The Method of Lagrange Multipliers

Phil Lucht
Rimrock Digital Technology, Salt Lake City, Utah 84103
last update: Oct 29, 2016
rimrock@xmission.com

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

The graphics look ratty in Windows Adobe PDF viewers when not scaled up, but look just fine in this excellent freeware viewer: https://www.tracker-software.com/product/pdf-xchange-editor .

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

Overview and Summary **2**

1. Finding the stationary points of a function subject to constraints **4**

2. The Method of Lagrange Multipliers **11**

3. Three simple examples **15**
3.1 Example 1: 3-dimensional hemisphere with one simple constraint 15
3.2 Example 2: 4-dimensional hemisphere with two simple constraints 20
3.3 Example 3: N-dimensional hemisphere of radius R in E^N with C simple constraints 23
3.4 Example 1 Revisited: What is the half-width of the peak? 24

4. Example 4: Extremal distances between an ellipse and a circle **27**

5. Example 5: The Boltzmann factor in Statistical Mechanics **34**
5.1 Statement of The Boltzmann Extremum Problem 34
5.2 Solution of The Boltzmann Extremum Problem 36
5.3 More details of the solution 40
5.4 Example: N particles in 3 states : a numerical example 46
5.5 Example: The Maxwell-Boltzmann Distribution 55

Section 6. Matrix proof of the Method of Lagrange Multipliers **68**
6.1. The R matrix and its Rank 68
6.2. Proof of Theorem 3 ($\Rightarrow$) 71
(a) Proof for N = 6 and S = 4 71
(b) Proof for general $S \leq N$ 75

Appendix A: Matrix rank equals the number of independent columns and rows **79**

Appendix B: Determinants **83**
B.1 Definition of the determinant 83
B.2 The permutation group, the permutation tensor ε, and expansions for det(M) 85
B.3 Minors, Cofactors, and the Cofactor Expansions of det(M) 91
B.4 Expressions for the inverse matrix M^{-1} and Cramer's Rule 93

Appendix C: The Intersection of Constraint Surfaces **96**

Appendix D: Lagrange Multipliers in Classical Mechanics **101**

References **116**

Overview and Summary

The Method of Lagrange Multipliers is used to determine the stationary points (including extrema) of a real function f(**r**) subject to some number of (holonomic) constraints.

The main purpose of this document is to provide a solid **derivation** of the method and thus to show why the method works. A geometric derivation is presented in Section 1 while an alternate matrix-based derivation appears in Section 6.

A secondary purpose is to provide some interesting and illustrative **examples**. Examples 1, 2 and 3 are geometric examples in 3, 4 and N dimensions. Example 4 concerns extremal distances between a circle and an ellipse. Extended Example 5 involves the Boltzmann distribution of statistical mechanics where the variables N_i are state population counts. Finally, Appendix D gives several examples from classical Lagrangian dynamics (with all supporting material) in which a functional F(**φ**) is rendered stationary rather than a function like f(**r**). In these examples the variables are functions φ_i known as generalized coordinates.

Here is a concise summary of the document:

Section 1 defines a stationary point **r** of a function f(**r**) subject to constraints. It then proves Theorem 1 which says that iff **r** is such a stationary point, then constants λ_i must exist so that a certain gradient statement is true. The λ_i are the "Lagrange Multipliers".

Section 2 presents the Method of Lagrange Multipliers as Theorem 2. This theorem is really just a recasting of Theorem 1, so Theorem 1 $\Leftrightarrow$ Theorem 2 and the derivation of Theorem 2 is trivial.

Section 3 gives three detailed examples involving hemispheres in 3,4 and N dimensions, with 1, 2 and S constraints respectively.

Section 4 does a numerical example to find the extremal distances between a circle and an ellipse which are coplanar.

Section 5 presents a more sophisticated physics example associated with the name Boltzmann wherein one determines the particle populations of energy levels in the context of statistical mechanics and thermal equilibrium. This self-contained section first treats the case of discrete energy levels with a numerical example involving three levels. It then examines a case with continuous energy levels and derives the well-known Maxwell-Boltzmann distribution. Numbers are obtained for a small box of helium atoms. It is shown for example that such atoms at room temperature have a mean speed of 2,780 mph.

Section 6 derives Theorem 1 in a manner totally different from the derivation presented in Section 1. The function f(**r**) and the constraint functions are treated as variables of a certain transformation which has a "differential matrix" called R(**r**). Theorem 3 is then proven, claiming that iff **r** is a stationary point, then the rank of the matrix R(**r**) falls below full rank and this in turn implies Theorem 1.

Appendix A proves for a general n x m matrix that column rank = row rank = rank. This fact is used in the proof of Theorem 3 in Section 6.

Appendix B derives a batch of facts about determinants some of which are used in Appendix A. This is a standalone section that might be a useful derivation resource for workers who deal with matrices.

Appendix C gives examples of intersecting constraint surfaces and discusses the notion of eliminating variables using constraint equations, as is done in the proof Section 6. It concludes with an explanation of why a constraint equation $a(\mathbf{r}) = 0$ with $\mathbf{r} = (x_1, x_2 ..... x_N)$ can be interpreted as a surface in E^N.

Appendix D presents two applications of Lagrange Multipliers in classical Lagrangian mechanics.

References (not many) 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.

We often use the abbreviations $\partial_i f \equiv \partial f / \partial x_i \equiv \frac{\partial f}{\partial x_i}$. Sometimes we also use $f_i \equiv \frac{\partial f}{\partial x_i}$.

1. Finding the stationary points of a function subject to constraints

Here we develop a method for finding the stationary points of a function f(**r**) subject to a set of constraint conditions $a_i(\mathbf{r}) = 0$. This method does involve "Lagrange multipliers", but the Method of Lagrange Multipliers is usually stated in the equivalent form shown in Section 2. Nevertheless, the proof of the method is presented below in this section.

In the following **r** = $(x_1,x_2.....x_N)$ is a point in Euclidian space E^N.

Preliminary Facts about Surfaces

Fact 1: The equation $F(x_1,x_2.....x_N) = K$ defines an N-1 dimensional surface in E^N . (1.1)

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 .

See the discussion starting at (C.16) in Appendix C for more details about Fact 1.

Fact 2: The vector $\nabla F(\mathbf{r})$ is always normal to the surface $F(\mathbf{r}) = K$ at point **r**. (1.2)

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 an *arbitrary* d**r** displacement on the surface, $\nabla F(\mathbf{r})$ must be locally normal to the surface at point **r**.

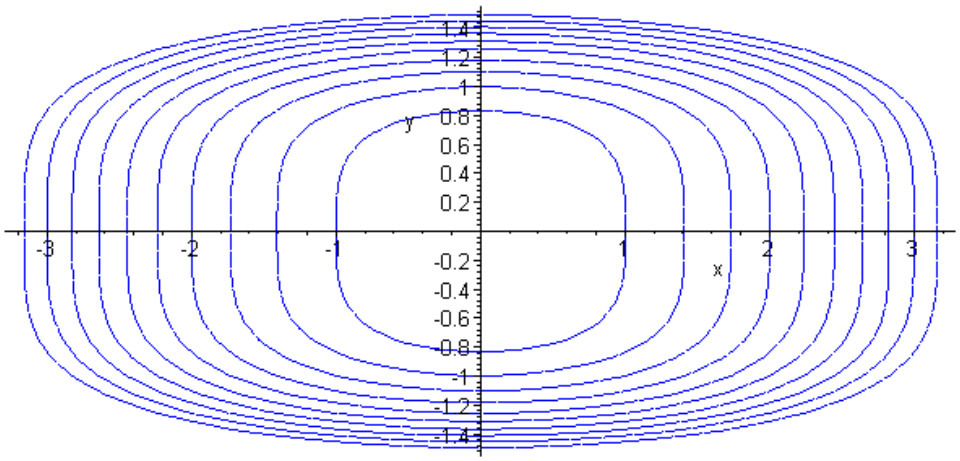
Definition: As K takes a set of different values $K_1, K_2.....K_M$, the equation $F(x_1,x_2.....x_n) = K$ describes a family of so-called **level surfaces (level sets)**. If M is large and the range of the K_i is small, these surfaces will be closely spaced. Adjacent surfaces then have nearly the same shape, although in general *all* the surfaces in the family of surfaces do not have the same shape. (1.3)

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);
```



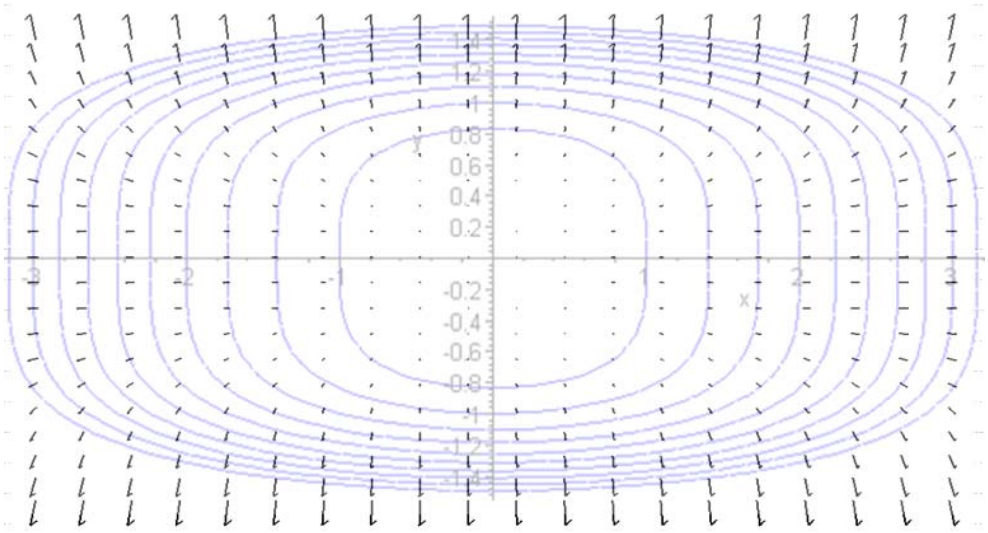
(1.4)

Fact 3: The gradient ∇F points in the direction in which F changes most rapidly. (1.5)

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").

For the above example $\nabla F = 2x\hat{\mathbf{x}} + 8y^3\hat{\mathbf{y}}$ we can superpose this gradient field on the above level curves using a Maple fieldplot routine,

```
fieldplot([2*x,8*y^3],x=-3..3, y=-1.5..1.5,thickness=2);
```



(1.6)

Wherever a gradient arrow base lies on a level curve, the arrow is normal to the curve and points in the direction in which the function x^2+2y increases most rapidly. Thinking of $z = f(x,y) = x^2+2y^4$, the blue lines above are the level curves of an oblong bowl into which the viewer is looking, and the gradient arrows all point uphill on the bowl surface. The arrows are longer where the level curves lines (contour lines, topo lines) are closer together.

The Intersection Constraint Surface A(**r**) = 0

Assume that a point **r** = (x_1, x_2 ... x_N) in E^N is required to satisfy C constraint equations,

$$a_i(\mathbf{r}) = 0 \qquad i = 1,2...C \ . \qquad \text{// constraints} \tag{1.7}$$

According to Fact 1 above, each constraint equation defines a surface of dimension N-1 in E^N. Our point **r** must lie simultaneously on all C constraint surfaces. The intersection of all these constraint surfaces is in fact an **intersection constraint surface** of dimension N-C in E^N whose constraint function we shall call A(**r**). In symbolic form we write

$$\begin{array}{lccccc} & (A(\mathbf{r}) = 0) & = & (a_1(\mathbf{r}) = 0) \cap (a_2(\mathbf{r}) = 0) \cap .....\cap (a_C(\mathbf{r}) = 0) \ . \\ \text{dimension :} & \text{N-C} & & \text{N-1} \quad \text{N-1} & & \text{N-1} \end{array} \tag{1.8}$$

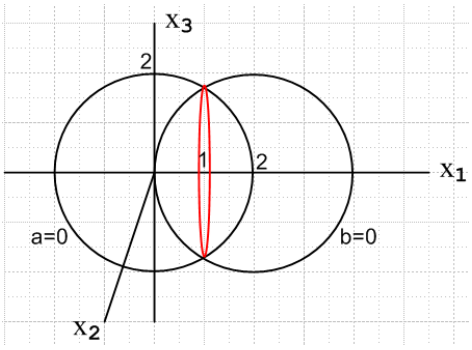
Barring degenerate cases (which we shall ignore in the following general discussion) each extra constraint lowers the dimensionality of the intersecting constraint surface by 1 degree of freedom. We assume that the intersection constraint surface is not null, otherwise our problem of interest would have no solutions. The intersection surface could in general have multiple disjoint pieces.

If there are no constraints, all points in E^N are "allowed". If there is one constraint $a_1(\mathbf{r}) = 0$, then $A(\mathbf{r}) = a_1(\mathbf{r})$. For two constraint surfaces $a_1(\mathbf{r}) = 0$ and $a_2(\mathbf{r}) = 0$, one must *compute* the intersecting constraint surface (A(**r**) = 0) which we symbolically write as $(a_1(\mathbf{r}) = 0) \cap (a_2(\mathbf{r}) = 0)$.

Example: N = 3 and C = 2 constraints. Suppose in E^3 there are two constraints as follows :

$$\begin{aligned} a_1(x_1,x_2,x_3) &= x_1^2 + x_2^2 + x_3^2 - 2^2 = 0 \\ a_2(x_1,x_2,x_3) &= (x_1-2)^2 + x_2^2 + x_3^2 - 2^2 = 0 \ . \end{aligned} \tag{1.9}$$

Each of these constraints describes a surface of dimension 2 in E^3 (a sphere of radius 2). These two spheres intersect in a circle $x_2^2 + x_3^2 = 3$, so the intersection surface has dimension 1 in E^3. This agrees with N-C = 1. Here is a crude drawing, where the intersection constraint surface (edge-on circle) is red,



(1.10)

In this example one has A(**r**) = $x_2^2 + x_3^2 - 3$ and then A(**r**) = 0 is the equation of the intersection constraint surface. If the origin of the second sphere were moved to $x_1 = 4$ (slide right sphere to the right 2 units), one has a degenerate case where the intersection constraint surface is just a point and so has dimension 0. Further separation of the spheres results in a null intersection surface. Setting $x_1 = 0$ aligns the two spheres for another degenerate case, resulting in the intersection surface having dimension 2.

Definitions: The tangent space and the perp space at **r**

Consider a point **r** located on (A(**r**)=0). As noted above, this intersection constraint surface (A(**r**)=0) is a surface of dimension N-C embedded in E^N. Therefore at **r** one can construct a set of N-C independent vectors d**r** such that **r**+d**r** also lies on the intersection constraint surface, so both A(**r**) = 0 and A(**r**+d**r**) = 0. These N-C independent vectors span a vector space (dimension N-C) at point **r** on the surface known as the **tangent space** at point **r**. These vectors d**r** are all "tangent to" the surface. (1.11)

There is another vector space within E^N at point **r** of dimension N - (N-C) = C which is orthogonal to the tangent space, and we shall call it the **perp space** at **r**. If d**r** is any vector in the tangent space at **r**, and **N** is a vector in the perp space at **r**, then **N** • d**r** = 0. So one can find C linearly independent vectors **N** such that **N** • d**r** = **0** and all these **N** vectors are normal to surface (A(**r**)=0) at **r**. (1.12)

As the point **r** moves on the surface (A(**r**)=0), the axes of both the tangent space and the perp space also move, perhaps rotating a bit. The union of the set of N-C basis vectors of the tangent space at **r** and the set of C basis vectors of the perp space at **r** forms a basis for E^N.

Summarizing the above:

Fact 4: Let M ≡ (A(**r**)=0) embedded in E^N represent the intersection constraint surface of C constraint surfaces (a_i(**r**) = 0). The dimension of surface M, and the dimension of the tangent space at any **r** on M, is N-C, while the dimension of the perp space at any **r** on M is C. (1.13)

Fact 5: If **A** • d**r** = 0 for all d**r** in the tangent space at **r**, then **A** lies in the perp space at **r**. (1.14)

Usually in such discussions one regards the surfaces involved as having smoothness properties which make them be **manifolds**. If the intersecting constraint surface is a manifold M, then the tangent space at **r** on M is often denoted as $T_r M$ and the perp space as $(T_r M)^\perp$.

For example, in Fig 1.10 for E^3 we started with two constraint spheres and the intersection constraint surface is the edge-on circle shown in red. Since this is a surface of dimension 1, the tangent space has dimension 1 and the perp space has dimension 2. At any point **r** on a circle embedded in E^3, there is only one ±d**r** direction which keeps one on the circle, but one could construct two linearly independent normal vectors at **r**. On the other hand, if either sphere were the only constraint in the problem, the tangent space would have dimension 2 and the perp space dimension 1. At a point on the sphere there is only one normal vector direction in E^3.

Fact 6: If d**r** lies in the tangent space of a point **r** on the intersection constraint surface, then $\nabla a_i(\mathbf{r}) \bullet d\mathbf{r} = 0$ and so the vectors $\nabla a_i(\mathbf{r})$ lie in the perp space at **r**. (1.15)

Proof: Let **r** be a point on (A(**r**)=0) and let d**r** be a vector in the tangent space at **r** on (A(**r**)=0). This means that both **r** and **r** + d**r** lie on the intersection constraint surface (A(**r**)=0), so dA = A(**r**+d**r**) = A(**r**) = 0. Since this surface is the intersection of all the (a_i(**r**)=0) component surfaces, it is also true that $da_i = a_i(\mathbf{r}+d\mathbf{r}) = a_i(\mathbf{r}) = 0$ for each surface i. But $da_i = \nabla a_i(\mathbf{r}) \bullet d\mathbf{r}$ so we have $0 = \nabla a_i(\mathbf{r}) \bullet d\mathbf{r}$. Therefore, if d**r** is any vector in the tangent space, then $\nabla a_i(\mathbf{r}) \bullet d\mathbf{r} = 0$. From Fact 5 one concludes that $\nabla a_i(\mathbf{r})$ must lie in the perp space at **r** on (A(**r**)=0).

Comment: Avoiding gradient confusion

Let **r** be a point in E^N. Consider the equation u = f(**r**) which maps E^N to E^1. Construct a space E^{N+1} having coordinates (**r**,u). Let g(**r**,u) = f(**r**) - u be a function on this new E^{N+1}. The equation u = f(**r**) is the same as the equation g(**r**,u) = 0. The equation g(**r**,u) = 0 is, according to Fact 1, a surface of dimension N in E^{N+1}. From Fact 2, a normal to this surface is given by $\nabla^{(N+1)} g(\mathbf{r},u)$ which is a vector in E^{N+1}. For any reasonable smooth surface, this normal vector cannot be null. On the other hand, the vector $\nabla^{(N)} f(\mathbf{r})$ is a vector in E^N and there is no reason why this vector cannot be null. We write this as $\nabla f(\mathbf{r})$ below. If this paragraph seems confusing, reread it thinking of N = 2 so u = f(**r**) = f(x,y) is a regular surface in E^3.

The Stationary Point Problem

STATIONARY POINT (NO CONSTRAINTS)

We know that if we move from **r** to **r**+d**r**, where d**r** is *any* d**r**, the differential change in f(**r**) is given by

$$df(d\mathbf{r}) = f(\mathbf{r}+d\mathbf{r}) - f(\mathbf{r}) = \nabla f(\mathbf{r}) \bullet d\mathbf{r}, \quad d\mathbf{r} = \text{any differential vector in } E^N \ . \tag{1.16}$$

If for some **r** it happens that $\nabla f(\mathbf{r}) = 0$, then the above says df = 0 and so **r** is a **stationary point** (also known as a **critical point**) of the function f(**r**). That is to say, the phrase "**r** is a stationary point of f(**r**)" means that df(d**r**) = 0 for any d**r**.

In this problem there are no constraints specifying legal values for **r**. All points **r** in E^N are legal. For **r** in E^2 the reader is no doubt familiar with the fact that, for u = f(**r**) plotted in E^3, such a stationary point could occur at a hill top, or a valley bottom, or a saddle point (which is neither a maximum or a minimum). For fun, go look up "monkey saddle".

STATIONARY POINT (WITH CONSTRAINTS)

Let **r** lie on the intersection constraint surface (A(**r**)=0). Now the only "allowed" d**r** are vectors in the tangent space for this **r**, because any other d**r** will violate at least one constraint since it will take us off the intersecting constraint surface. We modify the previous equation to say

$$df(d\mathbf{r}) = f(\mathbf{r}+d\mathbf{r}) - f(\mathbf{r}) = \nabla f(\mathbf{r}) \bullet d\mathbf{r}\ , \quad d\mathbf{r} = \text{any diff. vector lying in the tangent space at } \mathbf{r} \tag{1.17}$$

If one could find a value of **r** such that $\nabla f(\mathbf{r}) = 0$ *and* $\mathbf{A}(\mathbf{r}) = 0$, one would have df = 0 and such an **r** would be a stationary point of "f(**r**) subject to $\mathbf{A}(\mathbf{r}) = 0$", but this situation is generally very unlikely. Instead, we define a stationary point as follows:

Definition: If **r** lies on $A(\mathbf{r}) = 0$, then **r** is a "**stationary point** of f(**r**) subject to $A(\mathbf{r}) = 0$" if $df(d\mathbf{r}) = 0$ for any d**r** in the tangent space of $(A(\mathbf{r})=0)$ at point **r**. In simple terms, **r** is a stationary point if $df(d\mathbf{r}) = 0$ for all "constraint-allowed" d**r** displacements. (1.18)

Since $df(d\mathbf{r}) = \nabla f(\mathbf{r}) \bullet d\mathbf{r}$, this requirement will certainly be met in the special case just noted that $\nabla f(\mathbf{r}) = 0$ *and* $\mathbf{A}(\mathbf{r}) = 0$, but the above definition is more general and there will be stationary points for which $\nabla f(\mathbf{r}) \neq 0$.

Fact 7: If **r** is a stationary point of f(**r**), then $\nabla f(\mathbf{r})$ must lie in the perp space of $(A(\mathbf{r})=0)$ at **r**. (1.19)

Proof: If **r** is a stationary point, then by definition $df = \nabla f(\mathbf{r}) \bullet d\mathbf{r} = 0$ for any d**r** in the tangent space at **r**. Then by Fact 5 $\nabla f(\mathbf{r})$ must lie in the perp space at **r**.

In the special case $\nabla f(\mathbf{r}) = 0$ and $\mathbf{A}(\mathbf{r}) = 0$ we have noted that **r** is a stationary point, but we now look for stationary points for which $\nabla f(\mathbf{r}) \neq 0$.

Consider the set of C+1 vectors $\{\nabla a_i(\mathbf{r}), \nabla f(\mathbf{r})\}$. From Facts 6 and 7 we know that all C+1 vectors must lie in the perp space at **r**. According to Fact 4, the dimensionality of the perp space at **r** is C. Therefore, the vectors $\{\nabla a_i(\mathbf{r}), \nabla f(\mathbf{r})\}$ for i = 1 to C must be linearly *dependent* (see (6.1.6) below) -- one cannot have C+1 independent vectors in a space of dimension C. This means that one can find a set of constants λ'_i (not all 0) for i = 1 to C such that

$$\lambda'_0 \nabla f(\mathbf{r}) + \Sigma_{i=1}^{C} \lambda'_i \nabla a_i(\mathbf{r}) = 0 \ . \tag{1.20}$$

If $\lambda'_0 = 0$, then $\Sigma_{i=1}^{C} \lambda'_i \nabla a_i(\mathbf{r}) = 0$ which means the $\nabla a_i(\mathbf{r})$ are linearly dependent at point **r**. We explicitly rule out this case by requiring that the $a_i(\mathbf{r})$ are such that the C vectors $\nabla a_i(\mathbf{r})$ are linearly *in*dependent near any point **r** where we might have a stationary point.

So assuming then that $\lambda'_0 \neq 0$, define $\lambda_i \equiv \lambda'_i/\lambda'_0$ to get this new version of (1.20),

$$\nabla f(\mathbf{r}) + \Sigma_{i=1}^{C} \lambda_i \nabla a_i(\mathbf{r}) = 0 \ . \tag{1.21}$$

We have just proven the following theorem:

If a point **r** is a stationary point for f(**r**) subject to constraints $a_i(\mathbf{r}) = 0$, then the following are true :

(a) $a_i(\mathbf{r}) = 0$ for i = 1,2...C (point **r** must satisfy all the constraints)
(b) There must exist C constants λ_i such that $\nabla f(\mathbf{r}) + \Sigma_{i=1}^{C} \lambda_i \nabla a_i(\mathbf{r}) = 0$. (1.22)

Conversely, suppose we can find a set of λ_i such that the above two conditions are met for some **r**. Let d**r** be a vector in the tangent space at **r**. We then compute

$$df = \nabla f \bullet d\mathbf{r} = [- \Sigma_{i=1}^{C} \lambda_i \nabla a_i(\mathbf{r})] \bullet d\mathbf{r} = - \Sigma_{i=1}^{C} \lambda_i [\nabla a_i(\mathbf{r}) \bullet d\mathbf{r}] \quad . \tag{1.23}$$

But we know from Fact 6 that $\nabla a_i(\mathbf{r}) \bullet d\mathbf{r} = 0$ for d**r** in the tangent space, and therefore df = 0 for any such d**r**, and therefore **r** is a stationary point. Thus we improve the above Theorem to get

Theorem 1: A point **r** is a "stationary point for f(**r**) subject to constraints $a_i(\mathbf{r}) = 0$" $\Leftrightarrow$ (1.24)

(a) $a_i(\mathbf{r}) = 0$ for i = 1,2...C (point **r** must satisfy all the constraints)
(b) There exist C constants λ_i such that $\nabla f(\mathbf{r}) + \Sigma_{i=1}^{C} \lambda_i \nabla a_i(\mathbf{r}) = 0$.

Corollary: If **r** satisfies all the constraints [**r** lies on (A(**r**)=0)] but **r** is not a stationary point of f(**r**), then there exists no equation of the form $\nabla f(\mathbf{r}) + \Sigma_{i=1}^{C} \lambda_i \nabla a_i(\mathbf{r}) = 0$. That is to say, there is no set of real numbers λ_i which makes this equation be true. In this case the vectors $\{\nabla a_i(\mathbf{r}), \nabla f(\mathbf{r})\}$ are linearly independent. (1.25)

The constants λ_i are referred to as **Lagrange Multipliers**.

2. The Method of Lagrange Multipliers

The object is to solve this problem:

Find the stationary points $\mathbf{r}$ of function $f(\mathbf{r})$ subject to C constraints $a_i(\mathbf{r}) = 0$. (2.1)

Define the following function of N+C variables $\mathbf{r} = (x_1, x_2....x_N)$ and $\boldsymbol{\lambda} = (\lambda_1,\lambda_2....\lambda_C)$,

$$H(\mathbf{r},\boldsymbol{\lambda}) \equiv f(\mathbf{r}) + \Sigma_{i=1}^{C}\lambda_i a_i(\mathbf{r}) = f(r) + \boldsymbol{\lambda} \bullet \mathbf{a}(\mathbf{r}) \quad . \tag{2.2}$$

We then have the following theorem concerning this function H :

Theorem 2: Method of Lagrange Multipliers (2.3)

The stationary points $\mathbf{r}$ in E^N of the function $f(\mathbf{r})$ subject to C constraints $a_i(\mathbf{r}) = 0$ are the same as the $\mathbf{r}$ values obtained from finding stationary points $(\mathbf{r},\boldsymbol{\lambda})$ in E^{N+C} of the *unconstrained* function $H(\mathbf{r},\boldsymbol{\lambda})$.

Proof:

Compute the unconstrained stationary points $(\mathbf{r},\boldsymbol{\lambda})$ of H by setting all N+C partial derivatives to 0,

$$\begin{aligned} 0 &= \partial H/\partial\lambda_i = a_i(\mathbf{r}) && \text{for coordinates } \lambda_i \ \ i = 1..C \\ 0 &= \partial H/\partial x_k = \partial f(\mathbf{r})/\partial x_k + \Sigma_{i=1}^{C} \lambda_i \partial a_i(\mathbf{r})/\partial x_k && \text{for coordinates } x_k \ \ k = 1..N \ . \end{aligned} \tag{2.4}$$

Rewrite as

$$\begin{aligned} & a_i(\mathbf{r}) = 0 \quad \text{for } i = 1,2...C \\ & \nabla f(\mathbf{r}) + \Sigma_{i=1}^{C} \lambda_i \nabla a_i(\mathbf{r}) = 0 \ . \quad \text{// and therefore this equations exists} \end{aligned} \tag{2.5}$$

But we have already proven in Theorem 1 (1.24) that the conditions (2.5) are necessary and sufficient for $\mathbf{r}$ to be a stationary point of function $f(\mathbf{r})$ subject to C constraints $a_i(\mathbf{r}) = 0$, so that then concludes the proof of Theorem 2.

We now write out these equations in more detail:

$$\begin{aligned} & a_1(\mathbf{r}) = 0 \\ & a_2(\mathbf{r}) = 0 \\ & .... \\ & a_C(\mathbf{r}) = 0 \qquad \text{// C equations} \end{aligned} \tag{2.6}$$

$$\begin{aligned} & \partial_1 f(\mathbf{r}) + \lambda_1\partial_1 a_1(\mathbf{r}) + \lambda_2\partial_1 a_2(\mathbf{r}) + ... + \lambda_C \partial_1 a_C(\mathbf{r}) = 0 \\ & \partial_2 f(\mathbf{r}) + \lambda_1\partial_2 a_1(\mathbf{r}) + \lambda_2\partial_2 a_2(\mathbf{r}) + ... + \lambda_C \partial_2 a_C(\mathbf{r}) = 0 \\ & .... \\ & \partial_N f(\mathbf{r}) + \lambda_1\partial_N a_1(\mathbf{r}) + \lambda_2\partial_N a_2(\mathbf{r}) + ... + \lambda_C \partial_N a_C(\mathbf{r}) = 0 \ . \qquad \text{// N equations} \end{aligned} \tag{2.7}$$

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 stationary points **r** 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. The Method of Lagrange Multipliers provides **no 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 (2.7), it is possible to obtain expressions for the λ_i as follows. First, write out the equations (2.7) in matrix form,

$$-\begin{pmatrix} \partial_1 f \\ \partial_2 f \\ \dots \\ \partial_N f \end{pmatrix} = \begin{pmatrix} \partial_1 a_1 & \partial_1 a_2 & \dots & \partial_1 a_C \\ \partial_2 a_1 & \partial_2 a_2 & \dots & \partial_2 a_C \\ \dots & \dots & \dots & \dots \\ \partial_N a_1 & \partial_N a_2 & \dots & \partial_N a_C \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_C \end{pmatrix} \tag{2.8}$$

where the matrix has C columns (one column for each constraint) and N rows. In general one must have $N \geq C+1$ so the matrix is "tall" -- it has more rows than columns. Recall that the intersection constraint surface (A(**r**)=0) has dimension N-C. If one were to allow $N < C$, this surface would have negative dimension so the problem is overconstrained. If $N = C$, then (A(**r**)=0) has dimension 0 so the problem is constrained to a single point **r** in E^N. This point either is or is not a stationary point according to Theorem 1 so there is no real stationarity problem to solve.

Taking only the first C rows of the above matrix equation, one finds

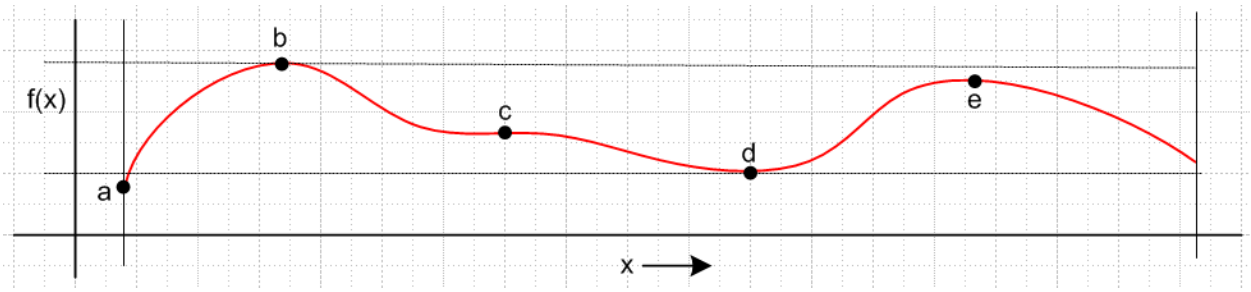
$$-\begin{pmatrix} \partial_1 f \\ \partial_2 f \\ \dots \\ \partial_C f \end{pmatrix} = \begin{pmatrix} \partial_1 a_1 & \partial_1 a_2 & \dots & \partial_1 a_C \\ \partial_2 a_1 & \partial_2 a_2 & \dots & \partial_2 a_C \\ \dots & \dots & \dots & \dots \\ \partial_C a_1 & \partial_C a_2 & \dots & \partial_C a_C \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_C \end{pmatrix} \tag{2.9}$$

where the matrix is now square, CxC. Assuming this matrix has non-zero determinant, one can invert (2.9) to obtain expressions for the Lagrange multipliers $\lambda_i(\mathbf{r})$ as functions of **r**. These λ_i functions can be inserted into equations (2.7) and then one can proceed to solve equations (2.6) and (2.7) for **r**.

By making different versions of (2.9) taking different sets of C rows from (2.8), one may produce several different sets of $\{\lambda_i(\mathbf{r})\}$ corresponding to different solutions **r**. Each stationary point will have its own set of Lagrange Multipliers $\{\lambda_i\}$, a fact which follows from Theorem 2. This will be clearly illustrated in Example 4 below.

Stationary Points vs. Extremal Points

The Method of Lagrange Multipliers identifies stationary points of f some of which *may* be extremal points. We have not mentioned the domain of $f : E^N \rightarrow$ reals, but in most problems there is some restricted domain of interest for **r** in f(**r**) and an extremum might occur on the boundary of that domain, in which case it won't be picked up by the Method of Lagrange Multipliers. The following 1D function f(x) plotted in red illustrates this point and shows various cases which can arise with stationary points :



(2.10)

Here the domain of interest is the portion of the x axis lying between the two thin 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 general $f(\mathbf{r})$.

Comments

1. For the Lagrange Multiplier method to be valid, the functions $f(\mathbf{r})$ and $a_i(\mathbf{r})$ must be **continuous** and **differentiable** in all arguments x_i *and* the derivatives must also be continuous. The functions are therefore C^0 and C^1. This ensures that the gradients in (2.4) are smooth continuous functions. The constraints are also assumed to be independent in that none can be written as a linear combination of the others, such as $a_3(\mathbf{r}) = 2a_1(\mathbf{r}) + a_2(\mathbf{r})$. In this case if $a_1(\mathbf{r}) = 0$ and $a_2(\mathbf{r}) = 0$, then $a_3(\mathbf{r}) = 0$ and a_3 adds nothing new.

2. The λ_i terms in (2.5) often appear with **minus signs** in place of plus signs. This removes the minus signs in (2.8) and (2.9). These signs are just a convention and have no effect on the solution points $\mathbf{r}$.

3. The function H in (2.2) is sometimes referred to as **the Lagrangian** and is written L. This Lagrangian *differs* from that which appears in the Lagrangian formulation of classical mechanics which we define in (D.38) of Appendix D. Here we show the connection.

In (2.2) the function H is defined by,

$$H(\mathbf{r},\boldsymbol{\lambda}) \equiv f(\mathbf{r}) + \Sigma_{i=1}^{c}\lambda_i a_i(\mathbf{r}) \quad . \qquad (2.2)$$

We require that

$$dH = 0 \qquad (2.11)$$

from which we conclude in (2.4) that $a_i(\mathbf{r}) = 0$ and

$$\frac{\partial H}{\partial x_k} = \frac{\partial f(\mathbf{r})}{\partial x_k} + \Sigma_{i=1}^{c} \lambda_i \frac{\partial a_i(\mathbf{r})}{\partial x_k} = 0 \qquad k = 1,2,....N \ . \tag{2.12}$$

The N coordinates x_k are the components of $\mathbf{r} = (x_1, x_2, ...x_N)$.

In Appendix D the object that most closely corresponds to H is the action S, where

$$S = \int_{t1}^{t2} dt\, L(q_k(t), \dot{q}_k(t), t) \ . \tag{2.13}$$

where L is the classical Lagrangian.

The N coordinates $q_k(t)$ are the generalized coordinates (functions) of the Lagrange problem.

We require, similar to (2.11), that

$$\delta S = 0 \ . \tag{2.14}$$

This produces a set of modified Euler-Lagrange equations (D.55) which closely ressemble those in (2.12),

$$\left(\frac{\partial L}{\partial q_k} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}_k}\right) + \Sigma_{i=1}^{c} \lambda_i \frac{\partial a_i}{\partial q_k} = 0 \qquad k = 1,2...N \tag{2.15}$$

$$\frac{\partial f}{\partial x_k} + \Sigma_{i=1}^{c} \lambda_i \frac{\partial a_i}{\partial x_k} = 0 \qquad k = 1,2,....N \ . \tag{*2.12*}$$

Here the role played by $\frac{\partial f}{\partial x_k}$ in (2.12) is assumed by $\left(\frac{\partial L}{\partial q_k} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}_k}\right)$ in (2.15). Just how this last expression arises from $\delta S = 0$ is shown at the end of Appendix D.

The main point is that H is not the Lagrangian L, but is rather analogous to the time integral of the Lagrangian L which is the action S of (2.13). In the case of H we do a simple calculus variation dH relative to the coordinates x_k, whereas for S we do a functional variation δS relative to the generalized coordinate functions $q_k(t)$.

It is of course OK to refer to H as "the Lagrangian" as long as one understands this differs from L.

3. Three simple examples

In this section the Method of Lagrange Multipliers is used to find stationary points for simple surfaces subject to simple constraints. Example 1 has **r** in E^2, while Example 2 has **r** in E^3. Note that the surface *graphs* for these examples, being of the form u = f(**r**), are in E^3 and E^4.

Each of these examples could be reviewed in perhaps two small paragraphs. Instead, we have chosen to pour out a plethora of descriptive text with drawings in an attempt to hammer home the basic ideas presented abstractly in the previous sections.

The final Example 3 has **r** in E^N and generalizes Examples 1 and 2.

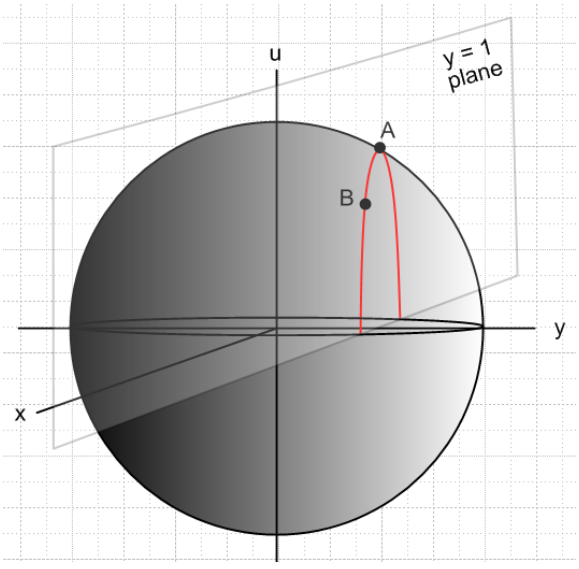
3.1 Example 1: 3-dimensional hemisphere with one simple constraint

The situation in E^3

Let **r** = (x,y) in E^2 and let $u = f(x,y) = \pm\sqrt{4 - x^2 - y^2}$ 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 wish to find (x,y) that maximizes f. If there are no constraints, then (1.16) says $\nabla f = 0$. The only place on our u = f surface having $\nabla f = 0$ is the north pole of the sphere. This is just a regular "critical point" where $\partial_x f = 0$ and $\partial_y f = 0$, so we are happy with this interpretation of (1.24b) with no constraints, C = 0.

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 which corresponds to a stationary point for 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 stationary point of this constrained problem will be at point **A**, and point **B** is not a stationary point .



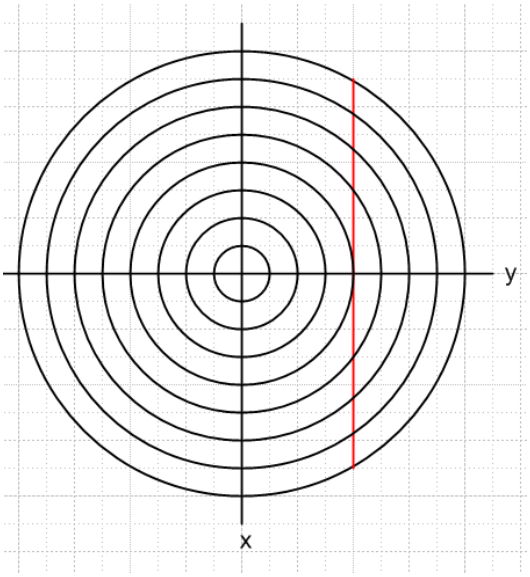
(3.1.1)

We pause to see how this example fits into the general formalism. The space shown above in which we are plotting the graph u = f(x,y) is E^{N+1} whereas the space which holds **r** = (x,y) is E^N, and N = 2. This **r**-space E^2 is the equatorial plane in (3.1.1) and it is in this plane that we work when doing the Lagrange Multiplier analysis.

What is potentially misleading is the notion of surface and gradient. A sphere of radius r is described by R(**r**) = r where R(**r**) = $\sqrt{x^2+y^2+u^2}$. In spherical coordinates one easily computes that $\nabla R = (\partial_{\mathbf{r}} R)\ \hat{\mathbf{r}} = \hat{\mathbf{r}}$, and this is indeed normal to the sphere at every point on it, as predicted by Fact 2 (1.2) . 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 Fig (3.1.1). It plays out, as just noted, in the u = 0 plane of Fig (3.1.1) which is E^2. The gradients of interest are 2D gradients in this plane, not the 3D gradients of surfaces in (3.1.1).

Here then is plane of interest, oriented with the x axis down,



(3.1.2)

We wish to maximize the function f(x,y) = $\sqrt{4 - x^2 - y^2}$ subject to the constraint y = 1. In (3.1.1) the spherical surface is $u^2+x^2+y^2 = R^2 = 4$, so f(x,y) is then height u. So the function $\sqrt{4 - x^2 - y^2} = \sqrt{4 - r^2}$ (r is now the 2D r) is the height of the hemisphere in Fig (3.1.1) 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 (3.1.2) which are circles with r = $\sqrt{4-K^2}$. On each circle in (3.1.2), the height of the sphere lying over the surface in (3.1.1) is constant -- that is why they are called level curves. Think of these as lines of latitude projected onto the u=0 equatorial plane.

In this problem there is only C = 1 constraint. In the general formalism we would refer to this constraint and its Lagrange multiplier as a_1 and λ_1, but here we shall just call them a and λ.

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 (3.1.2). We have,

$$f = \sqrt{4 - x^2 - y^2} = \sqrt{4 - r^2}$$
$$a = y - 1 \qquad (3.1.3)$$

so

$$\nabla f = \nabla\sqrt{4 - r^2} = [-r/\sqrt{4-r^2}\,]\,\hat{\mathbf{r}} \qquad \text{// 2D gradients}$$
$$\nabla a = \nabla(y-1) = \hat{\mathbf{y}}\,. \qquad (3.1.4)$$

For any point **r** in the disk of (3.1.2), ∇f therefore points toward the disk center, and this is then the uphill direction for the hemisphere in (3.1.1). In Cartesian coordinates,

$$\nabla f = \nabla\sqrt{4 - x^2 - y^2} = -(x/f)\hat{\mathbf{x}} - (y/f)\hat{\mathbf{y}} = -\frac{x}{\sqrt{4 - x^2 - y^2}}\,\hat{\mathbf{x}} - \frac{y}{\sqrt{4 - x^2 - y^2}}\,\hat{\mathbf{y}}\,. \qquad (3.1.5)$$

Now for just one constraint (1.24b) reads,

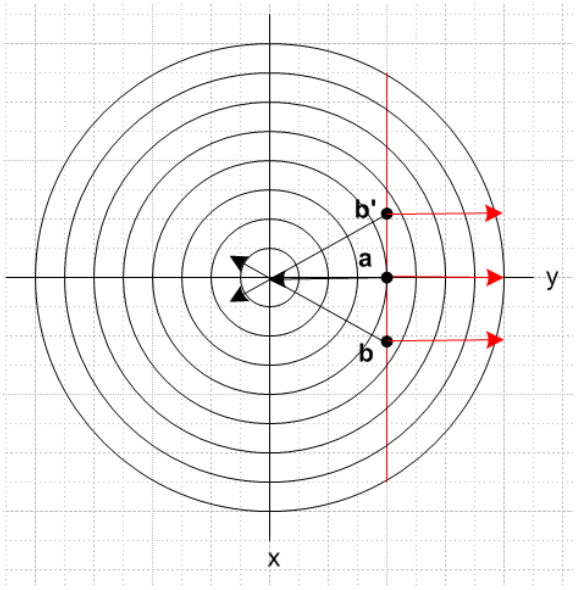
$$\nabla f(\mathbf{r}) = -\lambda \nabla a(\mathbf{r}). \qquad (3.1.6)$$

How do we make geometric sense of this equation? If **r** is stationary point, we should have df = 0 for any legal small displacement d**r** starting at this point **r**. Eq. (3.1.6) dotted into d**r** then says,

$$0 = df = \nabla f(\mathbf{r}) \bullet d\mathbf{r} = -\lambda \nabla a(\mathbf{r}) \bullet d\mathbf{r}\,. \qquad (3.1.7)$$

If we displace d**r** in any direction on the constraint surface (which from Fact 2 has normal $\nabla a(\mathbf{r})$), then $\nabla a(\mathbf{r}) \bullet d\mathbf{r} = \mathbf{0}$ and we find df = 0, so we are thus *at* a stationary point 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 with a computer program (e.g. GNU Octave freeware) to find a maximum (minimum) of f is called the **method of gradient ascent (descent)**.

In our example with only one constraint, (3.1.6) says that we have reached a stationary point when ∇f and ∇a are collinear. Let's draw onto (3.1.2) the points **a** and **b** which lie under points **A** and **B** in (3.1.1), and we include a point **b'** which is a mirror image of **b** :



(3.1.8)

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 (3.1.1). At point **a** we see that in fact the two gradients are collinear as required by (3.1.6) at a stationary point.

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

$$df(\mathbf{b}) = \nabla f \bullet d\mathbf{r} = [\, -(x/f)\hat{\mathbf{x}} - (y/f)\hat{\mathbf{y}} \,] \bullet |dx|(-\hat{\mathbf{x}}) = |dx|\,(x/f) > 0 \qquad \text{since } x>0 \text{ and } f>0 \ . \qquad (3.1.9)$$

Since $df > 0$, it is to our advantage in finding max f to move upwards in (3.1.8). Conversely, suppose we are instead at point **b'**. Then moving toward **a** gives,

$$df(\mathbf{b}') = \nabla f \bullet d\mathbf{r} = [\, -(x/f)\hat{\mathbf{x}} - (y/f)\hat{\mathbf{y}} \,] \bullet |dx|(\hat{\mathbf{x}}) = |dx|\,(-x/f) > 0 \qquad \text{since } x<0 \text{ and } f>0 \qquad (3.1.10)$$

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

In this Example $\nabla a = \nabla(y-1) = \hat{\mathbf{y}}$ spans the 1 dimensional perp space at any point **r** on the constraint surface $A(\mathbf{r}) = 0$, that is, $y = 1$. Similarly, $\hat{\mathbf{x}}$ spans the 1 dimensional tangent space at any **r** on $A(\mathbf{r}) = 0$.

Finding the stationary point

It is an easy matter to compute the solution value of the Lagrange Multiplier λ_1 and **r**. Inserting (3.1.4) into (3.1.6) gives

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

$$-\frac{x}{\sqrt{4 - x^2 - y^2}}\hat{\mathbf{x}} - \frac{y}{\sqrt{4 - x^2 - y^2}}\hat{\mathbf{y}} = -\lambda\,\hat{\mathbf{y}} \ . \qquad (3.1.11)$$

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

$$-\frac{1}{\sqrt{4 - 1}}\hat{\mathbf{y}} = -\lambda\,\hat{\mathbf{y}} \qquad \Rightarrow \qquad \lambda = 1/\sqrt{3} \ . \qquad (3.1.12)$$

We have thus found the following stationary point **r** and Lagrange multiplier λ for Example 1:

$$\mathbf{r} = (0,1) \qquad \lambda = 1/\sqrt{3} \ . \qquad (3.1.13)$$

Therefore the height of point A in Fig (3.1.1) is $u = \sqrt{4 - x^2 - y^2} = \sqrt{4 - 0^2 - 1^2} = \sqrt{3}$.

Now consider, setting $(x,y) = (x_1,x_2)$,

$$f = \sqrt{4 - x_1^2 - x_2^2}$$

$$\partial_i f = (1/2)\, f^{-1}\,(-2x_i) = -x_i/f$$

$$\partial^2_i f = -[f * 1 - x_i \partial_i f] / f^2 = -[f - x_i(-x_i/f)] / f^2 = -(1/f) - (x_i/f)^2 < 0 \qquad (3.1.14)$$

Thus $\partial^2_1 f < 0$ and $\partial^2_2 f < 0$ for all points (x_1,x_2) on the equatorial disk under the hemisphere, as is intuitively obvious. Thus these two inequalities apply at the obtained stationary point **r**, and this confirms that this stationary point is a maximum of the function f.

A look at H(x,y,λ) .

Theorem 2 of (2.3) claims that the unconstrained "Lagrangian" H(x,y,λ) = f(x,y) +λ(y-1) should have a stationary point at the point (x,y,λ) = (0,1,1/√3) shown in (3.1.13). Consider first,

$\partial H/\partial\lambda = (y-1) = 0$ at the point $(x,y,\lambda) = (0,1,1/\sqrt{3})$ // λ partial is 0 at stationary point

$\partial^2 H/\partial\lambda^2 = 0$ // since H is linear in λ // no λ curvature anywhere (3.1.15)

Next,

$\partial_x H = \partial_x f + \lambda\partial_x a = \partial_x f = -x/f = 0$ at $(0,1,1/\sqrt{3})$ // x partial is 0 at stationary point

$\partial_y H = \partial_y f + \lambda\partial_y a = -y/f + \lambda = -1/\sqrt{3} + 1/\sqrt{3} = 0$ // y partial is 0 at stationary point (3.1.16)

as expected. Just for fun, here is a plot of the function H with λ set to the solution value $\lambda = 1/\sqrt{3}$,

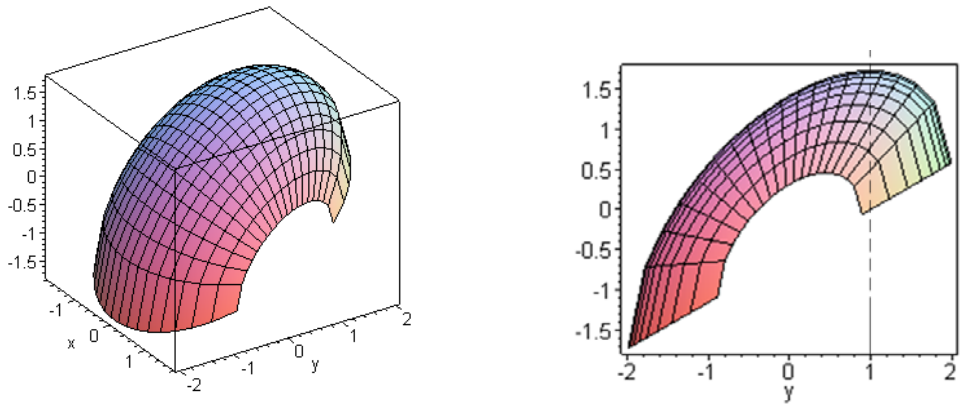
$H(x,y,\lambda=1/\sqrt{3}) = \sqrt{4 - x^2 - y^2} + (1/\sqrt{3})(y-1)$. (3.1.17)

Now $u = (1/\sqrt{3})(y-1)$ appearing in the second term of (3.1.17) is a plane sloping up to the right in Fig (3.1.1). This is different from the plane y = 1 which is the vertical constraint plane in that figure. In (3.1.17) 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 should have a stationary point at **r** = (0,1) :

```
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]);
```

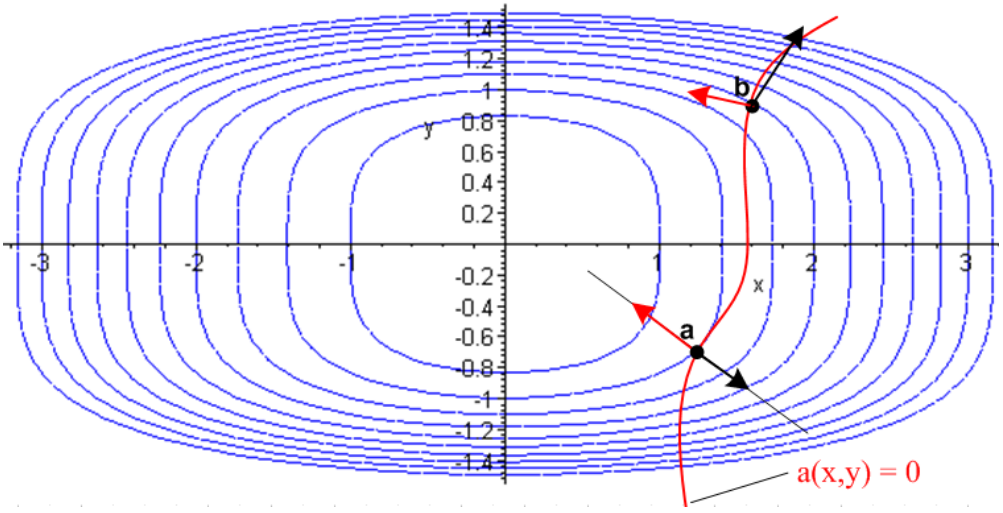


(3.1.18)

One can see from these plots that in fact **r** = (0,1) is a stationary point of H(x,y,λ=1/$\sqrt{3}$) and is in fact a maximum. In this example it just happens that the stationary point is a maximum for both f and H.

More complicated Examples of the same type

If we had a surface in Fig (3.1.1) more complicated than a sphere, and a constraint more complicated than y = 1, the nature of the interpretation of (3.1.2) does not change. For example, if Fig (3.1.1) showed an oblong bowl surface u = f(x,y) whose level curves were those of (1.4) (u = K values increasing toward the outside) and if the constraint a(x,y) = 0 were some arbitrary constraint curve (shown in red), we would have this picture:



(3.1.19)

The black arrows are the 2D "up-hill" bowl gradients ∇f (larger toward the outside) while the red arrows ∇a are normals to the constraint surface (∇a now varies along that surface). The solution point is **a** where ∇f and ∇a are collinear. At point **b** the gradient arrows do not line up, and there is advantage in this case for decreasing f by moving toward point **a**. One can see from the "topo lines" that **a** is indeed the point on the constraint curve that has the least value of u.

Notice that saying the two arrows in E^2 are collinear is to say they are linearly dependent.

3.2 Example 2: 4-dimensional hemisphere with two simple constraints

In order to better demonstrate the interpretation of (1.24b) concerning the Lagrange multiplier gradients, we upgrade Example 1 and then allow two constraints. We take the surface of Fig (3.1.1) to be the "upper half" of a 4D sphere of radius 2, u = f(x,y,z) = $\sqrt{4 - x^2 - y^2 - z^2}$. We must *imagine* a 4D drawing of this sphere which is the new Fig (3.1.1) with a vertical u axis and three "horizontal" axes x,y,z. We cannot draw this picture, but luckily the Lagrange Multiplier Show for this problem plays out in E^3 where we *can* draw pictures. Now with **r** = (x,y,z) and r = $\sqrt{x^2+y^2+z^2}$ the level surfaces (really surfaces this time) are given by u = $\sqrt{4 - r^2}$ = K for a set of constant K values. Since then r = $\sqrt{4-K^2}$, we see that these level surfaces are in fact a set of concentric spheres of radius $\sqrt{4-K^2}$ (so we shall restrict to |K| ≤ 2).

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

$$f(x,y,z) = \sqrt{4 - x^2 - y^2 - z^2} = \sqrt{4 - r^2}$$
$$a_1(x,y,z) = y-1$$
$$a_2(x,y,z) = x-1 \ . \qquad (3.2.1)$$

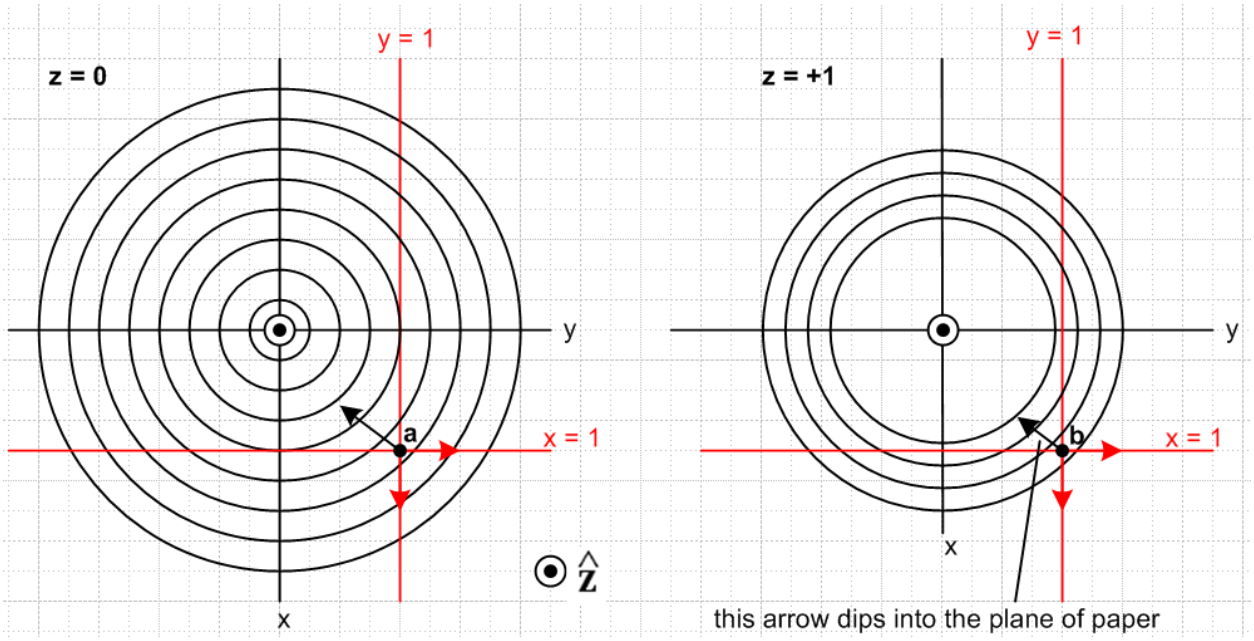
The Lagrange method gradients of interest are now 3D gradients in E^3 space, so

$$\nabla f = \nabla(\sqrt{4 - r^2}) = [-r/\sqrt{4-r^2}]\,\hat{\mathbf{r}} = -(x/f)\hat{\mathbf{x}} - (y/f)\hat{\mathbf{y}} - (z/f)\hat{\mathbf{z}} \qquad \text{points to sphere center}$$
$$\nabla a_1 = \nabla(y-1) = \hat{\mathbf{y}}$$
$$\nabla a_2 = \nabla(x-1) = \hat{\mathbf{x}} \ . \qquad (3.2.2)$$

Equation (1.24b) reads

$$\nabla f(\mathbf{r}) + \lambda_1 \nabla a_1(\mathbf{r}) + \lambda_2 \nabla a_2(\mathbf{r}) = 0 \ . \qquad (3.2.3)$$

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 in E^3 corresponding to Fig (3.1.2) in E^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 drawings) :



(3.2.4)

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 (z=2), 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 $\nabla a_1 = \hat{\mathbf{y}}$ and $\nabla a_2 = \hat{\mathbf{x}}$, 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 $\nabla f, \nabla a_1, \nabla a_2$ are therefore not coplanar, so they are linearly independent. That means that equation (3.2.3) cannot exist 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. This means the three vectors $\nabla f, \nabla a_1, \nabla a_2$ are coplanar in the z=0 plane so they are linearly *dependent*. For this point **a**, the equation (3.2.3) 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 (3.2.3),

$$df = \nabla f \bullet d\mathbf{r} \;=\; -\lambda_1[\nabla a_1(\mathbf{r}) \bullet d\mathbf{r}] \;-\; \lambda_2\,[\nabla a_2(\mathbf{r}) \bullet d\mathbf{r}] \;= -\lambda_1\,[0] \;-\lambda_2[0] \;= 0 \qquad (3.2.5)$$

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

$$df = \nabla f \bullet d\mathbf{r} \;=\; [\,-(x/f)\hat{\mathbf{x}} - (y/f)\hat{\mathbf{y}} \;-(z/f)\hat{\mathbf{z}}\,] \bullet (-|dr|\,\hat{\mathbf{z}}\,) = (z/f)dr \;> 0 \qquad (3.2.6)$$

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 (3.2.6) at z = 0 for point **a** again shows df = 0.

Finally, we solve the problem. Inserting (3.2.2) into (3.2.3) gives

$$\nabla f(\mathbf{r}) = -\lambda_1 \nabla a_1(\mathbf{r}) \;-\lambda_2\, \nabla a_2(\mathbf{r}) \qquad \Rightarrow$$

$$[-\frac{x}{\sqrt{4 - x^2 - y^2 - z^2}}\,\hat{\mathbf{x}} - \frac{y}{\sqrt{4 - x^2 - y^2 - z^2}}\,\hat{\mathbf{y}} \;- \frac{z}{\sqrt{4 - x^2 - y^2 - z^2}}\,\hat{\mathbf{z}}\,] = -\;\lambda_1\hat{\mathbf{y}} \;-\; \lambda_2\,\hat{\mathbf{x}}\,. \qquad (3.2.7)$$

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

$$\frac{x}{\sqrt{4 - x^2 - y^2}} = \lambda_2 \qquad \frac{y}{\sqrt{4 - x^2 - y^2}} = \lambda_1 \qquad . \qquad (3.2.8)$$

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

$$\frac{1}{\sqrt{4 - 1^2 - 1^2}} = \lambda_2 \qquad \frac{1}{\sqrt{4 - 1^2 - 1^2}} = \lambda_1 \qquad \Rightarrow \qquad \lambda_1 = \lambda_2 = 1/\sqrt{2}\;. \qquad (3.2.9)$$

Therefore the solution to Example 2 is this:

$$\mathbf{r} = (1,1,0) \qquad \lambda_1 \;=\; 1/\sqrt{2} \qquad \lambda_2 \;=\; 1/\sqrt{2}\;. \qquad (3.2.10)$$

The value of u at this point is $u = \sqrt{4 - x^2 - y^2 - z^2} \;= \sqrt{4 - 1^2 - 1^2 - 0^2} \;= \sqrt{2}\,.$

It is simple matter conceptually to generalize our Example 2 to more complicated functions f, a_1, a_2. For some arbitrary function $f(\mathbf{r})$ there will be a set of 2D surfaces in E^3 which are the level surfaces on which

$f(\mathbf{r}) = K$ 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 $a_1(x,y,z)$ and $a_2(x,y,z)$ with general functions which then result in arbitrarily shaped 2D constraint surfaces $a_1(x,y,z) = 0$ and $a_2(x,y,z) = 0$ in E^3. The gradients ∇a_1 and ∇a_2 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 $\mathbf{r}$ must lie on this curve. At each point $\mathbf{r}$ on this curve, the gradients $\nabla a_1(\mathbf{r})$ and $\nabla a_2(\mathbf{r})$ define a local plane. At a stationary point one will find that the gradient $\nabla f(r)$ lies in the same plane as $\nabla a_1(\mathbf{r})$ and $\nabla a_2(\mathbf{r})$ so the three gradients are linearly dependent and then $\nabla f(\mathbf{r}) + \lambda_1 \nabla a_1(\mathbf{r}) + \lambda_2 \nabla a_2(\mathbf{r}) = 0$ for some λ_1 and λ_2. At such a point $\mathbf{r}$, if one considers any $d\mathbf{r}$ which lies on both constraint surfaces, one will then find that $df = \nabla f \bullet d\mathbf{r} = 0$ so the point $\mathbf{r}$ is therefore a stationary point.

The situation is described by a 3D version of drawing (3.2.4).

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.

3.3 Example 3: N-dimensional hemisphere of radius R in E^N with C simple constraints

We now generalize the above two examples.

Consider an upper-half hypersphere which is an N-1 dimensional surface embedded in E^N. The sphere has radius R and there are C constraints $x_i = 1$ for i = 1,2..C.

Proceed as in the previous examples where now $f(\mathbf{r})$ is defined on E^N and the hypersphere surface "graph" is given by $u = f(\mathbf{r})$ in E^{N+1}.

$$\mathbf{r} = (x_1, x_2 .... x_N)$$
$$f(\mathbf{r}) = f(x_1, x_2 .... x_N) = +\sqrt{R^2 - \Sigma_{i=1}^{N} x_i^2} \qquad // \; u = f(\mathbf{r}) \text{ is upper-half hypersphere in } E^N$$
$$a_i = x_i - 1 \qquad i = 1,2..C \; . \qquad // \text{ C constraints are } x_i = 1 \text{ for first C coordinates} \qquad (3.3.1)$$

The gradients of interest (N-dimensional) are

$$\nabla f = - \Sigma_{i=1}^{N} (x_i/f) \hat{\mathbf{x}}_i$$
$$\nabla a_i = \hat{\mathbf{x}}_i \qquad i = 1,2..C \; . \qquad (3.3.2)$$

Insert these into (1.24b) ,

$$\nabla f(\mathbf{r}) = - \Sigma_{i=1}^{C} \lambda_i \nabla a_i(\mathbf{r}) \qquad \textit{(1.24b)}$$

so

$$- \Sigma_{i=1}^{N} (x_i/f) \hat{\mathbf{x}}_i = - \Sigma_{i=1}^{C} \lambda_i \hat{\mathbf{x}}_i$$

or

$$\Sigma_{i=1}^{C} (x_i/f) \hat{\mathbf{x}}_i + \Sigma_{i=C+1}^{N} (x_i/f) \hat{\mathbf{x}}_i = \Sigma_{i=1}^{C} \lambda_i \hat{\mathbf{x}}_i$$

or

$$- \Sigma_{i=1}^{C} [(x_i/f) - \lambda_i] \hat{\mathbf{x}}_i + \Sigma_{i=C+1}^{N} (x_i/f) \hat{\mathbf{x}}_i = 0 \; . \qquad (3.3.3)$$

Solve to get

$$\lambda_i = x_i/f \qquad i = 1,2..C$$
$$x_i = 0 \qquad i = C+1,....N \ . \tag{3.3.4}$$

Since $x_i = 1$ for i = 1 to C one has $\lambda_i = 1/f$ where

$$f = \sqrt{R^2 - \Sigma_{i=1}^{C} x_i^2} = \sqrt{R^2 - C} \ . \tag{3.3.5}$$

The stationary point is then determined by

$$\lambda_i = 1/\sqrt{R^2 - C} \qquad i = 1,2...C \qquad \text{// the C Lagrange multipliers are all the same}$$
$$\mathbf{r} = (1,1...1,0,0...0) \ . \qquad \text{// C ones followed by N-C zeros} \tag{3.3.6}$$

There is no stationary point if $C > R^2$ since then the λ_i become imaginary.

Finally, we verify that the results of Examples 1 and 2 are recovered:

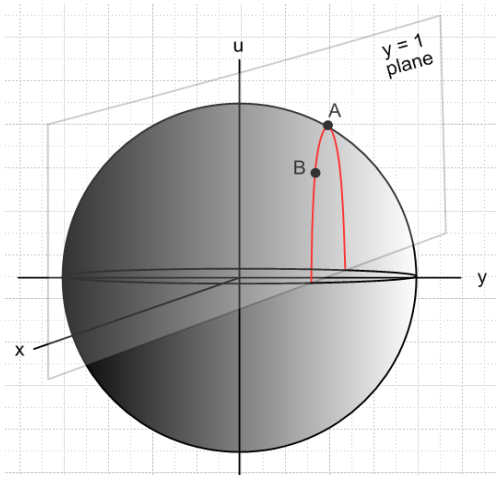
Example 1: N = 2, C = 1, R = 2, $\lambda_1 = 1/\sqrt{3}$, $\mathbf{r} = (1,0)$

Example 2: N = 3, C = 2, R = 2, $\lambda_i = 1/\sqrt{2}$, $\mathbf{r} = (1,1,0)$ (3.3.7)

3.4 Example 1 Revisited: What is the half-width of the peak?

The question pursued here seems of small interest, but it serves as a model for the situation which arises in Section 5.3 where we are interested in the width ΔN_i of the peak of the Boltzmann microstate distribution Ω in (5.3.16).

Recall from (3.1.1) where $u = f(x,y) = \sqrt{4 - x^2 - y^2}$,



(3.1.1)

The upper spherical surface is a graph in E^3 of $u = f(x,y) = \sqrt{4-x^2-y^2}$. The variables x,y are not independent. When the constraint y = 1 is applied we get $u = f(x,1) = \sqrt{3-x^2} \equiv F(x)$ which is the equation of the red curve shown in the picture. Here we have invented a new function name F(x) to represent the function f(x,y) after y has been eliminated by the constraint. One is left with only one independent variable which is x.

We could then ask about the half-width of the "peak" which is the red curve. As an estimate, we model the red curve as a parabola using a Taylor expansion around the peak point A,

$$F(x+\Delta x) \approx F(x) + (\partial F/\partial x)\,(\Delta x) + (1/2)\;(\partial^2 F/\partial x^2)\,(\Delta x)^2 \qquad (3.4.1)$$

Maple computes the two derivatives and then evaluates them at point A where x = 0:

```
F := sqrt(3-x^2);
```

$$F := \sqrt{3-x^2}$$

```
DF := diff(F,x);
```

$$DF := -\frac{x}{\sqrt{3-x^2}}$$

```
DDF := diff(F,x,x);
```

$$DDF := -\frac{x^2}{(3-x^2)^{\left(\frac{3}{2}\right)}} - \frac{1}{\sqrt{3-x^2}}$$

```
eval(F,x=0);
```

$$\sqrt{3}$$

```
eval(DF,x=0);
```

$$0$$

```
eval(DDF,x=0);
```

$$-\frac{1}{3}\sqrt{3} \qquad (3.4.2)$$

As expected, $(\partial F/\partial x) = 0$ since A is a stationary point for the function F(x) of one independent coordinate x. Using $(\partial^2 F/\partial x^2) = -1/\sqrt{3}$ we rearrange (3.4.1) to get

$$\Delta F = F(x+\Delta x) - F(x) \;\; = -(2\sqrt{3}\,)^{-1}(\Delta x)^2\,. \qquad (3.4.3)$$

To get the half-width of the peak we set (the half width at half maximum).

$$|\Delta F/F(0)| = 1/2 \qquad (3.4.4)$$

so that

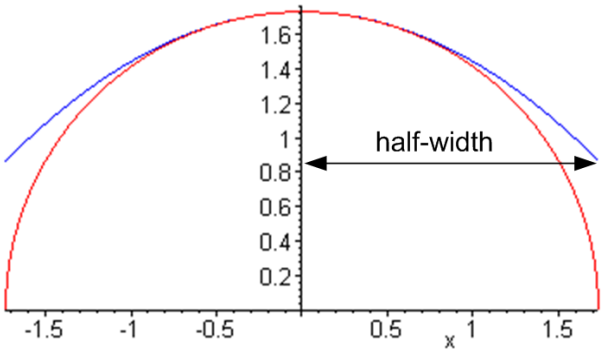
$$|\; -(2\sqrt{3}\,)^{-1}(\Delta x)^2 \;/\sqrt{3}\;| \;= 1/2 \qquad \Rightarrow \;\; (\Delta x) \approx \sqrt{3} \;= 1.73 \;\; . \qquad (3.4.5)$$

The following plot shows the parabolic approximation in blue and marks the estimated half width

```
Fapprox := sqrt(3)-(2*sqrt(3))^(-1)*x^2
```

$$Fapprox := \sqrt{3} - \frac{1}{6}\sqrt{3}\,x^2$$

```
plot([F,Fapprox],x=-sqrt(3)..sqrt(3), color = [red,blue],scaling = constrained);
```



(3.4.6)

As a percentage of the peak value, the half width here is 100%, so it is a fairly large half-width.

In the Boltzmann problem of Section 5, we shall start with a function $f(N_1,N_2.....N_m)$ of m variables which are not independent. We use up the two problem constraints to eliminate the variables N_1 and N_2 and then define a function $F(N_3,N_4.....N_M) \equiv f(N_1(N_3,...N_m), N_2(N_3,...N_m), N_3, ...N_m)$. In the space of the independent variables N_3, N_4....N_m we use a vector version of the above procedure to estimate the half width of the peak of the function F in any of the variables N_i, as illustrated in (5.3.18). For large N_i values, we find that the peak is very narrow compared to its height, unlike the peak shown in (3.4.6).

4. Example 4: 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. Example 4 is a bit more complex than the toy examples in texts and in our Section 3, and we obtain a numerical solution after carrying out the analytic steps of the Lagrange Multiplier method.

Example 4 has two constraints so C = 2 and 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.

We adopt several **new notations** in this section :

1. The constraint functions will be called a and b instead of a_1 and a_2, but the multipliers are still λ_1,λ_2.

2. The partial derivatives of a function $F(x_1, x_2 ....)$ shall be represented as F_i :

$$F_i \equiv \partial_i F(x_1, x_2 ....) \equiv \frac{\partial_i F(x_1, x_2 ....)}{\partial x_i}$$

which is a notation promoted by Buck with generalization $F_{ijk..} = \partial_i\partial_j\partial_k ... F(x_1, x_2 ....)$.

3. The components of the vector **r** will be called (x,y,x',y') instead of (x_1,x_2,x_3,x_4) .

Consider then these two curves in E^2:

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

$(x'-\alpha)^2 + (y'-\beta)^2 = R^2$. circle of radius R centered at (α,β) (4.1)

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

The distance squared is given by

$f = (x-x')^2 + (y-y')^2 = f(x,y,x',y')$ N = 4 coordinates (4.2)

while the constraint functions are

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

$b(x,y,x',y') = (x'-\alpha)^2 + (y'-\beta)^2 - R^2$ $b(x,y,x',y') = 0$. (4.3)

The various derivatives are,

$$
\begin{array}{lll}
f_1 = \partial_x f = 2(x-x') & a_1 = \partial_x a = 2x/A^2 & b_1 = \partial_x b = 0 \\
f_2 = \partial_y f = 2(y-y') & a_2 = \partial_y a = 2y/B^2 & b_2 = \partial_y b = 0 \\
f_3 = \partial_x, f = -2(x-x') & a_3 = \partial_x, a = 0 & b_3 = \partial_x, b = 2(x'-\alpha) \\
f_4 = \partial_y, f = -2(y-y') & a_4 = \partial_y, a = 0 & b_4 = \partial_y, b = 2(y'-\beta) \ .
\end{array} \qquad (4.4)
$$

The H function (2.2) is then,

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

$$
= (x-x')^2 + (y-y')^2 + \lambda_1 [x^2/A^2 + y^2/B^2 - 1] + \lambda_2 [(x'-\alpha)^2 + (y'-\beta)^2 - R^2] \ . \qquad (4.5)
$$

Compute the six H_i derivatives as prescribed in (2.4),

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

$$
\begin{array}{l}
H_5 = a \\
H_6 = b \ .
\end{array} \qquad (4.6)
$$

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

$$
\begin{array}{ll}
(x-x') + \lambda_1 x/A^2 = 0 & 1 \\
(y-y') + \lambda_1 y/B^2 = 0 & 2 \\
-(x-x') + \lambda_2 (x'-\alpha) = 0 & 3 \\
-(y-y') + \lambda_2 (y'-\beta) = 0 & 4 \\
 & \\
x^2/A^2 + y^2/B^2 - 1 = 0 & 5 \\
(x'-\alpha)^2 + (y'-\beta)^2 - R^2 = 0 & 6
\end{array} \qquad (4.7)
$$

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 (after some algebra) an equation for y of the form

$$
k_1 y^4 + k_2 y^3 + k_3 y^2 + k_4 y + k_4 = 0 \ . \qquad x = \pm A \sqrt{1-(y/B)^2} \qquad (4.8)
$$

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 (4.7)? A possible first step is to find a viable set of λ_i. From a version of (2.9) using the first and third rows of (2.8) one may write,

$$-\begin{pmatrix} f_1 \\ f_3 \end{pmatrix} = \begin{pmatrix} a_1 & b_1 \\ a_3 & b_3 \end{pmatrix}\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \equiv M\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \qquad (2.9)$$

where

$$M = \begin{pmatrix} a_1 & b_1 \\ a_3 & b_3 \end{pmatrix} = \begin{pmatrix} a_1 & 0 \\ 0 & b_3 \end{pmatrix} \qquad \text{// } b_1 \text{ and } a_3 \text{ vanish from (4.4)} \qquad (4.9)$$

so that

$$M^{-1} = \begin{pmatrix} b_3 & 0 \\ 0 & a_1 \end{pmatrix} / \det(M) = \frac{1}{a_1 b_3}\begin{pmatrix} b_3 & 0 \\ 0 & a_1 \end{pmatrix} . \qquad (4.10)$$

Therefore,

$$-\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = M^{-1}\begin{pmatrix} f_1 \\ f_3 \end{pmatrix} = \frac{1}{a_1 b_3}\begin{pmatrix} b_3 & 0 \\ 0 & a_1 \end{pmatrix}\begin{pmatrix} f_1 \\ f_3 \end{pmatrix} \qquad (4.11)$$

and thus

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

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

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

We then have a candidate solution set $\{\lambda_i\}$. The first four equations of (4.7) become

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

or

$$\begin{aligned} &(x-x') - (x-x') = 0 && 1 \\ &(y-y') - (A/B)^2(x-x')y/x = 0 && 2 \\ &-(x-x') + (x-x') = 0 && 3 \\ &-(y-y') + (y'-\beta)(x-x')/(x'-\alpha) = 0 && 4 \ . \end{aligned} \qquad (4.12)$$

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

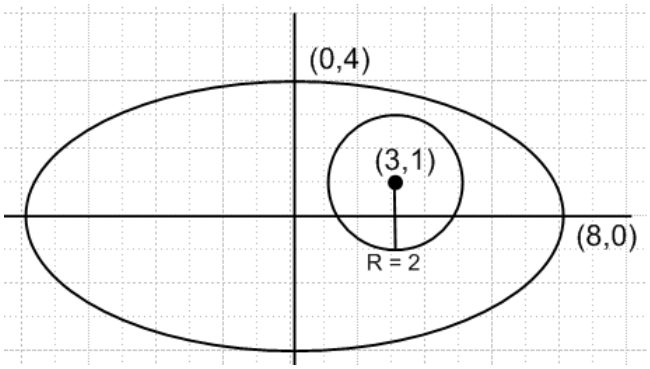
$$
\begin{aligned}
&(y-y') - (A/B)^2(x-x')y/x = 0 && 2\\
&-(y-y') + (y'-\beta)(x-x')/(x'-\alpha) = 0 && 4\\
&\text{or}\\
&(y-y')x - (A/B)^2(x-x')y = 0 && 2\\
&-(y-y')(x'-\alpha) + (y'-\beta)(x-x') = 0 && 4\ .
\end{aligned}
\tag{4.13}
$$

Including the last two equations of (4.7), one has four equations in four unknowns x,y,x',y' :

$$
\begin{aligned}
&(y-y')x - (A/B)^2(x-x')y = 0 && 2\\
&-(y-y')(x'-\alpha) + (y'-\beta)(x-x') = 0 && 4\\
&\\
&x^2/A^2 + y^2/B^2 - 1 = 0 && 5\\
&(x'-\alpha)^2 + (y'-\beta)^2 - R^2 = 0\ . && 6
\end{aligned}
\tag{4.14}
$$

This is a system of four 2nd degree polynomial equations in four variables. When we ask Maple to analytically solve these equations for the four unknowns x,y,x',y', it succeeds but 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,



(4.15)

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 (4.7) 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;
```

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

(4.16)

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

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

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

(4.17)

Without given a starting point, Maple finds the following solution

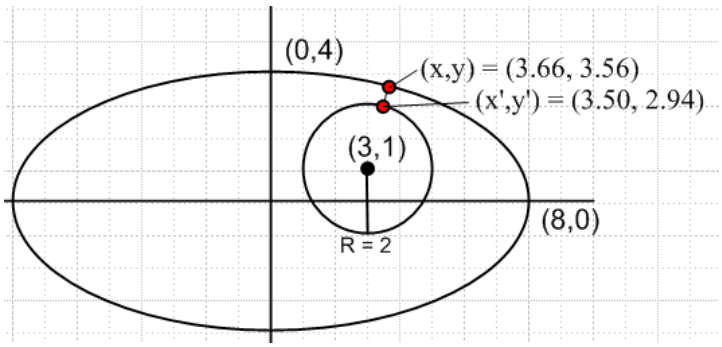
```
fsolve({e1,e2,e3,e4,e5,e6},{x,y,xp,yp,L1,L2});
```

$$\{L1 = -2.790619099,\ L2 = .3203235392,\ x = 3.657316705,\ xp = 3.497845176,\ y = 3.557528444,\ yp = 2.937046768\}$$

which we approximate as

$\lambda_1 = -2.8$ $\lambda_2 = .32$ (x,y) = (3.66, 3.56) (x',y') = (3.50, 2.94) . (4.18a)

Plotting this on our figure,



(4.18b)

we see that this solution corresponds to the minimum distance between the two curves. Maple computes this minimum distance to be $\sqrt{(x-x')^2+(y-y')^2}$ = .6406470846 units.

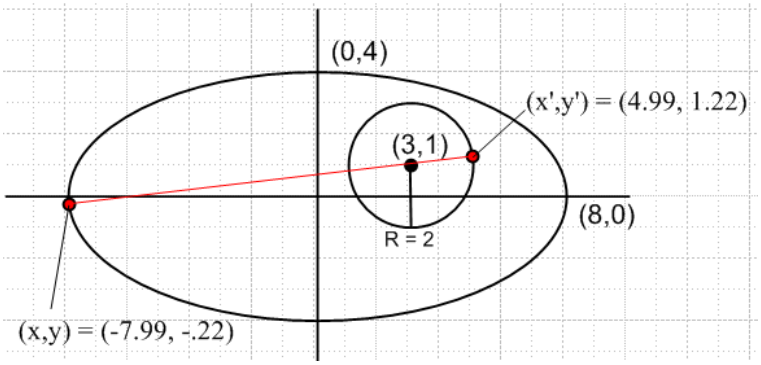
Another solution is found by giving Maple a search starting point (-8,0) on the ellipse and (5,1) on the circle :

```
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

$\lambda_1 = -104$ $\lambda_2 = -6.5$ (x,y) = (-7.99, -.22) (x',y') = (4.99, 1.22) . (4.19a)

Plotting this solution,



(4.19b)

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

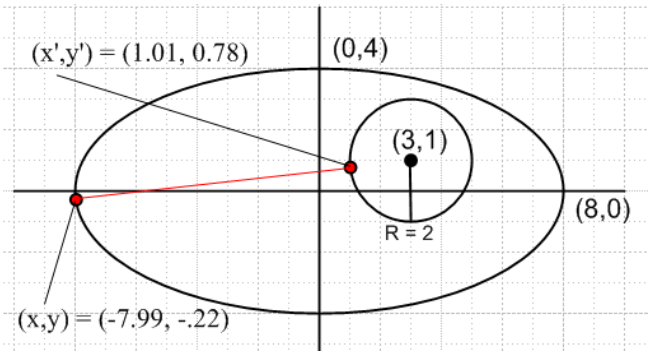
A third candidate solution is found by giving Maple a starting point (-8,0) on the ellipse and (1,1) on the circle:

```
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

$\lambda_1 = -72 \quad \lambda_2 = 4.5 \qquad (x,y) = (-7.99, -.22) \qquad (x',y') = (1.01, 0.78)$. (4.20a)

Plotting this solution,



(4.20b)

one sees that this solution represents a "stationary point" but is not in fact a solution to the problem. A slight d**r** = (dx,dy,dx',dy') variation along both constraint curves from this solution point gives df = 0 where f is the squared distance (4.2).

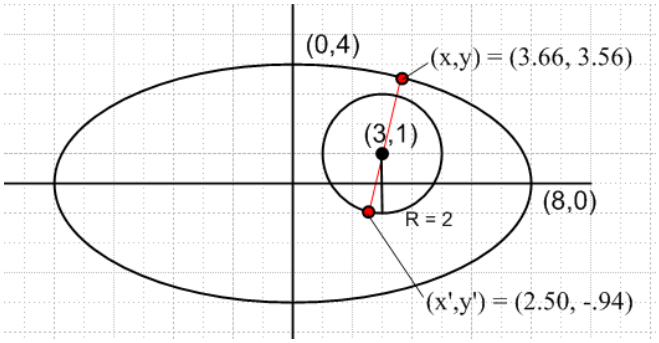
Finally, we find this fourth solution,

```
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

$\lambda_1 = -20.2 \quad \lambda_2 = -2.3 \qquad (x,y) = (3.66, 3.56) \qquad (x',y') = (2.50, -.94)$. (4.21a)

Plotting this solution,



(4.21b)

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

5. Example 5: The Boltzmann factor in Statistical Mechanics

In the following physics example the role of $\mathbf{r} = (x_1, x_2 \ldots x_m)$ is played by $\mathbf{N} = (N_1, N_2 \ldots N_m)$. The function f($\mathbf{r}$) to be maximized will be $\Omega(\mathbf{N}) = \frac{g_1^{N_1}}{N_1!}\frac{g_2^{N_2}}{N_2!} \ldots \frac{g_m^{N_m}}{N_m!}$. The two constraints are $a(\mathbf{N}) = \Sigma_i N_i - M = 0$ and $b(\mathbf{N}) = \Sigma_i \varepsilon_i N_i - U = 0$, all as described below.

5.1 Statement of The Boltzmann Extremum Problem

Imagine a physical system which has m **state levels** into which a set of M *identical* 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 = $(N_1, N_2 \ldots N_m) = \mathbf{N}$, 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 $\{N_i\} = \mathbf{N}$. 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 (hence "statistical" mechanics).

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

Placing N identical particles in the g degenerate states of a system at energy level ε is like placing N balls in a set of g boxes. The number of different "ways" of placing N indistinguishable balls into g boxes (bins) 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. The asterisks are the identical balls. 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). In this case the number of unique layouts is $\binom{9}{5} = \frac{9!}{5!4!} = \binom{9}{4}$. (5.1.1)

The count $\binom{N+g-1}{N}$ allows multiple balls (particles) to be in any given box (state). This count is then associated with so-called **Bose-Einstein statistics** -- multiple bosons are allowed in the same state. Bosons are particles with integral spin such as photons (S=1) and ground-state helium atoms (S=0).

If at most one ball is allowed in any given box, then one must have $N \leq g$ and the count of ways is just the number of ways to pick a committee of N boxes (one ball into each box) 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. Fermions are particles having half-integral spin. Because electrons (S=1/2) are fermions, atomic energy levels get "filled up" as N increases, allowing there to be some small number of electrons at the highest energy level ("valence electrons") which cause an atom of atomic number N (an "element") to have certain chemical properties. If electrons were bosons, they would all to go into the lowest or "ground" atomic state and the world would be a very different place. In other words, 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:

$$\text{ways} = \binom{N+g-1}{N} \approx \binom{g}{N} = \frac{g!}{N!\,(g-N)!} = \frac{g(g-1)..(g-N+1)}{N!} \approx \frac{g^N}{N!} \ . \qquad (5.1.2)$$

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

$$\Omega(N_1,N_2...N_m) = \frac{g_1^{\,N_1}}{N_1!}\frac{g_2^{\,N_2}}{N_2!} .... \frac{g_m^{\,N_m}}{N_m!} \ . \qquad \text{// see for example Zemansky Eq. (10-4)} \qquad (5.1.3)$$

Since the particles are identical, there is only one distinct way to partition the set of N particles into sets having counts N_1, N_2....N_m (prior to putting the particles of each set into that set's states as done above). For example, put the N particles in a straight line, and draw m-1 divider lines as needed to get the desired partition counts. So one could add an overall factor of "1" to (5.1.3) to denote this one partition.

Here then 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:

$$\Sigma_i N_i = M \qquad \text{and} \qquad \Sigma_i \varepsilon_i N_i = U \ . \qquad (5.1.4)$$

When M is a very large number (like Avogadro's number), it turns out (coming soon) 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 Sections 1 and 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,

$$M\varepsilon_{min} \leq U \leq M\varepsilon_{max} \ . \tag{5.1.5}$$

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,

$$u \equiv U/M \ = \text{average energy of a particle in the system} \tag{5.1.6}$$

and then (5.1.5) states the obvious fact that

$$\varepsilon_{min} \leq u \leq \varepsilon_{max} \ . \tag{5.1.7}$$

5.2 Solution of The Boltzmann Extremum Problem

Maximizing the state count Ω is the same as maximizing $f \equiv \ln\Omega$, and from (5.1.3),

$$f(N_1, N_2...N_m) \equiv \ln\Omega(N_1,N_2...N_m) = \Sigma_i [N_i \ln g_i] - \Sigma_i [\ln N_i!] \tag{5.2.1}$$

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

$$x! \approx \sqrt{2\pi x}\, x^x e^{-x} \quad \Rightarrow \quad \ln x! \approx \ln\sqrt{2\pi} + (1/2)\ln x + x\ln x - x \approx x\ln x - x, \quad x >> 1 \ . \tag{5.2.2}$$

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

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

$$= \Sigma_i [(1+\ln g_i)N_i - N_i \ln N_i] = \Sigma_i N_i [1 + \ln(g_i/N_i)] \ . \tag{5.2.3}$$

Here then is a restatement of the extremum problem given in (5.1.4):

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:

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

$b(\mathbf{N}) = \Sigma_i \varepsilon_i N_i - U = 0$. (5.2.4)

Application of the Method of Lagrange Multipliers

Following now the prescription of Section 2, we write our "Lagrangian" H of (2.2) as

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

$$= \Sigma_i [(1+\ln g_i)N_i - N_i \ln N_i] + \lambda_1 [\Sigma_i N_i - M] + \lambda_2 [\Sigma_i \varepsilon_i N_i - U] \ . \qquad (5.2.5)$$

The next step is to compute the derivatives with respect to N_i and λ_i and then set those derivatives to 0. We know that the two equations $\partial H/\partial\lambda_i = 0$ just replicate our two constraints, so it is the other m derivatives which are of interest. We again use the subscript notation for partial derivatives introduced at the start of Section 4, so now $F_i \equiv \partial F(\mathbf{N})/\partial N_i$. We need these three contributions to H_i,

$$f = \Sigma_i [(1+\ln g_i)N_i - N_i \ln N_i] \Rightarrow$$

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

$$a = \Sigma_i N_i - M \Rightarrow$$

$$a_i = \partial a/\partial N_i = 1$$

$$b = \Sigma_i \varepsilon_i N_i - U \Rightarrow$$

$$b_i = \partial b/\partial N_i = \varepsilon_i \ . \qquad i = 1,2...m \qquad (5.2.6)$$

Now set the H_i partials to 0 :

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

or

$$\ln(N_i/g_i) = \lambda_1 + \lambda_2 \varepsilon_i \ . \qquad i = 1,2...m \qquad (5.2.7)$$

Here then are the equations we need to solve,

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

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

$$\Sigma_i \varepsilon_i N_i = U \qquad \text{// } b(\mathbf{N}) = 0 \ . \qquad (5.2.8)$$

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

$$N_i = g_i\, e^{\lambda_1}\, e^{\lambda_2 \varepsilon_i} \quad . \tag{5.2.9}$$

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

$$\beta \equiv -\lambda_2$$

// derived Lagrange multipliers

$$A \equiv e^{\lambda_1} \tag{5.2.10}$$

and then (5.2.9) reads,

$$N_i = A\, g_i\, e^{-\beta \varepsilon_i} \quad . \qquad // = \frac{M}{Z}\, g_i\, e^{-\beta \varepsilon_i}, \quad \text{see below} \tag{5.2.11}$$

So of we can somehow determine A and β, the problem is solved and the solution $\mathbf{N} = (N_1, N_2 ... N_m)$ is known.

When a problem has a solution of the form (5.2.11) it is said to have **Maxwell-Boltzmann statistics**.

The notion of absolute temperature

Suppose $\varepsilon_i > \varepsilon_j$ and consider from (5.2.11) that $N_i/N_j = (g_i/g_j)\, e^{-\beta(\varepsilon_i - \varepsilon_j)}$. As $\beta \to +\infty$, $N_i/N_j \to 0$ and there are no particles in the upper state i compared to the lower state j. Basically, as $\beta \to +\infty$ all boson particles will inhabit only the lowest energy state (the "ground state") of a system. This is a situation one associates with absolute zero temperature T = 0. The system is "frozen" into its ground state. For fermions, the particles fill up the energy levels from the bottom up to some level, and above that level $N_i = 0$ (frozen atoms, white dwarfs, neutron stars, etc.). On the other hand, as $\beta \to 0$, we find that $N_i/N_j = g_i/g_j$ and particles are then distributed statistically according to state degeneracy counts without regard to the energy level difference between the states. This is the situation at an infinitely high temperature $T \to \infty$. So a simple way to *define* **absolute temperature** is $T \equiv C/\beta$ where C is any positive constant.

Measurements show that the T=0 situation exists when $T_C = -273.15$, where T_C is temperature in Celsius (centigrade) units. In order to have the size of one degree of absolute temperature units T be the same as that for T_C, one must write $T = T_C + 273.15$. The units of T are called K after Lord Kelvin, 1824-1907, aka William Thomson. Thus, the freezing point of water (triple point 0.01°C) corresponds to T = 273.16K. Room temperature $T_F = 70°F = 21°C$ corresponds to T = 294K.

In Section 5.4 below we shall show that the mean energy of a particle of ideal gas is given by $u = (3/2)(1/\beta)$. Writing this using $T \equiv C/\beta$ gives $u = (3/2)(1/C)T$. In effect, one can take a box of ideal gas at temperature T and measure the mean energy u, and this then determines the constant C. In practice this measurement of C is done indirectly, for example by measuring the speed of sound in argon gas, and C is now measured to about 7 decimal points of accuracy. Historically one deals with the inverse constant k =

1/C, and that constant k is known as **Boltzmann's constant**. The upshot of this discussion is that for systems in thermal equilibrium, one can associate the derived Lagrange multiplier $\beta = -\lambda_2$ with absolute temperature T according to :

$$\beta = 1/(kT) \qquad k = 1.380648 \times 10^{-23}\ \text{J/K} \ . \qquad \text{// Joules/Kelvin} \qquad (5.2.12)$$

Notice from (5.2.11) or $u = (3/2)(1/\beta)$ that the units of β must be inverse energy, so kT has energy units.

The factor $e^{-\beta\varepsilon_i}$ in (5.2.11) is referred to as **the Boltzmann factor**.

The quantity $S \equiv k \ln\Omega = kf$ is the **entropy** of a system, where k makes another appearance.

These matters are discussed in any thermodynamics or statistical mechanics text. See for example Zemansky p 258-272 and equation (10-17).

<u>How the derived Lagrange multiplier values β and A are determined.</u>

From (5.2.11) the probability of a particle having energy ε_i is given by

$$p_i = \frac{N_i}{M} = \frac{A}{M}\ g_i\, e^{-\beta\varepsilon_i} \ . \qquad (5.2.13)$$

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),

$$A = \frac{M}{\Sigma_i g_i\, e^{-\beta\varepsilon_i}} \ . \qquad // = \frac{M}{Z} \ , \quad Z \equiv \Sigma_i g_i e^{-\beta\varepsilon_i} \qquad (5.2.14)$$

The denominator in (5.2.14) is called **the partition function** Z for the system, so then A = M/Z.

Using (5.2.11) for N_i, the two constraint equations shown in (5.2.4) [$\Sigma_i N_i = M$ and $\Sigma_i N_i \varepsilon_i = U$] read

$$A\,\Sigma_i g_i e^{-\beta\varepsilon_i} = M \qquad // \ N_i = A\ g_i\, e^{-\beta\varepsilon_i}$$
$$A\,\Sigma_i \varepsilon_i g\, e^{-\beta\varepsilon_i} = U \ . \qquad (5.2.15)$$

Dividing these two equations and using (5.1.6) that u = U/M, one gets this self-consistent result,

$$u = \frac{U}{M} = \frac{\Sigma_i g_i \varepsilon_i e^{-\beta\varepsilon_i}}{\Sigma_i g_i e^{-\beta\varepsilon_i}} = \frac{(A/M)\Sigma_i g_i \varepsilon_i e^{-\beta\varepsilon_i}}{(A/M)\Sigma_i g_i e^{-\beta\varepsilon_i}} = \frac{\Sigma_i p_i \varepsilon_i}{1} = \Sigma_i p_i \varepsilon_i = \langle\varepsilon_i\rangle \ . \qquad (5.2.16)$$

To find the derived Lagrange multiplier constant $\beta = -\lambda_2$, write out the second equation of (5.2.15),

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

or

$$\Sigma_i\ g_i\ [\varepsilon_i - u]\, e^{-\beta\varepsilon_i} = 0 \ . \qquad (5.2.17)$$

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. It is possible then to adjust β to bring these positive and negative contributions into balance to generate a 0 sum. One can see from (5.2.17) that β is a function of mean energy u.

For general values of ε_i, equation (5.2.17) 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.

Summary

We have used the method of Lagrange multipliers to solve the Boltzmann extremum problem stated in either (5.1.4) or (5.2.4). 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 (5.2.17), and then A is determined by (5.2.14). Then the solution (extremum) vector $\mathbf{N} = \{N_i\}$ is given by (5.2.11) 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 for large N_i the microstate count $\Omega(\mathbf{N})$ 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.

5.3 More details of the solution

Recall from (5.2.3) the function for which we seek a stationary point,

$$f = \ln\Omega = \Sigma_{i=1}^{m} N_i\,[1 + \ln g_i - \ln N_i] \quad . \qquad \textit{(5.2.3)} \qquad (5.3.1)$$

The constraints $\Sigma_{i=1}^{m} N_i = M$ and $\Sigma_{i=1} N_i \varepsilon_i = U$ can be regarded as two equation in the two unknowns N_1 and N_2 ,

$$N_1 + N_2 = M - \Sigma_{i=3}^{m} N_i$$

$$\varepsilon_1 N_1 + \varepsilon_2 N_2 = U - \Sigma_{i=3}^{m} \varepsilon_i N_i \ , \qquad (5.3.2)$$

which are easily solved to obtain,

$$N_1 = +(\varepsilon_2-\varepsilon_1)^{-1}[\,\varepsilon_2 M - U + \Sigma_{i=3}^{m}(\varepsilon_i-\varepsilon_2)N_i] \quad = N_1(N_3,\ldots N_m)$$
$$N_2 = -(\varepsilon_2-\varepsilon_1)^{-1}[\,\varepsilon_1 M - U + \Sigma_{i=3}^{m}(\varepsilon_i-\varepsilon_1)N_i] \quad = N_2(N_3,\ldots N_m)\ . \qquad (5.3.3)$$

Notice that,

$$\frac{\partial N_1}{\partial N_i} = +(\varepsilon_2-\varepsilon_1)^{-1}\,[\,(\varepsilon_i-\varepsilon_2)\,] = \frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1} \qquad i = 3,4..m$$
$$\frac{\partial N_2}{\partial N_i} = -(\varepsilon_2-\varepsilon_1)^{-1}\,[\,(\varepsilon_i-\varepsilon_1)\,] = -\frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1} \qquad i = 3,4..m\ . \qquad (5.3.4)$$

Define function F of m-2 *independent* variables of $N_3...N_m$ as follows :

$$F(N_3,N_4.....N_m) \equiv f(N_1(N_3,\ldots N_m), N_2(N_3,\ldots N_m), N_3, ...N_m)\ . \qquad (5.3.5)$$

F has the same value as f but has a different functional form. We rewrite (5.3.1) as

$$F = N_1[\,1 + \ln(g_1) - \ln N_1] + N_2\,[\,1 + \ln(g_2) - \ln N_2] + \Sigma_{j=3}^{m} N_j\,[1 + \ln(g_j) - \ln N_j] \qquad (5.3.6)$$

where recall that $N_1 = N_1(N_3,\ldots N_m)$ and $N_2 = N_2(N_3,\ldots N_m)$.

First Derivative

Compute $\partial_i F$ from (5.3.6) as follows,

$$\frac{\partial F}{\partial N_i} = \frac{\partial N_1}{\partial N_i}\,[\,1 + \ln(g_1) - \ln N_1] + N_1\,[-\frac{1}{N_1}]\frac{\partial N_1}{\partial N_i}$$
$$+\frac{\partial N_2}{\partial N_i}\,[\,1 + \ln(g_2) - \ln N_2] + N_2\,[-\frac{1}{N_2}]\frac{\partial N_2}{\partial N_i}$$
$$+\Sigma_{j=3}^{m}\{\,\delta_{i,j}\,[\,1 + \ln(g_j) - \ln N_j] + N_j\,[\,-1/N_j]\,\delta_{i,j}\,\}$$

$$= \frac{\partial N_1}{\partial N_i}\,[\,\ln(g_1) - \ln N_1] + \frac{\partial N_2}{\partial N_i}\,[\,\ln(g_2) - \ln N_2] + [\,\ln(g_i) - \ln N_i]$$

$$= \frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1}\,[-\ln(N_1/g_1)] + (-\frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1})\,[-\ln(N_2/g_2)] + [-\ln(N_i/g_i)\,] \qquad //\ (5.3.4)$$

$$= -\frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1}\,\ln(N_1/g_1)] + \frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1}\,\ln(N_2/g_2) - \ln(N_i/g_i) \qquad i = 3,4 .... m\ . \qquad (5.3.7)$$

This result is valid for any values of variables N_3, N_4....N_m. In order to evaluate $\partial F/\partial N_i$ at the stationary point of interest, recall from (5.2.9) that at the stationary point,

$$\ln(N_i/g_i) = \lambda_1 + \lambda_2\varepsilon_i \qquad i = 1,2...m\ . \qquad \textit{(5.2.9)}$$

Inserting this into (5.3.7) one finds, adding factor $1 = \frac{\varepsilon_2-\varepsilon_1}{\varepsilon_2-\varepsilon_1}$ to the last term,

$$\frac{\partial F}{\partial N_i}\Big|^{stat} = -\frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1}(\lambda_1+\lambda_2\varepsilon_1) + \frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1}(\lambda_1+\lambda_2\varepsilon_2) - \frac{\varepsilon_2-\varepsilon_1}{\varepsilon_2-\varepsilon_1}(\lambda_1+\lambda_2\varepsilon_i)$$

$$= \frac{1}{\varepsilon_2-\varepsilon_1}\{-(\varepsilon_i-\varepsilon_2)(\lambda_1+\lambda_2\varepsilon_1) + (\varepsilon_i-\varepsilon_1)(\lambda_1+\lambda_2\varepsilon_2) - (\varepsilon_2-\varepsilon_1)(\lambda_1+\lambda_2\varepsilon_i)\} \quad . \qquad (5.3.8)$$

Inside the Curly Bracket all terms cancel, as Maple verifies

```
CB := -(ei-e2)*(L1+L2*e1)+(ei-e1)*(L1+L2*e2)-(e2-e1)*(L1+L2*ei);
```

$$CB := -(ei-e2)(L1+L2\,e1)+(ei-e1)(L1+L2\,e2)-(e2-e1)(L1+L2\,ei)$$

```
simplify(CB);
```

$$0$$

with the final result

$$\frac{\partial F}{\partial N_i}\Big|^{stat} = 0 \qquad i = 3,4...m \quad . \qquad (5.3.9)$$

This is the expected result since the solution $\{N_i\}$ should be a regular stationary point of the unconstrained function $F(N_3,N_4.....N_M)$.

Second Derivative

Start with (5.3.7) written this way,

$$\frac{\partial F}{\partial N_i} = -\frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1}[\ln N_1 - \ln g_1] + \frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1}[\ln N_2 - \ln g_2] - [\ln N_i - \ln g_i] \quad . \qquad (5.3.10)$$

Then

$$\frac{\partial^2 F}{\partial (N_i)^2} = -\frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1}\frac{1}{N_1}\frac{\partial N_1}{\partial N_i} + \frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1}\frac{1}{N_2}\frac{\partial N_2}{\partial N_i} - \frac{1}{N_i} \qquad i = 3,4...m$$

$$= -\frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1}\frac{1}{N_1}\frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1} + \frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1}\frac{1}{N_2}\left(-\frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1}\right) - \frac{1}{N_i} \qquad // (5.3.4)$$

$$= -\left[\left(\frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1}\right)^2\frac{1}{N_1} + \left(\frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1}\right)^2\frac{1}{N_2} + \frac{1}{N_i}\right] = \partial_i{}^2 F \quad . \qquad (5.3.11)$$

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 $\mathbf{N} = \{N_i\}$ for i = 3,4.... This is of course then true for the solution vector **N**. But for the

solution vector **N** we also have $\partial f/\partial N_i = 0$ from (5.3.9) and therefore the solution is a *maximum* of F and therefore of Ω. Note that $\partial^2{}_i F$ is independent of the degeneracy g_i (in our $g_i >> N_i$ limit of interest) .

Finally, we may *estimate* the width of the peak of F and then of Ω. We follow the procedure outlined earlier in Section 3.4. Taking a variation only in a particular N_i (in the "i" direction), we approximate

$$F(\mathbf{N} + \Delta N_i \hat{\mathbf{i}}) \approx F(\mathbf{N}) + (\partial_i F)\,\Delta N_i + (1/2)\,(\partial_i{}^2 F)\,(\Delta N_i)^2 \qquad i = 3,4...m \ . \tag{5.3.12}$$

But at the solution point $(\partial F/\partial N_i) = 0$ as in (5.3.9) so

$$\Delta F \ = \ F(\mathbf{N} + \Delta N_i \hat{\mathbf{i}}) - F(\mathbf{N}) \ \approx \ (1/2)\,(\partial_i{}^2 F)\,(\Delta N_i)^2 \ .$$

Thus

$$|\Delta N_i| \ \approx \sqrt{\frac{2\,|\Delta F|}{|\partial_i{}^2 F|}} \qquad i = 3,4...m \ . \tag{5.3.13}$$

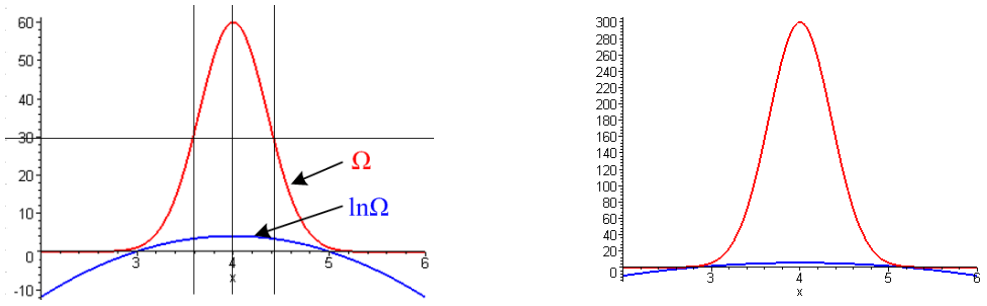
Suppose we are interested in finding ΔN_i such that the Ω(**N**) drops to half its peak value. Then

$$\Omega_{half}/\Omega_{peak} = 1/2 \quad \Rightarrow \quad \ln \Omega_{half} - \ln \Omega_{peak} = \ln(1/2) = -\ln(2) = F_{half} - F_{peak}$$

$$|\Delta F| = F_{peak} - F_{half} = \ln(2) = 0.7 \ . \tag{5.3.14}$$

Consider this toy example where Ω is taken to be a Gaussian function,

```
N:=60:
Omega := N*exp(-4*(x-4)^2):
plot([Omega,ln(Omega)],x=2..6, color= [red,blue],thickness = 2);
```



(5.3.15)

On the left one can see that the half point of Ω is about 0.7 the half point of lnΩ. The plot on the right has N = 300 and is more representative of what happens in statistical mechanics. The peak of Ω is usually a huge number (example in the next section) and lnΩ is barely visible if Ω and lnΩ are plotted on the same scale. But of course (5.3.14) is still valid.

Using |ΔF| = 0.7 in (5.3.13) along with (5.3.11) for the curvature, one finds

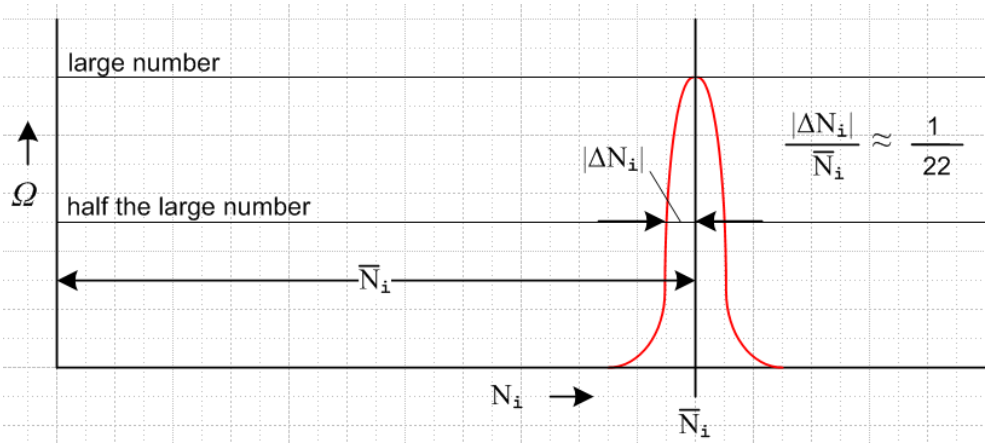
$$|\Delta N_i| \approx \sqrt{\frac{1.4}{|\partial_i{}^2 F|}} = \sqrt{\frac{1.4}{\left[\left(\frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1}\right)^2 \frac{1}{N_1} + \left(\frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1}\right)^2 \frac{1}{N_2} + \frac{1}{N_i} \right]}}$$

$$= \sqrt{\frac{1.4\, N_i}{\left(\frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1}\right)^2 (N_i/N_1) + \left(\frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1}\right)^2 (N_i/N_2) + 1}} \quad . \qquad (5.3.16)$$

The fractional width is then

$$\frac{|\Delta N_i|}{N_i} \approx \frac{1.2}{\sqrt{N_i}} \sqrt{\frac{1}{\left(\frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1}\right)^2 (N_i/N_1) + \left(\frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1}\right)^2 (N_i/N_2) + 1}} \quad . \qquad (5.3.17)$$

To clarify the meaning of this fractional width, let us temporarily indicate by $\{\bar{N}_k\}$ the components of the solution vector $\bar{\mathbf{N}}$ at which $\Omega(N_1,N_2 \ldots N_m)$ has its maximum value. Since we have taken N_1 and N_2 as dependent variables (functions of all the other N_i as in (5.3.3)), we can write $\Omega(N_1,N_2 \ldots N_m) = \mathit{\Omega}(N_3,N_4 \ldots N_m)$ where $\mathit{\Omega}$ is used only to distinguish the functional form. Here $N_3 \ldots N_m$ are all independent variables. Then consider $\mathit{\Omega}(\bar{N}_3,\bar{N}_4 \ldots N_i \ldots \bar{N}_m)$ where we treat N_i as the only variable and we allow it to vary through $\bar{N}_i$. Here is a sample plot illustrating the peak in $\mathit{\Omega}(\bar{N}_3,\bar{N}_4 \ldots N_i \ldots \bar{N}_m)$ versus N_i ,



(5.3.18)

Although the value at the peak might be a very large number (like 10^{1500}), the "narrowness" of the peak is really related to the ratio $|\Delta N_i|/\bar{N}_i$. If $\bar{N}_i \sim 10^{20}$, as might be typical for a macroscopic system, then $|\Delta N_i|/\bar{N}_i \lesssim 10^{-10}$ and the peak would be *extremely* narrow and sharp. Since this is true for any N_i one selects, one would in this large-$\bar{N}_i$ case conclude that in an ensemble of a billion systems, there is only a miniscule variation in the solution vectors $\bar{\mathbf{N}}$ for the different systems. This is so because for any system in the ensemble, the components N_i of the solution vector $\bar{\mathbf{N}}$ are extremely likely to be within a few ΔN_i of the solution value $\bar{N}_i$. The underlying assumption here is that a system is most likely to be found where its available microstate count is very large.

In the above analysis we took N_1 and N_2 to be the two dependent variables. Had we instead selected N_a and N_b as the dependent variables (those remaining then being independent), we would get these versions of equations above (italics) by taking 1→a and 2→b :

$$N_a = +(\varepsilon_b - \varepsilon_a)^{-1}[\,\varepsilon_b M - U + \Sigma_{i\neq a,b}\,(\varepsilon_i - \varepsilon_b)N_i] \;=\; N_a(N_{i\neq a,b})$$
$$N_b = -\,(\varepsilon_b - \varepsilon_a)^{-1}[\,\varepsilon_a M - U + \Sigma_{i\neq a,b}\,(\varepsilon_i - \varepsilon_a)N_i] \;=\; N_b(N_{i\neq a,b}) \qquad \textit{(5.3.3)}$$

$$\frac{\partial N_a}{\partial N_i} = \frac{\varepsilon_i - \varepsilon_b}{\varepsilon_b - \varepsilon_a} \qquad i = 1..m \text{ but } i \neq a,b$$
$$\frac{\partial N_b}{\partial N_i} = -\frac{\varepsilon_i - \varepsilon_a}{\varepsilon_b - \varepsilon_a} \qquad i = 1..m \text{ but } i \neq a,b \qquad \textit{(5.3.4)}$$

$$\frac{\partial F}{\partial N_i}\Big|^{stat} = 0 \qquad i = 1..m \text{ but } i \neq a,b \qquad \textit{(5.3.8)}$$

$$\frac{\partial^2 F}{\partial (N_i)^2} = -\left[\left(\frac{\varepsilon_i - \varepsilon_b}{\varepsilon_b - \varepsilon_a}\right)^2 \frac{1}{N_a} + \left(\frac{\varepsilon_i - \varepsilon_a}{\varepsilon_b - \varepsilon_a}\right)^2 \frac{1}{N_b} + \frac{1}{N_i}\right] = \partial_i{}^2 F \qquad \textit{(5.3.11)}$$

$$\frac{|\Delta N_i|}{N_i} \approx \frac{1.2}{\sqrt{N_i}} \sqrt{\frac{1}{\left(\frac{\varepsilon_i - \varepsilon_b}{\varepsilon_b - \varepsilon_a}\right)^2 (N_i/N_a) + \left(\frac{\varepsilon_i - \varepsilon_a}{\varepsilon_b - \varepsilon_a}\right)^2 (N_i/N_b) + 1}} \;. \qquad \textit{(5.3.17)} \qquad (5.3.19)$$

5.4 Example: N particles in 3 states : a numerical example

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

$$\Omega(N_1,N_2,N_3) = \frac{g_1^{N_1}}{N_1!}\frac{g_2^{N_2}}{N_2!}\frac{g_3^{N_3}}{N_3!} \qquad \textit{(5.1.3)} \qquad (5.4.1)$$

and correspondingly (using the Stirling approximation for $N_i!$),

$$f = \ln\Omega = \Sigma_{i=1}^{3} N_i\,[1 + \ln(g_i/N_i)]\,. \qquad \textit{(5.2.3)} \qquad (5.4.2)$$

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

$$\varepsilon_1 < \varepsilon_2 < \varepsilon_3 \qquad \text{and recall} \qquad u = U/M \quad \text{and } \varepsilon_1 \le u \le \varepsilon_3\ . \qquad (5.4.3)$$

We regard both N_1 and N_2 as functions of N_3 where, according to (5.3.3),

$$\begin{aligned} N_1 &= +(\varepsilon_2 - \varepsilon_1)^{-1}[\,\varepsilon_2 M - U + (\varepsilon_3 - \varepsilon_2)N_3] = N_1(N_3) \\ N_2 &= -(\varepsilon_2 - \varepsilon_1)^{-1}[\,\varepsilon_1 M - U + (\varepsilon_3 - \varepsilon_1)N_3] = N_2(N_3)\ . \end{aligned} \qquad \textit{(5.3.3)} \qquad (5.4.4)$$

It seems clear that one must have

$$\begin{aligned} &0 \le N_1 \le M \\ &0 \le N_2 \le M\,. \end{aligned} \qquad (5.4.5)$$

Using the expressions (5.4.4) in these inequalities, and making use of (5.4.3), one finds that all four inequalities can be summarized as just two inequalities which we write as,

$$\begin{aligned} &N_{3min} \le N_3 \le N_{3max} \\ &\quad N_{3min} = \max\left(\frac{u-\varepsilon_2}{\varepsilon_3-\varepsilon_2}M,\ 0\right) \\ &\quad N_{3max} = \frac{u-\varepsilon_1}{\varepsilon_3-\varepsilon_1}M\,. \end{aligned} \qquad (5.4.6)$$

The above fact is tedious to derive and we state it only because it is used in the code below for the cosmetic purpose of restricting the N_3 range of some graphs.

Recall from (5.2.10) that the derived Lagrange multipliers are $\beta \equiv -\lambda_2$ and $A \equiv e^{\lambda_1}$.

To find β we must solve (5.2.17),

$$\Sigma_{i=1}^{3}\,[\,\varepsilon_i - u\,]\,g_i\,e^{-\beta\varepsilon_i} = 0\,. \qquad \textit{(5.2.17)} \qquad (5.4.7)$$

Then the other derived multiplier A is given by (5.2.14)

$$A = \frac{M}{\Sigma_i g_i\, e^{-\beta\varepsilon_i}} = \frac{M}{Z} \quad \text{where} \quad Z \equiv \Sigma_i g_i e^{-\beta\varepsilon_i} \ . \qquad (5.2.14) \qquad (5.4.8)$$

The components N_i of the solution vector **N** are then given by (5.2.11),

$$N_i = A\ g_i\, e^{-\beta\varepsilon_i} \ . \qquad i = 1,2,3 \qquad (5.2.11) \qquad (5.4.9)$$

The curvature of f [= F as in (5.3.5)] at its peak is shown in (5.3.11),

$$\frac{\partial^2 F}{\partial^2 N_3} = -\left[\left(\frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1}\right)^2 \frac{1}{N_1} + \left(\frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1}\right)^2 \frac{1}{N_2} + \frac{1}{N_3} \right] . \qquad (5.3.11) \qquad (5.4.10)$$

The half-width of the Ω peak is then given by (5.3.16),

$$\Delta N_3 \approx \sqrt{\frac{1.4\, N_3}{\left(\frac{\varepsilon_i-\varepsilon_2}{\varepsilon_2-\varepsilon_1}\right)^2 (N_3/N_1) + \left(\frac{\varepsilon_i-\varepsilon_1}{\varepsilon_2-\varepsilon_1}\right)^2 (N_3/N_2) + 1}} \ . \qquad (5.3.16) \qquad (5.4.11)$$

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

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

$$\Omega := \frac{g_1^{N_1}\, g_2^{N_2}\, g_3^{N_3}}{N_1!\, N_2!\, N_3!} \qquad (5.4.12)$$

The quantities N_1 and N_2 are then replaced as in (5.4.4) [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]);
```

$$N_2 := -\frac{e_1 M - U + (e_3 - e_1) N_3}{e_2 - e_1} \qquad (5.4.13)$$

For a specific example we assume these values,

$\varepsilon_3 = 3$ M = 1000 = number of particles
$\varepsilon_2 = 2$ u = 1.5 = average energy
$\varepsilon_1 = 1$ $g_1 = g_2 = g_3 = 5000$ = degeneracy . (5.4.14)

Enter values and use them to compute N_{3min} and N_{3max} and then to restate $N_1(N_3)$ and $N_2(N_3)$,

```
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[2]);
```

$$N3min := 0$$

$$N3max := 250.0000000$$

```
N[1];
```

$$500.0 + N_3$$

```
N[2];
```

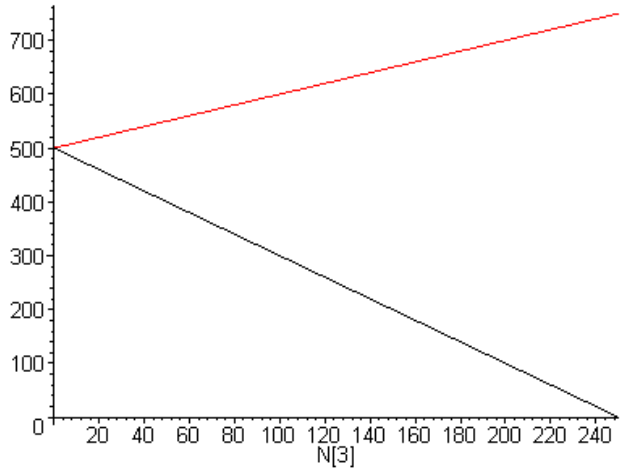
$$500.0 - 2\,N_3 \qquad (5.4.15)$$

For this example one then has

$$N_1 = 500 + N_3 \quad \Rightarrow \quad dN_1 = +1\ dN_3$$
$$N_2 = 500 - 2N_3 \quad \Rightarrow \quad dN_2 = -2\ dN_3 \qquad (5.4.16)$$

which we plot as follows with N_1 in red and N_2 in black,

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



(5.4.17)

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;
```

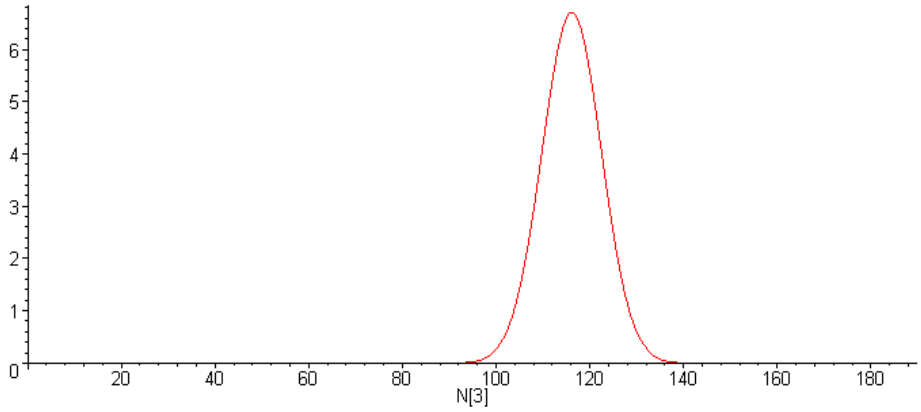
$$\frac{5000^{(500.0+N_3)}\ 5000^{(500.0-2N_3)}\ 5000^{N_3}}{(500.0+N_3)!\,(500.0-2\,N_3)!\,N_3!} \qquad (5.4.18)$$

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)

$$\Omega_{peak} = 5.71 \times 10^{1519} \ . \qquad (5.4.19)$$

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 scaled $\Omega(N_1(N_3),N_2(N_3),N_3)$,

```
OmegaScaled := (Omega/scale)*Heaviside(Omega/scale-.01):
xx := plot(  OmegaScaled,N[3] = N3min..190,numpoints=100  ):
display(xx);
```

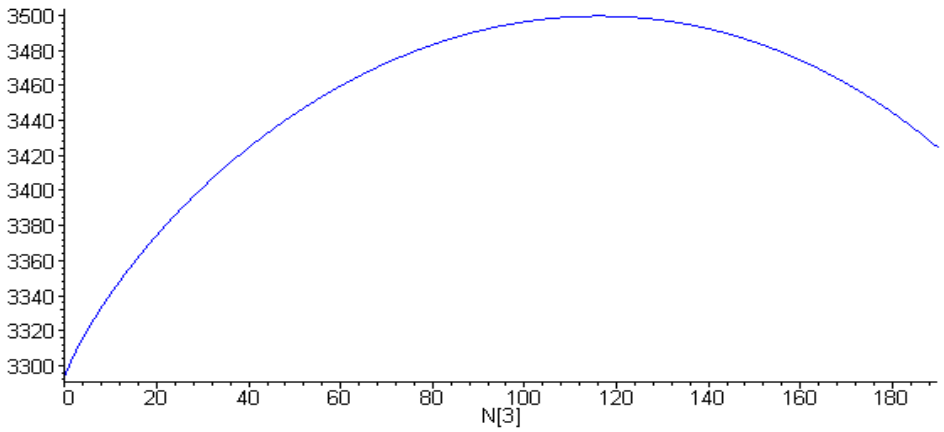


(5.4.20)

which may be compared with our generic plot of Fig (5.3.18). 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(Ω) has a much smoother and broader shape, reminiscent of Fig (5.3.15),

```
plot(ln(Omega),N[3] = N3min..190,numpoints=100, color = blue );
```



(5.4.21)

Recall from (5.4.10) that $\frac{\partial^2 F}{\partial^2 N_3}$ is everywhere negative, consistent with the cupping down seen in this plot.

Next, we have Maple solve (5.4.7) 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];
```

$$\beta := .8341151943 \tag{5.4.22}$$

With $w = e^{-\beta}$ the above "eq" equation (divided by 2500w) is just quadratic $3w^2+w-1 = 0$ which has solutions $w_\pm = (-1 \pm \sqrt{13})/6$. Since $\beta = -\ln w$, the w_- solution gives the complex (and therefore illegal) value .26 - iπ shown above in (5.4.22), while the w_+ solution is $\beta = -\ln[(-1 + \sqrt{13})/6] = .834$.

From this β value we compute the other Lagrange parameter A from (5.4.8),

```
Z := sum(g[i]*exp(-beta*e[i]),i=1..3);
```

$$Z := 3523.658588$$

```
A := M/Z;
```

$$A := .2837959397 \tag{5.4.23}$$

Once these parameters are known, we use (5.4.9) 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,
```

$$N_3 := 116.2040605$$

$$N_2 := 267.5918793$$

$$N_1 := 616.2040607 \tag{5.4.24}$$

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

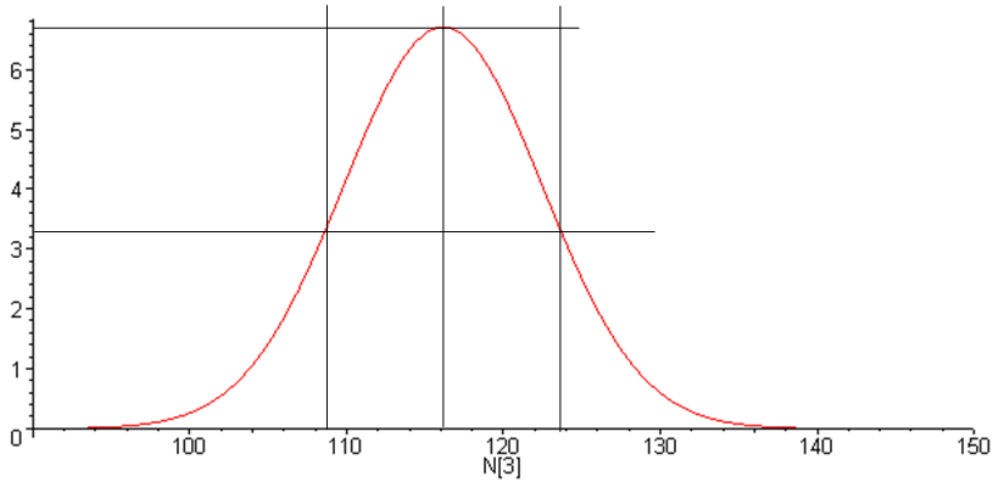
Our final task is to use (5.4.11) determine the half-width of the Ω peak,

```
D31 := ((e[3]-e[2])/(e[2]-e[1]))^2 * (N[3]/N[1]):
D32 := ((e[3]-e[1])/(e[2]-e[1]))^2 * (N[3]/N[2]):
D3 := D31+D32+1:
DN3 := sqrt(1.4*N[3]/D3);
```

$$DN3 := 7.457033725 \tag{5.4.25}$$

which says $\Delta N_3 \approx 7.5$. 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);
```



(5.4.26)

Notice that our Maple code does not use the Stirling approximation for $N_3!$ in the Ω of (5.4.18) but provides a continuous interpolation of $N_3!$ as $\Gamma(N_3+1)$. All the plots above should perhaps be indicated with dotted lines to emphasize the fact that the N_3 axis is really discrete integers.

The peak values of Ω and f = ln(Ω) are found to be,

```
evalf(Omega);
```

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

```
ln(evalf(Omega));
```

$$3499.530441$$

(5.4.27)

To see more directly where this huge Ω number is coming from, we have Maple compute the three factors in the expression (5.4.1) for Ω using the approximate solution values for N_1, N_2, N_3 shown in (5.4.20):

$$\Omega(N_1,N_2,N_3) = \frac{g_1^{N_1}}{N_1!}\frac{g_2^{N_2}}{N_2!}\frac{g_3^{N_3}}{N_3!} \qquad (5.4.1)$$

```
n1 := evalf(5000^616/ (616!));
```

$$n1 := .8227460158\ 10^{826}$$

```
n2 := evalf(5000^268/ (268!));
```

$$n2 := .2298609194\ 10^{456}$$

```
n3 := evalf(5000^116/ (116!));
```

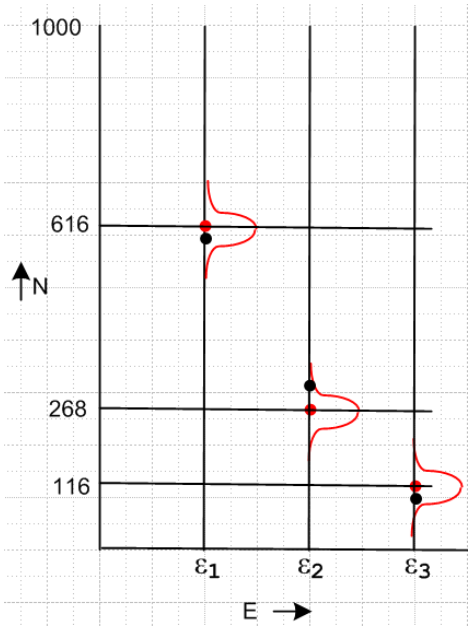
$$n3 := .3547502680\ 10^{239}$$

```
n1*n2*n3;
```

$$.6708936163\ 10^{1520}$$

(5.4.28)

Consider now the following schematic diagram:



(5.4.29)

The red dots show the solution values for the integers (N_1,N_2,N_3). These would be the values obtained by averaging the values measured in 1 billion systems in an ensemble. If we take N_3 to be the independent variable, it has a probability distribution (bell curve) shown in (5.4.20). For any point on this curve, $N_1(N_3)$ and $N_2(N_3)$ will also be displaced from their red-dot values. We roughly show this variation with little bell curves in the figure. For example, a particular system taken from the ensemble might have (N_1,N_2,N_3) values indicated by the black dots in the figure. The displacements shown are roughly these,

$$dN_3 = -50 \quad \Rightarrow \quad dN_1 = -50 \text{ and } dN_2 = +100 \quad . \qquad \text{// see (5.4.16)} \tag{5.4.30}$$

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

Geometric Interpretation

The above example is similar to our Example 2 of Section 3.2 involving a hypersphere. In both problems there are three components, here $\mathbf{N} = (N_1,N_2,N_3)$ and there $\mathbf{r} = (x,y,z)$. In both problems there are two constraints, here $\Sigma N_i = N$ and $\Sigma \varepsilon_i N_i = U$ and there x=1 and y=1. In both problems we can imagine plotting the function of interest in E^4, here $f = \Sigma_{i=1}^{3} N_i [1 + \ln(g_i/N_i)]$ and there $f = \sqrt{4 - x^2 - y^2 - z^2}$. In both problems the "intersection constraint surface" of Section 1 has dimension *N*-C = 3-2 = 1. For the hypersphere problem, that intersection constraint surface is the line which is the intersection of the x=1 and y=1 planes drawn in E^3. In our current problem the intersection constraint surface is the line which is the intersection of the two planes $N_1+N_2+N_3 = N$ and $\varepsilon_1N_1+\varepsilon_2N_2+\varepsilon_3N_3 = U$ in E^3. The first plane has

normal (1,1,1) and its closest approach to the origin is distance N. The second plane has normal ($\varepsilon_1,\varepsilon_2,\varepsilon_3$) and its closest approach to the origin is distance U.

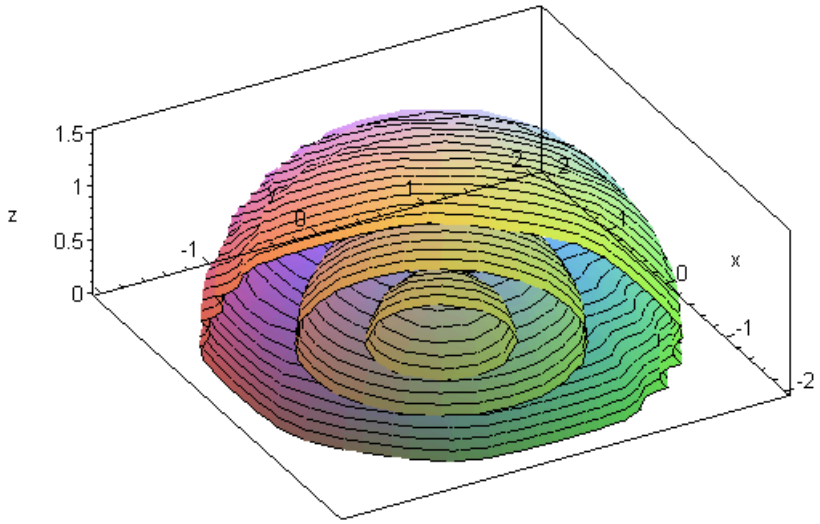
Before discussing our current problem, it is useful to review the hypersphere Example 2.

Recall that in Example 2 the "level surfaces" in E^3 are determined by $K = f(\mathbf{r}) = \sqrt{4 - x^2 - y^2 - z^2}$ for various values of K, and these surfaces are a set of concentric spheres. Here is some crude Maple code to display three of these spherical level surfaces :

```
f := sqrt(4-x^2-y^2-z^2);
```

$$f := \sqrt{4 - x^2 - y^2 - z^2}$$

```
implicitplot3d({Re(f) = 0.5,Re(f) = 1.6, Re(f) = 1.9}, x=-2..2,y=-2..2,z=0..1.5,
scaling = constrained,style = patchcontour,grid = [15,15,15]);
```



(5.4.31)

We have restricted the z range to obtain a cutaway view of the concentric spheres. In this picture the constraints are the planes x = 1 and y = 1 which intersect in a line parallel to the z axis. The solution must lie on this line, and there is one point on the line where the gradients to the planes x=1 and y=0 are coplanar with the gradient (normal) of one of the concentric spheres. As we saw, and as is clear here, that point will lie at the z = 0 plane and has the value **r** = (1,1,0). This point corresponds to the smallest radius concentric sphere touched by the constraint line and thus gives a maximum for $f(\mathbf{r}) = \sqrt{4 - x^2 - y^2 - z^2}$ subject to the constraints x=1 and y = 1.

In the current problem, the level surfaces are determined by

$$K = f(\mathbf{N}) = \Sigma_{i=1}^{3} N_i \, [1 + \ln(g_i/N_i)] \qquad (5.4.32)$$

for various values of K. For more familiarity, write this with $(N_1,N_2,N_3) = (x,y,z)$ to get

$$K = f(\mathbf{r}) = x[1+\ln(g_1/x)] + y[1+\ln(g_2/y)] + z[1+\ln(g_3/z)]$$

$$= (1+ \ln g_1)\, x + (1+ \ln g_2)y + (1+ \ln g_3)z - (x\ln x + y\ln y + z\ln z) \ . \qquad (5.4.33)$$

The two constraint planes are then

$$x_1+x_2+x_3 = N \qquad \text{and} \qquad \varepsilon_1 x_1+\varepsilon_2 x_2+\varepsilon_3 x_3 = U\ . \tag{5.4.34}$$

Using $g_i = 5000$ as in our current example, we have Maple plot some level surfaces . First, we construct the function f(**r**) as follows :

```
a := (1+ln(g1))*x - x*ln(x):
b := (1+ln(g2))*y - y*ln(y):
c := (1+ln(g3))*z - z*ln(z):
f := a+b+c;
```

$$f := (1+\ln(g1))\,x - x\ln(x) + (1+\ln(g2))\,y - y\ln(y) + (1+\ln(g3))\,z - z\ln(z)$$

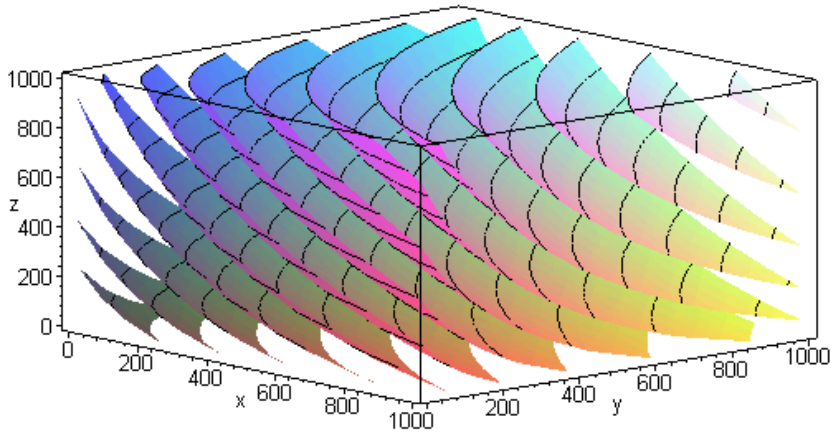
Then we set the g_i values and restate f(x,y,z),

```
g1 := 5000: g2 := 5000: g3 := 5000:
evalf(f);
```

$$9.517193191\,x - 1.\,x\ln(x) + 9.517193191\,y - 1.\,y\ln(y) + 9.517193191\,z - 1.\,z\ln(z) \tag{5.4.35}$$

Finally we display level surfaces f(x,y,z) = K for K = 1000, 1500, 2000 8000,

```
implicitplot3d({seq(f = 500+500*N,N=1..15)}, x=1..1000,y=1..1000,z=1..1000,
axes=boxed,style = patchcontour,grid = [15,15,15]);
```



(5.4.36)

Each level surface has the shape of a smooth scallop shell. The two constraint planes $x_1+x_2+x_3 = N$ and $\varepsilon_1 x_1+\varepsilon_2 x_2+\varepsilon_3 x_3 = U$ intersect in a skewed straight line in this figure (the "intersecting constraint surface"). There is only one point on this line where the constraint plane normals (1,1,1) and $(\varepsilon_1,\varepsilon_2,\varepsilon_3)$ are coplanar with the normal to one of the scallop surfaces. For $(\varepsilon_1,\varepsilon_2,\varepsilon_3) = (1,2,3)$ we found that the point is roughly $(x,y,z) = (N_1,N_2,N_3) = (616,238,116)$ and this is the point which maximizes the function

$f(\mathbf{N}) = \Sigma_{i=1}^{3} N_i\,[1 + \ln(g_i/N_i)]$ subject to the constraints $\Sigma N_i = 1000$ and $\Sigma \varepsilon_i N_i = 1000*1.5$. Eq. (5.4.27) shows that $f_{max} \approx 3500$.

5.5 Example: The Maxwell-Boltzmann Distribution

The System

In Sections 5.1-5.4 we dealt with a system having a finite number of discrete energy levels ε_i with degeneracy g_i. In the current section we shall consider a situation which has an infinite number of energy levels which form a continuum which has no upper limit. However, we shall treat this continuum as if it were a finely grained set of discrete energy levels so we can use the methods developed in Section 5.1. In the end, we shall apply quantum mechanics to show that in fact the energy levels really are discrete.

The system of interest is a small box of helium gas near room temperature. Our "particles" then are the individual gas atoms. The gas is treated as an "ideal gas" : atoms are indistinguishable (identical); there are no at-distance interactions between the atoms; the atoms are small compared to the distances between them; collisions between atoms and at the walls of the box are "elastic"; no energy is stored in any kind of rotational or vibrational modes of the atoms. The atoms have "spin" 0 so they are bosons and therefore any number of them can be placed into any given energy state, as in our previous examples.

The atoms are in a stable state of "thermal equilibrium" at temperature T. The box is sealed and insulated to prevent the transfer of particles or energy in or out. The box contains some total large fixed number of atoms M (to be computed below) and has some total energy U = Mu where u is the average energy of a particle.

Connection to Sections 5.1 and 5.2

An atom of helium gas has kinetic energy

$$\varepsilon_i = (1/2)mv^2 \equiv \varepsilon(v) . \qquad (5.5.1)$$

This is the total energy of a particle having speed $v = |\mathbf{v}|$ and mass m. There is no potential energy because the particles don't interact and because we ignore gravity which is only a miniscule effect for a small box of gas. In classical mechanics, this energy $\varepsilon(v)$ is continuous because the speed v is a continuous variable. So $\varepsilon(v) = (1/2)mv^2$ represents the continuum of "energy levels" for atoms in a box of helium.

Let dv be some small but finite speed difference. The number of atoms with speed in this range is taken to be N(v)dv, and the count of available states in the range we write as g(v)dv. The amount of energy in the atoms with speed between v and v+dv is $\varepsilon(v)N(v)dv$. We shall determine N(v) and g(v) below.

To get a number to represent the degeneracy g_i, we hop into velocity space where the axes are v_x, v_y, v_z and we note that the volume in this space corresponding to v in the range (v,v+dv) is $4\pi v^2 dv$, since this is the volume of a shell of radius v and thickness dv. As an artificial (at this point) construct, we shall

assume that velocity space is not continuous but is a 3D lattice of allowed values and that the volume of a lattice cube is V. Therefore, we write (at least for now),

$$g_i = [4\pi v^2 dv]/V = g(v)dv \qquad \text{where } g(v) = (4\pi v^2/V) . \tag{5.5.2}$$

All the g_i states in this thin shell have energy $\varepsilon_i = \varepsilon(v)$. Note that volume V is a tiny assumed cell size in velocity space, it is *not* the volume of our box of gas.

The number N_i of atoms having energy ε_i is now,

$$N_i = N(v)dv, \tag{5.5.3}$$

and then

$$g_i/N_i = [g(v)\, dv]/[N(v)dv] = g(v)/N(v) . \tag{5.5.4}$$

We have the same two constraints as in Section 5.1, but here they take the form

$$M = \Sigma_i N_i = \int_0^\infty N(v)dv$$
$$U = \Sigma_i \varepsilon_i N_i = \int_0^\infty \varepsilon(v)\, N(v)dv . \tag{5.5.5}$$

As noted, the distribution of atoms into the energy states is described by N(v). The elastic collisions between atoms and the walls do not affect N(v) but merely redirect an atom's velocity vector without changing the speed v. However, elastic collisions between atoms can and do alter velocities. Before a collision two atoms might have velocities $\mathbf{v}_1$ and $\mathbf{v}_2$, and after $\mathbf{v'}_1$ and $\mathbf{v'}_2$. But since the collisions are elastic one has $E_1 + E_2 = E'_1 + E'_2$, so the total energy U is not affected (nor is M affected). It is true that a collision rearranges the two atoms' positions in the energy level diagram since in general one won't have $E_1 = E_1'$ and $E_2 = E'_2$. But for every such rearrangement, the exact reverse rearrangement is equally likely in some other collision in the box (which contains perhaps 10^{20} atoms), so overall the inter-atomic collisions don't affect N(v).

Recall from (5.1.3) that we had this microstate count, assuming $g_i >> N_i$.

$$\Omega(\mathbf{N}) = \frac{g_1^{N_1}}{N_1!}\frac{g_2^{N_2}}{N_2!} \cdots \frac{g_m^{N_m}}{N_m!} = \Pi_{i=1}^{m} \frac{g_i^{N_i}}{N_i!} . \tag{5.1.3}$$

What happens to this expression for Ω when there are an infinite number of energy states which are spaced closely together? We might write

$$g_i^{N_i} = (g(v)dv)^{N(v)\,dv}$$

$$N_i! = [N(v)dv]! \tag{5.5.6}$$

and one gets a product of an infinite number of terms,

$$\Omega(\mathbf{N}) = \Pi_{n=0}^{\infty} \frac{(g(ndv))^{N(ndv)\,dv}}{(N(ndv)dv)!} \quad . \tag{5.5.7}$$

Just as an infinite sum $\Sigma_n a_n$ can converge only if $a_n \rightarrow 0$, an infinite product $\Pi_n a_n$ can only converge if $a_n \rightarrow 1$. There is a developed theory of such infinite products, but we leave this topic to the interested reader. Here are a few examples of infinite products,

$$\frac{c}{c-z} = \prod_{n=1}^{\infty} e^{\frac{1}{n}\left(\frac{z}{c}\right)^n}$$

$$\operatorname{sinc}(\pi z) = \prod_{n=1}^{\infty} \left(1 - \frac{z^2}{n^2}\right)$$

$$\frac{1}{\Gamma(z)} = z e^{\gamma z} \prod_{n=1}^{\infty} \left(1 + \frac{z}{n}\right) e^{-\frac{z}{n}}$$

https://en.wikipedia.org/wiki/Infinite_product (5.5.8)

and one sees that $a_n \rightarrow 1$ in all cases.

The object $f = \ln(\Omega)$ shown in (5.2.3) is a little more tractable since it is merely an infinite sum,

$$f = \Sigma_i N_i [1 + \ln(g_i/N_i)] \tag{5.2.3}$$

$$= \Sigma_{n=0}^{\infty} N(ndv)dv [\, 1 + \ln\{g(ndv)/N(ndv)\}] \quad . \tag{5.5.9}$$

We shall avoid dealing directly with Ω or $f = \ln(\Omega)$ and instead just make use of results derived in Sections 5.1 and 5.2 for a finite number of energy levels.

The Maxwell-Boltzmann speed distribution

Our main result of interest is (5.2.11) which says that the components N_i of the vector **N** which causes Ω to have a maximum are given by,

$$N_i = A\ g_i\, e^{-\beta \varepsilon_i} , \tag{5.2.11}$$

where $A \equiv e^{\lambda_1}$ and $\beta \equiv -\lambda_2$ are the derived Lagrange multipliers. Using (5.5.1,2,3) for ε_i, g_i and N_i we write the above as

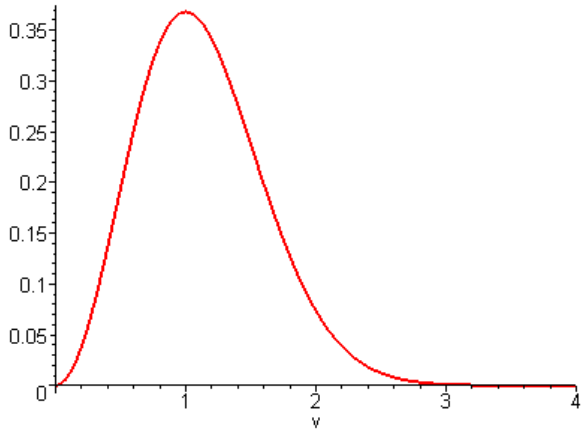
$$\{N(v)dv\} = A\ \{[4\pi v^2 dv]/V\}\ \ \{\exp(-\beta m v^2/2)\ \}$$

or

$$N(v) = 4\pi(A/V)\ v^2 \exp(-\beta m v^2/2) . \tag{5.5.10}$$

We thus arrive (with very little work) at the famous Maxwell-Boltzmann speed distribution. For now we ignore the constant $4\pi(A/V)$ -- it will be dealt with later. The general shape of this distribution is as follows :

```
plot(N,v=0..4,thickness = 2);
```



(5.5.11)

The quadratic left end of the plot arises from v^2 where exp ≈ 1, while the right end comes from the decaying exponential which here is just $\exp(-v^2)$.

Mean Values

Knowing the form (5.5.10) of N(v) allows us to compute various mean values.

The first item of interest is the **mean energy** of a particle in the gas,

$$\langle\varepsilon\rangle = u = \frac{\int_0^\infty (1/2)mv^2\, N(v)dv}{\int_0^\infty N(v)dv} = (m/2)\frac{\int_0^\infty v^4 \exp(-\beta mv^2/2)dv}{\int_0^\infty v^2 \exp(-\beta mv^2/2)dv} \quad .$$

We can make use of the following integral

$$\int_0^\infty v^n \exp(-av^2)dv = (1/2)\,\Gamma[(n+1)/2]\, a^{-(n+1)/2} \qquad \text{// Spiegel 15.77} \qquad (5.5.12)$$

with $a = \beta m/2$ to obtain

$$\langle\varepsilon\rangle = u = (m/2)\frac{\Gamma(5/2)(\beta m/2)^{-5/2}}{\Gamma(3/2)(\beta m/2)^{-3/2}} = (m/2)\frac{(3/4)\sqrt{\pi}}{(2/4)\sqrt{\pi}}(2/\beta m) = (3/2)(1/\beta) = (3/2)kT \ , \qquad (5.5.13)$$

a result we quoted earlier. This is an example of the "equipartition theorem" which says that each independent quadratic term in the energy expression ends up getting (1/2)kT of energy in equilibrium. In our case we have three terms since $\varepsilon(v) = (1/2)mv_x^2 + (1/2)mv_y^2 + (1/2)mv_z^2$.

Next we calculate the **mean speed** v of a particle,

$$v_{mean} = \langle v \rangle = \frac{\int_0^\infty v\, N(v)dv}{\int_0^\infty N(v)dv} = \frac{\int_0^\infty v^3 \exp(-\beta m v^2/2)dv}{\int_0^\infty v^2 \exp(-\beta m v^2/2)dv} = \frac{\Gamma(2)(\beta m/2)^{-2}}{\Gamma(3/2)(\beta m/2)^{-3/2}}$$

$$= \frac{1}{\sqrt{\pi}/2}(\beta m/2)^{-1/2} = (2/\sqrt{\pi})\sqrt{2/\beta m} = \sqrt{8/\beta\pi m} = \sqrt{8kT/\pi m}\ . \qquad (5.5.14)$$

The **root mean square speed** is given by

$$v_{rms} \equiv \sqrt{\langle v^2 \rangle} = \sqrt{\frac{\langle (1/2)mv^2 \rangle}{(1/2)m}} = \sqrt{\frac{2}{m}}\sqrt{\langle \varepsilon \rangle} = \sqrt{\frac{2}{m}}\sqrt{\frac{3}{2}kT} = \sqrt{\frac{3kT}{m}}\ . \qquad (5.5.15)$$

Finally, the **peak speed** (most likely speed) occurs where the distribution slope is zero :

```
N :=v^2*exp(-beta*m*v^2/2);
```

$$N := v^2\, e^{\left(-\frac{1}{2}\beta m v^2\right)}$$

```
diff(N,v);
```

$$2v\, e^{\left(-\frac{1}{2}\beta m v^2\right)} - v^3 \beta m\, e^{\left(-\frac{1}{2}\beta m v^2\right)}$$

```
solve(diff(N,v)=0,v);
```

$$0,\ \frac{\sqrt{2}\sqrt{\beta m}}{\beta m},\ -\frac{\sqrt{2}\sqrt{\beta m}}{\beta m} \qquad (5.5.16)$$

and the meaningful root is the second item, so

$$v_{peak} = \sqrt{\frac{2kT}{m}}\ . \qquad (5.5.17)$$

We now summarize our conclusions about mean values,

$$\langle \varepsilon \rangle = u = (3/2)kT$$

$$v_{peak} = \sqrt{\frac{2kT}{m}} = 1.41\sqrt{\frac{kT}{m}} = 1.00\, v_{peak}$$

$$v_{mean} = \sqrt{\frac{8kT}{\pi m}} = 1.60\sqrt{\frac{kT}{m}} = 1.13\, v_{peak}$$

$$v_{rms} = \sqrt{\frac{3kT}{m}} = 1.73\sqrt{\frac{kT}{m}} = 1.23\ v_{peak} \qquad (5.5.18)$$

The three speed values are not very far apart.

Determination of the derived Lagrange Multiplier A

Recall from (5.5.10) that.

$$N(v) = 4\pi(A/V)\ v^2 \exp(-\beta m v^2/2)\ . \qquad (5.5.10)$$

The total number constraint requires that

$$M = \int_0^\infty N(v)dv\ = 4\pi(A/V)\ \int_0^\infty v^2 \exp(-\beta m v^2/2) = 4\pi(A/V)\ (1/2)(\sqrt{\pi}/2)(\beta m/2)^{-3/2}$$

$$= 4\pi(A/V)(\sqrt{\pi}/4)\ (\beta m/2)^{-3/2}\ . \qquad (5.5.19)$$

Therefore

$$4\pi(A/V) = M\ (4/\sqrt{\pi})\ (\beta m/2)^{3/2}$$

and then

$$N(v) = (4M/\sqrt{\pi})\ (\beta m/2)^{3/2}\ v^2 \exp(-\beta m v^2/2) \qquad (5.5.20)$$

which agrees with Zemansky p. 159 Eq. (6-25). Notice that the lattice cell size V in velocity space does not appear in this result, so it applies whether V is finite or V→0 which is the classical limit.

In our study of the discrete-state Boltzmann problem in Section 5.2 it was noted that **N** is the "solution vector" which has "components" $N_i = Ag_i e^{-\beta\varepsilon_i}$. In the continuum problem the "solution vector" is the entire continuous function N(v) while the "components" are N(v)dv for specific values of v.

The other constraint $U = \int_0^\infty \varepsilon(v)\ N(v)dv$ tells us what we already know from (5.5.13),

$$U = \int_0^\infty \varepsilon(v)\ N(v)dv\ = <\varepsilon> \int_0^\infty N(v)dv\ = <\varepsilon>M\ = uM\ = (3/2)\ kT\ M\ . \qquad (5.5.21)$$

We would like now to verify that $N_i << g_i$ since (5.2.11) that $N_i = A\ \ g_i\ e^{-\beta\varepsilon_i}$ depends on this assumption. From (5.5.4) we must therefore show that N(v) << g(v). But so far we have $g(v) = (4\pi v^2/V)$ from (5.5.2) where V was our assumed lattice cube size in v-space. Taking V→0 results in g(v) = ∞ and then certainly N(v) << g(v). We now visit the quantum mechanics department to obtain the correct value for V.

Quantum Mechanics and Determination of V

To determine the quantum theory value of V in (5.5.2), we consider the problem of a single theoretical point particle in a cubic box of edge L. The Schrodinger equation says $\hat{H}\psi(\mathbf{r}) = \varepsilon\psi(\mathbf{r})$ for a single particle

wavefunction ψ, where the Hamiltonian operator $\hat{H}$ is just $\hat{p}^2/2m$. Here $\hat{\mathbf{p}} = (-\hbar/i)\boldsymbol{\nabla}$ (a key assumption of quantum theory), so $\hat{p}^2 = -\hbar^2\nabla^2$. The constant $\hbar$ is $h/(2\pi)$ where h is Planck's constant,

$h = 6.62607 \times 10^{-34}$ J s $\quad \hbar = 1.0545718 \times 10^{-34}$ J s $\quad \hbar = h/(2\pi)$ // Joule-sec .

The Schrodinger equation then reads $-(1/2m)\,\hbar^2\nabla^2\psi = \varepsilon\psi$. We seek a solution to this equation which vanishes at the 6 walls of a cubic box of edge L. This is so because ψ must be continuous and $\psi = 0$ everywhere outside the box since there is no probability $|\psi|^2$ that the particle is outside the box. That solution is

$$\psi_{n_x n_y n_z}(x,y,z) = K \sin(\pi n_x x/L)\sin(\pi n_y y/L)\sin(\pi n_z z/L) \qquad n_i = 1,2,3... \quad . \tag{5.5.22}$$

The constant K is determined by requiring that the probability of the particle being in the box is 1,

$$\int_0^L dx \int_0^L dy \int_0^L dz\, |\psi_{n_x n_y n_z}(x,y,z)|^2 = 1 \; . \tag{5.5.23}$$

Since $\int_0^L dx \sin^2(\pi n_x x/L) = (L/2)$ the above says $K^2(L/2)^3 = 1$ so

$$K = (2/L)^{3/2}. \tag{5.5.24}$$

In (5.5.22) setting $n_x = 0$ results in $\psi = 0$ which cannot be normalized to 1 and is a non-solution. A solution with $n_x = -2$ is minus of the solution with $n_x = +2$ and so these two solutions are really the same solution. Solutions must be linearly independent to be separately counted. That is why we have written $n_i = 1,2,3...$.

The Schrodinger equation then says

$$\frac{\hat{p}^2}{2m}\psi = \varepsilon\psi \quad \Rightarrow \quad -\frac{\hbar^2}{2m}\nabla^2\psi = \varepsilon\psi \quad \Rightarrow \quad -\hbar^2\nabla^2\psi = 2m\varepsilon\psi \quad \Rightarrow$$

$$-\hbar^2\left(\left[-(\pi n_x/L)^2 - (\pi n_y/L)^2 - (\pi n_z/L)^2 \right] \right)\psi = 2m\varepsilon\psi \qquad \text{// using (5.5.22) for } \psi$$

or

$$(\pi\hbar/L)^2 \left[n_x^2 + n_y^2 + n_z^2 \right] = 2m\varepsilon$$

or

$$\varepsilon = (2m)^{-1} (\pi\hbar/L)^2 \left[n_x^2 + n_y^2 + n_z^2 \right] \qquad n_i = 1,2,3... \quad . \tag{5.5.25}$$

We arrive at the interesting conclusion that the energy spectrum is not continuous but is discrete. It is quantized. Setting $\varepsilon = (1/2)mv^2$ we find that

$$v^2 = (\pi\hbar/mL)^2 \left[n_x^2 + n_y^2 + n_z^2 \right] \tag{5.5.26}$$

so the speed variable is also discrete, not continuous.

If we now think of velocity space with $\mathbf{v} = (v_x, v_y, v_z)$ then the above says

$$v_x^2 = (\pi\hbar/mL)^2 n_x^2 \qquad v_y^2 = (\pi\hbar/mL)^2 n_y^2 \qquad v_z^2 = (\pi\hbar/mL)^2 n_z^2 \tag{5.5.27}$$

so then

$$\mathbf{v} = (v_x, v_y, v_z) = (\pi\hbar/mL)(n_x, n_y, n_z) \ . \tag{5.5.28}$$

In the space (n_x, n_y, n_z) the lattice cube size is 1. Therefore, in v-space this cube size is $(\pi\hbar/mL)^3$. We have now found a value for the parameter V in (5.5.2) :

$$g_i = [4\pi v^2 dv]/V = g(v)dv \qquad \text{where } g(v) = (4\pi v^2/V) \ . \tag{5.5.2}$$

$$V = (\pi\hbar/mL)^3 \ . \tag{5.5.29}$$

But now we make a correction. Since $\mathbf{v} = (\pi\hbar/mL)(n_x, n_y, n_z)$ and since $n_i = 1,2,3...$, when we count states we should only include the first octant shell of v-space where v_x, v_y, v_z are all positive. Thus, we correct (5.5.2) as follows:

$$g_i = [(1/8)4\pi v^2 dv]/V = g(v)dv \qquad \text{where } g(v) = (1/8)(4\pi v^2/V) \ . \tag{5.5.2}_{corr}$$

$$V = (\pi\hbar/mL)^3 \ . \tag{5.5.30}$$

The degeneracy function is then

$$g(v) = \pi v^2/2V = (1/2)\pi v^2 (\pi\hbar/mL)^{-3} = (1/2)\pi v^2 (h/2mL)^{-3} = (1/2)\pi v^2 (2mL/h)^3$$

$$= 4\pi(mL/h)^3 v^2 \ . \tag{5.5.31}$$

As a check on this result, integrating it from 0 to v' gives

$$(4\pi/3)(mL/h)^3 (v'^2)^{3/2} = (4\pi/3)(mL/h)^3 (2\varepsilon'/m)^{3/2} = (4\pi/3)(L/h)^3 (2\varepsilon' m)^{3/2}$$

and this agrees with Zemansky p. 272 Problem (10-2) (answers on p 645).

At this point we have determined that,

$$N(v) = (4M/\sqrt{\pi})\,(\beta m/2)^{3/2}\; v^2 \exp(-\beta m v^2/2),\;\; \beta = 1/(kT) \qquad (5.5.20)$$

$$g(v) = 4\pi(mL/h)^3 v^2 \qquad (5.5.31) \qquad (5.5.32)$$

where $m = m_{He}$ is the mass of a helium atom.

Numeric Values for N(v) and g(v) and g(v)/N(v)

We now seek numeric values for N(v) and g(v) for a 1 cm^3 box of helium at room temperature. Before entering constants, we create expressions for N, g, g/N and v_m (v_{mean} from (5.5.14)) as follows, where the Maple v -> construct is used to define a function of v :

```
N := v-> M*(4/sqrt(Pi))*(beta*mHe/2)^(3/2)*v^2*exp(-beta*mHe*v^2/2);
```

$$N := v \to \frac{M\sqrt{2}\,(\beta\, mHe)^{\left(\frac{3}{2}\right)}\, v^2\, e^{\left(-\frac{1}{2}\beta\, mHe\, v^2\right)}}{\sqrt{\pi}}$$

```
g := v-> 4*Pi*(mHe*L/h)^3*v^2;
```

$$g := v \to 4\,\frac{\pi\, mHe^3\, L^3\, v^2}{h^3}$$

```
g_over_N := unapply(g(v)/N(v),v);    # note that the v^2's cancel
```

$$g\_over\_N := v \to 2\,\frac{\pi^{\left(\frac{3}{2}\right)}\, mHe^3\, L^3\sqrt{2}}{h^3\, M\,(\beta\, mHe)^{\left(\frac{3}{2}\right)}\, e^{\left(-\frac{1}{2}\beta\, mHe\, v^2\right)}}$$

```
vm := sqrt(8*k*T/(Pi*mHe));
```

$$vm := 2\sqrt{2}\,\sqrt{\frac{k\,T}{\pi\, mHe}} \qquad (5.5.33)$$

Detail: The strange "unapply" command causes g_over_N(v) to be a function of v, namely g(v)/N(v). If we don't do it this way, g_over_N(0) reports a divide by 0 error at v = 0 since N(0) = 0.

Notice that the g/N ratio takes its *smallest* value at v = 0 where $\exp(\beta m_{He} v^2/2) = 1$. Since we want to show that g/N >> 1, showing this at v = 0 guarantees that g/N >> 1 for all v.

Our first task is to compute the number M of helium atoms in the box. This can be found from the ideal gas law PV = NkT which applied to our situation says $PL^3 = MkT$. We display units as if they were Maple variables, which allows a "dimension check" for all our results.

```
L := 1e-2*m;                    # box edge length 1 cm
```

$$L := .01\ m$$

```
P := 1.01325e5*Pa;              # atmospheric pressure in Pascals
```

$$P := 101325.\ Pa$$

```
T :=  294 * K;                  # room temperature 70F = 21C
```

$$T := 294\ K$$

```
k := 1.38064852e-23*J/K;        # Boltzmann constant Joules/Kelvin
```

$$k := .138064852\ 10^{-22}\ \frac{J}{K}$$

```
J := kg*m^2/s^2;                # Joule converted to basic units
```

$$J := \frac{kg\ m^2}{s^2}$$

```
Pa := kg/(m*s^2);               # Pascal converted to basic units
```

$$Pa := \frac{kg}{m\ s^2}$$

```
M := P*L^3/(k*T);               # He atoms in the box; gas law says PV = MkT
```

$$M := .2496238921\ 10^{20} \qquad (5.5.34)$$

So there are M ≈ 2.5 x 10^{19} helium atoms in our 1 cm^3 box.

Next compute the derived Lagrange multiplier value β = $-\lambda_2$ = 1/(kT), and take a look at the basic energy scale for a helium atom kT:

```
unassign('J');
beta := 1/(k*T);                # derived Lagrange multiplier
```

$$\beta := .2463596270\ 10^{21}\ \frac{1}{J}$$

```
k*T;                            # (2/3) the average energy of an atom
```

$$.4059106649\ 10^{-20}\ J$$

```
subs(J = 1e7*erg,k*T);          # the above expressed in ergs
```

$$.4059106649\ 10^{-13}\ erg$$

```
subs(J = 6.242e+18*eV, k*T);    # the above expressed in electron-volts
```

$$.02533694370\ eV \qquad (5.5.35)$$

Since it takes 19.8 eV to excite helium out of its electronic ground state, and since (3/2)kT = .038 eV, the elastic collision aspect of our ideal gas assumption is well justified at room temperature.

Next, set the mass m_{He} of a helium atom,

```
mP := 1.6726219e-27*kg;         # mass of a proton
```

$$mP := .16726219\ 10^{-26}\ kg$$

```
mHe := (4*mP);                  # Helium mass = roughly 4 proton masses
```

$$mHe := .66904876\ 10^{-26}\ kg \qquad (5.5.36)$$

Recall that $g_i = g(v)dv$ and $N_i = N(v)dv$ so we expect g and N to have units of inverse velocity. Here then are the numeric values for v_m, $N(v_m)$ and $g(v_m)$:

```
vm: simplify(%);
```

$$1242.957854\ \frac{m}{s}$$

```
N(vm): simplify(%);
```

$$.1822716279\ 10^{17}\ \frac{s}{m}$$

```
g(vm);
```

$$.1998609382\ 10^{23}\ \frac{s}{m}$$

```
g(vm)/N(vm): simplify(%);
```

$$.1096500538\ 10^{7} \qquad (5.5.37)$$

The mean helium atoms are moving along at a respectable clip, 1243 m/sec (2,780 mph). Of course they don't go far between collisions.

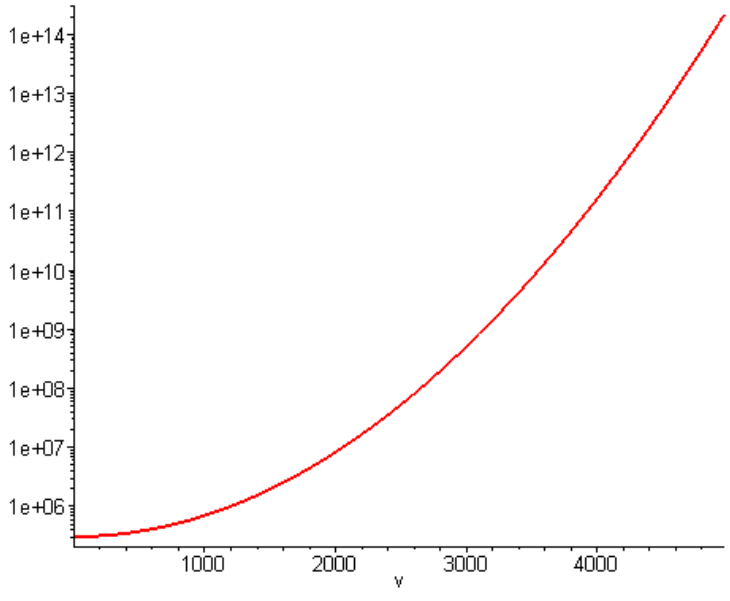
The ratio $g(v_m)/N(v_m)$ is thus about 10^6, which justifies the assumption that $g(v) >> N(v)$ and $g_i >> N_i$, at least at the mean velocity v_m. As noted above, the lower bound for this ratio occurs at v = 0 where we have,

```
g_over_N(0):simplify(%);
```

$$306936.0792 \qquad (5.5.38)$$

and even here we have g(v) >> N(v). Here is plot of g(v)/N(v) from v = 0 up to 4 times the mean velocity,

```
logplot(eval(g_over_N(v),[s=1,m=1,kg=1]),v = 1..4*eval(vm,[s=1,m=1]),thickness=2);
```

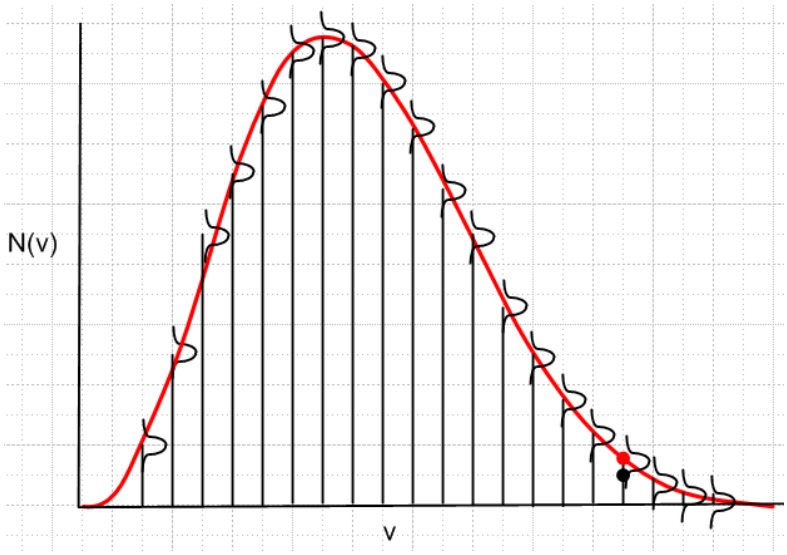


(5.5.39)

One can view the first octant of the velocity sphere as consisting of a large number of tiny cubes (size V) which are the available quantum states. Only a tiny fraction of these cubes are occupied by particles. For example, in the shell of the sphere at speed v = 2200 m/sec, the above graph shows that only about 1 in 10^7 cubes is occupied, all the rest are empty. The occupancy here is much sparser than in our 3-level example of Section 5.4 where the ε_2 state had $N_2 = 268$ out of $g_2 = 5000$ states occupied.

The ensemble

The smooth N(v) distribution shows the ensemble average for $N_i = N(v)dv$ at various speeds. If one were to examine a particular system in an ensemble of one billion 1 cm^3 boxes of helium, and if one were to divide the distribution into a set of vertical strips each having a large number of atoms, one would measure a set of N_i values that do not quite match the distribution curve. For each such strip there is a "bell curve" of the type shown in Fig (5.3.18) which has a fractional half-width on the order of $1/\sqrt{N_s}$ where N_s is the (large) number of atoms in a strip. We illustrate this idea in the following drawing,



(5.5.40)

which is analogous to Fig (5.4.29) for the 3-level system of Section 5.4. One could display on this picture a set of red dots right on the curve, and black dots within the bell curve at each v (we drew one pair of such dots only), as discussed relative to (5.4.29). In the current situation, if one were to dramatically increase the number of strips to drive N_s down to a much lower number, the bell curve widths would increase dramatically, and a pattern of black dots if connected by lines would have a very noise-like appearance, hardly resembling the red curve.

Summary

Applying the discrete-energy-level result $N_i = Ag_i e^{-\beta\varepsilon_i}$ developed in (5.2.11) (using the method of Lagrange multipliers) we have obtained a set of results concerning the nature of a small box of helium atoms at room temperature. The distribution of particle speeds has the Maxwell-Boltzmann shape shown in Fig (5.5.11) and the corresponding expression for the speed distribution N(v) is given in (5.5.32). The

distribution of energy level degeneracies is given by g(v) also in (5.5.32). We used quantum theory only to determine the tiny lattice size V in v-space which causes the speed and energy of a helium atom to be quantized. For this specific example we showed that $g_i >> N_i$ thus justifying the use of (5.2.11). We showed along the way that the average helium atom energy is (3/2)kT and hinted how one might derive the equipartition theorem of which this is an example. Expressions were found for the mean, rms and peak speeds of the N(v) distribution.

Section 6. Matrix proof of the Method of Lagrange Multipliers

In Section 2 it was shown that the Method of Lagrange Multipliers (Theorem 2) follows directly from Theorem 1, and is really just a restatement of Theorem 1 which we quote:

Theorem 1: A point **r** is a "stationary point for f(**r**) subject to constraints $a_i(\mathbf{r}) = 0$" $\Leftrightarrow$ *(1.24)*

(a) $a_i(\mathbf{r}) = 0$ for i = 1,2...C (point **r** must satisfy all the constraints)
(b) There exist C constants λ_i such that $\nabla f(\mathbf{r}) + \Sigma_{i=1}^{C} \lambda_i \nabla a_i(\mathbf{r}) = 0$.

Our proof of Theorem 1 was entirely geometric in nature. We talked about surfaces, normals to surfaces, tangent spaces, perp spaces, the intersection constraint surface, and so on. In this section, we shall rederive Theorem 1 using a certain "R matrix", though some geometry still appears. The reader should regard this simply as an alternative derivation of Theorem 1.

As in Section 1, a certain amount of background material is necessary to provide a framework for the proof. We define S = C+1 so the number of constraint functions is C = S-1, and we rename the constraint functions to be a,b,c....q instead of a_1, a_2, a_3....a_C. The purpose of this renaming is to avoid having double subscripts on a constraint function, one identifying it and one indicating a partial derivative.

6.1. The R matrix and its Rank

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

We seek stationary points 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 the stationary points of f(**r**) subject to constraints $\underset{1}{a(\mathbf{r})=0}, \underset{2}{b(\mathbf{r})=0}, \underset{3}{c(\mathbf{r})=0}, \ldots \underset{S-1}{q(\mathbf{r})=0}$. (6.1.1)

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 :

$$R = \begin{pmatrix} f_1 & f_2 & f_3 & \ldots & f_N \\ a_1 & a_2 & a_3 & \ldots & a_N \\ b_1 & b_2 & b_3 & \ldots & b_N \\ \ldots & & & & \\ q_1 & q_2 & q_3 & \ldots & q_N \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{a} \\ \mathbf{b} \\ \ldots \\ \mathbf{q} \end{pmatrix} = \begin{pmatrix} \nabla f \\ \nabla a \\ \nabla b \\ \ldots \\ \nabla q \end{pmatrix} \qquad (6.1.2)$$

For example, $f_2 \equiv \partial f(x_1, x_2, x_3 \ldots x_N)/\partial x_2$.
On the right we abbreviate each row of functions in **bold**, indicating a row vector like $\mathbf{f} = \nabla f$.

The R matrix has S rows and N columns. We shall only be interested in $N \geq S$ which is the same as saying $N \geq C+1$ or $N > C$. Recall from the discussion below (2.8) that if $N \leq C$, the problem is overconstrained and one cannot seek stationary points where df = 0. If N = S, the R matrix is square, otherwise for N > S it has more columns than rows, so it 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). S would then be the "full rank" of R.

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,

$$R = \begin{matrix} f_1 & f_2 & f_3 & f_4 & f_5 & f_6 \\ a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \\ b_1 & b_2 & b_3 & b_4 & b_5 & b_6 \\ c_1 & c_2 & c_3 & c_4 & c_5 & c_6 \end{matrix} \qquad (6.1.3)$$

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:

$$\text{red} = (f_1,f_2,f_3,f_4) \qquad \text{blue} = (f_2,f_3,f_4,f_5) \qquad \text{green} = (f_1,f_2,f_4,f_6)\ . \qquad (6.1.4)$$

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 in (6.1.2) 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.

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, $r \leq \min(m,n)$. (6.1.5)

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

$$k_1 \mathbf{r}_1 + k_2 \mathbf{r}_2 + ... + k_N \mathbf{r}_N = 0 \quad . \tag{6.1.6}$$

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 (6.1.6) 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 (6.1.5) has these properties for any m x n matrix A: (6.1.7)

(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]

In Appendix A below we provide our own proof of the claims of (6.1.7).

With the above as background, the following theorem is fundamental to our matrix proof of Theorem 1 :

Theorem 3. Point **r** is a "stationary point of f(**r**) subject to the constraints a=0,b=0...q=0" ⇔
rank[R(**r**)] < S . (6.1.8)

But, according to (6.1.7b), rank[R(**r**)] < S ⇔ the rows of R shown in (6.1.2) are linearly dependent since there are S rows in R. Therefore rank[R(**r**)] < S ⇔ $\nabla f(\mathbf{r}) + \lambda_1 \nabla a(\mathbf{r}) + \lambda_2 \nabla b(\mathbf{r}) + ... + \lambda_{S-1} \nabla q(\mathbf{r}) = 0$ for some λ_i. Thus Theorem 3 ⇔ Theorem 1 of (1.24). But we know from Section 2 that Theorem 1 ⇔ Theorem 2, so proving Theorem 3 provides an alternate proof of the Method of Lagrange Multipliers.

According to the rank definition (6.1.5),

rank(R) < S ⇔ all SxS submatrices must have zero determinant (6.1.9)

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

The point of Theorem 3 is that the solution point **r** of the constrained stationary point problem must be a point where the R(**r**) matrix drops below full rank. We shall prove Theorem 3 in the ⇒ direction in Section 6.2 by showing that, if **r** is a stationary point, 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 stationary point problem, one can restrict one's search to points **r** for which rank[R(**r**)] < S.

The proof of Theorem 3 in the ⇐ direction is fairly simple. If rank(R) < S, then the rows of R are linearly dependent by (6.1.7) and then $\nabla f(\mathbf{r}) + \lambda_1 \nabla a(\mathbf{r}) + \lambda_2 \nabla b(\mathbf{r}) + ... + \lambda_{S-1} \nabla q(\mathbf{r}) = 0$. This implies that

$$df(\mathbf{r}) = \nabla f(\mathbf{r}) \bullet d\mathbf{r} \; = \; -\lambda_1[\nabla a(\mathbf{r}) \bullet d\mathbf{r}] \; - \lambda_2 \, [\nabla b(\mathbf{r}) \bullet d\mathbf{r}] - ..... - \lambda_{S-1} \, [\nabla q(\mathbf{r}) \bullet d\mathbf{r}] \tag{6.1.10}$$

for any d**r**. Restricting to d**r** which don't violate any of the constraints, we know that each term on the right vanishes. For example, $\nabla a(\mathbf{r}) \bullet d\mathbf{r} = 0$ because $\nabla a(\mathbf{r})$ is normal to the surface $a(\mathbf{r}) = 0$, see (1.2). Since all the terms on the right above then vanish, we get $df(\mathbf{r}) = 0$ and **r** is a stationary point of f by our definition (1.18) of a constrained stationary point.

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

$$
\begin{array}{llll}
f = u_1(x_1, x_2, ...x_N) & R_{11} = \partial u_1/\partial x_1 = \partial f/\partial x_1 = f_1 & R_{12} = f_2 & \text{etc.} \\
a = u_2(x_1, x_2, ...x_N) & R_{21} = \partial u_2/\partial x_1 = \partial a/\partial x_1 = a_1 & R_{22} = a_2 & \text{etc} \\
b = u_3(x_1, x_2, ...x_N) & & & \\
... & & & \\
q = u_S(x_1, x_2, ...x_N) & & & \qquad (6.1.11)
\end{array}
$$

then the matrix shown in (6.1.2) 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.

6.2. Proof of Theorem 3 (⇒)

Theorem 3. Point **r** is a "stationary point of f(**r**) subject to the constraints a=0,b=0...q=0" ⇔
rank[R(**r**)] < S. *(6.1.8)*

(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 3 for the specific case shown above in Fig (6.1.3) 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,

$$
R = \begin{array}{cccccc} f_1 & f_2 & f_3 & f_4 & f_5 & f_6 \\ a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \\ b_1 & b_2 & b_3 & b_4 & b_5 & b_6 \\ c_1 & c_2 & c_3 & c_4 & c_5 & c_6 \end{array} = \begin{array}{c} \mathbf{f} \\ \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \end{array} \qquad (6.2.1)
$$

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

$$\begin{aligned} a(x_1, x_2, x_3, x_4, x_5, x_6) &= 0 \\ b(x_1, x_2, x_3, x_4, x_5, x_6) &= 0 \\ c(x_1, x_2, x_3, x_4, x_5, x_6) &= 0 \ . \end{aligned} \tag{6.2.2}$$

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

$$\begin{aligned} a(x_1, x_2, x_3, x_4, x_5, x_6) &= 0 \qquad & x_1 &= X^1(x_4, x_5, x_6) \\ b(x_1, x_2, x_3, x_4, x_5, x_6) &= 0 \qquad \Rightarrow & x_2 &= X^2(x_4, x_5, x_6) \\ c(x_1, x_2, x_3, x_4, x_5, x_6) &= 0 \qquad & x_3 &= X^3(x_4, x_5, x_6) \end{aligned} \tag{6.2.3}$$

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. See Appendix C for an elaboration of this point.

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

$$\begin{aligned} f(X^1(x_4, x_5, x_6), X^2(x_4, x_5, x_6), X^3(x_4, x_5, x_6), x_4, x_5, x_6) &\equiv F(x_4, x_5, x_6) \\ a(X^1(x_4, x_5, x_6), X^2(x_4, x_5, x_6), X^3(x_4, x_5, x_6), x_4, x_5, x_6) &\equiv A(x_4, x_5, x_6) \\ b(X^1(x_4, x_5, x_6), X^2(x_4, x_5, x_6), X^3(x_4, x_5, x_6), x_4, x_5, x_6) &\equiv B(x_4, x_5, x_6) \\ c(X^1(x_4, x_5, x_6), X^2(x_4, x_5, x_6), X^3(x_4, x_5, x_6), x_4, x_5, x_6) &\equiv C(x_4, x_5, x_6) \ . \end{aligned} \tag{6.2.4}$$

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

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

$$\frac{\partial F}{\partial x_4} = \frac{\partial f}{\partial x_1}\frac{\partial X^1}{\partial x_4} + \frac{\partial f}{\partial x_2}\frac{\partial X^2}{\partial x_4} + \frac{\partial f}{\partial x_3}\frac{\partial X^3}{\partial x_4} + \frac{\partial f}{\partial x_4} \ . \tag{6.2.5a}$$

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

$$F_4 = f_1 X^1{}_4 + f_2 X^2{}_4 + f_3 X^3{}_4 + f_4 \ . \tag{6.2.5b}$$

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

$$\begin{aligned} F_4 &= f_1 X^1{}_4 + f_2 X^2{}_4 + f_3 X^3{}_4 + f_4 \\ F_5 &= f_1 X^1{}_5 + f_2 X^2{}_5 + f_3 X^3{}_5 + f_5 \\ F_6 &= f_1 X^1{}_6 + f_2 X^2{}_6 + f_3 X^3{}_6 + f_6 \ . \end{aligned} \tag{6.2.6}$$

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

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

$$\begin{aligned} f_1X^1{}_4 + f_2X^2{}_4 + f_3X^3{}_4 + f_4 &= 0 \\ f_1X^1{}_5 + f_2X^2{}_5 + f_3X^3{}_5 + f_5 &= 0 \\ f_1X^1{}_6 + f_2X^2{}_6 + f_3X^3{}_6 + f_6 &= 0 \,. \end{aligned} \qquad (6.2.7)$$

Now apply the *same process* to the three constraint functions A,B,C. First compute the derivatives, and then set the derivatives to zero. One does this for $A(x_4, x_5, x_6)$, for example, because the constraint condition a = 0 says that that $A(x_4, x_5, x_6) = 0$ = a 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,

$$\begin{aligned} a_1X^1{}_4 + a_2X^2{}_4 + a_3X^3{}_4 + a_4 &= 0 \\ a_1X^1{}_5 + a_2X^2{}_5 + a_3X^3{}_5 + a_5 &= 0 \\ a_1X^1{}_6 + a_2X^2{}_6 + a_3X^3{}_6 + a_6 &= 0 \,. \end{aligned} \qquad (6.2.8)$$

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:

$$\begin{aligned} f_1X^1{}_4 + f_2X^2{}_4 + f_3X^3{}_4 + f_4 &= 0 \\ f_1X^1{}_5 + f_2X^2{}_5 + f_3X^3{}_5 + f_5 &= 0 \\ f_1X^1{}_6 + f_2X^2{}_6 + f_3X^3{}_6 + f_6 &= 0 \end{aligned}$$

$$\begin{aligned} a_1X^1{}_4 + a_2X^2{}_4 + a_3X^3{}_4 + a_4 &= 0 \\ a_1X^1{}_5 + a_2X^2{}_5 + a_3X^3{}_5 + a_5 &= 0 \\ a_1X^1{}_6 + a_2X^2{}_6 + a_3X^3{}_6 + a_6 &= 0 \end{aligned}$$

$$\begin{aligned} b_1X^1{}_4 + b_2X^2{}_4 + b_3X^3{}_4 + b_4 &= 0 \\ b_1X^1{}_5 + b_2X^2{}_5 + b_3X^3{}_5 + b_5 &= 0 \\ b_1X^1{}_6 + b_2X^2{}_6 + b_3X^3{}_6 + b_6 &= 0 \end{aligned}$$

$$\begin{aligned} c_1X^1{}_4 + c_2X^2{}_4 + c_3X^3{}_4 + c_4 &= 0 \\ c_1X^1{}_5 + c_2X^2{}_5 + c_3X^3{}_5 + c_5 &= 0 \\ c_1X^1{}_6 + c_2X^2{}_6 + c_3X^3{}_6 + c_6 &= 0 \,. \end{aligned} \qquad (6.2.9)$$

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,

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

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

$$
\begin{aligned}
&f_1 X^1{}_6 + f_2 X^2{}_6 + f_3 X^3{}_6 + f_6 = 0\\
&a_1 X^1{}_6 + a_2 X^2{}_6 + a_3 X^3{}_6 + a_6 = 0\\
&b_1 X^1{}_6 + b_2 X^2{}_6 + b_3 X^3{}_6 + b_6 = 0\\
&c_1 X^1{}_6 + c_2 X^2{}_6 + c_3 X^3{}_6 + c_6 = 0 \quad .
\end{aligned}
\tag{6.2.10}
$$

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,

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

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

$$
\begin{pmatrix} 0\\0\\0\\0 \end{pmatrix} = \begin{pmatrix} f_1 & f_2 & f_3 & f_6\\ a_1 & a_2 & a_3 & a_6\\ b_1 & b_2 & b_3 & b_6\\ c_1 & c_2 & c_3 & c_6 \end{pmatrix} \begin{pmatrix} X^1{}_6\\ X^2{}_6\\ X^3{}_6\\ 1 \end{pmatrix} = (f_1, f_2, f_3, f_6) \begin{pmatrix} X^1{}_6\\ X^2{}_6\\ X^3{}_6\\ 1 \end{pmatrix} \quad . \tag{6.2.11}
$$

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

$$
\begin{pmatrix} X^1{}_4\\ X^2{}_4\\ X^3{}_4\\ 1 \end{pmatrix} = (f_1, f_2, f_3, f_4)^{-1} \begin{pmatrix} 0\\0\\0\\0 \end{pmatrix} = \begin{pmatrix} 0\\0\\0\\0 \end{pmatrix} \quad . \tag{6.2.12}
$$

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

$$\det(f_1, f_2, f_3, f_m) = 0 \qquad m \neq 1,2,3 \ . \tag{6.2.13}$$

Therefore, we have shown that three of the SxS = 4x4 submatrices of Fig (1.3) vanish. But there are (6,4) = 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

$$\det(f_1, f_2, f_4, f_m) = 0 \qquad m \neq 1,2,4 \ . \tag{6.2.14}$$

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

$$\det(f_i, f_j, f_k, f_m) = 0 \qquad m \neq i,j,k \ . \tag{6.2.15}$$

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 stationary point of f($\mathbf{r}$) subject to the constraints. And then the conclusion is that, for such a solution point $\mathbf{r}$, one has rank[R($\mathbf{r}$)] < S.

To summarize, one theoretically eliminates S-1 of the N variables to create the functions F,A,B,C of (6.2.4). 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.

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 $\partial f/\partial x_2$.

$$R = \begin{matrix} f_1 & f_2 & f_3 & ... & f_N \\ a_1 & a_2 & a_3 & ... & a_N \\ b_1 & b_2 & b_3 & ... & b_N \\ .... & & & & \\ q_1 & q_2 & q_3 & ... & q_N \end{matrix} = \begin{matrix} \mathbf{f} \\ \mathbf{a} \\ \mathbf{b} \\ ... \\ \mathbf{q} \end{matrix} \qquad (6.2.1)'$$

Here are the S-1 constraint equations,

$$\begin{aligned} &a(x_1, x_2, ...x_N) = 0 \qquad \text{// S-1 constraint equations} \\ &b(x_1, x_2, ...x_N) = 0 \\ &... \\ &q(x_1, x_2, ...x_N) = 0 \ . \end{aligned} \qquad (6.2.2)'$$

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

$$\begin{aligned} &a(x_1, x_2, ...x_N) = 0 && x_1 = X^1(x_S, x_{S+1} ... x_N) \\ &b(x_1, x_2, ...x_N) = 0 \quad \Rightarrow && x_2 = X^2(x_S, x_{S+1} ... x_N) \\ &... && ... \\ &q(x_1, x_2, ...x_N) = 0 && x_{S-1} = X^{S-1}(x_S, x_{S+1} ... x_N) \ . \end{aligned} \qquad (6.2.3)'$$

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

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

(6.2.4)'

Now take first derivatives as shown in (6.2.5) to get this generalized version of (6.2.6),

$$\begin{aligned} F_S &= f_1X^1{}_S + f_2X^2{}_S + f_3X^3{}_S + ... + f_{S-1}X^{S-1}{}_S + f_S \\ F_{S+1} &= f_1X^1{}_{S+1} + f_2X^2{}_{S+1} + f_3X^3{}_{S+1} + ... + f_{S-1}X^{S-1}{}_{S+1} + f_{S+1} \\ F_{S+2} &= f_1X^1{}_{S+2} + f_2X^2{}_{S+2} + f_3X^3{}_{S+2} + ... + f_{S-1}X^{S-1}{}_{S+2} + f_{S+2} \\ &... \\ F_N &= f_1X^1{}_N + f_2X^2{}_N + f_3X^3{}_N + ... + f_{S-1}X^{S-1}{}_N + f_N \ . \qquad \text{// N-S+1 equations} \end{aligned} \qquad (6.2.6)'$$

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

$$
\begin{aligned}
&f_1X^1{}_S + f_2X^2{}_S + f_3X^3{}_S + \ldots + f_{S-1}X^{S-1}{}_S + f_S = 0\\
&f_1X^1{}_{S+1} + f_2X^2{}_{S+1} + f_3X^3{}_{S+1} + \ldots + f_{S-1}X^{S-1}{}_{S+1} + f_{S+1} = 0\\
&f_1X^1{}_{S+2} + f_2X^2{}_{S+2} + f_3X^3{}_{S+2} + \ldots + f_{S-1}X^{S-1}{}_{S+2} + f_{S+2} = 0\\
&\ldots\\
&f_1X^1{}_N + f_2X^2{}_N + f_3X^3{}_N + \ldots + f_{S-1}X^{S-1}{}_N + f_N = 0\ .\quad \text{// N-S+1 equations} \qquad (6.2.7)'
\end{aligned}
$$

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 (6.2.9). 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,

$$
\begin{aligned}
&f_1X^1{}_S + f_2X^2{}_S + f_3X^3{}_S + \ldots + f_{S-1}X^{S-1}{}_S + f_S = 0\\
&a_1X^1{}_S + a_2X^2{}_S + a_3X^3{}_S + \ldots + a_{S-1}X^{S-1}{}_S + a_S = 0\\
&b_1X^1{}_S + b_2X^2{}_S + b_3X^3{}_S + \ldots + b_{S-1}X^{S-1}{}_S + b_S = 0\\
&\ldots.\\
&q_1X^1{}_S + q_2X^2{}_S + q_3X^3{}_S + \ldots + q_{S-1}X^{S-1}{}_S + q_S = 0\ .\quad \text{// S equations} \qquad (6.2.10)'_S
\end{aligned}
$$

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

$$
\begin{aligned}
&f_1X^1{}_N + f_2X^2{}_N + f_3X^3{}_N + \ldots + f_{S-1}X^{S-1}{}_N + f_N = 0\\
&a_1X^1{}_N + a_2X^2{}_N + a_3X^3{}_N + \ldots + a_{S-1}X^{S-1}{}_N + a_N = 0\\
&b_1X^1{}_N + b_2X^2{}_N + b_3X^3{}_N + \ldots + b_{S-1}X^{S-1}{}_N + b_N = 0\\
&\ldots.\\
&q_1X^1{}_N + q_2X^2{}_N + q_3X^3{}_N + \ldots + q_{S-1}X^{S-1}{}_N + q_N = 0\ .\quad \text{// S equations} \qquad (6.2.10)'_N
\end{aligned}
$$

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. There are then N-S+1 matrix equations,

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

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

●●●●

$$
\begin{pmatrix} 0\\0\\\ldots\\0\\0 \end{pmatrix} = (f_1, f_2, f_3, \ldots f_{S-1}, f_N) \begin{pmatrix} X^1{}_N\\X^2{}_N\\\ldots\\X^{S-1}{}_N\\1 \end{pmatrix} \ . \qquad (6.2.11)'
$$

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

$$\det(f_1, f_2, f_3, \ldots f_{S-1}, f_m) = 0 \qquad m \neq 1,2,3\ldots S\text{-}1 \quad . \qquad (6.2.13)'$$

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

$$\det(f_i, f_j, f_k, \ldots f_r, f_m) = 0 \qquad m \neq i,j,k\ldots r \quad . \qquad (6.2.15)'$$

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 stationary point of f(**r**) subject to the constraints.

Appendix A: Matrix rank equals the number of independent columns and rows

This Appendix proves our claim of (6.1.7) 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.2.8) :

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

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

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):

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

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

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

Theorem 8 (Cramer's Rule):

$$\mathbf{y} = M\mathbf{x} \quad \Rightarrow \quad x_{\mathbf{s}} = \frac{\det(M[\mathbf{c_s} \rightarrow \mathbf{y}])}{\det(M)} \qquad \textit{(B.4.14)}$$

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 (6.1.6) 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 (6.1.6) 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 $\mathbf{c_j} = \Sigma_{i\neq j} k_i \mathbf{c_i}$. 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

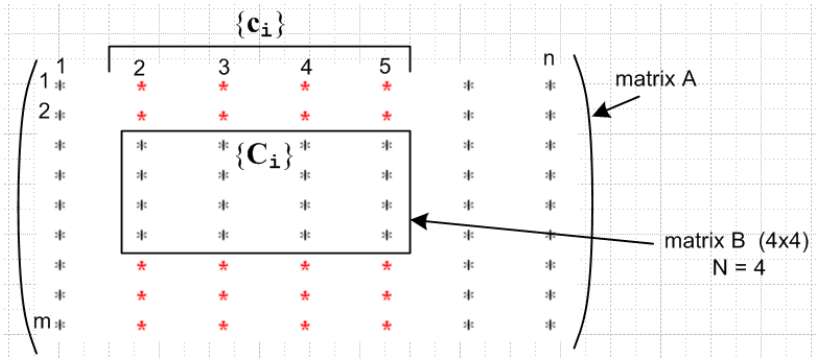
If the column vectors $\mathbf{c_i}$ of matrix M are linearly independent, then (6.1.6) says $\Sigma x_i \mathbf{c_i} = 0 \Rightarrow \mathbf{x} = 0$. Therefore we know that $\Sigma x_i M_{ji} = 0 \Rightarrow x_j = 0$ which is the same as $M\mathbf{x} = 0 \Rightarrow \mathbf{x} = 0$. Thinking of $M{:}E^n \rightarrow E^n$, 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 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)

Below we shall prove (b) which then implies (a). A picture is worth a thousand words. Here matrix B happens to be contiguous in both directions :



(A.3)

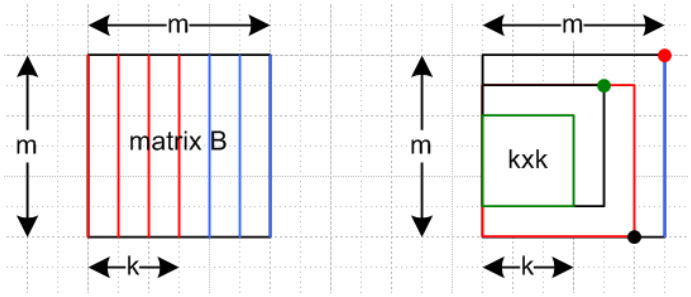
Proof of (b): If the set $\{\mathbf{c_i}\}$ is linearly dependent in A, then from (6.1.6) one can write $\Sigma_{\mathbf{i\in S}}\, k_i\mathbf{c_i} = 0$ with $\mathbf{k} \neq 0$. This is a set of m scalar equations (one for each row in A) which contains as a subset the set of N equations $\Sigma_{\mathbf{i\in S}}\, k_i\mathbf{C_i} = 0$ (one for each row in B). 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 $A \rightarrow A^{\mathbf{T}}$ 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). *(6.1.5)* (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) by (A.6).

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 $\mathbf{c_2} = \mathbf{\Sigma_{i=1}}^{\mathbf{k}}\ \mathbf{k_i c_i} = 1\ \mathbf{c_2}$. 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

An alterative title for this Appendix might be "More than you really want to know about determinants". We find it convenient to gather these Facts all in one place. Many of these Facts are used in Appendix A where we prove other Facts about matrix rank which in turn are used in Section 6 to support our alternate derivation of the Method of Lagrange Multipliers. It is hoped that the encapsulated presentation below might be useful to the reader for other applications of determinants. Of special interest to us is that permutations form a group which implies certain "rearrangement theorems" which in turn provide quick proofs of many well-known and some lesser-known Facts about determinants.

B.1 Definition of the determinant

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

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

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

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

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: $\text{Parity}(C\mathbf{a}) = \text{Parity}(CA\mathbf{z_0}) = (-1)^{S_\mathbf{c}+S_\mathbf{a}} = \text{Parity}(\mathbf{c})\,\text{Parity}(\mathbf{a}) = \text{Parity}(C^{-1}\mathbf{a})$ (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 means "product")

$$\Pi(\mathbf{a},\mathbf{b};\,M) \equiv M_{\mathbf{ab}} \equiv M_{\mathbf{a}_1\mathbf{b}_1}M_{\mathbf{a}_2\mathbf{b}_2} \ldots\ldots M_{\mathbf{a}_n\mathbf{b}_n} = \text{product of n factors} \quad . \tag{B.1.5}$$

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 $M_{\mathbf{a}_1\mathbf{b}_1}M_{\mathbf{a}_2\mathbf{b}_2} \ldots\ldots M_{\mathbf{a}_n\mathbf{b}_n}$ merely reorders the terms in the product but the product stays the same.

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

Proof: $\Pi(\mathbf{a},\mathbf{b};\,M^{\mathbf{T}}) = M^{\mathbf{T}}_{\mathbf{a}_1\mathbf{b}_1}M^{\mathbf{T}}_{\mathbf{a}_2\mathbf{b}_2} \ldots\ldots M^{\mathbf{T}}_{\mathbf{a}_n\mathbf{b}_n} = M_{\mathbf{b}_1\mathbf{a}_1}M_{\mathbf{b}_2\mathbf{a}_2} \ldots\ldots M_{\mathbf{b}_n\mathbf{a}_n} = \Pi(\mathbf{b},\mathbf{a};\,M)$.

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),

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

$$= \frac{1}{n!}\,\Sigma_{\mathbf{a}}\,\Sigma_{\mathbf{b}}\,(-1)^{S_{\mathbf{a}}+S_{\mathbf{b}}}\,M_{\mathbf{a}_1\mathbf{b}_1}M_{\mathbf{a}_2\mathbf{b}_2} \ldots\ldots M_{\mathbf{a}_n\mathbf{b}_n}\,. \tag{B.1.8}$$

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

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

We are now ready for our first determinant theorem:

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

Proof:

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

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

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

B.2 The permutation group, the permutation tensor ε, and expansions for det(M)

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. Here G is the set of 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 above 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 the 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, ε = 0 $\varepsilon_{..i..i..} = -\,\varepsilon_{..i..i..} = 0$ (B.2.8)

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

$= \Sigma_{\mathbf{a}}\,[\,\Sigma_{\mathbf{b}}\ \text{Parity}(\mathbf{a})\ \text{Parity}(\mathbf{b})\ \Pi(\mathbf{z}_0,\mathbf{a};X)\ \ \Pi(\mathbf{z}_0,\mathbf{b};Y)\,]$ // (B.1.4) used next line

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

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

$= \det(XY)$. // (B.2.5)

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

$\Sigma_{\mathbf{a}}\ \Pi(\mathbf{z}_0,\mathbf{a};X)\ \Pi(\mathbf{a},\mathbf{b};Y) = \Sigma_{a_1a_2\cdots a_n}\ X_{1a_1}X_{2a_2}\ \ldots\ldots\ X_{na_n}\ \ Y_{a_1b_1}Y_{a_2b_2}\ \ldots\ldots\ Y_{a_nb_n}$

$= \Sigma_{a_1a_2\cdots a_n}\,(X_{1a_1}Y_{a_1b_1})(X_{2a_21}Y_{a_2b_2})\ldots\ldots(X_{na_n}Y_{a_nb_n})$

$= (XY)_{1b_1}\,(XY)_{2b_2}\ldots(XY)_{nb_n} = \Pi(\mathbf{z}_0,\mathbf{b};XY)$

which in dense notation one would write as

$\Sigma_{\mathbf{a}}\,X_{z_0a}Y_{ab} = (XY)_{z_0b}$.

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) .

It is not hard to slightly generalize the results of Theorem 2 to obtain :

Theorem 2A: If **b** is a permutation of $\mathbf{z}_0$, then det(M) can be represented in these two ways:

$$\det(M) = \varepsilon_{b_1 b_2 \cdots b_n} \Sigma_{a_1 a_2 \cdots a_n} \varepsilon_{a_1 a_2 \cdots a_n} M_{b_1 a_1} M_{b_2 a_2} \cdots\cdots M_{b_n a_n} \quad \text{(B.2.17a)}$$

$$\det(M) = \varepsilon_{b_1 b_2 \cdots b_n} \Sigma_{a_1 a_2 \cdots a_n} \varepsilon_{a_1 a_2 \cdots a_n} M_{a_1 b_1} M_{a_2 b_2} \cdots\cdots M_{a_n b_n} \ . \quad \text{(B.2.17b)}$$

In dense notation, one could write the above as (**a** and **b** are now vectors),

$$\det(M) = \varepsilon_b \Sigma_a \varepsilon_a M_{ba} \quad \text{(B.2.18a)}$$

$$\det(M) = \varepsilon_b \Sigma_a \varepsilon_a M_{ab} \ . \quad \text{(B.2.18b)}$$

Since $\varepsilon_b^2 = 1$, one can move ε_b to the left side of these equations, so in the dense notation,

$$\varepsilon_b \det(M) = \Sigma_a \varepsilon_a M_{ba} \quad \text{(B.2.19a)}$$

$$\varepsilon_b \det(M) = \Sigma_a \varepsilon_a M_{ab} \ . \quad \text{(B.2.19b)}$$

Equations (B.2.19) are valid even if **b** is not a permutation of $\mathbf{z}_0$. In this case one has 0 = 0.
For the special case $\mathbf{b} = \mathbf{z}_0$ Theorem 2A reduces to Theorem 2 since $\varepsilon_{z_0} = \varepsilon_{12\ldots n} = 1$.

Proof of Theorem 2A: We shall prove (B.2.17a) and then (B.2.17b) follows by taking $M \to M^T$ and using $\det(M^T) = \det(M)$.

$\det(M) = \Sigma_a$ Parity(**a**) $\Pi(\mathbf{z}_0, \mathbf{a}; M)$ // line 6 of (B.2.5)

$= \Sigma_a$ Parity(**a**) $\Pi(B\mathbf{z}_0, B\mathbf{a}; M)$ // (B.1.6) with C = B

$= \Sigma_a$ Parity(B**a**) Parity(**b**) $\Pi(B\mathbf{z}_0, B\mathbf{a}; M)$ // Parity(B**a**) Parity(**b**) = Parity(**a**)

$= \Sigma_a$ Parity(**a**) Parity(**b**) $\Pi(B\mathbf{z}_0, \mathbf{a}; M)$ // (B.2.4) that $\Sigma_a f(B\mathbf{a}) = \Sigma_a f(\mathbf{a})$

$=$ Parity(**b**) Σ_a Parity(**a**) $\Pi(\mathbf{b}, \mathbf{a}; M)$ // $\mathbf{b} = B\mathbf{z}_0$

$= \varepsilon_{b_1 b_2 \cdots b_n} \Sigma_{a_1 a_2 \cdots a_n} \varepsilon_{a_1 a_2 \cdots a_n} M_{b_1 a_1} M_{b_2 a_2} \cdots M_{b_n a_n}$ // (B.1.5) and (B.2.9)

$= \varepsilon_b \Sigma_a \varepsilon_a M_{ba}$. // previous line in dense notation (B.2.20)

Here it is assumed that B and hence **b** is associated with a permutation of $\mathbf{z}_0$, In this case, $(\varepsilon_b)^2 = 1$ so one can move ε_b to the left side to get

$$\varepsilon_b \det(M) = \Sigma_a \varepsilon_a M_{ba} \ . \quad \text{(B.2.21)}$$

In this form, the equation is valid whether or not **b** is a permutation of $\mathbf{z_0}$. If in $\mathbf{b} = (b_1,b_2...b_n)$ the b_i are arbitrary elements of the set {1,2,...n} where two or more b_i are the same (that is, **b** is not a permutation of $\mathbf{z_0}$), then $\Sigma_a\ \varepsilon_{a_1a_2\cdots a_n}\ M_{b_1a_1}M_{b_2a_2}\ ....\ M_{b_na_n} = 0$ by symmetry. For example, if $b_1 = b_2$ then

$$\begin{aligned} S &\equiv \Sigma_a\ \varepsilon_{a_1a_2\cdots a_n}\ M_{b_1a_1}M_{b_1a_2}\ ....\ M_{b_na_n} \\ &= \Sigma_a\ \varepsilon_{a_2a_1\cdots a_n}\ M_{b_1a_2}M_{b_1a_1}\ ....\ M_{b_na_n} && \text{// rename dummy indices } a_1 \leftrightarrow a_2 \\ &= \Sigma_a\ [-\varepsilon_{a_1a_2\cdots a_n}]\ M_{b_1a_1}M_{b_1a_2}\ ....\ M_{b_na} && \text{// (B.2.8) for } \varepsilon \text{ and slide } M_{b_1a_2} \text{ to the right} \\ &= -\Sigma_a\varepsilon_{a_1a_2\cdots a_n}M_{b_1a_1}M_{b_1a_2}\ ....\ M_{b_na} = -S \qquad \Rightarrow\ S = 0\ . && \text{(B.2.22)} \end{aligned}$$

So in this case the right side of (B.2.21) is zero. But from (B.2.8) $\varepsilon_b = 0$ so the left side is also zero.

Reader Exercise: Prove that

$$\varepsilon_{1i_2\cdots i_n}\det(M) = \Sigma_A\ \text{parity(A)}\ M_{1,A(i_1)}M_{2,A(i_2)}\ .....M_{n,A(i_n)}\ . \qquad \text{(B.2.23)}$$

Proof in dense notation:

$$\begin{aligned} \varepsilon_i\det(M) &= \varepsilon_i\Sigma_a\ \varepsilon_a\ M_{z_0,a} && \text{// this is } \varepsilon_i \text{ times usual expansion (B.2.7a)} \\ &= \text{parity(I)}\ \Sigma_A\ \text{parity(A)}\ M_{z_0,Az_0} && \text{// } \mathbf{a} = A\mathbf{z_0} \\ &= \text{parity(I)}\ \Sigma_A\ \text{parity(AI)}\ M_{z_0,AIz_0} && \text{// rearrangement theorem } \Sigma_A f(A) = \Sigma_A f(AI) \\ &= \Sigma_A\ \text{parity(A)}\ M_{z_0,Ai} && \text{// } \mathbf{i} = I\mathbf{z_0} \qquad \text{QED} \end{aligned}$$

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 of M obtained by crossing out the r^{th} row and the s^{th} 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 element 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):

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

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

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

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

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

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

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

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.

Appendix C: The Intersection of Constraint Surfaces

In the proof of Theorem 3 in Section 6.2 (a) we rely on our ability, at least in theory, to eliminate x_1, x_2, x_3 from the three constraint equations shown in (6.2.3).

$$\begin{aligned} a(x_1, x_2, x_3, x_4, x_5, x_6) &= 0 \\ b(x_1, x_2, x_3, x_4, x_5, x_6) &= 0 \\ c(x_1, x_2, x_3, x_4, x_5, x_6) &= 0 \end{aligned} \quad \Rightarrow \quad \begin{aligned} x_1 &= X^1(x_4, x_5, x_6) \\ x_2 &= X^2(x_4, x_5, x_6) \\ x_3 &= X^3(x_4, x_5, x_6) \ . \end{aligned} \qquad (6.2.3)$$

How does one know this is possible? The equations might be very complicated, some of the coordinates might not appear in some of the equations, and so on.

We look first at three examples in E^3 then in Example 4 we return to our question.

Example 1: Consider these two constraint equations in E^3, [this is the same example as appears in (1.9)]

$$\begin{aligned} a(x_1,x_2,x_3) &= x_1^2 + x_2^2 + x_3^2 - 2^2 = 0 \\ b(x_1,x_2,x_3) &= (x_1-2)^2 + x_2^2 + x_3^2 - 2^2 = 0 \ . \end{aligned} \qquad (C.1)$$

Each constraint represents a sphere of radius 2 and so is a 2-dimensional smooth surface in E^3. The first sphere is centered at the origin, while the second has its center at (2,0,0). Subtracting the first equation from the second gives $-4x_1+4 = 0$ so $x_1 = 1$, then inserting this into both equations gives

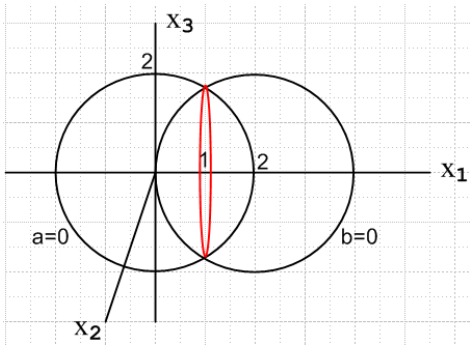
$$\begin{aligned} 1 + x_2^2 + x_3^2 - 2^2 &= 0 \\ 1 + x_2^2 + x_3^2 - 2^2 &= 0 \end{aligned} \quad \Rightarrow \ x_2^2 + x_3^2 = 3 \ \text{ and } x_1 = 1 \ . \qquad (C.2)$$

The intersection of the two constraint surfaces is a circle, a 1-dimensional smooth surface in E^3.

Given $a(x_1,x_2,x_3) = 0$ and $b(x_1,x_2,x_3) = 0$, is it possible to eliminate x_1 and x_2 from the two constraint equations? The answer is yes, as follows,

$$\begin{aligned} x_1 &= X^1(x_3) = 1 \\ x_2 &= X^2(x_3) = \pm\sqrt{3 - x_3^2} \ . \end{aligned} \qquad (C.3)$$

The following schematic drawing shows the intersection surface in red,



(C.4)

Notice in this example that there are two possible solutions for x_2. One can regard the intersection surface (the circle $x_2^2 + x_3^2 = 3$) as having two pieces (front half, back half) and the ± sign selects one of these pieces.

It is easy to imagine more complicated examples for two constraints in E^3 where perhaps one or both of the constraint surfaces contain disjoint pieces (e.g. a hyperboloid), and where the intersection surface also has several disjoint pieces (ellipsoid intersecting a hyperboloid) .

Example 2: Consider these two constraint equations in E^3 :

$$\begin{aligned} a(x_1,x_2,x_3) &= x_1^2 + x_2^2 + x_3^2 - 2^2 = 0 \\ b(x_1,x_2,x_3) &= \quad x_2^2 + x_3^2 - 1^2 = 0 \,. \end{aligned} \tag{C.5}$$

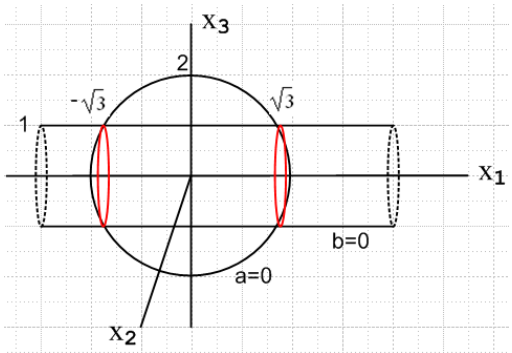
The second equation is missing the x_1 coordinate. The first equation describes the same origin-centered radius 2 sphere of the previous example. The second equation is that of a circle of radius 1, but as a function of three variables it is in fact a cylinder whose axis is the x_1 axis. If x_2 and x_3 lie on the circle, any value of x_1 satisfies the second equation. Missing coordinates result in surfaces which are "extruded" in the dimensions of the missing coordinates. Subtracting the second equation from the first gives $x_1^2 - 3 = 0$ so $x_1 = \pm\sqrt{3}$, then inserting this into both equations gives

$$\begin{aligned} 3 \; + x_2^2 + x_3^2 - 2^2 &= 0 \\ x_2^2 + x_3^2 - 1^2 &= 0 \qquad \Rightarrow \; x_2^2 + x_3^2 = 1 \;\text{ and }\; x_1 = \pm\sqrt{3} \,. \end{aligned} \tag{C.6}$$

Given $a(x_1,x_2,x_3) = 0$ and $b(x_1,x_2,x_3) = 0$, is it possible to eliminate x_1 and x_2 from the two constraint equations? The answer is yes, as follows (the signs are independent),

$$\begin{aligned} x_1 &= X^1(x_3) = \pm\sqrt{3} \\ x_2 &= X^2(x_3) = \pm\sqrt{1 - x_3^2} \;. \end{aligned} \tag{C.7}$$

The following schematic drawing shows the intersection surface in red,



(C.8)

In this example there are two possible solutions for x_1 and two for x_2 so overall there are four solutions. The intersection surface consists of the two circles each having a front half and a back half.

Example 3: Consider these two constraint equations in E^3 :

$$\begin{aligned} a(x_1,x_2,x_3) &= x_1^2 + x_2^2 + x_3^2 - 2^2 = 0 \\ b(x_1,x_2,x_3) &= x_1^2 + x_2^2 - 1^2 = 0 . \end{aligned} \qquad \text{(C.9)}$$

Now x_3 is missing from the second equation instead of x_1. Subtracting the second equation from the first gives $x_3^2 - 3 = 0$ so $x_3 = \pm\sqrt{3}$, then inserting this into both equations gives

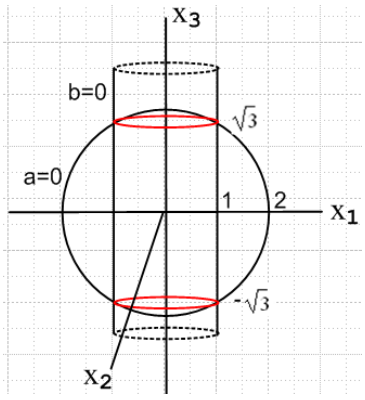
$$\begin{aligned} & x_1^2 + x_2^2 + 3 - 2^2 = 0 \\ & x_1^2 + x_2^2 - 1^2 = 0 \quad \Rightarrow \; x_1^2 + x_2^2 = 1 \;\text{ and }\; x_3 = \pm\sqrt{3} . \end{aligned} \qquad \text{(C.10)}$$

Given $a(x_1,x_2,x_3) = 0$ and $b(x_1,x_2,x_3) = 0$, is it possible to eliminate x_1 and x_2 from the two constraint equations? The answer is yes, as follows, where α is an arbitrary real parameter in [-1,1],

$$\begin{aligned} & x_1 = X^1(x_3 = \pm\sqrt{3}) = \alpha \qquad -1 \le \alpha \le 1 \\ & x_2 = X^2(x_3 = \pm\sqrt{3}) = \pm\sqrt{1 - \alpha^2} . \quad \text{// the two } \pm \text{ signs are independent} \end{aligned} \qquad \text{(C.11)}$$

In the previous two examples the argument x_3 was a free parameter whose variation mapped out the smooth constraint intersection surface piece(s). In Example 3 x_3 is fixed at $\pm\sqrt{3}$ and a new parameter α must be introduced to map out the intersection surfaces. There are again four solution half-circles.

The following schematic drawing shows the intersection surface in red,



(C.12)

Example 4: Now consider these three constraint equations in E^6 :

$$\begin{aligned} & a(x_1, x_2, x_3, x_4, x_5, x_6) = 0 \\ & b(x_1, x_2, x_3, x_4, x_5, x_6) = 0 \\ & c(x_1, x_2, x_3, x_4, x_5, x_6) = 0 . \end{aligned} \qquad \textit{(6.2.2)} \qquad \text{(C.13)}$$

We assume that each equation describes a smooth surface of dimension 5 in E^6 (each is a hypersurface) and each surface may have several pieces. In a problem with constraints, one *assumes* that there is some

non-null surface of intersection of all the constraint surfaces. This intersection surface may have multiple pieces and each piece (barring degenerate cases) is itself a smooth surface of dimension 3 in E^6. A candidate solution point **r** must lie on one of these intersection surface pieces. In general, each constraint knocks down the dimension of the intersection surface by one degree of freedom.

So let us assume that x_4, x_5, x_6 are the last three components of some point(s) on the overall intersection surface. Each such point of course has some x_1, x_2, x_3 components. If x_4, x_5, x_6 are varied slightly, x_1, x_2, x_3 will also vary slightly, and this is what we mean by writing the functions

$$\begin{aligned} x_1 &= X^1(x_4, x_5, x_6) \\ x_2 &= X^2(x_4, x_5, x_6) \\ x_3 &= X^3(x_4, x_5, x_6)\ . \end{aligned} \qquad \textit{(6.2.3)} \qquad \text{(C.14)}$$

If the intersection surface has multiple pieces which have the same x_4, x_5, x_6 value, then any of the three functions X^n above may be multi-valued, as occurred in our earlier examples. Conversely, if we select a triplet x_4, x_5, x_6 for which there are no points on the intersection surface, the solution x_1, x_2, x_3 does not exist. We don't care about such points in E^6.

So this then is what we mean in Section 6.2 (a) when we say above (6.2.3) that one can use the three constraint equations to eliminate the three variables x_1, x_2, x_3. In the concluding equations (6.2.11) one can regard the functions X^1, X^2, X^3 as being any of the multi-valued functions just discussed if the intersection surface has multiple pieces. For example, we had in (6.2.11) that

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} f_1 & f_2 & f_3 & f_4 \\ a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \end{pmatrix} \begin{pmatrix} X^1{}_4 \\ X^2{}_4 \\ X^3{}_4 \\ 1 \end{pmatrix} \qquad \textit{part of (6.2.11)} \qquad \text{(C.15)}$$

and here $X^1{}_4$ refers to $(\partial/\partial x_4)X^1(x_4, x_5, x_6)$ evaluated at coordinates x_4, x_5, x_6 for some candidate solution point **r** on the intersection surface. The fact that the matrix in (C.15) must have zero determinant is not affected by the possible existence of multiple solution X^1, X^2, X^3 functions.

Notice that the derivatives appearing in the matrix are also unaffected by the possibility of multi-valued functions for X^1, X^2, X^3.

In the same manner as above, we can "eliminate" any triplet of variables x_i, x_j, x_k ($i \neq j \neq k$) from the three constraint equations (C.13).

In the general case where there are S-1 constraints and **r** has N components $x_1, x_2 \ldots x_N$ with N > S, the dimensionality of the intersection surface is N - (S-1) in E^N. The conclusions of the Theorem 3 proof outlined in Section 6.2 (b) are similarly not affected by the possible existence of an intersection surface having multiple pieces and the functions X^n possibly having multiple values.

Footnote: Why is $a(x_1, x_2, x_3, ...x_N) = 0$ an (N-1)-dimensional surface in E^N ? (C.16)

This fact probably seems obvious to the reader and is certainly obvious in the first three examples above. Here we attempt a simple "engineering" explanation.

We *assume* that a constraint equation is a "smooth" equation, meaning it is continuous and differentiable in all its arguments except perhaps at certain isolated locations which can be handled on an *ad hoc* basis either by fiat or by taking limits. Consider that, since $a(\mathbf{r})$ = constant (namely 0), da = 0 so

$$0 = da = \Sigma_{i=1}^{N} a_i dx_i \ , \qquad // \ a_i = a_i(\mathbf{r}) = a_i(x_1, x_2, x_3, ...x_N) \tag{C.17}$$

where a_i is usual means $\partial a/\partial x_i$. Let $\mathbf{r}$ be some point for which $a(\mathbf{r}) = 0$. At this point we shall assume that there exists at least one $a_s(\mathbf{r})$ which is non-zero. If *all* $a_i(\mathbf{r}) = 0$, then from (1.2) that $\mathbf{r}$ is a point on a surface which has a null normal vector, which is not possible, so such an $\mathbf{r}$ value could not lie on $a(\mathbf{r}) = 0$. Then (C.17) may be written

$$dx_s = (1/a_s) \Sigma_{i \neq s} a_i dx_i \ . \tag{C.18}$$

Create a coordinate system whose origin is located at this point $\mathbf{r}$ for which $a(\mathbf{r}) = 0$, and whose axes are aligned with those of E^N. Now imagine an arbitrary tiny displacement of all the dx_i other than dx_s. For this set of N-1 dx_i there is only one possible dx_s which causes the point $\mathbf{r} + d\mathbf{r}$ to satisfy the equation $a(\mathbf{r}+d\mathbf{r}) = a(\mathbf{r}) + da = 0 + 0 = 0$, where $d\mathbf{r} \equiv (dx_1, dx_2.....dx_N)$. That one possible dx_s value is that given by (C.18). Imagine repeating this process for a continuum of values for the dx_i other than dx_s, and in each case we obtain the unique solution dx_s from (C.18). We can think of the x_s axis as being "vertical" and all the other axes being "horizontal" inasmuch as they are all perpendicular to the x_s axis. The vectors $d\mathbf{r}$ generated in this manner comprise a tiny patch of a "plane" of dimension N-1 (a hyperplane) in E^N which contains the point $\mathbf{r}$. This is so because $\Sigma_{i=1}^{N} a_i dx_i = \mathbf{a} \bullet d\mathbf{r} = 0$ is the equation of a plane in E^N passing through our origin at $\mathbf{r}$ with $\mathbf{a}$ being that plane's normal vector. More generally, $\hat{\mathbf{n}} \bullet \mathbf{r} = d$ is the equation of a hyperplane in E^n having a unit normal $\hat{\mathbf{n}}$ and whose closest approach to the origin is distance d.

Thus at a point $\mathbf{r}$ satisfying $a(\mathbf{r}) = 0$ we have constructed a tiny neighborhood of nearby points $\mathbf{r}$ which also satisfy $a(\mathbf{r}) = 0$. This neighborhood is a patch of a hyperplane in E^N which is certainly a piece of "surface" in E^N having dimension N-1. By repeating this process using a mesh of points $\mathbf{r}_i$ satisfying $a(\mathbf{r}_i) = 0$, we then map out a triangulated polyhedral surface of dimension N-1 in E^N. In the limit the mesh size goes to 0, we arrive at a smooth surface of dimension N-1 in E^N and that surface is $a(\mathbf{r}) = 0$. The tiny planar patch at any point on the surface is part of the "tangent plane" to the surface at that point.

The constraint $a(\mathbf{r}) = 0$ is assumed to be "locally smooth" in the immediate region of any point $\mathbf{r}$ on the operational constraint surface, so a small local planar region (an open set) can be constructed around any such point. This is the basic idea of a surface being a manifold M, and the set of vectors in the tangent plane of a point $\mathbf{r}$ on M comprise the "tangent space" of M at $\mathbf{r}$, usually denoted by $T_r M$.

(C.19)

Appendix D: Lagrange Multipliers in Classical Mechanics

Here is an outline of this Appendix showing the equation number nearest the start of each subsection:

Constraints and virtual displacements (D.1)
Assumption that constraints do no virtual work (D.11)
Comment on internal forces (D.13)
First Application of Lagrange Multipliers: the forces of constraint (D.13)
Summary of the above Lagrange Multiplier Application (D.25)
Generalized coordinates, generalized forces, and Lagrange's Equations (D.27)
Footnote: Derivation of the result (D.31) used above (D.40)
Second Application of Lagrange Multipliers: non-holonomic constraints (D.49)
The Case of C holonomic constraints treated as if they were non-holonomic (D.55)
Footnote: Carry out the $\delta S = 0$ functional variation of the action (D.56)

Our main focus in this document is the use of Lagrange Multipliers to find candidates $\mathbf{r}$ for which a scalar *function* $f(\mathbf{r})$ is stationary, $df(\mathbf{r}) = 0$, subject to constraints. In this Appendix, however, we consider applications of Lagrange Multipliers which relate to classical mechanics for a system of N particles in the presence of constraints. In this case the function $f(\mathbf{r})$ with $\mathbf{r}$ lying in $E^{\mathbf{N}}$ is replaced by a *functional* $f(\boldsymbol{\varphi})$ where the $\boldsymbol{\varphi}$ lie in a space of functions. The goal is to find the functions $\boldsymbol{\varphi}$ which create a stationary point of the functional, and one writes this as $\delta f(\boldsymbol{\varphi}) = 0$.

A certain amount of background information is required to put these applications into context, which we now present.

Constraints and virtual displacements

Let $\mathbf{r_k}(t)$ be the position of the $k^{\mathbf{th}}$ particle of an N particle system in 3 dimensional space. At any fixed time t, if the particles are unencumbered with constraints, the variables $(\mathbf{r_1}, \mathbf{r_2}..... \mathbf{r_N})$ can be regarded as a single point in the space $E^{\mathbf{3N}}$ and the "configuration space" of the system indicated by $(\mathbf{r_1}, \mathbf{r_2}..... \mathbf{r_N})$ is all of this space $E^{\mathbf{3N}}$. All points in $E^{\mathbf{3N}}$ are "legal" points for the system.

A **holonomic** constraint has the form

$$a(\mathbf{r_1}, \mathbf{r_2}..... \mathbf{r_N}, t) = 0. \qquad (D.1)$$

At a given time t, this equation represents a surface of dimension 3N-1 in $E^{\mathbf{3N}}$. If there are s such holonomic constraints

$$a_{\mathbf{i}}(\mathbf{r_1}, \mathbf{r_2}..... \mathbf{r_N}, t) = 0 \qquad i = 1,2...s \qquad (D.2)$$

then at time t the system vector $(\mathbf{r_1}, \mathbf{r_2}..... \mathbf{r_N})$ is constrained to lie on a surface of dimension 3N-s in $E^{\mathbf{3N}}$, as discussed in Appendix C, so each extra constraint lowers the operating surface dimension by 1. One can think of the term holonomic as meaning that the constraints are functions of "whole" coordinates like $\mathbf{r_2}$ as opposed to differential coordinates like $d\mathbf{r_2}$.

The constraints which appear in (1.7) are holonomic constraints of the form $a(x_1, x_2, ... x_n) = 0$.

A **non-holonomic** constraint is one that is not holonomic. An inequality constraint like $a(\mathbf{r}_1, \mathbf{r}_2 ..... \mathbf{r}_N, t) < 0$ which might bound the particles to one side of a surface is an example of a nonholonomic constraint, but our interest lies more in nonholonomic constraints of the form

$$\Sigma_{k=1}^{N} \mathbf{b}_k \bullet d\mathbf{r}_k + b_t dt = 0 \qquad (D.3)$$

in which differentials rather than "whole" coordinates appear. The "coefficients" $\mathbf{b}_k$ and b_t can in general be functions of $(\mathbf{r}_1, \mathbf{r}_2 ..... \mathbf{r}_k, t)$. If it happens that such a constraint can be integrated by some hook or crook to give an equation of the form $a(\mathbf{r}_1, \mathbf{r}_2 ..... \mathbf{r}_k, t) = 0$, then it is regarded as being a holonomic constraint, so a non-holonomic constraint must be non-integrable. If there are m such non-holonomic constraints imposed on a system, one could write them as (second line is a velocity constraint),

$$\begin{aligned} &\Sigma_{k=1}^{N} \mathbf{A}_{ik} \bullet d\mathbf{r}_k + A_{it} dt = 0 \qquad i = 1,2...m \\ \text{or} \qquad &\\ &\Sigma_{k=1}^{N} \mathbf{A}_{ik} \bullet \dot{\mathbf{r}}_k + A_{it} = 0 \qquad i = 1,2...m \ . \end{aligned} \qquad (D.4)$$

Here bolded $\mathbf{A}_{ik}$ represents 3 matrices each of which is m x N (rows x columns). Since these constraints are non-integrable, one cannot really described them in terms of surfaces in E^{3N}.

If one differentiates the holonomic constraints (D.2) one gets

$$\begin{aligned} &\Sigma_{k=1}^{N} [\nabla^{(k)} a_i] \bullet d\mathbf{r}_k + [\partial_t a_i] \, dt = 0 \qquad i = 1,2...s \\ \text{or} \qquad &\\ &\Sigma_{k=1}^{N} [\nabla^{(k)} a_i] \bullet \dot{\mathbf{r}}_k + [\partial_t a_i] = 0 \qquad i = 1,2...s \end{aligned} \qquad (D.5)$$

where $\nabla^{(k)} \equiv (\partial/\partial(\mathbf{r}_k)_1, \partial/\partial(\mathbf{r}_k)_2, \partial/\partial(\mathbf{r}_k)_3)$ is a gradient with respect to the components of $\mathbf{r}_k$. These equations have the same form as those in (D.4), but these are holonomic because they are integrable to give $a_i(\mathbf{r}_1, \mathbf{r}_2 ..... \mathbf{r}_N, t) = 0$.

For convenience below, we shall now combine (D.4) and (D.5) into a single set of equations,

$$\begin{aligned} &\Sigma_{k=1}^{N} \mathbf{B}_{ik} \bullet d\mathbf{r}_k + B_{it} dt = 0 \qquad i = 1,2....C \qquad // \ C = s+m \\ \text{or} \qquad &\\ &\Sigma_{k=1}^{N} \mathbf{B}_{ik} \bullet \dot{\mathbf{r}}_k + B_{it} = 0 \qquad i = 1,2....C \ , \end{aligned} \qquad (D.6)$$

where

$$\begin{aligned} \mathbf{B}_{ik} &= \mathbf{A}_{ik} \qquad B_{it} = \mathbf{A}_{it} \qquad \text{for } i = 1,2...s \qquad \text{non-holonomic} \\ \mathbf{B}_{ik} &= \nabla^{(k)} a_i \qquad B_{it} = \partial_t a_i \qquad \text{for } i = s+1, s+2....C \qquad \text{holonomic} \ . \end{aligned} \qquad (D.7)$$

In general, the functions a_i, $\mathbf{B}_{ik}$ and B_{it} are all functions of $(\mathbf{r}_1, \mathbf{r}_2 ..... \mathbf{r}_k, t)$.

In the above we have followed the notation of Ray and Shamanna (2006) and we continue our path in the manner they propose.

Suppose the set of N differentials $\{d\mathbf{r}_k\}$ satisfies all of equations (D.6) for some time interval dt. Such a set $\{d\mathbf{r}_k\}$ is then called a set of **allowed displacements**. Imagine that $\{d\mathbf{r'}_k\}$ is some other distinct allowed displacement set. Elements of the difference set defined as

$$\{\delta\mathbf{r}_k\} = \{d\mathbf{r}_k\} - \{d\mathbf{r'}_k\} \qquad k = 1,2...N \tag{D.8}$$

are then called **virtual displacements**. This is how Ray and Shamanna define the vague term virtual displacement which appears in many texts (about which they have much to say in their Section I A). It follows then that for any virtual displacement set $\{\delta\mathbf{r}_k\}$ equations (D.6) take this form

$$\Sigma_{k=1}^{N} \mathbf{B}_{ik} \bullet \delta\mathbf{r}_k = 0 \qquad i = 1,2....C \tag{D.9}$$

simply because the last term in (D.6) cancels out when one writes $\delta\mathbf{r}_k = d\mathbf{r}_k - d\mathbf{r'}_k$. Notice that the $\delta\mathbf{r}_k$ are not arbitrary differential displacements. They are differences of allowed displacements.

Meanwhile, Newton's Law for a system of N particles subject to constraints has the form

$$m_k \ddot{\mathbf{r}}_k = \mathbf{F}_k + \mathbf{R}_k \qquad k = 1,2...N \tag{D.10}$$

where $\mathbf{R}_k$ is the sum of all constraint forces applied to particle k and $\mathbf{F}_k$ is the sum of all *other* forces applied to particle k.

Assumption that constraints do no virtual work

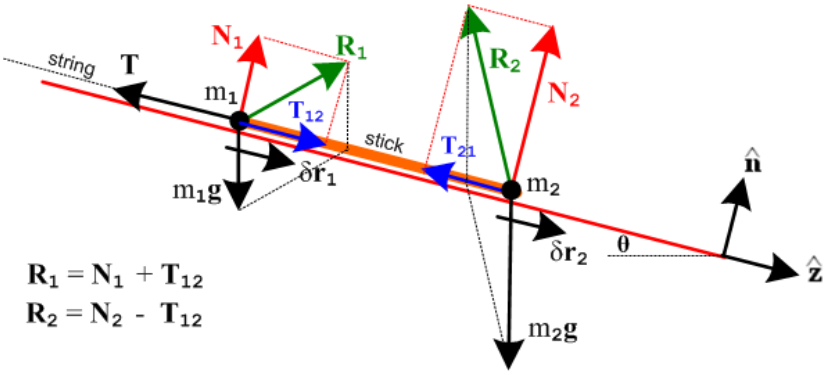
Next, *assume* that (Ray and Shamanna and others refer to this as being an "ideal constraint" situation),

$$0 = \Sigma_{k=1}^{N} \mathbf{R}_k \bullet \delta\mathbf{r}_k \; . \tag{D.11}$$

In Goldstein this equation corresponds to setting the second term in (1-40) to zero. The main justification for making this assumption is that it is valid for particles which form a rigid body and for many other constraint situations. There seems to be no general justification of this equation for *all* constraint situations, and it is certainly not valid if friction is present. The equation can be interpreted as saying that the **total virtual work** done by all the constraint forces on all the particles of the system vanishes.

As a first simple example, consider two particles 1 and 2 independently sliding down a frictionless static ramp. In this case for particle 1 the constraint force $\mathbf{R}_1$ is the normal force of the ramp acting on the particle, and this is clearly perpendicular to $d\mathbf{r}_1$ for any allowed displacement set $\{d\mathbf{r}_k\}$, and thus $\mathbf{R}_1$ is also perpendicular to any $\delta\mathbf{r}_1$ which is part of any virtual displacement set $\{\delta\mathbf{r}_k\}$. In this case we have $\mathbf{R}_1 \bullet \delta\mathbf{r}_1 = 0$ and $\mathbf{R}_2 \bullet \delta\mathbf{r}_2 = 0$ so the terms in (D.11) are separately zero.

A second example is more interesting. Particles 1 and 2 of arbitrary masses accelerate down (or up) a one dimensional ramp but are glued to the two ends of a massless stick so they comprise a rigid body. The upper mass is attached to a string which provides a constant tension T. Here is a picture (for $m_2 = 2m_1$),



$$m_1 N_2 = m_2 N_1$$

$$\mathbf{T}_{12} = -\frac{m_2}{m_1+m_2}\mathbf{T}$$

$$a_{1z} = a_{2z} = g\sin\theta - \frac{1}{m_1+m_2}|\mathbf{T}|$$

$$\hat{\mathbf{z}} \bullet \mathbf{R}_1 = \mathbf{T}_{12}$$

$$\hat{\mathbf{z}} \bullet \mathbf{R}_2 = -\mathbf{T}_{12}$$

(D.12)

Particle 2 exerts a constraint force $\mathbf{T}_{12}$ on particle 1 through the stick, and Particle 1 exerts a constraint force $\mathbf{T}_{21}$ on particle 2. Since the stick is massless, $\mathbf{T}_{21} = -\mathbf{T}_{12}$ ($\mathbf{T}_{12}+\mathbf{T}_{21} = m_{stick}\mathbf{a}$). All allowed displacement sets $\{d\mathbf{r}_1, d\mathbf{r}_2\}$ are of the form $\{d\mathbf{r}, d\mathbf{r}\}$, and therefore all *virtual* displacement sets are of the form $\{\delta\mathbf{r}_1, \delta\mathbf{r}_2\} = \{\delta\mathbf{r}, \delta\mathbf{r}\}$. In (D.12) we have drawn the two normal forces $\mathbf{N}_i$ and the two *total* constraint forces $\mathbf{R}_i$. It is no longer true that $\mathbf{R}_1 \bullet \delta\mathbf{r}_1 = 0$ and $\mathbf{R}_2 \bullet \delta\mathbf{r}_2 = 0$ as the picture shows. However, it is true that $\Sigma_{k=1}^{2}\, \mathbf{R}_k \bullet \delta\mathbf{r}_k = [\, \Sigma_{k=1}^{2}\, \mathbf{R}_k \,] \bullet \delta\mathbf{r} = 0$ because $\mathbf{R}_1$ and $\mathbf{R}_2$ have equal and opposite components along the ramp, so equation (D.11) is valid for this example.

Now suppose the ramp is translating in some direction not along the ramp. The displacements $d\mathbf{r}_1 = d\mathbf{r}_2$ are then no longer along the ramp so that $\Sigma_{k=1}^{2}\, \mathbf{R}_k \bullet d\mathbf{r}_k \neq 0$. However, any virtual displacement set would still have $\delta\mathbf{r}_1 = \delta\mathbf{r}_2$ along the ramp and then again $\Sigma_{k=1}^{2}\, \mathbf{R}_k \bullet \delta\mathbf{r}_k = 0$. Remember that a virtual displacement set is the difference of any two allowed displacement sets, and in this subtraction the effect of the ramp translation cancels out. So equation (D.11) is still valid.

Other examples are given in Ray and Shamanna Section III.

Reader Exercise: Derive the equations shown in (D.12). Show that things makes sense at $\theta = 0, \pi/2$, and also for cases $m_1 \to 0$ and $m_2 \to 0$. (Use $\mathbf{F}_1 = m_1\mathbf{g} + \mathbf{R}_1 + \mathbf{T}$, $\mathbf{F}_2 = m_2\mathbf{g} + \mathbf{R}_2$, $\mathbf{a}_1 = \mathbf{a}_2$.)

Comment on internal forces

In (D.10) we have classified the forces acting on particle k into two groups and we write $\mathbf{R}_k + \mathbf{F}_k$, where $\mathbf{R}_k$ are "constraint" forces and the $\mathbf{F}_k$ are any "other" forces. Goldstein refers to the $\mathbf{F}_k$ forces as "applied" forces in (1-39), while Ray and Shamanna call them "external" in Sec IV. Consider a system consisting of N masses all connected by a network of massless springs. The force acting on particle k due to all the springs to which it is attached would be best classified in the $\mathbf{F}_k$ "other" category. Were we to classify this force into the $\mathbf{R}_k$ group, we could not use the no-constraint-work property (D.11) since spring forces do work as the springs compress and expand. An example of such a system is the Solar System where the spring forces are represented by gravity. It would seem a misnomer to classify the internal forces in this case as "applied" or "external". However, if all the springs are slowly adjusted until they become infinitely stiff, in that limit it would be better to classify the internal spring forces into the $\mathbf{R}_k$ constraint forces group, since these springs do no work and (D.11) applies. The system is then a "rigid body" and in this case, with the internal forces classified as $\mathbf{R}_k$ constraint forces, one benefits from the many simplifications which arise for rigid body mechanics. The internal forces in our spring example fall into

the category where $\mathbf{F}_{ij} = -\mathbf{F}_{ji}$ and $\mathbf{F}_{ij}$ is in the direction $\mathbf{r}_i - \mathbf{r}_k$. When this is not the case, the rigid body limit does not apply, but the internal forces can still be classified as "constraint" plus "other".

First Application of Lagrange Multipliers: the forces of constraint

Recall the C = m+s constraint equations (D.9),

$$\Sigma_{k=1}^{N} \mathbf{B}_{ik} \bullet \delta\mathbf{r}_k = 0 \qquad i = 1,2...C . \qquad // \; C \equiv m+s = \text{total number of constraints} \qquad (D.9)$$

Multiply each of these C equations by a Lagrange Multiplier λ_i and then add the equations to get

$$0 = \Sigma_{i=1}^{C} \lambda_i \; [\Sigma_{k=1}^{N} \mathbf{B}_{ik} \bullet \delta\mathbf{r}_k] = 0$$

$$= \Sigma_{k=1}^{N} [\; \Sigma_{i=1}^{C} \lambda_i \; \mathbf{B}_{ik} \;] \bullet \; \delta\mathbf{r}_k . \qquad (D.13)$$

Now subtract equation (D.13) from (D.11) [$0 = \Sigma_{k=1}^{N} \mathbf{R}_k \bullet \delta\mathbf{r}_k$] to get

$$0 = \Sigma_{k=1}^{N} [\; \mathbf{R}_k - \Sigma_{i=1}^{C} \lambda_i \; \mathbf{B}_{ik}] \bullet \delta\mathbf{r}_k . \qquad (D.14)$$

We know that the N vectors in the set $\{\delta\mathbf{r}_k\}$ are not linearly independent because they must satisfy the equations (D.9), so we cannot claim from (D.14) that [$\mathbf{R}_k - \Sigma_{i=1}^{C} \lambda_i \; \mathbf{B}_{ik}$] = 0 for each k. But now write out equation (D.14) in full detail, showing all vector components,

$$0 = \Sigma_{k=1}^{N} \; \Sigma_{j=1}^{3} [\; (\mathbf{R}_k)_j - \Sigma_{i=1}^{C} \lambda_i \; (\mathbf{B}_{ik})_j] \, (\delta\mathbf{r}_k)_j . \qquad (D.15)$$

Next, relabel the 3N terms in this sum using a single index n defined in this way

$$\begin{array}{lllllll} (k,j) = (1,1), & (1,2), & (1,3), & (2,1), & (2,2), & (2,3), & .....(N,1), \quad (N,2), \quad (N,3) \\ n = 1, & 2, & 3, & 4, & 5, & 6 , & \qquad 3N\text{-}2, \quad 3N\text{-}1, \quad 3N \end{array}$$

so (D.16)

$$n = 3(k-1)+j \qquad k = 1+ \text{Int}[(n-1)/3] \qquad j = 1 + \text{Rem}[(n-1)/3] .$$

Then define (here the n value is implied by the k,j labels as per above)

$$x_n \equiv (\mathbf{r}_k)_j \qquad R_n \equiv (\mathbf{R}_k)_j \qquad B_{in} \equiv (\mathbf{B}_{ik})_j . \qquad (D.17)$$

Equation (D.15) can then be expressed as

$$0 = \Sigma_{n=1}^{3N} [\; R_n - \Sigma_{i=1}^{C} \lambda_i \; B_{in}] \; \delta x_n \qquad (D.18)$$

and (D.9) becomes $0 = \Sigma_{k=1}^{N} \; \Sigma_{j=1}^{3} \; (\mathbf{B}_{ik})_j \, (\delta\mathbf{r}_k)_j$ or

$$0 = \Sigma_{n=1}^{3N} B_{in} \; \delta x_n \qquad i = 1,2...C . \qquad (D.19)$$

Once again, we cannot set the square brackets in (D.18) to 0 because the 3N coordinates δx_n are not linearly independent due to the C equations (D.19).

At any given time t, due to these equations (D.19), there must exist at least one subset of the 3N variables δx_n which are independent and the remaining C δx_n are dependent. Assume that the dependent variables have indices $n = n_1, n_2,...n_C$. Consider then the following set of C equations involving the dependent-term square brackets in (D.18),

$$[\ R_n - \Sigma_{i=1}^{C} \lambda_i B_{in}] = 0 \qquad n = n_1, n_2,...n_C \ . \tag{D.20}$$

This represents C linear equations in the C parameters λ_i so, when all is said and done, there must exist a set of values $\{\lambda_i\}$ which make these C equations be true. We assume these values will be $\lambda_i(t)$ and this is justified below when the full set of equations is considered.

Now the terms in (D.18) indicated by $n = n_1, n_2,...n_C$ all vanish due to (D.20), so (D.18) can be written,

$$0 = \Sigma_{n \neq n_1, n_2, ... n_C} [\ R_n - \Sigma_{i=1}^{C} \lambda_i B_{in}] \delta x_n \ . \tag{D.21}$$

But all the δx_n appearing in (D.21) are independent, so the square brackets here must also vanish. We then end up with *all* the square brackets in (D.18) being 0 (assuming the correct solution values of λ_i),

$$[\ R_n - \Sigma_{i=1}^{C} \lambda_i B_{in}] = 0 \qquad n = 1,2....3N$$

or

$$R_n = \Sigma_{i=1}^{C} \lambda_i \ B_{in} \qquad n = 1,2....3N$$

or

$$(\mathbf{R}_k)_j = \Sigma_{i=1}^{C} \lambda_i \ (\mathbf{B}_{ik})_j \qquad k = 1,2..N \qquad j = 1,2,3$$

or

$$\mathbf{R}_k = \Sigma_{i=1}^{C} \lambda_i \ \mathbf{B}_{ik} \qquad k = 1,2...N \ . \tag{D.22}$$

Thus the constraint forces are certain linear combinations of the $\mathbf{B}_{ik}$ constraint functions shown in (D.7).

Meanwhile, Newton's Law (D.10) says that $m_k \mathbf{a}_k = \mathbf{F}_k + \mathbf{R}_k$ where the $\mathbf{F}_k$ are the non-constraint forces for the problem. Inserting (D.22) gives

$$m_k \ddot{\mathbf{r}}_k = \mathbf{F}_k + \Sigma_{i=1}^{C} \lambda_i \mathbf{B}_{ik} \ . \qquad \mathbf{B}_{ik} = \begin{cases} \mathbf{A}_{ik} & i = 1 \text{ to m (non-holonomic)} \\ [\nabla^{(k)} a_i] & i = m+1 \text{ to C (holonomic)} \end{cases} \tag{D.23}$$

The problem is then summarized in the following set of equations from (D.23) and (D.6),

$$m_k \dot{\mathbf{v}}_k(t) = \mathbf{F}_k(\mathbf{r}_1, \mathbf{r}_2..... \mathbf{r}_N, t) + \Sigma_{i=1}^{C} \lambda_i(t) \mathbf{B}_{ik}(\mathbf{r}_1, \mathbf{r}_2..... \mathbf{r}_N, t) \qquad k = 1..N \qquad 3N \text{ equations}$$

$$\dot{\mathbf{r}}_k(t) = \mathbf{v}_k(t) \qquad k = 1..N \qquad 3N \text{ equations}$$

$$\Sigma_{k=1}^{N} \mathbf{B}_{ik}(\mathbf{r}_1, \mathbf{r}_2..... \mathbf{r}_N, t) \bullet \dot{\mathbf{r}}_k(t) + B_{it}(\mathbf{r}_1, \mathbf{r}_2..... \mathbf{r}_N, t) = 0 \qquad i = 1,2....C \ . \qquad C \text{ equations}$$

(D.24)

This can be regarded as a set of 6N+C scalar first-order ODEs in the single variable t for the 6N+C unknown scalar functions $\mathbf{r}_k(t)$, $\mathbf{v}_k(t)$ and $\lambda_i(t)$. Although the first 3N equations happen to be linear in the $\lambda_i(t)$, the equations are in general non-linear in the $\mathbf{r}_k(t)$ so this is then a *non-linear* system of ODE's. Nevertheless, a well-known existence theorem (e.g., Ince Section 3.3) states that, provided the $\mathbf{F}_k$, $\mathbf{B}_{ik}$ and B_{it} functions are continuous and differentiable in all their arguments, a unique solution exists for any set of initial values $\mathbf{r}_k(0)$, $\mathbf{v}_k(0)$ and $\lambda_i(0)$. As (D.22) shows, the $\lambda_i(0)$ are related to the constraint forces at time 0.

The uniqueness of the solution agrees with our intuition that a physical system will evolve in a unique manner. Finding an analytic form for the solution might not be possible, but a numeric form can always be obtained.

Summary of the above Lagrange Multiplier Application

We have used a set of Lagrange Multipliers as an aid to solving a mechanics problem involving N particles with C constraints. As already noted, the above presentation is based on Section IV of Ray and Shamanna. The steps were as follows, considered more generically:

1. One has a sum $\Sigma_{n=1}^{N} R_n \delta x_n = 0$. (D.25)

2. One also has a set of sums $\Sigma_{n=1}^{N} B_{in} \delta x_n = 0$ for i = 1,2...C with C < N (constraints in the above case). This implies that one can find C of the δx_n which are linearly dependent on the other δx_n.

3. If λ_i are C arbitrary factors, certainly $\Sigma_{i=1}^{C} \lambda_i (\Sigma_{n=1}^{N} B_{in} \delta x_n) = 0$ based on 2. These λ_i are "Undetermined Lagrange Multipliers".

4. Subtract sum 3 from sum 1 to get the new sum $\Sigma_{n=1}^{N} [R_n - \Sigma_{i=1}^{C} \lambda_i B_{in}] \delta x_n = 0$.

5. The C λ_i are selected so that [..] = 0 in the C dependent δx_n terms of this sum. The remaining sum has only independent δx_n terms so [..] = 0 for all those terms as well. Then *all* [...] = 0 in the sum of step 4.

6. The conclusion is that $[R_n - \Sigma_i \lambda_i B_{in}] = 0$ for all n, assuming the solution λ_i values of step 5.

In contrast with our main topic of finding stationary points df = 0 of a function f(**r**), in the Lagrange Multiplier application just presented there is no explicit function f(**r**) we are making stationary. On the other hand, one can regard (D.11) that $\Sigma_k \mathbf{R}_k \bullet \delta \mathbf{r}_k = 0$ as the statement $\delta W = 0$ where $\delta W = \Sigma_k \mathbf{R}_k \bullet \delta \mathbf{r}_k$ is the total differential "virtual work" done by the constraint forces during a differential virtual displacement $\{\delta \mathbf{r}_k\}$. Using Newton's law (D.10) this equation can then be written

$$0 = \delta W = \Sigma_k [m_k \mathbf{a}_k - \mathbf{F}_k] \bullet \delta \mathbf{r}_k \qquad \text{// D'Alembert} \qquad \text{(D.26)}$$

in which form it is known as **D'Alembert's Principle** (1743, Goldstein (1-42)). The analog of f(**r**) is then an action-type integral W of the differential virtual work which integral one renders stationary to get δW = 0. The notion of "the action" is discussed at the end of this appendix.

We now present another application of the Lagrange Multiplier technique outlined in (D.25). As before, a certain amount of background is needed before the application can be demonstrated.

Generalized coordinates, generalized forces, and Lagrange's Equations

Usually the Cartesian components of the particle positions $\mathbf{r}_k$ are not the most convenient variables to use in solving a problem. Angles are often more useful. With C holonomic constraints one can replace the 3N variables $\{x_n\}$ in (D.17) with some set of 3N-C *independent* variables $\{q_n\}$ where

$$x_n = x_n(q_1, q_2, ...q_{3N-C}, t) \qquad n = 1,2..3N \qquad (D.27)$$

and one can then rework the above presentation in these new **generalized coordinates** q_j which then won't always have the dimensions of length. For example, one then has, for virtual displacements,

$$\delta x_n = \Sigma_{j=1}^{3N-C}(\partial x_n/\partial q_j)\,\delta q_j \qquad n = 1,2..3N$$

or

$$\delta \mathbf{r}_k = \Sigma_{j=1}^{3N-C}(\partial \mathbf{r}_k/\partial q_j)\,\delta q_j \qquad k = 1,2..N \qquad \text{// as in Goldstein (1-44) .} \qquad (D.28)$$

The usual presentation of Lagrangian dynamics with holonomic constraints *avoids* dealing with the forces of constraint and also avoids the need for using a set of Lagrange Multipliers. That presentation appears in Section V of the Ray and Shamanna and Goldstein p 17 (and many other places) and proceeds like this:

1. Start with $\Sigma_k \mathbf{R}_k \bullet \delta\mathbf{r}_k = 0$ in (D.11) which we write in the form of D'Alembert's Principle (D.26) mentioned above,

$$0 = \Sigma_{k=1}^{N} [m_k\mathbf{a}_k - \mathbf{F}_k] \bullet \delta\mathbf{r}_k \qquad \textit{(D.26)}$$

where $\mathbf{F}_k$ is the total non-constraint force acting on particle k.

2. Install (D.28) into D'Alembert's Principle to get

$$\begin{aligned} 0 &= \Sigma_{k=1}^{N} [m_k\mathbf{a}_k - \mathbf{F}_k] \bullet \Sigma_{j=1}^{3N-C}\left(\frac{\partial \mathbf{r}_k}{\partial q_j}\right)\delta q_j \\ &= \Sigma_{j=1}^{3N-C} \left\{ \Sigma_{k=1}^{N} m_k\mathbf{a}_k \bullet \frac{\partial \mathbf{r}_k}{\partial q_j} - \Sigma_{k=1}^{N} \mathbf{F}_k \bullet \left(\frac{\partial \mathbf{r}_k}{\partial q_j}\right) \right\} \delta q_j \\ &= \Sigma_{j=1}^{3N-C} \left\{ \Sigma_{k=1}^{N} m_k\mathbf{a}_k \bullet \frac{\partial \mathbf{r}_k}{\partial q_j} - Q_j \right\} \delta q_j \qquad (D.29) \end{aligned}$$

where

$$Q_j \equiv \Sigma_{k=1}^{N} \mathbf{F}_k \bullet \left(\frac{\partial \mathbf{r}_k}{\partial q_j}\right) . \qquad (D.30)$$

The Q_j defined above are the **generalized forces** associated with the generalized coordinates q_j .

3. Make use of the following non-obvious result (derived in Footnote below starting at (D.40)),

$$\Sigma_{k=1}^{N}\ m_k \mathbf{a}_k \bullet \frac{\partial \mathbf{r}_k}{\partial q_j} \;=\; \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_j}\right) - \frac{\partial T}{\partial q_j} \qquad \text{(D.31)}$$

where T is the total kinetic energy of the N particles,

$$T \equiv (1/2)\ \Sigma_{k=1}^{N}\ m_k\ (\mathbf{v}_k \bullet \mathbf{v}_k)\ . \qquad \text{(D.32)}$$

Then (D.29) reads,

$$0 = \Sigma_{j=1}^{3N-C}\ \{\ \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_j}\right) - \frac{\partial T}{\partial q_j} - Q_j\ \}\ \delta q_j\ . \qquad \text{(D.33)}$$

Since the δq_j are independent variables, {..} = 0 and we end up with one form of "Lagrange's Equations"

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_j}\right) - \frac{\partial T}{\partial q_j} \;=\; Q_j \qquad j = 1,2....3N\text{-}C\ . \qquad \text{// as in Goldstein (1-50)} \qquad \text{(D.34)}$$

If the "other" forces can be derived from a potential $\mathcal{V}(\mathbf{r}_1, \mathbf{r}_2, ... \mathbf{r}_N)$ such that

$$\mathbf{F}_k \;= -\ \nabla^{(k)}\, \mathcal{V}(\mathbf{r}_1, \mathbf{r}_2, ... \mathbf{r}_N) \qquad k = 1,2..N \qquad \text{// no t or } \dot{\mathbf{r}}_i \text{ dependence in } \mathcal{V} \qquad \text{(D.35)}$$

then from (D.30),

$$Q_j \equiv\ \Sigma_{k=1}^{N}\ \mathbf{F}_k \bullet \left(\frac{\partial \mathbf{r}_k}{\partial q_j}\right) = -\ \Sigma_{k=1}^{N}\ \left(\frac{\partial \mathcal{V}}{\partial \mathbf{r}_k}\right) \bullet \left(\frac{\partial \mathbf{r}_k}{\partial q_j}\right) \;= -\ \frac{\partial V}{\partial q_j} \qquad j = 1,2....3N\text{-}C \qquad \text{(D.36)}$$

where $V(q_i) \equiv \mathcal{V}(\mathbf{r}_k(q_i))$ is a function of the generalized coordinates q_i. Since V is not a function of the $\dot{q}_i$ we know that $\frac{\partial V}{\partial \dot{q}_j} = 0$. Using this fact, and (D.36) that $Q_j = -\ \frac{\partial V}{\partial q_j}$, (D.34) can be written

$$\frac{d}{dt}\left(\frac{\partial (T\text{-}V)}{\partial \dot{q}_j}\right) - \frac{\partial (T\text{-}V)}{\partial q_j} \;=\; 0\ . \qquad j = 1,2....3N\text{-}C \qquad \text{(D.37)}$$

The last step is to define the classical **Lagrangian**

$$L \equiv T - V \qquad \text{(D.38)}$$

to obtain

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_j}\right) - \frac{\partial L}{\partial q_j} = 0 \qquad j = 1,2....3N\text{-}C \quad . \qquad \text{// as in Goldstein (1-53)} \tag{D.39}$$

which is the more conventional form of "Lagrange's Equations" with C holonomic constraints. The equations of (D.39) are usually called the **Euler-Lagrange Equations**.

One can define **generalized (canonical, conjugate) momenta** $p_j \equiv \frac{\partial L}{\partial \dot{q}_j}$ so the Lagrange equations (D.39) are just $\dot{p}_j = \frac{\partial L}{\partial q_j}$ or $\dot{\mathbf{p}} = \nabla^{(q)} L$. If $V = V(q_i)$, then $p_j = \frac{\partial T}{\partial \dot{q}_j}$ and then if $T = \frac{1}{2}\Sigma_k m_k \dot{q}_k^2$ one has the familiar result $p_j = m_j \dot{q}_j$. When $T = T(\dot{q}_i)$ one finds $Q_j \equiv -\frac{\partial V}{\partial q_j} = \frac{\partial L}{\partial q_j} = \dot{p}_j$ so $\dot{\mathbf{p}} = \mathbf{Q}$, which is the **generalized Newton's Law** in terms of generalized coordinates and generalized forces.

Footnote: Derivation of the result (D.31) used above

Define the 3N-C independent generalized coordinates q_i as in (D.27) according to

$$\mathbf{r}_k = \mathbf{r}_k(q_1, q_2....q_{3N-C}, t) \equiv \mathbf{r}_k(q_i, t) \; . \tag{D.40}$$

Compute the total time derivative of $\mathbf{r}_k$,

$$\mathbf{v}_k = \dot{\mathbf{r}}_k = \frac{d\mathbf{r}_k}{dt} = \Sigma_j \frac{\partial \mathbf{r}_k}{\partial q_j}\frac{\partial q_j}{\partial t} + \frac{\partial \mathbf{r}_k}{\partial t} = \Sigma_j \frac{\partial \mathbf{r}_k}{\partial q_j}\dot{q}_j + \frac{\partial \mathbf{r}_k}{\partial t} \quad . \qquad \text{// } = \mathbf{v}_k(q_i, \dot{q}_i, t) \tag{D.41}$$

Then based on this result compute these two partial derivatives of $\mathbf{v}_k$,

$$\frac{\partial \mathbf{v}_k}{\partial q_i} = \Sigma_j \frac{\partial^2 \mathbf{r}_k}{\partial q_i \partial q_j}\dot{q}_j + \frac{\partial^2 \mathbf{r}_k}{\partial q_i \partial t} \tag{D.42}$$

$$\frac{\partial \mathbf{v}_k}{\partial \dot{q}_i} = \frac{\partial \mathbf{r}_k}{\partial q_i} \; . \tag{D.43}$$

Next, the total kinetic energy of all k particles is given by (D.32),

$$T \equiv (1/2)\,\Sigma_k \; m_k \,(\mathbf{v}_k \bullet \mathbf{v}_k) \tag{*D.32*}$$

so that

$$\frac{\partial T}{\partial \dot{q}_i} = \Sigma_k \, m_k \, \mathbf{v}_k \bullet \frac{\partial \mathbf{v}_k}{\partial \dot{q}_i} = \Sigma_k \, m_k \, \mathbf{v}_k \bullet \frac{\partial \mathbf{r}_k}{\partial q_i} \quad . \qquad \text{// using (D.43)} \tag{D.44}$$

Apply d/dt to the above, with $\mathbf{a_k} = \dot{\mathbf{r}}_\mathbf{k}$:

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_i}\right) = \Sigma_k\, m_k\, \mathbf{a_k} \bullet \frac{\partial \mathbf{r_k}}{\partial q_i} + \Sigma_k\, m_k\, \mathbf{v_k} \bullet \frac{d}{dt}\left(\frac{\partial \mathbf{r_k}}{\partial q_i}\right) \quad . \tag{D.45}$$

Compute the total time derivative appearing in the second term,

$$\frac{d}{dt}\left(\frac{\partial \mathbf{r_k}}{\partial q_i}\right) = \Sigma_j \frac{\partial^2 \mathbf{r_k}}{\partial q_j \partial q_i}\dot{q}_j + \frac{\partial^2 \mathbf{r_k}}{\partial t\, \partial q_i} \quad . \tag{D.46}$$

Meanwhile,

$$\frac{\partial T}{\partial q_i} = \Sigma_k\, m_k\, \mathbf{v_k} \bullet \frac{\partial \mathbf{v_k}}{\partial q_i}$$

$$= \Sigma_k\, m_k\, \mathbf{v_k} \bullet \left[\, \Sigma_j \frac{\partial^2 \mathbf{r_k}}{\partial q_i \partial q_j}\dot{q}_j + \frac{\partial^2 \mathbf{r_k}}{\partial q_i \partial t}\,\right] \qquad \text{// (D.42)}$$

$$= \Sigma_k\, m_k\, \mathbf{v_k} \bullet \frac{d}{dt}\left(\frac{\partial \mathbf{r_k}}{\partial q_i}\right) \quad . \qquad \text{// (D.46)} \tag{D.47}$$

Subtract (D.47) from (D.45) to get

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_i}\right) - \frac{\partial T}{\partial q_i} = \Sigma_k\, m_k\, \mathbf{a_k} \bullet \frac{\partial \mathbf{r_k}}{\partial q_i} \tag{D.48}$$

which is the result quoted above in (D.31).

Second Application of Lagrange Multipliers: non-holonomic constraints

This is our final Lagrange Multiplier example, and it exactly follows the generic prescription of (D.25). The general topic is well described in Goldstein Chapter 2.

As the starting point we invoke **Hamilton's Principle** which is this :

$$\delta S = 0 \qquad \text{where} \qquad S = \int_{t1}^{t2} dt\, L(q_n(t), \dot{q}_n(t), t) \qquad n = 1,2....M \tag{D.49}$$

where there are M independent generalized coordinates q_n. The notation is a shorthand to imply that L is

a function of *all* the $q_n(t)$ and *all* the $\dot{q}_n(t)$ functions. One might better write $L(\{q_n(t)\}, \{\dot{q}_n(t)\}, t)$ where the {..} indicate sets.

L = T - V is the Lagrangian (D.38) with $T(\dot{q}_i)$ and $V(q_i)$ being the kinetic and potential energies associated with the M generalized coordinates. We shall assume that M = N-D where there were D holonomic constraints on N initial problem coordinates. As shown in (D.61) below (or see Goldstein p 41 (2-20') based on p 37 (2-15)) the equation $\delta S = 0$ results in the following sum being 0,

$$1. \quad \Sigma_{n=1}^{M} \left[\int_{t1}^{t2} dt \left\{ \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_n}\right) - \frac{\partial L}{\partial q_n} \right\} \right] \delta q_n(t) = 0 \quad . \tag{D.50}$$

The number 1. on the left refers to the first step outlined in (D.25). In this equation the $\delta q_n(t)$ are M independent functions of t ("path variations"), restricted only by $\delta q_n(t_1) = \delta q_n(t_2) = 0$. In order for (D.50) to be true for M arbitrary independent *functions* $\delta q_n(t)$, one must have {} = 0,

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_n}\right) - \frac{\partial L}{\partial q_n} = 0 \qquad n = 1,2....M \tag{D.51}$$

which are the Euler-Lagrange Equations (D.39). Thus it is that the Lagrange Equations can be derived from Hamilton's Principle in the case of only holonomic constraints.

But suppose that in addition to the D holonomic constraints there are C non-holonomic constraints, so only M-C of the δq_n are independent. In this case the above conclusion (D.51) is incorrect.

2. The C non-holonomic constraints result in C sums $\Sigma_n A_{ni} \delta q_n = 0$ i = 1,2...C analogous to (D.9) above.

3. If λ_i are C arbitrary factors, certainly $\Sigma_i \lambda_i (\Sigma_n A_{in} \delta q_n) = 0$ based on 2. These λ_i are "Undetermined Lagrange Multipliers". Then $\int_{t1}^{t2} dt \ (\Sigma_i \lambda_i \Sigma_n A_{in} \delta q_n) = 0$ as well.

4. Subtract sum 3 from sum 1 to get the new sum

$$\Sigma_{n=1}^{M} \left[\int_{t1}^{t2} dt \left\{ \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_n}\right) - \frac{\partial L}{\partial q_n} - \Sigma_{i=1}^{C} \lambda_i A_{in} \right\} \right] \delta q_n(t) = 0 \quad . \quad \text{// as in Goldstein (2-26)} \tag{D.52}$$

5. The C λ_i are selected so that {..} = 0 in the C dependent $\delta q_n(t)$ terms of this sum. The remaining sum has only independent $\delta q_n(t)$ terms so {..} = 0 for all those terms as well. Then *all* {...} = 0 in the sum of 4.

6. The conclusion is that

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_n}\right) - \frac{\partial L}{\partial q_n} = \Sigma_{i=1}^{C} \lambda_i A_{in} \quad . \qquad n = 1,2....M \quad . \quad \text{// as in Goldstein (2-30)} \tag{D.53}$$

This then is the corrected form of the Euler-Lagrange Equations in the presence of C non-holonomic constraints. Recalling from (D.7) that $\mathbf{B_{ik}} = \mathbf{A_{ik}}$ for non-holonomic constraints, the second line of (D.22) with $B_{in} = A_{in}$ reads $R_n = \Sigma_{i=1}^{C} \lambda_i A_{in}$. Thus one can interpret the sum on the right of (D.53) as being the force of constraint R_n arising from the C non-holonomic constraints. The forces of the D holonomic constraints are already incorporated into the left side of (D.53) in the reduction of independent coordinates q_n from N to M = N-D.

The integral S shown in (D.49) is called "the action" and Hamilton's Principle is a classical mechanics incarnation of the **Principle of Least Action**. In relativistic quantum field theory the action is a spacetime integral of the Lagrangian density $\mathcal{L}(\varphi_n(x^\mu), \partial_\mu\varphi_n(x^\mu))$ where the coordinates q_n of L are replaced by fields $\varphi_n(x^\mu)$, and t is replaced by x^μ, a point in spacetime. In this case the Lagrange Equations (D.39) take the form (the second line is for comparison),

$$\frac{\partial}{\partial x^\mu}\left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu\varphi_n)}\right) - \frac{\partial \mathcal{L}}{\partial \varphi_n} = 0 \quad \text{where} \quad \partial_\mu \equiv \frac{\partial}{\partial x^\mu} \qquad \text{// Bjorken and Drell (11.30)} \tag{D.54}$$

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_n}\right) - \frac{\partial L}{\partial q_n} = 0 \qquad t \rightarrow x^\mu \qquad q_n(t) \rightarrow \varphi_n(x^\mu) \qquad \dot{q}_n(t) \rightarrow \partial_\mu\varphi_n(x^\mu) \qquad \textit{(D.39)}$$

<u>The Case of C holonomic constraints treated as if they were non-holonomic</u>

Suppose there are N generalized coordinates q_i with C holonomic constraints and 0 non-holonomic constraints. As noted in (D.6) and (D.7), the holonomic constraints can be written in differential form so they look just like non-holonomic constraints. For generalized coordinates q_n these two equations would appear as

$$\Sigma_{n=1}^{N} B_{in} dq_n + B_{it} dt = 0 \qquad i = 1,2....C \qquad \text{// } C = s+m$$

or (D.6)'

$$\Sigma_{n=1}^{N} B_{in} \dot{q}_n + B_{it} = 0 \qquad i = 1,2....C$$

where

$$\begin{array}{llll} B_{in} = A_{in} & B_{it} = \mathbf{A_{it}} & \text{for } i = 1,2...s & \text{non-holonomic} \\ B_{in} = \partial a_i/\partial q_n & B_{it} = \partial_t a_i & \text{for } i = s+1, s+2....C & \text{holonomic} \quad . \end{array} \tag{D.7)'}$$

Setting s = 0 we can carry out the procedure of the previous section with $\mathbf{A_{ik}}$ replaced by $\partial a_i/\partial q_k$ to obtain this version of (D.53) rendering the action S stationary,

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_n}\right) - \frac{\partial L}{\partial q_n} = \Sigma_{i=1}^{C} \lambda_i \frac{\partial a_i}{\partial q_n} \qquad n = 1,2....N \tag{D.55}$$

where the Lagrange multipliers λ_i are chosen to make this equation be valid for a set of C dependent q_n coordinates, as in Step 5 above. Only N-C of the coordinates q_n are independent but the equations are

written for all N coordinates. Here the functions $a_i(q_1, q_2, ...q_n, t) = 0$ are the C holonomic constraint equations.

Footnote: Carry out the $\delta S = 0$ functional variation of the action .

The variation in each function $q_n(t)$ is parameterized in terms of an arbitrary independent function $\eta_n(t)$ (but which vanishes at both time integration endpoints) and one scalar parameter α :

$$q_n(t,\alpha) = q_n(t,0) + \alpha\,\eta_n(t) \qquad \delta q_n(t) = \alpha\,\eta_n(t) \qquad \frac{\partial q_n}{\partial \alpha} = \eta_n(t) \qquad \eta_n(t_1) = \eta_n(t_2) = 0$$

$$\dot{q}_n(t,\alpha) = \dot{q}_n(t,0) + \alpha\,\dot{\eta}_n(t) \qquad \delta\dot{q}_n(t) = \alpha\,\dot{\eta}_n(t \qquad \frac{\partial \dot{q}_n}{\partial \alpha} = \dot{\eta}_n(t) \quad . \qquad \text{(D.56)}$$

Then

$$S(\alpha) \equiv \int_{t1}^{t2} dt\, L(q_n(t,\alpha), \dot{q}_n(t,\alpha), t) \qquad \text{// action} \qquad \textit{(D.49)}$$

$$\frac{\partial S}{\partial \alpha} = \int_{t1}^{t2} dt\ \Sigma_n \left(\frac{\partial L}{\partial q_n}\frac{\partial q_n}{\partial \alpha} + \frac{\partial L}{\partial \dot{q}_n}\frac{\partial \dot{q}_n}{\partial \alpha} \right) = \int_{t1}^{t2} dt\ \Sigma_n \left(\frac{\partial L}{\partial q_n}\,\eta_n + \frac{\partial L}{\partial \dot{q}_n}\frac{d\eta_n}{dt} \right) \quad . \qquad \text{(D.57)}$$

But for the second term do a parts integration,

$$\int_{t1}^{t2} dt \left(\frac{\partial L}{\partial \dot{q}_n}\frac{d\eta_n}{dt} \right) = \left[\frac{\partial L}{\partial \dot{q}_n}\eta_n(t) \right]_{t1}^{t2} - \int_{t1}^{t2} dt \left(\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_n} \right) \eta_n(t) \,. \qquad \text{(D.58)}$$

The "parts" terms vanish since η_n vanishes at both endpoints, so (D.57) becomes,

$$\frac{\partial S}{\partial \alpha} = \int_{t1}^{t2} dt\ \Sigma_n \left(\frac{\partial L}{\partial q_n} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}_n} \right) \eta_n(t) \qquad \text{(D.59)}$$

so

$$\delta S \equiv \frac{\partial S}{\partial \alpha}\, d\alpha = \left\{ \int_{t1}^{t2} dt\ \Sigma_n \left(\frac{\partial L}{\partial q_n} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}_n} \right) \eta_n(t) \right\} d\alpha . (\alpha/\alpha)$$

$$= \left\{ \int_{t1}^{t2} dt\ \Sigma_n \left(\frac{\partial L}{\partial q_n} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}_n} \right) \delta q_n(t) \right\} d\alpha/\alpha \quad . \qquad \text{(D.60)}$$

Then,

$$\delta S = 0 \quad \Leftrightarrow \quad \int_{t1}^{t2} dt \ \Sigma_n \left[\left(\frac{\partial L}{\partial q_n} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_n} \right) \eta_n(t) \right] = 0 .$$

or (D.61)

$$\delta S = 0 \quad \Leftrightarrow \quad \int_{t1}^{t2} dt \ \Sigma_n \left[\left(\frac{\partial L}{\partial q_n} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_n} \right) \delta q_n(t) \right] = 0 .$$

But the functions $\delta q_n(t,\alpha) = \alpha \eta_n(t)$ are arbitrary and independent functions of t, so the integral can only vanish if the function in the parentheses vanishes for each term in the sum on n, so

$$\delta S = 0 \quad \Leftrightarrow \quad \left(\frac{\partial L}{\partial q_n} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_n} \right) = 0, \quad n = 1,2...m \qquad (D.62)$$

which are the Euler-Lagrange Equations (D.39) and (D.51). For more detail see Goldstein Chapter 2.

References

The method of Lagrange multipliers often makes cameo appearances in textbooks on the calculus of multiple variables. There are also several pages and pdf's on the 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. The links below were last verified 29 Oct 2016.

J.D. Bjorken and S.D. Drell, *Relativistic Quantum Fields* (McGraw-Hill, New York, 1965).

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, but our page and equation references are to the 1950 edition.

E.L. Ince, *Ordinary Differential Equations* (Longmans, Green & Co. (Ltd.), London, 1927). This classic text became a Dover paperback in 1956 and is available online for $4.

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* (2016), 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).

S. Ray and J. Shamanna, "On Virtual Displacement and Virtual Work in Lagrangian Dynamics", European Journal of Physics, 27 (2006) 311-329. Also https://arxiv.org/abs/physics/0510204v2 .

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

M.R. Spiegel, *Mathematical Handbook of Formulas and Tables* (Schaum's Outline Series, McGraw-Hill, New York, 1968). This is now at 4th Ed (2012) with two coauthors and new reference numbers.

W. F. Trench, *The Method of Lagrange Multipliers* (2013), http://digitalcommons.trinity.edu/mono/7/ , Supplement 2 (31p). His Theorem 1 (Eq. 6) is basically the same as our Theorem 1 (1.24).

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

M.W. Zemansky, *Heat and Thermodynamics*, *5th Ed.* (McGraw-Hill, New York, 1968). Our page references are to this edition, but there is a 7th Ed. (1997) with coauthor R.H Dittman and an 8th Ed (2011) in the Special Indian Edition (SIE) series.