Optimization and Data Science

Lecture 21: Lagrange Multiplier Rule for Equality Constraints

Prof. Dr. Thomas Slawig

Kiel University - CAU Kiel
Dep. of Computer Science

Summer 2020

Contents

Lagrange multiplier rule

- For constrained problems, the Lagrange multiplier rule is the analogue to the first oder necessary condition

$$\nabla f(x^*) = 0 \tag{1}$$

  that we have for unconstrained problems.
- First, we consider problems with equality constraints only:

$$\min_{x \in \mathbb{R}^n} f(x) \ \text{ subject to } h(x) = 0, \quad \text{where } h = (h_i)_{i=1}^{p} : \mathbb{R}^n \to \mathbb{R}^p.$$

- The main idea is to replace f in (1) by another function.

Contents

Lagrange function and Lagrange mulitipliers

Definition

For the problem

$$\min_{x \in \mathbb{R}^n} f(x) \text{ subject to } h(x) = 0, \quad \text{where } h = (h_i)_{i=1}^p : \mathbb{R}^n \to \mathbb{R}^p,$$

we call the function $L : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$, defined by

$$L(x, \lambda) := f(x) + \lambda^\top h(x) = f(x) + \sum_{i=1}^{p} \lambda_i h_i(x),$$

the **Lagrange function** or **Lagrangian**.
The value λ_i is called **Lagrange multiplier** corresponding to the constraint $h_i(x) = 0$.

Example: Lagrange function and Lagrange mulitipliers

- Compute a rectangle with maximal area and given perimeter c:

$$\max_{x\in\mathbb{R}^2} x_1x_2 \ \text{ s.t. } \ 2(x_1+x_2)=c$$

- ... can be reformulated as:

$$\min_{x\in\mathbb{R}^2} (-x_1x_2) \ \text{ s.t. } \ h(x)=2(x_1+x_2)-c=0.$$

- We have $h:\mathbb{R}^n\to\mathbb{R}^p, p=1 \Rightarrow \lambda\in\mathbb{R}^p=\mathbb{R}$.

$\rightsquigarrow$ Lagrange function

$$L(x,\lambda)=f(x)+\lambda^\top h(x)=f(x)+\lambda h(x)=-x_1x_2+\lambda(2(x_1+x_2)-c).$$

Contents

Lagrange multiplier rule

Theorem

Let x^* be a regular point (w.r.t. $h(x) = 0$) and a local mimimizer (or maximizer) of f subject to the constraint $h(x) = 0$. Then, there exists a **Lagrange multiplier** $\lambda \in \mathbb{R}^p$ *with*

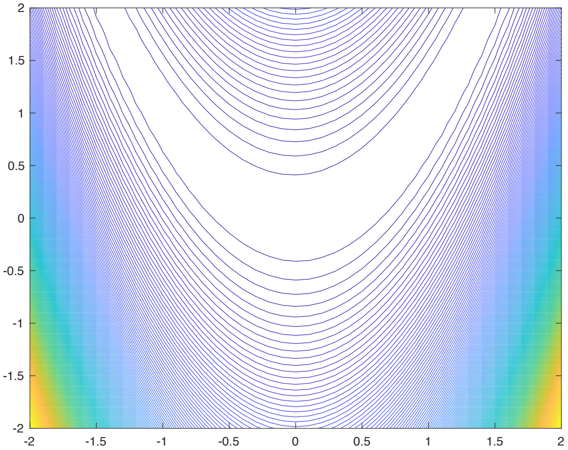
$$\nabla f(x^*) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) = 0.$$

Definition

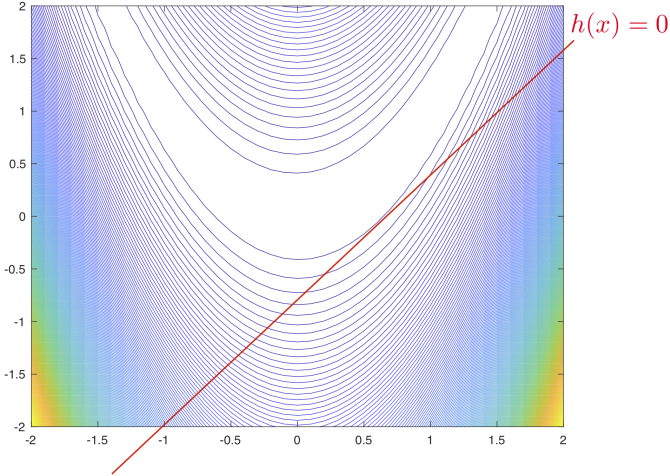
A point $x \in X_{ad}$ is called **regular** (w.r.t. the constraint $h(x) = 0$), if the set of gradients $\{\nabla h_i(x) : i = 1, \ldots, p\}$ is linear independent.

- The regularity condition above for the existence of Lagrange multipliers can be replaced/relaxed by a variety of other conditions that the constraints have to satisfy.
- These conditions are referred to as **constraint qualifications**.

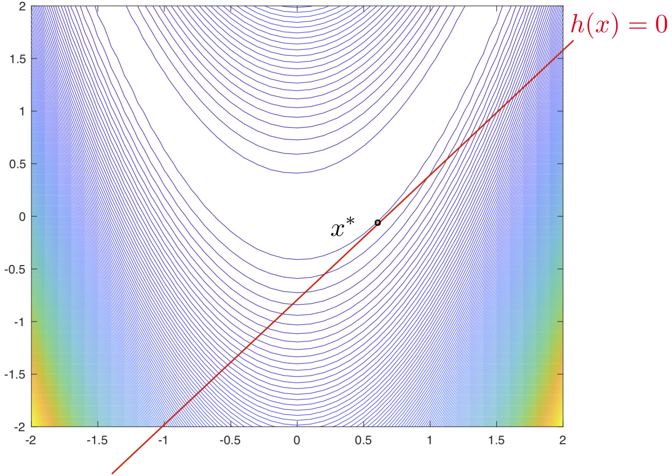
Example: Lagrange mulitiplier rule – Rosenbrock function



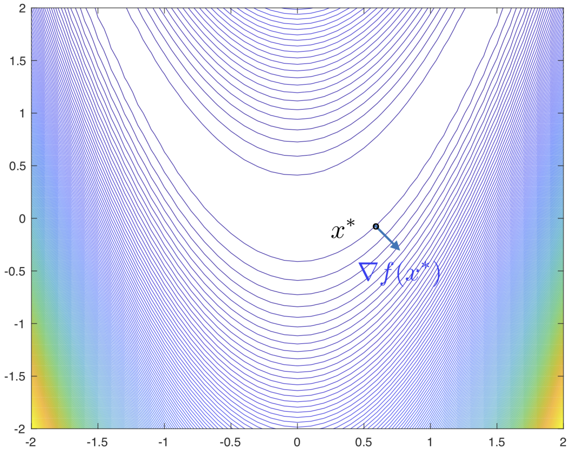
Rosenbrock function with equality constraint



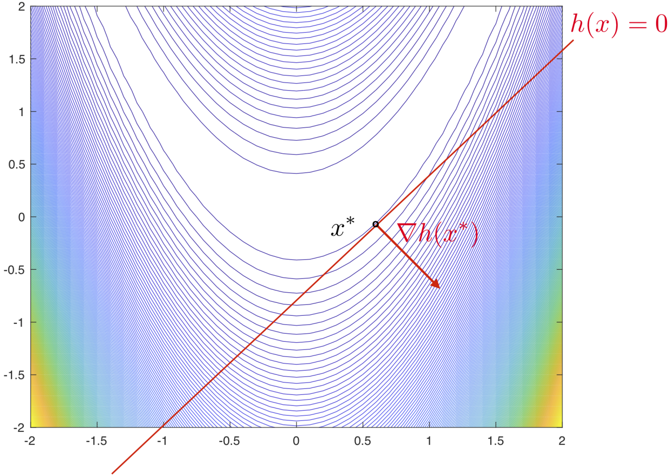
Minimizer x^* of the constrained problem



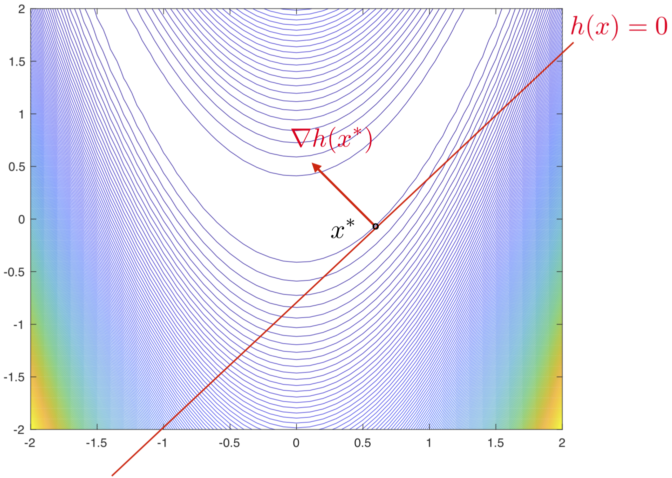
Gradient of the cost function at the minimizer



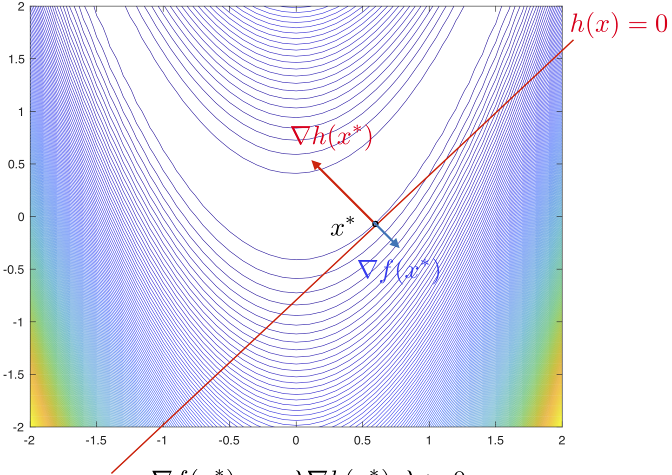
Gradient of the constraint at the minimizer ...



... or like this

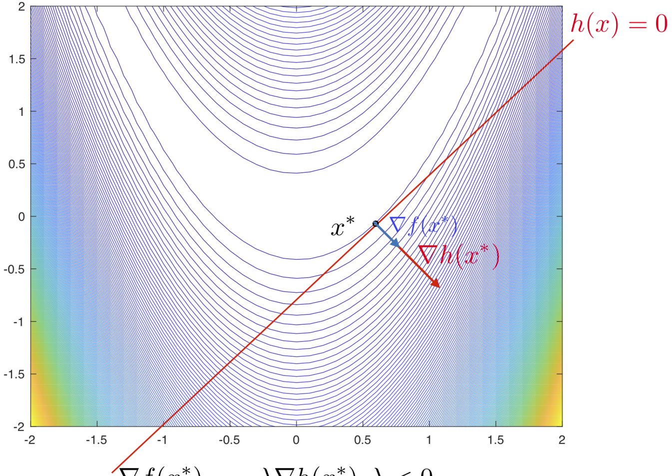


Lagrange mulitiplier rule: Both gradients are linear dependent ...



$$\nabla f(x^*) = -\lambda \nabla h(x^*), \lambda > 0$$

... or like this



$$\nabla f(x^*) = -\lambda \nabla h(x^*), \lambda < 0$$

Optimality system

- Summarizing, we look for a pair $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^p$ that satisfies

$$\begin{aligned} \nabla_x L(x,\lambda) = \quad & \nabla f(x) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x) && = 0 \\ \nabla_\lambda L(x,\lambda) = \quad & h(x) && = 0 \end{aligned}$$

- ... where:

$$L(x,\lambda) = f(x) + \lambda^\top h(x) = f(x) + \sum_{i=1}^{p} \lambda_i h_i(x).$$

- This system is called the **optimality system**, sometimes also **Karush-Kuhn-Tucker (KKT) conditions** or **KKT system**.
- It consists of $n+p$ equations for the $n+p$ unknowns (x, λ).

Example: Lagrange multiplier rule

- Compute a rectangle with maximal area and given perimeter c:

$$\min_{x\in\mathbb{R}^2} (-x_1x_2) \text{ s.t. } h(x) = 2(x_1+x_2) - c = 0.$$

- Lagrange function:

$$L(x,\lambda) = f(x) + \lambda h(x) = -x_1x_2 + \lambda(2(x_1+x_2) - c).$$

- Optimality system:

$$\begin{aligned}\nabla_x L(x,\lambda) &= \nabla f(x) + \lambda \nabla h(x) = 0\\ \nabla_\lambda L(x,\lambda) &= \qquad\qquad\quad h(x) = 0\end{aligned}$$

- Write down the optimality system for this example!

Example: Lagrange multiplier rule

- Compute a rectangle with maximal area and given perimeter c:

$$\min_{x\in\mathbb{R}^2} (-x_1x_2) \;\text{ s.t. }\; h(x) = 2(x_1 + x_2) - c = 0.$$

- Optimality system:

$$\left.\begin{aligned} \nabla_x L(x,\lambda) &= \nabla f(x) + \lambda \nabla h(x) = 0 \\ \nabla_\lambda L(x,\lambda) &= h(x) = 0 \end{aligned}\right\} \Longrightarrow \left\{ \begin{aligned} -x_2 + 2\lambda &= 0 \\ -x_1 + 2\lambda &= 0 \\ 2(x_1 + x_2) - c &= 0 \end{aligned}\right.$$

- Compute the solution (x_1, x_2, λ)!

Example: Lagrange Multiplier Rule/KKT System

- Optimality system:

$$\begin{aligned}\nabla_x L(x,\lambda) &= \nabla f(x) + \lambda \nabla h(x) = \begin{pmatrix} -x_2 + 2\lambda \\ -x_1 + 2\lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ \nabla_\lambda L(x,\lambda) &= \qquad\qquad h(x) = 2(x_1 + x_2) - c = 0\end{aligned}$$

- $\Rightarrow x_1 = x_2 = 2\lambda$ (square).
- Constraint:

$$2(x_1 + x_2) = 2(2\lambda + 2\lambda) = 8\lambda = c \Rightarrow \lambda = \frac{c}{8}$$

- $\Rightarrow x_1 = x_2 = \dfrac{c}{4}$.
- Is this point regular? Yes, since

$$\nabla h(x) = (2,2)^\top \neq 0 \text{ for all } x \in \mathbb{R}^2.$$

Contents

Why regular points?

Definition

A point $x \in X_{ad}$ is called **regular** (w.r.t. the constraint $h(x) = 0$), if the set of gradients $\{\nabla h_i(x) : i = 1, \dots, p\}$ is linear independent.

- The property of a point to be regular depends on the description of the admissible set (by the function h):

Example

- Consider as admissible set in $\mathbb{R}^2$ the x_1-axis, described as
$$X_{ad} = \{x \in \mathbb{R}^2 : h(x) = x_2 = 0\}.$$
- We may write the same set as
$$X_{ad} = \{x \in \mathbb{R}^2 : \tilde{h}(x) = x_2^2 = 0\}.$$

Compute the regular points in both cases.

Regular points

Example

For $h(x) = x_2 = 0$ we get:

- $\nabla h(x) = (0,1)^\top$ for all $x \in \mathbb{R}^2$
- ... also for all $x \in X_{ad}$.
- $\Rightarrow$ every point $x \in X_{ad}$ is regular w.r.t. $h(x) = x_2 = 0$.

For $\tilde{h}(x) = x_2^2 = 0$ we get:

- $\nabla \tilde{h}(x) = (0, 2x_2)^\top$ for $x \in \mathbb{R}^2$.
- For $x \in X_{ad}$ we have $x_2 = 0$
- ... and thus $\nabla \tilde{h}(x) = (0,0)^\top$ for all $x \in X_{ad}$.
- $\Rightarrow$ no point $x \in X_{ad}$ is regular w.r.t. $\tilde{h}(x) = x_2^2 = 0$.

Example: Lagrange multiplier rule and regular points

- Consider the two descriptions for the x_1-axis

$$X_{ad} = \{x \in \mathbb{R}^2 : h(x) = x_2 = 0\} = \{x \in \mathbb{R}^2 : \tilde{h}(x) = x_2^2 = 0\},$$

- ... and the cost function:

$$f(x_1, x_2) = (x_1 - 1)^2 + (x_2 + 1)^2.$$

- Minimizer of f in X_{ad} is $x^* = (1, 0)^\top$.
- Lagrange multiplier rule using h and $x_2 = 0$ in X_{ad} gives:

$$\nabla f(x) + \lambda \nabla h(x) = \begin{pmatrix} 2(x_1 - 1) \\ 2(x_2 + 1) \end{pmatrix} + \lambda \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \begin{cases} x_1 = 1 \\ \lambda = -2 \end{cases}$$

- ... using $\tilde{h}$ and $x_2 = 0$ in X_{ad}:

$$\nabla f(x) + \lambda \nabla \tilde{h}(x) = \begin{pmatrix} 2(x_1 - 1) \\ 2(x_2 + 1) \end{pmatrix} + \lambda \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \begin{cases} x_1 = 1 \\ 1 \neq 0 \quad \text{contradiction!!!} \end{cases}$$

Contents

Second order sufficient condition for equality-constrained problems

Theorem (Second order sufficient optimality condition)

Let

- f, h *be twice continuously differentiable,*
- x^* *be regular,*
- (x^*, λ) *satisfy the optimality system,*
- *the Hessian matrix (w.r.t. x),* $\nabla^2_{xx} L(x^*, \lambda)$ *be positive definite on the set*

$$S := \{d \in \mathbb{R}^n : \nabla h_i(x^*)^\top d = 0, i = 1, \ldots, p\},$$

i.e.,

$$d^\top \left(\nabla^2 f(x^*) + \sum_{i=1}^{p} \lambda_i \nabla^2 h_i(x^*) \right) d > 0 \qquad \forall d \in S \setminus \{0\}.$$

Then x^* is a strict local minimizer of f in X_{ad}.

Example: Lagrange function and Lagrange mulitipliers

- Rectangle with maximal area and given perimeter c:

$$\min_{x\in\mathbb{R}^2} (-x_1x_2) \;\text{ s.t. }\; h(x) = 2(x_1+x_2) - c = 0.$$

- Solution was $(x^*, \lambda) = (\frac{c}{4}, \frac{c}{4}, \frac{c}{8})$.
- Lagrange function and gradient w.r.t. x:

$$L(x,\lambda) = -x_1x_2 + \lambda(2(x_1+x_2) - c), \quad \nabla_x L(x,\lambda) = \begin{pmatrix} -x_2 + 2\lambda \\ -x_1 + 2\lambda \end{pmatrix}.$$

- Compute the Hessian of the Lagrangian w.r.t. x!

$$\nabla^2_{xx} L(x,\lambda) = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$$

- Is the Hessian positive definite?

Example: Lagrange function and Lagrange mulitipliers

- Hessian not positive definite: Take $d = (1,1)^\top$:

$$d^\top \nabla^2_{xx} L(x^*,\lambda) d = \begin{pmatrix} 1 \\ 1 \end{pmatrix}^\top \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}^\top \begin{pmatrix} -1 \\ -1 \end{pmatrix} = -2 < 0$$

- But: Has to be positive definite only on

$$S = \left\{ d \in \mathbb{R}^2 : \nabla h(x^*)^\top d = 0 \right\}.$$

- Constraint and gradient:

$$h(x) = 2(x_1 + x_2) - c = 0, \quad \nabla h(x) = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \text{ (constant).}$$

- We compute

$$S = \left\{ d \in \mathbb{R}^2 : 2d_1 + 2d_2 = 0 \right\} = \left\{ d \in \mathbb{R}^2 : d_2 = -d_1 \right\}.$$

Example: Lagrange function and Lagrange mulitipliers

- Is the Hessian

$$\nabla^2_{xx} L(x,\lambda) = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$$

positive definite on this set

$$S = \left\{ d \in \mathbb{R}^2 : d_2 = -d_1 \right\}?$$

- Compute for $d \in S$:

$$d^\top \nabla^2_{xx} L(x^*,\lambda) d$$

- Gives

$$d^\top \nabla^2_{xx} L(x^*,\lambda) d = \begin{pmatrix} d_1 \\ -d_1 \end{pmatrix}^\top \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} d_1 \\ -d_1 \end{pmatrix} = \begin{pmatrix} d_1 \\ -d_1 \end{pmatrix}^\top \begin{pmatrix} d_1 \\ -d_1 \end{pmatrix} = 2d_1^2 > 0.$$

- Second order sufficient condition is satisfied!

Contents

Sequential Quadratic Programming (SQP) method

- Idea: Use Newton or Quasi-Newton Method for optimality system:

$$\nabla L(x, \lambda) = 0.$$

- Recall interpretation of Newton's method: Minimization of a **sequence** of approximations of f by **quadratic functions**

$$f(x_k + d) \approx f_k(d) = f(x_k) + \nabla f(x_k)^\top d + \frac{1}{2} d^\top \nabla^2 f(x_k) d = \frac{1}{2} d^\top A d + b^\top d + c.$$

- Necessary optimality condition:

$$\nabla f_k(d) = Ad + b = \nabla^2 f(x_k) d + \nabla f(x_k) = 0.$$

- This gives d as solution of

$$\nabla^2 f(x_k) d = -\nabla f(x_k).$$

- $\rightsquigarrow$ d is the Newton direction.

Same for Lagrangian

- We approximate L in the vicinity of the current iterate (x_k, λ_k) by the quadratic function

$$L(x_k + d, \lambda_k + \delta) \approx L(x_k, \lambda_k) + \begin{pmatrix} \nabla_x L(x_k, \lambda_k) \\ \nabla_\lambda L(x_k, \lambda_k) \end{pmatrix}^\top \begin{pmatrix} d \\ \delta \end{pmatrix} + \frac{1}{2} \begin{pmatrix} d \\ \delta \end{pmatrix}^\top \nabla^2_{(x,\lambda)} L(x_k, \lambda_k) \begin{pmatrix} d \\ \delta \end{pmatrix}$$

$$= \frac{1}{2} \begin{pmatrix} d \\ \delta \end{pmatrix}^\top A \begin{pmatrix} d \\ \delta \end{pmatrix} + b^\top \begin{pmatrix} d \\ \delta \end{pmatrix} + c.$$

- We minimize this approximation w.r.t. (d, δ). Necessary optimality condition:

$$A \begin{pmatrix} d \\ \delta \end{pmatrix} + b = \nabla^2_{(x,\lambda)} L(x_k, \lambda_k) \begin{pmatrix} d \\ \delta \end{pmatrix} + \nabla_{(x,\lambda)} L(x_k, \lambda_k) = 0$$

- This gives (d, δ) as solution of

$$\nabla^2_{(x,\lambda)} L(x_k, \lambda_k) \begin{pmatrix} d \\ \delta \end{pmatrix} = -\nabla_{(x,\lambda)} L(x_k, \lambda_k).$$

$\leadsto$ (d, δ) is the Newton direction.

Gradient and Hessian of Lagrangian

$$\nabla_{(x,\lambda)} L(x,\lambda) = \begin{pmatrix} \nabla_x L(x,\lambda) \\ \nabla_\lambda L(x,\lambda) \end{pmatrix} = \begin{pmatrix} \nabla f(x) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x) \\ h(x) \end{pmatrix}$$

$$\nabla^2_{(x,\lambda)} L(x,\lambda) = \begin{pmatrix} \nabla^2_{xx} L & \nabla^2_{\lambda x} L \\ \nabla^2_{x\lambda} L & \nabla^2_{\lambda\lambda} L \end{pmatrix} (x,\lambda) = \begin{pmatrix} \nabla^2 f(x) + \sum_{i=1}^{p} \lambda_i \nabla^2 h_i(x) & \nabla h_1(x) \cdots \nabla h_p(x) \\ \nabla h_1(x)^\top & \\ \vdots & \vdots \\ \nabla h_p(x)^\top & \end{pmatrix}$$

Newton step:

$$\begin{pmatrix} \nabla^2_{xx} L & \nabla^2_{\lambda x} L \\ \nabla^2_{x\lambda} L & 0 \end{pmatrix} (x_k,\lambda_k) \begin{pmatrix} d_k \\ \delta_k \end{pmatrix} = - \begin{pmatrix} \nabla_x L(x_k,\lambda_k) \\ h(x_k) \end{pmatrix}$$

Example: SQP method

- $f(x) = -x_1 x_2$, $h(x) = 2(x_1 + x_2) - c$.
- Optimality system:

$$\nabla_{(x,\lambda)} L(x,\lambda) = \begin{pmatrix} \nabla f(x) + \lambda \nabla h(x) \\ h(x) \end{pmatrix} = \begin{pmatrix} -x_2 + 2\lambda \\ -x_1 + 2\lambda \\ 2(x_1 + x_2) - c \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

- Quadratic approximation:

$$L(x_k + d, \lambda_k + \delta) \approx \frac{1}{2} \begin{pmatrix} d \\ \delta \end{pmatrix}^\top A \begin{pmatrix} d \\ \delta \end{pmatrix} + b^\top \begin{pmatrix} d \\ \delta \end{pmatrix} + c.$$

  with

$$A = \nabla^2_{(x,\lambda)} L(x,\lambda) = \begin{pmatrix} \nabla^2 f(x) + \lambda \nabla^2 h(x) & \nabla h(x) \\ \nabla h(x)^\top & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 & 2 \\ -1 & 0 & 2 \\ 2 & 2 & 0 \end{pmatrix},$$

$$b = \nabla_{(x,\lambda)} L(x,\lambda) \text{ as above.}$$

Example from above: SQP method

- Newton step:

$$A\begin{pmatrix} d_1 \\ d_2 \\ \delta \end{pmatrix} = \begin{pmatrix} 0 & -1 & 2 \\ -1 & 0 & 2 \\ 2 & 2 & 0 \end{pmatrix}\begin{pmatrix} d_1 \\ d_2 \\ \delta \end{pmatrix} = -b = -\begin{pmatrix} -x_2 + 2\lambda \\ -x_1 + 2\lambda \\ 2(x_1 + x_2) - c \end{pmatrix}.$$

- Compute first Newton step for initial guess $(x_0, \lambda_0) = (0, 0, 0)$, i.e., solve

$$\begin{pmatrix} 0 & -1 & 2 \\ -1 & 0 & 2 \\ 2 & 2 & 0 \end{pmatrix}\begin{pmatrix} d_1 \\ d_2 \\ \delta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ c \end{pmatrix}.$$

- Gives:

$$d_1 = d_2 = 2\delta, 2d_1 + 2d_2 = 8\delta = c \;\Rightarrow\; \delta = \frac{c}{8}, d_1 = d_2 = \frac{c}{4} \;\Rightarrow\; x_1 = \left(\frac{c}{4}, \frac{c}{4}\right), \lambda_1 = \frac{c}{8}.$$

- This is the solution (in one step, since optimality system is linear).

Globalized SQP-Newton method

Algorithm (Globalized SQP-Newton method):

1. Choose initial guess $(x_0, \lambda_0) \in \mathbb{R}^n \times \mathbb{R}^p$.
2. For $k = 0, 1, \dots$:
   1. Compute Newton direction, i.e., solve
   $$\nabla^2_{(x,\lambda)} L(x_k, \lambda_k) \begin{pmatrix} d_k \\ \delta_k \end{pmatrix} = -\nabla_{(x,\lambda)} L(x_k, \lambda_k).$$
   2. If direction is not gradient-related, take negative gradient step for (d_k, δ_k).
   3. Choose an efficient step-size $\rho_k > 0$.
   4. Set $\begin{pmatrix} x_{k+1} \\ \lambda_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ \lambda_k \end{pmatrix} + \rho_k \begin{pmatrix} d_k \\ \delta_k \end{pmatrix}$,

   until a stopping criterion is satisfied.

- Quasi-Newton idea can be applied as well.

$\rightsquigarrow$ Quasi-Newton-SQP methods are more popular than Newton-SQP.

What is important

- The Lagrange multiplier rule is the generalization of the first order necessary condition $\nabla f(x^*) = 0$ for constrained problems.
- It introduces additional variables, the Lagrange multipliers, here called λ.
- The condition $\nabla f(x^*) = 0$ is replaced by $\nabla_{(x,\lambda)} L(x^*, \lambda) = 0$ for the Lagrange function or Lagrangian L.
- This gives a system of equations we call optimality system.
- An additional condition, the regularity or a constraint qualification is needed for the existence of the multipliers and the validity of the rule.
- Based on the Lagrangian, a second order sufficient condition for constrained problems can also be formulated.
- The optimality system can be solved by Newton or Quasi-Newton methods.