مبانی یادگیری ماشین

Intro to Machine Learning

بابک نجار اعرابی

دانشکده مهندسی برق و کامپیوتر دانشگاه تهران

نیم سال اول سال تحصیلی 1403-04

موضوع این جلسه

مروری بر روش های بهینه سازی

جلسه ششم

Babak Nadjar Araabi
School of Electrical & Computer Eng
University of Tehran

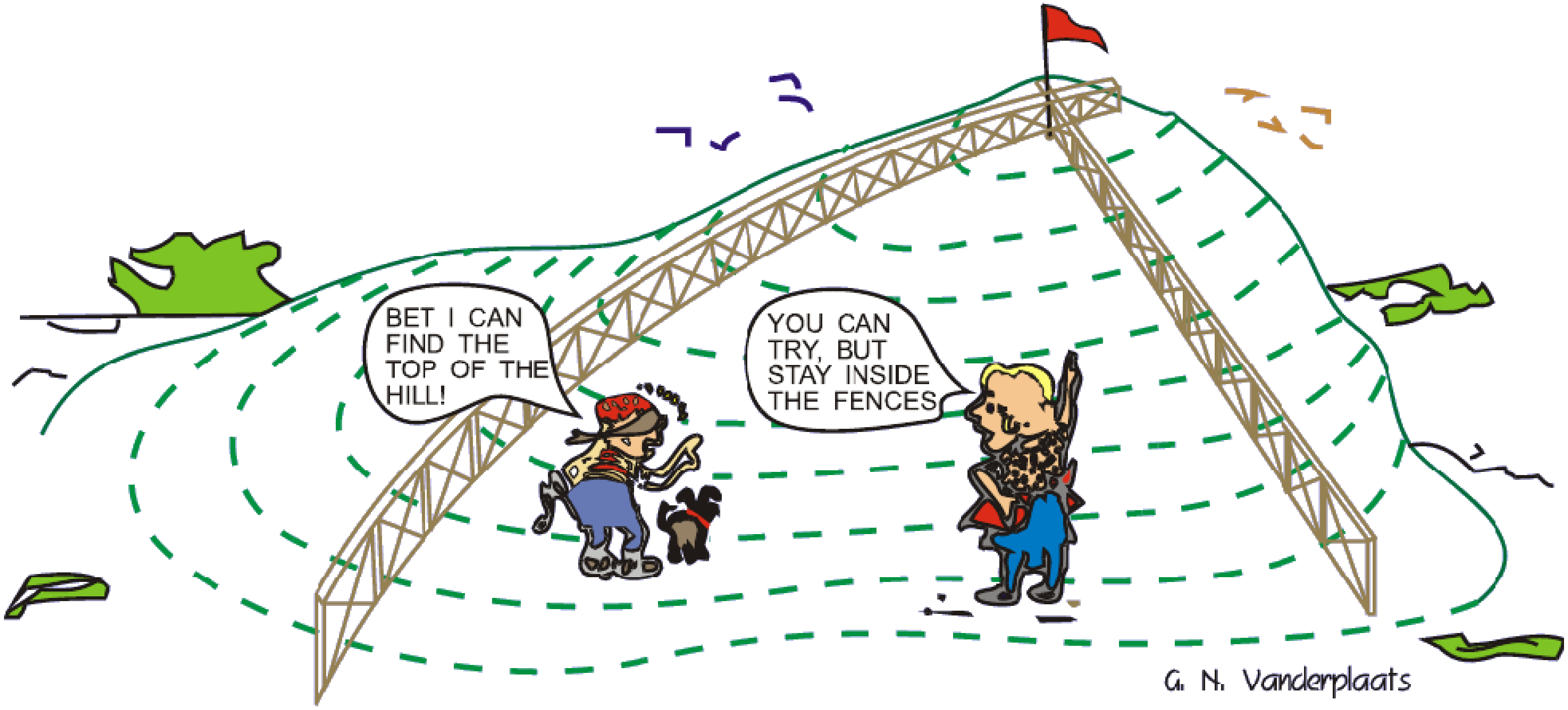
ECE-UT - Fall 2024

Constrained Optimization

Operations Research by Hamdy Taha

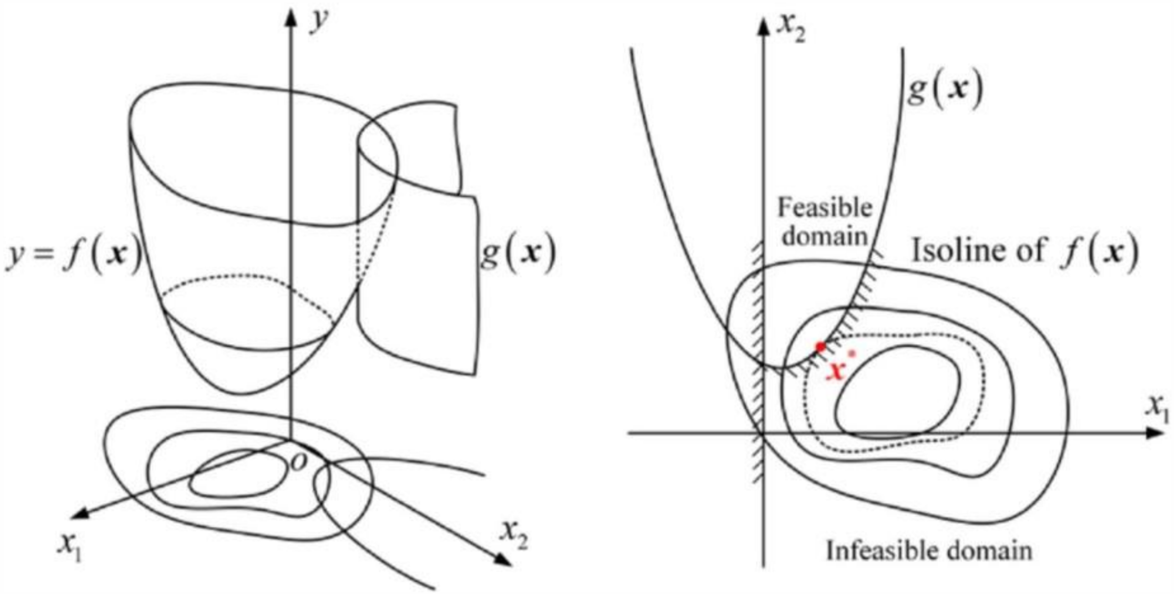
Ch 20, 10^{th} ed, 2017

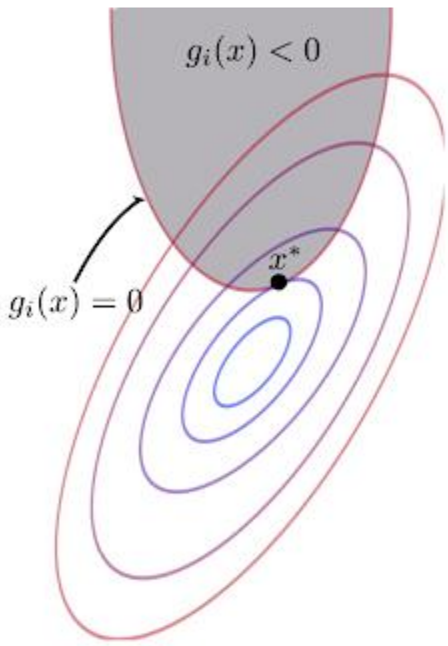
Feasible Solution

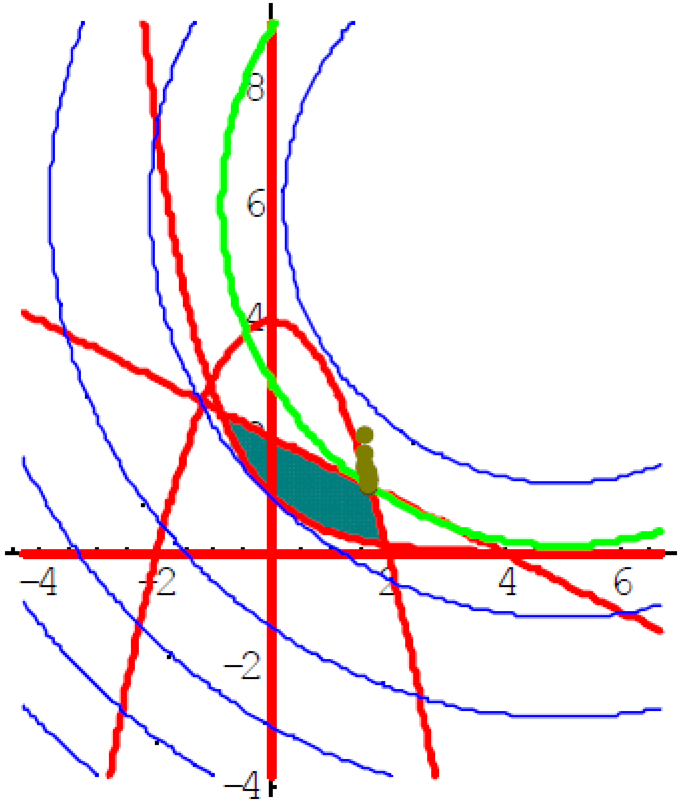


Review of optimization techniques by G. Venter, 2010

What is Constrained Optimization



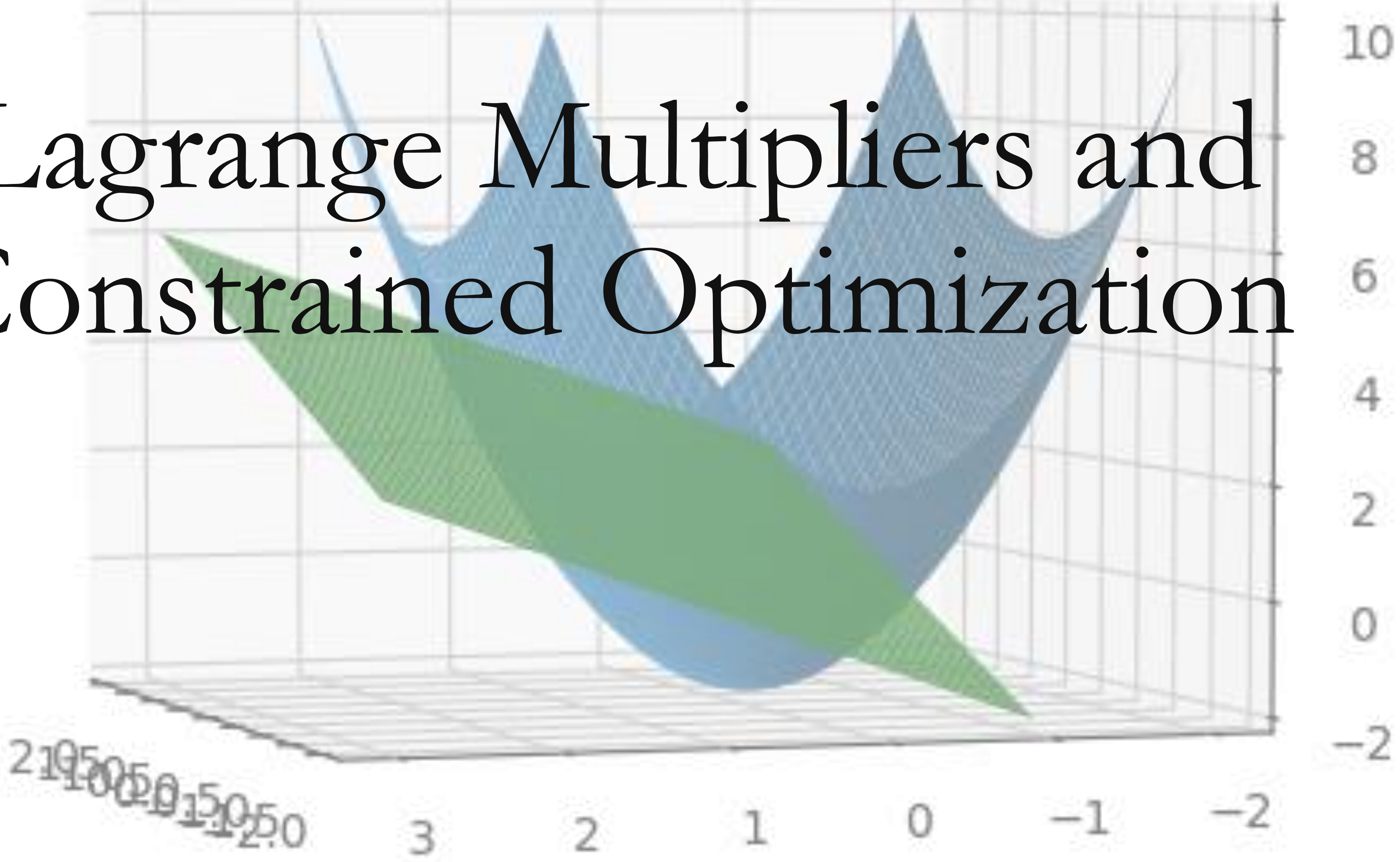

$g_i(x) < 0$
x^*
$g_i(x) = 0$


8
6
4
–4
–2
2
4
6
–2
–4

Belmann & Zadeh 1970

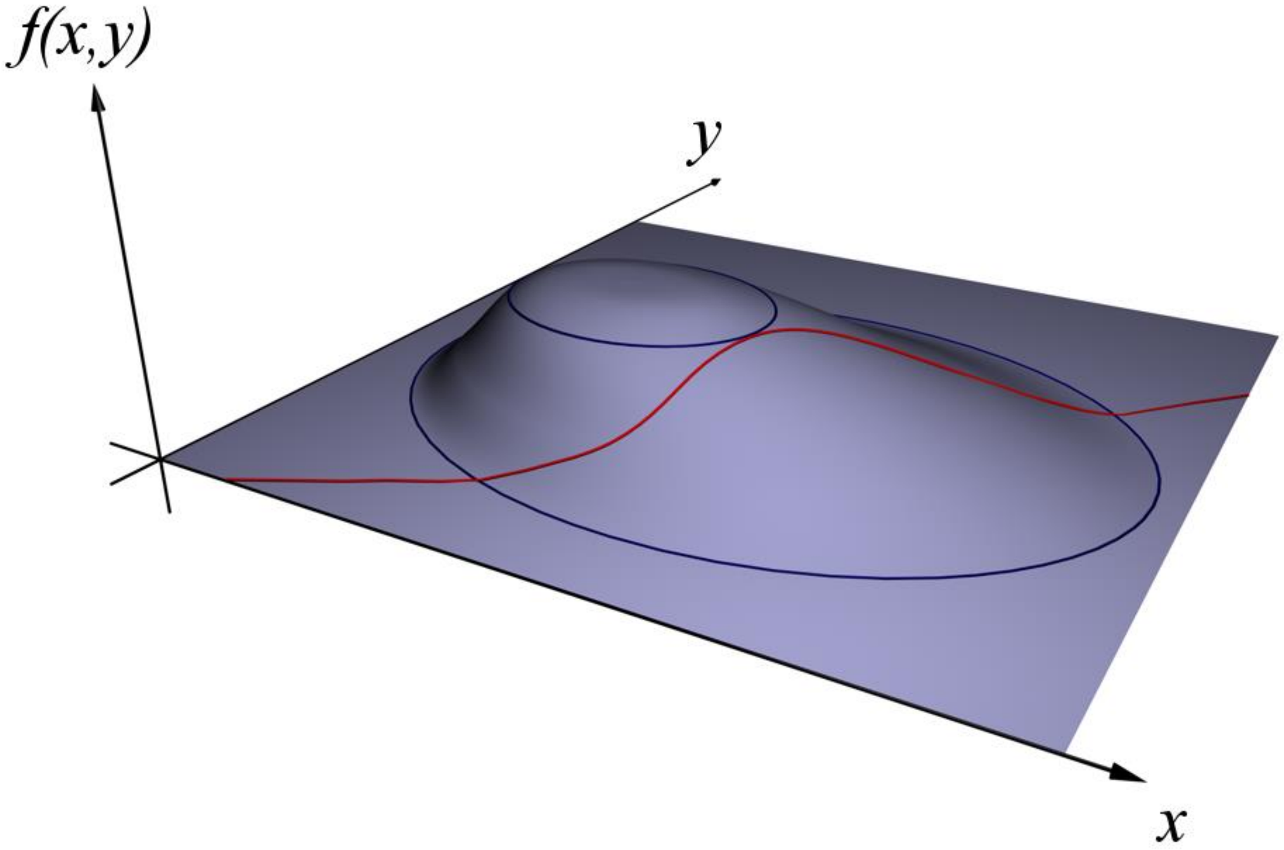
- Bellman, R.E. and Zadeh, L.A. (1970) Decision Making in a Fuzzy Environment. Management Sciences, 17, 141-164. http://dx.doi.org/10.1287/mnsc.17.4.B141

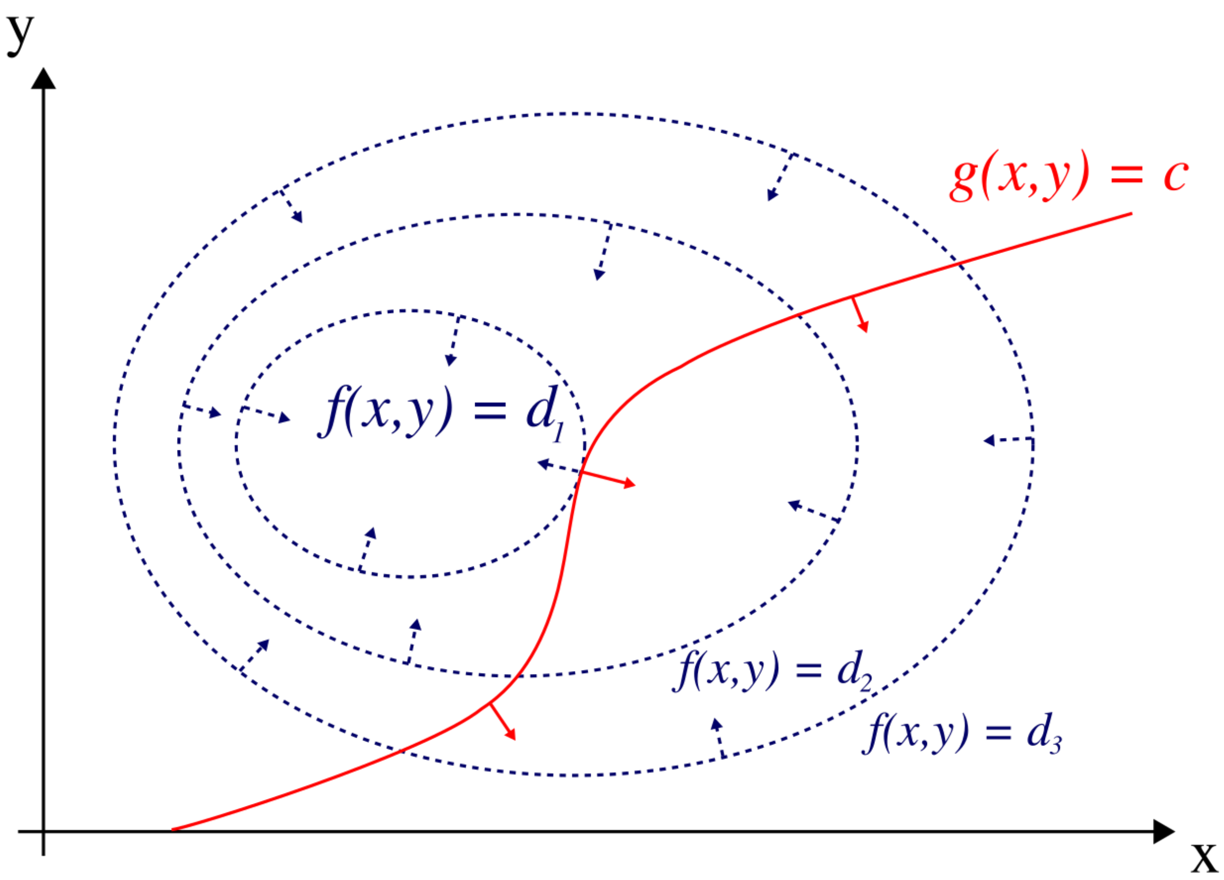
Lagrange Multipliers and Constrained Optimization



Basic Idea

- The basic idea is to convert a **constrained problem** into a form such that the derivative test of an **unconstrained problem** can still be applied. **The relationship between the gradient of the function and gradients of the constraints** rather naturally leads to a reformulation of the original problem, known as the **Lagrangian function or Lagrangian**.


f(x,y)
y
x



y
x
$g(x,y) = c$
$f(x,y) = d_1$
$f(x,y) = d_2$
$f(x,y) = d_3$

The method of Lagrange multipliers relies on the intuition that at a maximum, $f(x, y)$ cannot be increasing in the direction of any such neighboring point that also has $g = 0$. If it were, we could walk along $g = 0$ to get higher, meaning that the starting point wasn't actually the maximum. Viewed in this way, it is an exact analogue to testing if the derivative of an unconstrained function is 0, that is, we are verifying that the directional derivative is 0 in any relevant (viable) direction.

We can visualize contours of f given by $f(x, y) = d$ for various values of d, and the contour of g given by $g(x, y) = c$.

There are two ways this could happen:

1. We could touch a contour line of f, since by definition f does not change as we walk along its contour lines. This would mean that the tangents to the contour lines of f and g are parallel here.
2. We have reached a "level" part of f, meaning that f does not change in any direction.

To check the first possibility (we touch a contour line of f), notice that since the gradient of a function is perpendicular to the contour lines, the tangents to the contour lines of f and g are parallel if and only if the gradients of f and g are parallel. Thus we want points (x, y) where $g(x, y) = c$ and

$$\nabla_{x,y} f = \lambda \nabla_{x,y} g,$$

for some λ

where

$$\nabla_{x,y} f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right), \qquad \nabla_{x,y} g = \left(\frac{\partial g}{\partial x}, \frac{\partial g}{\partial y} \right)$$

are the respective gradients. The constant λ is required because although the two gradient vectors are parallel, the magnitudes of the gradient vectors are generally not equal. This constant is called the Lagrange multiplier. (In some conventions λ is preceded by a minus sign).

Notice that this method also solves the second possibility, that f is level: if f is level, then its gradient is zero, and setting $\lambda = 0$ is a solution regardless of $\nabla_{x,y} g$.

To incorporate these conditions into one equation, we introduce an auxiliary function

$$\mathcal{L}(x, y, \lambda) \equiv f(x, y) + \lambda \cdot g(x, y) \, ,$$

and solve

$$\nabla_{x,y,\lambda} \mathcal{L}(x, y, \lambda) = 0 \, .$$

Note that this amounts to solving three equations in three unknowns. This is the method of Lagrange multipliers.

Note that $\nabla_\lambda \mathcal{L}(x, y, \lambda) = 0$ implies $g(x, y) = 0$, as the partial derivative of $\mathcal{L}$ with respect to λ is $g(x, y)$.

To summarize

$$\nabla_{x,y,\lambda} \mathcal{L}(x, y, \lambda) = 0 \iff \begin{cases} \nabla_{x,y} f(x, y) = -\lambda \nabla_{x,y} g(x, y) \\ g(x, y) = 0 \end{cases}$$

Multiple Constraints

We are still interested in finding points where f does not change as
we walk, since these points might be (constrained) extrema. We therefore seek $\mathbf{x}$ such that any allowable direction of movement away from $\mathbf{x}$ is perpendicular to $\nabla f(\mathbf{x})$ (otherwise we could increase f by moving along that allowable direction). In other words, $\nabla f(\mathbf{x}) \in A^{\perp} = S$. Thus there are scalars
$\lambda_1, \lambda_2, \ \ldots, \lambda_M$ such that

$$\nabla f(\mathbf{x}) = \sum_{k=1}^{M} \lambda_k \nabla g_k(\mathbf{x}) \quad \Longleftrightarrow \quad \nabla f(\mathbf{x}) - \sum_{k=1}^{M} \lambda_k \nabla g_k(\mathbf{x}) = 0 \,.$$

These scalars are the Lagrange multipliers. We now have M of them, one for every constraint.

As before, we introduce an auxiliary function

$$\mathcal{L}\left(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_M\right) = f\left(x_1, \ldots, x_n\right) - \sum_{k=1}^{M} \lambda_k g_k\left(x_1, \ldots, x_n\right)$$

and solve

$$\nabla_{x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_M} \mathcal{L}(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_M) = 0 \iff \begin{cases} \nabla f(\mathbf{x}) - \sum_{k=1}^{M} \lambda_k \nabla g_k(\mathbf{x}) = 0 \\ g_1(\mathbf{x}) = \cdots = g_M(\mathbf{x}) = 0 \end{cases}$$

which amounts to solving $n + M$ equations in $n + M$ unknowns.

The constraint qualification assumption when there are multiple constraints is that the constraint gradients at the relevant point are linearly independent.

KKT Necessary Conditions for Optimality

KKT: Karush–Kuhn–Tucker conditions

Basic Idea

- the **Karush–Kuhn–Tucker (KKT) conditions**,
  also known as the **Kuhn–Tucker conditions**, are first derivative tests (sometimes called first-order **necessary conditions**) for a solution in nonlinear programming to be optimal,
  provided that some regularity conditions are satisfied.
- **Allowing inequality constraints**, the KKT approach to nonlinear programming generalizes the method of Lagrange multipliers, which allows only equality constraints.
- Harold W. Kuhn and Albert W. Tucker, 1951.
- William Karush, in his master's thesis, 1939.

Consider the following nonlinear optimization problem in standard form:

minimize $f(\mathbf{x})$

subject to

$$g_i(\mathbf{x}) \leq 0,$$
$$h_j(\mathbf{x}) = 0.$$

where $\mathbf{x} \in \mathbf{X}$ is the optimization variable chosen from a convex subset of $\mathbb{R}^n$, f is the objective or utility function, g_i $(i = 1, \ldots, m)$ are the inequality constraint functions and h_j $(j = 1, \ldots, \ell)$ are the equality constraint functions. The numbers of inequalities and equalities are denoted by m and ℓ respectively.

Corresponding to the constrained optimization problem one can form the Lagrangian function

$$\mathcal{L}(\mathbf{x}, \mu, \lambda) = f(\mathbf{x}) + \mu^\top \mathbf{g}(\mathbf{x}) + \lambda^\top \mathbf{h}(\mathbf{x}) = L(\mathbf{x}, \alpha) = f(\mathbf{x}) + \alpha^\top \begin{pmatrix} \mathbf{g}(\mathbf{x}) \\ \mathbf{h}(\mathbf{x}) \end{pmatrix}$$

where

$$\mathbf{g}(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_i(\mathbf{x}) \\ \vdots \\ g_m(\mathbf{x}) \end{bmatrix}, \quad \mathbf{h}(\mathbf{x}) = \begin{bmatrix} h_1(\mathbf{x}) \\ \vdots \\ h_j(\mathbf{x}) \\ \vdots \\ h_\ell(\mathbf{x}) \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_i \\ \vdots \\ \mu_m \end{bmatrix}, \quad \lambda = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_j \\ \vdots \\ \lambda_\ell \end{bmatrix} \quad \text{and} \quad \alpha = \begin{bmatrix} \mu \\ \lambda \end{bmatrix}.$$

The Karush–Kuhn–Tucker theorem

Theorem — (sufficiency) If $(\mathbf{x}^*, \alpha^*)$ is a saddle point of $L(\mathbf{x}, \alpha)$ in $\mathbf{x} \in \mathbf{X}$, $\mu \geq \mathbf{0}$, then $\mathbf{x}^*$ is an optimal vector for the above optimization problem.

(necessity) Suppose that $f(\mathbf{x})$ and $g_i(\mathbf{x})$, $i = 1, \ldots, m$, are convex in $\mathbf{X}$ and that there exists $\mathbf{x}_0 \in \mathbf{relint}(\mathbf{X})$ such that $\mathbf{g}(\mathbf{x}_0) < \mathbf{0}$ (i.e., Slater's condition holds). Then with an optimal vector $\mathbf{x}^*$ for the above optimization problem there is associated a vector $\alpha^* = \begin{bmatrix} \mu^* \\ \lambda^* \end{bmatrix}$ satisfying $\mu^* \geq \mathbf{0}$ such that $(\mathbf{x}^*, \alpha^*)$ is a saddle point of $L(\mathbf{x}, \alpha)$.[5]

KKT necessary conditions

The necessary conditions can be written with Jacobian matrices of the constraint functions. Let $\mathbf{g}(x) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be defined as $\mathbf{g}(x) = (g_1(x), \ldots, g_m(x))^\top$ and let $\mathbf{h}(x) : \mathbb{R}^n \rightarrow \mathbb{R}^\ell$ be defined as $\mathbf{h}(x) = (h_1(x), \ldots, h_\ell(x))^\top$. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\top$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_\ell)^\top$. Then the necessary conditions can be written as:

Stationarity

For maximizing $f(x)$: $\partial f(x^*) - D\mathbf{g}(x^*)^\top \boldsymbol{\mu} - D\mathbf{h}(x^*)^\top \boldsymbol{\lambda} = \mathbf{0}$

For minimizing $f(x)$: $\partial f(x^*) + D\mathbf{g}(x^*)^\top \boldsymbol{\mu} + D\mathbf{h}(x^*)^\top \boldsymbol{\lambda} = \mathbf{0}$

Primal feasibility

$$\mathbf{g}(x^*) \leq \mathbf{0}$$

$$\mathbf{h}(x^*) = \mathbf{0}$$

Dual feasibility

$$\boldsymbol{\mu} \geq \mathbf{0}$$

Complementary slackness

$$\boldsymbol{\mu}^\top \mathbf{g}(x^*) = 0.$$

Alternating Optimization

اشاره شد

What is Optimal?

Take home advice!

Living in a world doomed by uncertainty

What Is Optimal?

LOTFI A. ZADEH

How r solutions? designing fications. filtering this attitu componer fetish of one sense

a system. In any case, neither Wiener's theory nor the more sophisticated approaches of decision theory have resolved the basic problem of how to find a "best" or even a "good" system under uncertainty.

cations. ts, and, tem due m. system ero-sum rinciple nature ose loss

we are apt to place too much confidence in a system is the designer's gain. In a modification of the