APPENDIX

HINTS ON CONSTRAINED OPTIMIZATION

C

C.1 EQUALITY CONSTRAINTS

We will first focus on linear equality constraints and then generalize to the nonlinear case. The problem is cast as

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & J(\boldsymbol{\theta}), \\ \text{s.t.} \quad & A\boldsymbol{\theta} = \boldsymbol{b}, \end{aligned}$$

where A is an $m \times l$ matrix and $\boldsymbol{b}, \boldsymbol{\theta}$ are $m \times 1$ and $l \times 1$ vectors, respectively. It is assumed that the cost function $J(\boldsymbol{\theta})$ is twice continuously differentiable and it is, in general, a nonlinear function. Furthermore, we assume that the rows of A are linearly independent; hence A has full row rank. This assumption is known as the *regularity assumption*.

Let $\boldsymbol{\theta}_*$ be a local minimizer of $J(\boldsymbol{\theta})$ over the set $\{\boldsymbol{\theta}\colon A\boldsymbol{\theta} = \boldsymbol{b}\}$. It can be shown (e.g., [5]) that, at this point, there exists a $\boldsymbol{\lambda}$ such as the gradient of $J(\boldsymbol{\theta})$ is written as

$$\frac{\partial}{\partial \boldsymbol{\theta}} \big(J(\boldsymbol{\theta}) \big) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*} = A^T \boldsymbol{\lambda}, \tag{C.1}$$

where $\boldsymbol{\lambda} := [\lambda_1, \ldots, \lambda_m]^T$. Taking into account that

$$\frac{\partial}{\partial \boldsymbol{\theta}} (A\boldsymbol{\theta}) = A^T, \tag{C.2}$$

Eq. (C.1) states that, at a constrained minimum, the gradient of the cost function is a linear combination of the gradients of the constraints. We can get a better feeling of this result by mobilizing a simple example and exploiting geometry. Let us consider a single constraint,

$$\boldsymbol{a}^T \boldsymbol{\theta} = b.$$

Eq. (C.1) then becomes

$$\frac{\partial}{\partial \boldsymbol{\theta}} (J(\boldsymbol{\theta}_*)) = \lambda \boldsymbol{a},$$

where the parameter λ is now a scalar. Fig. C.1 shows an example of isovalue contours of $J(\boldsymbol{\theta}) = c$ in the two-dimensional space ($l = 2$). The constrained minimum coincides with the point where the straight line "meets" the isovalue contours for the first time, as one moves from small to large values of c. This is the point where the line is tangent to an isovalue contour; hence, at this point, the gradient of the cost function is in the direction of $\boldsymbol{a}$.

Let us now define the function

$$L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = J(\boldsymbol{\theta}) - \boldsymbol{\lambda}^T (A\boldsymbol{\theta} - \boldsymbol{b}) \tag{C.3}$$

$$= J(\boldsymbol{\theta}) - \sum_{i=1}^{m} \lambda_i (\boldsymbol{a}_i^T \boldsymbol{\theta} - b_i), \tag{C.4}$$

where $\boldsymbol{a}_i^T,\ i = 1, 2, \ldots, m$, are the rows of A; $L(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is known as the *Lagrangian function* and the coefficients $\lambda_i,\ i = 1, 2, \ldots, m$, as the *Lagrange multipliers*. The optimality condition (C.1), together with the constraints which the minimizer has to satisfy, can now be written in a compact form as

$$\nabla L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \boldsymbol{0}, \tag{C.5}$$

where ∇ denotes the gradient operation with respect to both $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$. Indeed, equating to zero the derivatives of the Lagrangian with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ gives, respectively,

$$\frac{\partial}{\partial \boldsymbol{\theta}} J(\boldsymbol{\theta}) = A^T \boldsymbol{\lambda},$$

$$A\boldsymbol{\theta} = \boldsymbol{b}.$$

The above is a set of $m + l$ unknowns, i.e., $(\theta_1, \ldots, \theta_l, \lambda_1, \ldots, \lambda_m)$, with $m + l$ equations, whose solution provides the minimizer $\boldsymbol{\theta}_*$ and the corresponding Lagrange multipliers.

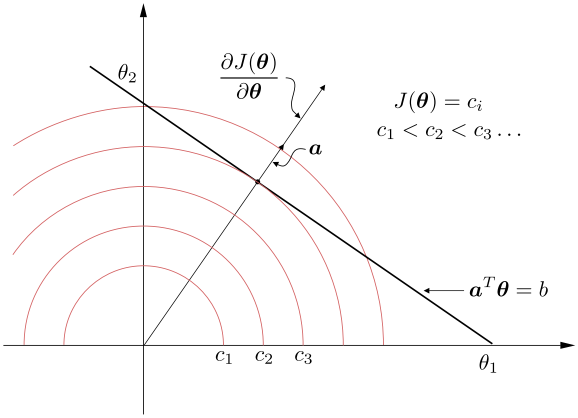
θ_2
$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$
$J(\boldsymbol{\theta}) = c_i$
$c_1 < c_2 < c_3 \ldots$
$\boldsymbol{a}$
$\boldsymbol{a}^T \boldsymbol{\theta} = b$
c_1
c_2
c_3
θ_1

FIGURE C.1

At the minimizer, the gradient of the cost function is in the direction of the gradient of the constraint function.

Similar arguments hold for nonlinear equation constraints. Let us consider the problem

$$\begin{aligned} \text{minimize} \quad & J(\boldsymbol{\theta}), \\ \text{subject to} \quad & f_i(\boldsymbol{\theta}) = 0, \quad i = 1, 2, \dots, m. \end{aligned}$$

The minimizer is again a *stationary point* of the corresponding Lagrangian

$$L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = J(\boldsymbol{\theta}) - \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{\theta})$$

and results from the solution of the set of $m + l$ equations

$$\nabla L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbf{0}.$$

The regularity condition for nonlinear constraints requires the gradients of the constraints $\frac{\partial}{\partial \boldsymbol{\theta}}(f_i(\boldsymbol{\theta}))$ to be linearly independent.

C.2 INEQUALITY CONSTRAINTS

The general problem can be cast as follows:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & J(\boldsymbol{\theta}), \\ \text{s.t.} \quad & f_i(\boldsymbol{\theta}) \geq 0, \quad i = 1, 2, \dots, m. \end{aligned} \tag{C.6}$$

Each one of the constraints defines a region in $\mathbb{R}^l$. The intersection of all these regions defines the area in which the constrained minimum, $\boldsymbol{\theta}_*$, must lie. This is known as the *feasible region* and the points in it (candidate solutions) as *feasible points*. The type of the constraints control the type of the feasible region, i.e., whether it is convex or not.

Assuming that each one of the functions in the constraints is concave, we can write each one of the constraints in Eq. (C.6) as $-f_i(\boldsymbol{\theta}) \leq 0$. Now, each of the constraints becomes a convex function and the inequalities define the respective zero level sets (Chapter 8); however, these are convex sets, and hence the feasible region is a convex one. For more on these issues, the interested reader may refer, for example, to [1].

Note that this is also valid for linear inequality constraints, because a linear function can be considered either convex or concave.

THE KARUSH–KUHN–TUCKER (KKT) CONDITIONS

This is a set of *necessary* conditions, which a local minimizer $\boldsymbol{\theta}_*$ of the problem given in Eq. (C.6) has to satisfy. If $\boldsymbol{\theta}_*$ is a point that satisfies the regularity condition, then there exists a vector $\boldsymbol{\lambda}$ of Lagrange multipliers so that the following are valid:

$$\begin{aligned}
&(1) \quad \frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\lambda})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*} = \mathbf{0}, \\
&(2) \quad \lambda_i \geq 0, \quad i = 1, 2, \ldots, m, \\
&(3) \quad \lambda_i f_i(\boldsymbol{\theta}_*) = 0, \quad i = 1, 2, \ldots, m.
\end{aligned} \tag{C.7}$$

Actually, there is a fourth condition concerning the Hessian of the Lagrangian function, which is not of interest to us. The above set of equations is also part of the sufficiency conditions; however, for the sufficient conditions, there are a few subtle points and the interested reader is referred to more specialized textbooks (e.g., [1,3,2,4,5]).

Conditions (3) in (C.7) are known as *complementary slackness conditions*. They state that at least one of the two factors in the products is zero. In the case where, in each one of the equations, only one of the two factors is zero, i.e., either λ_i or $f_i(\boldsymbol{\theta}_*)$, we talk about *strict complementarity*.

Having now discussed all these nice properties, the major question arises: how can one compute a constrained (local) minimum? Unfortunately, this is not always an easy task. A straightforward approach would be to assume that some of the constraints are active (equality to zero holds) and some inactive, and check if the resulting Lagrange multipliers of the active constraints are nonnegative. If not, then choose another combination of constraints and repeat the procedure until one ends up with nonnegative multipliers. However, in practice, this may require a prohibitive amount of computation. Instead, a number of alternative approaches have been proposed. To this end, we will review some basics from game theory and use these to reformulate the KKT conditions. This new setup can be useful in a number of cases in practice.

MINMAX DUALITY

Let us consider two players, namely, X and Y, playing a game. Player X will choose a strategy, say, x, and simultaneously player Y will choose a strategy y. As a result, X will pay to Y the amount $\mathcal{F}(x, y)$, which can also be negative, i.e., X wins. Let us now follow their thinking, prior to their final choice of strategy, assuming that the players are good professionals.

X: If Y knew that I was going to choose x, then, because he/she is a clever player, he/she would choose y to make her/his profit maximum, i.e.,

$$\mathcal{F}^*(x) = \max_y \mathcal{F}(x, y).$$

Thus, in order to make my *worst-case payoff* to Y minimum, I have to choose x so as to minimize $\mathcal{F}^*(x)$, i.e.,

$$\min_x \mathcal{F}^*(x).$$

This problem is known as the *minmax* problem because it seeks the value

$$\min_x \max_y \mathcal{F}(x, y).$$

Y: X is a good player, so if he/she knew that I was going to play y, he/she would choose x to make her/his payoff minimum, i.e.,

$$\mathcal{F}_*(y) = \min_x \mathcal{F}(x, y).$$

Thus, in order to make my *worst-case profit* maximum I must choose y that maximizes $\mathcal{F}_*(y)$, i.e.,

$$\max_y \mathcal{F}_*(y).$$

This is known as the *maxmin* problem, as it seeks the value

$$\max_y \min_x \mathcal{F}(x, y).$$

The two problems are said to be *dual to each other*. The first is known to be the *primal*, whose objective is to minimize $\mathcal{F}^*(x)$, and the second is the *dual* problem, with the objective to maximize $\mathcal{F}_*(y)$.

For any x and y, the following is valid:

$$\mathcal{F}_*(y) := \min_x \mathcal{F}(x, y) \leq \mathcal{F}(x, y) \leq \max_y \mathcal{F}(x, y) := \mathcal{F}^*(x), \tag{C.8}$$

which easily leads to

$$\max_y \min_x \mathcal{F}(x, y) \leq \min_x \max_y \mathcal{F}(x, y). \tag{C.9}$$

SADDLE POINT CONDITION

Let $\mathcal{F}(\boldsymbol{x}, \boldsymbol{y})$ be a function of two vector variables with $\boldsymbol{x} \in X \subseteq \mathbb{R}^l$ and $\boldsymbol{y} \in Y \subseteq \mathbb{R}^l$. If a pair of points $(\boldsymbol{x}_*, \boldsymbol{y}_*)$, with $\boldsymbol{x}_* \in X$, $\boldsymbol{y}_* \in Y$, satisfies the condition

$$\mathcal{F}(\boldsymbol{x}_*, \boldsymbol{y}) \leq \mathcal{F}(\boldsymbol{x}_*, \boldsymbol{y}_*) \leq \mathcal{F}(\boldsymbol{x}, \boldsymbol{y}_*), \tag{C.10}$$

for every $\boldsymbol{x} \in X$ and $\boldsymbol{y} \in Y$, we say that it satisfies the *saddle point condition*. It is not difficult to show (e.g., [5]) that a pair $(\boldsymbol{x}_*, \boldsymbol{y}_*)$ satisfies the saddle point conditions *if and only if*

$$\max_{\boldsymbol{y}} \min_{\boldsymbol{x}} \mathcal{F}(\boldsymbol{x}, \boldsymbol{y}) = \min_{\boldsymbol{x}} \max_{\boldsymbol{y}} \mathcal{F}(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{F}(\boldsymbol{x}_*, \boldsymbol{y}_*). \tag{C.11}$$

LAGRANGIAN DUALITY

We will now use all the above to formulate our original cost function minimization problem as a minmax task of the corresponding Lagrangian function. Under certain conditions, this formulation can lead to computational savings when computing the constrained minimum. The optimization task of interest is

$$\begin{aligned} &\text{minimize} \quad && J(\boldsymbol{\theta}), \\ &\text{subject to} \quad && f_i(\boldsymbol{\theta}) \geq 0, \quad i = 1, 2, \ldots, m. \end{aligned}$$

The Lagrangian function is

$$L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = J(\boldsymbol{\theta}) - \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{\theta}). \tag{C.12}$$

Let

$$L^*(\boldsymbol{\theta}) = \max_{\boldsymbol{\lambda}} L(\boldsymbol{\theta}, \boldsymbol{\lambda}). \tag{C.13}$$

However, because $\boldsymbol{\lambda} \geq \mathbf{0}$ and $f_i(\boldsymbol{\theta}) \geq 0$, the maximum value of the Lagrangian occurs if the summation in Eq. (C.12) is zero ($\lambda_i = 0,\ f_i(\boldsymbol{\theta}) = 0$, or both) and

$$L^*(\boldsymbol{\theta}) = J(\boldsymbol{\theta}). \tag{C.14}$$

Therefore, our original problem is equivalent to

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \max_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\boldsymbol{\theta}, \boldsymbol{\lambda}). \tag{C.15}$$

As we already know, the dual problem of the above is

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}} \min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\lambda}). \tag{C.16}$$

CONVEX PROGRAMMING

A large class of practical problems obeys the following two conditions:

$$(1) \quad J(\boldsymbol{\theta}) \ \text{ is convex,} \tag{C.17}$$

$$(2) \quad f_i(\boldsymbol{\theta}),\ i = 1, 2, \ldots, m, \quad \text{are concave.} \tag{C.18}$$

This class of problems turns out to have a very useful and mathematically tractable property.

Theorem C.1. *Let $\boldsymbol{\theta}_*$ be a minimizer of such a problem, which is also assumed to satisfy the regularity condition. Let $\boldsymbol{\lambda}_*$ be the corresponding vector of Lagrange multipliers. Then $(\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*)$ is a saddle point of the Lagrangian function, and, as we know, this is equivalent to*

$$L(\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*) = \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \min_{\boldsymbol{\theta}} \max_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\boldsymbol{\theta}, \boldsymbol{\lambda}). \tag{C.19}$$

Proof. Because $f_i(\boldsymbol{\theta})$ are concave, $-f_i(\boldsymbol{\theta})$ are convex, so the Lagrangian function

$$L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = J(\boldsymbol{\theta}) - \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{\theta}),$$

for $\lambda_i \geq 0$, is also convex. Note, now, that for concave function constraints of the form $f_i(\boldsymbol{\theta}) \geq 0$, the feasible region is convex (see comments made before). The function $J(\boldsymbol{\theta})$ is also convex. Hence, every local minimum is also a global one; thus for any $\boldsymbol{\theta}$

$$L(\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*) \leq L(\boldsymbol{\theta}, \boldsymbol{\lambda}_*). \tag{C.20}$$

Furthermore, the complementary slackness conditions suggest that

$$L(\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*) = J(\boldsymbol{\theta}_*), \tag{C.21}$$

and for any $\boldsymbol{\lambda} \geq \mathbf{0}$

$$L(\boldsymbol{\theta}_*, \boldsymbol{\lambda}) := J(\boldsymbol{\theta}_*) - \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{\theta}_*) \leq J(\boldsymbol{\theta}_*) = L(\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*). \tag{C.22}$$

Combining Eqs. (C.20) and (C.22) we obtain

$$L(\boldsymbol{\theta}_*, \boldsymbol{\lambda}) \leq L(\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*) \leq L(\boldsymbol{\theta}, \boldsymbol{\lambda}_*). \tag{C.23}$$

In other words, *the solution* $(\boldsymbol{\theta}_*, \boldsymbol{\lambda}_*)$ *is a saddle point.* □

This is a very important theorem and it states that the constrained minimum of a convex programming problem can also be obtained as a maximization task applied on the Lagrangian. This leads us to the following very useful formulation of the optimization task.

WOLFE DUAL REPRESENTATION

A convex programming problem is equivalent to

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}} \quad L(\boldsymbol{\theta}, \boldsymbol{\lambda}), \tag{C.24}$$

$$\text{s.t.} \quad \frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbf{0}. \tag{C.25}$$

The last equation guarantees that $\boldsymbol{\theta}$ is a minimum of the Lagrangian.

Example C.1. Consider the quadratic problem

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}, \\ \text{subject to} \quad & A\boldsymbol{\theta} \geq \boldsymbol{b}. \end{aligned}$$

This is a convex programming problem; hence, the Wolfe dual representation is valid, i.e.,

$$\begin{aligned} \text{maximize} \quad & \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} - \boldsymbol{\lambda}^T(A\boldsymbol{\theta} - \boldsymbol{b}), \\ \text{subject to} \quad & \boldsymbol{\theta} - A^T\boldsymbol{\lambda} = \mathbf{0}. \end{aligned}$$

For this example, the equality constraint has an analytic solution (this is not, however, always possible). Solving with respect to $\boldsymbol{\theta}$, we can eliminate it from the maximizing function and the resulting dual problem involves only the Lagrange multipliers,

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & -\frac{1}{2}\boldsymbol{\lambda}^T A A^T \boldsymbol{\lambda} + \boldsymbol{\lambda}^T \boldsymbol{b}, \\ \text{s.t.} \quad & \boldsymbol{\lambda} \geq \mathbf{0}. \end{aligned}$$

This is also a quadratic problem but the set of constraints is now simpler.

REFERENCES

[1] M.S. Bazaraa, C.M. Shetty, Nonlinear Programming: Theory and Algorithms, John Wiley, New York, 1979.
[2] D.P. Bertsekas, M.A. Belmont, Nonlinear Programming, Athenas Scientific, Belmont, MA, 1995.
[3] R. Fletcher, Practical Methods of Optimization, second ed., John Wiley, New York, 1987.
[4] D.G. Luenberger, Linear and Nonlinear Programming, Addison Wesley, Reading, MA, 1984.
[5] S.G. Nash, A. Sofer, Linear and Nonlinear Programming, McGraw-Hill, New York, 1996.