & c(\mathbf{r})=0, & \dots & q(\mathbf{r})=0 \\ 1 & 2 & 3 & & S-1 \end{matrix} \quad (1.1)$$

can be solved in the following manner:

(a) Define the following function H [the λ_i are the "Lagrange multipliers"]

$$H(\mathbf{r}, \boldsymbol{\lambda}) \equiv f(\mathbf{r}) + \lambda_1 a(\mathbf{r}) + \lambda_2 b(\mathbf{r}) + \dots + \lambda_{S-1} q(\mathbf{r}) \quad (2.1)$$

Notice that H is a function of $N + (S-1) = N+S-1$ variables which are the x_i and λ_i .

(b) Compute all $N+S-1$ partial derivatives of H , which as usual we denote by H_i .

For the first N partials $i = 1, 2, \dots, N$ one gets

$$H_i = f_i + \lambda_1 a_i + \lambda_2 b_i + \dots + \lambda_{S-1} q_i \quad i = 1, 2, \dots, N \quad (2.2)$$

For the final $S-1$ partials one finds,

$$\begin{aligned} H_{N+1} &= a \\ H_{N+2} &= b \\ &\dots \\ H_{N+S-1} &= q \end{aligned} \quad (2.3)$$

(c) Now set all $N+S-1$ partials of H to 0, that is, set $H_i = 0$ for $i = 1$ to $N+S-1$. Then (2.2) and (2.3) read

$$\begin{aligned} f_i(\mathbf{r}) + \lambda_1 a_i(\mathbf{r}) + \lambda_2 b_i(\mathbf{r}) + \dots + \lambda_{S-1} q_i(\mathbf{r}) &= 0 & i = 1, 2, \dots, N \\ a(\mathbf{r}) = 0 \quad b(\mathbf{r}) = 0 \quad \dots \quad q(\mathbf{r}) = 0 & & i = N+1 \dots N+S-1 \end{aligned} \quad (2.4)$$

The second group of $S-1$ equations is a statement of the $S-1$ constraints of the problem.

The first set of equations in (2.4) can be written in this vector notation

$$\mathbf{f}(\mathbf{r}) + \lambda_1 \mathbf{a}(\mathbf{r}) + \lambda_2 \mathbf{b}(\mathbf{r}) + \dots + \lambda_{S-1} \mathbf{q}(\mathbf{r}) = 0 \quad (2.5)$$

where each bolded vector is a row of the R matrix (1.2). This equation has the form (1.6) and certainly the coefficient of $\mathbf{f}(\mathbf{r})$ is non-zero, being one.

Equation (2.5) thus says that the S rows of the R matrix are linearly dependent! According to (1.7), this means that $\text{rank}[R(\mathbf{r})] < S$. If one had $\text{rank}(R) = S$, all S rows of R would be linearly independent and then an equation of the form (2.5) could not exist. According to **Theorem 1** of (1.8), any point $\mathbf{r}$ which makes

(2.5) be true is therefore a *candidate* for a solution to the constrained extremum problem (1.1). Such a solution value of $\mathbf{r}$ will be associated with solution values for the λ_i .

(d) It is then incumbent on the problem solver to solve the following $N+S-1$ equations for the $N+S-1$ unknowns which are $\mathbf{r} = (x_1, x_2, \dots, x_N)$ and $\lambda_1, \lambda_2, \dots, \lambda_{S-1}$. Here again are the equations to solve:

$$\begin{aligned}
 f_1(\mathbf{r}) + \lambda_1 a_1(\mathbf{r}) + \lambda_2 b_1(\mathbf{r}) + \dots + \lambda_{S-1} q_1(\mathbf{r}) &= 0 \\
 f_2(\mathbf{r}) + \lambda_1 a_2(\mathbf{r}) + \lambda_2 b_2(\mathbf{r}) + \dots + \lambda_{S-1} q_2(\mathbf{r}) &= 0 & // \text{recall: } b_2 \text{ means } \partial b / \partial x_2 \\
 &\dots \\
 f_N(\mathbf{r}) + \lambda_1 a_N(\mathbf{r}) + \lambda_2 b_N(\mathbf{r}) + \dots + \lambda_{S-1} q_N(\mathbf{r}) &= 0 & // N \text{ equations} \\
 \\
 a(\mathbf{r}) &= 0 \\
 b(\mathbf{r}) &= 0 \\
 &\dots \\
 q(\mathbf{r}) &= 0 & // S-1 \text{ constraint equations}
 \end{aligned} \tag{2.6}$$

Since all the functions in these equations are known (since $f(\mathbf{r})$ and all the constraints are known), it is in principle possible to solve the equations for $\mathbf{r}$ and the λ_i . Then those solution $\mathbf{r}$ values are candidate solutions to the constrained extremum problem and need to be studied more to see if they really solve the problem of interest (max, min, saddle, etc). In general the various functions of $\mathbf{r}$ are non-linear and not just polynomials, and there are likely to be multiple candidate solutions. A numeric solution may be required. At least this method very significantly reduces the search for solutions to the initial problem (1.1). The Method of Lagrange Multipliers does not provide a silver bullet for solving these equations. Most examples one sees in texts and on the web have very simple functions allowing a straightforward analytic solution.

Since the λ_i appear linearly in the first N equations in (2.6), it is possible to obtain solution values for the λ_i as follows. First, write the first N equations of (2.6) in matrix form,

$$- \begin{pmatrix} f_1 \\ f_2 \\ \dots \\ f_N \end{pmatrix} = \begin{pmatrix} a_1 & b_1 & \dots & q_1 \\ a_2 & b_2 & \dots & q_2 \\ \dots & \dots & \dots & \dots \\ a_N & b_N & \dots & q_N \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_{S-1} \end{pmatrix} \tag{2.7}$$

where the matrix has $S-1$ columns (one column for each constraint) and N rows. Since $S \leq N$, one has $S-1 < N$ so there are fewer columns than rows. This matrix is the transpose of the R matrix (1.2) with the f_i elements removed.

Taking only the first $S-1$ rows of the above matrix equation, one finds

$$- \begin{pmatrix} f_1 \\ f_2 \\ \dots \\ f_{S-1} \end{pmatrix} = \begin{pmatrix} a_1 & b_1 & \dots & q_1 \\ a_2 & b_2 & \dots & q_2 \\ \dots & \dots & \dots & \dots \\ a_{S-1} & b_{S-1} & \dots & q_{S-1} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_{S-1} \end{pmatrix} \tag{2.8}$$

where the matrix is now square. Assuming this matrix has non-zero determinant, one can invert (2.8) to obtain solution values of the Lagrange multipliers $\lambda_i(\mathbf{r})$ as functions of $\mathbf{r}$. These λ_i functions can be inserted into the first N equations of (2.6) and then one can proceed to solve those N equations for $\mathbf{r}$.

By making different versions of (2.8) taking different sets of S-1 rows from (2.7), one may produce several different sets of $\{\lambda_i(\mathbf{r})\}$ corresponding to different candidate solutions $\mathbf{r}$.

Comments

1. Some authors use the notation $f_{\mathbf{x}_i} \equiv \partial f / \partial x_i$ and the constraints are $g^{(s)}(\mathbf{x}) = 0$ where s is a label, Then the first derivative of a constraint function is $g^{(s)}_{\mathbf{x}_i}(\mathbf{x})$. Another choice is to refer to the constraint functions as $g_s(\mathbf{x})$ where s is a label, and then one just writes out $\frac{\partial g_s}{\partial x_i}$. We feel that our simple though less rigorous notation provides more clarity. Writing $\partial f / \partial x_i = f_i$ is a notation promoted by Buck.

2. The λ_i terms in (2.1) often appear with minus signs in place of plus signs. This then removes the minus sign in (2.7) and (2.8). These signs are just a convention and have no effect on the solution points $\mathbf{r}$.

3. The function H in (2.1) is sometimes referred to as "the Lagrangian" and is written as L. This Lagrangian differs from that which appears in the Lagrangian formulation of particle motion. In Lagrangian dynamics, the motion of a set of particles (perhaps comprising a rigid body) can be determined from a certain Lagrangian $L \equiv T - V$ where T and V are the kinetic and potential energies of the set of particles. L is a function of the "generalized coordinates" q_k and velocities $\dot{q}_k = dq_k/dt$ of the particles, and of time t, so one writes $L(q_k, \dot{q}_k, t)$. The equations of motion in the absence of constraints are known as the Euler-Lagrange equations, $(\frac{\partial L}{\partial q_k} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_k}) = 0$. One may add a set of differential constraints to the problem of the form $\sum_k a_{ik} dq_k = 0$ where the a_{ik} are functions causing the particles to be constrained in their motion. For example, the particles may be constrained to move only along certain curves like beads on a wire. The adjusted Euler-Lagrange equations then have the form $(\frac{\partial L}{\partial q_k} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_k}) + \sum_i \lambda_i a_{ik} = 0$. This equation is similar to the first line of our equation (2.4). See Goldstein page 41 (2.23) and (2.27). Goldstein refers to this method of incorporating constraints into particle dynamics as the method of Lagrange undetermined multipliers.

3. Proof of Theorem 1

Theorem 1. If the point $\mathbf{r}$ is a solution to the extremal problem (1.1), then $\text{rank}[\mathbf{R}(\mathbf{r})] < S$. (1.8)

(a) Proof for $N = 6$ and $S = 4$

The more general the proof is made, the less clear it becomes due to the notational baggage that must be added. Therefore we shall prove Theorem 1 for the specific case shown above in Fig (1.3) where $N = 6$ and $S = 4$ (3 constraints). The reader should have no trouble following each step for this sample case. When we reach the proof conclusion for the sample case, we start over again doing the general case and compare each step to the steps of this sample case. Although a lot of equations appear below (and a lot of column inches are consumed), everything is really quite simple.

So, for our sample case we start off with this $\mathbf{R}$ matrix, again with bolded row vectors on the right,

$$\mathbf{R} = \begin{array}{cccccc} f_1 & f_2 & f_3 & f_4 & f_5 & f_6 \\ a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \\ b_1 & b_2 & b_3 & b_4 & b_5 & b_6 \\ c_1 & c_2 & c_3 & c_4 & c_5 & c_6 \end{array} = \begin{array}{c} \mathbf{f} \\ \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \end{array} \quad (3.1)$$

We want then to find an extremum of $f(x_1, x_2, x_3, x_4, x_5, x_6)$ subject to these constraints:

$$\begin{aligned} a(x_1, x_2, x_3, x_4, x_5, x_6) &= 0 \\ b(x_1, x_2, x_3, x_4, x_5, x_6) &= 0 \\ c(x_1, x_2, x_3, x_4, x_5, x_6) &= 0. \end{aligned} \quad (3.2)$$

In practice, a person solving this problem might try to eliminate variables. For example, somehow with enough labor one should be able to "use up" the three constraint equations to eliminate three of the six variables x_i . There are of course (6,3) ways to select three variables for elimination.

For starters, imagine that we have used up the 3 constraints to eliminate variables x_1, x_2 and x_3 . Then we write

$$\begin{aligned} a(x_1, x_2, x_3, x_4, x_5, x_6) &= 0 & x_1 &= X^1(x_4, x_5, x_6) \\ b(x_1, x_2, x_3, x_4, x_5, x_6) &= 0 & \Rightarrow & x_2 = X^2(x_4, x_5, x_6) \\ c(x_1, x_2, x_3, x_4, x_5, x_6) &= 0 & & x_3 = X^3(x_4, x_5, x_6) \end{aligned} \quad (3.3)$$

where X^1, X^2 and X^3 are three resulting functions. This is a "theoretical" elimination, one does not actually have to do the process, one need only understand that in principle it could be done and the functions X^i could in principle be found.

Now rewrite the function f and the three constraint functions in this manner, substituting for example the function X^1 for x_1 everywhere it appears,

$$\begin{aligned}
f(X^1(x_4, x_5, x_6), X^2(x_4, x_5, x_6), X^3(x_4, x_5, x_6), x_4, x_5, x_6) &\equiv F(x_4, x_5, x_6) \\
a(X^1(x_4, x_5, x_6), X^2(x_4, x_5, x_6), X^3(x_4, x_5, x_6), x_4, x_5, x_6) &\equiv A(x_4, x_5, x_6) \\
b(X^1(x_4, x_5, x_6), X^2(x_4, x_5, x_6), X^3(x_4, x_5, x_6), x_4, x_5, x_6) &\equiv B(x_4, x_5, x_6) \\
c(X^1(x_4, x_5, x_6), X^2(x_4, x_5, x_6), X^3(x_4, x_5, x_6), x_4, x_5, x_6) &\equiv C(x_4, x_5, x_6) .
\end{aligned} \tag{3.4}$$

In this way we have defined four new functions F,A,B,C each of the 3 variables shown (those that were not eliminated).

Next, compute the first partial derivatives of F using the chain rule. For example

$$\frac{\partial F}{\partial x_4} = \frac{\partial f}{\partial x_1} \frac{\partial X^1}{\partial x_4} + \frac{\partial f}{\partial x_2} \frac{\partial X^2}{\partial x_4} + \frac{\partial f}{\partial x_3} \frac{\partial X^3}{\partial x_4} + \frac{\partial f}{\partial x_4} . \tag{3.5a}$$

In our compact notation where F_i means $\partial F / \partial x_i$ this can be restated as

$$F_4 = f_1 X^1_4 + f_2 X^2_4 + f_3 X^3_4 + f_4 . \tag{3.5b}$$

Doing this for each of the non-eliminated variables one gets,

$$\begin{aligned}
F_4 &= f_1 X^1_4 + f_2 X^2_4 + f_3 X^3_4 + f_4 \\
F_5 &= f_1 X^1_5 + f_2 X^2_5 + f_3 X^3_5 + f_5 \\
F_6 &= f_1 X^1_6 + f_2 X^2_6 + f_3 X^3_6 + f_6 .
\end{aligned} \tag{3.6}$$

Notice that the subscripts on the leftmost three f_i correspond to those of the eliminated variables $i = 1, 2, 3$.

Now, since we are seeking an extremal point of $F(x_4, x_5, x_6)$ with no constraints at all (the three initial constraints have all been absorbed in the process of eliminating three variables), we require that (x_4, x_5, x_6) be a "critical point" for the function F, which means we require that $F_4 = F_5 = F_6 = 0$. Then the above equations become,

$$\begin{aligned}
f_1 X^1_4 + f_2 X^2_4 + f_3 X^3_4 + f_4 &= 0 \\
f_1 X^1_5 + f_2 X^2_5 + f_3 X^3_5 + f_5 &= 0 \\
f_1 X^1_6 + f_2 X^2_6 + f_3 X^3_6 + f_6 &= 0 .
\end{aligned} \tag{3.7}$$

Now apply the *same process* to the three constraint functions A,B,C. First compute the derivatives, and then set the derivatives to zero. One does this for $A(x_4, x_5, x_6)$, for example, because the constraint condition $a = 0$ says that that $A(x_4, x_5, x_6) = 0 = a$ a constant, for all values of x_4, x_5, x_6 , so all first derivatives must vanish. Doing this for function A then gives these three equations,

$$\begin{aligned}
a_1 X^1_4 + a_2 X^2_4 + a_3 X^3_4 + a_4 &= 0 \\
a_1 X^1_5 + a_2 X^2_5 + a_3 X^3_5 + a_5 &= 0 \\
a_1 X^1_6 + a_2 X^2_6 + a_3 X^3_6 + a_6 &= 0 .
\end{aligned} \tag{3.8}$$

We have just taken the previous equations and replaced $f \rightarrow a$ everywhere. Then do this for B and C. One ends up then with this set of 12 equations:

$$\begin{aligned}
f_1X^1_4 + f_2X^2_4 + f_3X^3_4 + f_4 &= 0 \\
f_1X^1_5 + f_2X^2_5 + f_3X^3_5 + f_5 &= 0 \\
f_1X^1_6 + f_2X^2_6 + f_3X^3_6 + f_6 &= 0 \\
\\
a_1X^1_4 + a_2X^2_4 + a_3X^3_4 + a_4 &= 0 \\
a_1X^1_5 + a_2X^2_5 + a_3X^3_5 + a_5 &= 0 \\
a_1X^1_6 + a_2X^2_6 + a_3X^3_6 + a_6 &= 0 \\
\\
b_1X^1_4 + b_2X^2_4 + b_3X^3_4 + b_4 &= 0 \\
b_1X^1_5 + b_2X^2_5 + b_3X^3_5 + b_5 &= 0 \\
b_1X^1_6 + b_2X^2_6 + b_3X^3_6 + b_6 &= 0 \\
\\
c_1X^1_4 + c_2X^2_4 + c_3X^3_4 + c_4 &= 0 \\
c_1X^1_5 + c_2X^2_5 + c_3X^3_5 + c_5 &= 0 \\
c_1X^1_6 + c_2X^2_6 + c_3X^3_6 + c_6 &= 0 .
\end{aligned} \tag{3.9}$$

Here there are $S = 4$ equation groups, and each group has $N-S+1 = 6-4+1 = 3$ equations (this last 3 is the number of non-eliminated variables).

Next *reorder* these equations into three groups of four, for example taking the first equation in each group above to make the first new group below,

$$\begin{aligned}
f_1X^1_4 + f_2X^2_4 + f_3X^3_4 + f_4 &= 0 \\
a_1X^1_4 + a_2X^2_4 + a_3X^3_4 + a_4 &= 0 \\
b_1X^1_4 + b_2X^2_4 + b_3X^3_4 + b_4 &= 0 \\
c_1X^1_4 + c_2X^2_4 + c_3X^3_4 + c_4 &= 0 \\
\\
f_1X^1_5 + f_2X^2_5 + f_3X^3_5 + f_5 &= 0 \\
a_1X^1_5 + a_2X^2_5 + a_3X^3_5 + a_5 &= 0 \\
b_1X^1_5 + b_2X^2_5 + b_3X^3_5 + b_5 &= 0 \\
c_1X^1_5 + c_2X^2_5 + c_3X^3_5 + c_5 &= 0 \\
\\
f_1X^1_6 + f_2X^2_6 + f_3X^3_6 + f_6 &= 0 \\
a_1X^1_6 + a_2X^2_6 + a_3X^3_6 + a_6 &= 0 \\
b_1X^1_6 + b_2X^2_6 + b_3X^3_6 + b_6 &= 0 \\
c_1X^1_6 + c_2X^2_6 + c_3X^3_6 + c_6 &= 0 .
\end{aligned} \tag{3.10}$$

Now there are $N-S+1 = 3$ equation groups, and each group has $S = 4$ equations.

Next, write each of these three equation sets as a matrix equation,

$$\begin{aligned}
\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} &= \begin{pmatrix} f_1 & f_2 & f_3 & f_4 \\ a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \end{pmatrix} \begin{pmatrix} X_4^1 \\ X_4^2 \\ X_4^3 \\ 1 \end{pmatrix} = (f_1, f_2, f_3, f_4) \begin{pmatrix} X_4^1 \\ X_4^2 \\ X_4^3 \\ 1 \end{pmatrix} \\
\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} &= \begin{pmatrix} f_1 & f_2 & f_3 & f_5 \\ a_1 & a_2 & a_3 & a_5 \\ b_1 & b_2 & b_3 & b_5 \\ c_1 & c_2 & c_3 & c_5 \end{pmatrix} \begin{pmatrix} X_5^1 \\ X_5^2 \\ X_5^3 \\ 1 \end{pmatrix} = (f_1, f_2, f_3, f_5) \begin{pmatrix} X_5^1 \\ X_5^2 \\ X_5^3 \\ 1 \end{pmatrix} \\
\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} &= \begin{pmatrix} f_1 & f_2 & f_3 & f_6 \\ a_1 & a_2 & a_3 & a_6 \\ b_1 & b_2 & b_3 & b_6 \\ c_1 & c_2 & c_3 & c_6 \end{pmatrix} \begin{pmatrix} X_6^1 \\ X_6^2 \\ X_6^3 \\ 1 \end{pmatrix} = (f_1, f_2, f_3, f_6) \begin{pmatrix} X_6^1 \\ X_6^2 \\ X_6^3 \\ 1 \end{pmatrix} .
\end{aligned} \tag{3.11}$$

On the right we show our abbreviated notation for a matrix, just showing the first row.

Now we claim that the three matrices shown must have zero determinant! Suppose the first matrix equation had a non-zero determinant. It could then be inverted to give

$$\begin{pmatrix} X_4^1 \\ X_4^2 \\ X_4^3 \\ 1 \end{pmatrix} = (f_1, f_2, f_3, f_4)^{-1} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} . \tag{3.12}$$

But this produces a blatant contradiction that $1 = 0$ (not to mention the other rows), and therefore we must have

$$\begin{aligned}
\det(f_1, f_2, f_3, f_4) &= 0 \\
\det(f_1, f_2, f_3, f_5) &= 0 \\
\det(f_1, f_2, f_3, f_6) &= 0
\end{aligned}$$

or

$$\det(f_1, f_2, f_3, f_m) = 0 \quad m \neq 1, 2, 3 . \tag{3.13}$$

Therefore, we have shown that three of the $S \times S = 4 \times 4$ submatrices of Fig (1.3) vanish. But there are 15 such submatrices, and in order to show that $\text{rank}(R) < S$, we have to show that *all* 15 $S \times S$ determinants vanish.

But that can be demonstrated by starting over several times and each time choosing to eliminate three different variables. Notice in the $\det = 0$ equations above that the first three f_i values correspond to the three eliminated variables x_1, x_2 and x_3 . Had we instead chosen to eliminate x_1, x_2 and x_4 , we would have obtained,

$$\det(f_1, f_2, f_4, f_3) = 0$$

$$\det(f_1, f_2, f_4, f_5) = 0$$

$$\det(f_1, f_2, f_4, f_6) = 0$$

or

$$\det(f_1, f_2, f_4, f_m) = 0 \quad m \neq 1, 2, 4 \quad (3.14)$$

More generally, had we chosen to eliminate variables x_i, x_j and x_k , we would obtain

$$\det(f_i, f_j, f_k, f_m) = 0 \quad m \neq i, j, k \quad (3.15)$$

As i, j, k range over all possible values 1,2,3,4,5,6, we clearly hit all possible 4x4 subdeterminants, and thus we have shown that they all vanish, and therefore $\text{rank}(R) < S$, QED.

Lest it be overlooked, one must keep in mind that $\mathbf{r} = (x_1, x_2, x_3, x_4, x_5, x_6)$ must be a solution to the extremum problem (or at least a critical point as in (3.7)) including meeting all the constraints. And then the conclusion is that, for such a solution point $\mathbf{r}$, one has $\text{rank}[R(\mathbf{r})] < S$.

Having now shown that $\text{rank}[R(\mathbf{r})] < S$ for a solution point $\mathbf{r}$, we know from (1.7) (b) that the S rows of the R matrix must be linearly *dependent* for that $\mathbf{r}$. This means that one can find a set of constants $\lambda_1, \lambda_2, \lambda_3$ such that

$$\mathbf{f}(\mathbf{r}) + \lambda_1 \mathbf{a}(\mathbf{r}) + \lambda_2 \mathbf{b}(\mathbf{r}) + \lambda_3 \mathbf{c}(\mathbf{r}) = 0. \quad (2.5)$$

These constants *are* the Lagrange Multipliers and we see that the fact that they exist follows directly from the fact that $\text{rank}[R(\mathbf{r})] < S$.

To summarize, one theoretically eliminates $N-S+1$ of the N variables to create the functions F, A, B, C of (3.4). One then looks for the normal (unconstrained) critical points of F where all partials vanish. Since the constraints all have the form $A = 0, B = 0, C = 0$, their partials vanish as well. Writing all these equations in matrix form (for various sets of eliminated variables) leads to the conclusion that all the $S \times S$ minors of matrix R vanish so that $\text{rank}(R) < S$.

(b) Proof for general $S \leq N$

To generalize now to general N and S , we restate some of the previous equations (adding a prime to the equation number) using a more generalized notation. The reader will note how even our simple notation becomes rather unpleasant.

Comment: This complexity increases further if one denotes the constraint functions by a notation like $a = g^{(1)}$, $b = g^{(2)}$ $q = g^{(S-1)}$. We avoid the idea of $a = g_1, b = g_2$ etc because we want a subscript to indicate a partial derivative, not a label.

We start off with this R matrix which has N columns and $S \leq N$ rows. The $S-1$ constraint functions are $a, b, \dots, q$. Remember that f_2 means $\partial f / \partial x_2$.

$$\begin{array}{ccccccccc}
 & f_1 & f_2 & f_3 & \dots & f_N & & \mathbf{f} \\
 & a_1 & a_2 & a_3 & \dots & a_N & & \mathbf{a} \\
 R = & b_1 & b_2 & b_3 & \dots & b_N & = & \mathbf{b} \\
 & \dots & & & & & & \dots \\
 & q_1 & q_2 & q_3 & \dots & q_N & & \mathbf{q}
 \end{array} \quad (3.1)'$$

Here are the S-1 constraint equations,

$$\begin{aligned}
 a(x_1, x_2, \dots, x_N) &= 0 & // \text{ S-1 constraint equations} \\
 b(x_1, x_2, \dots, x_N) &= 0 \\
 \dots & \\
 q(x_1, x_2, \dots, x_N) &= 0 .
 \end{aligned} \quad (3.2)'$$

Eliminate the first S-1 variables $x_1, x_2, \dots, x_{S-1}$ using the S-1 constraint equations:

$$\begin{aligned}
 a(x_1, x_2, \dots, x_N) &= 0 & x_1 &= X^1(x_S, x_{S+1} \dots x_N) \\
 b(x_1, x_2, \dots, x_N) &= 0 & \Rightarrow & x_2 = X^2(x_S, x_{S+1} \dots x_N) \\
 \dots & & & \dots \\
 q(x_1, x_2, \dots, x_N) &= 0 & x_{S-1} &= X^{S-1}(x_S, x_{S+1} \dots x_N) .
 \end{aligned} \quad (3.3)'$$

Substitute to get the F,A,B...Q equations (a set now of S equations),

$$\begin{aligned}
 f(X^1(x_S, x_{S+1} \dots x_N), X^2(x_S, x_{S+1} \dots x_N), \dots, X^{S-1}(x_S, x_{S+1} \dots x_N), x_S, x_{S+1} \dots x_N) &\equiv F(x_S, x_{S+1} \dots x_N) \\
 a(X^1(x_S, x_{S+1} \dots x_N), X^2(x_S, x_{S+1} \dots x_N), \dots, X^{S-1}(x_S, x_{S+1} \dots x_N), x_S, x_{S+1} \dots x_N) &\equiv A(x_S, x_{S+1} \dots x_N) \\
 b(X^1(x_S, x_{S+1} \dots x_N), X^2(x_S, x_{S+1} \dots x_N), \dots, X^{S-1}(x_S, x_{S+1} \dots x_N), x_S, x_{S+1} \dots x_N) &\equiv B(x_S, x_{S+1} \dots x_N) \\
 \dots & \\
 q(X^1(x_S, x_{S+1} \dots x_N), X^2(x_S, x_{S+1} \dots x_N), \dots, X^{S-1}(x_S, x_{S+1} \dots x_N), x_S, x_{S+1} \dots x_N) &\equiv Q(x_S, x_{S+1} \dots x_N) .
 \end{aligned} \quad (3.4)'$$

Now take first derivatives as shown in (3.5) to get this generalized version of (3.6),

$$\begin{aligned}
 F_S &= f_1 X^1_S + f_2 X^2_S + f_3 X^3_S + \dots + f_{S-1} X^{S-1}_S + f_S \\
 F_{S+1} &= f_1 X^1_{S+1} + f_2 X^2_{S+1} + f_3 X^3_{S+1} + \dots + f_{S-1} X^{S-1}_{S+1} + f_{S+1} \\
 F_{S+2} &= f_1 X^1_{S+2} + f_2 X^2_{S+2} + f_3 X^3_{S+2} + \dots + f_{S-1} X^{S-1}_{S+2} + f_{S+2} \\
 \dots & \\
 F_N &= f_1 X^1_N + f_2 X^2_N + f_3 X^3_N + \dots + f_{S-1} X^{S-1}_N + f_N . \quad // \text{ N-S+1 equations}
 \end{aligned} \quad (3.6)'$$

Requiring $(x_S, x_{S+1} \dots x_N)$ to be a "critical point", we set all the these first derivatives to 0 to get,

$$\begin{aligned}
f_1 X_s^1 + f_2 X_s^2 + f_3 X_s^3 + \dots + f_{s-1} X_s^{s-1} + f_s &= 0 \\
f_1 X_{s+1}^1 + f_2 X_{s+1}^2 + f_3 X_{s+1}^3 + \dots + f_{s-1} X_{s+1}^{s-1} + f_{s+1} &= 0 \\
f_1 X_{s+2}^1 + f_2 X_{s+2}^2 + f_3 X_{s+2}^3 + \dots + f_{s-1} X_{s+2}^{s-1} + f_{s+2} &= 0 \\
\dots \\
f_1 X_N^1 + f_2 X_N^2 + f_3 X_N^3 + \dots + f_{s-1} X_N^{s-1} + f_N &= 0 \quad // \text{N-S+1 equations} \quad (3.7)'
\end{aligned}$$

A similar set of equations is obtained for a,b,c...q. Just replace f→a, then f→ b and so on. We shall not write out all these equations as we did in (3.9). We have then S sets of equations, each set containing N-S+1 equations. As in the sample case, we then reorder the equations to get first this group,

$$\begin{aligned}
f_1 X_s^1 + f_2 X_s^2 + f_3 X_s^3 + \dots + f_{s-1} X_s^{s-1} + f_s &= 0 \\
a_1 X_s^1 + a_2 X_s^2 + a_3 X_s^3 + \dots + a_{s-1} X_s^{s-1} + a_s &= 0 \\
b_1 X_s^1 + b_2 X_s^2 + b_3 X_s^3 + \dots + b_{s-1} X_s^{s-1} + b_s &= 0 \\
\dots \\
q_1 X_s^1 + q_2 X_s^2 + q_3 X_s^3 + \dots + q_{s-1} X_s^{s-1} + q_s &= 0 \quad // \text{S equations} \quad (3.10)'_s
\end{aligned}$$

The next group has S→S+1 in all subscript positions. The last group has S→N, and here is that last group,

$$\begin{aligned}
f_1 X_N^1 + f_2 X_N^2 + f_3 X_N^3 + \dots + f_{s-1} X_N^{s-1} + f_N &= 0 \\
a_1 X_N^1 + a_2 X_N^2 + a_3 X_N^3 + \dots + a_{s-1} X_N^{s-1} + a_N &= 0 \\
b_1 X_N^1 + b_2 X_N^2 + b_3 X_N^3 + \dots + b_{s-1} X_N^{s-1} + b_N &= 0 \\
\dots \\
q_1 X_N^1 + q_2 X_N^2 + q_3 X_N^3 + \dots + q_{s-1} X_N^{s-1} + q_N &= 0 \quad // \text{S equations} \quad (3.10)'_N
\end{aligned}$$

So there are now N-S+1 groups of equations and each group has S equations.

The next task is to write each set of S equations as a matrix equation. Here we do that just using the abbreviated notation for the matrix,

$$\begin{aligned}
\begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 0 \end{pmatrix} &= (f_1, f_2, f_3, \dots, f_{s-1}, \mathbf{f_s}) \begin{pmatrix} X_s^1 \\ X_s^2 \\ \dots \\ X_s^{s-1} \\ 1 \end{pmatrix} \\
\begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 0 \end{pmatrix} &= (f_1, f_2, f_3, \dots, f_{s-1}, \mathbf{f_{s+1}}) \begin{pmatrix} X_{s+1}^1 \\ X_{s+1}^2 \\ \dots \\ X_{s+1}^{s-1} \\ 1 \end{pmatrix} \\
\dots \\
\begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 0 \end{pmatrix} &= (f_1, f_2, f_3, \dots, f_{s-1}, \mathbf{f_N}) \begin{pmatrix} X_N^1 \\ X_N^2 \\ \dots \\ X_N^{s-1} \\ 1 \end{pmatrix} \quad . \quad (3.11)'
\end{aligned}$$

We then argue as before that, in order to avoid the contradiction $0 = 1$, we must have the determinants of these $S \times S$ submatrices be zero,

$$\begin{aligned} \det(f_1, f_2, f_3, \dots, f_{S-1}, f_S) &= 0 \\ \det(f_1, f_2, f_3, \dots, f_{S-1}, f_{S+1}) &= 0 \\ &\dots \\ \det(f_1, f_2, f_3, \dots, f_{S-1}, f_N) &= 0 \end{aligned}$$

or

$$\det(f_1, f_2, f_3, \dots, f_{S-1}, f_m) = 0 \quad m \neq 1, 2, 3, \dots, S-1 \quad . \quad (3.13)'$$

Finally, if we start off by eliminating the $S-1$ variables $x_i, x_j, x_k, \dots, x_r$, we would find that

$$\det(f_i, f_j, f_k, \dots, f_r, f_m) = 0 \quad m \neq i, j, k, \dots, r \quad . \quad (3.15)'$$

Since this includes any possible $S \times S$ subdeterminant of the R matrix, we have then shown that all (N, S) $S \times S$ subdeterminants vanish, and therefore $\text{rank}[R(\mathbf{r})] < S$ when $\mathbf{r}$ is a solution or critical point of the extremum problem.

Having now shown that $\text{rank}[R(\mathbf{r})] < S$ for a solution point $\mathbf{r}$, we know that the S rows of the R matrix must be linearly *dependent* for that $\mathbf{r}$. This means that one can find a set of constants $\lambda_1, \lambda_2, \lambda_3$ such that

$$\mathbf{f}(\mathbf{r}) + \lambda_1 \mathbf{a}(\mathbf{r}) + \lambda_2 \mathbf{b}(\mathbf{r}) + \dots + \lambda_{S-1} \mathbf{q}(\mathbf{r}) = 0 \quad . \quad (2.5)$$

These constants *are* the Lagrange Multipliers and we see that the fact that they exist follows directly from the fact that $\text{rank}[R(\mathbf{r})] < S$.

4. The Gradient Interpretation: Example 1

Recall that for $\mathbf{r}$ such that $\text{rank}[\mathbf{R}(\mathbf{r})] < S$ we can find Lagrange multipliers λ_i such that

$$\mathbf{f}(\mathbf{r}) + \lambda_1 \mathbf{a}(\mathbf{r}) + \lambda_2 \mathbf{b}(\mathbf{r}) + \dots + \lambda_{S-1} \mathbf{q}(\mathbf{r}) = 0 \quad (2.5)$$

where the bolded letters are the rows of the $\mathbf{R}$ matrix shown in (1.2). In component notation this equation reads

$$f_i(\mathbf{r}) + \lambda_1 a_i(\mathbf{r}) + \lambda_2 b_i(\mathbf{r}) + \dots + \lambda_{S-1} q_i(\mathbf{r}) = 0 \quad i = 1, 2, \dots, N \quad (4.1)$$

where recall $f_i = \partial f / \partial x_i$. With the usual gradient operator ∇ this can be written

$$\nabla f(\mathbf{r}) + \lambda_1 \nabla a(\mathbf{r}) + \lambda_2 \nabla b(\mathbf{r}) + \dots + \lambda_{S-1} \nabla q(\mathbf{r}) = 0 \quad \nabla = (\partial_1, \partial_2, \dots, \partial_N) \quad (4.2)$$

How might one interpret this gradient sum being 0? If we use $N = 2$ and $S = 2$, we can at least obtain some visualization, then we can generalize the conclusion in a logical fashion.

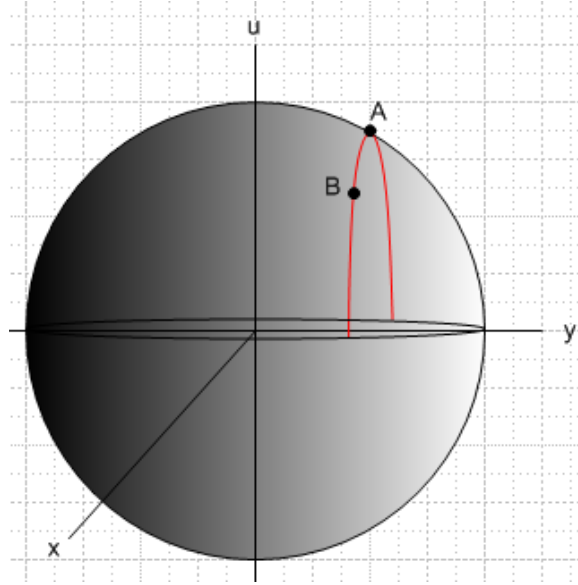
A Simple Example with One Constraint

Let $u = f(x, y)$ represent the surface of a sphere (radius $R = 2$) in E^3 , centered at the origin.

We consider only the upper half of this surface, and we want to find (x, y) that maximizes f . If there are no constraints, then (4.2) above says $\nabla f = 0$. The only place on our $u = f$ surface having $\nabla f = 0$ is the north pole of the sphere. This is just a regular "critical point" where $\partial_x f = 0$ and $\partial_y f = 0$, so we are happy with this interpretation of (4.2) with no constraints.

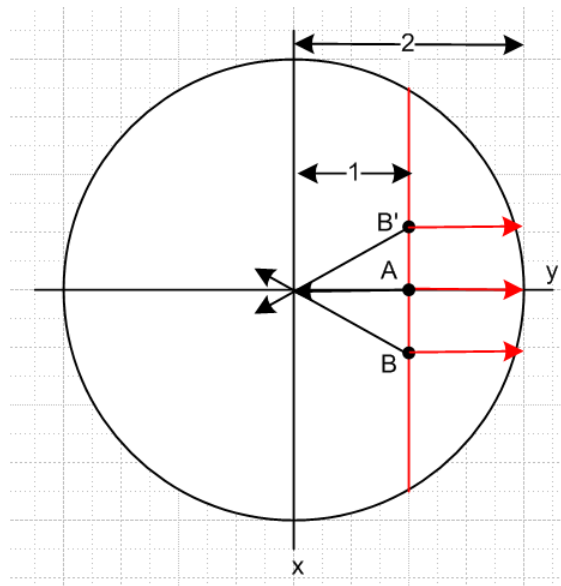
We now add a constraint $a(x, y) = 0$ where $a(x, y) = y - 1$. This constraint is the line $y = 1$ in the x - y plane E^2 . We can extrude this line into a *plane* $y = 1$ in E^3 . The hemispherical surface is a 2D surface in E^3 , and the extruded constraint is also a 2D surface in E^3 . These 2D surfaces intersect in a 1D surface which is a curve. There is hopefully some point on this curve that maximizes f .

Below is a picture of the sphere. The intersection of the upper spherical surface with the plane $y = 1$ is shown as a red curve (a half circle). Due to the constraint, we cannot get to the north pole so the maximum value of f is some value less than that $u = 2$ at the north pole. The extremum of this constrained problem will be at point A, and point B is not an extremum.



(4.3)

Since $N = 2$, the gradients in (4.2) are 2D gradients and so are parallel to the x,y plane. To view these gradients, we draw a top view of the sphere, looking toward the sphere center from the positive u axis:



(4.4)

The three black arrows show ∇f at the points A, B and B'. It seems clear that on the spherical surface ∇f always points toward the north-south axis of the sphere. We can put the gradient arrow tails at the points of interest, but the gradient arrows are all parallel to the x,y plane since for example $\nabla f = (\partial_{\mathbf{x}} f, \partial_{\mathbf{y}} f)$.

Meanwhile, the three red arrows show the direction of ∇a at points A, B and B'. Since $a = y-1$, this direction is of course $\nabla f = (\partial_{\mathbf{x}} a, \partial_{\mathbf{y}} a) = (0,1)$ which points to the right no matter where on the extruded constraint surface the arrow tail is placed.

As point B or B' moves toward point A, the black and red arrows become more aligned and finally at point A (the extremum) they are exactly collinear.

Equation (4.2) says that at the solution point $\mathbf{r} = A$ for the extremum, we have

$$\nabla f(\mathbf{r}=\mathbf{A}) + \lambda_1 \nabla a(\mathbf{r}=\mathbf{A}) = 0 \quad (4.5)$$

and indeed, this equation says that at the solution point the two gradients must be collinear.

More details of the Example 1.

$$f(x,y) = \sqrt{4 - x^2 - y^2} \quad \text{since the sphere has } x^2 + y^2 + u^2 = R^2 = 4$$

$$\nabla f(x,y) = - (x/f) \hat{\mathbf{x}} - (y/f) \hat{\mathbf{y}}$$

$$f(r,\theta) = \sqrt{4 - r^2} \quad \text{polar coordinates}$$

$$\nabla f(r,\theta) = [-r/\sqrt{4-r^2}] \hat{\mathbf{r}} \quad \text{points toward the sphere's vertical axis (and } \nabla f = 0 \text{ at } \mathbf{r} = 0, \text{ north pole)}$$

$$a(x,y) = y-1 = 0 \quad \text{the constraint}$$

$$\nabla a = [1] \hat{\mathbf{y}} \quad \text{points toward the right}$$

At point $\mathbf{A} = (x,y) = (0,1)$ one has $r = 1$, $\hat{\mathbf{r}} = \hat{\mathbf{y}}$ and $f = \sqrt{3}$. Then (4.5) reads

$$[-1/\sqrt{3}] \hat{\mathbf{y}} + \lambda_1 [1/\sqrt{3}] \hat{\mathbf{y}} = 0 \quad \Rightarrow \quad \lambda_1 = 1/\sqrt{3}. \quad (4.6)$$

At the solution point $\mathbf{A}$ the Lagrange multiplier λ_1 can thus be interpreted as the negative of the ratio of the two gradient vectors ∇f and ∇a . Note that at the solution point, the two gradients really are collinear.

At point $\mathbf{B}$, we can consider a differential movement $d\mathbf{r} = |dx|(-\hat{\mathbf{x}})$ along the constraint surface toward point $\mathbf{A}$ (in Fig (4.4) the x axis points down). We find that

$$df(\mathbf{B}) = \nabla f \bullet d\mathbf{r} = [- (x/f) \hat{\mathbf{x}} - (y/f) \hat{\mathbf{y}}] \bullet |dx|(-\hat{\mathbf{x}}) = |dx| (x/f) > 0 \quad \text{since } x > 0 \text{ and } f > 0. \quad (4.7)$$

Thus, as we move toward point $\mathbf{A}$, f increases since $df > 0$, suggesting that $\mathbf{A}$ is indeed a maximum.

Consider the mirror point $\mathbf{B}'$ located above point $\mathbf{A}$ in Fig (4.4). The differential vector $d\mathbf{r} = |dx| \hat{\mathbf{x}}$ lies along the constraint surface and points from $\mathbf{B}'$ toward $\mathbf{A}$. Then

$$df(\mathbf{B}') = \nabla f \bullet d\mathbf{r} = [- (x/f) \hat{\mathbf{x}} - (y/f) \hat{\mathbf{y}}] \bullet |dx| \hat{\mathbf{x}} = |dx| (-x/f) > 0 \quad \text{since } -x > 0 \text{ and } f > 0. \quad (4.8)$$

Once again, $df > 0$ showing indeed that point $\mathbf{A}$ is a maximum.

Finally, we have claimed that at the solution point, the Lagrangian function $H(x,y)$ should have a maximum, since the whole theory is based on H being the function to maximize without constraints. On has,

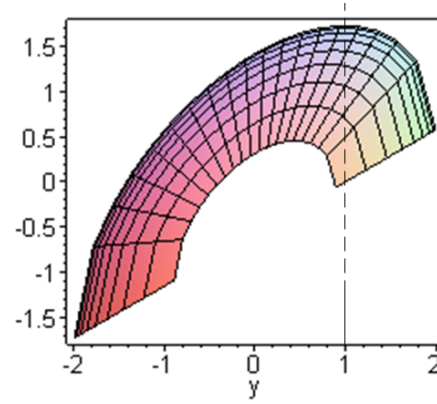
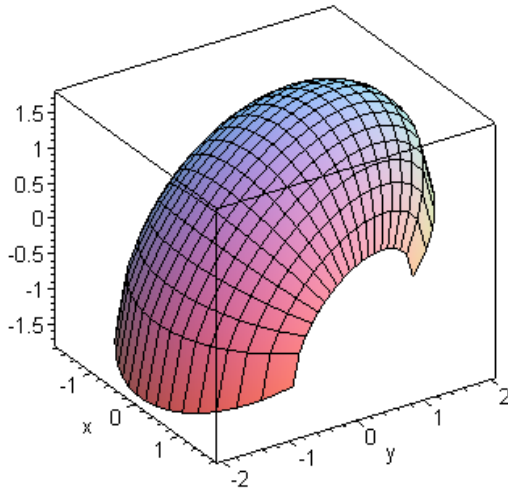
$$H(x,y) = f(\mathbf{r}) + \lambda_1 a(\mathbf{r}) = \sqrt{4 - x^2 - y^2} + (1/\sqrt{3}) (y-1). \quad (4.9)$$

Now $u = (1/\sqrt{3}) (y-1)$ is a plane sloping up to the right in Fig (4.3). This is different from the plane $y = 1$ which is a vertical plane in that figure. In (4.9) we are adding a spherical surface to a plane sloping up to the right, and the result is an ellipsoidal-like surface (really a quartic surface) which has a maximum at the point $A = (0,1)$. Here is a plot of that surface:

```
H := sqrt(4-x^2-y^2) + (1/sqrt(3))*(y-1);
```

$$H = \sqrt{4 - x^2 - y^2} + \frac{1}{3}\sqrt{3} (y-1)$$

```
plot3d(H,x=-2..2,y = -sqrt(4-x^2)+1e-8..sqrt(4-x^2)-1e-8,axes=boxed,grid = [20,20]);
```



(4.10)

A direct method of confirming the maximum is provided by examining H_{ii} :

$$H_i = f_i + \lambda_1 a_i \quad H_{ii} = f_{ii} + \lambda_1 a_{ii} \quad a_{ii} = \partial^2(y-1)/\partial x_i^2 = 0$$

$$f = \sqrt{4 - x^2 - y^2} = \sqrt{4 - x_1^2 - x_2^2} \quad \Rightarrow \quad f_i = -x_i/f \quad f > 0$$

$$f_{ii} = -[f * 1 - x_i f_i]/f^2 = -[f - x_i(-x_i/f)]/f^2 = -(1/f) - (x_i/f)^2$$

$$H_{ii} = f_{ii} + \lambda_1 a_{ii} = f_{ii} = -(1/f) - (x_i/f)^2 < 0 .$$

This shows that the function $H(x,y)$ is "cupping down" at all locations, as the graph suggests. The point A where $H_i = 0$ (see (4.6)) thus also has $H_{ii} < 0$ and is thus a maximum.

If we were to upgrade this example to $u = f(x,y,z)$ and have two constraints $a=0$ and $b=0$, the final intersection path of the hypersphere surface with the two constraint surfaces is again a 1D curve. If we consider $d\mathbf{r}$ along this curve, that $d\mathbf{r}$ will be perpendicular to both ∇a and ∇b , which says

$$(\lambda_1 \nabla a + \lambda_2 \nabla b) \bullet d\mathbf{r} = 0 . \quad (4.11)$$

At an extremum point, if (4.2) is valid, we find that

$$\nabla f \bullet \mathbf{dr} = -(\lambda_1 \nabla a + \lambda_2 \nabla b) \bullet \mathbf{dr} = 0 \quad . \quad (4.12)$$

Since this $\mathbf{dr}$ movement is perpendicular to ∇f , there can be no change in f by moving along the intersecting constraint surface, and that is why we are at an extremum.

So this then provides an interpretation for the general solution case where this equation is true:

$$\nabla f + \lambda_1 \nabla a + \dots + \lambda_{s-1} \nabla q = 0 \quad (4.2)$$

and thus,

$$\nabla f \bullet \mathbf{dr} = -(\lambda_1 \nabla a + \lambda_2 \nabla b + \dots + \lambda_{s-1} \nabla q) \bullet \mathbf{dr} \quad (4.13)$$

In general the intersection surface of f with all the constraint surfaces will be of dimension $N-S+1$ in E^n . A tiny displacement $\mathbf{dr}$ along this surface is perpendicular to *all* the local constraint surface gradients, so the right side of (4.8) is 0. At a solution point where equation (4.2) is true, we then have $\nabla f \bullet \mathbf{dr} = 0$ so a small displacement $\mathbf{dr}$ in any constraint-legal direction results in $df = 0$, hence we are at an extremum.

5. Example 2: Extremal distances between an ellipse and a circle

Many examples of the use of the Method of Lagrange Multipliers can be found in texts and on the web. See for example Trench or the wiki Lagrange Multiplier page.

Our example has 2 constraints so $S = 3$. There are four coordinates so $N = 4$. It happens that these four coordinates are not those of a single point in E^4 but represent two points in E^2 . Our example is a bit more complex than the toy examples in texts and we obtain a numerical solution after carrying out the analytic steps of the Lagrange Multiplier algorithm.

Consider then these two curves in E^2 :

$$\begin{aligned} x^2/A^2 + y^2/B^2 &= 1 && \text{ellipse centered at the origin, semimajor/minor axes A and B} \\ (x'-\alpha)^2 + (y'-\beta)^2 &= R^2 && \text{circle of radius R centered at } (\alpha, \beta) \end{aligned} \quad (5.1)$$

What is the minimum and maximum distance between these two curves?

The distance squared is given by

$$f = (x-x')^2 + (y-y')^2 = f(x, y, x', y') \quad N = 4 \text{ coordinates} \quad (5.2)$$

while the constraint functions are

$$\begin{aligned} a(x, y, x', y') &= x^2/A^2 + y^2/B^2 - 1 && a(x, y, x', y') = 0 \\ b(x, y, x', y') &= (x'-\alpha)^2 + (y'-\beta)^2 - R^2 && b(x, y, x', y') = 0 \end{aligned} \quad (5.3)$$

Notice that the vector $\mathbf{r}$ for this problem is $\mathbf{r} = (x, y, x', y')$.

The various derivatives are,

$$\begin{aligned} f_1 &= \partial_{\mathbf{x}} f = 2(x-x') & a_1 &= \partial_{\mathbf{x}} a = 2x/A^2 & b_1 &= \partial_{\mathbf{x}} b = 0 \\ f_2 &= \partial_{\mathbf{y}} f = 2(y-y') & a_2 &= \partial_{\mathbf{y}} a = 2y/B^2 & b_2 &= \partial_{\mathbf{y}} b = 0 \\ f_3 &= \partial_{\mathbf{x}'} f = -2(x-x') & a_3 &= \partial_{\mathbf{x}'} a = 0 & b_3 &= \partial_{\mathbf{x}'} b = 2(x'-\alpha) \\ f_4 &= \partial_{\mathbf{y}'} f = -2(y-y') & a_4 &= \partial_{\mathbf{y}'} a = 0 & b_4 &= \partial_{\mathbf{y}'} b = 2(y'-\beta) \end{aligned} \quad (5.4)$$

The H function of (2.1) is then,

$$\begin{aligned} H(x, y, x', y', \lambda_1, \lambda_2) &= f + \lambda_1 a + \lambda_2 b \\ &= (x-x')^2 + (y-y')^2 + \lambda_1 [x^2/A^2 + y^2/B^2 - 1] + \lambda_2 [(x'-\alpha)^2 + (y'-\beta)^2 - R^2] \end{aligned} \quad (5.5)$$

Compute the four first derivatives as in (2.2),

$$\begin{aligned}
H_1 &= f_1 + \lambda_1 a_1 + \lambda_2 b_1 = 2(x-x') + \lambda_1 2x/A^2 \\
H_2 &= f_2 + \lambda_1 a_2 + \lambda_2 b_2 = 2(y-y') + \lambda_1 2y/B^2 \\
H_3 &= f_3 + \lambda_1 a_3 + \lambda_2 b_3 = -2(x-x') + \lambda_2 2(x'-\alpha) \\
H_4 &= f_4 + \lambda_1 a_4 + \lambda_2 b_4 = -2(y-y') + \lambda_2 2(y'-\beta) .
\end{aligned} \tag{5.6}$$

Set these derivatives to 0 as in (2.4) to find this set of six equations (2.6)

$$\begin{aligned}
(x-x') + \lambda_1 x/A^2 &= 0 & 1 \\
(y-y') + \lambda_1 y/B^2 &= 0 & 2 \\
-(x-x') + \lambda_2 (x'-\alpha) &= 0 & 3 \\
-(y-y') + \lambda_2 (y'-\beta) &= 0 & 4 \\
x^2/A^2 + y^2/B^2 - 1 &= 0 & 5 \\
(x'-\alpha)^2 + (y'-\beta)^2 - R^2 &= 0 & 6
\end{aligned} \tag{5.7}$$

where the 6 unknowns are $x, y, x', y', \lambda_1, \lambda_2$.

Notice that a possible λ_1 solution is $\lambda_1 = \lambda_2 = 0$ in which case one gets $x = x'$ and $y = y'$. Certainly this is an extremum situation since distance squared is $f = (x-x')^2 + (y-y')^2 = 0$, a minimum. Insertion of $x' = x$ and $y' = y$ into 5 and 6 above and elimination of x yields an equation for y of the form

$$k_1 y^4 + k_2 y^3 + k_3 y^2 + k_4 y + k_5 = 0. \quad x = \pm A \sqrt{1 - (y/B)^2} \tag{5.8}$$

If the ellipse and circle intersect, one will find that the above equation has 1,2,3 or 4 real roots which then represent the intersection point(s) of the two curves. If the curves don't intersect, the solution y values will be complex so there are no physical solutions for $\lambda_1 = \lambda_2 = 0$.

So how does one go about solving the set of 6 equations shown in (5.7)? The possible first step is to find a viable set of λ_i . From a version of (2.8) using the first and third rows of (2.7) one may write,

$$-\begin{pmatrix} f_1 \\ f_3 \end{pmatrix} = \begin{pmatrix} a_1 & b_1 \\ a_3 & b_3 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \equiv M \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \tag{2.8}$$

where

$$M = \begin{pmatrix} a_1 & b_1 \\ a_3 & b_3 \end{pmatrix} = \begin{pmatrix} a_1 & 0 \\ 0 & b_3 \end{pmatrix} \tag{5.9}$$

so that

$$M^{-1} = \begin{pmatrix} b_3 & 0 \\ 0 & a_1 \end{pmatrix} / \det(M) = \frac{1}{a_1 b_3} \begin{pmatrix} b_3 & 0 \\ 0 & a_1 \end{pmatrix} . \tag{5.10}$$

Therefore,

$$-\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = M^{-1} \begin{pmatrix} f_1 \\ f_3 \end{pmatrix} = \frac{1}{a_1 b_3} \begin{pmatrix} b_3 & 0 \\ 0 & a_1 \end{pmatrix} \begin{pmatrix} f_1 \\ f_3 \end{pmatrix} \quad (5.11)$$

and thus

$$-\lambda_1 = \frac{1}{a_1 b_3} b_3 f_1 \quad -\lambda_2 = \frac{1}{a_1 b_3} a_1 f_3 \quad \Rightarrow$$

$$\lambda_1 = -f_1/a_1 = -2(x-x')/[2x/A^2] = -A^2(x-x')/x$$

$$\lambda_2 = -f_3/b_3 = 2(x-x')/[2(x'-\alpha)] = (x-x')/(x'-\alpha)$$

and we then have a candidate solution set $\{\lambda_i\}$. The first four equations of (5.7) become

$$\begin{aligned} (x-x') + [-A^2(x-x')/x] x/A^2 &= 0 & 1 \\ (y-y') + [-A^2(x-x')/x] y/B^2 &= 0 & 2 \\ -(x-x') + [(x-x')/(x'-\alpha)] (x'-\alpha) &= 0 & 3 \\ -(y-y') + 2[(x-x')/(x'-\alpha)] (y'-\beta) &= 0 & 4 \end{aligned}$$

or

$$\begin{aligned} (x-x') - (x-x') &= 0 & 1 \\ (y-y') - (A/B)^2(x-x')y/x &= 0 & 2 \\ -(x-x') + (x-x') &= 0 & 3 \\ -(y-y') + (y'-\beta)(x-x')/(x'-\alpha) &= 0 & 4 \end{aligned} \quad (5.12)$$

Two of these equations are identities, while the other two are

$$\begin{aligned} (y-y') - (A/B)^2(x-x')y/x &= 0 & 2 \\ -(y-y') + (y'-\beta)(x-x')/(x'-\alpha) &= 0 & 4 \end{aligned}$$

or

$$\begin{aligned} (y-y')x - (A/B)^2(x-x')y &= 0 & 2 \\ -(y-y')(x'-\alpha) + (y'-\beta)(x-x') &= 0 & 4 \end{aligned} \quad (5.13)$$

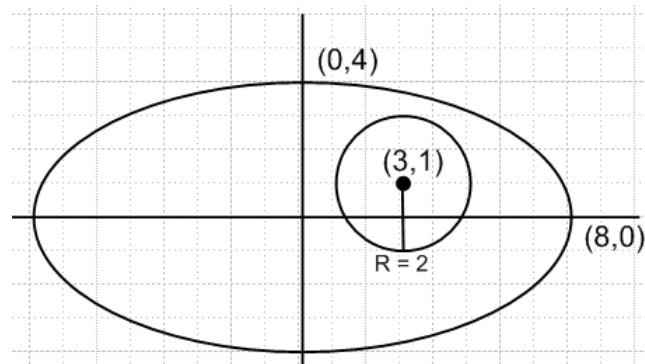
Including the last two equations (5.7), one has four equations in four unknowns x, y, x', y' :

$$\begin{aligned} (y-y')x - (A/B)^2(x-x')y &= 0 & 2 \\ -(y-y')(x'-\alpha) + (y'-\beta)(x-x') &= 0 & 4 \\ x^2/A^2 + y^2/B^2 - 1 &= 0 & 5 \\ (x'-\alpha)^2 + (y'-\beta)^2 - R^2 &= 0 & 6 \end{aligned} \quad (5.14)$$

This is a system of four 2nd degree polynomial equations in four variables. When we ask Maple to analytically solve these equations for the four unknowns x, y, x', y' the results are very messy and involve roots of fourth degree polynomials. In this situation, the analytic solution exists, but is so complicated it

hardly seems very useful. We selected this example just to show that even a relatively simple problem can be nearly intractable analytically.

We turn then to a numerical solution for the particular ellipse + circle case shown in this scaled figure where the background boxes are 1 unit squares,



(5.15)

The circle has radius $R = 2$ and is centered at $(\alpha, \beta) = (3, 1)$. The semi ellipse axes are $A = 8$ and $B = 4$. We enter into Maple the 6 original equations (5.7) in the 6 unknowns $x, y, x', y', \lambda_1, \lambda_2$ ($x_p = x'$, $L1 = \lambda_1$, etc) :

$$\begin{aligned}
 e1 &:= (x - x_p) + L1 * (x / A^2) = 0; \\
 e1 &:= x - x_p + \frac{L1 x}{A^2} = 0 \\
 e2 &:= (y - y_p) + L1 * (y / B^2) = 0; \\
 e2 &:= y - y_p + \frac{L1 y}{B^2} = 0 \\
 e3 &:= -(x - x_p) + L2 * (x_p - \alpha) = 0; \\
 e3 &:= -x + x_p + L2 (x_p - \alpha) = 0 \\
 e4 &:= -(y - y_p) + L2 * (y_p - \beta) = 0; \\
 e4 &:= -y + y_p + L2 (y_p - \beta) = 0 \\
 e5 &:= (x^2 / A^2) + (y^2 / B^2) - 1 = 0; \\
 e5 &:= \frac{x^2}{A^2} + \frac{y^2}{B^2} - 1 = 0 \\
 e6 &:= (x_p - \alpha)^2 + (y_p - \beta)^2 - R^2 = 0; \\
 e6 &:= (x_p - \alpha)^2 + (y_p - \beta)^2 - R^2 = 0
 \end{aligned}
 \tag{5.16}$$

We next set in the specific parameters for the figure shown above

$$\begin{aligned}
 A &:= 8; B := 4; R := 2; \alpha := 3; \beta := 1; \\
 A &:= 8 \\
 B &:= 4 \\
 R &:= 2 \\
 \alpha &:= 3 \\
 \beta &:= 1
 \end{aligned}
 \tag{5.17}$$

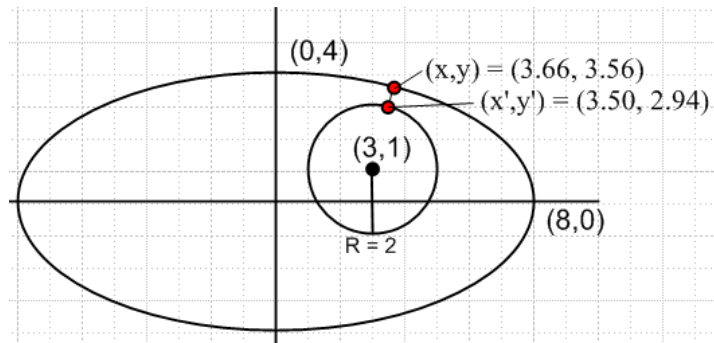
Without given a starting point, Maple finds the following solution

```
fsolve({e1,e2,e3,e4,e5,e6},{x,y,xp,yp,L1,L2});  
{L1 = -2.790619099, L2 = .3203235392, x = 3.657316705, xp = 3.497845176, y = 3.557528444, yp = 2.937046768}
```

which we approximate as

$$\lambda_1 = -2.8 \quad \lambda_2 = .32 \quad (x,y) = (3.66, 3.56) \quad (x',y') = (3.50, 2.94) . \quad (5.18a)$$

Plotting this on our figure,



(5.18b)

we see that this solution corresponds to the minimum distance between the two curves. Maple computes this minimum distance to be .6406470846 units.

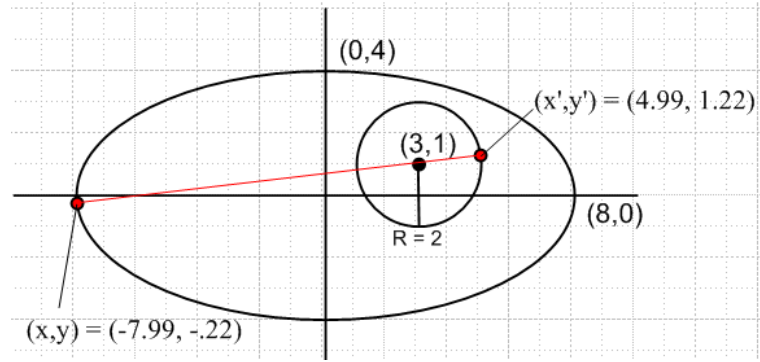
Another solution is found by giving Maple a search starting point $(-8,0)$ on the ellipse and $(5,1)$ on the circle :

```
fsolve({e1,e2,e3,e4,e5,e6},{x=-8,y=0,xp=5,yp=1,L1,L2});  
{x = -7.987657508, y = -.2221078222, L1 = -103.9635963, yp = 1.221087675, xp = 4.987742498, L2 = -6.527706691}
```

which we approximate as

$$\lambda_1 = -104 \quad \lambda_2 = -6.5 \quad (x,y) = (-7.99, -.22) \quad (x',y') = (4.99, 1.22) . \quad (5.19a)$$

Plotting this solution,



(5.19b)

we see that this solution corresponds to the maximum distance between the two curves. Maple computes this maximum distance to be 13.05541338 units.

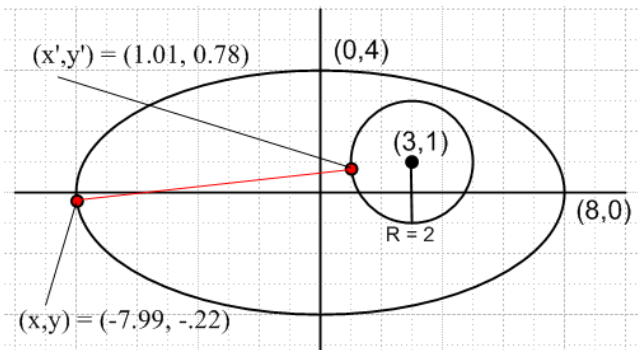
A third candidate solution is found by giving Maple a starting point $(-8,0)$ on the ellipse and $(1,1)$ on the circle:

```
fsolve({e1,e2,e3,e4,e5,e6},{x=-8,y=0,xp=1,yp=1,L1,L2});
{x = -7.987657508,y = -2.221078222,yp = .7789123247,L1 = -72.11057310,L2 = 4.527706691,xp = 1.012257502}
```

which we approximate as

$$\lambda_1 = -72 \quad \lambda_2 = 4.5 \quad (x,y) = (-7.99, -2.22) \quad (x',y') = (1.01, 0.78) . \quad (5.20a)$$

Plotting this solution,



(5.20b)

one sees that this solution represents a "stationary point" but is not in fact a solution to the problem.

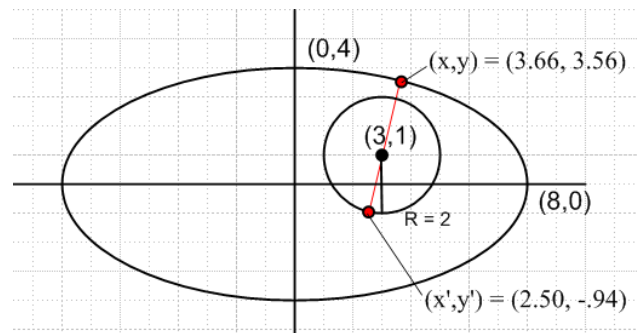
Finally, we find this fourth solution,

```
fsolve({e1,e2,e3,e4,e5,e6},{x=0,y=4,xp=3,yp=-1,L1,L2});
{xp = 2.502154824,yp = -.9370467677,L2 = -2.320323539,L1 = -20.21437201,x = 3.657316705,y = 3.557528444}
```

which we approximate as

$$\lambda_1 = -20.2 \quad \lambda_2 = -2.3 \quad (x,y) = (3.66, 3.56) \quad (x',y') = (2.50, -0.94) . \quad (5.21a)$$

Plotting this solution,



(5.21b)

we find a situation similar to the third candidate solution -- a stationary point which is neither a global maximum nor a global minimum.

6. Example 3: The Boltzmann factor in Statistical Mechanics

6.1 Statement of The Boltzmann Extremum Problem

Imagine a system which has m **state levels** into which particles can be placed. At a given energy level ϵ_i , the number of available **states** is g_i , known as the **degeneracy** at energy ϵ_i . A particular state of the entire system of particles is characterized by the number of particles N_i in each of the states, so system state $= (N_1, N_2, \dots, N_m) = \mathbf{N}$, a vector. The total number of particles in the system is fixed at $M = \sum_i N_i$, and the total energy is $U = \sum_i \epsilon_i N_i$. The system is isolated so its initial energy U does not change, nor can particles be created or destroyed, so U and M are both fixed constants. The system of particles (perhaps a box of gas atoms) is assumed to be at thermal equilibrium so all its macroscopic characteristics (like pressure and temperature) are stable.

Imagine a billion of these particle systems (an "ensemble"). Each system settles into its own partition set $\{N_i\} = \mathbf{N}$. If M is very large, one will find that those billion $\mathbf{N}$ vectors are all very similar. This is so because certain $\{N_i\}$ partitions are statistically favored over other partitions just because there are a lot more "states" available to a system with those favored $\{N_i\}$ partitions.

So, how many states are available ("accessible") to such a system characterized by vector $\mathbf{N}$ which represents some partitioning of the particles $\{N_i\}$?

The number of different "ways" of placing N indistinguishable balls into g boxes is a fascinating elementary problem and the answer is $\binom{N+g-1}{N}$, a binomial coefficient.

Footnote: One arrives at this answer by thinking of N balls and $g-1$ "partitions" between boxes, all of which have to be laid out in a row, such as $| * * * | * | | *$ for the case $N = 5$ and $g-1 = 4$. Here the leftmost bin is empty, as is the second bin from the right. Each layout of the 9 objects is characterized by picking a committee of 5 balls from the 9 positions (or a committee of 4 partitions). So in this case the number of unique layouts is $\binom{9}{5} = \frac{9!}{5!4!} = \binom{9}{4}$. (6.1.1)

The count $\binom{N+g-1}{N}$ allows multiple balls to be in any given box. This count is then associated with so-called **Bose-Einstein statistics** -- multiple bosons are allowed in the same state (for example, photons or any integral-spin particles). If at most one ball is allowed in any given box, the count of ways is just the number of ways to pick a committee of N boxes from the set of g boxes, and this count is $\binom{g}{N}$. This count is then associated with so-called **Fermi-Dirac statistics** -- only one fermion is allowed in a state (for example electrons or any half-integral spin particles). The fact that electrons are fermions is the reason the periodic table of elements exists.

However, we shall assume that $g_i \gg N_i$ (known as the Boltzmann limit), and this allows a simplification of the state count. Both the Bose and Fermi state counts reduce to the same count in this limit:

$$\text{ways} = \binom{N+g-1}{N} \approx \binom{g}{N} = \frac{g!}{N! (g-N)!} = \frac{g(g-1)\dots(g-N+1)}{N!} \approx \frac{g^N}{N!} . \quad (6.1.2)$$

So in our particle problem, the number of ways of having N_1 particles in states with energy ϵ_1 is $\frac{g_1^N}{N_1!}$.

For each of these ways, there are $\frac{g_2^N}{N_2!}$ ways of putting N_2 particles in energy level ϵ_2 . And so on. Thus, the total number of available states ("microstates") for a partition $\{N_i\}$ of the M particles is given by

$$\Omega(N_1, N_2, \dots, N_m) = \frac{g_1^{N_1}}{N_1!} \frac{g_2^{N_2}}{N_2!} \dots \frac{g_m^{N_m}}{N_m!} . \quad (6.1.3)$$

So here is our initial **extremum problem**:

$$\begin{aligned} \text{Find } \mathbf{N} \text{ which maximizes } \Omega(\mathbf{N}) &= \frac{g_1^{N_1}}{N_1!} \frac{g_2^{N_2}}{N_2!} \dots \frac{g_m^{N_m}}{N_m!} \text{ subject to these two constraints:} \\ \sum_i N_i &= M \quad \text{and} \quad \sum_i \epsilon_i N_i = U . \end{aligned} \quad (6.1.4)$$

When M is a very large number (like Avogadro's number), it turns out that $\Omega(\mathbf{N})$ has a very strong and sharp maximum at a certain $\mathbf{N}$ which is the solution value $\mathbf{N}$ for this extremum problem. Thus, when one examines an ensemble of such systems, this solution $\mathbf{N}$ is the overwhelmingly likely partitioning of the particles into $\{N_i\}$. Our task is then to solve this problem for $\mathbf{N}$.

This vector $\mathbf{N} = (N_1, N_2, \dots, N_m)$ is the vector $\mathbf{r} = (x_1, x_2, \dots, x_N)$ of our general Lagrange Multiplier presentation in Section 2, so in this application the number of components in the vector is $N = m$, and we don't want to confuse this previous use of symbol N with $\mathbf{N}$ or N_i of the present problem!

One might observe that the x_i in the general analysis were reals, whereas the N_i here are integers. We can dispense with this issue by simply writing $N_i! = \Gamma(N_i+1)$ which then serves to interpolate N_i between its integer values. Another approach is to replace N_i everywhere in the above problem statement by $n_i \equiv N_i/M$. Since M and the solution N_i values are very large integers, one can regard this n_i as essentially a continuous real (albeit rational) variable.

It should be noted that the value of U is restricted to a certain range.

$$M\epsilon_{\min} \leq U \leq M\epsilon_{\max} . \quad (6.1.5)$$

The limits represent the two extreme possible partitions $\{N_i\}$ of the particles (all in lowest energy state, or all in highest energy state). It is useful to define the average energy of a particle,

$$u \equiv U/M = \text{average energy of particle in the system} \quad (6.1.6)$$

and then (6.1.5) says,

$$\epsilon_{\min} \leq u \leq \epsilon_{\max} . \quad (6.1.7)$$

6.2 Solution of The Boltzmann Extremum Problem

Maximizing the state count Ω of (6.2) is the same as maximizing $f \equiv \ln \Omega$, and from (6.1.3),

$$f(N_1, N_2 \dots N_m) \equiv \ln \Omega(N_1, N_2 \dots N_m) = \sum_i [N_i \ln g_i] - \sum_i [\ln N_i!] \quad (6.2.1)$$

where $\sum_i$ means $\sum_{i=1}^m$. Assuming all the N_i are very large numbers, we may use Stirling's formula,

$$x! \approx \sqrt{2\pi x} x^x e^{-x} \Rightarrow \ln x! \approx \ln \sqrt{2\pi} + (1/2) \ln x + x \ln x - x \approx x \ln x - x, \quad x \gg 1. \quad (6.2.2)$$

Then $\ln N_i! \approx N_i \ln N_i - N_i$ so that,

$$\begin{aligned} f = \ln \Omega &= \sum_i [N_i \ln g_i] - \sum_i [\ln N_i!] \approx \sum_i [N_i \ln g_i] - \sum_i [N_i \ln N_i - N_i] \\ &= \sum_i [(1 + \ln g_i) N_i - N_i \ln N_i] = \sum_i N_i [1 + \ln(g_i/N_i)]. \end{aligned} \quad (6.2.3)$$

Here then is a restatement of the extremum problem given in (6.1.4):

Find $\mathbf{N}$ which maximizes $f(\mathbf{N}) = \sum_i [(1 + \ln g_i) N_i - N_i \ln N_i]$ subject to these two constraints:

$$\begin{aligned} a(\mathbf{N}) &= \sum_i N_i - M = 0 \\ b(\mathbf{N}) &= \sum_i \epsilon_i N_i - U = 0. \end{aligned} \quad (6.2.4)$$

Following now the prescription of Section 2, we write our "Lagrangian" H as

$$\begin{aligned} H(\mathbf{N}, \boldsymbol{\lambda}) &\equiv f(\mathbf{N}) + \lambda_1 a(\mathbf{N}) + \lambda_2 b(\mathbf{N}) \\ &= \sum_i [(1 + \ln g_i) N_i - N_i \ln N_i] + \lambda_1 [\sum_i N_i - M] + \lambda_2 [\sum_i \epsilon_i N_i - U]. \end{aligned} \quad (6.2.5)$$

The next step (b) is to compute the m derivatives with respect to N_i :

$$\begin{aligned} f_i &= \partial f / \partial N_i = (1 + \ln g_i) - N_i (1/N_i) - \ln N_i = \ln g_i - \ln N_i = \ln(g_i/N_i) \\ a_i &= \partial a / \partial N_i = \partial_i [\sum_j N_j - M] = 1 \\ b_i &= \partial b / \partial N_i = \partial_i [\sum_j \epsilon_j N_j - U] = \epsilon_i. \end{aligned} \quad (6.2.6)$$

In step (c) we set the partials of H to zero, so

$$\begin{aligned} 0 &= H_i(\mathbf{N}, \boldsymbol{\lambda}) = f_i + \lambda_1 a_i + \lambda_2 b_i = \ln(g_i/N_i) + \lambda_1 + \lambda_2 \epsilon_i \quad i = 1, 2, \dots, m \\ \text{or} \quad \ln(g_i/N_i) + \lambda_1 + \lambda_2 \epsilon_i &= 0 \quad i = 1, 2, \dots, m \\ \text{or} \end{aligned}$$

$$\ln(N_i/g_i) = \lambda_1 + \lambda_2 \varepsilon_i \quad i = 1, 2, \dots, m \quad (6.2.7)$$

Step (d) just displays the set of equations one needs to solve :

$$\begin{aligned} \ln(N_i/g_i) &= \lambda_1 + \lambda_2 \varepsilon_i & i &= 1, 2, \dots, m \\ \sum_i N_i &= M & // \quad a(\mathbf{N}) &= 0 \\ \sum_i \varepsilon_i N_i &= U & // \quad b(\mathbf{N}) &= 0 \end{aligned} \quad (6.2.8)$$

This is a system of $m+2$ equations in $m+2$ unknowns which are the N_i , λ_1 and λ_2 . To solve these equations, we first solve the first set of m equations for N_i , leaving λ_1 and λ_2 as unknowns:

$$\begin{aligned} \ln(N_i/g_i) &= \lambda_1 + \lambda_2 \varepsilon_i \\ \text{or} \\ N_i &= g_i e^{\lambda_1} e^{\lambda_2 \varepsilon_i} \end{aligned} \quad (6.2.9)$$

Already a **major result** has emerged from our Lagrange Multiplier analysis! The population N_i of states at energy level ε_i depends *exponentially* on ε_i . We intuitively expect higher energy levels to be less populated than lower ones, so we expect λ_2 to be negative. For this reason, and following tradition, we set $\lambda_2 = -\beta$ where $\beta > 0$. And while we're at it, we can set $e^{\lambda_1} = A$ to simplify notation. So

$$\begin{aligned} \beta &\equiv -\lambda_2 \\ A &\equiv e^{\lambda_1} \end{aligned} \quad (6.2.10)$$

and then (6.2.9) reads,

$$N_i = A g_i e^{-\beta \varepsilon_i} \quad (6.2.11)$$

A study of the statistical mechanics of thermal equilibrium in fact shows that

$$\beta = 1/kT \quad (6.2.12)$$

where T is **absolute temperature** and k is **Boltzmann's constant**. The factor $e^{-\beta \varepsilon_i}$ in (6.2.11) is referred to as **the Boltzmann factor**. The quantity $S = k \ln \Omega = k f$ is the **entropy** of a system. The idea that $\beta = 1/kT$ is really the *definition* of absolute temperature, but all these matters are not of immediate interest, so we continue onward.

From (6.2.11) the probability of a particle having energy ε_i is given by

$$p_i = \frac{N_i}{M} = \frac{A}{M} g_i e^{-\beta \varepsilon_i} \quad (6.2.13)$$

Since $\sum_i p_i = 1$ one finds that $1 = \frac{A}{M} [\sum_i g_i e^{-\beta \epsilon_i}]$ which then determines the derived Lagrange multiplier constant A (expressed in terms of β),

$$A = \frac{M}{\sum_i g_i e^{-\beta \epsilon_i}} . \quad (6.2.14)$$

The denominator $Z \equiv \sum_i g_i e^{-\beta \epsilon_i}$ is called the **partition function** for the system.

Using (6.2.11) for N_i , the two constraint equations shown in (6.2.4) [$\sum_i N_i = M$ and $\sum_i N_i \epsilon_i = U$] read

$$\begin{aligned} A \sum_i g_i e^{-\beta \epsilon_i} &= M \\ A \sum_i \epsilon_i g_i e^{-\beta \epsilon_i} &= U . \end{aligned} \quad (6.2.15)$$

Dividing these two equations and using (6.1.6) that $u = U/M$, one gets this self-consistent result,

$$u = \frac{U}{M} = \frac{\sum_i g_i \epsilon_i e^{-\beta \epsilon_i}}{\sum_i g_i e^{-\beta \epsilon_i}} = \frac{(A/M) \sum_i g_i \epsilon_i e^{-\beta \epsilon_i}}{(A/M) \sum_i g_i e^{-\beta \epsilon_i}} = \frac{\sum_i p_i \epsilon_i}{1} = \sum_i p_i \epsilon_i = \langle \epsilon_i \rangle . \quad (6.2.16)$$

To find the derived Lagrange multiplier constant $\beta = -\lambda_2$, write out the second equation of (6.2.15),

$$\sum_i g_i \epsilon_i e^{-\beta \epsilon_i} = U/A = (U/M)(M/A) = (u) (\sum_i g_i e^{-\beta \epsilon_i})$$

or

$$\sum_i [\epsilon_i - u] g_i e^{-\beta \epsilon_i} = 0 \quad (6.2.17)$$

so this is the equation one must solve for β . Recall that ϵ_i , g_i , and u are all known quantities. For the lowest energy states one will have $\epsilon_i < u$ (recall that $u = \langle \epsilon_j \rangle$) so $[\epsilon_i - u]$ will be negative. For high energy states $[\epsilon_i - u]$ will be positive, so that is why (6.2.17) at least stands a chance of having a solution for β . As noted in (6.2.12), the solution for β indicates the temperature of the system of particles. Changing u changes the solution β , so one concludes that the mean particle energy u is a function of temperature.

For general values of ϵ_i (6.2.17) is a transcendental equation which must be solved numerically for β . If it happens that all the ϵ_i are ratios of integers, the equation can be written as a polynomial equation, but since those integers are likely to be large, the order of this polynomial is large and again one must do a numerical solution.

Notice that the solution to our particle system problem has the same general form regardless of the number m of energy levels ϵ_i . If we take m very large and make the energy levels ϵ_i be very closely spaced, they approach a continuum of energy values and the conclusions apply to a classical system. For small finite m , the discrete energy levels indicate a quantum mechanical system.

To summarize, we have used the method of Lagrange multipliers to solve the Boltzmann extremum problem stated in either (6.1.4) or (6.2.4). Since the problem has two constraints, there are two Lagrange

multipliers λ_1 and λ_2 which we replaced with the numbers $A \equiv e^{\lambda_1}$ and $\beta \equiv -\lambda_2$. We showed how β is determined by numerically solving (6.2.17), and then A is determined by (6.2.14). Then the solution (extremum) vector $\mathbf{N} = \{N_i\}$ is given by (6.2.11) which says $N_i = A g_i e^{-\beta \epsilon_i}$.

What we have not shown is that this extremum solution is a maximum. nor have we shown the fact that $\Omega(\mathbf{N})$ has a very sharp peak at the solution value $\mathbf{N}$, which then statistically forces a system to assume a value $\mathbf{N}$ very close to this solution $\mathbf{N}$. We shall deal with these issues in the following section.

6.3 More details of the solution

Start with

$$f = \ln \Omega = \sum_i [(1 + \ln g_i) N_i - N_i \ln N_i] = \sum_i N_i [1 + \ln(g_i/N_i)] . \quad (6.2.3)$$

The constraints $\sum_i N_i = M$ and $\sum_i N_i \epsilon_i = U$ can be regarded as two equation in the two unknowns N_1 and N_2 which are easily solved to obtain,

$$\begin{aligned} N_1 &= + (\epsilon_2 - \epsilon_1)^{-1} [\epsilon_2 M - U + \sum_{i=3}^m (\epsilon_i - \epsilon_2) N_i] = N_1(N_3, \dots, N_m) \\ N_2 &= - (\epsilon_2 - \epsilon_1)^{-1} [\epsilon_1 M - U + \sum_{i=3}^m (\epsilon_i - \epsilon_1) N_i] = N_2(N_3, \dots, N_m) . \end{aligned} \quad (6.3.1)$$

Notice that, for $i = 3, 4, \dots, m$,

$$\begin{aligned} \frac{\partial N_1}{\partial N_i} &= + (\epsilon_2 - \epsilon_1)^{-1} [(\epsilon_i - \epsilon_2)] = \frac{\epsilon_i - \epsilon_2}{\epsilon_2 - \epsilon_1} \\ \frac{\partial N_2}{\partial N_i} &= - (\epsilon_2 - \epsilon_1)^{-1} [(\epsilon_i - \epsilon_1)] = - \frac{\epsilon_i - \epsilon_1}{\epsilon_2 - \epsilon_1} . \end{aligned} \quad (6.3.2)$$

Now think of f as function of $N_3, \dots, N_m$ (analogous to the first line of (3.4)),

$$f(N_1(N_3, \dots, N_m), N_2(N_3, \dots, N_m), N_3, \dots, N_m) . \quad (6.3.3)$$

Compute the total derivative df/dN_i for $i = 3, 4, \dots, n$:

$$\begin{aligned} \frac{df}{dN_i} &= f_1 \frac{\partial N_1}{\partial N_i} + f_2 \frac{\partial N_2}{\partial N_i} + f_i \quad \text{where } f_i \equiv \frac{\partial f}{\partial N_i} \\ &= f_1 \frac{\epsilon_i - \epsilon_2}{\epsilon_2 - \epsilon_1} - f_2 \frac{\epsilon_i - \epsilon_1}{\epsilon_2 - \epsilon_1} + f_i \end{aligned} \quad (6.3.2)$$

$$= \ln(g_1/N_1) \frac{\epsilon_i - \epsilon_2}{\epsilon_2 - \epsilon_1} - \ln(g_2/N_2) \frac{\epsilon_i - \epsilon_1}{\epsilon_2 - \epsilon_1} + \ln(g_i/N_i) . \quad (6.2.6) \quad (6.3.4)$$

Installing the solution values $\ln(N_i/g_i) = \lambda_1 + \lambda_2 \epsilon_i$ shown in (6.2.8) we continue,

$$\frac{df}{dN_i} = (\lambda_1 + \lambda_2 \epsilon_1) \frac{\epsilon_i - \epsilon_2}{\epsilon_2 - \epsilon_1} - (\lambda_1 + \lambda_2 \epsilon_2) \frac{\epsilon_i - \epsilon_1}{\epsilon_2 - \epsilon_1} + (\lambda_1 + \lambda_2 \epsilon_i)$$

$$\begin{aligned}
&= \lambda_1 \left[\frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} - \frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} + 1 \right] + \lambda_2 \left[\varepsilon_1 \frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} - \varepsilon_2 \frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} + \varepsilon_i \right] \\
&= \lambda_1 \left[\frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} - \frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} + \frac{\varepsilon_2 - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} \right] + \lambda_2 \left[\varepsilon_1 \frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} - \varepsilon_2 \frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} + \varepsilon_i \frac{\varepsilon_2 - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} \right] \\
&= \lambda_1 [0] + \lambda_2 [0] = 0 .
\end{aligned} \tag{6.3.5}$$

This is the expected result since the solution is supposed to be a critical point of f given in (6.3.3).

Our main interest is in the second derivative d^2f/dN_i^2 . If this comes out negative at the solution point, we know we have a maximum. Applying d/dN_i to the second line in (6.3.4),

$$\begin{aligned}
\frac{d^2f}{d^2N_i} &= \frac{df_1}{dN_i} \frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} - \frac{df_2}{dN_i} \frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} + \frac{df_i}{dN_i} \\
&= \left[f_{11} \frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} - f_{12} \frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} + f_{1i} \right] \frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} \\
&\quad - \left[f_{21} \frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} - f_{22} \frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} + f_{2i} \right] \frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} \\
&\quad + \left[f_{i1} \frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} - f_{i2} \frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} + f_{ii} \right] .
\end{aligned} \quad f_{ij} \equiv \frac{\partial^2 f}{\partial N_i \partial N_j} \quad i = 3, 4, \dots, m \tag{6.3.6}$$

Recall now that

$$f = \ln \Omega = \sum_i [(1 + \ln g_i) N_i - N_i \ln N_i] \tag{6.2.3}$$

$$f_i = \ln g_i - \ln N_i \tag{6.2.6}$$

$$f_{ij} = -\delta_{ij} (1/N_i) . \tag{6.3.7}$$

Therefore (6.3.6) simplifies to

$$\begin{aligned}
\frac{d^2f}{d^2N_i} &= f_{11} \left[\frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} \right]^2 + f_{22} \left[\frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} \right]^2 + f_{ii} \\
&= - \left\{ \frac{1}{N_1} \left[\frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} \right]^2 + \frac{1}{N_2} \left[\frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} \right]^2 + \frac{1}{N_i} \right\} .
\end{aligned} \tag{6.3.8}$$

This says that the curvature of f in all independent directions N_i for $i = 3, 4, \dots, m$ is negative and this is true for *all* vectors $\mathbf{N} = \{N_i\}$. This is of course then true for the solution vector $\mathbf{N}$. But for the solution vector $\mathbf{N}$ we also have $df/dN_i = 0$ from (6.3.5) and therefore the solution is a *maximum* of f and therefore of Ω .

Finally, we may *estimate* the width of the peak of f and then of Ω . The energy fractions in (6.3.8) are on the ballpark order of unity and the N_i are on the order of M , the total number of particles. If we do a Taylor expansion of $f(\mathbf{N})$ in the i direction only ($i = 3, 4, \dots, m$) about the solution $\mathbf{N}$, then

$$f(N_i + dN_i) \approx f(N_i) + (df/dN_i)dN_i + (1/2)(d^2f/d^2N_i)(dN_i)^2 . \quad (6.3.9)$$

But at the solution point $(df/dN_i) = 0$ so one finds

$$df = f(N_i + dN_i) - f(N_i) \approx (1/2)(d^2f/d^2N_i)(dN_i)^2 \quad (6.3.10)$$

so

$$(dN_i)^2 \approx \frac{2|df|}{|d^2f/d^2N_i|} . \quad (6.3.11)$$

Suppose we are interested in finding dN_i such that the $\Omega(\mathbf{N})$ drops to half its peak value. Then

$$\begin{aligned} \Omega_{\text{half}}/\Omega_{\text{peak}} &= 1/2 & \Rightarrow & \ln \Omega_{\text{half}} - \ln \Omega_{\text{peak}} = \ln(1/2) & \Rightarrow \\ f_{\text{half}} - f_{\text{peak}} &= \ln(1/2) & \Rightarrow & |df| = f_{\text{peak}} - f_{\text{half}} = \ln(2) = 0.7 . \end{aligned} \quad (6.3.12)$$

In this case, the half-width (6.3.11) of the Ω plot is

$$dN_i = \sqrt{\frac{1.4}{|d^2f/d^2N_i|}} . \quad (6.3.13)$$

For a ballpark estimate, we can interpret (6.3.8) as saying

$$\frac{d^2f}{d^2N_i} \sim -3/M \quad \Rightarrow \quad dN_i \approx \sqrt{\frac{1.4 M}{3}} \sim \sqrt{M} \quad (6.3.14)$$

so the fractional half-width is roughly.

$$\frac{dN_i}{M} \sim \sqrt{\frac{1}{M}} . \quad (6.3.15)$$

For a macroscopic box of gas particles, one might have $M = 10^{20}$ and then $\frac{dN_i}{M} \sim 10^{-10}$. Since this is the fractional half-width of the peak in Ω , the peak is *extremely* sharp. This then is why, in an ensemble of macroscopic particle systems, all systems will have almost the same $\mathbf{N}$ value.

6.4 A numerical example with plots

For this example we choose three energy levels ($m = 3$) so the state count is given by

$$\Omega(N_1, N_2, N_3) = \frac{g_1^{N_1}}{N_1!} \frac{g_2^{N_2}}{N_2!} \frac{g_3^{N_3}}{N_3!} \quad (6.1.3) \quad (6.4.1)$$

and correspondingly (this uses the Stirling approximation),

$$f = \ln \Omega = \sum_{i=1}^3 N_i [1 + \ln(g_i/N_i)] . \quad (6.2.3) \quad (6.4.2)$$

Assume the three energy levels are ordered in this manner, where ε_1 is the lowest energy state,

$$\varepsilon_1 < \varepsilon_2 < \varepsilon_3 . \quad \text{and recall} \quad u = U/M \quad \text{and} \quad \varepsilon_1 \leq u \leq \varepsilon_3 . \quad (6.4.3)$$

We regard both N_1 and N_2 as functions of N_3 where, according to (6.3.1),

$$\begin{aligned} N_1 &= +(\varepsilon_2 - \varepsilon_1)^{-1} [\varepsilon_2 M - U + (\varepsilon_3 - \varepsilon_2)N_3] = N_1(N_3) \\ N_2 &= -(\varepsilon_2 - \varepsilon_1)^{-1} [\varepsilon_1 M - U + (\varepsilon_3 - \varepsilon_1)N_3] = N_2(N_3) . \end{aligned} \quad (6.3.1) \quad (6.4.4)$$

It seems clear that one must have

$$\begin{aligned} 0 &\leq N_1 \leq M \\ 0 &\leq N_2 \leq M . \end{aligned} \quad (6.4.5)$$

Using the expressions (6.4.4) in these inequalities and making use of (6.4.3) and the fact that $\varepsilon_1 \leq u \leq \varepsilon_3$, one finds that all four inequalities can be summarized as just two inequalities which we write as,

$$\begin{aligned} N_{3\min} &\leq N_3 \leq N_{3\max} \\ N_{3\min} &= \max\left(\frac{u - \varepsilon_2}{\varepsilon_3 - \varepsilon_2} M, 0\right) \\ N_{3\max} &= \min\left(\frac{u - \varepsilon_1}{\varepsilon_3 - \varepsilon_1} M, M\right) = \frac{u - \varepsilon_1}{\varepsilon_3 - \varepsilon_1} M . \quad // \text{ since } u \leq \varepsilon_3 \end{aligned} \quad (6.4.6)$$

To find the Lagrange multiplier β we must solve (6.2.17),

$$\sum_{i=1}^3 [\varepsilon_i - u] g_i e^{-\beta \varepsilon_i} = 0 . \quad (6.2.17) \quad (6.4.7)$$

Then the other derived multiplier is given by (6.2.14)

$$A = \frac{M}{\sum_i g_i e^{-\beta \varepsilon_i}} = \frac{M}{Z} \quad \text{where} \quad Z \equiv \sum_i g_i e^{-\beta \varepsilon_i} . \quad (6.2.14) \quad (6.4.8)$$

The components N_i of the solution vector are then given by

$$N_i = A g_i e^{-\beta \epsilon_i} . \quad (6.2.11) \quad (6.4.9)$$

The curvature of f at its peak is,

$$\frac{d^2 f}{d^2 N_3} = - \left\{ \frac{1}{N_1} \left[\frac{\epsilon_3 - \epsilon_2}{\epsilon_2 - \epsilon_1} \right]^2 + \frac{1}{N_2} \left[\frac{\epsilon_3 - \epsilon_1}{\epsilon_2 - \epsilon_1} \right]^2 + \frac{1}{N_3} \right\} . \quad (6.3.8) \quad (6.4.10)$$

The half-width of the Ω peak is then given by

$$dN_i = \sqrt{\frac{1.4}{|d^2 f / d^2 N_i|}} . \quad (6.3.13) \quad (6.4.11)$$

For our specific Maple plots, we shall assume

$$\begin{array}{ll} \epsilon_3 = 3 & M = 1000 = \text{number of particles} \\ \epsilon_2 = 2 & u = 1.5 = \text{average energy} \\ \epsilon_1 = 1 & g_1 = g_2 = g_3 = 5000 = \text{degeneracy} \end{array} \quad (6.4.12)$$

Here then is our Maple code. We first enter Ω ,

```
restart; with(plots):
Omega := product(g[i]^N[i]/(N[i]!), i=1..3);
```

$$\Omega := \frac{g_1^{N_1} g_2^{N_2} g_3^{N_3}}{N_1! N_2! N_3!}$$

The quantities N_1 and N_2 are then replaced as in (6.4.4) [we use e_i in place of ϵ_i],

```
N[1] := (e[2]*M-U+(e[3]-e[2])*N[3])/(e[2]-e[1]);
```

$$N_1 := \frac{e_2 M - U + (e_3 - e_2) N_3}{e_2 - e_1}$$

```
N[2] := -(e[1]*M-U+(e[3]-e[1])*N[3])/(e[2]-e[1]);
```

$$N_2 := -\frac{e_1 M - U + (e_3 - e_1) N_3}{e_2 - e_1}$$

We next enter most of the data shown above and use it to compute $N_{3\min}$ and $N_{3\max}$

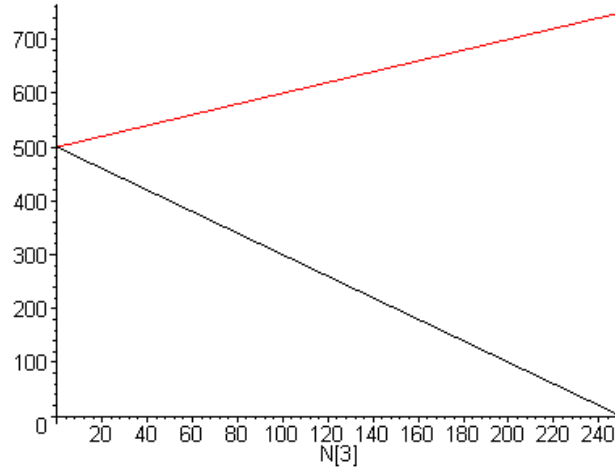
```
e[1] := 1: e[2] := 2: e[3] := 3: u := 1.5: M := 1000: U := M*u:
N3min := max(0, M*(u-e[2])/(e[3]-e[2]));
N3max := M*(u-e[1])/(e[3]-e[1]);
```

$$N3min := 0$$

$$N3max := 250.0000000$$

so the legal range of N_3 is (0,250). Here is a plot of N_1 (red) and N_2 (black) versus N_3 :

```
plot([N[1],N[2]],N[3] = N3min..N3max, color = [red,black]);
```



(6.4.13)

We then enter the degeneracies and take a look at Ω in its numerical form,

```
g[1] := 5000: g[2] := 5000: g[3] := 5000:
Omega;
```

$$\frac{5000^{(500.0 + N_3)} 5000^{(500.0 - 2 N_3)} N_3^{N_3}}{(500.0 + N_3)! (500.0 - 2 N_3)! N_3!}$$

(6.4.14)

As is typical in statistical mechanics, at its peak the value of Ω is a *very* large number. Even with the relatively small number of particles $M = 1000$ and degeneracy $g = 5000$, the peak value is (see below)

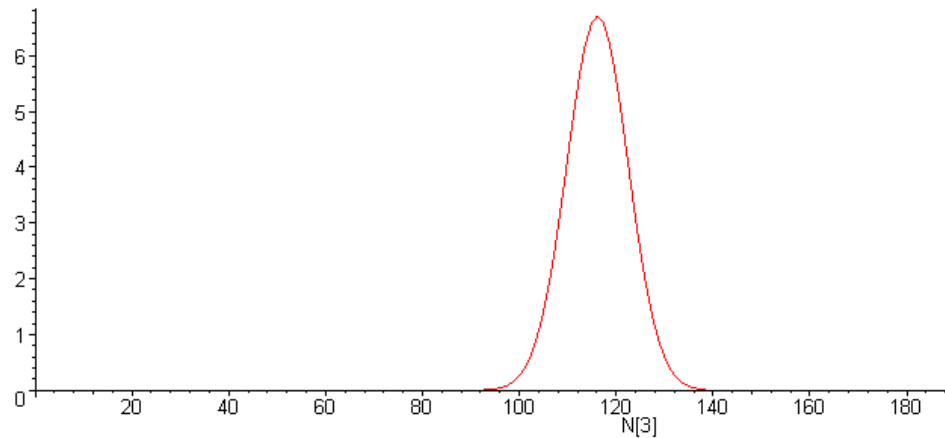
$$\Omega_{\text{peak}} = 6.71 \times 10^{1519} \quad (6.4.15)$$

Maple is uncomfortable plotting numbers larger than about 10^{40} so we pre-scale Ω down by 10^{1519} to make the scaled Ω be Maple-digestible. This scaling process in turn creates numbers smaller than 10^{-40} which are also rejected by the Maple plotter, so we use a Heaviside function to pin small numbers to 0 unless they are greater than the arbitrary value .01. Here then is a plot of $\Omega(N_3)$,

```

scale := 1e1519;
                                scale := .1 101520
OmegaScaled := (Omega/scale)*Heaviside(Omega/scale-.01):
xx := plot( OmegaScaled,N[3] = N3min..N3max,numpoints=100 ):
display(xx);

```



(6.4.16)

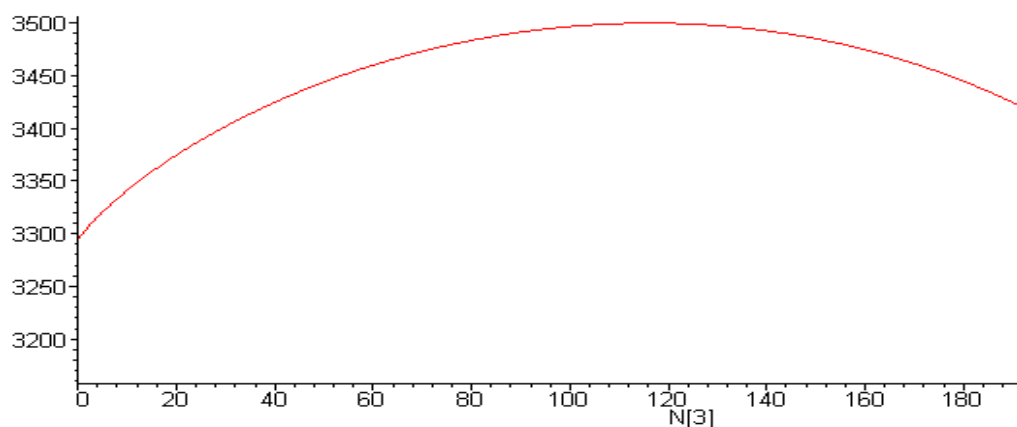
We broke the display command into two pieces: first create xx then plot xx. Replacing : with ; after the plot command gives one a list of numbers to be plotted, and from this list one can determine a reasonable scale factor. Once that is found, the plotting coordinates need not be displayed. The curve $\Omega(N_3)$ has the typical Gaussian (normal) shape which no doubt the energetic reader could show results from the central limit theorem.

A plot of $f = \ln(\Omega)$ has a much smoother and broader shape,

```

plot(ln(Omega),N[3] = N3min..N3max,numpoints=100 );

```



(6.4.17)

Recall from (6.3.8) that $\frac{d^2f}{d^2N_i}$ is everywhere negative, consistent with the cupping down seen in this plot.

Next, we have Maple solve (6.2.17) for the Lagrange multiplier β ,

```

eq := sum((e[i]-u)*g[i]*exp(-beta*e[i]), i=1..3) = 0;
      eq := -2500. e(-1. β) + 2500. e(-2. β) + 7500. e(-3. β) = 0
s := solve(eq, beta);
      s := .2644970943 - 3.141592654 I, .8341151943
beta := s[2];
      β := .8341151943

```

With $w = e^{-\beta}$ the above equation is $3w^2 + w - 1 = 0$ which has solutions $w = (-1 \pm \sqrt{13})/6$. The minus sign gives the complex β noted, while the plus sign gives the physical solution $\beta = -\ln[(-1 + \sqrt{13})/6] = .834$.

From this β value we compute the other Lagrange parameter A from (6.2.14),

```

Z := sum(g[i]*exp(-beta*e[i]), i=1..3);
      Z := 3523.658588
A := M/Z;
      A := .2837959397

```

Once these parameters are known, we use (6.2.11) to display the solution vector $\mathbf{N} = \{N_i\}$.

```

for i in [3,2,1] do N[i] := A*g[i]*exp(-beta*e[i]) od,
      N3 := 116.2040605
      N2 := 267.5918793
      N1 := 616.2040607

```

Since we set $u < \varepsilon_2$, we see here a normal distribution where higher energy states have fewer particles.

Our final task is to use (6.3.8) and (6.3.13) to determine the half-width of the Ω peak,

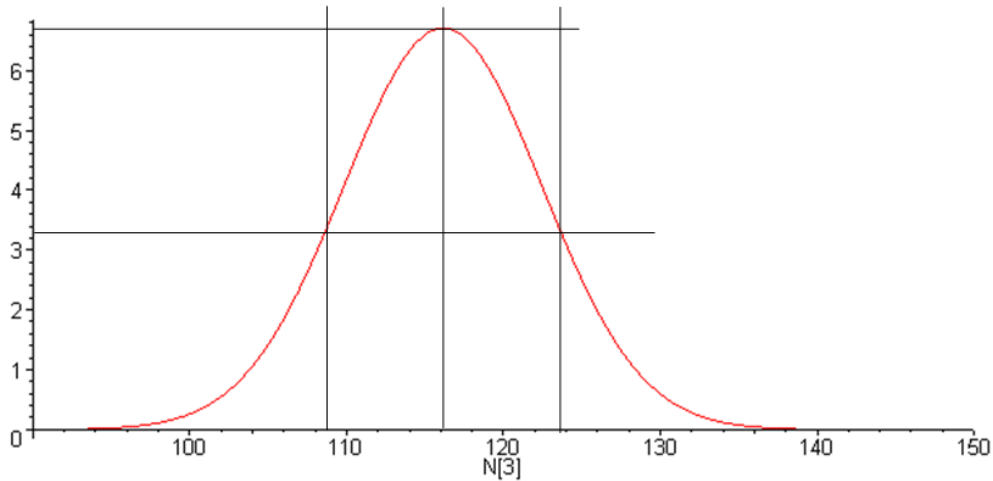
```

curve := (1/N[1])*((e[3]-e[2])/(e[2]-e[1]))^2 +
(1/N[2])*((e[3]-e[1])/(e[2]-e[1]))^2 + 1/N[3];
      curve := .02517652703
dN3 := sqrt(1.4/curve);
      dN3 := 7.457033725

```

which says $\Delta N_{1/2} \approx 7.5$. This width and the solution $N_3 = 116.2$ can be verified from this blowup plot of the Ω peak,

```
xx := plot( OmegaScaled, N[3] = 90..150, numpoints=100 ):
display(xx);
```



(6.4.18)

Notice that this Maple code does not use the Stirling approximation and provides a continuous interpolation of $N!$ as $\Gamma(N+1)$.

The peak value of Ω is found to be

```
evalf(Omega);
```

$.6710572393 \cdot 10^{1520}$

In the above example we used $u = 1.5 < \varepsilon_2 = 2$. If we instead use $u = 2.5 > \varepsilon_2 = 2$, we find that the state populations are as shown above but in reverse order. Such a inverted situation corresponds to a negative absolute temperature T ($\beta = -.834 = 1/(kT)$) and cannot therefore arise in thermal equilibrium. It does arise in a laser medium which is driven by some power source ("population inversion").

Appendix A : Proof that matrix rank is the number of independent columns and rows

Shilov proves these facts but the proof is a bit spread out over several sections. Here we provide an alternative proof that is perhaps more direct. Recall the definition (1.5) of the rank of a matrix :

Definition: The **rank** r of an $m \times n$ matrix A is the dimension of the largest non-vanishing subdeterminant (minor) within A . [Shilov 1.92] Thus, $r \leq \min(m,n)$. (1.5) (A.1)

The following well-known linear algebra fact will be used below (theorem and contrapositive):

$$\begin{aligned} \text{the columns/rows of a square matrix } M \text{ are linearly independent} &\Leftrightarrow \det(M) \neq 0 \\ \text{the columns/rows of a square matrix } M \text{ are linearly dependent} &\Leftrightarrow \det(M) = 0 . \end{aligned} \quad (\text{A.2})$$

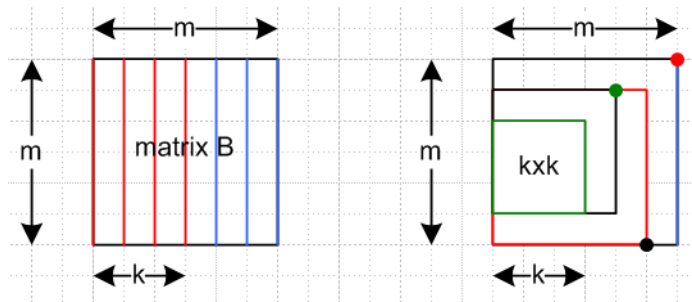
Roughly speaking, if the rows of M are linearly independent, we expect the system of linear equations $y = Mx$ to be solvable by Cramer's Rule $x = M^{-1}y = [\det(M)]^{-1} \text{cof}(M^T) y$ which requires $\det(M) \neq 0$. More obscurely, if a set of n column vectors c_i is linearly independent, we expect the volume of the n -piped spanned by these vectors in E^n to be non-zero, and that volume is given by $|\det(c_1, c_2, \dots, c_n)| = |\det(M)|$. See our Tensor Analysis document (B.6.10).

Once Theorem 1 below is established, the rest is easy.

Theorem 1. If all $k \times k$ minors in k columns of matrix A vanish, the k columns are linearly dependent. (A.3)

Contrapositive: If k columns are linearly independent, they must contain at least one non-zero $k \times k$ minor.

Proof: Assume matrix A has m rows and n columns. In order for $k \times k$ minors to exist, one must have $k \leq \min(m,n)$. Gather up the k columns and make them be the leftmost k columns of a new matrix B . Since these columns have m elements, and since $k \leq m$, add $m-k$ arbitrary new columns to the right of the k columns so that matrix B is then a square matrix. On the left below we show the k columns taken from matrix A in red, and then the arbitrary added columns are shown in blue.



(A.4)

Consider now the process of computing $\det(B)$ by going down the rightmost column using the standard cofactor sum formula. This $\det(B)$ is a linear combination of $(m-1) \times (m-1)$ minors all in the left $m-1$ columns. Consider one of these minors. If we evaluate it using the same cofactor formula (of one less dimension), it will be a linear combination of $(m-2) \times (m-2)$ minors in the left $m-2$ columns. We keep going until we are evaluating a set of $k \times k$ cofactors in the left k columns. But these all vanish by the

theorem premise. Thus, reversing this stack we conclude that $\det(B) = 0$. The picture on the right above shows one minor at each level of the descent just described (red dot goes with red minor, etc).

Now suppose it were possible that the k red columns were independent. Since the added blue columns are arbitrary, we could certainly find a set of added blue columns so that all m columns were independent. But then we would have $\det(B) \neq 0$. But we just showed that $\det(B) = 0$, so it must not be possible to have the k red columns be independent, so they must be dependent. QED.

Theorem 2. $\text{Rank}(A) = r \Rightarrow A$ has r linearly independent columns. (A.5)

Proof: If $\text{rank}(A) = r$, A must have at least one non-vanishing $r \times r$ minor. The r columns of this minor are therefore linearly independent (each of these mini-columns has r elements). This means that the r full columns containing these mini-columns are also independent (see * below). So matrix A has at least r independent columns. Now let $k = r+1$. We know that all $k \times k$ minors in A vanish. By Theorem 1, any set of k columns must be linearly dependent. Since $k = r+1$, any $r+1$ columns are linearly dependent, so the number of independent columns is r .

* Since the mini-columns C_i are independent, one cannot write $\sum_i k_i C_i = 0$ (with at least one $k_j \neq 0$). If the corresponding full columns c_i were dependent, one could write $\sum_i k_i c_i = 0$. But if $\sum_i k_i c_i = 0$, then one must have $\sum_i k_i C_i = 0$ since the latter is just a subset of the former equation set. But $\sum_i \lambda_i C_i = 0$ says that the mini-columns are dependent, which they are not. Thus, one cannot write $\sum_i k_i c_i = 0$ and thus the full columns are independent. (A.6)

Theorem 3. A has r linearly independent columns $\Rightarrow \text{Rank}(A) = r$ (A.7)

Proof: If there are r linearly independent columns, then at least one $r \times r$ minor within those columns must be non-zero (contrapositive of Theorem 1). Suppose there were an $(r+k) \times (r+k)$ non-vanishing minor. Then the $r+k$ mini-columns of that minor would be independent, and thus there would be $r+k$ independent full columns. But the theorem premise says there are only r independent columns, so all minors larger than $r \times r$ must vanish. Therefore $\text{rank}(A) = r$.

Making trivial replacements in the above three theorems and proofs (column $\rightarrow$ row, leftmost $\rightarrow$ topmost and so on), one finds that all three theorems are valid for rows as well as columns. Thus:

Theorem 1'. If all $k \times k$ minors in k rows of matrix A vanish, the k rows are linearly dependent. (A.8)

Theorem 2'. $\text{Rank}(A) = r \Rightarrow A$ has r linearly independent rows. (A.9)

Theorem 3'. A has r linearly independent rows $\Rightarrow \text{Rank}(A) = r$ (A.10)

Comment: Sometimes the number of linearly independent rows of a matrix is called the **row rank**, and the number of independent columns is called the **column rank**. Theorems 2 and 3 show that column rank = rank, while Theorems 1' and 2' show that row rank = rank. Therefore:

$$\text{column rank} = \text{row rank} = \text{rank} = \text{dimension of largest non-vanishing minor} \quad (\text{A.11})$$

Basis Minors and Basis Columns. If $\text{rank}(A) = r$, we know there must exist at least one non-vanishing $r \times r$ minor in A . Shilov refers to such a minor as a **basis minor** and the columns passing through this minor are called **basis columns**. We showed in Theorem 2 that the set of such basis columns is linearly independent (which is why they are called basis columns). Obviously any of these columns can be written as a linear combination of the basis columns, such as $\mathbf{c}_2 = \sum_{i=1}^r k_i \mathbf{c}_i = 1 \mathbf{c}_2$. If $k = r+1$, all $k \times k$ minors vanish and by Theorem 1 all sets of $k = r+1$ columns are linearly dependent. Thus, every non-basis column is a linear combination of the basis columns. We have thus proved Shilov's Basis Minor Theorem which we quote from section 1.93 of his book (p 25)

THEOREM (*Basis minor theorem*). *Any column of the matrix A is a linear combination of its basis columns.* (A.12)

In his proof that column rank = rank, Shilov uses the above theorem as a starting point.

References

The method of Lagrange multipliers often makes cameo appearances in texts on the calculus of multiple variables. There are also a moderate number of sites and pdf's on web, search on "Lagrange Multipliers". Whole books on the use the Lagrange multipliers for specialized applications may be found online.

R.C. Buck with E.F. Buck, *Advanced Calculus*, 2nd Ed. (McGraw-Hill, New York, 1965). The discussion of extremum problems with constraints occupies pp 359-363. In the 3rd Edition (McGraw-Hill, New York, 1978), reprinted by (Waveland Press, Long Grove IL, 2003), this discussion has moved to pages 536-540. This excellent and now classic book was first published in 1956.

H. Goldstein, *Classical Mechanics* (Addison-Wesley, Boston, 1950), another classic. There is a 3rd edition 2001 by Goldstein, Safco and Poole.

S. Jensen, *An Introduction to Lagrange Multipliers*,
<http://www.slimy.com/~steuard/teaching/tutorials/Lagrange.html> (includes video lectures)

P. Lucht, *Tensor Analysis and Curvilinear Coordinates* (2014), <http://user.xmission.com/~rimrock> . If not there, search on "Phil Lucht Documents".

G.E. Shilov, *Linear Algebra* (Dover Publications, 1977).

W. F. Trench, *The Method of Lagrange Multipliers*, <http://digitalcommons.trinity.edu/mono/7/> ,
Supplement 2 (31p). This paper also has a "Theorem 1" which is related to but different from ours.