Pictures at an Exhibition – A Lagrange Multiplier Gallery

Vivek S Borkar
Department of Computer Science and Automation
Indian Institute of Science
Bangalore 560 012, India

This article gives several pictorial interpretations of the Lagrange multiplier rule (a useful distraction in this age of divide and rule).

Go Forth and Multiply

While teaching optimization theory to engineering students, one quickly realizes that a picture is worth more than a thousand words (unless the words are in C++ or something). An important instance of this is the Lagrange multiplier rule. It is usually taught 'cook book' style: If you want to minimize a real-valued function f on R^d subject to the constraints $g_i(x) \leq 0, 1 \leq i \leq m$, and $h_j(x) = 0, 1 \leq j \leq k$, perform instead the unconstrained minimization of $f(\cdot) + \sum_{i=1}^{m} \lambda_i g_i(\cdot) + \sum_{j=1}^{k} \mu_j h_j(\cdot)$, where λ_i, μ_j (with $\lambda_i \geq 0$) are the Lagrange multipliers. The point x^* where the minimum is attained then satisfies

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) + \sum_{j=1}^{k} \mu_j \nabla h_j(x^*) = 0. \qquad (1)$$

This is the condition one must look for, along with the conditions: For all i

$$\lambda_i g_i(x^*) = 0. \qquad (2)$$

The students are usually quite proficient at going through these steps without understanding why. Well, here's a gallery of pictorial interpretations of the rule, notably of (1), done many different ways. Take your pick (or should I say 'pic'?).

A Touching Incident

For simplicity, consider only the inequality constraints. Let $D = \{x : g_i(x) \leq 0, 1 \leq i \leq m\}$, the 'feasible' region where the constraints are met. Consider the surface $S = \{x : f(x) = c\}$ with the scalar c chosen so that S intersects D.

As we reduce c, S moves in the direction of $-\nabla f(\cdot)$, till it is barely touching D at the point x^*. (If not, we could reduce c a wee bit more while still intersecting D - see *Figure 1*). This then is the solution we are after. *Figure 2 a, b* illustrate the cases $m = 1$ and 2. Note that $-\nabla f(x^*)$ must make an acute angle with each outward normal $\nabla g_i(x^*)$ to the constraint surfaces that are active (i.e, satisfy $g_i(x^*) = 0$. The j-th constraint is inactive at x^* if $g_j^*(x^*) < 0$.) Hence $-\nabla f(x^*)$ is a positive linear combination of the $\nabla g_i(x^*)$'s for active constraints, leading to (1), (2). (The mathematical gadget that sees this through is called the Farkas lemma.) It is not hard to throw in the equality constraints as well. (Try it!).

Getting Rid of the Dependents

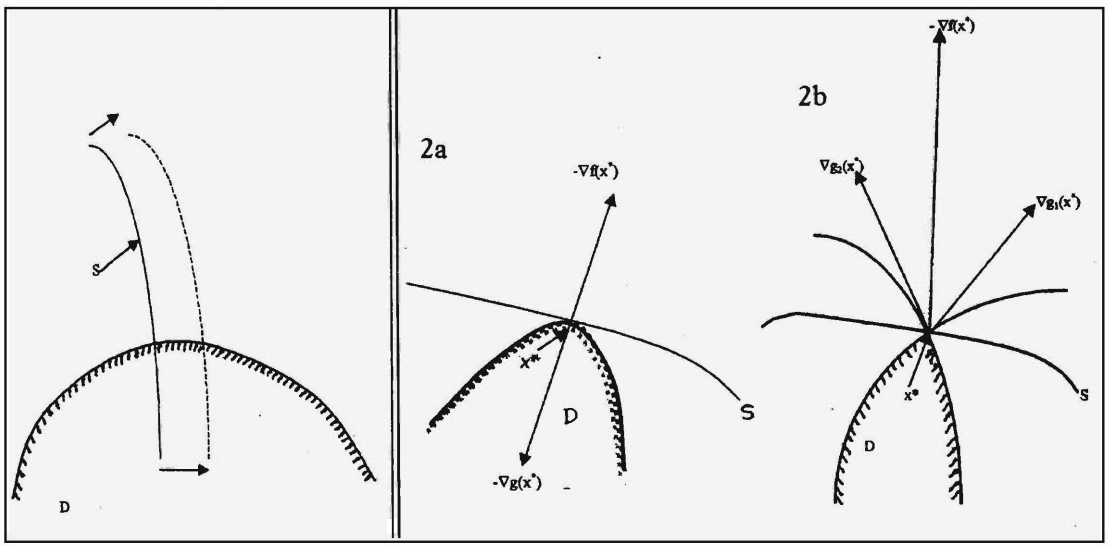
Consider only the equality constraints: For $k < d$,

$$h(x) = [h_1(x), \ldots, h_k(x)] = \theta \ (= [0, \ldots, 0]), \qquad (3)$$

which defines a surface S. Now (3) is a system of k equations in d unknowns. Suppose in a neighbourhood of x^*, we can solve for the dependent variables (say, $\hat{x} = [x_1, \ldots, x_k]$) in terms of the independent variables (say, $\overline{x} = [x_{k+1}, \ldots, x_d]$) as: $\hat{x} = \phi(\overline{x})$. Then (3) becomes

$$h(\phi(\overline{x}), \overline{x}) = 0.$$

Left: Figure1.
Right: Figure 2a,b.



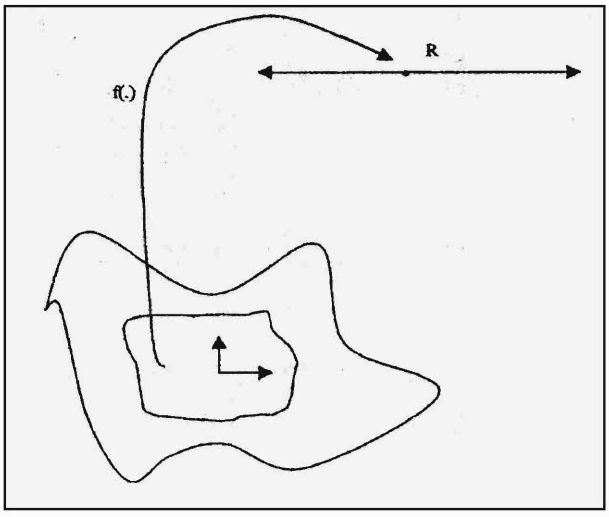


Figure 3.

Setting the gradient of this (in $\bar{x}$) equal to zero at x^*, some manipulation leads to (1) (See, e.g., [1], pp. 257-260.). What we are doing here is to consider the optimization problem on S itself, in a neighbourhood of x^* that looks like a patch of R^k, paying no heed to what else is out there (See *Figure 3*). (The mathematical mantra that allows this is the 'implicit function theorem'.) Most traditional textbook treatments of the Lagrange multiplier rule proceed this way. The procedure is less obvious for inequality constraints, but [3], pp. 315, provides a neat trick for taking care of that, a 'proof by contradiction' using the familiar argument that, if not, one could reduce f a bit more while still meeting the constraints.

Penalty for Excess Baggage

One can penalize the violation of constraints as follows: Let $g_i^+(x) = \max(g_i(x), 0)$ and minimize

$$F_N(x) = f(x) + ||x - x^*||^2 + N\left(\sum_{i=1}^{m} g_i^+(x)^2 + \sum_{j=1}^{k} h_j(x)^2\right), \tag{4}$$

for $N > 0$ 'large'. The second term penalizes deviation from x^* while the last term penalizes violation of the constraints.

For 'small' N, the minimum of $F_N(\cdot)$ will typically fall outsite the feasible region D (say, at x_N). As we increase N, x_N will move towards D and eventually converge to $x_\infty = x^*$

See *Figure 4*. (The second term in (4) ensures that the convergence is to x^* and not to any other point on the boundary of D.) Write out $\nabla F_N(x_N) = 0$ and let N tend to infinity. In the limit, the second term drops out and the rest can be arranged to look like (1) ([4], or [1], pp. 261-264). The most elementary proof of the Lagrange multiplier rule that I know follows this line of argument [4]. (Recall that 'elementary' does not mean short, just that it does not use any fancy machinery.)

A Matter of Inclination

The next picture applies to the case of inequality constraints when both f and g_i's are convex. (That is, a line joining any two points on the graph thereof lies above the graph.) Let $g(\cdot) = [g_1(\cdot), \ldots, g_m(\cdot)]$. Our problem is: minimize f on $\{x : g(x) \leq \theta\}$. Consider instead the family of problems:

$$\text{minimize } f(x), \text{ subject to } g(x) \leq z,$$

for $z \in R^m$, and let $w(z)$ denote the value of the minimum.

Then $z \rightarrow w(z)$ is decreasing in each of its arguments (obvious) and convex (easy to show). What we want is $w(\theta)$.

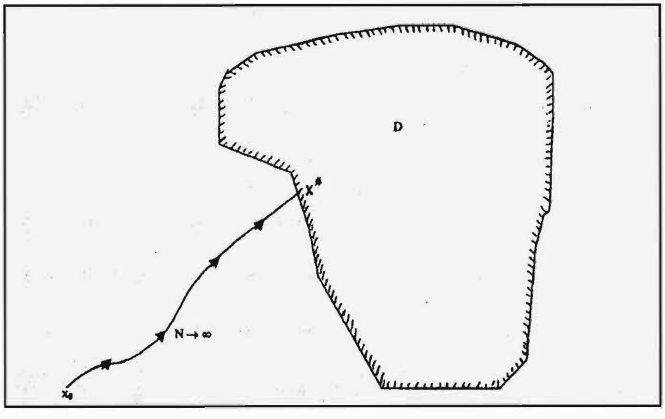


Figure 4.

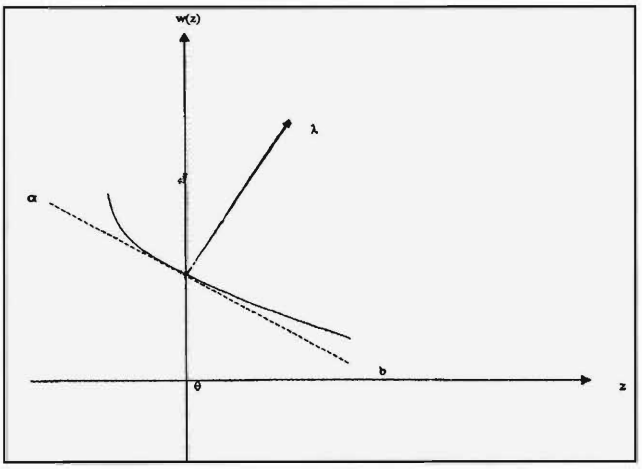


Figure 5.

Let S be a hyperplane that touches the graph of $w(\cdot)$ at $(\theta, w(\theta))$. (See *Figure 5.*) Then the point $(\theta, w(\theta))$ can be found by tilting the coordinates so that S becomes 'horizontal' and then looking for the minimum of the tilted graph of $w(\cdot)$.

This is attained at θ. But tilting thus amounts to adding a linear function to $w(\cdot)$, i.e., the whole operation is equivalent to minimizing $w(z) + \sum_{i=1}^{m} \lambda_i z_i$ over $z = [z_1, \ldots, z_m]$, where $\lambda_i \geq 0$. (This last bit is because $w(\cdot)$ is decreasing: Note that $\lambda = [\lambda_1, \ldots, \lambda_m, 1]$ is the normal to S as shown in Figure 5.)

That this tilting exercise does indeed correspond to the Lagrange multiplier rule takes some more manipulation:

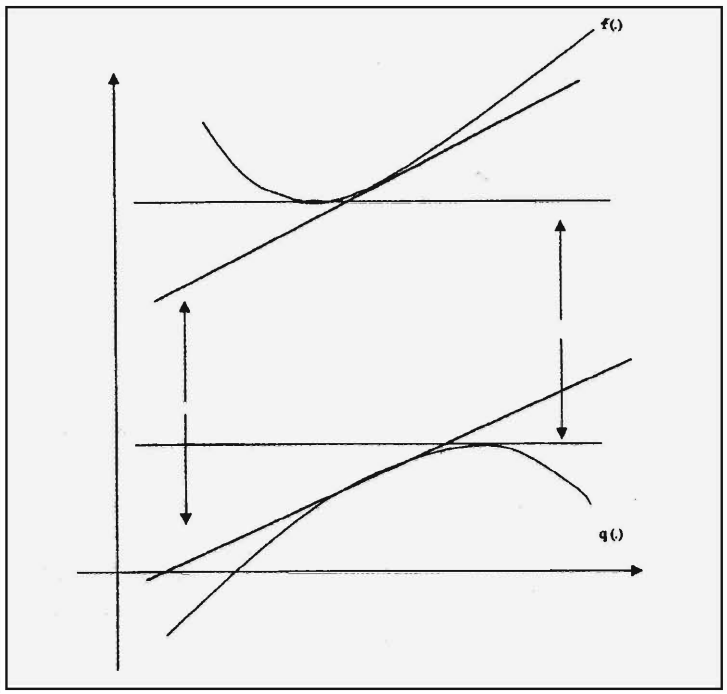
$$\begin{aligned} \min_z(w(z) + \lambda^T z) &= \min_z(\min_x\{f(x) : g(x) \leq z\} + \lambda^T z) \\ &= \min_z \min_x\{f(x) + \lambda^T z : g(x) \leq z\} \\ &= \min_x \min_z\{f(x) + \lambda^T z : g(x) \leq z\} \\ &= \min_x(f(x) + \lambda^T g(x)), \end{aligned}$$

and we are home ! This is the basis for the 'global' theory of large multipliers for convex problems.

Pairs that Separate

Suppose f is convex and $q : R^d \to R$ concave (i.e., $-q$ is convex), with $f(x) > q(x)$ for all x. Then $\min_x(f(x) - q(x))$

Figure 6.



is the minimum vertical separation between their graphs, which equals the maximum possible vertical separation between pairs of parallel hyperplanes separating the two graphs (See *Figure 6*). A formal statement of this is the celebrated 'Fenchel duality theorem' of optimization ([2], pp. 201) that associates to a minimization problem a 'dual' maximization problem.

Going back to our original problem, suppose we have $f(\cdot) > 0$. Let $q(x) = 0$ on D and $-\infty$ elsewhere (See *Figure 7*). Consider the pair of parallel hyperplanes separating their graphs, whose vertical separation equals the minimum vertical separation between f and q. Then the upper hyperplane touches the graph of f at $(x^*, f(x^*))$. The normals **a**, **b** to the hyperplanes as shown in the figure are parallel and oppositely oriented. Then so are their horizontal components, which turn out to be resp. $\nabla f(x^*)$ and a positive linear combination of the $\nabla g_i(x^*)$'s. Conclude from this, using the kind of intuition we used in *Figure 2*, that (1) holds for some $\lambda_i \geq 0$.

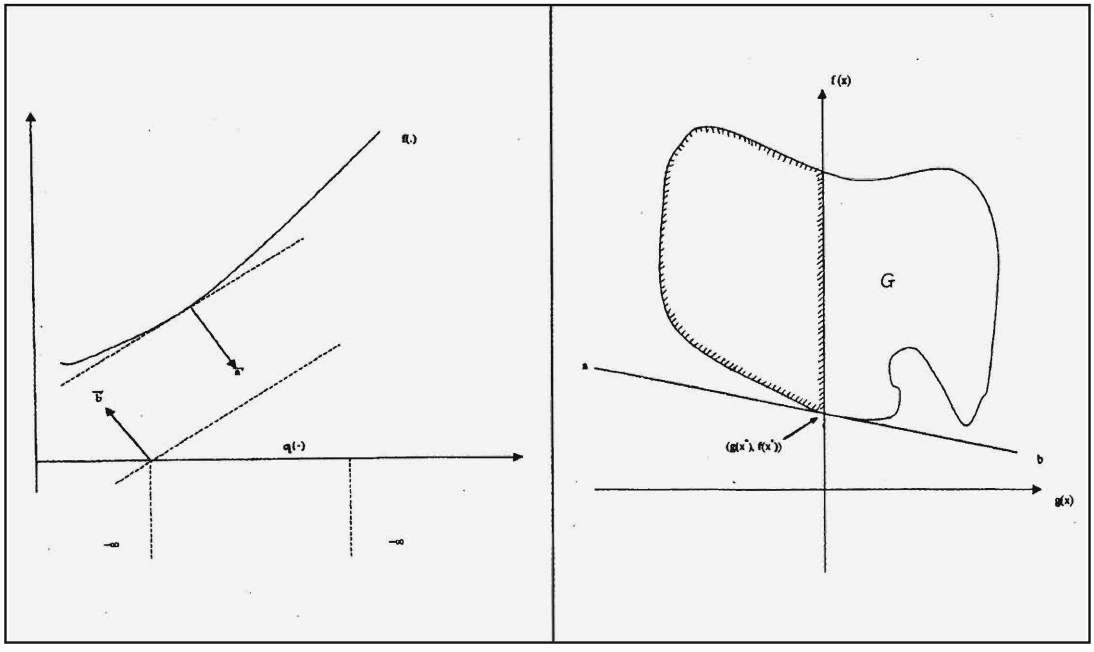


Figure 7. (left)
Figure 8. (right)

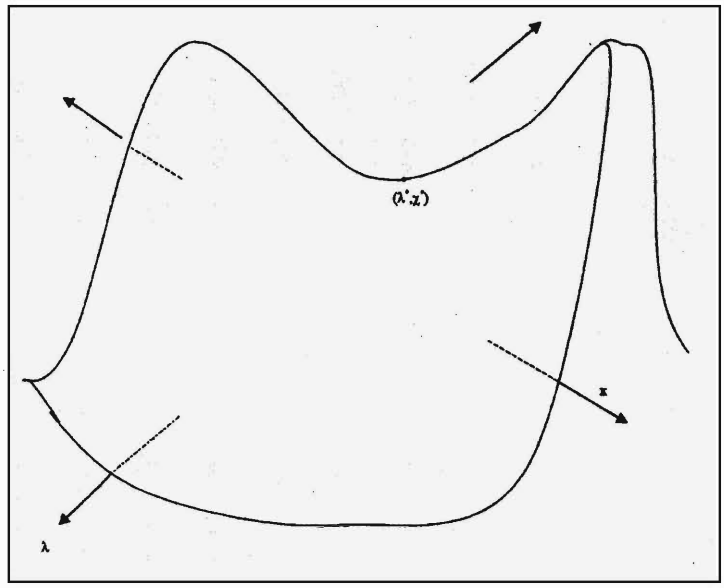
The 'Mona Lisa'

This section is titled thus because this picture, adapted from [1], is *the* picture to end all pictures. Consider the inequality constraints alone and consider the region $G = \{(g(x), f(x)) : x \in R^d\}$ in R^{m+1} (See *Figure 8*). Then the shaded portion corresponds to $g(x) \leq \theta$, with $(g(x^*), f(x^*))$ as marked.

Clearly, this is the minimum point of the region G if we tilt the axes so as to make the hyperplane $a - b$ horizontal. Now argue as in section 5.

The reason I rate this picture so high is that there are many details that I have been glossing over (note that I always took the 'nice' pictures, not the pathologies), that are far clearer with this picture than with any other. One is a technical condition we need for the above to go through, viz., that there be an x_0 with $g(x_0) < \theta$ (the 'Slater condition'). Another is the role of convexity. Things go through smoothly when f, g_i are convex. The reader is invited to think this through.

Figure 9.



Finally, there is the 'saddle point' theorem, which deserves mention. This theorem says that if $\{\lambda_i^*\}$ are the 'true' Lagrange multipliers, then the map $\lambda = [\lambda_1, \ldots, \lambda_m] \rightarrow f(x^*) + \sum_{i=1}^m \lambda_i g_i(x^*)$ is maximized over $\{\lambda : \lambda \geq \theta\}$ by λ^*. That is, the map $(x, \lambda) \rightarrow f(x) + \sum_{i=1}^m \lambda_i g_i(x)$ has a 'saddle point' at (x^*, λ^*). (See *Figure 9*).

Suggested Readings

[1] D P Bertsekas. *Nonlinear programming*, Athena Scientific. Belmont. Mass., 1995.

[2] D G Luenberger. Optimization by vector space methods. John Wiley and Sons. New York, 1967.

[3] D G Luenberger. *Linear and Nonlinear Programming*. 2nd edn. Addison Wesley Publishing Co. Reading. Mass., 1984.

[4] E J McShane. The Lagrange multiplier rule. *American Mathematical Monthly*. pp.922–925, 1973.

Going by the Rule

In the foregoing, I have taken so many liberties with mathematical rigour that I can almost see my 'pure' mathematician friends squirm in their seats. ("Whither epsilons? Whither the deltas?", I hear them moan, "At least take care of the epsilons and the deltas will take care of themselves!")

To appease them, I give here references to the precise statements of the rule. A nice 'local' statement (along the lines of (1)) is in [4], pp. 923. For 'global' theory in presence of convexity, see [2], pp. 217-219, which also contains the saddle point theorem. Alternatively, see [1], Chapters 3 and 5 for an extensive account.