Lagrange Multipliers

Robert Gilmore
Physics Department, Drexel University, Philadelphia, PA 19104

January 27, 2003

Abstract

The problem of determining the maximum value, minimum value, or the stationary values of a function of several variables when these variables are subject to one or more constraints occurs sufficiently often that special methods have been devised to address this problem. One special method involves the use of Lagrange multipliers. We address this problem from three points of view, two of them exploiting Lagrange multipliers.

1 Introduction

1.1 Statement of the Problem

It frequently happens that some function ("objective function") must be maximized (or minimized, or made stationary) subject to conditions. To be specific, a function of n variables must be optimized subject to k constraints. The function is expressed as a function

$$f(x^1, x^2, \cdots, x^n) \tag{1}$$

of n variables $x^1, x^2, \cdots, x^n$. The constraints are expressed in terms of k constraint equations

$$\begin{aligned} \phi_1(x^1, x^2, \cdots, x^n) &= c_1 \\ \phi_2(x^1, x^2, \cdots, x^n) &= c_2 \\ \vdots \qquad\qquad &\vdots \quad \vdots \\ \phi_k(x^1, x^2, \cdots, x^n) &= c_k \end{aligned} \tag{2}$$

1.2 Approaches to the Problem

There are three general approaches to treating this problem.

The first approach (Sec. 2) involves reducing the number of independent variables to $n - k$ and then solving the unconstrained equations by standard methods. The second approach (Sec. 3) involves retaining n independent variables and solving for them as functions of k auxiliary variables λ^α $(1 \leq \alpha \leq k)$, the Lagrange multipliers. The third approach (Sec. 4) involves augmenting the set of independent variables to $n + k$ and optimizing an unconstrained function of $n + k$ variables by the usual methods.

In order to illustrate each method we carry along a specific example. It is treated by each of these methods in turn. The common example is this: Optimize the nonsingular quadratic form

$$Ax^2 + By^2 \tag{3}$$

subject to the condition

$$ax + by = c \tag{4}$$

For this example $n = 2$ and $k = 1$.

For each of the three methods a flow diagram outlining the logic is presented and a Maple worksheet describing how this logic applies to the standard example is also provided.

Section	# Independent Variables
2	$n - k$
3	n
4	$n + k$

In Section 5 we provide a very useful result relating the rate of change of the value of the objective function at its critical point with the Lagrange multipliers and the differentials of the constraint values dc_i. There is a kind of duality between the two.

1.3 Applications

Applications of "Lagrange multiplier technology" given here fall into three groups. Two applications are given in the first group. One is drawn from classical physics. It involves computing the classical trajectory of a light ray under reflection and refraction conditions. This involves objective functions with square roots. These calculations are carried out in Section 6. The second example, presented in Section 7, is purely mathematical. It demonstrates how sophisticated methods (of algebraic topology) are used to compute the stationary functions of a polynomial objective function when the constraint is also polynomial. This example shows that there can be many critical values. It also describes how the number of critical values can change ("bifurcate") as the value of the constraint changes.

The second group of examples contains three classes of applications where the methods of linear algebra are used to give explicit closed forms for the solutions. These involve:

Section	Objective Function	Constraint(s)
8	Quadratic	Linear
9	Linear	Quadratic
10	Quadratic	Quadratic

Section 11 provides a nice application of these results to the problem of fitting a straight line to data when there are errors in the measurements of both the dependent (y) and independent (x) variables.

The final two sections deal with applications to Classical Thermodynamics (Section 12) and Statistical Mechanics (Section 13). In both cases the Lagrange multipliers have a natural interpretation as intensive thermodynamic variables that are conjugate to specific extensive thermodynamic variables.

2 Reducing The Number of Variables

2.1 The Procedure

In the first approach the k constraint equations are used to solve for k of the variables x^i in terms of the remaining $n-k$, which are then treated as independent unconstrained variables. These k expressions in $n-k$ independent variables are then plugged into the function to be optimized, which becomes a function of $n-k$ independent variables. This function of $n-k$ variables is optimized in the usual way: by taking partial derivatives, setting them equal to zero, and solving for the values of the $n-k$ independent variables that cause the function to be stationary. These values are used to determine the value of the k dependent variables at the critical point. The values of the $n-k$ independent variables and the k dependent variables are plugged back into the function to find its stationary values.

The logic of this procedure is summarized in the flow chart shown in Fig. 1.

2.2 Example

Application of this method of solution to the common example is summarized in the Maple worksheet shown in Fig. 2. We first solve the constraint equation for one of the variables in terms of the other: $y = (c - ax)/b$. This expression for y in terms of x is substituted into the quadratic form to provide an objective function of a single $(n = 2, k = 1, n - k = 1)$ variable x:

$$f(x) = Ax^2 + B[(c - ax)/b]^2$$

The critical point of this function of a single variable is determined by differentiating and setting the derivative equal to zero

$$2Ax + 2B[(c - ax)/b](-a/b) = 0$$

The value of x that solves this equation is determined: $x_c = acB/(a^2B + b^2A)$. The subscript $_c$ identifies this value of x as the x coordinate of the critical point. This value of x is used to determine the value of the dependent variable y at the critical point: $y_c = bcA/(a^2B + b^2A)$ (by symmetry!). The values of the independent variable x_c and the dependent variable y_c are next substituted into the function f to determine its value at the stationary point ("critical value"):

$$f_{\text{stat}}(x_c, y_c) = \frac{ABc^2}{a^2B + b^2A}$$

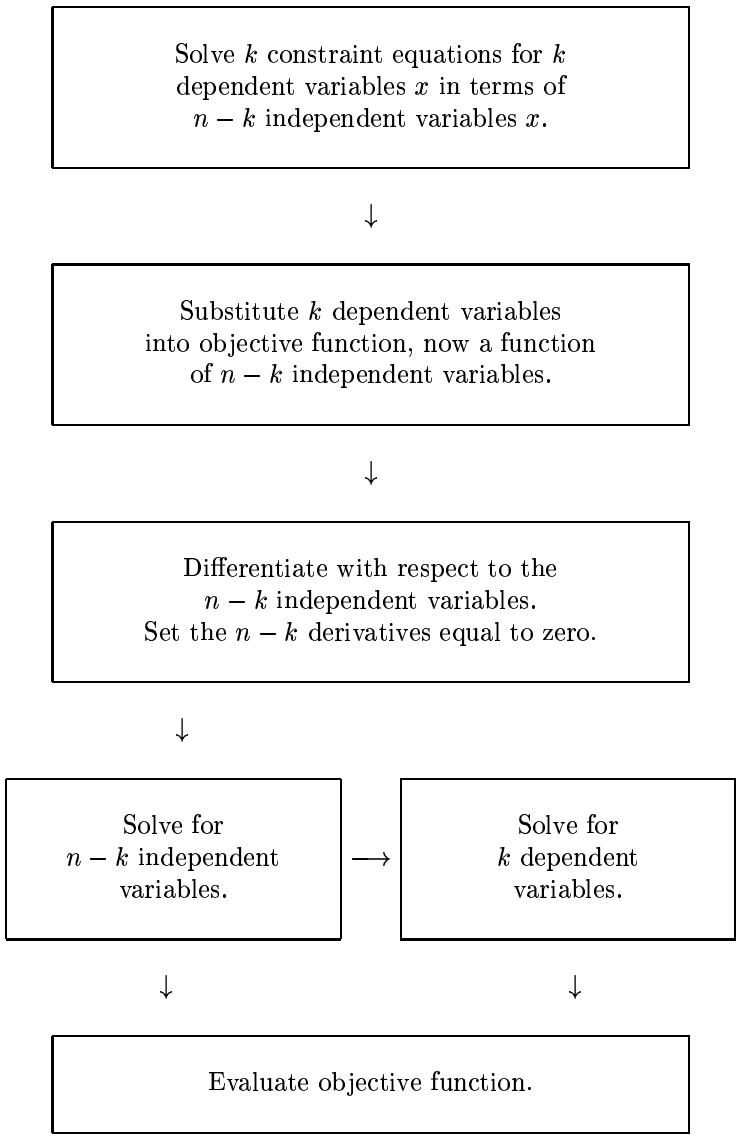


Figure 1: Flow diagram describing the logic of the method depending on $n-k$ independent variables.

```
> f:=A*x^2+B*y^2;          ## Expression to be made stationary
```

$$f := A x^2 + B y^2$$

```
> g:=a*x+b*y-c=0;          ## Constraint
```

$$g := a x + b y - c = 0$$

```
> y1:=solve(g,y);          ## Solve constraint for y
```

$$y1 := -\frac{a x - c}{b}$$

```
> f2:=subs(y=y1,f);        ## Substitute into f to get function of one
  variable
```

$$f2 := A x^2 + \frac{B (a x - c)^2}{b^2}$$

```
> h:=diff(f2,x);           ## Differentiate this function of x
>
```

$$h := 2 A x + \frac{2 B (a x - c) a}{b^2}$$

```
> x1:=solve(h=0,x);        ## Solve for x
```

$$x1 := \frac{B c a}{A b^2 + B a^2}$$

```
> y2:=subs(x=x1,y1);       ## Substitute value of x to dertermine value
  of y
```

$$y2 := -\frac{\frac{a^2 B c}{A b^2 + B a^2} - c}{b}$$

```
> z1:=subs(x=x1,y=y2,f);## plug values of x and y into f
```

$$z1 := \frac{A B^2 c^2 a^2}{(A b^2 + B a^2)^2} + \frac{B \left(\frac{a^2 B c}{A b^2 + B a^2} - c \right)^2}{b^2}$$

```
> answer:=simplify(z1);  ##  Incredible!  What a mess!!  Simplify.
```

$$answer := \frac{A B c^2}{A b^2 + B a^2}$$

```
~
```

Figure 2: Maple worksheet describing solution of constrained maximization using $n - k$ independent variables. The constraint equation is used to express a dependent variable y in terms of an independent variable x. The function to be optomized is a function of the single dependent variable x. The derivative is set equal to zero, and the value of x at the critical point is determined. This is used to determine y_c. The independent variable x_c and dependent variable y_c are used to determine the value of f at the critical point (critical value).

3 Lagrange Multipliers 1. As Basic Variables

3.1 Procedure

The second approach involves the introduction of a set of Lagrange multipliers, $\lambda^1, \lambda^2, \cdots, \lambda^k$. There is one for each constraint equation. A new objective function is created:

$$F(x^1, x^2, \cdots, x^n) = f(x^1, x^2, \cdots, x^n) + \sum_{\alpha=1}^{k} \lambda^\alpha (\phi_\alpha(x^1, x^2, \cdots, x^n) - c_\alpha) \tag{5}$$

The Lagrange multipliers "lift" the constraints. This new function ("modified objective function") is treated as an unconstrained function of the n independent variables x^i. It is treated in the usual way. The n partial derivatives $\partial F/\partial x^i$ are taken and set to zero, giving a set of n equations involving n variables x^i and k Lagrange multipliers λ^α. The n equations are used to solve for the n variables x^i in terms of the k Lagrange multipliers. These n expressions for the x^i are plugged back into the k constraint equations, and the values of the k Lagrange multipliers are determined from these k equations. These k values of the Lagrange multipliers at the critical point are plugged back into the expressions for the x^i. Finally, these critical values of the x^i are plugged back into the original function $f(x^1, x^2, \cdots, x^n)$ [or the modified objective function $F(x^1, x^2, \cdots, x^n)$] to determine its stationary value.

The logic of this procedure is summarized in the flow chart shown in Fig. 3.

3.2 Example

Application of this method of solution to the common example is summarized in the Maple worksheet shown in Fig. 4. A multiple of the constraint equation is added to the objective function to create a new objective function depending on the two variables (x, y) and the Lagrange multiplier λ

$$F(x, y) = Ax^2 + By^2 + \lambda(ax + by - c)$$

This is treated as an unconstrained maximization problem. The partial derivatives with respect to the two variables (x, y) are taken and set equal to zero. This gives two equations in the two unknowns. These equations are solved for x and y as a function of the Lagrange multiplier λ: $x = -\lambda a/2A$, $y = -\lambda b/2B$. These values of x and y are then substituted into the constraint equation to construct a single equation in the single parameter λ:

$$\left(\frac{a^2}{2A} + \frac{b^2}{2B}\right)\lambda + c = 0$$

This equation is solved for λ: $\lambda = -2ABc/(a^2B + b^2A)$. This value of λ is substituted back into the expressions for x and y to give their critical values x_c and y_c. As a final step, these critical values are substituted back into $f(x, y)$ or $F(x, y)$ to obtain the value of f at the critical point.

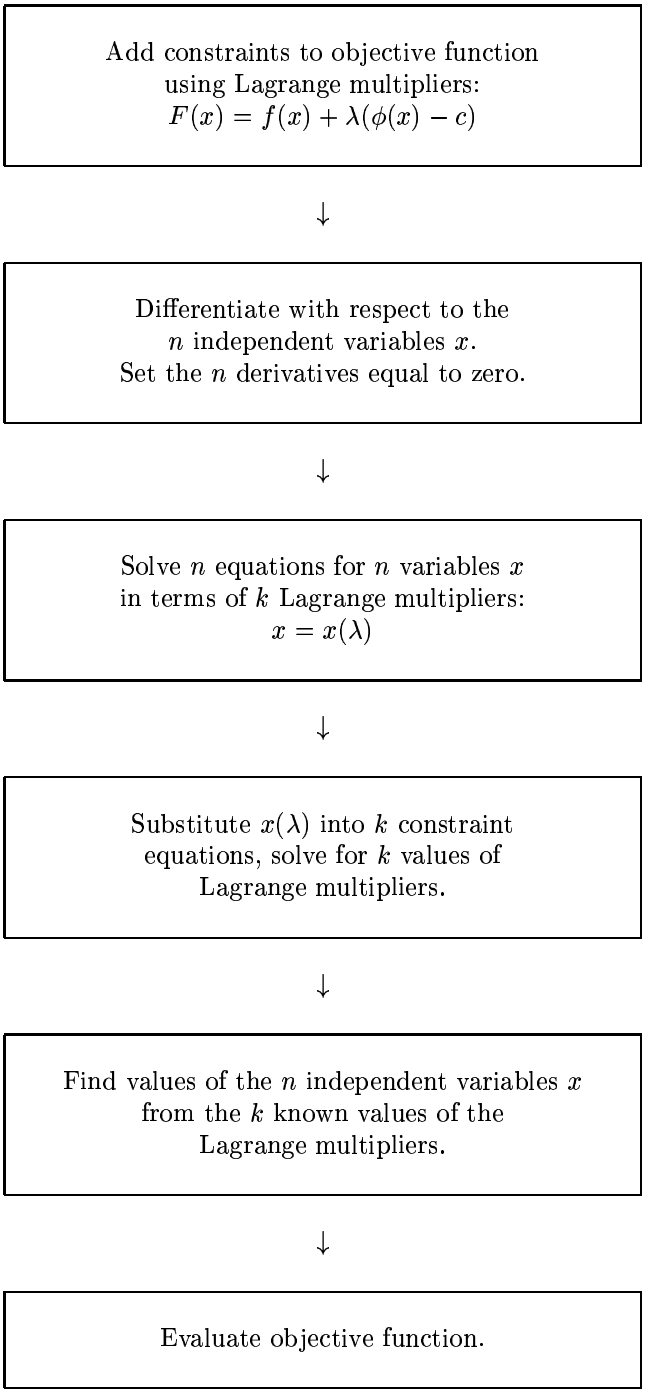


Figure 3: Flow diagram describing the logic of the method depending on n independent variables.

```
> ##  Use Lagrange Multipliers to maximize a function.
> f(x,y):=A*x^2+B*y^2;  ## Expression to be 'stationarized'
```

$$f(x,y) := Ax^2 + By^2$$

```
> g(x,y):=a*x+b*y-c;    ## Constraint
```

$$g(x,y) := ax + by - c$$

```
> F(x,y):=f(x,y)-lambda*g(x,y);  ## Unconstrained function
```

$$F(x,y) := Ax^2 + By^2 - \lambda(ax + by - c)$$

```
> F_x:=diff(F(x,y),x);  ## Partial derivative wrt x
```

$$F\_x := 2Ax - \lambda a$$

```
> F_y:=diff(F(x,y),y);  ## Partial derivative wrt y
```

$$F\_y := 2By - \lambda b$$

```
> solve({F_x=0,F_y=0},{x,y});  ## Solve for x,y in terms of lambda
```

$$\{y = \frac{1}{2}\frac{\lambda b}{B}, x = \frac{1}{2}\frac{\lambda a}{A}\}$$

```
> x1:=1/2/A*lambda*a;y1:=1/2/B*lambda*b;
>
```

$$x1 := \frac{1}{2}\frac{\lambda a}{A}$$

$$y1 := \frac{1}{2}\frac{\lambda b}{B}$$

```
> answer1:=subs(y=y1,x=x1,g(x,y));  ## Evaluate constrraint in terms
  of lambda
```

$$answer1 := \frac{1}{2}\frac{a^2\lambda}{A} + \frac{\frac{1}{2}b^2\lambda}{B} - c$$

```
> lambda1:=solve(answer1=0,lambda); ## Solve constraint for lambda
```

$$\lambda 1 := 2\frac{cAB}{a^2B + b^2A}$$

```
> x2:=subs(lambda=lambda1,x1);      ## Evaluaate x
```

$$x2 := \frac{cBa}{a^2B + b^2A}$$

```
> y2:=subs(lambda=lambda1,y1);      ## Evaluate y
```

$$y2 := \frac{cAb}{a^2B + b^2A}$$

```
> answer2:=subs(x=x2,y=y2,f(x,y)); ## Evaluate f(x,y) at stationary
  pt.
```

$$answer2 := \frac{Ac^2B^2a^2}{(a^2B + b^2A)^2} + \frac{Bc^2A^2b^2}{(a^2B + b^2A)^2}$$

```
> answer3:=simplify(answer2);       ## Simplify this mess
```

$$answer3 := \frac{Ac^2B}{a^2B + b^2A}$$

```
>
```

Figure 4: Maple worksheet describing solution of constrained maximization using n independent variables. The partial derivatives of $F(x, y)$ are set equal to zero and the values of the variables (x, y) are determined as a function of the Lagrange multiplier λ. These expressions for (x, y) are substituted into the constraint equations to determine the critical value of λ, λ_c. This value is used to determine critical values for x and y: (x_c, y_c). The critical values are plugged back into $f(x, y)$ to provide its value at the critical point.

4 Lagrange Multipliers 2. As Additional Variables

4.1 Procedure

In the third approach the Lagrange multipliers are adjoined to the initial set of n variables to construct a set of $n+k$ variables, where $x^{n+1} = \lambda^1$, $\cdots$, $x^{n+k} = \lambda^k$. The objective function is

$$F(x^1, \cdots, x^n; x^{n+1}, \cdots, x^{n+k}) = f(x^1, \cdots, x^n) + \sum_{\alpha=1}^{k} x^{n+\alpha}(\phi_\alpha(x^1, \cdots, x^n) - c_\alpha) \tag{6}$$

This function F of $n+k$ variables is treated as an unconstrained function of all its arguments. The first n partial derivatives $\partial F/\partial x^j = 0$ provide exactly the information presented by differentiating Eq(5) above. The remaining k partial derivatives are exactly the constraint equations: $\partial F/\partial x^{n+\alpha} = (\phi_\alpha(x^1, \cdots, x^n) - c_\alpha) = 0$. The resulting set of $n+k$ simultaneous equations in $n+k$ variables has only isolated solutions in general. The $n+k$ equations are solved for the set of isolated critical points, and these values are plugged back into the function $F(x^1, \cdots, x^n; x^{n+1}, \cdots, x^{n+k})$. Equivalently, and more simply, the first n components of the $n+k$ vector of solutions is plugged back into the starting function $f(x^1, \cdots, x^n)$.

The logic of this procedure is summarized in the flow chart shown in Fig. 5

4.2 Example

Application of this method of solution to the common example is summarized in the Maple worksheet shown in Fig. 6. From the function $f(x, y)$ to be optimized and the constraint equation we construct an objective function of *three* independent real variables according to

$$F(x, y, z) = Ax^2 + By^2 + z(ax + by - c)$$

We search for an extremum by computing the three partial derivatives and searching for solutions in the usual way:

$$\begin{array}{rcrcl} 2Ax & & + & az & = 0 \\ & 2By & + & bz & = 0 \\ & & & ax + by - c & = 0 \end{array}$$

These three equations are simultaneously solved for the three independent variables (x, y, z) to find (x_c, y_c, z_c). The first two coordinates are plugged back into the initial function $f(x, y)$ to obtain its value at the stationary point. Equivalently, the three coordinates (x_c, y_c, z_c) of the critical point are substituted back into the modified objective function $F(x, y, z)$ to obtain the same critical value.

4.3 Solutions Using Algebraic Geometry

Powerful tools exists for solving the $n+k$ equations resulting from this procedure when both the objective function and the constraints are polynomial functions. This method uses tools developed for the study of Algebraic Geometry.

Each equation defines a "codimension-1" surface in a space of dimension n. This jargon just means that the surface has one lower dimension (*co*dimension one) than the $n+k$ dimensional space in which the surface is embedded (e.g., the two-sphere S^2 defined by $x^2 + y^2 + z^2 = 1$ has one lower dimension than the 3-space R^3 in which it is embedded). Intersection of two such surfaces is a surface of codimension-2. Continuing, $n+k$ such surfaces intersect (if they intersect at all) at isolated points in R^{n+k}. Special powerful algorithms from algebraic geometry have been developed

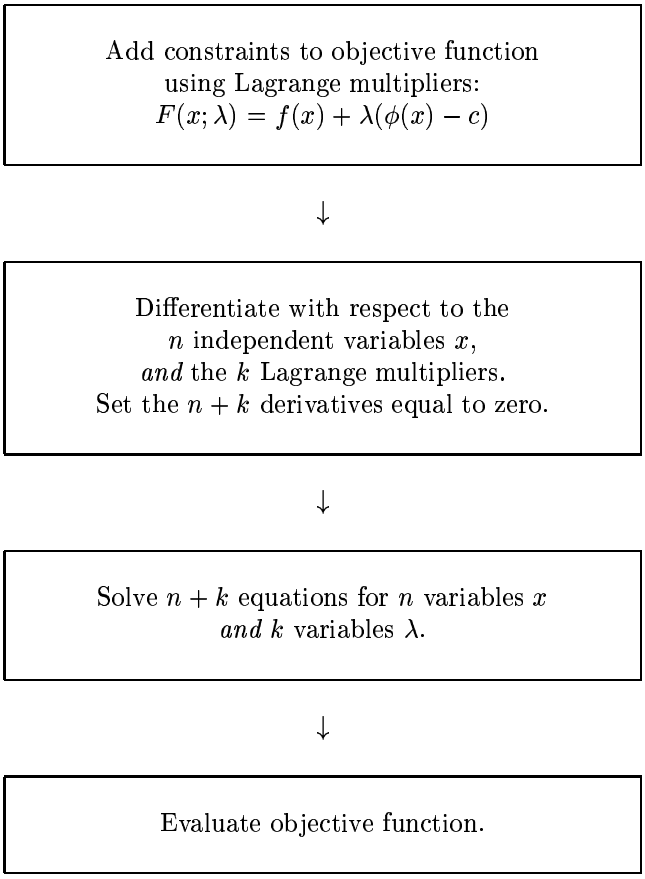


Figure 5: Flow diagram describing the logic of the method depending on $n+k$ independent variables.

```
> ##  Use Lagrange multipliers to maximize a function.
> ##  Exploit the maximum symmetry.
>
> f(x,y):=A*x^2+B*y^2;
```

$$f(x,y) := A x^2 + B y^2$$

```
> g(x,y):=a*x+b*y-c;
```

$$g(x,y) := a x + b y - c$$

```
> F(x,y,lambda):=f(x,y)-lambda*g(x,y);
```

$$F(x,y,\lambda) := A x^2 + B y^2 - \lambda (a x + b y - c)$$

```
> F_x:=diff(F(x,y,lambda),x);
```

$$F\_x := 2 A x - \lambda a$$

```
> F_y:=diff(F(x,y,lambda),y);
```

$$F\_y := 2 B y - \lambda b$$

```
> F_l:=diff(F(x,y,lambda),lambda);
```

$$F\_l := -a x - b y + c$$

```
> solve({F_x=0,F_y=0,F_l=0},{x,y,lambda});
```

$$\{y = \frac{c A b}{B a^2 + A b^2}, x = \frac{c B a}{B a^2 + A b^2}, \lambda = 2 \frac{B c A}{B a^2 + A b^2}\}$$

```
> x1:=c*B*a/(B*a^2+A*b^2);
```

$$x1 := \frac{c B a}{B a^2 + A b^2}$$

```
> y1:=1/(B*a^2+A*b^2)*c*A*b;
```

$$y1 := \frac{c A b}{B a^2 + A b^2}$$

```
> answer:=subs(x=x1,y=y1,f(x,y));
```

$$answer := \frac{A c^2 B^2 a^2}{(B a^2 + A b^2)^2} + \frac{B c^2 A^2 b^2}{(B a^2 + A b^2)^2}$$

```
> simplify(answer);
```

$$\frac{A c^2 B}{B a^2 + A b^2}$$

```
>
```

Figure 6: Maple worksheet describing solution of constrained maximization using $n + k$ independent variables. A new function with three independent variables, (x, y, z), is created from the original function to be optimized and the constraint equation. The critical values (x_c, y_c, z_c) are determined by setting the three partial derivatives $\partial F/\partial x, \partial F/\partial y, \partial F/\partial x$ equal to zero. The first two coordinates (x_c, y_c) are substituted back into the original function to obtain its value at the critical point. The third critical coordinate, z_c, is the value of the Lagrange multiplier at the critical point.

to determine the locations of these intersections. They are embedded in many symbol manipulation codes: the *grobner* package in Maple, for example.

Application of this method of solution to the common example is summarized in the Maple worksheet shown in Fig. 7. First, the *grobner* package is loaded. The function to be optimized and the constraint equation are introduced, and the modified objective function $F(x, y, z)$ of three variables is defined, as before. Partial derivatives with respect to the three independent variables (x, y, z) are computed, also as before. It is the solution method for this set of equations that differs. An algorithm is summoned (gbasis). Two arguments are sent into this algorithm. The first consists of the three simultaneous equations whose solutions are sought. This argument is sent in as an array. The second argument consists of the three variables whose values are sought. This argument is also sent in as an array: $[x, y, z]$. The output of the algorithm depends on the order in which this last array is introduced ($[z, y, x]$ results in an output that looks lots different). The output generated by this algorithm consists of one or more arrays. Each array represents a solution of the simultaneous equations. Each array (in Maple's implementation) should be read left to right. The leftmost entry is a function of the single variable x (because of ordering: x is the first argument in the array $[x, y, z]$). This expression is set to zero and solved for x: $x_c = acB/(a^2B + b^2A)$. The next argument in the output array involves both x and y. The value x_c is introduced, the expression is set equal to zero, and the result is solved for y_c: $y_c = bcA/(a^2B + b^2A)$. Things continue on along this line until all components of the input vector (x, y, z) are determined. If there is no interest in the value of the Lagrange multiplier $\lambda = z_c$, one can stop after the value of y_c is determined. The set (x_c, y_c) is backsubstituted into $f(x, y)$ to determine the value of this function at a critical point. If *gbasis* outputs another array, as it may if the input equations are nonlinear, this procedure is repeated. This continues until all output arrays have been solved. Since the input equations are linear, *gbasis* outputs a single array.

A nonsingular quadratic function under linear constraints will have a single critical point. It is for this reason that *gbasis* will output a single array for such problems.

5 A Differential Result

The value of the function (5) depends on the critical points $(x^1, x^2, \cdots, x^n)$, the values of the Lagrange multipliers λ^α, and the values of the constraining parameters c_α. The critical points (x) depend on the Lagrange multipliers λ, which in turn depend on the values of the constraints (c), so that in fact the stationary (or critical) values of the modified function F depend only on the values of the constraints. We show this now by computing the differential of F:

$$F(x, \lambda, c) = f(x) + \lambda^\alpha(\phi_\alpha(x) - c_\alpha) \tag{7}$$

$$dF(x, \lambda, c) = \underbrace{\left(\frac{\partial f(x)}{\partial x^i} + \lambda^\alpha \frac{\partial \phi_\alpha(x)}{\partial x^i}\right)}_{=0} dx^i + \underbrace{(\phi_\alpha(x) - c_\alpha)}_{=0} d\lambda^\alpha - \lambda^\alpha dc_\alpha = -\lambda^\alpha dc_\alpha \tag{8}$$

The underlined terms vanish, leaving only the dependence on the values of the constraints. In particular, the value of the function increases in the direction of the constraint c_α by an amount equal the the conjugate Lagrange multiplier $(\mp)\lambda^\alpha$, depending on the sign $(\pm)$ with which the constraints are added to the original objective function.

Applications

```
> with(grobner);
```

[finduni, finite, gbasis, gsolve, leadmon, normalf, solvable, spoly]

```
> f(x,y):=A*x^2+B*y^2;
```

$$f(x,y) := A\,x^2 + B\,y^2$$

```
> g(x,y):=a*x+b*y-c;
```

$$g(x,y) := a\,x + b\,y - c$$

```
> F(x,y,lam):=f(x,y)-lam*g(x,y);
```

$$F(x,y,lam) := A\,x^2 + B\,y^2 - lam\,(a\,x + b\,y - c)$$

```
> F_x:=diff(F(x,y,lam),x);
```

$$F\_x := 2\,A\,x - lam\,a$$

```
> F_y:=diff(F(x,y,lam),y);
```

$$F\_y := 2\,B\,y - lam\,b$$

```
> F_l:=diff(F(x,y,lam),lam);
```

$$F\_l := -a\,x - b\,y + c$$

```
> gbasis([F_x,F_y,F_l],[x,y,lam]);
```

$$[B\,a^2\,x + x\,A\,b^2 - B\,a\,c,\ B\,y\,a^2 + y\,A\,b^2 - b\,A\,c,\ B\,lam\,a^2 - 2\,B\,A\,c + A\,b^2\,lam]$$

```
> x1:=solve(B*a^2*x+x*A*b^2-B*a*c=0,x);
```

$$x1 := \frac{B\,a\,c}{B\,a^2 + A\,b^2}$$

```
> y1:=subs(x=x1,B*y*a^2+y*A*b^2-b*A*c);
```

$$y1 := B\,y\,a^2 + y\,A\,b^2 - b\,A\,c$$

```
> y2:=solve(B*y*a^2+y*A*b^2-b*A*c=0,y);
```

$$y2 := \frac{b\,A\,c}{B\,a^2 + A\,b^2}$$

```
> answer:=subs(x=x1,y=y2,f(x,y));
```

$$answer := \frac{A\,B^2\,a^2\,c^2}{(B\,a^2 + A\,b^2)^2} + \frac{B\,b^2\,A^2\,c^2}{(B\,a^2 + A\,b^2)^2}$$

```
> simplify(answer);
```

$$\frac{A\,B\,c^2}{B\,a^2 + A\,b^2}$$

```
>
```

Figure 7: Maple worksheet describing solution of constrained maximization using $n+k$ independent variables. The grobner package is first loaded. The algorithm proceeds as before to construct a set of $n+k$ simultaneous equations in $n+k$ unknowns. The equations are loaded into the *gbasis* callup in one array. The second argument in the *gbasis* routine is an array containing the independent variables. There is one output array for each solution set. The solutions of each output array are obtained explicitly by setting each component of the array equal to zero and working from left to right.

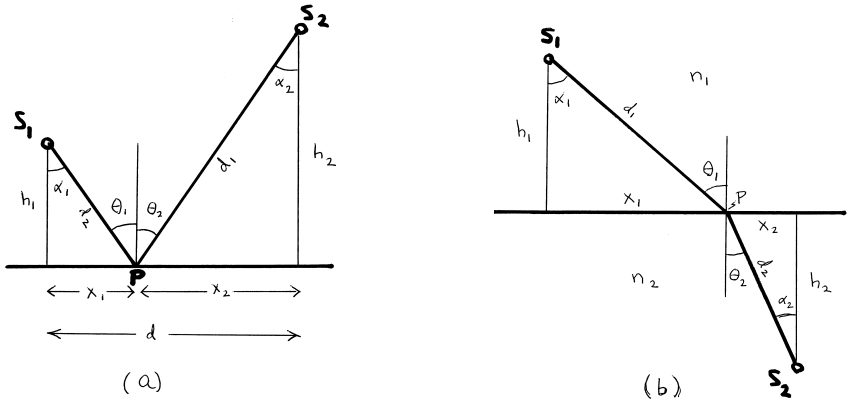


Figure 8: Reflection and refraction. (a) A light ray from S_1 to S_2 is reflected at P from a horizontal surface. The total *distance* traveled is minimum. (b) A light ray from S_1 in medium 1 with index of reflection n_1 travels to S_2 in a medium with index n_2. The total *time* traveled is minimum.

6 Reflection and Refraction

The classical problems of reflection and refraction of light rays can be treated simply using the technology of Lagrange multipliers.

6.1 Reflection

A light ray originates at the source S_1, is reflected by a surface, and observed at the sink S_2 (*or vice versa*). What path is followed?

The geometry is shown in Fig. 8(a). We assume that the ray hits the reflecting surface at the point P. This divides the interval between the perpendicular projections of S_1 and S_2 onto the reflecting plane (of length d) into two subintervals of lengths x_1 and x_2. The values of x_1 and x_2 are not known, but the constraint is $x_1 + x_2 = d$. It is assumed that the shortest path is determined by the following principle:

Principle of Least Distance: The shortest path covers the least distance.

If this principle is correct, the shortest path is determined by computing the shortest total distance subject to the constraint $x_1 + x_2 - d = 0$. Pythagoras tells us that the distance traveled during the first leg $(_1)$ of the trip is determined by $d_1^2 = h_1^2 + x_1^2$. A similar result holds for the second leg of the trip. The objective function to be optimized is

$$F(x_1, x_2; \lambda) = \sqrt{h_1^2 + x_1^2} + \sqrt{h_2^2 + x_2^2} - \lambda(x_1 + x_2 - d) \tag{9}$$

From this we find

$$\begin{aligned} \frac{\partial F}{\partial x_1} &= \frac{x_1}{\sqrt{h_1^2 + x_1^2}} - \lambda = 0 \\ \frac{\partial F}{\partial x_2} &= \frac{x_2}{\sqrt{h_2^2 + x_2^2}} - \lambda = 0 \end{aligned} \tag{10}$$

In this computation there is no need to evaluate the Lagrange multiplier, since the two ratios $x_i/\sqrt{h_i^2 + x_i^2}$ are equal. From the figure, these ratios are the sines of the enclosed angles: $x_i/\sqrt{h_i^2 + x_i^2} = \sin(\alpha_i)$. However, Euclid tells us that $\alpha_i = \theta_i$, so that the condition for reflection is

$$\sin\theta_1 = \sin\theta_2 \quad (= \lambda) \tag{11}$$

In short, the angle of incidence is equal to the angle of reflection.

Remark: Euclid tells us how to solve this problem without resorting to Calculus at all.

6.2 Refraction

The problem of refraction can be treated similarly. However, there is a slight problem. When we place a stick in water, it appears bent. Two interpretations are possible. One is that the stick bends but light travels in a straight line. Such a possibility is consistent with the Principle of Minimal Length. However, most people don't believe this explanation: We believe the stick remains straight and light bends as it passes from one medium to another. The geometry and the light ray trajectory are shown in Fig. 8(b).

In this case the Principle of Least Distance won't do: something else must be minimized. One reasonable guess is that the *time* it takes to go from the source S_1 to the sink S_2 is minimum.

Principle of Least Time: The shortest path takes the least time.

The speed of light in a medium with index of refraction n is c/n, where c is the speed of light in vacuum. The optimization problem then becomes

$$T = \frac{d_1}{c/n_1} + \frac{d_2}{c/n_2} - \lambda(x_1 + x_2 - d) \tag{12}$$

Proceeding as above, we find

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \quad (= \lambda) \tag{13}$$

Euclid would have a little more trouble deriving this result.

Remark: Although the Principle of Least Distance cannot be used to derive the refraction result, the Principle of Least Time *can* be used to derive the reflection result. We conclude that the latter is the more fundamental of the two Principles.

7 Polynomial Objective Function - Polynomial Constraint

When both the objective function and the constraints are polynomial functions the methods of algebraic topology can be used to find the stationary values. We illustrate with an example from Cox, Little, and O'Shea.

The objective function is the nonhomogeneous polynomial $f(x, y, z) = x^3 + 2xyz - z^2$. The stationary values of this function on the unit sphere $\phi(x, y, z) = x^2 + y^2 + z^2 = r^2 = 1$ are to be

determined. Setting up the problem in the usual way, $F(x,y,z;\lambda) = f(x,y,z) - \lambda\,(\phi(x,y,z) - 1)$, leads to the four equations

$$\begin{array}{lcccccccc} \partial F/\partial x & = & 3x^2 & + & 2yz & - & 2\lambda x & = & 0 \\ \partial F/\partial y & = & & + & 2xz & - & 2\lambda y & = & 0 \\ \partial F/\partial z & = & 2xy & - & 2z & - & 2\lambda z & = & 0 \\ \partial F/\partial \lambda & = & x^2 & + & y^2 & + & z^2 & = & 1 \end{array} \tag{14}$$

The call to gbasis reveals the structure of the solution:

$$gbasis([F_x, F_y, F_z, F_\lambda], [x, y, z, \lambda]);$$

$$\begin{array}{l} \lambda - \frac{3}{2}x - \frac{3}{2}yz - \frac{167616}{3835}z^6 - \frac{36717}{590}z^4 - \frac{134419}{7670}z^2 \\ x^2 + y^2 + z^2 - 1 \\ xy - \frac{19584}{3835}z^5 + \frac{1999}{295}z^3 - \frac{6403}{3835}z \\ xz + yz^2 - \frac{1152}{3835}z^5 - \frac{108}{295}z^3 + \frac{2556}{3835}z \\ y^3 + yz^2 - y - \frac{9216}{3835}z^5 + \frac{906}{295}z^3 - \frac{2562}{3835}z \\ y^2z - \frac{6912}{3835}z^5 + \frac{827}{295}z^3 - \frac{3839}{3835}z \\ yz^3 - yz - \frac{576}{59}z^6 + \frac{1605}{118}z^4 - \frac{453}{118}z^2 \\ z^7 - \frac{1763}{1152}z^5 + \frac{655}{1152}z^3 - \frac{11}{288}z \end{array}$$

The last factor is set equal to zero and the seven roots of z are determined. The roots are $0, \pm 1, \pm\frac{2}{3}, \pm\sqrt{22}/16$. These values are substituted into the previous equations to determine the values of x and y allowed for each of the seven possible values of z. Each solution set is inserted into the objective function to produce a critical value. The results follow:

z	y	x	$f_{\rm cr}$
0	0	± 1	± 1
0	± 1	0	0
± 1	0	0	-1
$\pm\frac{2}{3}$	$\pm\frac{1}{3}$	$-\frac{2}{3}$	$-\frac{28}{27}$
$\pm\frac{\sqrt{22}}{16}$	$\mp\frac{3\sqrt{22}}{16}$	$-\frac{3}{8}$	$\frac{7}{128}$

The maximum value of this function on the sphere is $+1$ at $(x,y,z) = (+1,0,0)$, while the minimum is "doubly degenerate" in the sense that the minimum at $-\frac{28}{27}$ occurs at the symmetry-related points $(-\frac{2}{3}, \pm\frac{1}{3}, \pm\frac{2}{3})$.

Remark: The solution set above has been determined for the unit sphere, $r = 1$. Solution sets also exist for other values of the sphere radius, r, so that solution sets are functions of r. The solution sets can be written as $(x(r), y(r), z(r))_\alpha$, where α indexes the different solutions. In the case above, $1 \le \alpha \le 10$. It is possible that as r ranges from 0^+ to "∞", two or more solution sets disappear in bifurcations, or additional solution sets appear. Since only one parameter r appears in this optimization problem, fold bifurcations (A_2) are possible. Since there is a symmetry ($(x,y,z) \to (x,-y,-z)$), cusp bifurcations (A_3) may also occur. No other bifurcations are generically possible.

Three Important Classes of Problems

Certain classes of problems involving Lagrange multipliers can be solved systematically and in closed form. These involve cases where the objective function is either linear or quadratic, and the constraints are linear or quadratic. In the following three sections we systematically treat the three cases of interest. These are as shown.

Objective Function	Constraint(s) Linear	Quadratic
Linear		Sec.9
Quadratic	Sec.8	Sec.10

8 Quadratic Form — Linear Constraints

An entire class of optimization problems can be treated by the methods of linear algebra. These problems involve finding the critical point and value of a quadratic form subject to linear constraints.

8.1 Result

We begin with a quadratic form in n variables

$$Q(x^1, \cdots, x^n) = \frac{1}{2} Q_{ij} x^i x^j \tag{15}$$

The matrix Q_{ij} is assumed to be nonsingular but not necessarily positive definite. This allows the possibility of finding maxima (if Q is negative definite) or saddles (if Q is indefinite).

Next we introduce k linear constraint equations

$$A_{\alpha j} x^j - c_\alpha = 0 \tag{16}$$

The quadratic form and constraint equations are combined to form a functon of $n + k$ variables $x \in R^n$ and $z \in R^k$ following the procedure introduced in Section 4:

$$F(x; z) = \frac{1}{2} Q_{ij} x^i x^j + z^\alpha (A_{\alpha j} x^j - c_\alpha) \tag{17}$$

The n partial derivatives with respect to the x^i and the k partial derivatives with respect to the z_α can be written in the forms

$$\begin{array}{lrcl} \partial_i : & Q_{ij} x^j + A_{\alpha i} z^\alpha & = & 0 \\ \partial_\alpha : & A_{\alpha j} x^j & = & c_\alpha \end{array} \qquad \text{or} \qquad \begin{bmatrix} Q & A^t \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ c \end{bmatrix} \tag{18}$$

In the matrix expression on the right, A is the $k \times n$ matrix of coefficients $A_{\alpha j}$ and A^t is its transpose. The column vectors $[x], [z], [0]$, and $[c]$ are of length n, k, n, and k. From these expressions we find simple relations for the Lagrange multipliers z and the original variables x at the critical point:

$$z_c = -(AQ^{-1}A^t)^{-1} c \qquad\qquad x_c = Q^{-1} A^t (AQ^{-1}A^t)^{-1} c \tag{19}$$

From here it is a simple step to evaluate the quadratic form at the critical point:

$$Q(x_c) = \frac{1}{2} c^t (AQ^{-1}A^t)^{-1} c \tag{20}$$

8.2 Example

For the common example in Eqs(3,4) we find

$$AQ^{-1}A^t = \left[\begin{array}{cc} a & b \end{array}\right] \left[\begin{array}{cc} \frac{1}{2A} & 0 \\ 0 & \frac{1}{2B} \end{array}\right] \left[\begin{array}{c} a \\ b \end{array}\right] = \frac{a^2B + b^2A}{2AB}$$

This yields immediately

$$Q(x_c) = \frac{ABc^2}{a^2B + b^2A} \qquad \lambda_c = z_c = -\frac{2ABc}{a^2B + b^2A} \tag{21}$$

Thus the theorem that $\partial Q(x_c)/\partial c_\alpha = \partial F/\partial c_\alpha = -\lambda_\alpha$ at the critical point is satisfied in this example.

9 Linear Form - Quadratic Constraint

9.1 Result

The objective function is assumed to be linear: $f(x) = A_i x^i = Ax$. Here A is a row vector. The constraint has the form $\phi(x) : \frac{1}{2}x^t Bx = \frac{1}{2}B_{ij}x^i x^j = c$. In matrix form, the objective function to be optimized is

$$F(x;\lambda) = Ax - \lambda(\frac{1}{2}x^t Bx - c) \tag{22}$$

This leads to the linear equation

$$\nabla F(x;\lambda) = A^t - \lambda Bx = 0 \Rightarrow x = \frac{1}{\lambda}B^{-1}A^t$$

The normalization condition is used to determine the Lagrange multiplier:

$$AB^{-1}A^t = 2c\lambda^2$$

From this we obtain immediately for the stationary value

$$f(x) \to \sqrt{2c}\,\sqrt{AB^{-1}A^t}$$

9.2 Example

As an example, we dualize the canonical example followed through Section 2 - Section 4. We will attempt to optimize the linear function $f(x,y) = ax + by$ subject to the quadratic constraint $Ax^2 + By^2 = c$. The modified objective function is

$$F(x,y,\lambda) = (ax + by) - \lambda(Ax^2 + By^2 - c)$$

From this we easily compute

$$\begin{array}{lccccccc} \frac{\partial}{\partial x}: & a & - & 2\lambda Ax & = & 0 & \Rightarrow & x = \frac{a}{2\lambda A} \\ \frac{\partial}{\partial y}: & b & - & 2\lambda By & = & 0 & \Rightarrow & y = \frac{b}{2\lambda B} \end{array}$$

The Lagrange multiplier λ is determined by plugging these expressions for x and y into the constraint equation:

$$A\left(\frac{a}{2\lambda A}\right)^2 + B\left(\frac{b}{2\lambda B}\right)^2 = c \quad \Rightarrow \quad \left(\frac{1}{2\lambda}\right)^2 = \frac{ABc}{a^2B + b^2A}$$

From this we easily determine the values of x and y in terms of the parameters $a, b; A, B$ and the value of the constraint c, and from these values the stationary value of the linear function:

$$f(x(c), y(c)) = \sqrt{c}\sqrt{\frac{a^2}{A} + \frac{b^2}{B}} = \frac{1}{2\lambda}$$

10 Quadratic Form — Quadratic Constraint

10.1 Result

In this case the objective function is the quadratic form in n variables $(x^1, x^2, \cdots, x^n)$ that is easily expressed in matrix form $f(x) = x^t Ax = A_{ij}x^i x^j$ The constraint is represented by the quadratic form $\phi(x) = x^t Bx - c = B_{ij}x^i x^j - c = 0$. To be specific we assume that both matrices A and B are positive definite. The optimization problem assumes canonical form

$$F(x; \lambda) = x^t Ax - \lambda(x^t Bx - c) = x^t(A - \lambda B)x + \lambda c \tag{23}$$

Taking the derivatives $\partial/\partial x^i$ leads to the eigenvalue equation

$$(A - \lambda B)x = 0 \tag{24}$$

We assume further that the positive definite matrices A and B have nondegenerate eigenvalues λ_α with corresponding eigenvectors $x(\alpha)$. These eigenvectors are orthogonal with respect to the matrices A and B (i.e., $x^t(\alpha)Ax(\beta) = 0, x^t(\alpha)Bx(\beta) = 0, \alpha \neq \beta$). We normalize the eigenvectors so they satisfy the constraint, so that $x^t(\alpha)Bx(\alpha) = c$. For the eigenvector $x(\alpha)$ normalized to satisfy the constraint, the eigenvalue equation (24) gives

$$x^t(\alpha)(A - \lambda_\alpha B)x(\alpha) = x^t(\alpha)Ax(\alpha) - x^t(\alpha)\lambda_\alpha Bx(\alpha) = x^t(\alpha)Ax(\alpha) - \lambda_\alpha c = 0 \tag{25}$$

The result is as follows: Each eigenvector of the eigenvalue equation $(A - \lambda B)x = 0$ provides a stationary value for the objective function. The critical value of the objective function, for the eigenvector $x(\alpha)$, is $\lambda_\alpha c$.

10.2 Solving the Eigenvalue Equation

The eigenvalue equation $(A - \lambda B)x = 0$ can be solved as follows. We first assume that the real symmetric $n \times n$ matrix B is nonsingular. This matrix is written as $B = B^{1/2}B^{1/2}$, where $B^{1/2} =$"$\sqrt{B}$" (see below). The eigenvalue equation can then be expressed

$$B^{1/2}(B^{-1/2}AB^{-1/2} - \lambda I)B^{1/2}x = 0 \tag{26}$$

The closely related eigenvalue equation

$$(A' - \lambda I)y = 0 \tag{27}$$

is solvable with standard routines. Here $A' = B^{-1/2}AB^{-1/2}$ and $y = B^{1/2}x$. The eigenvectors $y(\alpha)$ are orthogonal with respect to both matrices A' and I. The eigenvectors $x(\alpha) = B^{-1/2}y(\alpha)$ are orthogonal with respect to both matrices A and B.

10.3 Computing the Square Root of a Matrix

To compute the square root of the matrix B we proceed as follows. The eigenvectors $z(\alpha)$ of B (normalized to unity) are columns in the real orthogonal transformation S that diagonalizes B

$$S^{-1}BS = D \qquad\qquad D = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \tag{28}$$

By inverting this relation we find an expression for B in terms of an orthogonal matrix S and a diagonal matrix D: $B = SDS^{-1}$. In particular, $B^2 = SDS^{-1}SDS^{-1} = SD^2S^{-1}$ and $B^n = (SDS^{-1})^n = SD^nS^{-1}$. More generally

$$f(B) = f(SDS^{-1}) = Sf(D)S^{-1} = S \begin{bmatrix} f(\lambda_1) & & & \\ & f(\lambda_2) & & \\ & & \ddots & \\ & & & f(\lambda_n) \end{bmatrix} S^{-1}$$

The square root of B is simply obtained by taking the square roots of the eigenvalues on the diagonal of the diagonal matrix D. Since each eigenvalue has two square roots, the positive definite $n \times n$ matrix B has 2^n square roots, only one of which is positive definite.

10.4 Stability of Stationary Values

The objective function $f(x) = x^tAx$ has different stability properties in the neighborhood of each eigenvector $x(\alpha)$. Order the eigenvalues in a monotonic increasing way: $0 < \lambda_0 < \lambda_1 < \lambda_2 < \cdots < \lambda_{n-1}$. (**Beware:** Change in numbering!) In the neighborhood of the eigenvector $x(i)$, the function has critical value $c\lambda_i$ and is a Morse i-saddle. To show this, we observe that there are $n-1$ independent variables, since there are n variables and one constraint. We expand in the neighborhood of a critical point at $x(i)$ using the $n-1$ orthogonal eigenvectors:

$$X = rx(i) + \sum_{\alpha \neq i} a_\alpha x(\alpha) \tag{29}$$

The $n-1$ amplitudes a_α ($\alpha \neq i$) are the independent small coordinates of the objective function evaluated in the neighborhood of the critical point $x(i)$. The renormalization amplitude r has been introduced to preserve the normalization of the vector. Normalization requires

$$X^tBX = r^2 + \sum_{\alpha \neq i} a_\alpha^2 = c$$

In the neighborhood of the critical value $c\lambda(i)$ the shape of the function is

$$X^tAX = r^2\lambda_i + \sum_{\alpha \neq i} \lambda_\alpha a_\alpha^2 = c\lambda_i + c\sum_{\alpha \neq i} (\lambda_\alpha - \lambda_i)\ a_\alpha^2 \tag{30}$$

The quadratic form is a Morse i-saddle, as there are i $(= 0, 1, 2, \cdots)$ negative signs in this quadratic form: $\alpha = 0, 1, 2, \cdots, i-1$.

Example: For the matrices A and B

$$A = \begin{bmatrix} 2 & & \\ & 3 & \\ & & 5 \end{bmatrix} \qquad \text{and} \qquad B = \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix}$$

the eigenvalues and eigenvectors are

$$\begin{aligned} &\text{Eigenvalues}: \quad \lambda_0 = 2 \quad \lambda_1 = 3 \quad \lambda_2 = 5 \\ &\text{Eigenvectors}: \quad \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \end{aligned} \tag{31}$$

The stability properties associated with eigenvalue λ_1 are determined by perturbing around this eigenvector:

$$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \longrightarrow X = \begin{bmatrix} x \\ r \\ z \end{bmatrix}$$

The amplitudes x and z are assumed to be small (this is a *local* expansion), and the amplitude of the original eigenvector, r, is determined by $r^2 + x^2 + z^2 = 1$ ($c = 1$). The objective function in the neighborhood of the critical point $(0, 1, 0)^t$ is

$$X^t A X = 3r^2 + 2x^2 + 5z^2 = 3 + (2-3)x^2 + (5-3)z^2$$

In the neighborhood of λ_1 the quadratic form is locally a Morse 1-saddle.

11 Data Fitting

It happens often that we are called upon to find a "best fit" straight line to messy data. The old standby for this trial-by-fire is the Least Squares data fitting procedure. This works well when there are error measurements in the observed dependent variable (aka y) but not errors in the determination of the independent variable x.

This can be justified, to a very limited extent, when y is "determined" by x. However, it often happens that the two variables are *correlated*. In such cases both may possibly be determined by yet a third variable, not known. In such cases it is more than reasonable to assume that there are measurement errors in both x and y.

11.1 Measurement Errors: The Covariance Matrix

We will formulate the Least Squares fitting problem when measurement errors can be anticipated in all variables x^i. We assume that many measurements are made around a single data point whose "true value" $(x_0^1, x_0^2, \cdots, x_0^n)$ is not known but also doesn't vary from measurement to measurement. A series of measurements $x_{(\alpha)}$, $\alpha = 1, 2, \cdots, n \to \infty$ allows us to construct both an estimate for the coordinates of the "true" value

$$x_0^i = \lim_{n\to\infty} \frac{1}{n} \sum_{\alpha=1}^{n} x_{(\alpha)}^i \tag{32}$$

as well as a measurement covariance matrix based on this estimate

$$\langle \Delta x^i \Delta x^j \rangle = \lim_{n\to\infty} \frac{1}{n-1} \sum_{\alpha=1}^{n} ((x_{(\alpha)} - x_0)^i (x_{(\alpha)} - x_0)^j = \langle (x - x_0)^i (x - x_0)^j \rangle = g^{ij} \tag{33}$$

It will be assumed that the covariance matrix is unchanged throughout the region in which a linear model to describe the data is to be formulated.

Remark: The covariance matrix is actually a contravariant second order tensor. By rights it should be called a contravariance matrix.

11.2 Metric Matrix

Ultimately, we want to determine the distance between a point and a straight line. This means that we need to be able to measure distances in a plane. In short, we need to devise a metric tensor. Standard tensor calculus (covariance/contravariance) arguments suggest that g_{ij}, the inverse of the covariance matrix $g^{ij} = \langle \Delta x^i \Delta x^j \rangle$, is a suitable candidate. We adopt g_{ij} as the metric tensor in the measurement space. The square of the distance between two points $x_{(r)}$ and $x_{(s)}$ is

$$s^2 = g_{ij}(x_{(r)} - x_{(s)})^i (x_{(r)} - x_{(s)})^j$$

Our final result will depend on the covariance matrix g^{ij} and not its inverse, the metric tensor g_{ij}. The latter is simply a useful means to a "dimensionally correct" end.

11.3 Distance Between a Point and a Line

A line in a plane is determined by one constraint. More generally, an $n-k$ dimensional subspace in R^n is defined by k linear constraints. We illustrate how to compute the distance between a point and a linear subspace in the case that the subspace is defined by a single constraint. The generalization is straightforward.

The constraint is chosen in the form

$$A_i x^i + c = 0 \tag{34}$$

There are $n+1$ coefficients A_i and c in this constraint equation. The equation is scale-invariant, so some convenient relation can be placed on these coefficients at some later time.

The distance between a point $x_{(r)}$ and the surface $A_i x^i + c = 0$ is

$$s^2 = g_{ij}(x - x_{(r)})^i (x - x_{(r)})^j \tag{35}$$

This is minimized as usual. A Lagrange multiplier is used to introduce the constraint

$$s^2 = g_{ij}(x - x_{(r)})^i (x - x_{(r)})^j - \lambda(A_i x^i + c)$$

The equation is subtly rewritten for simplicity

$$s^2 = g_{ij}(x - x_{(r)})^i (x - x_{(r)})^j - \lambda(A_i (x - x_{(r)})^i + A_i x^i_{(r)} + c)$$

We introduce new variables $y^i = (x - x_{(r)})^i$ and $c_{(r)} = c + A_i x^i_{(r)}$ to rewrite this equation in simpler form

$$s^2 = g_{ij} y^i y^j + \lambda(A_i y^i + c_{(r)})$$

Proceeding in the usual fashion, it is a simple calculation to get the result

$$d_{(r)}^2 = \frac{c_{(r)}^2}{g^{ij}A_iA_j} \tag{36}$$

Here $d_{(r)}$ is the minimum distance of the measurement $x_{(r)}$ from the line $A_ix^i + c = 0$.

11.4 Least Squares

The line that minimizes the sum of the squares of the distances of the n observations $x_{(r)}$ from it is determined by minimizing the sum

$$\sum_{r=1}^{n} d_{(r)}^2 = \sum_{r=1}^{n} \frac{(A_ix_{(r)}^i + c)^2}{g^{ij}A_iA_j} \tag{37}$$

At this stage we can proceed by brute strength (ugly!) or elegantly. We choose the latter strategy.

- First, define the coordinates of the centroid of the measurements, so that

  $$\bar{x}^i = \frac{1}{n}\sum_{r=1}^{n} x_{(r)}^i \qquad\qquad i = 1, 2, \cdots, n$$

  Then we shift the origin of the coordinates to the centroid. In particular, this guarantees that the constant c is zero.

- Second, we use the observation that the linear constraint equation is scale-invariant to impose a nonlinear condition that scales the coefficients of the line. This is the quadratic constraint

  $$g^{ij}A_iA_j = 1$$

  This constraint will be imposed using Lagrange multipliers (a second time!).

The equation to be optimized is

$$F(x;\lambda) = \sum_{r=1}^{n}\left[A_i(x_{(r)}^i - \bar{x}^i)\right]^2 - \lambda\left(g^{ij}A_iA_j - 1\right) = \sum_{r=1}^{n}(x_{(r)} - \bar{x})^i(x_{(r)} - \bar{x})^jA_iA_j - \lambda\left(g^{ij}A_iA_j - 1\right) \tag{38}$$

Recalling that $g^{ij} = \langle \Delta x^i \Delta x^j \rangle$, the eigenvalue equation that results from searching for the stationary solutions of Eq(38) has the form

$$\left(\sum_{r=1}^{n}(x_{(r)} - \bar{x})^i(x_{(r)} - \bar{x})^j - \lambda\langle \Delta x^i \Delta x^j \rangle\right) A_j = 0 \tag{39}$$

The smallest eigenvalue λ_0 determines the eigenvector $A_i(0)$ that defines the line (hyperplane) through the origin (mean value of measurements) that minimizes the sum of the squared distances from the observations to that line. Since the (second) constraint is $+1$ ($g^{ij}A_iA_j = +1$), the eigenvalue *is* the sum of the squares of these distances.

11.5 Goodness of Fit Test

The eigenvalue λ_0 is actually the χ^2 value that should be used to test for goodness of fit. If there are n observations there are $n-2$ degrees of freedom (two points determine the line perfectly). Thus, if $\lambda_0 > \chi^2(n-2,p)$ the model can be rejected with confidence level p. Otherwise, the model cannot be rejected.

11.6 Elegant Formulation

These results can be put into elegant form by defining a second "measurement contravariance matrix" by

$$G^{ij} = \frac{1}{n}\sum_{r=1}^{n}(x_{(r)} - \bar{x})^i (x_{(r)} - \bar{x})^j \tag{40}$$

With these definitions the expression that determines both the best fit line and the statistic that determines how good the fit really is has the hard-to-forget form

$$\left(nG^{ij} - \lambda g^{ij}\right) A_j = 0 \tag{41}$$

Applications Involving Temperature

12 Equilibrium Thermodynamics

12.1 Formulation

Equilibrium Thermodynamics can be described in a number of different ways. In these sections we describe this subject in two different ways, one of which is more physical than the other. One involves minimizing energy subject to the constraint that entropy is fixed. The dual approach involves maximizing entropy subject to the constraint that energy is conserved. The second formulation is much more physical, since "energy is conserved, while entropy always increases." However, we will follow this approach in the following section, dealing with statistical mechanics. Therefore we adopt the less physical approach of minimizing energy subject to the constraint of fixed entropy. This is the traditional formulation of classical thermodynamics.

12.2 Fluctuations Around Equilibrium

We begin by placing two systems into contact with each other, and both inside a box. Nothing can get into or out of the box: not entropy (if you believe that ...), particles of any type, other extensive variables. The volume of the box is fixed. However, the two systems in the box can exchange entropy, volume, particles, other extensive variables.

The objective is to minimize the total energy, subject to the constraints that all other extensive variables are conserved. The energy is additive, so that we can write

$$U_{\rm tot} = U_{(1)}(S_{(1)}, V_{(1)}, N_{(1)}, E^{\alpha}_{(1)}) + U_{(2)}(S_{(2)}, V_{(2)}, N_{(2)}, E^{\alpha}_{(2)}) \tag{42}$$

Here $E^{\alpha}_{(1)}$ represents any other extensive variable required to specify the internal energy of the first system.

Exchanges of entropy, particle number, volume, etc. can take place between the two systems in the box. We assume that all the extensive quantities are conserved: whatever goes into one system comes out of the other. The fluctuation quantities obey, in an obvious notation:

$$\begin{array}{rcrcl} \delta S_{(1)} & + & \delta S_{(2)} & = & 0 \\ \delta V_{(1)} & + & \delta V_{(2)} & = & 0 \\ \delta N_{(1)} & + & \delta N_{(2)} & = & 0 \\ & & & \vdots & \\ \delta E^{\alpha}_{(1)} & + & \delta E^{\alpha}_{(2)} & = & 0 \end{array}$$

The energy minimization problem can now be formulated as follows. We allow the extensive arguments to fluctuate around their equilibrium values, and look for values of the fluctuations that minimize the internal energy. Specifically, the objective function is

$$\begin{array}{rcl} U_{\rm tot} & = & U_{(1)}(S_{(1)} + \delta S_{(1)}, V_{(1)} + \delta V_{(1)}, N_{(1)} + \delta N_{(1)}, E^{\alpha}_{(1)} + \delta E^{\alpha}_{(1)}) + \\ & & U_{(2)}(S_{(2)} + \delta S_{(2)}, V_{(2)} + \delta V_{(2)}, N_{(2)} + \delta N_{(2)}, E^{\alpha}_{(2)} + \delta E^{\alpha}_{(2)}) \\ & & -\lambda_S(\delta S_{(1)} + \delta S_{(2)} - 0) - \lambda_V(\delta V_{(1)} + \delta V_{(2)} - 0) \\ & & -\lambda_N(\delta N_{(1)} + \delta N_{(2)} - 0) - \cdots - \lambda_{\alpha}(\delta E^{\alpha}_{(1)} + \delta E^{\alpha}_{(2)} - 0) \end{array} \tag{43}$$

We proceed in the usual fashion. We note that $\partial U/\partial \delta S_{(1)} = \partial U/\partial S_{(1)}$, and so on for all the fluctuation quantitites. The following minimization conditions result:

$$\begin{array}{rcccl} \dfrac{\partial U_{(1)}}{\partial S_{(1)}} & = & \lambda_S & = & \dfrac{\partial U_{(2)}}{\partial S_{(2)}} \\ \dfrac{\partial U_{(1)}}{\partial V_{(1)}} & = & \lambda_V & = & \dfrac{\partial U_{(2)}}{\partial V_{(2)}} \\ \dfrac{\partial U_{(1)}}{\partial N_{(1)}} & = & \lambda_N & = & \dfrac{\partial U_{(2)}}{\partial N_{(2)}} \\ & & \vdots & & \\ \dfrac{\partial U_{(1)}}{\partial E^{\alpha}_{(1)}} & = & \lambda_{\alpha} & = & \dfrac{\partial U_{(2)}}{\partial E^{\alpha}_{(2)}} \end{array} \tag{44}$$

12.3 Lagrange Multipliers as Intensive Thermodynamic Variables

From this we conclude that under equilibrium conditions the "slopes" of the internal energy functions of the two systems must be the same for all intrinsic variables. The slopes are

$$\begin{array}{rcl} \dfrac{\partial U}{\partial S} & = & T \\ \dfrac{\partial U}{\partial V} & = & -P \\ \dfrac{\partial U}{\partial N} & = & \mu \\ \dfrac{\partial U}{\partial E^{\alpha}} & = & i_{\alpha} \end{array} \tag{45}$$

The derivative of the internal energy with respect to an extensive variable E^α is the conjugate intensive thermodynamic variable. These conjugate (Extensive, intensive) pairs of variables are $(S,T),(V,-P),(N,\mu),\cdots,(E^\alpha,i_\alpha)$, where T is the temperature, P the pressure, μ the chemical potential in the Energy representation. These identifications of intensive variables are made using the First and Second Laws of Thermodynamics:

$$dU = TdS - PdV + \mu dN + \cdots + i_\alpha dE^\alpha \tag{46}$$

and comparing the Lagrange multipliers with the partial derivatives of U: for example, $(\partial U/\partial S) = T$, etc. The partial derivatives are taken holding all other extensive thermodynamic variables constant.

12.4 Stability

The stability properties at an equilibrium are determined by expanding the potential function out to second order. We find

$$d^{(2)}U_{\rm tot} = \frac{\partial^2 U_{(1)}}{\partial E^\alpha_{(1)}\partial E^\beta_{(1)}}\delta E^\alpha_{(1)}\delta E^\beta_{(1)} + \frac{\partial^2 U_{(2)}}{\partial E^\alpha_{(2)}\partial E^\beta_{(2)}}\delta E^\alpha_{(2)}\delta E^\beta_{(2)} \tag{47}$$

Since $\delta E^\alpha_{(1)} = -\delta E^\alpha_{(2)} \stackrel{\rm def}{=} \delta E^\alpha$, the second variation can be written in the more streamlined form

$$d^{(2)}U_{\rm tot} = \left\{ \frac{\partial^2 U_{(1)}}{\partial E^\alpha_{(1)}\partial E^\beta_{(1)}} + \frac{\partial^2 U_{(2)}}{\partial E^\alpha_{(2)}\partial E^\beta_{(2)}} \right\} \delta E^\alpha \delta E^\beta \tag{48}$$

Both matrices within the curly brackets { } must be positive definite. This is a thermodynamic stability condition. The thermodynamic stability condition is the basis for all existing thermodynamic inequalities.

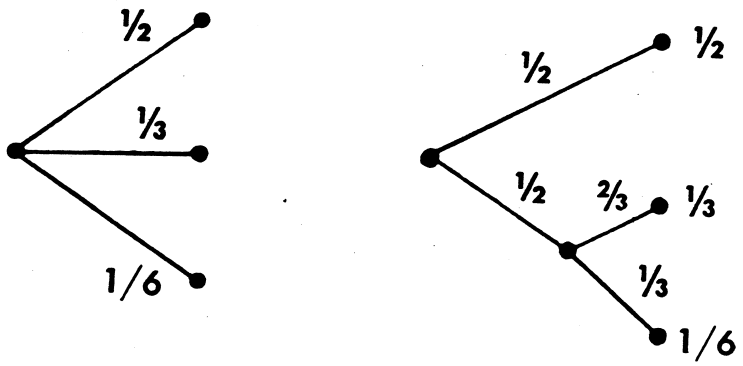
13 Statistical Mechanics

13.1 Shannon's Information Function

Shannon devised a measure of information for his study of the capacity of channels to transmit messages. He assumed that a finite number of states, $i = 1,2,\cdots,n$ were available to a system, and that state i could occur with probability p_i. He desired to construct an "information function," $H(p_1,p_2,\cdots,p_n)$ that measured the "information" in the system. This measures the amount of information required to gain perfect understanding: to move to a state where one probability is 1 and all the remaining probabilities are 0.

Shannon demanded that the function $H(p)$ satisfy three reasonable requirements:

1. H is a continuous function of all the probabilities.
2. If all the n probabilities are equal, $p_i = \frac{1}{n}$, H is a monotonic increasing function of n: $n' > n \Rightarrow H(n') > H(n)$.
3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H (H is subadditive).



$$H(\frac{1}{2},\frac{1}{3},\frac{1}{6}) = H(\frac{1}{2},\frac{1}{2}) + \frac{1}{2}H(\frac{2}{3},\frac{1}{3})$$

Figure 9: The function H obeys the subadditive property.

The last assumption is illustrated in Fig. 9. The only function that obeys these three properties is

$$H(p_1, p_2, \cdots, p_n) = -k \sum_{i=1}^{n} p_i \log p_i \qquad\qquad k > 0 \tag{49}$$

If the logarithms are taken base 2 and $k = 1$, information is measured in *binary digits*, or *bits*, a term due to Tukey. This is the preference of anyone working in information theory/cummunications theory. If natural logarithms are used and $k = k_B = 1.38 \times 10^{-16}$ erg/°K, H reduces to the statistical mechanical entropy function.

In the case of two states with $p_1 = p$ and $p_2 = q$, $p + q = 1$, Figure 10 provides a plot of $H(p, q)$. The maximum occurs at $p = q = \frac{1}{2}$ and is $H(\frac{1}{2}, \frac{1}{2}) = 1$ (bit).

13.2 Relation with Boltzmann's H-Function

Functions of the form $\sum p_i \log p_i$ had previously been studied. Boltzmann had introduced his "H-Function" to study the equilibration of a gas of atoms:

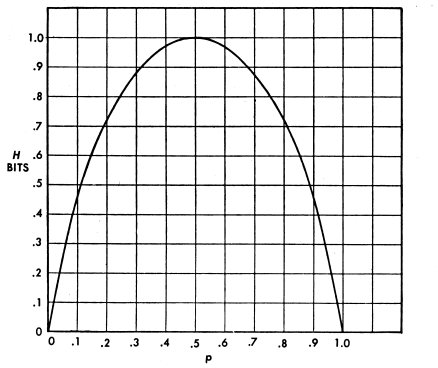
$$H(t) = \int d^3\mathbf{x} \int d^3\mathbf{v}\ f(\mathbf{x}, \mathbf{v}, t) \log f(\mathbf{x}, \mathbf{v}, t) \tag{50}$$

Here $f(\mathbf{x}, \mathbf{v}, t)$ is interpreted as the probability distribution function for a gas atom/molecule as a function of position $\mathbf{x}$, velocity $\mathbf{v}$, and time t. As a probability distribution, the function $f(\mathbf{x}, \mathbf{v}, t)$ satisfies

$$\int d^3\mathbf{x} \int d^3\mathbf{v}\ f(\mathbf{x}, \mathbf{v}, t) = 1 \tag{51}$$

As such, $f(\mathbf{x}, \mathbf{v}, t)$ has an interpretation as p_i above, with i now a continuous index. Boltzmann was able to "prove" the "Boltzmann H-Theorem":

$$\frac{dH(t)}{dt} \leq 0$$



$$H(p,q) = -(p\log_2(p) + q\log_2(q))$$

Figure 10: The function H has a maximum when all the probabilities are equal.

If Boltzmann's function were defined with a negative sign, this result would be interpreted as a "proof" that entropy is always increasing. This is not possible: entropy increases to the most probable state, and thereafter fluctuations from this most probable state act to (momentarily) decrease entropy. It is impossible for monotonic increase of entropy to be compatible with fluctuations around equilibrium.

13.3 Application to Statistical Mechanics

The most likely state of a physical system can be determined by maximizing the entropy $S = -k\sum p_i \log p_i$ subject to constraints. Four typical constraints are:

$$\begin{array}{llll} \sum & p_i & = & 1 \\ \sum & p_i \; E_i & = & \bar{E} \\ \sum & p_i \; V_i & = & \bar{V} \\ \sum & p_i \; N_i & = & \bar{N} \end{array} \tag{52}$$

The first constraint is that the probabilities p_i must behave like probabilities. The second, third, and fourth are of physical origin. The first of these assumes the average energy $\bar{E}$ is known. The other two assume that the average volume is $\bar{V}$ and the average particle number is $\bar{N}$. For example, if two systems are separated by a flexible membrane, different states of one system, the one of interest, could have different volumes, while the long term average volume is known.

These constraints on the entropy maximization problem are imposed in the canonical way. A modified objective function is established:

$$S = -k\sum p_i \ln p_i - (\lambda_0 - k)\left(\sum p_i - 1\right) - \lambda_E\left(\sum p_i E_i - \bar{E}\right) - \lambda_V\left(\sum p_i V_i - \bar{V}\right) - \lambda_N\left(\sum p_i N_i - \bar{N}\right) \tag{53}$$

A constant term $-k$ is added to the Lagrange multiplier λ_0 for later convenience. This expression is differentiated with respect to each p_i to obtain

$$\frac{\partial S}{\partial p_i} = -k(\ln(p_i) + 1) - (\lambda_0 - k) - \lambda_E E_i - \lambda_V V_i - \lambda_N N_i = 0 \tag{54}$$

This gives immediately an expression for $k \ln p_i$:

$$k \ln p_i = -\lambda_0 - \lambda_E E_i - \lambda_V V_i - \lambda_N N_i \tag{55}$$

From this the probabilities are immediate:

$$p_i = e^{-\lambda_0/k} \; e^{-(\lambda_E E_i + \lambda_V V_i + \lambda_N N_i)/k} \tag{56}$$

13.4 Partition Functions

The Lagrange multiplier λ_0 is a normalization constant, and may be determined by summing all the possibilities:

$$e^{\lambda_0/k} = \sum_i e^{-(\lambda_E E_i + \lambda_V V_i + \lambda_N N_i)/k} = \mathcal{Z}(\lambda_E, \lambda_V, \lambda_N) \tag{57}$$

The function $\mathcal{Z}$ resulting from the sum is called the *partition function*. The partition function can be used as a *generating function* for expectation values. It is a function of the Lagrange multipliers. For example, if we differentiate the (natural) logarithm of $\mathcal{Z}$ with respect to one of its arguments, we obtain

$$\frac{\partial \ln(\mathcal{Z})}{\partial \lambda_E} = \frac{\sum_i (-E_i/k) e^{-(\lambda_E E_i + \lambda_V V_i + \lambda_N N_i)/k}}{\sum_i e^{-(\lambda_E E_i + \lambda_V V_i + \lambda_N N_i)/k}} = -\frac{1}{k} \sum_i E_i \frac{e^{-(\lambda_E E_i + \lambda_V V_i + \lambda_N N_i)/k}}{\mathcal{Z}} = -\frac{1}{k} \sum_i p_i E_i = -\frac{\bar{E}}{k} \tag{58}$$

In simpler form, this is

$$-k \frac{\partial \ln(\mathcal{Z})}{\partial \lambda_E} = \bar{E} \tag{59}$$

Similar results hold for all the other Lagrange multipliers.

Since $\ln(\mathcal{Z}) = \lambda_0/k$, this last result becomes

$$-\frac{\partial \lambda_0}{\partial \lambda_E} = \bar{E} \tag{60}$$

These results lead to a very nice interpretation of the Lagrange multipliers.

13.5 Interpretation of the Lagrange Multipliers

The entropy function S is a function of the values of the constraints: $S = S(\bar{E}, \bar{V}, \bar{N})$. The differential of S can be expressed in terms of the differentials of the expectation values, each multiplied by the conjugate Lagrange multiplier, as shown in Section 5. We find immediately

$$dS = \lambda_E d\bar{E} + \lambda_V d\bar{V} + \lambda_N d\bar{N} \tag{61}$$

This can be compared with the equilibrium thermodynamic expression for the First and Second Laws of Thermodynamics (c.f., Equ (46))

$$dS = \frac{1}{T} d\bar{E} + \frac{P}{T} d\bar{V} - \frac{\mu}{T} d\bar{N} \tag{62}$$

to give an immediate interpretation of the Lagrange multipliers in the Entropy representation:

$$\begin{aligned} \lambda_E &= \frac{1}{T} \\ \lambda_V &= \frac{P}{T} \\ \lambda_N &= -\frac{\mu}{T} \\ \lambda_0 &= k \ln\left(\mathcal{Z}(\frac{1}{T}, \frac{P}{T}, \frac{-\mu}{T})\right) \end{aligned} \tag{63}$$

Remark: The Lagrange multipliers in (45) and (63) differ. In both cases they are intensive thermodynamic variables conjugate to extensive thermodynamic variables. However, they are conjugate in two different representations. In (45) they are conjugate in the Energy representation (46). In (63) they are conjugate in the Entropy representation (62). In either case the Second Law of Thermodynamics (46,62) is satisfied.

13.6 "Entropy of a State"

The expectation values of E, V, N, and the most probable value of the entropy, are

$$\begin{array}{llll} \sum_i & p_i & E_i & = \bar{E} \\ \sum_i & p_i & V_i & = \bar{V} \\ \sum_i & p_i & N_i & = \bar{N} \\ \sum_i & p_i & (-k\ln(p_i)) & = S_{\text{most likely}} \end{array} \tag{64}$$

In view of these analogies, it is not at all unreasonable to regard $-k\ln(p_i)$ as the "entropy" of the i^{th} state. We adopt this convention. Then from the differential relation (53) defining the constrained probabilities, we find

$$-k\ln(p_i) = S_i = \frac{1}{T}E_i + \frac{P}{T}V_i - \frac{\mu}{T}N_i \tag{65}$$

As a result, we can write the probability p_i in the unforgettable way as

$$p_i \simeq e^{-S_i/k} \tag{66}$$

The probabilities are normalized in the usual way

$$p_i = \frac{e^{-S_i/k}}{\mathcal{Z}} = e^{-S_i/k - \ln(\mathcal{Z})} \tag{67}$$

13.7 Quantum Statistical Mechanics

Relatively little has to be done to extend the results above, developed for classical statistical mechanics, into the realm of quantum statistical mechanics. The differences are

1. The probabilities p_i are replaced by a density operator $\hat{\rho}$.
2. Sums over probabilities are replaced by the traces of the product of the density operator with an operator representing the appropriate observable: $\sum_i p_i E_i \to \text{Tr}\ \hat{\rho}\hat{H}$, where $\hat{H}$ is the Hamiltonian (energy operator).

3. An entropy operator $\hat{S}$ is defined by

$$\hat{S} = -k \ln \hat{\rho} = \frac{1}{T}\hat{H} + \frac{P}{T}\hat{V} - \frac{\mu}{T}\hat{N} \tag{68}$$

4. The generalization of (66) is

$$\hat{\rho} = \frac{e^{-\hat{S}}}{\mathcal{Z}} = e^{-(\frac{1}{T}\hat{H} + \frac{P}{T}\hat{V} - \frac{\mu}{T}\hat{N})/k - \ln(\mathcal{Z})} \tag{69}$$

Remark: In the case that only the expectation of the energy is known, the density operator reduces to

$$\hat{\rho} = \frac{e^{-\hat{S}/k}}{\mathcal{Z}} = e^{-(\frac{1}{T}\hat{H})/k} / \mathrm{Tr}\ e^{-(\frac{1}{T}\hat{H})/k} \simeq e^{-\beta\hat{H}}$$

where $\beta = 1/kT$. In case $\hat{H}$ is diagonal, these are the usual Boltzmann factors.