3.7 Constrained Optimization and Lagrange Multipliers

Overview: Constrained optimization problems can sometimes be solved using the methods of the previous section, if the constraints can be used to solve for variables. Often this is not possible. Lagrange devised a strategy to turn constrained problems into the search for critical points by adding variables, known as Lagrange multipliers. This section describes that method and uses it to solve some problems and derive some important inequalities.

Often minimization or maximization problems with several variables involve constraints, which are additional relationships among the variables. For example, investments might be constrained by total assets; engineering improvements on an airplane might be constrained by costs or time to implement or weight or available workers; my maximum altitude on a hiking trail on a mountain is constrained by the trail itself and may not be the altitude of the top of the mountain. Each constraint equation reduces the dimension of the domain by one (in general). It is often inconvenient and sometimes not feasible to express the extremal problem in reduced coordinates. Lagrange found an alternative approach using what are now called **Lagrange multipliers**. Assuming the constraints are given as equations, Lagrange's idea is to solve an unconstrained problem in more variables! This section introduces his key idea and applies it. A project at the end of the chapter considers inequality constraints.

Let's start in two dimensions, with one constraint. As a first example, consider a geometric problem whose solution you probably know already: for a line l not through the origin, find the point on l nearest to the origin and compute its distance from the origin. We will solve this in three ways: first, geometrically; second, eliminating one variable and solving a 1-D minimization problem; and third, with Lagrange multipliers.

Geometrically, the closest point $p_* = (x_*, y_*)$ is where a (second) line through p_* and the origin is perpendicular to l. The Pythagorean theorem shows that p_* is closer to the origin than any other point on l. As an example, suppose that l has equation: $2x + 3y = 6$. The perpendicularity condition tells us that the closest point p_* has coordinates satisfying $x_* = 2t, y_* = 3t$ for some number $t \neq 0$. Requiring p_* to lie on l becomes $2(2t) + 3(3t) = 6$ so $t = \frac{6}{13}$. The distance is then found to be $\sqrt{x_*^2 + y_*^2} = \sqrt{\frac{36}{13}} = \frac{6}{\sqrt{13}}$. Note that we didn't use any calculus in this solution!

Our second method parameterizes the line and then uses one variable calculus. Use the constraint $2x + 3y = 6$ to solve for y in terms of x, namely $y = 2 - (2/3)x$ and then write the distance to the origin as $\sqrt{x^2 + y^2}$. It is easier to work with the distance squared (no denominators in derivatives), so we define $f(x, y) = x^2 + y^2$. Then we can reduce our constrained minimization of f on the line l to minimizing the function $g(x) = f(x, 2-(2/3)x) = x^2+(2-(2/3)x)^2$. We find that $g'(x) = 0$

only when $g'(x) = 2x + 2(2-(2/3)x)(-2/3) = 0$. Some algebra finds the critical point $x_* = \frac{12}{13}$ and then substitution yields y_* and the rest is as before.

A third approach, using Lagrange multipliers, solves three equations in three unknowns: to minimize $f(x,y) = x^2 + y^2$ with constraint $g(x,y) = 2x + 3y = 6$, Lagrange adds a third variable, which we will denote by a Greek lambda, λ and forms the function $L(x,y,\lambda) = f(x,y) - \lambda(g(x,y) - 6)$. The new problem seeks a critical point of L in the three(!) variables (x,y,λ). This leads to the following three equations in this case:

$$\begin{aligned}
\frac{\partial L}{\partial x} &= f_x(x,y) - \lambda g_x(x,y) = 2x - \lambda 2 = 0 \\
\frac{\partial L}{\partial y} &= f_y(x,y) - \lambda g_y(x,y) = 2y - \lambda 3 = 0 \\
\frac{\partial L}{\partial \lambda} &= g(x,y) - 6 = 2x + 3y - 6 = 0.
\end{aligned}$$

This is essentially the same system of equations (except for a factor of 2) as in our first approach except t is replaced by $\lambda/2$. The statement is that the gradients of f and g are multiples of each other, so they are aligned and therefore locally the level sets of f and the constraint set $g(x,y) = 0$ are lined up also, using the linear approximations.

Why did we do the last approach? Because it generalizes to any continuously differentiable function f with any continuously differentiable constraint $g(x,y) = c$ for some constant c, which by changing the definition of g if we wish, will become 0.

3.7.1 Lagrange's method in general, 2-D:

Define $L(x,y,\lambda) = f(x,y) - \lambda(g(x,y) - c)$. Then every critical point of L corresponds to a possible extremum of f with constraint $g(x,y) = c$. This leads to three equations in three unknowns!

$$\frac{\partial L}{\partial x} = f_x(x,y) - \lambda g_x(x,y) = 0, \quad \frac{\partial L}{\partial y} = f_y(x,y) - \lambda g_y(x,y) = 0, \quad \frac{\partial L}{\partial \lambda} = g(x,y) - c = 0.$$

Note that critical points of f are possible solutions **if they satisfy the constraints**, in which case the Lagrange multiplier λ is zero. But now there are new possibilities, as in our geometry problem.

Why is Lagrange correct? Thinking locally Recall how we thought about local maxima in the last section. The idea was to see that away from critical points, the gradient vector describes a direction for increasing the function (and its opposite direction decreases the function). Therefore points where the gradient was not zero couldn't be local extrema. A similar but more elaborate argument now shows that when Lagrange's equations do not hold at some point, that point is not a constrained local extremum.

More precisely, we show that if the gradient of f is not zero and not a multiple of the gradient of g at some location, then that location cannot be a constrained extrema. The idea is that in the tangent direction to the constraint (admissible variations) the directional derivative of f is non-zero, so without violating the constraint, there is an uphill and downhill direction within the level set of g. I find the picture of the level sets more revealing, so imagine a location which is not a critical point for f and where the constraint $g(x, y) = 0$ is not lined up with a level set of f. Then the allowed set crosses various levels of f and therefore we are not at a constrained maximum or minimum. In my hiking example, the trail has an uphill direction and a downhill direction in the opposite way, so I have not finished my climbing or descending at that location. but now the *directional derivative* in the directions allowed by the constraint is the relevant non-zero quantity. Notice that this is what the non-zero derivative in the second method above would be saying, where we were able to solve for y. The local level set plot for f is a family of parallel lines and the local level set plot for g is a single line, so the picture where Lagrange's equations do not hold is like the following picture, where the set $g = 0$ is the thicker line (red). Therefore any constrained local maximum or minimum would need to satisfy the Lagrange condition with some multiplier λ.

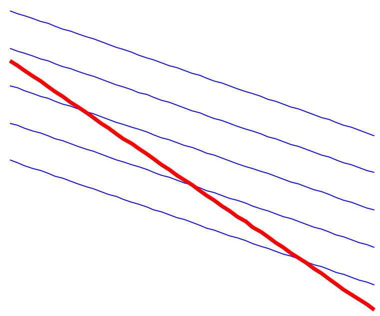

Figure 20: The level sets and constraint when Lagrange's condition *does not hold*: f (blue level sets) can change without changing g (red).

To confirm the usefulness of the local analysis, what do you think happens when we replace the line in our geometry problem by a general curve in the plane? Let's see what Lagrange's equations say in that case and if you are correct.

Example 3.43 Distance from origin to curve $g(x, y) = 0$ *The distance squared is again* $x^2 + y^2$ *and if we require* $g(x, y) = 0$*, then Lagrange's equations are:*

$$2x = \lambda g_x(x, y), \;\; 2y = \lambda g_y(x, y), \;\; g(x, y) = 0$$

which are three equations for x, y, λ. The geometric statement is that at the minimizer, the gradient of g will point directly towards or away from the origin at the minimizer (i.e. radially). Several examples of this are in the exercises for particular curves.

What does the number λ mean? λ refers to the relative magnitude of the two gradients at the critical point. In economics and related problems, an alternative description of λ as the partial derivative in the variable (now viewed in yet another extension) c of the (extended) function L. When the constraint is cost, this measures how much the critical value of L moves per unit cost, and this is in turn the change in f when the (varying!) constraint holds.

Some alternate language: In older books using differentials, the tangent vector to the constraint set is called *an allowed (or admissible) variation* and Lagrange's main condition says that the directional derivative of f in all allowed variations is zero. In physics and engineering, equilibrium configurations are often described as minimizers of energy. Thinking about the change near some location and deducing conditions for such a minimum is often called the principle of virtual work.

Lagrange's method extends to three or more dimensions, with one or more constraints.

3.7.2 3D case with one or two constraints

If we wish to extend our geometric example to 3D, there are two possible versions: the nearest point in a line in 3D and the nearest point in a plane in 3D. The restricted set is one dimensional for the line, two dimensional for the plane, so they are defined by two linear equations and one linear equation respectively.

The plane is the easier case, as it has one constraint: minimize the distance (or its square) from a point (chosen to be the origin) to a plane. The plane is assumed to satisfy $ax + by + cz = D$ so we are minimizing $x^2 + y^2 + z^2$ with constraint that points lie in the plane. Applying the idea of Lagrange multipliers, we find four equations in four unknowns:

$$2x - a\lambda = 0,\ 2y - b\lambda = 0,\ 2z - c\lambda = 0,\ ax + by + cz = D$$

and the minimum occurs when the vector $< x, y, z >$ is aligned with the normal vector to the plane $< a, b, c >$ with scalar factor $\frac{\lambda}{2}$. The minimum value of $x^2 + y^2 + z^2$ is therefore $\lambda^2(a^2 + b^2 + c^2)/4$, which must equal $\frac{D^2}{a^2+b^2+c^2}$ so the minimal distance is $\frac{|D|}{\sqrt{a^2+b^2+c^2}}$ and it occurs at the perpendicular to the plane.

The distance to the line is similar, but with the added notion that the line is the intersection of two planes. This gives us two linear constraints. The interpretation is now that we should look at level sets of the function f to be minimized near a possible extreme point. Then the level sets are again nearly planes and if the constraint sets cross each other transversely, then the level set of f near a minimizer should *include* the line representing both constraints simultaneously, for if the line crossed the level set of f then the point would not be an extreme point. In other words, the tangent vector to the curve comprising the constraint set should be orthogonal to the gradient of f, so that the level set of f is tangent to the constraint set. This says that the gradients of the two functions, say g_1 and g_2, which give the constraints are such that every vector orthogonal to both of them is also orthogonal to the gradient of f. This says that f is a blend of the two gradients — in equation form there are numbers λ_1 and λ_2 so that

$$\nabla \mathbf{f} = \lambda_{\mathbf{1}} \nabla \mathbf{g_1} + \lambda_{\mathbf{2}} \nabla \mathbf{g_2}$$

Your geometric intuition should connect the example to what you know. The closest point to the origin on a line in 3D is the point where the gradient field of the distance function, which as we know is the unit radial field, points orthogonally to the line.

In general, we can justify the Lagrange multiplier principle for two or more constraints by "thinking locally" and arguing that *if the condition fails*, we cannot be at a constrained local max or min. In 3D we say this two ways, depending on how we view the two constraints, which lead (typically) to a one dimensional restricted domain: first, in terms of the tangent vector to the restricted domain, namely we are decidedly not at a constrained extremum if the directional derivative of f in the direction of the tangent is non-zero. The second way uses the normals to the constraint domains, and says that if the level sets of f, which are locally like parallel planes, cross the constraints transversely then the constraints do not prohibit moving in directions that increase and decrease f, so we are not at a constrained extremal. The two views are linked since the tangent vector is orthogonal to both gradients of the constraints, this non-extremal condition is that the gradient of f has some component which is in the direction orthogonal to both constraint gradients. If the gradient of f has a part that is not a linear combination of the gradients of the constraints, then the level sets of f near the point in question will cross the constraint sets transversely and therefore there are values of f both higher and lower in any neighborhood of our point. This negative linear result says that the Lagrange multiplier condition must hold at any potential extreme point.

3.7.3 Some classic inequalities in 3D

Here are some classic mathematical inequalities that can be derived using Lagrange multipliers.
algebraic-geometric mean: For any three non-negative real numbers x, y, z, the geometric mean is less than or equal to the algebraic mean, which says:

$$\sqrt[3]{x\,y\,z} \leq \frac{x+y+z}{3}$$

To prove this we first note that if any of the three numbers is 0 then the root is 0 also and the inequality is true. Second, note that as all the numbers are multiplied by a constant, both sides are also multiplied by the same constant. So if we prove it when $x+y+z=1$, it is true for all values.

Now we can create a constrained maximization problem: find the largest value of $x\,y\,z$ when $x+y+z=1$. If this value is C, we then conclude by scaling that $\sqrt[3]{x\,y\,z} \leq \sqrt[3]{C}(x+y+z)$. Note that we dropped the cube root in our constrained problem — this just makes the algebra a little cleaner. Note also that this is a classic word problem, as in problem By Lagrange multipliers, the four equations for x, y, z, λ are:

$$yz - \lambda = 0, \;\; xz - \lambda = 0, \;\; xy - \lambda = 0, \;\; x+y+z = 1$$

Either by symmetry or by multiplication of the first three equations by x, y, z respectively, it follow that the maximum value occurs when $x = y = z = \frac{1}{3}$. Then the extreme value of the product xyz is $\frac{1}{27}$ and the desired result holds.

The generalization to a sum of n numbers $x_1, x_2, \ldots x_n$ states correctly that $\sqrt[n]{x_1\,x_2\,\cdots x_n} \leq \frac{x_1+x_2+\cdots+x_n}{n}$ and its proof is an exercise.

Cauchy-Schwarz inequality: If we look at a linear function $ax+by+cz$ for constants a, b, c, the Cauchy-Schwarz inequality states that

$$|ax+by+cz| \leq \sqrt{a^2+b^2+c^2}\sqrt{x^2+y^2+z^2}$$

with equality only when the vectors $< a, b, c >$ and $< x, y, z >$ are proportional. For $a = b = c = 0$, the result is clear, so we assume that the vector $< a, b, c >$ is non-zero. By scaling the vector $< x, y, z >$ by its length, which also scales the linear function, it is enough to show that for $x^2+y^2+z^2 = 1$, the corresponding inequality holds. This puts it in a form where constrained optimization applies.

If we consider the Lagrange function $ax+by+cz-\lambda\,(x^2+y^2+z^2-1)$, its gradient being zero leads to the equations:

$$a - 2\lambda x = 0, \;\; b - 2\lambda y = 0, \;\; c - 2\lambda z = 0$$

which says the vectors $< a, b, c >$ and $< x, y, z >$ are proportional at the critical points, with proportionality factor 2λ. Since the variable location $< x, y, z >$ is on the unit sphere, we find that $< x, y, z >= \pm\frac{<a,b,c>}{\sqrt{a^2+b^2+c^2}}$ and $2\lambda = \pm\sqrt{a^2+b^2+c^2}$, so the value of the function $ax+by+cz$ at its maximum is $\sqrt{a^2+b^2+c^2}$, as advertised.

The Cauchy-Schwarz inequality was used earlier in Chapter 1 when discussing the dot product of two vectors!Note that this proof did not use any trig (as did our earlier discussion) nor any algebra of quadratic function in all six variables. It also generalizes to the result known as Hölder's inequality, which we now discuss in a special case.

A special case of Hölder's inequality: If for some reason we don't want to treat the two pieces from the Cauchy-Schwarz equally, we can generalize to get Hölder's inequality, which in 3D says:

$$|ax+by+cz| \leq (|a|^p + |b|^p + |c|^p)^{\frac{1}{p}}\,(|x|^q + |y|^q + |z|^q)^{\frac{1}{q}}$$

provided $1 < p < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$.

To avoid dealing with absolute values, we will do the special case where $q = 4$ and we assume a, b, c are all non-negative. Thus we want to show:

$$|ax + by + cz| \le (a^{\frac{4}{3}} + b^{\frac{4}{3}} + c^{\frac{4}{3}})^{\frac{3}{4}} (x^4 + y^4 + z^4)^{\frac{1}{4}}.$$

Again, by scaling with $(x^4 + y^4 + z^4)^{\frac{1}{4}}$, we need only establish the inequality when this sum is equal to one: $x^4 + y^4 + z^4 = 1$ will be the constraint. The Lagrange function is now $ax + by + cz - \lambda (x^4 + y^4 + z^4 - 1)$ and the gradient equations are:

$$a - 4\lambda x^3 = 0, \;\; b - 4\lambda y^3 = 0, \;\; c - 4\lambda z^3 = 0$$

which are a bit nastier than our previous case. But we find that the vector $< x, y, z >$ is proportional to the vector of cube roots $< a^{\frac{1}{3}}, b^{\frac{1}{3}}, c^{\frac{1}{3}} >$ and using the constraint we find that the critical values for $< x, y, z >$ are at the locations $\pm \frac{< a^{\frac{1}{3}}, b^{\frac{1}{3}}, c^{\frac{1}{3}} >}{(a^{\frac{4}{3}} + b^{\frac{4}{3}} + c^{\frac{4}{3}})^{\frac{1}{4}}}$, and the maximal value of the linear function is $(a^{\frac{4}{3}} + b^{\frac{4}{3}} + c^{\frac{4}{3}})^{\frac{3}{4}}$ as advertised.

3.7.4 Quadratic functions on unit sphere

If we look at a quadratic function in three variables, it takes the form $\mathbf{r}^T H \mathbf{r}$ where as usual H is a symmetric 3×3 matrix and $\mathbf{r}$ is the position vector $\mathbf{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$. Such a function is continuously differentiable and typically grows unboundedly as $|\mathbf{r}|$ grows. But if we restrict to the unit sphere by requiring $|\mathbf{r}| = 1$, then such a function will achieve its maximal and minimal values somewhere on the sphere. Lagrange multipliers end up relating such locations to the **eigenvectors** of the symmetric matrix H as follows:

To find constrained extrema for the quadratic function $Q(x, y, z) = \mathbf{r}^T H \mathbf{r}$ with constraint $\mathbf{r}^T \mathbf{r} = x^2 + y^2 + z^2 = 1$, we introduce a Lagrange multiplier λ and seek the points where $\nabla(\mathbf{Q} - \lambda \nabla(\mathbf{x^2} + \mathbf{y^2} + \mathbf{z^2} - \mathbf{1})) = \mathbf{0}$, which becomes

$$2\, H\mathbf{r} - 2\lambda\, \mathbf{r} = \mathbf{0}.$$

Up to a factor of 2 which clearly drops out, this says the constrained extrema must be **eigenvectors** with λ being the value of the **eigenvalue**.

Looking forward: Theorems in advanced calculus guarantee the existence of the constrained maximum and minimum in any finite dimension and some extra work ends up showing that the **spectral theorem for symmetric matrices** holds. This theorem states that every symmetric matrix has a full set of eigenvalue-eigenvector pairs and that the eigenvectors may be chosen to be orthogonal (choice only happens when eigenvalues are repeated).

Looking forward: Constrained optimization is a large subject which is flourishing. Many modern technologies rely on the solution of large optimization problems and many business decisions factor in some major mathematical analysis of the situation. Applications include image processing, statistics, internet searching and routing, manufacturing engineering, economics, and data compression. Some constraints come as inequalities and many problems involve large scale computing to find approximate solutions.

3.7.5 EXERCISES

Writing: Explaining, Reacting, Questioning

1. Anita Solushun reasons as follows for the example of minimizing the distance from the origin to the line $2x + 3y = 6$: Consider the parallel line through the origin: $2x + 3y = 0$, then the distance between the lines is tied to the change of size 6 in the function values compared to the change in (x, y) in the maximal direction. This is the magnitude of the gradient, so the minimal distance is $6/\sqrt{2^2 + 3^2}$. Is she lucky or is this correct? If not correct, explain what is wrong. If correct, find the general case answer for a point at (x_1, y_1) and a line $ax + by = c$ by similar reasoning.

2. Consider two non-intersecting curves in the plane. The minimization of the distance between the curves requires some algebra ingredients to set it up. Before doing that, think through a local picture of what the condition should be geometrically. What do you think the condition will be and why? Now set up the problem, assuming the points on the first curve are called (x_1, y_1), those on the second curve are called (x_1, y_1), and the curves are given by relations $g_1(x_1, y_1) = 0$ and $g_1(x_1, y_1) = 0$. Minimize the square of the distance between the points. Deduce from the equations that your geometric intuition was correct (if it wasn't, fix it!).

3. **Another view of the distance from the origin to a curve:** View the constraint $g(x, y) = 0$ in polar coordinates, since then r represents the distance from the origin. If the constraint is viewed as defining $r(\theta)$ implicitly, then at a minimal distance, $r(\theta)$ achieves a local minimum, so $r'(\theta) = 0$ there. But by implicit differentiation, this says that in polar form, $g_\theta = 0$, which says the gradient at the constrained point is radial! Write out a complete version of this for the function $g(x, y) =$.