

This text is intended to function as an introduction to *Linear Programming (LP)* and the *Simplex algorithm*. The specific topics covered and the structure of the material is as follows:

- The LP formulation and the underlying assumptions
- Graphical solution of 2-var LP's
- Generalization to the  $n$ -var case: the “geometry” of the LP feasible region and the Fundamental Theorem of Linear Programming.
- An algebraic characterization of the solution search space: Basic Feasible Solutions
- The Simplex Algorithm

Most of the text material is presented inductively, by generalizing some introductory highlighting examples. Also, integrated with the text is a software package which provides the reader with the ability of running *customized interactive examples*. Specifically, the software interfaces are distributed at the end of key sections of the text, and they are intended to demonstrate/visualize key concepts and the functionality of the algorithms discussed in the text, by showing how these concepts and algorithms apply to reader-provided LP instances.

Finally, it is acknowledged that the basic structure of the material and many of the examples used in the text have been inspired by W. L. Winston's "*Introduction to Mathematical Programming*", ed. Duxbury, which was used as the class text during the two quarters that I taught the corresponding introductory LP course at the School of Industrial & Systems Engineering, at Georgia Tech.

## 1 The LP formulation and the underlying assumptions

A *Linear Programming* problem is a special case of a *Mathematical Programming* problem. From an analytical perspective, a mathematical program tries to identify an *extreme* (i.e., minimum or maximum) point of a function  $f(x_1, x_2, \dots, x_n)$ , which furthermore satisfies a set of constraints, e.g.,  $g(x_1, x_2, \dots, x_n) \geq b$ . Linear programming is the specialization of mathematical programming to the case where both, function  $f$  – to be called the *objective function* – and the problem constraints are *linear*.

From an applications perspective, mathematical (and therefore, linear) programming is an *optimization* tool, which allows the rationalization of many managerial and/or technological decisions required by contemporary technosocio-economic applications. An important factor for the applicability of the mathematical programming methodology in various application contexts, is the computational tractability of the resulting analytical models. Under the advent

of modern computing technology, this tractability requirement translates to the existence of effective and efficient algorithmic procedures able to provide a systematic and fast solution to these models. For Linear Programming problems, the *Simplex* algorithm, discussed later in the text, provides a powerful computational tool, able to provide fast solutions to very large-scale applications, sometimes including hundreds of thousands of variables (i.e., decision factors). In fact, the Simplex algorithm was one of the first Mathematical Programming algorithms to be developed (George Dantzig, 1947), and its subsequent successful implementation in a series of applications significantly contributed to the acceptance of the broader field of *Operations Research* as a scientific approach to decision making.

As it happens, however, with every modeling effort, the effective application of Linear Programming requires good understanding of the underlying modeling assumptions, and a pertinent interpretation of the obtained analytical solutions. Therefore, in this section we discuss the details of the LP modeling and its underlying assumptions, by means of the following example.

**A prototype LP problem:** Consider a company which produces two types of products  $P_1$  and  $P_2$ . Production of these products is supported by two workstations  $W_1$  and  $W_2$ , with each station visited by both product types. If workstation  $W_1$  is dedicated completely to the production of product type  $P_1$ , it can process 40 units per day, while if it is dedicated to the production of product  $P_2$ , it can process 60 units per day. Similarly, workstation  $W_2$  can produce daily 50 units of product  $P_1$  and 50 units of product  $P_2$ , assuming that it is dedicated completely to the production of the corresponding product. If the company's profit by disposing one unit of product  $P_1$  is \$200 and that of disposing one unit of  $P_2$  is \$400, and assuming that the company can dispose its entire production, how many units of each product should the company produce on a daily basis to maximize its profit?

**Solution:** First notice that this problem is an *optimization* problem. Our *objective* is to *maximize* the company's profit, which under the problem assumptions, is equivalent to maximizing the company's daily profit. Furthermore, we are going to maximize the company profit by adjusting the levels of the daily production for the two items  $P_1$  and  $P_2$ . Therefore, these daily production levels are the control/decision factors, the values of which we are called to determine. In the analytical formulation of the problem, the role of these factors is captured by modeling them as the problem *decision variables*:

- $X_1$  := number of units of product  $P_1$  to be produced daily
- $X_2$  := number of units of product  $P_2$  to be produced daily

In the light of the above discussion, the problem objective can be expressed analytically as:

$$\max f(X_1, X_2) := 200X_1 + 400X_2 \quad (1)$$

Equation 1 will be called the *objective function* of the problem, and the coefficients 200 and 400 which multiply the decision variables in it, will be called the *objective function coefficients*.

Furthermore, any decision regarding the daily production levels for items  $P_1$  and  $P_2$  in order to be realizable in the company's operation context must observe the production capacity of the two workstations  $W_1$  and  $W_2$ . Hence, our next step in the problem formulation seeks to introduce these *technological constraints* in it. Let's focus first on the constraint which expresses the finite production capacity of workstation  $W_1$ . Regarding this constraint, we know that one day's work dedicated to the production of item  $P_1$  can result in 40 units of that item, while the same period dedicated to the production of item  $P_2$  will provide 60 units of it. Assuming that production of one unit of product type  $P_i$ ,  $i = 1, 2$ , requires a constant amount of processing time  $\tau_{1i}$  at workstation  $W_1$ , it follows that:  $\tau_{11} = \frac{1}{40}$  and  $\tau_{12} = \frac{1}{60}$ . Under the further assumption that the combined production of both items has no side-effects, i.e., does not impose any additional requirements for production capacity of workstation  $W_1$  (e.g., zero set-up times), the total capacity (in terms of time length) required for producing  $X_1$  units of product  $P_1$  and  $X_2$  units of product  $P_2$  is equal to  $\frac{1}{40}X_1 + \frac{1}{60}X_2$ . Hence, the technological constraint imposing the condition that our total daily processing requirements for workstation  $W_1$  should not exceed its production capacity, is analytically expressed by:

$$\frac{1}{40}X_1 + \frac{1}{60}X_2 \leq 1.0 \quad (2)$$

Notice that in Equation 2 time is measured in days.

Following the same line of reasoning (and under similar assumptions), the constraint expressing the finite processing capacity of workstation  $W_2$  is given by:

$$\frac{1}{50}X_1 + \frac{1}{50}X_2 \leq 1.0 \quad (3)$$

Constraints 2 and 3 are known as the *technological constraints* of the problem. In particular, the coefficients of the variables  $X_i$ ,  $i = 1, 2$  in them,  $\frac{1}{\tau_{ji}}$ ,  $j, i = 1, 2$ , are known as the *technological coefficients* of the problem formulation, while the values on the right-hand-side of the two inequalities define the *right-hand side (rhs)* vector of the constraints.

Finally, to the above constraints we must add the requirement that any permissible value for variables  $X_i$ ,  $i = 1, 2$  must be nonnegative, i.e.,

$$X_i \geq 0 \quad i = 1, 2 \quad (4)$$

since these values express production levels. These constraints are known as the *variable sign restrictions*.

Combining Equations 1 to 4, the analytical formulation of our problem is as follows:

$$\max f(X_1, X_2) := 200X_1 + 400X_2$$

s.t.

$$\begin{aligned}\frac{1}{40}X_1 + \frac{1}{60}X_2 &\leq 1.0 \\ \frac{1}{50}X_1 + \frac{1}{50}X_2 &\leq 1.0 \\ X_i &\geq 0 \quad i = 1, 2\end{aligned}\tag{5}$$

□

**The general LP formulation** Generalizing formulation 5, the general form for a Linear Programming problem is as follows:

**Objective Function:**

$$\max / \min f(X_1, X_2, \dots, X_n) := c_1X_1 + c_2X_2 + \dots + c_nX_n\tag{6}$$

s.t.

**Technological Constraints:**

$$a_{i1}X_1 + a_{i2}X_2 + \dots + a_{in}X_n \begin{pmatrix} \leq \\ = \\ \geq \end{pmatrix} b_i, \quad i = 1, \dots, m\tag{7}$$

**Sign Restrictions:**

$$(X_j \geq 0) \text{ or } (X_j \leq 0) \text{ or } (X_j \text{ urs}), \quad j = 1, \dots, n\tag{8}$$

where “urs” implies *unrestricted in sign*.

The formulation of Equations 6 to 8 has the general structure of a mathematical programming problem, presented in the introduction of this section, but it is further characterized by the fact that the functions involved in the problem objective and the left-hand-side of the technological constraints are *linear*. It is the assumptions implied by linearity that to a large extent determine the applicability of the above model in real-world applications.

To provide a better feeling of the linearity concept, let us assume that the different decision variables  $X_1, \dots, X_n$  correspond to various activities from which any solution will be eventually synthesized, and the values assigned to the variables by any given solution indicate the activity level in the considered plan(s). For instance, in the above example, the two activities are the production of items  $P_1$  and  $P_2$ , while the activity levels correspond to the daily production volume. Furthermore, let us assume that each technological constraint of Equation 7 imposes some restriction on the consumption of a particular resource. Referring back to the prototype example, the two problem resources are the daily production capacity of the two workstations  $W_1$  and  $W_2$ . Under this interpretation, the linearity property implies that:

**Additivity assumption:** the total consumption of each resource, as well as the overall objective value are the aggregates of the resource consumptions and the contributions to the problem objective, resulting by carrying out each activity independently, and

**Proportionality assumption:** these consumptions and contributions for each activity are proportional to the actual activity level.

It is interesting to notice how the above statement reflects to the logic that was applied when we derived the technological constraints of the prototype example: (i) Our assumption that the processing of each unit of product at every station requires a constant amount of time establishes the *proportionality* property for our model. (ii) The assumption that the total processing time required at every station to meet the production levels of both products is the aggregate of the processing times required for each product if the corresponding activity took place independently, implies that our system has an *additive* behavior. It is also interesting to see how the linearity assumption restricts the modeling capabilities of the LP framework: As an example, in the LP paradigm, we cannot immediately model effects like economies of scale in the problem cost structure, and/or situations in which resource consumption by one activity depends on the corresponding consumption by another complementary activity. In some cases, one can approach these more complicated problems by applying some *linearization* scheme. The resulting approximations for many of these cases have been reported to be quite satisfactory.

Another approximating element in many real-life LP applications results from the so called *divisibility* assumption. This assumption refers to the fact that for LP theory and algorithms to work, the problem variables must be *real*. However, in many LP formulations, meaningful values for the levels of the activities involved can be only *integer*. This is, for instance, the case with the production of items  $P_1$  and  $P_2$  in our prototype example. Introducing integrality requirements for some of the variables in an LP formulation turns the problem to one belonging in the class of (*Mixed*) *Integer Programming (MIP)*. The complexity of a MIP problem is much higher than that of LP's. Actually, the general IP formulation has been shown to belong to the notorious class of *NP-complete* problems. (This is a class of problems that have been "formally" shown to be extremely "hard" computationally). Given the increased difficulty of solving IP problems, sometimes in practice, near optimal solutions are obtained by solving the LP formulation resulting by relaxing the integrality requirements – known as the *LP relaxation* of the corresponding IP – and (judiciously) rounding off the fractional values for the integral variables in the optimal solution. Such an approach can be more easily justified in cases where the typical values for the integral variables are in the order of tens or above, since the errors introduced by the rounding-off are rather small, in a relative sense.

We conclude our discussion on the general LP formulation, by formally defining the solution search space and optimality. Specifically, we shall define as

the *feasible region* of the LP of Equations 6 to 8, the entire set of vectors  $\langle X_1, X_2, \dots, X_n \rangle^T$  that satisfy the technological constraints of Eq. 7 and the sign restrictions of Eq. 8. An *optimal* solution to the problem is any feasible vector that further satisfies the optimality requirement expressed by Eq. 6. In the next section, we provide a geometric characterization of the feasible region and the optimality condition, for the special case of LP's having only two decision variables.

## 2 Graphical solution of 2-var LP's

In this section, we develop a solution approach for LP problems, which is based on a geometrical representation of the feasible region and the objective function. In particular, the space to be considered is the  $n$ -dimensional space with each dimension defined by one of the LP variables  $X_j$ . The objective function will be described in this  $n$ -dim space by its *contour plots*, i.e., the sets of points that correspond to the same objective value. To the extent that the proposed approach requires the visualization of the underlying geometry, it is applicable only for LP's with upto three variables. Actually, to facilitate the easy visualization of the ensuing discussion and the concepts involved, in the following we shall restrict ourselves to the two-dimensional case, i.e., to LP's with two decision variables. In the next section, we shall generalize the geometry introduced here for the 2-var case, to the case of LP's with  $n$  decision variables, providing more analytic (algebraic) characterizations of these concepts and properties.

**Feasible Regions of Two-Var LP's** The primary idea behind the geometrical representation adopted in the subsequent analysis, is to correspond every vector  $\langle X_1, X_2 \rangle^T$  denoting the variables of a 2-var LP, to the point with co-ordinates  $(X_1, X_2)$  in a 2-dim (planar) *Cartesian* system. Under this correspondence, the feasible region of a 2-var LP is depicted by the set of points the coordinates of which satisfy the LP constraints and the sign restrictions. Since all these constraints are expressed by *linear* inequalities, to geometrically characterize the feasible region, we must first characterize the set of points that constitute the solution space of a linear inequality. Then, the LP feasible region will result from the intersection of the solution spaces corresponding to each technological constraint and/or sign restriction.

**The solution space of a single equality constraint** We start our investigation regarding the geometrical representation of 2-var linear constraints by considering first constraints of the equality type, i.e.,

$$a_1X_1 + a_2X_2 = b \tag{9}$$

It is a well-known result that, assuming  $a_2 \neq 0$ , this equation corresponds to a *straight line* with *slope*  $s = \Leftrightarrow \frac{a_1}{a_2}$  and *intercept*  $d = \frac{b}{a_2}$ . In the special case

where  $a_2 = 0$ , the solution space (*locus*) of Equation 9 is still a straight line perpendicular to the  $X_1$ -axis, intersecting it at the point  $(\frac{b}{a_1}, 0)$ . Notice that the presence of an equality constraint restricts the dimensionality of the feasible solution space by one degree of freedom, i.e., it turns it from a planar area to a line segment.

**The solution space of a single inequality constraint** Consider the constraint:

$$a_1 X_1 + a_2 X_2 \begin{pmatrix} \leq \\ \geq \end{pmatrix} b \quad (10)$$

The solution space of this constraint is one of the closed *half-planes* defined by the equation:  $a_1 X_1 + a_2 X_2 = b$ . To show this, let us consider a point  $(X_1, X_2)$  which satisfies Equation 10 as equality, and another point  $(X'_1, X'_2)$  for which Equation 10 is also valid. For any such pair of points, it holds that:

$$a_1(X'_1 \ominus X_1) + a_2(X'_2 \ominus X_2) \begin{pmatrix} \leq \\ \geq \end{pmatrix} 0 \quad (11)$$

Interpreting the left side of Eq. 11 as the *inner (dot) product* of the two vectors  $\mathbf{a} := \langle a_1, a_2 \rangle^T$  and  $\Delta \mathbf{X} := \langle X'_1 \ominus X_1, X'_2 \ominus X_2 \rangle^T$ , and recognizing that  $\mathbf{a} \cdot \Delta \mathbf{X} = |\mathbf{a}| |\Delta \mathbf{X}| \cos(\angle(\mathbf{a}, \Delta \mathbf{X}))$ , it follows that line  $a_1 X_1 + a_2 X_2 = b$ , itself, can be defined by point  $(X_1, X_2)$  and the set of points  $(X'_1, X'_2)$  such that vector  $\Delta \mathbf{X}$  is at right angles with vector  $\mathbf{a}$ . Furthermore, the set of points  $(X'_1, X'_2)$  that satisfy the  $>$  ( $<$ ) part of Equation 11 have the vector  $\Delta \mathbf{X}$  forming an acute (obtuse) angle with vector  $\mathbf{a}$ , and therefore, they are “above” (“below”) the line. Hence, the set of points satisfying each of the two inequalities implied by Equation 10 is given by one of the two half-planes the boundary of which is defined by the corresponding equality constraint. Figure 1 summarizes the above discussion.

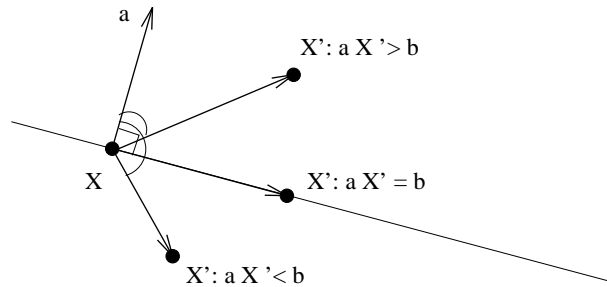


Figure 1: Half-planes: the feasible region of a linear inequality

An easy way to determine the half-plane depicting the solution space of a linear inequality, is to draw the line depicting the solution space of the corresponding equality constraint, and then test whether the point  $(0,0)$  satisfies the inequality. In case of a positive answer, the solution space is the half-space containing the origin, otherwise, it is the other one.

From the above discussion, it follows that the feasible region for the prototype LP of Equation 5 is the shaded area in the following figure:

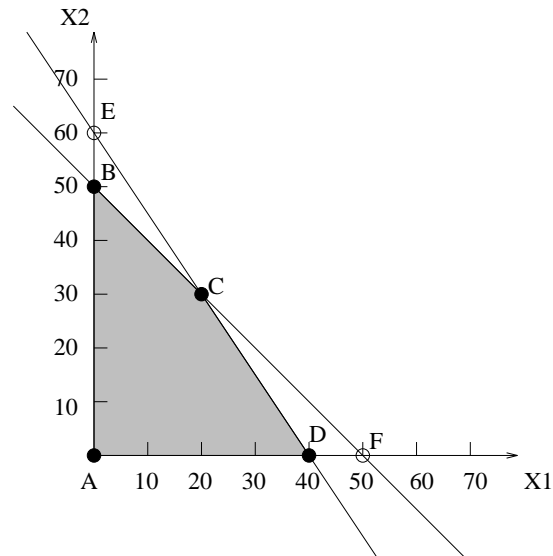


Figure 2: The feasible region of the prototype example LP

**Representing the Objective Function in the LP solution space** The most typical way to represent a two-variable function  $f(X_1, X_2)$  is to perceive it as a surface in an (orthogonal) three-dimensional space, where two of the dimensions correspond to the independent variables  $X_1$  and  $X_2$ , while the third dimension provides the function value for any pair  $(X_1, X_2)$ . However, in the context of our discussion, we are interested in expressing the information contained in the two-var LP objective function  $f(X_1, X_2)$  in the Cartesian plane defined by the two independent variables  $X_1$  and  $X_2$ . For this purpose, we shall use the concept of *contour plots*. Contour plots depict a function by identifying the set of points  $(X_1, X_2)$  that correspond to a constant value of the function  $f(X_1, X_2) = \alpha$ , for any given range of  $\alpha$ 's. The plot obtained for any fixed value of  $\alpha$  is a contour of the function. Studying the structure of a contour is expected to identify some patterns that essentially depict some useful properties of the

function.

In the case of LP's, the linearity of the objective function implies that any contour of it will be of the type:

$$c_1X_1 + c_2X_2 = \alpha \quad (12)$$

i.e., a straight line. For a maximization (minimization) problem, this line will be called an *isoprofit (isocost)* line. Assuming that  $c_2 \neq 0$  (o.w., work with  $c_1$ ), Equation 12 can be rewritten as:

$$X_2 = \Leftrightarrow \frac{c_1}{c_2}X_1 + \frac{\alpha}{c_2} \quad (13)$$

which implies that by changing the value of  $\alpha$ , the resulting isoprofit/isocost lines have constant slope and varying intercept, i.e, they are parallel to each other (which makes sense, since by the definition of this concept, isoprofit/isocost lines cannot intersect). Hence, if we continuously increase  $\alpha$  from some initial value  $\alpha_0$ , the corresponding isoprofit lines can be obtained by "sliding" the isoprofit line corresponding to  $f(X_1, X_2) = \alpha_0$  parallel to itself, in the direction of increasing or decreasing intercepts, depending on whether  $c_2$  is positive or negative.

**Graphical solution of the prototype example: a 2-var LP with a unique optimal solution** The "sliding motion" described above suggests a way for identifying the optimal values for, let's say, a max LP problem. The underlying idea is to keep "sliding" the isoprofit line  $c_1X_1 + c_2X_2 = \alpha_0$  in the direction of increasing  $\alpha$ 's, until we cross the boundary of the LP feasible region. The implementation of this idea on the prototype LP of Equation 5 is depicted in Figure 3.

From this figure, it follows that the optimal daily production levels for the prototype LP are given by the coordinates of the point corresponding to the intersection of line  $\frac{1}{50}X_1 + \frac{1}{50}X_2 = 0$  with the  $X_2$ -axis, i.e.,  $X_1^{opt} = 0$ ;  $X_2^{opt} = 50$ . The maximal daily profit is  $f(X_1^{opt}, X_2^{opt}) = 200 \cdot 0 + 400 \cdot 50 = 20,000$  (\$). Notice that the optimal point is one of the "corner" points of the feasible region depicted in Figure 3. Can you argue that for the geometry of the feasible region for 2-var LP's described above, if there is a bounded optimal solution, then there will be one which corresponds to one of the corner points? (This argument is developed for the broader context of n-var LP's in the next section.)

**2-var LP's with many optimal solutions** Consider our prototype example with the unit profit of item  $P_1$  being \$600 instead of \$200. Under this modification, the problem isoprofit lines become:

$$600X_1 + 400X_2 = \alpha \Leftrightarrow X_2 = \Leftrightarrow \frac{3}{2}X_1 + \frac{\alpha}{400}$$

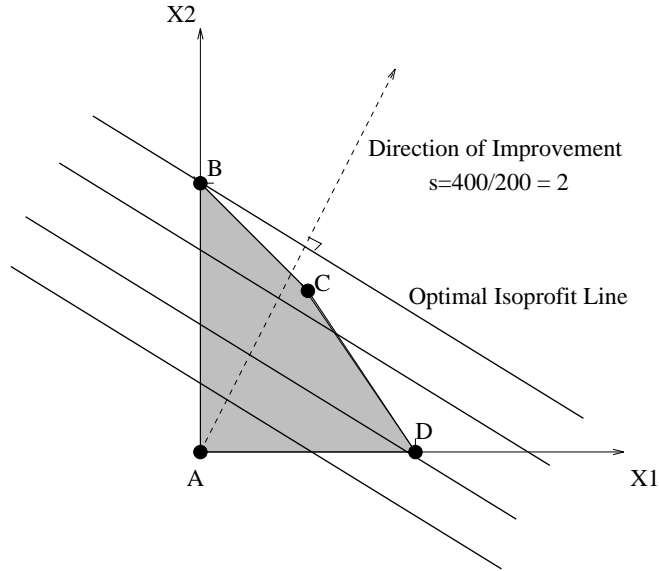


Figure 3: Graphical solution of the prototype example LP

and they are parallel to the line corresponding to the first problem constraint:

$$\frac{1}{40}X_1 + \frac{1}{60}X_2 = 1 \Leftrightarrow X_2 = \Leftrightarrow \frac{3}{2}X_1 + 60.$$

Therefore, if we try to apply the optimizing technique of the previous paragraph in this case, we get the situation depicted below, i.e., every point in the line segment  $CD$  is an optimal point, providing an optimal objective value of \$24,000. The situation is depicted in Figure 4.

It is worth-noticing that even in this case of many optimal solutions, we have two of them corresponding to “corner” points of the feasible region, namely points  $C$  and  $D$ .

**Infeasible 2-var LP’s** Consider again the original prototype example, modified by the additional requirements (imposed by the company’s marketing department) that the daily production of product  $P_1$  must be at least 30 units, and that of product  $P_2$  should exceed 20 units. These requirements introduce two new constraints into the problem formulation, i.e.,

$$X_1 \geq 30$$

$$X_2 \geq 20$$

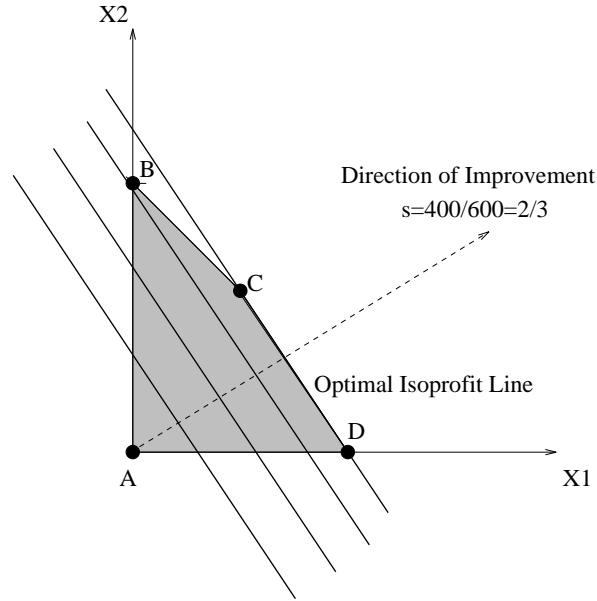


Figure 4: An LP with many optimal solutions

Attempting to plot the feasible region for this new problem, we get Figure 5, which indicates that there are no points on the  $(X_1, X_2)$ -plane that satisfy all constraints, and therefore our problem is *infeasible* (over-constrained).

**Unbounded 2-var LP's** In the LP's considered above, the feasible region (if not empty) was a bounded area of the  $(X_1, X_2)$ -plane. For this kind of problems it is obvious that all values of the LP objective function (and therefore the optimal) are bounded. Consider however the following LP:

$$\max f(X_1, X_2) := 2X_1 \Leftrightarrow X_2$$

s.t.

$$X_1 \Leftrightarrow X_2 \leq 1$$

$$2X_1 + X_2 \geq 6$$

$$X_1, X_2 \geq 0$$

The feasible region and the direction of improvement for the isoprofit lines for this problem are given in Figure 6.

It is easy to see that the feasible region of this problem is unbounded, and furthermore, the orientation of the isoprofit lines is such that no matter how

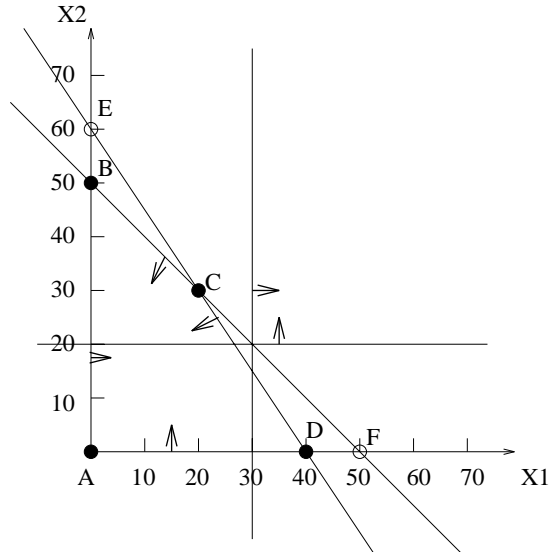


Figure 5: An infeasible LP

far we “slide” these lines in the direction of increasing the objective function, they will always share some points with the feasible region. Therefore, this is an example of a (2-var) LP whose objective function can take arbitrarily large values. Such an LP is characterized as *unbounded*. Notice, however, that even though an unbounded feasible region is a necessary condition for an LP to be unbounded, it is not sufficient; to convince yourself, try to graphically identify the optimal solution  $f$  for the above LP in the case that the objective function is changed to:  $\max f(X_1, X_2) := \Leftrightarrow X_2$ .

Summarizing the above discussion, we have shown that a 2-var LP can either

- have a *unique* optimal solution which corresponds to a “corner” point of the feasible region, or
- have *many* optimal solutions that correspond to an entire “edge” of the feasible region, or
- be *unbounded*, or
- be *infeasible*.

In the next section, we generalize this geometrical description of the LP solution space for the  $n$ -var LP case, and we provide a brief (informal) derivation

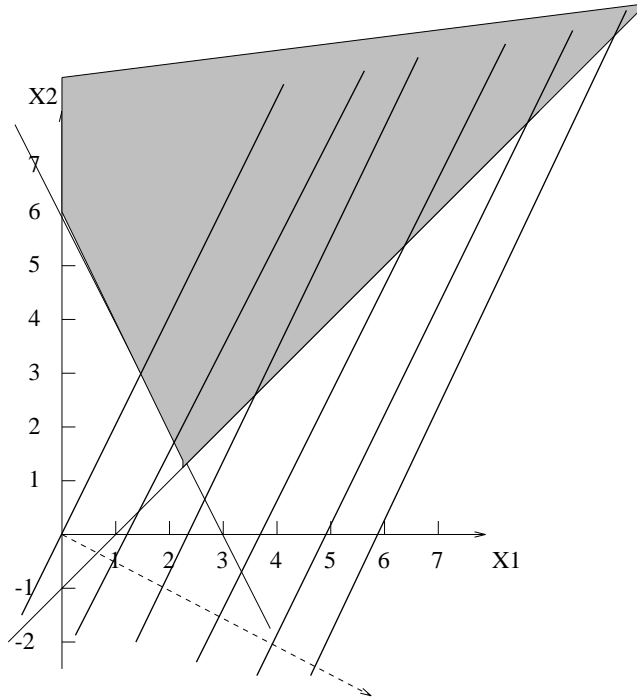


Figure 6: An unbounded LP

of the *Fundamental Theorem of Linear Programming*. The latter states that if an LP has a bounded optimal solution, then it must have one which is an *extreme point* of its feasible region. The Simplex algorithm to be discussed in the last section of this set of notes essentially exploits this fundamental result to reduce the space to be searched for an optimal solution.

### 3 Generalization to the $n$ -var case: the “geometry” of the LP feasible region and the Fundamental Theorem of Linear Programming

In this section we generalize the geometry of the 2-var LP’s, presented in the previous section, to LP’s with  $n$  decision variables. The study of the “geometric” properties of the LP feasible region for this general case, will eventually lead to the *Fundamental Theorem of Linear Programming*, which is at the basis of the *Simplex algorithm*.

**The geometry of  $n$ -var LP's**  $n$ -var LP's require an  $n$ -dimensional space to "geometrically" represent a vector corresponding to a pricing of their decision variables. However, the concepts and techniques that allowed the geometric representation of the 2-var LP's in the previous section, generalize quite straightforwardly in this more complicated case.

Hence, given a linear constraint:

$$a_1X_1 + a_2X_2 + \dots + a_nX_n \begin{pmatrix} \leq \\ = \\ \geq \end{pmatrix} b \quad (14)$$

and a point  $\mathbf{X}_0 = \langle X_{01}, X_{02}, \dots, X_{0n} \rangle^T$  satisfying Equation 14 as equality, we can perceive the solution space of

- Equation  $a_1X_1 + a_2X_2 + \dots + a_nX_n = b$  as the set of points  $\mathbf{X}$  for which the vector  $\mathbf{DX} = \mathbf{X} \ominus \mathbf{X}_0$  is at right angles with vector  $\mathbf{a} = \langle a_1, a_2, \dots, a_n \rangle$ ,
- Inequality  $a_1X_1 + a_2X_2 + \dots + a_nX_n > b$  as the set of points  $\mathbf{X}$  for which the vector  $\mathbf{DX} = \mathbf{X} \ominus \mathbf{X}_0$  forms an acute angle with vector  $\mathbf{a} = \langle a_1, a_2, \dots, a_n \rangle$ , and
- Inequality  $a_1X_1 + a_2X_2 + \dots + a_nX_n < b$  as the set of points  $\mathbf{X}$  for which the vector  $\mathbf{DX} = \mathbf{X} \ominus \mathbf{X}_0$  forms an obtuse angle with vector  $\mathbf{a} = \langle a_1, a_2, \dots, a_n \rangle$ .

From the above description, it follows that the solution space of the equation  $a_1X_1 + a_2X_2 + a_3X_3 = b$ , i.e., the 3-dim case, is a *plane* perpendicular to vector  $\langle a_1, a_2, a_3 \rangle^T$ , which is characterized as the plane *normal*. Since this normality concept carries over to the more general  $n$ -dim case, we characterize the solution space of an  $n$ -var linear equation as a *hyperplane*. Then, similar to the 2-var LP case, a hyperplane defined by the equation  $a_1X_1 + a_2X_2 + \dots + a_nX_n = b$ , divides the  $n$ -dim space into two *half-spaces*: one of them is the solution space of the inequality  $a_1X_1 + a_2X_2 + \dots + a_nX_n \geq b$ , and the other one is the solution space of the inequality  $a_1X_1 + a_2X_2 + \dots + a_nX_n \leq b$ .

Hence, the solution space (*feasible region*) of an  $n$ -var LP is geometrically defined by the intersection of a number of half-spaces and/or hyperplanes equal to the LP constraints, including the *sign restrictions*. Such a set is characterized as a *polytope*. In particular, a bounded polytope is called a *polyhedron*.

Two other concepts of interest in the following discussion are those of the *straight line* and the *line segment*. Given two points  $\mathbf{X}_1 = \langle X_{11}, X_{12}, \dots, X_{1n} \rangle^T$  and  $\mathbf{X}_2 = \langle X_{21}, X_{22}, \dots, X_{2n} \rangle^T$ , the *straight line* passing through them is algebraically defined by the set of points

$$\mathbf{X}_1 + \kappa(\mathbf{X}_2 \ominus \mathbf{X}_1), \quad \kappa \in \mathfrak{R}. \quad (15)$$

Figure 7 provides a visualization of this definition.

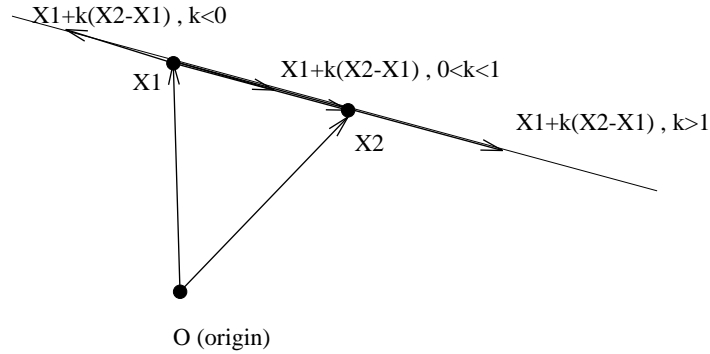


Figure 7: The equation of a straight line

As it is seen in the figure above, the *line segment* between points  $\mathbf{X}_1$  and  $\mathbf{X}_2$  is analytically defined by:

$$\mathbf{X}_1 + \kappa(\mathbf{X}_2 - \mathbf{X}_1), \quad \kappa \in [0, 1]. \quad (16)$$

Notice that Equation 16 can be rewritten as:

$$(1 - \kappa)\mathbf{X}_1 + \kappa\mathbf{X}_2, \quad \kappa \in [0, 1]. \quad (17)$$

or

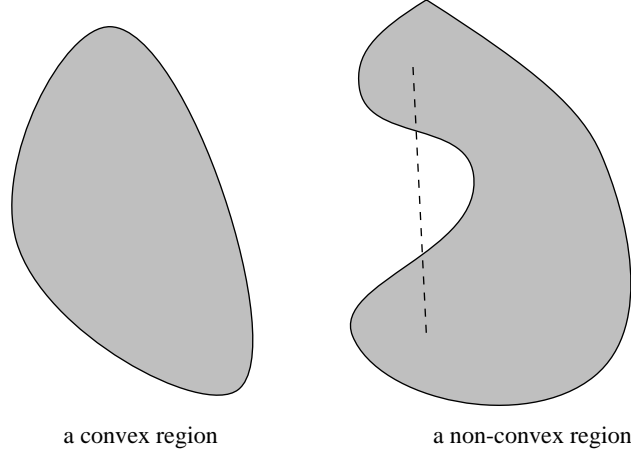
$$\begin{aligned} \mu\mathbf{X}_1 + \kappa\mathbf{X}_2 \\ \mu + \kappa = 1 \\ \mu, \kappa \geq 0 \end{aligned} \quad (18)$$

We say that Equations 17 and 18 define a *convex combination* of points  $\mathbf{X}_1$  and  $\mathbf{X}_2$ .

**Polytope Convexity and Extreme Points** In this paragraph we show that polytopes are *convex* sets. This property is important for eventually proving the Fundamental Theorem of LP. A set of points  $S$  in the  $n$ -dim space is *convex*, if the line segment connecting any two points  $\mathbf{X}_1, \mathbf{X}_2 \in S$ , belongs completely in  $S$ . According to the previous mathematical definition of line segments, this definition of convexity is mathematically expressed as:

$$\begin{aligned} S \text{ convex} &\iff \\ \forall \mathbf{X}_1, \mathbf{X}_2 \in S, \forall \kappa \in (0, 1) : &(1 - \kappa)\mathbf{X}_1 + \kappa\mathbf{X}_2 \in S \end{aligned} \quad (19)$$

Figure 8 depicts the concept.



a convex region

a non-convex region

Figure 8: Convex sets

To show that a polytope is a convex set, we first establish that the solution space of any linear constraint (i.e., hyperplanes and half-spaces) is a convex set. Since a polytope is the *intersection* of a number of hyperplanes and half-spaces, the convexity of the latter directly implies the convexity of the polytope (i.e., a line segment belonging to each defining hyperplane and/or half-space will also belong to the polytope).

To establish the convexity of the feasible region of a linear constraint, let's consider the constraint:

$$a_1 X_1 + a_2 X_2 + \dots + a_n X_n \begin{pmatrix} \leq \\ = \\ \geq \end{pmatrix} b \quad (20)$$

and two points  $\mathbf{X}_1 = \langle X_{11}, X_{12}, \dots, X_{1n} \rangle^T$ ,  $\mathbf{X}_2 = \langle X_{21}, X_{22}, \dots, X_{2n} \rangle^T$  satisfying it. Then, for any  $\kappa \in (0, 1)$ , we have:

$$a_1(\kappa X_{11}) + a_2(\kappa X_{12}) + \dots + a_n(\kappa X_{1n}) \begin{pmatrix} \leq \\ = \\ \geq \end{pmatrix} \kappa b \quad (21)$$

$$a_1[(1 \ominus \kappa)X_{21}] + a_2[(1 \ominus \kappa)X_{22}] + \dots + a_n[(1 \ominus \kappa)X_{2n}] \begin{pmatrix} \leq \\ = \\ \geq \end{pmatrix} (1 \ominus \kappa)b \quad (22)$$

Adding Equations 21 and 22, we get:

$$a_1[\kappa X_{11} + (1 \leftrightarrow \kappa)X_{21}] + a_2[\kappa X_{12} + (1 \leftrightarrow \kappa)X_{22}] + \dots + a_n[\kappa X_{1n} + (1 \leftrightarrow \kappa)X_{2n}] \begin{pmatrix} \leq \\ = \\ \geq \end{pmatrix} b \quad (23)$$

i.e., point  $\kappa X_1 + (1 \leftrightarrow \kappa)X_2$  belongs in the solution space of the constraint.

A last concept that we must define before the statement of the *Fundamental Theorem of LP*, is that of the *extreme point* of a convex set. Given a *convex* set  $S$ , point  $X_0 \in S$  is an *extreme point*, if each line segment that lies completely in  $S$  and contains point  $X_0$ , has  $X_0$  as an end point of the line segment. Mathematically,

$$X_0 \text{ is an extreme point of } S \iff$$

$$(X_0 = \kappa X_1 + (1 \leftrightarrow \kappa)X_2, X_1, X_2 \in S, \kappa \in (0, 1) \Rightarrow X_0 = X_1 = X_2) \quad (24)$$

**The Fundamental Theorem of Linear Programming** Having established all the necessary concepts and properties of the solution space of  $n$ -var LP's, we are now ready to discuss the Fundamental Theorem of Linear Programming. This theorem can be stated as follows:

**Theorem 1** *If an LP has a (bounded) optimal solution, then there exists an extreme point of the feasible region which is optimal.*

**Proof (Sketch):** We establish the validity of Theorem 1, through a series of observations:

1. First notice that according to the previous discussion, the feasible region of an LP is a *polytope*, and thus, *convex*.
2. Furthermore, since we assume that the LP has an optimal solution, let  $X_0$  denote such an optimal point. The optimal objective value will be denoted by  $z^* = \mathbf{c} \cdot X_0 = c_1 X_{01} + c_2 X_{02} + \dots + c_n X_{0n}$ .
3. Then notice that point  $X_0$  cannot be interior to a line segment that is not perpendicular to the direction of improvement to the "*isoprofit*" hyperplanes – w.l.o.g., let's assume a maximization LP for our discussion – defined by vector  $\mathbf{c}$ . Otherwise, by moving on this line segment in the direction of improvement of the "isoprofit" hyperplanes, we would be able to obtain another point  $X_1$  of the feasible region, such that  $\mathbf{c} \cdot X_1 > \mathbf{c} \cdot X_0$ . But this contradicts the assumption that  $X_0$  is an optimal point. Figure 9 depicts this argument.
4. However, point  $X_0$  can be interior to a line segment of the feasible solution space which is perpendicular to the direction of improvement of the optimal "isoprofit" hyperplane. This, in fact, corresponds to a situation

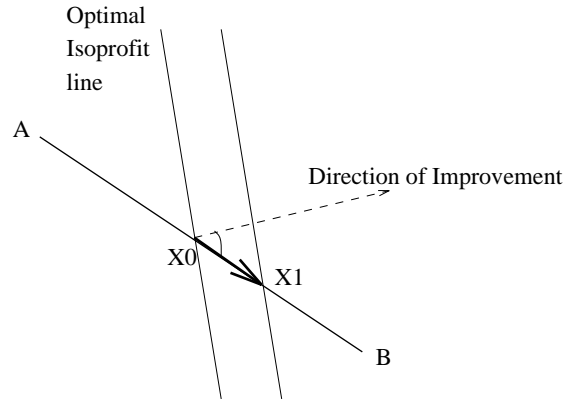


Figure 9: Why an optimal solution to an LP cannot be interior to a line segment not perpendicular to the optimal isoprofit line

of many optimal solutions. In this case, notice that this line segment must have at least one end  $\mathbf{X}_1$  defined by the fact that one (or more) additional constraint is binding at this point. Otherwise, the problem is ill-posed, since we can vary some variable(s) at will over  $(-\infty, \infty)$  with this variation affecting neither the constraints nor the objective.

5. Hence,  $\mathbf{X}_1$ , being on the optimal “isoprofit” hyperplane, is another optimal point at which an additional constraint is binding. Then, there are two possibilities: (i)  $\mathbf{X}_1$  is an extreme point of the feasible region, in which case we are done, or (ii)  $\mathbf{X}_1$  is interior point to another line segment lying in the optimal “isoprofit” hyperplane  $\mathbf{c} \cdot \mathbf{X}_0 = \mathbf{c} \cdot \mathbf{X}_1 = z^*$ , which binds, however, an additional constraint, compared to point  $\mathbf{X}_0$ . In this case, repeating the argument above, we establish the existence of another end point  $\mathbf{X}_2$ , determined by the binding of at least one more constraint. Then, we repeat the entire argument for  $\mathbf{X}_2$ , and so on. Figure 10 depicts this part of the proof.
6. Finally, notice that every time we bind an additional constraint, we restrict the (sub-)space of optimal solutions considered by one “degree of freedom”. Since an  $n$ -dim space has  $n$  “degrees of freedom”, the number of end points visited in the argument above before we find one that it is an extreme point is *finite*. Thus, this last observation establishes the existence of an optimal extreme point for the case of many optimal solution, and the proof is complete.

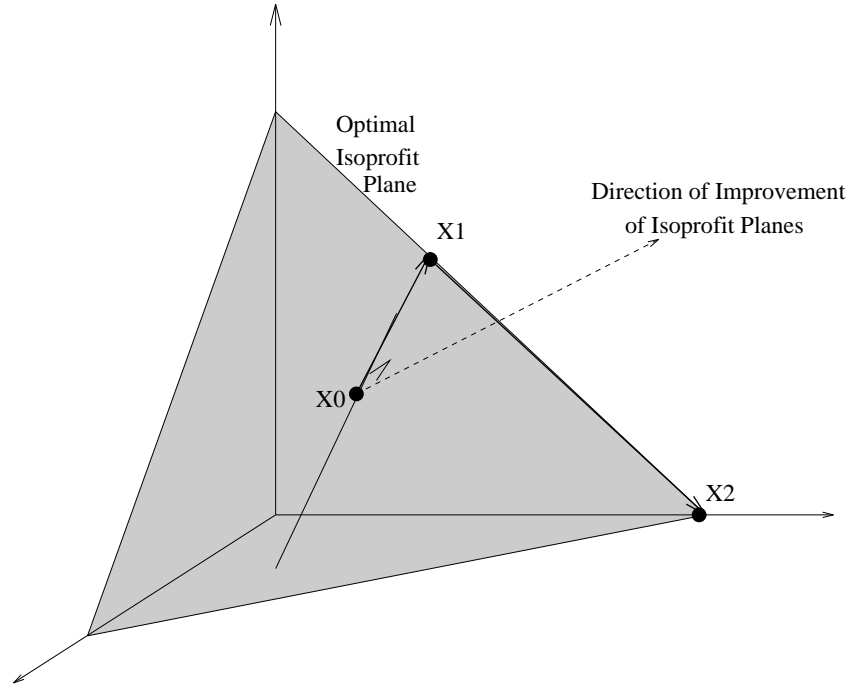


Figure 10: Identifying an optimal extreme point on the optimal isoprofit hyperplane

The discourse of the previous proof has also revealed a very important property of extreme points: At these points, the number of binding constraints is such that it allows *zero* “degrees of freedom”, or, in other words, these constraints define the point *uniquely*. Starting from this observation, in the next section we provide a series of *algebraic* characterizations of the *extreme points*, which will eventually allow us to analytically manipulate the set of extreme points of an LP, in the context of the Simplex algorithm. As it has been previously mentioned, this algorithm exploits the result stated in the *Fundamental Theorem* above, by limiting the search for an optimal solution over the set of extreme points of the polytope defining the LP feasible region. As we shall show in the next section, an important property of this set is that it is *finite* and *discrete*, so that it can even be exhaustively enumerated. Simplex algorithm provides an efficient way to search this set.

## 4 An algebraic characterization of the solution search space: Basic Feasible Solutions

In the previous section we showed that if an LP has a (bounded) optimal solution, then it has (at least) one which corresponds to an extreme point of its feasible region. To exploit this result algorithmically, we need an algebraic characterization of the *extreme point* concept. This is the topic of this section.

The starting point of this discussion is the observation made at the end of the previous section, that at an extreme point, the set of binding constraints is such that it characterizes the point *uniquely*. Let's try to investigate what is the algebraic structure implied by this statement. Also, staying close to the general spirit of our discussion, let's examine this issue in an *inductive* manner.

- We know that in the 1-dim space, i.e., on the line of real numbers, a point can be identified uniquely by an equation  $aX = b$ , where  $a \neq 0$ .
- In the 2-dim space, a linear equation  $a_1X_1 + a_2X_2 = b$  defines a line, i.e., a subspace with 1 “degree of freedom”, while the definition of a unique point requires a system of two linear equations:

$$a_{11}X_1 + a_{12}X_2 = b_1$$

$$a_{21}X_1 + a_{22}X_2 = b_2$$

with a *unique* solution, i.e., with

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0$$

Such a system of equations is characterized as *linearly independent*, and geometrically, it corresponds to two intersecting straight lines.

- In the 3-dim space, a linear equation  $a_1X_1 + a_2X_2 + a_3X_3 = b$  corresponds to a *plane* perpendicular to the vector  $\langle a_1, a_2, a_3 \rangle^T$ . A system of two linear equations:

$$a_{11}X_1 + a_{12}X_2 + a_{13}X_3 = b_1$$

$$a_{21}X_1 + a_{22}X_2 + a_{23}X_3 = b_2$$

for which:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0, \text{ or } \begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} \neq 0, \text{ or } \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} \neq 0$$

corresponds to the intersection of two planes, i.e., a *straight line*.

Defining a unique point in the 3-dim space requires three *linearly independent* equations, i.e.,

$$a_{11}X_1 + a_{12}X_2 + a_{13}X_3 = b_1$$

$$a_{21}X_2 + a_{22}X_2 + a_{23}X_3 = b_2$$

$$a_{31}X_2 + a_{32}X_2 + a_{33}X_3 = b_3$$

with

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \neq 0.$$

- In a similar fashion, in the  $n$ -dim space, a point is uniquely defined by  $n$  linear equations which are *linearly independent*, i.e.,

$$a_{11}X_1 + a_{12}X_2 + \dots + a_{1n}X_n = b_1$$

$$a_{21}X_2 + a_{22}X_2 + \dots + a_{23}X_n = b_2$$

⋮

$$a_{n1}X_2 + a_{n2}X_2 + \dots + a_{nn}X_n = b_3$$

with

$$\begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ & & \vdots & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} \neq 0.$$

**Example:** We demonstrate the findings of the above discussion on the feasible region of the prototype example, which for convenience is reproduced in Figure 11.

As we can see in the figure, each of the extreme points of this region corresponds to the binding of a pair of the LP constraints, i.e.,

- point A corresponds to the binding of the two sign restriction constraints,
- point B corresponds to the binding of the second technological constraint and the sign restriction of variable  $X_2$ ,
- point C corresponds to the binding of both technological constraints, and
- point D corresponds to the binding of the first technological constraint and the sign restriction of variable  $X_1$ .

Notice, however, that points E and F, even though they are defined by the binding of the constraint pairs (tech. con. 1, sign res. of  $X_2$ ) and (tech. con. 2, sign res. of  $X_1$ ), respectively, are not extreme points of the feasible region just because they are infeasible (i.e., some other LP constraints are violated). Hence, having  $n$  linearly independent constraints binding at certain point is a

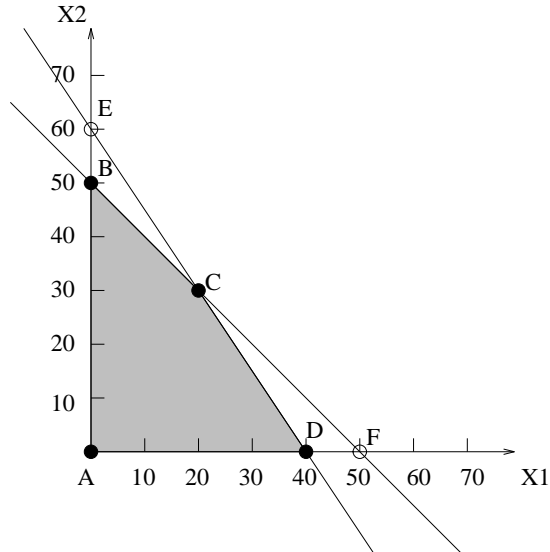


Figure 11: The feasible region of the prototype example LP

*necessary* condition for it to be an extreme point of the feasible region, but not *sufficient*.  $\square$

Finally, it should be easy to see that for an  $n$ -var LP,  $n$  is the *minimum* number of constraints binding at an extreme point. If more than this minimum number of constraints are binding to an extreme point, the point (and the corresponding solution) are characterized as *degenerate*. As we shall see in a later section, *degeneracy* can complicate the search for the optimal solution carried out by the Simplex method.

**LP's in "standard form"** To further exploit the previous characterization of extreme points as the solution to  $n$  binding linearly independent constraints, we must define the concept of LP's in "*standard form*". An LP is said to be in "*standard form*", if: (i) all *technological constraints* are *equality* constraints, and (ii) all the variables have a *nonnegativity sign restriction*.

Every LP can be brought into "standard form" through the following transformations:

- an inequality constraint:

$$a_1 X_1 + a_2 X_2 + \dots + a_n X_n \begin{pmatrix} \leq \\ \geq \end{pmatrix} b$$

can be converted into an equality one, through the introduction of a *slack* (*excess*) variable  $S$  ( $E$ )  $\geq 0$ :

$$a_1 X_1 + a_2 X_2 + \dots + a_n X_n + \begin{pmatrix} S \\ \Leftrightarrow E \end{pmatrix} = b.$$

- A variable  $X_i$  with sign restriction  $X_i \leq 0$  can substituted by  $X_i = \Leftrightarrow X'_i$  with  $X'_i \geq 0$ .
- Finally, a *urs* variable  $X_i$  can be substituted by  $X_i = X'_i \Leftrightarrow X''_i$  with  $X'_i, X''_i \geq 0$ .

**Basic Feasible Solutions: An algebraic characterization of extreme points for LP's in "standard form"** For LP's in "standard form" the previous characterization of extreme points as the solution of  $n$  linearly independent binding constraints, which is, furthermore, a feasible point, can become even more concise. Consider, for instance, the LP

$$\begin{aligned} \mathbf{AX} &\leq \mathbf{b} \\ \mathbf{X} &\geq 0 \end{aligned} \tag{25}$$

with  $m$  technological constraints and  $n$  variables. In "standard form", it becomes:

$$\begin{aligned} \mathbf{AX} + \mathbf{IS} &= \mathbf{b} \\ \mathbf{X}, \mathbf{S} &\geq 0 \end{aligned} \tag{26}$$

where  $I$  is the  $m \times m$  identity matrix, and  $\mathbf{S} = \langle S_1, S_2, \dots, S_m \rangle^T$  is the vector of *slack* variables. It is interesting to notice that for every binding constraint from the  $m+n$  constraints of the original formulation (Eq. 25), one of the  $m+n$  variables of the "standard form" formulation must be equal to 0. Specifically, if the  $i$ -th technological constraint is binding, the corresponding slack variable  $S_i = 0$ . Similarly, if one of the sign restriction constraints is binding, then the corresponding variable  $X_j = 0$ . Since an extreme point of the feasible region of formulation 25 requires  $n$  binding constraints for its definition, it follows that  $n$  from the  $n+m$  variables in the corresponding solution of the "standard form" formulation (Eq. 26) must be equal to zero. Furthermore, since this point is uniquely defined, the system of equations defined by the  $m$  technological constraints in the "standard form" formulation and the  $m$  remaining variables, must have a unique solution. In other words, the columns of the "standard form" formulation corresponding to these  $m$  variables must be linearly independent (and their determinant must have a nonzero value). Finally, since the extreme point considered belongs in the feasible region of the problem, it follows that the unique solution of the aforementioned system of  $m$  equations in the  $m$

nonzero variables must be nonnegative (to meet the sign restrictions required by “standard form”).

The structure of the “standard form” solutions corresponding to extreme points of the original feasible region (i.e., that defined with respect to the primary LP variables  $X_i$ ), described for the example above, actually applies to any other LP in “standard form”. We formally characterize this structure through the definition of *basic feasible solutions* for LP’s in “standard form” (taken from Winston, “*Introduction to Mathematical Programming*”).

**Definition 1** Consider the system  $AX = \mathbf{b}$  of  $m$  linear equations in  $N$  variables, corresponding to the technological constraints of an LP in “standard form”.

1. A basic solution to this system is obtained by setting  $N \ominus m$  variables equal to zero, and solving for the values of the remaining  $m$  variables. This assumes that setting the  $N \ominus m$  variables equal to zero yields unique values for the remaining  $m$  variables, or, equivalently, the columns in the  $A$  matrix for the remaining  $m$  variables are linearly independent. The  $m$  variables which are not bound to zero, are the basic variables (or equivalently, they define the basis) of the basic solution under consideration.
2. Any basic solution to  $AX = \mathbf{b}$  in which all variables are nonnegative is a basic feasible solution (bfs).

**Example:** Returning to the prototype example, it is easy to see that its “standard form” formulation is as follows:

$$\max f(X_1, X_2) := 200X_1 + 400X_2$$

s.t.

$$\begin{aligned} \frac{1}{40}X_1 + \frac{1}{60}X_2 + S_1 &= 1.0 \\ \frac{1}{50}X_1 + \frac{1}{50}X_2 + S_2 &= 1.0 \\ X_i, S_i &\geq 0 \quad i = 1, 2 \end{aligned} \tag{27}$$

This formulation involves four variables and two technological constraints. Therefore, any basic solution will be defined by selecting a basis of two variables, with the remaining two being set equal to zero. For example, selecting as basis the variable set  $\{X_1, X_2\}$  implies that  $S_1 = S_2 = 0$  (since they are the remaining non-basic variables). This further implies that the considered basic solution corresponds to the extreme point of the LP feasible region defined by the binding of the two technological constraints. Finally, the values of the two basic variables  $X_1$  and  $X_2$  are obtained by solving the system of equations:

$$\frac{1}{40}X_1 + \frac{1}{60}X_2 = 1.0$$

$$\frac{1}{50}X_1 + \frac{1}{50}X_2 = 1.0$$

which results from the technological constraints in “standard form”, by eliminating the non-basic variables.

On the other hand, consider the extreme point  $B$  of the feasible region, which is the optimal solution of this example LP. It has been already shown (cf. previous example) that this point is defined by the binding of the second technological constraint and the sign restriction imposed on variable  $X_1$ . Hence, at this point,  $S_2 = X_1 = 0$ , and the corresponding *basis* consists of variables  $X_2$  and  $S_1$ . To compute the values for these variables, we solve the system of equations:

$$\begin{aligned}\frac{1}{60}X_2 + S_1 &= 1.0 \\ \frac{1}{50}X_2 &= 1.0\end{aligned}$$

Finally, the basic solution defined by the basis  $\{X_1, S_1\}$  implies that  $X_2 = S_2 = 0$ , and therefore, the corresponding point on the  $(X_1, X_2)$ -plane,  $F$ , is defined by the binding of second technological constraint and the sign restriction of variable  $X_2$ . Solving the system of equations:

$$\begin{aligned}\frac{1}{40}X_1 + S_1 &= 1.0 \\ \frac{1}{50}X_1 &= 1.0\end{aligned}$$

we obtain:  $X_1 = 50$ ,  $S_1 = 1/4$ . This basic solution is not feasible, which is also reflected in Figure 2 by the fact that point  $F$  is not an extreme point of the feasible region.  $\square$ .

In the example above, extreme points  $B$  and  $C$  are *adjacent*, in the sense that they are linked by one edge of the feasible region. This reflects in the structure of the corresponding bases by the fact that they differ in only one binding constraint. This observation generalizes to the  $n$ -dimensional case: extreme points connected by “*edges*” of the feasible region have  $n \Leftrightarrow 1$  common binding constraints, and therefore, their corresponding *bases* will differ in one variable only. Hence, we have the following definition:

**Definition 2** *Two basic feasible solutions of an LP with  $m$  technological constraints in “standard form” is said to be adjacent, if their bases have  $m \Leftrightarrow 1$  variables in common.*

The characterization of the extreme points of the feasible region of an LP as basic feasible solutions for its “standard form” representation provides the analytical means for organizing the search for an optimal extreme point performed by the Simplex algorithm. The details of this algorithm is the topic of the next section.

## 5 The Simplex Algorithm

In the previous sections we have established the following two important results:

1. If an LP has a bounded optimal solution, then there exists an *extreme point* of the feasible region which is optimal.
2. Extreme points of the feasible region of an LP correspond to *basic feasible solutions* of its “standard form” representation.

The first of these results implies that in order to obtain an optimal solution of an LP, we can constrain our search on the set of the extreme points of this feasible region. The second result provides an algebraic characterization of this set: each of these points is determined by selecting a set of *basic* variables, with cardinality equal to the number of the technological constraints of the LP, and the additional requirement that the (uniquely determined) values of these variables are *nonnegative* (cf. discussion on basic feasible solutions). This further implies that the set of extreme points for an LP with  $m$  technological constraints and  $N$  variables in its “standard form” representation can have only a *finite* number of extreme points; specifically,  $\binom{N}{m} = \frac{N!}{m!(N-m)!}$  is an upper bound for the cardinality of this set.

The last observation would make one think that a (naive) approach to the problem would be to enumerate the entire set of extreme points, compare their corresponding objective values, and eventually select one which minimizes the objective function over this set. Such an approach would actually work for rather small formulations. But for reasonably sized LP’s, the set of extreme points, even though finite, can become extremely large. For example, a small LP with 10 variables (in “standard form”) and 3 technological constraints can have upto 120 extreme points, while an LP with 100 variables and 20 constraints can have upto  $5.36 \times 10^{20}$  extreme points. And yet, this is a rather small LP!

Hence, we need a more systematic approach to organize the search so that we manage the complexity resulting from the size of the search space. Such a systematic approach is provided by the *Simplex* algorithm. The basic logic of the algorithm is depicted in Figure 12.

The algorithm starts with an initial basic feasible solution (bfs) and tests its optimality. If some optimality condition is verified, then the algorithm terminates. Otherwise, the algorithm identifies an *adjacent* bfs, with a better objective value. The optimality of this new solution is tested again, and the entire scheme is repeated, until an optimal bfs is found. Since every time a new bfs is identified the objective value is improved (except from a certain pathological case that we shall see later), and the set of bfs’s is finite, it follows that the algorithm will terminate in a *finite* number of steps (iterations). It is also interesting to examine the *geometrical* interpretation of the behavior of Simplex algorithm. Given the above description of the algorithm and the correspondence

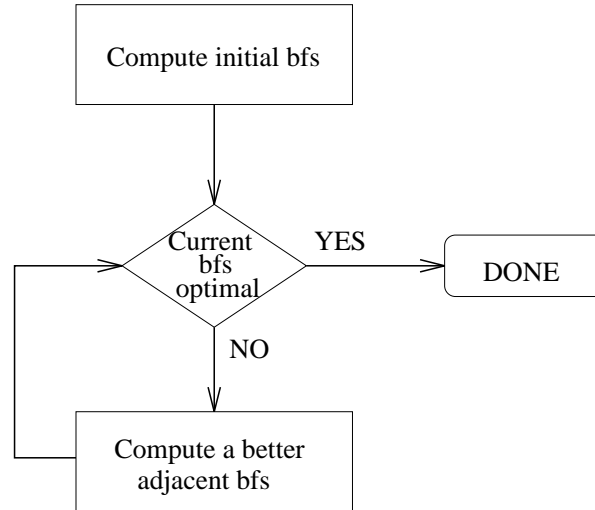


Figure 12: The basic Simplex logic

of bfs's to extreme points, it follows that Simplex essentially starts from some initial extreme point, and follows a path along the edges of the feasible region towards an optimal extreme point, such that all the intermediate extreme points visited are improving (more accurately, not worsening) the objective function.

In the following, we explain how the Simplex algorithm implements this logic in its computations by applying it on our prototype LP.

**The basic Simplex iteration through an example:** Consider our prototype LP in standard form, repeated below for convenience:

$$\max z = f(X_1, X_2) := 200X_1 + 400X_2$$

s.t.

$$\frac{1}{40}X_1 + \frac{1}{60}X_2 + S_1 = 1.0$$

$$\frac{1}{50}X_1 + \frac{1}{50}X_2 + S_2 = 1.0$$

$$X_i, S_i \geq 0 \quad i = 1, 2 \tag{28}$$

**Finding an initial bfs** To start the Simplex algorithm on this problem, we need to identify an initial bfs. For this particular problem, a bfs will have two basic variables, since we have two technological constraints. Taking a closer look to the structure of these constraints in Equation 28, it can be seen that a

convenient selection is  $B_0 = \{S_1, S_2\}$ , where  $B_0$  denotes the set of basic variables (*basis*). Indeed, setting  $X_1 = X_2 = 0$ , we readily obtain,  $S_1 = 1.0$ ;  $S_2 = 1.0$ . The easy computation of the values of the basic variables was the result of the fact that each of these variables could be associated with one and only one constraint. More specifically, (i) each of these variables showed up in only one constraint, (ii) the coefficient of the variable in that constraint was equal to 1.0, and (iii) any pair of basic variables showed up in different constraints. An LP the constraints of which satisfy these three properties with respect to a certain basis, is said to be in *canonical form* with respect to that basis. However, notice that the *feasibility* of basis  $B_0 = \{S_1, S_2\}$  was established by the fact that the *right-hand-side* coefficients of the constraints in their canonical form with respect to basis  $B_0$  are non-negative. We shall address the problem of how to compute an initial bfs in the more general case where one is not readily available by inspection, in a later section.

**Testing bfs optimality: Checking the sign of the objective function coefficients for nonbasic variables** Notice that the LP objective value corresponding to basis  $B_0$  is determined by the fact that  $X_1 = X_2 = 0$ , and therefore,  $z = 0$ . So, we pose the question: Is it possible that there exists another bfs with a better objective value? Obviously, any bfs for which  $X_1$  and/or  $X_2$  is a basic variable has a considerable chance of having a better (i.e., strictly positive) objective value, since the objective function coefficients for these variables are positive numbers. Hence, the answer to the previous question is that the *non-negativity* of the objective function coefficients of the nonbasic variables  $X_1, X_2$  implies that there is potential for improvement of the objective value,  $z$ , and both variables  $X_1, X_2$  are good candidates for entering the basis.

**Selecting the entering variable** Given that at every objective-improving iteration the Simplex algorithm considers *adjacent* bfs's, it follows that only one of the candidate nonbasic variables will eventually enter the basis. Typically, the variable selected to enter the basis is the one that will incur the maximum improvement to the objective value per unit of increase of the variable. In our case, this translates to selecting the (nonbasic) variable with the *most positive* coefficient in the objective function, i.e., variable  $X_2$ .

**Selecting the variable to leave the basis: the ratio test** Once we have selected the variable to enter the basis, we are faced with the question of which of the current basic variables will be dropped out of it, in order to obtain the improving adjacent bfs. The logic behind this step is as follows: Since increasing  $X_2$  from its current (zero) value improves the objective value, we would like to increase it as much as we can. What constrains us in this increase, is the requirement to meet the technological constraints:

$$\frac{1}{40}X_1 + \frac{1}{60}X_2 + S_1 = 1.0$$

$$\frac{1}{50}X_1 + \frac{1}{50}X_2 + S_2 = 1.0$$

as well as the sign (nonnegativity) restrictions imposed on the LP variables. In particular, since variable  $X_1$  will remain nonbasic, its value will remain equal to zero, and therefore it vanishes from the above set of equations. Hence, we have:

$$\frac{1}{60}X_2 + S_1 = 1.0 \Leftrightarrow S_1 = 1.0 \Leftrightarrow \frac{1}{60}X_2$$

$$\frac{1}{50}X_2 + S_2 = 1.0 \Leftrightarrow S_2 = 1.0 \Leftrightarrow \frac{1}{50}X_2$$

Notice that as  $X_2$  increases, both  $S_1$  and  $S_2$  are decreased. Obviously,  $X_2$  cannot increase beyond a value that makes any of  $S_1$  and  $S_2$  negative. So, the maximal allowable increase for  $X_2$  is obtained by solving the system of inequalities:

$$\begin{aligned} S_1 = 1.0 &\Leftrightarrow \frac{1}{60}X_2 \geq 0 \\ S_2 = 1.0 &\Leftrightarrow \frac{1}{50}X_2 \geq 0 \end{aligned} \tag{29}$$

Hence,

$$X_2 \leq \min \left\{ \begin{array}{l} \frac{1}{1/60} \\ \frac{1}{1/50} \end{array} \right. \tag{30}$$

i.e.,  $X_2 \leq 50$ . Equation 30 is known as the *ratio test* in the theory of Simplex algorithm, and for an entering variable  $X_j$ , its general form is:

$$X_j \leq \min_i \frac{b_i}{a_{ij}} : a_{ij} > 0 \tag{31}$$

The variable leaving the basis is anyone of those corresponding to a technological constraint with index  $i$  minimizing the ratio of Equation 31, since setting  $X_j$  to the min-ratio value drives them to zero. Hence, in our case, the variable to leave the basis is  $S_2$ , and the new bfs is  $B_1 = \{S_1, X_2\}$ . The new (improved) objective value is  $z = 400 \cdot 50 = 20,000$ .

**Obtaining the canonical form with respect to the new bfs: Pivoting the entering variable** At this point, we must reset to ourselves the question regarding the optimality of basis  $B_1$ . Notice, however, that the way that we addressed this, as well as the remaining set of questions above, was facilitated by the fact that the original set of technological constraints in Equation 28 were in *canonical form* with respect to the current basis  $B_0$ . To be able to pursue the same set of questions regarding basis  $B_1$ , we must transform the original set of constraints into canonical form with respect to this new basis. This is done by rewriting the original LP equations in the form:

$$\begin{array}{rccccccc} z & \Leftrightarrow & 200X_1 & \Leftrightarrow & 400X_2 & & = & 0.0 \\ & & \frac{1}{40}X_1 & & + \frac{1}{60}X_2 & + S_1 & = & 1.0 \\ & & \frac{1}{50}X_1 & & + \frac{1}{50}X_2 & & + S_2 & = & 1.0 \end{array} \tag{32}$$

This representation of the LP technological constraints and the objective-function equation is known as the *LP tableau*. In particular, the row corresponding to the objective-function equation is known as the *Row-0* of the tableau, and the coefficients of the (nonbasic) variables in that row are known as the *Row-0 coefficients*. Notice also that the right-hand-side entry of Row-0 provides the objective value of the current bfs.

Notice that under the above tableau representation, the columns corresponding to the basic variables  $S_1$  and  $S_2$  are essentially the elementary (unit) vectors:  $e_2 = [0 \ 1 \ 0]^T$  and  $e_3 = [0 \ 0 \ 1]^T$ , respectively, while the third unit vector  $e_1 = [1 \ 0 \ 0]^T$  is the column of the objective variable  $z$ . This is another way to characterize the fact that the above tableau is in canonical form with respect to variables  $S_1, S_2$ . To obtain the tableau corresponding to basis  $B_1 = \{S_1, X_2\}$ , we must convert the column of  $X_2$  in the above tableau to the unit vector  $e_3 = [0 \ 0 \ 1]^T$ , making sure that while we are doing so we do not alter the content of these three equations. This can be done by exploiting the following two properties of a system of linear equations (generally, known as *elementary row operations*):

- If we multiply any of the system equations with a *nonzero* constant, we obtain an *equivalent* system of equations (i.e., one with the same solution set).
- if we multiply one of the system equations with a constant and add it to a second equation, we obtain an equivalent system of equations.

Applying the first of these properties to the third of equations 32, with a coefficient of 50, we get the equivalent system of equations:

$$\begin{array}{rccccccc} z & \Leftrightarrow 200X_1 & \Leftrightarrow 400X_2 & & = & 0.0 & \\ & \frac{1}{40}X_1 & + \frac{1}{60}X_2 & + S_1 & & = & 1.0 \\ & X_1 & + X_2 & & + 50S_2 & = & 50.0 \end{array} \quad (33)$$

Multiplying the third equation above with  $-1/60$  and adding it to the second equation, we get:

$$\begin{array}{rccccccc} z & \Leftrightarrow 200X_1 & \Leftrightarrow 400X_2 & & = & 0.0 & \\ & \frac{1}{120}X_1 & & + S_1 & \Leftrightarrow \frac{5}{6}S_2 & = & \frac{1}{6} \\ & X_1 & + X_2 & & + 50S_2 & = & 50.0 \end{array} \quad (34)$$

Finally, multiplying the third equation with 400 and adding it to the first equation, we get:

$$\begin{array}{rccccccc} z & 200X_1 & & & 20,000S_2 & = & 20,000 \\ & \frac{1}{120}X_1 & & + S_1 & \Leftrightarrow \frac{5}{6}S_2 & = & \frac{1}{6} \\ & X_1 & + X_2 & & + 50S_2 & = & 50.0 \end{array} \quad (35)$$

This the new tableau is in canonical form with respect to basis  $B_1$ . As it was expected, the transformation provided also (automatically) the values of the

new basic variables, as well as the objective function value corresponding to the new basis  $B_1$  (i.e., the right-hand-side of Row-0). Finally, from the signs of the Row-0 coefficients, we can see that increasing any of the non-basic variables from their zero value will have a decreasing effect on the objective value. Hence, we can conclude that  $B_1$  is an *optimal* basis for our example LP, with the optimal values for the basic variables being:  $S_1 = 1/6$  and  $X_2 = 50.0$ . The optimal objective value is  $z^* = 20,000$ .

**Obtaining an initial bfs in the general case** As we saw in the previous example, if all the constraints in the original LP formulation are of the ' $\leq$ '-type, we can readily obtain an initial bfs for Simplex, consisting of all the *slack* variables in the corresponding "standard form" formulation. In this section we consider the more general case, where the original LP formulation might contain also ' $\geq$ '-type inequality as well as equality constraints. To facilitate the subsequent discussion, let's consider the following LP:

$$\min z = 2X_1 + 3X_2$$

s.t.

$$\frac{1}{2}X_1 + \frac{1}{4}X_2 \leq 4$$

$$X_1 + 3X_2 \geq 20$$

$$X_1 + X_2 = 10$$

$$X_1, X_2 \geq 0$$

which in "standard form" becomes:

$$\min z = 2X_1 + 3X_2$$

s.t.

$$\frac{1}{2}X_1 + \frac{1}{4}X_2 + S_1 = 4$$

$$X_1 + 3X_2 \Leftrightarrow E_2 = 20$$

$$X_1 + X_2 = 10$$

$$X_1, X_2 \geq 0 \tag{36}$$

For this LP,  $S_1$  can constitute an initial basic variable associated with the first constraint. However, the *excess* variable  $E_2$  cannot be a basic variable for constraint 2, even though it appears only in this constraint, since this would imply a negative value for it (i.e.,  $E_2 = \Leftrightarrow 20$ ), and the resulting basic solution would not be feasible. The last constraint (#3), does not even involve an auxiliary variable.

To overcome this problem, we "synthesize" an initial bfs by introducing additional (*artificial*) variables  $A_2, A_3$  for the two "problematic" constraints,

with  $A_2, A_3 \geq 0$ . Hence, our initial bfs is  $B_0 = \{S_1, A_2, A_3\}$ , with  $S_1 = 4, A_2 = 20, A_3 = 10$ . Notice, however, that even though we obtained a bfs for our set of constraints, we have altered the structure – and therefore, the content – of the original formulation. In fact, it is easy to check that this bfs (together with the implied zero values for the nonbasic variables) is not even feasible for the original LP! On the other hand, if we were able to obtain another bfs for the *modified* problem in which the introduced artificial variables are *nonbasic*, then this bfs would be also feasible for the original LP: the artificial variables, being nonbasic, would be equal to zero, and therefore, they would have no effective contribution in the corresponding “canonical form” representation. To obtain the effect just described, we try to drive the artificial variables to zero, by initially trying to achieve the following objective:

$$\min z = A_2 + A_3$$

s. t.

$$\begin{aligned} \frac{1}{2}X_1 + \frac{1}{4}X_2 + S_1 &= 4 \\ X_1 + 3X_2 \Leftrightarrow E_2 + A_2 &= 20 \\ X_1 + X_2 + A_3 &= 10 \\ X_1, X_2, A_2, A_3 &\geq 0 \end{aligned} \tag{37}$$

Synthesizing and solving the above LP is known as the *Phase I* - step of the Simplex algorithm. If the original LP (Eq. 36) has a feasible solution, then the nonnegativity of  $A_2, A_3$  implies that the *optimal* value of this new LP will be zero, and by the end of its solution we shall have also a feasible bfs for the original formulation. On the other hand, having a strictly positive optimal objective value for the LP of Equation 37 implies that the artificial variables are absolutely necessary to obtain a feasible solution for this set of constraints, and therefore, our original LP is *infeasible*. Hence, *infeasibility* is tested during the Phase-I step of the Simplex algorithm.

Applying the previously described Simplex algorithm on the Phase-I LP of Equation 37, we obtain the optimal tableau:

$z$	$X_1$	$X_2$	$S_1$	$E_2$	$A_2$	$A_3$	$RHS$
1	0	0	0	0	$\Leftrightarrow 1$	$\Leftrightarrow 1$	0
0	0	0	1	$\Leftrightarrow \frac{1}{8}$	$\frac{1}{8}$	$\Leftrightarrow \frac{5}{8}$	$\frac{1}{4}$
0	0	1	0	$\Leftrightarrow \frac{1}{2}$	$\frac{1}{2}$	$\Leftrightarrow \frac{1}{2}$	5
0	1	0	0	$\frac{1}{2}$	$\Leftrightarrow \frac{1}{2}$	$\frac{3}{2}$	5

(38)

Therefore, a feasible basis for the original LP is  $B_1 = \{S_1, X_2, X_1\}$ . The canonical form of the original tableau with respect to basis  $B_1$  is obtained by:

1. dropping the columns corresponding to the artificial variables  $A_2$ ,  $A_3$  from the tableau of Equation 38:

$$\begin{array}{cccccc}
 z & X_1 & X_2 & S_1 & E_2 & RHS \\
 1 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 1 & \Leftrightarrow \frac{1}{8} & \frac{1}{4} \\
 0 & 0 & 1 & 0 & \Leftrightarrow \frac{1}{2} & 5 \\
 0 & 1 & 0 & 0 & \frac{1}{2} & 5
 \end{array} \tag{39}$$

2. re-introducing in Row-0 of the resulting tableau the original LP objective:

$$\begin{array}{cccccc}
 z & X_1 & X_2 & S_1 & E_2 & RHS \\
 1 & \Leftrightarrow 2 & \Leftrightarrow 3 & 0 & 0 & 0 \\
 0 & 0 & 0 & 1 & \Leftrightarrow \frac{1}{8} & \frac{1}{4} \\
 0 & 0 & 1 & 0 & \Leftrightarrow \frac{1}{2} & 5 \\
 0 & 1 & 0 & 0 & \frac{1}{2} & 5
 \end{array} \tag{40}$$

3. and, finally, bringing the tableau of Equation 40 into canonical form by performing the appropriate elementary row operations (the Row-0 coefficients of the basic variables  $X_1$  and  $X_2$  must be zero in the corresponding canonical-form formulation):

$$\begin{array}{cccccc}
 z & X_1 & X_2 & S_1 & E_2 & RHS \\
 1 & 0 & 0 & 0 & \Leftrightarrow \frac{1}{2} & 25 \\
 0 & 0 & 0 & 1 & \Leftrightarrow \frac{1}{8} & \frac{1}{4} \\
 0 & 0 & 1 & 0 & \Leftrightarrow \frac{1}{2} & 5 \\
 0 & 1 & 0 & 0 & \frac{1}{2} & 5
 \end{array} \tag{41}$$

Then, we are ready to restart Simplex, but this time on the original LP formulation. In this particular case, it is easy to see that, since we have a minimization problem, the current basis  $B_1$  is already optimal (i.e., increasing  $E_2$  from its zero value will only increase the objective function).

**Unbounded LP's and LP's with many optimal solutions** In the previous section, we saw that Simplex detects infeasibility while trying to solve the Phase-I LP. It is instructive to consider how the algorithm behaves on unbounded LP's as well as on LP's with many optimal solutions. To understand these aspects of the algorithm, try to implement it on the corresponding examples provided in Section 2. What is happening?