FUNDAMENTALS OF OPTIMIZATION

LECTURE NOTES

2007

R. T. Rockafellar

Dept. of Mathematics

University of Washington

Seattle

CONTENTS

1. WHAT IS OPTIMIZATION?

Optimization problem: Maximizing or minimizing some function relative to some set, often representing a range of choices available in a certain situation. The function allows comparison of the different choices for determining which might be "best."

Common applications: Minimal cost, maximal profit, best approximation, optimal design, optimal management or control, variational principles.

Goals of the subject: Understanding the practical and theoretical aspects of:

Modeling issues: What to look for in setting up an optimization problem? What features are advantageous or disadvantageous? What aids to formulation are available? How can problems usefully be categorized?

Analysis of solutions: What is meant by a "solution?" When do solutions exist, and when are they unique? How can solutions be recognized and characterized? What happens to solutions under perturbations?

Numerical methods: How can solutions be determined by iterative schemes of computation? What modes of local simplification of a problem are appropriate? How can different solution techniques be compared and evaluated?

Distinguishing features of optimization as a mathematical discipline:

Descriptive math $\longrightarrow$ prescriptive math: Much of mathematics in the past has been devoted to describing how systems behave, e.g. in the laws of physics. The advent of computers, however, has brought an emphasis on using mathematics to make systems behave in *chosen* ways, and choices lead to questions of optimization.

Equations $\longrightarrow$ inequalities: Optimization typically deals with variables that have to lie within certain ranges, dictated by the limitations of what choices are allowable, and this leads to a dominance of relationships that are expressed by inequalities rather than equations.

Linear/nonlinear $\longrightarrow$ convex/nonconvex: The familiar division between linearity and nonlinearity is less important in optimization than the one between convexity and nonconvexity, which for its appreciation requires a new investment in concepts.

Differential calculus $\longrightarrow$ subdifferential calculus: The prevalence of inequalities, along with the special properties of "max" and "min" as operations, raise the need for a methodology that doesn't rely so much as classical mathematics on supposing surfaces to be smooth and functions to be differentiable everywhere.

Finite-dimensional optimization: The case where a choice corresponds to selecting the values of a finite number of real variables, called *decision variables*. For purposes of general discussion, such as now, the decision variables may be denoted by $x_1, \ldots, x_n$ and each allowable choice therefore identified with a point $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$.

Feasibility: The points x representing allowable choices are said to be *feasible*. They form the subset of $\mathbb{R}^n$ over which the maximization or minimization takes place.

Constraints: Conditions on the decision variables that are used to specify the set of feasible points x in $\mathbb{R}^n$.

Equality and inequality constraints: Conditions of the form $f_i(x) = c_i$, $f_i(x) \leq c_i$ or $f_i(x) \geq c_i$ for certain functions f_i on $\mathbb{R}^n$ and constants c_i in $\mathbb{R}$. (Note: Strict inequalities are avoided in constraints, for reasons which will be clear later.)

Range constraints: Conditions restricting the values of some decision variables to lie within certain closed intervals of $\mathbb{R}$. Very important in many situations, for instance, are *nonnegativity constraints*: some variables x_j may only be allowed to take values ≥ 0; the interval then is $[0, \infty)$. Range constraints can also arise from the desire to keep a variable between certain upper and lower bounds.

Linear constraints: This covers range constraints and conditions $f_i(x) = c_i$, $f_i(x) \leq c_i$, or $f_i(x) \geq c_i$, in which the function f_i is *linear*—in the standard sense of being expressible as sum of constant coefficients times the variables $x_1, \ldots, x_n$.

Data parameters: Problem statements usually involve not only decision variables but symbols designating "given" constants and coefficients. Conditions on such elements, such as the nonnegativity of a particular coefficient, are *not* among the "constraints" in a problem of optimization, however crucial they may be from some other perspective.

Knowns and unknowns: A good way to think about the important distinction between decision variables and data parameters is as follows. This may help you through confusing situations where certain "parameters" might actually be decision variables and certain "variables" actually data parameters. In practice, of course, all kinds of symbols besides x get used, and it can get hard to tell things apart.

The decision variables in an optimization problem are unknowns that are open to manipulation in the process of maximization or minimization, whereas the data parameters aren't open to manipulation when it comes to solving a particular problem, but instead would be furnished with specific numerical values as inputs to any solution procedure.

Mathematical programming: This is a synonym for finite-dimensional optimization. Its usage predates "computer programming," which actually arose from attempts at solving optimization problems on early computers. "Programming," in the sense of optimization, survives in problem classifications such as linear programming, quadratic programming, convex programming, integer programming, and so forth.

EXAMPLE 1: Engineering design

General description: In the design of some object, system or structure, the values of certain parameters can be chosen subject to some conditions expressing their ranges and interrelationships. The choice determines the values of a number of other variables on which the desirability of the end product depends, such as cost, weight, speed, bandwidth, reliability, ... Among the choices of the design parameters that meet certain performance specifications, what's "best" by some overall criterion?

An illustration, optimal proportions of a can: The following example, although toy-sized, brings out many features of optimization modeling. A cylindrical can of a given volume V_0 is to be proportioned in such a way as to minimize the total cost of the material in a box of 12 cans, arranged in a 3×4 pattern. The cost expression takes the form $c_1 S_1 + c_2 S_2$, where S_1 is the surface area of the 12 cans and S_2 is the surface area of the box. The coefficients c_1 and c_2 are nonnegative. A side requirement is that no dimension of the box can exceed a given size D_0.

design parameters: $r =$ radius of can, $h =$ height of can
volume constraint: $\pi r^2 h = V_0 \quad$ (or $\geq V_0$, see below!)
surface area of cans: $S_1 = 12(2\pi r^2 + 2\pi r h) = 24\pi r(r + h)$
box dimensions: $8r \times 6r \times h$
surface area of box: $S_2 = 2(48r^2 + 8rh + 6rh) = 4r(24r + 7h)$
size constraints: $8r \leq D_0, \quad 6r \leq D_0, \quad h \leq D_0$
nonnegativity constraints: $r \geq 0, \quad h \geq 0 \quad$ (!)

Summary: The decision variables are r and h, and the choices that are available can be identified with the set $C \subset \mathbb{R}^2$ consisting of all (r, h) satisfying the conditions

$$r \geq 0, \quad h \geq 0, \quad 8r \leq D_0, \quad 6r \leq D_0, \quad h \leq D_0, \quad \pi r^2 h = V_0.$$

The first five conditions can be thought of as range constraints and the sixth as an equality constraint on $f_1(r, h) = \pi r^2 h$. Over the set C we wish to minimize

$$f_0(r, h) = c_1\big[24\pi r(r + h)\big] + c_2\big[4r(24r + 7h)\big] = d_1 r^2 + d_2 rh,$$

where $d_1 = 24\pi c_1 + 96c_2$ and $d_2 = 24\pi c_1 + 28c_2$. The symbols V_0, D_0, c_1 and c_2, and ultimately d_1 and d_2, are data parameters. Although $c_1 \geq 0$ and $c_2 \geq 0$, these aren't "constraints" in the problem. As for S_1 and S_2, they were only introduced as temporary symbols and didn't end up as decision variables.

Redundant constraints: It is obvious that the condition $6r \leq D_0$ is implied by the other constraints and therefore could be dropped without affecting the problem. But in problems with many variables and constraints such redundancy may be hard to recognize. The elimination of redundant constraints could pose a practical challenge as serious as that of solving the optimization problem itself.

Inactive constraints: It could well be true that the optimal pair (r, h) (unique??) is such that either the condition $8r \leq D_0$ or the condition $h \leq D_0$ is satisfied as a strict inequality, or both. In that case the constraints in question are inactive in the local characterization of optimal point, although they do affect the shape of the set C. Again, however, there is little hope, in a problem with many variables and constraints, of determining by some preliminary procedure just which constraints will be active and which will not. This is the crux of the difficulty in many numerical approaches.

Redundant variables: It would be possible to solve the equation $\pi r^2 h = V_0$ for h in terms of r and thereby reduce the given problem to one in terms of just r, rather than (r, h). Fine—but besides being a technique that is usable only in special circumstances, the elimination of variables from (generally nonlinear) systems of equations is not necessarily helpful. There may be a trade-off between the lower dimensionality achieved in this way and other properties.

Inequalities versus equations: The constraint $\pi r^2 h = V_0$ could be written in the form $\pi r^2 h \geq V_0$ without affecting anything about the solution. This is because of the nature of the cost function; no pair (r, h) in the larger set C', obtained by substituting this weaker condition for the equation, can minimize f_0 unless actually $(r, h) \in C$. While it may seem instinctive to prefer the equation to the inequality in the formulation, the inequality turns out to be superior in the present case because the set C' happens to be "convex," whereas C isn't.

Convexity: Even with the reformulation just suggested, the problem wouldn't be fully of convex type because the function of r and h being minimized isn't itself "convex"; further maneuvers might get around that. The lesson is that the formulation of a problem can be subtle, when it comes to bringing out features of importance in optimization. Knowledge and experience play a valuable role.

EXAMPLE 2: Utilization of Resources

General description: Actions are to be taken that hopefully will result in profit, but these actions, which correspond to selecting the values of a number of variables, draw on some limited resources. What is the best way to allocate the resources so as to maximize the profit, or for that matter to optimize the actions with respect to some other criterion? There are many variants, such as using resources to meet given needs while minimizing cost or pollution. In this class of problems inequality constraints are especially prominent.

An illustration, tourist sales: We keep to toy reality for the purpose of concentrating better on key aspects of optimization modeling. The summer Olympic Games are coming to town, and a teenage entrepreneur is planning to make money off tourists by selling them souvenir sun hats and umbrellas. He can get plain sun hats and umbrellas at costs of c_1 and c_2 dollars each, respectively, and then use a kit he has in order to imprint them with a souvenir logo. With the logo he can sell them for p_1 and p_2 dollars each, in unlimited quantity.

He does have limitations from two directions, however. First, he has only k dollars to invest in the venture. Second, he has to store his items between the time he can get them and the time of the Games. He has a total of b cubic meters of space that he can use for free in his own basement, and there's an extra e cubic meters of space available in his neighbor's house—but to the extent that he uses that he'll have to pay d dollars per cubic meter over the storage period. Sun hats and umbrellas take up s_1 and s_2 cubic meters of storage space each, respectively. How many sun hats and umbrellas should he stock up, and how much space should he rent from the neighbor, if any? The "resource" limits here are in the capital k, the free space b, and the rental space e.

actions: $x_1 =$ hats ordered, $x_2 =$ umbrellas ordered, $x_3 =$ space rented

range constraints: $x_1 \geq 0,\ x_2 \geq 0,\ 0 \leq x_3 \leq e$

storage constraint: $s_1x_1 + s_2x_2 \leq b + x_3$

investment constraint: $c_1x_1 + c_2x_2 + dx_3 \leq k$

profit expression: $(p_1 - c_1)x_1 + (p_2 - c_2)x_2 - dx_3$

Summary: The decision variables are x_1, x_2 and x_3. Besides the range constraints there are two inequality constraints in terms of functions of these variables—this is better seen by writing the storage requirement in the form $s_1x_1 + s_2x_2 - x_3 \leq b$. All these constraints together determine a feasible set of points $(x_1, x_2, x_3) \in \mathbb{R}^3$. Over this set the function given by the profit expression is to be maximized.

Linearity: Note that in this formulation a linear function of the decision variables is being maximized subject to a system of linear constraints.

Continuous versus integer variables: It would make sense to restrict the variables x_1 and x_2 to have whole values only, since there's no meaning to selling a fraction of a hat or umbrella. That would have major effects on the ease with which the problem could be solved, however. It's customary therefore to represent discrete quantities by continuous variables, as a practical simplification as long as the magnitudes are large relative to the "quantum of discreteness."

In some situations we couldn't take that view. If the optimization problem involved a decision about whether to build a shopping center or not, and we modeled that as a *zero-one variable* (no=0, yes=1), we surely would want to steer away from fractions.

Integer programming: Problems with integer variables constitute a special, more difficult area of optimization which we won't be treating.

Breaks in cost structure: The way storage costs have been treated in this example deserves close attention. We could have set the problem up instead with x_1, x_2 and a decision variable y standing for the amount of storage space to be used; the relation between these variables would then be handled as an equation constraint, $s_1x_1 + s_2x_2 - y = 0$. Then y would be restricted to the interval $[0, b+e]$, and the profit expression to be maximized would be $(p_1-c_1)x_1+(p_2-c_2)x_2-g(y)$, where $g(y)$ denotes the storage cost for y cubic meters. What would the function g look like? We would have $g(y) = 0$ for $0 \le y \le b$ but $g(y) = d[y-b]$ for $b \le y \le b+e$. It wouldn't be linear, just "piecewise linear" with a breakpoint at $y = b$, where it would have a kink and not be differentiable.

In contrast, the formulation above achieved linearity. How? That was done by widening the range of possible actions. The entrepreneur was permitted (in the variable x_3) to rent space from his neighbor even if, on the basis of the quantities x_1 and x_2 he had chosen, he didn't need all of it. No constraint was added to prevent squandering money on useless space. Allowing such squandering was harmless because it would be eliminated in the course of *optimizing*.

Extension to uncertain prices, a worst-case model: Often in planning for the future there are uncertainties about some of the data parameters. That's a major issue which we can't get into here, but a simple tactic is worth discussing now for its mathematical features, which often come up in other circumstances as well. The tourist sales example is a good vehicle for explaining this.

The prices p_1 and p_2 for the sun hats and umbrellas could depend on how the weather turns out at the time of the Games. Let's imagine p_1 and p_2 correspond to normal weather, but there are two other possibilities: very sunny or very rainy. In the very sunny case the two prices would instead be p_1' and p_2' (higher for hats, lower for umbrellas), whereas in the very rainy case they would be p_1'' and p_2'' (lower for hats, higher for umbrellas). Then the entrepreneur would in fact face three possible sales outcomes, given by the expressions $p_1x_1 + p_2x_2$, $p_1'x_1 + p_2'x_2$ and $p_1''x_1 + p_2''x_2$. There might be a statistical approach to coping with this, but an alternative is to maximize, subject to the given constraints, the function

$$f(x_1, x_2, x_3) = \min\{p_1x_1 + p_2x_2,\ p_1'x_1 + p_2'x_2,\ p_1''x_1 + p_2''x_2\} - c_1x_1 - c_2x_2 - dx_3.$$

This function gives the worst of all the possible profit outcomes that could result from a particular choice of x_1, x_2 and x_3. (Note that "min" refers merely to the numerical operation of taking the lowest value from a collection of numbers, here a collection of three numbers which happen to depend on x_1, x_2 and x_3.)

Nonlinearity: Even though the function f is built up from linear expressions, it's not itself linear. Moreover it has kinks at which its differentiability fails.

Linearizing device for worst-case models: This example can be reformulated to get rid of the nonlinearity. The trick is to introduce another variable, u say, and think of the problem as one in x_1, x_2, x_3 and u in which the old constraints are kept, but there are new constraints $p_1x_1 + p_2x_2 - u \geq 0$, $p_1'x_1 + p_2'x_2 - u \geq 0$ and $p_1''x_1 + p_2''x_2 - u \geq 0$, and the expression to maximize is $u - c_1x_1 - c_2x_2 - dx_3$.

EXAMPLE 3: Management of Systems

General description: A sequence of decisions must be made in discrete time which will affect the operation of some kind of "system," often of an economic nature. As in Example 2, the decisions, each in terms of choosing the values of a number of variables, have to respect limitations in resources, and the aim is to minimize cost or maximize profit or efficiency, say, but over a particular time horizon.

An illustration, inventory management: A warehouse with total capacity a (in units of volume) is to be operated over time periods $t = 1, \ldots, T$ as the sole facility for the supply of a number of different commodities (or medicines, or equipment parts, etc.), indexed by $j = 1, \ldots, n$. The demand for commodity j during period t is the known amount $d_{tj} \geq 0$ (in volume units)—this is a *deterministic* approach

to modeling the situation. In each period t it is possible not only to fill demands but to acquire additional supplies up to certain limits, so as to maintain stocks. The problem is to plan the pattern of acquiring supplies in such a way as to maximize the net profit over the T periods, relative to the original inventory amounts and the desired terminal inventory amounts.

inventory variables: x_{tj} units of commodity j at the end of period t

inventory constraints: $x_{tj} \geq 0, \quad \sum_{j=1}^{n} x_{tj} \leq a$ for $t = 1, \ldots, T$

initial inventory: x_{0j} units of j given at the beginning

terminal constraints: $x_{Tj} = b_j$ (given amounts) for $j = 1, \ldots, n$

inventory costs: s_{tj} dollars per unit of j held from t to $t+1$

supply variables: u_{tj} units of j acquired during period t

supply constraints: $0 \leq u_{tj} \leq a_{tj}$ (given availabilities)

supply costs: c_{tj} dollars per unit of j acquired during t

dynamical constraints: $x_{tj} = \max\{0,\, x_{t-1,j} + u_{tj} - d_{tj}\}$

rewards: p_{tj} dollars per unit of filled demand

filled demand: $\min\{d_{tj},\, x_{t-1,j} + u_{tj}\}$ units of j during period t

net profit: $\sum_{t=1}^{T} \sum_{j=1}^{n} \left(p_{tj} \min\{ d_{tj},\, x_{t-1,j} + u_{tj} \} - s_{tj} x_{tj} - c_{tj} u_{tj} \right)$

Summary: The latter expression, as a function of x_{tj} and u_{tj} for $t = 1, \ldots, T$ and $j = 1, \ldots, n$ (these being the decision variables), is to be maximized subject to the inventory constraints, terminal constraints, supply constraints and the dynamical constraints, which are regarded as determining a feasible set of points in $\mathbb{R}^{2Tn}$. The symbols d_{tj}, a, x_{0j}, b_j, s_{ij}, a_{tj}, c_{tj} and p_j stand for data parameters.

Large-scale context: The number of variables and constraints that can be involved in a problem may well be very large, and the interrelationships may be too complex to appreciate in any direct manner. This calls for new ways of thinking and for more reliance on guidelines provided by theory.

Uncertainty: Clearly, the assumption that the demands d_{tj} are known precisely in advance is unrealistic for many applications, although by solving the problem in this case one might nonetheless learn a lot. To pass from deterministic modeling to *stochastic* modeling, where each d_{tj} is a random variable (and the same perhaps for other data elements like a_{tj}), it is necessary to expand the conceptual horizons considerably. The decision vector $(u_{t1}, \ldots, u_{tn})$ at time t must be viewed as an *unknown function* of the "information" available to the decision maker at that time, rather than just at the initial time, but this type of optimization is beyond

us here. A worst-case approach could be taken, although this likewise would seriously raise the level of complication.

Dependent variables: The values of the variables x_{tj} are determined by the values of the variables u_{tj} for $t = 1, \ldots, T$ and $j = 1, \ldots, n$ through the dynamical equations and the initial values. In principal, therefore, a specific expression in the latter variables could be substituted for each x_{tj}, and the dimensionality of the problem could thereby be cut sharply. But this trick, because it hides basic aspects of structure, could actually make the problem harder to analyze and solve.

Constraints versus penalties: The requirements that $\sum_{j=1}^{n} x_{tj} \leq a$ for $t = 1, \ldots, T$ and that $x_{Tj} = b_j$, although innocent-looking, are troublesome. Better modeling would involve some recourse in the eventuality of these conditions not being satisfied. For instance, instead of a constraint involving the capacity one could incorporate into the function being minimized a penalty term, which kicks in when the total inventory being stored rises above a (perhaps with the interpretation that extra storage space has to be rented).

Max and min operations: The "max" operation in the dynamics and the "min" operation in the expression of the net profit force the consideration of functions that aren't differentiable everywhere and thus don't submit to ordinary calculus. Sometimes this is unavoidable and points to the need for fresh developments in analysis. Other times it can be circumvented by reformulation. The present example fits with the latter. Really, it would be better to introduce more variables, namely v_{tj} as the amount of good j used to meet demands at time t. In terms of these additional variables, constrained by $0 \leq v_{tj} \leq d_{tj}$, the dynamics turn out to be *linear*,

$$x_{tj} = x_{t-1,j} + u_{tj} - v_{tj},$$

and so too does the profit expression, which is now

$$\sum_{t=1}^{T} \sum_{j=1}^{n} \big(p_{tj} v_{tj} - s_{tj} x_{tj} - c_{tj} u_{tj} \big).$$

Hidden assumptions: The alternative model just described with variables v_{tj} is better in other ways too. The original model had the hidden assumption that demands in any period should always be met as far as possible from the stocks on hand. But this might be disadvantageous if rewards will soon be higher but inventory can only be built up slowly due to the constraints on availabilities. The alternative model allows sales to be held off in such circumstances.

EXAMPLE 4: Identification of Parameter Values

General description: It's common in science and engineering to model behavior by a mathematical law, like an equation, which however can't be implemented without specifying the values of various parameters ("constants") that appear in it. A basic task is to determine the parameter values that provide the best fit to the available data, known through experiment or observation. This is central to statistics (regression, maximum likelihood), econometrics, error analysis, etc.

In speaking of "best" fit, reference is evidently being made to some criterion for optimization, but there isn't always just one. Note also a linguistic pitfall: "the" best fit suggests uniqueness of the answer being sought, but even relative to a single criterion there might be a tie, with different answers equally good.

This kind of optimization is entirely technical: the introduction of something to be optimized is just a mathematical construct. Still, in analyzing and computing solutions the challenges are the same as in other areas of optimization.

An illustration, "least squares" estimates: Starting out very simply, suppose that two variables x and y are being modeled as related by a linear law $y = ax+b$, either for inherent theoretical reasons or as a first-level approximation. The values of a and b are not known *a priori* but must be determined from the data, consisting of a large collection of pairs $(x_k, y_k) \in \mathbb{R}^2$ for $k = 1, \ldots, N$. These pairs have been gleaned from experiments (where random errors of measurement could arise along with other discrepancies due to oversimplifications in the model). The error expression

$$E(a,b) = \sum_{k=1}^{N} \big| y_k - (ax_k + b) \big|^2$$

is often taken as representing the goodness of the fit of the parameter pair (a, b). The problem is to minimize this over all $(a, b) \in \mathbb{R}^2$.

Note that, from the *optimization* perspective, a and b are decision variables, whereas the symbols x_k and y_k stand for data parameters. Words can be slippery!

Multidimensional extension: More generally, instead of a real variable x and a real variable y one could be dealing with a vector $x \in \mathbb{R}^n$ and a vector $y \in \mathbb{R}^m$, which are supposed to be related by a formula $y = Ax+b$ for a matrix $A \in \mathbb{R}^{m\times n}$ and a vector $b \in \mathbb{R}^m$. Then the error expression $E(A, b)$ would depend on the $m \times (n + 1)$ components of A and b.

Incorporation of constraints: The problem, as stated so far, concerns the *unconstrained* minimization of a certain quadratic function of a and b, but it's easy to find

situations where the parameters might be subject to side conditions. For instance, it may be known on the basis of theory for the model in question that $1/2 \leq a \leq 3/2$, while $b \geq -1$. In the multidimensional case of $y = Ax + b$, there could be conditions on A such as positive semidefiniteness. (In the applications that make use of least squares estimation, such conditions are sometimes neglected by practitioners, and the numerical answer obtained is simply "fixed up" if it doesn't have the right form. But that's clearly not good methodology.)

Nonlinear version: A so-called problem of *linear* least squares has been presented, but the same ideas can be used when the underlying relation between x and y is supposed to be nonlinear. For instance, a law of the form $y = e^{ax} - e^{bx}$ would lead to an error expression

$$E(a,b) = \sum_{k=1}^{N} \left| y_k - (e^{ax_k} - e^{bx_k}) \right|^2.$$

In minimizing this with respect to $(a,b) \in \mathbb{R}^2$, we would not be dealing with a quadratic function, but something much more complicated. The graph of E in a problem of nonlinear least squares could have lots of "bumps" and "dips," which could make it hard to find the minimum computationally.

Beyond squares: Many other expressions for error could be considered instead of a sum of squares. Back in the elementary case of $y = ax + b$, one could look at

$$E(a,b) = \sum_{k=1}^{N} \left| y_k - (ax_k + b) \right|.$$

A different (a,b) would then be "best." The optimization problem would have a technically different character as well, because E would lack differentiability at points (a,b) where $y_k - (ax_k + b) = 0$ for some k. Attractive instead, as a *worst case* approach, would be the error expression

$$E(a,b) = \max_{k=1,\ldots,N} \left| y_k - (ax_k + b) \right|.$$

The formula in this case means that the value assigned by E to the pair (a,b) is the largest value occurring among the errors $|y_k - (ax_k + b)|$, $k = 1, \ldots, N$. It's this maximum deviation that we wish to make as small as possible. Once more, E isn't a differentiable function on $\mathbb{R}^2$.

Tricks of reformulation: When $E(a,b)$ is given by the max, the problem of minimizing it can be simplified by introducing an another variable u and trying to minimize its value over the combinations $(a,b,u) \in \mathbb{R}^3$ such that $u \geq |y_k - (ax_k - b)|$ for $k = 1, \ldots, N$. Each inequality $u \geq |y_k - (ax_k - b)|$ can moreover be replaced by $u - y_k + ax_k - b \geq 0$ and $u + y_k - ax_k + b \geq 0$. Similarly, when $E(a,b) = \sum_{k=1}^{N} |y_k - (ax_k + b)|$ one can minimize $u_1 + \cdots + u_N$ over all $(a, b, u_1, \ldots, u_N)$ satisfying $u_k \geq |y_k - (ax_k + b)|$ for $k = 1, \ldots, N$.

Inverse problems: This term is often used for parameter identification problems that involve differential equations. There is a model of behavior in which a differential equation dictates what should happen, i.e., what "outputs" should be occur in response to given "inputs," but the coefficients in the equation aren't fully known and have to be inferred from experimental data.

Geological investigations: An example is that of exploring subterranean strata by sending signals (as inputs) which pass through the earth and, after being modified by that process, are picked up (as outputs) by sensors. The "coefficients" in this case are physical parameters, like densities, that describe what the underlying rocks are like and where the layers begin and end.

Tomography: A similar application is found in the medical technology of trying to reconstruct the internal geometry of a patient's body from data that has been collected by passing x-rays through the body from different angles.

Image reconstruction or enhancement: Data received through radar or a satellite camera may be fuzzy, distorted or marred by signal interference. How can one use it to best advantage in approximating the true image that gave rise to it? This is a kind of parameter identification problem in which the parameters measure the shades of grayness of the pixels in an image.

EXAMPLE 5: Variational Principles

General description: The kinds of equations that are the focus of much of numerical analysis are often associated in hidden ways with problems of optimization. A *variational principle* for an equation $F(x) = 0$, involving a mapping $F : \mathbb{R}^n \mapsto \mathbb{R}^n$, is an expression of F as the gradient mapping ∇f associated with function $f : \mathbb{R}^n \mapsto \mathbb{R}$. It leads to the interpretation of the desired x as satisfying a first-order optimality condition with respect to f. Sometimes there are reasons to conclude that that the equation's solution actually minimizes f, at least "locally." Then a way of solving $F(x) = 0$ by optimization is opened up.

Quite similar in concept are numerous examples where one wishes to solve an equation $A(u) = 0$ where u is some unknown *function* and A is a differential operator, so that an ordinary or partial differential equation is at issue. A variational principle characterizes the desired u as providing the minimum, say, of some expression. Many equations of physics have such an interpretation.

On a different front, conditions of price equilibrium in economics can sometimes be characterized as stemming from the actions of a multitude of "economic agents," like producers and consumers, all optimizing from their own perspectives. Yet again, the equilibrium state following the reactions which take place in a complicated chemical brew may be characterized through a variational principle as the configuration of substances that minimizes a certain energy function.

Status in optimization: In the study of variational principles, optimization theory can provide interesting insights quite independently of whether a numerical solution to a particular case is sought or not.

EXAMPLE 6: Optimal Control

General description: The evolution of a system in continuous time t can often be characterized by an ordinary differential equation $\dot{x}(t) = f(t, x(t))$ with $x(0) = x_0$ (initial condition), where $\dot{x}(t) = (dx/dt)(t)$. (Equations in higher derivatives are typically reducible to ones of first order.) Here $x(t)$, called the *state* of the system, is a point in $\mathbb{R}^n$. This is *descriptive* mathematics. We get *prescriptive* mathematics when the ODE involves parameters for which the values can be chosen as a function of time: $\dot{x}(t) = f(t, x(t), u(t))$, where $u(t) \in U \subset \mathbb{R}^m$.

Without going into the details necessary to provide a rigorous foundation, the idea can be appreciated that under certain assumptions there will be a mapping which assigns to each choice of a *control* function $u(\cdot)$ over a time interval $[0, T]$ a corresponding state trajectory $x(\cdot)$. Then, subject to whatever restrictions may be necessary or desirable on these functions, one can seek the choice of $u(\cdot)$ which is optimal according to some criterion. Such a problem would be infinite-dimensional, but a finite-dimensional version would arise as soon as the differential equation is approximated by a difference equation in discrete time.

Stochastic version: The system may be subject to random disturbances which the controller must react to. Further, there may be difficulty in knowing exactly what the state is at any time t, due to measurement errors and shortcomings of the sensors. Control must be framed in terms of feedback mappings, giving the response at time t to the information available right then about $x(t)$.

Adaptive version: Also intriguing as a mathematical challenge, but largely out of reach of current concepts and techniques, is *adaptive control*, where the controller has not only to react to events but learn the basics of the system being controlled as time goes on. A major difficulty in this area is deciding what to optimize and for that matter what can or can't be assumed about the imperfectly known system.

Control of PDE's: The state of a system may be given by an element of a function space rather than a point in $\mathbb{R}^n$, as for instance when the problem revolves around the temperature distribution at time t over a solid body represented by a closed, bounded region $\Omega \subset \mathbb{R}^3$. The temperature can be influenced by heating or cooling elements arrayed on the surface of the body. How should these elements be operated in order to bring the temperature of the body uniformly within a certain range—in the shortest possible time, or with the least expenditure of energy?

EXAMPLE 7: Combinatorial optimization

General description: Many problems involve the optimal ordering or arrangements of discrete objects or actions, or either-or choices which might be represented by zero-one variables. Examples are found in the scheduling of processing jobs in manufacturing, or the scheduling of airline flights and crews, but also in shortest-path problems and the like.

Models involving networks (directed graphs) are often useful in this regard and can produce enormous simplifications, but in general such optimization problems may be hard to solve and even intractable. Work on them goes on nonetheless because of their practical importance. Success is often measured in heuristic schemes that at least are able to generate improvements over the status quo.

Overview of where we are now headed: Because we'll be concentrating on finite-dimensional optimization (as in Examples 1, 2, 3, and 4), no infinite-dimensional applications to variational principles (as in Example 5) or optimal control (as in Example 6) will be covered directly, nor will special tools for combinatorial optimization (as in Example 7) be developed. The ideas we'll develop are highly relevant, though, as background for such other areas of optimization. In particular, infinite-dimensional problems are often approximated by finite-dimensional ones through some kind of discretization of time, space, or probability.

2. PROBLEM FORMULATION

To set the stage for solving a problem of optimization, it's necessary first to formulate it in a manner not only reflecting the situation being modeled, but so as to be amenable to computational techniques and theoretical analysis. This raises a number of fundamental issues, which range from the problem format to be adopted to criteria for when a problem is "well posed."

Basic problem: Minimize a function $f_0 : \mathbb{R}^n \to \mathbb{R}$, the *objective function*, over a specified set $C \subset \mathbb{R}^n$, the *feasible set*.

Max versus min: Maximizing a function g is equivalent to minimizing $-g$, so there's no loss of generality in concentrating on minimization. This is the convention in much of optimization theory.

Solution concepts. Different things can be sought in a problem of optimization. The following terms identify the main concepts.

Feasible solution: Any point x that belongs to C, regardless of the value it gives to f_0. Just finding such a point could be difficult numerically in cases where the constraints are complicated or numerous, and indeed, the very existence of a feasible solution may sometimes be an open question. This is just a first-level solution concept, but important nevertheless.

Optimal solution: A point $\bar{x}$ furnishing the minimum value of f_0 over C, i.e., a feasible solution such that $f_0(\bar{x}) \leq f_0(x)$ for all other feasible solutions x. This is more specifically what is called a *globally* optimal solution, when contrast must be made with the next concept.

Locally optimal solution: A point $\bar{x} \in C$ such that, for some neighborhood U of $\bar{x}$, one has $f_0(\bar{x}) \leq f_0(x)$ for all $x \in C \cap U$. Optimality in this case isn't asserted relative to C as a whole, but only relative to a sufficiently small ball U around $\bar{x}$. In practice it may be very hard to distinguish whether a numerical method has produced a globally optimal solution or just a locally optimal one, if that much.

Optimal set: The set of all (globally) optimal solutions (if any).

Optimal value: The greatest lower bound to the values of $f_0(x)$ as x ranges over C. There may or may not be a point $\bar{x} \in C$ at which f_0 actually attains this value. Furthermore, although the optimal value is always well defined, it could fail to be finite. It is $-\infty$ when f_0 is not bounded below on C, and on the other hand, it is ∞ by convention if $C = \emptyset$.

Constraint manipulation: Constraints can be expressed in more than one way, and some forms of expression may be more convenient in one context than another.

Function constraints: In conditions like $f_i(x) = c_i$, or $f_i(x) \leq c_i$, or $f_i(x) \geq c_i$, f_i is called a *constraint function*.

(1) An equality constraint $f_i(x) = c_i$ can be expressed equivalently, if desired, as a pair of inequality constraints: $f_i(x) \leq c_i$ and $f_i(x) \geq c_i$.

(2) An inequality constraint $f_i(x) \geq c_i$ can be expressed also as $-f_i(x) \leq -c_i$.

(3) An inequality constraint $f_i(x) \leq c_i$ can be expressed also as $-f_i(x) \geq -c_i$.

(4) An inequality constraint $f_i(x) \leq c_i$ can be expressed as an equality constraint $f_i(x) + s_i = c_i$ involving an additional decision variable s_i, itself constrained to be nonnegative. Such a variable is called a *slack variable*.

(5) Any constraint $f_i(x) = c_i$, or $f_i(x) \leq c_i$, or $f_i(x) \geq c_i$, can be expressed in terms of $g_i(x) = f_i(x) - c_i$ as $g_i(x) = 0$, or $g_i(x) \leq 0$, or $g_i(x) \geq 0$.

Geometric, or abstract constraints: For methodological purposes it's often convenient to represent only some of the constraints in a problem in terms of constraint functions f_i and to lump the rest together in the abstract form $x \in X$. This is especially handy for treating range constraints.

Example: For instance, a requirement on $x = (x_1, \ldots, x_n)$ that $0 \leq x_1 \leq 1$ *could* be represented by two function constraints $g_1(x) \geq 0$ and $g_2(x) \leq 1$ with $g_1(x) = g_2(x) = 1 \cdot x_1 + 0 \cdot x_2 + \cdots + 0 \cdot x_n$, but it could also be incorporated into the description of a set X to which x must belong.

Boxes: A set $X \subset \mathbb{R}^n$ is a *box* if it is a product $I_1 \times \cdots \times I_n$ of closed intervals $I_j \subset \mathbb{R}$. To require $x \in X$ is to require $x_j \in I_j$ for $j = 1, \ldots, n$. This is just what it means to have *range constraints* on x. Here I_j could be bounded or even consist of just one point, or it could be unbounded or even $(-\infty, \infty)$.

Nonnegative orthant: the box $\mathbb{R}^n_+ = [0, \infty) \times \cdots \times [0, \infty)$. For $X = \mathbb{R}^n_+$, the constraint $x \in X$ means that the components x_j of x have to be nonnegative.

Whole space: the box $\mathbb{R}^n = (-\infty, \infty) \times \cdots \times (-\infty, \infty)$. For $X = \mathbb{R}^n$, the constraint $x \in X$ trivializes; each component x_j of x is "free."

Linear and affine functions: A function g on $\mathbb{R}^n$ is called *affine* if it can be expressed in the form $g(x_1, \ldots, x_n) = d_0 + d_1 x_1 + \cdots + d_n x_n$ for some choice of constants $d_0, d_1, \ldots, d_n$. Many people simply refer to such a function as linear, and in this they are following a long tradition, but in higher mathematics the term *linear* is reserved

for the special case of such a function where the constant term vanishes: $d_0 = 0$. Thus, g is linear when there's a vector $d = (d_1, \ldots, d_n) \in \mathbb{R}^n$ such that

$$g(x) = d \cdot x \qquad \text{(the inner product, or dot product, of two vectors)}$$

Linear constraints: Conditions $f_i(x) = c_i$, $f_i(x) \leq c_i$ or $f_i(x) \geq c_i$ in which the function f_i is linear—or affine. Or, conditions $x \in X$ in which X is a box.

Standard problem format in finite-dimensional optimization:

$$(\mathcal{P}) \qquad \begin{array}{l} \text{minimize } f_0(x) \text{ over all } x = (x_1, \ldots, x_n) \in X \subset \mathbb{R}^n \text{ satisfying} \\ f_i(x) \begin{cases} \leq 0 & \text{for } i = 1, \ldots, s, \\ = 0 & \text{for } i = s+1, \ldots, m. \end{cases} \end{array}$$

The feasible set C for $(\mathcal{P})$ consists of all the points $x \in X$ that satisfy all the constraints $f_i(x) \leq 0$ or $f_i(x) = 0$. Here in particular X could be all of $\mathbb{R}^n$, in which case the condition $x \in X$ would impose no restriction whatever. More typically, however, X might be chosen to be some box other than $\mathbb{R}^n$, in which case the requirement that $x \in X$ is a way of representing range constraints on x that underlie $(\mathcal{P})$.

Unconstrained minimization: the case where $X = \mathbb{R}^n$ and "$m = 0$," i.e., no equality or inequality constraints are present, so that $C = \mathbb{R}^n$.

Linear programming: the case where a linear (or affine) function f_0 is minimized subject to linear constraints: the functions $f_1, \ldots, f_m$ are affine and the set X is a box (e.g. $X = \mathbb{R}^n$ or $X = \mathbb{R}^n_+$).

Quadratic programming: like linear programming, but the objective function f_0 is allowed to have quadratic terms, as long as it remains *convex*, as defined later. (Note: in quadratic programming the constraints are still only linear!)

Nonlinear programming: this term is used in contrast to linear programming, but a much more important watershed will eventually be seen in the distinction between *convex* programming and *nonconvex* programming.

Geometric considerations: In problems with a few, simple constraints, the feasible set C might be decomposable into a collection of "pieces," each of which could be inspected separately in an attempt to locate the minimum of the objective function. For instance, if C were a (solid) cube in $\mathbb{R}^3$, one could look at what happens at the 8 corners, along the 12 edges, on the 6 faces, and in the cube's interior.

For most problems of interest in modern applications, however, there is little hope in such an approach. The number of "pieces" would be astronomical, or there would be no easy organization or listing of them. A further difficulty would lie in identifying which of the constraints might be redundant. Then too, there could be problems of degeneracy, where the constraints line up in odd ways and spoil the possibility of a good description of the "pieces" of C. As if this weren't enough trouble, there is the real prospect that C might be disconnected. These considerations force a different perspective on the analyst, who must look instead for a new kind of geometric framework on which to base computational schemes.

Geometry of linear constraints: Initial insight into a kind of geometry that *does* provide important support in optimization can be gained through the following ideas, which will later be subsumed under the broader heading of "convexity."

Half-spaces and hyperplanes: Subsets of $\mathbb{R}^n$ of the form $\{x \mid d \cdot x = c\}$ for a vector $d = (d_1, \ldots, d_n) \neq (0, \ldots, 0)$ and some constant $c \in \mathbb{R}$ are called *hyperplanes*, while those of the form $\{x \mid d \cdot x \leq c\}$ or $\{x \mid d \cdot x \geq c\}$ are called *closed half-spaces*. (With strict inequality, the latter would be *open half-spaces*.) A linear equality or inequality constraint on x thus corresponds to making x belong to a certain hyperplane or closed half-space (unless the linear function is $\equiv 0$, in which case the set isn't a hyperplane or half-space but just $\emptyset$ or $\mathbb{R}^n$, depending on c).

Polyhedral sets: A set $C \subset \mathbb{R}^n$ is called *polyhedral* if it can be represented as the intersection of a collection of finitely many hyperplanes or closed half-spaces, or in other words, specified by a finite system of linear constraints. (The whole space $\mathbb{R}^n$ is regarded as fitting this description by virtue of being the intersection of the "empty collection" of hyperplanes. The empty set fits because it can be viewed as the intersection of two parallel hyperplanes with no point in common.)

Argument: When a set is specified by a collection of constraints, it is the intersection of the sets specified by each of these constraints individually.

Inequalities alone: In the definition of "polyhedral" it would be enough to speak just of closed half-spaces, inasmuch as any hyperplane is itself the intersection of the two closed half-spaces associated with it.

Boxes as a special case: Any box is in particular a polyhedral set, since it's determined by upper or lower bounds on coordinates x_j of $x = (x_1, \ldots, x_n)$, each of which could be expressed in the form $d \cdot x \leq c$ or $d \cdot x \geq c$ for a vector d lining up with some coordinate axis. A box is thus an intersection of certain closed half-spaces.

Linear subspaces as a special case: Polyhedral sets don't have to have "corners" or "edges." For instance, any subspace of $\mathbb{R}^n$ is polyhedral, since by linear algebra it can be specified by finitely many homogeneous linear equations.

Geometric interpretation of linear programming: The feasible set C in any linear programming problem is a certain *polyhedral* set. The function f_0 being minimized over C is a *linear* function, so (unless $f_0 \equiv 0$) its "isosurfaces" $\{x \mid f_0(x) = \alpha\}$, as α ranges over $\mathbb{R}$, form a family of *parallel hyperplanes* H_α. The gradient of f_0, which is the same at all points x, is a certain vector that points in a direction perpendicular to all these hyperplanes, toward increases in α.

In this picture, f_0 takes on a value α somewhere in C if and only if the hyperplane H_α meets C, i.e., has $H_\alpha \cap C \neq \emptyset$. In minimizing f_0 over C, one is seeking the "lowest" of these hyperplanes that still meets C.

Penalties and the choice of objectives: The standard problem format suggests that a modeler should approach a situation looking for a family of functions f_i of certain decision variables x_j, one of these functions being the objective function, and the rest, constraint functions. But reality can be murkier. The distinction between what should be set up as a constraint and what should be incorporated into the expression to be minimized may be quite subtle and even in some cases just a matter of the notation being adopted.

Hard versus soft constraints: Some kinds of constraints are "hard" in the sense of representing intrinsic conditions that can't be violated. For instance, a vector $(x_1, \ldots, x_n)$ may give a system of probabilities or weights through the stipulation that $x_1 \geq 0, \cdots, x_n \geq 0$ and $x_1 + \ldots + x_n = 1$. It wouldn't make sense to consider the alternative where $x_1 \geq -.0001$ and $x_1 + \cdots + x_n = .9999$. Constraints of such type are often built into the specification of the set X in problems in the standard format.

Other constraints may have quite a different, "soft" character. For instance, in asking that a mechanism under design have a strength coefficient of at least .78, the modeler may be expressing a general desire that could be changed a bit once the costs and trade-offs are better known. A coefficient value of .76 may be quite acceptable, once it is realized that the difference could cost a fortune.

Penalty expressions: In dealing with soft constraints $f_i(x) \leq 0$ or $f_i(x) = 0$, it may be better in many cases to introduce a penalty expression instead. Thus, instead of enforcing an exact constraint, a term $\varphi_i \circ f_i$ could be added to the objective where (for the inequality constraint) $\varphi_i(t) = 0$ when $t \leq 0$ but $\varphi_i(t) > 0$ when $t > 0$. A

popular choice is a "linear" approach to penalties,

$$\varphi_i(f_i(x)) = \alpha_i \max\big\{0, f_i(x)\big\} \text{ with penalty parameter } \alpha_i > 0,$$

but many other choices are available, with penalties growing in other ways. It's worthwhile sometimes to relax the requirement of φ_i to just $\varphi_i(t) \le 0$ for $t \le 0$, with a negative penalty interpreted as a reward (for satisfying the inequality with room to spare).

Multiple objectives: Contrary to what we hear every day, it is impossible to design something to be the quickest, the cheapest and the most convenient all at the same time. While a number of variables may be of keen interest in a situation, the best that can be done is to optimize one of them while keeping the others within reasonable ranges.

As a compromise, one can look to minimizing an expression like a weighted combination of the variables or more generally, in the case of variables given by functions $f_1, \ldots, f_m$, an expression $\varphi\big(f_1(x), \ldots, f_m(x)\big)$.

Max functions and nonsmoothness: A "max function" is a function defined as the pointwise maximum of a collection of other functions, for instance

$$g(x) = \max\big\{g_1(x), \ldots, g_r(x)\big\}.$$

Here the "max" refers to the operation of taking, as the definition of the value $g(x)$, the highest of the r values $g_1(x), \ldots, g_r(x)$ (not necessarily all different). Such a function g is generally *nonsmooth*; a function g is called *smooth* if its first partial derivatives exist everywhere and behave continuously, but the max expression goes against this. Sometimes it's preferable to deal with a nonsmooth function directly, but other times smoothness can be achieved by a reformulation.

Scheme for minimizing a max function: Consider the problem of minimizing, over some $X \subset \mathbb{R}^n$, a function g of the form just given. Suppose $g_1, \ldots, g_r$ are themselves smooth on $\mathbb{R}^n$. By introducing an additional variable $u \in \mathbb{R}$, we can re-express the problem equivalently as

$$\begin{aligned} &\text{minimize } f_0(x,u) := u \text{ over all } (x,u) \in X' = X \times \mathbb{R} \subset \mathbb{R}^{n+1} \text{ satisfying} \\ &\qquad f_k(x,u) := g_k(x) - u \le 0 \text{ for } k = 1, \ldots, r, \end{aligned}$$

where functions $f_0, f_1, \ldots, f_r$ are smooth on $\mathbb{R}^n \times \mathbb{R}$.

Scheme for handling linear penalties: Suppose, for the sake of illustrating a technique, that every constraint in $(\mathcal{P})$ has been replaced by a "linear penalty" expression. The task then is to minimize

$$f_0(x) + \sum_{i=1}^{s} \alpha_i \max\{0, f_i(x)\} + \sum_{i=s+1}^{m} \alpha_i |f_i(x)|$$

over all $x \in X \subset \mathbb{R}^n$, where the coefficients α_i are positive. With additional variables $u_i \in \mathbb{R}$ and the vector $(u_1, \ldots, u_m)$ denoted by $u \in \mathbb{R}^m$, we can write this as the problem of minimizing

$$\bar{f}_0(x, u) := f_0(x) + \sum_{i=1}^{m} \alpha_i u_i$$

over all $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$ satisfying $x \in X$ and

$$f_i(x) - u_i \leq 0 \text{ and } u_i \geq 0 \text{ for } i = 1, \ldots, s,$$
$$f_i(x) - u_i \leq 0 \text{ and } -f_i(x) - u_i \leq 0 \text{ for } i = s+1, \ldots, m.$$

Aspects of good problem formulation: In many areas of mathematics, a problem targeted for the application of a numerical method is not considered to be well formulated unless the *existence* of a *unique* solution is assured, and the solution is *stable* in the sense of being affected only slightly when the data elements of the problem are shifted slightly. In optimization, however, the goals of uniqueness and stability are unrealistic in the way they are usually interpreted, and that of existence has to be adapted to the multiple notions of what may be meant by a solution.

Issues to consider in the general framework of optimization theory:

Does a feasible solution exist?

Does an optimal solution exist (global optimality)?

Can there be more than one optimal solution?

What happens to the set of feasible solutions under perturbations of problem data?

What happens to the set of optimal solutions under perturbations of problem data?

What happens to the optimal value under perturbations of problem data?

Note: The optimal value always exists and is unique.

The role of sequences: These issues are all the more important in view of the fact that most problems have to be solved by a *numerical method*. Such methods don't just produce an answer, but instead (however this might be masked by the software) generate a sequence of solution *candidates* which, it is hoped, get closer and closer to something *perhaps acceptable in lieu of* a true optimal solution.

Unless certain basic conditions are fulfilled, in particular ensuring the existence of an optimal solution, the candidates might not get progressively closer to anything meaningful at all. Anyway, they might not satisfy the problem's constraints exactly. Many questions then arise.

Example: potential trouble in one-dimensional minimization. Even in the case of minimizing a function over an interval in $\mathbb{R}$, pitfalls are apparent. An optimal solution can fail to exist because the function is unbounded below, or because the optimal value can be approached only in an asymptotic sense (getting arbitrarily close, but without attainment), or simply because the function lacks continuity properties. Gradual changes in the shape of the graph of the function, in the case of multiple dips and humps, can induce jumps and multiplicities in the behavior of the optimal solution set. All of these phenomena can make trouble for methods that are supposed to generate a sequence of points tending somehow toward a minimizing point.

Example: potential trouble in linear programming. As explained earlier, the feasible set C in a linear programming problem is a certain polyhedral set. It could be empty if the constraints are improperly chosen or even if they are perturbed only slightly from their proper values. Furthermore, in minimizing a linear function over such a set one can obtain as the set of optimal solutions a "corner," an entire "edge" or "face," or other such portion. Indeed a gradual, continuous change in the coefficients of the linear objective function can induce jumps in the answer.

Thus, even in the most elementary so-called linear cases of optimization, there can be difficulties under all three of the headings of existence, uniqueness and stability of solutions. In particular, two numerical formulations of a problem that differ only in roundoff in input data—the number of decimal points allocated to the representation of the various coefficients—could in principle have unique *optimal solutions* very different from each other. Because only *linear* programming is involved, the two *optimal values* would be close together, according to theoretical results we haven't discussed. But for more general classes of problems there can be discontinuities even in the behavior of optimal values unless special assumptions are invoked.

Existence of optimal solutions: The good news is that readily verifiable criteria are available to ensure that *at least one* optimal solution exists. The goal here will be to develop such a criterion, not just for the sake of a bare existence result, but in a form suited to the analysis of sequential approaches to finding solutions. This obliges us to work with the possibility that, in the course of calculations, constraints might only be satisfied approximately.

Approximate feasibility: Of special concern in connection with the stability of the feasible set C is what happens when constraints are only required to be satisfied to within a certain error bound. For any $\varepsilon > 0$, the set of *ε-feasible solutions* to a problem in standard format is

$$C_\varepsilon := \Big\{ x \in X \;\Big|\; f_i(x) \le \varepsilon \text{ for } i = 1, \ldots, s;\ \big|f_i(x)\big| \le \varepsilon \text{ for } i = s+1 \ldots, m \Big\}.$$

Clearly C_ε includes C, so the minimum value of f_0 over C_ε is less than or equal to the minimum over C, the optimal value in the given problem.

Well posed problems: The key concept that we'll work with in connection with the existence and approximation of solutions is the following. An optimization problem in the standard format will be deemed to be *well posed* when:

(a) the set $X \subset I\!R^n$ is closed as well as nonempty,

(b) the functions f_0, $f_1, \ldots, f_m$ on $I\!R^n$ are continuous,

(c) For some $\varepsilon > 0$, the set C_ε defined above has the property that, for every value $\alpha \in I\!R$, the set $C_\varepsilon \cap \big\{ x \;\big|\; f_0(x) \le \alpha \big\}$ is bounded.

Easy special cases: Condition (a) is fulfilled when X is a nonempty box, or indeed any nonempty polyhedral set. Condition (b) is fulfilled when the functions are linear or affine, or are given by polynomial expressions in the variables x_j.

Condition (c) is satisfied when X itself is bounded (since $C_\varepsilon \subset X$), or alternatively if for every $\alpha \in I\!R$ the set $X \cap \big\{ x \;\big|\; f_0(x) \le \alpha \big\}$ is bounded. Also, (c) is sure to be satisfied if for some $\varepsilon > 0$ any one of the functions f_i for $i = 1, \ldots, s$ has the property that the set $\big\{ x \in X \;\big|\; f_i(x) \le \varepsilon \big\}$ is bounded, or one of the functions f_i for $i = s+1, \ldots, m$ is such that the set $\big\{ x \in X \;\big|\; |f_i(x)| \le \varepsilon \big\}$ is bounded.

Caution: This concept refers to the manner in which an application of optimization has been set up as a problem $(\mathcal{P})$ in standard format: it's a property of the problem's *formulation* and depends on the specification of the functions f_i, the index s and the set X. A given application might be formulated in various ways as $(\mathcal{P})$, not

only in the choice of decision variables but according to whether its requirements are taken as inequality constraints, equality constraints, or lumped into the abstract constraint $x \in X$. In some modes it could turn out to be well posed, but in others perhaps not. This is true in particular because the "perturbations" introduced in terms of ε affect the inequality and equality constraints differently, and don't affect the abstract constraint at all.

Review of terminology and notation for dealing with sequences: For the benefit of students wanting a refresher, we briefly go over some of the facts and ideas of advanced calculus now coming into play here. Throughout these notes, we use superscript ν, the Greek letter "nu", as the running index for sequences so as to avoid conflicts with the indices that may appear in reference to coordinates, powers, etc. For instance, a sequence in $\mathbb{R}^n$ will be comprised of points x^ν for $\nu = 1, 2, \ldots$, where $x^\nu = (x_1^\nu, \ldots, x_n^\nu)$. A "sequence" means an "infinite sequence," but the points don't all have to be different. As a special case, every x^ν could be the same point c, giving a *constant* sequence. Implicitly always $\nu \to \infty$ when we write things like $\lim_\nu x^\nu$.

Convergence: A sequence of points $x^\nu = (x_1^\nu, \ldots, x_n^\nu)$ in $\mathbb{R}^n$ is said to *converge* to a point $x = (x_1, \ldots, x_n)$ (and therefore be a *convergent sequence*) if for each coordinate index j one has $x_j^\nu \to x_j$ as $\nu \to \infty$, or equivalently

$$|x^\nu - x| \to 0, \text{ where } |x| := (x_1^2 + \cdots + x_n^2)^{1/2} \text{ (Euclidean norm).}$$

Then x is the *limit* of the sequence; this is written as $x^\nu \to x$, or $x = \lim_{\nu\to\infty} x^\nu$.

Continuity: A function $f : \mathbb{R}^n \to \mathbb{R}$ is *continuous* if whenever $x^\nu \to x$ in $\mathbb{R}^n$ one has $f(x^\nu) \to f(x)$. The standard ways of verifying continuity involve such facts as the sum, product, max, or composition of continuous functions being continuous, along with the knowledge that certain elementary functions, for instance polynomial functions, exponential functions, sine and cosine functions, etc., are continuous.

For a mapping $F : \mathbb{R}^n \to \mathbb{R}^m$, continuity is defined similarly by the condition that $x^\nu \to x$ implies $F(x^\nu) \to F(x)$. In terms of a coordinate representation $F(x) = \big(f_1(x), \ldots, f_m(x)\big)$, this is equivalent to each of the component functions $f_i : \mathbb{R}^n \to \mathbb{R}$ being continuous.

Closedness: A set $S \subset \mathbb{R}^n$ is said to be *closed* if for every sequence of points $x^\nu \in S$ ($\nu = 1, 2, \ldots$) that converges to a point $x \in \mathbb{R}^n$, one has $x \in S$. A set is *open* if its complement in $\mathbb{R}^n$ is closed. In the extreme cases of $S = \mathbb{R}^n$ and $S = \emptyset$, S is both open and closed at the same time.

Intersections and unions: The intersection of any family of closed sets is closed. The union of any family of *finitely many* closed sets is closed.

Example: Level sets of continuous functions. If a function f is continuous, then for every choice of $c \in \mathbb{R}$ the set $\{x \mid f(x) \leq c\}$ and the set $\{x \mid f(x) \geq c\}$ are both closed. So too is the set $\{x \mid f(x) = c\}$, which is their intersection.

Example: Closedness of feasible sets. In an optimization problem in standard format, the feasible set C is closed if the set X is closed and the functions $f_1, \ldots, f_m$ are continuous. This is because C is the intersection of X with m other sets of the form $\{x \mid f_i(x) \leq 0\}$ or $\{x \mid f_i(x) = 0\}$, each of which is itself closed by the foregoing. As for X being closed in this context, that could come for instance from X belonging to the next category.

Example: Boxes and other polyhedral sets. These are the feasible sets for systems of linear constraints, so they are closed because affine functions are continuous. A common case is that of the nonnegative orthant $\mathbb{R}^n_+$.

Boundedness: A set $S \subset \mathbb{R}^n$ is called *bounded* if it lies within some (large enough) ball, or in other words, if there exists $\rho \in (0, \infty)$ such that $|x| \leq \rho$ for all $x \in S$. An equivalent characterization in terms of the coordinates x_j of x is that there exist (finite) bounds α_j and β_j such that for every $x \in S$ one has $\alpha_j \leq x_j \leq \beta_j$ for $j = 1, \ldots, n$. As a special case, the empty subset $\emptyset$ of $\mathbb{R}^n$ is bounded.

Compactness: Closely related to closedness and boundedness is another property, which ultimately is crucial in any discussion of existence of solutions. A set S is called *compact* if every sequence $\{x^\nu\}_{\nu=1}^\infty$ of points in S has at least one subsequence $\{x^{\nu_\kappa}\}_{\kappa=1}^\infty$ that converges to a limit. The Heine-Borel Theorem asserts that a set $S \subset \mathbb{R}^n$ is compact if and only if S is both closed and bounded.

Cluster points: A point that is the limit of some *subsequence* of a given sequence, although not necessarily the limit of the sequence as a whole, is called a *cluster point* of the sequence. It follows from the theorem just quoted that *every bounded sequence in $\mathbb{R}^n$ has at least one cluster point* (since a bounded sequence can in particular be viewed as being in some large, closed ball, which by the theorem will be a compact set).

This leads to the occasionally useful criterion that a sequence $\{x^\nu\}_{\nu=1}^\infty$ in $\mathbb{R}^n$ converges (in its entirety) if and only if it is bounded and (because of certain circumstances) can't possibly have two different cluster points.

Standard criterion for the attainment of a minimum or maximum: A basic fact of calculus related to optimization is the following. If a *continuous* function is minimized over a *nonempty*, *compact* set in $\mathbb{R}^n$, the minimum value is attained at some point (not necessarily unique) in the set. Likewise, the maximum value is attained somewhere in the set.

Shortcomings for present purposes: This criterion could immediately be applied to optimization problems in standard format by making assumptions that guarantee not only the closedness of the feasible set C (as already discussed) but also its boundedness. Then, as long as the objective function f_0 being minimized over C is continuous, an optimal solution will exist. But in many problems the feasible set *isn't* bounded. For instance, in "unconstrained" optimization we have $C = \mathbb{R}^n$. Therefore, we need a result that's more general.

THEOREM 1 (existence of optimal solutions)**.** *Consider an optimization problem* $(\mathcal{P})$ *in standard format, and assume it is well posed. Then the feasible set is closed. If the feasible set is also nonempty, then the optimal set is nonempty and the optimal value is finite. Furthermore, the optimal set is compact.*

Proof. The assumption that the problem is well posed entails (in conditions (a) and (b) of the definition of that property) the closedness of X and continuity of $f_1, \ldots, f_m$. These properties have already been seen above to imply that the feasible set C is closed.

Under the assumption now that C is also nonempty, let $\tilde{x}$ denote any point of C and let $\tilde{\alpha} = f_0(\tilde{x})$. The problem of minimizing f_0 over C has the same optimal solutions, if any, as the problem of minimizing f_0 over $\tilde{C} = C \cap \{x \,|\, f_0(x) \leq \tilde{\alpha}\}$. The set $\tilde{C}$ is nonempty, because it contains $\tilde{x}$. It is closed by virtue of being the intersection of the closed set C and the set $\{x \,|\, f_0(x) \leq \tilde{\alpha}\}$, which is closed because f_0 is continuous by condition (b). Furthermore, it is bounded because of condition (c) in the definition of well posedness. Therefore, $\tilde{C}$ is compact. It follows from the standard criterion for the attainment of a minimum that the problem of minimizing f_0 over $\tilde{C}$ has an optimal solution. Hence the given problem, of minimizing f_0 over C, has an optimal solution as well.

Let $\bar{x}$ denote an optimal solution, not necessarily the only one, and let $\bar{\alpha} = f_0(\bar{x})$. Then $\bar{\alpha}$ is the optimal value in the problem, and because f_0 is a real-valued function this optimal value is finite. The optimal set is $\{x \in C \,|\, f_0(x) = \bar{\alpha}\}$, and this is the same as $\{x \in C \,|\, f_0(x) \leq \bar{\alpha}\}$ because strict inequality is impossible. The same argument applied to the set $\tilde{C}$ tells us that this set, like $\tilde{C}$, is compact. □

Example: problems with only an abstract constraint. As a simple case to which the existence criterion in Theorem 1 can be applied, consider the problem

$$\text{minimize } f_0(x) \text{ over all } x \in X \subset \mathbb{R}^n,$$

where there are no side conditions of the form $f_i(x) \leq 0$ or $f_i(x) = 0$. The basic criterion for good formulation comes down in this case to

(a) the set $X \subset \mathbb{R}^n$ is closed as well as nonempty,

(b) the function f_0 is continuous,

(c) the set $\{x \in X \mid f_0(x) \leq \alpha\}$ is bounded in $\mathbb{R}^n$ for all values of $\alpha \in \mathbb{R}$.

Under these conditions, therefore, f_0 attains its minimum over X, and the set of minimizing points is compact. In particular, (a) and (c) hold when X is compact, but the case of *unconstrained* optimization, where $X = \mathbb{R}^n$ is also covered, namely by having f_0 be continuous with the sets $\{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}$ all bounded.

Example: the gap distance between two sets. For two nonempty, closed sets C_1 and C_2 in $\mathbb{R}^n$, the *gap distance* between C_1 and C_2 is the optimal value in the problem of minimizing $|x_1 - x_2|$ (Euclidean distance) over all $x_1 \in C_1$ and $x_2 \in C_2$, or in other words, over all pairs (x_1, x_2) in the set $X = C_1 \times C_2$. *While this optimal value is well defined, as always, are there optimal solutions to this problem?* That is, do there exist pairs $(\bar{x}_1, \bar{x}_2) \in C_1 \times C_2$ for which $|\bar{x}_1 - \bar{x}_2|$ is the minimum value? A sufficient condition for the existence of such a pair, as provided by Theorem 1 through the preceding example, is the boundedness of the sets $\{(x_1, x_2) \in C_1 \times C_2 \mid |x_1 - x_2| \leq \rho\}$. This obviously holds if both C_1 and C_2 are bounded, but one can show actually that the boundedness of just one of the sets is enough.

If the gap distance between C_1 and C_2 is 0, does that mean these sets have at least one point in common? Not necessarily; this hinges on the existence of an optimal solution to the problem described. If $(\bar{x}_1, \bar{x}_2)$ is an optimal solution, and the optimal value is 0, then obviously $\bar{x}_1 = \bar{x}_2$, and this is a point in $C_1 \cap C_2$. An example where the gap distance is 0 but $C_1 \cap C_2 = \emptyset$ is furnished in $\mathbb{R}^2$ by taking C_1 to be a hyperbola having C_2 as one of its asymptotes.

Special existence criterion in linear and quadratic programming: When $(\mathcal{P})$ is a linear or quadratic programming problem, the existence of optimal solutions can be concluded more easily, without the burden of establishing well-posedness. It's enough to verify that the feasible set is nonempty and the optimal value is not $-\infty$. This criterion suffices because of the simple structure of such problems, which precludes "nasty asymptotic behavior," but a proof won't be given here.

Uniqueness of optimal solutions: The bad news is that there is no criterion, verifiable directly in terms of a problem's structure and data without going into computation, that can be imposed on a general *nonconvex* problem to ensure the existence of *at most one* optimal solution. This topic will be taken up later, after some theory of convexity has been built up.

Existence of feasible solutions: Generally speaking, there is no good criterion to apply to a system of constraints in order to ascertain on a theoretical level that there is at least one point satisfying the system. However, a numerical approach is possible. For a constraint system in the standard problem format, a numerical method of optimization could be applied to the auxiliary problem of minimizing the function

$$g(x) := \sum_{i=1}^{s} \alpha_i \max\{0, f_i(x)\} + \sum_{i=s+1}^{m} \alpha_i |f_i(x)|$$

over all $x \in X$, where the introduced penalty coefficients α_i are all positive (e.g. $\alpha_i = 1$). Obviously, $g(x) = 0$ for each $x \in X$ satisfying the desired constraints, whereas $g(x) > 0$ for all other choices of $x \in X$. Thus if the optimal value in the auxiliary problem is 0 the optimal solutions to the auxiliary problem are precisely the feasible solutions to the original problem, but if the optimal value in the auxiliary problem is positive, there are no feasible solutions to the original problem.

Note that the minimization of $g(x)$, as proposed, could be handled by the reformulation trick developed earlier for "linear" penalties in general.

Convergence to a solution: As groundwork for the consideration of numerical methods, it's important to broaden Theorem 1 to cover sequences such as could be generated by such methods. The ε provision in "well-posedness" will be utilized in this. We keep the discussion focused on a problem in standard format and notation.

Feasible sequence: A sequence of points x^ν all belonging to the feasible set, or in other words, satisfying $x^\nu \in X$, $f_i(x^\nu) \leq 0$ for $i \in [1, s]$, and $f_i(x^\nu) = 0$ for $i \in [s+1, m]$.

Optimizing sequence: A feasible sequence of points x^ν such that, for the optimal value $\bar{\alpha}$ in the problem, $f_0(x^\nu) \to \bar{\alpha}$. (Note that this property says nothing about the points x^ν themselves converging or even remaining bounded as $\nu \to \infty$!)

Asymptotically feasible sequence: A sequence of points $x^\nu \in X$ with the property that $\max\{0, f_i(x^\nu)\} \to 0$ for $i = 1, \ldots, s$, and $f_i(x^\nu) \to 0$ for $i = s+1, \ldots, m$. An equivalent description of this property is that $x^\nu \in C_{\varepsilon^\nu}$ for some choice of shrinking "error bounds" $\varepsilon^\nu \to 0$.

Asymptotically optimizing sequence: An asymptotically feasible sequence of points x^ν with the additional property that, in terms of the optimal value $\bar{\alpha}$ in the problem, one has $\max\{\bar{\alpha}, f_0(x^\nu)\} \to \bar{\alpha}$.

Comments: For some kinds of unconstrained problems or problems with linear constraints only, methods can be devised that generate an optimal sequence. Usually, however, the most one can hope for is an asymptotically optimizing sequence. The properties of such sequences are therefore of fundamental interest. Note that every optimizing sequence is in particular an asymptotically optimizing sequence.

THEOREM 2 (optimality from sequences)**.** *Consider an optimization problem* $(\mathcal{P})$ *in standard format, and suppose that it is well posed, and that its feasible set is nonempty. Then any (optimizing or) asymptotically optimizing sequence* $\{x^\nu\}_{\nu=1}^\infty$ *is bounded, and all of its cluster points (at least one exists) are optimal solutions to* $(\mathcal{P})$*. Furthermore, the sequence* $\{f_0(x^\nu)\}_{\nu=1}^\infty$ *of function values converges to the optimal value in* $(\mathcal{P})$*.*

If in fact $(\mathcal{P})$ *has a unique optimal solution* $\bar{x}$*, any (optimizing or) asymptotically optimizing sequence must converge (as a whole) to* $\bar{x}$*.*

Proof. Let ε be an error bound value for which condition (c) in the definition of well posedness is satisfied. Denote the optimal value by $\bar{\alpha}$, and consider any number $\alpha \in (\bar{\alpha}, \infty)$. For any asymptotically optimizing sequence $\{x^\nu\}_{\nu=1}^\infty$, we have $x^\nu \in X$ and there is an index $\bar{\nu}$ such that, for all $\nu \geq \bar{\nu}$, we have

$$f_i(x^\nu) \leq \varepsilon \text{ for } i \in [1, s], \quad |f_i(x^\nu)| \leq \varepsilon \text{ for } i \in [s+1, m], \quad f_0(x^\nu) \leq \alpha.$$

In other words, all the points x^ν with $\nu \geq \bar{\nu}$ lie in the set $\{x \in C_\varepsilon \,|\, f_0(x) \leq \alpha\}$. This set is bounded because of condition (c). Therefore, the sequence $\{x^\nu\}_{\nu=1}^\infty$ is bounded.

Let $\bar{x}$ denote any cluster point of the sequence; $\bar{x} = \lim_{\kappa\to\infty} x^{\nu_\kappa}$ for some subsequence $\{x^{\nu_\kappa}\}_{\kappa=1}^\infty$. Because X is closed by condition (a), and $x^{\nu_\kappa} \in X$, we have $\bar{x} \in X$. From the continuity of the functions f_i in condition (b), we have $f_i(x^{\nu_\kappa}) \to f_i(\bar{x})$, and through the asymptotic optimality of the sequence $\{x^\nu\}_{\nu=1}^\infty$ this implies that

$$f_i(\bar{x}) \leq 0 \text{ for } i \in [1, s], \quad f_i(\bar{x}) = 0 \text{ for } i \in [s+1, m], \quad f_0(\bar{x}) \leq \bar{\alpha}.$$

Since no feasible solution can give the objective function a value lower than the optimal value $\bar{\alpha}$, we conclude that $\bar{x}$ is an optimal solution.

To justify the final assertion of the theorem, we note that in the case described the asymptotically optimizing sequence, which is bounded, can't have two different cluster points, so it must converge. The only limit candidate is the unique optimal solution. □

Sad truth: It merits emphasis that the case at the end of Theorem 2 is generally the *only* one in which a proposed numerical method that is truly able to generate an asymptotically optimizing sequence (some don't even claim to do that) is *guaranteed* to produce points converging to a particular optimal solution.

Puncturing an illusion: Isn't the last assertion of Theorem 2 *always* true, at least for *optimizing* sequences? No, it relies on $(\mathcal{P})$ being well posed. Otherwise, there are counterexamples even for unconstrained minimization. For instance, consider minimizing $f_0(x) = x^2/(1+x^4)$ over $x \in \mathbb{R}$. Obviously $f_0(0) = 0$ but $f_0(x) > 0$ when $x \neq 0$, so $\bar{x} = 0$ is the unique optimal solution. But the graph of f_0 is asymptotic to the x-axis in both directions, so any sequence of points that tends to $+\infty$, or to $-\infty$, is an optimizing sequence which fails to converge to $\bar{x}$.

Boundedness avoids the trouble: In the proof of Theorem 2, the role of condition (c) in the definition of $(\mathcal{P})$ being well posed is to guarantee that $\{x^\nu\}_{\nu=1}^\infty$ is bounded. The same conclusions would hold if the boundedness of this sequence were *assumed outright*, even if (c) might not be satisfied (but (a) and (b) in the definition are satisfied). To put this another way, asymptotically optimizing sequences can't really get into trouble unless they are unbounded.

Comparison of possible solution methods: Different numerical techniques for trying to solve an optimization problem can be quite different in their behavior.

Is the method valid for the application at hand? Has it rigorously been justified, and are the assumptions on which its justification depends satisfied for the problem being solved? If not, the output may be worthless. In this respect a frequent abuse is the neglect of convexity or differentiability assumptions.

What does the method actually claim to find, and how reliable is the claim? Methods that are said to "solve" a problem often just try to approximate a point for which various conditions associated mathematically with optimality, but not necessarily guaranteeing optimality, are fulfilled. Some methods generate more information about a problem and its potential solutions than do others, and this auxiliary information could be useful on the side.

How robust is the method? Here the issue is whether the technique works dependably or is liable to get stuck on problems for which the "flavor" isn't quite to its liking.

Rates of convergence: An important consideration in comparing methods is their speed in producing an "answer." Real conclusions can only come from experiments with well selected test problems, but there's a theoretical discipline which makes comparisons on the basis of the rate at which the distance of $f_0(x^\nu)$ to the optimal value $\bar{\alpha}$ decreases, or the distance of x^ν from an optimal solution $\bar{x}$ decreases.

Linear convergence: For instance, one can try to ascertain the existence of a constant c such that $|x^{\nu+1} - \bar{x}| \leq c|x^\nu - \bar{x}|$ for all indices ν beyond some $\bar{\nu}$. Then the method is said to *converge linearly at the rate c*. If two methods both guarantee convergence if this type, but one method typically yields a lower value of c than the other, then that method can be expected to find solutions faster—at least if computations are continued until the convergence behavior takes over.

Note: Rates of convergence usually apply only to some undetermined "tail portion" of the sequence that's generated. They say nothing about behavior early on.

Quadratic convergence: Modes of convergence can also differ more profoundly. A method with $|x^{\nu+1} - \bar{x}| \leq c|x^\nu - \bar{x}|^2$ for all indices ν beyond some $\bar{\nu}$ is said to *converge quadratically at the rate c*. Such a method is likely to be much quicker than one that converges only linearly, but it may carry additional overhead or be applicable only in special situations.

Finite convergence: A method may *converge finitely* in the sense of being sure to terminate after only finitely many iterations (no infinite sequence then being involved at all). That's true for example in some methods of solving linear programming problems or optimization problems involving flows in networks. For such numerical procedures still other forms of theoretical comparison have been developed in terms of computational "complexity."

Nonstandard formats for optimization: Although the standard format dominates the way most people think about optimization, other approaches can also be useful, or from a theoretical angle, sometimes better.

Abstract format for optimization: Sometimes it's good to think of an optimization problem with respect to $x \in \mathbb{R}^n$ as simply a problem of minimizing some function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ over all of $\mathbb{R}^n$, where $\overline{\mathbb{R}} = [-\infty, \infty]$. Ordinarily $-\infty$ doesn't come in as a value that f takes on, but ∞ can serve as a an infinite penalty. For instance, in the case of a problem already formulated in the standard way as $(\mathcal{P})$, we can define

$$f(x) = \begin{cases} f_0(x) & \text{if } x \text{ is feasible,} \\ \infty & \text{if } x \text{ is not feasible.} \end{cases}$$

Minimizing f over all $x \in \mathbb{R}^n$ can be viewed then as equivalent to $(\mathcal{P})$. This is primarily useful in contexts of theory, of course; when it comes to computation, the structure of f must anyway be explicit. The function f just defined is known as the *essential* objective function in $(\mathcal{P})$.

Epigraphical geometry: The geometry in this case centers not on the graph of f but on its *epigraph*, which is the set $\big\{(x,\alpha) \in \mathbb{R}^n \times \mathbb{R} \,\big|\, f(x) \le \alpha\big\}$. The importance of considering epigraphs, instead of just the traditional graphs of functions, has already been seen in the study of "max functions"

Composite format in optimization: An approach to problem formulation that's richer in possibilities than the standard format draws on the ideas of the essential objective function by looking to

$$\text{minimize } f(x) = \varphi\big(f_0(x), f_1(x), \ldots, f_m(x)\big) \text{ over all } x \in X \subset \mathbb{R}^n$$

for a choice of functions $f_0, f_1, \ldots, f_m$ and a "modeling" function $\varphi : \mathbb{R}^{m+1} \to \overline{\mathbb{R}}$. This includes the standard format as the special case where

$$\varphi(u_0, u_1, \ldots, u_m) = \begin{cases} u_0 & \text{if } u_i \le 0 \text{ for } i \in [1,s] \text{ and } u_i \le 0 \text{ for } i \in [s+1,m], \\ \infty & \text{otherwise.} \end{cases}$$

It also, however, very easily covers penalty representations of constraints in a direct manner.

Overview of the different formats: While the standard format has a long tradition behind it and has become almost synonymous with "optimization," the theoretical power and modeling advantages of the composite format are now encouraging a trend in that direction, at least among specialists. The abstract format is primarily a tool for thinking about problems in various ways for theoretical purposes, rather than an approach to modeling, but for that it is often very helpful. In this introductory course the standard format will receive the main attention, but the impression should be resisted that all optimization models should be forced into that channel. In the long run it's better to have a broader perspective.

3. UNCONSTRAINED MINIMIZATION

Many of the complications encountered in problems of optimization are due to the presence of constraints, but even when there are no constraints a number of important issues arise as to the nature of optimal solutions and the possible ways they might be determined. In treating these issues in the case of a smooth objective function, we will want to take full advantage of the properties incorporated into the standard definition of differentiability for a function of n-variables.

Vector notation: The inner product (or dot product) of two vectors is the value

$$z \cdot w = z_1 w_1 + \cdots + z_n w_n \text{ for } z = (z_1, \ldots, z_n) \text{ and } w = (w_1, \ldots, w_n),$$

as we've already been using. In books with extensive linear algebra this is often expressed instead by $z^T w$ under the convention that vectors are interpreted special matrices—"column vectors"—unless the transpose operation (indicated by a superscript T) turns them into "row vectors." Here we follow the typographically preferable pattern of always writing vectors horizontally but viewing them as "column vectors" in formulas where they get multiplied by matrices.

Angles between vectors: When $z \neq 0$ and $w \neq 0$, one has $z \cdot w = |z||w| \cos\theta$, where θ is the angle between z and w (and $|z|$ and $|w|$ their lengths). Thus, $z \cdot w$ is positive, zero, or negative according to whether the angle is acute, right, or obtuse.

Review of differentiability: The differentiability of a function f on $\mathbb{R}^n$ means more than the existence of its first partial derivatives, which after all would just refer to behavior along various lines parallel to the n coordinate axes. Rather, it's a property expressing the possibility of a kind of approximation of f (namely, "linearization") that is present regardless of any change of coordinates that might be introduced. For our purposes here, we'll avoid subtle distinctions by keeping to the mainstream case where differentiability can be combined with continuity.

Standard differentiability classes: A function f on $\mathbb{R}^n$ is *continuously differentiable*, or of class $\mathcal{C}^1$, if its first partial derivatives exist everywhere and are continuous everywhere. It's *twice continuously differentiable*, or of class $\mathcal{C}^2$, if this holds for second partial derivatives, and in general of class $\mathcal{C}^k$ its kth partial derivatives exist and are continuous everywhere. Then actually f and all its partial derivatives of orders less than k must be continuous as well.

Localization: Similarly one can speak of f as being a $\mathcal{C}^k$ function relative to some open set, for instance an open neighborhood of some point $\bar{x}$.

Gradient vectors and Hessian matrices: If f is of class $\mathcal{C}^1$ on a neighborhood of a point $\bar{x}$ it has the *first-order expansion*

$$f(x) = f(\bar{x}) + \nabla f(\bar{x}) \cdot [x - \bar{x}] + o\big(|x - \bar{x}|\big),$$

where the vector $\nabla f(\bar{x})$ has the partial derivatives $(\partial f/\partial x_j)(\bar{x})$ as its components and is called the *gradient* of f at $\bar{x}$. The classical "o(t)" notation refers to an error term with the property that $o(t)/t \to 0$ as $t \to 0$. This notation is often confusing to students, but really it is just a sort of code for writing down, in a manner deemed more convenient, the assertion that

$$\lim_{\substack{x \to \bar{x} \\ x \neq \bar{x}}} \frac{f(x) - f(\bar{x}) - \nabla f(\bar{x}) \cdot [x - \bar{x}]}{|x - \bar{x}|} = 0,$$

which says that the affine function $l(x) = f(\bar{x}) + \nabla f(\bar{x}) \cdot [x - \bar{x}]$ furnishes a *first-order approximation* $f \approx l$ at $\bar{x}$. Likewise, if f is of class $\mathcal{C}^2$ on a neighborhood of $\bar{x}$ it further has the *second-order expansion*

$$f(x) = f(\bar{x}) + \nabla f(\bar{x}) \cdot [x - \bar{x}] + \tfrac{1}{2}[x - \bar{x}] \cdot \nabla^2 f(\bar{x})[x - \bar{x}] + o(|x - \bar{x}|^2),$$

where $\nabla^2 f(\bar{x})$ is the $n \times n$ matrix with the partial derivatives $(\partial^2 f/\partial x_i \partial x_j)(\bar{x})$ as its components and is called the *Hessian* of f at $\bar{x}$. This time the "o" notation is code for the assertion that

$$\lim_{\substack{x \to \bar{x} \\ x \neq \bar{x}}} \frac{f(x) - f(\bar{x}) - \nabla f(\bar{x}) \cdot [x - \bar{x}] - \frac{1}{2}[x - \bar{x}] \cdot \nabla^2 f(\bar{x})[x - \bar{x}]}{|x - \bar{x}|^2} = 0.$$

The quadratic function $q(x) = f(\bar{x}) + \nabla f(\bar{x}) \cdot [x - \bar{x}] + \frac{1}{2}[x - \bar{x}] \cdot \nabla^2 f(\bar{x})[x - \bar{x}]$ then furnishes a *second-order approximation* $f \approx q$ at $\bar{x}$.

Hessian symmetry: When f_0 is of class $\mathcal{C}^2$, the matrix $\nabla f_0(\bar{x})$ is sure to be symmetric, but otherwise this might fail.

Vector-valued functions and Jacobian matrices: A mapping, or vector-valued function, $F : \mathbb{R}^n \to \mathbb{R}^m$ with $F(x) = \big(f_1(x), \ldots, f_m(x)\big)$ is of class $\mathcal{C}^k$ when its component functions f_i are. As long as F is of class $\mathcal{C}^1$ around $\bar{x}$ it has the first-order expansion

$$F(x) = F(\bar{x}) + \nabla F(\bar{x})[x - \bar{x}] + o\big(|x - \bar{x}|\big),$$

where $\nabla F(\bar{x})$ is the $m \times n$ matrix with the partial derivatives $(\partial f_i/\partial x_j)(\bar{x})$ as its components and is called the *Jacobian* of F at $\bar{x}$.

A connection: In the case of a function f of class $\mathcal{C}^2$ on a neighborhood of $\bar{x}$, the gradient mapping $\nabla f : x \to \nabla f(x)$ has Jacobian $\nabla(\nabla f)(\bar{x}) = \nabla^2 f(\bar{x})$ at $\bar{x}$.

Local information: For a $\mathcal{C}^2$ function f_0 on $\mathbb{R}^n$, the gradient $\nabla f_0(\bar{x})$ and Hessian $\nabla^2 f_0(\bar{x})$ provide local information about f_0 at $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)$ which can well be put to use in numerical methods for minimizing f_0. The main idea is to consider what happens to f_0 along various lines through $\bar{x}$. Any such line can represented parametrically as the set of points of the form $\bar{x} + \tau w$ for $-\infty < \tau < \infty$ for some vector $w \neq 0$, The direction of $w = (w_1, \ldots, w_n)$ gives the direction of the line.

The values of f_0 along such a line can be investigated parametrically through the expression $\varphi(\tau) := f_0(\bar{x} + \tau w)$, where $\varphi(0) = f_0(\bar{x})$. In particular, one can try to minimize $\varphi(\tau)$ in τ, at least part way, in order to come up with a point $\bar{x} + \tau w$ yielding a lower value of f_0 than does $\bar{x}$. A crucial fact in this respect is that

$$\varphi'(0) = \frac{d}{d\tau} f_0(\bar{x} + \tau w)\Big|_{\tau=0} = \nabla f_0(\bar{x}) \cdot w,$$

this being known as the *directional derivative* at f_0 relative to w. By looking at the sign of this quantity we can tell whether the values of f_0 will go up or down as we start to move away from $\bar{x}$ in the direction of w. On the other hand, one has

$$\varphi''(0) = \frac{d^2}{d\tau^2} f_0(\bar{x} + \tau w)\Big|_{\tau=0} = w \cdot \nabla^2 f_0(\bar{x}) w,$$

and this quantity can be crucial in determining second-order effects.

Geometric interpretation of gradients: A well known fact coming out of this is that $\nabla f(\bar{x})$, when it isn't the zero vector, points in the direction in which f_0 increases at the fastest rate, and that its length gives that rate. Likewise, $-\nabla f(\bar{x})$ points in the direction in which f_0 decreases at the fastest rate.

Why? Divide $\nabla f_0(\bar{x})$ by its length $|\nabla f_0(\bar{x})|$ to get a vector $\bar{w}$ that points in the same direction but has length 1. For any vector w of length 1, the rate of change of f_0 at $\bar{x}$ in the direction of w is $\nabla f_0(\bar{x}) \cdot w$ according to the analysis above, and in terms of $\bar{w}$ this comes out as $|\nabla f_0(\bar{x})| \bar{w} \cdot w = |\nabla f_0(\bar{x})| \cos\theta$, where θ is the angle between $\bar{w}$ and w. This rate equals $|\nabla f_0(\bar{x})|$ when w lines up with $\bar{w}$, but otherwise falls short.

Optimality considerations, first-order: Such parametric analysis along lines leads to concepts that are fundamental to understanding whether f_0 has a local minimum at $\bar{x}$, or if not, how to proceed to a point where the value of f_0 is lower than at $\bar{x}$.

Descent vectors: A vector w is called a *descent vector* for f_0 at $\bar{x}$ if $\nabla f_0(\bar{x}) \cdot w < 0$, so that the function $\varphi(\tau) = f_0(\bar{x} + \tau w)$ has $\varphi'(0) < 0$. This ensures that $\varphi(\tau) < \varphi(0)$ for all τ in some interval $(0, \varepsilon)$, and therefore $f_0(\bar{x} + \tau w) < f_0(\bar{x})$ for $0 < \tau < \varepsilon$. When the object is to minimize f_0, we therefore get an improvement in replacing $\bar{x}$ by $\bar{x} + \tau w$ for any value $\tau > 0$ that isn't too large.

Stationary points: A point $\bar{x}$ is *stationary* for f_0 if $\nabla f_0(\bar{x}) = 0$. This is equivalent to the condition that $\nabla f_0(\bar{x}) \cdot w = 0$ for *every* vector w, and it thus means that no descent vector exists at $\bar{x}$.

Basic first-order condition: If f_0 has a local minimum at $\bar{x}$, then $\nabla f_0(\bar{x}) = 0$. Indeed, otherwise there would exist a descent vector w at $\bar{x}$ (for instance $w = -\nabla f_0(\bar{x})$). Thus, every point giving a local minimum must be a stationary point for f_0.

A trap not to fall into: The converse is false: a stationary point doesn't always provide a local minimum. This is obvious as soon as attention is focused on it, since the one-dimensional case already provides numerous examples, but many people slip up on it nonetheless.

Optimality considerations, second-order: The stationary-point condition just discussed only involves first partial derivatives, as embodied in the gradient $\nabla f_0(\bar{x})$. In order to obtain conditions that are sufficient for local optimality, rather than just necessary, it's essential to work with second partial derivatives, unless a property called "convexity" (soon to be treated) comes to the rescue. Because the second partial derivatives of f_0 at $\bar{x}$ are embodied in the Hessian $\nabla^2 f_0(\bar{x})$, the following notions of linear algebra come into play. Recall that a matrix $A \in \mathbb{R}^{n \times n}$ is

$$\text{*positive definite* if } w \cdot Aw > 0 \text{ for all } w \neq 0,$$
$$\text{*positive semidefinite* if } w \cdot Aw \geq 0 \text{ for all } w.$$

THEOREM 3 (local optimality conditions without constraints). *For a function $f_0 : \mathbb{R}^n \to \mathbb{R}$ of class $\mathcal{C}^2$, consider the problem of minimizing f_0 over all $x \in \mathbb{R}^n$.*

(a) (necessary condition). *If $\bar{x}$ is a locally optimal solution, then $\nabla f_0(\bar{x}) = 0$ and $\nabla^2 f_0(\bar{x})$ is positive semidefinite.*

(b) (sufficient condition). *If $\bar{x}$ is such that $\nabla f_0(\bar{x}) = 0$ and $\nabla^2 f_0(\bar{x})$ is positive definite, then $\bar{x}$ is a locally optimal solution. Moreover, in these circumstances the local optimality of $\bar{x}$ is strict, in the sense that there exists a $\delta > 0$ such that*

$$f_0(x) > f_0(\bar{x}) \text{ for all points } x \text{ with } 0 < |x - x_0| < \delta.$$

Proof. (a) If $\bar{x}$ is locally optimal for f_0 on $\mathbb{R}^n$, then in particular it will be true for each choice of $w \in \mathbb{R}^n$ that the function $\varphi(\tau) := f_0(\bar{x} + \tau w)$ has a local minimum at $\tau = 0$. Since $\varphi'(0) = \nabla f_0(\bar{x}) \cdot w$ and $\varphi''(0) = w \cdot \nabla^2 f_0(\bar{x}) w$, as noted earlier, we conclude that $\nabla f_0(\bar{x}) \cdot w = 0$ for every $w \in \mathbb{R}^n$ and $w \cdot \nabla^2 f_0(\bar{x}) w \geq 0$ for every $w \in \mathbb{R}^n$. This means that $\nabla f_0(\bar{x}) = 0$ and $\nabla^2 f_0(\bar{x})$ is positive semidefinite.

(b) The reverse argument is more subtle and can't just be reduced to one dimension, but requires utilizing fully the second-order expansion of f_0 at $\bar{x}$. Our assumptions give for $A := \nabla^2 f_0(\bar{x})$ that $f_0(x) = f_0(\bar{x}) + \frac{1}{2}[x - \bar{x}] \cdot A[x - \bar{x}] + o\big(|x - \bar{x}|^2\big)$. According to the meaning of this, we can find for any $\varepsilon > 0$ a $\delta > 0$ such that

$$\frac{\big|f_0(x) - f_0(\bar{x}) - \frac{1}{2}[x - \bar{x}] \cdot A[x - \bar{x}]\big|}{|x - \bar{x}|^2} < \varepsilon \text{ when } 0 < |x - \bar{x}| < \delta,$$

and in particular,

$$f_0(x) - f_0(\bar{x}) \ > \ \tfrac{1}{2}[x - \bar{x}] \cdot A[x - \bar{x}] - \varepsilon |x - \bar{x}|^2 \text{ when } 0 < |x - \bar{x}| < \delta.$$

Because A is positive definite, $\frac{1}{2} w \cdot Aw$ is positive when $w \neq 0$; it depends continuously on w and therefore achieves its minimum over the closed, bounded set consisting of the vectors w with $|w| = 1$ (the unit sphere in $\mathbb{R}^n$). Denoting this minimum by λ, we have $\lambda > 0$ and $\frac{1}{2}[\tau w] \cdot A[\tau w] \geq \lambda \tau^2$ for all $\tau \in \mathbb{R}$ when $|w| = 1$. Since any difference vector $x - \bar{x} \neq 0$ can be written as τw for $\tau = |x - \bar{x}|$ and $w = [x - \bar{x}]/|x - \bar{x}|$, we have $\frac{1}{2}[x - \bar{x}] \cdot A[x - \bar{x}] \geq \lambda |x - \bar{x}|^2$ for all x. The estimate from twice differentiability then yields

$$f_0(x) - f_0(\bar{x}) \ > \ (\lambda - \varepsilon)|x - \bar{x}|^2 \text{ when } 0 < |x - \bar{x}| < \delta.$$

Recalling that ε could have been chosen arbitrarily small, in particular in the interval $(0, \lambda)$, we conclude that there's a $\delta > 0$ such that $f_0(x) > f_0(\bar{x})$ when $0 < |x - \bar{x}| < \delta$. Thus, f has a local minimum at $\bar{x}$. □

Local versus global optimality: Because the results in Theorem 3 relate only to the properties of f_0 in some neighborhood of $\bar{x}$, and give no estimate for the size of that neighborhood (it might be tiny, for all we know), an important question is left unanswered. How can we tell whether a given point $\bar{x}$ furnishes a global minimum to f_0? The best approach to answering this question, and often in practice the only one, is through the concept of convexity.

Convex functions: A function f on $\mathbb{R}^n$ is *convex* if for every choice of points x_0 and x_1 with $x_0 \neq x_1$, and every choice of $\tau \in (0, 1)$, one has

$$f\big((1 - \tau)x_0 + \tau x_1\big) \leq (1 - \tau) f\big(x_0\big) + \tau f\big(x_1\big) \text{ for all } \tau \in (0, 1).$$

Interpretation: The expression $x(\tau) := (1-\tau)x_0 + \tau x_1 = x_0 + \tau(x_1 - x_0)$ parameterizes the line through x_0 in the direction of $w = x_1 - x_0$, with $x(0) = x_0$ and $x(1) = x_1$. When $0 < \tau < 1$, $x(\tau)$ is an intermediate point on the line segment joining x_0 with x_1, specifically the point reached in moving the fraction τ of the distance from x_0 to x_1 along this segment. The inequality says that the value of f at this intermediate point doesn't exceed the interpolated value obtained by going the fraction τ of the way from the value $f(x_0)$ to the value $f(x_1)$ (whichever direction that might involve, depending on which of the two values might be larger).

Relativization to lines: Since the condition involves only three collinear points at a time, we have the principle that *f is convex on $\mathbb{R}^n$ if and only if for every line L in $\mathbb{R}^n$, f is convex relative to L*; in fact, instead of lines it would be enough to speak of line segments. Here a *line* is a set of points in $\mathbb{R}^n$ that can be expressed as $\{x + \tau w \mid -\infty < \tau < \infty\}$ for some x and w with $w \neq 0$, whereas a *line segment* is the same thing but with $0 \leq \tau \leq 1$.

Related properties: A function f is called

strictly convex: if the inequality always holds with $<$,

concave: if the inequality always holds with $\geq$,

strictly concave: if the inequality always holds with $>$.

Affine functions as an example: It can be shown that f is affine on $\mathbb{R}^n$, as already defined, if and only if f is simultaneously convex and concave.

Jensen's inequality: The definition of convexity implies more generally that for any points $x_k \in \mathbb{R}^n$ and weights $\lambda_k \geq 0$ for $k = 0, 1, \ldots, p$ with $\sum_{k=0}^{p} \lambda_k = 1$, one has

$$f\big(\lambda_0 x_0 + \lambda_1 x_1 + \cdots + \lambda_p x_p\big) \;\leq\; \lambda_0 f\big(x_0\big) + \lambda_1 f\big(x_1\big) + \cdots + \lambda_p f\big(x_p\big).$$

Tests for convexity using derivatives: The following facts help in identifying examples of convex functions that are differentiable. Later there will be other tactics available, which can be used to ascertain the convexity of functions that have been put together in certain ways from basic functions whose convexity is already known.

Monotonicity of first derivatives in one dimension: For f differentiable on $\mathbb{R}$,

$$\begin{aligned}
f \text{ is convex} &\iff f' \text{ is nondecreasing,}\\
f \text{ is strictly convex} &\iff f' \text{ is increasing,}\\
f \text{ is concave} &\iff f' \text{ is nonincreasing,}\\
f \text{ is strictly concave} &\iff f' \text{ is decreasing,}\\
f \text{ is affine} &\iff f' \text{ is constant.}
\end{aligned}$$

Incidentally, these generalize also to functions of a single variable that merely have a right derivative and a left derivative at every point. For instance, a piecewise linear cost function is convex if and only if the slope values for consecutive pieces form an increasing sequence. Also as first derivative conditions,

$$\begin{aligned} f \text{ is convex} &\iff f(y) \geq f(x) + f'(x)(y-x) \text{ for all } x \text{ and } y, \\ f \text{ is strictly convex} &\iff f(y) > f(x) + f'(x)(y-x) \text{ for all } x \text{ and } y,\ x \neq y. \end{aligned}$$

Signs of second derivatives in one dimension: For f twice differentiable on $\mathbb{R}$,

$$\begin{aligned} f \text{ is convex} &\iff f''(x) \geq 0 \text{ for all } x, \\ f \text{ is strictly convex} &\impliedby f''(x) > 0 \text{ for all } x. \end{aligned}$$

Notice that the final condition is *not* an equivalence but only an implication in one direction! An example is $f(x) = x^4$, with $f''(x) = 12x^2$. This function is strictly convex on $\mathbb{R}$ because $f'(x) = 4x^3$ is an increasing function. But $f''(x)$ fails to be positive everywhere: $f''(0) = 0$.

THEOREM 4 (derivative tests for convexity in higher dimensions). *For a once differentiable function f on $\mathbb{R}^n$,*

$$\begin{aligned} f \textit{ is convex} &\iff f(y) \geq f(x) + \nabla f(x) \cdot [y-x] \textit{ for all } x \textit{ and } y, \\ f \textit{ is strictly convex} &\iff f(y) > f(x) + \nabla f(x) \cdot [y-x] \textit{ for all } x \textit{ and } y,\ x \neq y, \end{aligned}$$

and in the case of f being twice differentiable function, also

$$\begin{aligned} f \textit{ is convex} &\iff \nabla^2 f(x) \textit{ is positive semidefinite for all } x, \\ f \textit{ is strictly convex} &\impliedby \nabla^2 f(x) \textit{ is positive definite for all } x. \end{aligned}$$

Proof. The trick in every case is to reduce to the corresponding one-dimensional criterion through the principle that f has the property in question if and only if it has it relative to every line segment. The first of the conditions will suffice in illustrating this technique. To say that f is convex is to say that for every choice of points x_0 and x_1 with $x_0 \neq x_1$ the function $\varphi(\tau) := f\big((1-\tau)x_0 + \tau x_1\big) = f\big(x_0 + \tau(x_1 - x_0)\big)$ is convex on the interval $(0,1)$. From the chain rule one calculates that

$$\varphi''(\tau) = w \cdot \nabla^2 f(x) w \quad \text{for } x = (1-\tau)x_0 + \tau x_1, \quad w = x_1 - x_0.$$

The convexity of f is thus equivalent to having $w \cdot \nabla^2 f(x) w \geq 0$ for every possible choice of x and $w \neq 0$ such that x is an intermediate point of some line segment in the direction of w. This holds if and only if $\nabla^2 f(x)$ is positive semidefinite for every x. The arguments for the other three conditions are very similar in character. □

Local strict convexity: A function f is strictly convex *locally* around $\bar{x}$ if the strict convexity inequality holds over the line segment joining x_0 and x_1 whenever these points lie within a certain neighborhood of $\bar{x}$.

The proof of Theorem 3 show that for this to be true in the case of a function f of class $\mathcal{C}^2$ it suffices to have the Hessian $\nabla^2 f(x)$ be positive definite at all points x in some neighborhood of $\bar{x}$. In fact it suffices to have $\nabla^2 f(\bar{x})$ itself be positive definite, because any matrix having entries close enough to those of a positive definite matrix must likewise be positive definite, and here the entries of $\nabla^2 f(x)$ depend continuously on x. (The stability of positive definiteness under perturbations follows from identifying the positive definiteness of a matrix A with the positivity of the function $q(w) = w \cdot Aw$ on the compact set consisting of the vectors w with $|w| = 1$.)

Tests of positive definiteness: For a symmetric matrix $A \in \mathbb{R}^{n \times n}$ there are many tests of whether A is positive definite or positive semidefinite, as may be found in texts on linear algebra, but they aren't always easy to apply. Computer tests are available as well. Perhaps the main thing to remember is that any symmetric matrix A is similar to a diagonal matrix having as its diagonal entries the n eigenvalues of A (with multiplicities). *Positive definiteness holds if and only if all the eigenvalues are positive, whereas positive semidefiniteness holds if and only of all the eigenvalues are nonnegative.*

Two-dimensional criterion: A matrix $A \in \mathbb{R}^{2 \times 2}$ is positive definite if and only if its determinant and its trace (the sum of its diagonal entries) are both positive; it is positive semidefinite if and only if both are nonnegative. (For $n \times n$ matrices with $n > 2$, these conditions are still necessary, but no longer sufficient.)

The important consequences of convexity in unconstrained minimization:

Global optimality of stationary points: For a $\mathcal{C}^1$ function f_0 on $\mathbb{R}^n$ that's convex, the condition $\nabla f_0(\bar{x}) = 0$ implies that $\bar{x}$ gives the *global minimum* of f_0 on $\mathbb{R}^n$.

Argument: Convexity implies by Theorem 4 that $f_0(x) \geq f_0(\bar{x}) + \nabla f_0(\bar{x}) \cdot [x - \bar{x}]$ for all x. When $\nabla f_0(\bar{x}) = 0$, this reduces to having $f_0(x) \geq f_0(\bar{x})$ for all x.

Uniqueness from strict convexity: If a strictly convex function f_0 has its minimum at $\bar{x}$, then $\bar{x}$ is the *only* point where f_0 has its minimum. In fact, for this conclusion it's enough that f_0 be a convex function that's strictly convex locally around $\bar{x}$.

Argument: If there were another point $\hat{x}$ where f_0 had its minimum value, say α, the intermediate points $x_\tau = (1 - \tau)\bar{x} + \tau\hat{x}$ for $\tau \in (0, 1)$ would have

$f_0(x_\tau) \le (1-\tau)f_0(\bar{x}) + \tau f_0(\hat{x}) = (1-\tau)\alpha + \tau\alpha = \alpha$, hence $f_0(x_\tau) = \alpha$, since nothing lower than α is possible. Then f_0 would be constant on the line segment joining $\bar{x}$ and $\hat{x}$, so it couldn't be strictly convex any portion of it.

Convexity of quadratic functions: If $f(x) = \frac{1}{2}x\cdot Ax + b\cdot x + c$ for a symmetric matrix $A \in \mathbb{R}^{n\times n}$, a vector $b \in \mathbb{R}^n$, and a constant $c \in \mathbb{R}$, we have $\nabla f(x) = Ax + b$ and $\nabla^2 f(x) = A$ for all x. Therefore, such a function is convex if and only if A is positive semidefinite. It is strictly convex if and only if A is positive definite. This second assertion doesn't fully follow from the second-order condition in Theorem 4, which only gives the implication in one direction, but it can be deduced from the *first-order* condition for strict convexity.

Minimizing a quadratic function: A quadratic function can't attain its minimum anywhere if it isn't a convex function. It attains its minimum at a unique point if and only if it's strictly convex—with positive definite Hessian.

Argument: If a quadratic function q attains its minimum at a point $\bar{x}$, its Hessian at $\bar{x}$ must be positive semidefinite by Theorem 3. But, because it's quadratic, q has this same Hessian at every point. Then by Theorem 4, q is convex. If the Hessian matrix is A, the fact that the gradient of q at $\bar{x}$ is 0 means we have the expansion $q(x) = q(\bar{x}) + \frac{1}{2}[x-\bar{x}]\cdot A[x-\bar{x}]$. Under the assumption that the minimum is attained uniquely at $\bar{x}$ there can't be a vector $x - \bar{x} \neq 0$ such that $A[x-\bar{x}] = 0$. Then A is nonsingular. But from linear algebra, a positive semidefinite matrix is nonsingular if and only if it's positive definite. Then q is strictly convex by Theorem 4.

Conversely, if q has Hessian A, it has the expression $q(x) = \frac{1}{2}x\cdot Ax + b\cdot x + c$ for $b = \nabla q(0)$ and $c = q(0)$. If A is positive definite there is a $\lambda > 0$ such that $\frac{1}{2}x\cdot Ax \ge \lambda|x|^2$ for all x, by reasoning given in the proof of part of Theorem 3(b). Then $|q(x)| \ge \lambda|x|^2 - |b||x| - |c|$, so that for any $\rho > 0$ the norms $|x|$ of the vectors in the level set $\{x \mid q(x) \le \rho\}$ all lie in the interval $\{t \mid \lambda t^2 - |b|t - [c+\rho] \le 0\}$, which is bounded because of λ being positive. These level sets are therefore all bounded, so the problem of minimizing q is well posed and by Theorem 1 has a solution.

Applications to numerical optimization: Most numerical methods for the unconstrained minimization of a twice continuously differentiable function rely on the facts we've been developing. Here, for purposes of illustration, we'll look at some of the most popular approaches based on utilization of local information.

Descent methods: A large class of methods for minimizing a smooth function f_0 on $\mathbb{R}^n$ fits the following description. A sequence of points x^ν such that

$$f_0(x^0) > f_0(x^1) > \cdots > f_0(x^\nu) > f_0(x^{\nu+1}) > \cdots$$

is generated from a chosen starting point x^0 by selecting, through some special scheme in each iteration, a descent vector w^ν and a corresponding value $\tau^\nu > 0$, called a *step size*, such that $f_0(x^\nu + \tau^\nu w^\nu) < f_0(x^\nu)$. The improved point $x^\nu + \tau^\nu w^\nu$ is taken to be the successor point $x^{\nu+1}$.

Of course, if in a given iteration no descent vector exists at all, this means that $\nabla f_0(x^\nu) = 0$. In that case the method terminates with $\bar{x} = x^\nu$ as a stationary point. Usually, however, an infinite sequence $\{x^\nu\}_{\nu=1}^\infty$ is generated, and the question for analysis is whether this is an optimizing sequence, or more soberly in the nonconvex case, at least a sequence for which every cluster point $\bar{x}$ is a stationary point.

Line search: One way to choose a step size τ^ν yielding $f_0(x^\nu + \tau^\nu w^\nu) < f(x^\nu)$ is to execute some kind of *line search* in the direction of w^ν, which refers to an exploration of the values of f_0 along the half-line emanating from x^ν in the direction of w^ν. In the notation $\varphi^\nu(\tau) := f(x^\nu + \tau w^\nu)$ we have $(\varphi^\nu)'(0) < 0$, and the task is to select a value $\tau^\nu > 0$ such that $\varphi^\nu(\tau^\nu) < \varphi^\nu(0)$, yet not one so small that progress might stagnate.

Exact line search: An approach with natural appeal is to choose τ^ν to be a value of τ that minimizes φ^ν on the interval $[0, \infty)$. Techniques are available for carrying out the one-dimensional minimization of φ^ν to whatever accuracy is desired, at least when f_0 is convex. Of course, in numerical work hardly anything is really "exact."

Backtracking line search: Professional opinion now favors a different approach, which depends on a choice of parameters β and γ with $0 < \beta < \gamma < 1$. Take τ^ν to be the first γ^k in the sequence $\gamma, \gamma^2, \gamma^3, \ldots \to 0$, such that $\big[\varphi^\nu(\gamma^k) - \varphi^\nu(0)\big]/\gamma^k < \beta(\varphi^\nu)'(0)$. (Such a γ^k exists, because $\big[\varphi^\nu(\tau) - \varphi^\nu(0)\big]/\tau$ tends to $(\varphi^\nu)'(0)$ as τ decreases to 0, and $(\varphi^\nu)'(0) < \beta(\varphi^\nu)'(0) < 0$.)

Cauchy's method, or steepest descent: A descent method can be obtained with $w^\nu = -\nabla f_0(x^\nu)$ in every iteration (as long as that vector is nonzero), since this choice makes $\nabla f_0(x^\nu) \cdot w^\nu = -\big|\nabla f_0(x^\nu)|^2 < 0$ when $\nabla f_0(x^\nu) \neq 0$. As observed earlier, this vector points in the direction in which f_0 has the biggest rate of decrease at $\bar{x}$. (One could equally well take $w^\nu = -\nabla f_0(x^\nu)/|\nabla f_0(x^\nu)|$, and for use in line search this can be helpful numerically with scaling.)

Newton's method in optimization: From the definition of twice differentiability we know that the quadratic function

$$q^\nu(x) := f_0(x^\nu) + \nabla f_0(x^\nu)\cdot[x - x^\nu] + \tfrac{1}{2}[x - x^\nu]\cdot\nabla^2 f_0(x^\nu)[x - x^\nu]$$

(whose Hessian everywhere is $A = \nabla^2 f_0(x^\nu)$) furnishes a second-order local approximation of f_0 around x^ν. This suggests that by investigating the minimum of $q^\nu(x)$ we can learn something about where to look in trying to minimize f_0. Specifically, *assume that $q^\nu(x)$ attains its minimum at a unique point different from x^ν*, this point being denoted by $\hat{x}^\nu$; then $\hat{x}^\nu \neq x^\nu$, and $q^\nu(\hat{x}^\nu) < q^\nu(x^\nu) = f(x^\nu)$. (From the above, this is equivalent to assuming that the matrix $\nabla^2 f(x^\nu)$ is positive definite, hence in particular nonsingular, while the vector $\nabla f(x^\nu)$ isn't 0.) The vector $w^\nu = \hat{x}^\nu - x^\nu \neq 0$ is then called the *Newton vector* for f_0 at $\bar{x}$. It satisfies

$$w^\nu = -\nabla^2 f_0(x^\nu)^{-1}\nabla f_0(x^\nu).$$

It is a descent vector, and descent methods based on using it are called versions of *Newton's method* in optimization.

Argument: Because $\hat{x}^\nu$ minimizes q^ν, it must be a stationary point of q^ν:

$$0 = \nabla q^\nu(\hat{x}^\nu) = \nabla f_0(x^\nu) + \nabla^2 f_0(x^\nu)[\hat{x}^\nu - x^\nu] = \nabla f_0(x^\nu) + \nabla^2 f_0(x^\nu)w^\nu.$$

In solving this equation for w^ν, utilizing our assumption, which implies that the inverse matrix $\nabla^2 f_0(x^\nu)^{-1}$ exists, we get the formula claimed. To verify that w^ν is then a descent vector, observe that because $q^\nu(\hat{x}^\nu) < q^\nu(x^\nu)$ we have $\nabla f_0(x^\nu)\cdot w^\nu + \frac{1}{2}w^\nu\cdot\nabla^2 f_0(x^\nu)w^\nu < 0$. We wish to conclude that $\nabla f_0(x^\nu)\cdot w^\nu < 0$. If this weren't true, we'd have to have from the preceding inequality that $w^\nu\cdot\nabla^2 f_0(x^\nu)w^\nu < 0$. But this would contradict the positive definiteness of $\nabla^2 f_0(x^\nu)$, which was observed to follow from our assumption about q^ν attaining its minimum at a unique point.

Relation to Newton's method for equation solving: Newton's method in classical form refers not to minimizing a function but solving an equation $F(x) = 0$ for a smooth mapping $F : \mathbb{R}^n \to \mathbb{R}^n$. In principle, a sequence $\{x^\nu\}_{\nu\in\mathbb{N}}$ is generated from an initial point x^0 as follows. In iteration ν, the given equation is replaced by its first-order approximation $F(x^\nu) + \nabla F(x^\nu)(x - x^\nu) = 0$. The unique solution to this approximate equation is $\hat{x}^\nu = -\nabla F(x^\nu)^{-1}F(x^\nu)$, as long as the inverse matrix $\nabla F(x^\nu)^{-1}$ exists, and one takes $x^{\nu+1} = \hat{x}^\nu$.

Newton's method in optimization corresponds closely to the case of this where $F(x) = \nabla f_0(x)$. It resembles applying the classical form of Newton's method to solving the equation $\nabla f_0(x) = 0$. But it differs in not just automatically taking $x^{\nu+1} = \hat{x}^\nu = x^\nu + w^\nu$ but $x^{\nu+1} = x^\nu + \tau^\nu w^\nu$ for some step size τ^ν determined through a form of line search.

Effectiveness and validation of descent methods: Describing a numerical approach to minimizing a function, or for that matter solving a vector equation, is a far cry from establishing the circumstances in which it can be counted upon to work effectively, or providing an analysis that helps in comparing it with other approaches. For such purposes it is essential at the very least to make use of the conditions characterizing a point at which a minimum occurs as well as, in some situations, aspects of convexity.

Convergence questions: The theory of numerical methods of optimization and why (or whether) they work is full of ingenious ideas and pleasing geometry, as well as rigorous, technical developments. For a small taste of what it involves, let's consider more closely the question of whether a descent method (with a particular scheme for choosing decent vectors and executing line searches) for the unconstrained minimization of a function f_0 generates a sequence of points x^ν that in some way "solves" the problem.

Any such method does generate a decreasing sequence of function values $f_0(x^0) > f_0(x^1) > f_0(x^2) \ldots$, and any decreasing sequence of real numbers does have a limit $\alpha \in \overline{\mathbb{R}}$, but unfortunately α could fall short of furnishing the optimal value in the problem unless f_0 has certain rather special properties. Nonetheless we can search for guidance on when a method can sensibly be implemented and what it might accomplish even if it doesn't determine an optimal or locally optimal solution.

Applications to well posed problems: As noted after Theorem 1, an unconstrained problem of minimizing f_0 over $\mathbb{R}^n$ is well posed as long as f_0 is continuous and all its level sets $\left\{x \,\middle|\, f_0(x) \leq \alpha\right\}$ are bounded. Certainly f_0 is continuous when it's differentiable, as in the descent methods we've been investigating.

In unconstrained minimization there's no distinction between feasible and asymptotically feasible sequences (every sequence is such), nor any between optimal and asymptotically optimizing sequences. In these circumstances we know from Theorem 2 that every optimizing sequence is bounded, with all its cluster points being optimal solutions. However, this doesn't necessarily make it easier to *generate* an optimizing sequence.

THEOREM 5 (convergence of descent methods; exact line search). *Consider a well posed problem of minimizing a function f_0 over $\mathbb{R}^n$, with f_0 not just continuous but differentiable, and let S be the set of all stationary points of f_0 (the points $\bar{x}$ where $\nabla f_0(\bar{x}) = 0$). Consider a descent method that starts from a point x^0 and generates subsequent points by exact line search relative to vectors w^ν determined by a formula $w^\nu = D(x^\nu)$ having the property that, for each $x \notin S$ with $f_0(x) \leq f_0(x^0)$, $D(x)$ is a uniquely determined descent vector for f_0 at x, and $D(x)$ depends continuously on x. (The method terminates if $x^\nu \in S$.)*

(a) *If the method generates an infinite sequence $\{x^\nu\}_{\nu=0}^\infty$ (by not attaining a point of S in finitely many iterations), this sequence must be bounded, and all of its cluster points must belong to S.*

(b) *If actually there is only one point $\bar{x} \in S$ with $f_0(\bar{x}) \leq f_0(x^0)$, the sequence is indeed optimizing and converges to $\bar{x}$, this being the unique optimal solution.*

Proof. In each iteration with $x^\nu \notin S$, the vector w^ν is well defined according to our hypothesis, and it is not the zero vector (because it is a descent vector). We minimize $\varphi^\nu(\tau) := f_0(x^\nu + \tau w^\nu)$ over $\tau \in [0, \infty)$ to get τ^ν and then set $x^{\nu+1} = x^\nu + \tau^\nu w^\nu$. This line search subproblem is itself a well posed problem of optimization because the sets $\big\{\tau \geq 0 \,\big|\, \varphi^\nu(\tau) \leq \alpha\big\}$ are all bounded by virtue of the level sets of f_0 all being bounded. Thus it does have an optimal solution τ^ν (perhaps not unique) by Theorem 1.

From the definition of w^ν being a descent vector, we know moreover that $f_0(x^{\nu+1}) < f_0(x^\nu)$ always. Thus the sequence $\{f_0(x^\nu)\}_{\nu=1}^\infty$ is decreasing and therefore converges to some value α. Also, the sequence $\{x^\nu\}_{\nu=1}^\infty$ is contained in the set $\big\{x \,\big|\, f_0(x) \leq f_0(x^0)\big\}$, which by hypothesis is bounded. Consider any cluster point $\bar{x}$ of this sequence; there is a subsequence $\{x^{\nu_\kappa}\}_{\kappa=1}^\infty$ such that $x^{\nu_\kappa} \to \bar{x}$ as $\kappa \to \infty$. In particular we have $f_0(\bar{x}) = \alpha$, because f_0 is continuous. We wish to show that $\bar{x} \in S$ in order to establish (a).

Suppose $\bar{x} \notin S$. Then the vector $\bar{w} := D(\bar{x})$ is a descent vector for f_0 at $\bar{x}$, and the vectors $w^{\nu_\kappa} := D(x^{\nu_\kappa})$ are such that $w^{\nu_\kappa} \to \bar{w}$ (by our assumption in (a) that the mapping D specifying the method is well defined everywhere outside of S and continuous there). Because $\bar{w}$ is a descent vector, we know there is a value $\bar{\tau} > 0$ such that

$$f_0(\bar{x} + \bar{\tau}\bar{w}) < f_0(\bar{x}).$$

On the other hand, for each κ we know that $f_0(x^{\nu_\kappa+1}) = f_0(x^{\nu_\kappa} + \tau^{\nu_\kappa} w^{\nu_\kappa}) \leq f_0(x^{\nu_\kappa} + \bar{\tau} w^{\nu_\kappa})$ because $\bar{\tau}$ is one of the candidates considered in the minimization subproblem solved by τ^{ν_κ}. Taking the limit in the outer expressions in this inequality, we get

$$\alpha \leq f_0(\bar{x} + \bar{\tau}\bar{w})$$

because $f_0(x^\nu) \to \alpha$, $x^{\nu_\kappa} \to \bar{x}$ and $w^{\nu_\kappa} \to \bar{w}$ (again by the continuity of f_0). This is incompatible with the fact that $f_0(\bar{x} + \bar{\tau}\bar{w}) < f_0(\bar{x}) = \alpha$. The contradiction yields (a).

The extra assumption in (b) gives the existence of a unique optimal solution to the unconstrained minimization problem, because (1) an optimal solution exists by Theorem 1, (2) any optimal solution must in particular belong to S by Theorem 3, and of course any optimal solution must belong to the set $\{x \,|\, f_0(x) \leq f_0(x^0)\}$. From (a), this optimal solution $\bar{x}$ is the only candidate for a cluster point of $\{x^\nu\}_{\nu=1}^\infty$. As noted earlier, a bounded sequence with no more than one cluster point must be convergent. Thus, $x^\nu \to \bar{x}$ and, by the continuity of f_0, also $f_0(x^\nu) \to f_0(\bar{x})$. Since $f_0(\bar{x})$ is the optimal value in the problem, we conclude in this case that the sequence is optimal. □

Other convergence theorems: Theorem 5 offers just one example, drawn from wealth of results about descent methods and their variants. In particular, convergence theorems are also available for backtracking line search, which, as already mentioned, is regarded as superior for real calculations.

Specializations: Particular applications of the convergence result in Theorem 5 are obtained by considering various choices of the mapping D.

Cauchy's method with exact line search: Under the assumption that f_0 is a $\mathcal{C}^1$ function, so that $f_0(x)$ and $\nabla f_0(x)$ depend continuously on x, let $D(x) = -\nabla f_0(x)$. This is a descent vector as long as x is not a stationary point (cf. Example 1). The assumptions of Theorem 5 are satisfied, and we can conclude that if all the level sets $\{x \,|\, f_0(x) \leq \alpha\}$ are bounded the method will generate a bounded sequence $\{x^\nu\}_{\nu=1}^\infty$, all of whose cluster points are stationary points of f_0. If in addition f_0 is convex, these stationary points give the global minimum of f_0. In that case the method has generated an optimizing sequence $\{x^\nu\}_{\nu=1}^\infty$.

Newton's method with exact line search: Under the assumption that f_0 is a $\mathcal{C}^2$ function, so that $\nabla^2 f_0(x)$ too depends continuously on x, let $D(x)$ denote the Newton vector—*under the supposition that it's well defined* for every $x \notin S$ having $f_0(x) \leq f_0(x^0)$; we've seen this is tantamount to $\nabla^2 f_0(x)$ being positive definite for all such x. Then $D(x) = -\nabla^2 f_0(x)^{-1} \nabla f_0(x)$, so $D(x)$ depends continuously on x. (If a nonsingular matrix varies continuously, its inverse varies continuously, a fact derivable from determinant formulas for the inverse). As long as the level sets $\{x \,|\, f_0(x) \leq \alpha\}$ of f_0 are bounded, so that the problem is well posed, Theorem 5 is applicable and tells us that the method will generate a bounded sequence $\{x^\nu\}_{\nu=1}^\infty$, all of whose cluster points are stationary points of f_0. Because of the positive definiteness of the Hessians, any cluster point must be a *locally optimal*

solution to the problem of minimizing f_0, due to Theorem 3(b). Around any such cluster point, f_0 is strictly convex, so if f_0 is convex as a whole there can only be one cluster point, $\bar{x}$, this being the only point where f_0 attains its minimum. Then the sequence $\{x^\nu\}_{\nu=1}^\infty$ is optimal and converges to $\bar{x}$.

Comparison: Cauchy's method works quite generally, whereas Newton's method needs positive definiteness of the Hessians and therefore local strict convexity. But Newton's method has a much better *rate* of convergence than Cauchy's method: typically Newton's method converges *quadratically*, whereas Cauchy's method only converges *linearly*. We won't go into the theory of that here, however.

Compromise: Because Cauchy's method and Newton's method have complementary strengths and weaknesses, they are often combined in a single descent method in which, roughly speaking, the Cauchy descent vector is used early on, but eventually a switch is made to the Newton descent vector. In some versions, this approach would likewise fit into Theorem 5 for a certain formula for $D(x)$.

Further ideas in numerical approaches: Beyond the "pure" versions of Cauchy's method and Newton's method in optimization that we've been discussing, there are variants that have been developed not only in terms of backtracking line search but with other attractive features as well. Two of these will be described briefly.

Newton-like methods: These popular procedures, also called *matrix-secant* methods try to span between the properties of Cauchy's method and Newton's method of optimization in an especially interesting way. They select the direction vector by $w^\nu = -A^\nu \nabla f_0(x^\nu)$, where the matrix A^ν, generated in each iteration by some further rule, is symmetric and positive semidefinite. The case of $A^\nu = I$ gives Cauchy's method, while the case of $A^\nu = \nabla^2 f_0(x^\nu)^{-1}$ (when this matrix is positive definite) gives Newton's method.

For reasons already suggested, a simple choice like the one for Cauchy's method is favored as long as the current point is likely to be far from the solution, in order to take advantage of the global convergence properties of that method without making too many demands on f_0. But a choice close to the one for Newton's method is favored near a locally optimal solution $\bar{x}$ at which $\nabla^2 f_0(\bar{x})$ is positive definite (cf. the sufficient second-order optimality condition in Theorem 3(b)). A central question is how to select and update A^ν by gleaning information about second-derivative properties of f_0 that may be present in the computations carried out up to a certain stage. This is a big topic, with many clever schemes having been developed through years of research.

Trust region methods: Newton's method gets the descent vector w^ν from the fact that $x^\nu + w^\nu$ is the point that minimizes the quadratic function q^ν giving the local second-order approximation to f_0 at x^ν, as explained earlier. Instead of minimizing q^ν over all of $\mathbb{R}^n$, however, one can minimize it over some bounded neighborhood X^ν of x^ν, which is called a *trust region*. In denoting a minimizing point by $\hat{x}^\nu$ and defining $w^\nu = \hat{x}^\nu - x^\nu$, one obtains a descent vector w^ν. The trust region can in particular be specified by linear constraints, like upper and lower bounds on the variables to keep their values near the component values in the current vector x^ν, and the subproblem for producing w^ν is then one of *quadratic programming*. This scheme can be hybridized with Newton-like methods.

Optimization versus equation solving: The equation $\nabla f_0(x) = 0$ is a first step toward identifying points that minimize a smooth function f_0. This leads to the notion that solving an unconstrained optimization problem might be reduced to solving a system of equations, which has some degree of merit, but tends to misguide beginners, to whom equation-solving is a more familiar idea. The best approaches to equation-solving are through optimization, rather than the other way around.

Linear equations: In numerical analysis, the solving of $Ax = b$ when A is *symmetric* and *positive semidefinite* (and possibly quite large) plays a big role. Such an equation gives the condition that is both necessary and sufficient for the minimization of the quadratic convex function $f_0(x) = \frac{1}{2}x \cdot Ax - b \cdot x$ over $\mathbb{R}^n$. Thus, this branch of numerical analysis is in truth a branch of numerical optimization.

Linear least squares: The solving of a $Ax = b$ when A isn't necessarily symmetric or positive semidefinite can be approached as that of minimizing the function $f(x) = \frac{1}{2}|Ax - b|^2$. This has the advantage of making sense even when there are more equations than unknowns. Here f is a convex function having $\nabla f_0(x) = A^*[Ax - b]$ and $\nabla^2 f_0(x) = A^*A$ (with A^* the transpose of A). Thus, solving $\nabla f_0(x) = 0$ means solving $A^*Ax = A^*b$, where the matrix A^*A is symmetric and positive semidefinite.

Nonlinear least squares: An approach often taken to solving $F(x) = 0$ in the case of a general smooth mapping $F : \mathbb{R}^n \to \mathbb{R}^n$, with $F(x) = \big(f_1(x), \ldots, f_n(x)\big)$, is to translate it to minimizing $f(x) := \frac{1}{2}|F(x)|^2 = \frac{1}{2}f_1(x)^2 + \cdots + \frac{1}{2}f_n(x)^2$.

Remark: While this device is useful, it would be off-track if the equation $F(x) = 0$ *already* corresponds to an optimization problem because $F = \nabla f_0$ for some function f_0, which could be minimized instead.

4. CONSTRAINED MINIMIZATION

The minimization of a function f_0 over a set $C \subset \mathbb{R}^n$ can be much harder than the minimization of f_0 over all of $\mathbb{R}^n$, and it raises a host of issues. Some of these concern the ways a numerical method might be able to maneuver around in C or near to it, while others come up in characterizing optimality itself.

Feasible directions: A vector $w \neq 0$ is said to give a *feasible direction* into C at a point $x \in C$ if there's an $\varepsilon > 0$ such that the line segment $\{x + \tau w \,|\, 0 \leq \tau \leq \varepsilon\}$ lies in C.

Descent methods with constraints: The general class of descent methods described for unconstrained optimization could be adapted to the minimization of f_0 over C if the nature of C is such that feasible directions can readily be found. The rough idea is this. Starting from a point $x^0 \in C$, a sequence of points is generated by the scheme that, when at $x^\nu \in C$, a descent vector w^ν is chosen for f_0 which at the same time gives a feasible direction into C at x^ν. (If no such w^ν exists, the method terminates.) Next, some sort of line search is executed to produce a value $\tau^\nu > 0$ such that both $x^\nu + \tau^\nu w^\nu \in C$ and $f_0(x^\nu + \tau^\nu w^\nu) < f_0(x^\nu)$. The next point is taken then to be $x^{\nu+1} := x^\nu + \tau^\nu w^\nu$. In particular, one can imagine choosing τ^ν to be a value that minimizes $\varphi(\tau) = f_0(x^\nu + \tau w^\nu)$ over the set $\{\tau \geq 0 \,|\, x^\nu + \tau w^\nu \in C\}$; this would be the analogue of exact line search.

Pitfalls: Many troubles can plague this scheme, unless the situation is safeguarded by rather special features. Finding a descent vector w^ν that gives a feasible direction may be no easy matter, and even if one is found, there may be difficulties in using it effectively because of the need to keep within the confines of C. A phenomenon called *jamming* is possible, where progress is stymied by frequent collisions with the boundary of C and the method "gets stuck in a corner" of C or makes too many little zigzagging steps.

Lack of feasible directions at all: Of course, this kind of approach doesn't make much sense unless one is content to regard, as a "quasi-solution" to the problem, any point $x \in C$ at which there is no descent vector for f_0 giving a feasible direction into C. That may be acceptable for some sets C such as boxes, but not for sets C in which curvature dominates. For example, if C is a curved surface there may be *no* point $x \in C$ at which there's a feasible direction into C, because feasible directions refer only to movement along straight line segments within in C. Then there would be no point of C from which progress could be made by a modified descent method in minimizing f_0.

Variational geometry: Whether or not some form of descent method might be made to work, it's essential to have a solid grip on the geometry of the feasible set C. Classical geometry doesn't meet the needs, so new concepts have to be brought in. To get rolling, we need a notion of tangency which includes the tangent spaces long associated with "linearization" of curved structures, but at the same time covers vectors giving feasible directions.

Tangent vectors: For a closed set $C \subset \mathbb{R}^n$ and a point $\bar{x} \in C$, a vector w is said to be *tangent* to C at $\bar{x}$ if there exists a sequence of vectors $x^\nu \to \bar{x}$ in C along with a sequence of scalars $\tau^\nu \searrow 0$ such that the vectors $w^\nu = (x^\nu - \bar{x})/\tau^\nu$ (defined so that $x^\nu = \bar{x} + \tau^\nu w^\nu$) satisfy $w^\nu \to w$. (**Notation**: $\tau^\nu \searrow 0$ means $\tau^\nu \to 0$ with $\tau^\nu > 0$.)

Interpretation: The tangent vectors w to C at $\bar{x}$, apart from $w = 0$ (which corresponds in the definition to $x^\nu \equiv \bar{x}$, $w^\nu \equiv 0$), are the vectors pointing in a possibly asymptotic direction from which a sequence of points $x^\nu \in C$ can converge to $\bar{x}$. The direction of x^ν as seen from $\bar{x}$, which is the direction of w^ν, is not necessarily that of w, but gets closer and closer to it as $\nu \to \infty$.

Relation to feasible directions: Every vector $w \neq 0$ giving a feasible direction into C at $\bar{x}$ is a tangent vector to C at $\bar{x}$. Indeed, for such a vector w one can take $x^\nu = \bar{x} + \tau^\nu w$ for *any* sequence of values $\tau^\nu \searrow 0$ sufficiently small and have for $w^\nu = (x^\nu - \bar{x})/\tau^\nu$ that $w^\nu \equiv w$, hence trivially $w^\nu \to w$. Note from this that tangent vectors, in the sense defined here, can well point right into the interior of C, if that is nonempty; they don't have to lie along the boundary of C.

Relation to classical tangent spaces: When C is a "nice two-dimensional surface" in $\mathbb{R}^3$, the tangent vectors w to C at a point $\bar{x}$ form a two-dimensional linear subspace of $\mathbb{R}^3$, which gives the usual *tangent space* to C at $\bar{x}$. When C is a "nice one-dimensional curve," a one-dimensional linear subspace is obtained instead. Generalization can be made to tangent spaces to "nice smooth manifolds" of various dimensions m in $\mathbb{R}^n$, with $0 < m < n$. These comments are offered here on a preliminary level for geometric motivation, but a rigorous version will be developed in the next chapter in terms of sets defined by equality constraints.

Tangent cone at a point: The set of all vectors w that are tangent to C at a point $\bar{x} \in C$ is called the *tangent cone* to C at $\bar{x}$ and is denoted by $T_C(\bar{x})$.

Basic properties: Always, the set $T_C(\bar{x})$ contains the vector 0. Further, for every vector $w \in T_C(\bar{x})$ and every scalar $\lambda > 0$, the vector λw is again in $T_C(\bar{x})$ (because of the arbitrary size of the scaling factors τ^ν in the definition). But in many situations

one can have $w \in T_C(\bar{x})$, yet $-w \notin T_C(\bar{x})$. In particular, $T_C(\bar{x})$ can well be something other than a linear subspace of $\mathbb{R}^n$.

Cones: A subset of $\mathbb{R}^n$ is called a *cone* if it contains the zero vector and contains with each of its vectors all positive multiples of that vector. Geometrically, this means that a cone, unless it consists of 0 alone, is the union of a family of rays.

Limits of tangent vectors: The limit of any convergent sequence of vectors $w^\nu \in T_C(\bar{x})$ is another vector $w \in T_C(\bar{x})$. In other words, $T_C(\bar{x})$ is always a closed set. (This can readily be gleaned from the definition of $T_C(\bar{x})$ and the general properties of convergent sequences.) In particular, if the vectors w^ν give feasible directions to C at $\bar{x}$ and $w^\nu \to w$, then $w \in T_C(\bar{x})$.

Extreme examples of tangent cones: If $\bar{x}$ lies in the interior of C, then $T_C(\bar{x}) = \mathbb{R}^n$. In particular, $T_C(\bar{x}) = \mathbb{R}^n$ for all $\bar{x}$ when $C = \mathbb{R}^n$. At the opposite extreme, if C is a "singleton" (one-element) set $\{a\}$ and $\bar{x} = a$, then $T_C(\bar{x}) = \{0\}$. More generally, $T_C(\bar{x}) = \{0\}$ whenever $\bar{x}$ is an *isolated* point of C (in the sense that there is no sequence of points $x^\nu \neq \bar{x}$ in C with $x^\nu \to \bar{x}$).

Fundamental role of tangent cones in optimality: Much of the motivation for the study of tangent vectors comes from the following observation.

Basic necessary condition for a constrained minimum: If a function f_0 (of class $\mathcal{C}^1$, say) has a local minimum over C at $\bar{x}$, then f_0 has no descent vector $w \in T_C(\bar{x})$, i.e.,

$$\nabla f_0(\bar{x}) \cdot w \geq 0 \quad \text{for every} \quad w \in T_C(\bar{x}).$$

Argument: Consider any $w \in T_C(\bar{x})$. By definition there exist sequences $x^\nu \to \bar{x}$ in C and $\tau^\nu \searrow 0$ such that $(x^\nu - \bar{x})/\tau^\nu \to w$. Because $\bar{x}$ is locally optimal, we must have $f_0(x^\nu) - f_0(\bar{x}) \geq 0$ once x^ν is within a certain neighborhood of $\bar{x}$, hence for all indices ν beyond some ν_0. Our differentiability assumption implies that $f_0(x^\nu) - f_0(\bar{x}) = \nabla f_0(\bar{x}) \cdot [x^\nu - \bar{x}] + o(|x^\nu - \bar{x}|)$. By setting $w^\nu = (x^\nu - \bar{x})/\tau^\nu$, so that $x^\nu = \bar{x} + \tau^\nu w^\nu$ with $w^\nu \to w$, we see that $0 \leq \nabla f_0(\bar{x}) \cdot \tau^\nu w^\nu + o(\tau^\nu |w^\nu|)$ when $\nu \geq \nu_0$. Dividing by τ^ν and taking the limit as $\nu \to \infty$, we get $0 \leq \nabla f_0(\bar{x}) \cdot w$.

Tangents to boxes: If $X = I_1 \times \cdots \times I_n$ in $\mathbb{R}^n$ with I_j a closed interval in $\mathbb{R}$, then at any point $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n) \in X$ (the component $\bar{x}_j$ lying in I_j) one has

$$T_X(\bar{x}) = T_{I_1}(\bar{x}_1) \times \cdots \times T_{I_n}(\bar{x}_n), \text{ where}$$

$$T_{I_j}(\bar{x}_j) = \begin{cases} [0, \infty) & \text{if } \bar{x}_j \text{ is the left endpoint (only) of } I_j, \\ (-\infty, 0] & \text{if } \bar{x}_j \text{ is the right endpoint (only) of } I_j, \\ (-\infty, \infty) & \text{if } \bar{x}_j \text{ lies in the interior of } I_j, \\ [0, 0] & \text{if } I_j \text{ is a one-point interval, consisting just of } \bar{x}_j. \end{cases}$$

In other words, the condition $w \in T_X(\bar{x})$ for $w = (w_1, \ldots, w_n)$ amounts to restricting w_j to lie in a one of the intervals $(-\infty, 0]$, $[0, \infty)$, $(-\infty, \infty)$, or $[0, 0]$ (the one-point interval consisting just of 0). The particular interval for an index j depends on the location of $\bar{x}_j$ relative to the endpoints of I_j.

Tangents to the nonnegative orthant: The set $X = \mathbb{R}^n_+$ is a box, the product of the intervals $I_j = [0, \infty)$. In this case one has $w \in T_X(\bar{x})$ if and only if $w_j \geq 0$ for all indices j with $\bar{x}_j = 0$. (For indices j with $\bar{x}_j > 0$, w_j can be any real number.)

Feasible directions in a box: When X is a box, not only does every vector $w \neq 0$ giving a feasible direction in X at $\bar{x}$ belong to $T_X(\bar{x})$, but conversely, every vector $w \neq 0$ in $T_X(\bar{x})$ gives a feasible direction in X at $\bar{x}$. This important property generalizes as follows to *any* set specified by finitely many linear constraints.

Tangents to polyhedral sets: For a polyhedral set C, like a box, simple formulas describe all the tangent cones. As a specific case that's ripe for many applications (and fits with the description of "linear constraints" on page 17), suppose

$$x \in C \iff \begin{cases} a_i \cdot x \leq b_i & \text{for } i \in [1, s], \\ a_i \cdot x = b_i & \text{for } i \in [s+1, m], \\ x \in X & \text{with } X \text{ a box (possibly } X = \mathbb{R}^n). \end{cases}$$

Let $\bar{x} \in C$, and call $i \in [1, s]$ *active* if $a_i \cdot \bar{x} = b_i$, but *inactive* if $a_i \cdot \bar{x} < b_i$. Then

$$w \in T_C(\bar{x}) \iff \begin{cases} a_i \cdot w \leq 0 & \text{for active } i \in [1, s], \\ a_i \cdot w = 0 & \text{for } i \in [s+1, m], \\ w \in T_X(\bar{x}) & \text{(this cone being again a box).} \end{cases}$$

Argument: Let K denote the set of vectors w described by the final conditions (on the right). The question is whether $K = T_C(\bar{x})$. For a vector w and scalar $\tau > 0$, one has $\bar{x} + \tau w \in C$ if and only if $\bar{x} + \tau w \in X$ and $a_i \cdot w \leq [b_i - a_i \cdot \bar{x}]/\tau$ for $i \in [1, s]$, whereas $a_i \cdot w = 0$ for $i \in [s+1, m]$. Here for the indices $i \in [1, s]$ we have $[b_i - a_i \cdot \bar{x}] = 0$ if i is active, but $[b_i - a_i \cdot \bar{x}] > 0$ if i is inactive; for the latter, $[b_i - a_i \cdot \bar{x}]/\tau^\nu \nearrow \infty$ whenever $\tau^\nu \searrow 0$. In light of the previously derived formula for the tangent cone $T_X(\bar{x})$ to the box X, it's clear then that the vectors expressible as $w = \lim_\nu w^\nu$ with $\bar{x} + \tau^\nu w^\nu \in C$, $\tau^\nu \searrow 0$ (i.e., the vectors $w \in T_C(\bar{x})$) are none other than the vectors $w \in K$, as claimed.

Special tangent cone properties in the polyhedral case: When C is polyhedral, the tangent cone $T_C(\bar{x})$ at any point $\bar{x} \in C$ is polyhedral too. Moreover it fully describes the geometry of C around $\bar{x}$ in the sense that there is a $\rho > 0$ for which

$$\big\{x - \bar{x} \,\big|\, x \in C,\ |x - \bar{x}| \leq \rho\big\} \;=\; \big\{w \in T_C(\bar{x}) \,\big|\, |w| \leq \rho\big\}.$$

Reason: In accordance with the definition of C being polyhedral, there's no loss of generality in taking C to have a representation of the sort just examined. The corresponding representation for $T_C(\bar{x})$ then immediately supports these conclusions. Since $a_i \cdot w \leq |a_i||w|$, the claimed equation holds for any $\rho > 0$ small enough that $\rho \leq [b_i - a_i \cdot \bar{x}]/|a_i|$ for every inactive $i \in [1, s]$ (with $a_i \neq 0$).

Feasible directions in a polyhedral set: When C is polyhedral, the vectors w giving feasible directions into C at a point $\bar{x} \in C$ are exactly the vectors $w \neq 0$ in the tangent cone $T_C(\bar{x})$. This is clear from the observation just made about how, in the polyhedral case, $T_C(\bar{x})$ captures the local geometry of C near $\bar{x}$.

Characterizations of optimality: To what extent can the necessary and sufficient conditions for local optimality in unconstrained minimization in Theorem 3 be extended to minimization over a set C? This is a complex matter, because not only the "curvature" of f_0 as embodied in its Hessian matrices, but also that of the boundary of C can be crucial, yet we don't have any handle so far on analyzing the latter. Nonetheless, a substantial extension can already be stated for the case where C lacks curved portions because of being polyhedral.

THEOREM 6 (local optimality conditions on a polyhedral set). *Consider the problem of minimizing f_0 over a polyhedral set C, with f_0 of class $\mathcal{C}^2$. Let $\bar{x} \in C$.*

(a) (necessary). *If $\bar{x}$ is a locally optimal solution, then*

$$\nabla f_0(\bar{x}) \cdot w \geq 0 \text{ for every } w \in T_C(\bar{x}),$$
$$w \cdot \nabla^2 f_0(\bar{x}) w \geq 0 \text{ for every } w \in T_C(\bar{x}) \text{ satisfying } \nabla f_0(\bar{x}) \cdot w = 0.$$

(b) (sufficient). *If $\bar{x}$ has the property that*

$$\nabla f_0(\bar{x}) \cdot w \geq 0 \text{ for every } w \in T_C(\bar{x}),$$
$$w \cdot \nabla^2 f_0(\bar{x}) w > 0 \text{ for every } w \in T_C(\bar{x}) \text{ satisfying } \nabla f_0(\bar{x}) \cdot w = 0,\ w \neq 0,$$

then $\bar{x}$ is a locally optimal solution. Moreover, in these circumstances the local optimality of $\bar{x}$ is strict, in the sense that there exists a $\delta > 0$ such that

$$f_0(x) > f_0(\bar{x}) \text{ for all points } x \in C \text{ with } 0 < |x - \bar{x}| < \delta.$$

Proof. The argument is an adaptation of the one for Theorem 3. To set the stage, we invoke the polyhedral nature of C to get the existence of $\rho > 0$ such that the points $x \in C$ with $0 < |x - \bar{x}| \leq \rho$ are the points expressible as $\bar{x} + \tau w$ for some vector $w \in T_C(\bar{x})$ with $|w| = 1$ and scalar $\tau \in (0, \rho]$. Then too, for any $\delta \in (0, \rho)$, the points $x \in C$ with

$0 < |x - \bar{x}| \leq \delta$ are the points expressible as $\bar{x} + \tau w$ for some $w \in T_C(\bar{x})$ with $|w| = 1$ and some $\tau \in (0, \delta]$. Next we use the twice differentiability of f_0 to get second-order estimates around $\bar{x}$ in this notation: for any $\varepsilon > 0$ there is a $\delta > 0$ such that

$$\left| f_0(\bar{x} + \tau w) - f_0(\bar{x}) - \tau \nabla f_0(\bar{x}) \cdot w - \frac{\tau^2}{2} w \cdot \nabla^2 f_0(\bar{x}) w \right| \leq \varepsilon \tau^2 \quad \text{for all } \tau \in [0, \delta] \text{ when } |w| = 1.$$

In (a), the local optimality of $\bar{x}$ gives in this setting the existence of $\bar{\delta} > 0$ such that $f_0(\bar{x} + \tau w) - f_0(\bar{x}) \geq 0$ for $\tau \in [0, \bar{\delta}]$ when $w \in T_C(\bar{x})$, $|w| = 1$. For such w and any $\varepsilon > 0$ we then have through second-order expansion the existence of $\delta > 0$ such that

$$\tau \nabla f_0(\bar{x}) \cdot w + \frac{\tau^2}{2} \Big[w \cdot \nabla^2 f_0(\bar{x}) w + 2\varepsilon \Big] \geq 0 \text{ for all } \tau \in [0, \delta].$$

This condition implies that $\nabla f_0(\bar{x}) \cdot w \geq 0$, and if actually $\nabla f_0(\bar{x}) \cdot w = 0$ then also that $w \cdot \nabla^2 f_0(\bar{x}) w + 2\varepsilon \geq 0$. Since ε can be chosen arbitrarily, it must be true in the latter case that $w \cdot \nabla^2 f_0(\bar{x}) w \geq 0$. Thus, the claim in (a) is valid for all $w \in T_C(\bar{x})$ with $|w| = 1$. It is also valid then for positive multiples of such vectors w, and hence for all $w \in T_C(\bar{x})$.

In (b), the desired conclusion corresponds to the existence of $\delta > 0$ such that $f_0(\bar{x} + \tau w) - f_0(\bar{x}) > 0$ for $\tau \in (0, \delta]$ when $w \in T_C(\bar{x})$, $|w| = 1$. Through the second-order expansion it suffices to demonstrate the existence of $\varepsilon > 0$ and $\delta' > 0$ such that

$$\tau \nabla f_0(\bar{x}) \cdot w + \frac{\tau^2}{2} \Big[w \cdot \nabla^2 f_0(\bar{x}) w - 2\varepsilon \Big] > 0 \quad \text{for all } \tau \in (0, \delta'] \text{ when } w \in T_C(\bar{x}),\ |w| = 1.$$

Pursuing an argument by contradiction, let's suppose that such ε and δ' don't exist. Then, for any sequence $\varepsilon^\nu \searrow 0$ there must be sequences $\tau^\nu \searrow 0$ and $w^\nu \in T_C(\bar{x})$ with $|w^\nu| = 1$ and

$$\tau^\nu \nabla f_0(\bar{x}) \cdot w^\nu + \frac{\tau^{\nu 2}}{2} \Big[w^\nu \cdot \nabla^2 f_0(\bar{x}) w^\nu - 2\varepsilon^\nu \Big] \leq 0.$$

Because the sequence of vectors w^ν is bounded, it has a cluster point w; there is a subsequence $w^{\nu_\kappa} \to w$ as $\kappa \to \infty$. Then $|w| = 1$ (because the norm is a continuous function), and $w \in T_C(\bar{x})$ (because the tangent cone is a closed set). Rewriting our inequality as

$$w^{\nu_\kappa} \cdot \nabla^2 f_0(\bar{x}) w^{\nu_\kappa} - 2\varepsilon^{\nu_\kappa} \leq -2 \nabla f_0(\bar{x}) \cdot w^{\nu_\kappa} / \tau^{\nu_\kappa},$$

where $\nabla f_0(\bar{x}) \cdot w^{\nu_\kappa} \geq 0$ under the assumption of (b), we see when $\kappa \to \infty$ with

$$\nabla f_0(\bar{x}) \cdot w^{\nu_\kappa} \to \nabla f_0(\bar{x}) \cdot w, \qquad w^{\nu_\kappa} \cdot \nabla^2 f_0(\bar{x}) w^{\nu_\kappa} + 2\varepsilon^{\nu_\kappa} \to w \cdot \nabla^2 f_0(\bar{x}) w,$$

that $w \cdot \nabla^2 f_0(\bar{x}) w \leq 0$, yet also $\nabla f_0(\bar{x}) \cdot w = 0$ (for if $\nabla f_0(\bar{x}) \cdot w > 0$ the right side of the inequality would go to $-\infty$). This mix of properties of w is impossible under the assumption of (b). The contradiction finishes the proof. □

Remark: In the case where $C = \mathbb{R}^n$, so that $T_C(\bar{x}) = \mathbb{R}^n$, the assertions in Theorem 6 turn precisely into the ones for unconstrained minimization in Theorem 3. Thus, the version of optimality conditions just obtained subsumes the earlier one.

Minimization subject to linear constraints: When the polyhedral set C is expressed in terms of a system of linear constraints, an expression for $T_C(\bar{x})$ in terms of an associated system of linear constraints is achieved in the manner explained on page 52. This associated system of constraints on w can be substituted then for the condition $w \in T_C(\bar{x})$ wherever that appears in Theorem 6.

Optimality over a box: In particular Theorem 6 applies to a standard problem $(\mathcal{P})$ without constraint functions f_i, but just an abstract constraint $x \in X$ in which X is a box—specifying upper and/or lower bounds (or nonnegativity conditions) on the variables x_j. Then $C = X$.

Example: Consider a problem in which $f_0(x_1, x_2)$ is minimized over the quadrant $X = \big\{(x_1, x_2) \,\big|\, x_1 \geq 0,\, x_2 \geq 0\big\}$. What can be said about the possible attainment of the minimum at $\bar{x} = (0, 1)$? The tangent cone $T_X(\bar{x})$ at this point consists of all vectors $w = (w_1, w_2)$ with $w_1 \geq 0$. In both (a) and (b) of Theorem 6, the first-order condition on $\nabla f_0(\bar{x})$ comes down to the requirement that $(\partial f_0/\partial x_1)(\bar{x}) \geq 0$ but $(\partial f_0/\partial x_2)(\bar{x}) = 0$. In the case where $(\partial f_0/\partial x_1)(\bar{x}) > 0$, the second-order necessary condition in (a) is $(\partial^2 f_0/\partial x_2^2)(\bar{x}) \geq 0$, while the second-order sufficient condition in (b) is $(\partial^2 f_0/\partial x_2^2)(\bar{x}) > 0$. When $(\partial f_0/\partial x_1)(\bar{x}) = 0$, however, the condition in (a) is $w \cdot \nabla^2 f_0(\bar{x}) w \geq 0$ for all $w = (w_1, w_2)$ with $w_1 \geq 0$, while the condition in (b) is $w \cdot \nabla^2 f_0(\bar{x}) w > 0$ for all such $w \neq (0, 0)$.

Convexity in optimization: In constrained as well as in unconstrained minimization, convexity is a watershed concept. The distinction between problems of "convex" and "nonconvex" type is much more significant in optimization than that between problems of "linear" and "nonlinear" type.

Convex sets: A set $C \subset \mathbb{R}^n$ is *convex* if for every choice of $x_0 \in C$ and $x_1 \in C$ with $x_0 \neq x_1$ and every $\tau \in (0, 1)$ the point $(1 - \tau)x_0 + \tau x_1$ belongs to C.

Interpretation: This means that C contains with every pair of points the line segment joining them. Although "convex" in English ordinarily refers to a "bulging" appearance, the mathematical meaning is that there are no dents, gaps or holes.

Elementary rules for manipulating convex sets: The following facts about how convexity is preserved under various operations can readily be established from the definition of convexity.

Intersections: If C_i is a convex set in $\mathbb{R}^n$ for $i = 1, \ldots, r$, then $C_1 \cap \cdots \cap C_r$ is a convex set in $\mathbb{R}^n$. (This is true in fact not just for a finite intersection but the intersection of an arbitrary infinite family of convex sets.)

Products: If C_i is a convex set in $\mathbb{R}^{n_i}$ for $i = 1, \ldots, r$, then $C_1 \times \cdots \times C_r$ is a convex set in the space $\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_r} = \mathbb{R}^{n_1 + \cdots + n_r}$.

Images: If C is a convex set in $\mathbb{R}^n$, A is a matrix in $\mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, then the set $D = \{Ax + b \,|\, x \in C\}$ is convex in $\mathbb{R}^m$.

Inverse images: If D is a convex set in $\mathbb{R}^m$, A is a matrix in $\mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, then the set $C = \{x \,|\, Ax + b \in D\}$ is convex in $\mathbb{R}^n$.

Basic examples of convex sets:

Extremes: The whole space $C = \mathbb{R}^n$ is convex. On the other hand, the empty set $C = \emptyset$ is convex. (It satisfies the definition "vacuously.") Likewise, sets $C = \{a\}$ consisting of just a single point (*singleton* sets) are convex sets.

Linear subspaces: Any subspace of $\mathbb{R}^n$ (as in linear algebra) is closed and convex.

Intervals: The convex subsets of the real line $\mathbb{R}$ are the various *intervals*, whether bounded, unbounded, open, closed, or a mixture.

Boxes and orthants: As a product of closed intervals, any box is a closed, convex set. For instance the nonnegative orthant $\mathbb{R}^n_+$, being a box, is closed and convex. So too is the nonpositive orthant $\mathbb{R}^n_-$, which is defined analogously.

Hyperplanes and half-spaces: All such sets are closed and convex—as an elementary consequence of their definition.

Polyhedral sets: As the intersection of a family of hyperplanes or closed half-spaces, any polyhedral set is closed and convex.

Euclidean balls: For any point $\bar{x} \in \mathbb{R}^n$ and radius value $\rho \in (0, \infty)$, the *closed ball* of radius ρ around $\bar{x}$ consists of the points x with $|x - \bar{x}| \leq \rho$. The corresponding *open ball* or radius ρ around $\bar{x}$ is defined in the same way, but with strict inequality. Both kinds of balls are examples of convex sets in $\mathbb{R}^n$.

Argument. For the case of $C = \{x \,|\, |x - \bar{x}| \leq \rho\}$, consider x_0 and x_1 in C and $\tau \in (0, 1)$. For $x = (1-\tau)x_0 + \tau x_1$, we can use the fact that $\bar{x} = (1-\tau)\bar{x} + \tau\bar{x}$

to write $|x-\bar{x}| = \big|(1-\tau)(x_0-\bar{x})+\tau(x_1-\bar{x})\big| \le (1-\tau)|x_0-\bar{x}|+\tau|x_1-\bar{x}| \le (1-\tau)\rho+\tau\rho = \rho$, from which we conclude that $x \in C$.

Epigraphs of convex functions: A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex *if and only if* its epigraph $E = \big\{(x,u) \in \mathbb{R}^n \times \mathbb{R} \,\big|\, u \ge f(x)\big\}$ is convex as a subset of $\mathbb{R}^{n+1}$. This fact is the source of a great many insights.

Tangents to convex sets: For a convex set $C \subset \mathbb{R}^n$, the tangent cone $T_C(\bar{x})$ at any point $\bar{x} \in C$ consists of the zero vector and all vectors $w \neq 0$ expressible as limits of sequences of vectors $w^\nu \neq 0$ giving feasible directions into C at $\bar{x}$.

Argument. This follows from the observation that when C is convex, the presence of $\bar{x}+\tau w$ in C entails that the entire line segment from $\bar{x}$ to $\bar{x}+\tau w$ lies in C. Likewise, if $\bar{x}+\tau^\nu w^\nu \in C$ then $\bar{x}+\tau w^\nu \in C$ for all $\tau \in [0,\tau^\nu]$.

Convexity of tangent cones to convex sets: When C is a convex set, $T_C(\bar{x})$ is always a convex cone. This follows, for instance, from the characterization just given.

THEOREM 7 (optimality over a convex set). *Consider a problem of minimizing a function f_0 of class $\mathcal{C}^1$ over a convex set $C \subset \mathbb{R}^n$. Let $\bar{x}$ be a point of C.*

(a) (necessary). *If $\bar{x}$ is locally optimal, then $\nabla f_0(\bar{x})\cdot w \ge 0$ for all $w \in T_C(\bar{x})$, and this is equivalent in fact to the condition that*

$$\nabla f_0(\bar{x})\cdot[x-\bar{x}] \ge 0 \text{ for all } x \in C.$$

(b) (sufficient). *If this condition holds and f_0 is convex on C, then $\bar{x}$ is globally optimal.*

Proof. In (a), consider first any point $x \in C$ different from $\bar{x}$. The line segment joining $\bar{x}$ with x lies in C by convexity; the vector $w = x-\bar{x}$ gives a feasible direction into C at $\bar{x}$. The function $\varphi(\tau) = f(\bar{x}+\tau w)$ then has a local minimum at $\tau = 0$ relative to $0 \le \tau \le 1$. Hence $0 \le \varphi'(0) = \nabla f_0(\bar{x})\cdot w = \nabla f_0(\bar{x})\cdot[x-\bar{x}]$. From this we see further that $\nabla f_0(\bar{x})\cdot w \ge 0$ for *all* vectors w giving feasible directions into C at $\bar{x}$, inasmuch as these are positive multiples $\lambda(x-\bar{x})$ of vectors of the form $x-\bar{x}$ with $x \in C$.

As noted above, any vector $w \neq 0$ in $T_C(\bar{x})$ is a limit of vectors w^ν giving feasible directions (which themselves are in $T_C(\bar{x})$ as well). From having $\nabla f_0(\bar{x})\cdot w^\nu \ge 0$ for all ν we get in the limit as $w^\nu \to w$ that $\nabla f_0(\bar{x})\cdot w \ge 0$. Therefore, $\nabla f_0(\bar{x})\cdot w \ge 0$ for all $w \in T_C(\bar{x})$, and this is equivalent to having $\nabla f_0(\bar{x})\cdot[x-\bar{x}] \ge 0$ for all $x \in C$.

In (b), we have $f_0(x) - f_0(\bar{x}) \ge \nabla f_0(\bar{x})\cdot[x-\bar{x}]$ for all $x \in \mathbb{R}^n$ by the convexity of f_0 (Theorem 4). If also $\nabla f_0(\bar{x})\cdot[x-\bar{x}] \ge 0$ for all $x \in C$, we get $f_0(x) - f_0(\bar{x}) \ge 0$ for all $x \in C$, which means $\bar{x}$ is globally optimal in the minimization of f_0 over C. □

Interpretation via linearization: In terms of the function $l(x) = f_0(\bar{x}) + \nabla f_0(\bar{x}) \cdot [x - \bar{x}]$ giving the first-order expansion of f_0 at $\bar{x}$, the necessary condition for optimality in Theorem 7 is equivalent to saying that $l(x) \geq l(\bar{x})$ for all $x \in C$, or in other words, that l attains its global minimum over C at $\bar{x}$.

Variational inequalities: The condition in Theorem 7 that $\nabla f_0(\bar{x}) \cdot [x - \bar{x}] \geq 0$ for all $x \in C$ is known as the *variational inequality* for the mapping ∇f_0 and the convex set C, the point $\bar{x}$ being a *solution* to it. Variational inequalities can be studied also with ∇f_0 replaced by some other vector-valued mapping F from C into $\mathbb{R}^n$. They have an interesting and significant place in optimization theory beyond merely the characterization of points at which a minimum is attained.

Convex functions on convex sets: The convexity of a function f on $\mathbb{R}^n$ has already been defined in terms of the inequality $f\big((1-\tau)x_0 + \tau x_1\big) \leq (1-\tau)f(x_0) + \tau f(x_1)$ holding for all x_0 and x_1 in $\mathbb{R}^n$ and $\tau \in (0,1)$. The concept can be generalized now to the convexity of f *on a convex set* $C \subset \mathbb{R}^n$: the same inequality is used, but x_0 and x_1 are restricted to C. Similarly one speaks of f being strictly convex, or concave, or strictly concave on C. (Note that for these concepts to make sense f only has to be defined on C itself; values of f outside of C have no effect, because the convexity of C ensures that the relevant points $(1-\tau)x_0 + \tau x_1$ all belong to C.)

Derivative tests: The tests in Theorem 4 apply equally well to the convexity or strict convexity of a differentiable function f on any *open* convex set $O \subset \mathbb{R}^n$.

Epigraphical test: The definition of f being convex on C corresponds geometrically to the convexity of the set $\big\{(x,\alpha) \in C \times \mathbb{R} \,\big|\, \alpha \geq f(x)\big\}$.

Convexity-preserving operations on functions: Derivative tests are by no means the only route to verifying convexity or strict convexity. Often it's easier to show that a given function is convex because it is constructed by convexity-preserving operations from other functions, already known to be convex (or, as a special case, affine). The following operations are convexity-preserving on the basis of elementary arguments using the definition of a convexity. Here C denotes a general convex set.

Sums: If f_1 and f_2 are convex functions on C, then so is $f_1 + f_2$. (This can be extended to a sum of any number of convex functions.) Moreover, if one of the functions in the sum is *strictly* convex, then the resulting function is strictly convex.

Multiples: If f is convex on C and $\lambda \geq 0$, then λf is convex on C. (In combination with the preceding, this implies that any linear combination $\lambda_1 f_1 + \cdots + \lambda_r f_r$ of convex functions with coefficients $\lambda_i \geq 0$ is convex.) Again, if one of the functions

in the sum is *strictly* convex, and the associated coefficient is positive, then the function expressed by the sum is strictly convex.

Compositions I: If f is convex on C, then any function of the form $g(x) = \theta\big(f(x)\big)$ is convex on C, provided that the function θ on $\mathbb{R}^1$ is convex *and nondecreasing*.

Example: If f is convex in $\mathbb{R}^n$ *and* $f \geq 0$ *everywhere*, then the function $g(x) := f(x)^2$ is convex on $\mathbb{R}^n$, because $g(x) = \theta\big(f(x)\big)$ for θ defined by $\theta(t) = t^2$ when $t \geq 0$, but $\theta(t) = 0$ when $t \leq 0$. This follows because θ is convex and nondecreasing (the convexity can be verified from the fact that $\theta'(t) = \max\{0, 2t\}$, which is nondecreasing). The tricky point is that unless $f \geq 0$ everywhere it would not be possible to write f as composed from this θ. Composition with $\theta(t) := t^2$ for all t wouldn't do, because this θ, although convex, isn't nondecreasing as a function on all of $\mathbb{R}^1$.

Compositions II: If f is convex on C, then $g(x) := f(Ax + b)$ is convex on $D := \big\{x \,\big|\, Ax + b \in C\big\}$ for any matrix $A \in \mathbb{R}^{n\times n}$ and vector $b \in \mathbb{R}^n$.

Pointwise max: If f_i is convex on C for $i = 1, \ldots, r$ and $f(x) = \max\big\{f_1(x), \ldots, f_r(x)\big\}$, then f is convex on C. Likewise for strict convexity.

Level sets of convex functions: If f is a convex function on $\mathbb{R}^n$, then for any $\alpha \in \mathbb{R}$ the sets $\big\{x \,\big|\, f(x) \leq \alpha\big\}$ and $\big\{x \,\big|\, f(x) < \alpha\big\}$ are convex. (Similarly, if f is a concave function the sets $\big\{x \,\big|\, f(x) \geq \alpha\big\}$ and $\big\{x \,\big|\, f(x) > \alpha\big\}$ are convex.)

Argument: These facts can be deduced by applying the definition of convexity to the sets in question and appealing to the convexity inequality satisfied by f.

Convex constraints: A *convex* constraint is a condition of the form $f_i(x) \leq c_i$ with f_i convex, or $f_i(x) \geq c_i$ with f_i concave, or $f_i(x) = c_i$ with f_i affine. Also, a condition of the form $x \in X$ is called a convex constraint if X is convex. Thus, a system of the form

$$x \in X \text{ and } f_i(x) \begin{cases} \leq 0 & \text{for } i = 1, \ldots, s, \\ = 0 & \text{for } i = s+1, \ldots, m, \end{cases}$$

is a system of convex constraints when X is convex, f_i is convex for $i = 1, \ldots, s$, and f_i is affine for $i = s+1, \ldots, m$. *Any set C defined by a system of convex constraints is a convex set*, because each separate constraint requires x to belong to a certain convex set, and C is the intersection of these sets.

Convex programming: An optimization problem in standard format is called a *convex programming problem* if the constraints are convex, as just described, and also the objective function f_0 is convex.

Extension: This term is also used even if the objective and inequality constraint functions aren't convex over all of $\mathbb{R}^n$, as long as they are convex on the convex set X. The feasible set C is still convex in that case.

Linear programming: This has been defined already, but we can now interpret it as the case of convex programming where the objective function and all the constraint functions are actually affine, and the set X is a box.

Quadratic programming: This too is a special case of convex programming; it is just like linear programing in the constraints and objective, except that the objective function can include a positive semidefinite quadratic term. (But this view is slipping, and nowadays people often speak of quadratic programming even when a *nonconvex* polynomial function of degree two is minimized subject to linear constraints.)

THEOREM 8 (special characteristics of convex optimization). *In a problem of minimizing a convex function f_0 over a convex set $C \subset \mathbb{R}^n$ (and thus any problem of convex programming) the following properties hold.*

(a) (local is global) *Any locally optimal solution is a globally optimal solution. Moreover, the set of all optimal solutions (if any) is convex.*

(b) (uniqueness criterion) *Strict convexity of the objective function f_0 implies there cannot be more than one optimal solution.*

Proof. (a) Suppose the point $\bar{x} \in C$ is locally optimal, i.e., there is an $\varepsilon > 0$ such that $f_0(x) \geq f_0(\bar{x})$ for all $x \in C$ satisfying $|x - \bar{x}| < \varepsilon$. Suppose also that $\tilde{x} \in C$, $\tilde{x} \neq \bar{x}$. Our aim is to show that $f_0(\tilde{x}) \geq f_0(\bar{x})$, thereby establishing the global optimality of $\bar{x}$ relative to C. For any $\tau \in (0,1)$ we know that $f_0\big((1-\tau)\bar{x} + \tau\tilde{x}\big) \leq (1-\tau)f_0(\bar{x}) + \tau f_0(\tilde{x})$. By choosing τ small enough, we can arrange that the point $x_\tau := (1-\tau)\bar{x} + \tau\tilde{x}$ (which still belongs to C by the convexity of C) satisfies $|x_\tau - \bar{x}| < \varepsilon$. (It suffices to take $\tau < \varepsilon/|\tilde{x} - \bar{x}|$.) Then the left side of the convexity inequality, which is $f_0(x_\tau)$, cannot be less than $f_0(\bar{x})$ by the local optimality of $\bar{x}$. We deduce that $f_0(\bar{x}) \leq f_0(x_\tau) \leq (1-\tau)f_0(\bar{x}) + \tau f_0(\tilde{x})$, which from the outer terms, after rearrangement, tells us that $f_0(\bar{x}) \leq f_0(\tilde{x})$, as needed.

Having determined that $\bar{x}$ is globally optimal, we can apply the same argument for arbitrary $\tau \in (0,1)$, without worrying about any ε. If $\tilde{x}$ is another optimal solution, of course, we have $f_0(\tilde{x}) = f_0(\bar{x})$, so that the right side of the double inequality $f_0(\bar{x}) \leq f_0(x_\tau) \leq (1-\tau)f_0(\bar{x}) + \tau f_0(\tilde{x})$ reduces to $f_0(\bar{x})$ and we can conclude that $f_0(x_\tau) = f_0(\bar{x})$ for all $\tau \in (0,1)$. In other words, the entire line segment joining the two optimal solutions $\bar{x}$ and $\tilde{x}$ must consist of optimal solutions; the optimal set is convex.

(b) Looking at the displayed inequality in the first part of the proof of (a) in the case where f_0 is strictly convex, and $\tilde{x}$ is again just any point of C different from the optimal solution $\bar{x}$, we get strict inequality. This leads to the conclusion that $f_0(\bar{x}) < f_0(\tilde{x})$. It's impossible, therefore, for $\tilde{x}$ to be optimal as well as $\bar{x}$. □

Convexity in estimating progress toward optimality: Another distinguishing feature of optimization problems of convex type is that in numerical methods for solving such problems it's usually possible to devise tests of how close one is getting to optimality—global optimality—as the method progresses. By contrast, for most other kinds of optimization problems one has hardly any handle on this important issue, and the question of a *stopping criterion* for an iterative procedure can only be answered in an ad hoc manner.

Upper and lower bounds on the optimal value: A simple example of the kind of estimate that can be built into a stopping criterion can be derived from the linearization inequality for convex functions in Theorem 4. Consider a problem of minimizing a differentiable convex function f_0 over a nonempty, closed set $C \subset \mathbb{R}^n$ that's also bounded, and imagine that a numerical method has generated in iteration ν a point $x^\nu \in C$. The affine function $l^\nu(x) = f_0(x^\nu) + \nabla f_0(x^\nu) \cdot [x - x^\nu]$ has the property that $l^\nu(x) \leq f_0(x)$ for all x, and $l^\nu(x^\nu) = f_0(x^\nu)$. It follows that

$$\min_{x \in C} l^\nu(x) \;\leq\; \min_{x \in C} f_0(x) \;\leq\; f_0(x^\nu),$$

where the middle expression is the optimal value $\bar{\alpha}$ in the given problem, but the left expression, let's denote it by β^ν, is the optimal value in the possibly very easy problem of minimizing l^ν instead of f_0 over C. If C were a box, for instance, β^ν could instantly be calculated. While β^ν furnishes a current lower bound to $\bar{\alpha}$, the objective value $\alpha^\nu = f_0(x^\nu)$ furnishes a current upper bound. The difference $\alpha^\nu - \beta^\nu$ provides a measure of how far the point x^ν is from being optimal.

Duality: Optimization in a context of convexity is distinguished further by a pervasive phenomenon of "duality," in which a given problem of minimization ends up being paired with some problem of maximization in entirely different "dual" variables. Many important schemes of computation are based on this curious fact, or other aspects of convexity. In particular, almost all the known methods for breaking a large-scale problem down iteratively into small-scale problems, which perhaps could be solved in parallel, require convexity in their justification. This topic will be taken up later, after Lagrange multipliers for constraints have been introduced.

Minimization of nonsmooth convex functions: Not every convex function of interest in optimization is continuously differentiable.

Piecewise linear costs: Cost functions of a single variable often take the form of piecewise linear functions with increasing slope values. Such functions are convex. Specifically, suppose that a closed interval $C \subset \mathbb{R}^1$ is partitioned into a finite sequence of closed subintervals $C_1, C_2, \ldots, C_r$, and that the function $f : C \to \mathbb{R}$ is given on these subintervals by expressions

$$f(x) = a_i x + b_i \text{ when } x \in C_i$$

which agree at the points where the consecutive intervals join (so that f is continuous) and have $a_1 \leq a_2 \leq \cdots \leq a_r$. Then f is convex on C by the criterion for "pointwise max." In fact it can be seen in the given circumstances that

$$f(x) = \max\left\{a_1 x + b_1, \ldots, a_r x + b_r\right\} \text{ for all } x \in C.$$

Piecewise linear approximations: A smooth convex function f on $\mathbb{R}^n$ can be approximated from below by a nonsmooth convex function in a special way. We've already noted in connection with obtaining lower bounds on the optimal value in a problem of minimizing a convex function that the linearization (first-order Taylor expansion) of f at any point provides an affine lower approximation which is exact at the point in question. That degree of approximation is crude in itself, but imagine now what might be gained by linearizing f at more than one point. Specifically, consider a collection of finitely many points x_k, $k = 1, \ldots, r$, and at each such point the corresponding affine function obtained by linearization, namely

$$l_k(x) = f(x_k) + \nabla f(x_k) \cdot [x - x_k].$$

The function $g(x) = \max\{l_1(x), \ldots, l_r(x)\}$ is convex on $\mathbb{R}^n$ (because the pointwise max of finitely many convex functions is always convex), and it satisfies

$$g(x) \leq f(x) \text{ for all } x, \text{ with } g(x_k) = f(x_k) \text{ for } k = 1, \ldots, r.$$

The convex function g is termed "piecewise linear" because its epigraph, as the intersection of the epigraphs of the l_k's, each of which is an upper closed half-space in $\mathbb{R}^{n+1}$, is a polyhedral subset of $\mathbb{R}^{n+1}$.

Cutting plane methods: An interesting class of numerical methods in convex programming relies on replacing the objective function and the inequality constraint functions, to the extent that they aren't merely affine, by such piecewise linear approximations. The finite collection of points in $\mathbb{R}^n$ on which the approximations

are based is generated as the iterations proceed. These methods are called *cutting plane* methods because each new affine function entering one of the approximations cuts away part of the epigraph from the proceeding approximation.

Remark: Cutting plane methods tend to be quite slow in comparison with typical descent methods, but they are useful nonetheless in a number of situations where for some reason it's tedious or expensive to generate function values and derivatives, and approaches requiring line search are thereby precluded.

Minimizing a max of convex functions: In problems where a function of the form $f_0(x) = \max\{g_1(x), \ldots, g_r(x)\}$ is to be minimized over a set C specified by convex constraints, the case where each function g_k is convex and smooth is especially amenable to treatment. Then f_0 is convex, and although it isn't smooth itself the usual device of passing to an epigraphical formulation retains convexity while bringing the smoothness of the g_k's to the surface. When an extra variable u is added, and the problem is viewed as one of minimizing the value of u over all choices of $(x_1, \ldots, x_n, u) \in C \times \mathbb{R}$ such that $g_k(x_1, \ldots, x_n) - u \leq 0$ for $k = 1, \ldots, r$, it is seen that all constraints are convex.

Minimizing a max of affine functions over a polyhedral set: As a special case, if C is polyhedral and the functions g_k affine in the foregoing, the set $C \times \mathbb{R}$ will be polyhedral and the constraints $g_k(x_1, \ldots, x_n) - u \leq 0$ are linear. In expressing C itself by a system of linear constraints, one sees that reformulated problem isn't just one of convex programming, but of *linear* programming.

Application to cutting plane methods: The subproblems generated in the cutting plane scheme of piecewise linear approximation, as described above, can, after epigraphical reformulation, be solved as linear programming problems if the set over which f_0 is to be minimized is specified by linear constraints. More generally such a reduction to solving a sequence of linear programming problems is possible even if C is specified just by convex constraints over a box X, as long as the convex functions giving inequality constraints are smooth. The extension to this case involves generating piecewise linear approximations to those functions f_i along with the one to f_0 as computations proceed.

Norms: As another reminder that derivative tests aren't the only route to verifying convexity, consider any *norm* on $\mathbb{R}^n$, that is, a real-valued expression $\|x\|$ with the following properties, which generalize those of the Euclidean norm $|x|$.

(a) $\|x\| > 0$ for all $x \neq 0$,

(b) $\|\lambda x\| = |\lambda|\|x\|$ for all x and all λ,

(c) $\|x + y\| \leq \|x\| + \|y\|$ for all x and y.

The function $f(x) = \|x\|$ is convex, because for $0 < \tau < 1$ we have

$$f\Big((1-\tau)x_0 + \tau x_1\Big) = \big\|(1-\tau)x_0 + \tau x_1\big\| \\ \leq (1-\tau)\|x_0\| + \tau\|x_1\| = (1-\tau)f(x_0) + \tau f(x_1).$$

Commonly seen in problems of approximation are the *l^p-norms* for $p \in [1, \infty]$: for a point $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, one defines

$$\|x\|_1 := |x_1| + \cdots + |x_n|,$$
$$\|x\|_p := \Big(|x_1|^p + \cdots + |x_n|^p\Big)^{1/p} \text{ with } 1 < p < \infty \text{ (where } \|x\|_2 = |x|),$$
$$\|x\|_\infty := \max\Big\{|x_1|, \ldots, |x_n|\Big\}.$$

We won't try here to verify that these expressions do indeed give norms, i.e., that they satisfy the three axioms above, although this is elementary for $p = 1$ and $p = \infty$.

Example: Consider a parameter identification problem in which we wish to minimize an error expression of the form

$$E\big(r(a,b)\big) = \Big\|\big(r_1(a,b), \ldots, r_N(a,b)\big)\Big\|_p$$

with $r_k(a,b) := y_k - (ax_k + b)$ for $k = 1, \ldots, N$. (The unknowns here are the two parameter values a and b.) This expression is convex as a function of a and b, because it is obtained by composing an l^p-norm with an affine mapping.

Piecewise linear norms: The functions $f(x) = \|x\|_1$ and $f(x) = \|x\|_\infty$ are piecewise linear, in that each can be expressed as the pointwise max of a finite collection of affine (in fact linear) functions. Specifically, in terms of the vectors $e_j \in \mathbb{R}^n$ having coordinate 1 in jth position but 0 in all other positions, $\|x\|_1$ is the maximum of the 2^n linear functions $[\pm e_1 \pm e_2 \cdots \pm e_n]\cdot x$, whereas $\|x\|_\infty$ is the maximum of the $2n$ linear functions $\pm e_j \cdot x$. In contrast, the norm function $f(x) = \|x\|_2 = |x|$ can usually be treated in terms of its square, which is a simple quadratic convex function.

5. LAGRANGE MULTIPLIERS

Optimality with respect to minimization over a set $C \subset \mathbb{R}^n$ has been approached up to now in terms of *tangent* vectors to C at a point $\bar{x}$. Further progress depends now on developing a complementary notion of *normal* vectors to C at $\bar{x}$. Formulas for normal vectors in terms of a constraint representation for C will yield special coefficients, called Lagrange multipliers, which not only serve in the statement of necessary and sufficient conditions for optimality but take on an intriguing life of their own.

Traditional perspective: In calculus it's customary to treat tangents and normals to surfaces and curves. At a point of a surface in $\mathbb{R}^3$ there's a tangent plane and, perpendicular to it, a normal line. At a point on a curve in $\mathbb{R}^3$, there's instead a tangent line and a normal plane. The higher dimensional analogs of curves and surfaces are "smooth manifolds" in $\mathbb{R}^n$ defined by systems of equality constraints. At a point of such a smooth manifold of dimension m it's possible to speak similarly of a tangent space of dimension m and a normal space of dimension $n - m$, these being linear subspaces of $\mathbb{R}^n$ that are orthogonal to each other.

Broadened perspective: For purposes of optimization we need to work with inequality as well as equality constraints as well as with abstract constraints which perhaps could have a more general character. This forces us to adopt a much broader approach to "normality," leading to normal *cones* instead of subspaces.

Normal vectors: For a closed set $C \subset \mathbb{R}^n$ and a point $\bar{x} \in C$, a vector v is said to be *normal to C at $\bar{x}$ in the regular sense* (a "regular normal") if

$$v \cdot w \leq 0 \text{ for every } w \in T_C(\bar{x}).$$

It is *normal to C in the general sense* (a "general normal," or just a "normal") if it can be approximated by normals in the regular sense: namely, there exist sequences $v^\nu \to v$ and $x^\nu \to \bar{x}$ such that v^ν is a regular normal to C at x^ν.

Interpretation of regular normals: The normal vectors v to C at $\bar{x}$ in the regular sense, apart from $v = 0$, are the vectors that make a right or obtuse angle with every tangent vector w to C at $\bar{x}$. It's not hard to show that this holds if and only if

$$v \cdot (x - \bar{x}) \leq o\big(|x - \bar{x}|\big) \text{ for } x \in C.$$

The definition of a regular normal vector v could therefore just as well be given in terms of this inequality property.

Regular normals as general normals: Any normal vector v in the regular sense is in particular a normal vector in the general sense. (Consider $v^\nu \equiv v$ and $x^\nu \equiv \bar{x}$).

Regularity of a set: The set C is called (Clarke) *regular* at $\bar{x}$ (one of its points) if every normal at $\bar{x}$ in the general sense is in fact a normal at $\bar{x}$ in the regular sense, i.e., if the limit process in the definition doesn't yield any more normals than are already obtained by the angle condition in the definition of regular normals.

Example of irregularity: A heart-shaped set C fails to be regular at its "inward corner point," although it's regular everywhere else.

Normal cone at a point: For a closed set C, the set of all vectors v that are normal to C at a point $\bar{x} \in C$ in the general sense is called the *normal cone* to C at $\bar{x}$ and is denoted by $N_C(\bar{x})$.

Basic properties: Always, $N_C(\bar{x})$ contains the vector 0. For all $v \in N_C(\bar{x})$ and $\lambda > 0$, the vector λv is in $N_C(\bar{x})$. Thus, $N_C(\bar{x})$ is a "cone." When C is regular at $\bar{x}$, $N_C(\bar{x})$ is furthermore convex and contains for each v and v' also $v + v'$.

Extreme cases: If $C = \mathbb{R}^n$, then $N_C(\bar{x}) = \{0\}$. Indeed, one has $N_C(\bar{x}) = \{0\}$ whenever C is a set with $\bar{x}$ in its interior. On the other hand, if C is just a one-element set $\{a\}$, and $\bar{x} = a$, then $N_C(\bar{x}) = \mathbb{R}^n$.

Limits of general normals: The cone $N_C(\bar{x})$ always closed, and a stronger property even holds: if $v^\nu \to v$ with $v^\nu \in N_C(x^\nu)$ and $x^\nu \to \bar{x}$, then $v \in N_C(\bar{x})$.

Argument: By its definition, the set of $(x, v) \in \mathbb{R}^n \times \mathbb{R}^n$ such that v is a general normal to C at x is the closure of the set of (x, v) such that v is a regular normal to C at x. Hence in particular, it's a closed set in $\mathbb{R}^n \times \mathbb{R}^n$.

Normals to convex sets: A closed, convex set $C \subset \mathbb{R}^n$ is regular at any of its points $\bar{x}$, and its normal cone $N_C(\bar{x})$ consists of the vectors v such that

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

Proof. Previously we characterized $T_C(\bar{x})$ in the convex case as consisting of 0 and all vectors obtainable as limits of vectors $w \neq 0$ that give feasible directions in C at $\bar{x}$. On that basis, v is a *regular* normal at $\bar{x}$ if and only if $v \cdot w \leq 0$ for all such w. By convexity, a vector $w \neq 0$ gives a feasible direction in C at $\bar{x}$ if and only if $\bar{x} + \tau w \in C$ for some $\tau > 0$, or equivalently, $w = \lambda(x - \bar{x})$ for some $x \in C$ and $\lambda > 0$. Hence v is a *regular* normal if and only if $v \cdot (x - \bar{x}) \leq 0$ for all $x \in C$.

All that remains, then, is verifying that every normal at $\bar{x}$ is a regular normal. Consider a sequence of points $x^\nu \to \bar{x}$ in C and regular normals v^ν to C at these points, with v^ν converging to a vector v. We have to show that v too is a *regular* normal to C at $\bar{x}$. From the preceding, we know that for each ν we have

$v^\nu \cdot (x - x^\nu) \leq 0$ for all $x \in C$. Taking the limit in this inequality as $\nu \to \infty$ with x fixed, we get $v \cdot (x - \bar{x}) \leq 0$. This being true for an arbitrary choice of $x \in C$, we conclude, by the same criterion, that v is a regular normal to C at $\bar{x}$. □

Supporting half-space property: This characterization means that, in the case of a closed, *convex* set C and a point $\bar{x} \in C$, the nonzero vectors in $N_C(\bar{x})$ are the nonzero vectors v such that C lies in the half-space $H = \{x \in \mathbb{R}^n \,|\, v \cdot x \leq v \cdot \bar{x}\}$. A half-space H with this property is said to *support* C at $\bar{x}$. Note that $\bar{x}$ itself is on the boundary hyperplane of H.

Normals to linear subspaces: For a subspace M of $\mathbb{R}^n$, one has at every point $\bar{x} \in M$ that $N_M(\bar{x}) = M^\perp$, the set of vectors v such that $v \cdot w = 0$ for all $w \in M$. This is immediate from the fact that M is convex, using the characterization above.

Tangents versus normals under regularity: When a closed set C is regular at $\bar{x}$, the geometric relationship between tangents and normals is beautifully symmetric:

$$\begin{aligned} N_C(\bar{x}) &= \{v \,|\, v \cdot w \leq 0 \text{ for all } w \in T_C(\bar{x})\}, \\ T_C(\bar{x}) &= \{w \,|\, v \cdot w \leq 0 \text{ for all } v \in N_C(\bar{x})\}. \end{aligned}$$

Proof: The first equation merely combines the definition of regular normals with the assumption that all normals are regular. It yields the '$\subset$' half of the second equation. To show that $\supset$ holds in that equation as well, we fix any vector $\bar{w} \notin T_C(\bar{x})$ and aim at demonstrating the existence of a vector $\bar{v} \in N_C(\bar{x})$ such that $\bar{v} \cdot \bar{w} > 0$.

Replacing C by its intersection with some closed ball around $\bar{x}$ if necessary (which involves no loss of generality, since the generation of normal vectors depends only on a neighborhood of $\bar{x}$), we can suppose that C is compact. Let B stand for some closed ball around $\bar{w}$ that doesn't meet $T_C(\bar{x})$ (as exists because $T_C(\bar{x})$ is closed). The definition of $T_C(\bar{x})$, in conjunction with having $T_C(\bar{x}) \cap B = \emptyset$, implies the existence of an $\varepsilon > 0$ such that the compact, convex set $S = \{\bar{x} + \tau w \,|\, w \in B,\ \tau \in [0, \varepsilon]\}$ meets C only at $\bar{x}$. Consider any sequence $\varepsilon^\nu \searrow 0$ with $\varepsilon^\nu < \varepsilon$ along with the compact, convex sets $S^\nu = \{\bar{x} + \tau w \,|\, w \in B,\ \tau \in [\varepsilon^\nu, \varepsilon]\}$, which are disjoint from C.

The function $h(x, u) = \frac{1}{2}|x - u|^2$ attains its minimum over the compact set $C \times S^\nu$ at some point (x^ν, u^ν). In particular, x minimizes $h(x, u^\nu)$ with respect to $x \in C$, so by Theorem 9 the vector $-\nabla_x h(x^\nu, u^\nu) = u^\nu - x^\nu$ belongs to $N_C(x^\nu)$. Likewise, the vector $-\nabla h_u(x^\nu, u^\nu) = x^\nu - u^\nu$ belongs to $N_{S^\nu}(u^\nu)$. Necessarily $x^\nu \neq u^\nu$ because $C \cap S^\nu = \emptyset$, but $x^\nu \to \bar{x}$ and $u^\nu \to \bar{x}$ because the sets S^ν increase to S (the closure of their union), and $C \cap S = \{\bar{x}\}$.

Let $v^\nu = (u^\nu - x^\nu)/|u^\nu - x^\nu|$, so $|v^\nu| = 1$, $v^\nu \in N_C(x^\nu)$, $-v^\nu \in N_{S^\nu}(u^\nu)$. The sequence of vectors v^ν being bounded, it has a cluster point $\bar{v}$, $|\bar{v}| = 1$; without loss of generality (by passing to a subsequence if necessary) we can suppose for simplicity that $v^\nu \to \bar{v}$. Along with the fact that $v^\nu \in N_C(x^\nu)$ and $x^\nu \to \bar{x}$, this implies that $\bar{v} \in N_C(\bar{x})$. Because $-v^\nu \in N_{S^\nu}(x^\nu)$ and S^ν is convex, we also have $-v^\nu \cdot [u - u^\nu] \le 0$ for all $u \in S^\nu$. Since S^ν increases to S while $u^\nu \to \bar{x}$, we obtain in the limit that $-\bar{v} \cdot [u - \bar{x}] \le 0$ for all $u \in S$. Recalling the construction of S, we note that among the vectors $u \in S$ are all vectors of the form $\bar{x} + \varepsilon w$ with $w \in B$. Further, B is the closed ball of a certain radius $\delta > 0$ around $\bar{w}$, so its elements w have the form $\bar{w} + \delta z$ with $|z| \le 1$. Plugging these expressions into the limiting inequality that was obtained, we get $-\bar{v} \cdot \varepsilon[\bar{w} + \delta z] \le 0$ for all z with $|z| \le 1$. In particular we can take $z = -\bar{v}$ (since $|\bar{v}| = 1$) and see that $-\bar{v} \cdot \bar{w} + \delta |\bar{v}|^2 \le 0$. This reveals that $\bar{v} \cdot \bar{w} \ge \delta$, hence $\bar{v} \cdot \bar{w} > 0$ as desired. □

Polar cones: When two cones in $\mathbb{R}^n$, say T and N, are in this symmetric relationship, with $N = \{v \,|\, v \cdot w \le 0 \text{ for all } w \in T\}$ and $T = \{w \,|\, v \cdot w \le 0 \text{ for all } v \in N\}$, they are said to be *polar* to each other. Then, incidentally, they both must be convex. Polarity of cones generalizes orthogonality of linear subspaces.

Normals to boxes: If $X = I_1 \times \cdots \times I_n$ for closed intervals I_j in $\mathbb{R}$, then X is regular at any of its points $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)$ (by virtue of closedness and convexity), and

$$N_X(\bar{x}) = N_{I_1}(\bar{x}_1) \times \cdots \times N_{I_n}(\bar{x}_n), \text{ where}$$
$$N_{I_j}(\bar{x}_j) = \begin{cases} (-\infty, 0] & \text{if } \bar{x}_j \text{ is the left endpoint (only) of } I_j, \\ [0, \infty) & \text{if } \bar{x}_j \text{ is the right endpoint (only) of } I_j, \\ [0, 0] & \text{if } \bar{x}_j \text{ lies in the interior of } I_j, \\ (-\infty, \infty) & \text{if } I_j \text{ is a one-point interval, consisting just of } \bar{x}_j. \end{cases}$$

In other words, the condition $z \in N_X(\bar{x})$ for $z = (z_1, \ldots, z_n)$ constitutes a list of *sign restrictions* on the coordinates of z. Depending on the mode in which $\bar{x}_j$ fulfills the constraint $x_j \in I_j$, each z_j is designated as nonpositive, nonnegative, zero, or free.

Normals to the nonnegative orthant: As an important example, the normal vectors $z = (z_1, \ldots, z_n)$ to the box $\mathbb{R}^n_+$, at $\bar{x} = (\bar{x}_1, \ldots \bar{x}_n)$ are given by

$$\begin{aligned} z \in N_{\mathbb{R}^n_+}(\bar{x}) \quad &\Longleftrightarrow \quad \begin{cases} z_j \le 0 & \text{for all } j \text{ with } \bar{x}_j = 0, \\ z_j = 0 & \text{for all } j \text{ with } \bar{x}_j > 0 \end{cases} \\ &\Longleftrightarrow \quad z_j \le 0, \quad \bar{x}_j \ge 0, \quad z_j \cdot \bar{x}_j = 0 \quad \text{for all } j. \end{aligned}$$

THEOREM 9 (fundamental normal cone condition for optimality). *Consider the problem of minimizing a function f_0 of class $\mathcal{C}^1$ over a closed set $C \subset \mathbb{R}^n$. Let $\bar{x} \in C$.*

(a) (necessary). *If $\bar{x}$ is locally optimal, then*

$$-\nabla f_0(\bar{x}) \in N_C(\bar{x}).$$

(b) (sufficient). *If this relation holds with C and f_0 convex, then $\bar{x}$ is globally optimal.*

Proof. To prove (a), we need to recall the fundamental role of tangent cones in optimality, as explained early in Chapter 4: the local optimality of $\bar{x}$ implies that $\nabla f_0(\bar{x}) \cdot w \geq 0$ for every $w \in T_C(\bar{x})$. Then, by definition, $-\nabla f_0(\bar{x})$ is actually a *regular* normal to C at $\bar{x}$. Hence $-\nabla f_0(\bar{x}) \in N_C(\bar{x})$. To prove (b) we simply note that, in view of the description of normal cones to convex sets given above, this merely restates the sufficient condition of Theorem 7(b). □

Connection with variational inequalities: The condition $-\nabla f_0(\bar{x}) \in N_C(\bar{x})$ for *convex* C is equivalent to the "variational inequality" condition in Theorem 7, as seen from the above characterization of normal cones to convex sets.

Versatility of the fundamental normal cone condition: The condition in Theorem 9 has conceptual and technical advantages over a tangent cone condition like the one in Theorem 7. Such a tangent cone condition seemingly refers to an infinite family of inequalities, which is cumbersome to think about in contrast to thinking about how a cone $N_C(\bar{x})$ might look. The examples of normal cones already given, and more to come, will help in applying Theorem 9 to particular situations and will lead to rules involving Lagrange multipliers.

Fermat's rule as a special case: At any point $\bar{x}$ lying in the interior of C, one has $N_C(\bar{x}) = \{0\}$; there's no normal other than the zero vector. Therefore, if f_0 has a local minimum relative to C at such a point, it's necessary that $\nabla f_0(\bar{x}) = 0$.

Optimality relative to a box: In the case of minimizing a function f_0 over a box X, the condition $-\nabla f_0(\bar{x}) \in N_X(\bar{x})$ that comes out of Theorem 9 can be combined with the description of the cones $N_X(\bar{x})$ that was furnished above. It thereby translates into sign conditions on the components of $\nabla f_0(\bar{x})$, i.e., on the partial derivatives $(\partial f_0/\partial x_j)(\bar{x})$.

Normal cone condition in maximization: Although our convention is to reduce everything to minimization, it's worth noting that when a function f_0 is maximized over C instead of minimized, the necessary condition corresponding to the one in Theorem 7 comes out as $\nabla f_0(\bar{x}) \in N_C(\bar{x})$.

Determining normal vectors from constraint representations: To get more out of the optimality condition in Theorem 9, we need to develop formulas for $N_C(\bar{x})$ in terms of constraint representations of C.

Standard constraint structure: Our attention will be concentrated henceforth on the case where C is the set of feasible solutions to a problem $(\mathcal{P})$ in standard format:

$$x \in C \iff x \in X \text{ and } \begin{cases} f_i(x) \leq 0 & \text{for } i \in [1,s], \\ f_i(x) = 0 & \text{for } i \in [s+1,m]. \end{cases}$$

For short, we'll refer to this as the case where C is a *standard feasible set*.

Constraint qualifications: Success in obtaining a formula for normal cones $N_C(\bar{x})$ by way of a constraint representation of C typically depends on some assumption about the nature of the constraints, such as a gradient condition or something involving convexity or linearity. Such an assumption is called a "constraint qualification." When there are linear constraints *only*, it will turn out that this complication can be avoided.

Lagrange multipliers for standard constraints: Our analysis of normal vectors to a standard feasible set C will require us to consider certain expressions in which each of the constraint functions f_i is assigned a coefficient y_i, called its *Lagrange multiplier*. The "multiplier vectors" $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$ comprised of such coefficients will have a grand role. To get the stage ready for them, we now introduce (for reasons not yet explained but soon to come out) the *standard multiplier cone*

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

We also introduce, for each point $\bar{x} \in C$, a special subset of this multiplier cone which is tailored to the *inactive* inequality constraints at $\bar{x}$ (with $f_i(\bar{x}) < 0$), namely

$$Y(\bar{x}) = \big\{ y \in Y \,\big|\, y_i = 0 \text{ for } i \in [1,s] \text{ inactive at } \bar{x} \big\},$$

so that, in other words,

$$y \in Y(\bar{x}) \iff \begin{cases} y_i = 0 & \text{for } i \in [1,s] \text{ with } f_i(\bar{x}) < 0, \\ y_i \geq 0 & \text{for } i \in [1,s] \text{ with } f_i(\bar{x}) = 0, \\ y_i \text{ free} & \text{for } i \in [s+1,m]. \end{cases}$$

Complementary slackness: These conditions can equally well be written in another way, which is more symmetric:

$$y \in Y(\bar{x}) \iff \begin{cases} \text{for } i \in [1,s]: & y_i \geq 0,\ f_i(\bar{x}) \leq 0,\ y_i f_i(\bar{x}) = 0, \\ \text{for } i \in [s+1,m]: & y_i \text{ free},\ f_i(\bar{x}) = 0. \end{cases}$$

The rules for $i \in [1,s]$ are called *complementary slackness conditions* because they insist that for each pair of inequalities, $y_i \geq 0$, $f_i(\bar{x}) \leq 0$, if either one of them is slack (i.e., satisfied as a strict inequality), then the other one must be tight (i.e., satisfied as an equation).

Cone properties of the multiplier sets: Both Y and $Y(\bar{x})$ are closed, convex cones, moreover of the kind described only by "sign conditions," such as have been encountered in the study of tangent and normal cones to boxes. This isn't just a coincidence. It's easy to see from our formulas for normal cones to boxes that $\bar{u} = (\bar{u}_1, \ldots, \bar{u}_m)$ where $\bar{u}_i = f_i(\bar{x})$, we actually have

$$Y(\bar{x}) = N_K(\bar{u}) \quad \text{for } \bar{u}_i = f_i(\bar{x}) \text{ and the box } K \text{ defined by}$$
$$K = J_1 \times \cdots \times J_m \text{ with } J_i = \begin{cases} (-\infty, 0] & \text{for } i \in [1,s], \\ [0,0] & \text{for } i \in [s+1,m], \end{cases}$$

Also, there's the interesting—and later very useful—observation, through the complementary slackness formulation of the condition for having $y \in Y(\bar{x})$, that

$$y \in Y(\bar{x}) \iff (f_1(\bar{x}), \ldots, f_m(\bar{x})) \in N_Y(y).$$

THEOREM 10 (normals and tangents to standard feasible sets). *Let C be standard feasible set coming from $\mathcal{C}^1$ functions $f_1, \ldots, f_m$ and a closed set $X \subset \mathbb{R}^n$. Let $\bar{x}$ be a point of C (and hence of X) such that X is regular at $\bar{x}$ (as holds for instance when X is convex, and in particular when it is a box). Suppose that the following assumption, to be called the standard constraint qualification at $\bar{x}$, is satisfied:*

$$(\star) \begin{cases} \text{there is no vector } y = (y_1, \ldots, y_m) \in Y(\bar{x}) \text{ satisfying} \\ -\big[y_1 \nabla f_1(\bar{x}) + \cdots + y_m \nabla f_m(\bar{x})\big] \in N_X(\bar{x}), \text{ except } y = 0. \end{cases}$$

Then C is regular at $\bar{x}$, and the normal cone at $\bar{x}$ has the formula

$$v \in N_C(\bar{x}) \iff \begin{cases} v = y_1 \nabla f_1(\bar{x}) + \cdots + y_m \nabla f_m(\bar{x}) + z \text{ for some} \\ (y_1, \ldots, y_m) = y \in Y(\bar{x}) \text{ and } z \in N_X(\bar{x}), \end{cases}$$

whereas the tangent cone at $\bar{x}$ has the formula

$$w \in T_C(\bar{x}) \iff w \in T_X(\bar{x}) \text{ with } \begin{cases} \nabla f_i(\bar{x}) \cdot w \text{ free} & \text{for } i \in [1,s] \text{ with } f_i(\bar{x}) < 0, \\ \nabla f_i(\bar{x}) \cdot w \leq 0 & \text{for } i \in [1,s] \text{ with } f_i(\bar{x}) = 0, \\ \nabla f_i(\bar{x}) \cdot w = 0 & \text{for } i \in [s+1,m]. \end{cases}$$

Note: When $X = \mathbb{R}^n$, or for that matter whenever $\bar{x}$ lies in the interior of X, the normal cone $N_X(\bar{x})$ consists of just 0, so $(\star)$ revolves around $y_1 \nabla f_1(\bar{x}) + \cdots + y_m \nabla f_m(\bar{x}) = 0$, and the z term in the formula for $v \in N_C(\bar{x})$ drops out.

Proof. This level of result, in the framework of normal cones, tangent cones and regularity, can't yet be found in textbooks on optimization. The argument, although lengthy, is elementary in that it only uses basic facts about sequences and continuity. It has big advantages over traditional treatments of Lagrange multipliers, because it readily extends to sets C with other, "nonstandard" constraint representations.

Let $K = (-\infty, 0]^s \times [0,0]^{m-s}$, this being the special box in $\mathbb{R}^m$ that was introduced just ahead of the theorem in the normal cone interpretation of the multiplier set $Y(\bar{x})$. In terms of this set K and $F(x) = (f_1(x), \ldots, f_m(x))$, we have

$$C = \big\{x \in X \,\big|\, F(x) \in K\big\}, \qquad Y(\bar{x}) = N_K\big(F(\bar{x})\big).$$

That mode of expression will help us by tying the behavior of multiplier vectors y into that of normal vectors to K. In this setting, we will be able to utilize the fact that K is closed and convex, hence regular at any of its points, such as $F(\bar{x})$.

The $F(x)$ notation will benefit us further by making it possible to write the gradient sums in the theorem in the Jacobian form $\nabla F(\bar{x})^* y$. To appreciate this, recall that $\nabla F(\bar{x})$ is the $m \times n$ matrix having the gradients $\nabla f_i(\bar{x})$ as its *rows*; the transpose matrix $\nabla F(\bar{x})^*$ has them as its *columns*, so that

$$\nabla F(\bar{x})^* y \;=\; y_1 \nabla f_1(\bar{x}) + \cdots + y_m \nabla f_m(\bar{x}).$$

Note also that for any $y = (y_1, \ldots, y_m)$ and $w = (w_1, \ldots, w_n)$ we have

$$y \cdot \nabla F(\bar{x}) w \;=\; \sum_{i=1,j=1}^{m,n} y_i \frac{\partial f_i}{\partial x_j}(\bar{x}) w_j \;=\; [\nabla F(\bar{x})^* y] \cdot w.$$

Part 1. We'll demonstrate first that the implication "$\Rightarrow$" is valid in the tangent cone formula. A vector $w \in T_C(\bar{x})$ is by definition the limit of a sequence of vectors $w^\nu = (x^\nu - \bar{x})/\tau^\nu$ formed with $x^\nu \in C$, $x^\nu \to \bar{x}$ and $\tau^\nu \searrow 0$. Then, since $\bar{x} \in X$ and $x^\nu \in X$ in particular, we have $w \in T_X(\bar{x})$, as required on the right side of the claimed formula.

Since also $F(\bar{x}) \in K$ and $F(x^\nu) \in K$, and F is $\mathcal{C}^1$, we furthermore have $F(x^\nu) \to F(\bar{x})$ with $F(x^\nu) = F(\bar{x}) + \nabla F(\bar{x})(x^\nu - \bar{x}) + o(|x^\nu - \bar{x}|)$, hence

$$\frac{F(x^\nu) - F(\bar{x})}{\tau^\nu} \;=\; \nabla F(\bar{x}) w^\nu + \frac{o\big(|x^\nu - \bar{x}|\big)}{|x^\nu - \bar{x}|} |w^\nu| \;\to\; \nabla F(\bar{x}) w.$$

By thinking of this in terms of $\bar{u} = F(\bar{x})$ and $u^\nu = F(x^\nu)$ we can interpret it as referring to a situation where we have $\bar{u} \in K$, $u^\nu \in K$, $u^\nu \to \bar{u}$, and $\tau^\nu \searrow 0$ such that vectors

$h^\nu = [u^\nu - \bar{u}]/\tau^\nu$ converge to $\bar{h} = \nabla F(\bar{x})w$. That, of course, says that $\nabla F(\bar{x})w \in T_K(\bar{u})$. Because K is the product of s copies of $(-\infty, 0]$ and $m - s$ copies of $(-\infty, \infty)$, we know from our earlier analysis of tangent cones to boxes that

$$(h_1, \ldots, h_m) \in T_K(\bar{u}) \iff \begin{cases} h_i \text{ free} & \text{for } i \in [1, s] \text{ with } \bar{u}_i < 0, \\ h_i \geq 0 & \text{for } i \in [1, s] \text{ with } \bar{u}_i = 0, \\ h_i = 0 & \text{for } i \in [s+1, m]. \end{cases}$$

Applying this to $\bar{h} = \nabla F(\bar{x})w$, which has components $\bar{h}_i = \nabla f_i(\bar{x}) \cdot w$, and recalling that $\bar{u}$ has components $\bar{u}_i = f_i(\bar{x})$, we see that the sign conditions on the right side of the tangent cone formula must hold.

Part 2. We'll demonstrate next that the implication "$\Leftarrow$" is valid in the normal cone formula, and moreover that every vector v having a representation of the indicated kind must actually be a *regular* normal vector to C at $\bar{x}$. This means verifying that $v \cdot w \leq 0$ for every $w \in T_C(\bar{x})$. In the notation we've introduced, the indicated representation of v can be written as $v = \nabla F(\bar{x})^* y + z$ for some $y \in N_K\big(F(\bar{x})\big)$ and $z \in N_X(\bar{x})$, so that, in view of $[\nabla F(\bar{x})^* y] \cdot w$ being the same as $y \cdot \nabla F(\bar{x})w$, we have

$$v \cdot w \;=\; y \cdot \nabla F(\bar{x})w + z \cdot w.$$

Here y is known to be a regular normal to K at $\bar{u} = F(\bar{u})$, while z is a regular normal to X at $\bar{x}$ by our regularity assumption on X. On the other hand, every $w \in T_C(\bar{x})$ satisfies $\nabla F(\bar{x})w \in T_K(\bar{u})$ and $w \in T_X(\bar{x})$, as seen in Part 1. Then, from the regularity of y and z, we have $y \cdot \nabla F(\bar{x})w \leq 0$ and $z \cdot w \leq 0$, hence $v \cdot w \leq 0$, as needed.

Part 3. The next task on our agenda is showing that "$\Leftarrow$" is valid in the normal cone formula. Then it will also follow, from what we saw in Part 2, that every vector $v \in N_C(\bar{x})$ is actually a regular normal, so that C is regular at $\bar{x}$, as claimed. The argument will pass, however, through an intermediate stage involving vector representations that are only "approximately" of the form in question. In this approximation stage, we focus temporarily on a vector $v \in N_C(\bar{x})$ which we *assume* to be a regular normal, and in fixing an arbitrary $\varepsilon > 0$, we demonstrate the existence of

$$\left.\begin{aligned} x \in I\!R^n \text{ with } |x - \bar{x}| < \varepsilon \\ u \in K \text{ with } \big|u - F(x)\big| < \varepsilon \\ w \in I\!R^n \text{ with } |w| < \varepsilon \\ y \in N_K(u),\ z \in N_X(x) \end{aligned}\right\} \text{ such that } v = \nabla F(x)^* y + z + w.$$

The demonstration won't yet use the constraint qualification ($\star$) or the assumption about regularity. We take any sequence $\tau^\nu \searrow 0$ and define the functions φ^ν on $I\!R^n \times I\!R^m$ by

$$\varphi^\nu(x, u) = \frac{1}{2\tau^\nu}\big|x - (\bar{x} + \tau^\nu v)\big|^2 + \frac{1}{2\tau^\nu}\big|F(x) - u\big|^2 \;\geq 0.$$

These functions are continuously differentiable in x and u. Our strategy will be based on analyzing the problem of minimizing φ^ν over the closed set $X \times K$ in $\mathbb{R}^n \times \mathbb{R}^m$.

Part 3a. We have to know, in order to get started on this track, that for each ν an optimal solution (x^ν, u^ν) to the auxiliary minimization problem does exist. We'll confirm it by establishing that the level set $S = \big\{(x,u) \in X \times K \,\big|\, \varphi^\nu(x,u) \le \alpha\big\}$ is bounded for any $\alpha \ge 0$. Then the problem of minimizing φ^ν over $X \times K$ is well posed, and the existence of an optimal solution is assured by Theorem 1.

Clearly, any point $(x,u) \in S$ has both $|x-(\bar{x}+\tau^\nu v)|^2 \le 2\tau^\nu\alpha$ and $|u-F(x)|^2 \le 2\tau^\nu\alpha$. Then $|x| \le \lambda := |\bar{x}|+\tau^\nu|v|+\sqrt{2\tau^\nu\alpha}$. Over the closed ball $\big\{x \,\big|\, |x| \le \lambda\big\}$ there is a maximum to the possible values of $\big|F(x)\big|$ (an expression that is continuous in x); say $\big|F(x)\big| \le \sigma$ for all such x. The inequality $|u - F(x)|^2 \le 2\tau^\nu\alpha$ then yields $|u| \le \mu := \sigma + \sqrt{2\tau^\nu\alpha}$. Since every element $(x,u) \in S$ has $|x| \le \lambda$ and $|u| \le \mu$, the level set S is indeed bounded.

Part 3b. We now have license to denote by (x^ν, u^ν) for each ν an optimal solution (not necessarily unique) to the problem of minimizing the function φ^ν over $X \times K$; we denote the optimal value in this problem by α^ν. Obviously

$$0 \le \alpha^\nu = \varphi^\nu(x^\nu, u^\nu) \le \varphi^\nu\big(\bar{x}, F(\bar{x})\big) = \frac{\tau^\nu}{2}|v|^2 \to 0.$$

The inequalities deduced in our investigation of level sets of φ^ν tell us at the same time that neither $|x^\nu - (\bar{x}+\tau^\nu v)|^2$ nor $|u^\nu - F(x^\nu)|^2$ can exceed $2\tau^\nu\alpha^\nu$. Since $\alpha^\nu \le (\tau^\nu/2)|v|^2$, as just seen, we must have $\big|x^\nu - (\bar{x}+\tau^\nu v)\big| \le \tau^\nu|v|$ and $\big|u^\nu - F(x^\nu)\big| \le \tau^\nu|v|$. Therefore, $x^\nu \to \bar{x}$ and $\big|u^\nu - F(x^\nu)\big| \to 0$. In addition the sequence of vectors $w^\nu = (x^\nu - \bar{x})/\tau^\nu$ is bounded because the inequality $\big|x^\nu - (\bar{x}+\tau^\nu v)\big| \le \tau^\nu|v|$, when divided by τ^ν, gives $|w^\nu - v| \le |v|$. This sequence therefore has a cluster point w.

Any such cluster point w belongs by definition to the tangent cone $T_C(\bar{x})$, so w satisfies $v \cdot w \le 0$, because, in this stage, v has been assumed to be a regular normal at $\bar{x}$. But at the same time w satisfies $|w - v| \le |v|$. In squaring the latter we see that $|w|^2 - 2v\cdot w + |v|^2 \le |v|^2$, which implies $|w|^2 \le 2v\cdot w \le 0$. Hence actually $w = 0$. Thus, the only possible cluster point of the sequence of vectors w^ν is 0, and we conclude that $w^\nu \to 0$. Eventually then, once ν is large enough, we'll have

$$|x^\nu - \bar{x}| < \varepsilon, \qquad |u^\nu - F(x^\nu)| < \varepsilon, \qquad |w^\nu| < \varepsilon.$$

Next we use the fact that (x^ν, u^ν) minimizes φ^ν over $X \times K$. In particular the minimum of $\varphi^\nu(x^\nu, u)$ over $u \in K$ is attained at u^ν, whereas the minimum of $\varphi^\nu(x, u^\nu)$ over $x \in X$ is attained at x^ν. From part (a) of Theorem 9, therefore, we have

$$-\nabla_u\varphi^\nu(x^\nu, u^\nu) \in N_K(u^\nu), \qquad -\nabla_x\varphi^\nu(x^\nu, u^\nu) \in N_X(x^\nu),$$

these being *regular* normal vectors. Let $y^\nu = -\nabla_u \varphi^\nu(x^\nu, u^\nu)$ and $z^\nu = -\nabla_x \varphi^\nu(x^\nu, u^\nu)$, so $y^\nu \in N_K(u^\nu)$ and $z^\nu \in N_X(x^\nu)$. In calculating these vectors from the formula for φ^ν by differentiating separately in u and then in x, we get $y^\nu = [F(x^\nu) - u^\nu]/\tau^\nu$ and $z^\nu = -w^\nu + v - \nabla F(x^\nu)^* y^\nu$. Thus, $v = \nabla F(x^\nu)^* y^\nu + z^\nu + w^\nu$. It follows that when ν is taken sufficiently large the elements $x = x^\nu$, $u = u^\nu$, $y = y^\nu$ and $w = w^\nu$ furnish the kind of approximate representation of v that was required.

Part 3c. Now we are ready for the last stage of demonstrating that "$\Rightarrow$" holds in the normal cone formula. We consider a general vector $v \in N_C(\bar{x})$ and aim at proving the existence of $y \in N_K(F(\bar{x}))$ and $z \in N_X(\bar{x})$ such that $v = \nabla F(\bar{x})^* y + z$.

Fix any sequence of values $\varepsilon^\nu \searrow 0$. From the definition of normal vectors in general sense, we know there exist sequences $\bar{x}^\nu \to \bar{x}$ in C and $v^\nu \to v$ with v^ν a regular normal to C at $\bar{x}^\nu$. Then, on the basis of the intermediate fact about "approximation" that has just been established, there exist for each ν

$$\left.\begin{aligned} x^\nu \in I\!R^n \text{ with } |x^\nu - \bar{x}^\nu| < \varepsilon^\nu \\ u^\nu \in K \text{ with } \big|u^\nu - F(\bar{x}^\nu)\big| < \varepsilon^\nu \\ w^\nu \in I\!R^n \text{ with } |w^\nu| < \varepsilon^\nu \\ y^\nu \in N_K(u^\nu),\ z^\nu \in N_X(x^\nu) \end{aligned}\right\} \text{ such that } v^\nu = \nabla F(x^\nu)^* y^\nu + z^\nu + w^\nu.$$

There are two cases to distinguish: either (1) the sequence of vectors y^ν has a cluster point y, or (2) it has no bounded subsequences at all, meaning that $|y^\nu| \to \infty$.

In the case (1) of a cluster point y, we have in the limit that $y \in N_K\big(F(\bar{x})\big)$, inasmuch as $F(\bar{x}^\nu) \to F(\bar{x})$ and $\nabla F(\bar{x}^\nu) \to \nabla F(\bar{x})$ by the continuity of F and its first partial derivatives. The corresponding subsequence of the z^ν's is then bounded and must have a cluster point too, say z. We have $z \in N_X(\bar{x})$ and consequently from taking limits in the equation $v^\nu = \nabla F(x^\nu)^* y^\nu + z^\nu + w^\nu$ that $v = \nabla F(\bar{x})^* y + z$, as needed.

In the case (2) of $|y^\nu| \to \infty$, the vectors $\bar{y}^\nu = y^\nu/|y^\nu|$ and $\bar{z}^\nu = z^\nu/|y^\nu|$, which like y^ν and z^ν belong to $N_K\big(F(x^\nu)\big)$ and $N_X(x^\nu)$, have $|\bar{y}^\nu| = 1$ and $|\bar{z}^\nu| \le |\nabla F(x^\nu)\bar{y}^\nu| + \delta^\nu$, where $\delta^\nu := |v^\nu - w^\nu|/|y^\nu| \to 0$ and $\nabla F(x^\nu) \to \nabla F(\bar{x})$. Therefore, the sequence of pairs $(\bar{y}^\nu, \bar{z}^\nu)$ is bounded has a cluster point $(\bar{y}, \bar{z})$ with $|\bar{y}| = 1$. Again we get $\bar{y} \in N_K\big(F(\bar{x})\big)$ and $\bar{z} \in N_X(\bar{x})$. In dividing the equation $v^\nu = \nabla F(x^\nu)^* y^\nu + z^\nu + w^\nu$ by $|y^\nu|$ and taking the limit as $\nu \to \infty$, we see that $\nabla F(\bar{x})^* \bar{y} + \bar{z} = 0$. But $\bar{y} \neq 0$, so this is impossible under the constraint qualification ($\star$). Thus, only the cluster point case (1) is viable.

Part 4. Our final task is to confirm that "$\Leftarrow$" holds in the tangent cone formula. Let w have the properties on the right side of the formula. In Part 1 we saw this meant having $\nabla F(\bar{x})w \in T_K(F(\bar{x}))$, along with $w \in T_X(\bar{x})$. Through the regularity of C at

$\bar{x}$ (established in Part 3), $T_C(\bar{x})$ consists of the vectors w satisfying $v \cdot w \leq 0$ for every $v \in N_C(\bar{x})$; we know those vectors have the form $\nabla F(\bar{x})^* y + z$ with $y \in N_K(F(\bar{x}))$ and $z \in N_X(\bar{x})$. Then $v \cdot w = y \cdot \nabla F(\bar{x}) w + z \cdot w$, as seen in Part 2. Is it true that this expression is ≤ 0 for $y \in N_K(F(\bar{x}))$ and $z \in N_X(\bar{x})$ when $\nabla F(\bar{x}) w \in T_K(F(\bar{x}))$ and $w \in T_X(\bar{x})$? Yes, by the regularity of K at $F(\bar{x})$ and X at $\bar{x}$. □

Classical example of a set defined by equations only: Suppose that

$$C = \big\{ x \in I\!R^n \;\big|\; f_i(x) = 0 \text{ for } i = 1, \ldots, m \big\}$$

for $\mathcal{C}^1$ functions f_i. What does Theorem 10 say then at a point $\bar{x} \in C$? Here $X = I\!R^n$, so $N_X(\bar{x}) = \{0\}$, while $Y(\bar{x}) = Y = I\!R^m$. The standard constraint qualification $(\star)$ in the assumptions of Theorem 10 asks in these circumstances that there be

$$\begin{cases} \text{no coefficient vector } (y_1, \ldots, y_m) \in I\!R^m \text{ has the property that} \\ y_1 \nabla f_1(\bar{x}) + \cdots + y_m \nabla f_m(\bar{x}) = 0, \text{ except } (y_1, \ldots, y_m) = (0, \ldots, 0). \end{cases}$$

In other words, it demands the *linear independence* of the gradient vectors $\nabla f_i(\bar{x})$, $i = 1, \ldots, m$, or equivalently, requires the $m \times n$ Jacobian matrix $\nabla F(\bar{x})$, with these vectors as its rows, to have rank m. The normal cone formula we obtain from Theorem 10 under this assumption, comes out as

$$v \in N_C(\bar{x}) \iff \begin{cases} v = y_1 \nabla f_1(\bar{x}) + \cdots + y_m \nabla f_m(\bar{x}) \\ \text{for some } (y_1, \ldots, y_m) \in I\!R^m, \end{cases}$$

which says that $N_C(\bar{x})$ is *the subspace of $I\!R^n$ spanned by* all the vectors $\nabla f_i(\bar{x})$. The corresponding tangent cone formula we obtain from Theorem 10 is

$$w \in T_C(\bar{x}) \iff \nabla f_i(\bar{x}) \cdot w = 0 \text{ for } i = 1, \ldots, m,$$

which says that $T_C(\bar{x})$ is *the subspace of $I\!R^n$ orthogonal to* all the vectors $\nabla f_i(\bar{x})$.

Smooth manifold interpretation: Sets defined by systems of one or more equations fit a long-standing tradition in calculus and differential geometry. In $I\!R^2$, a single equation can be imagined as specifying a curve, whereas in $I\!R^3$ a single equation would give a surface and two equations would be needed to get a curve. Such sets and their higher-dimensional analogs are called *smooth manifolds*, and we can think of them as the models for what we hope to get as C, when C is a standard feasible set coming from equation constraints only.

Sobering realities must be remembered, however. In $\mathbb{R}^3$, for instance, two equations don't necessarily determine a "curve," even if each, by itself, determines a "surface." The surfaces might just touch in a point, and if flat portions of them come up against each other, their intersection could be something weirder. We can't expect to get a smooth manifold without imposing an assumption on the equations, and moreover focusing just on a neighborhood of a point $\bar{x}$.

The classical assumption for ensuring that C is a smooth manifold around one of its points $\bar{x}$ is precisely the linear independence condition, or rank condition, encountered here. As we've seen, it results in the normal cone $N_C(\bar{x})$ and tangent cone $T_C(\bar{x})$ being *linear subspaces* that are *orthogonal* to each other and moreover complementary in dimension (the normal subspace having dimension m and the tangent subspace having dimension $n - m$).

The role of the standard constraint qualification: Even without any recourse to condition $(\star)$, the regularity assumption in Theorem 10 guarantees that $N_C(\bar{x})$ *includes* all the vectors v expressible as a sum in the manner prescribed. This was demonstrated the first part of the proof of Theorem 10. Condition $(\star)$ was essential however in confirming, in the much harder second part of the proof, that the inclusion is really an equation. There, $(\star)$ entered at the very end.

Importance for handling equality constraints: The example of smooth manifolds, just described, where $(\star)$ came down to a classical assumption, makes clear that even when working only with equations, a constraint qualification may be needed.

Importance for handling inequality constraints: An example of a shortfall in the normal cone formula in a case of inequality constraints will underscore the need for $(\star)$ more generally. Let $C = \{x \mid f_1(x) \leq 0,\ f_2(x) \leq 0\}$ for $x = (x_1, x_2) \in \mathbb{R}^2$ and the functions $f_1(x) = (x_1 - 1)^2 + (x_2 - 1)^2 - 2$ and $f_2(x) = (x_1 + 1)^2 + (x_2 + 1)^2 - 2$. Here C is the intersection of two disks, both of radius $\sqrt{2}$ but centered at $(1, 1)$ and $(-1, -1)$ respectively. The disks touch only at the origin, so C consists of just $\bar{x} = (0, 0)$. Then $N_C(\bar{x}) = \mathbb{R}^2$. But $Y(\bar{x}) = \mathbb{R}^2_+$, $\nabla f_1(\bar{x}) = (-1, -1)$ and $\nabla f_2(\bar{x}) = (1, 1)$ so the vectors of the form $v = y_1 \nabla f_1(\bar{x}) + y_2 \nabla f_2(\bar{x})$ with $(y_1, y_2) \in Y(\bar{x})$ are collinear and don't fill up $\mathbb{R}^2$.

Exceptional status of linear constraints: For a standard feasible set C coming from *affine* functions f_i and a *box* X, the standard constraint qualification $(\star)$ isn't needed. The formulas we've developed for $N_C(\bar{x})$ and $T_C(\bar{x})$ are valid then without it, as will be established in Theorem 12.

THEOREM 11 (first-order optimality for problems in standard format). *Let $\bar{x}$ be a feasible solution to problem $(\mathcal{P})$, with X closed, regular at $\bar{x}$, and every f_i of class $\mathcal{C}^1$.*

(a) (necessary). *If $\bar{x}$ is locally optimal and the standard constraint qualification $(\star)$ is fulfilled at $\bar{x}$, then there must be a multiplier vector $\bar{y} = (\bar{y}_1, \ldots, \bar{y}_m)$ such that*

$$-\big[\nabla f_0(\bar{x}) + \bar{y}_1 \nabla f_1(\bar{x}) + \cdots + \bar{y}_m \nabla f_m(\bar{x})\big] \in N_X(\bar{x}) \quad \textit{with } \bar{y} \in Y(\bar{x}).$$

(b) (sufficient). *If such a multiplier vector $\bar{y}$ exists, and f_0 and C are convex (as in the case of convex programming), then $\bar{x}$ is globally optimal.*

Proof. This combines Theorem 9 with Theorem 10. Saying that $-\nabla f_0(\bar{x})$ can be represented by $-\nabla f_0(\bar{x}) = \bar{y}_1 \nabla f_1(\bar{x}) + \cdots + \bar{y}_m \nabla f_m(\bar{x}) + \bar{z}$ for some $\bar{y} \in Y(\bar{x})$ and $\bar{z} \in N_X(\bar{x})$ is the same as saying that $-\big[\nabla f_0(\bar{x}) + \bar{y}_1 \nabla f_1(\bar{x}) + \cdots + \bar{y}_m \nabla f_m(\bar{x})\big]$ belongs to $N_X(\bar{x})$ for some $\bar{y} \in Y(\bar{x})$. The reason the constraint qualification $(\star)$ isn't needed in (b) is that a representation of this kind for $-\nabla f_0(\bar{x})$ guarantees, even without it, that $-\nabla f_0(\bar{x}) \in N_C(\bar{x})$. (As explained earlier, the role of $(\star)$ was only to make sure that the vectors with such a representation fill up all of $N_C(\bar{x})$; they always do lie within $N_C(\bar{x})$.) □

Common simplification: When $\bar{x}$ lies in the interior of X, so that $N_X(\bar{x}) = \{0\}$, the gradient condition in Theorem 11 becomes $\nabla f_0(\bar{x}) + \bar{y}_1 \nabla f_1(\bar{x}) + \cdots + \bar{y}_m \nabla f_m(\bar{x}) = 0$.

Problems with an abstract constraint only: If there are no constraints in function form, so $(\mathcal{P})$ consists just of minimizing $f_0(x)$ over all $x \in X$ (this set being identical then with C), the optimality condition furnished by Theorem 11 reduces in essence to the one in Theorem 9. At any locally optimal solution $\bar{x}$ we must have $-\nabla f_0(\bar{x}) \in N_X(\bar{x})$, and this is sufficient for global optimality when f_0 and X are convex.

Abstract constraints without convexity: Instead of bothering with a regularity assumption in Theorem 11, we could just have taken X to be a box, or at least a closed, *convex* set, since such sets are regular at all of their points. That would have covered most direct applications. A plus for Theorem 11 in its broader formulation, however, is that it allows X to be specified by an additional system of possibly nonconvex constraints which we don't need to pin down immediately.

Kuhn-Tucker conditions: For a problem $(\mathcal{P})$ in standard format with $\mathcal{C}^1$ functions f_i, a vector pair $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ is said to satisfy the *Kuhn-Tucker conditions* for $(\mathcal{P})$ (and be a *Kuhn-Tucker pair*) if $\bar{x}$ is a feasible solution to $(\mathcal{P})$ and $\bar{y}$ is an associated Lagrange multiplier vector as in Theorem 11, i.e.,

$$-\big[\nabla f_0(\bar{x}) + \bar{y}_1 \nabla f_1(\bar{x}) + \cdots + \bar{y}_m \nabla f_m(\bar{x})\big] \in N_X(\bar{x}) \quad \text{with } \bar{y} \in Y(\bar{x}).$$

To summarize in this language, the Kuhn-Tucker conditions are the first-order optimality conditions for such problems $(\mathcal{P})$. Under the standard constraint qualification $(\star)$, the local optimality of $\bar{x}$ (along with regularity of X at $\bar{x}$) implies the existence of a $\bar{y}$ such that $(\bar{x}, \bar{y})$ is a Kuhn-Tucker pair. On the other hand, the global optimality of $\bar{x}$ is assured when $(\bar{x}, \bar{y})$ is a Kuhn-Tucker pair and $(\mathcal{P})$ is a problem of convex programming, even without $(\star)$.

Lagrangian function: An elegant expression of the Kuhn-Tucker conditions as a whole can be achieved in terms of the *Lagrangian* for problem $(\mathcal{P})$, which is the function

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

where $Y = \mathbb{R}^s_+ \times \mathbb{R}^{m-s}$. Then $\nabla_x L(x, y) = \nabla f_0(x) + y_1 \nabla f_1(x) + \cdots + y_m \nabla f_m(x)$, whereas $\nabla_y L(x, y) = \big(f_1(x), \ldots, f_m(x)\big)$.

Lagrangian form of the Kuhn-Tucker conditions: A pair of vectors $\bar{x}$ and $\bar{y}$ satisfies the Kuhn-Tucker conditions for $(\mathcal{P})$ if and only if $\bar{x} \in X$, $\bar{y} \in Y$, and

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

This is immediate from the observations just made about L and the alternative expression for $Y(\bar{x})$ furnished just before the statement of Theorem 10.

Mathematical history: Louis Lagrange, one of the greatest mathematicians of the 18th century, only considered minimization subject to equality constraints, i.e., in our framework the case of $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$. He found that a locally optimal solution $\bar{x}$ would have to satisfy, along with $f_i(\bar{x}) = 0$ for $i = 1, \ldots, m$, the condition $\nabla f_0(\bar{x}) + \bar{y}_1 \nabla f_1(\bar{x}) + \cdots + \bar{y}_m \nabla f_m(\bar{x}) = 0$. He realized that in terms of the function L these requirements could be written as $\nabla L(\bar{x}, \bar{y}) = 0$, a vector equation comprised of $n+m$ scalar equations in $n+m$ unknowns $\bar{x}_j$ and $\bar{y}_i$! These are the Kuhn-Tucker conditions when $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$.

Like many mathematicians of his day, Lagrange operated on a heuristic level, and in this case he even fell into the trap of claiming the equation $\nabla L(\bar{x}, \bar{y}) = 0$ meant that L had its minimum at $(\bar{x}, \bar{y})$. Inequality constraints didn't get serious attention until Kuhn and Tucker, around 1950. They got as far as covering abstract constraints $x \in X$ with $X = \mathbb{R}^r \times \mathbb{R}^{n-r}$, thus allowing for a distinction between variables x_j required to be nonnegative for $j = 1, \ldots, r$ but free for $j = r+1, \ldots, n$. The treatment of more general X had to wait however for the theory of normal cones in the '70's and '80's.

Taking advantage of linear constraints: When a standard feasible set C is specified through linear constraints, at least in part, there are important simplifications in the analysis of normals and tangents, because an assumption as strong as the standard constraint qualification $(\star)$ isn't really required. The formulas in Theorem 10 remain valid under alternative assumptions tailored to special features of linearity. Those assumptions help also in applying the necessary condition in Theorem 11.

THEOREM 12 (constraint qualification refinements). *Let C be a standard feasible set coming from constraint functions f_i of class $\mathcal{C}^1$ and an underlying set X that is a* **box**.

(a) *When some of the functions f_i are affine, the following milder assumption, to be called the refined constraint qualification, can be substituted in Theorems 10 and 11 for the standard constraint qualification $(\star)$:*

$$(\star\star)\begin{cases} \text{there is no vector } y = (y_1, \ldots, y_m) \in Y(\bar{x}) \text{ satisfying} \\ -\big[y_1 \nabla f_1(\bar{x}) + \cdots + y_m \nabla f_m(\bar{x})\big] \in N_X(\bar{x}) \text{ except} \\ \text{with } y_i = 0 \text{ for every } i \text{ such that } f_i \text{ is } \textbf{not affine.} \end{cases}$$

Thus, when C is specified by a linear constraint system, where every f_i is affine, no substitute for $(\star)$ is needed at all, and the constraint qualification can be dropped entirely.

(b) *When C is specified by a convex constraint system, where f_i is convex for $i \in [1, s]$ and affine for $i \in [s+1, m]$, the following assumption of a different character, to be called the refined Slater condition, can be substituted in Theorems 10 and 11 for the standard constraint qualification $(\star)$:*

$$(\star\star\star)\begin{cases} \text{there is a point } \tilde{x} \in C \text{ having } f_i(\tilde{x}) < 0 \\ \text{for every } i \in [1, s] \text{ such that } f_i \text{ is } \textbf{not affine.} \end{cases}$$

Proof. The assertion in (a) will be established in two stages, starting with the case where every f_i is affine and then going on to the case where only some of the functions f_i may be affine. Then (b) will be derived as a consequence of (a). In each instance, we are trying to show that the indicated refinement of $(\star)$ is adequate for the normal cone formula in Theorem 10 to be correct, since the other claims in Theorem 10 all follow from that formula, as does the application made of it in Theorem 11.

Part 1. Suppose that every f_i is affine: $f_i(x) = a_i \cdot x - b_i$ for certain vectors $a_i \in \mathbb{R}^n$ and scalars $b_i \in \mathbb{R}$, so that $\nabla f_i(x) = a_i$ for all x. For the purpose at hand, only the active inequality constraints at a point $\bar{x} \in C$ really matter. The inactive inequality constraints have no effect on the representation of normal vectors, since the multipliers y_i for their

gradients have to be 0. Therefore, with out loss of generality in our argument, we can just assume that there are *no* inactive constraints. This is aimed at notational simplification; it makes $a_i \cdot \bar{x} = b_i$ for all i. In fact we can simplify still further, however. Through the change of variables $x' = x - \bar{x}$, we can reduce from $a_i \cdot x = b_i$ to $a_i \cdot x' = 0$. In this way, we can pass without loss of generality (by reverting notation from x' back to x) to the case where $b_i = 0$, so that $f_i(x) = a_i \cdot x$ for all i, and our focus is the point $\bar{x} = 0$ of

$$C = \big\{x \,\big|\, a_i \cdot x \leq 0 \text{ for } i \in [1, s],\ a_i \cdot x = 0 \text{ for } i \in [s+1, m]\big\}.$$

In this simplified setting, we have $Y(\bar{x}) = Y = I\!R^s_+ \times I\!R^{m-s}$. Our goal is to verify that

$$v \in N_C(\bar{x}) \iff v = y_1 a_1 + \cdots + y_m a_m \text{ for some } y \in Y.$$

We already know, though, that "$\Leftarrow$" is valid without any constraint qualification (as shown in Part 2 of the proof of Theorem 10). Thus, only "$\Rightarrow$" really requires our attention. Our tactic will be to demonstrate that C can equally well be expressed in the form

$$C = \big\{x \,\big|\, a_i \cdot x \leq 0 \text{ for } i \in I_1,\ a_i \cdot x = 0 \text{ for } i \in I_2\big\}$$

for certain index sets $I_1 \subset [1, s]$ and $I_2 \subset [1, m] \backslash I_1$ having the property that this smaller constraint representation (with the indices renumbered if one prefers) does satisfy the corresponding instance of the standard constraint qualification ($\star$) (with $X = I\!R^n$, $N_X(\bar{x}) = \{0\}$). That will guarantee through Theorem 10 that any $v \in N_C(\bar{x})$ can be expressed as a sum of terms $y_i a_i$ for $i \in I_1 \cup I_2$, with $y_i \geq 0$ when $i \in I_1$. We'll build then on that to get the kind of representation we need for v in order to conclude the validity of "$\Rightarrow$".

Proceeding in this direction, we let $Y_0 = \big\{y \in Y \,\big|\, y_1 a_1 + \cdots + y_m a_n = 0\big\}$ and take I_0 to be the set of indices $i \in [1, s]$ such that there exists $y \in Y_0$ with $y_i > 0$. There exists then a vector $\tilde{y} \in Y_0$ having $\tilde{y}_i > 0$ for every $i \in I_0$ (which we can get by choosing for each $i \in I_0$ a $y^i \in Y_0$ with $y^i_i > 0$ and taking $\tilde{y}$ to be the sum of these vectors y^i for $i \in I_0$). The vector $\tilde{v} = \tilde{y}_1 a_1 + \cdots + \tilde{y}_m a_m$ belongs to $N_C(\bar{x})$ (because of the validity of "$\Leftarrow$" in the normal cone formula), and therefore (from the characterization of normals to convex sets) we have $\tilde{v} \cdot [x - \bar{x}] \leq 0$ for all $x \in C$. In other words (since $\bar{x} = 0$),

$$\tilde{y}_1 [a_1 \cdot x] + \cdots + \tilde{y}_m [a_m \cdot x] \leq 0 \text{ for all } x \in C.$$

But $x \in C$ means that $a_i \cdot x \leq 0$ for $i \in [1, s]$, while $a_i \cdot x = 0$ for $i \in [s+1, m]$. From the special construction of the multiplier vector $\tilde{y}$, we see this implies $a_i \cdot x = 0$ for all $x \in C$

when $i \in I_0$. Thus, C wouldn't be changed if the inequality constraints $a_i \cdot x \leq 0$ for $i \in I_0$ were switched over to being equality constraints $a_i \cdot x = 0$; we have

$$C = \big\{x \,\big|\, a_i \cdot x \leq 0 \text{ for } i \in [1,s]\backslash I_0,\ a_i \cdot x = 0 \text{ for } i \in [s+1,m] \cup I_0\big\}.$$

On this basis, therefore, let $I_1 = [1,s]\backslash I_0$. Choose a maximal linearly independent subset of $\big\{a_i \,\big|\, i \in [s+1,m] \cup I_0\big\}$. Then $\big\{a_i \,\big|\, i \in I_2\big\}$ is a basis for the subspace of $\mathbb{R}^n$ spanned by $\big\{a_i \,\big|\, i \in [s+1,m] \cup I_0\big\}$, so every a_i for i in $[s+1,m] \cup I_0$ but not in I_2 can be expressed as a linear combination of the vectors in this basis. Utilizing this to substitute for such a_i in the equation $\tilde{y}_1 a_1 + \cdots + \tilde{y}_m a_m = 0$ we have for our $\tilde{y} \in Y_0$, we can get (for later purposes) a vector $\tilde{y}' \in Y_0$ such that $\tilde{y}'_i = 0$ for each i in $[s+1,m] \cup I_0$ that's not in I_2, so that

$$\sum_{i \in I_0} \tilde{y}'_i a_i + \sum_{i \in I_1} \tilde{y}'_i a_i + \sum_{i \in I_2} \tilde{y}'_i a_i = 0 \quad \text{with } \tilde{y}'_i \geq 0 \text{ for } i \in I_0 \text{ and } \tilde{y}'_i > 0 \text{ for } i \in I_1.$$

In particular, the basis property of $\big\{a_i \,\big|\, i \in I_2\big\}$ implies that the set of x satisfying $a_i \cdot x = 0$ for all $i \in [s+1,m] \cup I_0$ is the same as the set of x satisfying $a_i \cdot x = 0$ for all $i \in I_2$. Thus, we have an expression of C as the feasible set with respect to the constraints $a_i \cdot x \leq 0$ for $i \in I_1$ and $a_i \cdot x = 0$ for $i \in I_2$, as we mentioned in our plan. We want to confirm that this reduced constraint system satisfies the corresponding version of the standard constraint qualification ($\star$). That means showing that an equation of the form $\sum_{i \in I_1} y_i a_i + \sum_{i \in I_2} y_i a_i = 0$ with $y_i \geq 0$ for $i \in I_1$ can't hold unless $y_i = 0$ for all $i \in I_1 \cup I_2$. Is this fulfilled?

Given such coefficients y_i, we can augment them by $y_i = 0$ for $i \notin I_1 \cup I_2$ to get a vector $y \in Y$ satisfying $y_1 a_1 + \cdots + y_m a_m = 0$. Then, by passing to $y' = y + \lambda \tilde{y}'$ for some $\lambda > 0$ sufficiently large, we can obtain $y' \in Y$ such that $y'_1 a_1 + \cdots + y'_m a_m = 0$, but with $y'_i \geq 0$ for all $i \in I_0 \cup I_1 = [1,s]$, not merely for $i \in I_1$. Both $\tilde{y}'$ and this vector y' belong to the set Y_0 involved in the definition of I_0 and I_1; according to that definition, we must have both $\tilde{y}'_i$ and $y'_i = 0$ for all $i \in I_1$. Hence $y_i = 0$ as well for all $i \in I_1$, and in the equation $y_1 a_1 + \cdots + y_m a_m = 0$ the nonzero coefficients, if any, can only occur for indices $i \in [s+1,m] \cup I_0$. But for such indices that aren't in I_2, we have $y_i = 0$. On the other hand, being left now with the equation coming down to $\sum_{i \in I_2} y_i a_i = 0$ we can conclude that even for $i \in I_2$ we have $y_i = 0$, by the linear independence of $\big\{a_i \,\big|\, i \in I_2\big\}$.

Having established that the reduced constraint representation for C in terms of the index sets I_1 and I_2 does fit the assumptions of Theorem 10, we obtain from that theorem, for any $v \in N_C(\bar{x})$, an expression of the form $v = \sum_{i \in I_1} y_i a_i + \sum_{i \in I_2} y_i a_i$ with $y_i \geq 0$ when $i \in I_1$. Again, we can augment these by taking $y_i = 0$ for $i \notin I_1 \cup I_2$. Next, take

$y' = y + \lambda\tilde{y}$ for $\lambda > 0$ large enough that $y'_i \geq 0$ for all $i \in I_0 \cup I_1 = [1, s]$, which is possible because $\tilde{y}_i > 0$ for $i \in I_0$ and $\tilde{y}_i \geq 0$ for $i \in I_1$). Then, since $\tilde{y}_1 a_1 + \cdots + \tilde{y}_m a_m = 0$, we will have $y' \in Y$ and $v = y'_1 a_1 + \cdots + y'_m a_m$, which is the type of expression we had to achieve.

Part 2. Next, we tackle the general claim in (a). We can suppose the notation is chosen so that the indices of the affine inequality constraints are $i \in [1, q]$ while those of the affine equality constraints are $i \in [r + 1, m]$. Let

$$X' = \big\{x \in X \,\big|\, f_i(x) \leq 0 \text{ for } i \in [1, q],\ f_i(x) = 0 \text{ for } i \in [r + 1, m]\big\},$$

so that C can be interpreted equally well as consisting of all x satisfying

$$x \in X' \text{ and } f_i(x) \begin{cases} \leq 0 & \text{for } i = q + 1, \ldots, s, \\ = 0 & \text{for } i = s + 1, \ldots, r. \end{cases}$$

Let $\bar{x}$ be a point where the refined constraint qualification ($\star\star$) holds.

Because X' has been specified by a linear constraint system, the normal cone formula of Theorem 10 applies to it without any need for a constraint qualification; this is what we've just established in Part 1. Thus, $N_{X'}(\bar{x})$ consists of the vectors of the form

$$z' = \sum_{i=1}^{q} y_i \nabla f_i(\bar{x}) + \sum_{i=r+1}^{m} y_i \nabla f_i(\bar{x}) \text{ with } \begin{cases} y_i \geq 0 & \text{for } i \in [1, q] \text{ active at } \bar{x}, \\ y_i = 0 & \text{for } i \in [1, q] \text{ inactive at } \bar{x}, \\ y_i \text{ free} & \text{for } i \in [r + 1, m]. \end{cases}$$

We wish is to apply Theorem 10 to the alternative constraint representation of C, using this information. Note that X' closed as well as regular at all of its points, because X' is polyhedral (in particular convex). The standard constraint qualification ($\star$) for the alternative constraint representation of C obliges us to examine the possibilities of having

$$-\sum_{i=q+1}^{r} y_i \nabla f_i(\bar{x}) \in N_{X'}(\bar{x}) \text{ with } \begin{cases} y_i \geq 0 & \text{for } i \in [q + 1, s] \text{ active at } \bar{x}, \\ y_i = 0 & \text{for } i \in [q + 1, s] \text{ inactive at } \bar{x}, \\ y_i \text{ free} & \text{for } i \in [s + 1, r]. \end{cases}$$

The issue is whether this necessitates $y_i = 0$ for all $i \in [q+1, r]$. It does, through the direct combination of the formula for vectors $z' \in N_{X'}(\bar{x})$ and the refined constraint qualification ($\star\star$) being satisfied at $\bar{x}$. It follows then from Theorem 10 that $N_C(\bar{x})$ consists of the vectors v that can be expressed in the form

$$v = \sum_{i=q+1}^{r} \bar{y}_i \nabla f_i(\bar{x}) + z' \text{ with } \begin{cases} z' \in N_{X'}(\bar{x}) & \\ \bar{y}_i \geq 0 & \text{for } i \in [q + 1, s] \text{ active at } \bar{x}, \\ \bar{y}_i = 0 & \text{for } i \in [q + 1, s] \text{ inactive at } \bar{x}, \\ \bar{y}_i \text{ free} & \text{for } i \in [s + 1, r]. \end{cases}$$

Invoking the formula for vectors $z' \in N_{X'}(\bar{x})$ once more, we end up with the additional Lagrange multipliers needed to see that these vectors v are the the same as the ones expressible by the normal cone formula in Theorem 10.

Part 3. Our task now is the confirmation of (b). Consider any $\bar{x} \in C$. We'll show that under $(\star\star\star)$ we have $(\star\star)$ fulfilled at $\bar{x}$, that the conclusion we want follows from (a). Let $\tilde{x}$ have the property in $(\star\star\star)$. By the convexity of each f_i in (b), we have

$$\begin{cases} 0 > f_i(\tilde{x}) - f_i(\bar{x}) \geq \nabla f_i(\bar{x}) \cdot [\tilde{x} - \bar{x}] & \text{for } i \in [1, s] \text{ active nonaffine at } \bar{x}, \\ 0 \geq f_i(\tilde{x}) - f_i(\bar{x}) = \nabla f_i(\bar{x}) \cdot [\tilde{x} - \bar{x}] & \text{for } i \in [1, s] \text{ active affine at } \bar{x}, \\ 0 = f_i(\tilde{x}) - f_i(\bar{x}) = \nabla f_i(\bar{x}) \cdot [\tilde{x} - \bar{x}] & \text{for } i \in [s+1, m]. \end{cases}$$

Suppose that $y = (y_1, \ldots, y_m) \in Y(\bar{x})$ with $-\big[y_1 \nabla f_1(\bar{x}) + \cdots + y_m \nabla f_m(\bar{x})\big] \in N_X(\bar{x})$. Since $\tilde{x} \in X$ in particular, and X is convex, we must have

$$0 \geq -\big[y_1 \nabla f_1(\bar{x}) + \cdots + y_m \nabla f_m(\bar{x})\big] \cdot [\tilde{x} - \bar{x}] = -\sum_{i=1}^{m} y_i \nabla f_i(\bar{x}) \cdot [\tilde{x} - \bar{x}].$$

In this sum the terms for the equality constraints and the inactive inequality constraints drop out. The remaining terms, for the active inequality constraints, all have $y_i \geq 0$ and $\nabla f_i(\bar{x}) \cdot [\tilde{x} - \bar{x}] \leq 0$, moreover with the latter inequality strict when f_i is not affine. The sum would therefore come out > 0 if any of the active inequality constraints with f_i not affine had $y_i > 0$. Hence $y_i = 0$ for all such constraints, and $(\star\star)$ is indeed satisfied. □

Original Slater condition: The refined Slater condition in (b) is notable for providing a test, on the constraints only, that yields a conclusion applicable simultaneously to *every* point $\bar{x} \in C$. The original version of it, going back to the early history of convex programming, concerned convex constraint systems with $X = \mathbb{R}^n$ and inequality constraints only, and made no distinction between affine and nonaffine convex constraints.

Key consequences for linear and convex programming: When $(\mathcal{P})$ is a problem of convex programming with *linear constraints only* (as in the case of linear programming), the Kuhn-Tucker conditions are *always necessary* for the optimality of $\bar{x}$. This is true since, by Theorem 12(a), no constraint qualification is required in Theorem 11(a) if the constraint system is linear.

When $(\mathcal{P})$ is a problem of convex programming with some *nonlinear convex constraints*, one can't get away with no constraint qualification at all. However, the refined Slater condition in Theorem 12(b) has the remarkable ability to confirm the necessity of the Kuhn-Tucker conditions *without requiring any specific knowledge about $\bar{x}$ in advance of computation.*

Second-order conditions for optimality: Until now in the study of Lagrange multipliers, we have been occupied with first-order conditions only. The full theory of second-order necessary conditions and sufficient conditions for local optimality is subtle and complicated. (It likewise can be formulated with tangent and normal cones but can involve juggling several different Lagrange multiplier vectors at the same point $\bar{x}$.) Here we'll be content with looking only at a particular form of *sufficient* condition that's especially favored in the development of numerical methods.

THEOREM 13 (second-order optimality for problems in standard format). *In problem $(\mathcal{P})$, suppose that X is polyhedral and the functions $f_0, f_1, \ldots, f_m$ are of class $\mathcal{C}^2$. Suppose $\bar{x}$ is a feasible solution at which the Kuhn-Tucker conditions hold for a Lagrange multiplier vector $\bar{y}$ such that, in addition,*

$$w \cdot \nabla^2_{xx} L(\bar{x}, \bar{y}) w > 0 \text{ for every } w \in T \text{ satisfying } \nabla f_0(\bar{x}) \cdot w = 0,\ w \neq 0,$$

where T is the set consisting of vectors w such that

$$w \in T_X(\bar{x}) \text{ with } \begin{cases} \nabla f_i(\bar{x}) \cdot w \text{ free} & \text{for } i \in [1, s] \text{ with } f_i(\bar{x}) < 0, \\ \nabla f_i(\bar{x}) \cdot w \leq 0 & \text{for } i \in [1, s] \text{ with } f_i(\bar{x}) = 0, \\ \nabla f_i(\bar{x}) \cdot w = 0 & \text{for } i \in [s+1, m]. \end{cases}$$

Then $\bar{x}$ is a locally optimal solution to problem $(\mathcal{P})$. Moreover, in these circumstances the local optimality of $\bar{x}$ is strict, in the sense that there exists a $\delta > 0$ such that

$$f_0(x) > f_0(\bar{x}) \text{ for all points } x \in C \text{ with } 0 < |x - \bar{x}| < \delta.$$

Note: The set T here can be identified with the tangent cone $T_C(\bar{x})$ by Theorem 10 under the standard constraint qualification $(\star)$ or one of its substitutes $(\star\star)$ or $(\star\star\star)$.

Proof. For a value $\rho > 0$ of magnitude yet to be determined, consider the problem

$$\text{minimize } f(x, u) := f_0(x) + \sum_{i=1}^{m} \bar{y}_i \big[f_i(x) - u_i \big] + \frac{\rho}{2} \sum_{i=1}^{m} \big[f_i(x) - u_i \big]^2 \text{ over } X \times K,$$

where K is again the box formed by product of s intervals $(-\infty, 0]$ and $m - s$ intervals $[0, 0]$. This is a problem in which a $\mathcal{C}^2$ function f is minimized over a polyhedral set, and the sufficient condition in Theorem 6(b) is therefore applicable for establishing local optimality. For $\bar{u}_i = f_i(\bar{x})$, is this condition satisfied at $(\bar{x}, \bar{u})$? It turns out that under our hypothesis it is, provided ρ is high enough. This will be verified shortly, but first suppose it's true, in order to see where it leads.

Suppose, in other words, that $f(x,u) \geq f(\bar{x},\bar{u})$ for all $(x,u) \in X \times K$ in some neighborhood of $(\bar{x},\bar{u})$. Then for feasible $x \in X$ near enough to $\bar{x}$, the vector $u(x) := \big(f_1(x),\ldots,f_m(x)\big)$ in K will (by the continuity of the f_i's) be near to $u(\bar{x}) = \bar{z}$ with $f(x,u(x)) \geq f(\bar{x},\bar{u})$. But $f(x,u(x)) = f_0(x)$ when x is feasible, and in particular $f(\bar{x},\bar{u}) = f_0(\bar{x})$. It follows that $f_0(x) \geq f_0(\bar{x})$ for all feasible x in some neighborhood of $\bar{x}$, and we conclude that $\bar{x}$ is locally optimal in the given problem.

We proceed now with verifying that for large values of ρ the sufficient condition in Theorem 6(b) is satisfied for the local optimality of $(\bar{x},\bar{u})$ in the problem of minimizing f over $X \times K$. The condition in question involves first and second partial derivatives of f as well as the tangent cone to the box $X \times K$, which from the characterization given earlier for tangents to boxes can be expressed in the product form $T_{X\times K}(\bar{x},\bar{u}) = T_X(\bar{x}) \times T_K(\bar{u})$. Specifically, the condition requires that

$$\begin{aligned}
&\nabla f(\bar{x},\bar{u})\cdot(w,z) \geq 0 \text{ for all } (w,z) \text{ in } T_X(\bar{x}) \times T_K(\bar{u}),\\
&(w,z)\cdot\nabla^2 f(\bar{x},\bar{w})(w,z) > 0 \text{ for all } (w,z) \neq (0,0) \text{ in } T_X(\bar{x}) \times T_K(\bar{u})\\
&\qquad\qquad \text{with } \nabla f(\bar{x},\bar{u})\cdot(w,z) = 0.
\end{aligned}$$

The first partial derivatives are

$$\begin{aligned}
\frac{\partial f}{\partial x_j}(x,u) &= \frac{\partial L}{\partial x_j}(x,\bar{y}) + \rho\sum_{i=1}^m \big[f_i(x) - u_i\big]\frac{\partial f_i}{\partial x_j}(x),\\
\frac{\partial f}{\partial u_i}(x,u) &= -\bar{y}_i - \rho\big[f_i(x) - u_i\big],
\end{aligned}$$

while the second partial derivatives are

$$\begin{aligned}
&\frac{\partial^2 f}{\partial x_k \partial x_j}(x,u) = \frac{\partial^2 L}{\partial x_k \partial x_j}(x,\bar{y}) + \rho\sum_{i=1}^m \big[f_i(x) - u_i\big]\frac{\partial^2 f_i}{\partial x_k \partial x_j}(x) + \rho\frac{\partial f_i}{\partial x_k}(x)\frac{\partial f_i}{\partial x_j}(x),\\
&\frac{\partial^2 f}{\partial u_l \partial x_j}(x,u) = -\rho\frac{\partial f_l}{\partial x_j}(x), \qquad \frac{\partial^2 f}{\partial x_k \partial u_i}(x,u) = -\rho\frac{\partial f_i}{\partial x_k}(x),\\
&\frac{\partial^2 f}{\partial u_l \partial u_i}(x,u) = \begin{cases} \rho & \text{if } l = i\\ 0 & \text{if } l \neq i. \end{cases}
\end{aligned}$$

Because $f_i(\bar{x}) - \bar{u}_i = 0$, we have $\nabla f(\bar{x},\bar{u})\cdot(w,z) = \nabla_x L(\bar{x},\bar{y})\cdot w - \bar{y}\cdot z$ and on the other hand $(w,z)\cdot\nabla^2 f(\bar{x},\bar{u})(w,z) = w\cdot\nabla^2_{xx}L(\bar{x},\bar{y})w + \rho\sum_{i=1}^m \big[\nabla f_i(\bar{x})\cdot w - z_i\big]^2$. The sufficient condition we wish to verify (for ρ large) thus takes the form:

$$\begin{aligned}
&\nabla_x L(\bar{x},\bar{y})\cdot w \geq 0 \text{ for all } w \in T_X(\bar{x}), \qquad \bar{y}\cdot z \leq 0 \text{ for all } z \in T_K(\bar{u}),\\
&w\cdot\nabla^2_{xx}L(\bar{x},\bar{y})w + \rho\sum_{i=1}^m \big[\nabla f_i(\bar{x})\cdot w - z_i\big]^2 > 0 \text{ for all } (w,z) \neq (0,0) \text{ with}\\
&\qquad w \in T_X(\bar{x}), \quad \nabla_x L(\bar{x},\bar{y})\cdot w = 0, \quad z \in T_K(\bar{u}), \quad \bar{y}\cdot z = 0.
\end{aligned}$$

Here the first-order inequalities merely restate the relations $-\nabla_x L(\bar{x},\bar{y}) \in N_X(\bar{x})$ and $\bar{y} \in N_K(\bar{u})$ (equivalent to $\bar{u} \in N_Y(\bar{y})$, as seen before), which hold by assumption. In the second-order condition we obviously do have strict inequality when $w = 0$ and $z \neq 0$, since the quadratic expression reduces in that case to $\rho|z|^2$. Therefore, we can limit attention to demonstrating strict inequality in cases where $w \neq 0$, or more specifically (through rescaling), where $|w| = 1$. From the form of K and $\bar{u}$ we know

$$\left.\begin{array}{l} z \in T_K(\bar{u}) \\ \bar{y}\cdot z = 0 \end{array}\right\} \iff \left\{\begin{array}{ll} z_i \text{ free} & \text{for inactive } i \in [1,s], \\ z_i \leq 0 & \text{for active } i \in [1,s] \text{ with } \bar{y}_i = 0, \\ z_i = 0 & \text{for active } i \in [1,s] \text{ with } \bar{y}_i > 0 \text{ and for } i \in [s+1,m], \end{array}\right.$$

so for any $w \neq 0$ in $T_X(\bar{x})$ the minimum of the quadratic expression with respect to $z \in T_K(\bar{u})$ with $\bar{y}\cdot z = 0$ will be attained when $z_i = z_i(w)$ with

$$z_i(w) = \left\{\begin{array}{ll} \nabla f_i(\bar{x})\cdot w & \text{for inactive } i \in [1,s], \\ \min\big\{0, \nabla f_i(\bar{x})\cdot w\big\} & \text{for active } i \in [1,s] \text{ with } \bar{y}_i = 0, \\ 0 & \text{for active } i \in [1,s] \text{ with } \bar{y}_i > 0 \text{ and for } i \in [s+1,m]. \end{array}\right.$$

Thus, we can even limit attention to pairs (w,z) having both $|w| = 1$ and $z_i = z_i(w)$. We'll suppose the claim for this special case is false and argue toward a contradiction.

If the claim is false, there has to be a sequence of values $\rho^\nu \to \infty$ along with vectors $w^\nu \in T_X(\bar{x})$ with $|w^\nu| = 1$ such that $\nabla_x L(\bar{x},\bar{y})\cdot w^\nu = 0$ and

$$w^\nu \cdot \nabla^2_{xx} L(\bar{x},\bar{y}) w^\nu + \rho^\nu \sum_{i=1}^m \Big[\nabla f_i(\bar{x})\cdot w^\nu - z_i(w^\nu)\Big]^2 \leq 0,$$

and hence in particular

$$\begin{aligned} w^\nu \cdot \nabla^2_{xx} L(\bar{x},\bar{y}) w^\nu &\leq 0, \\ \Big[\nabla f_i(\bar{x})\cdot w^\nu - z_i(w^\nu)\Big]^2 &\leq -\frac{1}{\rho^\nu} w^\nu \cdot \nabla^2_{xx} L(\bar{x},\bar{y}) w^\nu \text{ for all } i. \end{aligned}$$

Because the sequence $\{w^\nu\}_{\nu=1}^\infty$ is bounded, it has a cluster point $\bar{w}$. By the continuity of the expressions involved, and the closedness of tangent cones, we get in the limit that

$$\begin{aligned} &\bar{w} \in T_X(\bar{x}), \qquad |\bar{w}| = 1, \qquad \nabla_x L(\bar{x},\bar{y})\cdot\bar{w} = 0, \qquad \bar{y}\cdot\bar{z}(\bar{w}) = 0, \\ &\bar{w}\cdot\nabla^2_{xx} L(\bar{x},\bar{y})\bar{w} \leq 0, \qquad \big[\nabla f_i(\bar{x})\cdot\bar{w} - z_i(\bar{w})\big]^2 \leq 0 \text{ for all } i. \end{aligned}$$

The final batch of inequalities says that $\nabla f_i(\bar{x})\cdot\bar{w} = z_i(\bar{w})$ for all i, which means

$$\nabla f_i(\bar{x})\cdot w \left\{\begin{array}{ll} \leq 0 & \text{for active } i \in [1,s] \text{ with } \bar{y}_i = 0, \\ = 0 & \text{for inactive } i \in [1,s] \text{ with } \bar{y}_i > 0 \text{ and for } i \in [s+1,m]. \end{array}\right.$$

These conditions along with the fact that $F(\bar{x}) \in N_Y(\bar{y})$ and

$$\nabla_x L(\bar{x}, \bar{y}) \cdot \bar{w} \;=\; \nabla f_0(\bar{x}) \cdot \bar{w} + \bar{y}_1 \nabla f_1(\bar{x}) \cdot \bar{w} + \cdots + \bar{y}_m \nabla f_m(\bar{x}) \cdot \bar{w}$$

also imply $\nabla f_0(\bar{x}) \cdot \bar{w} = 0$. We have arrived therefore at a vector $\bar{w} \neq 0$ for which the second-order condition in the theorem is violated. This finishes the proof. □

Critical cone: In Theorem 13, the set of all w satisfying $w \in T$ and $\nabla f_i(\bar{x}) \cdot w = 0$ is called the *critical* cone at $\bar{x}$. It's truly a cone, and moreover it's polyhedral, because T itself is described by a certain system of linear constraints on w (inasmuch as $T_X(\bar{x})$ is polyhedral by our assumption that X is polyhedral).

Connection with earlier second-order conditions: How is the sufficient condition for optimality in Theorem 13 related to the earlier one in Theorem 6(b)? The comparison is facilitated by the observation, made just after the statement of Theorem 13, that T is the same as $T_C(\bar{x})$ under one of the constraint qualifications $(\star)$, $(\star\star)$, or $(\star\star)$. In Theorem 6(b), a very similar condition was stated in terms of $T_C(\bar{x})$. The important difference, though, is that in Theorem 6(b) the strict inequality was invoked for the Hessian matrix $\nabla^2 f_0(\bar{x})$, whereas in Theorem 13 the matrix that appears is

$$\nabla^2_{xx} L(\bar{x}, \bar{y}) = \nabla^2 f_0(\bar{x}) + \bar{y}_1 \nabla^2 f_1(\bar{x}) + \cdots + \bar{y}_m \nabla^2 f_m(\bar{x}).$$

Another difference, however, is that C was assumed to be polyhedral in Theorem 6, whereas now it might not even be convex.

Of course, in the more general framework we have now, if the constraint functions $f_1, \ldots, f_m$ are actually affine, then C is indeed polyhedral (because the underlying set X was assumed in Theorem 13 to be polyhedral). The Hessian matrices for those functions vanish in that case, so that in fact $\nabla^2_{xx} L(\bar{x}, \bar{y}) = \nabla^2 f_0(\bar{x})$. In recalling that no constraint qualification is needed when dealing with linear constraints, we see that Theorem 13 reduces exactly to Theorem 6(b) in the special case where $f_1, \ldots, f_m$ are affine. The formulation in Theorem 6(b) doesn't work for sets C that aren't polyhedral, because such sets have "curvature" that must be taken into account. That's done through the extra terms in $\nabla^2_{xx} L(\bar{x}, \bar{y})$.

Corresponding second-order necessary condition? Caution: It might be imagined that by weakening the "$>$" in Theorem 13 to "$\geq$" the second-order sufficient condition could be transformed into a second-order necessary condition for optimality. But that's not true even under the assumption that the standard constraint qualification $(\star)$ holds in $(\mathcal{P})$. To establish a second-order necessary condition, one can't really

get by with only a single Lagrange multiplier vector $\bar{y}$. Instead, several such vectors may be needed, and it's the max of $w \cdot \nabla^2_{xx} L(\bar{x}, \bar{y}) w$ over those different $\bar{y}$'s that has to be positive for each w.

Generalized constraint systems: A powerful feature of Theorem 10, the key to the approach we've taken to Lagrange multipliers, is the fact that it extends, in the right interpretation, to all sets C that can be represented in the form

$$C = \big\{ x \in X \,\big|\, F(x) \in K \big\}, \text{ with } F(x) = (f_1(x), \ldots, f_m),$$

for closed sets $X \subset \mathbb{R}^n$, $K \subset \mathbb{R}^m$ and $\mathcal{C}^1$ functions f_i on $\mathbb{R}^n$. For a standard feasible set, K is the special box described before the statement of Theorem 10; it makes $N_K(F(\bar{x}))$ be $Y(\bar{x})$. But the proof of Theorem 10 only requires K to be a closed set that is regular at $F(\bar{x})$. For such K, the normal cone formula in Theorem 10 remains valid when $Y(\bar{x})$ is replaced there, and in the constraint qualification ($\star$), by $N_K(F(\bar{x}))$, whatever that might be (depending on the choice of K). For tangents, the corresponding formula then is $T_C(\bar{x}) = \big\{ w \in T_X(\bar{x}) \,\big|\, \nabla F(\bar{x}) w \in T_K\big(F(\bar{x})\big) \big\}$.

Example in variational geometry: For the theoretically minded, here's an illustration of the far-reaching consequences of this observation. Suppose $C = C_1 \cap C_2$ for closed sets C_1 and C_2 in $\mathbb{R}^n$. What can be said about $N_C(\bar{x})$ and $T_C(\bar{x})$ in terms of normals and tangents to C_1 and C_2 themselves?

We can approach this by taking $X = \mathbb{R}^n$, $K = C_1 \times C_2 \subset \mathbb{R}^n \times \mathbb{R}^n$, and defining $F(x) = (x, x)$. Then C consists of all $x \in X$ such that $F(x) \in K$. When the generalization of Theorem 10 just described is invoked in this framework, the constraint qualification turns into the condition that there be no choice of $v_1 \in N_{C_1}(\bar{x})$ and $v_1 \in N_{C_1}(\bar{x})$ such that $v_1 + v_2 = 0$. When that holds, he conclusion obtained is that

$$T_C(\bar{x}) = T_{C_1}(\bar{x}) \cap T_{C_2}(\bar{x}), \qquad N_C(\bar{x}) = N_{C_1}(\bar{x}) + N_{C_2}(\bar{x}),$$

where the set sum refers to the set $\big\{ v_1 + v_2 \,\big|\, v_1 \in N_{C_1}(\bar{x}),\ v_2 \in N_{C_2}(\bar{x}) \big\}$.

6. GAMES AND DUALITY

In trying to understand how Lagrange multiplier vectors $\bar{y}$ might be generated or utilized in computational schemes, the case of linear programming is instructive. Inspection of the Kuhn-Tucker conditions in that case reveals that these vectors solve a mysterious problem of optimization inextricably tied to the given one. Pursuing the mystery further, we are led to an interesting branch of modern mathematics: game theory. We'll look briefly at this theory and use it to develop the fact that for convex programming problems quite generally the Lagrange multiplier vectors associated with optimality can be obtained in principle by solving an auxiliary problem said to be "dual" to the given "primal" problem.

Kuhn-Tucker conditions in linear programming: Consider an optimization problem in the so-called *primal canonical format in linear programming*:

$$(\mathcal{P}_{\text{lin}}) \qquad \begin{aligned} &\text{minimize } \textstyle\sum_{j=1}^n c_j x_j \text{ subject to} \\ &\textstyle\sum_{j=1}^n a_{ij} x_j \geq b_i \text{ for } i = 1, \ldots, m, \\ &x_j \geq 0 \text{ for } j = 1, \ldots, n. \end{aligned}$$

This corresponds to a problem $(\mathcal{P})$ standard format with objective function $f_0(x) = \sum_{j=1}^n c_j x_j$ and constraints $x \in X = \mathbb{R}^n_+$ and $0 \geq f_i(x) = b_i - \sum_{j=1}^n a_{ij} x_j$ for $i = 1, \ldots, m$. We have $s = m$, so $Y = \mathbb{R}^m_+$. The Lagrangian is therefore

$$\begin{aligned} L(x, y) &= \textstyle\sum_{j=1}^n c_j x_j + \sum_{i=1}^m y_i \Big(b_i - \sum_{j=1}^n a_{ij} x_j \Big) \\ &= \textstyle\sum_{i=1}^m b_i y_i + \sum_{j=1}^n x_j \Big(c_j - \sum_{i=1}^m y_i a_{ij} \Big) \text{ on } \mathbb{R}^n_+ \times \mathbb{R}^m_+. \end{aligned}$$

The Kuhn-Tucker conditions on $\bar{x}$ and $\bar{y}$, which as we know can be written in general as $-\nabla_x L(\bar{x}, \bar{y}) \in N_X(\bar{x})$ and $\nabla_y L(\bar{x}, \bar{y}) \in N_Y(\bar{y})$, come out for this L, X and Y as the following conditions, coordinate by coordinate:

$$\begin{cases} \bar{x}_j \geq 0, \ \Big(c_j - \sum_{i=1}^m \bar{y}_i a_{ij} \Big) \geq 0, \ \bar{x}_j \Big(c_j - \sum_{i=1}^m \bar{y}_i a_{ij} \Big) = 0 & \text{for } j = 1, \ldots, n, \\ \bar{y}_i \geq 0, \ \Big(\sum_{j=1}^n a_{ij} \bar{x}_j - b_i \Big) \geq 0, \ \bar{y}_i \Big(\sum_{j=1}^n a_{ij} \bar{x}_j - b_i \Big) = 0 & \text{for } i = 1, \ldots, m. \end{cases}$$

Complementary slackness: These relationships are called the *complementary slackness conditions* in linear programming. They list for each index j or i a pair of inequalities (one on $\bar{x}$ and one on $\bar{y}$), requiring that at most one of the two can be "slack," i.e., satisfied with strict inequality. By Theorems 11 and 12, $\bar{x}$ is optimal in $(\mathcal{P}_{\text{lin}})$ if and only if these relationships hold for some $\bar{y}$.

Dual linear programming problem: The symmetry in these conditions is tantalizing. There turns out to be a connection with the following problem, said to be *dual* to $(\mathcal{P}_{\text{lin}})$, which is in the so-called *dual canonical format in linear programming*:

$$(\mathcal{D}_{\text{lin}}) \qquad \begin{aligned} &\text{maximize } \textstyle\sum_{i=1}^{m} b_i y_i \text{ subject to} \\ &\textstyle\sum_{i=1}^{m} y_i a_{ij} \le c_j \text{ for } j = 1, \ldots, n, \\ &\qquad y_i \ge 0 \text{ for } i = 1, \ldots, m. \end{aligned}$$

To see the formal relationship between $(\mathcal{D}_{\text{lin}})$ and $(\mathcal{P}_{\text{lin}})$, begin by converting $(\mathcal{D}_{\text{lin}})$ from dual canonical format to primal canonical format:

$$\begin{aligned} &\text{minimize } \textstyle\sum_{i=1}^{m} [-b_i] y_i \text{ subject to} \\ &\textstyle\sum_{i=1}^{m} y_i [-a_{ij}] \ge [-c_j] \text{ for } j = 1, \ldots, n, \\ &\qquad y_i \ge 0 \text{ for } i = 1, \ldots, m. \end{aligned}$$

As an act of faith, permit the symbols $\bar{x}_j$ to be used for the Lagrange multipliers associated with optimality in this problem, without presupposing for now any tie to the previous $\bar{x}$. The Kuhn-Tucker conditions for this problem in their complementary slackness formulation emerge then as

$$\begin{cases} \bar{y}_i \ge 0, \;\; \Big([-b_i] - \sum_{j=1}^{n} [-a_{ij}] \bar{x}_j\Big) \ge 0, \;\; \bar{y}_i \Big([-b_i] - \sum_{j=1}^{n} [-a_{ij}] \bar{x}_j\Big) = 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } i = 1, \ldots, m, \\ \bar{x}_j \ge 0, \;\; \Big(\sum_{i=1}^{m} \bar{y}_i [-a_{ij}] - [-c_j]\Big) \ge 0, \;\; \bar{x}_j \Big(\sum_{i=1}^{m} \bar{y}_i [-a_{ij}] - [-c_j]\Big) = 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } j = 1, \ldots, n. \end{cases}$$

But these are identical to the complementary slackness conditions we had before. Problems $(\mathcal{P}_{\text{lin}})$ and $(\mathcal{D}_{\text{lin}})$ thus turn out to share the very same optimality conditions! Neither can be solved without somehow, explicitly or implicitly, solving the other as well. This astounding observation, which at first seems to lack a meaningful basis, leads to important consequences.

Symmetry: The reason for treating $(\mathcal{D}_{\text{lin}})$ as a problem of maximization instead of minimization is to bring out not only the symmetry in this switch of signs, but to promote a relationship of optimal values which is given in the theorem below.

It's interesting to note that after $(\mathcal{D}_{\text{lin}})$ has been converted to a problem in primal canonical form—denoted say by $(\mathcal{P}'_{\text{lin}})$—it will in turn have an associated dual problem—say $(\mathcal{D}'_{\text{lin}})$. But this can be seen to be none other than the problem obtained by converting $(\mathcal{P}_{\text{lin}})$ from primal canonical form to dual canonical form.

Canonical formats: As noted, any linear programming problem in primal canonical format can be transformed into dual canonical format and vice versa. In fact, any linear programming problem at all can be recast in either framework; either can serve as a means for standardization in formulating problems. This is true because equation constraints can be replaced by pairs of inequality constraints, while any free variable can be modeled as a difference of nonnegative variables.

THEOREM 14 (duality in linear programming). *For a linear programming problem* $(\mathcal{P}_{\text{lin}})$ *in primal canonical format and the associated linear programming problem* $(\mathcal{D}_{\text{lin}})$ *in dual canonical format, the following properties of a pair of vectors* $\bar{x}$ *and* $\bar{y}$ *are equivalent to each other (so that if any one of them holds, they all hold):*

(a) $\bar{x}$ *is an optimal solution to* $(\mathcal{P}_{\text{lin}})$*, and* $\bar{y}$ *is an associated Lagrange multiplier vector in the Kuhn-Tucker conditions for* $(\mathcal{P}_{\text{lin}})$ *at* $\bar{x}$*;*

(b) $\bar{y}$ *is an optimal solution to* $(\mathcal{D}_{\text{lin}})$*, and* $\bar{x}$ *is an associated Lagrange multiplier vector in the Kuhn-Tucker conditions for* $(\mathcal{D}_{\text{lin}})$ *at* $\bar{y}$*;*

(c) $\bar{x}$ *and* $\bar{y}$ *are optimal solutions to* $(\mathcal{P}_{\text{lin}})$ *and* $(\mathcal{D}_{\text{lin}})$*, respectively;*

(d) $\bar{x}$ *is a feasible solution to* $(\mathcal{P}_{\text{lin}})$*,* $\bar{y}$ *is a feasible solution to* $(\mathcal{D}_{\text{lin}})$*, and the objective function values at these points are equal:* $\sum_{j=1}^{n} c_j \bar{x}_j = \sum_{i=1}^{m} b_i \bar{y}_i$*;*

(e) $\bar{x}$ *and* $\bar{y}$ *satisfy the complementary slackness conditions*

$$\begin{cases} \bar{x}_j \geq 0, \ \left(c_j - \sum_{i=1}^{m} \bar{y}_i a_{ij}\right) \geq 0, \ \bar{x}_j \left(c_j - \sum_{i=1}^{m} \bar{y}_i a_{ij}\right) = 0 & \text{for } j = 1, \ldots, n, \\ \bar{y}_i \geq 0, \ \left(\sum_{j=1}^{n} a_{ij} \bar{x}_j - b_i\right) \geq 0, \ \bar{y}_i \left(\sum_{j=1}^{n} a_{ij} \bar{x}_j - b_i\right) = 0 & \text{for } i = 1, \ldots, m. \end{cases}$$

Furthermore, if either $(\mathcal{P}_{\text{lin}})$ *or* $(\mathcal{D}_{\text{lin}})$ *has an optimal solution, then optimal solutions* $\bar{x}$ *and* $\bar{y}$ *must exist for both problems, thereby triggering through (c) all the other properties, including through (d) the conclusion that the min in the primal problem equals the max in the dual problem, i.e.,*

$$\left[\text{optimal value in } \left(\mathcal{P}_{\text{lin}}\right)\right] \ = \ \left[\text{optimal value in } \left(\mathcal{D}_{\text{lin}}\right)\right].$$

Proof. The equivalence of (a) and (b) with (e) has been ascertained in the preceding discussion. We also know through Theorems 11 and 12 that the optimality of $\bar{x}$ in $(\mathcal{P}_{\text{lin}})$ is equivalent to the existence of a vector $\bar{y}$ satisfying these conditions, and likewise that the optimality of $\bar{y}$ in $(\mathcal{D}_{\text{lin}})$ is equivalent to the existence of a vector $\bar{x}$ satisfying these conditions. Hence if either problem has an optimal solution, the other must have one as

well. Observe next that the complementary slackness conditions in (e) entail

$$\sum_{j=1}^{n} c_j \bar{x}_j - \sum_{i=1}^{m} \bar{y}_i b_i \;=\; \sum_{j=1}^{n} \Big[c_j - \sum_{i=1}^{m} \bar{y}_i a_{ij} \Big] \bar{x}_j + \sum_{i=1}^{m} \bar{y}_i \Big[\sum_{j=1}^{n} a_{ij} \bar{x}_j - b_i \Big] \;=\; 0.$$

It follows that when optimal solutions exist the optimal values in the two problems coincide. This leads to the further equivalence of (a)–(b)–(e) with (c) and with (d). □

Correctly drawing conclusions about linear programming duality: It must be emphasized that the initial part of Theorem 14 *doesn't* say *directly* whether $(\mathcal{P}_{\rm lin})$ and $(\mathcal{D}_{\rm lin})$ actually enjoy the properties in (a), (b), (c), (d), and (e), or that the optimal values in these problems are equal. It merely asserts an equivalence; to get in the door, you have to have $\bar{x}$ and $\bar{y}$ being known to satisfy at least one of the five conditions. The final part of Theorem 14, however, provides an easy test. To obtain the existence of an $\bar{x}$ and $\bar{y}$ pair satifying all these conditions, and to conclude that the optimal value equation holds, it's enough to verify that one of the two problems has an optimal solution.

Showing that either $(\mathcal{P}_{\rm lin})$ and $(\mathcal{D}_{\rm lin})$ has an optimal solution could be accomplished through well-posedness and an application of Theorem 1, but remember that for the special case of linear programming problems a much easier criterion is available. *A linear programming problem is sure to have an optimal solution if its feasible set is nonempty and its optimal value isn't infinite.* The optimal value in $(\mathcal{P}_{\rm lin})$ can't be $-\infty$ if a finite lower bound exists to its objective function over its feasible set. Sometimes such a bound is very easy to produce (as, for instance, when the objective function is nonnegative over that set). Likewise, the optimal value in $(\mathcal{D}_{\rm lin})$ can't be ∞ if a finite upper bound exists to its objective function over its feasible set.

Existence through double feasibility: It follows from the special existence criterion for optimal solutions in linear programming that, for a primal-dual pair of problems $(\mathcal{P}_{\rm lin})$ and $(\mathcal{D}_{\rm lin})$, if feasible solutions exist to both problems then optimal solutions exist to both problems.

Argument: In a slight extension of the proof of Theorem 14, we observe that for any feasible solutions to the two problems we have

$$\sum_{j=1}^{n} c_j x_j - \sum_{i=1}^{m} y_i b_i \;=\; \sum_{j=1}^{n} \Big[c_j - \sum_{i=1}^{m} y_i a_{ij} \Big] x_j + \sum_{i=1}^{m} y_i \Big[\sum_{j=1}^{n} a_{ij} x_j - b_i \Big] \;\geq\; 0.$$

Then the value $\sum_{i=1}^{m} y_i b_i$ is a lower bound to the optimal value in $(\mathcal{P}_{\rm lin})$ (implying it must be finite), while at the same time $\sum_{j=1}^{n} c_j x_j$ is an upper bound to the optimal value in $(\mathcal{D}_{\rm lin})$ (implying that it too must be finite).

Knowing a priori that the optimal value is finite: The double feasibility test gives another way for ascertaining, without first solving a linear programming problem $(\mathcal{P}_{\text{lin}})$, that its optimal value must be finite: it's enough to verify that $(\mathcal{P}_{\text{lin}})$ and the associated $(\mathcal{D}_{\text{lin}})$ both possess feasible solutions.

Dantzig's simplex method in linear programming: The complementary slackness conditions are the algebraic foundation for an important numerical technique for solving linear programming problems, the very first, which actually was the breakthrough that got optimization rolling around 1950 as a modern subject with impressive practical applications. In theory, the task of finding an optimal solution $\bar{x}$ to $(\mathcal{P}_{\text{lin}})$ is equivalent to that of finding a pair of vectors $\bar{x}$ and $\bar{y}$ for which these conditions hold. Trying directly to come up with solutions to systems of linear inequalities is a daunting challenge, but solving linear equations is more attractive, and this can be made the focus through the fact that the complementary slackness conditions require at least one of the inequalities for each index j or i to hold as an equation.

Let's approach this in terms of selecting of two index sets $I \subset \{i = 1, \ldots, m\}$ and $J \subset \{j = 1, \ldots, n\}$, and associating with the following system of $n + m$ equations in the $n + m$ unknowns $\bar{x}_j$ and $\bar{y}_i$:

$$(I, J) \qquad \begin{aligned} &\textstyle\sum_{j=1}^{m} a_{ij}\bar{x}_j - b_i = 0 \text{ for } i \in I, \qquad \bar{y}_i = 0 \text{ for } i \notin I, \\ &c_j - \textstyle\sum_{i=1}^{m} \bar{y}_i a_{ij} = 0 \text{ for } j \in J, \qquad \bar{x}_j = 0 \text{ for } j \notin J. \end{aligned}$$

To say that $\bar{x}$ and $\bar{y}$ satisfy the complementary slackness conditions is to say that for *some* choice of I and J, this (I, J)-system of linear equations will have a solution $(\bar{x}, \bar{y})$ for which the inequalities

$$\begin{aligned} &\textstyle\sum_{j=1}^{m} a_{ij}\bar{x}_j - b_i \geq 0 \text{ for } i \notin I, \qquad \bar{y}_i \geq 0 \text{ for } i \in I, \\ &c_j - \textstyle\sum_{i=1}^{m} \bar{y}_i a_{ij} \geq 0 \text{ for } j \notin J, \qquad \bar{x}_j \geq 0 \text{ for } j \in J \end{aligned}$$

happen to be satisfied as well. The crucial observation is that there is only a *finite collection* of (I, J)-systems, because there are only finitely many ways of choosing the index sets I and J.

Moreover, it can be shown that one can limit attention to (I, J)-systems that are nondegenerate, i.e., have nonsingular matrix so that there's a unique corresponding pair $(\bar{x}, \bar{y})$. The prospect is thereby raised of searching by computer through a finite list of possible candidate pairs $(\bar{x}, \bar{y})$, each obtained by solving a certain system of linear equations, checking each time whether the desired inequalities are satisfied too,

until a pair is found that meets the test of the complementary slackness conditions and thereby provides an optimal solution to the linear programming problem $(\mathcal{P}_{\text{lin}})$.

Of course, even with this idea one is far from a practical method of computation, because the number of (I, J)-systems that would have to be inspected is likely to be awesomely astronomical, far beyond the capability of thousands of the fastest computers laboring for thousands of years! But fortunately it's not necessary to look at all such systems. There are ways of starting with one such (I, J)-system and then modifying the choice of I and J in tiny steps in such a manner that "improvements" are continually made.

We won't go into this scheme further here, but it's the contribution of Dantzig that so changed the world of optimization at the beginning of the computer age. Nowadays there are other ways of solving linear programming problems, but Dantzig's so-called "simplex method" is still competitive in some situations and continues to be used.

Duality more generally, via game theory: The facts about the tight relationship between the linear programming problems $(\mathcal{P}_{\text{lin}})$ and $(\mathcal{D}_{\text{lin}})$ raise more questions than they answer. What is the "explanation" for this phenomenon, and what significance, not just technical, can be ascribed to the Lagrange multiplier values that appear?

How, for instance, might they be interpreted in a particular application? These issues go far beyond linear programming. To lay the groundwork for their analysis, we need to spend some time with elementary game theory. That branch of mathematics is interesting anyway in its own right, because of connections with economics and sociology.

Lagrangian framework as the key: Duality in linear programming, in its manifestation so far, has grown out of complementary slackness form of the Kuhn-Tucker conditions that characterize optimality. As we move from linear programming to greater generality, we'll be focusing more on the Lagrangian form of the Kuhn-Tucker conditions and on the Lagrangian L itself as a special function on a certain product set $X \times Y$.

Cases where a pair $(\bar{x}, \bar{y})$ constitutes a "Lagrangian saddle point" will be especially important. For the time being, though, we'll profit from allowing the L, X, Y, notation to be used in a broader context, where they need not signify anything about a Lagrangian.

Two-person zero-sum games: Consider *any* function $L : X \times Y \to \mathbb{R}$ for *any* nonempty sets X and Y, not necessarily even in $\mathbb{R}^n$ and $\mathbb{R}^m$. There's an associated game, played by two agents, called Player 1 and Player 2. Player 1 selects some $x \in X$, while Player 2 selects some $y \in Y$. The choices are revealed simultaneously. Then Player 1 must pay $\$L(x, y)$ to Player 2—that's all there is to it.

Direction of payment: Because no restriction has been placed on the sign of $L(x, y)$, the game is *not* necessarily loaded against Player 1 in favor of Player 2. The payment of a negative amount $\$L(x, y)$ from Player 1 to Player 2 is code for money actually flowing in the opposite direction.

Terminology: In general, X and Y are called the "strategy sets" for Players 1 and 2, while L is the "payoff function."

Generality of the game concept: This abstract model of a game may appear too special to be worthy of the name, although it does clearly furnish a model of conflict of interest between two "agents." Yet appearances can be deceiving: games such as *chess* and even *poker* can in principle be covered. We won't be concerned with those games here, but it's worth sketching them into the picture anyway along with an example where, on the other hand, the game concept is useful artificially as a way of representing uncertainty in decision making.

Chess as a two-person, zero-sum game. In chess, each element x of the set X at the disposal of Player 1, with the white pieces, is a particular policy which specifies encyclopedically what Player 1 should do within all possible circumstances that might arise. For instance, just one tiny part of such a policy x would be a prescription like "if after 17 moves the arrangement of the pieces on the board is such-and-such, and the history of play that brought it about is such-and-such, then such-and-such move should be made." The elements $y \in Y$ have similar meaning as policies for Player 2, with the black pieces.

Clearly, in choosing such policies independently the two players are merely deciding in advance how they will respond to whatever may unfold as the chess game is played. From (x, y) the outcome is unambiguously determined: checkmate for Player 1, checkmate for Player 2, or a draw. Define $L(x, y)$ to be -1, 1 or 0 according to these three cases. The chess game is represented then by X, Y, and L. It's *not* claimed that this is a *practical* representation, because the sets X and Y are impossibly large, but conceptually it's correct.

Poker as a two-person zero-sum game. The game of poker can be handled in similar fashion. It's payoffs are probabilistic "expectations," since they depend to a certain extent on random events beyond the players' control. Nonetheless, features of poker like bluffing can be captured in the model and evaluated for their effects. Poker for more than two players can be treated as an N-person game.

Games against nature: A useful approach to decision-making under uncertainty in many situations is to put yourself in the place of Player 1 in a game in which you choose an x but Player 2 is "nature," out to get you by choosing, as y, an unfavorable environment of circumstances (weather, interest rates, ..., over which you have no control but will effect the "cost" you will have to pay).

Saddle points: The term "saddle point" has various usages in mathematics, but in game theory and optimization it always refers to the following concept, which in the particular context of a two-person zero-sum game expresses a kind of solution to the conflict between the two players. A *saddle point* of a function L over a product set $X \times Y$ is a pair $(\bar{x}, \bar{y}) \in X \times Y$ such that

$$L(x, \bar{y}) \geq L(\bar{x}, \bar{y}) \geq L(\bar{x}, y) \text{ for all } x \in X \text{ and } y \in Y.$$

This means that the minimum of $L(x, \bar{y})$ with respect to $x \in X$ is attained at $\bar{x}$, whereas the maximum of $L(\bar{x}, y)$ with respect to $y \in Y$ is attained at $\bar{y}$.

Equilibrium interpretation: In the setting of a game with strategy sets X and Y and payoff function L, a saddle point $(\bar{x}, \bar{y})$ of L over $X \times Y$ captures a situation in which neither player has any incentive for deviating from the strategies $\bar{x}$ and $\bar{y}$. In selecting $\bar{x}$, Player 1 can guarantee that the amount paid to Player 2 won't exceed $\$L(\bar{x}, \bar{y})$, even if Player 2 were aware in advance that $\bar{x}$ would be chosen. This results from the fact that $L(\bar{x}, \bar{y}) \geq L(\bar{x}, y)$ for all $y \in Y$ (as half of the defining property of a "saddle point"). At the same time, in selecting $\bar{y}$, Player 2 can guarantee that the amount received from Player 1 won't fall short of $\$L(\bar{x}, \bar{y})$, regardless of whether Player 1 acts with knowledge of this choice or not. This is because $L(\bar{x}, \bar{y}) \leq L(x, \bar{y})$ for all $x \in X$.

Existence of saddle points: Without crucial assumptions being satisfied by L, X and Y, there might not be a saddle point. (Typically such assumptions include the convexity of X and Y and convexity/concavity properties of L, among other things. The existence of saddle points is studied in *minimax theory*, but we'll be looking at some special cases in the context of optimization theory.

Chess and poker: No one knows whether a saddle point exists for chess in the game formulation we've given. What's known is that if "randomized" policies involving probabilistic play are introduced in a certain way, then a saddle point *does* exist. Correspondingly, the payoffs are modified to reflect *chances* of winning.) Whether the equilibrium value $L(\bar{x}, \bar{y})$ in such a model is 0, testifying that chess is a fair game when randomized policies are admitted, is unknown. The theory of poker is in a similar state.

Gradient condition associated with a saddle point: In the important case of a game in which X and Y happen to be closed subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$ while L is a function of class $\mathcal{C}^1$, our work on constrained optimization gives some insights into a saddle point. In that case a first-order *necessary* condition for $(\bar{x}, \bar{y})$ to be a saddle point of L over $X \times Y$ is that

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

This follows from applying Theorem 9(a) first to the minimization of $L(x, \bar{y})$ in $x \in X$ and then to the maximization of $L(\bar{x}, y)$ in $y \in Y$. By Theorem 9(b) we see that, conversely, this gradient condition is *sufficient* for a saddle point in situations where X and Y are convex sets, $L(x, y)$ is convex in x for each $y \in Y$, and $L(x, y)$ is concave in y for each $x \in X$.

Preview about Lagrangians: This gradient condition turns into the Lagrangian form of the Kuhn-Tucker conditions when L is the Lagrangian associated with an optimization problem $(\mathcal{P})$ in standard format. This is the channel down which we are eventually headed.

Broader game models: In an *N-person game*, each Player k for $k = 1, \ldots, N$ selects an element $x_k \in X_k$, and the choices are revealed simultaneously. Then each Player k must pay the (positive, zero, or negative) amount $L_k(x_1, \ldots, x_N)$—to the "great accountant in the sky," or whatever. The game is *zero-sum* if

$$\sum_{k=1}^{N} L_k(x_1, \ldots, x_N) = 0,$$

or in other words the total of positive payments exactly balances the total of negative payments (in which case one can think of the positive payments as going into a pot and then being restributed as the negative payments). In the two-person, zero-sum case one has $L_1(x_1, x_2) = -L_2(x_1, x_2) = L(x, y)$.

For games that aren't zero-sum, no actual exchange of anything needs to be envisioned, and there is no reason to focus on a medium of exchange, such as money.

The units in which the values of L_k are measured can be personal to Player k and different from those for the other players. This makes it possible to use game theory in the modeling of social and economic situations that don't necessarily reduce to competition alone. For instance, *cooperative* games can be studied, where everyone can win if the actions are properly coordinated. Theory must then address the mechanisms of achieving coordination among players.

The central feature of any N-person game is that the consequences for Player k of choosing x_k depend not only on x_k itself, over which Player k has control, but also on the decisions made by the other players.

Equilibrium idea: A basic concept of "solution" in the theory of N-person games is that of *Nash equilibrium*. This refers to having $(\bar{x}_1, \ldots, \bar{x}_N) \in X_1 \times \cdots \times X_N$ such that Player 1 faces

$$L_1(x_1, \bar{x}_2, \ldots, \bar{x}_N) \geq L_1(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_N) \text{ for all } x_1 \in X_1,$$

while Player 2 faces

$$L_2(\bar{x}_1, x_2, \ldots, \bar{x}_N) \geq L_2(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_N) \text{ for all } x_2 \in X_2,$$

and similarly for Player 3 to Player N (when present). As with a saddle point, the interpretation is that no single player would have incentive for *unilaterally* making a different decision. For a two-person zero-sum game, Nash equilibrium reduces to a saddle point. Other interesting concepts of equilibrium have also been explored theoretically, for instance ones involving the formation of "coalitions" among the players, reinforced perhaps by side payments to keep members in line, with these side payments coming out of the total proceeds of the coalition.

Game theory applications: The greatest interest in developing game theory has come from economists, who have used it to study markets and competition. In fact, the subject was invented expressly for that purpose by John von Neumann, one of the most remarkable mathematicians of the 20th century.

Optimization problems derived from a game: Associated with any two-person, zero-sum game specified by a general choice of X, Y and L, there are two complementary problems of optimization. The study of these problems provides further insight into the role of saddle points and, eventually, the ways that Lagrange multiplier vectors can be determined, interpreted and utilized. The problems in question result from adopting a very conservative approach to playing the game, amounting to little more than a worst-case analysis. Whether or not one would really be content with playing the game in this manner in practice, the approach does lead to impressive mathematical results.

Strategy optimizing problem for Player 1: To determine $\bar{x}$, Player 1 should solve

$$(\mathcal{P}_1) \qquad \text{minimize } f(x) \text{ over all } x \in X, \text{ where } f(x) := \sup_{y \in Y} L(x, y).$$

In other words, for each $x \in X$, Player 1 should look at the value $f(x)$, which indicates the worst that could possibly happen if x were the element selected (the measure being in terms of how high a payment might have to be made to Player 2, in the absence of any way to predict what Player 2 will actually do). The choice of $\bar{x}$ should be made to ameliorate this as far as possible.

Strategy optimizing problem for Player 2: To determine $\bar{y}$, Player 2 should solve

$$(\mathcal{P}_2) \qquad \text{maximize } g(y) \text{ over all } y \in Y, \text{ where } g(y) := \inf_{x \in X} L(x, y).$$

In other words, for each $y \in Y$, Player 2 should look at the value $g(y)$, which indicates the worst that could possibly happen if y were the element selected (the measure being in terms of how low a payment might be forthcoming from Player 1, in the absence of any way to predict what Player 1 will actually do). The choice of $\bar{y}$ should be made to ameliorate this as far as possible.

Fundamental relation in optimal values: Let v_1 denote the optimal value in Player 1's min problem and v_2 the optimal value in Player 2's max problem. Then

$$v_1 \geq v_2.$$

This is evident on intuitive grounds, even without assurance that optimal solutions exist to the two problems. Namely, for any $\alpha > v_1$, Player 1 can choose $x \in X$ with $f(x) \leq \alpha$ and thereby be certain of not having to pay more than the amount α. On the other hand, for any $\beta < v_2$, Player 2 can choose $y \in Y$ with $g(y) \geq \beta$ and be certain of getting at least the amount β from Player 1.

Seen more fully, the definition of f gives $f(x) \geq L(x,y)$ for any $y \in Y$, while the definition of g gives $g(y) \leq L(x,y)$ for any $x \in X$, so that

$$f(x) \geq L(x,y) \geq g(y) \qquad \text{for all } x \in X,\ y \in Y.$$

Each value $g(y)$ for $y \in Y$ thus provides a lower bound to the values of the function f on X, so v_2 likewise provides a lower bound, the best that can be deduced from the various values $g(y)$. This can't be more than the greatest lower bound for f on X, which by definition is v_1.

The value of a game: It's possible that $v_1 > v_2$, but if these optimal values in the strategy problems for the two players are equal, then the common amount $v = v_1 = v_2$ is said to be the game's *value*. It quantifies the degree in which the game is biased (toward Player 2 if positive, and toward Player 1 if negative) and can be interpreted as how much Player 2 should be willing to pay, and how much Player 1 should ask to be paid, for participating in the game.

Fairness: A game is called *fair* if its value v exists by this definition and equals 0.

Minimax issues: The question of whether $v_1 = v_2$ can be viewed in terms of the extent to which the order of operations of minimization and maximization (in separate arguments) can be interchanged: by definition we have

$$v_1 = \inf_{x \in X} \sup_{y \in Y} L(x,y), \qquad v_2 = \sup_{y \in Y} \inf_{x \in X} L(x,y).$$

Because "inf" and "sup" can often be replaced by "min" and "max," results giving conditions under which $v_1 = v_2$ are generally called *minimax theorems*. Although special assumptions are needed to guarantee the validity of such an exchange, some cases of which will emerge as we proceed, the fact that $v_1 \geq v_2$, noted above, reveals that one always does at least have "inf sup $\geq$ sup inf."

THEOREM 15 (basic characterization of saddle points). *In any two-person, zero-sum game, the following conditions on a pair $(\bar{x}, \bar{y})$ are equivalent to each other and ensure that $L(\bar{x}, \bar{y})$ is the optimal value in both players' problems.*

(a) *$(\bar{x}, \bar{y})$ is a saddle point of $L(x,y)$ on $X \times Y$.*

(b) *$\bar{x}$ is an optimal solution to problem $(\mathcal{P}_1)$, $\bar{y}$ is an optimal solution to problem $(\mathcal{P}_2)$, and the optimal values in these two problems agree.*

(c) *$\bar{x} \in X$, $\bar{y} \in Y$, and the objective values $f(\bar{x})$ in $(\mathcal{P}_1)$ and $g(\bar{y})$ in $(\mathcal{P}_2)$ agree.*

Proof. The equivalence is obvious from the general inequality $f(x) \geq L(x,y) \geq g(y)$ just noted above and the fact that the saddle point condition can, by its definition and that of f and g, be written as $f(\bar{x}) = L(\bar{x}, \bar{y}) = g(\bar{y})$. □

Saddle point sets as product sets: This theorem reveals that the components $\bar{x}$ and $\bar{y}$ of a saddle point $(\bar{x}, \bar{y})$ have a sort of independent character. They can be obtained by solving one optimization to get $\bar{x}$ and another to get $\bar{y}$. The set of saddle points is therefore always a *product set* within $X \times Y$. If (x, y) and (x', y') are saddle points, then so too are (x', y) and (x, y').

Application of game theory to Lagrangians: Returning to the case of an optimization problem $(\mathcal{P})$ in standard format and its associated Lagrangian L on $X \times Y$, we ask a question that at first could seem frivolous. If we think of these elements L, X, and Y as specifying a certain two-person, zero-sum game, what would we get? As a matter of fact, in this game the strategy problem $(\mathcal{P}_1)$ for Player 1 turns out to be $(\mathcal{P})$! Indeed, the formula for the function that is to be minimized in $(\mathcal{P}_1)$,

$$f(x) := \sup_{y \in Y} L(x, y) = \sup_{y \in Y} \Big\{ f_0(x) + \sum\nolimits_{i=1}^{m} y_i f_i(x) \Big\},$$

gives none other than the *essential* objective function in $(\mathcal{P})$: in terms of the feasible set $C = \big\{ x \in X \,\big|\, f_i(x) \le 0 \text{ for } i \in [1, s],\ \ f_i(x) = 0 \text{ for } i \in [s+1, m] \big\}$, we have

$$C = \big\{ x \in X \,\big|\, f(x) < \infty \big\}, \qquad f(x) = \begin{cases} f_0(x) & \text{if } x \in C, \\ \infty & \text{if } x \notin C. \end{cases}$$

Thus, in tackling problem $(\mathcal{P})$ as our own, we are tacitly taking on the role of Player 1 in a certain game. Little did we suspect that our innocent project would necessarily involve us with an opponent, a certain Player 2!

This will take time to digest, but a place to begin is with our knowledge, through game theory, that the strategy problem $(\mathcal{P}_2)$ that Player 2 is supposed to want to solve is the one in which the function $g(y) := \inf_{x \in X} L(x, y)$ is maximized over $y \in Y$. We define this to be the problem "dual" to $(\mathcal{P})$ in the Lagrangian framework.

Lagrangian dual problem: The *dual problem* of optimization associated with $(\mathcal{P})$ (the latter being called the *primal problem* for contrast) is

$$(\mathcal{D}) \qquad \begin{array}{l} \text{maximize } g(y) \text{ over all } y \in Y, \text{ where} \\ \qquad g(y) := \displaystyle\inf_{x \in X} L(x, y) = \inf_{x \in X} \Big\{ f_0(x) + \sum\nolimits_{i=1}^{m} y_i f_i(x) \Big\}. \end{array}$$

Here g has the general status that f did in $(\mathcal{P})$ as the *essential* objective function in $(\mathcal{D})$, because $g(y)$ might be $-\infty$ for some choices of $y \in Y$. The feasible set in problem $(\mathcal{D})$ therefore isn't actually the set Y, but the set D that we define by

$$D = \big\{ y \in Y \,\big|\, g(y) > -\infty \big\}.$$

In other words, a vector $y \in Y$ isn't regarded as a feasible solution to $(\mathcal{D})$ unless $g(y)$ is finite for this y. Of course, a closer description of D and g can't emerge until more information about X and the functions f_i has been supplied in a given case, so that the calculation of $g(y)$ can be carried out further.

Example: Lagrangian derivation of linear programming duality: For a linear programming problem $(\mathcal{P}_{\text{lin}})$ as considered earlier, the Lagrangian is

$$\begin{aligned} L(x,y) &= \sum\nolimits_{j=1}^{n} c_j x_j + \sum\nolimits_{i=1}^{m} y_i \Big(b_i - \sum\nolimits_{j=1}^{n} a_{ij} x_j \Big) \\ &= \sum\nolimits_{i=1}^{m} b_i y_i + \sum\nolimits_{j=1}^{n} x_j \Big(c_j - \sum\nolimits_{i=1}^{m} y_i a_{ij} \Big) \text{ on } X \times Y = \mathbb{R}^n_+ \times \mathbb{R}^m_+. \end{aligned}$$

Using the second version of the formula to calculate the essential objective in $(\mathcal{D})$, we get for arbitrary $y \in \mathbb{R}^m_+$ that

$$\begin{aligned} g(y) = \inf_{x \in \boldsymbol{R}^n_+} L(x,y) &= \inf_{\substack{x_j \ge 0 \\ j=1,\dots,n}} \Big\{ \sum\nolimits_{i=1}^{m} b_i y_i + \sum\nolimits_{j=1}^{n} x_j \Big(c_j - \sum\nolimits_{i=1}^{m} y_i a_{ij} \Big) \Big\} \\ &= \begin{cases} \sum_{i=1}^{m} b_i y_i & \text{when } c_j - \sum_{i=1}^{m} y_i a_{ij} \ge 0 \text{ for } j = 1, \dots, n, \\ -\infty & \text{otherwise.} \end{cases} \end{aligned}$$

The Lagrangian dual problem $(\mathcal{D})$, where $g(y)$ is maximized over $y \in Y$, comes out therefore as the previously identified dual problem $(\mathcal{D}_{\text{lin}})$.

Convexity properties of the dual problem: Regardless of whether any more detailed expression is available for the feasible set D and objective function g in problem $(\mathcal{D})$, it's always true that D is a *convex* set with respect to which g is *concave*. This problem, therefore, falls within the realm of optimization problems of convex type.

Argument: Consider any points y_0 and y_1 in D along with any $\tau \in (0,1)$. Let $y_\tau = (1-\tau)y_0 + \tau y_1$. From the definition of D and g we have for each $x \in X$ that $-\infty < g(y_0) \le L(x,y_0)$ and $-\infty < g(y_1) \le L(x,y_1)$, hence

$$-\infty \;<\; (1-\tau)g(y_0) + \tau g(y_1) \;\le\; (1-\tau)L(x,y_0) + \tau L(x,y_1) = L(x,y_\tau),$$

where the equation is valid because $L(x,y)$ is affine with respect to y. Since this holds for arbitrary $x \in X$, while $g(y_\tau) = \inf_{x \in X} L(x,y_\tau)$, we obtain the concavity inequality $(1-\tau)g(y_0) + \tau g(y_1) \le g(y_\tau)$ along with the guarantee that $y_\tau \in D$.

Basic relationship between the primal and dual problems: A number of facts about $(\mathcal{P})$ and $(\mathcal{D})$ follow at once from game theory, without any need for additional assumptions. It's always true that

$$\big[\text{optimal value in } (\mathcal{P})\,\big] \;\ge\; \big[\text{optimal value in } (\mathcal{D})\,\big].$$

When these two optimal values coincide, as they must by Theorem 15 if a saddle point exists for the Lagrangian L on $X \times Y$, the saddle points are the pairs $(\bar{x}, \bar{y})$ such that $\bar{x}$ solves $(\mathcal{P})$ while $\bar{y}$ solves $(\mathcal{D})$. The sticking point, however, is whether a saddle point does exist. For that we need to draw once more on convexity.

THEOREM 16 (duality in convex programming). *Consider an optimization problem $(\mathcal{P})$ in standard format along with its Lagrangian dual $(\mathcal{D})$ in the case where the set X is closed and convex, the functions f_i are $\mathcal{C}^1$ and convex for $i = 0, 1, \ldots, s$, and affine for $i = s+1, \ldots, m$. Then the Lagrangian $L(x, y)$ is convex in x for each $y \in Y$ as well as affine in y for each $x \in X$, and the following properties are equivalent:*

(a) *$\bar{x}$ is an optimal solution to $(\mathcal{P})$, and $\bar{y}$ is a Lagrange multiplier vector associated with $\bar{x}$ by the Kuhn-Tucker conditions for $(\mathcal{P})$;*

(b) *$(\bar{x}, \bar{y})$ is a saddle point of the Lagrangian $L(x, y)$ over $X \times Y$;*

(c) *$\bar{x}$ and $\bar{y}$ are optimal solutions to $(\mathcal{P})$ and $(\mathcal{D})$, respectively, and*

$$\big[\text{optimal value in } \big(\mathcal{P}\big)\ \big] \;=\; \big[\text{optimal value in } \big(\mathcal{D}\big)\ \big].$$

In particular, therefore, this equation must hold if $(\mathcal{P})$ has an optimal solution for which the Kuhn-Tucker conditions are fulfilled.

Note: In connection with the final assertion, remember that the Kuhn-Tucker conditions are sure to be fulfilled at an optimal solution $\bar{x}$ to $(\mathcal{P})$ if the standard constraint qualification $(\star)$ in Theorem 10, or one of its substitutes $(\star\star)$ or $(\star\star\star)$ in Theorem 12, is satisfied. Moreover, $(\star\star\star)$ doesn't require specific knowledge about $\bar{x}$.

Proof. Since $L(x, y) = f_0(x) + y_1 f_1(x) + \cdots + y_m f_m(x)$, we always have $L(x, y)$ affine in y for fixed x. Because $Y = \mathbb{R}^s_+ \times \mathbb{R}^{m-s}$, the vectors $y \in Y$ have components y_i that are nonnegative for $i \in [1, s]$, and the convex programming assumptions on $(\mathcal{P})$ therefore ensure that $L(x, y)$ is convex in x when $y \in Y$. Saddle points $(\bar{x}, \bar{y})$ of L over the convex product set $X \times Y$ are characterized therefore by the gradient relations $-\nabla_x L(\bar{x}, \bar{y}) \in N_X(\bar{x})$ and $\nabla_y L(\bar{x}, \bar{y}) \in N_Y(\bar{y})$ as noted in our discussion of saddle points. But these relations give the Lagrangian form of the Kuhn-Tucker conditions for $(\mathcal{P})$. We conclude from this that (a) and (b) are equivalent. The equivalence of (b) and (c) is based on the general principles of Theorem 15, which is applicable because $(\mathcal{P})$ and $(\mathcal{D})$ have been identified as the strategy problems for the two players in the game specified by the Lagrangian triplet L, X, Y. □

Comparison with duality in linear programming: Convex programming covers linear programming as a special case, but the results in Theorem 16 are generally not as sharp or symmetric as those in Theorem 14. Nonetheless they do offer something new even for $(\mathcal{P}_{\text{lin}})$ and $(\mathcal{D}_{\text{lin}})$. We now know that the pairs $(\bar{x}, \bar{y})$ constituting optimal solutions to those linear programming problems are precisely the saddle points over $X \times Y = \mathbb{R}^n_+ \times \mathbb{R}^m_+$ of the function

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

Understanding the meaning of duality through perturbations: Optimal solutions to the dual problem $(\mathcal{D})$ are explained by Theorem 16, in the case of a convex programming problem $(\mathcal{P})$ that has an optimal solution $\bar{x}$ at which the Kuhn-Tucker conditions hold, as the Lagrange multiplier vectors $\bar{y}$ that appear in such conditions. In the background there must somehow also be an interpretation of such vectors $\bar{y}$ in the context of a "game," since after all that's how the primal/dual setup was derived. We'll come to such interpretations soon, but first it will be helpful, on a deeper level, to develop a connection between dual optimal solutions $\bar{y}$ and the effects of certain kinds of "perturbations" of problem $(\mathcal{P})$.

For this purpose we don't need to assume that $(\mathcal{P})$ is a convex programming problem and can instead proceed very generally, looking for exactly the stage where convexity might have to be brought in. Taking $(\mathcal{P})$ to be any optimization problem in standard format, with constraint functions $f_1, \ldots, f_m$ as usual, let's consider for each vector $u = (u_1, \ldots, u_m)$ the problem

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

The given problem corresponds to $u = 0$, i.e., $u_i = 0$ for all i; we have $(\mathcal{P}(0)) = (\mathcal{P})$. Our attention will be focused on the behavior, around $u = 0$, of the special function

$$p(u) := \ \big[\text{ optimal value in } (P(u)) \ \big].$$

The aim is achieve some kind of quantitative understanding of how $p(u)$ shifts from $p(0)$, the optimal value in the given problem, as u is "perturbed" away from 0.

Connection with constants in the constraints: When $f_i(x) = b_i - g_i(x)$, so that the constraints have the form $g_i(x) \geq b_i$ for $i \in [1, s]$ but $g_i(x) = b_i$ for $i \in [s+1, m]$, the replacement of f_i by $f_i + u_i$ in $(\mathcal{P}(u))$ corresponds to the perturbation of b_i to $b_i + u_i$. If instead $f_i(x) = h_i(x) - c_i$, we are perturbing c_i to $c_i - u_i$.

THEOREM 17 (role of dual optimal solutions in perturbations)**.** *A vector $\bar{y}$ furnishes for the optimal value function p the inequality*

$$p(u) \geq p(0) + \bar{y}\cdot u \quad \text{for all } u \in I\!R^m, \text{ with } p(0) \text{ finite,}$$

if and only if $\bar{y}$ is an optimal solution to $(\mathcal{D})$ and the optimal value in $(\mathcal{D})$ agrees with the optimal value in $(\mathcal{P})$, these values being finite. In particular, therefore, in situations where it can be guaranteed that the two optimal values are equal (such as in Theorem 16), this inequality completely characterizes the optimal solution set in $(\mathcal{D})$.

Proof. The targeted inequality can be identified with the condition that the minimum of $p(u) - \bar{y}\cdot u$ over all $u \in I\!R^m$ is attained when $u = 0$. To get a handle on this, let $C(u)$ denote the feasible set in problem $(\mathcal{P}(u))$ and let

$$h(x,u) = \begin{cases} f_0(x) - \bar{y}\cdot u & \text{if } x \in C(u), \\ \infty & \text{if } x \notin C(u). \end{cases}$$

Define the function ψ on $I\!R^m$ by $\psi(u) = \inf_x h(x,u)$ and the function φ on $I\!R^n$ by $\varphi(x) = \inf_u h(x,u)$. Then $\psi(u) = p(u) - \bar{u}\cdot y$ by the definition of $p(u)$, whereas on general grounds we have

$$\inf_u \psi(u) = \inf_{x,u} h(x,u) = \inf_x \varphi(x).$$

Our task reduces therefore to demonstrating that $\inf_x \varphi(x) = p(0)$ with $p(0)$ finite if and only if $\bar{y}$ is an optimal solution to $(\mathcal{D})$ and the optimal value in $(\mathcal{D})$ is finite and agrees with the optimal value in $(\mathcal{P})$. From the definition of $C(u)$ and the fact that $Y = I\!R^s_+ \times I\!R^{m-s}$ it's easy to see that

$$\varphi(x) = \begin{cases} L(x,\bar{y}) & \text{if } x \in X \text{ and } \bar{y} \in Y, \\ -\infty & \text{if } x \in X \text{ and } \bar{y} \notin Y, \\ \infty & \text{if } x \notin X \end{cases}$$

and consequently, through the definition of the dual objective function g in $(\mathcal{D})$, that

$$\inf_x \varphi(x) = \begin{cases} g(\bar{y}) & \text{if } \bar{y} \text{ is feasible in } (\mathcal{D}), \\ -\infty & \text{if } \bar{y} \text{ is not feasible in } (\mathcal{D}). \end{cases}$$

Thus $\inf_x \varphi(x) = p(0)$ with $p(0)$ finite if and only if $\bar{y}$ is a feasible solution to $(\mathcal{D})$ such that $g(\bar{y}) = p(0)$. Because the optimal value in $(\mathcal{D})$ can never be greater than the optimal value in $(\mathcal{P})$, which is $p(0)$, this yields the desired characterization of $\bar{y}$. □

The role of convexity: The inequality in Theorem 17 says that the affine function $l(u) = p(0) + \bar{y}\cdot u$ satisfies the conditions $l \leq p$, $l(0) = p(0)$ (finite). These conditions means geometrically that the half-space in $I\!R^{m+1}$ that's the epigraph of l "supports"

the epigraph of p at the point $(0, p(0))$ (not just in a local sense but in a global sense, because the inequality is supposed to hold for all u, not just for u in some neighborhood of 0). The equivalence in Theorem 17 tells us therefore that to say that $(\mathcal{D})$ has an optimal solution, and the optimal values in $(\mathcal{D})$ and $(\mathcal{P})$ are equal, is to assert the existence of such a "support."

When does such a "support" exist? Well, this is closely tied to the epigraph of p being a convex set—i.e,. p being a convex function—plus perhaps some other condition being satisfied. (We know that a convex set has a supporting half-space at any of its boundary points—and $(0, p(0))$ is certainly a boundary point of the epigraph of p—but to know that a supporting half-space is the epigraph of an affine function l requires knowing that it's not "vertical." Constraint qualifications, in particular the original Slater condition, can be seen as assumptions that make it possible to obviate such "verticality.")

But with that insight, what can we assume in order to make sure that the optimal value function p is convex? The answer is basically this: when $(\mathcal{P})$ is a *convex programming* problem as in Theorem 16, p is convex, but there's no other good test of convexity in terms of the data elements X, $f_0, f_1, \ldots, f_m$ that go into $(\mathcal{P})$.

Lack of duality in nonconvex programming: That's the sad verdict, that aside from the case of convex programming we really can't expect to have Lagrange multipliers correspond to saddle points or even to have the optimal values in $(\mathcal{P})$ and $(\mathcal{D})$ equal, except by lucky accident.

Lagrange multipliers as rates of change: If the function p happens somehow to be differentiable at $u = 0$, the inequality in Theorem 17 implies that the gradient $\nabla p(0)$ is $\bar{y}$; we have

$$\bar{y}_i = \frac{\partial p}{\partial u_i}(0) \text{ for } i = 1, \ldots, m.$$

Then $\bar{y}_i$ gives the rate of change of the optimal value in $(\mathcal{P})$ with respect to perturbations in the parameter u_i away from u_i=0.

More broadly, in convex programming, where p turns out always to be a convex function, the inequality in Theorem 17 is said to mean, by definition, that $\bar{y}$ is a *subgradient* of p at $u = 0$. It's known in that context that the differentiability of p at $u = 0$ corresponds to there being a *unique* subgradient (which is then the gradient). Hence for convex programming problems (with linear programming problems as a special case), p is differentiable at $u = 0$ if and only if the optimal values in $(\mathcal{P})$ and $(\mathcal{D})$ are finite and equal, and $(\mathcal{D})$ has a *unique* optimal solution $\bar{y}$. Such uniqueness is then the key to interpreting Lagrange multipliers $\bar{y}_i$ as partial derivatives.

Game-theoretic interpretation of duality: A puzzle remains. If the consideration of a problem $(\mathcal{P})$ leads inevitably to involvement in a game in which an opponent is trying to solve another problem $(\mathcal{D})$, just what is the game, and what does it signify? In principle, the game is of course as follows. Player 1, with whom we identify ourselves, chooses a vector $x \in X$ (ignoring all other constraints!) while the sinister Player 2 chooses a multiplier vector $y = (y_1, \ldots, y_m) \in Y$. Then Player 1 must pay the amount $L(x, y) = f_0(x) + y_1 f_1(x) + \ldots + y_m f_m(x)$ to Player 2. What can this mean? An important clue can usually be found in analyzing the units in which this payment is made and trying through that to interpret the quantity $L(x, y)$. We'll take this approach now in the context of an example in which the variables can be given an insightful interpretation.

An example with Lagrange multipliers as prices: Consider the following case of $(\mathcal{P})$ in which there are no equality constraints (i.e., $s = m$):

$$(\mathcal{P}_0) \qquad \begin{array}{l} \text{minimize } f_0(x) \text{ over all } x \in X \text{ satisfying} \\ \qquad f_i(x) = g_i(x) - c_i \leq 0 \quad \text{for } i = 1, \ldots, m, \end{array}$$

where the elements $x \in X$ represent decisions about the operation of, say, a manufacturing plant (e.g., how much to make of various products, how to employ the work force in various tasks, ...), and $f_0(x)$ is the overall cost in dollars that would result from decision x (with profit terms incorporated as negative costs). Interpret $g_i(x)$ as the amount of resource i that would be consumed by x, and c_i as the total amount of resource i that is at hand, with these amounts being measured in units appropriate for resource i (tons, hours, ...). The constraints $f_i(x) \leq 0$ mean that the only decisions $x \in X$ deemed feasible are those that operate without exceeding the available resource amounts. The Lagrangian comes out as

$$L(x, y) = f_0(x) + y_1[g_1(x) - c_1] + y_2[g_2(x) - c_2] + \cdots + y_m[g_m(x) - c_m]$$

with $y = (y_1, \ldots, y_m) \in Y = \mathbb{R}^m_+$. What does this stand for?

For the units of measurement to come out consistently, $L(x, y)$ must be in dollars, like $f_0(x)$, and the same for each of the terms $y_i[g_i(x) - c_i]$. Therefore, the units for y_i must be dollars per unit of resource i (thus \$/ton, \$/hour, ...). In other words, the coefficients y_i convert resource amounts into money. They serve somehow as *prices*. In the game framework, whereas Player 1 makes a decision $x \in X$ *without worrying about whether the resource constraints* $g_i(x) \leq c_i$ *are satisfied*, it seems now that Player 2 selects a vector $y = (y_1, \ldots, y_m) \geq 0$ of prices for the resources.

These choices lead to Player 1 paying Player 2 the amount of dollars specified by expression $L(x, y)$. For this there ought to be a market interpretation.

Let's imagine a market in which the resources $i = 1, \ldots, m$ can be bought or sold in arbitrary quantities at the prices y_i. If the choice x results in consuming more of a particular resource i than was at hand, so that $0 < g_i(x) - c_i$, the excess has to be purchased from the market, and the cost of that purchase with respect to the price y_i, namely $y_i[g_i(x) - c_i]$, is added to the basic cost $f_0(x)$ incurred by x. If, on the other hand, x results in consuming less of resource i than was at hand, so that there's an amount $c_i - g_i(x)$ left over, that surplus is sold off in the market and the proceeds from the sale, namely $y_i[c_i - g_i(x)]$, are subtracted from the basic cost $f_0(x)$ (or equivalently, the negative dollar amount $y_i[g_i(x) - c_i]$ is added).

We see then that $L(x, y)$ stands for the net cost to Player 1 after paying $f_0(x)$ and accounting for the purchases and sales of resources that result from choosing x, when the market prices are given by y.

The market as the opponent: Player 2 is seen now as the market itself, acting to keep Player 1 from making undue profit out of prices that don't reflect the true values of the resources. The market acts to set the prices so to get the highest net amount out of Player 1 that is consistent with the situation. Indeed, in the dual problem $(\mathcal{D}_0)$ of maximizing $g(y)$ over $y \in Y = \mathbb{R}^m_+$, the objective value

$$g(y) := \inf_{x \in X} \Big\{ f_0(x) + y_1[g_1(x) - c_1] + \cdots + y_m[g_m(x) - c_m] \Big\}$$

gives the dollar amount that the market is sure to receive when the price vector is y, regardless of Player 1's eventual decision.

Equilibrium prices: Under certain assumptions, including *convexity* and a bit more, our theory tells us there will be a saddle point $(\bar{x}, \bar{y})$. Then $\bar{x}$ and $\bar{y}$ will be optimal solutions to $(\mathcal{P}_0)$ and $(\mathcal{D}_0)$, and moreover the price vector $\bar{y}$ will have the property that

$$\text{the minimum of } L(x, \bar{y}) \text{ over } x \in X \text{ is achieved at } \bar{x}.$$

In other words, the prices $\bar{y}_i$ will induce Player 1, in the market framework that has been described, to respect the resource constraints $g_i(x) \leq c_i$ even though these constraints have been supplanted in that framework by the introduction of buying and selling of resources! To have such a powerful effect they must provide an evaluation of resources that's just right for balancing out any incentive for buying or selling. That information could be valuable in many ways.

Decentralization of decision making: Although the market interpretation of the Lagrangian in the preceding example may seem rather artificial, the revelation that Lagrange multipliers can be viewed as prices associated with constraints has far-reaching consequences in subjects ranging from economics to computation. As a bridge toward the computational end, let's look now at an extended version of the example in which $\bar{y}$ will emerge as a vehicle for reducing a single difficult problem to a collection of easy problems.

Consider a situation where a number of separate agents, indexed by $k = 1, \ldots, r$, would solve problems of optimization independent of each other if it weren't for the necessity of sharing certain resources. Agent k would prefer just to minimize $f_{0k}(x_k)$ over all x_k in a set $X_k \subset \mathbb{R}^{n_k}$, but there are mutual constraints

$$f_{i1}(x_1) + \cdots + f_{ir}(x_r) \leq c_i \text{ for } i = 1, \ldots, m$$

requiring some coordination. Furthermore, there's community interest in having the coordination take place in such a way that the overall sum $f_{01}(x_1) + \cdots + f_{0r}(x_r)$ is kept low. In other words, the problem as seen from the community is to

$$\begin{aligned} \text{minimize } f_0(x) &:= f_{01}(x_1) + \cdots + f_{0r}(x_r) \text{ subject to} \\ f_i(x) &:= f_{i1}(x_1) + \cdots + f_{ir}(x_r) - c_i \leq 0 \text{ for } i = 1, \ldots, m, \\ x &:= (x_1, \ldots, x_r) \in X := X_1 \times \cdots \times X_r. \end{aligned}$$

This obviously fits as a special case of problem $(\mathcal{P}_0)$ above in which we have additional structure relative to vector components x_k of x and take

$$g_i(x) = f_{i1}(x_1) + \cdots + f_{ir}(x_r) \text{ for } i = 1, \ldots, m.$$

In these circumstances a central authority could solve the overall community problem and tell the different agents what they should do. But is there a way instead to take advantage of the energies and interests of these agents?

Decentralization through resource allocation: In one approach, a "coordinator" would allocate the available resource amounts c_i to the agents k in quantities c_{ik} adding up to c_i and leave the individual agents solve separate problems of the form

$$\begin{aligned} &\text{minimize } f_{0k}(x_k) \text{ over all } x_k \in X_k \text{ satisfying} \\ &f_{ik}(x_k) - c_{ik} \leq 0 \text{ for } i = 1, \ldots, m \end{aligned}$$

for $k = 1, \ldots, r$. Of course, that would raise the issue of how to determine a resource allocation under which the resulting decisions $\bar{x}_k$ of the individual

agents, when put together as $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_r)$, would really furnish an optimal solution to the community problem.

Decentralization through pricing: In another approach, Lagrange multipliers are featured. For it to work most effectively, let's put ourselves back in the picture of convex programming—where the sets X_k are convex and the functions f_{ik} are convex. Take the objective terms $f_{0k}(x_k)$ even to be strictly convex; then the Lagrangian $L(x, y)$ is strictly convex in x for each $y \in Y = \mathbb{R}^m_+$. Assume further that we have a case where a saddle point $(\bar{x}, \bar{y})$ of L on $X \times Y$ is sure to exist (as provided by Theorem 16 etc.). Then the subproblem

$$\text{minimize } L(x, \bar{y}) \text{ over all } x \in X$$

has *a unique optimal solution, which must equal* $\bar{x}$ *and thus solve the community problem.* This follows from the strict convexity. Observe next that

$$\begin{aligned} L(x, y) :&= \sum\nolimits_{k=1}^{r} f_{0k}(x_k) + \sum\nolimits_{i=1}^{m} y_i \Big(\sum\nolimits_{k=1}^{r} f_{ik}(x_k) - c_i \Big) \\ &= \sum\nolimits_{k=1}^{r} L_k(x_k, y) - y \cdot c, \quad \text{where} \\ L_k(x_k, y) &:= f_{0k}(x_k) + \sum\nolimits_{i=1}^{m} y_i f_{ik}(x_k), \quad c = (c_1, \ldots, c_m). \end{aligned}$$

Minimizing $L(x, \bar{y})$ over $x \in X$ is the same as minimizing $L_1(x_1, \bar{y}) + \cdots + L_r(x_r, \bar{y})$ over all $(x_1, \ldots, x_r) \in X_1 \times \cdots \times X_r$. But from this way of posing the subproblem it's clear that there is no interaction between the different components x_k: the minimum is achieved by minimizing each term $L_k(x_k, \bar{y})$ separately with respect to $x_k \in X_k$. Thus, by letting the agents k solve the separate problems

$$\text{minimize } f_{0k}(x_k) + \sum\nolimits_{i=1}^{m} \bar{y}_i f_{ik}(x_k) \text{ over all } x_k \in X_k$$

for $k = 1, \ldots, r$ to get vectors $\bar{x}_k$, and then putting these together as $\bar{x}$, the community is able to arrive at an optimal solution to its problem without having to dictate actions or resort to an allocation of the resources!

There's still the question of how to come up with a vector $\bar{y}$ that will do the job, but this time we immediately know the answer, at least in principle. Determining $\bar{y}$ is solving the associated dual problem. That is what now becomes the task of a coordinator, which could then report the prices $\bar{y}_i$ to the agents k and let them carry on. This time, however, the coordinator might be envisioned as a *market* mechanism rather than a community "authority."

Economic interpretation: In the price model, each agent k solves a problem in which there are no constraints on shared resources, only the condition that $x_k \in X_k$ (where X_k incorporates merely the "local" circumstances faced by agent k). The interpretation again is that resources can be purchased at the specified prices. Agent k, in choosing x_k and thereby consuming resource amounts $f_{ik}(x_k)$ for $i = 1, \ldots, m$, must pay the basic cost $f_{0k}(x_k)$ plus the purchase costs $\bar{y}_k f_{ik}(x_k)$. The prices $\bar{y}_k$ are able to do all the coordinating. The individual agents optimize in their own *self-interest*, according to the costs they perceive, and yet, as if by a miracle, their separate actions result in a vector $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_r)$ that solves the *community* problem.

This is the kernel of a famous principle in economics that goes back centuries to Adam Smith and underlies most discussions of "free markets." More complicated versions have been developed in terms of N-person games, where the players are producers, consumers and other economic agents.

Abuses of the notion of price decentralization: Free-market enthusiasts have captured the world's attention with the idea that economic behavior that's optimal from society's interest can be induced simply by letting markets set the proper prices. For instance, if every form of pollution has its price, and the price is right, companies will keep pollution within society's desired bounds just by acting out of their own self interest. But the underpinnings of such claims rest on mathematical models in which *convexity* properties are essential. Roughly speaking, such models can be valid in the setting of an economy built on a huge mass of infinitesimal agents (family farmers was the classical example), but not a modern economy with only a few big agents in each industry.

Lagrangian relaxation: These economic examples bring out the importance of studying, in connection with an optimization problem $(\mathcal{P})$ in standard format, subproblems of the form

$$(\mathcal{P}^y) \qquad \text{minimize } f_0(x) + y_1 f_1(x) + \cdots + y_m f_m(x) = L(x, y) \text{ over } x \in X$$

for various choices of $y \in Y$. Such a subproblem $(\mathcal{P}^y)$ is called a *Lagrangian relaxation* of $(\mathcal{P})$, because it "relaxes" the constraints $f_i(x) \leq 0$ or $f_i(x) = 0$ by incorporating them into the objective function in a manner that we have come to understand as a kind of "pricing." After such modification of the objective the minimization is carried out subject only to the requirement that $x \in X$.

In the examples, we have been concentrating on the consequences of the existence of a saddle point $(\bar{x}, \bar{y})$. The multiplier vector $\bar{y}$ then gives us a relaxed problem $(\mathcal{P}^{\bar{y}})$

that has among its optimal solutions (if it has more than one) the optimal solution $\bar{x}$ to $(\mathcal{P})$. (In general it is known that the optimal solutions to $(\mathcal{P})$ are the optimal solutions to $(\mathcal{P}^{\bar{y}})$ that are feasible solutions to $(\mathcal{P})$; this follows from additional analysis of the saddle point condition.) But there are other interesting aspects of Lagrangian relaxation besides this.

Lower bounds on the optimal value in the primal problem: Recall that the optimal value in $(\mathcal{P})$ can never be less than the optimal value in $(\mathcal{D})$, where $g(y)$ is maximized over $y \in Y$. Furthermore, $g(y)$ is by its definition the optimal value in the problem of minimizing $L(x, y)$ over $x \in X$, which is $(\mathcal{P}^y)$. We can summarize this as follows:

$$[\text{ optimal value in } (\mathcal{P}) \text{ }] \;\geq\; [\text{ optimal value in } (\mathcal{P}^y) \text{ }] \quad \text{for any } y \in Y,$$

$$\sup_{y \in Y} [\text{ optimal value in } (\mathcal{P}^y) \text{ }] \;=\; [\text{ optimal value in } (\mathcal{D}) \text{ }].$$

Thus, by choosing a $y \in Y$ and solving the relaxed problem $(\mathcal{P}^y)$ we always get a *lower bound* to the optimal value in $(\mathcal{P})$ and the greatest lower bound that can be derived this way is the optimal value in $(\mathcal{D})$. These assertions don't depend on convexity, but unless $(\mathcal{P})$ is a convex programming problem there is little likelihood that the greatest lower bound will actually equal the optimal value in $(\mathcal{P})$—there could be a discrepancy, called a *duality gap*. Anyway, regardless of the presence of a duality gap, such lower bounds can be useful in methodology for solving $(\mathcal{P})$, for instance in indicating when computations can be halted because not very much more can be gained by continuing with them.

Solving the dual problem: Lagrangian relaxation is typically built into procedures involving a whole sequence of multiplier vectors $y^\nu \in Y$ with the aim of getting this sequence $\{y^\nu\}$ to be an optimizing sequence for the dual problem $(\mathcal{D})$. In each iteration ν, the relaxed problem $(\mathcal{P}^{y^\nu})$ is solved to obtain a vector x^ν. The information generated out of that process, along with the information generated in previous iterations perhaps, is used to "update" y^ν to a new multiplier vector $y^{\nu+1}$, and the step then is repeated. In some situations where $(\mathcal{P})$ is a convex programming problem, the auxiliary sequence $\{x^\nu\}$, or one readily constructed from it, can turn out to be asymptotically optimizing in $(\mathcal{P})$. The details of this subject are too much to go into here, however.

Decomposition of large-scale problems: When problems of optimization are very large in dimension—with thousands of variables and possibly a great many constraints to go with them—there's strong interest in devising methods of computation

able to take advantage of special structure that might be present. An attractive idea is that of breaking a problem $(\mathcal{P})$ down into much smaller subproblems, to be solved independently, maybe by processing them in parallel.

Of course, there's no way of achieving such a decomposition once and for all, because that would presuppose knowing unrealistically much about the problem before it's even been tackled. Rather, one has to envision schemes in which an initial decomposition of $(\mathcal{P})$ into subproblems yields information leading to a better decomposition into modified subproblems, and so forth iteratively in a manner that can be justified as eventually generating an asymptotically optimal sequence $\{x^\nu\}_{\nu \in N}$. As long as $(\mathcal{P})$ is well posed, an asymptotically optimal sequence can be deemed to solve it in the sense of Theorem 2.

Dantzig-Wolfe decomposition: A prime case where decomposition can be effective has already been encountered in our example of decentralization through pricing. When $(\mathcal{P})$ has *separable* structure, which is the name for the breakdown of X and the functions f_i in that example into separate terms for components x_k of x, then each step in a Lagrangian relaxation procedure involving a sequence of multiplier vectors y^ν comes down to solving a collection of relaxed subproblems

$$\text{minimize } f_{0k}(x_k) + \sum\nolimits_{i=1}^{m} y_i^\nu f_{ik}(x_k) \text{ over } x_k \in X_k$$

to get x_k^ν and putting these elements together to get x^ν. George Dantzig and Philip Wolfe hit on this idea originally in a context of linear programming and expressed it in terms of the "simplex method" for solving linear programming problems, but it has since come to be understood far more generally.

Frank-Wolfe decomposition: For problems of minimizing a nonlinear, differentiable convex function f_0 over a set C specified by linear constraints, a possible way to generate a sequence of feasible solutions x^ν from an initial point $x^0 \in C$ is as follows. Having arrived at x^ν, form the linearized function

$$l^\nu(x) := f_0(x^\nu) + \nabla f_0(x^\nu) \cdot [x - x^\nu]$$

and minimize l^ν over C to get a point $\hat{x}^\nu$. The information provided by this subproblem can be used in various ways. For instance, $d^\nu = \hat{x}^\nu - x^\nu$ turns out to be a descent vector for f_0, moreover one giving a feasible direction into C at x^ν (unless x^ν itself already minimizes l^ν over C, in which case x^ν already has to be optimal). We won't go into this here. The main thing is that the subproblem of minimizing l^ν over C is one of linear programming, in contrast to

that of minimizing f_0 over C. In special situations, such as when C is a box or a even just a product of polyhedral sets of low dimension, it breaks down into still smaller subproblems solvable in closed form or in parallel.

Benders decomposition: This term was originally attached to a scheme in linear programming, but the concept can be explained much more broadly. Imagine in $(\mathcal{P})$ that the vector x is split into two vector components, $x = (x', x'')$ with $x' \in \mathbb{R}^{n'}$ and $x'' \in \mathbb{R}^{n''}$, representing the "hard" and "easy" parts of $(\mathcal{P})$ in the following sense. For any *fixed* choice of x', the residual problem of minimizing $f_0(x', x'')$ over all x'' such that $(x', x'') \in C$ is "easy" because of its special structure. Let $\varphi(x')$ denote the optimal value in this subproblem, with x' as parameter, and let B denote the set of all x' for which at least one x'' exists with $(x', x'') \in C$. In principle then, $(\mathcal{P})$ can be solved by minimizing $\varphi(x')$ over all $x' \in B$ to get $\bar{x}'$, and then solving the "easy" subproblem associated with this $\bar{x}'$ to get $\bar{x}''$; the pair $\bar{x} = (\bar{x}', \bar{x}'')$ will be optimal in $(\mathcal{P})$. Once more, this is just the skeleton of an approach which has to be elaborated into an iterative procedure to make it practical. Typically, duality enters in representing φ approximately in such a manner that it can minimized effectively.

Decomposition through block coordinate minimization: Sometimes $(\mathcal{P})$ has many easy parts, if they could be treated separately. Suppose $x = (x_1, \ldots, x_r)$ with vector components $x_k \in \mathbb{R}^{n_k}$; each x_k designates a *block* of coordinates of x in general. For any $\hat{x} \in C$, denote by $(\mathcal{P}_k(\hat{x}))$ for $k = 1, \ldots, r$ the subproblem

$$\begin{gathered}\text{minimize } f_0(\hat{x}_1, \ldots, \hat{x}_{k-1}, x_k, \hat{x}_{k+1}, \ldots, \hat{x}_r) \text{ over all } x_k \text{ such that} \\ (\hat{x}_1, \ldots, \hat{x}_{k-1}, x_k, \hat{x}_{k+1}, \ldots, \hat{x}_r) \in C.\end{gathered}$$

The basic idea is to generate a sequence $\{x^\nu\}_{\nu=0}^\infty$ in C from an initial x^0 as follows. Having reached x^ν, choose an index k^ν in $\{1, \ldots, r\}$ and solve subproblem $(\mathcal{P}_{k^\nu}(x^\nu))$, obtaining a vector $\bar{x}_{k^\nu}$. Replace the k^ν-component of x^ν by this vector, leaving all the other components as they were, to obtain the next point $x^{\nu+1}$. Obviously, if this is to work well, care must be exercised that the same index in $\{1, \ldots, r\}$ isn't always chosen; some scheme in which every index repeatedly gets its turn is essential. But what's not realized by many people is that, although the method may lead to lower values of f_0, never higher, it can get hung up and fail to produce an optimal sequence x^ν unless strong assumptions are fulfilled. Typically f_0 must not only be differentiable, it must be strictly convex, and the feasible set C should be convex and have the product form $C_1 \times \cdots \times C_r$.

ASSIGNMENT 1

Formulate the following as optimization problems with *linear* constraints and a *linear* function to be minimized or maximized. In your exposition, clearly distinguish between "decision variables" and "data parameters." Indicate the meaning of all symbols you introduce. (For additional pointers, see the next page.)

1. A new kind of hog feed is to be blended together from n different kinds of grain. Each kilogram of grain of type j (for $j = 1, \ldots, n$) costs d_j dollars and contains a_{ij} milligrams of "ingredient" i, where "ingredients" can be vitamins, fats, proteins, or anything else of interest in a hog's diet, whether desirable or undesirable. Each kilogram of the feed must contain at least b_i milligrams of ingredient i, but no more than c_i, for $i = 1, \ldots, m$. To get the cheapest blend that is compatible with these requirements, what fraction of each kilogram of the hog feed should be grain of type j?

2. A manufacturing company has a permit to operate for T seasons, after which (in season $T + 1$) it is only allowed to sell any leftover products. It is able to manufacture m different products, each requiring n different types of processing. Product i (for $i = 1, \ldots, m$) costs c_i dollars/liter to make and requires h_{ij} hours/liter of processing of type j (for $j = 1, \ldots, n$). Due to equipment limitations, the total time available for type j processing of all products during season t is H_{jt} hours (for $t = 1, \ldots, T$). (Potential complications about the order of processing are being ignored here.)

All the processing of an amount of product i must be completed in one season (it's not possible to start with some of the types of processing in one season and then finish with the others in the next), and that liter can then be sold from the *next* season onward. To sell a liter of product i in season t requires e_{it} hours of marketing effort (labor). This labor can be hired in season t at the cost of d_t dollars/hour at the ordinary rate, up to a total of a_t hours. For additional labor beyond that, a higher rate of D_t dollars/hour must be paid. (There is no limit on the amount of hours at this higher rate or on the amounts of sales, which are regarded as a sure consequence of the marketing effort.)

The selling price for product i in season t is p_{it} dollars/liter. If a quantity of product i is available for sale during a season, but is not sold then, the manufacturer has to pay q_i dollars/liter to store it and keep it in shape for possible sale in the next season. An alternative to marketing or storing is to donate quantities to charity. For this there is no monetary cost or reward. All products must be disposed of by the end of period $T + 1$.

What should the manufacturer do to maximize net profit over the entire period?

3. A hydrological model is sought to give a simple rule for predicting the water runoff in a given month on the basis of knowing the recent amounts of precipitation. It is believed that a linear equation of the form $r = b_0p_0 + b_1p_1 + b_2p_2$ should work adequately as a rough tool of prediction, where r is the runoff, p_0, p_1, p_2, are the observed precipitation amounts in the current month and the two preceding months, and b_0, b_1, and b_2 are coefficients to be determined. These coefficients are nonnegative weights (adding up to 1) which must satisfy $b_0 \geq b_1 \geq b_2$.

According to historical data for the locality in question, month t (for $t = 1, \ldots, 12$) typically has the precipitation amount q_t and the runoff amount r_t. What choice of the coefficients b_i will best fit these figures in the sense of minimizing the error expression

$$E(b_0, b_1, b_2) := \max_{t=1,\ldots,12} \left| r_t - b_0q_t - b_1q_{t-1} - b_2q_{t-2} \right| ?$$

(Here the month preceding $t = 1$ is to be interpreted as $t = 12$, and so forth.)

Caution in these exercises: Don't overlook the stipulation that the problems are to be set up with *linear* constraints and a *linear* objective function (not merely "piecewise linear"). At first this linearity may appear to be lacking, for instance in the broken costs in Exercise 2, where higher rates come in beyond a certain level, or in the error expression in Exercise 3 because of the absolute values and the "max" operation (which corresponds to a worst-case approach to errors). In all cases, however, linearity can be achieved through such tricks as introducing additional variables.

For Exercise 2, it is wise to familiarize yourself thoroughly with Example 3 in the lecture notes. For Exercise 3 you especially need to understand Example 4 of the lecture notes, but aspects of Examples 2 and 3 enter as well. Recall that an inequality of the form $y \geq |z|$ is equivalent to the pair of inequalities $y \geq z$ and $y \geq -z$.

Try to avoid writing constraints as equations when in reality inequalities would do just as well (and provide additional flexibility, or in some cases even a way of getting linearity). Be sure that no constraints have been missed, such as nonnegativity conditions on certain decision variables. Constraints can be grouped together; it will greatly help the grading and your own sense of the problem if you provide a "name" for each group.

In the end, each constraint other than a range constraint on a decision variable should involve only *one* equality or inequality sign, and it should be written with the decision variables on the left and only a constant on the right. For instance, a condition that you might be tempted to write as $1 \leq x_1 + x_2 \leq x_3$ ought finally to be presented instead as the pair of conditions $x_1 + x_2 \geq 1$ and $x_1 + x_2 - x_3 \leq 0$.

ASSIGNMENT 2

1. An elementary but helpful introduction to questions of optimality will be gained by investigating a small problem in standard format:

$$\begin{aligned}&\text{minimize } f_0(x_1, x_2) := 3x_1^2 + 3x_1x_2 + \frac{1}{2}x_2^2\\ &\text{over all } x \in X := \big\{(x_1, x_2) \,\big|\, -1 \le x_1 \le 1,\ -1 \le x_2 \le 1\big\}\\ &\text{satisfying } f_1(x_1, x_2) := 2x_1 + x_2 - 1 \le 0.\end{aligned}$$

(a) Draw a picture in $\mathbb{R}^2$ of the feasible set C.

(b) Verify that this problem is "well posed"—in the precise sense defined in class.

(c) Show that at the point $(0, 0)$, which lies in the interior of C, the partial derivatives of f_0 with respect to both x_1 and x_2 vanish, and yet f_0 has neither a local minimum nor a local maximum at $(0, 0)$. (Consider the behavior of f_0 along various lines through $(0, 0)$, not only the two axes. If you need help in recalling how to do this, see the notes after (d).)

(d) Determine the optimal value and the optimal solution(s). Also identify the points, if any, that are just locally optimal. (In this two-dimensional situation you can arrive at the answers by applying what you know about one-dimensional minimization. The idea is to study what happens to f_0 along families of parallel lines as they intersect C, these lines being chosen for instance to line up with one of the edges of C.)

Reduction to one-dimensional minimization: The line through a point (u_1, u_2) in the direction of a vector $(w_1, w_2) \neq (0, 0)$ consists of all $(x_1, x_2) = (u_1 + tw_1, u_2 + tw_2)$ for $t \in \mathbb{R}$. The segment of it that lies in C will, in the present case, be obtained by restricting t to a certain closed, bounded *interval* I (determined from the constraints that specify C). You can study the behavior of f_0 over this segment by looking at what happens to the function $\varphi(t) = f_0(u_1 + tw_1, u_2 + tw_2)$ as t ranges over I. By varying (u_1, u_2) while keeping (w_1, w_2) fixed, you get a family of parallel line segments.

Remember that in minimizing a differentiable function $\varphi(t)$ over a closed, bounded interval $I = [a, b]$, there is sure to be at least one point at which the global minimum is attained. You can identify such points by first identifying the points that satisfy the necessary conditions for a local minimum of φ on I and then checking the values of φ at those points. Having a local minimum at a point t with $a < t < b$ requires having $\varphi'(t) = 0$, whereas having a local minimum at a merely requires having $\varphi'(a) \ge 0$, and having a local minimum at $t = b$ merely requires having $\varphi'(b) \le 0$.

2. The two small problems that follow in (a) and (b) will get you acquainted with solving linear and quadratic programming problems on the computer using MATLAB. They are chosen to be extremely simple, so that you can anticipate the answer and check that you get it. (Harder problems will come later.) Your task is to input the data, apply the appropriate MATLAB tool, and get a printed output. You should turn in this output, clearly noting on it in each case both the optimal value and optimal solution you obtain, and sketching a picture of the feasible set and result.

For both problems, the set of feasible solutions is the polyhedral set $C \subset I\!R^3$ consisting of the points $x = (x_1, x_2, x_3)$ such that $x_i \geq 0$ for $i = 1, 2, 3$, and $x_1 + x_2 + x_3 \geq 1$.

(a) Minimize $6x_1 + 2x_2 + 9x_3$ over C.

(b) Find the point of C that is nearest to $a = (0, 0, -.5)$ and its distance.

3. The issues behind the basic theorems on well-posedness will be explored through the properties of the following problem in standard format:

$$\begin{aligned} \text{minimize } f_0(x_1, x_2) &:= x_1 + x_2^2 \text{ over all } (x_1, x_2) = x \in X := I\!R^2 \text{ satisfying} \\ f_1(x_1, x_2) &:= -\frac{2x_1}{1 + x_1^2} + x_2 \leq 0 \\ f_2(x_1, x_2) &:= -\frac{2x_1}{1 + x_1^2} - x_2 \leq 0. \end{aligned}$$

(a) Draw a picture of the feasible solution set C. Determine what the ε-feasible set $C_\varepsilon := \{x \in I\!R^2 \mid f_1(x) \leq \varepsilon,\ f_2(x) \leq \varepsilon\}$ will look like for small $\varepsilon > 0$. Caution: there is a "phantom" portion of C_ε which has no counterpart in C. (Hint: write the f_1 constraint in the form $x_2 \leq g(x_1)$; use calculus to graph g and understand its asymptotic properties.)

(b) Verify that the optimal value in this problem is 0 and the unique optimal solution is $\bar{x} = (0, 0)$. Establish moreover that every optimizing sequence $\{x^\nu\}$ must converge to $\bar{x}$. (Be careful to keep to the precise *definitions* of optimal value, optimal solution, and optimizing sequence.)

(c) Show that this problem *isn't* well posed. Produce in fact an *asymptotically feasible* sequence of points x^ν with the property that $f_0(x^\nu) \to -\infty$; this will then be an *asymptotically optimal* sequence that doesn't converge to the unique optimal solution $\bar{x}$ to the problem. (Verify that your sequence does fit these terms *as technically defined*.)

(d) Show that the addition of the constraint $f_3(x_1, x_2) := -x_1 - 1 \leq 0$ would cause the problem to be well posed after all. What could be said then about the convergence of asymptotically optimizing sequences?

ASSIGNMENT 3

1. This exercise concerns various properties of partial derivatives and convexity in unconstrained optimization.

(a) For the function $f(x) = f(x_1, x_2) := x_1^3 + x_2^3 - (x_1 + x_2)^2$ on $\mathbb{R}^2$, obtain formulas for $\nabla f(x)$ and $\nabla^2 f(x)$ and say what you can about where f might have a local or global minimum or maximum (unconstrained).

(b) Figure out what you can about the convexity or strict convexity of the function $g(x) = g(x_1, x_2) := 3 + x_2 + \frac{1}{2}(x_1 - x_2)^2 + \frac{1}{4}x_1^4$ on $\mathbb{R}^2$. Does this function attain a local or global minimum on $\mathbb{R}^2$?

(c) Show that if $h(x) = \max\{h_1(x), h_2(x)\}$ for convex functions h_i on $\mathbb{R}^n$, then h is convex as well. Moreover, if both h_1 and h_2 are strictly convex, then h is strictly convex. (Argue right from the definitions.)

2. The purpose of this exercise is to get you started in thinking about the systematic expression of optimality conditions by focusing on the problem of minimizing a differentiable function f_0 over the set $C = \mathbb{R}^n_+$, which is the "orthant" consisting of all $x = (x_1, \ldots, x_n)$ such that $x_j \geq 0$ for $j = 1, \ldots, n$.

(a) Show that if $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)$ is a locally optimal solution to this problem, then the following must be satisfied at $\bar{x}$ (and thus is a *necessary* condition for local optimality):

$$\begin{aligned} &(\partial f_0/\partial x_j)(\bar{x}) = 0 \text{ for each index } j \text{ with } \bar{x}_j > 0, \text{ whereas} \\ &(\partial f_0/\partial x_j)(\bar{x}) \geq 0 \text{ for each index } j \text{ with } \bar{x}_j = 0. \end{aligned} \tag{1}$$

(Get this by reduction to elementary calculus; fix all but one of the x_j's and let the remaining one vary. It might help if you first think about the cases of the given minimization problem in which $n = 1$, or $n = 2$.) Verify that (1) can be expressed by

$$\bar{x}_j \geq 0, \quad (\partial f_0/\partial x_j)(\bar{x}) \geq 0, \quad \bar{x}_j \cdot (\partial f_0/\partial x_j)(\bar{x}) = 0 \quad \text{for } j = 1, \ldots, n. \tag{1'}$$

(b) In terms of the gradient vector $\nabla f_0(x) = \big((\partial f_0/\partial x_1)(x), \ldots, (\partial f_0/\partial x_n)(x)\big)$ and the notation $y \cdot z = y_1 z_1 + \cdots + y_n z_n$ for the scalar product of two vectors $y = (y_1, \ldots, y_n)$ and $z = (z_1, \ldots, z_n)$, show that condition (1) is *equivalent* to having, for $C = \mathbb{R}^n_+$,

$$\nabla f_0(\bar{x}) \cdot [x - \bar{x}] \geq 0 \text{ for all } x \in C. \tag{2}$$

(Don't forget that "equivalence" refers to *two* implications, one in each direction.)

(c) Show that when f_0 is an *convex* function, condition (2) is in fact also *sufficient* for the *global* optimality of a point $\bar{x} \in C$: it guarantees that $f_0(x) \geq f_0(\bar{x})$ for all $x \in C$.

3. Formulate the following problem in terms of a *linear* objective function and *linear* constraints, and moreover express your answer in the *standard format* explained in class, using the set X to express any upper or lower bounds on the decision variables and introducing constraint functions for all other constraints. It will help your effort and the grading if in the final statement you give a name for each group of variables and group of constraints. (Remark: This is a problem relative to a single time period—one year. The model is entirely deterministic. A more realistic model would attempt to cope with uncertainties in prices and yields.)

A farmer can lease land up to 1000 acres. She has to pay \$6 per acre (per year) if she leases up to 600 acres. For any land beyond 600 acres, she can lease at \$8 per acre.

She grows corn on the land. She can grow corn at *normal* level or at an *intense* level (more fertilizer, frequent irrigation, etc.) Normal level yields 70 bushels/acre. Intense level yields 100 bushels/acre. The normal and intense levels require, respectively, 7 and 9 hours of labor per acre, and \$20 and \$30 in materials (such as seed, fertilizer, water, etc.) per acre. (On each acre, some amount can be at the normal level and some at the intense level.) Harvesting requires 0.1 hours of labor per bushel harvested. Harvested corn can be sold up to any amount at the rate of \$2.50 per bushel.

The farmer can raise poultry at the same time. Poultry is measured in poultry units. To raise one poultry unit requires 24 bushels of corn, 20 hours of labor, and 25 ft^3 of shed space. She can either use the corn she herself has grown or buy corn from the retail market. She gets corn at the rate of \$3.20/bushel from the retail market. She can sell at the price of \$180 per poultry unit in the wholesale market up to 200 units. Any amount of poultry over 200 units sells for just \$165 per unit. She has only one shed for raising poultry, and it has 15,000 ft^3 of space.

She and her commune can contribute 4000 hours of labor per year at no cost. If she needs more labor, she can hire it at \$4/hour up to 3000 hours. For any amount of labor over 3000 hours, she has to pay \$6 per hour.

The issue is what plan to execute in order to maximize net profit.

Note: At this stage you are only being asked to formulate this problem, but eventually you will be asked to solve it using MATLAB. By achieving a clear and accurate statement now, you'll facilitate your work later.

ASSIGNMENT 4

1. Consider the problem of minimizing over $x = (x_1, x_2) \in \mathbb{R}^2$ the function $f_0(x_1, x_2) := (x_1 - x_2)^2 + \frac{1}{4}(x_1 + x_2)^2$. The unique optimal solution is obviously $(0, 0)$. The level sets $f_0 = \text{const.}$ are ellipses having major axis along the line $x_2 = x_1$ and minor axis along the line $x_2 = -x_1$, the major axis being twice as long as the minor axis.

(a) Illustrate and explain (justify through basic calculus) the following fact relevant to any descent method for solving this problem: If at a point x^ν a descent vector w^ν for f_0 is chosen, and an exact line search relative to this vector is carried out, then $x^{\nu+1}$ will be the point obtained by moving away from x^ν in the direction indicated by w^ν until an ellipse in the nest of level sets of f_0 is reached that is tangent to the line being traced. (Consider angles made by the gradient vector as you move along the line.)

(b) Apply Cauchy's method of steepest descent to this problem using exact line search, starting from $x^0 = (3, 1)$. Carry out four iterations by following (roughly) the *graphical rule in part* (a) instead of actual numerical calculation; connect the points of the sequence you generate by the line segments used, so as to make the pattern of progress clearer.

(c) What happens if instead Newton's method (optimization version) is applied with exact line search to the same problem? (And, why is this outcome "totally expected"?)

(d) As a different approach, show graphically what happens from the same starting point under *coordinate-wise* minimization, where you first minimize in x_1 with x_2 fixed, then in x_2 with x_1 fixed, and keep alternating that way between the two variables.

2. Returning to the hydrological model in Exercise 3 of Assignment 1, but with 4 seasons instead of 12 months (for simplicity), suppose the data elements are

$$(q_1, q_2, q_3, q_4) = (0.7, 3.3, 5.2, 1.9) \text{ (precipitation in acre-inches)},$$
$$(r_1, r_2, r_3, r_4) = (1.5, 2.5, 4.1, 3.1) \text{ (runoff in acre-inches)}.$$

Season indices cycle as earlier: when $t = 1$, interpret $t - 1$ as 4, and so forth. Determine by MATLAB the coefficients b_0, b_1 and b_2 that minimize (subject to the conditions in the original problem) the error expression in each of the following two cases:

(a) $E_2(b_0, b_1, b_2) := \sum_{t=1}^{4} \left(r_t - b_0 q_t - b_1 q_{t-1} - b_2 q_{t-2}\right)^2$ (least squares error).

(b) $E_\infty(b_0, b_1, b_2) := \max_{t=1,\ldots,4} \left|r_t - b_0 q_t - b_1 q_{t-1} - b_2 q_{t-2}\right|$ (Chebyshev error).

(Note: Some "personal arithmetic" in (a) can be avoided by introducing variables u_t constrained to equal the expressions $r_t - b_0 q_t - b_1 q_{t-1} - b_2 q_{t-2}$. Then you can reduce to quadratic programming without having to process the given data tediously.)

3. This exercise concerns an approach, called the *cutting plane* method, to minimizing a convex function over a box. It will give you a taste of the thinking that goes into a numerical method and justifying it. The given problem is:

$$(\mathcal{P}) \qquad \text{minimize } f_0(x) \text{ over all } x \in X,$$

where f_0 is of class $\mathcal{C}^1$ and *convex*, while X is a nonempty box that is *bounded*. The method proceeds by iterations indexed with $\nu = 1, 2, \ldots$, starting from any initial point $x^0 \in X$ and generating each time a new point $x^\nu \in X$. It makes use of the "linearizations" of f_0 at these points, which are the affine functions $l^\nu(x) := f_0(x^\nu) + \nabla f_0(x^\nu) \cdot [x - x^\nu]$.

In the first iteration, the function $f_0^1(x) = l^0(x)$ is minimized over X; α^1 denotes the optimal value in this subproblem and x^1 the optimal solution found for it. In general for iteration ν, points $x^0, x^1, \ldots, x^{\nu-1}$ are available and the subproblem $(\mathcal{P}^\nu)$ is solved in which the function

$$f_0^\nu(x) := \max\Big\{l^0(x), l^1(x), \ldots, l^{\nu-1}(x)\Big\}$$

is minimized over X. The optimal solution found for $(\mathcal{P}^\nu)$ is taken then as x^ν, and the optimal value in $(\mathcal{P}^\nu)$ is denoted by α^ν. The following questions concern the sequences $\{x^\nu\}_{\nu=1}^\infty$ and $\{\alpha^\nu\}_{\nu=1}^\infty$ and the extent to which they serve to solve the given problem $(\mathcal{P})$.

(a) Draw a picture (or pictures) of how the method works in the one-dimensional case ($n = 1$), where X is merely a closed, bounded interval. (Indicate for several iterations the successive points x^ν and functions f_0^ν.)

(b) Verify (henceforth in the general n-dimensional case) that the functions f_0^ν are convex and that they satisfy $f_0^\nu(x) \le f_0(x)$ for all x. Establish moreover that the epigraph of f_0^ν (the set of points lying on or above the graph) is expressible by a system of finitely many *linear* constraints in $\mathbb{R}^{n+1}$, and explain how the problem of minimizing f_0^ν over X can therefore be formulated as a *linear programming* problem in $\mathbb{R}^{n+1}$ for the purpose of applying software to calculate x^ν and α^ν.

(c) Show that $\alpha^1 \le \cdots \le \alpha^\nu \le \alpha^{\nu+1} \le \cdots \le \bar{\alpha}$, where $\bar{\alpha}$ denotes the optimal value in $(\mathcal{P})$ itself. Show further that in every iteration κ after a given iteration ν one will have $l^\nu(x) \le f_0^\kappa(x)$ for all x, and in particular therefore (through the case of $x = x^\kappa$) that

$$f_0(x^\nu) + \nabla f_0(x^\nu) \cdot [x^\kappa - x^\nu] \;\le\; \bar{\alpha}.$$

(d) Show that if the sequence $\{x^\nu\}_{\nu=1}^\infty$ happens to converges to some $\bar{x}$, then $\bar{x}$ must be an optimal solution to $(\mathcal{P})$. Show that the same holds even if $\bar{x}$ is just a cluster point of this sequence. Does it follow that $\{x^\nu\}_{\nu=1}^\infty$ is *always an optimizing sequence* for $(\mathcal{P})$? (Hint: Consider taking limits twice in (c), first letting $\kappa \to \infty$ and then letting $\nu \to \infty$. Utilize the closedness of X and the continuity of $f_0(x)$ and $\nabla f_0(x)$ with respect to x.)

ASSIGNMENT 5

1. This returns to the optimization example in Exercise 1 of Assignment 2 for a closer look, now analyzing optimality systematically on the basis of Theorem 6.

(a) Determine all points $\bar{x}$ that satisfy the first-order necessary condition for optimality in terms of $\nabla f_0(\bar{x})$ and $T_C(\bar{x})$. In a diagram of the feasible set C, indicate at each of these points $\bar{x}$ both the tangent cone $T_C(\bar{x})$ and the reverse gradient $-\nabla f_0(\bar{x})$. (Consider separately, one by one, the nine possible situations corresponding to the interior, the four sides and the four corners of this polyhedral set C.)

(b) Check whether the first-order points $\bar{x}$ identified in (a) satisfy, in addition, the second-order necessary or sufficient condition involving also $\nabla^2 f_0(\bar{x})$.

(c) Apply MATLAB to this problem. Discuss the extent to which the solution output matches, or falls short of, what you have determined about optimality. As an experiment, see if you get anything different if you call up MATLAB more than once on the same data, or on the other hand, if you insist on some particular choice of MATLAB's starting point, for instance $(0, 0)$.

(d) Add the constraint $f_2(x_1, x_2) := -3x_1 + 3x_2 - 4 \leq 0$ to the problem and redo (a) and (b), concentrating your update on the places where the new constraint makes a difference. What now is locally or globally optimal? Once again apply MATLAB and compare the solution output with what you know.

2. The following tasks provide some ways to get acquainted with basic facts about convex sets and their tangent cones, including connections with feasible directions.

(a) Show that for a convex set C and a point $\bar{x} \in C$, the vectors giving feasible directions into C at $\bar{x}$ are the vectors w that can be expressed in the form $w = \lambda(x - \bar{x})$ for some $x \in C$ other than $\bar{x}$ and some $\lambda > 0$. (Note that this is a two-way assertion.)

(b) Verify that the set $C = \big\{(x_1, x_2, x_3) \,\big|\, x_2 \geq 0,\ |x_3| \leq 1 - x_1^2 - x_2^2\big\}$ is convex in $\mathbb{R}^3$. (Simplify your task by making use of various rules for deriving convex sets from other convex sets or convex functions.) What does this set C look like? (A picture isn't required, but one could help you in understanding parts (c) and (d) that follow.)

(c) For the set C in (b) and the point $\bar{x} = (1, 0, 0)$, show that if a vector $(w_1, w_2, w_3) \neq (0, 0, 0)$ gives a feasible direction into C at $\bar{x}$, then it satisfies $w_2 \geq 0$ and $|w_3| \leq -2w_1$. On the other hand, show that if a vector $(w_1, w_2, w_3) \neq (0, 0, 0)$ satisfies $w_2 \geq 0$ and $|w_3| < -2w_1$, then it gives a feasible direction into C at $\bar{x}$. (In each case rely on the definition of "feasible directions" and the constraints specifying C in (b).)

Note: The first argument develops a *necessary* condition for a feasible direction in

this special setting, whereas the second argument develops a *sufficient* condition.)

(d) Derive from the necessary and sufficient conditions on feasible directions in (c) the fact that, for the set C and point $\bar{x}$ in question, the tangent cone is given by

$$T_C(\bar{x}) = \big\{(w_1, w_2, w_3) \,\big|\, w_2 \geq 0,\ |w_3| \leq -2w_1\big\}$$

and is a convex set, moreover polyhedral.

3. Projections onto sets play a big role in many applications. You may already have run into the idea of projecting onto a subspace of $\mathbb{R}^n$, but here we look more generally at projecting onto a convex set. This often arises as a kind of approximation. It also comes up in technical ways, for instance when C is the set of feasible solutions of some system of equations and inequalities and, in trying to minimize something over C, it's hard to stay within C, so there's a constant need for corrections to get back into C.

(a) Let C be a nonempty, closed, convex subset of $\mathbb{R}^n$, and let a be any point of $\mathbb{R}^n$ that's not in C. Consider the problem of finding a point of C that's nearest to a with respect to the Euclidean distance (where the distance between points x and y in $\mathbb{R}^n$ is given by the Euclidean norm $|x - y|$). Setting this up as a certain optimization problem in standard format (some approaches are more convenient than others!), establish that a solution always *exists* and is *unique*.

Note: This solution point is called the projection of a on C and denoted by $P_C(a)$.

(b) By applying Theorem 7 to this situation (assumptions satisfied?), show that a point $b \in C$ is the projection $P_C(a)$ if and only if the vector $c = a - b$ has the property that the linear function $l(x) = c \cdot x$ achieves its maximum over C at b. Draw a figure illustrating this fact; indicate in it the gradient of l and the level set $\big\{x \,\big|\, l(x) = l(b)\big\}$. (Hint: Before getting too involved with the linear function in question, work out the details of the optimality condition supplied by Theorem 7, trying to write it in terms of the vector c.)

(c) Let $C = C_1 \times \cdots \times C_r$ for closed convex sets $C_k \in \mathbb{R}^{n_k}$ with $n_1 + \cdots + n_r = n$, and let a be expressed accordingly as $a = (a_1, \ldots, a_r)$ with components $a_k \in \mathbb{R}^{n_k}$. Show in this setting that the projection of a on C is $b = (b_1, \ldots, b_r)$, where b_k is the projection of a_k on C_k for $k = 1, \ldots, r$. (Write the problem in terms of the corresponding vector components of $x = (x_1, \ldots, x_r)$, and demonstrate that the minimum is achieved when x_k is taken to equal b_k for every k.) Next, illustrate this principle by applying it to the case where C is actually a box in $\mathbb{R}^n$, showing it leads to a simple rule with respect to coordinates that yields the projection b of any a. In particular, what's the rule when $C = \mathbb{R}^n_+$?

ASSIGNMENT 6

1. Solve the farmer's problem from Assignment 3 by means of MATLAB. Indicate the solution on the printout, but also translate the results back into the original language of the problem. In other words, write out an explanation of just what the farmer should do to maximize profit, and say how much the profit will be.

Furthermore, discuss the answer, noting in particular which of the inequality constraints are "active" at the solution (i.e., satisfied as equations) in contrast to which are "inactive." How sensitive is the solution to the amount of land or shed space available, or to the price of poultry going up or down? And so forth.

2. This concerns the separation of two sets C_1 and C_2 by a hyperplane, an issue of interest not only theoretically but in various applications such as data classification and image identification. The exercise furnishes a chance to work with the normal vector condition for optimality in Theorem 9, which will be utilized in constructing such a hyperplane.

Here C_1 and C_2 are nonempty, closed, *convex* subsets of $\mathbb{R}^n$ with $C_1 \cap C_2 = \emptyset$, and C_1 is *bounded* (but C_2 could be unbounded). These sets might be specified by systems of constraints, but that doesn't play a direct role for now, even though it could eventually be essential in passing to a numerical application of the ideas that follow.

Recall that a *hyperplane* in $\mathbb{R}^n$ generalizes a line in $\mathbb{R}^2$ and a plane in $\mathbb{R}^3$; by definition it is a set H expressible in the form $\{x \in \mathbb{R}^n \mid a \cdot x = \alpha\}$ for some nonzero vector $a \in \mathbb{R}^n$ and scalar $\alpha \in \mathbb{R}$. Any such hyperplane H is associated with two *closed half-spaces*, namely $\{x \in \mathbb{R}^n \mid a \cdot x \geq \alpha\}$ and $\{x \in \mathbb{R}^n \mid a \cdot x \leq \alpha\}$. It is said to *separate* C_1 and C_2 if C_1 lies in one of these closed half-spaces while C_2 lies in the other. (For your own understanding, draw yourself some pictures, making sure you see how the vector a would be oriented relative to H and the half-spaces, which are level sets of the linear function $l(x) = a \cdot x$ having $\nabla l(x) \equiv a$.)

The focus henceforth is on the problem of minimizing $f_0(x_1, x_2) = \frac{1}{2}|x_1 - x_2|^2$ subject to $x_1 \in C_1$ and $x_2 \in C_2$. (Note that x_1 and x_2 denote *vectors*: $x_1 = (x_{11}, \ldots, x_{1n})$ and $x_2 = (x_{21}, \ldots, x_{2n})$. The problem is $2n$-dimensional, even though it's pictured in $\mathbb{R}^n$!)

(a) Interpreting the problem as one of form $(\mathcal{P})$ in which f_0 is minimized over $X = C_1 \times C_2$, this being a nonempty, closed, convex subset of $\mathbb{R}^{2n}$, show that it is *well posed*, so that (by Theorem 1) it has at least one optimal solution. Illustrate by a counterexample that the optimal solution need not be unique. (A picture will suffice.)

(b) In preparation for applying Theorem 9 to this problem, verify, from what's known in general about normal vectors to convex sets, that $N_X(\bar{x}) = N_{C_1}(\bar{x}_1) \times N_{C_2}(\bar{x}_2)$ when $X = C_1 \times C_2$ and $\bar{x} = (\bar{x}_1, \bar{x}_2)$.

(c) Investigate the gradient condition furnished by Theorem 9 at an optimal solution $\bar{x}$, showing that it yields a vector $a \neq 0$ in $\mathbb{R}^n$ along with scalars $\alpha_1 < \alpha_2$ such that $a \cdot x_1 \leq \alpha_1$ for all $x_1 \in C_1$ and $a \cdot x_2 \geq \alpha_2$ for all $x_2 \in C_2$. Establish from this the existence of a hyperplane H in $\mathbb{R}^n$ that separates C_1 and C_2 in the sense defined, which moreover can be chosen to have no points in common with either C_1 or C_2. (Provide a specific formula for such a hyperplane.)

(Conclusion: a separating hyperplane is thus obtainable "constructively" by solving the minimization problem that has been posed, once the descriptions of C_1 and C_2 have adequately been fleshed out.)

3. This exercise explores the role of Lagrange multipliers in optimality. It concerns the set $C \subset \mathbb{R}^2$ consisting of all $x = (x_1, x_2) \in X = \mathbb{R}^2_+$ that satisfy

$$\begin{cases} 0 \geq f_1(x_1, x_2) = x_1 + x_2 - 5, \\ 0 \geq f_2(x_1, x_2) = x_1 - x_2 - 1, \\ 0 \geq f_3(x_1, x_2) = 2x_1 - x_2 - 4. \end{cases}$$

(a) Draw a picture of C, indicating the tangent and normal cones to C at representative points $\bar{x}$ and giving an algebraic expression in terms of Lagrange multipliers for the vectors $v = (v_1, v_2)$ that belong to the normal cone $N_C(\bar{x})$ in each case. (Hint: Use Theorem 10. There's no need to be concerned about the standard constraint qualification ($\star$) there, because this is a system of linear constraints only; cf. Theorem 12(a).) The description of $N_C(\bar{x})$ yields equations for v_1 and v_2 in terms of y_1, y_2, y_3, z_1, and z_2 as parameters (these being the coordinates of the vectors y and z there). Having written down these equations in general, specialize them at each $\bar{x}$ by suppressing any parameters that have to equal 0 and indicating the sign restrictions, if any, on the remaining parameters.

(b) For the problem of minimizing $f_0(x_1, x_2) = (x_1 - 1)^2 - x_2$ over C, add to your picture in (a) some indication of representative level sets $f_0(x_1, x_2) = \alpha$. Determine all locally and globally optimal solutions $\bar{x}$ "by hand" through Theorem 9. (Note: this amounts to checking where the Kuhn-Tucker conditions are satisfied; justifiable shortcuts based on the problem's structure are encouraged.) For each such $\bar{x}$ give the specific parameter values y_1, y_2, y_3, z_1, and z_2 in part (a) for the normal vector representation of $-\nabla f_0(\bar{x})$.

(c) Solve the minimization problem in (b) using the routine in MATLAB for solving quadratic programs. Implement that routine so that *Lagrange multipliers* are also returned. Do the latter correspond to the parameter values you calculated in (b)?

ASSIGNMENT 7

1. Consider an optimization problem $(\mathcal{P})$ in standard format in which the functions f_i are all of class C^1 and $X = \mathbb{R}^n$. Let $\bar{x}$ denote any *feasible* solution.

(a) Show that the standard constraint qualification $(\star)$ is satisfied at $\bar{x}$ if the following conditions, comprising the *Mangasarian-Fromovitz* constraint qualification, hold:

$$\begin{cases} (1) \text{ the vectors } \nabla f_i(\bar{x}) \text{ for } i \in [s+1, m] \text{ are linearly independent, and} \\ (2) \text{ there exists a vector } w \text{ such that } \nabla f_i(\bar{x}) \cdot w \begin{cases} < 0 & \text{for } i \in [1, s] \text{ } \textit{active} \text{ at } \bar{x}, \\ = 0 & \text{for } i \in [s+1, m]. \end{cases} \end{cases}$$

(Remark: The converse is true as well; these two constraint qualifications are *equivalent*—for problems with $X = \mathbb{R}^n$. But you're not being asked to prove it.)

(b) Let $s = m$, so that the constraints in $(\mathcal{P})$ are $f_i(x) \leq 0$ for $i = 1, \ldots, m$. Suppose the following condition, called the *Slater* constraint qualification, is satisfied: all the constraint functions f_i are *convex* and there is a point $\tilde{x}$ with $f_i(\tilde{x}) < 0$ for $i = 1, \ldots, m$. Show that then the Mangasarian-Fromovitz constraint qualification (in which now only the inequality part counts) is satisfied at $\bar{x}$ (no matter how $\bar{x}$ is located relative to $\tilde{x}$).

(c) Go back to the set C and point $\bar{x}$ in Assignment 5, Exercise 2(b)(c)(d), viewing this set as given in the manner above by three constraints $f_i(x) \leq 0$. (The condition involving $|x_3|$ converts into two of these). Determine the normal cone $N_C(\bar{x})$, taking care to check somehow that the assumptions behind the formula you are using are fulfilled.

2. Here is an elementary example of linear programming duality which provides hands-on experience with that concept.

(a) Consider as problem $(\mathcal{P}_{\text{lin}})$ (in primal canonical format):

$$\begin{cases} \text{minimize} \quad 3x_1 + 5x_2 \quad \text{over all } x = (x_1, x_2) \in \mathbb{R}^2_+ \\ \text{such that} \quad 2x_1 + x_2 \geq 4, \quad x_1 + x_2 \geq 3. \end{cases}$$

Sketch the feasible set in $(\mathcal{P}_{\text{lin}})$ and use this to locate an optimal solution $\bar{x}$ graphically. Passing through an interpretation of $(\mathcal{P}_{\text{lin}})$ as a problem in standard format, determine specific values $\bar{y}_1$ and $\bar{y}_2$ for the multipliers y_1 and y_2 associated with two inequality constraints in $(\mathcal{P}_{\text{lin}})$ such that the Kuhn-Tucker conditions at $\bar{x}$ are fulfilled.

(b) Write down the corresponding dual problem $(\mathcal{D}_{\text{lin}})$ (in dual canonical format). Then carry out the same procedure for that, but with a shift in notation: this time, with y_1 and y_2 as the decision variables, denote the multipliers by x_1 and x_2. You should end up with the same vectors $\bar{x} = (\bar{x}_1, \bar{x}_2)$ and $\bar{y} = (\bar{y}_1, \bar{y}_2)$ as in (a), and also the same optimal value in the two problems, thereby confirming the assertions of Theorem 14.

3. The following is a simplified mathematical model for the economical management of electrical power dispatch. There are n power generating plants indexed by $j = 1, \ldots, n$. Plant j is capable of producing any power amount x_j in a fixed interval $l_j \leq x_j \leq u_j$, where $0 \leq l_j < u_j < \infty$, the cost being $\varphi_j(x_j) = c_j x_j + \frac{1}{2} b_j x_j^2$ with known coefficients $c_j > 0$ and $b_j > 0$. The plants are all connected to the same transmission network, so that when each produces an amount x_j the sum of these amounts enters the network; but due to transmission losses, which involve electrical interactions, the amount of power actually made available to customers is not this sum but

$$h(x) = h(x_1, \ldots, x_n) \;:=\; \sum\nolimits_{j=1}^{n} x_j - \tfrac{1}{2} \sum\nolimits_{j=1,\, k=1}^{n,\, n} a_{jk} x_j x_k$$

for a certain symmetric, positive definite matrix $A \in \mathbb{R}^{n \times n}$ with entries $a_{jk} \geq 0$.

It is assumed in the model that the entries a_{jk} are small enough that the partial derivatives $(\partial h / \partial x_j)(x)$ are positive at all vectors $x = (x_1, \ldots, x_n)$ having $l_j \leq x_j \leq u_j$. This ensures that h is an increasing function with respect to each variable over these ranges; in other words, an increase in power at one of the plants always results in an increase in power available from the network. Note that the highest value h can achieve is $h(u) = h(u_1, \ldots, u_n)$, whereas the lowest is $h(l) = h(l_1, \ldots, l_n)$.

The exercise revolves around the following problem in these circumstances: for a given load demand d (power to be withdrawn from the network), with $h(l) < d < h(u)$, determine a scheduling vector $x = (x_1, \ldots, x_n)$ that meets this demand as cheaply as possible.

(a) Express this as a problem $(\mathcal{P})$ in standard format with one constraint function f_1. Is this quadratic programming? convex programming? Is the corresponding Lagrangian $L(x, y)$ convex in $x \in X$ for each $y \in Y$ as well as affine in $y \in Y$ for each $x \in X$?

(b) Does $(\mathcal{P})$ have at least one optimal solution? At most one optimal solution?

(c) Show that, by virtue of the assumptions in the model, the standard constraint qualification is satisfied at every feasible solution $\bar{x}$.

(d) If the Kuhn-Tucker conditions for $(\mathcal{P})$ hold at $\bar{x}$, can you legitimately conclude that $\bar{x}$ is optimal? On the other hand, if they don't hold at $\bar{x}$, might $\bar{x}$ be optimal anyway?

(e) Show that the Kuhn-Tucker conditions come down to relations between the single Lagrange multiplier $\bar{y} = \bar{y}_1$ and the ratio of $\varphi_j'(\bar{x}_j)$ to $(\partial h / \partial x_j)(\bar{x}_1, \ldots, \bar{x}_n)$ for $j = 1, \ldots, n$, and moreover that they imply $\bar{y}$ has to be positive.

(f) What must be the units in which $L(x, y)$ and y are measured, considering that the costs $\varphi_j(x_j)$ are in dollars, whereas $h(x)$ and the x_j's are in "power units"? In such terms, try to interpret the ratio relations required by the Kuhn-Tucker conditions. What do they tell you about schedule levels $\bar{x}_j$ that are optimal with respect to the given demand d?

ASSIGNMENT 8

Here first is some additional background on a basic class of two-person zero-sum games called *matrix* games. A major part of this assignment will be concerned with such games.

In a matrix game in its underlying "pure" form, Player 1 has only finitely many "pure strategies" available, indexed by $j = 1, \ldots, n$, whereas Player 2 has "pure strategies" indexed by $i = 1, \ldots, m$. When Player 1 selects strategy j and Player 2 selects strategy i, the payoff from Player 1 to Player 2 is d_{ij}; these amounts give the entries of the *payoff* matrix D associated with the game. (Any matrix D corresponds to a game in this manner. One can think of Player 1 as choosing a column of D and Player 2 as choosing a row.)

Instead of keeping merely with the pure form of a matrix game, however, one enlarges the spaces of actions available to the two players to include "mixed strategies," which have a randomized nature. This means that the strategy set X for Player 1 is really taken to consist of all $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ such that $x_j \geq 0$ and $\sum_{j=1}^n x_j = 1$, with the interpretation that x_j is the *probability* that Player 1 will select the pure strategy j.

The idea is that, in choosing x, Player 1 arranges to make a last-minute decision on which pure strategy j to use by means of a lottery device with the x_j's as its probabilities. (The device might for instance be a wheel marked into sectors $j = 1, \ldots, n$ having angular sizes proportional to the desired probabilities $x_1, \ldots, x_n$; the mixed strategy x would correspond to a spin of this wheel to get a particular j.) In this framework the original pure strategy for $j = 1$ is identified with the probability vector $x = (1, 0, \ldots, 0)$, and so forth.

Similarly, the strategy set Y for Player 2 is taken to consist of all $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$ such that $y_i \geq 0$ and $\sum_{i=1}^m y_i = 1$, with y_i interpreted as the probability assigned to i. The payoff is $L(x, y) = \sum_{i=1,\, j=1}^{m,n} d_{ij} y_i x_j = y \cdot Dx$, this being the *expected* amount that will paid by Player 1 to Player 2 when the probability vectors are x and y (inasmuch as the probability of the combination (i, j) coming up, with payment d_{ij}, is then $y_i x_j$).

1. A particular example to be explored is the game called *Penny-Nickel-Dime*. Each player has a penny, a nickel and a dime, and a pure strategy consists of selecting one of these coins and displaying it. If the sum of the cents (on the two displayed coins) is odd, Player 1 wins Player 2's coin, but if the sum is even, Player 2 wins Player 1's coin.

(a) Construct the 3×3 payoff matrix D for Penny-Nickel-Dime. Verify that the vectors $\bar{x} = (1/2, 0, 1/2)$ and $\bar{y} = (10/11, 0, 1/11)$ are optimal strategies for the two players, with the equilibrium value $L(\bar{x}, \bar{y})$ being 0 (so the game is fair despite its asymmetry!). Do this via Theorem 15 by checking that $(\bar{x}, \bar{y})$ is a saddle point of $L(x, y)$ with respect to $X \times Y$.

(b) Derive for a matrix game with *general* pay-off matrix D the fact that the strategy optimizing problem $(\mathcal{P}_1)$ for Player 1 comes out as having the form

$$(\mathcal{P}_1) \qquad \text{minimize } f(x) \text{ over } x \in X, \text{ with } f(x) = \max_{i=1,\ldots,m} \{d_{i1}x_1 + \cdots + d_{in}x_n\},$$

whereas the strategy optimizing problem $(\mathcal{P}_2)$ for Player 2 comes out as having the form

$$(\mathcal{P}_2) \qquad \text{maximize } g(y) \text{ over } y \in Y, \text{ with } g(y) = \min_{j=1,\ldots,n} \{y_1 d_{1j} + \cdots + y_m d_{mj}\}.$$

(Get these expressions for $f(x)$ and $g(y)$ by using the definitions of the two problems in the general theory of two-person zero-sum games.)

(c) In Penny-Nickel-Dime there is actually never an advantage to either player in displaying the nickel—the consequences of displaying the penny are always at least as good, and in some cases better. Making use of that observation in specializing $(\mathcal{P}_1)$ to Penny-Nickel-Dime, one can reduce from minimizing over $x \in X$ to minimizing over $t \in [0,1]$ under the parameterization $x = (1-t, 0, t)$. Graph the objective function in this one-dimensional subproblem and find an optimal $\bar{t}$, confirming in this way that the mixed strategy $\bar{x}$ for Player 1 that was seen in (a) is indeed an optimal solution to $(\mathcal{P}_1)$. Then proceed the same way with $(\mathcal{P}_2)$.

2. The duality relationships between linear programming problems $(\mathcal{P}_{\text{lin}})$ and $(\mathcal{D}_{\text{lin}})$ in canonical format can be extended to include equality constraints as well as variables that aren't restricted to be nonnegative. Let's say that a linear programming problem is in *extended* primal canonical format if it is written as

$$(\bar{\mathcal{P}}_{\text{lin}}) \qquad \begin{array}{l} \text{minimize } c_1x_1 + \cdots + c_nx_n \text{ subject to} \\ a_{i1}x_1 + \cdots + a_{in}x_n \begin{cases} \geq b_i & \text{for } i = 1, \ldots, s, \\ = b_i & \text{for } i = s+1, \ldots, m, \end{cases} \\ x_j \begin{cases} \geq 0 & \text{for } j = 1, \ldots, r, \\ \text{free} & \text{for } j = r+1, \ldots, n. \end{cases} \end{array}$$

(a) Working with the Lagrangian for this problem, when rewritten in standard format as $(\mathcal{P})$ for a certain choice of X and functions f_i, derive from the general theory of duality the associated Lagrangian dual problem $(\mathcal{D})$, showing that it can be identified with the following linear programming problem in *extended* dual canonical format:

$$(\bar{\mathcal{D}}_{\text{lin}}) \qquad \begin{array}{l} \text{maximize } b_1y_1 + \cdots + b_my_m \text{ subject to} \\ y_1a_{1j} + \cdots + y_ma_{mj} \begin{cases} \leq c_j & \text{for } j = 1, \ldots, r, \\ = c_j & \text{for } j = r+1, \ldots, n, \end{cases} \\ y_i \begin{cases} \geq 0 & \text{for } i = 1, \ldots, s, \\ \text{free} & \text{for } i = s+1, \ldots, m. \end{cases} \end{array}$$

(b) Show that these problems, like the earlier $(\mathcal{P}_{\text{lin}})$ and $(\mathcal{D}_{\text{lin}})$, share the same optimality conditions—a certain modification of those in part (e) of Theorem 14. (Develop the Kuhn-Tucker conditions for standard format version of $(\bar{\mathcal{P}}_{\text{lin}})$, and then do the same for $(\bar{\mathcal{D}}_{\text{lin}})$, likewise by passing through a re-expression in standard format.)

Note: With this modified statement of (e) Theorem 14 remains valid in its entirety for these extended problems. The previous arguments for proving Theorem 14 readily go through again in this setting. You're not being asked to provide the details of that, but you are free anyway to use this extended Theorem 14 in the exercise that follows.

3. The facts developed in Exercise 2 will be applied now to general matrix games.

(a) Show how the strategy optimizing problem $(\mathcal{P}_1)$ in 1(b) for a general matrix game with pay-off matrix D can be reformulated as a linear programming problem $(\bar{\mathcal{P}}_{\text{lin}})$ in the extended primal canonical format. Passing through the scheme in Exercise 2, determine the corresponding problem $(\bar{\mathcal{D}}_{\text{lin}})$. Show that this $(\bar{\mathcal{D}}_{\text{lin}})$ can moreover be identified with what one gets by reformulating the general strategy optimizing problem $(\mathcal{P}_2)$ in (1b) as a linear programming problem in the extended dual canonical format.

(b) On the basis of 1(b) and the identifications 3(a), prove the following famous fact (a theorem of Von Neumann): In any *matrix* game, optimal strategies (in the *mixed* sense) exist for both of the players, and there is a number V (the *value* of the game) such that Player 1 can guarantee not having to pay more than V (as a probabilistic "expectation"), while Player 2 can guarantee receiving at least V (as an "expectation").

(c) Utilize the reformulation technique in (a) to determine by MATLAB the value V of the matrix game with the following pay-off matrix D, along with optimal strategies $\bar{x}$ and $\bar{y}$ for the two players:

$$D = \begin{bmatrix} 1 & 0 & 3 & 2 \\ 4 & 2 & 0 & 2 \\ 2 & 1 & 5 & 0 \end{bmatrix}.$$

(If you've fully understood things, you'll see that you only need to apply a MATLAB tool once, not twice for two problems. Don't be thrown off by a MATLAB glitch which may or not have gotten fixed: Lagrange multipliers for *equation* constraints could be reported with the sign wrong.)

Note: This pay-off matrix was selected arbitrarily. You are now in the position of being able to "solve" *any* matrix game! You might be interested, for instance, to try the extensions of Penny-Nickel-Dime to Penny-Nickel-Dime-Quarter and Penny-Nickel-Dime-Quarter-Dollar, seeing whether these games likewise are fair.