SIAM J. CONTROL AND OPTIMIZATION
Vol. 25, No. 3, May 1987

010

DUAL TECHNIQUES FOR MINIMAX*

WILLIAM W. HAGER† AND DWAYNE L. PRESLER‡

Abstract. A dual formulation for a convex minimax problem is presented and the notion of reducibility is introduced. For nonconvex problems, Rockafellar's augmented Lagrangian is used to close the duality gap. We show how mathematical programming algorithms can be applied to the minimax problem and we develop a special algorithm for reducible minimax problems.

Key words. minimax, duality, augmented Lagrangians

AMS(MOS) subject classifications. 65K05, 90C30

1. Introduction. A dual minimax problem is formulated and the concept of reducibility is introduced. Since the dual problem may not solve the primal problem when the cost lacks convexity, § 3 develops augmented Lagrangian techniques to bridge the duality gap. In particular, an analogue of Rockafellar's augmented Lagrangian is applied to the minimax problem. Abstractly, the minimax problem is a constrained optimization problem with special structure. In § 4 we show how mathematical programming algorithms such as those in [18] can be embedded into an algorithm to solve the minimax problem. Section 5 presents a special scheme for solving reducible minimax problems. Other algorithms that have been proposed for minimax problems include those in [3], [4], [8], [9], [11]-[14], [20], [21], [24], [26] and [34]. Most of these algorithms are either "primal" in nature or the algorithm addresses problems where the maximization phase of the minimax problem is restricted to a finite set. Our methods, on the other hand, are dual methods derived from an augmented Lagrangian and the maximization can be performed over an infinite set. Another approach to minimax problems utilizes nondifferentiable optimization techniques. This family of methods is described in [23] by Kiwiel and in [32] by Shor. As discussed later, perhaps the algorithm [26] of Murray and Overton is the closest to the approach proposed in § 4. Our algorithm for reducible minimax problems seems to be distinct from other algorithms for the minimax problem.

2. Abstract dual. Given sets X and Y and given a real-valued function f defined on $X \times Y$, we consider the problem

$$\underset{x \in X}{\text{minimize}}\ \underset{y \in Y}{\text{maximum}}\, f(x, y). \tag{2.1}$$

In other words, if $\Phi: X \to R$ is the real-valued function defined by

$$\Phi(x) = \text{supremum}\,\{f(x, y): y \in Y\},$$

we are concerned with the problem

$$\text{minimize}\,\{\Phi(x): x \in X\}. \tag{2.2}$$

To derive a dual to (2.1), let us write (2.2) in the form

$$\begin{aligned} &\text{minimize } \rho \\ &\text{subject to } \Phi(x) - \rho \leqq 0, \quad x \in X, \quad \rho \in R. \end{aligned} \tag{2.3}$$

* Received by the editors July 16, 1985; accepted for publication (in revised form) March 27, 1986. This work was supported by National Science Foundation grants MCS-8101892, DMS-8401758 and DMS-8520926 and by Air Force Office of Scientific Research grant AFOSR-ISSA-860091.

† Department of Mathematics, The Pennsylvania State University, University Park, Pennsylvania 16802.
‡ INTER-NATIONAL Research Institute, Newport News, Virginia 23602.

The collection $\mathbf{P}$ of functionals that are nonnegative everywhere on Y is a cone. When g is a functional defined on Y, we write $g \geqq 0$ if $g \in \mathbf{P}$. Similarly the relation $g \leqq 0$ means that $-g \in \mathbf{P}$. Since $f(x, y) - \rho$ is a real-valued function of y for each fixed $x \in X$, and for each fixed $\rho \in R$, (2.3) can be expressed as follows:

$$(2.4) \qquad \text{minimize}\, \{\rho : f(x, \cdot) - \rho\mathbf{1}(\cdot) \leqq 0, x \in X, \rho \in R\}$$

where $\mathbf{1}: Y \to R$ is the function defined by $\mathbf{1}(y) = 1$ for every $y \in Y$. That is, $\mathbf{1}$ is identically equal to one on Y.

Let $\mathbf{Y}$ be a normed vector space consisting of functionals defined on Y and suppose that $\mathbf{Y}$ contains $f(x, \cdot) - \rho\mathbf{1}(\cdot)$ for every $x \in X$ and $\rho \in R$. The dual problem associated with (2.4) involves the dual space $\mathbf{Y}^*$. Given a linear functional $\lambda \in \mathbf{Y}^*$, let $\langle \lambda, p\rangle$ denote the value of λ at p and write $\lambda \geqq 0$ if $\langle \lambda, p\rangle \geqq 0$ for every $p \in \mathbf{P} \cap \mathbf{Y}$. Then the dual to (2.4) is

$$(2.5) \qquad \text{maximize}\, \{\mathbf{l}(\lambda) : \lambda \in \mathbf{Y}^*, \lambda \geqq 0\}$$

where the dual functional $\mathbf{l}$ is defined by

$$(2.6) \qquad \mathbf{l}(\lambda) = \text{infimum}\, \{\rho + \langle \lambda, f(x, \cdot) - \rho\mathbf{1}(\cdot)\rangle : x \in X, \rho \in R\}.$$

Since

$$\rho + \langle \lambda, f(x, \cdot) - \rho\mathbf{1}(\cdot)\rangle = \langle \lambda, f(x, \cdot)\rangle + \rho(1 - \langle \lambda, \mathbf{1}\rangle),$$

it follows that $\mathbf{l}(\lambda)$ is $-\infty$ unless $\langle \lambda, \mathbf{1}\rangle = 1$. Moreover, if $\langle \lambda, \mathbf{1}\rangle = 1$, then the dual functional can be expressed

$$\mathbf{l}(\lambda) = \text{infimum}\, \{\langle \lambda, f(x, \cdot)\rangle : x \in X\}.$$

Now let us consider the standard question in duality theory: When does there exist a solution to (2.5) and when is the maximum in (2.5) equal to the minimum in (2.2)? Applying [17, Thm. A.1], we have the following.

THEOREM 2.1. *Whenever ρ and x are feasible for* (2.3) *and λ is feasible for* (2.5), *we have $\mathbf{l}(\lambda) \leqq \rho$. Moreover, if ρ^* and x^* are feasible for* (2.3), *λ^* is feasible for* (2.5), *and $\mathbf{l}(\lambda^*) = \rho^*$, then ρ^* and x^* are optimal in* (2.3), *λ^* is optimal in* (2.5), *and the complementary slackness condition $\langle \lambda^*, f(x^*, \cdot)\rangle = \rho^*$ holds. On the other hand, suppose that X is a convex subset of a vector space and for each fixed $y \in Y$, $f(x, y)$ is a convex function of $x \in X$. If there exists a ball $\mathbf{B} \subset \mathbf{Y}$ with center at the origin such that*

$$(2.7) \qquad \sup_{\phi \in \mathbf{B}} \inf_{x \in X} \sup_{y \in Y} \{f(x, y) - \phi(y)\} < \infty,$$

then (2.5) *has a solution λ^* and*

$$\mathbf{l}(\lambda^*) = \inf_{x \in X} \sup_{y \in Y} f(x, y).$$

For illustration, suppose that Y is the finite set $\{1, \cdots, m\}$ and let $f_i(x)$ denote the function $f(x, i)$. In this case, both $\mathbf{Y}$ and $\mathbf{Y}^*$ can be identified with R^m, the space of m-tuples of real numbers, and the functional $\langle \cdot, \cdot\rangle$ is the usual dot product in R^m. Hence, the feasibility condition "$\lambda \in \mathbf{Y}^*$, $\lambda \geqq 0$, and $\langle \lambda, \mathbf{1}\rangle = 1$" associated with the dual problem (2.5) is equivalent to saying that $\lambda \in R^m$, $\lambda_i \geqq 0$ for $i = 1, \cdots, m$, and $\lambda_1 + \cdots + \lambda_m = 1$. Observe that in this finite-dimensional framework, assumption (2.7) is satisfied trivially. If both X and the f_i are convex, then Theorem 2.1 tells us that there exists an optimal solution λ^* to (2.5) and

$$\mathbf{l}(\lambda^*) = \inf \left\{ \sum_{i=1}^{m} \lambda_i^* f_i(x) : x \in X \right\} = \inf_{x \in X} \max\, \{f_i(x) : i = 1, \cdots, m\}.$$

For a second illustration, let us consider the case where Y is a compact subset of a normed vector space, $\mathbf{Y}$ is the space of continuous real-valued functions defined on Y, and the norm $\|\cdot\|$ for $\mathbf{Y}$ is defined by

$$\|\phi\| = \text{maximum}\,\{|\phi(y)|: y \in Y\}.$$

Again (2.7) is satisfied trivially. Furthermore, if Y is an interval $[a, b] \subset R$, then $\mathbf{Y}^*$ is the space of functions with bounded variation on $[a, b]$ and the complementary slackness condition can be expressed using a Stieltjes integral:

$$\int_a^b (f(x^*, y) - \rho^*)\, d\lambda^*(y) = 0. \tag{2.8}$$

In the proof of [17, Lemma 5.2], we note that when λ has bounded variation, the inequality $\lambda \geqq 0$ is equivalent to saying that $\lambda(y)$ is a nondecreasing function of y. Hence, if ρ^* and x^* are feasible for (2.3) and λ^* is feasible for (2.5), then $f(x^*, y) - \rho^* \leqq 0$ for every $y \in Y$ and (2.8) implies that $\lambda^*(\cdot)$ is constant on each subinterval of $[a, b]$ where $f(x^*, \cdot) < \rho^*$.

In many applications, one discovers that a solution pair (ρ^*, x^*) for (2.3) has the property that $f(x^*, y) < \rho^*$ except for y in a finite set $\{y_1, \cdots, y_m\} \subset Y$. In this case, the complementary slackness condition (2.8) tells us that $\lambda^*(\cdot)$ is constant on each open interval (y_i, y_{i+1}) and $\lambda^*(y_i^+) \geqq \lambda^*(y_i^-)$. If this jump set is known in advance and if we restrict our attention to λ's which are constant on each interval (y_i, y_{i+1}), then the dual functional can be expressed

$$\mathbf{l}(\lambda) = \inf\left\{\sum_{i=1}^{m} \lambda_i f(x, y_i): x \in X\right\}$$

where $\lambda_i = \lambda(y_i^+) - \lambda(y_i^-)$ is the jump at y_i. Generally, the y_i are unknown and the dual functional must be maximized over both the y_i and the λ_i. To summarize, if Y is the interval $[a, b]$, $\mathbf{Y}$ is the space of continuous real-valued functions defined on $[a, b]$, and there exists both a solution ρ^* and x^* to (2.3) and a solution λ^* to (2.5) such that $\rho^* = \mathbf{l}(\lambda^*)$ and $f(x^*, y) = \rho^*$ for finitely many y, then the continuous dual problem (2.5) can be replaced by the discrete dual

$$\begin{aligned} &\text{maximize } \mathbf{l}(\lambda_1, \cdots, \lambda_N, y_1, \cdots, y_N) \\ &\text{subject to } \lambda_i \geqq 0 \quad \text{and} \quad y_i \in Y \quad \text{for } i = 1, \cdots, N, \quad \sum_{i=1}^{N} \lambda_i = 1, \end{aligned} \tag{2.9}$$

where

$$\mathbf{l}(\lambda_1, \cdots, \lambda_N, y_1, \cdots, y_N) = \inf\left\{\sum_{i=1}^{N} \lambda_i f(x, y_i): x \in X\right\}$$

and where N is any integer greater than or equal to the number of y for which $f(x^*, y) = \rho^*$. The optimal λ_i and y_i in (2.9) correspond to the size and the location of the jumps in λ^*. The discrete dual (2.9) will now be studied in a general setting.

The generalized finite sequence space of Charnes, Cooper and Kortanek (see [5]–[7]) is a natural setting for the discrete dual. Let Λ denote a vector space of real-valued functions defined on Y where $\lambda \in \Lambda$ if and only if $\lambda(y) = 0$ for all but a finite number of y in Y. Instead of writing $\lambda(y)$, we write λ_y and we consider λ_y the yth component of λ. Given $\lambda \in \Lambda$, the collection of y in Y for which $\lambda_y \neq 0$ is the

support of λ. In [5] the space Λ is called a ***generalized finite sequence space.*** Given $\lambda \in \Lambda$, let us define the dual functional

$$\mathrm{l}(\lambda) = \inf\left\{\sum_{y \in Y} \lambda_y f(x, y) \colon x \in X\right\}.$$

The dual problem is

$$\text{maximize}\left\{\mathrm{l}(\lambda) \colon \lambda \in \Lambda, \lambda \geqq 0, \sum_{y \in Y} \lambda_y = 1\right\}. \tag{2.10}$$

THEOREM 2.2. *If λ^* is feasible for* (2.10), *$x^* \in X$, and*

$$\mathrm{l}(\lambda^*) = \sup\{f(x^*, y) \colon y \in Y\}, \tag{2.11}$$

then λ^* is optimal in (2.10), *x^* is optimal in* (2.2), *and $f(x^*, y) = \mathrm{l}(\lambda^*)$ for each y in the support of λ^*. Moreover, letting S denote the support of λ^*, we have*

$$\min_{x \in X} \max_{y \in S} f(x, y) = \min_{x \in X} \max_{y \in Y} f(x, y). \tag{2.12}$$

Proof. If the components of λ are nonnegative and sum to 1 and if $\Gamma \subset Y$ is any finite subset which contains the support of λ, then the following relations hold for any $x \in X$:

$$\begin{aligned} \mathrm{l}(\lambda) &= \inf\left\{\sum_{y \in \Gamma} \lambda_y f(z, y) \colon z \in X\right\} \\ &\leqq \sum_{y \in \Gamma} \lambda_y f(x, y) \\ &\leqq \left(\sum_{y \in \Gamma} \lambda_y\right) \max\{f(x, y) \colon y \in \Gamma\} \\ &= \max\{f(x, y) \colon y \in \Gamma\} \\ &\leqq \sup\{f(x, y) \colon y \in Y\} \\ &= \Phi(x). \end{aligned} \tag{2.13}$$

(The last equality is the definition of $\Phi(x)$.) Hence, $\mathrm{l}(\lambda) \leqq \Phi(x)$ whenever λ is feasible for (2.10) and $x \in X$. Since $\mathrm{l}(\lambda^*) = \Phi(x^*)$ by (2.11), we conclude from (2.13) that λ^* is optimal in (2.10) and x^* is optimal in (2.2). If S is the support of λ^*, then (2.13) also tells us that

$$\mathrm{l}(\lambda^*) \leqq \sum_{y \in S} \lambda_y^* f(x^*, y) \leqq \Phi(x^*). \tag{2.14}$$

But $\mathrm{l}(\lambda^*) = \Phi(x^*)$ and the inequalities in (2.14) are equalities. Since $\lambda_y^* > 0$ and $f(x^*, y) \leqq \Phi(x^*)$ for every $y \in S$ and since the compoments of λ^* sum to one, it follows from (2.14) that

$$\mathrm{l}(\lambda^*) = \Phi(x^*) = f(x^*, y) \quad \text{for every } y \in S. \tag{2.15}$$

Finally, let us consider (2.12). Relation (2.13) implies that

$$\mathrm{l}(\lambda^*) \leqq \max\{f(x, y) \colon y \in S\} \tag{2.16}$$

for every x in X. Taking the infimum over $x \in X$, (2.16) tells us that

$$\mathrm{l}(\lambda^*) \leqq \inf_{x \in X} \max_{y \in S} f(x, y). \tag{2.17}$$

Combining (2.15) and (2.17), we have

$$\Phi(x^*) = \max_{y\in S} f(x^*, y) = \min_{x\in X} \max_{y\in S} f(x, y),$$

which completes the proof of (2.12). □

For any $S \subset Y$, we have the trivial relation

$$\min_{x\in X} \max_{y\in S} f(x, y) \leqq \min_{x\in X} \max_{y\in Y} f(x, y).$$

Theorem 2.2 tells us that if the value of the dual problem (2.10) is equal to the value of the primal problem (2.2), then there exists a finite set $S \subset Y$ such that (2.12) holds. Therefore, one strategy for solving the mimimax problem (2.1) is to start with a finite set S and adjust it until (2.12) is satisfied. The simplified problem

$$\underset{x\in X}{\text{minimize}}\ \underset{y\in S}{\text{maximum}}\, f(x, y)$$

is often easier to solve than the original problem (2.1). This idea is developed further in § 5. When there exists a finite set S satisfying (2.12), we say that the minimax problem is *reducible.*

How often is a minimax problem reducible? Let us consider the following example:

$$\underset{x\in R}{\text{minimize}}\ \underset{y\in R}{\text{maximum}}\ \beta x^2 + 2\alpha xy + y^2. \tag{2.18}$$

Since the maximum is attained at $y = \alpha x$, the function Φ is given by

$$\Phi(x) = (\alpha^2 + \beta)x^2.$$

When $\alpha^2 + \beta$ is nonnegative, $\Phi(x)$ achieves its minimum at $x = 0$ and when x is 0, the maximizing value of y in (2.18) is $y = 0$. On the other hand, suppose that $S = \{0\}$ and let us consider the restricted minimax problem

$$\underset{x\in R}{\text{minimize}}\ \underset{y=0}{\text{maximum}}\ \beta x^2 + 2\alpha xy - y^2,$$

which is equivalent to

$$\underset{x\in R}{\text{minimize}}\ \beta x^2.$$

For $\beta \geqq 0$, the minimum is attained at $x = 0$ while for $\beta < 0$, βx^2 has no minimum. In summary, for $\beta < -\alpha^2$, Φ has no minimum. For $\beta \in [-\alpha^2, 0)$, no set S satisfies (2.12). And for $\beta \geqq 0$, the minimax problem is reducible with $S = \{0\}$.

Now suppose that $f: R^2 \to R$ is an arbitrary twice continuously differentiable function. We assume that there exists a solution x^* to the minimax problem (2.1) and there exists y^* such that

$$f(x^*, y^*) = \text{maximum}\, \{f(x^*, y) \colon y \in R\}$$

and

$$\frac{\partial^2 f}{\partial y^2}(x^*, y^*) < 0. \tag{2.19}$$

Then for x in a neighborhood of x^*, the implicit function theorem gives us a differentiable function $y(\cdot)$ such that $y(x^*) = y^*$ and

$$\frac{\partial f}{\partial y}(x, y(x)) = 0. \tag{2.20}$$

By (2.19), $y(x)$ is a local maximizer of $f(x,\cdot)$ for x near x^*. Let us assume that $y(x)$ is the global maximizer of $f(x,\cdot)$. Since x^* minimizes $\Phi(x)$, we know that $\Phi''(x^*)\geqq 0$. Applying the chain rule to $\Phi(x)=f(x, y(x))$ and utilizing (2.20), it can be shown that

$$\frac{\partial^2\Phi}{\partial x^2}(x^*)=\frac{\partial^2 f}{\partial x^2}(x^*,y^*)-\frac{((\partial^2 f/\partial x\,\partial y)(x^*,y^*))^2}{(\partial^2 f/\partial y^2)(x^*,y^*)}. \tag{2.21}$$

If X is restricted to a neighborhood of x^*, then (2.12) holds for $S=\{y^*\}$ provided

$$\frac{\partial^2 f}{\partial x^2}(x^*,y^*)>0. \tag{2.22}$$

On the other hand, by (2.21), the inequality $\Phi''(x^*)\geqq 0$ only guarantees that

$$\frac{\partial^2 f}{\partial x^2}(x^*,y^*)\geqq\frac{((\partial^2 f/\partial x\,\partial y)(x^*,y^*))^2}{(\partial^2 f/\partial y^2)(x^*,y^*)}. \tag{2.23}$$

In other words, relation (2.22) is sufficient for the minimax problem to be reducible when X is a neighborhood of x^* while the fact that x^* minimizes Φ only implies (2.23).

Returning to example (2.18), observe that the range of α and β for which the minimax problem is reducible is larger than the range for which the minimax problem is not reducible. Consequently, if α and β are chosen randomly and if the minimax problem (2.18) has a solution, then the problem is probably reducible. The problem (see [19]) of optimally coating a surface to minimize the maximum reflection associated with incoming waves is an example of a very complicated nonconvex problem that is reducible even though $\mathrm{l}(\lambda^*)<\Phi(x^*)$. Based on these observations, we feel that algorithms which search for a set S satisfying (2.12) will apply to a broad class of problems.

Although a general minimax problem is not necessarily reducible, a convex finite-dimensional minimax problem is always reducible. Dem'yanov and Malozemov [13] establish this fact in the following setting: X is a closed convex subset of R^n, Y is a compact subset of R^m, $f(x,y)$ is continuous and continuously differentiable with respect to x on $\overset{\circ}{X}\times Y$ where $\overset{\circ}{X}$ is an open set containing X, $f(\cdot,y)$ is convex for each $y\in Y$, and there exists a solution to (2.1). One can also establish (2.12) under weaker assumptions using Clarke's result [10, Thm. 2.1] and properties of subgradients found in Rockafellar's book [29]. The analysis of Charnes, Cooper and Kortanek [7] also appears applicable. For completeness, we now derive (2.12) using Theorem 2.1 and results from [29]. First, let us consider the case where Y is a finite set.

LEMMA 2.3. *Suppose that Y is a finite set, X is a nonempty convex subset of R^n, and $f(\cdot,y)$ is convex for each $y\in Y$. If there exists a solution x^* to the primal problem* (2.2), *then there exists a solution λ^* to the dual problem* (2.10) *and* $\mathrm{l}(\lambda^*)=\Phi(x^*)$. *Moreover, λ^* can be chosen so that its support has at most $n+1$ elements.*

(It follows from Theorem 2.2 that under the hypotheses of Lemma 2.3, (2.12) holds for some set S which has at most $n+1$ elements.)

Proof. By Theorem 2.1 and by the observations that follow the theorem, there exists a solution λ^* to the dual problem (2.10) and $\mathrm{l}(\lambda^*)=\Phi(x^*)$. Let m denote the number of elements in Y and assume for convenience that $Y=\{1,\cdots,m\}$. The equality $\mathrm{l}(\lambda^*)=\Phi(x^*)$ combined with (2.13) tell us that

$$\mathrm{l}(\lambda^*)=\underset{x\in X}{\text{minimum}}\sum_{i=1}^{m}\lambda_i^* f_i(x)=\sum_{i=1}^{m}\lambda_i^* f_i(x^*) \tag{2.24}$$

where $f_i(\cdot)$ denotes $f(\cdot,i)$. Let $\partial f_i(x)$ denote the collection of subgradients of f_i at x.

By [29, Thm. 27.4] and by (2.24), there exists $g_i \in \partial f_i(x^*)$ such that

$$\left\langle \sum_{i=1}^{m} \lambda_i^* g_i,\, x - x^* \right\rangle \geqq 0 \quad \text{for every } x \in X. \tag{2.25}$$

Define the set $P = \{i \in [1, m]\colon \lambda_i^* > 0\}$. Since $\lambda^* \geqq 0$ and $\lambda_1^* + \cdots + \lambda_m^* = 1$, it follows from [29, Thm. 17.1] that there exists nonnegative scalars μ_i for $i \in P$ such that the support of μ has at most $n+1$ elements,

$$\sum_{i \in P} \mu_i g_i = \sum_{i=1}^{m} \lambda_i^* g_i \quad \text{and} \quad \sum_{i \in P} \mu_i = 1.$$

Hence, (2.25) yields

$$\left\langle \sum_{i \in P} \mu_i g_i,\, x - x^* \right\rangle \geqq 0 \quad \text{for every } x \in X. \tag{2.26}$$

Again by [29, Thm. 27.4] and by (2.26), we have

$$\sum_{i \in P} \mu_i f_i(x^*) = \underset{x \in X}{\text{minimum}} \sum_{i \in P} \mu_i f_i(x). \tag{2.27}$$

The identity $\mathbf{l}(\lambda^*) = \Phi(x^*)$ combined with Theorem 2.2 imply that $f_i(x^*) = \Phi(x^*)$ for every $i \in P$. Since the μ_i sum to one, it follows from (2.27) that

$$\Phi(x^*) = \underset{x \in X}{\text{minimum}} \sum_{i \in P} \mu_i f_i(x) = \mathbf{l}(\mu).$$

By Theorem 2.2, μ is a solution to the dual problem. □

THEOREM 2.4. *Suppose that Y is a nonempty compact subset of a normed space, X is a nonempty compact, convex subset of R^n, and f is a real-valued function defined on $\mathring{X} \times Y$ where $\mathring{X}$ is a relatively open set containing X. If $f(\cdot, y)$ is convex and lower semicontinuous for each y in Y and $f(x, \cdot)$ is continuous for each x in $\mathring{X}$, then there exists a solution x^* to* (2.2), *there exists a solution λ^* to* (2.10), *and* $\mathbf{l}(\lambda^*) = \Phi(x^*)$. *Moreover, λ^* can be chosen so that its support has at most $n+1$ elements.*

Proof. Let $\{y_1, y_2, \cdots\}$ be a dense subset of Y and define $\Phi^N\colon X \to R$ by

$$\Phi^N(x) = \text{maximum}\, \{f(x, y_i)\colon i = 1, \cdots, N\}.$$

Since $f(\cdot, y)$ is lower semicontinuous, Φ^N is lower semicontinuous. Thus the compactness of X guarantees the existence of $x^N \in X$ such that

$$\Phi^N(x^N) = \underset{x \in X}{\text{minimum}}\, \Phi^N(x).$$

In addition, the compactness of X implies that a subsequence of the x^N converges to some $x^* \in X$. By [29, Thm. 10.8] and the assumption that $f(x, \cdot)$ is continuous and $\{y_1, y_2, \cdots\}$ is a dense subset of Y, Φ^N converges to Φ uniformly on X. Consequently, we have

$$\Phi(x^*) = \text{minimum}\, \{\Phi(x)\colon x \in X\}.$$

Referring to Lemma 2.3, there exists a set $Y^N \subset \{y_1, \cdots, y_N\}$ where Y^N has at most $n+1$ elements and there exists a corresponding set of nonnegative scalars $\{\lambda_y^N\colon y \in Y^N\}$ which sum to one and which satisfy the relation

$$\underset{x \in X}{\text{minimum}} \sum_{y \in Y^N} \lambda_y^N f(x, y) = \Phi^N(x^N). \tag{2.28}$$

Since the sets Y^N and the scalars λ_y^N lie in compact sets, we can extract convergent subsequences. Assume for convenience that x^N converges to x^*, λ^N converges to λ^* and Y^N converges to Y^*. For any $x \in X$,

$$\lim_{N\to\infty} \sum_{y\in Y^N} \lambda_y^N f(x, y) = \sum_{y\in Y^*} \lambda_y^* f(x, y) \tag{2.29}$$

since $f(x, \cdot)$ is continuous. Minimizing the left side of (2.29) over $x \in X$ yields

$$\lim_{N\to\infty} \mathbf{l}(\lambda^N) \leqq \sum_{y\in Y^*} \lambda_y^* f(x, y)$$

for every $x \in X$. Since $\mathbf{l}(\lambda^N)$ is equal to $\Phi(x^N)$ and since $\Phi(x^N)$ approaches $\Phi(x^*)$ as N tends to infinity, we conclude that

$$\Phi(x^*) \leqq \sum_{y\in Y^*} \lambda_y^* f(x, y) \tag{2.30}$$

for every $x \in X$. Minimizing the right side of (2.30) over x in X yields the relation $\Phi(x^*) \leqq \mathbf{l}(\lambda^*)$ and by (2.13), λ^* is a solution to the dual problem (2.10). □

In Theorem 2.4, we can replace the assumption that X is compact with the assumption that X is closed, however, the existence of x^* is lost.

COROLLARY 2.5. *Suppose that Y is a nonempty compact subset of a normed space, X is a nonempty closed, convex subset of R^n and f is a real-valued function defined on $\overset{\circ}{X} \times Y$ where $\overset{\circ}{X}$ is a relatively open set containing X. If $f(\cdot, y)$ is convex and lower semicontinuous for each y in Y and $f(x, \cdot)$ is continuous for each x in $\overset{\circ}{X}$, then there exists a solution λ^* to* (2.10), *and*

$$\mathbf{l}(\lambda^*) = \inf_{x\in X} \Phi(x).$$

Moreover, λ^* can be chosen so that its support has at most $n+1$ elements.

Proof. Given an integer N, define the set

$$X^N = \{x \in X : \|x\| \leqq N\}$$

where $\|\cdot\|$ is any norm for R^n. We assume that N is large enough that X^N is nonempty. By Theorem 2.4, there exists $\lambda^N \in \Lambda$ such that

$$\mathbf{l}^N(\lambda^N) = \underset{x\in X^N}{\text{minimum}}\, \Phi(x)$$

where $\mathbf{l}^N$ is defined by

$$\mathbf{l}^N(\lambda) = \underset{x\in X^N}{\text{minimum}} \sum_{y\in Y} \lambda_y f(x, y)$$

and where the support Y^N of λ^N has at most $n+1$ elements. From (2.10), the components of λ^N are nonnegative and sum to one. Since Y^N and the components of λ^N lie in compact sets, there exists a subsequence of the λ^N which converges to some λ^*. Let Y^* denote the support of λ^* and assume for convenience that the entire sequence $\{\lambda^N\}$ converges to λ^*. For each $x \in X$, the continuity of $f(x, \cdot)$ implies that

$$\lim_{N\to\infty} \sum_{y\in Y^N} \lambda_y^N f(x, y) = \sum_{y\in Y^*} \lambda_y^* f(x, y). \tag{2.31}$$

We minimize the left side of (2.31) over $x \in X^N$ to obtain

$$\lim_{N\to\infty} \mathbf{l}^N(\lambda^N) \leqq \sum_{y\in Y^*} \lambda_y^* f(x, y). \tag{2.32}$$

Minimizing the right side of (2.32) over $x \in X$ yields

$$\mathbf{l}(\lambda^*) \geqq \lim_{N\to\infty} \mathbf{l}^N(\lambda^N),$$

and since $\Phi(x^N)$ is equal to $\mathbf{l}^N(\lambda^N)$, we have

$$\mathbf{l}(\lambda^*) \geqq \lim_{N\to\infty} \Phi(x^N) = \inf_{x\in X} \Phi(x).$$

Finally, (2.13) tells us that λ^* is a solution to the dual problem (2.10). □

3. Augmented Lagrangians. A nonconvex problem often has a duality gap and the value of the dual problem is strictly less than the value of the primal problem. A strategy for bridging this gap emanates from work of Arrow and Solow [1], Hestenes [22], and Powell [28]. The basic idea is to augment the ordinary Lagrangian with a penalty term. To introduce this penalized dual approach, we first consider a finite-dimensional mathematical program with equality constraints:

$$\text{(3.1)} \qquad \begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } h(x) = 0, \qquad x \in R^n \end{aligned}$$

where $f\colon R^n \to R$ and $h\colon R^n \to R^m$. (Mathematical programs in a Hilbert space setting are studied in [16] and [27].) The ordinary Lagrangian corresponding to (3.1) is $\mathbf{l}(\lambda, x) = f(x) + \lambda^T h(x)$. Letting r be a positive scalar and letting $|\cdot|$ denote the Euclidean norm, the augmented Lagrangian corresponding to the penalty term $r|h(x)|^2$ is

$$\text{(3.2)} \qquad \mathbf{L}(\lambda, x) = f(x) + \lambda^T h(x) + r|h(x)|^2.$$

To illustrate the type of results that can be proved about the augmented Lagrangian, we state the following theorem which is extracted from Bertsekas [2]. In stating this theorem, our convention is that the gradient ∇ is a row vector and the gradient ∇h of the vector valued function h is a $m \times n$ matrix with ith row ∇h_i for $i = 1$ to m. Also, we let ∇^2 denote the Hessian matrix of second partial derivatives and the phrase "x^* is a local minimizer for (3.1)" means that $h(x^*) = 0$ and $f(x^*) \leqq f(x)$ whenever $h(x) = 0$ and x is near x^*.

THEOREM 3.1. *Suppose that x^* is a local minimizer for* (3.1), *both f and h are twice continuously differentiable in a neighborhood of x^*, and the rows of $\nabla h(x^*)$ are linearly independent. If $\lambda = \lambda^*$ is the solution to the equation*

$$\text{(3.3)} \qquad \nabla f(x^*) + \lambda^T \nabla h(x^*) = 0$$

and $\nabla_x^2 \mathbf{l}(\lambda^*, x^*)$ is positive definite in the null space of $\nabla h(x^*)$, then there exists a parameter s and a neighborhood $\mathbf{N}$ *of x^* such that the problem*

$$\text{minimize } \{\mathbf{L}(\mu, x)\colon x \in \mathbf{N}\}$$

has a unique minimizer $x_{\mu,r}$ whenever $r \geqq s$ and $|\lambda^* - \mu| \leqq r/s$. Moreover, there exists a constant c, independent of r and μ, such that

$$\text{(3.4)} \qquad |x_{\mu,r} - x^*| + |\lambda_{\mu,r} - \lambda^*| \leqq c|\mu - \lambda^*|/r$$

where $\lambda_{\mu,r} := \mu + 2rh(x_{\mu,r})$.

Now let us consider the inequality constrained problem

$$\text{(3.5)} \qquad \begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } g(x) \leqq 0, \qquad x \in R^n \end{aligned}$$

where $g\colon R^n \to R^l$. Rockafellar's augmented Lagrangian (see [30]) is obtained by converting the inequality constraints to equality constraints using slack variables, forming the augmented Lagrangian corresponding to these equality constraints, and minimizing over the slack variables to obtain

$$\mathbf{L}(\lambda, x) = f(x) + \sum_{i\in I_+} (\lambda_i g_i(x) + r g_i(x)^2) - \frac{1}{4r} \sum_{i\in I_-} \lambda_i^2 \tag{3.6}$$

where the sets I_+ and I_- are defined by

$$I_+ = \{i \in [1, l]: 2rg_i(x) + \lambda_i \geqq 0\} \quad \text{and} \quad I_- = \{i \in [1, l]: 2rg_i(x) + \lambda_i < 0\}.$$

Thus the part of the Lagrangian (3.6) corresponding to indices $i \in I_+$ resembles the equality Lagrangian (3.2) while the part of the Lagrangian corresponding to indices $i \in I_-$ is locally independent of x. Theorem 3.1 also applies to inequality constrained problems since an inequality can be converted to an equality using Valentine's device (see [2] and [33]).

Augmented Lagrangians are now applied to the minimax problem (2.1). Let us consider the case where the set Y connected with the maximization is the integers $\{1, 2, \cdots\}$ and at some solution x^* to (2.2), we have

$$f(x^*, i) \geqq f(x^*, i+1)$$

for each i. We assume that $f(x^*, i) = \Phi(x^*)$ for $i = 1, \cdots, N$ while $f(x^*, i) < \Phi(x^*)$ for $i > N$. Recall from § 2 that the set $S = \{1, \cdots, N\}$ is usually the support of a dual multiplier λ^*. If a good estimate for x^* is known, then S is known and, at least locally, x^* is a solution to the equality constrained problem

$$\text{minimize}\,\{\rho: f(x) = \rho\mathbf{1}, x \in X, \rho \in R\} \tag{3.7}$$

where $\mathbf{1}$ denotes the vector in R^N with every component equal to 1 and $f(x)$ denotes the vector-valued function with ith component $f_i(x)$ equal to $f(x, i)$ for $i = 1, \cdots, N$. The corresponding augmented Lagrangian is

$$\mathbf{L}(\lambda, x, \rho) = \rho + \lambda^T(f(x) - \rho\mathbf{1}) + r|f(x) - \rho\mathbf{1}|^2. \tag{3.8}$$

In practice, the support set S for the minimax problem is not known, and we must use the inequality constrained formulation

$$\text{minimize}\,\{\rho: f(x) \leqq \rho\mathbf{1}, x \in X, \rho \in R\}.$$

Since this formulation is equivalent to

$$\text{minimize}\,\{\rho: f(x) + z = \rho\mathbf{1}, z \geqq 0, x \in X, \rho \in R, z \in R^N\}, \tag{3.9}$$

the analogue of Rockafellar's augmented Lagrangian is

$$\mathbf{L}(\lambda, x, \rho) = \text{minimum}\,\{\rho + \lambda^T(f(x) + z - \rho\mathbf{1}) + r|f(x) + z - \rho\mathbf{1}|^2: z \geqq 0, z \in R^N\}. \tag{3.10}$$

Since the extremand in (3.10) is a strictly convex function of z, there exists a unique z which attains the minimum, and by (3.6), the augmented Lagrangian (3.10) can be expressed:

$$\mathbf{L}(\lambda, x, \rho) = \rho + \sum_{i\in I_+} \{\lambda_i(f_i(x) - \rho) + r(f_i(x) - \rho)^2\} + \frac{1}{4r} \sum_{i\in I_-} \lambda_i^2$$

where

$$I_+ = \{i \in [1, N]: 2r(f_i(x) - \rho) + \lambda_i \geqq 0\} \quad \text{and} \quad I_- = \{i \in [1, N]: 2r(f_i(x) - \rho) + \lambda_i < 0\}.$$

When using an augmented Lagrangian to solve the constrained optimization problem (3.9), we minimize $\mathbf{L}(\lambda, x, \rho)$ over x in X and ρ in R to obtain the dual functional $\mathbf{L}(\lambda)$. Then the dual functional is maximized over λ to obtain a solution λ^* to the dual problem

$$\text{maximize } \{\mathbf{L}(\lambda)\colon \lambda \in R^N\}.$$

As we will show shortly, the minimum of $\mathbf{L}(\lambda, x, \rho)$ over ρ can be computed explicitly. Let $\mathbf{L}(\lambda, x)$ denote the partly minimized functional defined by

$$\mathbf{L}(\lambda, x) = \text{minimum } \{\mathbf{L}(\lambda, x, \rho)\colon \rho \in R\}. \tag{3.11}$$

LEMMA 3.2. *There exists a unique ρ, which attains the minimum in* (3.11).

Proof. Defining the parameter

$$\rho_1 = \text{maximum } \{f_i(x) + \lambda_i/2r\colon i = 1, \cdots, N\},$$

observe that $\mathbf{L}(\lambda, x, \rho) = \rho$ plus a constant (independent of ρ) for $\rho \geqq \rho_1$. Thus the minimum of $\mathbf{L}(\lambda, x, \cdot)$ occurs on the interval $[-\infty, \rho_1]$. By [17, Cor. A6], the derivative of $\mathbf{L}(\lambda, x, \rho)$ with respect to ρ is a Lipschitz continuous function of ρ on bounded intervals. Since the second derivative of $\mathbf{L}(\lambda, x, \cdot)$ is at least $2r$ on $(-\infty, \rho_1]$, we conclude that $\mathbf{L}(\lambda, x, \cdot)$ is strictly convex on $(-\infty, \rho_1]$ and there exists a unique minimum. □

To compute the minimum for $\mathbf{L}(\lambda, x, \cdot)$, we define the parameters

$$\rho_i = f_i(x) + \lambda_i/2r$$

for $i = 1, \cdots, N$ and we reindex the components of f and λ so that

$$\rho_1 \geqq \rho_2 \geqq \cdots \geqq \rho_N.$$

Since $\mathbf{L}(\lambda, x, \cdot)$ is strictly convex on $(-\infty, \rho_1]$, the derivative of $\mathbf{L}(\lambda, x, \cdot)$ is monotone increasing (with slope at least $2r$). For ρ between ρ_{j+1} and ρ_j, the derivative of $\mathbf{L}(\lambda, x, \cdot)$ is given by

$$\frac{d}{d\rho}\mathbf{L}(\lambda, x, \rho) = 1 + 2rj\rho - \sum_{i=1}^{j} (\lambda_i + 2rf_i(x)). \tag{3.12}$$

With the convention that ρ_{N+1} is $-\infty$, there exists an interval $[\rho_{j+1}, \rho_j]$ where the derivative changes sign. Since $\mathbf{L}(\lambda, x, \cdot)$ is a quadratic on this interval, the minimizer ρ^* of the quadratic is easily evaluated:

$$\rho^* = \frac{-1 + \sum_{i=1}^{j} (\lambda_i + 2rf_i(x))}{2rj}.$$

Since the computer time to sort $f_i(x) + \lambda_i/2r$ into decreasing order is proportional to $N \log_2 N$ (see [25]) while the time to evaluate the derivative (3.12) for $\rho = \rho_1$ through $\rho = \rho_N$ is proportional to N, the computer time required to minimize $\mathbf{L}(\lambda, x, \cdot)$ is proportional to $N \log_2 N$.

4. General minimax problems.

Now let us return to the minimax problem

$$\underset{x \in X}{\text{minimize}}\ \underset{y \in Y}{\text{maximum}}\, f(x, y) \tag{4.1}$$

where f is a real-valued function defined on $X \times Y$. As demonstrated in the proof of Theorem 2.4, one method to solve the minimax problem is to introduce a set $Y^N \subset Y$ with N elements and to consider the approximation

$$\underset{x \in X}{\text{minimize}}\ \underset{y \in Y^N}{\text{maximum}}\, f(x, y).$$

If $x^N \in X$ has the property that

$$\underset{y \in Y^N}{\text{maximum}}\, f(x^N, y) = \underset{x \in X}{\text{minimum}}\ \underset{y \in Y^N}{\text{maximum}}\, f(x, y)$$

where $Y^1 \subset Y^2 \subset Y^3 \subset \cdots$ and the union of Y^N over N is a dense subset of Y, then under the hypotheses of Theorem 2.4, every convergent subsequence of $\{x^N\}$ approaches a solution to (4.1). Moreover, defining $\Phi^N : X \to R$ by

$$\Phi^N(x) = \underset{y \in Y^N}{\text{maximum}}\, f(x, y),$$

Dem'yanov and Malozemov [13] show that if x^N is an extreme point of Φ^N, then every convergent subsequence of $\{x^N\}$ approaches an extreme point of the function

$$\Phi(x) = \underset{y \in Y}{\text{maximum}}\, f(x, y).$$

Their assumptions are that $X \subset R^n$, $Y \subset R^m$, Y is compact, X is closed and convex, and $f(x, y)$ is continuous and continuously differentiable with respect to x on $\overset{\circ}{X} \times Y$ where $\overset{\circ}{X}$ is an open set containing X.

The principal difficulty involved with primal algorithms for minimax problems is that the function Φ is almost always nondifferentiable at its minimum. Ways to circumvent this lack of smoothness are developed in the algorithms of Dem'yanov [12] and others. Unlike the primal function, the dual function is usually smooth at the solution to the dual problem. Nonetheless, as we now show, the dual problem can be ill conditioned and algorithms for solving the dual problem must deal with this conditioning. In describing the ill conditioning associated with the dual problem, we assume for simplicity that X is R^n. The dual functional corresponding to Y^N is

$$\mathbf{l}(\lambda) = \underset{x \in X}{\text{infimum}}\, \mathbf{l}(\lambda, x) \tag{4.2}$$

where the Lagrangian $\mathbf{l}$: $R^N \times R^n \to R$ is defined by

$$\mathbf{l}(\lambda, x) = \sum_{y \in Y^N} \lambda_y f(x, y).$$

Suppose that $x = x^*$ attains the minimum in (4.2) when $\lambda = \lambda^*$, that $f(x, y)$ is twice continuously differentiable with respect to x for every $y \in Y^N$, and that the Hessian

$$\sum_{y \in Y^N} \lambda_y^* \nabla_x^2 f(x^*, y) \tag{4.3}$$

is positive definite. Since x^* attains the minimum in (4.2) when $\lambda = \lambda^*$, the gradient of the Lagrangian with respect to x is zero at $x = x^*$: $\nabla_x \mathbf{l}(\lambda^*, x^*) = 0$. Since the Hessian (4.3) is nonsingular, the implicit function theorem tells us that for λ near λ^*, there exists an $x(\lambda)$ that satisfies the equation

$$\nabla_x \mathbf{l}(\lambda, x(\lambda)) = 0, \tag{4.4}$$

and by the second order sufficiency condition, $x(\lambda)$ is a local minimizer for $\mathbf{l}(\lambda, \cdot)$. Let us assume that $x(\lambda)$ is also a global minimizer for $\mathbf{l}(\lambda, \cdot)$. By the chain rule and (4.4), we have

$$\frac{\partial \mathbf{l}}{\partial \lambda_y}(\lambda) = f(x(\lambda), y) + \nabla_x \mathbf{l}(\lambda, x(\lambda)) \frac{\partial x}{\partial \lambda_y}(x(\lambda), \lambda) = f(x(\lambda), y). \tag{4.5}$$

Differentiating (4.4) with respect to λ_z yields

$$\frac{\partial x}{\partial \lambda_z}(\lambda) = -\nabla_x^2 \mathbf{l}(\lambda, x(\lambda))^{-1} \nabla_x f(x(\lambda), z)^T,$$

and differentiating (4.5) with respect to λ_z gives us

$$\frac{\partial^2 \mathbf{l}}{\partial \lambda_y\, \partial \lambda_z}(\lambda) = -\nabla_x f(x(\lambda), y) \nabla_x^2 \mathbf{l}(\lambda, x(\lambda))^{-1} \nabla_x f(x(\lambda), z)^T.$$

Hence, the Hessian of the dual functional has the form

$$\frac{\partial^2 \mathbf{l}}{\partial \lambda^2}(\lambda) = -GF^{-1}G^T$$

where F is the $n \times n$ matrix given by

$$F = \sum_{y \in Y^N} \lambda_y \nabla_x^2 f(x(\lambda), y)$$

and G is the $N \times n$ matrix whose yth row is $\nabla_x f(x(\lambda), y)$ for $y \in Y^N$.

Remember that if λ is feasible in the dual problem, then the components of λ sum to one. If P is a $(N-1) \times N$ matrix whose rows are a basis in R^N for the space orthogonal to the vector with every component equal to one, then the convergence speed of steepest ascent applied to the dual functional is related to the distribution of eigenvalues for the Hessian of $\mathbf{l}$ evaluated in the row space of P. The Hessian of $\mathbf{l}(P^T\mu)$ with respect to μ is given by $-(PG)F^{-1}(PG)^T$. Since PG is $(N-1)\times n$ while the Hessian of $\mathbf{l}$ with respect to μ is $(N-1)\times(N-1)$, we conclude that the Hessian is singular whenever $N-1$ is greater than n, or equivalently, whenever N is greater than $n+1$.

Now consider the strategy of Theorem 2.4 where we introduce a set $Y^N \subset Y$ for which

$$\lim_{N \to \infty} = \inf \{|y-z| : z \in Y^N\} = 0$$

for every $y \in Y$. By its structure, the Hessian of $\mathbf{l}(\lambda)$ is singular whenever the support of λ has more than $n+1$ elements. The convergence speed of numerical schemes (like steepest ascent) for solving the dual problem is governed by the ratio between the absolute largest eigenvalue and the absolute smallest eigenvalue of the Hessian, and as the ratio tends to infinity, the convergence speed approaches zero. If the support of λ has more than $n+1$ elements, then the smallest eigenvalue is zero, the ratio is infinity, and convergence is slow. In other words, asymptotically, it is impractical to maximize the dual functional using say steepest descent (or almost any standard algorithm) when N is large.

The augmented Lagrangian is subject to similar instabilities. For the inequality constrained problem (3.5) and the augmented Lagrangian (3.6), Rockafellar shows in an appropriate setting (see [31]) that

$$\frac{\partial^2 \mathbf{L}}{\partial \lambda_+^2}(\lambda^*) = -\nabla g_+(x^*) F_r^{-1} \nabla g_+(x^*)^T, \quad \frac{\partial^2 \mathbf{L}}{\partial \lambda_-^2}(\lambda^*) = -\frac{1}{2r} I, \quad \frac{\partial^2 \mathbf{L}}{\partial \lambda_+ \partial \lambda_-}(\lambda^*) = 0 \tag{4.6}$$

where $\mathbf{L}(\lambda) = \inf\{\mathbf{L}(\lambda, x) : x \in R^n\}$, $\lambda_\pm$ and $g_\pm$ denote the components of λ and g corresponding to indices $i \in I_\pm$, and

$$F_r = \nabla^2 f(x^*) + \sum_{i \in I_+} (\lambda_i^* \nabla^2 g_i(x^*) + 2r \nabla g_i(x^*)^T \nabla g_i(x^*)).$$

Hence, the Hessian (4.6) is singular when the number of elements in I_+ is greater than n.

Recall that the minimax problem corresponding to Y^N can be written as the inequality constrained problem

$$\begin{aligned} &\text{minimize } \rho \\ &\text{subject to } x \in X, \quad \rho \in R, \quad f(x, y) \leqq \rho \quad \text{for every } y \in Y^N. \end{aligned} \tag{4.7}$$

Thus the yth component of g in (3.5) is identified with $f(x, y) - \rho$. Since the independent variables in (4.7) are x and ρ, the primal problem (4.7) is formulated in R^{n+1} when $X \subset R^n$ and the Hessian

$$\frac{\partial^2 \mathbf{L}}{\partial \lambda^2}(\lambda^*)$$

is singular when the number of elements in I_+ is greater than $n+1$. Observe that the augmented Lagrangian is better conditioned than the ordinary Lagrangian, since the part of the Hessian corresponding to the second partial derivative with respect to λ_- is a multiple of the identity matrix which is perfectly conditioned. Nonetheless, as N grows, the Hessian can still become singular.

Now let us develop an algorithm to solve the minimax problem. Given $x \in X$, let $y_1(x)$, $y_2(x), \cdots$ denote the local maxima of $f(x, \cdot)$ on Y. Our algorithm for solving the minimax problem has two phases. In both phases, we utilize the inquality formulation (4.7). However, in *phase one*, Y^N is a fixed set $\{y_1, \cdots, y_N\}$ contained in Y and N is "large." In *phase two*, Y^N has the form $\{y_1(x), \cdots, y_N(x)\}$ and N is "small." If $f(x, \cdot)$ has a finite number of local maxima on Y, then the phase two problem

$$\text{(4.8)} \qquad \begin{aligned} &\text{minimize } \rho \\ &\text{subject to } x \in X, \quad \rho \in R, \quad f(x, y_i(x)) \leqq \rho \quad \text{for } i = 1, \cdots, N \end{aligned}$$

is usually equivalent to (4.1) for N sufficiently large. Since (4.8) involves tracking the peaks $y_i(\cdot)$, solving (4.8) is more difficult than solving (4.7). Hence, phase two should only be activated when the algorithm applied to (4.8) converges rapidly. For many mathematical programming algorithms, rapid convergence only occurs in a *neighborhood* of an optimum. For this reason, it is more efficient to apply an unsophisticated algorithm to the ill conditioned problem (4.7) generating a starting guess for a fast algorithm that solves (4.8).

Let us now show in detail how an algorithm such as [18, Algorithm 5.2] or any other algorithm with similar structure can be used to solve either (4.7) or (4.8). Each iteration of Algorithm 5.2 has the following steps: A restoration step where the equality and binding inequality constraints are partially satisfied, a multiplier update where an improved approximation to the optimal dual multipliers is generated, an unconstrained minimization step where the augmented Lagrangian is minimized using (for example) several preconditioned conjugate gradient iterations, and an adjustment to the penalty when a minimizer of the augmented Lagrangian has essentially been computed. This algorithm monitors the convergence of the iterations to a Kuhn-Tucker point and typically, both the restoration step and the multiplier update are only activated in a neighborhood of an optimum. In other words, unless the iterations are in a neighborhood of an optimum, Algorithm 5.2 is essentially a preconditioned conjugate gradient method applied to an augmented Lagrangian. In [18] we show that Algorithm 5.2 is globally convergent while the iterations are locally quadratically convergent. When applying Algorithm 5.2 to either (4.7) or (4.8), the following four issues must be considered:

(1) *The initialization of phase one and phase two.* That is, given a guess for a solution to (4.1), what is the corresponding starting guess for the multipliers? Given an approximation to a solution to (4.7), what is the starting guess to (4.8)?

(2) *The addition and deletion of constraints.* After each iteration of an algorithm applied to either (4.7) or (4.8), we must delete "unnecessary" elements from Y^N and we must add "significant" elements to Y^N. The augmented Lagrangian will help to determine which elements to delete and which elements to add.

(3) *The elimination of* ρ. We introduced the parameter ρ to convert the minimax problem into an inequality constrained mathematical program. When applying a mathematical programming algorithm to either (4.7) or (4.8), we would like to eliminate the artificial variable ρ so that the iterations are expressed in terms of x and λ.

(4) *The computation of the gradient of the augmented Lagrangian.* When using the conjugate gradient method or any other gradient-based scheme to minimize the augmented Lagrangian, we need a formula for the gradient of the augmented Lagrangian with respect to x. Clarke's result [10, Thm. 2.1] can be used to compute this gradient.

To begin, let us consider the initialization of phase one. If x_1 is the starting guess in phase one, then in the absence of better information, let Y^N be the maximizers of $f(x_1, \cdot)$ on Y. In other words, $\eta \in Y^N$ if and only if

$$f(x, \eta) = \underset{y \in Y}{\text{maximum}}\, f(x, y).$$

In the absence of better information, the starting guess λ_1 for the multipliers corresponding to the constraint $f(x, y) \leqq \rho$ is $\lambda_{1y} = 1/N$ for each $y \in Y^N$. The x starting guess for phase two is simply the final iteration x_k of phase one. To initialize the phase two multipliers, we collapse the components of the phase one multipliers around the nearest peak. That is, in phase one we generate a multiplier λ_k with support Y^N. Given an element y in Y^N, the index $\nu(y)$ of the nearest peak is

$$\nu(y) = \arg\min \{\|y - y_i(x_k)\| : i = 1, 2, \cdots\}.$$

When more than one index achieves the minimum, let $\nu(y)$ be any one of them. Then the ith component of λ_1, the starting guess for phase two, is the sum of the phase one multiplier components that correspond to elements of Y^N closest to $y_i(x_k)$:

$$\lambda_{1i} = \sum_{\substack{y \in Y^N \\ \nu(y) = i}} \lambda_{ky}.$$

Moreover, the starting set Y^N for phase two consists of those $y_i(\cdot)$ for which λ_{1i} is positive.

To reduce the computing time associated with algorithms to solve (4.7) or (4.8), we wish to keep N as small as possible. After each complete iteration of [18, Algorithm 5.2], we will drop those constraints that appear to be nonbinding and we will add constraints where the inequality $f(x, y) \leqq \rho$ seems to be violated significantly. Let us now explain more precisely when to delete or add constraints. Given a finite set $S \subset Y$ and a multiplier λ with support in S, the augmented Lagrangian introduced in § 3 is

$$\mathbf{L}(\lambda, S, x) = \text{minimum}\, \{\mathbf{L}(\lambda, S, x, \rho) : \rho \in R\} \tag{4.9}$$

where

$$\mathbf{L}(\lambda, S, x, \rho) = \rho + \sum_{y \in S_+} \{\lambda_y (f(x, y) - \rho) + r(f(x, y) - \rho)^2\} - \frac{1}{4r} \sum_{y \in S_-} \lambda_y^2.$$

As usual, the limits for the summations above are

$$\begin{aligned} S_+ &= \{y \in S : 2r(f(x, y) - \rho) + \lambda_y \geqq 0\} \quad \text{and} \\ S_- &= \{y \in S : 2r(f(x, y) - \rho) + \lambda_y < 0\}. \end{aligned} \tag{4.10}$$

PROPOSITION 4.1. *For fixed* r, λ, S, *and* x, *suppose that* $\rho = \rho^*$ *attains the minimum in* (4.9), *and let* S_+ *be the corresponding set given in* (4.10). *Then we have*

$$\mathbf{L}(\lambda, S, x) = \mathbf{L}(\lambda_+, S_+, x) - \frac{1}{4r} \sum_{y \in S_-} \lambda_y^2$$

where λ_+ denotes the vector formed from λ by extracting those components λ_y corresponding to $y \in S_+$.

Proof. The identity

$$0=\frac{d}{d\rho}\mathbf{L}(\lambda, S, x, \rho^*)=\frac{d}{d\rho}\mathbf{L}(\lambda_+, S_+, x, \rho^*)$$

implies that ρ^* also minimizes $\mathbf{L}(\lambda_+, S_+, x, \cdot)$. □

Since $\mathbf{L}(\lambda, S, x)$ just differs from $\mathbf{L}(\lambda_+, S_+, x)$ by a constant, Proposition 4.1 implies that, at least locally, the constraints $f(x, y) \leqq \rho$ corresponding to $y \in S_-$ can be dropped. Consequently, our rule for deleting elements from Y^N can be stated:

Constraint Deletion

Let λ_k and x_k denote the approximations generated by one complete iteration of say [18, Algorithm 5.2]. Delete from Y^N those elements corresponding to $y \in Y^N_-$.

Now consider the addition of constraints. Again, the augmented Lagrangian helps us decide when the N in (4.8) must be increased. Suppose that η is not an element of S and $f(x, \eta)<\rho^*$ where $\rho=\rho^*$ attains the minimum in (4.9). Letting S_η denote $S \cup \{\eta\}$, we now show that $\mathbf{L}(\lambda, S_\eta, x)$ is locally equal to $\mathbf{L}(\lambda, S, x)$ if f is continuous and λ_η is zero. By the definition of the augmented Lagrangian, we have the inequality $\mathbf{L}(\lambda, S_\eta, x) \geqq \mathbf{L}(\lambda, S, x)$ whenever $\lambda_\eta=0$. Since $\mathbf{L}(\lambda, S_\eta, x, \rho^*)=\mathbf{L}(\lambda, S, x, \rho^*)$, it follows that $\mathbf{L}(\lambda, S_\eta, x)=\mathbf{L}(\lambda, S, x)$ whenever $\lambda_\eta=0$. Since the inequality $f(x, \eta)<\rho^*$ is preserved for small perturbations in x when f is continuous, we conclude that $\mathbf{L}(\lambda, S_\eta, x)$ is locally equal to $\mathbf{L}(\lambda, S, x)$. Conversely, suppose that $f(x, \eta)>\rho^*$ and $\lambda_\eta=0$. Since $\mathbf{L}(\lambda, S_\eta, x, \rho) \geqq \mathbf{L}(\lambda, S, x, \rho)$ for every ρ when $\lambda_\eta=0$ and since $\mathbf{L}(\lambda, S_\eta, x, \rho^*)>\mathbf{L}(\lambda, S, x, \rho^*)$, it follows from the uniqueness result Lemma 3.2 that $\mathbf{L}(\lambda, S_\eta, x)>\mathbf{L}(\lambda, S, x)$. To summarize, if $f(x, \eta)>\rho^*$, then $\mathbf{L}(\lambda, S_\eta, x)$ is larger than $\mathbf{L}(\lambda, S, x)$ and the gap between the value of the primal problem (4.1) and the value of the dual problem

$$\underset{\lambda, S}{\text{maximize}}\ \underset{x \in X}{\text{minimium}}\ \mathbf{L}(\lambda, S, x)$$

may be reduced by inserting η into S. These observations lead us to the following rule for adding constraints in phase one:

Constraint Addition in Phase One

Let λ_k and x_k denote the approximations generated by one complete iteration of say [18, Algorithm 5.2] and let $\rho=\rho^*$ minimize $\mathbf{L}(\lambda_k, Y^N, x_k, \rho)$ over ρ. Insert $y_i(x_k)$ into Y^N if $f(x_k, y_i(x_k))>\rho^*$ and the distance between $y_i(x_k)$ and Y^N is greater than some fixed predetermined constant Δ.

Since phase one approximates the solution to the minimax problem, the local maximizer $y_i(x_k)$ appearing in the constraint addition step of phase one does not need to be computed very accurately. The positive parameter Δ introduced above prevents points in Y^N from clustering together. As the number of points in Y^N increases, the time to evaluate $\mathbf{L}$ increases and the Hessian of $\mathbf{L}$ becomes ill conditioned. Since it helps to keep the number of points in Y^N small, we exclude those local maxima which are already near elements of Y^N. In numerical experiments, the convergence speed is not very sensitive to the choice of Δ. In phase two, the elements of Y^N are local maxima instead of fixed elements in Y. Hence, the analogous rule for adding constraints in phase two can be stated:

CONSTRAINT ADDITION IN PHASE TWO

Let λ_k and x_k denote the approximations generated by one complete iteration of say [18, Algorithm 5.2] and let $\rho = \rho^*$ minimize $\mathrm{L}(\lambda_k, Y^N, x_k, \rho)$ over ρ. Insert $y_i(\cdot)$ into Y^N if $f(x_k, y_i(x_k)) > \rho^*$.

Up to here, we have explained how to initialize a mathematical programming algorithm to solve either (4.7) or (4.8) and we have explained how to add or delete constraints at the end of each iteration in the algorithm. Now let us consider the details of an interation. In formulation (4.7) and (4.8), a parameter ρ is introduced and the number of independent variables is increased by one. For many algorithms, the artificial variable ρ can be eliminated and the iterations can be expressed in terms of x and λ. For notational convenience, we assume X is R^n. Let x_k be the kth approximation to a solution to the minimax problem and suppose that after deleting and adding constraints at the end of iteration k, we have $Y^N = \{y_1, \cdots, y_N\}$. In [18, Algorithm 5.2], we estimate the multipliers corresponding to the constraints of (4.7) or (4.8) by computing the least squares solution λ to the system of equations

$$\sum_{i=1}^{N} \lambda_i = 1, \qquad \sum_{i=1}^{N} \lambda_i \nabla f(x_k, y_i) = 0. \tag{4.11}$$

Computing the least squares solution to this system of $n+1$ equations in N unknowns is equivalent to computing the pseudoinverse of a $(n+1) \times N$ matrix. As an alternative to this procedure, we suggest the following: Solve the first equation in (4.11) for λ_1 in terms of λ_2 through λ_N and substitute into the second relation to obtain n equations in $N-1$ unknowns:

$$\sum_{i=2}^{N} \lambda_i (\nabla_x f(x_k, y_i) - \nabla_x f(x_k, y_1)) = -\nabla_x f(x_k, y_1).$$

The least squares solution to this system gives us an estimate for λ_2 through λ_N while λ_1 is determined from the relation $\lambda_1 = 1 - \lambda_2 - \lambda_3 - \cdots - \lambda_N$.

The procedure outlined above to estimate the multipliers is quite effective in phase two. On the other hand, in phase one a simpler strategy involving the gradient approximation to the multipliers (see [2]) is often just as effective. Let λ_k be the kth approximation to the multipliers and let x_k be the corresponding approximation to a minimizer of the augmented Lagrangian $\mathrm{L}(\lambda_k, Y^N, \cdot)$. Set $\lambda_{k+1,i} = \lambda_{ki} + 2r(f(x_k, y_{ki}) - \rho_k)$ for $i \in Y^N_+$ and set $\lambda_{k+1,i} = 0$ for $i \in Y^N_-$ if this rule generates a λ_{k+1} with the property that (λ_{k+1}, x_k) is a better approximation to a Kuhn-Tucker point for (4.7) than (λ_k, x_k). Otherwise, set $\lambda_{k+1} = \lambda_k$. Here ρ_k denotes the minimizer in (4.9) corresponding to $\lambda = \lambda_k$, $S = Y^N$, and $x = x_k$. A technique for measuring the distance to a Kuhn-Tucker point is developed in [18].

In [18, Algorithm 5.2], the restoration step is essentially a Newton iteration applied to the system of N equations

$$f(x, y_i) = \rho \quad \text{for } i = 1 \text{ to } N \tag{4.12}$$

where the starting guess is x_k and the corresponding ρ_k generated the previous iteration. Since the N equations (4.12) are equivalent to the $N-1$ equations

$$f(x, y_i) - f(x, y_1) = 0 \quad \text{for } i = 2 \text{ to } N, \tag{4.13}$$

an alternative procedure is to apply one Newton iteration to the system (4.13). Observe that this Newton iteration involves computing the pseudoinverse of the same matrix used in the multiplier estimate.

In the minimization step of [18, Algorithm 5.2], we use a preconditioned conjugate gradient method to minimize $\mathbf{L}(\lambda_k, Y^N, x)$ over x. (Here, λ_k denotes the multiplier associated with iteration k.) Near the optimum, the preconditioner is chosen to project the gradients into the null space of the binding constraints. Therefore, near the optimum, the preconditioner projects the gradients into the null space of the $(N-1)\times n$ matrix with rows

$$\nabla_x f(x_k, y_i) - \nabla_x f(x_k, y_1) \quad \text{for } i = 2 \text{ to } N.$$

Far from the optimum, the preconditioner is chosen to mitigate the ill conditioning due to penalty terms in the augmented Lagrangian. At the start of iteration $k+1$, the inequalities $f(x, y_i) \leqq \rho$ are viewed as equalities and it follows from (4.9) that for x near x_k,

$$\mathbf{L}(\lambda_k, Y^N, x) = \sum_{i=1}^{N} \lambda_{ki} f(x, y_i) + r \sum_{i=1}^{N} f(x, y_i)^2 - \frac{r}{N}\left(\sum_{i=1}^{N} f(x, y_i)\right)^2.$$

The identity

$$\sum_{i=1}^{N} f(x, y_i)^2 - \frac{1}{N}\left(\sum_{i=1}^{N} f(x, y_i)\right)^2 = \sum_{i=1}^{N}\left(f(x, y_i) - \frac{1}{N}\sum_{j=1}^{N} f(x, y_j)\right)^2,$$

combined with the preconditioning theory developed in [18, § 4], tells us that a natural preconditioner for the minimax problem is the matrix $H = (I + B^T B)^{-1}$, where B is the $N \times n$ matrix with ith row

$$\nabla_x f(x_k, y_i) - \frac{1}{N}\sum_{j=1}^{N} \nabla_x f(x_k, y_i).$$

Observe that the rows of B are linearly dependent since their sum is zero. Let V denote the matrix $I - vv^T$, where

$$v = \frac{1}{\sqrt{\sqrt{N}(1+\sqrt{N})}}\begin{bmatrix} 1+\sqrt{N} \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Since the first row of V is a multiple of $\mathbf{1}$, the first row of VB is zero. Let W be the matrix obtained by deleting the first row of VB. Since V is orthogonal, we have

$$B^T B = (VB)^T VB = W^T W.$$

Applying the Woodbury formula [15, p. 3], the preconditioner H can be written

$$H = (I + rW^T W)^{-1} = I - W^T(r^{-1} I + WW^T)^{-1} W.$$

When using any gradient technique to minimize $\mathbf{L}(\lambda_k, Y^N, \cdot)$, we must compute the gradient of the augmented Lagrangian with respect of x. By [10, Thm. 2.1] this gradient can be expressed

$$\nabla_x \mathbf{L}(\lambda, S, x) = \sum_{y \in S_+} (\lambda_y + 2r(f(x, y) - \rho^*))\nabla_x f(x, y)$$

where ρ^* attains the minimum in (4.9). This formula for the gradient is also valid when the elements of S depend on x (as in (4.8)) provided these elements are local extreme points of $f(x,\cdot)$ on Y and (for example) $\nabla_x f(x, y)$ is a continuous function of x and y.

Comparing our approach to the minimax problem to the approach of Murray and Overton [26], some similarities are that we both reformulate the minimax problem as a mathematical program with an extra unknown and we both estimate simultaneously the primal solution and the Lagrange multipliers. Some differences in our methods are the following: (i) In [26] Y is finite. (ii) We utilize an augmented Lagrangian while [26] considers the ordinary Lagrangian. (iii) Our strategy for adding and deleting constraints is different from [26]—our strategy ties in with the augmented Lagrangian. (iv) With our approach, nonlinear constraints contained in X can be incorporated in the augmented Lagrangian just as easily as the constraints $f(x, y) \leqq \rho$.

5. Reducible minimax problems. So far we have viewed the minimax problem as an optimization problem with inequality constraints and we have applied a constrained optimization algorithm. Now let us develop an algorithm that is specially tailored to reducible minimax problems. That is, we assume that there exists a finite set $Y^* \subset Y$ and a x^* in X such that

$$\min_{x\in X} \max_{y\in Y^*} f(x, y) = \max_{y\in Y^*} f(x^*, y) = \max_{y\in Y} f(x^*, y) = \Phi(x^*),$$

and we search for the set Y^*. It is also assumed that there exists a real number r and a multiplier λ^* with support in Y^* such that $\mathbf{L}(\lambda^*, Y^*) = \Phi(x^*)$ where $\mathbf{L}(\cdot,\cdot)$ denotes the dual functional defined by

$$\mathbf{L}(\lambda, S) = \inf\{\mathbf{L}(\lambda, S, x): x \in X\}.$$

Given an approximation $Y_k = \{y_{k1}, \cdots, y_{kN}\}$ to Y^* and given an approximation λ_k to λ^*, the rules for computing Y_{k+1} and λ_{k+1} are the following:

PEAK CHASING ALGORITHM

(a) If x_k minimizes $\mathbf{L}(\lambda_k, Y_k, x)$ over $x \in X$, then set $\lambda_{k+1,i} = \lambda_{ki} + 2r(f(x_k, y_{ki}) - \rho_k)$ for $i \in S_+$ and $\lambda_{k+1,i} = 0$ for $i \in S_-$ where ρ_k attains the minimum in (4.9) corresponding to $\lambda = \lambda_k$ and $S = Y_k$.

(b) If x_{k+1} minimizes $\mathbf{L}(\lambda_{k+1}, Y_k, \cdot)$ over X and if $Z_k = \{z_1, \cdots, z_N\}$ denotes a collection of local maxima for $f(x_{k+1}, \cdot)$ on Y where z_i is the closest local maximizer to y_{ki}, then we set $Y_{k+1} = \beta Y_k + (1-\beta) Z_k$ where

$$\beta = \arg\max_{0\leqq\alpha\leqq 1} \mathbf{L}(\lambda_{k+1}, \alpha Y_k + (1-\alpha) Z_k).$$

Step (a) is the usual gradient step for an augmented Lagrangian (see [2]). Since this algorithm is linearly convergent, the parameter β of step (b) can be imprecise. In practice, we find that the maximizer of the interpolating quadratic that agrees with $\mathbf{L}(\lambda_{k+1}, \alpha Z_k + (1-\alpha) Y_k)$ at $\alpha = 0$, at $\alpha = \frac{1}{2}$, and at $\alpha = 1$ works well. To show that the peak chasing algorithm is locally convergent, we verify that each iteration increases the value of the dual functional. That is, $\mathbf{L}(\lambda_{k+1}, Y_{k+1}) \geqq \mathbf{L}(\lambda_k, Y_k)$ with equality only possible at $\lambda_k = \lambda^*$ and at $Y_k = Y^*$. In order to show that step (a) is an ascent step, let us first consider the equality constrained problem

$$\text{(5.1)} \qquad \begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } h(x) = 0, \qquad x \in R^n \end{aligned}$$

where f is quadratic: $f(x)=x^TAx+a^Tx$ and h is linear: $h(x)=Bx-b$. Here A is an $n\times n$ matrix, B is an $m\times n$ matrix, and a and b are vectors in R^n and R^m, respectively. The augmented Lagrangian corresponding to (5.1) is

$$\mathbf{L}(\lambda, x)=f(x)+\lambda^Th(x)+r|h(x)|^2.$$

LEMMA 5.1. *Suppose that the rows of B are linearly independent and A is positive definite in the null space of B. Then there exist positive parameters α and s such that $A+rB^TB\geqq \alpha I$ for every $r\geqq s$. If $\lambda\in R^n$ and z minimizes $\mathbf{L}(\lambda,\cdot)$ over R^n, then we have*

$$\mathbf{L}(\lambda+2rh(z), y)\geqq \mathbf{L}(\lambda, z)+\tfrac{1}{2}r|h(z)|^2+\alpha|y-z|^2 \tag{5.2}$$

for every $r\geqq 3s$ and for every $y\in R^n$.

(If M_1 and M_2 are symmetric matrices of the same dimension, then the notation $M_1>M_2$ means that M_1-M_2 is positive definite.)

Proof. In [16, Lemma 2.6] we determine a parameter $s<\infty$ with the property that $A+rB^TB$ is positive definite for $r\geqq s$. Let μ denote $\lambda+2rh(z)$. Expanding the Lagrangian in a Taylor series, we have

$$\mathbf{L}(\mu, y)=\mathbf{L}(\mu, z)+\nabla_x\mathbf{L}(\mu, z)(y-z)+\tfrac{1}{2}\nabla_x^2\mathbf{L}(\mu, \xi)(y-z)^2 \tag{5.3}$$

where ξ lies on the line segment connecting y and z. The relation $\mu=\lambda+2rh(z)$ implies that

$$\mathbf{L}(\mu, z)=\mathbf{L}(\lambda, z)+2r|h(z)|^2 \tag{5.4}$$

and

$$\nabla_x\mathbf{L}(\mu, z)(y-z)=\nabla_x\mathbf{L}(\lambda, z)(y-z)+2rh(z)^T\nabla h(z)(y-z). \tag{5.5}$$

(Note that $\nabla_x\mathbf{L}(\mu, z)$ is not equal to the gradient of the right side of (5.4) since μ is treated as a constant when computing $\nabla_x\mathbf{L}(\mu, z)$.) If z minimizes $\mathbf{L}(\lambda,\cdot)$, then $\nabla_x\mathbf{L}(\lambda, z)$ is zero and by (5.5), we have

$$\nabla_x\mathbf{L}(\mu, z)(y-z)=2rh(z)^T\nabla h(z)(y-z). \tag{5.6}$$

By the definition of f and h, it follows that $\nabla h=B$ and $\nabla_x^2\mathbf{L}=2(A+rB^TB)$. Combining (5.3), (5.4), and (5.6) gives us

$$\mathbf{L}(\mu, y)=\mathbf{L}(\lambda, z)+2r|h(z)|^2+2rh(z)^TB(y-z)+(y-z)^T(A+rB^TB)(y-z).$$

Utilizing the inequality

$$ab\leqq \tfrac{3}{4}a^2+\tfrac{1}{3}b^2$$

where we identify a with $|h(z)|$ and b with $|B(y-z)|$ yields

$$\mathbf{L}(\mu, y)\geqq \mathbf{L}(\lambda, z)+\tfrac{1}{2}r|h(z)|^2+(y-z)^T(A+\tfrac{1}{3}rB^TB)(y-z). \tag{5.7}$$

Hence, (5.2) holds for $r\geqq 3s$. □

For a general f and h, the same argument employed in the proof of Lemma 5.1 can also be applied to a neighborhood of a local optimum. Removing the restriction that f is quadratic and h is linear, we have the following.

THEOREM 5.2. *Suppose that x^* and λ^* satisfy the hypotheses of Theorem 3.1. Then there exists a neighborhood $\mathbf{N}_x$ of x^*, a neighborhood $\mathbf{N}_\lambda$ of λ^*, and positive parameters α and s such that*

$$\nabla_x^2\mathbf{l}(\lambda, x)+2r\nabla h(x)^T\nabla h(x)>\alpha I \tag{5.8}$$

whenever $r \geqq s$, $\lambda \in \mathbf{N}_\lambda$, *and* $x \in \mathbf{N}_x$. *Moreover, for s sufficiently large, a parameter c can be chosen so that* $\mathbf{L}(\lambda, \cdot)$ *has a unique local minimizer* $x(\lambda)$ *inside* $\mathbf{N}_x$ *whenever* $|\lambda - \lambda^*| \leqq cr$ *and* $r \geqq s$. *And if* λ *and* μ *lie in* $\mathbf{N}_\lambda$, $y \in \mathbf{N}_x$ *with* $|y - x^*| \leqq c|\lambda - \lambda^*|/r$, *and* $r \geqq s$, *then we have*

$$\mathbf{L}(\mu, y) \geqq \mathbf{L}(\lambda, z) + \tfrac{1}{2}(r|h(z)|^2 + (\alpha - \delta|\lambda - \lambda^*|)|y - z|^2) \tag{5.9}$$

where $z = x(\lambda)$, $\mu = \lambda + 2rh(z)$, *and* δ *is a constant that is independent of* λ *and* y.

Proof. By [16, Lemma 2.6], $\nabla_x^2 \mathbf{L}(\lambda^*, x^*)$ is positive definite for r sufficiently large, and by [16, Lemma 6.5], there exists a neighborhood of (λ^*, x^*) where (5.8) holds for α sufficiently small and r sufficiently large. The statement concerning the existence of a locally unique minimizer $x(\lambda)$ for $\mathbf{L}(\lambda, \cdot)$ is established (for example) in [2]. To prove (5.9), we expand $\mathbf{L}$ in a Taylor series giving us the following analogue of (5.7):

$$\mathbf{L}(\mu, y) \geqq \mathbf{L}(\lambda, z) + \tfrac{1}{2}r|h(z)|^2 + \tfrac{1}{2}(y - z)^T \nabla_x^2 \mathbf{L}(\mu, \xi)(y - z) - \tfrac{2}{3}r|\nabla h(z)(y - z)|^2 \tag{5.10}$$

where ξ lies between y and z. Utilizing the inequality

$$|x|^2 \leqq (|x - y| + |y|)^2 \leqq 5|x - y|^2 + \tfrac{5}{4}|y|^2$$

where x is identified with $\nabla h(z)(y - z)$ and y is identified with $\nabla h(\xi)(y - z)$, the last two terms in (5.10) satisfy the relation

$$\begin{aligned} &\tfrac{1}{2}(y - z)^T \nabla_x^2 \mathbf{L}(\mu, \xi)(y - z) - \tfrac{2}{3}r|\nabla h(z)(y - z)|^2 \\ &\quad \geqq \tfrac{1}{2}(y - z)^T \nabla_x^2 \mathbf{l}(\mu, \xi)(y - z) + \tfrac{1}{6}r|\nabla h(\xi)(y - z)|^2 \\ &\qquad - \tfrac{10}{3}r|(\nabla h(\xi) - \nabla h(z))(y - z)|^2 \\ &\qquad + r(y - z)^T\left(\sum_{i=1}^m h_i(\xi)\nabla^2 h_i(\xi)\right)(y - z). \end{aligned} \tag{5.11}$$

By (5.8), we have

$$\tfrac{1}{2}(y - z)^T \nabla_x^2 \mathbf{l}(\mu, \xi)(y - z) + \tfrac{1}{6}r|\nabla h(\xi)(y - z)|^2 \geqq \tfrac{1}{2}\alpha|y - z|^2 \tag{5.12}$$

provided r is sufficiently large, $\mu \in \mathbf{N}_\lambda$, and $\xi \in \mathbf{N}_x$. By Theorem 3.1, $|z - x^*|$ is bound by a constant times $|\lambda - \lambda^*|/r$ and by assumption, $|y - x^*|$ is bound by a constant times $|\lambda - \lambda^*|/r$. Since $h(x^*)$ is zero, there exists a constant δ such that

$$\begin{aligned} &\tfrac{10}{3}r|(\nabla h(\xi) - \nabla h(z))(y - z)|^2 + r(y - z)^T\left(\sum_{i=1}^m h_i(\xi)\nabla^2 h_i(\xi)\right)(y - z) \\ &\quad \leqq \delta|\lambda - \lambda^*|\,|y - z|^2 \end{aligned} \tag{5.13}$$

for λ near λ^*. Combining (5.10)-(5.13), the proof is complete. □

Theorem 5.2 implies that if λ is near λ^*, then for $\mu = \lambda + 2rh(x(\lambda))$, $\mathbf{L}(\mu, x(\mu))$ is equal to $\mathbf{L}(\lambda, x(\lambda))$ only if $h(x(\lambda)) = 0$. Since $x(\lambda)$ minimizes $\mathbf{L}(\lambda, \cdot)$ over $\mathbf{N}_x$, we conclude that if $\mathbf{L}(\mu, x(\mu)) = \mathbf{L}(\lambda, x(\lambda))$, then $x = x(\lambda)$ is a solution to the problem: minimize $f(x)$ subject to $x \in \mathbf{N}_x$ and $h(x) = 0$. Therefore, $x(\lambda) = x^*$, the local minimizer corresponding to λ^*. Theorem 5.2 also applies to problems of the form

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } h(x) = 0, \qquad x \in X \end{aligned} \tag{5.14}$$

where X is a convex set. The proof of Theorem 5.2 in this more general setting involves an analogue of Theorem 3.1 that applied to (5.14). See Bertsekas [2] for the extension of Theorem 3.1 to problems with the constraint $x \in X$. Furthermore, referring to the proof of Lemma 5.1, the constraint $x \in X$ alters the treatment of the term $\nabla_x \mathbf{L}(\lambda, z)(y-z)$. When X is R^n, $\nabla_x \mathbf{L}(\lambda, z)$ is zero and the term $\nabla_x \mathbf{L}(\lambda, z)(y-z)$ can be dropped. But for an arbitrary convex set, the corresponding relation is

$$\nabla_x \mathbf{L}(\lambda, z)(y-z) \geqq 0.$$

This inequality has the right direction so that the term $\nabla_x \mathbf{L}(\lambda, z)(y-z)$ can still be dropped without affecting (5.7).

Although Theorem 5.2 is established for an equality constraint, it also applies to the inequality constraint $g(x) \leqq 0$ provided λ_i^* is positive whenever $g_i(x^*)$ is zero. This follows from [31, Thm. 5.1] where Rockafellar proves that near λ^*, minimizing $\mathbf{L}(\lambda, \cdot)$ is equivalent to minimizing the augmented Lagrangian corresponding to an equality constraint $h(x)=0$—the components h_i of h are the components g_i of g for which $\lambda_i^* > 0$. Hence, by Theorem 5.2, step (a) of the peak chasing algorithm is an ascent step under appropriate assumptions. That is, $\mathbf{L}(\lambda_{k+1}, Y_k) \geqq \mathbf{L}(\lambda_k, Y_k)$ with equality only possible when the x_k which locally minimizes $\mathbf{L}(\lambda_k, Y_k, \cdot)$ on X is a local minimizer for the problem

$$\underset{x \in X}{\text{minimize}}\ \underset{y \in Y_k}{\text{maximum}}\, f(x, y).$$

Now consider step (b) of the peak chasing algorithm. As noted above, in a neighborhood of an optimum, the augmented Lagrangian corresponding to an inequality constrained problem is the same as the augmented Lagrangian corresponding to an equality constrained problem. For this reason, we focus attention on the augmented Lagrangian

$$\mathbf{L}(\lambda, S, x) = \sum_{y \in S} \lambda_y f(x, y) + r \sum_{y \in S} f(x, y)^2 - \frac{r}{N}\left(\sum_{y \in S} f(x, y)\right)^2$$

which corresponds to the equality constrained problem

$$\text{minimize } \rho$$

$$\text{subject to } x \in X, \quad \rho \in R, \quad f(x, y) = \rho \quad \text{for every } y \in S.$$

And we establish the following property for the dual functional:

LEMMA 5.3. *Let λ be a fixed vector in R^N with nonnegative components, let $S_0 = \{y_1, \cdots, y_N\}$ and let $S_1 = \{z_1, \cdots, z_N\}$ be subsets of Y, and suppose that for $i=0$ and for $i=1$, x_i minimizes $\mathbf{L}(\lambda, S_i, x)$ over x in X. Then we have*

$$\mathbf{L}(\lambda, S_1) \geqq \mathbf{L}(\lambda, S_0) + \sum_{i=1}^{N} (\lambda_i + 2r(f(x_1, y_i) - \rho_1))(f(x_1, z_i) - f(x_1, y_i)) \tag{5.15}$$

where

$$\rho_1 = \frac{1}{N} \sum_{i=1}^{N} f(x_1, y_i).$$

Proof. Let $Q: R^N \to R$ be the quadratic defined by

$$Q(p) = \sum_{i=1}^{N} \lambda_i p_i + r p_i^2 - \frac{r}{N}\left(\sum_{i=1}^{N} p_i\right)^2.$$

The Hessian of Q is

$$\nabla^2 Q = 2r\left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^T\right),$$

which is positive semidefinite by Gerschgorin's theorem. Hence, Q is a convex function which satisfies the standard inequality [29, p. 242]:

$$Q(p)-Q(q) \geqq \nabla Q(q)(p-q)=\sum_{i=1}^{N}(\lambda_i+2r(q_i-\rho))(p_i-q_i) \tag{5.16}$$

where

$$\rho=\frac{1}{N}\sum_{i=1}^{N} q_i.$$

Since x_0 minimizes $\mathbf{L}(\lambda, S_0, x)$ over $x \in X$, it follows that $\mathbf{L}(\lambda, S_0) \leqq \mathbf{L}(\lambda, S_0, x_1)$, or equivalently,

$$\mathbf{L}(\lambda, S_1)-\mathbf{L}(\lambda, S_0) \geqq \mathbf{L}(\lambda, S_1, x_1)-\mathbf{L}(\lambda, S_0, x_1). \tag{5.17}$$

Applying (5.16) to the right side of (5.17) where $p_i=f(x_1, z_i)$ and $q_i=f(x_1, y_i)$ yields (5.15). □

Lemma 5.3 can be used to show that under appropriate assumptions, step (b) of the peak chasing algorithm is an ascent step. Let S_α denote $\alpha S_1+(1-\alpha)S_0$. Suppose that for α between zero and one, the minimum of $\mathbf{L}(\lambda, S_\alpha, x)$ over $x \in X$ is attained at a point labeled x_α which is a continuous function of α. Let λ^* maximize $\mathbf{L}(\lambda, S_0)$ over λ. Typically the components of λ^* are positive. By Theorem 3.1, the vector μ_0 with components

$$\mu_{0i}=\lambda_i+2r(f(x_0, y_i)-\rho_0), \qquad \rho_0=\frac{1}{N}\sum_{i=1}^{N} f(x_0, y_i)$$

satisfies the inequality $|\mu_0-\lambda^*| \leqq c|\lambda-\lambda^*|/r$ for some constant c. Hence, the components of μ_0 are positive for r sufficiently large. Letting μ_α be the vector defined by

$$\mu_{\alpha i}=\lambda_i+2r(f(x_\alpha, y_i)-\rho_\alpha), \qquad \rho_\alpha=\frac{1}{N}\sum_{i=1}^{N} f(x_\alpha, y_i),$$

it follows that the components of μ_α are positive for α sufficiently small. If z_i is a local maximizer of $f(x_0, \cdot)$, then we expect that $f(x_0, \cdot)$ is locally concave near z_i. Assuming that $f(z_\alpha, \cdot)$ is concave for α near zero on the line segment connecting y_i and z_i, we have

$$f(x_\alpha, \alpha z_i+(1-\alpha)y_i) \geqq \alpha f(x_\alpha, z_i)+(1-\alpha)f(x_\alpha, y_i). \tag{5.18}$$

Combining (5.15) and (5.18) gives us

$$\mathbf{L}(\lambda, S_\alpha) \geqq \mathbf{L}(\lambda, S_0)+\alpha\left\{\sum_{i=1}^{N} \mu_{\alpha i}(f(x_\alpha, z_i)-f(x_\alpha, y_i))\right\}. \tag{5.19}$$

Hence, for α sufficiently small, $\mathbf{L}(\lambda, S_\alpha)$ is strictly larger than $\mathbf{L}(\lambda, S_0)$ unless the y_i are equal to the z_i. Now let us state a more precise convergence result.

THEOREM 5.4. *We make the following assumptions*:

I. *X is R^n, Y is a convex, compact subset of a vector space, and $f(x, y)$ is a concave function of y for each fixed x. There exists x^* in X, a finite set $Y^*=\{y_1^*, \cdots, y_N^*\}$ contained in Y, and a multiplier λ^* with support equal to Y^* such that*

$$\mathbf{L}(\lambda^*, Y^*)=\underset{y \in Y^*}{\text{maximum}}\, f(x^*, y)=\underset{y \in Y}{\text{maximum}}\, f(x^*, y).$$

II. *The Hessian* $\nabla_x^2 f(x, y)$ *exists and depends continuously on* x *near* x^* *and on* y *near* y_i^* *for each* i *between* 1 *and* N. *Moreover, the mathematical program*

$$\textit{minimize } \rho$$

$$\textit{subject to } x \in R^n, \quad f(x, y_i^*) - \rho = 0 \quad \textit{for } i = 1, \cdots, N$$

satisfies the assumptions of Theorem 3.1 *at the optimum* $x = x^*$ *and* $\rho = f(x, y_i^*)$ *and if* **N** *is the neighborhood of* x^* *introduced in Theorem* 3.1, *then we have*

$$\mathbf{L}(\lambda, S) = \inf\{\mathbf{L}(\lambda, S, x) \colon x \in \mathbf{N}\}$$

for λ *and* S *in some neighborhood* **W** *of* (λ^*, Y^*).

III. *For* $x \in \mathbf{N}$ *there exist local maxima* $y_i(x)$ *of* $f(x, \cdot)$ *on* Y *such that* $y_i(x)$ *approaches* y_i^* *as* x *approaches* x^* *for* $i = 1, \cdots, N$. *Furthermore,* $y_i(x)$ *is the locally unique maximizer of* $f(x, \cdot)$ *for* x *near* x^* *and for* x *near* x^*, *we have*

$$\underset{1 \leq i \leq N}{\text{maximum}}\, f(x, y_i(x)) = \underset{y \in Y}{\text{maximum}}\, f(x, y).$$

IV. $\mathbf{L}(\lambda, S) < \mathbf{L}(\lambda^*, Y^*)$ *whenever* $(\lambda, S) \in \mathbf{W}$, $\lambda \neq \lambda^*$, *and* $S \neq Y^*$.

Under assumptions I–IV *and for* r *large enough, the peak chasing algorithm converges to* λ^* *and* Y^* *starting from any point sufficiently close to* λ^* *and* Y^*.

Proof. We just sketch the proof. For λ near λ^* and for S near Y^*, assumption II implies that there exists $x(\lambda, S)$ which minimizes $\mathbf{L}(\lambda, S, x)$ over x in a neighborhood of x^* and $x(\lambda, S)$ depends continuously on λ and S. By Theorem 5.2, step (a) of the peak chasing algorithm is an ascent step for r sufficiently large. Since step (b) of the peak chasing algorithm does not decrease the value of the dual functional, assumption IV implies that if the iterations start near λ^* and Y^*, then the iterations remain near λ^* and Y^*. Since λ_k, Y_k, and x_k lie in compact sets, we can extract subsequences converging to limits λ_∞, Y_∞, and x_∞, respectively. For convenience, these subsequences are also denoted λ_k, Y_k, and x_k. Since x_k minimizes $\mathbf{L}(\lambda_k, Y_k, x)$ over $x \in X$, we conclude that x_∞ minimizes $\mathbf{L}(\lambda_\infty, Y_\infty, x)$ over $x \in X$. Since $\mathbf{L}(\lambda_k, Y_k)$ is bound above by $\mathbf{L}(\lambda^*, Y^*)$, the difference $\mathbf{L}(\lambda_{k+1}, Y_{k+1}) - \mathbf{L}(\lambda_k, Y_k)$ approaches zero as k increases. Hence, Theorem 5.2 implies that $f(x_\infty, y_{\infty i}) - \rho_\infty$ is zero for each i. Also, it follows from (5.19) that the elements of Y_∞ are local maxima of $f(x_\infty, \cdot)$ on Y. Combining these relations, we conclude that

$$\mathbf{L}(\lambda_\infty, Y_\infty) = \underset{y \in Y_\infty}{\text{maximum}}\, f(x_\infty, y) = \underset{y \in Y}{\text{maximum}}\, f(x_\infty, y). \tag{5.20}$$

The first equality in (5.20) implies through duality that x_∞ is the solution to the discrete minimax problem

$$\text{minimize } \rho$$

$$\text{subject to } x \in R^n, \quad f(x, y_{\infty i}) \leq \rho \quad \text{for } i = 1, \cdots, N.$$

And the second equality in (5.20) implies that x_∞ is a solution to the continuous minimax problem

$$\underset{x \in R^n}{\text{minimize}}\ \underset{y \in Y}{\text{maximum}}\, f(x, y).$$

By assumption IV, λ^* and Y^* are locally unique maxima of $\mathbf{L}$. Consequently, $\lambda_\infty = \lambda^*$,

$Y_\infty = Y^*$, and $x_\infty = x^*$. It then follows from the ascent property that the original sequence (not just the extracted subsequence) converges to λ^*, Y^*, and x^*. □

Acknowledgment. We wish to thank the referee for a suggestion that led to a shorter proof for Lemma 3.2 and a more direct treatment of the algorithm to minimize $\mathbf{L}(\lambda, x, \rho)$ over ρ.

REFERENCES

[1] K. J. ARROW AND R. M. SOLOW, *Gradient methods for constrained maxima, with weakened assumptions*, in Studies in Linear and Nonlinear Programming, K. Arrow, L. Hurwicz and H. Uzawa, eds., Stanford University Press, Stanford, CA, 1958.

[2] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[3] R. W. CHANEY, *A method of centers algorithm for certain minimax problems*, Math. Programming, 22 (1982), pp. 202-226.

[4] C. CHARALAMBOUS AND A. R. CONN, *An efficient method to solve the minimax problem directly*, SIAM J. Numer. Anal., 15 (1978), pp. 162-187.

[5] A. CHARNES, W. W. COOPER AND K. KORTANEK, *Duality in semi-infinite programs and some works of Haar and Carathéodory*, Management Sci., 9 (1963), pp. 209-228.

[6] ———, *On representations of semi-infinite programs which have no duality gaps*, Management Sci., 12 (1965), pp. 113-121.

[7] ———, *On the theory of semi-infinite programming and a generalization of the Kuhn-Tucker saddle point theorem for arbitrary convex functions*, Naval Res. Logist. Quart., 16 (1969), pp. 41-51.

[8] J. A. CHATELON, D. W. HEARN AND T. J. LOWE, *A subgradient algorithm for certain minimax and minisum problems*, Math. Programming, 15 (1978), pp. 130-145.

[9] ———, *A subgradient algorithm for certain minimax and minisum problems—the constrained case*, this Journal, 20 (1982), pp. 455-469.

[10] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247-262.

[11] T. F. COLEMAN, *A note on 'New Algorithms for constrained minimax optimization,'* Math. Programming, 15 (1978), pp. 239-242.

[12] V. F. DEM'YANOV, *Algorithms for some minimax problems*, J. Comput. System Sci., 2 (1968), pp. 342-380.

[13] V. F. DEM'YANOV AND V. N. MALOZEMOV, *Introduction to Minimax*, D. Louvish, transl., John Wiley, New York, 1974.

[14] S. R. K. DUTTA AND M. VIDYASAGAR, *New algorithms for constrained minimax optimization*, Math. Programming, 13 (1977), pp. 140-155.

[15] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.

[16] W. W. HAGER, *Approximations to the multiplier method*, SIAM J. Numer. Anal., 22 (1985), pp. 16-46.

[17] W. W. HAGER AND G. D. IANCULESCU, *Dual approximations in optimal control*, this Journal, 22 (1984), pp. 423-465.

[18] ———, *Dual techniques for constrained optimization*, J. Optim. Theory Appl., to appear.

[19] W. W. HAGER AND R. ROSTAMIAN, *Optimal coatings, bang-bang controls, and gradient techniques*, in Optimal Control: Applications and Methods, to appear.

[20] J. HALD AND K. MADSEN, *Combined LP and quasi-Newton methods for minimax optimization*, Math. Programming, 20 (1981), pp. 49-62.

[21] S. P. HAN, *Variable metric methods for minimizing a class of nondifferentiable functions*, Math. Programming, 20 (1981), pp. 1-13.

[22] M. R. HESTENES, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303-320.

[23] K. C. KIWIEL, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Mathematics, 1133, Springer-Verlag, New York, 1985.

[24] R. KLESSIG AND E. POLAK, *A method of feasible directions using function approximations, with applications to min max problems*, J. Math. Anal. Appl., 41 (1973), pp. 583-602.

[25] D. E. KNUTH, *The Art of Computer Programming, Volume* III: *Sorting and Searching*, Addison-Wesley, Reading, MA, 1973.

[26] W. MURRAY AND M. L. OVERTON, *A projected Lagrangian algorithm for nonlinear minimax optimization*, SIAM J. Sci. Statist. Comput., 1 (1980), pp. 345-370.

[27] B. T. POLYAK AND N. V. TRET'YAKOV, *The method of penalty estimates for conditional extremum problems*, Zh. Vychisl. Mat. i Mat. Fiz., 13 (1973), pp. 34-46 (translated in U.S.S.R. Comput. Math. and Math. Phys., 13 (1973), pp. 42-58).

[28] M. J. D. POWELL, *A method for nonlienar constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, 1972.

[29] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.

[30] ———, *The multiplier method of Hestenes and Powell applied to convex programming*, J. Optim. Theory Appl., 12 (1973), pp. 555-562.

[31] ———, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Math. Programming, 5 (1973), pp. 354-373.

[32] N. Z. SHOR, *Minimization Methods for Nondifferentiable Functions*, Springer-Verlag, Berlin, New York, 1985.

[33] F. A. VALENTINE, *The problem of Lagrange with differential inequalities as added side conditions*, in Contributions to the Calculus of Variations, Univ. of Chicago Press, Chicago, 1937, pp. 407-448.

[34] A. VARDI, *A new minimax algorithm*, Report no. 84-25, Institute for Computer Applications in Science and Engineering, Hampton, VA, 1984.