Theory of Constrained Optimization

Signal Processing Graduate Seminar III

Lucio Azzari

lucio.azzari@tut.fi

Tampere University of Technology

February 06, 2013

Introduction

- Optimization problems are widely common in Signal Processing, as well as other fields (economics, etc.).
- It can occur the necessity to optimize some functions under some constrains, e.g., the minimization of a function that relates the temperature of a gas (under some hypothesis) to its pressure, and we want that the solution has to be found for $t \geq -273.15^o$.
- For constrained optimization we intend that the solution has to be found in a region arbitrarily decided, or, equivalently, that the solution has to satisfy the constrains in addition to the optimization.

Outline

Constrained optimization

- We refer to optimization problems that can be expressed as

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{h}(\mathbf{x}) = \mathbf{0}, \\ & \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \\ & \mathbf{x} \in \Omega \subset \mathbb{R}^n, \end{array} \tag{1}$$

  with $\mathbf{h} = (h_1, h_2, \ldots, h_m)$, and $\mathbf{g} = (g_1, g_2, \ldots, g_p)$.
- $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ are said *equality* constrains.
- $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$ are said *inequality* constrains.
- A point $\mathbf{x} \in \Omega$ is said *feasible* if it satisfies all the constraints.
- We consider $h_i, g_i \in C^1$ (they are smooth).

Equality constrain - Lagrange multipliers

Problem

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{h}(\mathbf{x}) = \mathbf{0}. \end{array} \tag{2}$$

Solution

Use of Lagrange multipliers.
We build, and solve the system of equations

$$\begin{aligned} \nabla_{\mathbf{x}} f(\mathbf{x}) + \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}) \boldsymbol{\lambda} &= \mathbf{0}, \\ \mathbf{h}(\mathbf{x}) &= \mathbf{0}, \qquad \text{with } \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m). \end{aligned} \tag{3}$$

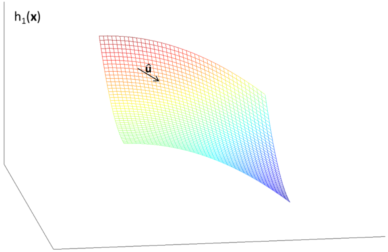
Being $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$, and $\mathbf{h}(\mathbf{x})$ composed by m equations, the total number of equations that we have is $n + m$, that is equal to the number of unknown.

Geometrical interpretation of Lagrange multipliers

- A local minimum/maximum (without any constrains) of a function $f(\mathbf{x})$ occurs when $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{0}$.
- When considering a constrained min/max along the constrain $h_1(\mathbf{x}) = 0$, the min/max point satisfies the condition

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}|_{\hat{u}}} = \nabla_{\mathbf{x}} f(\mathbf{x}) \cdot \hat{u} = \mathbf{0}, \tag{4}$$

where $\hat{u}$ is a versor pointing in any directions tangent to the curve $h_1(\mathbf{x}) = 0$.



Geometrical interpretation of Lagrange multipliers

- From Equation (4) we have that if $\mathbf{x}^*$ is a min/max point, then $\nabla_{\mathbf{x}} f(\mathbf{x}^*) \perp \tan h_1(\mathbf{x}^*)$.
- We know another vector that is *always* orthogonal to $h_1(\mathbf{x})$:

$$\nabla_{\mathbf{x}} h_1(\mathbf{x}) \perp \tan h_1(\mathbf{x}), \quad \forall \mathbf{x} \quad \in \Omega.$$

- Therefore, in a min/max point $\mathbf{x}^*$

$$\nabla_{\mathbf{x}} f(\mathbf{x}^*) \ /\!/ \ \nabla_{\mathbf{x}} h_1(\mathbf{x}^*).$$

- Imposing the condition

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \lambda_1 \nabla_{\mathbf{x}} h_1(\mathbf{x}),$$

  with the constrain

$$h_1(\mathbf{x}) = 0,$$

  we can find the min/max point $\mathbf{x}^*$.

General formula for Lagrange multipliers

- Considering multiple equality constrains $\mathbf{h} = (h_1, h_2, \ldots, h_m) = \mathbf{0}$

$$\nabla_{\mathbf{x}} f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \nabla_{\mathbf{x}} h_i(\mathbf{x}) = \nabla_{\mathbf{x}} f(\mathbf{x}) + \nabla_{\mathbf{x}} \mathbf{h}\left(\mathbf{x}\right) \lambda = \mathbf{0}.$$

- It is possible to express the system of equations in a compact form using the *Lagrangian*

$$L\left(\mathbf{x}, \lambda\right) = f(\mathbf{x}) + \mathbf{h}\left(\mathbf{x}\right) \lambda$$

in the following way:

$$\begin{aligned} \nabla_{\mathbf{x}} L\left(\mathbf{x}, \lambda\right) &= \mathbf{0}, \\ \nabla_{\lambda} L\left(\mathbf{x}, \lambda\right) &= \mathbf{0}. \end{aligned} \tag{5}$$

Second order conditions

Problem

If $f(\mathbf{x})$ is not *convex* or *concave* in the set of points $\mathbf{x} \in \Omega : \mathbf{h}(\mathbf{x}) = \mathbf{0}$, $\mathbf{x}^*$ can be a *minimum*, *maximum*, or a *saddle*.

We need information about the extremum.

Solution

We introduce the second order conditions in order to determinate weather an extremum is a minimum, a maximum or neither.

In practice we exploit the information about the second derivative (similarly to the unconstrained optimization), in order to determine the nature of $\mathbf{x}^*$.

Second derivative of the Lagrangian

- Let us denote the $n \times n$ Hessian of $f(\mathbf{x})$ as

$$(F(\mathbf{x}))_{i,j} = \left[\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}\right], \quad i,j = 1, \ldots, n \ .$$

- The constraining part of the Lagrangian can be expressed as

$$(H(\mathbf{x}, \lambda))_{i,j} = \sum_{k=1}^{m} \frac{\partial^2 h_k(\mathbf{x})}{\partial x_i \partial x_j} \lambda_k, \quad i,j = 1, \ldots, n \ .$$

- We can finally introduce the second partial derivative of the Lagrangian

$$\mathbf{L}(\mathbf{x}) = F(\mathbf{x}) + H(\mathbf{x}, \lambda). \tag{6}$$

Condition for minimum/maximum

Theorem

If $\mathbf{x}^*$ *satisfies* $\mathbf{h}(\mathbf{x}^*) = \mathbf{0}$*, and there is a* λ *such that*

$$\nabla_{\mathbf{x}} f(\mathbf{x}^*) + \nabla_{\mathbf{x}} \mathbf{h}\left(\mathbf{x}^*\right) \lambda = \mathbf{0},$$

and there is a matrix $\mathbf{L}(\mathbf{x}^*, \lambda)$ *that is positive definite in the tangent plane* P *of* $\mathbf{h}(\mathbf{x})$ *at* $\mathbf{x}^*$*, then* $\mathbf{x}^*$ *is a strict local minimum of* $f(\mathbf{x}^*)$*.*
On the contrary, if $\mathbf{L}(\mathbf{x}^*, \lambda)$ *is negative definite,* $\mathbf{x}^*$ *is a maximum of* $f(\mathbf{x}^*)$*.*

Reminder

$$\frac{d}{d\xi} f\left(\mathbf{x}\left(\xi\right)\right) = \frac{df(\mathbf{x})}{d\mathbf{x}} \frac{d\mathbf{x}\left(\xi\right)}{d\xi} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{dx_i}{d\xi} = \left(\nabla_{\mathbf{x}} f(\mathbf{x})\right)^T \dot{\mathbf{x}}.$$

$$\frac{d^2}{d\xi^2} f\left(\mathbf{x}\left(\xi\right)\right) = \frac{d}{d\xi}\left(\frac{d}{d\xi} f\left(\mathbf{x}\left(\xi\right)\right)\right) = \frac{d}{d\xi}\left(\frac{df(\mathbf{x})}{d\mathbf{x}} \frac{d\mathbf{x}(\xi)}{d\xi}\right) = \frac{df(\mathbf{x})}{d\mathbf{x}} \ddot{\mathbf{x}} + \frac{d}{d\xi}\left(\frac{df(\mathbf{x})}{d\mathbf{x}}\right) \dot{\mathbf{x}} = \left(\nabla_{\mathbf{x}} f\left(\mathbf{x}\right)\right)^T \ddot{\mathbf{x}} + \dot{\mathbf{x}}^T F \dot{\mathbf{x}}.$$

Condition for minimum/maximum

Proof.

Let $\mathbf{x}(\xi)$ be a curve on the surface $\mathbf{h}(\mathbf{x}) = \mathbf{0}$, with $\mathbf{x}(0) = \mathbf{x}^*$. We can write the Taylor series of $\mathbf{f}(\mathbf{x})$ and $h_i(\mathbf{x})$, $i = 1, \ldots, m$, around $\mathbf{x}^*$:

$$\begin{aligned} f\left(\mathbf{x}\left(\xi\right)\right) = \quad & f\left(\mathbf{x}^*\right) + \xi(\nabla_{\mathbf{x}} f\left(\mathbf{x}\left(0\right)\right))^T \dot{\mathbf{x}} + \\ & + \frac{\xi^2}{2}\left[\left(\nabla_{\mathbf{x}} f\left(\mathbf{x}^*\right)\right)^T \ddot{\mathbf{x}}\left(0\right) + \dot{\mathbf{x}}(0)^T F \dot{\mathbf{x}}\left(0\right)\right] + o\left(\xi^2\right). \end{aligned}$$

$$\begin{aligned} h_i\left(\mathbf{x}\left(\xi\right)\right) = \quad & h_i\left(\mathbf{x}^*\right) + \xi(\nabla_{\mathbf{x}} h_i\left(\mathbf{x}\left(0\right)\right))^T \dot{\mathbf{x}} + \\ & + \frac{\xi^2}{2}\left[\left(\nabla_{\mathbf{x}} h_i\left(\mathbf{x}^*\right)\right)^T \ddot{\mathbf{x}}\left(0\right) + \dot{\mathbf{x}}(0)^T H_i \dot{\mathbf{x}}\left(0\right)\right] + o\left(\xi^2\right). \end{aligned}$$

We now multiply $h_i\left(\mathbf{x}\left(\xi\right)\right)$ by their respective λ_i, and then sum them to $f\left(\mathbf{x}\left(\xi\right)\right)$.

Taking into account that the sum of the first derivatives is zero (by hypothesis), we obtain

$$f\left(\mathbf{x}\left(\xi\right)\right) - f\left(\mathbf{x}^*\right) = \frac{\xi^2}{2} \dot{\mathbf{x}}(0)^T \mathbf{L}\left(\mathbf{x}^*\right) \dot{\mathbf{x}}\left(0\right) + o\left(\xi^2\right).$$

Being $\mathbf{L}$ positive definite, $f\left(\mathbf{x}^*\right)$ is a local minimum. $\square$

Tangent plane P

- P is defined as
$$P = \left\{ \mathbf{y} : \nabla \mathbf{h}(\mathbf{x}^*)^T \mathbf{y} = 0 \right\}.$$
- The product $\mathbf{Ly}$ in general light not be in P. Therefore, we have to project $\mathbf{Ly}$ into P.
- We can find a orthonormal basis for P $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{n-m}$, and we can define the matrix $E = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{n-m}]$.
- Any point $\mathbf{y}$ can be written as $\mathbf{y} = E\mathbf{z}$, for some $\mathbf{z} \in \mathbb{R}^{n-m}$.
- The projection of $\mathbf{Ly}$ on P is
$$EE^T \mathbf{Ly} = EE^T \mathbf{L} E \mathbf{z}.$$
$E^T \mathbf{L} E$ gives the coordinates of the projection in terms of basis E.
- The positive or negative definiteness of $\mathbf{L}$ can be determined by finding the eigenvalues of $E^T \mathbf{L} E$.

Interpretation of Lagrange multipliers

Problem

We want to investigate the relation between the function to minimize and the constrains.

Solution

Let us consider the constrain as $\mathbf{h}(\mathbf{x}) = \mathbf{c}$, and the optimal solution $\mathbf{x}^*$ is found when $\mathbf{c} = \mathbf{0}$, with the respective Lagrange multipliers $\boldsymbol{\lambda}^*$.

For $\mathbf{c}$ sufficiently small around $\mathbf{0}$, we can consider the solution to the min/max problem continuously dependent of $\mathbf{c}$. Let us denote this solution with $\mathbf{x}(\mathbf{c})$.

We want to compute the derivative of $f(\mathbf{x})$ respect to the variable $\mathbf{c}$, in order to study the *sensitivity* of the minimum/maximum respect to some *relaxation* of the constrains.

Computation of $\nabla_{\mathbf{c}} \mathbf{h}(\mathbf{x}(\mathbf{c}))$

- Let us note that $\nabla_{\mathbf{c}} \mathbf{h}(\mathbf{x}(\mathbf{c})) = I$.
-
$$\nabla_{\mathbf{c}} h_i(\mathbf{x}(\mathbf{c})) = \begin{bmatrix} \sum_{i=1}^{m} \frac{\partial}{\partial x_k} h_i(\mathbf{x}(\mathbf{c})) \frac{\partial x_k}{\partial c_1} \\ \vdots \\ \sum_{i=1}^{m} \frac{\partial}{\partial x_k} h_i(\mathbf{x}(\mathbf{c})) \frac{\partial x_k}{\partial c_m} \end{bmatrix} = \begin{bmatrix} \Delta_{c_1}(\mathbf{x}(\mathbf{c}))^T \nabla_{\mathbf{x}} h_i(\mathbf{x}(\mathbf{c})) \\ \vdots \\ \Delta_{c_m}(\mathbf{x}(\mathbf{c}))^T \nabla_{\mathbf{x}} h_i(\mathbf{x}(\mathbf{c})) \end{bmatrix},$$

  with the notation $\Delta_{c_i}(\mathbf{x}(\mathbf{c}))^T = \left[\frac{\partial x_1}{\partial c_i}, \ldots, \frac{\partial x_n}{\partial c_i}\right]$.
- Stacking these together we have

$$\nabla_{\mathbf{c}} \mathbf{h}(\mathbf{x}(\mathbf{c})) = \begin{bmatrix} \Delta_{c_1}(\mathbf{x}(\mathbf{c}))^T \nabla_{\mathbf{x}} h_1(\mathbf{x}(\mathbf{c})) & \cdots & \Delta_{c_1}(\mathbf{x}(\mathbf{c}))^T \nabla_{\mathbf{x}} h_m(\mathbf{x}(\mathbf{c})) \\ \vdots & \ddots & \vdots \\ \Delta_{c_m}(\mathbf{x}(\mathbf{c}))^T \nabla_{\mathbf{x}} h_1(\mathbf{x}(\mathbf{c})) & \cdots & \Delta_{c_m}(\mathbf{x}(\mathbf{c}))^T \nabla_{\mathbf{x}} h_m(\mathbf{x}(\mathbf{c})) \end{bmatrix}.$$

- We can conclude that

$$\Delta_{c_i}(\mathbf{x}(\mathbf{c}))^T \nabla_{\mathbf{x}} h_j(\mathbf{x}(\mathbf{c})) = \delta_{i,j}.$$

Computation of $\nabla_{\mathbf{c}} f(\mathbf{x}(\mathbf{c}))$

- The i-th derivative can be expressed as

$$[\nabla_{\mathbf{c}} f(\mathbf{x}(\mathbf{c}))]_i = \sum_{k=1}^{n} \frac{\partial}{\partial x_k} f(\mathbf{x}(\mathbf{c})) \frac{\partial x_k}{\partial c_i} = \Delta_{c_i}(\mathbf{x}(\mathbf{c}))^T \nabla_{\mathbf{x}} f(\mathbf{x}(\mathbf{c})).$$

- Using the constrain

$$[\nabla_{\mathbf{c}} f(\mathbf{x}(\mathbf{c}))]_i = -\Delta_{c_i}(\mathbf{x}(\mathbf{c}))^T \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}(\mathbf{c}))\, \boldsymbol{\lambda}.$$

- As shown in the previous slide

$$\Delta_{c_i}(\mathbf{x}(\mathbf{c}))^T \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}(\mathbf{c})) = \mathbf{e}_i,$$

  where $\mathbf{e}_i$ is a zero vector with one 1 in the i-th position.
- Finally

$$\nabla_{\mathbf{c}} f(\mathbf{x}(\mathbf{c})) = -\boldsymbol{\lambda}.$$

  This means that the Lagrange multipliers indicates the variation rate of the minimum/maximum when the constrains are slightly changed from their optimal values.

Kuhn-Tucker conditions

Problem

Let us now consider the inequality constrained minimization

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{h}(\mathbf{x}) = \mathbf{0}, \\ & \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \end{array}$$

with m equality and p inequality constrains.
Only active constrains affect directly the solution.

Solution (Kuhn-Tucker conditions)

Let $\mathbf{x}^*$ be a local minimum that satisfies the above conditions, then there are a $\boldsymbol{\lambda} \in \mathbb{R}^m$, and a $\boldsymbol{\mu} \in \mathbb{R}^p$, for which

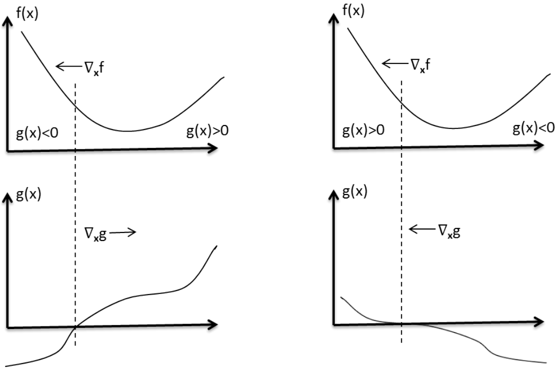
$$\boldsymbol{\mu} \geq \mathbf{0}, \tag{7}$$

$$\mathbf{g}(\mathbf{x}^*)^T \boldsymbol{\mu} = 0, \tag{8}$$

$$\nabla f(\mathbf{x}^*) + \nabla \mathbf{h}(\mathbf{x}^*)\, \boldsymbol{\lambda} + \nabla \mathbf{g}(\mathbf{x}^*)\, \boldsymbol{\mu} = \mathbf{0}. \tag{9}$$

Equation (8) is called *complementary* condition: μ_i is zero when the constrain is not (when the constrain is inactive).

Interpretation of inequality constrains



- An inactive constrain has $\mu = 0$.
- In order to be active, the gradient of the constrain has to have opposite sign of the gradient of $f(x)$:

$$\nabla f(x) + \mu \nabla g(x) = 0,$$

therefore $\mu \geq 0$.

Procedure

The solution is found using various combinations of active constrains:

1. Set one or more coefficients $\mu_i = 0$;
2. Solve the system of equations and find the other $\mu_k,\ k \neq i$;
3. If $\boldsymbol{\mu} \geq 0$ we found a solution, otherwise go to point 1 and zero other coefficients.

Example

$$\begin{array}{ll} \text{minimize} & f(x_1, x_2) = 3x_1^2 + 4x_2^2 + 6x_1x_2 - 6x_1 - 8x_2 \\ \text{subjectto} & g_1(x_1, x_2) = x_1^2 + x_2^2 - 9 \leq 0, \\ & g_2(x_1, x_2) = 2x_1 - x_2 - 4 \leq 0. \end{array}$$

- Trying g_1 active ($\mu_1 > 0$) and g_2 inactive ($\mu_2 = 0$) leads to negative value for μ_1.
- Analogously $\mu_2 < 0$ if we set $\mu_1 = 0$.
- If consider both active ($\mu_i > 0$), we find that both are negative.
- We can conclude that are both inactive!
- We set both equal to zero and we find the unconstrained minimum.

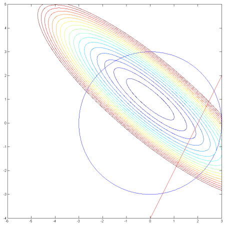


Figure: It is possible to notice that the minimum of the function is already in the feasible region.

Second order inequality conditions

Theorem

Let $\mathbf{x}^*$ be a regular point of $\mathbf{h}$ and the active constrains in $\mathbf{g}$, and let P be the tangent space of the active constrains at $\mathbf{x}^*$.
Then, if $\mathbf{x}^*$ is a local minimum of f,

$$\mathbf{L}(\mathbf{x}^*) = F(\mathbf{x}^*) + \mathbf{H}(\mathbf{x}^*, \lambda) + \mathbf{G}(\mathbf{x}^*, \boldsymbol{\mu})$$

is positive semidefinite on the tangent subspace of the active constrains.
In the above notation

$$\begin{aligned}
[\mathbf{H}(\mathbf{x}^*, \lambda)]_{\forall i,j} &= \sum_{k=1}^{m} \frac{\partial^2 h_k(\mathbf{x}^*)}{\partial x_i \partial x_i} \lambda_k, \\
[\mathbf{G}(\mathbf{x}^*, \boldsymbol{\mu})]_{\forall i,j} &= \sum_{k=1}^{p} \frac{\partial^2 g_k(\mathbf{x}^*)}{\partial x_i \partial x_i} \mu_k.
\end{aligned}$$

Exercise

Problem

Let us suppose to have n transmission channels, each one corrupted by WGN with different variance $(N_1, N_2, \ldots, N_n)$.
How can we distribute the available power P along each channel in an optimal way?

Solution (1/2)

Knowing that the capacity of a channel i is

$$\frac{1}{2}\log\left(1+\frac{P_i}{N_i}\right),$$

we have the following constrained optimization problem of the form

$$L = \sum_{i=1}^{n}\frac{1}{2}\log\left(1+\frac{P_i}{N_i}\right) + \lambda\left(\sum_{i=1}^{n}P_i - P\right) + \sum_{i=1}^{n}\mu_i P_i.$$

Exercise

Solution (2/2)

If $P_i > 0$ (inactive constrain), then $\mu_i = 0$,

$$\frac{\partial L}{\partial P_i} = \frac{1}{2}\frac{1}{N_i + P_i} + \lambda = 0,$$

that gives

$$P_i + N_i = -\frac{1}{2\lambda}.$$

If $P_i = 0$ (active constrain), $\mu_i \leq 0$,

$$N_i = -\frac{1}{2(\lambda+\mu_i)} \quad \Rightarrow \quad N_i \geq -\frac{1}{2\lambda} \quad ,$$

with the constrain $\sum_{i=active} P_i = P$.

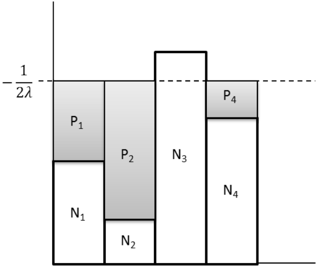


Figure: *Waterfall* solution to the maximization problem.