The Method of Lagrange Multipliers

Phil Lucht
Rimrock Digital Technology, Salt Lake City, Utah 84103
last update: June 13, 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 ..... **10**
(a) Proof for N = 6 and S = 4 10
(b) Proof for general S ≤ N 14
4. The Gradient Interpretation: Examples 1 and 2 ..... **18**
4.1 Setup 18
4.2 Example 1: Half sphere with one simple constraint 19
4.3 Example 2: Half 4D sphere with two simple constraints 24
4.4 The General Case 27
5. Example 3: Extremal distances between an ellipse and a circle **28**
6. Example 4: The Boltzmann factor in Statistical Mechanics **35**
6.1 Statement of The Boltzmann Extremum Problem 35
6.2 Solution of The Boltzmann Extremum Problem 37
6.3 More details of the solution 40
6.4 A numerical example with plots 43
Appendix A : Matrix rank equals the number of independent columns and rows **49**
Appendix B : Determinants **53**
B.1 Definition of the determinant 53
B.2 The permutation group and the permutation tensor ε 55
B.3 Minors, Cofactors, and the Cofactor Expansions of det(M) 59
B.4 Expressions for the inverse matrix M^{-1} and Cramer's Rule 61
References **64**

Overview and Summary

The Method of Lagrange Multipliers is used to determine extrema of a real function subject to some number of constraint equations. The main purpose of this document is to provide a solid derivation of the method and thus to show why the method works. A secondary purpose is to provide some interesting and illustrative examples.

Although we discuss the relevant gradient approach in Section 4, it seems more convincing to derive the method using a matrix approach where the candidate 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.

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

Here is a concise summary of the document:

Section 1 gives a statement of Theorem 1 in the context of a certain R matrix

Section 2 shows how Theorem 1 provides an immediate derivation of the Method of Lagrange Multipliers.

Section 3 then proves Theorem 1.

Section 4 gives the gradient interpretation with two visualizable examples.

Section 5 presents another example which, though simple, requires numeric work to solve.

Section 6 provides an extended standard example involving statistical mechanics.

Appendix A proves for a general matrix that column rank = row rank = rank.

Appendix B derives a batch of facts about determinants which are used in Appendix A.

References (a very short list) are provided on the last page.

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.

1. The R matrix and its Rank

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

We seek an extremum of a real function f(**r**) subject to S-1 constraints a(**r**) = 0, b(**r**) = 0, ... q(**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(**r**) = 0 and b(**r**) = 0. So here is the problem:

find extremum of f(**r**) subject to constraints $\underset{1}{a(\mathbf{r})=0}, \underset{2}{b(\mathbf{r})=0}, \underset{3}{c(\mathbf{r})=0}, \ldots. \underset{S-1}{q(\mathbf{r})=0}$. (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 $x_1, x_2, x_3 \ldots x_N$ the partial is with respect to. So here is that matrix, which we shall call R :

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

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

On the far right we abbreviate each row of functions in **bold**, indicating a row vector like **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 **r**, we really have R(**r**). As the point **r** is varied, the elements of the matrix R(**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 SxS "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 SxS. For example, for N = 6 and S = 4 we have three constraints and the R matrix is this,

$$R = \begin{pmatrix} f_1 & f_2 & f_3 & f_4 & f_5 & f_6 \\ a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \\ b_1 & b_2 & b_3 & b_4 & b_5 & b_6 \\ c_1 & c_2 & c_3 & c_4 & c_5 & c_6 \end{pmatrix} \tag{1.3}$$

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:

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

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 N columns and S 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 $u = f(x_1,x_2, x_3....x_N)$ 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 $a(x_1,x_2, x_3....x_N) = 0$ 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 extruded 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 intersection 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 (Example 1).

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

Definitions: A **minor** of matrix A is the determinant of any square submatrix of A obtained by crossing out some rows and columns The **rank** r of an m x n matrix A is the dimension of the largest non-vanishing minor within A. [Shilov 1.92] Thus, $r \leq \min(m,n)$. (1.5)

Definitions: [Shilov 2.21] Consider an m x n matrix A. Let $\mathbf{r}_1$, $\mathbf{r}_2$, ...$\mathbf{r}_N$ 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

$$k_1 \mathbf{r}_1 + k_2 \mathbf{r}_2 + ... + k_N \mathbf{r}_N = 0 \ . \tag{1.6}$$

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) is linearly independent.

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

Facts: The rank r as defined in (1.5) has these properties for any m x 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 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 **r** is a solution to the extremum problem (1.1), then rank[R(**r**)] < S. (1.8)

According to the rank definition (1.5),

rank(R) < S ⇔ all SxS submatrices must have zero determinant (1.9)

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

The point of Theorem 1 is that the solution point **r** of the constrained extremum problem must be a point where the R(**r**) matrix drops below full rank. We shall prove this theorem in Section 3 by showing that, if **r** is a solution to the extremum problem, then all those SxS submatrices discussed above have zero determinant and thus rank(R) < S. If one is searching for candidate solutions **r** to the extremum problem, one can restrict one's search to points **r** for which rank[R(**r**)] < S.

Comment on the R matrix. Imagine a general transformation from x-space $\mathbf{x} = (x_1, x_2, ...x_N)$ to u-space with coordinates $\mathbf{u} = (u_1, u_2, ... u_S)$. One might write such a transformation as $\mathbf{u} = \mathbf{F}(\mathbf{x})$ where $F:E^N \rightarrow E^S$. Although this transformation in general is non-linear, in a tiny neighborhood of point **x** in x-space (and the corresponding point **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

$f = u_1(x_1, x_2, ...x_N)$ $\quad R_{11} = \partial u_1/\partial x_1 = \partial f/\partial x_1 = f_1$ $\quad R_{12} = f_2$ etc.

$a = u_2(x_1, x_2, ...x_N)$ $\quad R_{21} = \partial u_2/\partial x_1 = \partial a/\partial x_1 = a_1$ $\quad R_{22} = a_2$ etc

$b = u_3(x_1, x_2, ...x_N)$

...

$q = u_S(x_1, x_2, ...x_N)$ (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 $F:E^N \rightarrow E^N$ 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),

find extremum of f(**r**) subject to constraints $a(\mathbf{r})=0,\ b(\mathbf{r})=0,\ c(\mathbf{r})=0,\ \ldots.\ q(\mathbf{r})=0$. *(1.1)*
$\quad\quad\quad

(2.5) be true is therefore a *candidate* for a solution to the constrained extremum problem (1.1). Such a solution value of **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 **r** = (x_1, x_2 .....x_N) and λ_1, λ_2.....λ_{S-1} . Here again are the equations to solve:

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

Since all the functions in these equations are known (since f(**r**) and all the constraints are known), it is in principle possible to solve the equations for **r** and the λ_i. Then those solution **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 **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 expressions for the λ_i as follows. First, write the first N equations of (2.6) in matrix form,

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

where the matrix has S-1 columns (one column for each constraint) and N rows. Since S ≤ 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

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

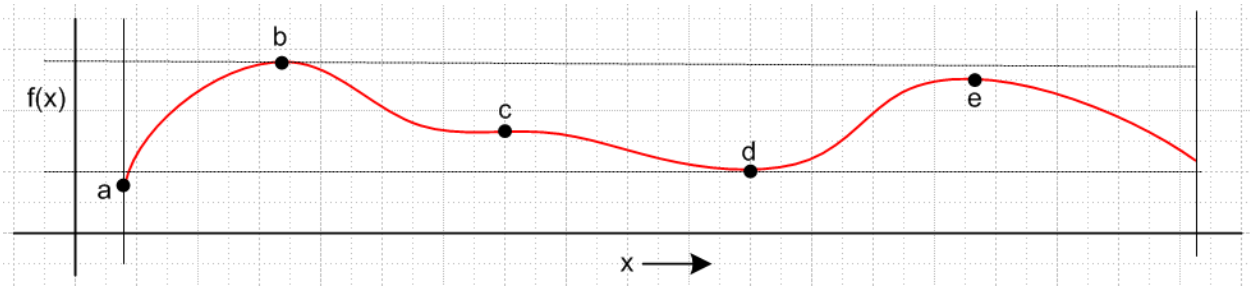
where the matrix is now square. Assuming this matrix has non-zero determinant, one can invert (2.8) to obtain expressions for the Lagrange multipliers $\lambda_i(\mathbf{r})$ as functions or **r**. These λ_i functions can be inserted into the first N equations of (2.6) and then one can proceed to solve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

Stationary Points vs. Extremum Points

The Method of Lagrange Multipliers really identifies stationary points of f some of which may be extremum points. A **stationary point** (**critical point**) of $f(x_1,x_2, x_3....x_N)$ is one where all the partial derivatives vanish. In the notation above, we would write $f_i = 0$ for i = 1,2...N or just **f** = 0. The Lagrange Multiplier method looks for stationary points of the auxiliary function H where **H** = 0 as in (2.5). For f(x,y) the reader well knows that a stationary point where $\partial_x f=0$ and $\partial_y f = 0$ could be a saddle point which is neither a minimum nor a maximum. There could be many stationary points for a given problem, some of which are extrema, and some of which are not (see Example 3 in Section 5). Each stationary point will have its own set of Lagrange Multipliers $\{\lambda_i\}$. Moreover, a true extremal point may occur at a boundary of the domain of interest, and would then not be detected by the Lagrange Multiplier method.

The following 1D function f(x) illustrates many of these points.



(2.9)

Here the domain of interest is the portion of the x axis lying between the two vertical bars. Point a is the true minimum and occurs at a boundary. Point b is the true global maximum, while point e is only a local maximum. Point d wants to be the global minimum but is trumped by point a. Point c has $\partial_x f = 0$ but is neither a local minimum nor a local maximum -- it is the 1D analog of a saddle point. Had points b and e the same height, there would be two equal global maxima. All these situations can occur for the function H for the constrained extremum problem and so analysis is required after the stationary points of H are located.

Comments

1. For the Lagrange Multiplier method to be valid, the functions f(**r**), a(**r**), b(**r**)....q(**r**) must be **continuous** and **differentiable** in all arguments x_i *and* their derivatives must also be continuous. The functions are therefore C^0 and C^1. This ensures that the R(**r**) matrix in (1.2) behaves smoothly (is continuous) as a function of **r**.

2. **Notation**: Some authors use the notation $f_{x_i} \equiv \partial f/\partial x_i$ and the constraints are $g^{(s)}(\mathbf{x}) = 0$ where s is a label, Then the first derivative of a constraint function is $g^{(s)}{}_{x_i}(\mathbf{x})$. Another choice is to refer to the constraint functions as $g_s(x)$ 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.

3. The λ_i terms in (2.1) often appear with **minus signs** in place of plus signs. This removes the minus sign in (2.7) and (2.8). These signs are just a convention and have no effect on the solution points **r**.

4. 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, $\left(\frac{\partial L}{\partial q_k} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_k} \right) = 0$. One may add a set of differential constraints to the problem of the form $\Sigma_k a_{ik} dq_k = 0$ 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 $\left(\frac{\partial L}{\partial q_k} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_k} \right) + \Sigma_i \lambda_i a_{ik} = 0$. 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 **r** is a solution to the extremal problem (1.1), then rank[R(**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 R matrix, again with bolded row vectors on the right,

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

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:

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

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

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

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,

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

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

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

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

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

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

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

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,

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

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 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,

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

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

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

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,

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

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,

$$\begin{pmatrix}0\\0\\0\\0\end{pmatrix} = \begin{pmatrix}f_1 & f_2 & f_3 & f_4\\ a_1 & a_2 & a_3 & a_4\\ b_1 & b_2 & b_3 & b_4\\ c_1 & c_2 & c_3 & c_4\end{pmatrix}\begin{pmatrix}X^1{}_4\\ X^2{}_4\\ X^3{}_4\\ 1\end{pmatrix} = (f_1, f_2, f_3, f_4)\begin{pmatrix}X^1{}_4\\ X^2{}_4\\ X^3{}_4\\ 1\end{pmatrix}$$

$$\begin{pmatrix}0\\0\\0\\0\end{pmatrix} = \begin{pmatrix}f_1 & f_2 & f_3 & f_5\\ a_1 & a_2 & a_3 & a_5\\ b_1 & b_2 & b_3 & b_5\\ c_1 & c_2 & c_3 & c_5\end{pmatrix}\begin{pmatrix}X^1{}_5\\ X^2{}_5\\ X^3{}_5\\ 1\end{pmatrix} = (f_1, f_2, f_3, f_5)\begin{pmatrix}X^1{}_5\\ X^2{}_5\\ X^3{}_5\\ 1\end{pmatrix}$$

$$\begin{pmatrix}0\\0\\0\\0\end{pmatrix} = \begin{pmatrix}f_1 & f_2 & f_3 & f_6\\ a_1 & a_2 & a_3 & a_6\\ b_1 & b_2 & b_3 & b_6\\ c_1 & c_2 & c_3 & c_6\end{pmatrix}\begin{pmatrix}X^1{}_6\\ X^2{}_6\\ X^3{}_6\\ 1\end{pmatrix} = (f_1, f_2, f_3, f_6)\begin{pmatrix}X^1{}_6\\ X^2{}_6\\ X^3{}_6\\ 1\end{pmatrix} \quad . \qquad (3.11)$$

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 (see (B.4.8)) to give

$$\begin{pmatrix}X^1{}_4\\ X^2{}_4\\ X^3{}_4\\ 1\end{pmatrix} = (f_1, f_2, f_3, f_4)^{-1}\begin{pmatrix}0\\0\\0\\0\end{pmatrix} = \begin{pmatrix}0\\0\\0\\0\end{pmatrix} \quad . \qquad (3.12)$$

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

$$\det(f_1, f_2, f_3, f_4) = 0$$
$$\det(f_1, f_2, f_3, f_5) = 0$$
$$\det(f_1, f_2, f_3, f_6) = 0$$

or

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

Therefore, we have shown that three of the SxS = 4x4 submatrices of Fig (1.3) vanish. But there are 15 such submatrices, and in order to show that rank(R) < S, we have to show that *all* 15 SxS 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

$$\det(f_1, f_2, f_4, f_m) = 0 \qquad m \neq 1,2,4 \ . \tag{3.14}$$

More generally, had we chosen to eliminate variables x_i, x_j and x_k ($i \neq j \neq k$), we would have obtained

$$\det(f_i, f_j, f_k, f_m) = 0 \qquad m \neq i,j,k \ . \tag{3.15}$$

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 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 **r**, one has rank[R(**r**)] < S .

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

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

These constants *are* the Lagrange Multipliers and we see that the fact that they exist follows directly from the fact that rank[R(**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 SxS minors of matrix R vanish so that rank(R) < S.

(b) Proof for general S ≤ 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 ≤ N rows. The S-1 constraint functions are a,b,....q. Remember that f_2 means $\partial f/\partial x_2$.

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

Here are the S-1 constraint equations,

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

Eliminate (theoretically) the first S-1 variables $x_1, x_2 .... x_{S-1}$ using the S-1 constraint equations:

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

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

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

$$(3.4)'$$

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

$$\begin{aligned} F_S &= f_1X^1{}_S + f_2X^2{}_S + f_3X^3{}_S + ... + f_{S-1}X^{S-1}{}_S + f_S \\ F_{S+1} &= f_1X^1{}_{S+1} + f_2X^2{}_{S+1} + f_3X^3{}_{S+1} + ... + f_{S-1}X^{S-1}{}_{S+1} + f_{S+1} \\ F_{S+2} &= f_1X^1{}_{S+2} + f_2X^2{}_{S+2} + f_3X^3{}_{S+2} + ... + f_{S-1}X^{S-1}{}_{S+2} + f_{S+2} \\ &... \\ F_N &= f_1X^1{}_N + f_2X^2{}_N + f_3X^3{}_N + ... + f_{S-1}X^{S-1}{}_N + f_N \ . \qquad \text{// N-S+1 equations} \end{aligned} \qquad (3.6)'$$

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

$$
\begin{aligned}
&f_1X^1{}_S + f_2X^2{}_S + f_3X^3{}_S + \ldots + f_{S-1}X^{S-1}{}_S + f_S = 0\\
&f_1X^1{}_{S+1} + f_2X^2{}_{S+1} + f_3X^3{}_{S+1} + \ldots + f_{S-1}X^{S-1}{}_{S+1} + f_{S+1} = 0\\
&f_1X^1{}_{S+2} + f_2X^2{}_{S+2} + f_3X^3{}_{S+2} + \ldots + f_{S-1}X^{S-1}{}_{S+2} + f_{S+2} = 0\\
&\ldots\\
&f_1X^1{}_N + f_2X^2{}_N + f_3X^3{}_N + \ldots + f_{S-1}X^{S-1}{}_N + f_N = 0\ .\quad \text{// N-S+1 equations} \qquad (3.7)'
\end{aligned}
$$

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,

$$
\begin{aligned}
&f_1X^1{}_S + f_2X^2{}_S + f_3X^3{}_S + \ldots + f_{S-1}X^{S-1}{}_S + f_S = 0\\
&a_1X^1{}_S + a_2X^2{}_S + a_3X^3{}_S + \ldots + a_{S-1}X^{S-1}{}_S + a_S = 0\\
&b_1X^1{}_S + b_2X^2{}_S + b_3X^3{}_S + \ldots + b_{S-1}X^{S-1}{}_S + b_S = 0\\
&\ldots.\\
&q_1X^1{}_S + q_2X^2{}_S + q_3X^3{}_S + \ldots + q_{S-1}X^{S-1}{}_S + q_S = 0\ .\quad \text{// S equations} \qquad (3.10)'_S
\end{aligned}
$$

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

$$
\begin{aligned}
&f_1X^1{}_N + f_2X^2{}_N + f_3X^3{}_N + \ldots + f_{S-1}X^{S-1}{}_N + f_N = 0\\
&a_1X^1{}_N + a_2X^2{}_N + a_3X^3{}_N + \ldots + a_{S-1}X^{S-1}{}_N + a_N = 0\\
&b_1X^1{}_N + b_2X^2{}_N + b_3X^3{}_N + \ldots + b_{S-1}X^{S-1}{}_N + b_N = 0\\
&\ldots.\\
&q_1X^1{}_N + q_2X^2{}_N + q_3X^3{}_N + \ldots + q_{S-1}X^{S-1}{}_N + q_N = 0\ .\quad \text{// S equations} \qquad (3.10)'_N
\end{aligned}
$$

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

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,

$$
\begin{pmatrix} 0\\0\\\ldots\\0\\0 \end{pmatrix} = (f_1, f_2, f_3, \ldots f_{S-1}, f_S) \begin{pmatrix} X^1{}_S\\X^2{}_S\\\ldots\\X^{S-1}{}_S\\1 \end{pmatrix}
$$

$$
\begin{pmatrix} 0\\0\\\ldots\\0\\0 \end{pmatrix} = (f_1, f_2, f_3, \ldots f_{S-1}, f_{S+1}) \begin{pmatrix} X^1{}_{S+1}\\X^2{}_{S+1}\\\ldots\\X^{S-1}{}_{S+1}\\1 \end{pmatrix}
$$

●●●●

$$
\begin{pmatrix} 0\\0\\\ldots\\0\\0 \end{pmatrix} = (f_1, f_2, f_3, \ldots f_{S-1}, f_N) \begin{pmatrix} X^1{}_N\\X^2{}_N\\\ldots\\X^{S-1}{}_N\\1 \end{pmatrix} \quad . \qquad (3.11)'
$$

We then argue as before that, to avoid the contradiction 0 = 1, the determinants of these SxS submatrices must be zero,

$$\det(f_1, f_2, f_3, \ldots f_{S-1}, f_S) = 0$$
$$\det(f_1, f_2, f_3, \ldots f_{S-1}, f_{S+1}) = 0$$
$$\ldots$$
$$\det(f_1, f_2, f_3, \ldots f_{S-1}, f_N) = 0$$

or

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

Finally, if we start off by eliminating the S-1 variables x_i, x_j, x_k.... x_r where $i \neq j \neq k \ldots \neq r$, we would find that

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

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

Having now shown that rank[R(**r**)] < S for a solution point **r**, we know that the S rows of the R matrix must be linearly *dependent* for that **r**. This means that one can find a set of constants $\lambda_1, \lambda_2, \ldots \lambda_{S-1}$ such that

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

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

4. The Gradient Interpretation: Examples 1 and 2

4.1 Setup

Recall that, for **r** such that rank[R(**r**)] < S, one can find Lagrange multipliers λ_i such that

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

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

$$f_i(\mathbf{r}) + \lambda_1 a_i(\mathbf{r}) + \lambda_2\, b_i(\mathbf{r}) + ... + \lambda_{S-1}\, q_i(\mathbf{r}) = 0 \qquad i = 1,2...N \qquad (4.1.1)$$

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

$$\nabla f(\mathbf{r}) + \lambda_1\nabla a(\mathbf{r}) + \lambda_2\,\nabla b(\mathbf{r}) + ... + \lambda_{S-1}\,\nabla q(\mathbf{r}) = 0 \qquad \nabla = (\partial_1,\partial_2...\partial_N)\ . \qquad (4.1.2)$$

Having obtained this equation by the matrix rank method, we shall now show that it makes perfect sense in terms of the gradients. We shall consider two Examples below in order to build up to the general claim made at the end in Section 4.4. The reader uninterested in examples may skip to that section.

Before continuing, here are a few Facts which perhaps are "old hat" to the reader but need stating:

Fact 1: The equation $F(x_1,x_2.....x_n) = K$ defines an n-1 dimensional surface in E^n . (4.1.3)

For example, $F(x,y,z) = x^2+y^2+z^2$ with $F = R^2$ describes a spherical 2D surface of radius R in E^3.
$F(x,y) = x^2+y^2$ with $F= R^2$ describes a 1D surface (curve) in E^2, a circle

Fact 2: $\nabla F(\mathbf{r})$ is always locally *normal* to the surface $F(\mathbf{r}) = K$, where $\mathbf{r} = (x_1,x_2.....x_n)$. (4.1.4)

Proof: Start at point **r** on the surface $F(\mathbf{r}) = K$ and move a small distance d**r** in an arbitrary direction along the surface. Since one stays on the surface, $F(\mathbf{r}+d\mathbf{r}) = K$. Then $dF = F(\mathbf{r}+d\mathbf{r})- F(\mathbf{r}) = 0$. But one knows that $dF = \nabla F \bullet d\mathbf{r}$ so, in order that $dF = 0$ for *all* d**r** displacements on the surface, $\nabla F(\mathbf{r})$ must be locally normal to the surface at point **r**.

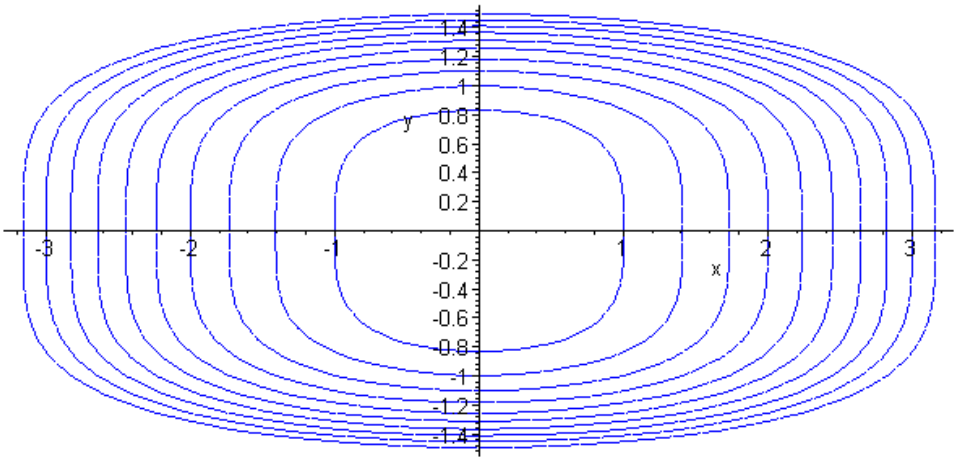
Fact 3: As K takes a set of different values K_1, K_2.....K_S, the equation $F(x_1,x_2.....x_n) = K$ describes a family of so-called **level surfaces (level sets)**. If S is large and the range of the K_i is small, these surfaces will be closely spaced. Adjacent surfaces then have nearly the same shape, although in general *all* the surfaces in the family of surfaces do not have the same shape. (4.1.5)

Example: F(x,y) = x^2+2y^4 and F= K for K = 1 to 10. Here "level surfaces" means "level curves".

```
restart;with(plots):
F := x^2+2*y^4;
```

$$F := x^2 + 2y^4$$

```
implicitplot({seq(F = K,K=1..10)},x=-4..4,y=-2..2,scaling=constrained,numpoints=1000,
color = blue);
```



(4.1.6)

Fact 4: The gradient ∇F points in the direction in which F changes most rapidly. (4.1.7)

Proof: For tiny displacement d**r**, $dF = \nabla F \bullet d\mathbf{r}$ is largest positive when d**r** points in the direction of ∇F ("uphill"). And $dF = \nabla F \bullet d\mathbf{r}$ is largest negative when d**r** points opposite the direction of ∇F ("downhill").

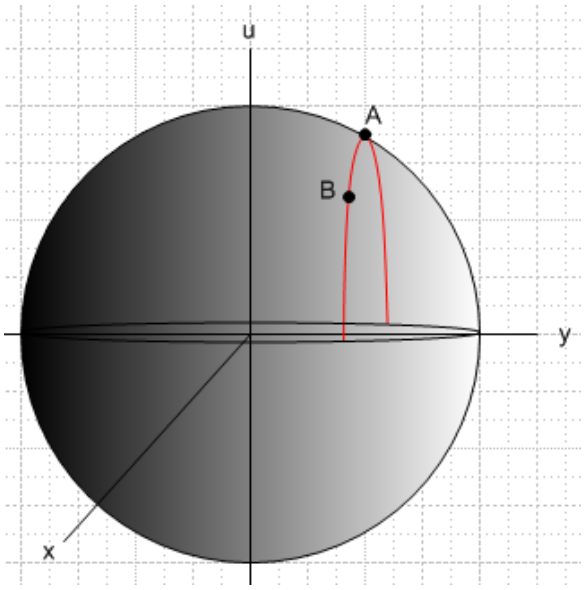
With these Facts established, we now consider a simple Lagrange Multiplier example.

4.2 Example 1: Half sphere with one simple 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 (4.1.2) 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 $\partial_{\mathbf{x}} f = 0$ and $\partial_{\mathbf{y}} f = 0$, so we are happy with this interpretation of (4.1.2) 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 plane 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.

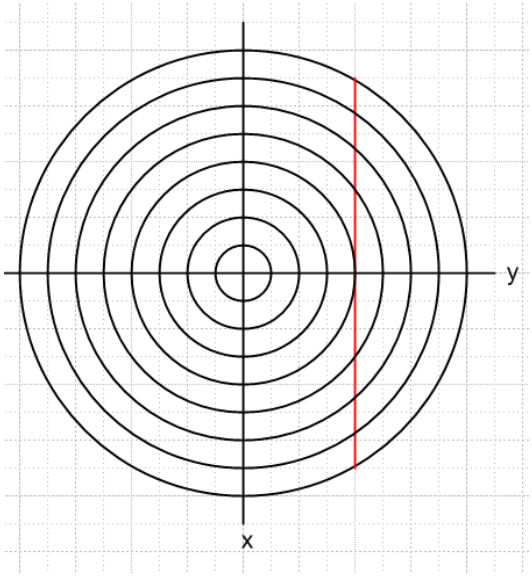


(4.2.1)

What is potentially misleading in this example (so far) is the notion of surface and gradient. For the above sphere, the surface is described by r = 2 in spherical coordinates, and then if f(**r**) = r, one finds that $\nabla f = \partial_{\mathbf{r}} f\ \hat{\mathbf{r}} = \hat{\mathbf{r}}$, and this is indeed normal to the surface at every point on it, as predicted by Fact 2 above. Similarly, the y=1 plane constraint function a(x,y) = y-1 = 0 has $\nabla a = \hat{\mathbf{y}}$ and this is everywhere normal to the constraint plane.

The potential confusion is that the Lagrange Multiplier Show does not play out on the stage shown in (4.2.1). It plays out in a space of one lower dimension. For Example 1, the Lagrange Multiplier Show plays out on the disk which lies at u = 0 in (4.2.1). The gradients involved with the Lagrange Multiplier method for Example 1 are 2D gradients which lie in this disk, not the 3D gradients to surfaces appearing in (4.2.1).

So let us move then to the proper setting which is this u=0 disk in the x,y plane: (viewed from above)



(4.2.2)

We wish to maximize the function $f(x,y) = \sqrt{4 - x^2 - y^2}$ subject to the constraint y = 1. In (4.2.1) the spherical surface is $u^2+x^2+y^2 = R^2 = 4$, so f(x,y) is then height u. So the function $\sqrt{4 - x^2 - y^2} = \sqrt{4 - r^2}$ is the height of the hemisphere in Fig (4.2.1) above the point (x,y). If we consider f(x,y) = K for a set of K values, we get a set of level curves in (4.2.2) which are circles with $r = \sqrt{4-K}$. On each circle in (4.2.2), the height of the sphere lying over the surface in (4.2.1) is constant -- that is why they are called level curves.

The constraint is a(x,y) = 0 with constraint function a(x,y) = y-1, and the constraint y = 1 is shown as the red line in (4.2.2). We compute,

$$f = \sqrt{4 - x^2 - y^2} = \sqrt{4 - r^2}$$
$$a = y-1 \qquad (4.2.3)$$

so

$$\nabla f = \nabla\sqrt{4 - r^2} = [-r/\sqrt{4-r^2}\,]\,\hat{\mathbf{r}}$$
$$\nabla a = \nabla(y-1) = \hat{\mathbf{y}}\,. \qquad (4.2.4)$$

For any point **r** in the disk of (4.2.2), ∇f therefore points toward the disk center, and this is then the uphill direction for the hemisphere in (4.2.1). In Cartesian coordinates,

$$\nabla f = \nabla\sqrt{4 - x^2 - y^2} = -(x/f)\hat{\mathbf{x}} - (y/f)\hat{\mathbf{y}} = -\frac{x}{\sqrt{4 - x^2 - y^2}}\,\hat{\mathbf{x}} - \frac{y}{\sqrt{4 - x^2 - y^2}}\,\hat{\mathbf{y}}\,. \qquad (4.2.5)$$

Now, equation (4.1.2) says

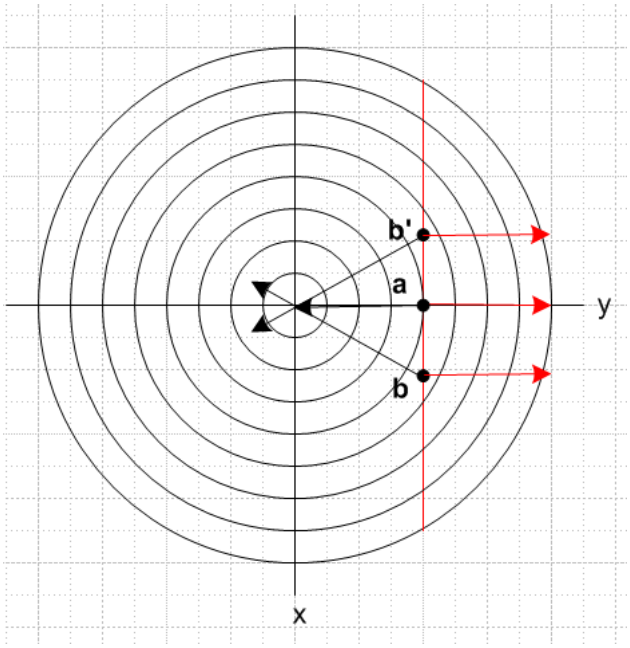
$$\nabla f(\mathbf{r}) = -\lambda_1 \nabla a(\mathbf{r}) = 0\,. \qquad (4.2.6)$$

How do we make geometric sense of this equation? If **r** is an "extremum" location, we should have df = 0 for any legal small displacement d**r** starting at this point **r**. Eq. (4.2.6) dotted into d**r** then says,

$$0 = df = \nabla f(\mathbf{r}) \bullet d\mathbf{r} = -\lambda_1 \nabla a(\mathbf{r}) \bullet d\mathbf{r}\,. \qquad (4.2.7)$$

If we displace d**r** in any direction on the constraint surface (which from Fact 2 has normal $\nabla a(\mathbf{r})$), then $\nabla a(\mathbf{r}) \bullet d\mathbf{r} = 0$ and we find df = 0, so we are thus *at* an extremum of f. For other values of position **r** , we will find either that df > 0 or df < 0 if we move d**r** along the constraint surface. If we are looking for a maximum of f, for example, then it is to our advantage to move a little d**r** in a direction for which df > 0. Doing this repeatedly to find a maximum (minimum) of f is called the **method of gradient ascent** (**descent**)

In our example with only one constraint, (4.2.6) says that we have reached an extremum when ∇f and ∇a are collinear. Let's draw onto (4.2.2) the points **a** and **b** which lie under points **A** and **B** in (4.2.1), and we include a point **b**' which is a mirror image of **b** :



(4.2.8)

At each of the points **a**,**b**,**b'** we show ∇f in black and ∇a in red. At point **a** we have reached the closest distance to the center that is allowed for points on the red constraint line, so this will correspond to the maximum value of f, which is point **A** in (4.2.1). At point **a** we see that in fact the two gradients are collinear as required by (4.1.2) at an extremum.

Suppose we are at point **b**. We may compute df for a small displacement upwards (minus x direction)

$$df(\mathbf{b}) = \nabla f \bullet d\mathbf{r} \;\; = [\, -(x/f)\hat{\mathbf{x}} - (y/f)\hat{\mathbf{y}}\,] \bullet |dx|(-\hat{\mathbf{x}}) = |dx|\,(x/f) \;> 0 \qquad \text{since } x>0 \text{ and } f>0 \;. \qquad (4.2.9)$$

Since $df > 0$, it is to our advantage in finding max f to move upwards in (4.2.8). Conversely, suppose we are instead at point **b'**. Then moving toward **a** gives,

$$df(\mathbf{b}') = \nabla f \bullet d\mathbf{r} \;\; = [\, -(x/f)\hat{\mathbf{x}} - (y/f)\hat{\mathbf{y}}\,] \bullet |dx|(\hat{\mathbf{x}}) = |dx|\,(-x/f) \;> 0 \qquad \text{since } x<0 \text{ and } f>0 \;. \qquad (4.2.10)$$

and again we find that $df > 0$. From either starting position **b** or **b'**, moving toward point **a** is a win.

It is an easy matter to compute the solution value of the Lagrange Multiplier λ_1 and **r**. Inserting (4.2.4) into (4.2.6) gives

$$\nabla f(\mathbf{r}) = -\lambda_1 \nabla a(\mathbf{r}) \qquad \Rightarrow$$

$$-\frac{x}{\sqrt{4 - x^2 - y^2}}\,\hat{\mathbf{x}} \;-\frac{y}{\sqrt{4 - x^2 - y^2}}\,\hat{\mathbf{y}} \;= -\lambda_1\,\hat{\mathbf{y}} \;. \qquad (4.2.11)$$

This says $x = 0$, and then using the constraint $y = 1$,

$$-\frac{1}{\sqrt{4 - 1}}\,\hat{\mathbf{y}} \;= -\lambda_1\,\hat{\mathbf{y}} \qquad \Rightarrow \qquad \lambda_1 = 1/\sqrt{3}\;. \qquad (4.2.12)$$

The solution to the Example 1 extremum problem is then

$$\mathbf{r} = (0,1) \qquad \lambda_1 = 1/\sqrt{3}\ . \tag{4.2.13}$$

We have claimed in (2.1) 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. One has,

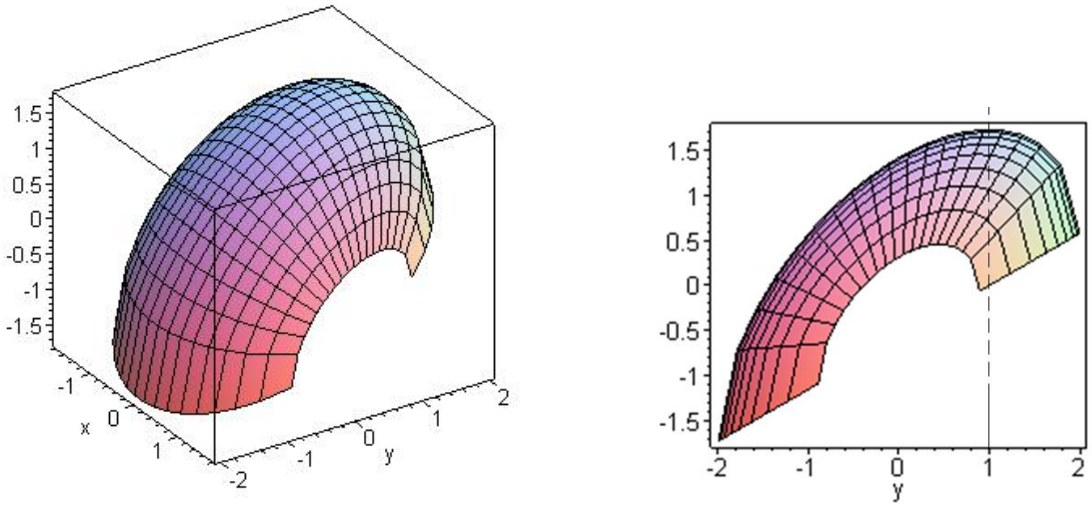
$$H(x,y) = f(\mathbf{r}) + \lambda_1 a(\mathbf{r}) = \sqrt{4 - x^2 - y^2} + (1/\sqrt{3}\ )\,(y-1). \tag{4.2.14}$$

Now $u = (1/\sqrt{3}\ )\,(y-1)$ is a plane sloping up to the right in Fig (4.2.1). This is different from the plane y = 1 which is the vertical constraint plane in that figure. In (4.2.14) 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]);
```



(4.2.15)

One can see from these plots that the maximum of H(x,y) occurs at x = 0 and y = 1.

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

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

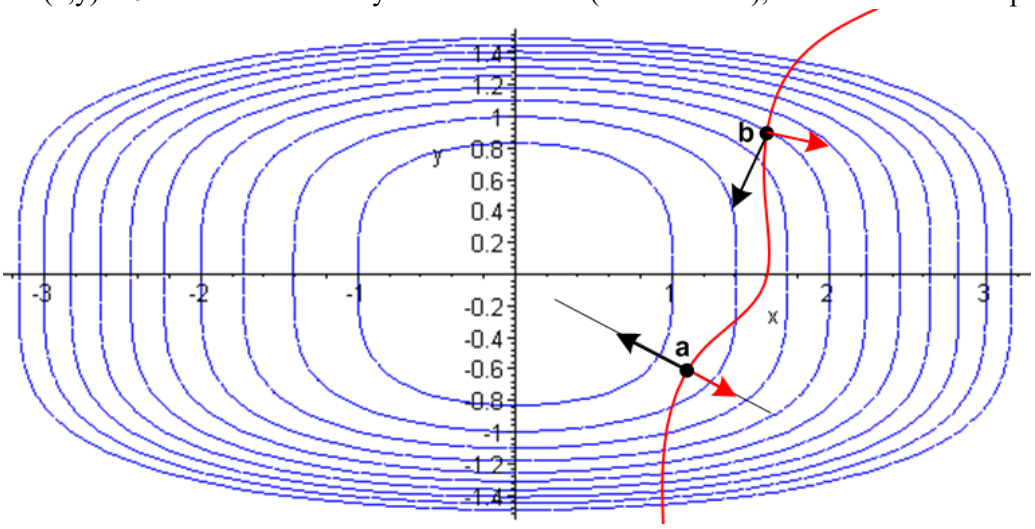
$$f = \sqrt{4 - x^2 - y^2} = \sqrt{4 - x_1^2 - x_2^2} \qquad \Rightarrow \qquad f_i = -x_i/f \qquad f > 0$$

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

$$H_{ii} = f_{ii} + \lambda_1 a_{ii} = f_{ii} = -(1/f) - (x_i/f)^2 < 0\ . \tag{4.2.16}$$

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 (2.4)) thus also has $H_{ii} < 0$ and is thus a maximum.

If we had a surface in (4.2.1) more complicated than a sphere, and a constraint more complicated than y = 1, the nature of the interpretation of (4.1.2) does not change. For example, if (4.2.1) contained a surface u = f(x,y) whose level curves were those of (4.1.6) with f = K values increasing toward the center, and if the constraint a(x,y) = 0 were some arbitrary constraint curve (shown in red), we would have this picture:



(4.2.17)

In this case the red arrow ∇a changes direction as one moves along the red curve, but the solution point is shown as the black dot **a** where the red arrow and the black arrow representing ∇f are collinear. At point **b** the gradient arrows do not line up, and there is advantage for increasing f by moving toward point **a**.

Notice that saying the two arrows are collinear is to say they are linearly dependent. Equation (4.1.2) says that the gradients are linearly dependent in the general case as well as this simple case.

4.3 Example 2: Half 4D sphere with two simple constraints

In order to better demonstrate the interpretation of (4.1.2) concerning the gradients. we upgrade Example 1 and then allow two constraints. We take the surface of (4.2.1) to be the "upper half" of a 4D sphere, whatever that means. Since we cannot draw such a thing in E^4 even in a projection drawing like (4.2.1), we don't even try. However, the Lagrange Multiplier scenario (the analog of (4.2.2)) exists in E^3, not E^4, and we can at least draw that. In Example 1 the level curves were a set of concentric circles in E^2. For Example 2 the level surfaces are a set of concentric spheres in E^3 each labeled by a value of K, where f(x,y,z) = K and where $f(x,y,z) = \sqrt{4 - x^2 - y^2 - z^2}$. We are not proving this claim, it is just based on analogy: if the slice of a 3D sphere is a 2D sphere (a disk), then a slice of a 4D sphere ought to be a 3D sphere.

We take as our two constraints the equations y = 1 and x = 1, just to keep it simple. Thus

$$f(x,y,z) = \sqrt{4 - x^2 - y^2 - z^2} = \sqrt{4 - r^2}$$
$$a(x,y,z) = y-1$$
$$b(x,y,z) = x-1 \ . \qquad (4.3.1)$$

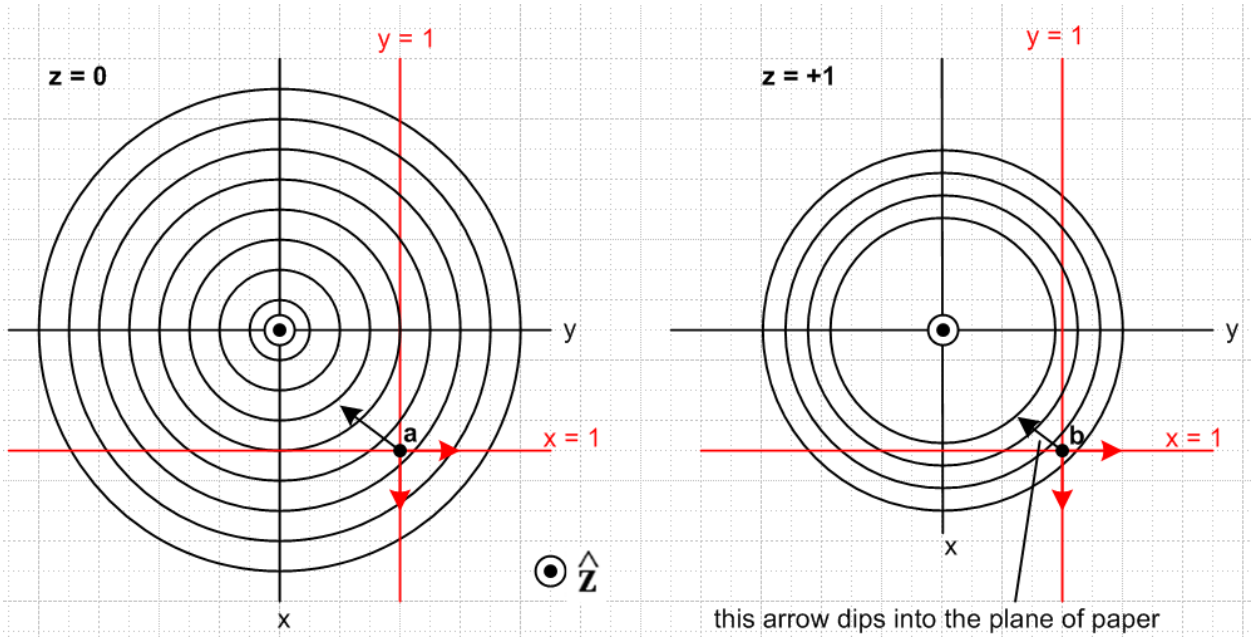
The gradients of interest are now 3D gradients in E^3 space, so

$$\nabla f = \nabla(\sqrt{4 - r^2}) = [-r/\sqrt{4-r^2}]\,\hat{\mathbf{r}} = -(x/f)\hat{\mathbf{x}} - (y/f)\hat{\mathbf{y}} - (z/f)\hat{\mathbf{z}}$$ points to sphere center
$$\nabla a = \nabla(y-1) = \hat{\mathbf{y}}$$
$$\nabla b = \nabla(x-1) = \hat{\mathbf{x}} \ . \qquad (4.3.2)$$

Equation (4.1.2) reads

$$\nabla f(\mathbf{r}) + \lambda_1 \nabla a(\mathbf{r}) + \lambda_2 \nabla b(\mathbf{r}) = 0 \ . \qquad (4.3.3)$$

This says that the three gradients must be linearly dependent, which means they must be coplanar! Before solving the problem, we would like to draw a picture corresponding to (4.2.2) for Example 1. Although one could in fact draw such a picture with some effort, we shall decline this task and instead draw two z = constant slices of the desired picture. On the left below is a slice at z = 0, while on the right is a slice at z = 1 (these are not precision pictures) :



(4.3.4)

In these pictures, we are viewing the two red constraint planes y = 1 and x = 1 edge on. The surface which satisfies both constraints is a line normal to the plane of paper which is the intersection of the two constraint planes. The circles on the left are slices of the family of concentric spheres taken at the equator z = 0. At z = +1 since our slicing plane is closer to the north pole, some of the inner spheres no longer cut the plane z = 1, and those that still do have smaller diameter circles.

Consider now the point **r** = **b** shown as a black dot on the right. The two red arrows are $\nabla a = \hat{\mathbf{y}}$ and $\nabla b = \hat{\mathbf{x}}$, each normal to its constraint plane. The black arrow dips into the plane of paper because $\nabla f = [-r/\sqrt{4-r^2}\,]\ \hat{\mathbf{r}}$ which points toward the sphere center. The three arrows ∇f, ∇a,∇b are therefore not coplanar, so they are linearly independent. That means that equation (4.3.3) cannot exist [see (1.6)] for this point **r**. This black point **b** is located outside the r = 7 units sphere (in our crude picture).

Consider next the point **r** = **a** shown as a black dot on the left. Since this is the equatorial slice, the black arrow ∇f lies in the plane of paper so the three arrows are coplanar. This means the three vectors ∇f, ∇a,∇b are coplanar so they are linearly *dependent*. For this point **a**, the equation (4.3.3) can and does exist, and therefore point **a** is the problem solution. This point is located inside the r = 6 unit sphere.

Once again, at an extremum point we should have df = 0. Moving an amount d**r** which is consistent with both constraints (meaning d**r** is in the z direction) we then have from (4.3.3),

$$df = \nabla f \bullet d\mathbf{r} \;=\; -\lambda_1[\nabla a(\mathbf{r})\bullet d\mathbf{r}] \;-\; \lambda_2\,[\nabla b(\mathbf{r})\bullet d\mathbf{r}] \;= -\lambda_1\,[0] \;-\lambda_2[0] \;= 0 \qquad (4.3.5)$$

and sure enough, df = 0. At point **b** one finds for a d**r** pointed down toward the equatorial plane,

$$df = \nabla f \bullet d\mathbf{r} \;=\; [\,-(x/f)\hat{\mathbf{x}} - (y/f)\hat{\mathbf{y}} \;-(z/f)\hat{\mathbf{z}}\,] \bullet (-|dr|\,\hat{\mathbf{z}}\,) = (z/f)dr \;> 0 \qquad (4.3.6)$$

and so it is advantageous to move $d\mathbf{r} = |dr|\hat{\mathbf{z}}$ toward **a** and thereby increase f, so **b** is not an extremum. Evaluating (4.3.6) at z = 0 for point **a** again shows df = 0.

Finally, we solve the problem. Inserting (4.3.2) into (4.3.3) gives

$$\nabla f(\mathbf{r}) = -\lambda_1\nabla a(\mathbf{r}) \;-\lambda_2\,\nabla b(\mathbf{r}) \qquad \Rightarrow$$

$$[-\frac{x}{\sqrt{4 - x^2 - y^2 - z^2}}\,\hat{\mathbf{x}} - \frac{y}{\sqrt{4 - x^2 - y^2 - z^2}}\,\hat{\mathbf{y}} \;-\frac{z}{\sqrt{4 - x^2 - y^2 - z^2}}\,\hat{\mathbf{z}}\,] = -\;\lambda_1\hat{\mathbf{y}} \;-\; \lambda_2\,\hat{\mathbf{x}}\,. \qquad (4.3.7)$$

We see at once that z = 0 and then the above becomes

$$\frac{x}{\sqrt{4 - x^2 - y^2}} = \lambda_2 \qquad \frac{y}{\sqrt{4 - x^2 - y^2}} = \lambda_1 \qquad . \qquad (4.3.8)$$

But the constraints say x = 1 and y = 1 so,

$$\frac{1}{\sqrt{4 - 1^2 - 1^2}} = \lambda_2 \qquad \frac{1}{\sqrt{4 - 1^2 - 1^2}} = \lambda_1 \qquad \Rightarrow \qquad \lambda_1 = \lambda_2 = 1/\sqrt{2}\ \ .1 \qquad (4.3.9)$$

Therefore the solution to Example 2 is this:

$$\mathbf{r} = (1,1,0) \qquad \lambda_1 \;=\; 1/\sqrt{2} \qquad \lambda_2 \;=\; 1/\sqrt{2} \qquad (4.3.10)$$

It is simple matter conceptually to generalize our Example 2 to more complicated functions f,a,b. For some arbitrary function f(**r**) there will be a set of 2D surfaces in E^3 which are the level surfaces on which f(**r**) = K for various values of K. These level surfaces then replace the set of concentric spheres of our simple case. We then imagine replacing our simple planar constraint functions a(x,y,z) and b(x,y,z) with

general functions which then result in arbitrarily shaped 2D constraint surfaces a(x,y,z) = 0 and b(x,y,z) = 0 in E^3. The gradients ∇a and ∇b vary over their respective surfaces (being normal vectors). The intersection of these two 2D constraint surfaces in E^3 will be a curve in E^3 (1D surface in E^3). Potential solution points **r** must lie on this curve. At each point **r** on this curve, the gradients ∇a(**r**) and ∇b(**r**) define a local plane. At a stationary point one will find that the gradient ∇f(r) lies in the same plane as ∇a(**r**) and ∇b(**r**) so the three gradients are linearly dependent and then $\nabla f(\mathbf{r}) + \lambda_1 \nabla a(\mathbf{r}) + \lambda_2 \nabla b(\mathbf{r}) = 0$ for some λ_1 and λ_2. At such a point **r**, if one considers any d**r** which lies on both constraint surfaces, one will then find that df = ∇f •d**r** = 0 so the point **r** is therefore a stationary point.

The situation is described by a 3D version of drawing (4.2.17).

Note that the intersection of the two constraint surfaces might result in multiple curves, and each must then be considered. For example, the intersection of two thin (prolate) ellipsoids might be two closed curves. If the intersection of the constraint surfaces is null (1st ellipsoid inside 2nd), the problem has no solutions.

4.4 The General Case

Here we summarize what has been demonstrated in the above examples. We have the general gradient equation (4.1.2) emerging from the Lagrange Multiplier analysis,

$$\nabla f(\mathbf{r}) + \lambda_1 \nabla a(\mathbf{r}) + \lambda_2 \nabla b(\mathbf{r}) + ... + \lambda_{s-1} \nabla q(\mathbf{r}) = 0 \qquad \nabla = (\partial_1, \partial_2 ... \partial_N) \ . \qquad (4.1.2)$$

This equation only exists if **r** is an extremum!! Solving for ∇f(**r**) and dotting with d**r** gives

$$df(\mathbf{r}) = \nabla f(\mathbf{r}) \bullet d\mathbf{r} = -\lambda_1 [\nabla a(\mathbf{r}) \bullet d\mathbf{r}] - \lambda_2 [\nabla b(\mathbf{r}) \bullet d\mathbf{r}] - ..... - \lambda_{s-1} [\nabla q(\mathbf{r}) \bullet d\mathbf{r}] \ . \qquad (4.4.1)$$

If d**r** at point **r** is chosen so *all* the constraints are respected (that is, we stay on all the constraint surfaces as we displace d**r** from **r**) , then $\nabla a(\mathbf{r}) \bullet d\mathbf{r} = 0$, $\nabla b(\mathbf{r}) \bullet d\mathbf{r} = 0$ ..... and $\nabla q(\mathbf{r}) \bullet d\mathbf{r} = 0$. Then

$$df(\mathbf{r}) = \nabla f(\mathbf{r}) \bullet d\mathbf{r} = -\lambda_1 [0] - \lambda_2 [0] - ..... - \lambda_{s-1} [0] = 0 \ . \qquad (4.4.2)$$

Since df(**r**) = 0, the point **r** is a stationary point of f(**r**) subject to these constraints. The point **r** is then a candidate for being an extremum point of f(**r**).

5. Example 3: 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 3 has two 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. This 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:

$$x^2/A^2 + y^2/B^2 = 1 \qquad \text{ellipse centered at the origin, semimajor/minor axes A and B}$$

$$(x'-\alpha)^2 + (y'-\beta)^2 = R^2 \; . \qquad \text{circle of radius R centered at } (\alpha,\beta) \tag{5.1}$$

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

The distance squared is given by

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

while the constraint functions are

$$a(x,y,x',y') \; = \; x^2/A^2 + y^2/B^2 - 1 \qquad a(x,y,x',y') \; = 0$$

$$b(x,y,x',y') \; = \; (x'-\alpha)^2 + (y'-\beta)^2 - R^2 \qquad b(x,y,x',y') = 0 \quad . \tag{5.3}$$

Notice that the vector **r** for this problem is **r** = (x,y,x',y').

The various derivatives are,

$$\begin{array}{lll} f_1 = \partial_x f \; = 2(x-x') & a_1 = \partial_x a \; = 2x/A^2 & b_1 = \partial_x b = 0 \\ f_2 = \partial_y f \; = 2(y-y') & a_2 = \partial_y a \; = 2y/B^2 & b_2 = \partial_y b = 0 \\ f_3 = \partial_{x'} f = -2(x-x') & a_3 = \partial_{x'} a \; = 0 & b_3 = \partial_{x'} b = 2(x'-\alpha) \\ f_4 = \partial_{y'} f = -2(y-y') & a_4 = \partial_{y'} a \; = 0 & b_4 = \partial_{y'} b = 2(y'-\beta) \; . \end{array} \tag{5.4}$$

The H function of (2.1) is then,

$$H(x,y,x',y',\lambda_1,\lambda_2) = \; f \; + \; \lambda_1 a + \lambda_2 \, b$$

$$= \; (x-x')^2 + (y-y')^2 \; + \lambda_1 \, [x^2/A^2 + y^2/B^2 - 1] + \lambda_2 \, [(x'-\alpha)^2 + (y'-\beta)^2 - R^2] \; . \tag{5.5}$$

Compute the four first derivatives as in (2.2),

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

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

$$(x-x') + \lambda_1 x/A^2 = 0 \qquad 1$$
$$(y-y') + \lambda_1 y/B^2 = 0 \qquad 2$$
$$-(x-x') + \lambda_2 (x'-\alpha) = 0 \qquad 3$$
$$-(y-y') + \lambda_2 (y'-\beta) = 0 \qquad 4$$

$$x^2/A^2 + y^2/B^2 - 1 = 0 \qquad 5$$
$$(x'-\alpha)^2 + (y'-\beta)^2 - R^2 = 0 \qquad 6 \qquad (5.7)$$

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

Notice that a possible λ_i 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

$$k_1y^4 + k_2y^3 + k_3 y^2 + k_4 y + k_4 = 0. \qquad x = \pm A\sqrt{1-(y/B)^2} \qquad (5.8)$$

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,

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

where

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

so that

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

Therefore,

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

and thus

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

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

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

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

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

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

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

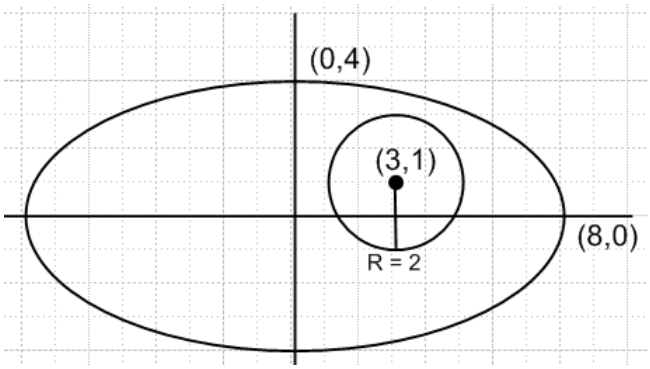
Including the last two equations (5.7), one has four equations in four unknowns x,y,x',y' :

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

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 (α,β) = (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',λ_1,λ_2 (xp = x', L1 = λ_1, etc) :

```
e1 := (x-xp) + L1*(x/A^2) = 0;
```

$$e1 := x - xp + \frac{L1\,x}{A^2} = 0$$

```
e2 := (y-yp) + L1*(y/B^2) = 0;
```

$$e2 := y - yp + \frac{L1\,y}{B^2} = 0$$

```
e3 := -(x-xp)+ L2*(xp-alpha) = 0;
```

$$e3 := -x + xp + L2\,(xp - \alpha) = 0$$

```
e4 := -(y-yp)+ L2*(yp-beta) = 0;
```

$$e4 := -y + yp + L2\,(yp - \beta) = 0$$

```
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 := (xp-alpha)^2 + (yp-beta)^2 - R^2 = 0;
```

$$e6 := (xp - \alpha)^2 + (yp - \beta)^2 - R^2 = 0$$

(5.16)

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

```
A := 8; B := 4; R := 2; alpha := 3; beta := 1;
```

$$A := 8$$
$$B := 4$$
$$R := 2$$
$$\alpha := 3$$
$$\beta := 1$$

(5.17)

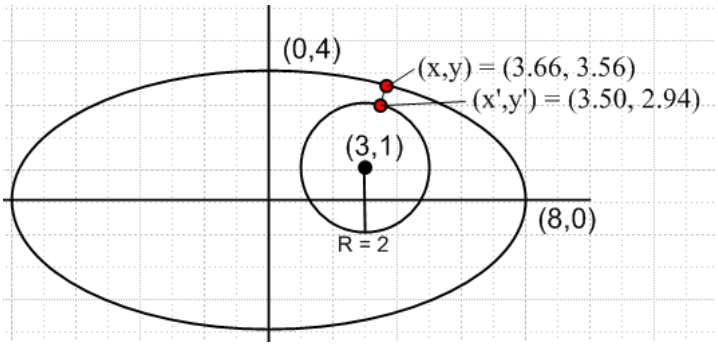
Without given a starting point, Maple finds the following solution

```
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

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

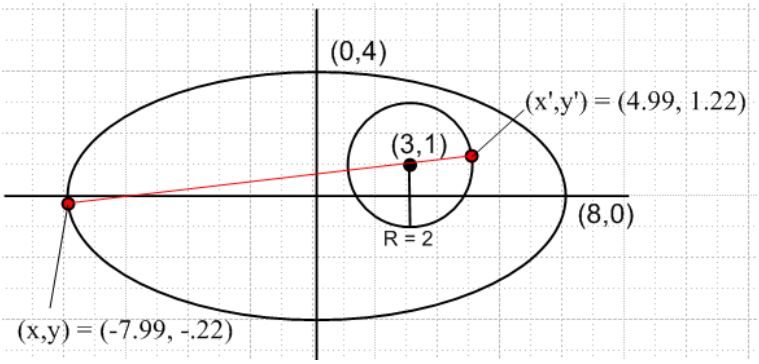
Another solution is found by giving Maple a search starting point (-8,0) on the ellipse and (5,1) on the circle :

```
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

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

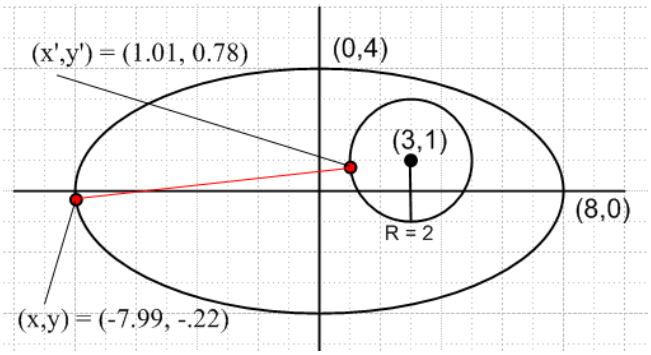
A third candidate solution is found by giving Maple a starting point (-8,0) on the ellipse and (1,1) on the circle:

```
fsolve({e1,e2,e3,e4,e5,e6},{x=-8,y=0,xp=1,yp=1,L1,L2});
{x = -7.987657508, y = -.2221078222, yp = .7789123247, L1 = -72.11057310, L2 = 4.527706691, xp = 1.012257502}
```

which we approximate as

$\lambda_1 = -72 \qquad \lambda_2 = 4.5 \qquad (x,y) = (-7.99, -.22) \qquad (x',y') = (1.01, 0.78)$. (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.

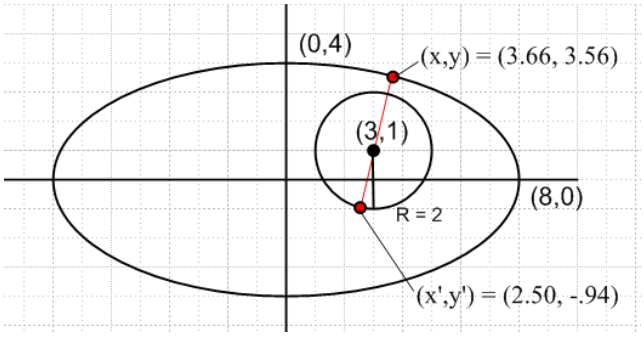
Finally, we find this fourth solution,

```
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

$\lambda_1 = -20.2 \quad \lambda_2 = -2.3 \qquad (x,y) = (3.66, 3.56) \qquad (x',y') = (2.50, -.94)$. (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 4: 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 = $(N_1,N_2...N_m)$ = **N**, a vector. The total number of particles in the system is fixed at $M = \Sigma_i N_i$, and the total energy is $U = \Sigma_i \varepsilon_i N_i$. 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 $\{N_i\}$ = **N**. If M is very large, one will find that those billion **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 **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 $g_i >> N_i$ (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:

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

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

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

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

Find **N** which maximizes $\Omega(\mathbf{N}) = \frac{g_1^{N_1}}{N_1!}\frac{g_2^{N_2}}{N_2!} ....\frac{g_m^{N_m}}{N_m!}$ subject to these two constraints:

$$\Sigma_i N_i = M \qquad \text{and} \qquad \Sigma_i \varepsilon_i N_i = U \ . \qquad (6.1.4)$$

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 **N** which is the solution value **N** for this extremum problem. Thus, when one examines an ensemble of such systems, this solution **N** is the overwhelmingly likely partitioning of the particles into $\{N_i\}$. Our task is then to solve this problem for **N**.

This vector $\mathbf{N} = (N_1, N_2...N_m)$ is the vector $\mathbf{r} = (x_1,x_2...x_N)$ 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 **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.

$$M\varepsilon_{min} \le U \le M\varepsilon_{max} \ . \qquad (6.1.5)$$

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,

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

and then (6.1.5) says,

$$\varepsilon_{min} \le u \le \varepsilon_{max} \ . \qquad (6.1.7)$$

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),

$$f(N_1, N_2...N_m) \equiv \ln\Omega(N_1,N_2...N_m) = \Sigma_i [N_i \ln g_i] - \Sigma_i [\ln N_i!] \qquad (6.2.1)$$

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

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

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

$$f = \ln\Omega = \Sigma_i [N_i \ln g_i] - \Sigma_i [\ln N_i!] \approx \Sigma_i [N_i \ln g_i] - \Sigma_i [N_i \ln N_i - N_i]$$

$$= \Sigma_i [(1+\ln g_i)N_i - N_i \ln N_i] = \Sigma_i N_i [1 + \ln(g_i/N_i)] . \qquad (6.2.3)$$

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

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

$a(\mathbf{N}) = \Sigma_i N_i - M = 0$

$b(\mathbf{N}) = \Sigma_i \varepsilon_i N_i - U = 0$. (6.2.4)

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

$$H(\mathbf{N},\boldsymbol{\lambda}) \equiv f(\mathbf{N}) + \lambda_1 a(\mathbf{N}) + \lambda_2 b(\mathbf{N})$$

$$= \Sigma_i [(1+\ln g_i)N_i - N_i \ln N_i] + \lambda_1 [\Sigma_i N_i - M] + \lambda_2 [\Sigma_i \varepsilon_i N_i - U] . \qquad (6.2.5)$$

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

$$f_i = \partial f/\partial N_i = (1+\ln g_i) - N_i (1/N_i) - \ln N_i = \ln g_i - \ln N_i = \ln(g_i/N_i)$$

$$a_i = \partial a/\partial N_i = \partial_i [\Sigma_j N_j - M] = 1$$

$$b_i = \partial b/\partial N_i = \partial_i [\Sigma_j \varepsilon_j N_j - U] = \varepsilon_i . \qquad (6.2.6)$$

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

$$0 = H_i(\mathbf{N},\boldsymbol{\lambda}) = f_i + \lambda_1 a_i + \lambda_2 b_i = \ln(g_i/N_i) + \lambda_1 + \lambda_2 \varepsilon_i \quad i = 1,2...m$$

or

$$\ln(g_i/N_i) + \lambda_1 + \lambda_2 \varepsilon_i = 0 \qquad i = 1,2...m$$

or

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

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

$$\ln(N_i/g_i) = \lambda_1 + \lambda_2\varepsilon_i \qquad i = 1,2...m$$

$$\Sigma_i N_i = M \qquad // \; a(\mathbf{N}) = 0$$

$$\Sigma_i \varepsilon_i N_i = U \qquad // \; b(\mathbf{N}) = 0 \quad . \qquad (6.2.8)$$

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:

$$\ln(N_i/g_i) = \lambda_1 + \lambda_2\varepsilon_i$$

or

$$N_i = g_i \, e^{\lambda_1} \, e^{\lambda_2\varepsilon_i} \quad . \qquad (6.2.9)$$

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

$$\beta \equiv -\lambda_2$$

$$A \equiv e^{\lambda_1} \qquad (6.2.10)$$

and then (6.2.9) reads,

$$N_i = A \; g_i \, e^{-\beta\varepsilon_i} \, . \qquad (6.2.11)$$

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

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

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 $S = k \ln\Omega = kf$ 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

$$p_i = \frac{N_i}{M} = \frac{A}{M} \; g_i \, e^{-\beta\varepsilon_i} \, . \qquad (6.2.13)$$

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

$$A = \frac{M}{\Sigma_i g_i\ e^{-\beta\varepsilon_i}}\ . \tag{6.2.14}$$

The denominator -is called the **partition function** for the system.

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

$$\begin{aligned} A\ \Sigma_i g_i e^{-\beta\varepsilon_i}\ &= M \\ A\ \Sigma_i \varepsilon_i g\ e^{-\beta\varepsilon_i}\ &= U\ . \end{aligned} \tag{6.2.15}$$

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

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

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

$$\Sigma_i g_i\varepsilon_i\ e^{-\beta\varepsilon_i}\ = U/A = (U/M)(M/A) = (u)\ (\ \Sigma_i g_i\ e^{-\beta\varepsilon_i}) \tag{6.2.17}$$

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 $\varepsilon_i < u$ (recall that $u = \langle\varepsilon_j\rangle$) so $[\varepsilon_i - u]$ will be negative. For high energy states [$\varepsilon_i - u$] 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 $N_i = A\ g_i\ e^{-\beta\varepsilon_i}$.

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 **N**, which then statistically forces a system to assume a value **N** very close to this solution **N**. We shall deal with these issues in the following section.

6.3 More details of the solution

Start with

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

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

$$\begin{aligned} N_1 &= +(\varepsilon_2 - \varepsilon_1)^{-1}[\varepsilon_2 M - U + \Sigma_{i=3}^{m} (\varepsilon_i - \varepsilon_2)N_i] &&= N_1(N_3, \ldots N_m) \\ N_2 &= -(\varepsilon_2 - \varepsilon_1)^{-1}[\varepsilon_1 M - U + \Sigma_{i=3}^{m} (\varepsilon_i - \varepsilon_1)N_i] &&= N_2(N_3, \ldots N_m) . \end{aligned} \qquad (6.3.1)$$

Notice that, for i = 3,4...m,

$$\begin{aligned} \frac{\partial N_1}{\partial N_i} &= +(\varepsilon_2 - \varepsilon_1)^{-1} [(\varepsilon_i - \varepsilon_2)] = \frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} \\ \frac{\partial N_2}{\partial N_i} &= -(\varepsilon_2 - \varepsilon_1)^{-1} [(\varepsilon_i - \varepsilon_1)] = -\frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} . \end{aligned} \qquad (6.3.2)$$

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

$$f(N_1(N_3, \ldots N_m), N_2(N_3, \ldots N_m), N_3, \ldots .N_m) . \qquad (6.3.3)$$

Compute the total derivative df/dN_i for i = 3,4...n:

$$\begin{aligned} \frac{df}{dN_i} &= f_1 \frac{\partial N_1}{\partial N_i} + f_2 \frac{\partial N_2}{\partial N_i} + f_i \qquad \text{where } f_i \equiv \frac{\partial f}{\partial N_i} \\ &= f_1 \frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} - f_2 \frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} + f_i \qquad (6.3.2) \\ &= \ln(g_1/N_1) \frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} - \ln(g_2/N_2) \frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} + \ln(g_i/N_i) . \qquad (6.2.6) \end{aligned} \qquad (6.3.4)$$

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

$$\begin{aligned} \frac{df}{dN_i} &= (\lambda_1 + \lambda_2 \varepsilon_1) \frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} - (\lambda_1 + \lambda_2 \varepsilon_2) \frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} + (\lambda_1 + \lambda_2 \varepsilon_i) \\ &= \lambda_1 \left[\frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} - \frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} + 1 \right] + \lambda_2 \left[\varepsilon_1 \frac{\varepsilon_i - \varepsilon_2}{\varepsilon_2 - \varepsilon_1} - \varepsilon_2 \frac{\varepsilon_i - \varepsilon_1}{\varepsilon_2 - \varepsilon_1} + \varepsilon_i \right] \end{aligned}$$

$$= \lambda_1 \left[\frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1} - \frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1} + \frac{\varepsilon_2-\varepsilon_1}{\varepsilon_2-\varepsilon_1} \right] + \lambda_2 \left[\varepsilon_1 \frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1} - \varepsilon_2 \frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1} + \varepsilon_i \frac{\varepsilon_2-\varepsilon_1}{\varepsilon_2-\varepsilon_1} \right]$$

$$= \lambda_1 [0] + \lambda_2[0] = 0 \ . \qquad (6.3.5)$$

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), and then using the first line of (6.3.4) sequentially for $f \rightarrow f_1$, f_2 and f_i , one gets

$$\frac{d^2f}{d^2N_i} = \frac{df_1}{dN_i} \frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1} - \frac{df_2}{dN_i} \frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1} + \frac{df_i}{dN_i}$$

$$= \left[f_{11} \frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1} - f_{12} \frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1} + f_{1i} \right] \frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1} \qquad f_{ij} \equiv \frac{\partial^2 f}{\partial N_i \partial N_j} = f_{ji}$$

$$- \left[f_{21} \frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1} - f_{22} \frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1} + f_{2i} \right] \frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1}$$

$$+ \left[f_{i1} \frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1} - f_{i2} \frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1} + f_{ii} \right] . \qquad i = 3,4....m \qquad (6.3.6)$$

Recall now that

$$f = \ln\Omega = \Sigma_i [(1+\ln g_i)N_i - N_i \ln N_i] \qquad \textit{(6.2.3)}$$

$$f_i = \ln g_i - \ln N_i \qquad \textit{(6.2.6)}$$

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

Therefore (6.3.6) simplifies to

$$\frac{d^2f}{d^2N_i} = f_{11} \left[\frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1} \right]^2 + f_{22} \left[\frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1} \right]^2 + f_{ii}$$

$$= - \left\{ \frac{1}{N_1} \left[\frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1} \right]^2 + \frac{1}{N_2} \left[\frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1} \right]^2 + \frac{1}{N_i} \right\} \ . \qquad (6.3.8)$$

This says that the curvature of f in all independent directions N_i for i = 3,4...m is negative and this is true for *all* vectors $\mathbf{N} = \{N_i\}$. This is of course then true for the solution vector **N**. But for the solution vector **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(**N**) in the i direction only (i = 3,4...m) about the solution **N**, then

$$f(N_i + dN_i) \approx f(N_i) + (df/dN_i)dN_i + (1/2)(d^2f/d^2N_i)\,(dN_i)^2 \ . \tag{6.3.9}$$

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

$$df = f(N_i + dN_i) - f(N_i) \approx (1/2)(d^2f/d^2N_i)\,(dN_i)^2 \tag{6.3.10}$$

so

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

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

$$\Omega_{half}/\Omega_{peak} = 1/2 \qquad \Rightarrow \qquad \ln \Omega_{half} - \ln \Omega_{peak} = 1/2 \qquad \Rightarrow$$

$$f_{half} - f_{peak} = \ln(1/2) \qquad \Rightarrow \qquad |df| = f_{peak} - f_{half} = \ln(2) = 0.7 \ . \tag{6.3.12}$$

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

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

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

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

so the fractional half-width is roughly.

$$\frac{dN_i}{M} \sim \sqrt{\frac{1}{M}} \ . \tag{6.3.15}$$

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 **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

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

and correspondingly (this uses the Stirling approximation),

$$f = \ln\Omega = \Sigma_{i=1}^{3} N_i\,[1 + \ln(g_i/N_i)] \,. \qquad \textit{(6.2.3)} \qquad (6.4.2)$$

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

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

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

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

It seems clear that one must have

$$\begin{aligned} &0 \le N_1 \le M \\ &0 \le N_2 \le M \,. \end{aligned} \qquad (6.4.5)$$

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

$$N_{3min} \le N_3 \le N_{3max}$$

$$N_{3min} = \max\left(\frac{u-\varepsilon_2}{\varepsilon_3-\varepsilon_2} M,\, 0\right)$$

$$N_{3max} = \min\left(\frac{u-\varepsilon_1}{\varepsilon_3-\varepsilon_1} M,\, M\right) = \frac{u-\varepsilon_1}{\varepsilon_3-\varepsilon_1} M \,. \qquad \text{// since } u \le \varepsilon_3 \qquad (6.4.6)$$

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

$$\Sigma_{i=1}^{3}\,[\,\varepsilon_i - u\,]\, g_i\, e^{-\beta\varepsilon_i} = 0 \,. \qquad \textit{(6.2.17)} \qquad (6.4.7)$$

Then the other derived multiplier is given by (6.2.14)

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

The components N_i of the solution vector are then given by

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

The curvature of f at its peak is,

$$\frac{d^2f}{d^2N_3} = -\{\ \frac{1}{N_1}\ [\frac{\varepsilon_3-\varepsilon_2}{\varepsilon_2-\varepsilon_1}\ ]^2 + \frac{1}{N_2}[\ \frac{\varepsilon_3-\varepsilon_1}{\varepsilon_2-\varepsilon_1}\ ]^2 + \frac{1}{N_3}\ \}\ . \qquad (6.3.8) \qquad (6.4.10)$$

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

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

For our specific Maple plots, we shall assume

$\varepsilon_3 = 3$	M = 1000 = number of particles	
$\varepsilon_2 = 2$	u = 1.5 = average energy	
$\varepsilon_1 = 1$	$g_1 = g_2 = g_3 = 5000$ = degeneracy	(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_{3min} and N_{3max}

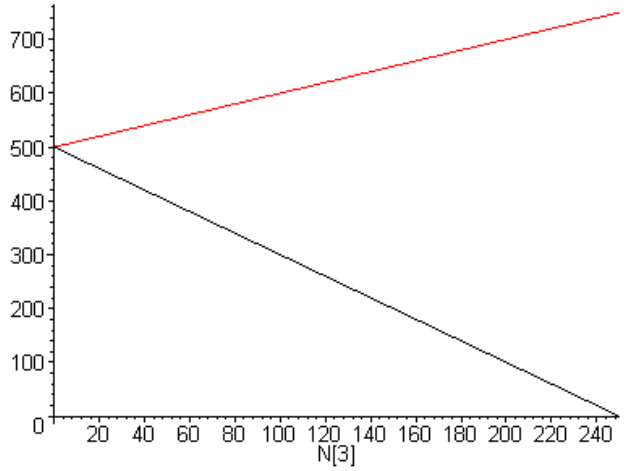
```
e[1] := 1: e[2] := 2: e[3] := 3: u := 1.5: M := 1000: U := M*u:
N3min := max(0,M*(u-e[2])/(e[3]-e[2]));
N3max := M*(u-e[1])/(e[3]-e[1]);
```

$$N3min := 0$$

$$N3max := 250.0000000$$

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;
```

$$\frac{5000^{(500.0+N_3)}\ 5000^{(500.0-2N_3)}\ 5000^{N_3}}{(500.0+N_3)!\,(500.0-2N_3)!\,N_3!} \tag{6.4.14}$$

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)

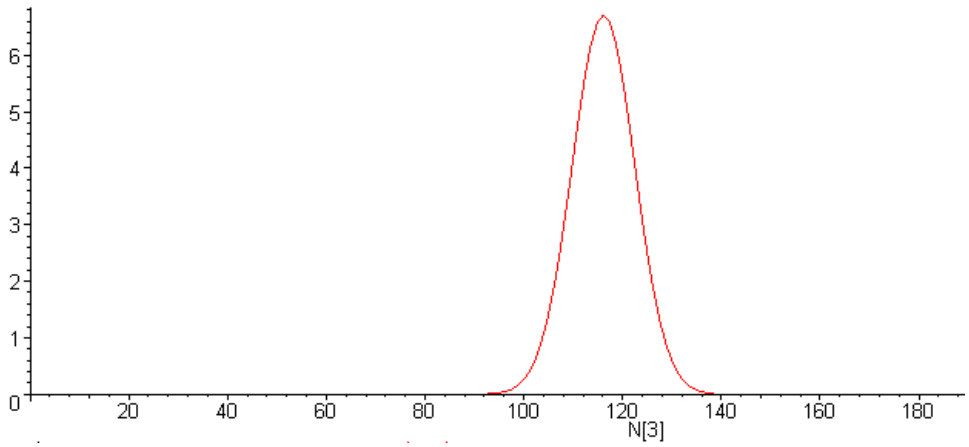
$$\Omega_{peak} = 6.71 \times 10^{1519} \tag{6.4.15}$$

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 Ω(N_3),

```
scale := 1e1519;
```

$$scale := .1\ 10^{1520}$$

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

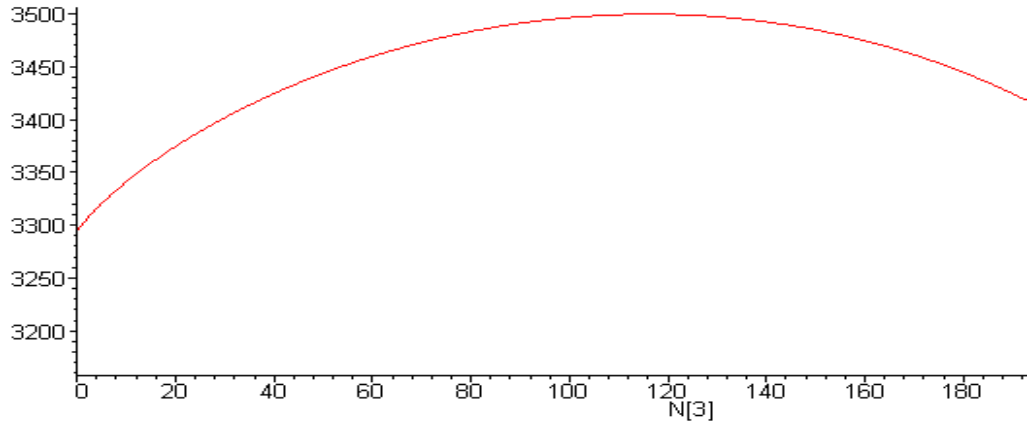


(6.4.16)

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 $\frac{d^2 f}{d^2 N_i}$ 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;
```

$$eq := -2500.\, e^{(-1.\,\beta)} + 2500.\, e^{(-2.\,\beta)} + 7500.\, e^{(-3.\,\beta)} = 0$$

```
s := solve(eq,beta);
```

$$s := .2644970943 - 3.141592654\, I,\ .8341151943$$

```
beta := s[2];
```

$$\beta := .8341151943$$

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,
```

$$N_3 := 116.2040605$$

$$N_2 := 267.5918793$$

$$N_1 := 616.2040607$$

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];
```

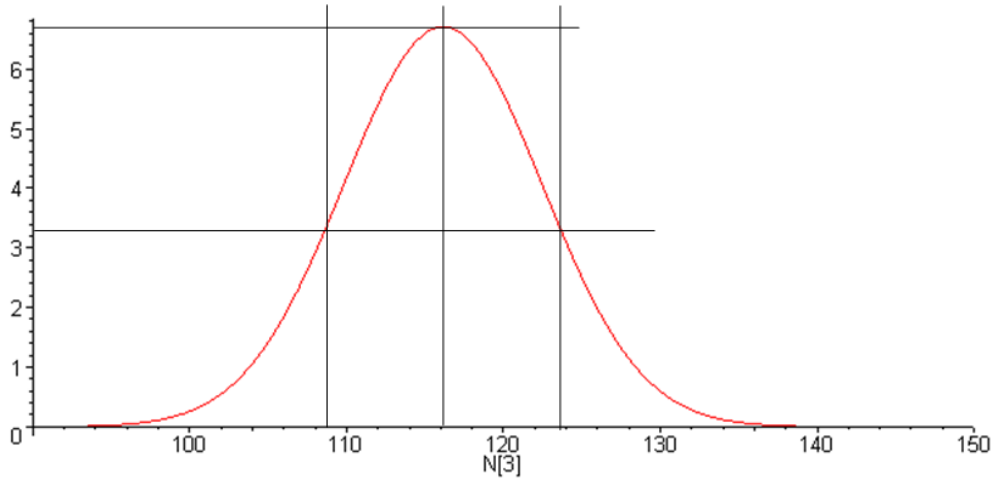
$$curve := .02517652703$$

```
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 Γ(N+1).

The peak value of Ω is found to be

```
evalf(Omega);
```

$$.6710572393\ 10^{1520}$$

In the above example we used u = 1.5 < ε_2 = 2. If we instead use u = 2.5 > ε_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 (β = -.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 : Matrix rank equals the number of independent columns and rows

This Appendix proves that, for a general m x n matrix M, the number of linearly independent columns and the number of linearly independent rows are the same, and both numbers are equal to the rank of the matrix. The reader will appreciate that this fact is non-obvious. Shilov proves it but the proof is a bit spread out over several sections. Here we provide a self-contained alternate proof.

We first quote a series of square-matrix theorems that are proved from scratch in our Appendix B. These are doubtless very familiar to the reader, but we just want to get them stated. The derivations of these theorems may be less familiar.

Theorem 1: $\det(M^{\mathbf{T}}) = \det(M)$. Switching rows with columns does not change a determinant. *(B.1.10)*

Theorem 2: det(M) can be represented in these two ways, where ε is the permutation tensor (B.2.1) :

$$\det(M) = \Sigma_{a_1 a_2 \cdots a_n} \varepsilon_{a_1 a_2 \cdots a_n} M_{1a_1} M_{2a_2} \cdots\cdots M_{na_n} \qquad \textit{(B.2.10a)}$$

$$\det(M) = \Sigma_{a_1 a_2 \cdots a_n} \varepsilon_{a_1 a_2 \cdots a_n} M_{a_1 1} M_{a_2 2} \cdots\cdots M_{a_n n} \ . \qquad \textit{(B.2.10b)}$$

Theorem 3: For a square matrix M, adding a multiple of one row to another does not change det(M). The same is true for adding a multiple of one column to another. *(B.2.12)*

Theorem 4: Swapping two rows (columns) of square matrix M causes $\det(M) \rightarrow -\det(M)$. *(B.2.13)*

Corollary 4: If two rows (columns) of square matrix M are the same, then det(M) = 0. *(B.2.14)*

Theorem 5: det(AB) = det(A)det(B) *(B.2.15)*

Corollary 5: det(ABC) = det(A)det(B)det(C) and so on. *(B.2.16)*

Theorem 6 (Cofactor Expansions):

$$\det(M) = \Sigma_n M_{sn} \operatorname{cof}(M_{sn}) = \Sigma_n (\mathbf{r}_s)_n \operatorname{cof}(M_{sn}) \quad \text{work across row s} \quad s = 1,2....n \qquad \textit{(B.3.16a)}$$

$$\det(M) = \Sigma_n M_{ns} \operatorname{cof}(M_{ns}) = \Sigma_n (\mathbf{c}_s)_n \operatorname{cof}(M_{ns}) \quad \text{work down column s} \quad s = 1,2....n \qquad \textit{(B.3.16b)}$$

Theorem 7: $$M^{-1} = \frac{C^{\mathbf{T}}}{\det(M)} = \frac{[\operatorname{cof}(M)]^{\mathbf{T}}}{\det(M)} = \frac{\operatorname{cof}(M^{\mathbf{T}})}{\det(M)}$$ for square matrix M. *(B.4.8)*

Theorem 8 (Cramer's Rule):

$$\mathbf{y} = M\mathbf{x} \quad \Rightarrow \quad x_s = \frac{\det(M[\mathbf{c}_s \rightarrow \mathbf{y}])}{\det(M)} \qquad \textit{(B.4.14)}$$

To this list, we shall now add four new Theorems which will be proven right here. The definition of **linear independence** and **linear dependence** is given in the discussion surrounding (1.6) and won't be repeated here.

Theorem 9: Columns of square matrix M are linearly dependent $\Leftrightarrow$ det(M) = 0 (A.1)
Contrapositives: Columns of square matrix M are linearly independent $\Leftrightarrow$ det(M) $\neq$ 0

If we can prove Theorem 9 as stated, then the theorem is also true for rows. The reason is that swapping rows and columns corresponds to M $\leftrightarrow$ $M^{\mathbf{T}}$ and **Theorem 1** says det($M^{\mathbf{T}}$) = det(M).

Proof of $\Rightarrow$: According to (1.6) and nearby discussion, our premise of linear dependence says that *either* one or more columns of M are zero, *or* at least one column can be written as a linear combination of the other columns, so perhaps $\mathbf{c_j} = \Sigma_i k_i \mathbf{c_i}$. If a column is zero, det(M) = 0 from **Theorem 6** going down that column. If $\mathbf{c_j} = \Sigma_{i\neq j} k_i \mathbf{c_i}$, one can add $-\Sigma_{i\neq j} k_i \mathbf{c_i}$ to $\mathbf{c_j}$ without changing det(M) according to **Theorem 3**. But this makes the new column j vanish, so again det(M) = 0.

Proof of $\Leftarrow$: Here we shall prove the contrapositive instead of the claim:

claim:	det(M) = 0 $\Rightarrow$ the columns of M are linearly dependent
contrapositive:	the columns of M are linearly independent $\Rightarrow$ det(M) $\neq$ 0

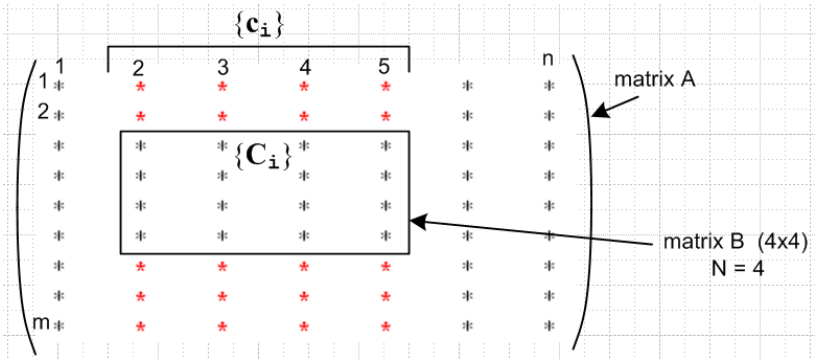
If the column vectors $\mathbf{c_i}$ of matrix M are linearly independent, then (1.6) says $\Sigma x_i \mathbf{c_i} = 0 \Rightarrow \mathbf{x} = 0$. Therefore we know that $\Sigma x_i M_{ji} = 0 \Rightarrow x_j = 0$ which is the same as $M\mathbf{x} = 0 \Rightarrow \mathbf{x} = 0$. Thinking of $M{:}E^n \rightarrow E^n$, we claim that mapping $M\mathbf{x} = \mathbf{y}$ is one-to-one. Certainly each **x** goes into a single **y**, Could a **y** map back into two different values of **x**, call them **x** and **x'** with $\mathbf{x} \neq \mathbf{x}'$? Then $M\mathbf{x} = \mathbf{y}$ and $M\mathbf{x}' = \mathbf{y}$. Subtract to get $M(\mathbf{x}-\mathbf{x}') = 0$. But since $M\mathbf{z} = 0 \Rightarrow \mathbf{z} = 0$ (as just shown), we find $\mathbf{x} = \mathbf{x}'$. Thus, the mapping $M\mathbf{x} = \mathbf{y}$ really is one-to-one, and that means it is invertible and the inverse is unique. From **Theorem 7** the inverse is in fact given by $M^{-1} = \text{cof}(M^{\mathbf{T}})/\det(M)$. For M^{-1} to exist, one must have det(M) $\neq$ 0.

Another proof is very simple but perhaps less convincing. One can show (see our Tensor Analysis document (B.6.10) in Refs) that if a set of column vectors $\mathbf{c_i}$ of matrix M spans an n-piped in E^n, the "volume" of that n-piped is given by V = |det(M)|. If those vectors are linearly independent, then V $\neq$ 0. This is obvious in 2D (volume = area) and 3D but less obvious for general E^n. For example, in 3D if a third vector lies in the plane of the other two (is dependent), there is no volume.

Theorem 10 (mini-columns theorem). Let R be a subset of N rows of nxm matrix A, and let S be a subset of columns $\{\mathbf{c_i}\}$ in A. The matrix elements of A included in the intersection of these two sets form a set of "mini-columns" $\{\mathbf{C_i}\}$. This set of mini-columns forms a matrix B within A. Matrix B could be contiguous, or it could be non-contiguous in one or both directions. The claim and its contrapositive are:

(a) $\{\mathbf{C_i}\}$ linearly independent in B $\Rightarrow$ $\{\mathbf{c_i}\}$ linearly independent in A
(b) $\{\mathbf{c_i}\}$ linearly dependent in A $\Rightarrow$ $\{\mathbf{C_i}\}$ linearly dependent in B (A.2)

A picture is worth a thousand words. Here matrix B happens to be contiguous in both directions.



(A.3)

Proof of (b): If the set $\{\mathbf{c}_i\}$ is linearly dependent in A, then from (1.6) one can write $\Sigma_{i\in S}\, k_i\mathbf{c}_i = 0$ with $\mathbf{k} \neq 0$. This is a set of m equations which contains as a subset the set of N equations $\Sigma_{i\in S}\, k_i\mathbf{C}_i = 0$. This last equation then says that the $\{\mathbf{C}_i\}$ are linearly dependent in B. There is of course a similar "mini rows theorem".

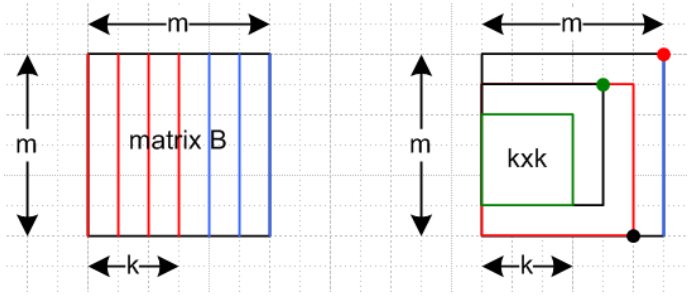
Theorem 11. If all kxk minors in k columns of matrix A vanish, the k columns are linearly dependent.

(A.4)

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

Once proven for columns, replacing $A \rightarrow A^T$ gives the same theorem for rows.

Proof: Assume matrix A has m rows and n columns. In order for kxk 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.5)

Consider now the process of computing det(B) by going down the rightmost column using the standard cofactor sum formula of **Theorem 6**. This det(B) is a linear combination of (m-1)x(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)x(m-2) minors in the left m-2 columns. We keep going until we are evaluating a set of kxk cofactors in the left k columns. But these all vanish by the theorem premise. Thus, reversing this logic 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) ≠ 0 by **Theorem 9**. 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.

We are finally ready for the Main Act. Recall first that :

Definition: The **rank** r of an m x n matrix A is the dimension of the largest non-vanishing minor within A. [Shilov 1.92] Thus, r ≤ min(m,n). *(1.5)* (A.6)

Theorem 12. For a general nxm matrix A, (A.7)

rank(A) = r ⇔ A has exactly r linearly independent columns

Once this theorem is proved for columns, it is also true for rows since rank(A^T) = rank(A).

Proof of ⇒: If rank(A) = r, A must have at least one non-vanishing rxr minor. Think of this minor being a set of mini-columns as in Theorem 10. Since then det(minor) ≠ 0, this set of mini-columns is linearly independent according to **Theorem 9**. Thus, the set of corresponding full columns (those that pass down through the minor) are linearly independent from **Theorem 10** (a). So matrix A has at least these r independent columns.

Now let k = r+1. Since rank(A) = r we know that all kxk minors in A vanish. By **Theorem 11**, 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 just r, the number of columns passing down through the minor.

Proof of ⇐: If there are r linearly independent columns, then at least one rxr minor within those columns must be non-zero (contrapositive of **Theorem 11**). Suppose there were an (r+k)x(r+k) non-vanishing minor. Then the r+k mini-columns of that minor would be independent, and thus by **Theorem 10** there would be r+k independent full columns. But the theorem premise says there are only r independent columns, so all minors larger than rxr must vanish. Therefore rank(A) = r.

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**. Theorem 12 and its row version show that

column rank = row rank = rank = dimension of largest non-vanishing minor (A.8)

Basis Minors and Basis Columns. If rank(A) = r, we know there must exist at least one non-vanishing rxr 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 10** 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 $\mathbf{c_2} = \sum_{i=1}^{k} k_i \mathbf{c_i} = 1\ \mathbf{c_2}$. If k = r+1, all kxk minors vanish and by **Theorem 11** 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.9)

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

Appendix B : Determinants

There must be hundreds of definitions and theorems inhabiting Matrix World. Here we provide our own "take" on a small number of them relating to the determinant of a square matrix. The reader might be surprised to see "group theory" playing a major role in the development below. There are many ways to skin a cat, and the following thread is just one of those ways. Most of the Theorems proven below are used in Appendix A.

B.1 Definition of the determinant

First, we define $\mathbf{z_0}$ to be the n-component vector of increasing integers 1 to n,

$$\mathbf{z_0} \equiv \begin{pmatrix} 1 \\ 2 \\ ... \\ n \end{pmatrix} . \tag{B.1.1}$$

Let **a** be some **permutation** (reordering) of these integers, so write

$$\mathbf{a} \equiv \begin{pmatrix} a_1 \\ a_2 \\ ... \\ a_n \end{pmatrix} = A \begin{pmatrix} 1 \\ 2 \\ ... \\ n \end{pmatrix} = A\,\mathbf{z_0} \tag{B.1.2}$$

where A is an nxn matrix which has 1's in the right places to create this permutation vector **a**. For example,

$$\mathbf{a} = A\mathbf{z_o} \qquad \leftrightarrow \qquad \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} .$$

The vector **a** can be obtained from the vector $\mathbf{z_0}$ by making $S_\mathbf{a}$ pairwise swaps of elements of $\mathbf{z_0}$. Although $S_\mathbf{a}$ is not unique, the number $(-1)^{S_\mathbf{a}}$ is unique. For example, to get from (1,2,3) to (1,3,2) one could swap the second pair so $S_\mathbf{a} = 1$, but one could then swap the first pair twice and then $S_\mathbf{a} = 3$. This number $(-1)^{S_\mathbf{a}}$ is of course ±1 and we shall call it the **parity** of the permutation **a**,

$$\text{Parity}(\mathbf{a}) \equiv (-1)^{S_\mathbf{a}} . \tag{B.1.3}$$

Fact: $$\text{Parity}(C\mathbf{a}) = \text{Parity}(CA\mathbf{z_0}) = (-1)^{S_\mathbf{c}+S_\mathbf{a}} = \text{Parity}(C^{-1}\mathbf{a}) \tag{B.1.4}$$

Proof: Doing permutation CA involves first doing A with its $S_\mathbf{a}$ swaps, and then doing C with its $S_\mathbf{c}$ swaps, for a total of $S_\mathbf{a}+S_\mathbf{c}$ swaps. The inverse permutation C^{-1} obviously involves the same number of swaps as the permutation C.

Moving now toward the definition of the determinant of a matrix M, define (Π meaning "product")

$$\Pi(\mathbf{a},\mathbf{b};\,M) \equiv M_{\mathbf{ab}} \equiv M_{a_1b_1}M_{a_2b_2} \,.....\, M_{a_nb_n} \;=\; \text{product of n factors} \;. \qquad \text{(B.1.5)}$$

The notation $\Pi(\mathbf{a},\mathbf{b};\,M)$ is easier to deal with than $M_{\mathbf{ab}}$, but we shall use both these notations.

Suppose $\mathbf{c} = C\mathbf{z_0}$ is some arbitrary permutation of $\mathbf{z_0}$. Then:

Fact: $\Pi(C\mathbf{a},C\mathbf{b};\,M) = \Pi(\mathbf{a},\mathbf{b};\,M)$ (B.1.6)

Proof: Applying the same permutation to both the a_k and b_k indices of $M_{a_1b_1}M_{a_2b_2} \,.....\, M_{a_nb_n}$, just reorders the terms in the product but the product stays the same.

Fact: $\Pi(\mathbf{a},\mathbf{b};\,M^T) = \Pi(\mathbf{b},\mathbf{a};\,M)$ (B.1.7)

Proof: $\Pi(\mathbf{a},\mathbf{b};\,M^T) = M^T_{a_1b_1}M^T_{a_2b_2} \,....\, M^T_{a_nb_n} \;=\; M_{b_1a_1}M_{b_2a_2} \,....\, M_{b_na_n} \;=\; \Pi(\mathbf{b},\mathbf{a};\,M)$

We shall now **define the determinant** of an n x n matrix M in the following admittedly obscure manner (later we will show that it reduces to more familiar forms),

$$\det(M) \;\equiv\; \frac{1}{n!}\,\Sigma_{\mathbf{a}}\,\Sigma_{\mathbf{b}}\;(-1)^{S_{\mathbf{a}}+S_{\mathbf{b}}}\;\Pi(\mathbf{a},\mathbf{b};\,M) \;=\; \frac{1}{n!}\,\Sigma_{\mathbf{a},\mathbf{b}}\;(-1)^{S_{\mathbf{a}}+S_{\mathbf{b}}}\;M_{\mathbf{ab}}$$

$$=\; \frac{1}{n!}\,\Sigma_{\mathbf{a}}\,\Sigma_{\mathbf{b}}\,(-1)^{S_{\mathbf{a}}+S_{\mathbf{b}}}\,M_{a_1b_1}M_{a_2b_2} \,.....\, M_{a_nb_n}\;. \qquad \text{(B.1.8)}$$

Here $\Sigma_{\mathbf{a}}$ means the sum over all permutations $\mathbf{a}$ of $\mathbf{z_0}$. In (B.1.8) the columns and rows of M are on a completely equal footing. Notice that $\mathbf{a}$ and $\mathbf{b}$ are in effect dummy summation indices. If we do $\mathbf{a}\leftrightarrow\mathbf{b}$, the expression for det(M) is unchanged. Thus one can rewrite (B.1.8) as

$$\det(M) \;=\; \frac{1}{n!}\,\Sigma_{\mathbf{b}}\,\Sigma_{\mathbf{a}}\;(-1)^{S_{\mathbf{b}}+S_{\mathbf{a}}}\;\;\Pi(\mathbf{b},\mathbf{a};\,M)\;. \qquad \text{(B.1.9)}$$

We are now ready for our first determinant theorem:

Theorem 1: $\det(M^T) = \det(M)$. Switching rows with columns does not change a determinant. (B.1.10)

Proof:

$$\det(M^T) \;=\; \frac{1}{n!}\,\Sigma_{\mathbf{b}}\,\Sigma_{\mathbf{a}}\,(-1)^{S_{\mathbf{b}}+S_{\mathbf{a}}}\;\;\Pi(\mathbf{b},\mathbf{a};\,M^T) \qquad \text{// (B.1.9) applied to } M^T$$

$$=\; \frac{1}{n!}\,\Sigma_{\mathbf{b}}\,\Sigma_{\mathbf{a}}\,(-1)^{S_{\mathbf{b}}+S_{\mathbf{a}}}\;\;\Pi(\mathbf{a},\mathbf{b};\,M) \qquad \text{// (B.1.7) applied to } M^T$$

$$= \det(M) \qquad \text{// (B.1.8)}$$

B.2 The permutation group and the permutation tensor ε

Definition. A set of elements $\{g_i\}$ form a **group** G if :

- $g_i g_j$ is also in the group (closure)
- g_i^{-1} exists for each g_i in G, where g_i^{-1} is also in G (inverse)
- $(g_i g_j)g_k = g_i(g_j g_k)$ (associative) (B.2.1)

Comment: It is implicit in the above definition that the group has some "operation" which gives meaning to $g_i g_j$, which we shall just think of as "multiplication". When group elements are represented by matrices, that operation is multiplication of those matrices, and that will apply to the permutation group below.

Fact: $g_i G = G$ (the rearrangement theorem) (B.2.2)

This says that multiplication of all the elements of a group by an element g_i in the group creates a reordering of the group elements.

Proof. Consider $g_i G = g_i\ [g_1, g_2....g_n] = [g_i g_1, g_i g_2....g_i g_n]$ = set of n elements. Unless two elements are the same, this must exhaust the entire group. How do we know that $g_i g_1$ and $g_i g_2$ might not be the same? Apply g_i^{-1} from the left and that would say $g_1 = g_2$ which is not the case.

Fact: $\Sigma_g f(g) = \Sigma_g f(g_1 g)$ if Σ_g runs over the entire group G (B.2.3)

Proof: In $\Sigma_g f(g_1 g)$, as g runs over G, the argument $g' \equiv g_1 g$ runs over G by the rearrangement theorem. Thus, the sum $\Sigma_g f(g_1 g)$ is just a reordering of the terms in the sum $\Sigma_g f(g)$.

It is easy to show that the set of permutations of $\mathbf{z}_0$ forms a group and therefore the above facts can be used. For example, the product of two permutations is a permutation, and every permutation clearly has an inverse, and (AB)C = A(BC) for any matrices. Fact (B.2.3) can be written

$$\Sigma_B f(B\mathbf{z}_o) = \Sigma_B f(CB\mathbf{z}_0) \qquad // \ \mathbf{b} = B\mathbf{z}_0$$

Here the sum Σ_B is over all permutation matrices which correspond to permutations **b** of $\mathbf{z}_0$, and C is some arbitrary permutation matrix associated with some permutation **c** of $\mathbf{z}_0$. An equivalent way of stating the above involves a direct summation Σ_b over all permutation vectors **b** of $\mathbf{z}_0$,

$$\Sigma_b f(\mathbf{b}) = \Sigma_b f(C\mathbf{b}) . \qquad // \text{ permutation sum rearrangement theorem} \qquad (B.2.4)$$

Throwing in the arbitrary permutation C merely causes a reordering of the sum. This is an extremely powerful and useful fact. Consider then :

$$\det(M) \equiv \frac{1}{n!}\,\Sigma_{\mathbf{a}}\,[\Sigma_{\mathbf{b}}\,(-1)^{S_{\mathbf{a}}+S_{\mathbf{b}}}\;\Pi(\mathbf{a},\mathbf{b};M)\,] \qquad \text{// (B.1.8)}$$

$$= \frac{1}{n!}\,\Sigma_{\mathbf{a}}\;[\Sigma_{\mathbf{b}}\,(-1)^{S_{\mathbf{a}}+S_{\mathbf{b}}}\;\;\Pi(A^{-1}\mathbf{a},A^{-1}\mathbf{b};M)] \qquad \text{// (B.1.6) with } C = A^{-1};\; A^{-1}\mathbf{a} = A^{-1}A\mathbf{z}_0 = \mathbf{z}_0$$

$$= \frac{1}{n!}\,\Sigma_{\mathbf{a}}\,[\,\Sigma_{\mathbf{b}}\;\text{Parity}(A^{-1}\mathbf{b})\;\;\Pi(\mathbf{z}_0,A^{-1}\mathbf{b};M)\,] \qquad \text{// (B.1.4) with } C = A^{-1} \text{ and } \mathbf{a} = \mathbf{b}$$

$$= \frac{1}{n!}\,\Sigma_{\mathbf{a}}\,[\,\Sigma_{\mathbf{b}}\;\text{Parity}(\mathbf{b})\;\;\Pi(\mathbf{z}_0,\mathbf{b};M)\,] \qquad \text{// (B.2.4) with } C = A^{-1} \text{ (key step)}$$

$$= \Sigma_{\mathbf{b}}\;\text{Parity}(\mathbf{b})\;\;\Pi(\mathbf{z}_0,\mathbf{b};M) \qquad \text{// n! identical terms in } \Sigma_{\mathbf{a}}$$

$$= \Sigma_{\mathbf{a}}\;\text{Parity}(\mathbf{a})\;\;\Pi(\mathbf{z}_0,\mathbf{a};M) \qquad \text{// rename dummy sum variable}$$

$$= \Sigma_{\mathbf{a}}\;\text{Parity}(\mathbf{a})\;\;M_{1a_1}M_{2a_2}\;\ldots\ldots\;M_{na_n}\,. \qquad \text{// (B.1.5)} \qquad \text{(B.2.5)}$$

Since $\det(M^{\mathbf{T}}) = \det(M)$ from Theorem 1 (B.1.10), we can also write this result as

$$\det(M) = \Sigma_{\mathbf{a}}\;\text{Parity}(\mathbf{a})\;\;M_{a_1 1}M_{a_2 2}\;\ldots\ldots\;M_{a_n n} \qquad \text{(B.2.6)}$$

Proof: $\det(M) = \det(M^{\mathbf{T}}) = \Sigma_{\mathbf{a}}\;\text{Parity}(\mathbf{a})\;\;M^{\mathbf{T}}_{1a_1}M^{\mathbf{T}}_{2a_2}\;\ldots\ldots\;M^{\mathbf{T}}_{na_n} = \Sigma_{\mathbf{a}}\;\text{Parity}(\mathbf{a})\;\;M_{a_1 1}M_{a_2 2}\;\ldots\ldots\;M_{a_n n}$

Here then are last two results: (recall that $S_{\mathbf{a}}$ is the number of pairwise swaps to get from $\mathbf{z}_0$ to $\mathbf{a}$)

$$\det(M) \equiv \Sigma_{\mathbf{a}}\;\text{Parity}(\mathbf{a})\;\;M_{1a_1}M_{2a_2}\;\ldots\ldots\;M_{na_n} \qquad \text{Parity}(\mathbf{a}) = (-1)^{S_{\mathbf{a}}} \qquad \text{(B.2.7a)}$$

$$\det(M) \equiv \Sigma_{\mathbf{a}}\;\text{Parity}(\mathbf{a})\;\;M_{a_1 1}M_{a_2 2}\;\ldots\ldots\;M_{a_n n} \qquad \text{Parity}(\mathbf{a}) = (-1)^{S_{\mathbf{a}}} \qquad \text{(B.2.7b)}$$

Definition: The **permutation tensor** $\varepsilon_{ijk..}$ (n subscripts) :

- $\varepsilon_{12\ldots n} = 1$
- for any index swap, ε changes sign: $\varepsilon_{..i..j..} = -\varepsilon_{..j..i..}$
- therefore, if any two indices are the same, $\varepsilon = 0$ $\quad \varepsilon_{..i..i..} = -\varepsilon_{..i..i..} = 0 \qquad \text{(B.2.8)}$

Fact: If $\mathbf{a}$ is a permutation of $\mathbf{z}_0$, then $\varepsilon_{a_1a_2\cdots a_n} = \text{Parity}(\mathbf{a}) = (-1)^{S_{\mathbf{a}}} \equiv \varepsilon_{\mathbf{a}}$. $\qquad \text{(B.2.9)}$

Proof: Since indices a_i represent a permutation of $\mathbf{z}_0$, it takes $S_{\mathbf{a}}$ swaps to get from $\varepsilon_{a_1a_2\cdots a_n}$ to $\varepsilon_{123\ldots n}$ by the definition of ε. In dense notation, one could say $\varepsilon_{\mathbf{a}} = (-1)^{S_{\mathbf{a}}}\,\varepsilon_{\mathbf{z}_0}$.

Using (B.2.9) in (B.2.7) then gives these two classic det(M) expressions,

Theorem 2: det(M) can be represented in these two ways:

$$\det(M) = \Sigma_{a_1a_2\cdots a_n}\,\varepsilon_{a_1a_2\cdots a_n}\,M_{1a_1}M_{2a_2}\;\ldots\ldots\;M_{na_n} \qquad \text{(B.2.10a)}$$

$$\det(M) = \Sigma_{a_1a_2\cdots a_n}\,\varepsilon_{a_1a_2\cdots a_n}\,M_{a_1 1}M_{a_2 2}\;\ldots\ldots\;M_{a_n n}\,. \qquad \text{(B.2.10b)}$$

In dense notation, one could write the above as (**a** is now a vector),

$$\det(M) = \Sigma_{\mathbf{a}} \varepsilon_{\mathbf{a}} M_{\mathbf{z_0 a}} \qquad \text{(B.2.10a)'}$$

$$\det(M) = \Sigma_{\mathbf{a}} \varepsilon_{\mathbf{a}} M_{\mathbf{a z_0}} \; . \qquad \text{(B.2.10b)'}$$

In most applications, the following simpler notation suffices:

$$\det(M) = \Sigma_{\mathbf{abc..q}} \varepsilon_{\mathbf{abc..q}} M_{\mathbf{1a}} M_{\mathbf{2b}} M_{\mathbf{3c}} .... M_{\mathbf{nq}} \qquad \text{(B.2.11a)}$$

$$\det(M) = \Sigma_{\mathbf{abc..q}} \varepsilon_{\mathbf{abc..q}} M_{\mathbf{a1}} M_{\mathbf{b2}} M_{\mathbf{c3}} .... M_{\mathbf{qn}} \qquad \text{(B.2.11b)}$$

Theorem 3: For a square matrix M, adding a multiple of one row to another does not change det(M). The same is true for adding a multiple of one column to another. (B.2.12)

Proof: (rows) Suppose we replace $\mathbf{r_3} \rightarrow \mathbf{r_3} + \alpha \mathbf{r_2}$. This says $M_{\mathbf{3i}} \rightarrow M_{\mathbf{3i}} + \alpha M_{\mathbf{2i}}$. Eq (B.2.11a) says :

$$\det(M') = \Sigma_{\mathbf{abc..q}} \varepsilon_{\mathbf{abc..q}} M_{\mathbf{1a}} M_{\mathbf{2b}} (M_{\mathbf{3c}} + \alpha M_{\mathbf{2c}}) .... M_{\mathbf{nq}}$$

$$= \det(M) + \alpha \; \Sigma_{\mathbf{abc..q}} \varepsilon_{\mathbf{abc..q}} \; M_{\mathbf{1a}} M_{\mathbf{2b}} M_{\mathbf{2c}} .... M_{\mathbf{nq}} \; .$$

Since $M_{\mathbf{1a}} M_{\mathbf{2b}} M_{\mathbf{2c}} .... M_{\mathbf{nq}}$ is symmetric under b↔c while $\varepsilon_{\mathbf{abc..q}}$ is anti-symmetric, the extra α term vanishes. That is to say, sum = $\Sigma_{\mathbf{bc}} A_{\mathbf{bc}} S_{\mathbf{bc}} = \Sigma_{\mathbf{cb}} A_{\mathbf{cb}} S_{\mathbf{cb}} = \Sigma_{\mathbf{cb}} (-A_{\mathbf{bc}})(S_{\mathbf{bc}}) = -$ sum = 0. Using (B.2.11b) this argument shows that $\mathbf{c_3} \rightarrow \mathbf{c_3} + \alpha \mathbf{c_2}$ similarly does not alter det(M).

Theorem 4: Swapping two rows (columns) of square matrix M causes det(M) → - det(M). (B.2.13)

Proof: Let's swap rows 1 and 3 in M to get M'. Then from (B.2.11a),

$$\det(M') = \Sigma_{\mathbf{abc..q}} \varepsilon_{\mathbf{abc..q}} M_{\mathbf{3a}} M_{\mathbf{2b}} M_{\mathbf{1c}} .... M_{\mathbf{nq}}$$

$$= \Sigma_{\mathbf{abc..q}} [- \varepsilon_{\mathbf{cba..q}}] M_{\mathbf{3a}} M_{\mathbf{2b}} M_{\mathbf{1c}} .... M_{\mathbf{nq}} \qquad \text{// swap a↔c on } \varepsilon$$

$$= - \Sigma_{\mathbf{abc..q}} \varepsilon_{\mathbf{abc..q}} M_{\mathbf{3c}} M_{\mathbf{2b}} M_{\mathbf{1a}} .... M_{\mathbf{nq}} \qquad \text{// dummy rename a↔c}$$

$$= - \Sigma_{\mathbf{abc..q}} \varepsilon_{\mathbf{abc..q}} M_{\mathbf{1a}} M_{\mathbf{2b}} M_{\mathbf{3c}} .... M_{\mathbf{nq}} \qquad \text{// reorder product}$$

$$= - \det(M) \; .$$

Using Theorem 1, we then know that $\det(M'^{\mathbf{T}}) = - \det(M^{\mathbf{T}})$, so swapping columns 1 and 3 negates det(M). An obvious corollary follows if the two swapped columns have identical data :

Corollary 4: If two rows (columns) of square matrix A are the same, then det(M) = 0. (B.2.14)

Here is one more well-known determinant theorem which again demonstrates the power of the rearrangement theorem (B.2.4).

Theorem 5: det(AB) = det(A)det(B) (B.2.15)

Proof: Since we already use A in $\mathbf{a} = A\mathbf{z}_0$, we shall prove det(XY) = det(X)det(Y) to avoid overloading symbols. Note that Parity(**a**) Parity(**b**) = $(-1)^{s_a}(-1)^{s_b} = (-1)^{s_a+s_b}$ = Parity(A**b**) from (B.1.4). Then :

$\det(X)\det(Y) = [\,\Sigma_{\mathbf{a}}\ \text{Parity}(\mathbf{a})\ \Pi(\mathbf{z}_0,\mathbf{a};X)]\,[\,\Sigma_{\mathbf{b}}\ \text{Parity}(\mathbf{b})\ \Pi(\mathbf{z}_0,\mathbf{b};Y)]$ // (B.2.5) twice

$= \Sigma_{\mathbf{a}}\,[\,\Sigma_{\mathbf{b}}\ \text{Parity}(\mathbf{a})\ \text{Parity}(\mathbf{b})\ \Pi(\mathbf{z}_0,\mathbf{a};X)\ \ \Pi(\mathbf{z}_0,\mathbf{b};Y)\,]$

$= \Sigma_{\mathbf{a}}\,[\,\Sigma_{\mathbf{b}}\ \ \text{Parity}(A\mathbf{b})\ \ \Pi(\mathbf{z}_0,\mathbf{a};X)\ \ \Pi(A\mathbf{z}_0,A\mathbf{b};Y)\,]$ // (B.1.6) with C = A

$= \Sigma_{\mathbf{a}}\,[\,\Sigma_{\mathbf{b}}\ \ \text{Parity}(A\mathbf{b})\ \ \Pi(\mathbf{z}_0,\mathbf{a};X)\ \ \Pi(\mathbf{a},A\mathbf{b};Y)\,]$ // $A\mathbf{z}_0 = \mathbf{a}$

$= \Sigma_{\mathbf{a}}\,[\,\Sigma_{\mathbf{b}}\ \ \text{Parity}(\mathbf{b})\ \ \Pi(\mathbf{z}_0,\mathbf{a};X)\ \ \Pi(\mathbf{a},\mathbf{b};Y)\,]$ // (B.2.4) with C = A

$= \Sigma_{\mathbf{b}}\ \ \text{Parity}(\mathbf{b})\ \Sigma_{\mathbf{a}}\ \Pi(\mathbf{z}_0,\mathbf{a};X)\ \Pi(\mathbf{a},\mathbf{b};Y)$ // move $\Sigma_{\mathbf{a}}$

$= \Sigma_{\mathbf{b}}\ \ \text{Parity}(\mathbf{b})\ \Sigma_{\mathbf{a}}\ \Pi(\mathbf{z}_0,\mathbf{b};XY)$

$= \det(XY)$.

The idea here is to get **a** in the right place on both Π's so that

$\Sigma_{\mathbf{a}}\ \Pi(\mathbf{z}_0,\mathbf{a};X)\ \Pi(\mathbf{a},\mathbf{b};Y) = \Sigma_{\mathbf{a}_1\mathbf{a}_2\cdots\mathbf{a}_n}\ X_{1\mathbf{a}_1}X_{2\mathbf{a}_2}\ldots\ldots X_{n\mathbf{a}_n}\ \ Y_{\mathbf{a}_1\mathbf{b}_1}Y_{\mathbf{a}_2\mathbf{b}_2}\ldots\ldots Y_{\mathbf{a}_n\mathbf{b}_n}$

$= \Sigma_{\mathbf{a}_1\mathbf{a}_2\cdots\mathbf{a}_n}(X_{1\mathbf{a}_1}Y_{\mathbf{a}_1\mathbf{b}_1})(X_{2\mathbf{a}_2 1}Y_{\mathbf{a}_2\mathbf{b}_2})\ldots\ldots(X_{n\mathbf{a}_n}Y_{\mathbf{a}_n\mathbf{b}_n})$

$= (XY)_{1\mathbf{b}_1}\,(XY)_{2\mathbf{b}_2}\ldots(XY)_{n\mathbf{b}_n} = \Pi(\mathbf{z}_0,\mathbf{b};XY)$

which in dense notation one would write as

$\Sigma_{\mathbf{a}}\,X_{\mathbf{z}_0\mathbf{a}}Y_{\mathbf{a}\mathbf{b}} = (XY)_{\mathbf{z}_0\mathbf{b}}$.

Corollary 5: det(ABC) = det(A)det(B)det(C) and so on. (B.2.16)

Proof: det(ABC) = det(A[BC]) = det(A)det(BC) = det(A)det(B)det(C) .

B.3 Minors, Cofactors, and the Cofactor Expansions of det(M)

Definition: The **minor** of a matrix element M_{rs} is the determinant of the submatrix obtained by crossing out the rth row and the sth column. It is a challenge, however, to write this out in symbols. Let's start with a more detailed version of (B.2.11a), where the a_i are *column* summation indices,

$$\det(M) = \Sigma_{a_1 a_2 a_3 a_4 a_5 \ldots a_n} \varepsilon_{a_1 a_2 a_3 a_4 a_5 \ldots a_n} M_{1a_1} M_{2a_2} M_{3a_3} M_{4a_4} M_{5a_5} \ldots\ldots M_{na_n} . \qquad \text{(B.3.1)}$$

We shall make the following conjecture for the form of minor(M_{23})

$$\text{minor}(M_{32}) = (-1)^{3-2} \Sigma_{a_1 a_2 a_4 a_5 \ldots a_n} \varepsilon_{a_1 a_2 2 a_4 a_5 \ldots a_n} M_{1a_1} M_{2a_2} M_{4a_4} M_{5a_5} \ldots\ldots M_{na_n} \qquad \text{(B.3.2)}$$

Compared to det(M) shown in (B.3.1), we have made these changes :

- Removed the factor M_{3a_3} (since row-3 matrix elements cannot appear in minor(M_{32}))
- Removed the sum over a_3.
- Replaced a_3 by the number 2 on the ε tensor.
- added a sign factor $(-1)^{3-2}$.

Notice that since there is a 2 on the ε tensor, any time a summation index $a_i = 2$ there is no contribution since then the ε tensor has two indices the same. so in effect the value 2 has been removed from all the residual summations a_i. That is good, since column 2 is supposedly "crossed out" in minor(M_{32}). The factor $(-1)^{3-2}$ is added so that the "diagonal term" in the minor will be positive. This sign $(-1)^{3-2}$ gets used up if we slide the "2" to its natural position (position 2) on the ε tensor,

$$\text{minor}(M_{32}) = \Sigma_{a_1 a_2 a_4 a_5 \ldots a_n} \varepsilon_{a_1 2 a_2 a_4 a_5 \ldots a_n} M_{1a_1} M_{2a_2} M_{4a_4} M_{5a_5} \ldots\ldots M_{na_n} . \qquad \text{(B.3.3)}$$

The diagonal term in minor(M_{32}) is then positive, matching the mechanical crossing-out method,

$$\varepsilon_{12345..n} M_{11} M_{22} M_{44} M_{55} \ldots\ldots M_{nn} = + M_{11} M_{22} M_{44} M_{55} \ldots\ldots M_{nn} . \qquad \text{(B.3.4)}$$

A more compact notation for (B.3.2) would be

$$\text{minor}(M_{32}) = (-1)^{3-2} \Sigma_{a_i, i\neq 3} \varepsilon_{a_i, a_3 = 2} \Pi_{i\neq 3} (M_{i,a_i}) . \qquad \text{(B.3.5)}$$

Starting over, we could show similarly that

$$\text{minor}(M_{42}) = (-1)^{4-2} \Sigma_{a_i, i\neq 4} \varepsilon_{a_i, a_4 = 2} \Pi_{i\neq 4} (M_{i,a_i}) \qquad \text{(B.3.6)}$$

We now have a sign factor $(-1)^{4-2}$ because the "2" on ε has to be slid 2 positions to get to its natural location (position 4 on ε). More generally we can say

$$\text{minor}(M_{s2}) = (-1)^{s-2} \Sigma_{a_i, i\neq s} \varepsilon_{a_i, a_s = 2} \Pi_{i\neq s} (M_{i,a_i}) \qquad \text{(B.3.7)}$$

where the slide is now s-2 places. Still more generally we find, replacing 2 by r,

$$\text{minor}(M_{sr}) = (-1)^{s\pm r}\ \Sigma_{a_i,\,i\neq s}\ \varepsilon_{a_i,a_s=r}\ \Pi_{i\neq s}\,(M_{i,a_i})\ . \qquad \text{(B.3.8)}$$

This then is our "best form" expression for a minor of matrix element M_{sr}. Either sign will do, since

$$(-1)^{s-r} = (-1)^{s-r}(-1)^{2r} = (-1)^{s+r} \qquad \text{since } (-1)^{2r} = 1$$

Choosing the + sign in (B.3.8) and putting $(-1)^{s+r}$ on the left side we get,

$$(-1)^{s+r}\ \text{minor}(M_{sr}) = \Sigma_{a_i,\,i\neq s}\ \varepsilon_{a_i,a_s=r}\ \Pi_{i\neq s}\,(M_{i,a_i})\ . \qquad \text{(B.3.9)}$$

The left side here is called the **cofactor** of M_{sr}, written $\text{cof}(M_{sr})$. Thus we have shown that

$$\text{cof}(M_{sr}) \equiv (-1)^{s+r}\ \text{minor}(M_{sr}) = \Sigma_{a_i,\,i\neq s}\ \varepsilon_{a_i,a_s=r}\ \Pi_{i\neq s}\,(M_{i,a_i})\ . \qquad \text{(B.3.10)}$$

Fact: Neither $\text{minor}(M_{sr})$ nor $\text{cof}(M_{sr})$ are functions of the M_{sr} matrix elements of M! In (B.3.10) this is so because: (1) row s is excluded in $\Pi_{i\neq s}(M_{i,a_i})$; (2) $a_i = r$ is excluded by the factor $\varepsilon_{a_i,a_s=r}$. More intuitively, this is so because to get $\text{minor}(M_{sr})$ we "cross out" row s and column r. Thus,

$$\frac{\partial \text{cof}(M_{sr})}{\partial M_{sr}} = \frac{\partial \text{minor}(M_{sr})}{\partial M_{sr}} = 0 \qquad \text{for any pair r,s in 1,2...n .} \qquad \text{(B.3.11)}$$

Finally, suppose we replace r with an integer which we call a_s. Doing this causes

$$\varepsilon_{a_i,a_s=r} \rightarrow \varepsilon_{a_i,a_s=a_s} = \varepsilon_{a_i} = \varepsilon_{a_1a_2a_4a_5\ldots a_n} = \text{the normal } \varepsilon \text{ tensor form .} \qquad \text{(B.3.12)}$$

Then (B.3.10) becomes the following,

$$\text{cof}(M_{sa_s}) \equiv (-1)^{s+a_s}\ \text{minor}(M_{sa_s}) = \Sigma_{a_i,\,i\neq s}\ \varepsilon_{a_i}\ \Pi_{i\neq s}\,(M_{i,a_i})\ . \qquad \text{(B.3.13)}$$

Now start again with det(M) of (B.3.1) and rewrite it in our compact notation:

$$\det(M) = \Sigma_{a_1a_2a_3a_4a_5\ldots a_n}\,\varepsilon_{a_1a_2a_3a_4a_5\ldots a_n}\,M_{1a_1}M_{2a_2}M_{3a_3}M_{4a_4}M_{5a_5}\ldots\ldots M_{na_n}$$

$$= \Sigma_{a_i}\,\varepsilon_{a_i}\,\Pi_i(M_{i,a_i}) \qquad \text{// next, extract the } a_s \text{ sum and its } M_{sa_s} \text{ factor :}$$

$$= \Sigma_{a_s}\,M_{sa_s}\,[\Sigma_{a_i,\,i\neq s}\ \varepsilon_{a_i}\,\Pi_{i\neq s}\ (M_{i,a_i})] \qquad \text{// next use (B.3.12) to get}$$

$$= \Sigma_{a_s}\,M_{sa_s}\,[\text{cof}(M_{sa_s})] \qquad \text{// next, change summation index from } a_s \text{ to n}$$

$$= \Sigma_n\,M_{sn}\,\text{cof}(M_{sn})\ . \qquad \text{// valid for any s in 1,2...n} \qquad \text{(B.3.14)}$$

This is the **cofactor expansion** of det(M) where one "works across row s" and n is a column index. Another form of this expansion is

$$\det(M) = \det(M^{\mathbf{T}}) = \Sigma_{\mathbf{n}}\, M^{\mathbf{T}}{}_{\mathbf{sn}}\, \mathrm{cof}(M^{\mathbf{T}}{}_{\mathbf{sn}})$$

$$= \Sigma_{\mathbf{n}}\, M_{\mathbf{ns}}\, \mathrm{cof}(M_{\mathbf{ns}}) \qquad \text{(B.3.15)}$$

and in this form one "works down column s" where n is now a row index. We have just proven :

Theorem 6 (Cofactor Expansions):

$$\det(M) = \Sigma_{\mathbf{n}}\, M_{\mathbf{sn}}\, \mathrm{cof}(M_{\mathbf{sn}}) = \Sigma_{\mathbf{n}}\, (\mathbf{r}_{\mathbf{s}})_{\mathbf{n}}\, \mathrm{cof}(M_{\mathbf{sn}}) \qquad \text{work across row s} \qquad s = 1,2....n \qquad \text{(B.3.16a)}$$

$$\det(M) = \Sigma_{\mathbf{n}}\, M_{\mathbf{ns}}\, \mathrm{cof}(M_{\mathbf{ns}}) = \Sigma_{\mathbf{n}}\, (\mathbf{c}_{\mathbf{s}})_{\mathbf{n}}\, \mathrm{cof}(M_{\mathbf{ns}}) \qquad \text{work down column s} \qquad s = 1,2....n \qquad \text{(B.3.16b)}$$

The term "cofactor" presumably arose because each sum term consists of a "factor" $M_{\mathbf{ij}}$ and a "co-factor" which cooperates with the factor to make the sum, as coauthors cooperate to write a book. A common notation is to define $M^{\mathbf{ij}} \equiv \mathrm{cof}(M_{\mathbf{ij}})$ and then the cofactor expansions become

$$\det(M) = \Sigma_{\mathbf{n}}\, M_{\mathbf{sn}} M^{\mathbf{sn}}$$
$$\det(M) = \Sigma_{\mathbf{n}}\, M_{\mathbf{ns}} M^{\mathbf{ns}} .$$

Although compact, this is not so handy for tensor notation with *its* up and down indices (see Lucht Ref).

B.4 Expressions for the inverse matrix M^{-1} and Cramer's Rule

Consider the expansion (B.3.16a) working across row s,

$$\det(M) = \Sigma_{\mathbf{n}}\, M_{\mathbf{sn}}\, \mathrm{cof}(M_{\mathbf{sn}}) . \qquad \text{(B.4.1)}$$

Suppose we replace the elements of row s ($M_{\mathbf{sn}}$) with the elements of some other row r in M ($M_{\mathbf{rn}}$). In doing so we have created a new matrix, call it M'. Notice that $\mathrm{cof}(M'_{\mathbf{sn}}) = \mathrm{cof}(M_{\mathbf{sn}})$ *because*, although going from M to M' we have altered row s, we have not altered $\mathrm{cof}(M_{\mathbf{sn}})$ since this depends only on the entries in all the rows *other than* row s (think "crossing out" row s for minor($M_{\mathbf{sn}}$)). See Fact (B.3.11) and text above. Therefore we find,

$$\det(M') = \Sigma_{\mathbf{n}}\, M'_{\mathbf{sn}}\, \mathrm{cof}(M'_{\mathbf{sn}}) = \Sigma_{\mathbf{n}}\, M_{\mathbf{rn}}\, \mathrm{cof}(M_{\mathbf{sn}}) . \qquad \text{(B.4.2)}$$

But by Corollary 4 (B.2.14) det(M') = 0 since two rows of M' are the same. Thus

$$0 = \Sigma_{\mathbf{n}}\, M_{\mathbf{rn}}\, \mathrm{cof}(M_{\mathbf{sn}}) . \qquad r \neq s \qquad \text{(B.4.3)}$$

Combining this with (B.4.1) gives,

$$\Sigma_{\mathbf{n}}\, M_{\mathbf{rn}}\, \mathrm{cof}(M_{\mathbf{sn}}) = \det(M)\delta_{\mathbf{r,s}} . \qquad \text{(B.4.4)}$$

Now for clarity define a cofactor matrix C in this manner

$$C_{sn} \equiv \text{cof}(M_{sn}) \; . \tag{B.4.5}$$

In matrix notation, we could define the matrix cof(M) to be matrix C and then

$$C = \text{cof}(M) \qquad \Rightarrow \qquad C_{sn} = [\text{cof}(M)]_{sn} = \text{cof}(M_{sn}) \; . \tag{B.4.6}$$

Now (B.4.4) says

$$\Sigma_n M_{rn} \, C_{sn} = \det(M) \delta_{r,s}$$

or

$$\Sigma_n M_{rn} \, C^T_{ns} = \det(M) \delta_{r,s}$$

and finally in matrix notation,

$$MC^T = \det(M) \, 1$$

or

$$M \left[\frac{C^T}{\det(M)} \right] = 1 \; . \tag{B.4.7}$$

Therefore we find this classic square matrix inversion formula,

Theorem 7: $$M^{-1} = \frac{C^T}{\det(M)} = \frac{[\text{cof}(M)]^T}{\det(M)} = \frac{\text{cof}(M^T)}{\det(M)} \tag{B.4.8}$$

To verify the last equality in (B.4.8) consider :

$= [\text{cof}(M)]^T_{ns} = [\text{cof}(M)]_{sn}$ // meaning of transpose

$= \text{cof}(M_{sn})$ // (B.4.6)

$= \text{cof}(M^T_{ns})$

$= [\text{cof}(M^T)]_{ns}$ // (B.4.6) applied to M^T

and therefore we have this matrix identity,

$$[\text{cof}(M)]^T = [\text{cof}(M^T)] \; . \tag{B.4.9}$$

Using (B.4.8) it is trivial to solve a non-singular (detM≠0) system of linear equations $\mathbf{y} = M\mathbf{x}$:

$$\mathbf{y} = M\mathbf{x} \qquad \Rightarrow \qquad \mathbf{x} = M^{-1} \mathbf{y} = \frac{1}{\det(M)} \, \text{cof}(M^T) \, \mathbf{y} = \frac{1}{\det(M)} \, [\text{cof}(M)]^T \mathbf{y} \; . \tag{B.4.10}$$

In components,

$$x_s = \frac{1}{\det(M)} \Sigma_n \, ([\mathrm{cof}(M)]^T)_{sn} \, y_n = \frac{1}{\det(M)} \Sigma_n \, y_n \, [\mathrm{cof}(M)]_{ns} = \frac{1}{\det(M)} \Sigma_n \, y_n \, \mathrm{cof}(M_{ns}) \, . \qquad \text{(B.4.11)}$$

Now recall the cofactor expansion (B.3.16b),

$$\det(M) = \Sigma_n \, M_{ns} \, \mathrm{cof}(M_{ns}) = \Sigma_n \, (\mathbf{c}_s)_n \, \mathrm{cof}(M_{ns}) \, . \qquad \text{(B.4.12)}$$

If one replaces column $\mathbf{c}_s$ in M by $\mathbf{y}$ one gets

$$\det(M[\mathbf{c}_s \rightarrow \mathbf{y}]) = \Sigma_n \, y_n \, \mathrm{cof}(M_{ns}) \qquad \text{// (B.4.12)}$$

$$= \det(M) \, x_s \, . \qquad \text{// (B.4.11)} \qquad \text{(B.4.13)}$$

Solving this for x_s we obtain another classic result,

Theorem 8 (Cramer's Rule, 1750): // Gabriel Cramer (1704-1752), Swiss, made it to age 47

$$\mathbf{y} = M\mathbf{x} \quad \Rightarrow \quad x_s = \frac{\det(M[\mathbf{c}_s \rightarrow \mathbf{y}])}{\det(M)} \qquad \text{(B.4.14)}$$

Comment: Names of Matrices

Historically the matrix $C^T = [\mathrm{cof}(M)]^T = [\mathrm{cof}(M^T)]$ was called the **adjoint** of matrix M and was often denoted by $\hat{M}$ and then (B.4.8) reads $M^{-1} = \hat{M} / |M|$. But this then conflicted with another usage, namely the matrix $M^\dagger \equiv (M^T)^* = (M^*)^T$ being the **adjoint** of M (the conjugate transpose of M). When $M^\dagger = M$, M is said to be **self-adjoint** and has significance in quantum (matrix) mechanics: the quantum operators of all physical observables are self-adjoint and thus have real eigenvalues. Nowadays the matrix $[\mathrm{cof}(M)]^T$ is called the **classical adjoint** of M, while $M^\dagger$ is then the **Hermitian adjoint** (or **Hermitian conjugate**) of M and is sometimes written M^H. If $M^\dagger = M$ then M is said to be **Hermitian**. Historically when $[\mathrm{cof}(M)]^T$ was the adjoint of M, $M^\dagger$ was called the associate of M. The classical adjoint sometimes appears as the adjugate or adj(M). Transpose matrices M^T are often denoted $\tilde{M}$.

On a related matter, some older authors used the word **minor** to refer to what we now call a cofactor. See for example Morse & Feshbach page 509. These authors never use the word cofactor.

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 pages and pdf's on web, search on "Lagrange Multipliers". Whole books on the use the Lagrange multipliers for specialized applications may be found at the usual book vendor websites.

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".

P.M. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, New York, 1953).

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.

wiki on Lagrange Multipliers: https://en.wikipedia.org/wiki/Lagrange_multiplier