Optimization

2. Constraint optimization and Lagrange multipliers

Andrew Lesniewski

Baruch College
New York

Fall 2018

Outline

1. Constraint optimization problems

2. Numerical methods

Formulation of the problem

- In most financial applications, the variables in an optimization problem are restricted to vary in a subset Ω of $\mathbb{R}^n$ rather than in the entire space $\mathbb{R}^n$.
- Examples:
  - (i) The correlation parameter in a stochastic volatility model has to lie within the range $[-1, 1]$.
  - (ii) The volatility parameters in a term structure model should be all positive, while the unconstrained optimizer stubbornly produces negative values.
  - (iii) Portfolio manager's mandate is to be long only, meaning that the weights of all assets in the portfolio are positive.
  - (iv) The share of each position in the portfolio cannot exceed 5% of its total value.
  - (v) No industry group should take more than 20% share of the total portfolio value.
- Oftentimes we are "lucky" and get the right results using an unconstrained method (like in data fitting). Otherwise, we have to resort to the specialized methods developed to handle constrained problems.

Formulation of the problem

- For tractability, we shall assume that the subset Ω is specified in terms of a number of functions $c_i(x)$, $i = 1, \ldots, m$, which define *equality constraints* or *inequality constraints*.
- Specifically, given an objective function $f(x)$, we formulate the following problem:

$$\min_{x \in \mathbb{R}^n} f(x), \quad \text{subject to} \begin{cases} c_i(x) = 0, & \text{if } i \in \mathcal{E}, \\ c_i(x) \leq 0, & \text{if } i \in \mathcal{I}. \end{cases} \tag{1}$$

  Here, $\mathcal{E}$ and $\mathcal{I}$ are disjoint subsets of the set of indices $1, \ldots, m$, such that $\mathcal{E} \cup \mathcal{I} = \{1, \ldots, m\}$.
- A point $x \in \mathbb{R}^n$ is called *feasible*, if it satisfies all the constraints. We can thus characterize the subset Ω as the set of all feasible points of the problem:

$$\Omega = \{x \in \mathbb{R}^n : \ c_i(x) = 0, \text{if } i \in \mathcal{E}; c_i(x) \leq 0, \text{if } i \in \mathcal{I}\}. \tag{2}$$

Formulation of the problem

- The constraint optimization problem can thus be formulated as follows:

$$\min_{x\in\Omega} f(x). \tag{3}$$

- A point x^* is a *local solution* to (3), if it is feasible, and $f(x)\geq f(x^*)$, for all x in a neighborhood of x^* (which is contained in Ω).
- It is a strict local solution if $f(x) > f(x^*)$, for all $x \neq x^*$ in a neighborhood of x^*.
- Solving constraint optimization problems is challenging and, before developing general methodology, we discuss a few examples.

Example

- *Example 1.* Consider the problem:

$$\min x_1 + x_2, \quad \text{subject to} \begin{cases} x_1 x_2 = a^2, \text{ where } a > 0, \\ x_1, x_2 \geq 0. \end{cases} \tag{4}$$

  In this case, $c_1(x) = x_1 x_2 - a^2, \mathcal{E} = \{1\}$, $c_2(x) = -x_1, c_3(x) = -x_2, \mathcal{I} = \{2, 3\}$, and the feasible set is the hyperbola $x_1 x_2 = a^2$ in the first quadrant.

- The special feature of this problem is that the constraints can be *solved*. Namely, $x_2 = a^2/x_1$, which reduces the problem to minimizing a function of one variable:

$$g(x_1) = x_1 + \frac{a^2}{x_1}.$$

- Setting the derivative of $g(x_1)$ to zero we find that $x_1^* = a$ (the solution $-a$ is rejected because it is not in the feasible set). That means that $x_2^* = a$, and inspection shows that $x^* = (a, a)$ is, in fact, a global minimum.

- Situations in which the constraints can be solved are extremely rare. In general, solving the constraints is either impossible or it leads to cumbersome calculations.

Example

- *Example 2.* Consider the problem:

$$\min x_1 + x_2, \quad \text{subject to } x_1^2 + x_2^2 = 2. \tag{5}$$

In this case, $c_1(x) = x_1^2 + x_2^2 - 2, \mathcal{E} = \{1\}, \mathcal{I} = \emptyset$, and the feasible set consists of the circle of radius $\sqrt{2}$.

- Inspection shows that the solution to this problem is

$$x^* = -\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

- We note that

$$\nabla f(x^*) = -\frac{1}{2} \nabla c_1(x^*), \tag{6}$$

i.e., at the solution, the gradient of the objective function is proportional to the gradient of the constraint function.

Example

- This is not a coincidence! We will see that this is a general fact: at a local minimizer, the gradient of the objective function is a linear combination of the gradients of the constraints.
- We rewrite (6) in the form:

$$\nabla f(x^*) + \frac{1}{2}\nabla c_1(x^*) = 0. \tag{7}$$

The proportionality coefficient $\lambda^* = 1/2$ (in this example) is called the *Lagrange multiplier*.

Example

- We can interpret this observation analytically as follows. If x is a feasible point that is *not* a solution to (5), then there is a small vector h such that $x + h$ is feasible and $f(x + h) < f(x)$:

$$\begin{aligned} 0 &= c_1(x + h) \approx \nabla c_1(x)^{\mathrm{T}} h, \\ 0 &> f(x + h) - f(x) \approx \nabla f(x)^{\mathrm{T}} h. \end{aligned}$$

- Therefore, for x to be a solution, there *cannot* exist a vector (direction) d such that the conditions below hold simultaneously:

$$\begin{aligned} \nabla c_1(x)^{\mathrm{T}} d &= 0, \\ \nabla f(x)^{\mathrm{T}} d &< 0. \end{aligned}$$

- This is possible if and only if $\nabla f(x)$ and $\nabla c_1(x)$ are parallel. In other words, there has to exist a scalar $-\lambda$ such that $\nabla f(x) = -\lambda \nabla c_1(x)$.

Equality constraints and Lagrange Multiplier Theorem

- Let us now consider the general constrained optimization problem with equality constraints only (i.e. $\mathcal{I} = \emptyset$).
- Reasoning along the lines of Example 2, we argue that a feasible point x is a solution to (3), provided that there is no vector $d \in \mathbb{R}^n$ with the properties:

$$\begin{aligned} \nabla c_i(x)^{\mathrm{T}} d &= 0, \text{ for } i = 1, \ldots, m, \\ \nabla f(x)^{\mathrm{T}} d &< 0. \end{aligned}$$

- The first of these conditions defines the subspace of *first order feasible variations*:

$$V(x) = \{d : \ \nabla c_i(x)^{\mathrm{T}} d = 0, \text{ for } i = 1, \ldots, m\}. \tag{8}$$

- The second of the conditions states that the gradient $\nabla f(x)$ of the objective function has to be perpendicular to $V(x)$ (since its inner product with any nonzero vector in $V(x)$ is nonzero). Consequently, it must be a linear combination of the constraint gradients $\nabla c_i(x)$, which span the subspace perpendicular to $V(x)$, i.e.

$$\nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla c_i(x) = 0,$$

with some scalars λ_i.

Equality constraints and Lagrange Multiplier Theorem

- The *Lagrange Multiplier Theorem* formulated below states necessary conditions for local minima of (3). It puts the informal reasoning above on a rigorous basis.
- The formulation of the theorem requires a technical assumption. Namely, we shall call a feasible point x *regular*, if the vectors $\nabla c_1(x), \ldots, \nabla c_m(x)$ are linearly independent.
- *Lagrange Multiplier Theorem. Let x^* be a regular local minimizer of $f(x)$ subject to $c_i(x) = 0$, for $i = 1, \ldots, m$. Then:*
  (i) There exists a unique vector $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*)$ of *Lagrange multipliers*, such that

  $$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla c_i(x^*) = 0 \tag{9}$$

  (ii) If, in addition, $f(x)$ and $c_i(x)$ are twice continuously differentiable, then

  $$d^{\mathrm{T}} \big(\nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 c_i(x^*)\big) d \geq 0, \tag{10}$$

  for all $d \in V(x^*)$.

Lagrangian function

- It is convenient (and customary) to write the necessary conditions in terms of the *Lagrangian function*. This is a function of $n + m$ variables defined as follows:

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i c_i(x). \tag{11}$$

- The necessary condition of the Lagrange Multiplier Theorem can be parsimoniously formulated as follows:

$$\begin{aligned} \nabla_x L(x^*, \lambda^*) &= 0, \\ \nabla_\lambda L(x^*, \lambda^*) &= 0, \\ d^{\mathrm{T}} \nabla^2_{xx} L(x^*, \lambda^*) d &\geq 0, \text{ for all } d \in V(x^*). \end{aligned} \tag{12}$$

- We emphasize that these conditions are necessary but not sufficient: a solution to the system above may not represent a local minimum.

Example

- *Example 3.* Consider the problem:

$$\min \frac{1}{2}(x_1^2 + x_2^2 + x_3^2), \qquad \text{subject to } x_1 + x_2 + x_3 = 3.$$

- The first order necessary conditions read:

$$\begin{aligned} x_i + \lambda &= 0, \text{ for } i = 1, 2, 3, \\ x_1 + x_2 + x_3 &= 3. \end{aligned}$$

  Solving this system yields

$$x^* = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \lambda^* = -1.$$

- Note also that the second order (positive definiteness) condition holds, as $\nabla^2_{xx} L(x^*, \lambda^*) = I_3$ (the 3 × 3 identity matrix). In fact, x^* is a global minimum.

Sufficient conditions

- The theorem below gives a sufficient condition for the existence of a local minimizer.
- *Second Order Sufficient Conditions.* Let $f(x)$ and $c_i(x)$ be twice continuously differentiable and let $x^* \in \Omega \subset \mathbb{R}^n$, $\lambda^* \in \mathbb{R}^m$ be such that

$$\begin{aligned} \nabla_x L(x^*, \lambda^*) &= 0, \\ \nabla_\lambda L(x^*, \lambda^*) &= 0, \end{aligned} \tag{13}$$

and

$$d^{\mathrm{T}} \nabla^2_{xx} L(x^*, \lambda^*) d > 0, \text{ for all } d \in V(x^*). \tag{14}$$

Then x^* is a strict local minimizer of $f(x)$ subject to equality constraints $c_i(x), i = 1, \ldots, m$.

Example

- *Example 4.* Consider the quadratic optimization problem:

$$\min f(x) = \frac{1}{2}\, x^{\mathrm{T}} A x + x^{\mathrm{T}} b, \quad \text{subject to } x^{\mathrm{T}} c = 1, \tag{15}$$

  where $b, c \in \mathbb{R}^n$ are constant vectors, and $A \in \mathrm{Mat}_n(\mathbb{R})$ is positive definite.
- First and second order conditions yield

$$\begin{aligned} \lambda^* &= -\frac{b^{\mathrm{T}} A^{-1} c + 1}{c^{\mathrm{T}} A^{-1} c}\,, \\ x^* &= -A^{-1}(b + \lambda^* c), \\ \nabla^2_{xx} L(x^*, \lambda^*) &= A. \end{aligned}$$

- Note that, since A is positive definite, condition (14) holds for all d (and in particular, for $d \in V(x^*) = \{d : \ c^{\mathrm{T}} d = 0\}$). Consequently, x^* is a strict (global) minimizer.

Sensitivity analysis

- Lagrange multipliers have often intuitive interpretation, depending on the specific problem at hand. In general, they can be interpreted as rates of change of the objective function as the constraint functions are varied.
- Let x^* and λ^* be a local minimizer and the corresponding Lagrange multiplier, respectively, of a constrained optimization problem for $f(x)$.
- Consider now the following family of constrained optimization problems, parameterized by a vector $u = (u_1, \ldots, u_m) \in \mathbb{R}^m$:

$$\min_{x \in \mathbb{R}^n} f(x), \quad \text{subject to } c_i(x) = u_i. \tag{16}$$

- Then, at least for small values of u, the solution $x(u)$ with the corresponding Lagrange multipliers $\lambda(u)$ form continuously differentiable functions, such that

$$\begin{aligned} x(0) &= x^*, \\ \lambda(0) &= \lambda^*. \end{aligned}$$

Sensitivity analysis

- Observe that we have the following equation:

$$\nabla f(x(u)) + \sum_{i=1}^{m} \lambda_i(u) \nabla c_i(x(u)) = 0, \tag{17}$$

  which is simply the first order condition for (16).

- We denote

$$p(u) = f(x(u)), \tag{18}$$

  the minimum of the objective function as a function of the parameters u. The function $p(u)$ is called the *primal function*.

- Then, for all u where $x(u)$ is defined, we have the following relation:

$$\nabla p(u) = -\lambda(u). \tag{19}$$

Sensitivity analysis

- The proof is a straightforward calculation. From the chain rule and (17), we get for each $j = 1, \ldots, m$,

$$\begin{aligned}\nabla_{u_j} p(u) &= \nabla_{u_j} f(x(u))\\ &= \sum_{k=1}^{n} \nabla_{x_k} f(x(u)) \nabla_{u_j} x_k(u)\\ &= -\sum_{i=1}^{m} \sum_{k=1}^{n} \lambda_i(u) \nabla_{x_k} c_i(x(u)) \nabla_{u_j} x_k(u).\end{aligned}$$

- However

$$\begin{aligned}\sum_{k=1}^{n} \nabla_{x_k} c_i(x(u)) \nabla_{u_i} x_k(u) &= \nabla_{u_j} c_i(x(u))\\ &= \nabla_{u_j} u_i\\ &= \delta_{ji}.\end{aligned}$$

Sensitivity analysis

- As a consequence,

$$\begin{aligned}\nabla_{u_j} p(u) &= -\sum_{i=1}^{m} \lambda_i(u)\delta_{ji}\\ &= -\lambda_j(u),\end{aligned}$$

which proves equation (19).

Inequality constraints

- We will now move on to the case of inequality constraints. We begin with a motivational example.
- *Example 5.* Consider the slightly modified Example 2:

$$\min x_1 + x_2, \qquad \text{subject to } x_1^2 + x_2^2 \leq 2. \tag{20}$$

  In this case, $c_1(x) = x_1^2 + x_2^2 - 2, \mathcal{E} = \emptyset, \mathcal{I} = \{1\}$, and the feasible set consists of the unit circle and its interior.
- Inspection shows that the solution to this problem continues to be

$$x^* = -\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

- Notice that condition (6) continues to hold. We will argue that, in case of an inequality constraint, the *sign* of the Lagrange multiplier is not a coincidence.

Inequality constraints

- Assume that a feasible point $x \in \mathbb{R}^2$ is *not* a local minimizer. Arguing along the lines of Example 2, this is possible, if there *cannot* exist a (small) vector d satisfying both conditions below hold simultaneously:

$$\begin{aligned} c_1(x) + \nabla c_1(x)^{\mathrm{T}} d &\leq 0, \\ \nabla f(x)^{\mathrm{T}} d &< 0. \end{aligned} \tag{21}$$

  We consider two cases.

- *Case 1.* x lies inside the circle, i.e. $c_1(x) < 0$ (the constraint is inactive). In this case, assuming that $\nabla f(x) \neq 0$, the vector $d = -\varepsilon \nabla f(x)$ satisfies both conditions, provided that $\varepsilon > 0$ is sufficiently small. This vector does not work if $\nabla f(x) = 0$.

Inequality constraints

- *Case 2.* x lies on the boundary of the circle, i.e. $c_1(x) = 0$ (the constraint is active).
- In this case, the conditions read

$$\begin{aligned} \nabla c_1(x)^{\mathrm{T}} d &\leq 0, \\ \nabla f(x)^{\mathrm{T}} d &< 0. \end{aligned} \tag{22}$$

- The first of these conditions defines a closed half-plane, while the second one defines an open half-plane. The intersection of these two half-planes should be empty!
- A reflection shows that this is possible only if there is a *positive* constant λ such that

$$\nabla f(x) = -\lambda \nabla c_1(x).$$

Inequality constraints

- We can formulate the results of the analysis of Cases 1 and 2 in the following elegant way using the Lagrange function.
- If x^* is a local minimizer of $f(x)$ (no feasible descent direction d is possible), then

  $$\nabla_x L(x^*, \lambda^*) = 0, \text{ for some } \lambda^* \geq 0,$$

  with the additional requirement that

  $$\lambda^* c_1(x^*) = 0.$$

- The latter condition is called the *complementary slackness condition*. It means that λ^* *can be strictly positive, only if the constraint* $c_1(x)$ *is active*.
- In Case 1, $c_1(x^*) < 0$, and so $\lambda^* = 0$, while in Case 2 $\lambda^* = 1/2$ is positive.

Inequality constraints and the KKT conditions

- Consider now the general problem (1), in which both equality and inequality constraints are present.
- For any feasible point $x \in \Omega$, we define the set of active inequality constraints by

$$\mathcal{A}(x) = \{i \in \mathcal{I} : \; c_i(x) = 0\}. \tag{23}$$

  A constraint $i \in \mathcal{I}$ for which $c_i(x) < 0$ is called *inactive* at x.
- If x^* is a local minimum solution to (1), then it is also a local minimum to the same problem with all constraints inactive at x^* ignored. Thus the inactive constraints do not matter.
- The active constraints can be treated as equality constraints. A local minimum solution to (1) solves thus the equality constraints problem:

$$\min_{x \in \mathbb{R}^n} f(x), \quad \text{subject to } c_i(x) = 0, \text{ if } i \in \mathcal{E} \cup \mathcal{A}(x). \tag{24}$$

Inequality constraints and the KKT conditions

- Consequently, assuming that x^* is regular, there exist Lagrange multipliers λ_i^*, $i \in \mathcal{E} \cup \mathcal{A}(x)$, so that

$$\nabla f(x^*) + \sum_{i \in \mathcal{E} \cup \mathcal{A}(x)} \lambda_i^* \nabla c_i(x^*) = 0.$$

- For convenience, we also introduce the Lagrange multiplier $\lambda_i^* = 0$ for each inactive constraint i. We can thus write the above condition compactly as

$$\begin{aligned} \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla c_i(x^*) &= 0, \\ \lambda_i^* c_i(x^*) &= 0. \end{aligned} \tag{25}$$

- An important fact about the Lagrange multipliers corresponding to the inequality constraints is that they are nonnegative.

Inequality constraints and the KKT conditions

- This can be seen as follows. If the *i*-th constraint is replaced by a weaker constraint $c_i(x) \leq u_i$, where $u_i > 0$, then the optimal objective function will not increase. The reason for this is that the feasible will become larger.
- From the sensitivity equation (19):

$$\begin{aligned}\lambda_i^* &= -\lim_{u_i \downarrow 0} \frac{\partial}{\partial u_i}\, p(u) \\ &\geq 0.\end{aligned}$$

- The arguments above are not exactly a proof, but are convincing enough to understand the following necessary condition.

Inequality constraints and the KKT conditions

- As in the case of equality constraints only, it is convenient to work with the Lagrangian function:
$$L(x,\lambda)=f(x)+\sum_{i=1}^{m}\lambda_i c_i(x). \tag{26}$$
- *Karush-Kuhn-Tucker Necessary Conditions.* Let x^* be a local minimum solution to (1), and assume that x^* is regular. Then there exists a unique vector of Lagrange multipliers λ_i^*, $i=1,\ldots m$, such that
$$\begin{aligned}\nabla_x L(x^*,\lambda^*)&=0,\\ \lambda_i^* c_i(x^*)&=0,\\ \lambda_i^*&\geq 0,\ \text{for } i\in\mathcal{I}.\end{aligned} \tag{27}$$
If, additionally, $f(x)$ and $c_i(x)$ are twice continuously differentiable, then
$$d^{\mathrm{T}}\nabla_{xx}^2 L(x^*,\lambda^*)d\geq 0, \tag{28}$$
for all d, such that $\nabla c_i(x^*)^{\mathrm{T}}d=0$, $i\in\mathcal{E}\cup\mathcal{A}(x)$.

Inequality constraints and the KKT conditions

- Note that the *complementary slackness condition*

  $$\lambda_i^* c_i(x^*) = 0 \tag{29}$$

  is just a compact way of stating the fact $\lambda_i^* = 0$, if $c(x^*)$ is not active, and it can be nonzero (and, in fact, nonnegative), if the constraint is active.

Example

- *Example 6.* Consider the problem:

$$\min \frac{1}{2}(x_1^2 + x_2^2 + x_3^2), \qquad \text{subject to } x_1 + x_2 + x_3 \leq -3.$$

- The first order necessary conditions read:

$$x_i + \lambda = 0, \text{ for } i = 1, 2, 3.$$

There are two possibilities:

(i) The constraint is inactive, $x_1^* + x_2^* + x_3^* < -3$, in which case $\lambda^* = 0$. This implies $x_i^* = 0$, for $i = 1, 2, 3$, which contradicts the constraint. Thus $(0, 0, 0)$ *is not* a minimizer.

(ii) The constraint is active, $x_1^* + x_2^* + x_3^* = -3$, in which case $\lambda^* = 1$, and

$$x^* = -\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Note also that the second order (positive definiteness) condition holds, as $\nabla_x^2 L(x^*, \lambda^*) = I_3$. A reflection shows that x^* is, in fact, a global minimizer.

Sufficient conditions

- Assume now that the functions $f(x)$ and $c_i(x), i = 1, \ldots, m$ are twice continuously differentiable, and let $x^* \in \mathbb{R}^n, \lambda^* \in \mathbb{R}^m$ be such that

$$\begin{aligned} \nabla_x L(x^*, \lambda^*) &= 0, \\ c_i(x^*) &= 0, \text{ if } i \in \mathcal{E}, \\ c_i(x^*) &\leq 0, \text{ if } i \in \mathcal{I}, \\ \lambda_i^* c_i(x^*) &= 0, \\ \lambda_i^* &\geq 0, \text{ for } i \in \mathcal{I}, \end{aligned} \tag{30}$$

  and

$$d^{\mathrm{T}} \nabla_{xx}^2 L(x^*, \lambda^*) d > 0, \tag{31}$$

  for all d, such that $\nabla c_i(x^*)^{\mathrm{T}} d = 0,\ i \in \mathcal{E} \cup \mathcal{A}(x)$.

- Assume also that $\lambda_i^* > 0$ for all active inequality constraints (*strict complementary slackness condition*). Then ξ^* is a strict local minimizer of $f(x)$ subject to the constraints.

Penalty function method

- A practical way of solving constrained optimization problems is by adding a *penalty function* per constraint to the objective function. These penalty functions attempt to mimic the effect of the constraint functions, i.e. discourage the optimizer to search outside of the feasible set.
- Typically, one considers a sequence of such modified objective functions, in which the penalty functions are multiplied by larger and larger positive coefficients, thus forcing the optimizer to stay closer and closer to the feasible set.
- Two commonly used penalty function methods are:
  - (i) *Exterior penalty functions* impose a penalty for violation of constraints.
  - (ii) *Interior penalty functions* impose a penalty for approaching the boundary of an inequality constraint.

Penalty function method

- The simplest exterior penalty method is the *quadratic penalty function*.
- The quadratic penalty function applied to constraint optimization problems with equality constraints only, $c_i(x) = 0$, for $i = 1, \ldots, m$, is defined as

$$Q(x, \mu) = f(x) + \frac{\mu}{2} \sum_{i=1}^{m} c_i(x)^2, \tag{32}$$

  where $\mu > 0$ is the *penalty parameter*.
- As $\mu \to \infty$, the presence of the penalty term imposes more and more severe penalty for constraints violations.

Penalty function method

- Algorithmically, we consider a sequence $\mu_0 < \ldots < \mu_k \to \infty$, and exit criteria

$$\|Q(x,\mu)\|_\infty \leq \varepsilon_k, \tag{33}$$

  where $\varepsilon_k > 0$, $k = 0, 1, \ldots,$ is a sequence with $\varepsilon_k \to 0$, as $k \to \infty$.
- We start the search with an initial guess x_0.
- For $k = 0, 1, \ldots,$ we find a minimizer x_k^* of $Q(x, \mu_k)$ satisfying the exit criterion (37).
- In the next step, we increase μ_k to μ_{k+1}, initialize the search with x_k^* ("warm start"), and find a minimizer x_{k+1}^* satisfying (37).
- We exit, once the final convergence test is satisfied.
- The expectation is that, as $k \to \infty$, x_k^* converges to the minimizer of the original problem, and thus, for μ_k sufficiently large, x_k^* will be a good approximation to x^*.

Example

- *Example 7.* Consider again the problem:

$$\min x_1 + x_2, \quad \text{subject to } x_1^2 + x_2^2 - 2 = 0.$$

- The quadratic penalty function is given by

$$Q(x, \mu) = x_1 + x_2 + \frac{\mu}{2}(x_1^2 + x_2^2 - 2)^2.$$

- We find that the gradient and Hessian of $Q(x, \mu)$ are

$$\nabla Q(x, \mu) = \begin{pmatrix} 1 + \mu x_1(x_1^2 + x_2^2 - 2) \\ 1 + \mu x_2(x_1^2 + x_2^2 - 2) \end{pmatrix},$$

  and

$$\nabla^2 Q(x, \mu) = \mu \begin{pmatrix} 3x_1^2 + x_2^2 - 2 & 2x_1 x_2 \\ 2x_1 x_2 & x_1^2 + 3x_2^2 - 2 \end{pmatrix},$$

  respectively.

Example

- We choose $\mu_0 = 1$ and perform Newton's search with the initial guess $x_0 = (0, 0.1)$.
- The table below shows the values of x_k^* corresponding to $\mu_k = k$.

μ_k	x_k^*
1	(-1.19148788, -1.19148788)
2	(-1.10715987, -1.10715987)
3	(-1.07474445, -1.07474445)
4	(-1.05745377, -1.05745377)
$\vdots$	$\vdots$
100	(-1.00249069, -1.00249069)

Penalty function method

- The sequence of μ_k can be chosen adaptively, based on the difficulty of minimizing $Q(x, \mu_k)$. When minimizing $Q(x, \mu_k)$ turns out to be difficult, in the next step, we choose μ_{k+1} to be modestly larger than μ_k. Otherwise, we can take μ_{k+1} significantly larger than μ_k.
- As $\mu_k \to \infty$, the search proves in general to get increasingly harder due to the fact that the Hessian of $Q(x, \mu_k)$ becomes close to ill conditioned. This phenomenon can be seen in Example 7, where the magnitude of the Hessian increases in proportion to μ.
- One can prove the following general theorem [2]: If x_k^* is the exact *global* minimizer of $Q(x, \mu_k)$, then every limit point x^* of x_k^* is a global solution to the original problem.

Penalty function method

- The quadratic penalty function can be defined for a problem with both equality and inequality constraints as follows:

$$Q(x,\mu)=f(x)+\frac{\mu}{2}\sum_{i\in\mathcal{E}}c_i(x)^2+\frac{\mu}{2}\sum_{i\in\mathcal{I}}\big(c_i(x)^+\big)^2, \tag{34}$$

  where $a^+=\max(a,0)$.

Augmented Lagrangian method

- A refinement of the quadratic penalty method is the *augmented Lagrange method*.
- It reduces the possibility of ill conditioning discussed above by introducing approximate Lagrange multipliers into the objective function.
- We discuss the case equality constraints only, see [2] and [1] for the general case.
- We consider the augmented Lagrange function

  $$L_a(x,\lambda,\mu)=f(x)+\sum_{m=1}^{m}\lambda_i c_i(x)+\frac{\mu}{2}\sum_{i=1}^{m}c_i(x)^2. \tag{35}$$

  It differs form the standard Lagrange function (11) by the presence of the quadratic penalty term.
- Note that the λ_i's are not, strictly speaking, the Lagrange multipliers, as the critical points of $L_a(x,\lambda,\mu)$ are not feasible points for the original problem.

Augmented Lagrangian method

- We do expect that as $\mu_k \to \infty$, the augmented Lagrange function will tend to the standard Lagrange function $L_a(x_k^*, \lambda, \mu_k) \to L(x^*, \lambda^*)$.
- Heuristically, the first order condition reads:

$$\begin{aligned} 0 &\approx \nabla_x L_a(x, \lambda, \mu) \\ &= \nabla f(x) + \sum_{m=1}^{m} (\lambda_i + \mu c_i(x)) \nabla c_i(x), \end{aligned}$$

from which we deduce that

$$\lambda_i^* \approx \lambda_i + \mu c_i(x),$$

or

$$c_i(x) \approx \frac{\lambda_i^* - \lambda_i}{\mu}.$$

Augmented Lagrangian method

- In other words, if λ_i is close to λ_i^*, the infeasibility of x will be much smaller than $1/\mu$.
- This suggest that the value of λ_i at each iteration step k(denoted by $\lambda_{i,k}$) is updated according to the following rule:

$$\lambda_{i,k+1} = \lambda_{i,k} + \mu_k c_i(x_k^*), \tag{36}$$

for all $i = 1, \ldots, m$.

Augmented Lagrangian method

- Algorithmically, we consider a sequence $\mu_0 < \ldots < \mu_k \to \infty$, and exit criteria

$$\|L_a(x, \lambda, \mu)\|_\infty \leq \varepsilon_k, \tag{37}$$

  where $\varepsilon_k > 0$, $k = 0, 1, \ldots$, is a sequence with $\varepsilon_k \to 0$, as $k \to \infty$.
- We start the search with an initial guess x_0 and λ_0.
- For $k = 0, 1, \ldots$, we find a minimizer x_k^* of $L_a(x, \lambda_k, \mu_k)$ satisfying the exit criterion (37).
- In the next step, we increase μ_k to μ_{k+1}, update λ_k to λ_{k+1} according to the rule (36), initialize the search with x_k^*, and find a minimizer x_{k+1}^* satisfying (37).
- We exit, once the final convergence test is satisfied.
- The expectation is that, as $k \to \infty$, x_k^* converges to the minimizer of the original problem *without* the necessity of sending $\mu_k \to \infty$. This should mitigate the danger of ill conditioning and speed up the convergence.

Example

- Consider again Example 7. The augmented Lagrange function reads

$$L(x,\lambda\mu)=x_1+x_2+\lambda(x_1^2+x_2^2-2)+\frac{\mu}{2}\,(x_1^2+x_2^2-2)^2.$$

- We find that the gradient and Hessian of $L(x,\lambda\mu)$ are

$$\nabla L(x,\lambda\mu)=\begin{pmatrix}1+2\lambda x_1+\mu x_1(x_1^2+x_2^2-2)\\ \\ 1+2\lambda x_2+\mu x_2(x_1^2+x_2^2-2)\end{pmatrix},$$

and

$$\nabla^2 L(x,\lambda\mu)=\begin{pmatrix}2\lambda+\mu(3x_1^2+x_2^2-2) & 2\mu x_1x_2\\ \\ 2\mu x_1x_2 & 2\lambda+\mu(x_1^2+3x_2^2-2)\end{pmatrix},$$

respectively.

Example

- We choose $\mu_0 = 1$ and perform Newton's search with the initial guess $x_0 = (0, 0.1)$ and $\lambda_0 = 0.3$.
- The table below shows the values of x_k^* and λ_k corresponding to $\mu_k = k$.

μ_k	λ_k	x_k^*
1	0.62706599	(-1.07867187, -1.07867187)
2	0.40227548	(-0.97149492, -0.97149492)
3	0.58299634	(-1.01494835, -1.01494835)
4	0.28520661	(-0.9901962, -0.9901962)
5	0.69491662	(-1.02027986, -1.02027986)
$\vdots$	$\vdots$	$\vdots$
10	0.31472584	(-0.99044963, -0.99044963)

Barrier methods

- Exterior penalty methods allow for generating infeasible points during the search. Therefore they not suitable when feasibility has to be strictly enforced.
- This could be the case if the objective function is undefined outside of the feasible set.
- Barrier methods are similar in spirit to the external penalty method. They generate a sequence of unconstrained modified differentiable objective functions whose unconstrained minimizers are expected to converge to the solution of the constrained problem in the limit.

Barrier methods

- *Barrier methods* belong to the category of interior penalty function methods and apply to inequality constraint optimization problems, i.e. $\mathcal{E} = \emptyset$ in (1).
- The essence of the method is to add a penalty term $B(x)$ to the objective function, which has the following properties:
  (i) It is defined and continuous whenever $c_i(x) < 0$, for all $i = 1, \ldots, m$.
  (ii) It goes to $+\infty$, whenever $c_j(x) \uparrow 0$, for any i.
- Commonly used barrier functions are:

$$B(x) = -\sum_{i=1}^{m} \log\big(-c_i(x)\big) \quad \text{(logarithmic)}, \tag{38}$$

and

$$B(x) = -\sum_{i=1}^{m} \frac{1}{c_i(x)} \quad \text{(inverse)}. \tag{39}$$

Barrier methods

- Algorithmically, barrier methods work similarly to the penalty function methods.
- We select a sequence of positive numbers $\mu_0 > \ldots > \mu_k \to 0$, and consider a sequence of functions:

$$R(x, \mu_k) = f(x) + \mu_k B(x). \tag{40}$$

- We find a local minimizer x_k^*, and use it as the initial guess for the next iteration.
- A general convergence theorem guarantees that any limit point of this sequence is a global minimizer of the original constrained problem.
- Since the barrier function is defined only in the interior of the feasible set, any successive iteration must also be an interior point.
- The barrier term $\mu_k B(x)$ goes to zero for interior points as $\mu_k \to 0$. The barrier term becomes thus increasingly irrelevant for interior points, while allowing x_k^* to move closer to the boundary of Ω. This behavior is expected if the solution to the original constraint problem lies on the boundary.

Example

- *Example 8.* Consider the problem:

$$\min \frac{1}{2}\,(x_1^2 + x_2^2), \qquad \text{subject to } x_1 \geq 2.$$

- Its optimal solution is clearly $x^* = (2, 0)$.
- The function

$$R(x, \mu_k) = \frac{1}{2}\,(x_1^2 + x_2^2) - \mu_k \log(x_1 - 2)$$

  has a global minimum at

$$x_k^* = (1 + \sqrt{1 + \mu_k}, 0).$$

- As μ_k decreases to zero, the unconstrained minimum x_k^* converges to the constrained minimum x^*.

References

- Bertsekas, D. P.: *Nonlinear Programming*, Athena Scientific (2016).
- Nocedal, J., and Wright, S. J.: *Numerical Optimization*, Springer (2006).