1. Parameter Optimization Problems:

1.2. Equality Constrained Optimization

©Erik I. Verriest*

November 19, 2008

1 Parameter Optimization with Equality Constraints

Consider the standard constraint minimization problem

$$\min L(y) \;\; \text{such that} \;\; f(y) = c$$

Let it be given that the scalar functions f and L are differentiable (hence automatically continuous). Assume that the parameter y lives in a parameter space Θ, whose dimension, n, may be infinity, but is at least equal to 2, in order to avoid the trivial case. Let also $f : \mathcal{D} \subset \mathbf{R}^n \to \mathbf{R}^m$. We shall refer to this problem as *Problem EC* (for Equality Constraint). Assume that the problem has a solution, for $y = y^*$. The corresponding value of L, namely $V = L(y^*)$ will be called the *value* of the minimization problem.

1.1 The Problem EC: Necessary Conditions

We analyzed the *unconstrained* parameter optimization problem:

$$\min L(y) \;\; \text{with} \;\; y \in \mathcal{D}$$

where $\mathcal{D}$ is a compact (i.e. closed and bounded) set in the parameter space $\mathbf{R}^n$. If L is differentiable, we know by Weierstrass's theorem that the minimum of L exists, and either is achieved in the interior Int $\mathcal{D}$ or on the boundary $\partial\mathcal{D}$. A necessary condition for a point $y^* \in \mathrm{Int}\,\mathcal{D}$ to be a minimum is that for a perturbation in *any* arbitrary direction dy, one has

$$dL = \frac{\partial L}{\partial y} dy = 0. \tag{1}$$

*School of ECE, Georgia Tech, Atlanta, GA, USA. Presently with SCD, ESAT, KULeuven.

But this can only be true for all infinitesimal dy if $\frac{\partial L}{\partial y} = 0$. It follows that an equivalent condition is $\frac{\partial L}{\partial y} = 0$. For the problem EC this is no longer the case. This is so because now the perturbations are no longer completely arbitrary. Only a perturbation in a direction that leaves f invariant is allowed. But the manifold $f(y) = c$ has dimension $n - m$, where m is the number of independent constraints, $f_i(y) = c_i$, constituting the vector constraint equation $f(y) = c$.

We invoke now a very fundamental theorem from multivariate analysis: The *Implicit Function Theorem*, stating in its elementary form that under some mild conditions any implicit function, $f(y_1, \ldots, y_n) = 0$, in n variables can be written in a form where one of the variables is an explicit function of the $n - 1$ remaining variables. Before giving the precise statement, some definitions are in order.

Definitions:
Let $\mathcal{D}$ be an *open* set in $\mathbf{R}^n$.

- A map $f : \mathcal{D} \subset \mathbf{R}^n \to \mathbf{R}^m$ is *differentiable at a point* $y_0 \in \mathcal{D}$ if the partial derivatives $\frac{\partial f_i}{\partial y_j}$ evaluated at the point y_0 exist for all $i = 1, \ldots, m$ and $j = 1, \ldots n$.
- A map $f : \mathcal{D} \subset \mathbf{R}^n \to \mathbf{R}^m$ is called *differentiable* in $\mathcal{D}$ if it is differentiable at every point in $\mathcal{D}$.
- The map f : open $U \subset \mathbf{R}^n \to$ open $V \subset \mathbf{R}^m$ is called a *diffeomorphism*, if it is bijective (one-to-one and onto), and both f and f^{-1} are differentiable in the appropriate domains.
- The *Jacobian matrix* of f at $y_0 \in \mathcal{D}$ is the $m \times n$ matrix whose ij-th entry is $\frac{\partial f_i}{\partial y_j}$ evaluated at the point y_0. Consistent with our notation, the Jacobian is the gradient matrix $\frac{\partial f}{\partial y}$, which we shall also denote by the simple form Df (the latter notation is standard in the theory of functions in several variables and differential geometry).
- The *rank* of a *differentiable* map $f : \mathcal{D} \subset \mathbf{R}^n$ at a point $y_0 \in \mathcal{D}$ is the rank of the Jacobian matrix Df evaluated at the point y_0.
- The differentiable map $f : \mathcal{D} \subset \mathbf{R}^n \to \mathbf{R}^m$ is said to have constant rank in $U \in \mathcal{D}$, if for every point $y \in U$, the rank is k, for some $k \leq \min(m, n)$.

Inverse Function Theorem
Let f be a differentiable map from an open domain $\mathcal{D}$ in $\mathbf{R}^n$ to $\mathbf{R}^n$. If the Jacobian of f at y_0 is nonsingular, then there exists an open neighborhood U of y_0 in $\mathcal{D}$ such that $V = f(U)$ is open in $\mathbf{R}^n$ and the restriction of f to U is a diffeomorfism onto V.

Rank Theorem
Let f be a differentiable map from an open domain $\mathcal{A} \subset \mathbf{R}^n$ to $\mathcal{B} \subset \mathbf{R}^m$. If the rank of f in $\mathcal{A}$ is k, then there exists for each point $y_0 \in \mathcal{A}$ a neighborhood $\mathcal{A}_0$ of y_0 in $\mathcal{A}$, a neighborhood $\mathcal{B}_0$ of $f(y_0)$ in $\mathcal{B}$, two open sets $U \subset \mathbf{R}^n, V \subset \mathbf{R}^m$, and two diffeomorphisms $g : U \to \mathcal{A}_0$ and

$h : \mathcal{B}_0 \to V$ such that for all $y \in U$, $(h \circ f \circ g)(y) = [y_1, \ldots, y_k, 0, \ldots, 0]'$.

Implicit Function Theorem
Let $\mathcal{A} \subset \mathbf{R}^m$ and $\mathcal{B} \subset \mathbf{R}^p$ be open sets. Let $f : \mathcal{A} \times \mathcal{B} \to \mathbf{R}^m$ be a differentiable map such that the Jacobian $\frac{\partial f}{\partial x}$ at the point $y_0' = (x_0', u_0') \in \mathcal{A} \times \mathcal{B}$ is nonsingular, and $f([x_0', u_0']') = 0$, then there exist neighborhoods V of x_0 in $\mathcal{A}$, and U of u_0 in $\mathcal{B}$ and a unique *differentiable map $g : U \to V$ such that*

$$f(g(u), u) = 0 \qquad , \qquad \forall\, u \in U.$$

For the proofs of these theorems, we refer to Boothby, Dieudonné or Spivak.

Let us now resume the problem EC. First of all notice that if $m \geq n$ we do not have an optimization problem anymore, so we shall always assume that $n > m$. Let us also assume that f has full rank in the domain of interest, if not we could have reduced the number of constraints until we end up with a set of independent constraints. Without loss of generality (e.g. by proper relabeling of variables), we can assume then that the first m columns of the Jacobian Df constitute a nonsingular submatrix. Let the corresponding parameters be denoted by x, and the remaining ones by u, thus $x_i = y_i$, for $i = 1, \ldots, n-m$; and $u_j = y_{m+j}$ for $j = 1, \ldots, n-m$. With some abuse of notation, let us rewrite the constraint as $f(x, u) = 0$. The *necessary* condition for minimality is that $dL = 0$ for arbitrary du, *while holding* $df = 0$. Thus,

$$dL = L_x dx + L_u du = 0 \tag{2}$$

for all u that satisfy

$$df = f_x dx + f_u du = 0. \tag{3}$$

Since f_x is by assumption invertible, u determines x, i.e., $dx = -f_x^{-1} f_u du$, from which also

$$dL = (L_u - L_x f_x^{-1} f_u) du \tag{4}$$

follows. In this expression the perturbation du is *completely arbitrary*. Thus this is as in the unconstrained problem. If the above is to hold for all du, the equivalent condition is that the term within brackets vanishes. Define now

$$\lambda = -L_x f_x^{-1},$$

the *condition* for optimality is then

$$L_u + \lambda f_u = 0, \tag{5}$$

while the *definition* of λ can be restated as

$$L_x + \lambda f_x = 0. \tag{6}$$

Together (5) and (6) can be restated as the following

1.2 Method of Lagrange Multipliers

Define a *Hamiltonian*: $H(y, \lambda) = L(y) + \lambda f(y)$, containing $n+m$ parameters (n components of y, and m Lagange multipliers λ). Consider now the *unconstrained* optimization problem of H over the parameters y and λ. The necessary conditions are $[H_y, H_\lambda] = 0$, from which thus $H_y = 0$ yields the necessary conditions (5) and (6), and $H_\lambda = 0$ yields precisely the constraint equations $f = 0$.

The Lagrange multiplier technique 'reduces' the constrained optimization problem to an unconstrained optimization problem, albeit of larger dimension.

Another approach to the same problem can be obtained as follows:
Consider again

$$\begin{aligned} dL &= L_x dx + L_u du \\ df &= f_x dx + f_u du. \end{aligned}$$

In matrix form, the necessary conditions for an extremum of the constrained problem are

$$\begin{bmatrix} L_x & L_u \\ f_x & f_u \end{bmatrix} \begin{bmatrix} dx \\ du \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{7}$$

Consistency requires that the rank of the above matrix is less than $n+1$, otherwise the only solution is $du = 0$ and $dx = 0$. Thus, there must be linear combinations of the rows adding to zero. Since *row*-operations are performed by left multiplications, there exists a vector $[1, \lambda]$ such that

$$[1, \lambda] \begin{bmatrix} L_y \\ f_y \end{bmatrix} = 0,$$

which gives the usual form

$$L_y + \lambda f_y = 0.$$

This expresses the fact that *at the extremum, the gradient of the performance index has to lie in the subspace spanned by the gradients of the constraints.* The Lagrange multipliers are then exactly the coefficients in this linear combination.

In the following subsection, we will derive another interpretation of the Lagrange multipliers.

1.3 Embedding of the Problem EC

Consider the problem EC, but with the constraint equation now parametrized by $\omega \in \Omega$ This creates a *family* of EC problems, which we shall refer to as EC_Ω. For each $\omega \in \Omega$, there will be a solution of the EC-optimization problem EC_ω. The corresponding optimizing point and value will then be parametrized by this ω as well, i.e., we get $y^*(\omega)$ and $V(\omega) = L(y^*(\omega))$.

We shall refer to this $V(\omega)$ as the *value function*. A central question is now: "What nice properties does $V(\omega)$ have as function of ω?" Typical questions are:

1. Is $V(\omega)$ *finite* in the neighborhood of some ω_0?
2. Is $V(\omega)$ *continuous* near ω_0?
3. Is $V(\omega)$ *differentiable* near ω_0?

Note that even if we are only interested in one particular optimization problem, say EC_{ω_0}, it still is of interest to consider a family in the neighborhood of ω_0 since each of the above cited properties deals with a particular aspect of the practical solution of EC_{ω_0}.

The first has to with *solvability*, or, in the appropriate context *controllability*. Indeed, if $f(y, \omega_0) = 0$ has no solution, then the admissible set of parameters is empty. The usual convention is to define the infimum of L over the empty set as ∞. (If there does not exist a control such that $x(t_f) = 0$, then the system is uncontrollable. Any finite amount of control energy (however measured) will not do the job, hence the minimal control energy or cost must be infinite.)

The second has to do with *well-posedness* of the solution. The continuity of V near ω_0 is a necessary statement to display a "nice" dependence on the parameters of the problem. Different types of continuity can be defined. For instance Lipshitz continuity, $\mid V(\omega) - V(\omega_0) \mid \leq K \mid \omega - \omega_0 \mid$, is a *stability* property that is useful in error estimates if not EC_{ω_0} but EC_{ω} is to be solved, but only $V(\omega_0)$ is evaluated.

The third property leads to *asymptotic analysis* or *perturbation methods*. If EC_{ω_0} has a simple solution, but EC_{ω} does not, then if differentiability holds, we can evaluate

$$V(\omega) = V(\omega_0) + <(\omega - \omega_0), V_{\omega}(\omega_0)> + o(\omega - \omega_0). \tag{8}$$

The bracketts denote the inner product of $(\omega - \omega_0)$ with the gradient of V with respect to ω for parameter ω_0.
Other appropriate and nice properties one may want to impart on the value function are discussed by Clarke.

1.4 The Value Function

We will now look at the *value function* $V(\omega)$ as function of ω. To fix on the ideas without unnecessary detail, we will assume that there is only one scalar constraint $f = c$, so $m = 1$, and the constraint parameter $\omega = c$. We define the *image set* as the set of pairs $\{(f(y), L(y)) \mid y \in \Theta\}$. Note that this set lies in $\mathbf{R}^2$, even if Θ is infinite dimensional. Since $L(y)$ and $f(y)$ are both continuous, the image set is a smooth curve in $\mathbf{R}^2$.

The value of $V(c)$ corresponds to the ordinate of the *lowest* point on the image set and the vertical $f(y) = c$. It is clear from figure 1 that the value function will be nondifferentiable with respect to c.

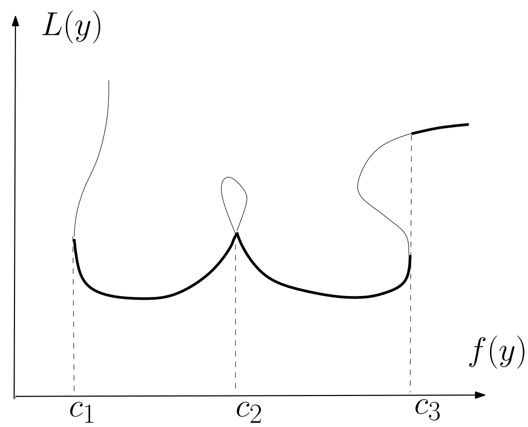

Figure 1: Illustrating the Value Function

In the open interval (c_1, c_2) the value function is continuous and differentiable. At c_1, the value is finite, but to the left of c_1, the value is ∞, since there are no points in Θ for which $f(y) = c$. At c_2, the value function $V(c)$ is continuous, but not differentiable: It has a "corner". In (c_2, c_3) the $V(c)$ is again differentiable, whereas at c_3 it is discontinuous, while being finite locally.

Thus, we conclude from this example that in order to study the differential properties of V we must either find and impose a priori conditions that will imply the smoothness of V at a given point, or else deal with and confront the nondifferentiability of it. Recently, new techniques have been developed to do exactly that.

Reference: Frank H. Clarke, *Methods of Dynamic and Nonsmooth Optimization*, SIAM CBMS-NSF No. 57, 1989.

Problem: Try to find a simple example, suitable for class presentation of a constraint optimization problem illustrating such nondifferentiable behaviors of the value function. Can you find an example with a value function that is *nowhere* differentiable?

If $V(c)$ is a value function for the above problem, then we must have, for any y:

$$V(f(y)) \leq L(y). \tag{9}$$

If y^* solves the optimization problem, then $V(f(y^*)) = L(y^*)$. Combining both observations, it follows that y^* is the unconstrained optimizer for the function $H : y \to L(y) - V(f(y))$

A necessary condition is therefore that at y^*, we have $\frac{dH(y)}{dy} = 0$, i.e.,

$$L_y - V'(c))f_y = 0. \tag{10}$$

We recognize the Lagrange multiplier as $\lambda = -V'(c)$.

This gives a nice re-interpretation of the Lagrange multiplier, but it does not provide a "proof" for the method, since it has a fatal flaw: We have assumed the differentiability of $V(c)$, and we know now better than to do that!

We will next derive more rigorously this second interpretation of the Lagrange multipliers. Consider again a family of minimization problems EC_Ω. The value function V is, as seen before, parametrized by $\omega \in \Omega$. We shall proceed *assuming that the value function is differentiable*, just to interpret the Lagrange multiplier λ, but heed the warnings we made about this. We have in general:

$$df = f_x\, dx + f_u\, du + f_\omega\, d\omega \tag{11}$$

and just as before, $dL = L_x\, dx + L_u\, du$. Note that the increment in the constraint due to a change of the parameter is $df|_{(u,x)} = f_\omega\, d\omega$. The notation $df|_{(x,u)}$ emphasizes that the differential is computed *keeping u and x constant*, thus $du = 0$ and $dx = 0$. At each neighboring ω we need to fulfill the constraint for this ω, that is $f(x, u, \omega) \equiv 0$, since to compute the optimal solution at each ω, the ω is kept constant. Thus we have, for fixed $d\varpi$, the constraint:

$$dx = -(f_x)^{-1} f_u\, du + (f_x)^{-1}\, f_\omega\, d\varpi. \tag{12}$$

Substituted in the expression for $dV = dL\,|_{\text{opt}}$ this gives

$$\begin{aligned} dV &= L_u\, du - L_x (f_x)^{-1} f_u\, du - L_x (f_x)^{-1}\, f_\omega\, d\varpi \\ &= [L_u + \lambda\,|_{\omega_0}] du + \lambda\,|_{\omega_0}\; f_\omega\, d\varpi \end{aligned} \tag{13}$$

The bracketted term on the right hand side was set to zero to find the optimal $u^*_{\omega_0}$ and $x^*_{\omega_0}$. So it remains that

$$dV(\omega_0) = dL^*(\omega_0) = \lambda\,|_{\omega_0}\; f_\omega\, d\varpi, \tag{14}$$

hence

$$\left.\frac{dL^*}{df}\right|_{x,u} = \lambda\,|_{\omega_0}\ . \tag{15}$$

If the value function is differentiable, the Lagrange multiplier gives the *influence* or *sensitivity* of the value *with respect to a change in the constraint.* For this reason the Lagrange multipliers are sometimes referred to as the sensitivity or influence functions (they are functions of ω.)

2 Sufficient Conditions for Constrained Optimality

Up to second order, one has

$$dL = [L_x, L_u] \begin{bmatrix} dx \\ du \end{bmatrix} + \frac{1}{2}[dx', du'] \begin{bmatrix} L_{xx} & L_{xu} \\ L_{ux} & L_{uu} \end{bmatrix} \begin{bmatrix} dx \\ du \end{bmatrix} + \cdots , \tag{16}$$

$$df = [f_x, f_u] \begin{bmatrix} dx \\ du \end{bmatrix} + \frac{1}{2}[dx', du'] \begin{bmatrix} f_{xx} & f_{xu} \\ f_{ux} & f_{uu} \end{bmatrix} \begin{bmatrix} dx \\ du \end{bmatrix} + \cdots . \tag{17}$$

Combining, and evaluating at the extremal point:

$$\begin{aligned} dL + \sum_{i=1}^{n} \lambda_i \, df^{(i)} &= \frac{1}{2}[dx', du'] \begin{bmatrix} L_{xx} + \sum_{i=1}^{n} \lambda_i f_{xx}^{(i)} & L_{xu} + \sum_{i=1}^{n} \lambda_i f_{xu}^{(i)} \\ L_{ux} + \sum_{i=1}^{n} \lambda_i f_{xu}^{(i)} & L_{uu} + \sum_{i=1}^{n} \lambda_i f_{uu}^{(i)} \end{bmatrix} \begin{bmatrix} dx \\ du \end{bmatrix} + \cdots \\ &= \frac{1}{2}[dx', du'] \begin{bmatrix} H_{xx} & H_{xu} \\ H_{ux} & H_{uu} \end{bmatrix} \begin{bmatrix} dx \\ du \end{bmatrix} + \cdots \end{aligned} \tag{18}$$

Thus sufficient conditions for a *local minimum* are the stationarity conditions:

$$f(x, u) = 0 \tag{19}$$

$$\frac{\partial H}{\partial x} = 0 \tag{20}$$

$$\frac{\partial H}{\partial u} = 0, \tag{21}$$

together with the the *positive definiteness* condition of the matrix

$$\begin{bmatrix} H_{xx} & H_{xu} \\ H_{ux} & H_{uu} \end{bmatrix} > 0. \tag{22}$$

Note that if we keep the constraint satisfied, i.e., $dx = -f_x^{-1} f_u \, du$ we get

$$\begin{aligned} dL &= \frac{1}{2} du' [-f_u' f_x^{-T}, I] \begin{bmatrix} H_{xx} & H_{xu} \\ H_{ux} & H_{uu} \end{bmatrix} \begin{bmatrix} -f_x^{-1} f_u \\ I \end{bmatrix} du + \cdots , \\ &= \frac{1}{2} du' \left[\frac{\partial^2 L}{\partial u^2} \right]_{f=0} du + \cdots , \end{aligned} \tag{23}$$

where

$$\left. \frac{\partial^2 L}{\partial u^2} \right]_{f=0} = H_{uu} - f_u' f_x^{-T} H_{xu} - H_{ux} f_x^{-1} f_u + f_u' f_x^{-T} H_{xx} f_x^{-1} f_u.$$

The sufficient conditions for a local minimum may thus be expressed as:

$$\begin{aligned} H_u &= 0 \\ H_x &= 0 \\ H_\lambda &= 0 \end{aligned} \qquad \frac{\partial^2 L}{\partial u^2}_{f=0} > 0. \tag{24}$$

3 Neighboring Solutions

Suppose that the minimum solution to $L(y)$ under the constraint $f(y) = 0$ was found, say y^0. Now we perturb the constraint to $f(y) = df$, for small df. The new optimum solution is located at y^*. We want now to relate y^* to y^0.

The new Hamiltonian is $H^*(x, u, \lambda) = L(x, u) + \lambda' f(x, u) - df$, with the new optimum

$$\begin{aligned} x^* &= x^0 + dx && (25) \\ u^* &= u^0 + du && (26) \\ \lambda^* &= \lambda^0 + d\lambda. && (27) \end{aligned}$$

Obviously, the stationarity conditions still need to be satisfied, thus

$$\left.\begin{aligned} H_x^* &= 0 \\ H_u^* &= 0 \\ H_\lambda^* &= 0 \end{aligned}\right\} \quad \text{for} \quad (x, u, \lambda) = (x^*, u^*, \lambda^*). \tag{28}$$

If $L(\cdot)$ and $f(\cdot)$ are sufficiently smooth, then for small df one expects small dx, du and $d\lambda$, hence using a Taylor expansion about (x^0, u^0, λ^0),

$$\begin{aligned} H_x^* &= H_x^0 + (H_{xx}^0 dx)' + (H_{xu}^0 du)' + (f_x^0 d\lambda)' = 0 && (29) \\ H_u^* &= H_u^0 + (H_{ux}^0 dx)' + (H_{uu}^0 du)' + (f_u^0 d\lambda)' = 0 && (30) \\ H_\lambda^* &= H_\lambda^0 + [f_x^0 dx + f_u^0 du]' = 0 && (31) \end{aligned}$$

Noting that H_x^0, H_u^0 and H_λ^0 are all necessarily zero, since (x^0, u^0, λ^0) is an extremum, and $df = [f_x^0 dx + f_u^0 du]$. Thus:

$$\begin{aligned} dH_x^0 &= H_{xx}^0 dx + H_{xu}^0 du + f_x^0 d\lambda = 0 && (32) \\ dH_u^0 &= H_{ux}^0 dx + H_{uu}^0 du + f_u^0 d\lambda = 0 && (33) \\ df &= f_x^0 dx + f_u^0 du. && (34) \end{aligned}$$

From which:

$$dx = (f_x^0)^{-1} df - (f_x^0)^{-1} f_u^0 du, \tag{35}$$

and

$$d\lambda = -(f_x^0)^{-1}(H_{xx}^0 dx + H_{xu}^0 du). \tag{36}$$

Substituting in the second equation:

$$\begin{aligned} &(H_{ux}^0 (f_x^0)^{-1} - f_u^{0'} (f_x^0)^{-T} H_{xx}^0 (f_x^0)^{-1}) df + && (37) \\ +(H_{uu}^0 - H_{ux}^0 (f_x^0)^{-1} f_u^0 - f_u^{0'} (f_x^0)^{-T} H_{xx}^0 (f_x^0)^{-1} f_u^0) du &= 0 && (38) \end{aligned}$$

or:

$$du = -\left(\frac{\partial^2 L}{\partial u^2}\right)_{f=0}^{-1} (H_{ux}^0 - f_u^{0'} (f_x^0)^{-T} H_{xx}^0)(f_x^0)^{-1} df. \tag{39}$$

Hence, knowing the optimum (u^0, x^0) for the constraint $f = 0$, we can compute directly the neighboring optimum for the perturbed constraint $f = df$, without computing the optimum by resolving the equations $H_x^* = 0, H_u^* = 0, H_\lambda^* = 0$. All that is needed is the evaluation of dx and du at the *nominal* point (u^0, x^0), then

$$x^* = x^0 + dx \tag{40}$$

$$u^* = u^0 + du \tag{41}$$

Notice again that we assumed the differentiablility of the value function in this analysis. Furthermore, writing dx and du in terms of df yields:

$$dL = -\lambda^{0'} df + \frac{1}{2} df' M^0 df + \cdots , \tag{42}$$

where M^0 is positive definite if the Hessian of H is. This last formula expresses again the change in L from u^0 to u^* due to a change in the constraint of df. It is in fact the change in the *value function*. Thus, once again $\frac{\partial L_{\min}}{\partial f} = -\lambda'$. The Lagrange multiplier expresses the sensitivity of $L_{\min}$ with respect to the constraint.