Chapter 12 Solving Linear Equations

An Introduction to Optimization

Spring, 2014

Wei-Ta Chu

Least-Squares Analysis

- Consider a system of linear equations $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A} \in R^{m \times n}$ and $\boldsymbol{b} \in R^m$, $m \geq n$, and $rank(\boldsymbol{A}) = n$. Note that the number of unknowns, n, is no larger than the number of equations, m.
- If $\boldsymbol{b}$ does not belong to the range of $\boldsymbol{A}$, that is, $\boldsymbol{b} \notin \mathcal{R}(\boldsymbol{A})$, then this system of equations is said to be ***inconsistent*** or ***overdetermined***.
- Our goal is to find the vector(s) $\boldsymbol{x}$ minimizing $\|\boldsymbol{Ax} - \boldsymbol{b}\|^2$. This problem is a special case of the nonlinear least-squares problem discussed in Section 9.4.

Least-Squares Analysis

- Let $\boldsymbol{x}^*$ be a vector that minimizes $\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2$; that is, for all $\boldsymbol{x} \in R^n$

$$\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2 \geq \|\boldsymbol{A}\boldsymbol{x}^*-\boldsymbol{b}\|^2$$

- We refer to the vector $\boldsymbol{x}^*$ as a least-squares solution to $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$. In the case where $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ has a solution, then the solution is a least-squares solution. Otherwise, a least-squares solution minimizes the norm of the difference between the left- and right-hand sides of the equation $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$

Least-Squares Analysis

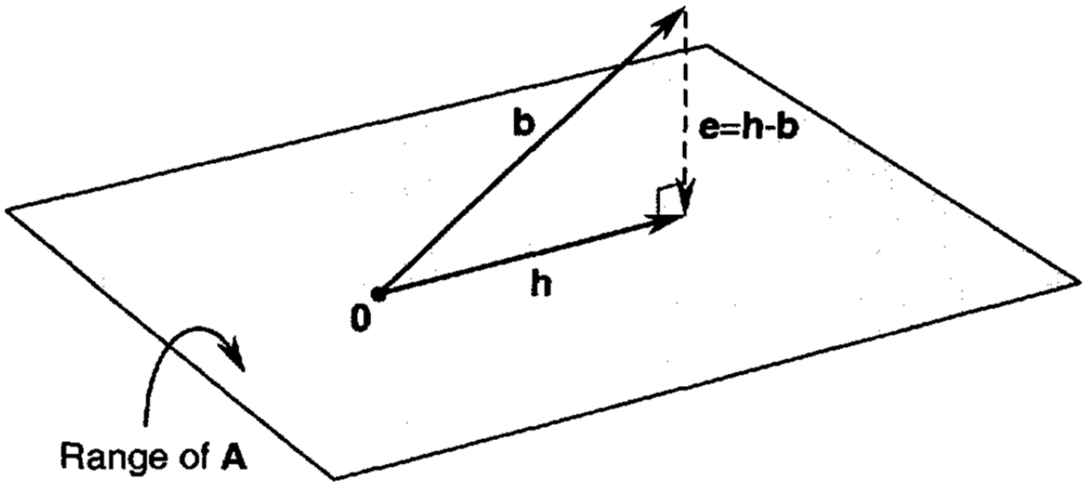
- Lemma 12.1: Let $\boldsymbol{A} \in R^{m \times n}$, $m \geq n$. Then, $rank(\boldsymbol{A}) = n$ if and only if $rank(\boldsymbol{A}^T\boldsymbol{A}) = n$ (i.e., the square matrix $\boldsymbol{A}^T\boldsymbol{A}$ is nonsingular).
- Theorem 12.1: The unique vector $\boldsymbol{x}^*$ that minimizes $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$ is given by the solution to the equation $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{b}$; that is,

$$\boxed{\boldsymbol{x}^* = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}}$$

- The columns of $\boldsymbol{A}$ span the range $\mathcal{R}(\boldsymbol{A})$ of $\boldsymbol{A}$, which is an n-dimensional subspace of R^m. The equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ has a solution if and only if $\boldsymbol{b} \in \mathcal{R}(\boldsymbol{A})$.
- If $m = n$, then $\boldsymbol{b} \in \mathcal{R}(\boldsymbol{A})$ always, and the solution is $\boldsymbol{x}^* = \boldsymbol{A}^{-1}\boldsymbol{b}$

Least-Squares Analysis

- Suppose now that $m > n$. We would expect the likelihood of $\boldsymbol{b} \in \mathcal{R}(\boldsymbol{A})$ to be small, because the subspace spanned by the columns of $\boldsymbol{A}$ is very “thin.”
- Suppose that $\boldsymbol{b} \notin \mathcal{R}(\boldsymbol{A})$. We wish to find a point $\boldsymbol{h} \in \mathcal{R}(\boldsymbol{A})$ that is “closest” to $\boldsymbol{b}$. Geometrically, the point $\boldsymbol{h}$ should be such that the vector $\boldsymbol{e} = \boldsymbol{h} - \boldsymbol{b}$ is orthogonal to the subspace $\mathcal{R}(\boldsymbol{A})$
- We call $\boldsymbol{h}$ the ***orthogonal projection*** of $\boldsymbol{b}$ onto the subspace $\mathcal{R}(\boldsymbol{A})$ It turns out that $\boldsymbol{h} = \boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$.



Least-Squares Analysis

- Write $\boldsymbol{A} = [\boldsymbol{a}_1, ..., \boldsymbol{a}_n]$, where $\boldsymbol{a}_1, ..., \boldsymbol{a}_n$ are the columns of $\boldsymbol{A}$
- The vector $\boldsymbol{e}$ is orthogonal to $\mathcal{R}(\boldsymbol{A})$ if and only if it is orthogonal to each of the columns $\boldsymbol{a}_1, ..., \boldsymbol{a}_n$ of $\boldsymbol{A}$.
- Note that $\langle \boldsymbol{e}, \boldsymbol{a}_i \rangle = 0, i = 1, ..., n$ if and only if for any set of scalars $\{x_1, x_2, ..., x_n\}$, we also have

$$\langle \boldsymbol{e}, x_1\boldsymbol{a}_1 + \cdots + x_n\boldsymbol{a}_n \rangle = 0$$

  Any vector in $\mathcal{R}(\boldsymbol{A})$ has the form $x_1\boldsymbol{a}_1 + \cdots + x_n\boldsymbol{a}_n$

Least-Squares Analysis

- Proposition 12.1: Let $\boldsymbol{h} \in \mathcal{R}(\boldsymbol{A})$ be such that $\boldsymbol{h} - \boldsymbol{b}$ is orthogonal to $\mathcal{R}(\boldsymbol{A})$. Then, $\boldsymbol{h} = \boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$
- Note that the matrix

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix} \langle \boldsymbol{a}_1, \boldsymbol{a}_1 \rangle & \cdots & \langle \boldsymbol{a}_n, \boldsymbol{a}_1 \rangle \\ \vdots & & \vdots \\ \langle \boldsymbol{a}_1, \boldsymbol{a}_n \rangle & \cdots & \langle \boldsymbol{a}_n, \boldsymbol{a}_n \rangle \end{bmatrix}$$

plays an important role in the least-squares solution. This matrix is often called the ***Gram matrix*** (or ***Grammian***).

Example

- Suppose that you are given two different types of concrete. The first type contains 30% cement, 40% gravel, and 30% sand (all percentages of weight). The second type contains 10% cement, 20% gravel, and 70% sand. How many pounds of each type of concrete should you mix together so that you get a concrete mixture that has as close as possible to a total of 5 pounds of cement, 3 pounds of gravel, and 4 pounds of sand?
- The problem can be formulated as a least-squares problem with

$$\boldsymbol{A} = \begin{bmatrix} 0.3 & 0.1 \\ 0.4 & 0.2 \\ 0.3 & 0.7 \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} 5 \\ 3 \\ 4 \end{bmatrix}$$

where the decision variable is $\boldsymbol{x} = [x_1, x_2]^T$ and x_1 and x_2 are the amounts of concrete of the first and second types, respectively.

Example

- The problem can be formulated as a least-squares problem with

$$\boldsymbol{A} = \begin{bmatrix} 0.3 & 0.1 \\ 0.4 & 0.2 \\ 0.3 & 0.7 \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} 5 \\ 3 \\ 4 \end{bmatrix}$$

  where the decision variable is $\boldsymbol{x} = [x_1, x_2]^T$ and x_1 and x_2 are the amounts of concrete of the first and second types, respectively.

- After some algebra, we obtain the solution:

$$\boldsymbol{x}^* = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$$

$$= \frac{1}{(0.34)(0.54) - (0.32)^2} \begin{bmatrix} 0.54 & -0.32 \\ -0.32 & 0.34 \end{bmatrix} \begin{bmatrix} 3.9 \\ 3.9 \end{bmatrix} = \begin{bmatrix} 10.6 \\ 0.961 \end{bmatrix}$$

Example

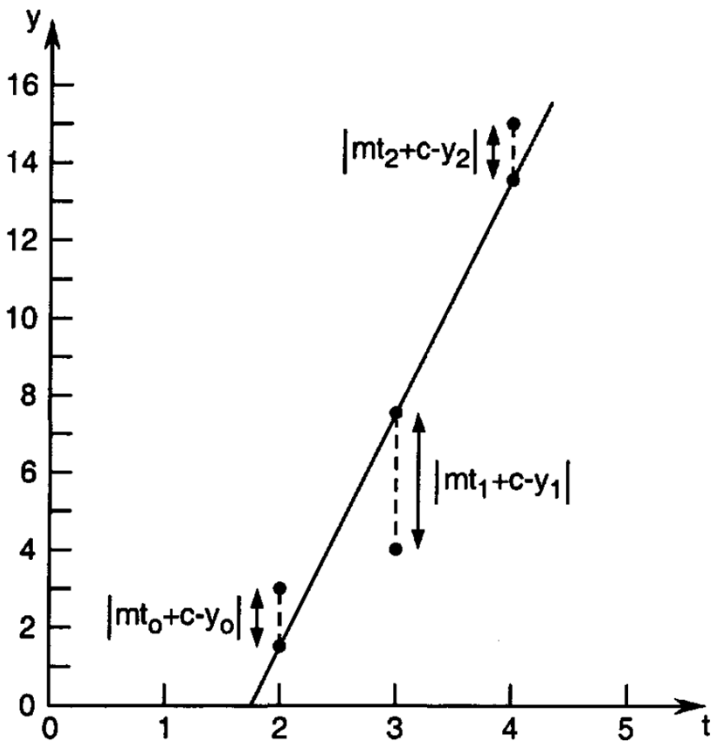
- ***Line Fitting***. Suppose that a process has a single input $t \in R$ and a single output $y \in R$. Suppose that we perform an experiment on the process, resulting in a number of measurements.

i	0	1	2
t_i	2	3	4
y_i	3	4	15

The ith measurement results in the input labeled t_i and the output labeled y_i. We would like to find a straight line given by $y = mt + c$ that fits the experimental data.
In other words, we wish to find two numbers, m and c, such that $y_i = mt_i + c, i = 0, 1, 2$.

Example

- However, it is apparent that there is no choice of m and c that results in the requirement above. Therefore, we would like to find the values of m and c that best fit the data.



i	0	1	2
t_i	2	3	4
y_i	3	4	15

Example

- We can represent our problem as a system of linear equations of the form

$$2m + c = 3$$
$$3m + c = 4$$
$$4m + c = 15$$

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} 3 \\ 4 \\ 15 \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} m \\ c \end{bmatrix}$$

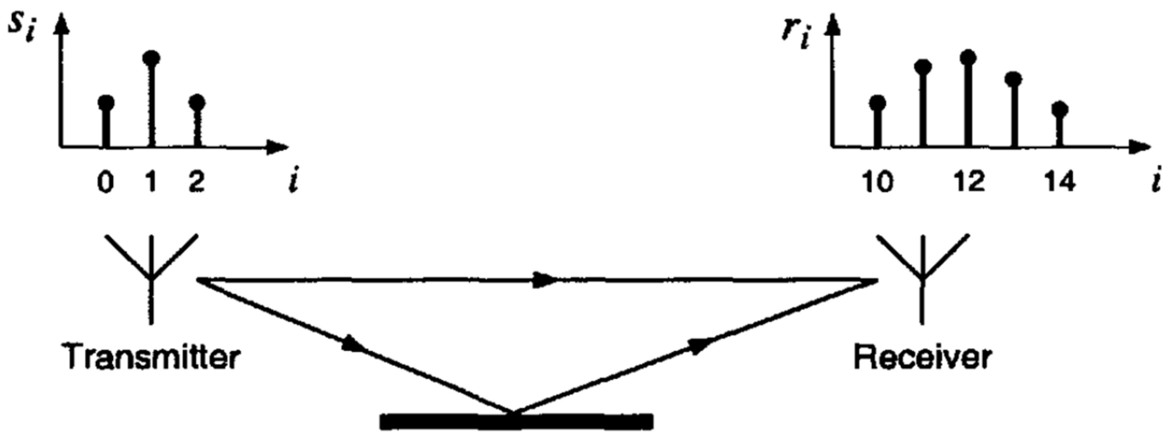
Notice that since $rank(\boldsymbol{A}) < rank([\boldsymbol{A}, \boldsymbol{b}])$, the vector $\boldsymbol{b}$ does not belong to the range of $\boldsymbol{A}$. Thus, this system is inconsistent.

- The solution to this least-squares problem is

$$\boldsymbol{x}^* = \begin{bmatrix} m^* \\ c^* \end{bmatrix} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b} = \begin{bmatrix} 6 \\ -32/3 \end{bmatrix}$$

Example

- ***Attenuation Estimation***. A wireless transmitter sends a discrete-time signal $\{s_0, s_1, s_2\}$ (of duration 3) to a receiver. The real number s_i is the value of the signal at time i.
- The transmitted signal takes two paths to the receiver: a direct path, with delay 10 and attenuation factor a_1, and an indirect (reflected) path, with delay 12 and attenuation factor a_2. The received signal is the sum of the signals from these two paths, with their respective delays and attenuation factors.



Example

- Suppose that the received signal is measured from times 10 through 14 as $r_{10}, r_{11}, \ldots, r_{14}$. We wish to compute the least-squares estimates of a_1 and a_2, based on the following values

s_0	s_1	s_2	r_{10}	r_{11}	r_{12}	r_{13}	r_{14}
1	2	1	4	7	8	6	3

- The problem can be posed as a least-squares problem with

$$a_1 s + a_2 s' = r$$

$$\Rightarrow \begin{aligned} a_1 s_0 + 0 &= r_{10} \\ a_1 s_1 + 0 &= r_{11} \\ a_1 s_2 + a_2 s_0 &= r_{12} \\ 0 + a_2 s_1 &= r_{13} \\ 0 + a_2 s_2 &= r_{14} \end{aligned}$$

$$\boldsymbol{A} = \begin{bmatrix} s_0 & 0 \\ s_1 & 0 \\ s_2 & s_0 \\ 0 & s_1 \\ 0 & s_2 \end{bmatrix} \quad \boldsymbol{x} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} r_{10} \\ r_{11} \\ r_{12} \\ r_{13} \\ r_{14} \end{bmatrix}$$

Example

- The least-squares estimate is given

$$\boldsymbol{x}^* = \begin{bmatrix} a_1^* \\ a_2^* \end{bmatrix} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$$

$$= \begin{bmatrix} \|\boldsymbol{s}\|^2 & s_0s_2 \\ s_0s_2 & \|\boldsymbol{s}\|^2 \end{bmatrix}^{-1} \begin{bmatrix} s_0r_{10} + s_1r_{11} + s_2r_{12} \\ s_0r_{12} + s_1r_{13} + s_2r_{14} \end{bmatrix}$$

$$= \begin{bmatrix} 6 & 1 \\ 1 & 6 \end{bmatrix}^{-1} \begin{bmatrix} 4+14+8 \\ 8+12+3 \end{bmatrix}$$

$$= \frac{1}{35}\begin{bmatrix} 133 \\ 112 \end{bmatrix}$$

Example

- ***Discrete Fourier Series***. Suppose that we are given a discrete time signal, represented by the vector

$$\boldsymbol{b} = [b_1, b_2, \ldots, b_m]^T$$

We wish to approximate this signal by a sum of sinusoids. Specifically, we approximate $\boldsymbol{b}$ by the vector

$$y_0\boldsymbol{c}^{(0)} + \sum_{k=1}^{n}\left(y_k\boldsymbol{c}^{(k)} + z_k\boldsymbol{s}^{(k)}\right)$$

where $y_0, y_1, \ldots, y_n, z_1, \ldots, z_n \in R$ and the vectors $\boldsymbol{c}^{(k)}$ and $\boldsymbol{s}^{(k)}$ are given by

$$\boldsymbol{c}^{(0)} = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, \ldots, \frac{1}{\sqrt{2}}\right]^T$$

$$\boldsymbol{c}^{(k)} = \left[\cos\left(1\frac{2k\pi}{m}\right), \cos\left(2\frac{2k\pi}{m}\right), \ldots, \cos\left(m\frac{2k\pi}{m}\right)\right]^T, k = 1, \ldots, n$$

$$\boldsymbol{s}^{(k)} = \left[\sin\left(1\frac{2k\pi}{m}\right), \sin\left(2\frac{2k\pi}{m}\right), \ldots, \sin\left(m\frac{2k\pi}{m}\right)\right]^T, k = 1, \ldots, n$$

Example

- We call the sum of sinusoids above a ***discrete Fourier series***. We wish to find $y_0, y_1, \ldots, y_n, z_1, \ldots, z_n$ such that
$$\left\| \left(y_0 \boldsymbol{c}^{(0)} + \sum_{k=1}^{n} \left(y_k \boldsymbol{c}^{(k)} + z_k \boldsymbol{s}^{(k)} \right) \right) - \boldsymbol{b} \right\|^2$$
is minimized.
- To proceed, we define
$$\boldsymbol{A} = \left[\boldsymbol{c}^{(0)}, \boldsymbol{c}^{(1)}, \ldots, \boldsymbol{c}^{(n)}, \boldsymbol{s}^{(1)}, \ldots, \boldsymbol{s}^{(n)} \right]$$
$$\boldsymbol{x} = \left[y_0, y_1, \ldots, y_n, z_1, \ldots, z_n \right]^T$$
Our problem can be reformulated as minimizing
$$\| \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \|^2$$

Example

- We assume that $m \geq 2n+1$. To find the solution, we first compute $\boldsymbol{A}^T\boldsymbol{A}$. We make use of the following trigonometric identities: for any nonzero integer k that is not an integral multiple of m, we have

$$\sum_{i=1}^{m} \cos\left(i\frac{2k\pi}{m}\right) = 0$$

$$\sum_{i=1}^{m} \sin\left(i\frac{2k\pi}{m}\right) = 0$$

  with the aid of these identities, we can verify that

$$\boldsymbol{c}^{(k)T}\boldsymbol{c}^{(j)} = \begin{cases} m/2 & \text{if } k = j \\ 0 & \text{otherwise} \end{cases}$$

$$\boldsymbol{s}^{(k)T}\boldsymbol{s}^{(j)} = \begin{cases} m/2 & \text{if } k = j \\ 0 & \text{otherwise} \end{cases}$$

$$\boldsymbol{c}^{(k)T}\boldsymbol{s}^{(j)} = 0 \text{ for any } k, j$$

Example

- Hence, $\boldsymbol{A}^T\boldsymbol{A} = \frac{m}{2}\boldsymbol{I}_{2n+1}$, which is clearly nonsingular, with inverse $(\boldsymbol{A}^T\boldsymbol{A})^{-1} = \frac{2}{m}\boldsymbol{I}_{2n+1}$
  Therefore, the solution to our problem is

$$\begin{aligned}\boldsymbol{x}^* &= [y_0^*, y_1^*, \ldots, y_n^*, z_1^*, \ldots, z_n^*]^T \\ &= (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b} \\ &= \tfrac{2}{m}\boldsymbol{A}^T\boldsymbol{b}\end{aligned}$$

  We represent the solution as

$$y_0^* = \frac{\sqrt{2}}{m}\sum_{i=1}^m b_i \qquad y_k^* = \frac{2}{m}\sum_{i=1}^m b_i \cos\left(i\frac{2k\pi}{m}\right) \qquad k = 1, \ldots, n$$

$$z_k^* = \frac{2}{m}\sum_{i=1}^m b_i \sin\left(i\frac{2k\pi}{m}\right) \qquad k = 1, \ldots, n$$

  We call these ***discrete Fourier coefficients***.

Example

- ***Orthogonal Projections***. Let $\mathcal{V} \subset R^n$ be a subspace. Given a vector $\boldsymbol{x} \in R^n$, we write the orthogonal decomposition of $\boldsymbol{x}$ as
$$\boldsymbol{x} = \boldsymbol{x}_{\mathcal{V}} + \boldsymbol{x}_{\mathcal{V}^\perp}$$
where $\boldsymbol{x}_{\mathcal{V}} \in \mathcal{V}$ is the orthogonal projection of $\boldsymbol{x}$ onto $\mathcal{V}$ and $\boldsymbol{x}_{\mathcal{V}^\perp} \in \mathcal{V}^\perp$ is the orthogonal projection of $\boldsymbol{x}$ onto $\mathcal{V}^\perp$. We can write $\boldsymbol{x}_{\mathcal{V}} = \boldsymbol{P}\boldsymbol{x}$ for some matrix $\boldsymbol{P}$ called ***orthogonal projector***. In the following, we derive expressions for $\boldsymbol{P}$ for the case where $\mathcal{V} = \mathcal{R}(\boldsymbol{A})$ and the case where $\mathcal{V} = \mathcal{N}(\boldsymbol{A})$
- Consider a matrix $\boldsymbol{A} \in R^{m \times n}, m \geq n$, and $rank(\boldsymbol{A}) = n$. Let $\mathcal{V} = \mathcal{R}(\boldsymbol{A})$ be the range of $\boldsymbol{A}$. In this case we can write an expression for $\boldsymbol{P}$ in terms of $\boldsymbol{A}$

Example

- By Proposition 12.1 we have $\boldsymbol{x}_{\mathcal{V}} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{x}$, whence $\boldsymbol{P} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$. Note that by Proposition 12.1, we may also write $\boldsymbol{x}_{\mathcal{V}} = \arg\min_{\boldsymbol{y}\in\mathcal{V}}\|\boldsymbol{y}-\boldsymbol{x}\|$
- Next, consider a matrix $\boldsymbol{A} \in R^{m\times n}, m \leq n$, and $rank(\boldsymbol{A}) = m$. Let $\mathcal{V} = \mathcal{N}(\boldsymbol{A})$ be the nullspace of $\boldsymbol{A}$. To derive an expression for the orthogonal projector $\boldsymbol{P}$ in terms of $\boldsymbol{A}$ for this case, we use the formula derived above and the identity $\mathcal{N}(\boldsymbol{A})^{\perp} = \mathcal{R}(\boldsymbol{A}^T)$ (see Theorem 3.4).
  Indeed, if $\mathcal{U} = \mathcal{R}(\boldsymbol{A}^T)$, then the orthogonal decomposition with respect to $\mathcal{U}$ is $\boldsymbol{x} = \boldsymbol{x}_{\mathcal{U}} + \boldsymbol{x}_{\mathcal{U}^{\perp}}$, where $\boldsymbol{x}_{\mathcal{U}} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\boldsymbol{x}$

Example

- Because $\mathcal{N}(\boldsymbol{A})^{\perp} = \mathcal{R}(\boldsymbol{A}^T)$, we deduce that $\boldsymbol{x}_{\mathcal{V}^{\perp}} = \boldsymbol{x}_{\mathcal{U}} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\boldsymbol{x}$
  Hence,
  $$\boldsymbol{x}_{\mathcal{V}} = \boldsymbol{x} - \boldsymbol{x}_{\mathcal{V}^{\perp}} = \boldsymbol{x} - \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{A})\boldsymbol{x}$$
  Thus, the orthogonal projector in this case is
  $$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{A}$$

The Recursive Least-Squares Algorithm

- Assume that we are originally given three experimental results $(t_0, y_0), (t_1, y_1), (t_2, y_2)$, and we find the parameters m^* and c^* of the straight line that best bits these data. Suppose that we are now given an extra measurement point (t_3, y_3) . We can use previous calculations of m^* and c^* for the three data points to calculate the parameters for the four data points. This procedure is called the ***recursive least-squares*** (RLS) algorithm.

The Recursive Least-Squares Algorithm

- Consider the problem of minimizing $\|\boldsymbol{A}_0\boldsymbol{x} - \boldsymbol{b}^{(0)}\|^2$. The solution is given by $\boldsymbol{x}^{(0)} = \boldsymbol{G}_0^{-1}\boldsymbol{A}_0^T\boldsymbol{b}^{(0)}$, where $\boldsymbol{G}_0 = \boldsymbol{A}_0^T\boldsymbol{A}_0$
  Consider now the problem of minimizing
  $$\left\| \begin{bmatrix} \boldsymbol{A}_0 \\ \boldsymbol{A}_1 \end{bmatrix} \boldsymbol{x} - \begin{bmatrix} \boldsymbol{b}^{(0)} \\ \boldsymbol{b}^{(1)} \end{bmatrix} \right\|^2$$
  The solution is given by
  $$\boldsymbol{x}^{(1)} = \boldsymbol{G}_1^{-1} \begin{bmatrix} \boldsymbol{A}_0 \\ \boldsymbol{A}_1 \end{bmatrix}^T \begin{bmatrix} \boldsymbol{b}^{(0)} \\ \boldsymbol{b}^{(1)} \end{bmatrix} \qquad \boldsymbol{G}_1 = \begin{bmatrix} \boldsymbol{A}_0 \\ \boldsymbol{A}_1 \end{bmatrix}^T \begin{bmatrix} \boldsymbol{A}_0 \\ \boldsymbol{A}_1 \end{bmatrix}$$
  Our goal is to write $\boldsymbol{x}^{(1)}$ as a function of $\boldsymbol{x}^{(0)}$, $\boldsymbol{G}_0$, and the new data $\boldsymbol{A}_1$ and $\boldsymbol{b}^{(1)}$

The Recursive Least-Squares Algorithm

- To this end, we first write $\boldsymbol{G}_1$ as

$$\boldsymbol{G}_1 = [\boldsymbol{A}_0^T \boldsymbol{A}_1^T] \begin{bmatrix} \boldsymbol{A}_0 \\ \boldsymbol{A}_1 \end{bmatrix} = \boldsymbol{A}_0^T \boldsymbol{A}_0 + \boldsymbol{A}_1^T \boldsymbol{A}_1 = \boldsymbol{G}_0 + \boldsymbol{A}_1^T \boldsymbol{A}_1$$

Next, we write

$$\begin{bmatrix} \boldsymbol{A}_0 \\ \boldsymbol{A}_1 \end{bmatrix}^T \begin{bmatrix} \boldsymbol{b}^{(0)} \\ \boldsymbol{b}^{(1)} \end{bmatrix} = [\boldsymbol{A}_0^T \ \ \boldsymbol{A}_1^T] \begin{bmatrix} \boldsymbol{b}^{(0)} \\ \boldsymbol{b}^{(1)} \end{bmatrix} = \boldsymbol{A}_0^T \boldsymbol{b}^{(0)} + \boldsymbol{A}_1^T \boldsymbol{b}^{(1)}$$

To proceed further, we write $\boldsymbol{A}_0^T \boldsymbol{b}^{(0)}$ as

$$\boldsymbol{A}_0^T \boldsymbol{b}^{(0)} = \boldsymbol{G}_0 \boldsymbol{G}_0^{-1} \boldsymbol{A}_0^T \boldsymbol{b}^{(0)} = \boldsymbol{G}_0 \boldsymbol{x}^{(0)}$$

$$= (\boldsymbol{G}_1 - \boldsymbol{A}_1^T \boldsymbol{A}_1) \boldsymbol{x}^{(0)} = \boldsymbol{G}_1 \boldsymbol{x}^{(0)} - \boldsymbol{A}_1^T \boldsymbol{A}_1 \boldsymbol{x}^{(0)}$$

The Recursive Least-Squares Algorithm

- Combining these formulas, we see that we can write $\boldsymbol{x}^{(1)}$ as

$$\boldsymbol{x}^{(1)} = \boldsymbol{G}_1^{-1} \begin{bmatrix} \boldsymbol{A}_0 \\ \boldsymbol{A}_1 \end{bmatrix}^T \begin{bmatrix} \boldsymbol{b}^{(0)} \\ \boldsymbol{b}^{(1)} \end{bmatrix} = \boldsymbol{G}_1^{-1} \left(\boldsymbol{G}_1 \boldsymbol{x}^{(0)} - \boldsymbol{A}_1^T \boldsymbol{A}_1 \boldsymbol{x}^{(0)} + \boldsymbol{A}_1^T \boldsymbol{b}^{(1)} \right)$$
$$= \boldsymbol{x}^{(0)} + \boldsymbol{G}_1^{-1} \boldsymbol{A}_1^T \left(\boldsymbol{b}^{(1)} - \boldsymbol{A}_1 \boldsymbol{x}^{(0)} \right)$$

  where $\boldsymbol{G}_1$ can be calculated using $\boldsymbol{G}_1 = \boldsymbol{G}_0 + \boldsymbol{A}_1^T \boldsymbol{A}_1$

- With this formula, $\boldsymbol{x}^{(1)}$ can be computed using only $\boldsymbol{x}^{(0)}$, $\boldsymbol{A}_1$, $\boldsymbol{b}^{(1)}$ and $\boldsymbol{G}_0$. Hence, we have a way of using our previous efforts in calculating $\boldsymbol{x}^{(0)}$ to compute $\boldsymbol{x}^{(1)}$. Observe that if the new data are consistent with the old data, that is, $\boldsymbol{A}_1 \boldsymbol{x}^{(0)} = \boldsymbol{b}^{(1)}$, then the correction term is 0 and the updated solution $\boldsymbol{x}^{(1)}$ is equal to the previous solution $\boldsymbol{x}^{(0)}$.

The Recursive Least-Squares Algorithm

- At the $(k+1)$th iteration, we have

$$\boldsymbol{G}_{k+1} = \boldsymbol{G}_k + \boldsymbol{A}_{k+1}^T \boldsymbol{A}_{k+1}$$

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{G}_{k+1}^{-1} \boldsymbol{A}_{k+1}^T \left(\boldsymbol{b}^{(k+1)} - \boldsymbol{A}_{k+1} \boldsymbol{x}^{(k)} \right)$$

  The vector $\boldsymbol{b}^{(k+1)} - \boldsymbol{A}_{k+1}\boldsymbol{x}^{(k)}$ is often called the ***innovation***. As before, observe that if the innovation is zero, then the updated solution $\boldsymbol{x}^{(k+1)}$ is equal to the previous solution $\boldsymbol{x}^{(k)}$

- We can see that to compute $\boldsymbol{x}^{(k+1)}$ we need $\boldsymbol{G}_{k+1}^{-1}$ rather than $\boldsymbol{G}_{k+1}$. It turns out that we can derive an update formula for $\boldsymbol{G}_{k+1}^{-1}$ itself.

The Recursive Least-Squares Algorithm

- Lemma 12.2: Let $\boldsymbol{A}$ be a nonsingular matrix. Let $\boldsymbol{U}$ and $\boldsymbol{V}$ be matrices such that $\boldsymbol{I}+\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U}$ is nonsingular. Then, $\boldsymbol{A}+\boldsymbol{U}\boldsymbol{V}$ is nonsingular, and

$$(\boldsymbol{A}+\boldsymbol{U}\boldsymbol{V})^{-1}=\boldsymbol{A}^{-1}-(\boldsymbol{A}^{-1}\boldsymbol{U})(\boldsymbol{I}+\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}(\boldsymbol{V}\boldsymbol{A}^{-1})$$

- By Lemma 12.2, we get

$$\begin{aligned}\boldsymbol{G}_{k+1}^{-1}&=\left(\boldsymbol{G}_k+\boldsymbol{A}_{k+1}^T\boldsymbol{A}_{k+1}\right)^{-1}\\&=\boldsymbol{G}_k^{-1}-\boldsymbol{G}_k^{-1}\boldsymbol{A}_{k+1}^T(\boldsymbol{I}+\boldsymbol{A}_{k+1}\boldsymbol{G}_k^{-1}\boldsymbol{A}_{k+1}^T)^{-1}\boldsymbol{A}_{k+1}\boldsymbol{G}_k^{-1}\end{aligned}$$

For simplicity of notation, we rewrite $\boldsymbol{G}_k^{-1}$ as $\boldsymbol{P}_k$. We summarize by writing the RLS algorithm using $\boldsymbol{P}_k$

$$\boldsymbol{P}_{k+1}=\boldsymbol{P}_k-\boldsymbol{P}_k\boldsymbol{A}_{k+1}^T(\boldsymbol{I}+\boldsymbol{A}_{k+1}\boldsymbol{P}_k\boldsymbol{A}_{k+1}^T)^{-1}\boldsymbol{A}_{k+1}\boldsymbol{P}_k$$

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}+\boldsymbol{P}_{k+1}\boldsymbol{A}_{k+1}^T\left(\boldsymbol{b}^{(k+1)}-\boldsymbol{A}_{k+1}\boldsymbol{x}^{(k)}\right)$$

The Recursive Least-Squares Algorithm

- In the special case where the new data at each step are such that $\boldsymbol{A}_{k+1}$ is a matrix consisting of a single row, $\boldsymbol{A}_{k+1} = \boldsymbol{a}_{k+1}^T$, and $\boldsymbol{b}^{(k+1)}$ is a scalar, $\boldsymbol{b}^{(k+1)} = b_{k+1}$, we get

$$\boldsymbol{P}_{k+1} = \boldsymbol{P}_k - \frac{\boldsymbol{P}_k \boldsymbol{a}_{k+1} \boldsymbol{a}_{k+1}^T \boldsymbol{P}_k}{1 + \boldsymbol{a}_{k+1}^T \boldsymbol{P}_k \boldsymbol{a}_{k+1}}$$

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{P}_{k+1} \boldsymbol{a}_{k+1} \left(b_{k+1} - \boldsymbol{a}_{k+1}^T \boldsymbol{x}^{(k)} \right)$$

Example

- Let
$$\boldsymbol{A}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \qquad \boldsymbol{b}^{(0)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$
$$\boldsymbol{A}_1 = \boldsymbol{a}_1^T = \begin{bmatrix} 2 & 1 \end{bmatrix} \qquad \boldsymbol{b}^{(1)} = b_1 = [3]$$
$$\boldsymbol{A}_2 = \boldsymbol{a}_2^T = \begin{bmatrix} 3 & 1 \end{bmatrix} \qquad \boldsymbol{b}^{(2)} = b_2 = [4]$$
First compute the vector $\boldsymbol{x}^{(0)}$ minimizing $\|\boldsymbol{A}_0\boldsymbol{x} - \boldsymbol{b}^{(0)}\|^2$. Then, use the RLS algorithm to find $\boldsymbol{x}^{(2)}$ minimizing
$$\left\| \begin{bmatrix} \boldsymbol{A}_0 \\ \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{bmatrix} \boldsymbol{x} - \begin{bmatrix} \boldsymbol{b}^{(0)} \\ \boldsymbol{b}^{(1)} \\ \boldsymbol{b}^{(2)} \end{bmatrix} \right\|^2$$
We have
$$\boldsymbol{P}_0 = (\boldsymbol{A}_0^T \boldsymbol{A}_0)^{-1} = \begin{bmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{bmatrix}$$
$$\boldsymbol{x}^{(0)} = \boldsymbol{P}_0 \boldsymbol{A}_0^T \boldsymbol{b}^{(0)} = \begin{bmatrix} 2/3 \\ 2/3 \end{bmatrix}$$

Example

- Applying the RLS algorithm twice, we get

$$\boldsymbol{P}_1 = \boldsymbol{P}_0 - \frac{\boldsymbol{P}_0 \boldsymbol{a}_1 \boldsymbol{a}_1^T \boldsymbol{P}_0}{1 + \boldsymbol{a}_1^T \boldsymbol{P}_0 \boldsymbol{a}_1} = \begin{bmatrix} 1/3 & -1/3 \\ -1/3 & 2/3 \end{bmatrix}$$

$$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \boldsymbol{P}_1 \boldsymbol{a}_1 \left(b_1 - \boldsymbol{a}_1^T \boldsymbol{x}^{(0)} \right) = \begin{bmatrix} 1 \\ 2/3 \end{bmatrix}$$

$$\boldsymbol{P}_2 = \boldsymbol{P}_1 - \frac{\boldsymbol{P}_1 \boldsymbol{a}_2 \boldsymbol{a}_2^T \boldsymbol{P}_1}{1 + \boldsymbol{a}_2^T \boldsymbol{P}_1 \boldsymbol{a}_2} = \begin{bmatrix} 1/6 & -1/4 \\ -1/4 & 5/8 \end{bmatrix}$$

$$\boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)} + \boldsymbol{P}_2 \boldsymbol{a}_2 \left(b_2 - \boldsymbol{a}_2^T \boldsymbol{x}^{(1)} \right) = \begin{bmatrix} 13/12 \\ 5/8 \end{bmatrix}$$

Solution to A Linear Equation with Minimum Norm

- Consider now a system of linear equations $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A} \in R^{m \times n}$, $\boldsymbol{b} \in R^{m}$, $m \leq n$, and $rank(\boldsymbol{A}) = m$. Note that the number of equations is no longer than the number of unknowns. There may exist an infinite number of solutions to this system of equations.
- However, as we shall see, there is only one solution that is closest to the origin: the solution to $\boldsymbol{Ax} = \boldsymbol{b}$ whose norm $\|\boldsymbol{x}\|$ is minimal.
- Let $\boldsymbol{x}^*$ be this solution that is $\boldsymbol{Ax}^* = \boldsymbol{b}$ and $\|\boldsymbol{x}^*\| \leq \|\boldsymbol{x}\|$ for any $\boldsymbol{x}$ such that $\boldsymbol{Ax} = \boldsymbol{b}$. In other words, $\boldsymbol{x}^*$ is the solution to the problem

$$\begin{aligned} &\text{minimize } \|\boldsymbol{x}\| \\ &\text{subject to } \boldsymbol{Ax} = \boldsymbol{b} \end{aligned}$$

Solution to A Linear Equation with Minimum Norm

- Theorem 12.2: The unique solution $\boldsymbol{x}^*$ to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ that minimizes the norm $\|\boldsymbol{x}\|$ is given by

$$\boldsymbol{x}^* = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{b}$$

- Example: Find the point closest to the origin of R^3 on the line of intersection of the two planes defined by the following two equations:

$$x_1 + 2x_2 - x_3 = 1$$

$$4x_1 + x_2 + 3x_3 = 0$$

Note that this problem is equivalent to the problem

$$\begin{aligned} &\text{minimize } \|\boldsymbol{x}\| \\ &\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{aligned} \qquad \boldsymbol{A} = \begin{bmatrix} 1 & 2 & -1 \\ 4 & 1 & 3 \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\boldsymbol{x}^* = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{b} = \begin{bmatrix} 0.0952 \\ 0.3333 \\ -0.2381 \end{bmatrix}$$

Kaczmarz's Algorithm

- Kaczmarz's algorithm converges to the vector $\boldsymbol{x}^* = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{b}$ without explicitly having to invert the matrix $\boldsymbol{A}\boldsymbol{A}^T$. This is important from a practical point of view, especially when $\boldsymbol{A}$ has many rows.
- Let $\boldsymbol{a}_j^T$ denote the jth row of $\boldsymbol{A}$, and b_j the jth component of $\boldsymbol{b}$ and μ a positive scalar, $0 < \mu < 2$. Kaczmarz's algorithm is:
  - 1. Set $i := 0$, initial condition $\boldsymbol{x}^{(0)}$
  - 2. For $j = 1, \ldots, m$, set
  $$\boldsymbol{x}^{(im+j)} = \boldsymbol{x}^{(im+j-1)} + \mu(b_j - \boldsymbol{a}_j^T \boldsymbol{x}^{(im+j-1)})\frac{\boldsymbol{a}_j}{\boldsymbol{a}_j^T \boldsymbol{a}_j}$$
  - 3. Set $i := i + 1$; go to step 2.

Kaczmarz's Algorithm

- For the first m iterations, we have

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \mu(b_{k+1} - \boldsymbol{a}_{k+1}^T \boldsymbol{x}^{(k)})\frac{\boldsymbol{a}_{k+1}}{\boldsymbol{a}_{k+1}^T \boldsymbol{a}_{k+1}}$$

where, in each iteration, we use rows of $\boldsymbol{A}$ and corresponding components of $\boldsymbol{b}$ successively. For the $(m+1)$ th iteration, we revert back to the first row of $\boldsymbol{A}$ and the first component of $\boldsymbol{b}$; that is,

$$\boldsymbol{x}^{(m+1)} = \boldsymbol{x}^{(m)} + \mu(b_1 - \boldsymbol{a}_1^T \boldsymbol{x}^{(m)})\frac{\boldsymbol{a}_1}{\boldsymbol{a}_1^T \boldsymbol{a}_1}$$

We continue with the $(m+2)$ th iteration using the second row of $\boldsymbol{A}$ and the second component of $\boldsymbol{b}$, and so on, repeating the cycle every m iterations. The reason for $0 < \mu < 2$ will become apparent from the convergence anlaysis.

Kaczmarz's Algorithm

- Theorem 12.3: In Kaczmarz's algorithm, if $\boldsymbol{x}^{(0)} = \mathbf{0}$, then $\boldsymbol{x}^{(k)} \to \boldsymbol{x}^* = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{b}$ as $k \to \infty$.
- Example: Let

$$\boldsymbol{A} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

  In this case $\boldsymbol{x}^* = [5, 3]^T$. This figure shows a few iterations of Kaczmarz's algorithm with $\mu = 1$ and $\boldsymbol{x}^{(0)} = \mathbf{0}$. We have $\boldsymbol{a}_1^T = [1, -1]$ $\boldsymbol{a}_2^T = [0, 1]$, $b_1 = 2$, $b_2 = 3$. The diagonal line passing through the point $[2, 0]^T$ corresponds to the set $\{\boldsymbol{x} : \boldsymbol{a}_1^T\boldsymbol{x} = b_1\}$, and the horizontal line passing through the point $[0, 3]^T$ corresponds to the set $\{\boldsymbol{x} : \boldsymbol{a}_2^T\boldsymbol{x} = b_2\}$.

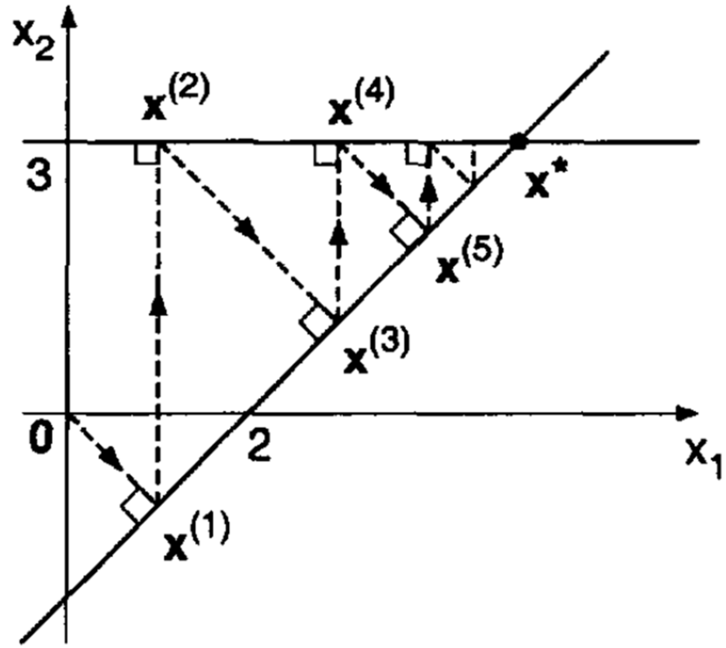
Kaczmarz's Algorithm

- We perform three iterations:

$$\boldsymbol{x}^{(1)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + (2-0)\tfrac{1}{2}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$\boldsymbol{x}^{(2)} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} + (3-(-1))\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$$\boldsymbol{x}^{(3)} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} + (2-(-2))\tfrac{1}{2}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$



Solving Linear Equations in General

- Consider a system of linear equations $\boldsymbol{Ax} = \boldsymbol{b}$, where $\boldsymbol{A} \in R^{m \times n}$ and $rank(\boldsymbol{A}) = r$. Note that we always have $r \leq \min\{m, n\}$. In the case $\boldsymbol{A} \in R^{n \times n}$ and $rank(\boldsymbol{A}) = n$, the unique solution to the equation above has the form $\boldsymbol{x}^* = \boldsymbol{A}^{-1}\boldsymbol{b}$. Thus, to solve the problem in this case it is enough to know the inverse $\boldsymbol{A}^{-1}$.
- A general approach to solving $\boldsymbol{Ax} = \boldsymbol{b}$. The approach involves defining a ***pseudoinverse*** or ***generalized inverse*** of a given matrix $\boldsymbol{A} \in R^{m \times n}$, which plays the role of $\boldsymbol{A}^{-1}$ when $\boldsymbol{A}$ does not have an inverse. In particular, we discuss the ***Moore-Penrose inverse*** of a given matrix $\boldsymbol{A}$, denoted $\boldsymbol{A}^{\dagger}$.

Solving Linear Equations in General

- Lemma 12.3 ***Full-Rank Factorization***: Let $\boldsymbol{A} \in R^{m\times n}$, $rank(\boldsymbol{A}) = r \le \min\{m, n\}$. Then, there exist matrices $\boldsymbol{B} \in R^{m\times r}$ and $\boldsymbol{C} \in R^{r\times n}$ such that $\boldsymbol{A} = \boldsymbol{BC}$, where

$$rank(\boldsymbol{A}) = rank(\boldsymbol{B}) = rank(\boldsymbol{C}) = r$$

- Proof: Because $rank(\boldsymbol{A}) = r$, we can find r linearly independent columns of $\boldsymbol{A}$. Without loss of generality, let $\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_r$ be such columns, where $\boldsymbol{a}_i$ is the i th column of $\boldsymbol{A}$. The remaining columns of $\boldsymbol{A}$ can be expressed as linear combination of $\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_r$. Thus, a possible choice for $\boldsymbol{B}$ and $\boldsymbol{C}$ are

$$\boldsymbol{B} = [\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_r] \in R^{m\times r} \qquad \boldsymbol{C} = \begin{bmatrix} 1 & \cdots & 0 & c_{1,r+1} & \cdots & c_{1,n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & c_{r,r+1} & \cdots & c_{r,n} \end{bmatrix} \in R^{r\times n}$$

where the entries $c_{i,j}$ are such that for each $j = r+1, ..., n$, we have $\boldsymbol{a}_j = c_{1,j}\boldsymbol{a}_1 + \cdots + c_{r,j}\boldsymbol{a}_r$. Thus, $\boldsymbol{A} = \boldsymbol{BC}$

Solving Linear Equations in General

- Note that if $m < n$ and $rank(\boldsymbol{A}) = m$, then we take $\boldsymbol{B} = \boldsymbol{I}_m$ and $\boldsymbol{C} = \boldsymbol{A}$
- If, on the other hand, $m > n$ and $rank(\boldsymbol{A}) = n$, then we can take $\boldsymbol{B} = \boldsymbol{A}$ and $\boldsymbol{C} = \boldsymbol{I}_n$
- Example: Let

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 & -2 & 5 \\ 1 & 0 & -3 & 2 \\ 3 & -1 & -13 & 5 \end{bmatrix} \qquad rank(\boldsymbol{A}) = 2$$

We can write a full-rank factorization of $\boldsymbol{A}$ based on the proof of Lemma 12.3

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 1 & 0 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -3 & 2 \\ 0 & 1 & 4 & 1 \end{bmatrix} = \boldsymbol{BC}$$

Solving Linear Equations in General

- Consider the matrix equation $\boldsymbol{AXA} = \boldsymbol{A}$, where $\boldsymbol{A} \in R^{m\times n}$ is a given matrix and $\boldsymbol{X} \in R^{n\times m}$ is a matrix we wish to determine. Observe that if $\boldsymbol{A}$ is a nonsingular square matrix, then the equation above has the unique solution $\boldsymbol{X} = \boldsymbol{A}^{-1}$
- Definition 12.1: Given $\boldsymbol{A} \in R^{m\times n}$, a matrix $\boldsymbol{A}^{\dagger} \in R^{n\times m}$ is called a ***pseudoinverse*** of the matrix $\boldsymbol{A}$ if $\boldsymbol{AA}^{\dagger}\boldsymbol{A} = \boldsymbol{A}$, and there exist matrices $\boldsymbol{U} \in R^{n\times n}$ and $\boldsymbol{V} \in R^{m\times m}$ such that $\boldsymbol{A}^{\dagger} = \boldsymbol{UA}^{T}$ and $\boldsymbol{A}^{\dagger} = \boldsymbol{A}^{T}\boldsymbol{V}$

Solving Linear Equations in General

- The requirement $\boldsymbol{A}^{\dagger} = \boldsymbol{U}\boldsymbol{A}^{T} = \boldsymbol{A}^{T}\boldsymbol{V}$ can be interpreted as follows. Each row of the pseudoinverse matrix $\boldsymbol{A}^{\dagger}$ of $\boldsymbol{A}$ is a linear combination of the rows of $\boldsymbol{A}^{T}$, and each column of $\boldsymbol{A}^{\dagger}$ is a linear combination of the columns of $\boldsymbol{A}^{T}$
- For the case which a matrix $\boldsymbol{A} \in R^{m \times n}$ with $m \geq n$ and $rank(\boldsymbol{A}) = n$, we can easily check that the following is a pseudoinverse of $\boldsymbol{A}$:

$$\boldsymbol{A}^{\dagger} = (\boldsymbol{A}^{T}\boldsymbol{A})^{-1}\boldsymbol{A}^{T}$$

- Indeed, $\boldsymbol{A}(\boldsymbol{A}^{T}\boldsymbol{A})^{-1}\boldsymbol{A}^{T}\boldsymbol{A} = \boldsymbol{A}$, and if we define $\boldsymbol{U} = (\boldsymbol{A}^{T}\boldsymbol{A})^{-1}$ and $\boldsymbol{V} = \boldsymbol{A}(\boldsymbol{A}^{T}\boldsymbol{A})^{-1}(\boldsymbol{A}^{T}\boldsymbol{A})^{-1}\boldsymbol{A}^{T}$, then $\boldsymbol{A}^{\dagger} = \boldsymbol{U}\boldsymbol{A}^{T} = \boldsymbol{A}^{T}\boldsymbol{V}$
- Note that we have $\boldsymbol{A}^{\dagger}\boldsymbol{A} = \boldsymbol{I}_{n}$. For this reason, $(\boldsymbol{A}^{T}\boldsymbol{A})^{-1}\boldsymbol{A}^{T}$ is often called the ***left pseudoinverse*** of $\boldsymbol{A}$. This formula also appears in least-squares analysis (Sec. 12.1)

Solving Linear Equations in General

- For the case which a matrix $\boldsymbol{A} \in R^{m \times n}$ with $m \leq n$ and $rank(\boldsymbol{A}) = m$, we can easily check that the following is a pseudoinverse of $\boldsymbol{A}$:

$$\boldsymbol{A}^{\dagger} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}$$

- Note that we have $\boldsymbol{A}\boldsymbol{A}^{\dagger} = \boldsymbol{I}_m$. For this reason, $\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}$ is often called the ***right pseudoinverse*** of $\boldsymbol{A}$. This formula also appears in the problem of minimizing $\|\boldsymbol{x}\|$ subject to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$

Solving Linear Equations in General

- Theorem 12.4: Let $\boldsymbol{A} \in R^{m \times n}$. If a pseudoinverse $\boldsymbol{A}^{\dagger}$ of $\boldsymbol{A}$ exists, then it is unique.
- Our goal now is to show that the pseudoinverse matrix always exists. In fact, we show that the pseudoinverse of any given matrix $\boldsymbol{A}$ is given by the formula

$$\boldsymbol{A}^{\dagger} = \boldsymbol{C}^{\dagger}\boldsymbol{B}^{\dagger}$$

  where $\boldsymbol{B}^{\dagger}$ and $\boldsymbol{C}^{\dagger}$ are the pseudoinverse of the matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ that form a full-rank factorization of $\boldsymbol{A}$; that is, $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C}$, where $\boldsymbol{B}$ and $\boldsymbol{C}$ are of full rank (Lemma 12.3)
- Note that we already know how to compute $\boldsymbol{B}^{\dagger}$ and $\boldsymbol{C}^{\dagger}$:

$$\boldsymbol{B}^{\dagger} = (\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T \qquad \boldsymbol{C}^{\dagger} = \boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}$$

Solving Linear Equations in General

- Theorem 12.5: Let a matrix $\boldsymbol{A} \in R^{m\times n}$ have a full-rank factorization $\boldsymbol{A} = \boldsymbol{BC}$, with $rank(\boldsymbol{A}) = rank(\boldsymbol{B}) = rank(\boldsymbol{C}) = r$ $\boldsymbol{B} \in R^{m\times r}$, $\boldsymbol{C} \in R^{r\times n}$, then

$$\boldsymbol{A}^\dagger = \boldsymbol{C}^\dagger \boldsymbol{B}^\dagger$$

(Does not necessarily hold if $\boldsymbol{A} = \boldsymbol{BC}$ is not a full-rank factorization)

- Example:

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 & -2 & 5 \\ 1 & 0 & -3 & 2 \\ 3 & -1 & -13 & 5 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 0 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -3 & 2 \\ 0 & 1 & 4 & 1 \end{bmatrix} = \boldsymbol{BC}$$

$$\boldsymbol{B}^\dagger = (\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T = \frac{1}{27}\begin{bmatrix} 5 & 2 & 5 \\ 16 & 1 & -11 \end{bmatrix}$$

$$\boldsymbol{C}^\dagger = \boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1} = \frac{1}{76}\begin{bmatrix} 9 & 5 \\ 5 & 7 \\ -7 & 13 \\ 23 & 17 \end{bmatrix}$$

$$\boldsymbol{A}^\dagger = \boldsymbol{C}^\dagger\boldsymbol{B}^\dagger = \frac{1}{2052}\begin{bmatrix} 123 & 23 & -10 \\ 137 & 17 & -52 \\ 173 & -1 & -178 \\ 387 & 63 & -72 \end{bmatrix}$$

Solving Linear Equations in General

- We can simplify the expression

$$\boldsymbol{A}^{\dagger} = \boldsymbol{C}^{\dagger}\boldsymbol{B}^{\dagger} = \boldsymbol{C}^{T}(\boldsymbol{C}\boldsymbol{C}^{T})^{-1}(\boldsymbol{B}^{T}\boldsymbol{B})^{-1}\boldsymbol{B}^{T}$$

  to

$$\boldsymbol{A}^{\dagger} = \boldsymbol{C}^{T}(\boldsymbol{B}^{T}\boldsymbol{A}\boldsymbol{A}^{T})^{-1}\boldsymbol{B}^{T}$$

  This is easily verified by substituting $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C}$

- Theorem 12.6: Consider a system of linear equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ $\boldsymbol{A} \in R^{m\times n}$, $rank(\boldsymbol{A}) = r$. The vector $\boldsymbol{x}^{*} = \boldsymbol{A}^{\dagger}\boldsymbol{b}$ minimizes $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^{2}$ on R^{n} . Furthermore, among all vectors in R^{n} that minimizes $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^{2}$, the vector $\boldsymbol{x}^{*} = \boldsymbol{A}^{\dagger}\boldsymbol{b}$ is the unique vector with minimal norm.

Solving Linear Equations in General

- The generalized inverse has the following useful properties
  - $(\boldsymbol{A}^T)^\dagger = (\boldsymbol{A}^\dagger)^T$
  - $(\boldsymbol{A}^\dagger)^\dagger = \boldsymbol{A}$
- These two properties are similar to those that are satisfied by the usual matrix inverse. However, the property $(\boldsymbol{A}_1\boldsymbol{A}_2)^\dagger = \boldsymbol{A}_2^\dagger\boldsymbol{A}_1^\dagger$ does not hold in general.