The Lagrange multiplier rule revisited

Jan Brinkhuis

Econometric Institute, Erasmus University Rotterdam

Vladimir Protasov

Department of Mechanics and Mathematics, Moscow State University

Econometric Institute Report EI 2008-08

April 2008

Abstract

In this note we give a short novel proof of the well-known Lagrange multiplier rule, discuss the sources of the power of this rule and consider several applications of this rule. The new proof does not use the implicit function theorem and combines the advantages of two of the most well-known proofs: it provides the useful geometric insight of the elimination approach based on differentiable curves and technically it is not more complicated than the simple penalty approach. Then we emphasize that the power of the rule is the reversal of order of the natural tasks, elimination and differentiation. This turns the hardest task, elimination, from a nonlinear problem into a linear one. This phenomenon is illustrated by several convincing examples of applications of the rule to various areas. Finally we give three hints on the use of the rule.

Keywords: Optimization, Lagrange multiplier rule, compactness

1 Introduction

Useful geometric insight and technical simplicity. In this note we give an elementary and natural proof of the Lagrange multiplier rule that uses only the two main principles of continuous optimization: 'local descent' and 'existence of solutions for optimization of a continuous function on a nonempty compact (closed and bounded) set'. This makes for example a course on optimization self-contained. In particular, the implicit function theorem, which itself requires a relatively difficult proof, is not used. The present proof gives a fuller insight than the two most well-known proofs: the derivation from the implicit function theorem (called the elimination approach or the feasible direction viewpoint) and the penalty proof. In fact, it combines the following advantages of these proofs: the technical simplicity of the penalty proof and the useful geometric insight of the elimination proof (these advantages are for example formulated explicitly by Bertsekas [2] (p.282 and p.377)).

Comparison with elimination proof. The proof given in this paper, and written down in a rigorous analytical style, is based on a simple geometric intuition. The idea is to carry out the main task of the elimination proof—producing admissible points close to the considered admissible point of the given minimization problem with equality constraints—by presenting these points as solutions of auxiliary optimization problems. Let us be more precise and sketch the strategies of the usual elimination proof and the one given in the present paper in geometric language, and compare them.

Beginning of both proofs. The beginnings of the two proofs run parallel. Let $f_j(x),\ 0 \le j \le m$, be continuously differentiable functions of n variables. The problem (P) to minimize $f_0(x)$ subject to the equality constraints $f_j(x) = 0,\ 1 \le j \le m$, is considered, together with an admissible point $\widehat{x}$ of (P). The rule is proved by making the assumption that its conclusion does not hold at $\widehat{x}$—this means that the gradients $\nabla f_j(\widehat{x}),\ 0 \le j \le m$, are linearly independent—and by deriving from this assumption that $\widehat{x}$ is not a point of local minimum for (P). We assume, without loss of generality for the argument, that $\widehat{x} = 0$ and $f_0(\widehat{x}) = 0$. Let W be the linear span of the gradients $\nabla f_j(\widehat{x}),\ 1 \le j \le m$, let U be the the orthogonal complement of W in $\mathbb{R}^n$, and let $\bar{u}$ be the orthogonal projection of $-\nabla f_0(\widehat{x})$ on U. By the assumption, $\bar{u}$ is nonzero. Therefore, by the definition of $\bar{u}$, the directional derivative of f_0 at $\widehat{x}$ in direction $\bar{u}$ is negative. Let V be the linear span of the gradients $\nabla f_j(\widehat{x}),\ 0 \le j \le m$.

Usual elimination proof. The usual elimination proof continues by applying the implicit function theorem to conclude that the intersection with V of the admissible set of (P) is, in the neighborhood of the point $\widehat{x}$, a curve and that $\bar{u}$ is a tangent vector to this curve at the point $\widehat{x}$ (moreover, U is the tangent space at the point $\widehat{x}$ to the admissible set of (P)). It follows that, if one starts from the point $\widehat{x}$ and follows this curve in the direction of $\bar{u}$, then initially f_0 decreases monotonically. Therefore, $\widehat{x}$ is not a local minimum of (P), as required.

Novel proof. The proof that is given in the present paper continues by defining the ellipsoid E in W by the inequality $\sum_{j=1}^{m}\langle \nabla f_j(\widehat{x}), h\rangle^2 \le 1$ in the variable vector $h \in W$. Let u be a positive scalar multiple of $\bar{u}$ that is so large that the linear function $\nabla f_0(\widehat{x})^T x$ assumes only values < -1 for all x belonging to the shifted ellipsoid $u + E$. One chooses a sufficiently small $t > 0$. One considers the steepest descent curve in the affine space $\widehat{x} + tu + W$ with an arbitrary point of the ellipsoid $E_t = \widehat{x} + t(u + E)$ as its starting point, of the function $g = \sum_{j=1}^{m} f_j^2$, the sum of the squares of the functions that define the equality constraints of (P). One can verify that this curve is contained in the ellipsoid E_t and that it converges to an interior point $\widehat{x}_t$ of E_t at which the value of g is zero, that is, $\widehat{x}_t$ is an admissible point of (P) (moreover, it could be proved that the endpoint of the steepest descent curve, $\widehat{x}_t$, does not depend on the starting point of the curve). One can check that the ellipsoid E_t is so small that the objective function of (P) takes a smaller value at $\widehat{x}_t$ than at $\widehat{x}$. As $\widehat{x}_t$ tends to $\widehat{x}$ as $t \downarrow 0$, it follows that $\widehat{x}$ is not a local minimum of (P), as required. In the rigorous write-up of this proof, the existence of the point $\widehat{x}_t$ will be produced by means of compactness as the point of minimum on the ellipsoid E_t for the function g.

Additional insights. This paper presents the result of our attempts to 'reach the essence' of the multiplier rule. These attempts led not only to the novel proof, but also to some additional insights: into the power of the rule and into the use of the rule. It appeared to be useful to include these insights in this paper: we could not find our viewpoint on the power of this rule in the literature, and our hints on the use of the rule are known to users but are usually not pointed out.

On the power of the rule. We try to give some insight into the source of the power of the Lagrange multiplier rule. We present our viewpoint that this source is the *reversal of order of the natural tasks*, elimination and differentiation. *This turns the hardest task, elimination, from a nonlinear problem into a linear one.* In particular, the role of the multipliers is shown to be not essential. Moreover, we present some convincing examples of problems that illustrate the power of the Lagrange multiplier rule.

Hints on the use of the rule. Finally, we give three useful hints on the use of the multiplier rule, and illustrate these by applications from various fields such as transportation theory (prediction of flows of cargo by the RAS-model), investment theory (the Markowitz problem), production economics (theorems of Gossen on the optimal allocation of money and time), location theory, bargaining (Nash bargaining), algorithms (self-concordancy properties of the logarithmic barrier for semidefinite programming problems), and ergodic theory.

Comparison with the literature. The Lagrange multiplier rule is a standard tool. The 'usual proof' by means of the implicit function theorem is for example given by Duistermaat and Kolk [7]. A technically simple proof, by means of the penalty approach is given by Hestenes [9] and by Bertsekas [2]. A proof that gives insight, based on differentiable curves, constructed using the implicit function theorem is given by Hestenes [9] and by Luenberger [11]. These two proofs are compared by Bertsekas [2]: "This approach (*using differentiable curves*) is insightful, but is considerably more complicated than the penalty approach we have followed'. *The proof given in the present paper* *appears to have the advantages of both these proofs: it is insightful and simple.*

Organization paper. The organization of this paper is as follows. In section 2, the rule is proved. In section 3, an illustration of the power of the rule is given, as well as insight into the source of this power. In section 4, three hints on the use of the multiplier rule are given and these are illustrated by applications from various fields.

2 A proof of the multiplier rule based on compactness

Statement of the rule. Let continuously differentiable functions $f_j,\ 0 \leq j \leq m$, of n variables be given. Consider the following minimization problem (P) with equality constraints (maximization problems can be written as minimization problems by multiplying the objective function by -1):

$$\min f(x), \quad x \in \mathbb{R}^n,\ f_j(x) = 0,\ 1 \leq j \leq m.$$

Let $\widehat{x}$ be an admissible point of (P). *The Lagrange multiplier rule states that if $\widehat{x}$ is a local minimum of (P), then there exists a nonzero row vector $\lambda = (\lambda_0, \ldots, \lambda_m) \in \mathbb{R}^{m+1}$ such that*

$$\sum_{j=0}^{m} \lambda_j \nabla f_j(\widehat{x}) = 0.$$

To prove the rule, we will argue by contradiction. We will assume that the conclusion of the rule does not hold, that is, that the gradients $\nabla f_j(\widehat{x})$ of the functions $f_j,\ 0 \leq j \leq m$, at the point $\widehat{x}$ are linearly independent. To prove the rule, it suffices to show that $\widehat{x}$ is not a local minimum of (P). For the sake of simplicity, we put $\widehat{x} = 0$ and $f_0(\widehat{x}) = 0$. This is always possible by shifting the coordinates and adding a constant to f_0.

Local descent of the objective function on a family of ellipsoids $(E_t)_t$. The argument is carried out in the subspace V of $\mathbb{R}^n$ spanned by the gradients $\nabla f_j(0),\ 0 \leq j \leq m$. Let W be the hyperplane in V that is spanned by the gradients $\nabla f_j(0),\ 1 \leq j \leq m$, and let L be the line in V through the origin that is orthogonal to W. Let E be the ellipsoid in W given by the inequality $\sum_{j=1}^{m} \langle \nabla f_j(0), h \rangle^2 \leq 1$ in the variable vector $h \in W$. A point u on the line L is chosen in the following way: on the same side of W as $-\nabla f_0(0)$ and so far away from the origin that

$$\langle \nabla f_0(0), x \rangle < -1 \tag{1}$$

for all $x \in u + E$. Now we verify that for sufficiently small numbers $t > 0$, the objective function f_0 assumes at 0 a higher value than at any point of the ellipsoid $E_t = t(u + E)$:

$$f_0(t(u+h)) = t\langle \nabla f_0(0), u + h \rangle + o(t) < -t + o(t) < 0$$

for each $h \in E$, if $t > 0$ is sufficiently small. The equality follows from the definition of the derivative, the first inequality from (1) and the second inequality from the definition of the small Landau-o symbol.

Analysis of auxiliary minimization problems. Now we consider for each such t the auxiliary minimization problem (Q_t) to minimize $g = \sum_{j=1}^m f_j^2$, the sum of the squares of the functions $f_j,\ 1 \leq j \leq m$, on the ellipsoid E_t. The problem (Q_t) has a global solution $\widehat{x}_t$, as its admissible set is nonempty, closed and bounded, and its objective function g is continuous. Now we calculate the gradient of the objective function at a point $t(u+h)$ of E_t:

$$(\sum_{j=1}^m f_j(t(u+h)+k)^2) - (\sum_{j=1}^m f_j(t(u+h))^2) = 2\langle \sum_{j=1}^m f_j(t(u+h))\nabla f_j(t(u+h)), k\rangle + o(k).$$

That is, the required gradient is the orthogonal projection on W of the following vector:

$$2\sum_{j=1}^m f_j(t(u+h))\nabla f_j(t(u+h)). \tag{2}$$

This calculation reveals two properties of this gradient. The first one is that it is nonzero at points where the objective function itself takes a nonzero value. Indeed, by assumption, the vectors $\nabla f_j(0),\ 1 \leq j \leq m$, are linearly independent, and $\lim_{t\downarrow 0} t(u+h) = 0$; therefore, for sufficiently small $t > 0$, the vectors $\nabla f_j(t(u+h)),\ 1 \leq j \leq m$, are linearly independent; it follows from (2) that the gradient is nonzero if one of the numbers $f_j(t(u+h)),\ 1 \leq j \leq m$, is nonzero, that is, if $g(t(u+h)$ is nonzero.

The second one is that at boundary points of the ellipsoid E_t, this gradient and the outward normal to the ellipsoid make an acute angle. Indeed, for each boundary point $\bar{h}$ of the ellipsoid E, an outward normal for the boundary point $t(u+\bar{h})$ of the ellipsoid E_t is given by the expression $\sum_{j=1}^m \langle \nabla f_j(0), \bar{h}\rangle \nabla f_j(0)$. This expression is clearly orthogonal to the boundary of the ellipsoid E_t at the point $t(u+\bar{h})$. To see that, moreover, this expression cannot be zero, we recall that the gradients $\nabla f_j(0),\ 1 \leq j \leq m$, are linearly independent and that at least one of the numbers $\langle \nabla f_j(0), \bar{h}\rangle,\ 1 \leq j \leq m$ is nonzero, as $\sum_{j=1}^m \langle \nabla f_j(0), \bar{h}\rangle^2 = 1$. The gradient of g at $t(u+\bar{h})$, given by (2) with $h = \bar{h}$, equals

$$2(t\sum_{j=1}^m \langle \nabla f_j(0), \bar{h}\rangle \nabla f_j(0)) + o(t).$$

It follows that the inner product of the gradient of g at $t(u+\bar{h})$ and the outward normal above equals $2t|\sum_{j=1}^m \langle \nabla f_j(0), \bar{h}\rangle \nabla f_j(0)|^2 + o(t)$—where $|\cdot|$ denotes the euclidian norm on $\mathbb{R}^n$—and so that this inner product is positive for sufficiently small $t > 0$, as required; that is, these two vectors make an acute angle.

Conclusion of the proof. It follows from the two properties above that if one moves in the affine subspace $tu + W$, starting from a boundary point of the ellipsoid E_t or from an interior point of the ellipsoid in which the value of g is nonzero, in a straight line in the direction of minus the gradient of g, then one stays initially inside this ellipsoid and, moreover, the value of g decreases. Therefore, $\widehat{x}_t$, being a point of global minimum of g on E_t, is an interior point of E_t and the value of g at $\widehat{x}_t$ is zero. That is, the sum of the squares of the functions f_j, $1 \le j \le m$ at the point $\widehat{x}_t$ equals zero, that is, $\widehat{x}_t$ is an admissible point for the original problem (P). As the point $\widehat{x}_t$ is contained in the ellipsoid $E_t = t(u + E)$, we have the inequality $f_0(\widehat{x}_t) < f_0(0)$ and the property that $\widehat{x}_t \to 0$ for $t \downarrow 0$. It follows that $\widehat{x} = 0$ is not a local minimum of (P), as required.

3 The source of the power of the multiplier rule

The idea that makes the Lagrange multiplier rule work, is the simple but clever trick to *reverse the natural order of the main tasks, elimination and differentiation. This turns the hardest task, elimination, from a nonlinear problem into a linear one.* To illustrate this idea, we compare with the natural order: first use the constraints to eliminate variables, then put derivatives equal to zero.

Let us consider the simplest case, a problem of type $f(x, y) \to \min,\ g(x, y) = 0$. The natural order would be to try to solve first the nonlinear problem of eliminating y from the constraint. The Lagrange method prescribes to differentiate first; this leads to the equation $dg(x, y) = g_x(x, y)dx + g_y(x, y)dy = 0$, in x, y, dx and dy, that is linear in dx and dy, so a *linear* elimination problem remains (then combination with the stationarity equation $df(x, y) = 0$ gives the multiplier rule, formulated without multipliers). Thus the source of the power of the multiplier rule has been revealed to be this simple idea. The role of multipliers is just to make the execution of the tasks in the reversed order slightly more convenient.

To illustrate the advantage over the natural order, consider a problem of the type $f(x, y) = a_0x^2 + b_0xy + c_0y^2 \to \min,\ g(x, y) = a_1x^2 + b_1xy + c_1y^2 - d = 0$. When one tries to solve this problem using the Fermat method, one gets stuck: elimination is possible but then differentiation gives an intractable equation. The Lagrange equations lead to a quadratic equation in $\frac{y}{x}$. Substituting its solutions into the constraint we obtain several suspicious points (x, y), from which we get by comparison the point(s) of minimum and maximum.

Moreover, it is natural to ask whether there are problems that can be solved by the Lagrange method in a shortest and most natural way. An example of this is the following theorem of Steiner that solved a celebrated problem from antiquity: a quadrangle with given lengths of the sides has maximal area if its four vertices lie on a circle. The corresponding optimization problem can be written as

$$(1) \qquad \begin{cases} S(\alpha,\beta) \;=\; \frac{1}{2}\Big(ab\sin\alpha + cd\sin\beta\Big) \quad \rightarrow \quad \max \\ a^2+b^2-2ab\cos\alpha \;=\; c^2+d^2-2cd\cos\beta \end{cases}$$

here a, b, c, d are the sides of the quadrangle, α is the angle between a and b, β is the angle between c and d. The Lagrange equations

$$(2) \qquad \begin{array}{l} \frac{1}{2}\lambda_0 ab\cos\alpha + 2\lambda_1 ab\sin\alpha = 0 \\ \frac{1}{2}\lambda_0 cd\cos\beta - 2\lambda_1 cd\sin\beta = 0 \end{array}$$

give $\tan\alpha = -\tan\beta$ and hence $\alpha = \pi - \beta$. This means that the quadrangle $abcd$ is inscribed in a circle.

4 Special tricks

Let us now give some special tricks, which are common knowledge among users of the multiplier rule, but which are usually not written down. In all applications we will put without comment $\lambda_0 = 1$. This is justified: although there exist optimization problems for which λ_0 can be zero, one can in each application of interest of the multiplier rule that we know of easily exclude that λ_0 is zero, by means of the Lagrange equations and the equality constraints. Then the multipliers can be normalized by putting $\lambda_0 = 1$. Therefore, we choose not to display the routine verifications that $\lambda_0 \neq 0$ and will always put $\lambda_0 = 1$.

First trick. *Find all variables, in which both the objective function and the constraints can be expressed in a simple and symmetric way.*

In the following example, the simplest version of the very flexible RAS-model from transportation theory, all constraints are linear. Therefore, it is possible to solve this problem by carrying out the main tasks in the natural order: by eliminating first and then differentiating. However, the multiplier rule gives an advantage, in particular because it keeps the symmetry of the problem.

In this and all other examples we make tacit use of the existence of global maxima and minima for continuous functions on nonempty compact (closed and bounded) subsets of $\mathbb{R}^n$ (the theorem of Weierstrass).

Example 4.1 *[Prediction of flows of cargo.] An investor wants to have information about the n^2 flows of cargo, measured in containers, within an area consisting of n zones, including the flows within each zone. The problem is that insufficient information is available. For all zones only data are available to him for the total flow originating in this zone, O_i, $1 \le i \le n$, and for the total* *flow with destination in this zone, D_j, $1 \leq j \leq n$. However, the investor wants to have at least an estimation for T_{ij}, the flow from zone i to zone j for all $i, j \in \{1, \ldots, n\}$. For this one can take as an estimate the distribution matrix T_{ij} with the highest probability, given the available data and assuming that all units of cargo are distributed over the n^2 possibilities with equal probability. The logarithm of the probability of a distribution matrix can for large n be approximated by the formula $C - \sum_{i,j}[T_{ij}(\ln T_{ij}) - T_{ij}]$, where C is a constant which does not depend on the choice of T.*

Thus we are led to the problem

$$\sum_{i,j}[T_{ij}(\ln T_{ij}) - T_{ij}] \to \min, \ \sum_{j=1}^{n} T_{ij} = O_i, \ \sum_{i=1}^{n} T_{ij} = D_j, \ T_{ij} > 0 \ \forall i, j.$$

Solution.

The Lagrange equations $\lambda_0 \ln T_{ij} - \lambda_i - \lambda'_j = 0$ give, after putting $\lambda_0 = 1$, that $T_{ij} = e^{\lambda_i} e^{\lambda'_j}$ and so the matrix T has rank one. This leads to the following solution $T_{ij} = \frac{O_i D_j}{S}$, which is the required estimate. $\diamond$

The second example is the celebrated Markowitz problem [12] from finance. It provides the foundation for single period investment theory.

Example 4.2 *Which portfolio of a number of assets with known mean rate of return and known covariances minimizes risk while yielding a desired expected return?*

Solution. We consider the following optimization problem

$$\frac{1}{2} \sum_{i,j=1}^{n} w_i w_j \sigma_{ij} \to \min, \ \sum_{i=1}^{n} w_i \bar{r}_i = \bar{r}, \ \sum_{i=1}^{n} w_i = 1.$$

Here n is the number of assets, the mean rates of return are $\bar{r}_i$, $1 \leq i \leq n$, the covariances are σ_{ij}, $1 \leq i, j \leq n$, and the weights of the portfolio are w_i, $1 \leq i \leq n$, and the mean value of the portfolio is fixed at $\bar{r}$. The Lagrange method leads to the following system of equations:

$$\sum_{j=1}^{n} \sigma_{ij} w_j - \lambda \bar{r}_i - \mu = 0, \ 1 \leq i \leq n, \ \sum_{i=1}^{n} w_i \bar{r}_i = \bar{r}, \ \sum_{i=1}^{n} w_i = 1.$$

All $n + 2$ equations are linear, so this system can be solved with linear algebra methods. $\diamond$

Second trick. *In the solution of the Lagrange equations, one should ask oneself whether it is really necessary to compute the Lagrange multipliers.*

As the first example we offer two of Gossen's fundamental theorems on optimal allocations.

Example 4.3 1. ***Optimal allocation of money.*** *In production economies the only items produced are the ones that have the largest marginal profits; the marginal profits are also equal, and maximal, among all items.*

2. ***Optimal allocation of time.*** *In order to optimize the use of time, one should only spent time on activities that give the largest marginal utility.*

Solution. Both results are proved in the same way. Here we prove the first one. We consider the problem of maximizing profits under a given budget:

$$\pi(x_1, \ldots, x_n) \to \max,\ \sum_{i=1}^{n} p_i x_i = B,\ x_i \geq 0,\ 1 \leq i \leq n.$$

Here n is the number of all items, the prices of the items are $p_i,\ 1 \leq i \leq n$, the budget is B and the profit function is π. To prove the theorem, we may leave the items that are zero in the optimum out of consideration. This is the same as saying that to prove the theorem, it suffices to show that in the solution of the problem $\pi(x_1, \ldots, x_n) \to \max,\ \sum_{i=1}^{n} p_i x_i = B$ the marginal profits $\frac{\partial \pi}{\partial x_i}/p_i$ are equal. The Lagrange method gives the equations $\frac{\partial \pi}{\partial x_i}/p_i = \lambda,\ 1 \leq i \leq n$, where λ is the multiplier. That is, all marginal profits are equal to the value of the Lagrange multiplier. This establishes Gossen's fundamental theorem on the optimal allocation of money. $\diamond$

In the following example the Lagrange equations give the information that all the variables x_i can be seen to satisfy the same polynomial equation (with unknown coefficients!), so each one is contained in the set of roots of this equation. If the degree of this equation is smaller than the number of variables, then some of these variables have the same value. This argument is very useful in many applications of the Lagrange method.

This example illustrates at the same time that most—maybe all—inequalities, such as those given in [8], can be derived in a standard way by optimization methods. We illustrate this by the selfconcordancy inequalities for the logbarrier function for semidefinite programming problems. In the seminal monograph [14] it is shown how selfconcordancy inequalities lead to efficient interior point algorithms.

Example 4.4 *The logarithmic barrier on the positive semidefinite matrices* $b(X) = -\ln \det X$ *is a* ν*-self-concordant barrier for some positive number* ν*, that is, for all positive definite matrices* X *and all symmetric matrices* H*, the following inequalities hold true:*

1. $|b'''(X)[H,H,H]| \leq 2(b''(X)[H,H])^{\frac{3}{2}}$,
2. $|b'(X)[H]| \leq \nu(b''(X)[H,H])^{\frac{1}{2}}$.

Proof. To establish the first inequality we consider the problem

$$f(X,H) = b'''(X)[H,H,H] \to \max\,(\min),\ b''(X)[H,H] = c$$

for an arbitrary nonnegative constant c. Writing α_i for the eigenvalues of $(\sqrt{X})^{-1}H(\sqrt{X})^{-1}$, this problem is seen to be equivalent to the problem $g(\alpha) = \sum_{i=1}^n \alpha_i^3 \to \text{extr}$, $\sum_{i=1}^n \alpha_i^2 = d$ for some nonnegative constant d. The Lagrange equations show that for a solution α of this problem, all its coordinates are roots of the same linear equation. Therefore, they are all equal. This leads to the solutions of the problem and to its extremal values; this establishes the first inequality. The second inequality can be derived in the same way. $\diamond$

Third trick. *Do not use second order conditions.*

Second order conditions lead to longwinded computations of minors of bordered hessians. The reward of these computations is meagre: these conditions allow us to distinguish between *local* minima and maxima; they give no global information. However—almost—always one can avoid these conditions, complementing the multiplier rule with the Weierstrass theorem. This is even possible if the feasible region R is not closed or unbounded, a useful remark. For example, if the feasible set is the entire space $\mathbb{R}^n$ and f is coercive (that is $|f(x)| \to +\infty$ for $|x| \to +\infty$), then for $M > 0$ sufficiently large, adding the constraint $|x| \leq M$ does not change the, possibly empty, solution set of the original problem; then the Weierstrass theorem can be applied and the required existence of a solution of the original problem follows. In other cases, one can often show in a similar way that there exists a number C for which the level set $\{x \in R : f(x) \leq C\}$ is nonempty, closed and bounded.

The following example plays a role in ergodic theory, the study of dynamical systems and classical mechanics (see, for instance, [7]).

Example 4.5 *[Birkhoff theorem]. For an arbitrary bounded convex body in $\mathbb{R}^2$ with a smooth boundary and for any $n \geq 3$ there exists a billiard with n vertices (a billiard is a polygon having its vertices on the boundary and possessing the property that two sides going from each vertex form equal angles with the boundary at this vertex).*

Proof. Denote the body by M and its boundary by ∂M. Consider the set of all polygons having n vertices, all lying on ∂M. Obviously this set is compact if we allow vertices to coincide and consecutive sides to lie on a common straight line. Therefore there exists a polygon of maximal perimeter. This is a desirable billiard. In the first place, it has exactly n different vertices, otherwise one can add extra vertices and the perimeter increases. Take now an arbitrary triple of consecutive vertices x_1, x_2, x_3 of this polygon and denote by l the tangent line to the curve ∂M at the point x_2. The point x_2 is a solution for the following maximization problem:

$f(x) = |x - x_1| + |x - x_3| \to \max,\ x \in \partial M$. Solving this in the same way as in example 4.6, we obtain that the vectors $x - x_1$ and $x - x_3$ form equal angles with l. Therefore this polygon is a billiard. $\diamond$

Remark. Without the assumption of smoothness of the boundary it is not known whether billiards exist (not even for triangles).

The following example, which arises in location theory—for instance in determining a location on a highway for a facility such as a fast food restaurant or a gas station—illustrates how the Weierstrass theorem can be used even if the feasible set is not compact.

Example 4.6 *Let a straight line l and three points x_1, x_2, x_3 be given on the plane. Find (or characterize) the point on the line for which the sum of the distances from this point to the three given points is minimal.*

Solution. We write the condition $x \in l$ as a constraint $\langle x - x_0, n \rangle = 0$, where x_0 is a point on l, n is a vector orthogonal to l, and $\langle \cdot, \cdot \rangle$ is the standard inner product. Thus we have

$$\left\{ \begin{array}{l} f(x) \;=\; |x - x_1| + |x - x_2| + |x - x_3| \;\to\; \min \\ \langle x - x_0, n \rangle = 0 \end{array} \right. \tag{3}$$

By Weierstrass this problem has a solution; in order to achieve boundedness one may add the constraint $|x| \leq M$ for sufficiently large M. Differentiating the Lagrangian, we get that the sum of the unit vectors $u_i = \frac{x - x_i}{|x - x_i|}$, $i = 1, 2, 3$ equals $-\lambda n$. This is the same as saying that the sum of the projections of the vectors u_1, u_2, u_3 onto l (or the sum of cosines of angles formed by these vectors with the line l) is zero. This property characterizes the desirable point x. $\diamond$

The solution remains the same for an arbitrary number of points $x_1, \ldots, x_k$. In particular, for $k = 2$ we obtain a well-known elementary high-school problem. For $k \geq 3$ the solution, in general, cannot be constructed by compasses and ruler, and can only be characterized as we did above. The same principle of solution is illustrated by the problems of minimization of the distance from a point on a plane to k given points on the plane or to three given points in three-dimensional space.

The references below contain many not very well-known examples where the power of the multiplier rule can be demonstrated.

References

[1] O.Bottema, R.Z.Djordjevic, R.R.Janic, D.S.Mitrinovic, P.M.Vasic, *Geometric inequalities*, Wolters-Noordhoff Publishing, Groningen, 1969, 151 pp.

[2] Dimitri P.Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, 2004, 287 pp.

[3] K.Binmore, *Fun and Games, A Text on Game Theory*, Lexington, 1992.

[4] J.Brinkhuis, V.Protasov, *Theory of extremum in simple examples*, Mathematical Education, 9 (2005), pp. 32-55 (in Russian).

[5] J.Brinkhuis, V.Tikhomirov, *Optimization: Insights and Applications*, Princeton Series in Applied Mathematics, Princeton, N.J., 2005, 682 pp.

[6] H.S.M.Coxeter, S.Greitzer, *Geometry revisited*, The L. W. Singer Company, Random House, New York , 1967, 193 p.

[7] J.J.Duistermaat and J.A.C. Kolk, *Multidimensional Real Analysis I*, Cambridge University Press, Cambridge, 2003, 412 pp.

[8] G.Hardy, J.E.Littlewood, G.Polya, *Inequalities*, Cambridge University Press, 1934.

[9] M.R.Hestenes, *Optimization Theory: The Finite Dimensional Case*, Wiley, New York, 1975.

[10] A.N.Kolmogorov, S.V.Fomin, *Introductory real analysis*, Dover Publications inc., New York, 1975, 403 pp.

[11] D.G.Luenberger, *Introduction to Linear and Nonlinear Programming*, (2nd Ed.), Addison-Wesley, Reading, 1984.

[12] H.Markowitz, *Portfolio Selection: efficient diversification of investments.* John Wiley, 1959.

[13] J.F.Nash, Jr, *The bargaining problem*, Econometrica 18: 155-162 (1950).

[14] Y.Nesterov and A.Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Studies in Applied Mathematics 13, Philadelphia (1994).

[15] Ya.G.Sinai, *Introduction to ergodic theory*, Translated by V. Scheffer. Mathematical Notes, 18. Princeton University Press, Princeton, N.J., 1976, 144 pp.

[16] K.Sydsaeter, P.Hammond, *Essential Mathematics for Economic Analysis*, 1995.

[17] V.M.Tikhomirov, *Stories about Maxima and Minima*, American Mathematical Association, 1990.

[18] A.M.Yaglom, I.M.Yaglom, *Challenging mathematical problems with elementary solutions. Vol. II. Problems from various branches of mathematics*, Translated from the Russian by James McCawley, Jr. Reprint of the 1967 edition. Dover Publications, Inc., New York, 1987. 214 pp.

Jan Brinkhuis, Erasmus Universiteit Rotterdam, Econometrisch Instituut, Faculteit der Economische Wetenschappen, H 11-16, Postbus 1738, 3000 DR Rotterdam, The Netherlands, e-mail: brinkhuis@few.eur.nl

Vladimir Protasov, Department of Mechanics and Mathematics, Moscow State University, Vorobyovy Gory 3, Moscow, 119992, Russia, e-mail: vladimir_ protassov@yahoo.com