CS 521: Linear Programming **Rutgers: Fall 2022**

Lecture 2

September 22, 2022

Instructor: Sepehr Assadi *Scribe: Sepehr Assadi*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Topics of this Lecture

1 Application: Linear Regression

Our goal throughout the semester will be to include various applications of linear programming in different lectures. For this lecture, we consider their application to *linear regression*.

In linear regression, we have a set of m points from the n-dimensional space $\mathbb{R}^n$:

$$(a_{1,1}, a_{1,2}, \ldots, a_{1,n}), (a_{2,1}, a_{2,2}, \ldots, a_{2,n}) \cdots, (a_{m,1}, a_{m,2}, \ldots, a_{m,n}).$$

We can denote these points by a matrix $A \in \mathbb{R}^{m \times n}$. The points are associated with real numbers in $\mathbb{R}$:

$$(b_1, b_2, \ldots, b_m).$$

Similarly, we can denote these numbers by a vector $b \in \mathbb{R}^m$.

Think of each column of this matrix as a "feature" or an "attribute" of the points, and think of each real number b_i as the "label" of the point $a_i := (a_{i,1}, \ldots, a_{i,n})$. Our goal is to find an "explanation" of these labels in terms of the feature using a *linear* function. Formally, we would like to find a *hyperplane* $x \in \mathbb{R}^n$ such that given any point a_i, we can recover the label of a_i, namely, b_i, via $\langle a_i \, , \, x \rangle$.

Of course, such a hyperplane can only exist if the system of linear equations $A \cdot x = b$ has a solution. However, this is not generally guaranteed (especially because we often have $m \gg n$). Thus, the goal is to find a hyperplane x with minimal "distance" from b. This leads to the family of regression problems wherein the goal is to solve the following optimization problem:

$$\min_{x \in \mathbb{R}^n} \|A \cdot x - b\|,$$

where we can pick the $\|\cdot\|$ differently depending on the application. Two popular choices of the norm are ℓ_2-norm and ℓ_1-norm that we review in this lecture.

1.1 Detour: a closed-form solution for ℓ_2-regression

Recall that for any vector $v \in \mathbb{R}^m$, the ℓ_2-norm of v is defined as:

$$\|v\|_2 := \sqrt{\langle v \,,\, v\rangle} = \sqrt{\sum_{i=1}^{m} v_i^2}.$$

In the ℓ_2-regression problem, we would like to find $x \in \mathbb{R}^n$ as follows:

$$\min_{x \in \mathbb{R}^n} \|A \cdot x - b\|_2.$$

Fix a choice of A, b. Consider the function $f(x) := \|Ax - b\|_2^2$. Minimizing f then solves our problem (note that here minimizing $\|A \cdot x - b\|_2^2$ is the same as $\|A \cdot x - b\|_2$). Since f is a convex function[1], we have that minimum of f happens on a point where

$$f'(x) := \frac{\partial f(x)}{\partial x} = 0,$$

where ∂ is the *gradient* of f. Computing the gradient of f is out of the scope of this course, but it is a basic exercise in calculus as follows:

$$\begin{aligned}
f(x) &= (A \cdot x - b)^T \cdot (A \cdot x - b) && (\text{as } \|v\|_2^2 = \langle v \,,\, v\rangle = v^T \cdot v)\\
&= (x^T \cdot A^T - b^T) \cdot (A \cdot x - b) && (\text{as } (w - v)^T = (w^T - v^T) \text{ and } (wv)^T = v^T w^T)\\
&= x^T A^T A x - b^T A x - x^T A^T b + b^T b.
\end{aligned}$$

Using the rules of gradients and the fact that $B := A^T A$ is a symmetric matrix, namely, $B^T = B$, we have,

$$f'(x) = \frac{\partial f(x)}{\partial x} = 2 \cdot A^T A x - 2A^T b.$$

Thus, when $f'(x) = 0$, we have,

$$A^T A x = 2A^T b,$$

which means that as long as $A^T A$ is non-singular, we have,

$$x = (A^T A)^{-1} A^T b.$$

This gives a closed-form solution for ℓ_2-regression[2]

1.2 Main application: a linear program for ℓ_1-regression

The ℓ_1-norm of a vector $v \in \mathbb{R}^m$ is defined as:

$$\|v\|_1 := \sum_{i=1}^{n} |v_i|.$$

Another form of linear regression is to instead minimize the ℓ_1-norm, i.e., find $x \in \mathbb{R}^n$ with

$$\min_{x \in \mathbb{R}^n} \|A \cdot x - b\|_1.$$

One reason to prefer ℓ_1-regression over ℓ_2-regression is that it is more "robust" in the presence of outliers (but this is a topic beyond the scope of our course). Unfortunately however, ℓ_1-regression does not have a closed-form solution and we instead use linear programming to solve it. Writing this problem as a LP is not entirely trivial so we are going to do it step by step.

[1] We revisit convexity later in the course; for this detour, we use this standard fact without proving it (or even defining it).

[2] Why should we expect $A^T A$ be non-singular? Well, for $A^T A$ to be singular, we need to have some $x \neq 0$ such that $A^T A x = 0$, which also implies $x^T A^T A x = \|Ax\|_2^2 = 0$, which means $Ax = 0$. This means that the *column space* of A is not full rank. In other words, one of the "features" we have picked is entirely useless as it can be expressed as a linear combination of the previous features. Thus, we can simply discard this feature (and any other dependent ones) to obtain a full column rank matrix A, solve the problem there, and then extend it to the original dimension using the fact that the discarded features were all linearly dependent on the rest and thus this does not introduce any error.

A more direct (and equivalent) approach here is to simply compute $x = (A^T A)^{\dagger} A^T b$ where for a matrix M, $M^{\dagger}$ is the pseudo-inverse of M.

Step 1. The ℓ_1-regression problem is the following optimization problem:

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{m} |\langle a_i\,,\, x\rangle - b_i|.$$

This is of course yet not a linear program because the objective function is not a linear function. But we made the problem somewhat simpler without changing it at all (it is straightforward to check the problems are equivalent so far).

Step 2. We are going to define m new variables $z_1, \ldots, z_m$, one per each row of the matrix A. We write our optimization problem now as:

$$\begin{aligned} \min_{x\in\mathbb{R}^n, z\in\mathbb{R}^m} \quad & \sum_{i=1}^{m} z_i \\ \text{subject to} \quad & z_i = \max\left(\langle a_i\,,\, x\rangle - b_i, b_i - \langle a_i\,,\, x\rangle\right) \quad \forall i \in [m]. \end{aligned}$$

This is still equivalent with the previous problem because for every $i \in [m]$,

$$|\langle a_i\,,\, x\rangle - b_i| = \max\left(\langle a_i\,,\, x\rangle - b_i, b_i - \langle a_i\,,\, x\rangle\right),$$

by definition; enforcing z_i's being equal to this thus ensures that the problem remains the same. We are now one step closer since our objective function now is linear although we still have many non-linear constraints that we have to handle.

Step 3. The only (slightly) non-trivial step is this one where we are going to relax the constraints a bit without violating feasibility as follows:

$$\begin{aligned} \min_{x\in\mathbb{R}^n, z\in\mathbb{R}^m} \quad & \sum_{i=1}^{m} z_i \\ \text{subject to} \quad & z_i \geqslant \max\left(\langle a_i\,,\, x\rangle - b_i, b_i - \langle a_i\,,\, x\rangle\right) \quad \forall i \in [m]. \end{aligned}$$

It may now be clear entirely that this problem is the same as above, in particular because we actually *expanded* the feasible region. However, we still have the following claim.

Claim 1. *In any* optimal *solution of this problem, we have that for every $i \in [m]$,*

$$z_i = \max\left(\langle a_i\,,\, x\rangle - b_i, b_i - \langle a_i\,,\, x\rangle\right);$$

Proof. Suppose some z_i is strictly larger; then reduce it by a tiny $\varepsilon > 0$ which does not violate the constraint but reduces the objective function, contradicting the optimality. □

This implies that optimal solutions of the problems in steps 2 and 3 coincide and thus we can still focus on solving this problem. We are still not done because our constraints are not yet linear.

Step 4. Finally, we can "linearize" the constraints simply as follows:

$$\begin{aligned} \min_{x\in\mathbb{R}^n, z\in\mathbb{R}^m} \quad & \sum_{i=1}^{m} z_i \\ \text{subject to} \quad & z_i \geqslant \langle a_i\,,\, x\rangle - b_i \quad \text{and} \quad z_i \geqslant b_i - \langle a_i\,,\, x\rangle \quad \forall i \in [m]. \end{aligned}$$

This is now truly a linear program and is clearly equivalent to the previous step because for all $i \in [m]$:

$$z_i \geqslant \max\left(\langle a_i\,,\, x\rangle - b_i, b_i - \langle a_i\,,\, x\rangle\right) \iff z_i \geqslant \langle a_i\,,\, x\rangle - b_i \quad \text{and} \quad z_i \geqslant b_i - \langle a_i\,,\, x\rangle$$

Wrap-up. Using the above four steps, we can write any ℓ_1-regression problem with m points in n dimension as a LP with $n+m$ variables and $2m$ constraints. Solving this LP then gives us the optimal solution to the ℓ_1-regression problem as well.

2 "Basic" Definitions in Linear Programs

Given that linear programming involves optimization over real numbers, it is not clear a priori that we can even have a *finite time* algorithm for it (no matter how inefficient) – after all, unlike most combinatorial problems, here we cannot simply enumerate all solutions until we find the right answer given that the number of all possible solutions is infinite. The goal in the rest of this lecture is to develop some basic understanding of linear programs and their structure, which along the way also addresses this question.

We start with the following definition.

Definition 2. We say that a LP is in **equational form** iff it is stated as

$$\max_{x \in \mathbb{R}^n} \quad c^T x \quad \text{subject to} \quad Ax = b \quad \text{and} \quad x \geqslant 0.$$

As we shall see in the rest of this lecture (and subsequent ones), working with LPs in equational form can be easier. The following simple result shows that every LP can be stated in the equational form without increasing its size by much.

Proposition 3. *Any n-variable m-constraint LP*

$$\max_{x \in \mathbb{R}^n} \quad c^T x \quad \text{subject to} \quad Ax \geqslant b$$

can be stated in the equational form with $n' = 2n + m$ variables and $m' = m$ equality constraints

$$\max_{x' \in \mathbb{R}^{n'}} \quad c'^T \cdot x' \quad \text{subject to} \quad A'x' = b' \quad \textit{and} \quad x' \geqslant 0,$$

so that the objective value of both LPs are equal and answer to the latter LP can be uniquely mapped to the answer in the original LP.

Proof. Firstly, for any constraint $\langle a_i \, , \, x \rangle \geqslant b_i$ in the first LP, define a new variable s_i in the new LP and add the constraints $\langle a_i \, , \, x \rangle - s_i = b_i$ and $s_i \geqslant 0$. For instance

$$x_1 + 2x_2 \geqslant 3 \implies x_2 + 2x_2 - s_1 = 3 \qquad \text{and} \qquad s_1 \geqslant 0.$$

Secondly, for any variable x_j in the first LP define two variables y_j, z_j and change any occurrence of x_j in the equations with $y_j - z_j$. For instance

$$x_2 + 2x_2 - s_1 = 3 \implies y_1 - z_1 + 2 \cdot (y_2 - z_2) - s_1 = 3.$$

Add the constraints $y_i \geqslant 0$ and $z_i \geqslant 0$ to the new LP. (This way, x_1 having value, say, 10 will be the same as setting $y_1 = 10$ and $z_1 = 0$, while x_1 being -5 corresponds to $y_1 = 0$ and $z_1 = 5$).

This new LP is now in the equational form and has $2n + m$ variables and m equations. We omit the straightforward but rather tedious task of verifying the equivalence of these two LPs. □

For the rest of this lecture, we work with an LP with n variables and m equations for $n \geqslant m$ in the equational form:

$$\max_{x \in \mathbb{R}^n} \quad c^T x \quad \text{subject to} \quad Ax = b \quad \text{and} \quad x \geqslant 0.$$

By Proposition 3, every LP can be stated as the above one (including the extra $n \geqslant m$ condition). We further have the following two assumptions:

- **Assumption** (i)**:** The linear system $Ax = b$ has at least one solution. (This is without loss of generality since we can always check if $Ax = b$ has any solution using Gaussian elimination and if not we know that our LP is not feasible and thus there is nothing else for us to solve[3]).
- **Assumption** (ii)**:** The matrix A has full rank. (This is without loss of generality since we can always remove any linearly dependent row of A, solve the problem on the remaining equations, and then find the unique value obtained for the dependent rows and check with the corresponding values on b).

We are now ready to state the main definition of this lecture: while in general there are infinitely many feasible solutions to an LP, we are almost exclusively interested in a finite number of them specified by the following definition (the reason we are only interested in these will become shortly):

Definition 4. Consider an LP in the equational form under assumptions (i) and (ii). We say that a feasible solution $x \in \mathbb{R}^n$ is a **basic feasible solution** iff there exists a set B of m columns of A such that (i) A_B has full rank and (ii) $x_{-B} = 0$.[a] We refer to B as a **basis** for the basic feasible solution x.

[a] A_B is the sub-matrix of A on the columns specified by B and $x_{-B} := x_{[n]\setminus B}$, namely, the columns of x not in B.

We have the following simple claims regarding basic feasible solutions.

Claim 5. *Any set $B \subseteq [n]$ of m columns of A can be a basis for at most basic feasible solution.*

Proof. Consider a basic feasible solution $x \in \mathbb{R}^n$ with basis B. We have that $x_{-B} = 0$, thus

$$Ax = A_B x_B + A_{-B} x_{-B} = A_B x_B.$$

Since $Ax = b$, we also get that $A_B x_B = b$. Since A_B has full rank, the system of linear equations $A_B x_B = b$ has a *unique* solution, thus x_B is determined uniquely for the basis B. □

Claim 6. *Given any feasible solution $x \in \mathbb{R}^n$, let*

$$S := \text{supp}(x) = \{j \in [n] \mid x_j > 0\}.$$

If A_S has full rank, then x is a basic feasible solution.

Proof. Since row rank equals column rank and row rank of A_S is at most m, if A_S is full rank it necessarily means $|S| \leqslant m$. If $|S| = m$, we get that x is a basic feasible solution by Definition 4.

Suppose now that $|S| < m$. Since we assumed A has full row rank, we know the columns in S can be extended to m columns that are linearly independent. Let B be these columns. We get that x is a basic feasible solution with a basis S. □

Remark. We used the following two linear algebraic facts in the above proofs.

Fact 7. *For any matrix $M \in \mathbb{R}^{m \times m}$ which has full rank, for any $b \in \mathbb{R}^m$, the system of linear equations $Ax = b$ has a unique solution $x = A^{-1} \cdot b$.*

Fact 8. *In any matrix $M \in \mathbb{R}^{m \times n}$, the row rank of M is equal to the column rank of M.*

In the next section, we prove a theorem that clarifies our interest in basic feasible solutions.

[3] We emphasize that this condition is necessary for feasibility of the LP but is not sufficient.

3 Optimum Solutions Happen on Basic Feasible Solutions

The following theorem is the main result of today's lecture.

Theorem 9. *Consider an LP in the equational form under assumptions (i) and (ii). Suppose the LP has an optimal solution; then, it also has an optimal solution which is a basic feasible solution.*

Proof. We prove the following statement:

> *Suppose the objective value of the LP on every feasible solution is bounded from above. Then, for any feasible solution y, there exists a basic feasible solution x such that $c^T x \geqslant c^T y$.*

This statement then implies the theorem because the first part is a necessary condition for the LP to have an optimal solution, and we can then apply this statement to y as any optimal solution.

We prove this statement as follows. Since y is a feasible solution and thus $y \geqslant 0$, we have,

$$S := \text{supp}(y) = \{j \in [n] \mid y_j > 0\}\,.$$

Among all choices of y with the maximum value of $c^T y$, find the one with the smallest support $\text{supp}(y)$.

Firstly, suppose A_S has full rank. In this case, by Claim 6 we are done as x itself is a basic feasible solution.

We now consider the case when A_S does not have full rank. This means that the kernel of A_S is non-empty, i.e., there exists some $w \in \mathbb{R}^{|S|}$ such that $A_S \cdot w = 0$ even though $w \neq 0$. It is without loss of generality to assume that $c^T w \geqslant 0$ as otherwise we can replace w with $-w$, which is fine because $A_S \cdot (-w) = 0$ also. We now consider two cases:

- **Case I:** There exists some $j \in S$ such that $w_j < 0$. Define $z \in \mathbb{R}^n$ such that $z_S = w$ and $z_{-S} = 0$. For any $t \geqslant 0$, we have that

$$\begin{aligned} A \cdot (y + t \cdot z) &= A \cdot y + t \cdot A \cdot z \\ &= A \cdot y + t \cdot A_S w && \text{(as } w = z_S \text{ and } z_{-S} = 0) \\ &= A \cdot y && \text{(as } A_S w = 0) \\ &= b. && \text{(as } y \text{ is a feasible solution)} \end{aligned}$$

  Now note that by increasing t slightly, we eventually reach a point that for some $j \in S$, $y_j + t \cdot z_j = y_j + t \cdot w_j = 0$ (as $w_j < 0$), while at the same time, all other $y_k \geqslant 0$. This implies that at this point, $y + t \cdot z \geqslant 0$, $A \cdot (y + t \cdot z) = b$, and $c^T(y + t \cdot z) \geqslant c^T y$; this is in contradiction with y having the smallest support as the vector $(y + t \cdot z)$ has one more zero entry.

- **Case II:** For every $j \in S$, we have $w_j > 0$. If $c^T w = 0$, we can simply take $-w$ again which takes us to Case I above and we will be done. So, we have $c^T w > 0$. Define the vector z as in the previous case. For every $t > 0$, we have $A \cdot (y + t \cdot z) = b$ (as proven above), $y + t \cdot z \geqslant 0$ always, (making $y + t \cdot z$ a feasible solution), and $c^T \cdot (y + t \cdot z) = c^T y + t \cdot c^T z$; since $c^T z > 0$, making t larger and larger, takes the value of the objective function to infinity, contradicting the assumption that the objective value of the LP on every feasible solution is bounded from above

This concludes the proof. □

We will use Theorem 9 repeatedly throughout the rest of the lecture. For now, we simply mention that a trivial application of this theorem (plus Claim 5) is an (inefficient) algorithm for solving LPs: iterate over all $\binom{n}{m}$ m-subsets B of $[n]$ to find all basic feasible solutions. Find the one that maximizes the value of the objective function. This gives an $(e \cdot n/m)^m \cdot \text{poly}(n)$ time algorithm for solving LPs (this is finite time thus addressing our motivating question from earlier, in particular, exponential time, but of course terribly slow).