Revista de Matemática: Teoría y Aplicaciones **31**(2): 195-229
CIMPA - UCR ISSN: 1409-2433(Print), 2215-3373(Online)
DOI: https://doi.org/10.15517/rmta.v31i2.56792

# GRADIENTS AND OPTIMIZATION WITH CONSTRAINTS IN ECONOMICS AND SOCIAL SCIENCES

# GRADIENTES Y OPTIMIZACIÓN CON RESTRICCIONES EN ECONOMÍA Y CIENCIAS SOCIALES

Sergio A. Pernice[1]

*Received: 07/Nov/2023; Accepted: 16/May/2024*



[1] Universidad del CEMA, Buenos Aires, Argentina. E-Mail: sp@ucema.edu.ar

**Abstract**

Despite their widespread use in advanced analytical and numerical techniques, gradient field methods are often underrepresented in the foundational training of economists and social scientists. As machine learning and sophisticated analytical and numerical approaches gain traction, the importance of gradient methods in optimization processes becomes increasingly apparent. This oversight in academic and practical toolsets is suboptimal. This paper aims to address this gap by introducing gradient field methods both intuitively and rigorously, situating them within the context of problems commonly encountered by economists and social scientists, with a particular focus on equality constrained optimization.

**Keywords:** minimization with constraints; Lagrange multipliers; gradient fields algorithms.

**Resumen**

A pesar de su uso generalizado en técnicas analíticas y numéricas avanzadas, los métodos de campo de gradientes suelen estar subrepresentados en la formación básica de economistas y científicos sociales. A medida que el aprendizaje automático y los enfoques analíticos y numéricos sofisticados ganan terreno, la importancia de los métodos de gradiente en los procesos de optimización se vuelve cada vez más evidente. Esta falta en las herramientas académicas y prácticas es subóptima. Este artículo tiene como objetivo abordar esta brecha introduciendo los métodos de campo de gradientes tanto de manera intuitiva como rigurosa, situándolos en el contexto de problemas comúnmente encontrados por economistas y científicos sociales, con un enfoque particular en la optimización con restricciones de igualdad.

**Palabras clave:** minimización con restricciones; multiplicadores de Lagrange; algoritmos con campos de gradientes.

**Mathematics Subject Classification:** Primary 97M40; Secondary 49-01, 49K05, 49K10, 49K21, 97G70, 97H60, 97I99, 97M70, 97P99.

## 1. Introduction

Central to economics lies a foundational problem: the rational choice between two goods given a limited budget. This quintessential economic dilemma provides an ideal backdrop for introducing gradient-based methods, especially as they apply to constrained optimization problems. Not only are these methods versatile, easily generalizing to situations with multiple variables and constraints, but they also retain their intuitive appeal across these complexities when presented judiciously. Beyond the inherent significance of constrained optimization, these gradient-based methods underpin the surging machine learning revolution with far-reaching societal implications.

Within the broader scope of social sciences, optimization is not merely an exercise in mastering the procedural steps. Rather, it delves deeper into compre-

hending the intricacies of decision-making amidst resource constraints—a pervasive challenge confronting governments, organizations, and individuals. Regrettably, traditional pedagogical approaches, as showcased by texts like [9], [13], and [24], just to name three, often convey these concepts in a rather prescriptive and mechanistic manner.

The traditional approach to introducing optimization not only potentially obstructs a richer understanding but also forgoes an excellent opportunity to acquaint economics and social science students with state-of-the-art optimization methods pivotal in today's machine learning landscape. These methods are well-documented in classic textbooks such as [4], [5], [6], [7], [11], and [20], with related optimization techniques detailed in [10], [15], and [23].

Given this context, the primary objective of this paper is to recalibrate this perspective. We aspire to provide students and practitioners with an intuitive grasp of optimization under constraints. In doing so, our aim is to endow them with a toolkit that not only bridges the divide between classical mathematical economics concepts and contemporary methodologies but also aligns with the ongoing shift towards data-centric decision-making and machine learning, as highlighted by [1], [2], [3], [8], [12], [16], [18], [19], [21], [22], [25] among many others.

In this work, we illustrate the prowess of gradient field methods within two specific realms: the unconstrained gradient descent minimization and the quintessential Lagrange multiplier technique for optimization challenges bound by equality constraints. The former, especially its stochastic variants, arguably constitutes the primary utilization of gradient field methods in contemporary machine learning. Conversely, the latter boasts a dual advantage. On one side, students often already possess familiarity with its mechanics, and on the other, it tends to be presented in textbooks in a rather prescriptive manner, justified predominantly by its ability to generate correct equations. By emphasizing an intuitive comprehension of the Lagrange multiplier technique, we underscore its applicability and relevance, especially in high-dimensional problems pivotal in advanced machine learning. Furthermore, this method offers a simpler alternative to the Karush-Kuhn-Tucker conditions associated with inequality constraints, even though the core principles rooted in gradient fields remain essentially the same [14], [17].

To achieve these objectives, this paper will follow a methodical trajectory, beginning with a concise overview of optimization under equality constraints. This will pave the way for an in-depth, yet intuitively grasped, exposition on the gradient "field" – a pivotal mathematical instrument that has, regrettably, been overshadowed in the foundational learning of economics and social science students. The primary thrust of our presentation will be geared towards cultivating a profound, intuitive grasp of the core principles.

Much as in unconstrained optimization, where first derivatives are rendered null to pinpoint critical points for both maximization and minimization tasks, the same happens in constrained optimization. While higher order derivatives are typically

employed to differentiate between constrained maximization, minimization, and general critical points, our current analysis will limit itself to first-order conditions exclusively. The reason of this approach is our intent to familiarize readers to gradient methods, thereby sidestepping an intricate exploration of higher-order derivatives at this juncture.

The subsequent sections of this paper are systematically organized to facilitate understanding. Section 2 provides an overview of conventional techniques deployed to address optimization with equality constraints. We'll explore direct solutions in 2.1, graphical methods in 2.2, and the customary approach to the Lagrange multipliers method in 2.3. Section 3 presents the prerequisites concepts of orthogonality and orthonormal bases in vector spaces in 3.1, and the gradient field and its relationship with level curves and hypersurfaces in 3.2. Section 4 elaborates on the utility of the gradient field in automating the detection of local optima. We emphasize its centrality, especially of its stochastic version, during the machine learning training phase—even in sophisticated neural network configurations. Section 5 shows how gradients illuminate, and generalizes the graphical intuition, of the often mechanistic Lagrange multiplier method. First in problems with two variables and a single constraint in 5.1, then in problems with $n$-variables with a single constraint in 5.2, and finally in multi-dimensional, $n$-variable systems with several constraints in 5.3. Section 6 presents two numerical algorithms to find local optima with constraints. Section 7 finishes our discussion, providing a synthesis and conclusion of our exploration. Through this structured walkthrough, we aim to offer readers an intuitive and comprehensive understanding of our subject matter.

## 2. Standard ways of solving optimization problems with equality constraints

The basic problem of maximizing utility with a budget constraint is this: there are two goods, say food and clothes, a unit of food costs $P_f$ and a unit of clothes costs $P_c$, the budget is $i$, how much can the agent buy of each good so as to maximize her utility?

Suppose the utility function is:

$$U(f, c) = f^{\alpha} c^{\beta}, \tag{2.1}$$

where $f$ represents the units of food and $c$ the units of clothes that the person consumes. We assume that $f$ and $c$ can be any real, positive number. The exponents $\alpha$ and $\beta$ can also be any positive number in principle. However if we assume that they are numbers between zero and one, we have the more or less standard situation for normal goods in which, on the one hand, the more the better, and on the other, the more the person consumes of one item, the less utility she gets from an additional unit of that same item.

Assume she has $i$ dollars to spend on food and clothing, and define the budget function $I$ of $f$ units of food and $c$ units of clothes as:

$$I(f,c) = fP_f + cP_c. \tag{2.2}$$

We are modeling a one-period problem and we assume that the agent does not get any utility from keeping unspent money, therefore we will choose $f$ and $c$ such that $I(f,c) = i$.

The problem then is to choose $(f,c)$ so as to maximize the utility (2.1) subject to the budget constraint:

$$\max_{(f,c)} U(f,c), \tag{2.3}$$

$$I(f,c) = i. \tag{2.4}$$

Mathematically, problem (2.3-2.4) is a simple two dimensional problem of maximization with constraints, and it can be solved in many different ways. In the rest of this section we present three standard ways of solving it. Each of these methods has its merits and can be applied effectively under certain conditions. However, they also have limitations in terms of complexity and intuitive appeal. Understanding these methods, and their shortcomings, sets the stage for the development of a more intuitive approach to constrained optimization problems - a gradient-based approach that is at the heart of this paper.

### 2.1. Direct solution.

The direct way, given the particular form of the constraint, is to use (2.2) and (2.4) to find an explicit expression of $c$ in terms of $f$:

$$c = -\frac{P_f}{P_c} f + \frac{i}{P_c}. \tag{2.5}$$

Replacing (2.5) in (2.1) and rearranging:

$$U(f, c(f)) = \left(\frac{P_f}{P_c}\right)^{\beta} f^{\alpha} \left(-f + \frac{i}{P_f}\right)^{\beta}, \tag{2.6}$$

we end up with a trivial one-dimensional maximization problem. In Figure 1, $U(f,c(f))$ is plotted for $\alpha = \beta = 1/2$, $P_f = 1$, $P_c = 2$, $I = 10$. The symmetry of the graph makes clear that the maximization happens at $f = 5$. (2.5) then implies that $c = 2.5$, and of course $I = 5 \times 1 + 2.5 \times 2 = 10$, satisfying the constraint.

In general, the maximization of $U(f,c(f))$, where $c(f)$ is an *explicitly known* function of $f$, corresponds to the value of $f$ such that $dU/df = 0$:

$$\frac{dU(f,c(f))}{df} = \frac{\partial U(f,c)}{\partial f} + \frac{\partial U(f,c)}{\partial c}\frac{dc}{df} = 0. \tag{2.7}$$

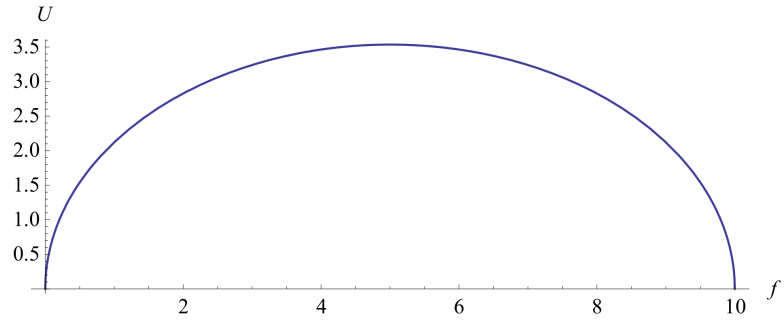


**Figure 1:** $U(f, c(f))$ in (2.6) for $\alpha = \beta = 1/2$, $P_f = 1$, $P_c = 2$, $I = 10$.

Given equation (2.5), we deduce that $dc/df = -P_f/P_c$. At the point of maximal utility, relationship (2.7) transforms into:

$$\frac{\partial U(f,c)}{\partial f} - \frac{\partial U(f,c)}{\partial c}\frac{P_f}{P_c} = 0. \tag{2.8}$$

Reconfiguring this equation, we derive:

$$\frac{\frac{\partial U(f,c)}{\partial f}}{P_f} = \frac{\frac{\partial U(f,c)}{\partial c}}{P_c}. \tag{2.9}$$

The implication of equation (2.9) is that, at the optimal utility juncture, the marginal utility per unit of expenditure remains consistent across all goods. In essence, if an individual reaches peak utility, the incremental joy or satisfaction they derive from allocating an additional dollar towards food precisely mirrors the joy from directing that dollar towards clothes.

Another representation, which we will refer to in subsequent sections, is depicted as:

$$\frac{\frac{\partial U(f,c)}{\partial f}}{\frac{\partial U(f,c)}{\partial c}} = \frac{P_f}{P_c}. \tag{2.10}$$

Equation (2.10) encapsulates the idea of the "marginal rate of substitution" in the context of prices. In essence, at the utility optimum, the ratio between the marginal utility derived from food and the marginal utility from clothes aligns perfectly with the ratio between the price of food and the price of clothes.

The direct method, employed for solving maximization problems with constraints, has the merit of necessitating only a basic understanding of calculus. However, its utility diminishes when grappling with problems encompassing more than two goods. This is due to the increasing challenges in graphical representation and the complications arising in resolving constraints, often diverging from the straightforward nature of equation (2.5).

### 2.2. Graphical method for maximization with constraints.

In the context of elementary two-dimensional problems, such as those described by equations (2.3-2.4), a graphical approach frequently offers an intuitive insight into the crux of the issue, facilitating an intuitive derivation of the pertinent equations. This section delves into this graphical technique, a staple in foundational economic textbooks, to shed light on maximization with constraints.

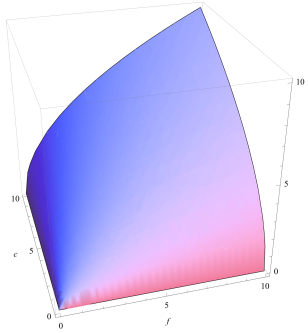

**Figure 2:** 3-D representation of the utility function (2.1), $\alpha = \beta = 1/2$, $P_f = 1$, $P_c = 2$, $I = 10$.

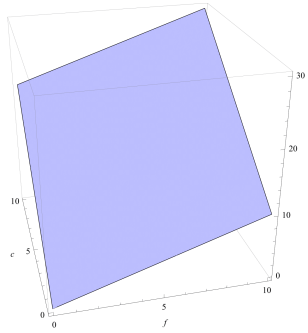

**Figure 3:** 3-D representation of the budget function (2.2), $\alpha = \beta = 1/2$, $P_f = 1$, $P_c = 2$, $I = 10$.

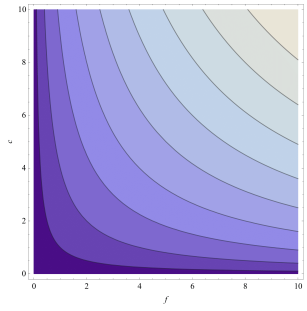

**Figure 4:** Utility function (2.1) level curves.

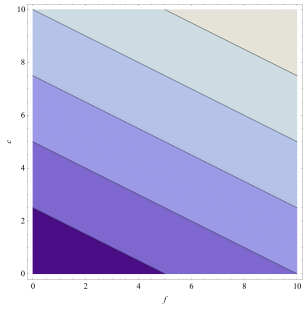

**Figure 5:** Budget function (2.2) level curves.

The utility function (2.1) and the budget function (2.2) can be portrayed in three dimensions, as in Figures 2-3, or in two dimensions, using their level curves, as in Figures 4-5[1]. The level curves representation proves particularly insightful in this context.

[1]The level curves of a function $G(f, c)$ represent the curves in the $(f, c)$ plane defined implicitly by the equation $G(f, c) = g$, where $g$ is a constant.

Comparing Figure 2 with Figure 4, and Figure 3 with Figure 5, it is apparent that in both cases, as we move in the up-right direction, the level curves denote increasing utility (Figure 4) and budget (Figure 5), respectively.

Let's recall the level curve for the budget function (2.2) corresponding to $I(f,c) = 10$ discussed in the previous subsection. We determined that utility is maximized for $f = 5$ and $c = 5/2$, resulting in a utility $U = \sqrt{5 \times 5/2} = 5/\sqrt{2}$. Hence, the corresponding budget and utility curves in the $(f,c)$ plane are, respectively:

$$2 \times c + 1 \times f = 10 \quad \Rightarrow \quad c = -\frac{1}{2} f + 5, \tag{2.11}$$

$$\sqrt{fc} = \frac{5}{\sqrt{2}} \quad \Rightarrow \quad c = \frac{25}{2} \frac{1}{f}. \tag{2.12}$$

In Figure 6, three utility level curves ($U = 5/\sqrt{2} - 0.5$, $U = 5/\sqrt{2}$, and $U = 5/\sqrt{2} + 0.5$), along with the budget line $I = 10$, are depicted in the $(f,c)$ plane.

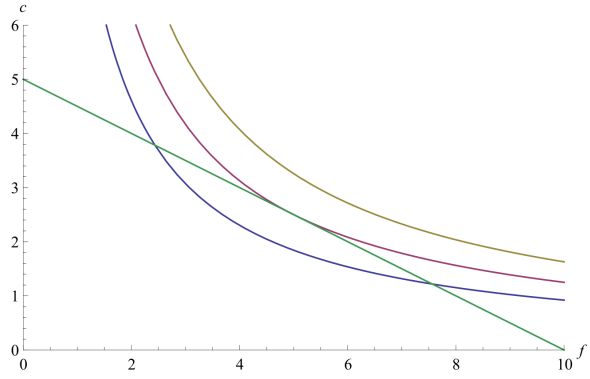


**Figure 6:** $U = 5/\sqrt{2} - 0.5$ (blue), $U = 5/\sqrt{2}$ (purple), $U = 5/\sqrt{2} + 0.5$ (yellow) and $I = 10$ (green).

Upon examining Figure 6, given the convex form of the level curves of the utility function, it's evident that the maximization of utility constrained by the budget will occur at the point on the $(f,c)$ plane where the budget line is tangent to a level curve of the utility function, intercepting it exactly once.

If the level curve of the utility function intercepts the budget line at two points, as exemplified by the blue curve, we can enhance utility by opting for higher level curves. If it doesn't intersect the budget line, as with the yellow curve, it isn't compatible with our budget constraint. Hence, the optimal scenario is the level curve of the utility function intercepting the budget line only once, as demonstrated by the purple curve. Given that the level curves are smooth, at this point, the budget line must align with the tangent line of the utility function's level curve.

Now let's convert this graphical insight into mathematical equations. The level curves of the utility function in the $(f, c)$ plane represent curves $c(f)$ implicitly defined by the equation $U(f, c(f)) = u$, where $u$ is a constant. According to the rules of differentiation for implicit functions, the slope $dc/df$ of such a curve is given by:

$$\frac{dc}{df} = -\frac{\frac{\partial U}{\partial f}}{\frac{\partial U}{\partial c}}. \tag{2.13}$$

The graphical interpretation suggests that at the optimal point, this slope matches the slope of the budget line (2.5),

$$\frac{dc}{df} = -\frac{P_f}{P_c}. \tag{2.14}$$

By equating equations (2.13) and (2.14), we deduce:

$$\frac{\frac{\partial U}{\partial f}}{\frac{\partial U}{\partial c}} = \frac{P_f}{P_c}. \tag{2.15}$$

In Section 2.1, we reduced the problem to a one-dimensional optimization scenario, which is addressed by equation (2.10). The graphical method employed in this section yields equation (2.15). At first glance, this appears analogous to (2.10). However, it's important to note that (2.15) resides within the two-dimensional $(f, c)$ plane, thus requiring an extra equation to conclusively determine a solution.

To discover this secondary equation, let's reconsider Figures 4-5 and 6. All level lines of the budget function (represented by (2.5) for varying $i$ values) share the same slope $-P_f/P_c$ (see Figure 5). As such, the right side of (2.15) does not uniquely specify our particular budget constraint. From this, it becomes evident that the second equation should be the explicit budget constraint itself.

Note that every level curve of the utility function (as shown in Figure 4) has a point where the tangent line bears a slope of $-P_f/P_c$. Therefore, equation (2.15) defines a curve $c^*(f)$, where the points signify the optimal utility for all potential budget values. To pinpoint the optimal budget, we must identify where this curve intersects with our budget line.

Hence, our constrained maximization problem is resolved by finding the solution to a pair of equations:

$$\frac{\frac{\partial U}{\partial f}}{\frac{\partial U}{\partial c}} = \frac{P_f}{P_c}, \tag{2.16}$$

$$fP_f + cP_c = i. \tag{2.17}$$

Equation (2.17) and the right-hand side of equation (2.16) explicitly rely on the specific structure of the budget function (2.2). However, the concept that the level

curves $c_U(f)$ of the function $U(f,c)$ being maximized and the level curves $c_I(f)$ of the constraining function $I(f,c)$ should be tangent at the constrained optimum is more general than our particular example might suggest. It applies to generic, smooth functions $U(f,c)$ and generic, smooth constraints $I(f,c) = i$.

Let's consider a new problem with the functions $U$ and $I$ given by

$$\max_{(f,c)} U = \sqrt{uf}, \tag{2.18}$$

$$I(f,c) = \frac{f^2}{5\sqrt{2}} + c = \frac{15}{2\sqrt{2}}. \tag{2.19}$$

In this case, $U$ is the same as before, but we treat it as a generic function to be maximized under the generic constraint $I(f,c) = 15/(2\sqrt{2})$. Figure 7 illustrates the same level curves of $U(f,c)$ (blue, purple, and yellow) as shown in Figure 6, along with the green curve corresponding to the level curve $I(f,c) = 15/(2\sqrt{2})$.

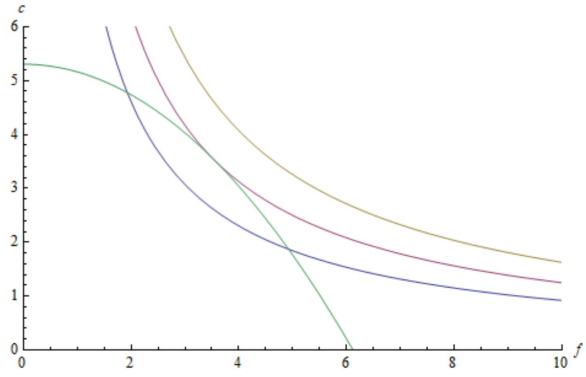


**Figure 7:** $U = 5/\sqrt{2} - 0.5$ (blue), $U = 5/\sqrt{2}$ (purple), $U = 5/\sqrt{2} + 0.5$ (yellow) and $I = 15/(2\sqrt{2})$ given in equation (2.19).

It's evident that the point of tangency between the purple level curve of $U$ and the green level curve of $I$, which has coordinates

$$f = c = \frac{5}{\sqrt{2}}, \tag{2.20}$$

represents the maximum of $U$ subject to the constraint $I(f,c) = 15/(2\sqrt{2})$. As we move along the allowed curve $c = -f^2/(5\sqrt{2}) + 15/(2\sqrt{2})$ from left to right, the values of $U$ increase up to the tangent point in (2.20), after which they start to decrease. Since $U$ and $I$ are smooth functions around the tangent point, this implies that $U$ should not change in value (to first order) under small enough displacements along the level curve of $I$ at that point. This means that small displacements along the level curve of $I$ at the tangent point are also along a level curve of $U$, indicating that the level curves of $I$ and $U$ coincide at the tangent point.

We now need to generalize the right-hand side of (2.16) to include the slope of a generic constraint given implicitly by the level curve $I(f,c) = i$ of a function $I(f,c)$. But we know from (2.13) that the slope of any level curve $c_I(f)$ defined implicitly by the function $I(f,c)$ is given by

$$\frac{dc_I}{df} = -\frac{\frac{\partial I}{\partial f}}{\frac{\partial I}{\partial c}}. \tag{2.21}$$

Thus, the generalization of (2.16) is

$$\frac{\frac{\partial U}{\partial f}}{\frac{\partial U}{\partial c}} = \frac{\frac{\partial I}{\partial f}}{\frac{\partial I}{\partial c}}. \tag{2.22}$$

The generalization of equation (2.17) is as necessary in the generic case as it was in the specific one. The single equation (2.22), with two variables $f$ and $c$, can't pinpoint a unique solution. Therefore, we explicitly need to impose the equation $I(f,c) = i$, for the value $i$ corresponding to our constraint. Thus, the generalization of the system (2.16-2.17) is

$$\frac{\frac{\partial U}{\partial f}}{\frac{\partial U}{\partial c}} = \frac{\frac{\partial I}{\partial f}}{\frac{\partial I}{\partial c}}, \tag{2.23}$$

$$I(f,c) = i. \tag{2.24}$$

The graphical method discussed in this section holds an advantage due to its intuitive nature. However, in its current form, it does not generalize to higher, arbitrary dimensions, where we can no longer "visualize" the problem. Nonetheless, as we will delve into in Section 5, the introduction of some mathematical concepts, often not covered in typical courses for economists and other social scientists, will allow us to extend our intuition to any number of dimensions.

### 2.3. The Lagrange multipliers method: Standard treatment.

The two methods we've discussed thus far struggle to generalize to multiple dimensions and constraints. It so happens that the most critical economic applications involving maximization with constraints typically encompass many dimensions and constraints. Thus, we'll introduce a method conceived by the Italian mathematician Joseph-Louis Lagrange (born on January 25, 1736), which effectively generalizes to multiple dimensions and constraints.

Let's consider the function:

$$L(f,c,\lambda) = U(f,c) - \lambda(I(f,c) - i), \tag{2.25}$$

where $\lambda$ is a new variable, termed the Lagrange multiplier. We aim to optimize $L$ concerning the *three* variables $f$, $c$, and $\lambda$:

$$\frac{\partial L}{\partial f} = \frac{\partial U}{\partial f} - \lambda \frac{\partial I}{\partial f} = 0, \tag{2.26}$$

$$\frac{\partial L}{\partial c} = \frac{\partial U}{\partial c} - \lambda \frac{\partial I}{\partial c} = 0, \tag{2.27}$$

$$\frac{\partial L}{\partial \lambda} = I(f, c) - i = 0. \tag{2.28}$$

These are three equations with three unknowns. In all but special cases, there will be isolated solutions. The solutions for $f$ and $c$ correspond to the optimum of $U(f, c)$ under the constraint $I(f, c) = i$. To see this, simply shift the $\lambda$ term in (2.26) and (2.27) to the right, and divide (2.26) by (2.27), giving us (2.23). And (2.28) is identical to (2.24).

This method reproduces the correct equations and readily generalizes to multiple dimensions and constraints, as we'll explore later. While standard textbooks often present and justify this method in a formal manner—primarily noting that it does reproduce the correct equations for simple problems, just as we've demonstrated—they seldom elaborate on why it works or explain the necessity of the additional $\lambda$ variable.

In the remainder of this paper, we aim to introduce the concept of gradient field, which among other things will help us fill the Lagrange multipliers method with the same intuitive appeal as the graphical method discussed in Section 2.2. Furthermore, with a touch of abstraction, we will extend this geometric intuition to multiple dimensions, even where traditional visualization is virtually impossible.

Throughout the subsequent sections of this paper, our primary objective is to introduce the concept of the gradient field. This will, in turn, enrich the Lagrange multipliers method with an intuitive clarity akin to the graphical approach delineated in Section 2.2. Moreover, by incorporating a modicum of abstraction, we aim to extrapolate this geometric intuition to multi-dimensional and multi constraints contexts, even in instances where conventional visual representation is inherently unfeasible.

## 3. Some mathematical background

To realize the objectives of this paper, it is imperative to understand the notions of orthogonality, basis vectors, and gradients. Although often given cursory treatment in many standard textbooks tailored for economists and social scientists (gradients in particular), these topics carry profound importance across various domains, notably in machine learning and data analysis. This section aims to offer a succinct overview of these fundamental concepts.

In Section 2, we adopted the symbols $f$ (for food) and $c$ (for clothes) as our independent variables, serving to contextualize the archetypical economic problem

delineated therein. Moving forward, we will transition to the more generic notation $x_1, x_2, \ldots, x_n$ when referencing independent variables, barring specific examples. Such a shift will facilitate the broadened application of the concepts and results to a wide range of economic and mathematical problems.

### 3.1. Orthogonality and orthonormal basis.

Let's consider two two-dimensional vectors (represented in bold typeface):

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}. \tag{3.1}$$

When plotting these vectors on a two-dimensional plane, it is easy to visually check if they are perpendicular, as their angle can be 90° or 270°. However, determining perpendicularity becomes more challenging as the vectors involve higher dimensions. Thankfully, we can employ an accessible algebraic structure to automate this determination, and we will refer to this concept as "orthogonality".

The scalar product (or "dot product") of two vectors, denoted as $\mathbf{a}\cdot\mathbf{b}$, is defined as the sum of the products of their corresponding coordinates. For the vectors in (3.1), it can be expressed as:

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2. \tag{3.2}$$

This definition straightforwardly extends to two $n$-dimensional vectors as follows:

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i. \tag{3.3}$$

Now, consider the two standard vectors $\hat{\mathbf{e}}_\mathbf{1} = (1, 0)$ and $\hat{\mathbf{e}}_\mathbf{2} = (0, 1)$. These vectors are orthogonal to each other and have unit length. Correspondingly:

$$\hat{\mathbf{e}}_\mathbf{1} \cdot \hat{\mathbf{e}}_\mathbf{1} = 1 \times 1 + 0 \times 0 = 1, \quad \hat{\mathbf{e}}_\mathbf{2} \cdot \hat{\mathbf{e}}_\mathbf{2} = 0 \times 0 + 1 \times 1 = 1, \tag{3.4}$$

and

$$\hat{\mathbf{e}}_\mathbf{1} \cdot \hat{\mathbf{e}}_\mathbf{2} = 1 \times 0 + 0 \times 1 = 0. \tag{3.5}$$

Equation (3.4) is not a coincidence, the scalar product of any vector with itself equals the square of its length, which we denote as $|\boldsymbol{a}|^2$:

$$\mathbf{a} \cdot \mathbf{a} = |\boldsymbol{a}|^2, \tag{3.6}$$

and (3.5) is not a coincidence either, the scalar product of two vectors equals zero if and only if they are orthogonal. This algebraic generalization captures the notion of orthogonality we discussed earlier.

Additionally, for any two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$:

$$\mathbf{a} \cdot \mathbf{b} = |\boldsymbol{a}||\boldsymbol{b}| \cos\theta, \tag{3.7}$$

where $\theta$ is the angle between them. Since $-1 \leq \cos\theta \leq 1$, fixing the lengths of the vectors $|\boldsymbol{a}|$ and $|\boldsymbol{b}|$, we can observe the following:

1. The maximum value of $\mathbf{a}\cdot\mathbf{b} = |\boldsymbol{a}||\boldsymbol{b}|$ is achieved when $\theta = 0$ (they are aligned).
2. The minimum value of $\mathbf{a}\cdot\mathbf{b} = -|\boldsymbol{a}||\boldsymbol{b}|$ occurs when $\theta = 180°$ (they are anti-aligned).
3. $\mathbf{a}\cdot\mathbf{b} = 0$ when $\theta = 90°$ or $\theta = 270°$, indicating orthogonality.

The notions of alignment, anti-alignment, orthogonality, or, in general, the angle $\theta$ between two vectors is generalized and automated by (3.7) for any dimension of the underlying vector space.

In an $n$-dimensional vector space, any set of $n$ linearly independent vectors $\{\boldsymbol{v_i}\}$, $\boldsymbol{i} = 1, \ldots, n$, forms a basis, and any vector $\boldsymbol{w}$ can be uniquely expressed in this basis as

$$\boldsymbol{w} = \sum_{i=1}^{n} w_i \boldsymbol{v_i}, \tag{3.8}$$

where the numbers $w_i$ are the "coordinates" of $\boldsymbol{w}$ in the basis $\{\boldsymbol{v_i}\}$.

The basis $\boldsymbol{v_i}$ is called "orthonormal" if the basis vectors are orthogonal to each other and have unit length:

$$\boldsymbol{v_i}\cdot\boldsymbol{v_j} = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases} \tag{3.9}$$

From (3.8), (3.9), and the linearity of the scalar product, it follows that if the basis is orthonormal, the $i$th coordinate of a vector $\boldsymbol{w}$ in this basis is given by:

$$w_i = \boldsymbol{v_i}\cdot\boldsymbol{w}. \tag{3.10}$$

This generalizes the standard notion of, for example, the vector $\boldsymbol{w} = (2, 3)$ having a coordinate "2" along the basis vector $\hat{\boldsymbol{e}}_{\mathbf{1}} = (1, 0)$ and coordinate "3" along the basis vector $\hat{\boldsymbol{e}}_{\mathbf{2}} = (0, 1)$, as clearly $\hat{\boldsymbol{e}}_{\mathbf{1}}\cdot\boldsymbol{w} = 2$ and $\hat{\boldsymbol{e}}_{\mathbf{2}}\cdot\boldsymbol{w} = 3$.

### 3.2. Gradient and level curves and hypersurfaces.

The "gradient" of a sufficiently smooth function of $n$ variables $U(x_1, \ldots, x_n)$, denoted by $\vec{\nabla}U$, is a vector "field" defined as:

$$\vec{\nabla}U(x_1, \ldots, x_n) = \begin{pmatrix} \frac{\partial U}{\partial x_1} \\ \vdots \\ \frac{\partial U}{\partial x_n} \end{pmatrix}. \tag{3.11}$$

In essence, the $i$th component of the vector $\vec{\nabla}U$ at the point $(x_1, \ldots, x_n)$ represents the $i$th derivative of $U$ evaluated at that point. Normally we think of vectors as having a common origin, associated with the zero vector. We use the term "vector field" rather than the standard "vector space", when there is a vector associated with each point in $(x_1, x_2, \ldots, x_n)$, as is the case in (3.11).

For instance, consider the function $U(x_1, x_2) = 50-(x_1^2+x_2^2)$, its level curves, and the corresponding implicit curves in the $(x_1, x_2)$ plane, as illustrated in Figure 8.

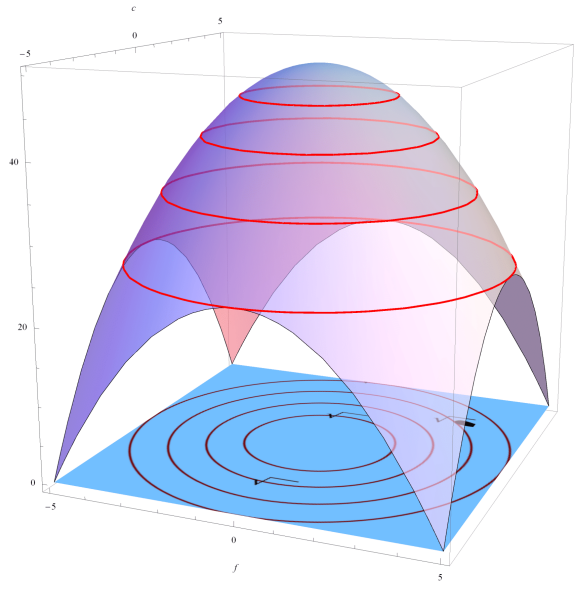


**Figure 8:** $U(x_1, x_2) = 50 - (x_1^2 + x_2^2)$ and its level curves $U(x_1, x_2) = u$, with $u = 46, 41, 34, 25$. In the plane $(f, c)$, the curves implicitly defined by these level curves are shown.

The gradient of this function is given by:

$$\vec{\nabla}U(x_1, x_2) = \begin{pmatrix} -2x_1 \\ -2x_2 \end{pmatrix}. \tag{3.12}$$

Thus, at the point $(2, 3)$ in the $(x_1, x_2)$ plane, the gradient is the vector $(-4, -6)$, and at $(1, -1)$, it is $(-2, +2)$.

In Figure 9, we present the gradient (3.12), rescaled for visual clarity, and the level curves of $U$ in the $(x_1, x_2)$ plane implicitly defined by $U(x_1, x_2) = u$ for various values of $u$.

Looking at Figure 9 it becomes evident that the gradient vector field of this function exhibits the following properties:

1. It is perpendicular (orthogonal) to the level curves at every point.
2. The direction of the gradient is the one of the steepest increase in $U$, known as the direction of steepest ascent. Conversely, the opposite direction is the one of the steepest decrease in $U$, known as the direction of steepest descent.

3. The magnitude of the gradient is proportional to the "steepness" of $U$ at each point.

4. In particular, at a local maximum or minimum, where the steepness is zero, the gradient is also zero.

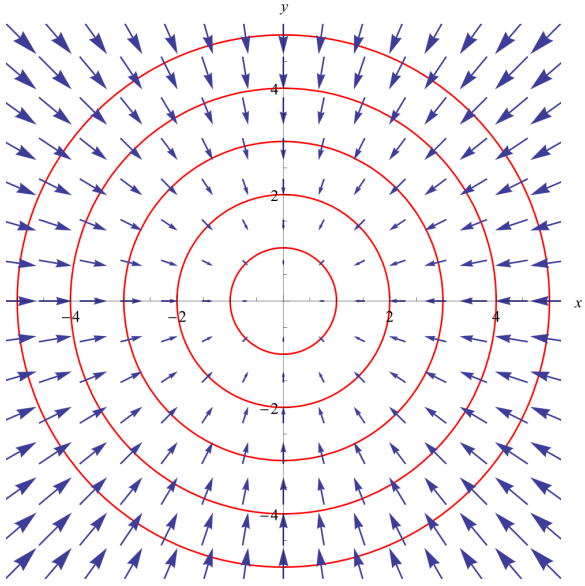


**Figure 9:** $\vec{\nabla}U(x_1, x_2) = (-2x_1, -2x_2)$ (rescaled), and curves in the $(x_1, x_2)$ plane defined by $U(x_1, x_2) = u$ for different values of $u$.

These properties are valid for every sufficiently smooth function $U(x_1, x_2)$, and they extend to any number of variables. All of these properties can be easily proved with the concepts presented in this section. Let's explore how.

If at point $(x_1, x_2)$, $U(x_1, x_2) = u$, for any small enough displacement $(dx_1, dx_2)$, basic calculus indicates that the change in $U$ is well approximated at first order by:

$$dU = \frac{\partial U}{\partial x_1} dx_1 + \frac{\partial U}{\partial x_2} dx_2, \tag{3.13}$$

where the partial derivatives are evaluated at the same point $(x_1, x_2)$.

From (3.2), (3.7), and (3.11), we can see that the right-hand side of (3.13) can be interpreted as the scalar product between the gradient $\vec{\nabla}U$ and the displacement vector $\boldsymbol{dx} = (dx_1, dx_2)$:

$$dU = \begin{pmatrix} \partial U/\partial x_1 \\ \partial U/\partial x_2 \end{pmatrix} \cdot \begin{pmatrix} dx_1 \\ dx_2 \end{pmatrix} = \vec{\nabla}U \cdot \boldsymbol{dx} = |\vec{\nabla}U||\boldsymbol{dx}| \cos\theta, \tag{3.14}$$

where $\theta$ is the angle between the gradient vector $\vec{\nabla}U$ at $(x_1, x_2)$ and the displacement vector, that can have a priori an arbitrary direction.

Now, let's focus on the right-hand side of (3.14):

1. If we want to our displacement to lie within the same level curve $U(x_1, x_2) = u$, the only way to do so is to make a displacement orthogonal to the gradient $\vec{\nabla}U$ at that point, so that $\cos\theta = 0$ and $dU = 0$. This proves the first point.

2. Fixing the magnitude of the displacement $|\boldsymbol{dx}|$, $dU$ is maximized when $\theta = 0$ ($\cos\theta = 1$), meaning the displacement is in the same direction as the gradient. Conversely, $dU$ is minimized when $\theta = 180°$ ($\cos\theta = -1$), indicating that the displacement is in the opposite direction of the gradient. This proves the second and third points.

3. If the gradient is zero, $dU = 0$ for small enough displacements regardless of their direction. This behavior is expected at local maximum or minimum points, as well as in general at critical points where all the first derivatives vanish. This proves the last point.

One can begin to appreciate the power of vector calculus when realizing that even though the first equality in (3.14) is explicitly in 2 dimensions, the last two equalities hold in any dimension. Indeed, since (3.7) holds for any pair of vectors in any-dimensional vector space:

$$dU = \sum_{i=1}^{n} \frac{\partial U}{\partial x_i} dx_i = \vec{\nabla}U \cdot \boldsymbol{dx} = |\vec{\nabla}U||\boldsymbol{dx}| \cos\theta. \tag{3.15}$$

Thus, although initially motivated in the two-dimensional case shown in Figure 9, the gradient is orthogonal to the "level hyper-surface" $U(x_1, \ldots, x_n) = u$ of any sufficiently smooth function $U$ in any dimension.

## 4. Gradient descent and related optimization algorithms

Finding the local maximum of a function $f$ mirrors the process of locating the local minimum of $-f$. Given this, our focus in this section will primarily be on identifying local minima.

In elementary calculus involving a single variable, students are introduced to the concept that identifying critical points involves taking derivatives, setting them equal to zero, and subsequently solving the resultant equation. To illustrate this with a rudimentary example, consider the function

$$f(x) = \frac{x^2}{2a}, \quad a > 0. \tag{4.1}$$

From this, we deduce

$$f' = \frac{x}{a} = 0, \tag{4.2}$$

which gives $x = 0$, representing a minimum.

In a similar vein, for functions of multiple variables, say $f(x_1, \ldots, x_n)$, critical points arise where all first partial derivatives are zero—that is, where the gradient vanishes. Yet, when dealing with functions of several variables or even intricate single-variable functions, deriving an analytical solution to the corresponding equations is typically elusive. Consequently, there's a need for algorithms that can approximate these local minima effectively.

Many modern algorithms, especially those employed in machine learning, stem from a direct implication of the four gradient field properties elaborated in the preceding section. To determine the local minimum of a function $f(x_1, \ldots, x_n)$, the following steps can be adopted:

1. Choose an initial point $\mathbf{x_0} = (x_{01}, \ldots, x_{0n})$ based on whichever criteria best suits the problem at hand.
2. Move in tiny steps, which are negatively proportional to the gradient direction ($\alpha, \varepsilon > 0$):
$$\text{if } |\nabla f(\mathbf{x_t})| > \varepsilon, \quad \mathbf{x_{t+1}} = \mathbf{x_t} - \alpha \nabla f(\mathbf{x_t}).$$
3. Conclude the process when the gradient magnitude is sufficiently small, i.e., if $|\nabla f(\mathbf{x_t})| \leq \varepsilon$.

This is the famous "gradient descent" minimization algorithm.

Taking, for instance, the rudimentary function $f(x) = \frac{x^2}{2a}$, where the gradient is simply the derivative $\nabla f = f' = \frac{x}{a}$, we get:

1. $\min_x f(x) = \frac{x^2}{2a}, \quad \nabla f = f' = \frac{x}{a}$.
2. Select an initial point, $x_0$.
3. $$x_{t+1} = x_t - \alpha \frac{x_t}{a} = \left(1 - \frac{\alpha}{a}\right) x_t, \quad \text{or equivalently} \quad x_t = \left(1 - \frac{\alpha}{a}\right)^t x_0.$$
4. End the process when
$$|\nabla f(x_{t+1})| = \left| \left(1 - \frac{\alpha}{a}\right)^t \frac{x_0}{a} \right| \leq \varepsilon.$$

The termination criterion implies that the process should be halted after:

$$t = \frac{\ln\left(\frac{|x_0|}{a\varepsilon}\right)}{\ln\left(\frac{1}{|1-\frac{\alpha}{a}|}\right)} \approx \ln\left(\frac{|x_0|}{a\varepsilon}\right) \frac{a}{\alpha}, \quad \text{for } \alpha \ll a. \tag{4.3}$$

Note that the third point implies that $x_t \to 0$ as $t \to \infty$ if $\alpha$ is sufficiently small. In essence, the algorithm converges to the correct global minimum. Equation (4.3) suggests that the number of iterations required to achieve a specific accuracy increases logarithmically with the inverse of the desired error $\varepsilon$ (i.e., as the desired error decreases, the number of iterations increase), and linearly with the inverse of the step size $\alpha$ (the smaller the step size, the greater the number of iterations required).

Further, if $\alpha$ is too large, the algorithm will diverge. Specifically, the third point insinuates that if $|1 - \alpha/a| > 1$, which occurs when $\alpha > 2a$ (keeping in mind that both $a$ and $\alpha$ are positive), the algorithm will diverge.

Although the details can vary depending on the function and the number of variables, generally, if the algorithm converges, the iterations needed to approximate a (local) minimum increase with the inverse of the desired accuracy and the inverse of the step size. This presents a trade-off: a smaller step size may require more computational time to converge to the desired accuracy, whereas a larger one might prevent convergence altogether.

It is crucial to understand that the algorithm will not always find the *global* minimum. This is only assured for *convex optimization* [7]. Typically, the algorithm may get caught in a local minimum, as seen in neural network training in machine learning. For more complex scenarios, pinpointing the global minimum can be computationally hard.

An essential property of the gradient is its linearity:

$$\vec{\nabla}(af(\boldsymbol{x}) + bg(\boldsymbol{x})) = a\vec{\nabla}f(\boldsymbol{x}) + b\vec{\nabla}g(\boldsymbol{x}). \tag{4.4}$$

In machine learning, the training often entails minimizing a "cost function" that is the mean of cost functions for each training set element:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) = \min_{\boldsymbol{x}} \frac{1}{m} \sum_{i=1}^{m} f_i(\boldsymbol{x}). \tag{4.5}$$

Due to linearity, the gradient of $f(\boldsymbol{x})$ is the mean of the gradients of each $f_i(\boldsymbol{x})$. Therefore, step 2 of the "gradient descent" minimization algorithm can be reexpressed as:

$$\boldsymbol{x_{t+1}} = \boldsymbol{x_t} - \alpha\vec{\nabla}f(\boldsymbol{x_t}) = \boldsymbol{x_t} - \alpha\frac{1}{m}\sum_{i=1}^{m}\vec{\nabla}f_i(\boldsymbol{x}). \tag{4.6}$$

The gradient of $f$ has a magnitude and direction equivalent to the mean of $m$ vectors $\vec{\nabla}f_i(\boldsymbol{x})$, which generally point in varying directions. If $m$ is large, computing the average can be computationally costly. It is expected that if we randomly split the $m$ training set elements into $s$ "mini-batches" of $b$ elements, with $m = s \times b$, and take $s$ steps in the direction of each mini-batch's average gradient, the trajectory $\boldsymbol{x_t}$ will, on average, align with the trajectory in equation (4.6).

Ultimately, the choice is between $T$ "total gradient" steps vs. $sT$ "mini-batch gradient" steps. Empirically, many practical situations, particularly when minimizing deep neural networks' cost functions, favor the latter. This method is termed "stochastic gradient descent" and forms the foundation of cutting-edge minimization techniques in machine learning.

## 5. Gradients and the intuition behind the Lagrange multiplier method

### 5.1. For two variables, one constraint.

In Section 2.3, we derived Lagrange's equations (2.26-2.28) to solve the maximization problem for the function $U(x_1, x_2)$, now employing the generic independent variables $x_i$, subject to the constraint $I(x_1, x_2) = i$. As noted earlier, many economics textbooks present these equations as a formal procedure, closely mirroring the outcomes from the graphical method outlined in Section 2.2. However, they often overlook the innate gradient-field based intuition behind why the Lagrange multiplier technique is effective, particularly its seamless adaptability to scenarios with higher dimensions and multiple constraints—a feature the graphical method lacks.

The goal of this section is to elucidate the underlying intuition behind the Lagrange multipliers using the mathematical preliminaries we introduced earlier.

Equations (2.26-2.27) can be succinctly represented as:

$$\vec{\nabla} U = \lambda \vec{\nabla} I. \tag{5.1}$$

This equation unveils a profound geometric insight: at the point of constrained optimization, the gradients of $U$ and $I$ exhibit proportionality. Specifically, depending on the sign of the Lagrange multiplier $\lambda$ (either positive or negative), these gradients align in parallel or antiparallel directions. In essence, the Lagrange multiplier serves as the constant of proportionality linking the gradients of $U$ and $I$ at the optimal point.

Even in the context of the two-dimensional analysis of Section 2.2, the compact form given by equation (5.1) provides deeper insights than the observation made there: Namely, that the level curves of both $U$ and $I$ coincide at the constrained optimum. When these level curves align, their tangent vectors will be oriented in the same direction at this optimum. However, we know that $\vec{\nabla} U$ is orthogonal to $U$'s level curves, and when viewing $I(x_1, x_2)$ as a distinct function, $\vec{\nabla} I$ is orthogonal to its level curves. In a two-dimensional space, given a direction tangent to the level curves, there's only one orthogonal direction. This implies that the gradients $\vec{\nabla} U$ and $\vec{\nabla} I$ must be proportional to each other.

For instance, at the constrained optimum (2.20), the gradients of the functions $U$ and $I$ given in equations (2.18-2.19) are:

$$\vec{\nabla}U = \frac{1}{2}\begin{pmatrix}\sqrt{x_2/x_1}\\ \sqrt{x_1/x_2}\end{pmatrix} = \begin{pmatrix}0.5\\ 0.5\end{pmatrix}, \tag{5.2}$$

$$\vec{\nabla}I = \begin{pmatrix}\sqrt{2}\,x_1/5\\ 1\end{pmatrix} = \begin{pmatrix}1\\ 1\end{pmatrix}. \tag{5.3}$$

Consequently, we find:

$$\vec{\nabla}U = 0.5\,\vec{\nabla}I. \tag{5.4}$$

This implies that $\lambda = 0.5$ in equation (5.1).

This observation elucidates why we require the new variable $\lambda$, which remained somewhat obscure in Section 2.2. Indeed, the preceding argument indicates that the *directions* of $\vec{\nabla}U$ and $\vec{\nabla}I$ must coincide, but it does not necessarily imply any specific relationship between their magnitudes. The gradient's geometric property highlighted in point 1 of Section 3.2, that it is orthogonal to the level curves at every point, does not provide information about the relationship between the steepness of $U$ and $I$ at the constrained optimum. Consequently, a new variable becomes necessary to establish this relationship: $\vec{\nabla}U$ and $\vec{\nabla}I$ should be proportional to each other, as we see in (5.4), but no assumptions about the magnitudes come into play, hence equation (5.1).

Equation (5.1) offers a meaningful interpretation of the Lagrange multiplier $\lambda$: when the constraint $I = i$ is relaxed to $I = i + di$, we have

$$dI = \vec{\nabla}I \cdot \boldsymbol{dx} = \frac{\vec{\nabla}U}{\lambda} \cdot \boldsymbol{dx} = \frac{dU}{\lambda}, \quad \text{or} \quad dU = \lambda\, dI. \tag{5.5}$$

In simple terms, if the constraint $I = i$ is slightly relaxed to $I = i + di$, the change in the value of $U$ at the new optimal point is proportional to the change in the constraint, and the constant of proportionality is the Lagrange multiplier. In the context of $U$ representing utility and $I$ as the budget function, $\lambda$ signifies the change in utility per unit change in the budget constraint.

### 5.2. For $n$ variables, one constraint.

The real strength and beauty of the Lagrange multiplier method emerge when we generalize it to functions with a higher number of variables and more constraints. Consider a function $U$ with $n$ variables subject to a single constraint:

$$\max_{(x_1,\ldots,x_n)} U(x_1,\ldots,x_n), \tag{5.6}$$

$$I(x_1,\ldots,x_n) = i. \tag{5.7}$$

To employ the Lagrange multiplier method, we formulate the Lagrangian:

$$L(x_1,\ldots,x_n,\lambda) = U(x_1,\ldots,x_n) - \lambda\,(I(x_1,\ldots,x_n) - i)\,. \tag{5.8}$$

The Lagrangian $L$ is a function of $n+1$ variables, which includes the $n$ original variables $x_1, \ldots, x_n$ and the Lagrange multiplier $\lambda$. Our goal is to optimize this Lagrangian with respect to all these variables, resulting in the following system of equations:

$$\vec{\nabla} U = \lambda \vec{\nabla} I, \tag{5.9}$$

$$I = i. \tag{5.10}$$

Equation (5.9) is an expansion of our earlier geometric insight from equation (5.1). In this generalized scenario, the gradients in (5.9) each represent a system of $n$ equations—one for each variable.

The constraint given by equation (5.10) implicitly defines an $n-1$ dimensional hypersurface. In theory, echoing our approach from Section 2.1, one could attempt to solve explicitly for $x_n$ in terms of $x_1, \ldots, x_{n-1}$, substitute this expression into equation (5.9), and then determine the gradients utilizing the chain rule. However, this technique frequently falters, especially when confronting intricate functions.

Another strategy, reminiscent of our approach in Section 2.2, is to examine the tangency between the hypersurfaces associated with the constraint and the objective function at the optimum. Yet, a predicament arises: When two $n-1$ dimensional hypersurfaces are tangent, they touch along infinitely many directions within the $n-1$ dimensional tangent hyperplane. This fact deviates from our simpler observation in Section 5.1 for the two-dimensional case, where there's solely one tangent direction. Describing these infinite tangential directions proves intricate.

Instead, a more fruitful approach is to discern that all these infinite directions within the $n-1$ dimensional tangent hyperplane are orthogonal to a *singular* direction, the one perpendicular to the tangent hyperplane. Given that the gradient of $I$ is perpendicular to the level hypersurface $I = i$, and likewise, the gradient of $U$ is orthogonal to the level hypersurface of $U$ correlated with the constrained critical point—keeping in mind that these level hypersurfaces touch each other at the critical point in such a way that they share the aforementioned tangent hyperplane—it follows that the gradients $\vec{\nabla} U$ and $\vec{\nabla} I$ must align in proportion. Note that there's no inherent need for their magnitudes to correlate, which guides us to the necessity of the Lagrange multiplier in equation (5.9), adjusting for possible gradient magnitude disparities.

Such insights illuminate why the Lagrange method extends gracefully to $n$ dimensions, providing a cohesive algebraic framework to capture the underlying geometric relationships.

Let us consider a simple example of minimization with three variables and a linear constraint:

$$\min U(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2, \tag{5.11}$$

$$I(x_1, x_2, x_3) = x_1 = 0.8. \tag{5.12}$$

In Figure 10 we see the spherical level surfaces of $U$ and its radial gradient $\vec{\nabla}U = (2x_1, 2x_2, 2x_3)$, and in Figure 11 we see the planar level surfaces of $I$ and its gradient $\vec{\nabla}I = (1, 0, 0)$.

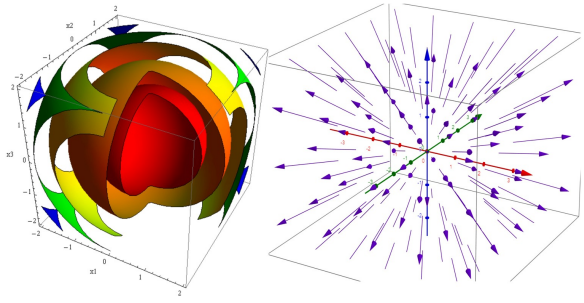

**Figure 10:** Left: level surfaces of the function $U = x_1^2 + x_2^2 + x_3^2$. Right: gradient field of $U$, $\vec{\nabla}U = (2x_1, 2x_2, 2x_3)$.

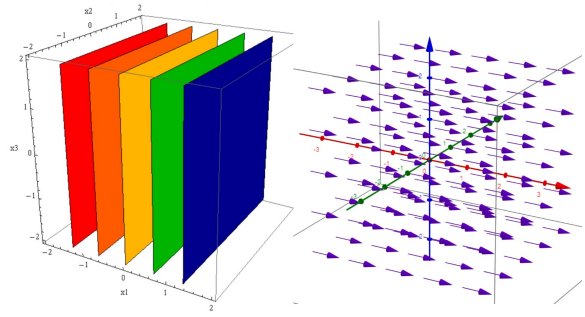

**Figure 11:** Left: level surfaces of the function $I = x_1$. Right: gradient field of $I$, $\vec{\nabla}I = (1, 0, 0)$.

The gradient equation (5.9) becomes

$$\begin{pmatrix} 2x_1 \\ 2x_2 \\ 2x_3 \end{pmatrix} = \lambda \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \Rightarrow x_2 = x_3 = 0 \; x_1 = \frac{\lambda}{2}. \tag{5.13}$$

The equations $x_2 = x_3 = 0$, $x_1 = \lambda/2$ determine the infinite minima for all possible values of the constraint $I(x_1, x_2, x_3) = x_1 = i$. For the specific value $x_1 = 0.8$ in (5.12), $\lambda$ is fixed to 1.6, where the spherical level surface of $U$ is tangent to the plane $x_1 = 0.8$, see Figure 12, and their gradients are parallel:

$$\vec{\nabla}U(0.8, 0, 0) = \begin{pmatrix} 1.6 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{\nabla}I(0.8, 0, 0) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}. \tag{5.14}$$

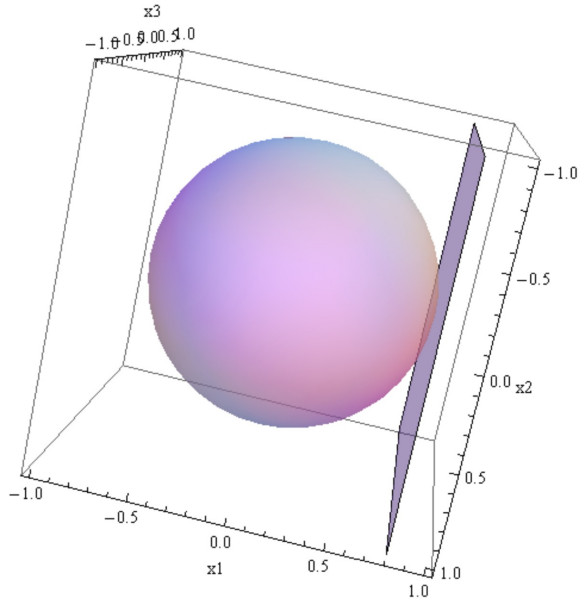


**Figure 12:** The spherical level surface of $U$ is tangent to the plane $x_1 = 0.8$. Note that there are infinitely many directions to reach the critical point from the tangent plane $x_1 = 0.8$.

The example of equations (5.11-5.12) correspond to a problem of minimization with a linear constraint, but the method works just as easily for nonlinear constraints. Consider for example the problem

$$\begin{aligned} \min U(x_1, x_2, x_3) &= x_1^2 + x_2^2 + x_3^2, && (5.15) \\ I(x_1, x_2, x_3) &= (x_1 - 1)^2 + x_2^2 + x_3^2 = 0.4^2 = 0.16. && (5.16) \end{aligned}$$

In Figure 13 different level surfaces of $U$ and the level surface constraint (5.16) are shown.

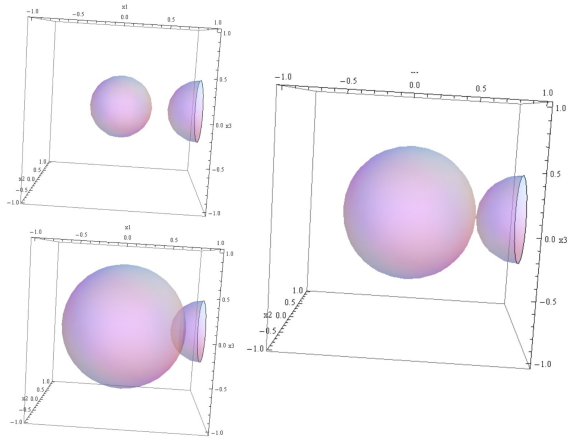

**Figure 13:** Level surfaces of $U$: top left: $U = 0.4^2 = 0.16$, bottom left: $U = 0.8^2 = 0.64$, right: $U = 0.6^2 = 0.36$. In all of them half of the level surface constraint $I = 0.4^2 = 0.16$ is shown.

A detailed observation of this figure clearly suggest that the constrained minimum is at $x_1 = 0.6$, $x_2 = x_3 = 0$, see the right part of Figure 13. Points in level surfaces corresponding to smaller values of $U$, as in the upper left part of Figure 13, are incompatible with the constraint (5.16), while points in level surfaces corresponding to greater values of $U$, as in the lower left part of Figure 13 are compatible with the constraint but do not minimize $U$. Note also that at the constrained minimum $x_1 = 0.6$, $x_2 = x_3 = 0$ the level surfaces of $U$ and $I$ are tangent to each other.

Let us see what the gradient equation (5.9) says:

$$\begin{pmatrix} 2x_1 \\ 2x_2 \\ 2x_3 \end{pmatrix} = \lambda \begin{pmatrix} 2(x_1 - 1) \\ 2x_2 \\ 2x_3 \end{pmatrix}. \tag{5.17}$$

The last two equations seem to imply $\lambda = 1$, and arbitrary $x_2$ and $x_3$. But if $\lambda = 1$, the first equation becomes $x_1 = x_1 - 1$, or $0 = -1$, which is clearly a contradiction. A more careful analysis of the last two equations is required.

Consider for example the second one: $2x_2 = 2\lambda x_2$, this implies $\lambda = 1$ if and only if $x_2 \neq 0$, but since this leads to a contradiction, this means that $x_2$ must be zero, which is also a solution of the equation. The same argument leads to $x_3 = 0$. If $x_2 = x_3 = 0$, no conditions on the value of $\lambda$ arises from the equations (5.14-5.15), and the equation (5.13) is satisfied for $\lambda = x_1/(x_1 - 1)$ as long as $x_1 \neq 1$.

With $x_2 = x_3 = 0$, the equation (5.16) becomes

$$(x_1 - 1)^2 = 0.4^2, \quad \text{or} \quad x_1 - 1 = \pm 0.4. \tag{5.18}$$

The minus sign leads to $x_1 = 0.6$, which is the solution that the right hand side of Figure 13 visually suggest. $x_1 = 0.6$ implies $\lambda = x_1/(x_1 - 1) = -1.5$, so the gradients of $U$ and $I$ are antiparallel:

$$\vec{\nabla} U = \begin{pmatrix} 1.2 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{\nabla} I = \begin{pmatrix} -0.8 \\ 0 \\ 0 \end{pmatrix}. \tag{5.19}$$

The reason they are antiparallel is that, as we have seen, the gradient always point in the direction of maximum increase of the respective function, and while $U$ increases to the right at $(0.6, 0, 0)$, $I$ increase to the left at that point.

Equation (5.18) implies another solution: $x_1 = 1.4$, $x_2 = x_3 = 0$. Figure 13 shows the level surfaces of $U$ and $I$ only in the range $-1 \leq x_i \leq +1$, $i = 1, 2, 3$, so this solution is not visible.

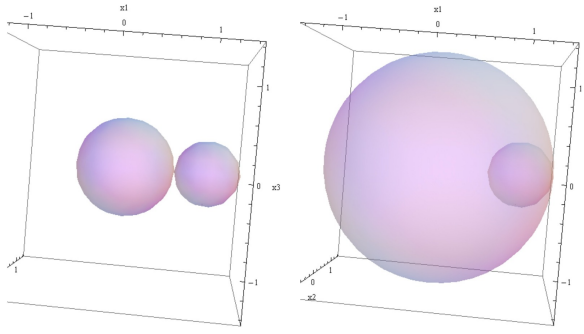

**Figure 14:** Left: level surfaces $U = 0.6^2 = 0.66$ and $I = 0.4^2 = 0.16$. Right: level surfaces $U = 1.4^2 = 1.96$ and $I = 0.4^2 = 0.16$.

In Figure 14 we extend the range up $-1.4 \leq x_i \leq +1.4$, $i = 1, 2, 3$. In the right side of this Figure 14 it can be appreciated that the new solution is a *maximum* of $U$ under the constraint (5.16).

As mentioned in the introduction, the first order Lagrange equation (5.9) does not distinguish between constrained maxima, minima, or general critical points. That would require higher order analysis which are not treated in this paper. That is why the constrained maximum solution shows up.

### 5.3. For $n$ variables, many constraints.

When $U$ is optimized under more than one constraint $I_j = i_j$, $j = 1, \ldots, m$, Lagrange's recipe consist of defining the Lagrangian as

$$L(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m) = U(x_1, \ldots, x_n) - \sum_{j=i}^{m} \lambda_j (I_j(x_1, \ldots, x_n) - i_j), \tag{5.20}$$

where there are as many new variables (Lagrange multipliers) $\lambda_j$ as there are constraints, and equating to zero the first derivatives with respect to all the variables $x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m$. In terms of the gradients, these equations become:

$$\vec{\nabla} U = \sum_{j=i}^{m} \lambda_j \vec{\nabla} I_j, \tag{5.21}$$

$$I_j = i_j. \tag{5.22}$$

The gain intuition for equations (5.21-5.22) let us consider first the case of two constraints, $m = 2$. As discussed in Section 5.2, in $n$ dimensions, a constraint like $I_1(x_1, \ldots, x_n,) = i_1$ defines implicitly an $n-1$ dimensional hypersurface, and at each point of this hypersurface there is only one orthogonal direction, the direction of the gradient $\vec{\nabla} I_1$. A second constraint, $I_2(x_1, \ldots, x_n,) = i_2$, also defines implicitly an $n-1$ dimensional hypersurface, and at each point the gradient $\vec{\nabla} I_2$ is orthogonal to it.

The constrained optimum, which must satisfy both the constraints $I_1 = i_1$ and $I_2 = i_2$, is located at the intersection of the corresponding $n-1$ dimensional hypersurfaces. Typically, this intersection is an $n-2$ dimensional hypersurface. The orthogonal complement of this intersecting hypersurface is spanned by the gradients $\vec{\nabla} I_1$ and $\vec{\nabla} I_2$, respectively.

The same arguments we have used many times now indicate that the constrained optimum of $U$ has to be positioned on an $n-1$ dimensional level hypersurface of $U$ that is tangential to the aforementioned $n-2$ dimensional hypersurface. This means that the gradient of $U$, which is orthogonal to its level hypersurface, should belong to the subspace spanned by the gradients $\vec{\nabla} I_1$ and $\vec{\nabla} I_2$. This idea is encapsulated precisely by equation (5.21) for the $m = 2$ constraint scenario.

For a number of constraints $m > 2$ the argument is essentially the same once one notes that the optimum must lie in the generically $n-m$ dimensional hypersurface intersection of the $m$, $(n-1)$ dimensional level hypersurfaces determined by the constraints $I_j = i_j$, $j = 1, \ldots, m$, and that the gradients $\vec{\nabla} I_j$ span the orthogonal complement of this intersecting hypersurface.

To visualize these ideas consider the following problem of minimization of the same function $U$ but now with two constraints:

$$\min U(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2, \tag{5.23}$$

$$I_1 = x_1 = 0.7, \tag{5.24}$$

$$I_2 = x_2 = 0.5. \tag{5.25}$$

In Figure 15 we can see that the constrained minimum of $U$ lies in the level surface $U = 0.74$, tangent to the line $x_1 = 0.7, x_2 = 0.5$ ($x_3$ arbitrary) given by the intersection of the two constraints (5.24-5.25).

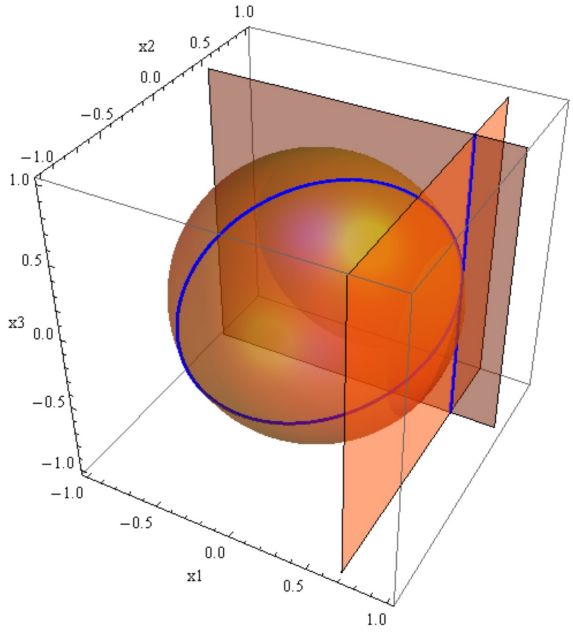


**Figure 15:** Level surface of $U = 0.7^2 + 0.5^2 = 0.74$ and the constraints $x_1 = 0.7$ and $x_2 = 0.5$.

Equations (5.21-5.22) become

$$\begin{pmatrix} 2x_1 \\ 2x_2 \\ 2x_3 \end{pmatrix} = \begin{pmatrix} 1.4 \\ 1 \\ 2x_3 \end{pmatrix} = \lambda_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}. \tag{5.26}$$

Which imply $x_3 = 0$, $\lambda_1 = 1.4$ and $\lambda_2 = 1.0$. So, the gradient of the function $U$ being minimized, $\vec{\nabla}U = (1.4, 1, 0)$, is a linear combination of the of the gradient of the constraining functions $\vec{\nabla}I_1 = (1, 0, 0)$ and $\vec{\nabla}I_2 = (0, 1, 0)$ at the constrained optimum $(0.7, 0.5, 0)$.

As explained above, the reason for this is that line $x_1 = 0.7, x_2 = 0.5$, $x_3$ arbitrary, being the $n - m = 3 - 2 = 1$ dimensional intersection of the plane $x_1 = 0.5$ and the plane $x_2 = 0.5$, and therefore belonging to both of them, must be orthogonal to the vectors orthogonal to the respective planes. This means that it must be orthogonal to their respective gradients. And since the level surface of $U$ should be tangent to this intersecting line at the constrained optimum, and its gradient $\vec{\nabla}U$ is orthogonal to this level surface, $\vec{\nabla}U$ must lie in the two dimensional subspace orthogonal to the intersection of the constraints, which implies that $\vec{\nabla}U$ should be a linear combination of $\vec{\nabla}I_1$ and $\vec{\nabla}I_2$. This is what equation (5.26) is saying.

## 6. Numerical optimization algorithms with equality constraints

The geometrical arguments developed in the previous section suggest the exploration of points that satisfy the constraints. Among these points, it is generally true that at the constrained minima, the gradients of $U$ and $I$ will be linearly dependent, see equations (5.9-5.10).

Numerically, finding a point that satisfies the constraints is nontrivial. Furthermore, moving in search of the constrained minima while remaining on the allowed hypersurface is challenging unless one can express one variable explicitly as a function of the others. This task becomes particularly difficult when the constraint is complex or when there are many variables. Equations (5.9) and (5.10) are best viewed as necessary conditions for the destination, rather than the path, to the constrained minima.

The difficulty of numerically following trajectories that satisfy the constraint can be seen in equation (3.15), which we used to demonstrate that level hypersurfaces of any function are locally orthogonal to the gradient of that function. This proof was based on a first-order approximation to the change in the function. However, numerically, every displacement is finite and inevitably deviates from this first-order approximation. The issue is not merely the deviation from the first-order approximation, but the lack of a mechanism to stabilize these deviations. Therefore, we need a mechanism that ensures these level hypersurfaces are stable. That is, even if our trajectory towards the constrained minimum deviates from the level hypersurface, some additional mechanism should return us to it.

With that objective in mind, a primary strategy is to modify the function $U$ to be minimized into a *sequence* of functions that increasingly penalize deviations from the constraint. A very effective approach is to use a quadratic penalty:

$$\min_{\mathbf{x}} F_n(x_1, x_2, x_3, \nu_n) = U + \nu_n (I - i)^2, \quad \nu_n \to \infty. \tag{6.1}$$

By minimizing each $F_n$ without constraints (for example, using the algorithm in Section 4) and using the final point $\mathbf{x}_n^*$ of the $n$th minimization as the starting point for the $(n+1)$th minimization, we can find a local constrained minimum under mild assumptions. Note that $\nu_n(I-i)^2$ is zero if $\mathbf{x}$ satisfies the constraint, and positive otherwise. In other words, $F_n = U$ for points $\mathbf{x}$ that satisfy the constraint.

The quadratic penalty function $(I - i)^2$, which generalizes to the norm square of the vector of constraints $\|\mathbf{I} - \mathbf{i}\|^2$ if there is more than one equality constraint, works well in many practical situations. The general idea is to allow unconstrained values of the independent variables $x_i$, and impose an increasingly heavier cost for violating the constraint. Consequently, the minimizing trajectory eventually ends up as close as desired to satisfying the constraints.

In the example (5.11-5.12), we have

$$F_n(x_1, x_2, x_3, \nu_n) = x_1^2 + x_2^2 + x_3^2 + \nu_n (x_1 - 0.8)^2 \tag{6.2}$$

and its gradient is

$$\vec{\nabla} F_n(x_1, x_2, x_3, \nu_n) = \begin{pmatrix} 2x_1 + 2\nu_n(x_1 - 0.8) \\ 2x_2 \\ 2x_3 \end{pmatrix}. \tag{6.3}$$

The $n$th minimization happens at

$$x^*_{1,n} = \frac{\nu_n}{1+\nu_n} \cdot 0.8 \to 0.8, \text{ as } \nu_n \to \infty, \quad x^*_{2,n} = 0, \quad x^*_{3,n} = 0. \tag{6.4}$$

Thus, we recover the correct result (see equations 5.13-5.14).

In general, to determine a local minimum of a function $U(x_1, \ldots, x_n)$ with equality constraints $I_j = i_j$, $j = 1, \ldots, k$, the following steps can be adopted:

1. Choose an increasing, positive sequence $\{\nu_n\}$ such that $\nu_n \to \infty$ as $n \to \infty$, and select a desired (small) error $\epsilon$.
2. Choose an initial point $\mathbf{x_0} = (x_{01}, \ldots, x_{0n})$ based on criteria that best suit the problem at hand.
3. With the $n$th element of the sequence $\nu_n$, find the unconstrained minimum of the function
$$F_n(\mathbf{x}, \nu_n) = U(\mathbf{x}) + \nu_n \sum_{j=1}^{k} (I_j - i_j)^2 \tag{6.5}$$
using the algorithm in Section 4 or your preferred unconstrained minimization algorithm. Use the final point of the $(n-1)$th iteration, denoted as $\mathbf{x}^*_{n-1}$, as the initial point for the $n$th iteration.
4. If $\|\vec{\nabla} F_n(\mathbf{x}, \nu_n)\| < \epsilon$, finish the process.
5. Select $\nu_{n+1}$ and return to step 3.

The analysis in Section 4 applies to the number of steps in each iteration.

For the selection of the sequence $\{\nu_n\}$, one must consider the following tradeoff: generally, at least for the initial iterations, the faster $\nu_n$ grows with $n$ (i.e., the greater the difference between $\nu_{n+1}$ and $\nu_n$, with $\nu_{n+1} > \nu_n$), the greater the difference between $F_{n+1}$ and $F_n$. Consequently, the distance between successive minima $\|\mathbf{x}^*_{n+1} - \mathbf{x}^*_n\|$ will also be greater, causing step 3 to take longer to converge. However, since the cost of violating the constraint will be higher, fewer iterations will generally be needed. Therefore, selecting the optimal sequence $\{\nu_n\}$ is a nontrivial problem. Nonetheless, if $\nu_n \to \infty$ as $n \to \infty$, we will find the local minima under very general assumptions.

Moreover, the algorithm will converge closer to the local minima satisfying the Lagrange conditions (5.9-5.10) if a smaller error $\epsilon$ is chosen. However, the smaller this error, the larger the number of iterations required.

One final remark about this algorithm: although it works well under fairly general assumptions, the condition $\nu_n \to \infty$ as $n \to \infty$ should raise some concern. Numerically, it is never advisable to rely on a very large number. Is there an alternative algorithm that does not require this condition?

Before addressing this question, it is worth noting that the Lagrange multipliers did not play any explicit role in the algorithm. It is only after the process halts that we should check whether conditions (5.9-5.10) are indeed satisfied. It would be beneficial to understand if and how the Lagrange multipliers $\lambda$ can be recovered from the algorithm, beyond merely verifying at the end.

In this direction, note that from equation (6.1), we have that at the unconstrained minimum of $F_n$,

$$\vec{\nabla} F_n(\mathbf{x}_n^*, \nu_n) = \vec{\nabla} U(\mathbf{x}_n^*) + 2\nu_n(I(\mathbf{x}_n^*) - i)\vec{\nabla} I(\mathbf{x}_n^*) = 0. \tag{6.6}$$

From this and equation (5.9), we can infer that

$$\lambda_n \equiv -2\nu_n(I(\mathbf{x}_n^*) - i), \tag{6.7}$$

should converge to the correct $\lambda$ as $n \to \infty$.

In the example (5.11-5.12), where $I = x_1$, $i = 0.8$, and $\mathbf{x}_n^*$ is given in equation (6.4), we have

$$-2\nu_n(I(\mathbf{x}_n^*) - i) = -2\nu_n\left(\frac{\nu_n}{1+\nu_n} \cdot 0.8 - 0.8\right) = 1.6\frac{\nu_n}{1+\nu_n}, \tag{6.8}$$

which indeed converges to $\lambda = 1.6$ as $\nu_n \to \infty$. In fact, $\lambda_n$ in equation (6.7) does converge to $\lambda$ under fairly general assumptions.

Equations (6.6)-(6.7) suggest a generalization of the five-step algorithm described earlier. Instead of equation (6.1), define the "augmented Lagrangian" function

$$L_n(x_1, x_2, x_3, \lambda_n, \nu_n) = U - \lambda_n'(I - i) + \nu_n(I - i)^2, \tag{6.9}$$

whose unconstrained minima will satisfy

$$\vec{\nabla} L_n(x_1, x_2, x_3, \lambda_n', \nu_n) = 0, \tag{6.10}$$

or

$$\vec{\nabla} U = (\lambda_n' - 2\nu_n(I(\mathbf{x}_n^*) - i))\,\vec{\nabla} I \equiv \lambda_n \vec{\nabla} I. \tag{6.11}$$

Returning to the example (5.11-5.12), the optimal point at iteration $n$ satisfies

$$\begin{pmatrix} 2x_{1,n} \\ 2x_{2,n} \\ 2x_{3,n} \end{pmatrix} = (\lambda_n' - 2\nu_n(x_{1,n} - 0.8))\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \tag{6.12}$$

which implies

$$x_{1,n} = \frac{1.6\nu_n + \lambda'_n}{2(1+\nu_n)}, \; x_{2,n} = x_{3,n} = 0. \tag{6.13}$$

Note that $x_{1,n} \to 0.8$ if $\lambda'_n \to 1.6$, even if $\nu_n \to$ any finite positive value!

This result is valid under fairly mild assumptions: if we manage to find a sequence $\{\lambda_n\}$ that converges to the correct Lagrange multiplier, the coefficient $\nu_n$ of the stabilizer term $\nu_n(I-i)^2$ in the augmented Lagrangian function (6.9) does not need to become excessively large for convergence. This avoids many potential problems associated with dealing with large numbers and, in addition, tends to converge faster to the correct constrained minima. The terms $\lambda_n(I-i)+\nu_n(I-i)^2$ in the augmented Lagrangian extend naturally if there are multiple equality constraints.

But how do we find a sequence $\{\lambda_n\}$ that converges to the right Lagrange multiplier? Not surprisingly, (6.11) suggests the following iterative form for updating $\lambda_n$:

$$\lambda_{n+1} = \lambda_n - 2\nu_n(I(\mathbf{x}^*_n) - i), \tag{6.14}$$

where the sequence $\{\nu_n\}$ no longer has to diverge for large $n$.

The iterative rule (6.14), together with the quadratic "penalty", or "stabilizer" function, defines a famous algorithm known as the "Method of Multipliers". It consists of the 5 steps as before, but minimizing without constraints $L_n$ given in (6.9) instead of $F_n$ in (6.5), and no longer requiring the weak point of the previous algorithm, namely, that $\nu_n \to \infty$ as $n \to \infty$.

Returning to our example (5.11-5.12), (6.13) and (6.14) imply

$$\lambda_{n+1} = \lambda_n - 2\nu_n\left(\frac{1.6\nu_n + \lambda_n}{2(1+\nu_n)} - 0.8\right) = \frac{\lambda_n}{1+\nu_n} + 1.6\frac{\nu_n}{1+\nu_n}. \tag{6.15}$$

Knowing that $\lambda = 1.6$, and assuming for simplicity a constant $\nu$, the error in the $n$th estimation of $\lambda$ is

$$\lambda_n - 1.6 = \frac{1}{1+\nu}(\lambda_{n-1} - 1.6) \tag{6.16}$$

$$= \frac{1}{(1+\nu)^n}(\lambda_0 - 1.6), \tag{6.17}$$

where, as can easily be checked, the second line is the solution of the iterative rule of the first line.

Equation (6.17) shows an exponentially fast convergence as long as $\nu > 1$, but still, the convergence is faster if $\nu$ is greater. This is true in general, but if the problem is not convex, $\nu$ could need to be greater than 1 to ensure convergence.

The general rule is that it is still the case that greater $\nu$ generally leads to faster convergence, but $\nu$ should not be too large to ensure convergence and avoid potential numerical instabilities associated with very large numbers.

This is just a tiny part of an enormous body of extremely powerful algorithms to minimize functions with not only equality constraints, but also inequality ones, see for example [4], [6], [7], [20].

## 7. Conclusions

The advent of the machine learning revolution has ushered in transformative shifts across multiple disciplines, including economics and the broader realm of social sciences. However, there remains a noticeable lag: the foundational quantitative training in these fields has not fully adapted to these novel developments.

In this review paper, we sought to address a small part of this lacuna, focusing on the vital but often overlooked concept of gradient fields. Despite its pivotal role in comprehending numerous traditional problems within economics and social sciences, gradient fields remain underrepresented in many foundational curricula and methodologies.

By situating the concept of gradient fields within core contexts – namely, optimization, both constrained and unconstrained – we endeavored to illuminate the intuitive power of gradient methods. Furthermore, we emphasized its relevance in the contemporary landscape, spotlighting its role in modern applications that harness the capabilities of machine learning.

The flexible Lagrange multiplier method to solve constrained optimization problems is often presented in a prescriptive and mechanistic manner. This represents an ideal context in which to convey the power of gradient field methods. Together with the generalization of orthogonality provided by the scalar product, gradient field methods are an ideal tool to extend our spatial intuitions into realms highly disconnected from our everyday visual experience. And when intuition is at work, solutions to problems often present themselves almost automatically.

Unfortunately, these kinds of mathematical tools that dramatically increase the reach of intuition as a tool to help solve problems are notoriously underemphasized in the quantitative training of economists and other social scientists. This review work tries to contribute to correcting at least part of this problem, which acquires increasingly larger dimensions given the usefulness of these methods in machine learning.

We hope that this exposition serves as a bridge for students and scholars in the mentioned disciplines, connecting traditional mathematical approaches with the emergent techniques that are increasingly essential in today's ever-evolving academic and practical landscapes.

## References


[1] S. Athey, *The Impact of Machine Learning on Economics*. The Economics of Artificial Intelligence: An Agenda. University of Chicago Press, 2019, 507–547. DOI: 10.7208/chicago/9780226613475.003.0021

[2] S. Athey, G. W. Imbens, *Machine Learning Methods That Economists Should Know About*. Annual Review of Economics **11**(2019), 685–725. DOI: 10.1146/annurev-economics-080217-053433

[3] A. Belloni, V. Chernozhukov, C. Hansen, *High-dimensional methods and inference on structural and treatment effects*. Journal of Economic Perspectives **28**(2014), no. 2, 29–50. DOI: 10.1257/jep.28.2.29

[4] D. Bertsekas, *Nonlinear Programming*. Athena scientific optimization and computation series. Athena Scientific, 2016. DOI: 10.1057/palgrave.jors.2600425

[5] C. Bishop, *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer New York, 2016.

[6] J. Bonnans, J. Gilbert, C. Lemarechal, C. Sagastizábal, *Numerical Optimization: Theoretical and Practical Aspects*. Universitext. Springer Berlin Heidelberg, 2013. DOI: 10.1007/978-3-540-35447-5

[7] S. Boyd, L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004. DOI: 10.1017/CBO9780511804441

[8] V. Chernozhukov et al., *Double/debiased machine learning for treatment and structural parameters*. The Econometrics Journal **21**(2018), no. 1, C1–C68. DOI: 10.1111/ectj.12097

[9] A. Chiang, K. Wainwright, *Fundamental Methods of Mathematical Economics*. McGraw-Hill international edition. McGraw-Hill Education, 2005.

[10] J. Duchi, E. Hazan, Y. Singer, *Adaptive Subgradient Methods for Online Learning and Stochastic Optimization*. Journal of Machine Learning Research **12**(2011), no. 61, 2121–2159.

[11] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*. Adaptive Computation and Machine Learning series. MIT Press, 2016. DOI: 10.1007/s10710-017-9314-z

[12] J. Grimmer, M. E. Roberts, B. M. Stewart, *Machine learning for social science: An agnostic approach*. Annual Review of Political Science **24**(2021), 395–419. DOI: 10.1146/annurev-polisci-053119-015921

[13] M. Hoy et al., *Mathematics for Economics, third edition*. The MIT Press. MIT Press, 2011.

[14] W. Karush, *Minima of Functions of Several Variables with Inequalities as Side Conditions*. G. Giorgi, T. H. Kjeldsen (Eds.). Traces and Emergence of Nonlinear Programming. Springer Basel, Basel, 2014, 217–245. DOI: 10.1007/978-3-0348-0439-4_10

[15] D. P. Kingma, J. Ba, *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980 (2014). DOI: 10.48550/arXiv.1412.6980

[16] J. Kleinberg, J. Ludwig, S. Mullainathan, C. R. Sunstein, *Discrimination in the Age of Algorithms*. Journal of Legal Analysis **10**(2019), 113–174. DOI: 10.1093/jla/laz001

[17] H. Kuhn, A. Tucker, *Nonlinear Programming Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. 1951.

[18] R. P. Masini, M. C. Medeiros, E. F. Mendes, *Machine learning advances for time series forecasting*. Journal of economic surveys **37**(2023), no. 1, 76–111. DOI: 10.1111/joes.12429

[19] S. Mullainathan, J. Spiess, *Machine learning: an applied econometric approach*. Journal of Economic Perspectives **31**(2017), no. 2, 87–106. DOI: 10.1257/jep.31.2.87

[20] J. Nocedal, S. Wright, *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2006. DOI: 10.1007/b98874

[21] S. Noy, W. Zhang, *Experimental evidence on the productivity effects of generative artificial intelligence*. Science **381**(2023), no. 6654, 187–192. DOI: 10.2139/ssrn.4375283

[22] D. Rolnick et al., *Tackling climate change with machine learning*. ACM Computing Surveys (CSUR) **55**(2022), no. 2, 1–96. DOI: 10.1145/3485128

[23] S. Ruder, *An overview of gradient descent optimization algorithms*. arXiv preprint arXiv:1609.04747 (2016). DOI: 10.48550/arXiv.1609.04747

[24] C. Simon, L. Blume, *Mathematics for Economists*. Norton, 1994.

[25] H. R. Varian, *Big data: New tricks for econometrics*. Journal of economic perspectives **28**(2014), no. 2, 3–28. DOI: 10.1257/jep.28.2.3