BASIC ISSUES IN LAGRANGIAN OPTIMIZATION

R. Tyrrell Rockafellar

Dept. of Applied Mathematics
University of Washington FS–20
Seattle, WA 98195, USA

Abstract. These lecture notes review the basic properties of Lagrange multipliers and constraints in problems of optimization from the perspective of how they influence the setting up of a mathematical model and the solution technique that may be chosen. Conventional problem formulations with equality and inequality constraints are discussed first, and Lagrangian optimality conditions are presented in a general form which accommodates range constraints on the variables without the need for introducing constraint functions for such constraints. Particular attention is paid to the distinction between convex and nonconvex problems and how convexity can be recognized and taken advantage of.

Extended problem statements are then developed in which penalty expressions can be utilized as an alternative to black-and-white constraints. Lagrangian characterizations of optimality for such problems closely resemble the ones for conventional problems and in the presence of convexity take a saddle point form which offers additional computational potential. Extended linear-quadratic programming is explained as a special case.

1. FORMULATION OF OPTIMIZATION PROBLEMS

Everywhere in applied mathematics the question of how to choose an appropriate mathematical model has to be answered by art as much as by science. The model must be rich enough to provide useful qualitative insights as well as numerical answers that don't mislead. But it can't be too complicated or it will become intractable for analysis or demand data inputs that can't be supplied. In short, the model has to reflect the right balance between the practical issues to be addressed and the mathematical approaches that might be followed.

This means, of course, that to do a good job of formulating a problem a modeler needs to be aware of the pros and cons of various problem statements that might serve as templates, such as standard linear programming, quadratic programming, and the like. Knowledge of which features are advantageous, versus which are potentially troublesome, is essential. In optimization the difficulties can be all the greater because the key ideas are often different from the ones central to rest of applied mathematics. For instance, in many subjects the crucial division is between linear and nonlinear models, but in optimization it is between convex and nonconvex. Yet convexity is not a topic much treated in a general mathematical education.

Problems of optimization always focus on the maximization or minimization of some function over some set, but the way the function and set are specified can have a great impact. One distinction is whether the decision variables involved are "discrete" or "continuous." Discrete variables with integer values, in particular logical variables which can only have the values 0 or 1, are appropriate in circumstances where a decision has to be made whether to build a new facility, or to start up a process with fixed initial costs. But the introduction of such variables in a model is a very serious step; the problem may become much harder to solve or even to analyze. Here we'll concentrate on continuous variables.

The conventional way to think about an optimization problem in finitely many continuous variables is that a function $f_0(x)$ is to be minimized over all the points $x = (x_1, \ldots, x_n)$ in some subset C of the finite-dimensional real vector space $\mathbb{R}^n$. (Maximization is equivalent to minimization through multiplication by -1.) The set C is considered to be specified by a number of side conditions on x which are called constraints, the most common form being equality constraints $f_i(x) = 0$ and inequality constraints $f_i(x) \leq 0$. As a catch-all for anything else, there may be an "abstract constraint" $x \in X$ for some subset $X \subset \mathbb{R}^n$. For instance, X can be thought of as indicating nonnegativity conditions, or upper and lower bounds, on some of the variables x_j appearing as components of x. Such conditions

could be translated one by one into the form $f_i(x) \leq 0$ for additional functions f_i, but this may not be convenient.

The conventional statement of a general problem of optimization from this point of view is

$$(\mathcal{P}) \qquad \begin{array}{l} \text{minimize } f_0(x) \text{ over all } x \in X \\ \text{such that } f_i(x) \begin{cases} \leq 0 & \text{for } i = 1, \ldots, s, \\ = 0 & \text{for } i = s+1, \ldots, m. \end{cases} \end{array}$$

The points x satisfying the constraints in $(\mathcal{P})$ are called the *feasible solutions* (i.e., candidates for solutions) to the problem. They form a certain set $C \subset \mathbb{R}^n$, and it is over this that the function f_0 is to be minimized. A point $\bar{x} \in C$ is a (globally) *optimal solution* to $(\mathcal{P})$ if $f_0(\bar{x}) \leq f_0(x)$ for all $x \in C$. It is a *locally optimal solution* if there is a neighborhood V of $\bar{x}$ such that $f_0(\bar{x}) \leq f_0(x)$ for all $x \in C \cap V$. The *optimal value* in $(\mathcal{P})$ is the minimum value of the objective function f_0 over C, as distinguished from the point or points where it's attained, if any.

In dealing with a problem in the format of $(\mathcal{P})$, people usually take for granted that the functions $f_0, f_1, \ldots, f_m$ are second-order smooth (i.e., have continuous second partial derivatives). We'll do that too, but there are important modeling issues here that shouldn't be swept under the rug. We'll return to them in Section 3 in discussing how penalty expressions may in some situations be a preferable substitute for "exact" equality or inequality constraints of the sort in $(\mathcal{P})$.

Concerning the set X, we'll assume here for simplicity that it's *polyhedral*, or in other words, definable in terms of a finite system of linear constraints, these being conditions that *could*, if we so wished, be written in the form $f_i(x) \leq 0$ or $f_i(x) = 0$ for additional functions f_i that are *affine* (linear-plus-constant). The main example we have in mind is the case where X is a *box*, $X = X_1 \times \cdots \times X_n$ with X_j a closed (nonempty but not necessarily bounded) interval in $\mathbb{R}$. Then, of course, the condition $x \in X$ reduces to $x_j \in X_j$ for $j = 1, \ldots, n$. If $X_j = [0, \infty)$ the condition $x_j \in X_j$ requires x_j to be nonnegative. If $X_j = [a_j, b_j]$, it requires x_j to lie between the bounds a_j and b_j. If $X_j = (-\infty, \infty)$ it places no restriction on x_j. The latter case is a reminder that even the general condition $x \in X$ in $(\mathcal{P})$ doesn't necessarily restrict x, because we can always take X to be all of $\mathbb{R}^n$ when we want to deal in effect with constraints of type $f_i(x) \leq 0$ or $f_i(x) = 0$ only. The whole space $\mathbb{R}^n$ is considered to be a polyhedral subset of $\mathbb{R}^n$, as is the empty set $\emptyset$; singleton sets (consisting of exactly one point) are polyhedral as well.

A technical point that shouldn't be overlooked in setting up a model is the existence of a solution. If that isn't guaranteed by the formulation, something's wrong; note that

the issue isn't whether the "application" has a solution in some sense (e.g. the existence in principle of a best mode of operating a given system), but whether the mathematical description of the problem is adequate. Under the assumptions we have given for $(\mathcal{P})$ a simple condition guaranteeing the existence of at least one optimal solution, provided there is at least one feasible solution, is the *boundedness* of the set X. (Boundedness of X means that for each coordinate x_j of x, there is an upper bound to x_j as x ranges over X and also a lower bound.) A more flexible criterion would be the boundedness, for each $\mu > 0$, of the set of all $x \in X$ satisfying $f_i(x) \leq \mu$ for $i = 0, 1, \ldots, s$ and $|f_i(x)| \leq \mu$ for $i = s+1, \ldots, m$.

Convexity has already been mentioned as a critical property in optimization which needs to be recognized and taken advantage of as far as possible when it is present. A set $C \subset \mathbb{R}^n$ is said to be *convex* if it contains along with any two different points the line segment joining those points:

$$x \in C,\ x' \in C,\ 0 < t < 1 \quad \Longrightarrow \quad (1-t)x + tx' \in C. \tag{1.1}$$

(In particular, the empty set is convex, as are sets consisting of a single point.) A real-valued function f on $\mathbb{R}^n$ is called *convex* if it satisfies the inequality

$$f\big((1-t)x + tx'\big) \leq (1-t)f(x) + tf(x') \text{ for any } x \text{ and } x' \text{ when } 0 < t < 1. \tag{1.2}$$

It's *concave* if the opposite inequality always holds, and *affine* under equality; the affine functions $f : \mathbb{R}^n \to \mathbb{R}$ have the form $f(x) = v \cdot x + \text{const}$. Finally, f is *strictly* convex if "$\leq$" can be replaced by "$<$" in (1.2); it's strictly concave in the case of "$>$."

Convexity is a large subject which can barely be touched on here; a book with many details is [1]. The importance of convexity in optimization comes from the following crucial properties.

Theorem 1.1.

(a) *In minimizing a convex function f_0 over a convex set C, every locally optimal solution $\bar{x}$ (if there is one) is globally optimal.*

(b) *In minimizing a strictly convex function f_0 over a convex set C, there can be no more than one optimal solution.*

In contrast to these properties of "convex optimization," two major difficulties with "nonconvex optimization" stand out. First, there is *virtually no way* to arrive for sure at a globally optimal solution. There are some global optimization techniques, more or less amounting in practice to forms of random search, but even with these one generally has

to be content merely with a statistical prospect of probably locating the true minimum eventually through persistence. In practice when applying an optimization package, for instance, one should be skeptical about any claims an optimal solution has been found, in the absence of convexity. Just because a sequence of points generated by a method seems to settle down and converge to something, that doesn't necessarily mean that an optimal solution is being approximated. This is a central issue in the analysis of algorithms. At best, with well designed methods that are soundly based on theoretical principles, one can hope that a locally optimal solution has been located, but that would still leave open the possibility that some other locally optimal solution—a better one—exists nearby.

Second, there is *virtually no way* to know that a problem has a *unique* optimal solution, apart from the strict convexity criterion just offered. This is even true in convex optimization. It's not wise therefore to speak of "the" solution to a problem of general type. Of course, a problem may well turn out to have a unique solution; the trouble is that we can't know that in advance, nor in the nonconvex case can we even hope to check whether an optimal solution already found (if that were possible) is unique.

These observations about what can go wrong without convexity might be regarded as raising false issues, in a sense. For some practitioners, it may be enough just to use optimization methodology to achieve improvements. Achieving the "ultimate" doesn't really matter. That's true to a degree, but only in the background of a method that provides a sequence of feasible points that get better and better. Most computational methods for problems with nonlinear constraints only approach feasibility in the limit, and that can open the door to various dangers. Another thing to remember is that optimization methods often entail the repeated solution of certain subproblems, such as in determining a good direction in which to search for improvement. One has to be careful that if such subproblems aren't solved to full optimality the method is still valid.

How do the properties in Theorem 1.1 connect with problem $(\mathcal{P})$? We'll refer to the *convex case* of $(\mathcal{P})$ when the objective and inequality constraint functions $f_0, f_1, \ldots, f_s$ are convex and the equality constraint functions $f_{s+1}, \ldots, f_m$ are affine.

Theorem 1.2. *In the convex case of $(\mathcal{P})$ the feasible set C is convex, so the property in Theorem 1.1(a) holds. If f_0 is not just convex but strictly convex, the property in Theorem 1.1(b) holds also.*

The next results review some criteria for a function to be convex. We denote by $\nabla f(x)$ the gradient of f at x, which is the vector of first partial derivatives. Similarly, we let $\nabla^2 f(x)$ stand for the square matrix of second partial derivatives, called the Hessian matrix of f at x. Recall that a matrix $H \in \mathbb{R}^{n\times n}$ is *positive semidefinite* when $w \cdot Hw \geq 0$

or all $w \in \mathbb{R}^n$. It is *positive definite* when $w \cdot Hw > 0$ for all $w \in \mathbb{R}^n$, except $w = 0$.

Proposition 1.3. *Let f be a function on $\mathbb{R}^n$ with continuous second derivatives.*

(a) *If f is convex, then $\nabla^2 f(x)$ is positive semidefinite for all x.*

(b) *If $\nabla^2 f(x)$ is positive semidefinite for all x, then f is convex.*

(c) *If $\nabla^2 f(x)$ is positive definite for all x, then f is strictly convex.*

Proposition 1.4.

(a) *If f_1 and f_2 are convex, then $f_1 + f_2$ is convex. If in addition either f_1 or f_2 is strictly convex, then $f_1 + f_2$ is strictly convex.*

(b) *If f is convex and $\lambda \geq 0$, then λf is convex. If f is strictly convex and $\lambda > 0$, then λf is strictly convex.*

(c) *If $f(x) = \phi(g(x))$ with g convex on $\mathbb{R}^n$ and ϕ is convex and nondecreasing on $\mathbb{R}$, then g is convex on $\mathbb{R}^n$. If in addition g is strictly convex and ϕ is increasing, then g is strictly convex.*

(d) *If $f(x) = g(Ax + b)$ for a convex function g on $\mathbb{R}^m$ and an matrix $A \in \mathbb{R}m \times n$ and a vector $b \in \mathbb{R}^n$, then f is convex on $\mathbb{R}^n$. If g is strictly convex and A has rank n, then f is strictly convex.*

(e) *If $f(x) = \sup_{s \in S} g_s(x)$ for a finite or infinite collection $\{g_s\}_{s \in S}$ of convex functions on $\mathbb{R}^n$, then f is convex on $\mathbb{R}^n$.*

Later we will use these criteria in verifying the convexity of expressions defined with penalty terms that aren't differentiable.

2. OPTIMALITY CONDITIONS

First-order optimality conditions for problem $(\mathcal{P})$ will now be stated in terms of Lagrange multipliers. In order to get a simple form of expression that will later be extendible to problems with penalty functions, we use the concept and notation of normal vectors. A more general exposition of the material in this section, complete with proofs, is available in the expository article [2].

Definition 2.1. *The (outward) normal vectors to the polyhedral set X at a point $\bar{x} \in X$ are the vectors v such that*

$$v \cdot (x - \bar{x}) \leq 0 \text{ for all } x \in X.$$

The set of all these vectors is called the normal cone to X at and is denoted by $N_X(\bar{x})$.

The term "cone" refers to the fact that for any $v \in N_X(\bar{x})$ and $\lambda \geq 0$, then $\lambda v \in N_X(\bar{x})$. In other words, $N_X(\bar{x})$ is a bundle of rays emanating from the origin—unless $\bar{x}$ is an interior point of X, in which case $N_X(\bar{x})$ consists of the zero vector alone. In the case where X is a box, the normal cone condition is especially easy to understand: in terms of $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$ we have

$$\left\{ \begin{array}{l} \text{if } \bar{x} \in X = X_1 \times \cdots \times X_n,\ \bar{x} = (\bar{x}_1, \ldots, \bar{x}_n), \text{ then} \\ v \in N_X(\bar{x}) \iff v_j \in N_{X_j}(\bar{x}_j) \text{ for } j = 1, \ldots, n. \end{array} \right. \tag{2.1}$$

When X_j is closed interval with lower bound a_j and upper bound b_j (these bounds possibly being infinite), we get that

$$v_j \in N_{X_j}(\bar{x}_j) \text{ means } \left\{ \begin{array}{ll} v_j \geq 0 & \text{if } a_j < \bar{x}_j = b_j, \\ v_j \leq 0 & \text{if } a_i = \bar{x}_j < b_i, \\ v_j = 0 & \text{if } a_i < \bar{x}_j < b_i, \\ v_j \text{ unrestricted} & \text{if } a_i = \bar{x}_j = b_i. \end{array} \right. \tag{2.2}$$

In order to state the main result about first-order optimality conditions in problem $(\mathcal{P})$, we'll need a condition on the constraints. This condition will involve normal vectors to the set

$$D = \big\{\, u = (u_1, \ldots, u_m) \,\big|\, u_i \leq 0 \text{ for } i \in [1, s],\ u_i = 0 \text{ for } i \in [s+1, m] \,\big\}. \tag{2.3}$$

The constraints in $(\mathcal{P})$ can be written as

$$x \in X,\ F(x) \in D, \text{ where } F(x) = \big(f_1(x), \ldots, f_m(x)\big). \tag{2.4}$$

Note that D is another polyhedral set, actually a box:

$$D = D_1 \times \cdots \times D_m \text{ with } D_i = \begin{cases} (-\infty, 0] & \text{for } i \in [1, s], \\ [0, 0] & \text{for } i \in [s+1, m]. \end{cases} \tag{2.5}$$

Definition 2.2. *The basic constraint qualification at a feasible solution $\bar{x}$ to problem $(\mathcal{P})$ is the condition:*

$$(\mathcal{Q}) \qquad \begin{cases} \text{there is no vector } \bar{y} = (\bar{y}_1, \ldots, \bar{y}_m) \text{ other than } \bar{y} = 0 \text{ such that} \\ \bar{y} \in N_D\big(F(\bar{x})\big), \qquad -\big[\bar{y}_1 \nabla f_1(\bar{x}) + \cdots + \bar{y}_m \nabla f_m(\bar{x})\big] \in N_X(\bar{x}). \end{cases}$$

This condition is needed to rule out situations where the constraints fail to give a robust representation of the feasible set C around $\bar{x}$. From the product form of D in (2.5), it's clear that

$$\bar{y} \in N_D\big(F(\bar{x})\big) \iff \begin{cases} \bar{y}_i = 0 & \text{for } i \in [1, s] \text{ with } f_i(\bar{x}) < 0, \\ \bar{y}_i \geq 0 & \text{for } i \in [1, s] \text{ with } f_i(\bar{x}) = 0, \\ \bar{y}_i \text{ unrestricted} & \text{for } i \in [s+1, m]. \end{cases} \tag{2.6}$$

Writing these sign conditions as $\bar{y} \in N_D\big(F(\bar{x})\big)$ is not only convenient but leads the way to a statement of optimality conditions that can be extended later to problems incorporating penalty expressions. Observe that if $\bar{x}$ belongs to the interior of X (as is certainly true when $X = \mathbb{R}^n$), the gradient condition in $(\mathcal{Q})$ reduces to $\sum_{i=1}^m \bar{y}_i \nabla f_i(\bar{x}) = 0$. Then $(\mathcal{Q})$ becomes the John constraint qualification [3], which is the dual form of the Mangasarian-Fromovitz constraint qualification [4].

Optimality conditions for $(\mathcal{P})$ involve the *Lagrangian function*

$$L(x, y) = f_0(x) + y_1 f_1(x) + \cdots + y_m f_m(x) \text{ for } x \in X \text{ and } y \in Y, \tag{2.7}$$

where

$$Y = \mathbb{R}^s_+ \times \mathbb{R}^{m-s} = \big\{ y = (y_1, \ldots, y_m) \,\big|\, y_i \geq 0 \text{ for } i \in [1, s] \big\}. \tag{2.8}$$

Observe that Y too is a box, and

$$\bar{y} \in N_D\big(F(\bar{x})\big) \iff F(\bar{x}) \in N_Y\big(\bar{y}\big). \tag{2.9}$$

This equivalence is clear from (2.6) and the fact that

$$u \in N_Y\big(\bar{y}\big) \iff \begin{cases} u_i \geq 0 & \text{for } i \in [1, s] \text{ with } \bar{y}_i = 0, \\ u_i = 0 & \text{for } i \in [1, s] \text{ with } \bar{y}_i > 0, \\ & \quad \text{and for } i \in [s+1, m]. \end{cases} \tag{2.10}$$

Theorem 2.3. *If $\bar{x} \in X$ is a locally optimal solution to $(\mathcal{P})$ at which the basic constraint qualification $(\mathcal{Q})$ is satisfied, there must exist a vector $\bar{y} \in Y$ such that*

$$(\mathcal{L}) \qquad -\nabla_x L(\bar{x}, \bar{y}) \in N_X(\bar{x}), \qquad \nabla_y L(\bar{x}, \bar{y}) \in N_Y(\bar{y}).$$

Condition $(\mathcal{L})$ is the *Lagrange multiplier rule* in general form. Since $\nabla_y L(\bar{x}, \bar{y}) = F(\bar{x})$, the second part of $(\mathcal{L})$ is simply another statement of the sign conditions in (2.6) on the multipliers $\bar{y}_i$, but again one which will lead to extensions. The first part of $(\mathcal{L})$ becomes the equation $\nabla f_0(\bar{x}) + \sum_{i=1}^m \bar{y}_i \nabla f_i(\bar{x}) = 0$ when $\bar{x}$ is an interior point of X.

Although the normal cone notation here is a recent development, and the incorporation of the abstract constraint $x \in X$ a novel feature, the first-order optimality conditions in Theorem 2.3 are basically the ones found in every textbook on optimization. They are commonly called the *Kuhn-Tucker conditions* because of the 1951 paper of Kuhn and Tucker [4], but it's now known that the same conditions were derived in the 1939 master's thesis of Karush [5], which however was never published. For this reason they are also referred to as the *Karush-Kuhn-Tucker conditions*.

In many applications linear constraints are very important, and then the following variant of Theorem 2.3 is useful.

Theorem 2.4. *The assertion of Theorem 2.3 remains valid when the basic constraint qualification $(\mathcal{Q})$ at $\bar{x}$ is replaced by*

$$(\mathcal{Q}') \qquad \begin{cases} \text{the nonzero vectors } \bar{y} \in Y \text{ satisfying} \\ \bar{y} \in N_D\big(F(\bar{x})\big), \qquad -\big[\bar{y}_1 \nabla f_1(\bar{x}) + \cdots + \bar{y}_m \nabla f_m(\bar{x})\big] \in N_X(\bar{x}), \\ \text{if any, have } \bar{y}_i = 0 \text{ for each index } i \text{ such that } f_i \text{ is not affine.} \end{cases}$$

On the basis of this theorem, for instance, the multiplier rule $(\mathcal{L})$ is always necessary for optimality in problems having only linear constraints.

What dangers are there in applying Lagrangian optimization methodology in the absence of being able to verify, when nonlinear constraints are present, that the constraint qualification $(\mathcal{Q})$ or (Q') is definitely satisfied at an optimal solution? This depends on the solution technique involved, but the main difficulty is that if a technique can at best identify points singled out as candidates by the Lagrange multiplier rule, but the desired solution is not such a point, then the technique has no hope of finding it. The technique in combination with some kind of global search could seem indicate a particular point as the best, but only because it is blind to the real solution. Another possibility is that numerical instabilities may be experienced. However, there is good theoretical support for the notion that the Lagrange multiplier rule "usually" is necessary for optimality.

Lagrange multipliers have special properties under convexity which lead to another level of practical usage.

Theorem 2.5. *In the convex case of problem* $(\mathcal{P})$*, the Lagrangian* $L(x, y)$ *is convex in* x *and concave (actually affine) in* y*. The multiplier rule* $(\mathcal{L})$ *in Theorem 2.3 is equivalent then to the condition that*

$$\begin{cases} \text{the minimum of } L(x, \bar{y}) \text{ in } x \in X \text{ is attained at } \bar{x}, \\ \text{the maximum of } L(\bar{x}, y) \text{ in } y \in Y \text{ is attained at } \bar{y}. \end{cases} \tag{2.11}$$

This theorem supports—up to a certain degree–a popular approach called *Lagrangian relaxation*, which is especially attractive in connection with ideas of decomposing a large-scale problem by introducing appropriate "prices" to achieve a decentralization of the decision process. Under this approach, a vector $\hat{y}$ is selected, and then a vector $\hat{x}$ is obtained by minimizing $L(x, \hat{y})$ subject to $x \in X$. It is hoped that through a good choice of $\hat{y}$ a nearly optimal solution $\hat{x}$ to $(\mathcal{P})$ itself will generated. But is this hope justified?

According to Theorem 2.5, if an optimal solution $\bar{x}$ *exists and satisfies the Lagrange multiplier rule* $(\mathcal{L})$ *along with some vector* $\bar{y}$ (the latter being true when $(\mathcal{Q})$ or $(\mathcal{Q}')$ holds at $\bar{x}$), and if one is dealing with a *convex* case of $(\mathcal{P})$, then $\bar{x}$ will be *among* the vectors $\hat{x}$ obtainable under the Lagrangian relaxation approach if, through luck or design, $\hat{y}$ can be chosen equal to $\bar{y}$. With strict convexity of the objective function f_0, $\hat{x}$ will have to be $\bar{x}$, the unique optimal solution to $(\mathcal{P})$, in these circumstances. But without strict convexity of f_0, even with everything else holding, $\hat{x}$ *might not even satisfy the constraints* of $(\mathcal{P})$.

In the nonconvex case of $(\mathcal{P})$, unfortunately, just about everything can go wrong in Lagrangian relaxation. A vector $\hat{x}$ obtained in this manner, even from some "ideal" choice of $\hat{y}$, need have no relation to optimality. All that can be said then is that the minimizing value of $L(x, \hat{y})$ as x ranges over X will be a *lower bound* for the optimal value (number) associated with $(\mathcal{P})$—provided that this minimizing value is *global*, which as noted earlier is very hard to guarantee without convexity. The vector $\hat{x}$ offers nothing.

Incidentally, it's interesting to note that this negative conclusion from theory doesn't stop economists, especially in today's political climate, from flirting with the idea that if only the right markets and prices could be introduced, decisions could effectively be decentralized and society could function more efficiently. Theory provides no backing for this concept in situations where convexity is absent, such as characterize much of the real world. (It's known that in the case of a very large number of small agents, such in classical free markets, a kind of convexification is approached, but this is far from the actual economies of developed countries.)

Lagrangian relaxation can be understood further in connection with saddle points and dual problems. A pair of elements $\bar{x}$ and $\bar{y}$ is said to give a *saddle point* of L on $X \times Y$ when (2.11) holds; this can also be written as

$$L(x, \bar{y}) \geq L(\bar{x}, \bar{y}) \geq L(\bar{x}, y) \text{ for all } x \in X,\ y \in Y \quad (\text{where } \bar{x} \in X,\ \bar{y} \in Y). \tag{2.12}$$

This relation has a life of its own as an equilibrium condition for certain "games," and it leads to further properties of Lagrange multipliers which are of prime importance for many applications. In particular it gives rise to the notion of *duality* in optimization. To appreciate the meaning of duality, let's first note that problem $(\mathcal{P})$ can be viewed as the problem of minimizing over all $x \in X$ the function f defined by

$$f(x) = \begin{cases} f_0(x) & \text{if } x \in C, \\ \infty & \text{if } x \notin C, \end{cases} \tag{2.13}$$

where C is the set of feasible solutions to $(\mathcal{P})$, and that f has the Lagrangian representation

$$f(x) = \sup_{y \in Y} L(x, y) = \sup_{y \in Y} \Big\{ f_0(x) + y_1 f_1(x) + \ldots + y_m f_m(x) \Big\} \text{ for } x \in X, \tag{2.14}$$

where the restriction of y to Y in taking the "sup" means that the coefficients y_i can be chosen arbitrarily for the terms indexed by $i = s+1, \ldots, m$, but must be nonnegative for $i = 1, \ldots, s$. By analogy in reversing the roles of x and y, we can state the problem:

$$(\mathcal{D}) \qquad \text{maximize } g(y) = \inf_{x \in X} \Big\{ f_0(x) + y_1 f_1(x) + \ldots + y_m f_m(x) \Big\} \text{ over } y \in Y.$$

This is the optimization problem *dual* to problem $(\mathcal{P})$ in the Lagrangian sense.

Observe that for each vector $\hat{y}$ the subproblem solved to get the value $g(\hat{y})$ of the essential objective function g in $(\mathcal{D})$ is precisely the one indicated in the Lagrangian relaxation approach. In general g might, like f in (2.13), be extended-real-valued. To learn more about the nature of the dual problem $(\mathcal{D})$ in a given case, with particular structure assigned to X and the f_i's, we would have to identify the set of points y where $g(y) > -\infty$ and regard that as the feasible set in $(\mathcal{D})$. Examples will be considered below, but we first record the main facts relating problems $(\mathcal{P})$ and $(\mathcal{D})$.

Theorem 2.6. *In the convex case of $(\mathcal{P})$, the existence for $\bar{x}$ of a multiplier vector $\bar{y}$ satisfying the Lagrange multiplier rule $(\mathcal{L})$ is sufficient for $\bar{x}$ to be a globally optimal solution to $(\mathcal{P})$. The vectors $\bar{y}$ that appear in this condition along with $\bar{x}$ are then precisely the optimal solutions to the dual problem $(\mathcal{D})$, and the optimal values in the two problems agree: one has*

$$\min(\mathcal{P}) \ = \ \max(\mathcal{D}).$$

The final equation in Theorem 2.6 confirms that, for any $\hat{y}$, the value $g(\hat{y})$ is a lower bound for the optimal value $\min(\mathcal{P})$, and under the right circumstances of convexity, this lower bound can be elevated to the degree that it actually equals the desired optimal value. Furthermore, Theorem 2.6 give the theoretical prescription to be used in designing an algorithm to product a multiplier vector $\bar{y}$ for which the equality holds. Once again, though, without the convexity the equation between $\min(\mathcal{P})$ and $\max(\mathcal{D})$ could very well become a strict inequality $>$. Then no amount of fiddling with the values of Lagrange multipliers could be expected to produce approximate optimal solutions fo $(\mathcal{P})$ through Lagrangian relaxation.

The best known and most highly successful example of duality in optimization occurs in *linear programming*, which is the case of problem $(\mathcal{P})$ where the objective function is linear, all the constraints are linear, and

$$X = \mathbb{R}^r_+ \times \mathbb{R}^{n-r} = \big\{ x = (x_1, \ldots, x_n) \,\big|\, x_j \geq 0 \text{ for } j \in [1, r] \big\}. \tag{2.15}$$

Adopting the notation

$$\begin{aligned} f_0(x) &= c_1 x_1 + \cdots + c_n x_n, \\ f_i(x) &= b_i - a_{i1} x_1 - \cdots - a_{in} x_n \text{ for } i = 1, \ldots m, \end{aligned}$$

we can express the problem in this special case as

$$(\mathcal{P}_{\text{lin}}) \qquad \begin{array}{c} \text{minimize } c_1 x_1 + \cdots + c_n x_n \text{ subject to } x_j \geq 0 \text{ for } j = 1, \ldots, r, \\ a_{i1} x_1 + \cdots + a_{in} x_n \begin{cases} \geq b_i & \text{for } i = 1, \ldots s, \\ = b_i & \text{for } i = s+1, \ldots, m. \end{cases} \end{array}$$

The Lagrangian function is

$$L(x, y) = \sum\nolimits_{j=1}^{n} c_j x_j + \sum\nolimits_{i=1}^{m} y_i b_i - \sum\nolimits_{i=1, j=1}^{m,n} y_i a_{ij} x_j, \tag{2.16}$$

which exhibits the same kind of symmetry between the x and y arguments as appears in the choice of X and Y. To obtain the problem dual to this, we must determine the function g defined in $(\mathcal{D})$ for this Lagrangian and see where it is finite or infinite. Elementary calculations show that $g(y) = \sum_{i=1}^m y_i b_i$ if $c_j - \sum_{i=1}^m y_i a_{ij} \geq 0$ for $j = 1, \ldots, r$ and $c_j - \sum_{i=1}^m y_i a_{ij} = 0$ for $j = r+1, \ldots, n$, whereas $g(y) = -\infty$ if y does not satisfy these constraints. The dual problem therefore comes out as

$$(\mathcal{D}_{\text{lin}}) \qquad \begin{array}{c} \text{maximize } y_1 b_1 + \cdots + y_m b_m \text{ subject to } y_i \geq 0 \text{ for } i = 1, \ldots, s, \\ y_1 a_{1j} + \cdots + y_m a_{mj} \begin{cases} \leq c_j & \text{for } j = 1, \ldots r, \\ = c_j & \text{for } j = r+1, \ldots, n. \end{cases} \end{array}$$

From all this symmetry it emerges that not only do the Lagrange multiplier vectors associated with an optimal solution to $(\mathcal{P}_{\text{lin}})$ have an interpretation as optimal solutions $\bar{y}$ to $(\mathcal{D}_{\text{lin}})$, but by the same token, the Lagrange multiplier vectors associated with an optimal solution to $(\mathcal{D}_{\text{lin}})$ have an interpretation as optimal solutions $\bar{x}$ to $(\mathcal{P}_{\text{lin}})$. Each of these problems furnishes the multipliers for the other.

Corollary 2.7 (Gale-Kuhn-Tucker Theorem [6])**.** *If either of the linear programming problems* $(\mathcal{P}_{\text{lin}})$ *or* $(\mathcal{D}_{\text{lin}})$ *has an optimal solution, then so does the other, and*

$$\min(\mathcal{P}_{\text{lin}}) = \max(\mathcal{D}_{\text{lin}}).$$

The pairs $(\bar{x}, \bar{y})$ *such that* $\bar{x}$ *solves* $(\mathcal{P}_{\text{lin}})$ *and* $\bar{y}$ *solves* $(\mathcal{D}_{\text{lin}})$ *are precisely the ones that, for the choice of* L, X *and* Y *corresponding to these problems, satisfy the Lagrange multiplier rule* $(\mathcal{L})$*, or equivalently, give a saddle point of* L *on* $X \times Y$*.*

Even for the nonconvex case of $(\mathcal{P})$, the dual problem $(\mathcal{D})$ has significance.

Proposition 2.8. *Regardless of whether* $(\mathcal{P})$ *is of convex type or not, the function* g *being maximized over the polyhedral set* Y *in* $(\mathcal{D})$ *is concave. For each* $y \in Y$ *the value* $g(y)$ *is a lower bound to the value* $\min(\mathcal{P})$*. The greatest of such lower bounds obtainable this way is* $\max(\mathcal{D})$*.*

In other words, by selecting any $y \in Y$ and then minimizing $L(x, y)$ over $x \in X$, one obtains a number denoted by $g(y)$ with the property that $g(y) \leq f_0(x)$ for every feasible solution x to problem $(\mathcal{P})$. This number may be useful in estimating how far a particular point $\hat{x}$ already calculated in $(\mathcal{P})$, and satisfying the constraints of $(\mathcal{P})$, may be from optimality. One will have

$$0 \leq f_0(\hat{x}) - \min(\mathcal{P}) \ \leq \ f_0(\hat{x}) - g(y), \tag{2.17}$$

so that if $f_0(\hat{x}) - g(y)$ is less than some threshold value ε, the decision can be made that $\hat{x}$ is good enough, and further computations aren't worth the effort. By applying an optimization technique to $(\mathcal{D})$, it may be possible to get better estimates of such sort. The best would be a dual optimal solution $\bar{y}$, for which $g(\bar{y}) = \max(\mathcal{D})$; then the estimate would take the form

$$0 \leq f_0(\hat{x}) - \min(\mathcal{P}) \ \leq \ f_0(\hat{x}) - \max(\mathcal{D}). \tag{2.18}$$

But in nonconvex problems where $\min(\mathcal{P}) > \max(\mathcal{D})$, the bound on the right can't be brought to 0 no matter how much effort is expended. The technique is therefore limited

in its ability to estimate optimality of $\hat{x}$. Another pitfall is that the estimates only make sense if the exact value of $g(y)$ can be computed for a given $y \in Y$, or at least a lower estimate c for $g(y)$ (then one gets $f_0(\hat{x}) - c$ as an upper bound to substitute for the right side in (2.17)). But in the nonconvex case of $(\mathcal{P})$ the expression $L(x, y)$ being minimized over $x \in X$ to calculate the value $g(y)$ may be nonconvex in x, yet the minimization must be *global*. Then, as already explained in Section 1, it may be difficult or impossible to know when the global minimum has been attained.

For more on duality in convex optimization, see [1], [7], [8]. For the theory of the *augmented Lagrangian* function for $(\mathcal{P})$, which makes saddle point characterizations of optimality possible even without convexity, see [2]. Second-order optimality conditions are discussed in [2] also.

3. EXTENDED PROBLEM MODELS

The conventional problem statement $(\mathcal{P})$ doesn't fully convey the range of possibilities available in setting up a mathematical model in optimization. First, it gives the impression that as modelers we won't have trouble distinguishing between objectives and constraints. We're supposed to know what should be minimized and be able to express it by a *smooth* function. All other features of the situation being addressed must be formulated as "black-and-white" constraints—side conditions that have to be satisfied exactly, or we'll be infinitely unhappy. No gray areas are allowed.

The real modeling context is often very different. There may well be some conditions that the variables must satisfy exactly, because otherwise the model doesn't make sense. For instance, a nonnegativity condition $x_j \geq 0$ may fall in this category: we wouldn't know how to interpret a negative value of x_j physically and aren't in the least interested in relaxing the constraint $x_j \geq 0$ to $x_j \geq -\varepsilon$, say. Other examples of such black-and-white constraints are defining relationships between variables. A condition like $x_3 - x_1 - x_2^2 = 1$ could simply indicate the definition of x_3 in terms of x_1 and x_2, and we wouldn't want to consider relaxing it. But many of the constraints may have a "soft" character. We might write $4.3x_1 + 2.7x_2 + x_3 \leq 5.6$ as a constraint because we desire the expression on the left not to exceed 5.6, but a sort of guesswork is involved. We could be quite content when the expression on the left had the value 5.9 if that resulted in substantial benefits in other respects. Another source of fuzziness might be that coefficients like 4.3 are just estimates, or worse. Then it seems foolish to insist on the inequality being satisfied without error.

In fact, a fair description of the difficulty often faced in reality may be that there are several expressions $f_0(x), f_1(x), \ldots, f_m(x)$ of interest to the modeler, who is seeking a sort of "ideal combination" subject to the trade-offs that may be involved. Somewhat

arbitrarily, one of these expressions is selected as the one to optimize while the others are held in fixed ranges, but after the optimization has been carried out, there may be second thoughts inspired by knowledge generated during the optimization process, and a modified optimization formulation may then be tested out. Besides choosing one of the functions as the objective and putting constraint bounds on the others, it's possible of course to form some combination. Examples to consider might be the minimization, subject to the underlying hard constraints on x, of a weighted sum $f(x) = f_0(x)+c_1f_1(x)+\cdots+c_mf_m(x)$ or a weighted max

$$f(x) = f_0(x) + \max\big\{c_1f_1(x), \ldots, c_mf_m(x)\big\}. \tag{3.1}$$

Or, taking as reference the minimization of "cost," with $f_0(x)$ expressing certain costs directly, we could consider for each other function f_i a nonlinear rescaling function ρ_i that converts the value $f_i(x)$ into an associated cost $\rho_i\big(f_i(x)\big)$. Then we would want to minimize

$$f(x) = f_0(x) + \rho_1\big(f_1(x)\big) + \cdots + \rho_m\big(f_m(x)\big). \tag{3.2}$$

Although the given functions $f_0, f_1, \ldots, f_m$ may be smooth, the function f obtained in such a manner may be nonsmooth. Problems of minimizing such a function aren't well covered by the standard theory for $(\mathcal{P})$.

To get around this difficulty and enhance the possibilities for optimization modeling, we direct our attention to the following *extended* problem formulation, which was introduced in [2]:

$$(\overline{\mathcal{P}}) \qquad \text{minimize } f(x) = f_0(x) + \rho\big(F(x)\big) \text{ over } x \in X, \text{ where } F(x) = \big(f_1(x), \ldots, f_m(x)\big).$$

In this the functions $f_0, f_1, \ldots, f_m$ will still be assumed to be smooth, and the set X to be closed, but the function ρ need not be smooth and can even take on the value ∞.

For a sense of what $(\overline{\mathcal{P}})$ covers, let's consider first the cases where ρ is *separable*, i.e.,

$$\rho(u) = \rho(u_1, \ldots, u_m) = \rho_1(u_1) + \cdots + \rho_m(u_m), \tag{3.3}$$

so that $(\overline{\mathcal{P}})$ takes the form of minimizing an expression of the form (3.2) over X. Right away we can observe that $(\overline{\mathcal{P}})$ contains $(\mathcal{P})$ as corresponding to the choice:

$$\begin{aligned} &\text{for } i = 1, \ldots, s: && \rho_i(u_i) = \begin{cases} 0 & \text{if } u_i \le 0, \\ \infty & \text{if } u_i > 0, \end{cases} \\ &\text{for } i = s+1, \ldots, m: && \rho_i(u_i) = \begin{cases} 0 & \text{if } u_i = 0, \\ \infty & \text{if } u_i \ne 0. \end{cases} \end{aligned} \tag{3.4}$$

This gives for $f(x)$ in (3.2) the value $f_0(x)$ when the point $x \in X$ is feasible in $(\mathcal{P})$, but the value ∞ if x is not feasible. As we saw earlier, the minimization of this "essential objective" function f over X is equivalent to the minimization of $f_0(x)$ subject to x being feasible. This example may create some discomfort with its use of ∞, but it also serves as a reminder of the true nature of the modeling represented by the conventional problem $(\mathcal{P})$. There is an infinite penalty if the stated x conditions are violated, but no gray area allowing for "approximate" satisfaction.

Obviously, the function f in this infinite penalty case of $(\overline{\mathcal{P}})$ corresponding to $(\mathcal{P})$ is far from smooth and even is discontinuous, but even a finite penalty approach may be incompatible with constructing a *smooth* function to minimize. For example, in the pure *linear penalty case* of $(\overline{\mathcal{P}})$ the choice is

$$\begin{aligned} &\text{for } i = 1, \dots, s: && \rho_i(u_i) = \begin{cases} 0 & \text{if } u_i \le 0, \\ d_i u_i & \text{if } u_i > 0, \end{cases} \\ &\text{for } i = s+1, \dots, m: && \rho_i(u_i) = \begin{cases} 0 & \text{if } u_i = 0, \\ d_i |u_i| & \text{if } u_i \ne 0, \end{cases} \end{aligned} \tag{3.5}$$

with positive constants d_i. These functions have "kinks" at the origin which prevent f from being smooth. The pure *quadratic penalty case* of $(\overline{\mathcal{P}})$ instead takes

$$\begin{aligned} &\text{for } i = 1, \dots, s: && \rho_i(u_i) = \begin{cases} 0 & \text{if } u_i \le 0, \\ \frac{1}{2} d_i u_i^2 & \text{if } u_i > 0, \end{cases} \\ &\text{for } i = s+1, \dots, m: && \rho_i(u_i) = \begin{cases} 0 & \text{if } u_i = 0, \\ \frac{1}{2} d_i u_i^2 & \text{if } u_i \ne 0, \end{cases} \end{aligned} \tag{3.6}$$

with coefficients $d_i > 0$. Penalty functions of this type are first-order smooth, yet discontinuous in their second derivatives.

To illustrate the modeling considerations, consider a situation where the demand for a certain commodity is $d > 0$, and this is to be met by producing amounts $x_j \ge 0$ at plants $j = 1, \dots, n$, the costs being $\phi_j(x_j)$. One formulation as a problem of optimization would be to minimize $f_0(x) = \phi_1(x_1) + \cdots + \phi_n(x_n)$ over all vectors $x \in X = \mathbb{R}^n_+$ satisfying $d - x_1 - \cdots - x_n = 0$. But this could be a fragile approach, since it takes the target to be exact and makes no provision for not meeting it precisely. A better model could be to minimize $f_0(x) + \rho\big(d - x_1 - \cdots - x_n\big)$, where $\rho(u) = ru$ when $u \ge 0$ and $\rho(u) = q|u|$ when $u < 0$, where the parameter values r and q are positive. This would correspond to a penalty rate of r per unit of overproduction, but a penalty rate of q per unit of underproduction. The function ρ in this case is finite but has a kink at the origin. An alternative might be a formula for ρ that maintains the positive slope r for significantly positive u and the

negative slope $-q$ for significantly negative u, but introduces a quadratic rounding between the two linear segments of the graph in order to do away with the kink.

A wide and flexible class of functions ρ_i, which aren't necessarily just penalty functions in the traditional sense, has been proposed by Rockafellar and Wets [10], [11], for modeling purposes in dynamic and stochastic programming. These are functions describable with four parameters β_i, $\hat{y}_i$, $\hat{y}_i^+$, $\hat{y}_i^-$, where

$$0 \le \beta_i < \infty, \qquad -\infty < \hat{y}_i < \infty, \qquad -\infty \le \hat{y}_i^- \le \hat{y}_i \le \hat{y}_i^+ \le \infty.$$

The formula falls into three pieces and is best described by first introducing the auxiliary function $\hat{\rho}_i(u_i) = \hat{y}_i u_i + (1/2\beta_i)u_i^2$, this being the unique quadratic function with the property that $\hat{\rho}_i(0) = 0$, $\hat{\rho}_i'(0) = \hat{y}_i$, and $\hat{\rho}_i''(0) = 1/\beta_i$. Let $\hat{u}_i^+$ be the unique value such that $\hat{\rho}_i'(\hat{u}_i^+) = \hat{y}_i^+$, and similarly let $\hat{u}_i^-$ be the unique value such that $\hat{\rho}_i'(\hat{u}_i^-) = \hat{y}_i^-$. Then

$$\rho_i(u_i) = \begin{cases} \hat{\rho}_i(\hat{u}_i^+) + \hat{y}_i^+(u_i - \hat{u}_i^+) & \text{when } u_i \ge \hat{u}_i^+, \\ \hat{\rho}_i(u_i) & \text{when } \hat{u}_i^- \le u_i \le \hat{u}_i^+, \\ \hat{\rho}_i(\hat{u}_i^-) + \hat{y}_i^-(u_i - \hat{u}_i^-) & \text{when } u_i \le \hat{u}_i^-. \end{cases} \tag{3.7}$$

In other words, ρ_i agrees with the quadratic function $\hat{\rho}_i$, except that it extrapolates linearly to the right from the point where the slope of $\hat{\rho}_i$ is the specified value $\hat{y}_i^+$, and linearly to the left from the point where the slope is $\hat{y}_i^-$. If $\hat{y}_{i+} = \infty$, this is taken to mean that the quadratic graph is followed forever to the right without switching over to a linear expression; the interpretation for $\hat{y}_i^- = -\infty$ is analogous. The case of $\beta_i = 0$ is taken to mean that there is no quadratic middle piece at all: the function is given by $\hat{y}^+ u_i$ when $u_i > 0$ and by $\hat{y}^- u_i$ when $u_i < 0$.

The functions in (3.5) and (3.6), and even (3.4), can be interpreted as special cases of (3.7). So too can the ρ function in the small modeling illustration. (The initial version of *rho* in the illustration would correspond to $\hat{y}_+ = r$, $\hat{y}_+ = -q$, $\hat{y} = 0$, and $\beta = 0$; the rounded version would differ only in having $\beta = \varepsilon > 0$.) This form also covers expressions that arise in augmented Lagrangian theory [2]. An example not of this kind, and not having the separable structure in (3.3), is

$$\rho(u) = \rho(u_1, \ldots, u_m) = \max\{u_1, \ldots, u_m\}. \tag{3.8}$$

This corresponds in $(\overline{\mathcal{P}})$ to the minimization of the expression $f(x)$ in (3.1).

Of course, a problem can also be formulated with a mixture of expressions like these. For each f_i one can decide whether to incorporate it into the model with an exact equality or inequality constraint, in effect by choosing the corresponding ρ_i as in (3.4), or one can

associate it with a ρ_i conforming to the prescription in (3.5), (3.6), or more generally (3.7). Certain functions can be lumped together by a "max" expression as in (3.7), and so forth.

It may seem that the level of generality being suggested is too complicated to be usable in practice. But things are simpler than might first be imagined. All these examples a show an underlying pattern which we capture by the following condition.

Definition 3.1. *The function ρ on $\mathbb{R}^m$ will be said to have an elementary dual representation if it can be expressed alternatively by*

$$\rho(u) \;=\; \sup_{y\in Y}\Big\{y{\cdot}u - k(y)\Big\}, \tag{3.9}$$

where Y is some nonempty polyhedral set in $\mathbb{R}^m$ and k is some linear-quadratic convex function on $\mathbb{R}^m$. (Possibly $Y=\mathbb{R}^m$, or $k\equiv 0$. "Linear-quadratic" refers to a polynomial expression with no terms of degree higher than 2.) In the separable case (3.3) this comes down to whether each function ρ_i on $\mathbb{R}$ can be expressed alternatively by

$$\rho_i(u_i) \;=\; \sup_{y_i\in Y_i}\Big\{y_iu_i - k_i(y_i)\Big\}, \tag{3.10}$$

where Y_i is some nonempty closed interval in $\mathbb{R}$ and k is some linear-quadratic convex function on $\mathbb{R}$.

Let's verify that the examples given do fit this. The case where $(\overline{\mathcal{P}})$ reduces to $(\mathcal{P})$ corresponds to $Y=\mathbb{R}^s_+\times\mathbb{R}^{m-s}$ and $k\equiv 0$; in other words, the functions ρ_i in (3.4) achieve the representation (3.10) through

$$\begin{array}{lll} \text{for } i=1,\dots,s: & k_i(y_i)\equiv 0, & Y_i=[0,\infty), \\ \text{for } i=s+1,\dots,m: & k_i(y_i)\equiv 0, & Y_i=(-\infty,\infty). \end{array} \tag{3.11}$$

The pure linear penalty case (3.5) corresponds instead to

$$\begin{array}{lll} \text{for } i=1,\dots,s: & k_i(y_i)\equiv 0, & Y_i=[0,d_i], \\ \text{for } i=s+1,\dots,m: & k_i(y_i)\equiv 0, & Y_i=[-d_i,d_i]. \end{array} \tag{3.12}$$

The pure quadratic penalty case (3.6) is represented by

$$\begin{array}{lll} \text{for } i=1,\dots,s: & k_i(y_i)=(1/2d_i)y_i^2, & Y_i=[0,\infty), \\ \text{for } i=s+1,\dots,m: & k_i(y_i)=(1/2d_i)y_i^2, & Y_i=(-\infty,\infty). \end{array} \tag{3.13}$$

The more general kind of ρ_i function in (3.6) corresponds to

$$k_i(y_i)=(\beta_i/2)|y_i-\hat{y}_i|^2, \qquad Y_i=[\hat{y}_i^-,\hat{y}_i^+]. \tag{3.14}$$

Finally, the max function case in (3.1) and (3.8)—which is not separable—arises from

$$k(y) \equiv 0, \qquad Y = \big\{\, y \,\big|\, y_i \geq 0,\ y_1 + \cdots + y_m = 1 \,\big\}. \tag{3.15}$$

Definition 3.2. *By the extended Lagrangian function corresponding to the extended problem $(\overline{\mathcal{P}})$ in the case where ρ has an elementary dual representation in the sense of Definition 3.1 for a set Y and function k, we shall mean the function*

$$\overline{L}(x, y) = f_0(x) + y_1 f_1(x) + \cdots + y_m f_m(x) - k(y) \ \text{ on } \ X \times Y.$$

Optimality conditions generalizing the ones for $(\mathcal{P})$ will be stated for $(\overline{\mathcal{P}})$ in terms of X, Y, and $\overline{\mathcal{L}}$. For this we need to develop the correct analog of the constraint qualification $(\mathcal{Q})$ that was used for $(\mathcal{P})$.

Proposition 3.3. *When the function ρ has an elementary dual representation as in Definition 3.1, the set*

$$D = \big\{\, u = (u_1, \ldots, u_m) \,\big|\, \rho(u) < \infty \,\big\}$$

is nonempty and polyhedral in $\mathbb{R}^m$. On D, ρ is a finite convex function, in fact ρ is continuous and piecewise linear-quadratic on D. The feasible set in $(\overline{\mathcal{P}})$, defined to be the set of point $x \in X$ where $f(x) < \infty$, is given by

$$C = \big\{\, x \in X \,\big|\, F(x) \in D \,\big\}.$$

These properties are easy to see except for the polyhedral nature of D and the piecewise linear-quadratic nature of ρ on D, which are proved in [12].

This view of feasibility in problem $(\overline{\mathcal{P}})$ makes it possible to state the constraint qualification for the extended problem in the same manner as for the original problem.

Definition 3.4. *The basic constraint qualification at a feasible solution $\bar{x}$ to problem $(\overline{\mathcal{P}})$, when ρ has an elementary dual representation, is the condition:*

$$(\overline{\mathcal{Q}}) \qquad \begin{cases} \text{there is no vector } \bar{y} = (\bar{y}_1, \ldots, \bar{y}_m) \text{ other than } \bar{y} = 0 \text{ such that} \\ \bar{y} \in N_D\big(F(\bar{x})\big), \qquad -\big[\bar{y}_1 \nabla f_1(\bar{x}) + \cdots + \bar{y}_m \nabla f_m(\bar{x})\big] \in N_X(\bar{x}), \end{cases}$$

where D is the polyhedral set in Proposition 3.3.

The main result about first-order optimality conditions in problem $(\overline{\mathcal{P}})$ can now be given. For details, see Rockafellar [2].

Theorem 3.5. *Suppose in $(\overline{\mathcal{P}})$ that ρ has an elementary dual representation in the sense of Definition 3.1 for a certain set Y and function k. If $\bar{x} \in X$ is a locally optimal solution to $(\overline{\mathcal{P}})$ at which the basic constraint qualification $(\overline{\mathcal{Q}})$ is satisfied, there must exist a vector $\bar{y} \in Y$ such that*

$$(\overline{\mathcal{L}}) \qquad -\nabla_x \overline{L}(\bar{x}, \bar{y}) \in N_X(\bar{x}), \qquad \nabla_y \overline{L}(\bar{x}, \bar{y}) \in N_Y(\bar{y}).$$

Convexity is important in $(\overline{\mathcal{P}})$ just as it was in $(\mathcal{P})$. We'll speak of the *convex case* of $(\overline{\mathcal{P}})$ when the extended Lagrangian $\overline{L}(x, y)$ is convex with respect to $x \in X$ for each $y \in Y$. (It's always concave in $y \in Y$ for each $x \in X$ by its definition.)

Theorem 3.6. *In the convex case of problem $(\overline{\mathcal{P}})$, the multiplier rule $(\overline{\mathcal{L}})$ in Theorem 3.5 is equivalent then to the saddle point condition*

$$\begin{cases} \text{the minimum of } \overline{L}(x, \bar{y}) \text{ in } x \in X \text{ is attained at } \bar{x}, \\ \text{the maximum of } \overline{L}(\bar{x}, y) \text{ in } y \in Y \text{ is attained at } \bar{y}. \end{cases} \tag{3.16}$$

This saddle point condition leads to a dual problem. The extended Lagrangian has been introduced in just such a way that the function f being minimized over X in $(\overline{\mathcal{P}})$ has the representation

$$f(x) = \sup_{y \in Y} \overline{L}(x, y) = \sup_{y \in Y} \Big\{ f_0(x) + y_1 f_1(x) + \ldots + y_m f_m(x) - k(y) \Big\} \text{ for } x \in X. \tag{3.17}$$

We therefore introduce as the *extended dual problem* associated with $(\overline{\mathcal{P}})$ (when ρ has an elementary dual representation) the problem

$$(\overline{\mathcal{D}}) \qquad \begin{array}{l} \text{maximize } \overline{g}(y) \text{ over } y \in Y, \text{ where} \\ \overline{g}(y) = \inf\limits_{x \in X} \overline{L}(x, y) = \inf\limits_{x \in X} \Big\{ f_0(x) + y_1 f_1(x) + \ldots + y_m f_m(x) - k(y) \Big\}. \end{array}$$

The results for $(\mathcal{P})$ and $(\mathcal{D})$ carry over to this more general pair of primal and dual problems.

Theorem 3.7. *In the convex case of $(\overline{\mathcal{P}})$, the existence for $\bar{x}$ of a multiplier vector $\bar{y}$ satisfying the extended Lagrange multiplier rule $(\overline{\mathcal{L}})$ is sufficient for $\bar{x}$ to be a globally optimal solution to $(\overline{\mathcal{P}})$. The vectors $\bar{y}$ that appear in this condition along with $\bar{x}$ are then precisely the optimal solutions to the dual problem $(\overline{\mathcal{D}})$, and the optimal values in the two problems agree: one has*

$$\min(\overline{\mathcal{P}}) \;=\; \max(\overline{\mathcal{D}}).$$

As a special case of this duality, of course, we have the earlier duality between $(\mathcal{P})$ and $(\mathcal{D})$, which corresponds to taking the function ρ to be given by the exact penalty expressions in (3.4). But the example we want to emphasize now is *extended linear-quadratic programming*, which will generalize the linear programming duality in Section 2. We take this term as referring to the case where the extended Lagrangian has the form

$$\overline{L}(x,y) = c{\cdot}x + \tfrac{1}{2}x{\cdot}Cx + b{\cdot}y - \tfrac{1}{2}y{\cdot}By - y{\cdot}Ax \tag{3.18}$$

where the matrices $C \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$ are symmetric and positive *semi*-definite (possibly 0). To give a general expression to the two problems in this case, we use the notation

$$\rho_{YB}(u) = \sup_{y\in Y}\Big\{y{\cdot}u - \tfrac{1}{2}y{\cdot}By\Big\}, \qquad \rho_{XC}(v) = \sup_{x\in X}\Big\{v{\cdot}x - \tfrac{1}{2}x{\cdot}Cx\Big\}. \tag{3.19}$$

These functions can be made more specific according to the particular choices made of the polyhedral sets X and Y along with the matrices C and B, in accordance with the examples we have been discussing. Especially to be noted is the case where X and Y are *boxes* and C and B are *diagonal*, because then the expressions in (3.19) break down component by component.

The primal and dual problems of extended linear-quadratic programming come out in this notation as:

$$(\mathcal{P}_{\rm elq}) \qquad \text{minimize } c{\cdot}x + \tfrac{1}{2}x{\cdot}Cx + \rho_{YB}\big(b - Ax\big) \text{ over } x \in X,$$

$$(\mathcal{D}_{\rm elq}) \qquad \text{maximize } b{\cdot}y - \tfrac{1}{2}y{\cdot}By - \rho_{XC}\big(A^*y - c\big) \text{ over } y \in Y,$$

where A^* denotes the transpose of the matrix A. The linear programming problems $(\mathcal{P}_{\rm lin})$ and $(\mathcal{D}_{\rm lin})$ correspond to

$$X = \mathbb{R}^r_+ \times \mathbb{R}^{n-r}, \qquad Y = \mathbb{R}^s_+ \times \mathbb{R}^{m-s}, \qquad C = 0, \qquad B = 0.$$

Theorem 3.8. *If either of the extended linear-quadratic programming problems $(\mathcal{P}_{\rm elq})$ or $(\mathcal{D}_{\rm elq})$ has an optimal solution, then so does the other, and*

$$\min(\mathcal{P}_{\rm elq}) = \max(\mathcal{D}_{\rm elq}).$$

The pairs $(\bar{x},\bar{y})$ such that $\bar{x}$ solves $(\mathcal{P}_{\rm elq})$ and $\bar{y}$ solves $(\mathcal{D}_{\rm elq})$ are precisely the ones that, for the choice of $\overline{L}$, X and Y corresponding to these problems, satisfy the extended Lagrange multiplier rule $(\overline{\mathcal{L}})$, or equivalently, give a saddle point of $\overline{L}$ on $X \times Y$.

This theorem was proved in [10]. The subject is elaborated and applied to dynamic modeling in [12]. Extended linear-quadratic programming models in multistage stochastic programming are described in [13].

Numerical approaches to extended linear-quadratic programming have been developed in Rockafellar and Wets [10], Rockafellar [14], Zhu and Rockafellar [15], Zhu [16], [17], and Chen and Rockafellar [16] for various purposes. No doubt much more could be done, so the brief review of this work that follows should be seen as merely suggestive of some of the possibilities.

Paper [10] explains how, as a fallback option, any problem of extended linear-quadratic programming can be reformulated as one of ordinary quadratic programming—through the introduction of many extra variables. That technique, while offering reassurance that exotic new codes are not necessarily needed to get numerical answers, suffers however from two drawbacks. It greatly increases the dimension of the problem to be solved and at the same time may disrupt special structure in the objective and constraints. Ideally, such structure should instead be put to use in computation, at least if the aim is to cope with the huge optimization models that can arise from dynamics and stochastics. But for problems of modest size the reduction technique may be adequate.

Closely related is the technique of rewriting the optimality condition ($\overline{\mathcal{L}}$) as a linear "variational inequality" as described in [14]. This can in turn be translated into a complementarity relation to which algorithms for linear complementarity problems can be applied. The symmetry between the primal and dual problems is thereby preserved, although dimensionality is again increased. No write-ups are yet available on this, but numerical experiments conducted by S. J. Wright at Argonne National Laboratories near Chicago on solving extended linear-quadratic programming problems through interior-point methods for linear complementarity problems appear very promising.

Most of the algorithmic development has been undertaken in the *strictly quadratic* case, i.e., with the assumption that both of the matrices B and C in ($\mathcal{P}_{\rm elq}$) are positive definite. While this assumption apparently excludes linear programming and even the standard form of quadratic programming, it's not as severe as it first may seem. A number of approaches to solving large-scale problems introduce "proximal terms" in the objective. These are regularizing terms in the form of a strictly quadratic (although possibly small) penalty for deviation from a current estimate of the solution. They are moved and updated as computations proceed. Each subproblem with such a term does have, in effect, a positive definite matrix B. It turns out that proximal terms can be added iteratively in the dual variables of the Lagrangian as well as the primal variables, and in that way a sequence of regularized subproblems is generated in which the associated C is positive definite too. By solving the subproblems, one obtains sequences of primal and dual vectors which, in the limit, solve ($\mathcal{P}_{\rm elq}$) and ($\mathcal{D}_{\rm elq}$).

From this perspective, the solution of strictly quadratic problems is the key to the solution of more general problems. Such an approach with proximal terms has, for instance, been explored in some detail in the context of two-stage stochastic programming in [10].

In [14], a novel class of algorithms for solving strictly quadratic problems ($\mathcal{P}_{\mathrm{elq}}$) and ($\mathcal{D}_{\mathrm{elq}}$) has been developed in terms of "envelope representations" of the essential objective functions f and g. The partial approximation of f and g by an "envelope representation" is somewhat kin to using a pointwise maximum of affine functions to represent a convex function from below (which can be seen as a cutting-plane idea), but the approximations are piecewise linear-quadratic rather than just piecewise affine. The envelope representations are generated by iterative application of steps in which the Lagrangian $L(x, y)$ is minimized in $x \in X$ for fixed y, or maximized in $y \in Y$ for fixed x. For many large-scale problems arising in applications, such steps are easy to carry out, because the models can be set up in such a way that $L(x, y)$ is separable in x and y—separately, cf. [13].

Results of numerical experiments using envelope methods to solve extended linear-quadratic programming problems are reported in [15]. In particular, that paper develops special methods called primal-dual projected gradient algorithms. These methods are characterized by having two procedures go on at once—one in the primal problem and one in the dual problem—with a kind of information feedback between them. The feedback is the source of dramatic improvements in the rate of convergence. Besides being effective for moderately sized problems, the algorithms have successfully been used to solve problems in as many as 100,000 primal and 100,000 dual variables in a stable manner. This line of research has been carried further in [15] and [16].

While envelope methods take advantage of possible decomposability of large-scale problem structure through separate separability of the Lagrangian in the primal and dual variables, another form of decomposition is exploited by the Lagrangian "splitting methods" introduced in [16]. These are aimed at problems in which the Lagrangian is a kind of sum of independent sub-Lagrangians coming from prospective subproblems and a bilinear linking expression. Examples are furnished in [13], but they also arise in finite-element models for partial differential equations and associated variational inequalities. Iterations proceed with an alternation between "backward steps" which can be calculated by assigning each subproblem to a separate processor, and "forward steps" which are analogous to integrating dynamics, or calculating conditional expectations.

REFERENCES

1. R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
2. R. T. Rockafellar, "Lagrange multipliers and optimality," SIAM Review, 1993.
3. F. John, "Extremum problems with inequalities as subsidiary conditions," in *Studies and Essays, Courant Anniversary Volume*, Interscience, New York, 1948.
4. O. L. Mangasarian and S. Fromovitz, "The Fritz John conditions in the presence of equality and inequality constraints," J. Math. Anal. Appl. **17** (1967), 73–74.
5. H. W. Kuhn and A. W. Tucker, "Nonlinear programming," Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability (J. Neyman, ed.), Univ. Calif. Press, Berkeley, 1951, 481–492.
6. W. Karush, "Minima of functions of several variables with inequalities as side conditions," master's thesis, Dept. of Mathematics, Univ. of Chicago, 1939.
7. D. Gale, H. W. Kuhn, and A. W. Tucker, "Linear programming and the theory of games," in *Activity Analysis of Production and Allocation* (T. C. Koopmans, ed.), Wiley, 1951, 317–328.
8. R. T. Rockafellar, *Conjugate Duality and Optimization*, Regional Conference Series No. 16, SIAM Publications, 1974.
9. R. T. Rockafellar, *Network Flows and Monotropic Optimization*, Wiley, 1984.
10. R. T. Rockafellar and R. J-B Wets, "A Lagrangian finite generation technique for solving linear-quadratic problems in stochastic programming," Math. Programming Studies **28** (1986), 63–93.
11. R. T. Rockafellar and R. J-B Wets, *Linear-quadratic problems with stochastic penalties: the finite generation algorithm*, in: *Numerical Techniques for Stochastic Optimization Problems*, Y. Ermoliev and R. J-B Wets (eds.), Springer-Verlag Lecture Notes in Control and Information Sciences No. 81, 1987, 545-560.
12. R. T. Rockafellar, "Linear-quadratic programming and optimal control," SIAM J. Control Opt. **25** (1987), 781–814.
13. R. T. Rockafellar and R. J-B Wets, "Generalized linear-quadratic problems of deterministic and stochastic optimal control in discrete time," SIAM J. Control Opt. **320** (1990), 810–822.
14. R. T. Rockafellar, "Computational schemes for solving large-scale problems in extended linear-quadratic programming," Math. Prog., Ser. B **48** (1990), 447–474.

15. C. Zhu and R. T. Rockafellar, "Primal-dual projected gradient algorithms for extended linear-quadratic programming," SIAM J. Optimization, 1993.

16. C. Zhu, "On the primal-dual steepest descent algorithm for extended linear-quadratic programming," preprint.

17. C. Zhu, "Solving large-scale minimax problems with the primal-dual steepest descent algorithm," preprint.

18. H. G. Chen and R. T. Rockafellar, "Forward-backward splitting methods in Lagrangian optimization," SIAM J. Optimization, 1993.