4
Inverses, Solvability, and Rank

Updated 2/5/2018.

4.1 Matrix Inverses

So far we've demonstrated how Gaussian elimination can solve linear systems of the form $\mathbf{Ax} = \mathbf{y}$. Gaussian elimination involves a series of elementary row operations to transform the coefficient matrix $\mathbf{A}$ into the identity matrix. While Gaussian elimination works well, our initial goal of defining an algebra for vectors requires something stronger – a multiplicative inverse. For vectors, the multiplicative inverse is called a *matrix inverse*. For any square matrix, a matrix inverse (if it exists) is a square matrix $\mathbf{A}^{-1}$ such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1}$$

This definition is analogous to the field axiom that there exists a^{-1} such that $a^{-1}a = 1$ for all nonzero a. Since scalar multiplication always commutes, $a^{-1}a = aa^{-1}$. Since matrix multiplication doesn't commute, we need to state this property separately.

If we could prove the existence of a *matrix inverse* for $\mathbf{A}$, we could solve a wide variety of linear systems, including $\mathbf{Ax} = \mathbf{y}$.

$$\begin{aligned}
\mathbf{Ax} &= \mathbf{y} \\
\mathbf{A}^{-1}\mathbf{Ax} &= \mathbf{A}^{-1}\mathbf{y} \\
\mathbf{I}\,\mathbf{x} &= \mathbf{A}^{-1}\mathbf{y} \\
\mathbf{x} &= \mathbf{A}^{-1}\mathbf{y}
\end{aligned}$$

The existence of the matrix inverse and being amenable to Gaussian elimination are related. While the end result is the same (a transformation of the coefficient matrix into the identity matrix), the processes are different. The Gaussian elimination algorithm applies a series of elementary row operations, including pivoting to avoid numerical issues. For a matrix inverse to exists, it must be able to capture all of these operations in a single matrix multiplication. This condensing of Gaussian elimination is not a trivial task.

Our first goal of this chapter is to prove the existence of the matrix inverse for any coefficient matrix that can be solved by Gaussian elimination. Then we will derive a method to construct the matrix inverse if it exists. Finally, this chapter formalizes the requirements for solvability of a linear system of equations, which is related to the existence of the matrix inverse.

4.2 Elementary Matrices

Before proving the existence of the matrix inverse, we need to add another matrix manipulation tool to our arsenal – the *elementary matrix*. An elementary matrix is constructed by applying any single elementary row operation to the identity matrix. For example, consider swapping the second and third rows of the 3×3 identity matrix:

We use the notation $\mathbf{E}_r$ to denote an elementary matrix constructed using the elementary row operation r.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{R_2 \leftrightarrow R_3} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} = \mathbf{E}_{R_2 \leftrightarrow R_3}$$

Notice what happens when we left multiply a matrix with an elementary matrix.

$$\mathbf{E}_{R_2 \leftrightarrow R_3} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \\ 4 & 5 & 6 \end{pmatrix}$$

Multiplication by the elementary matrix exchanges the second and third rows – the same operation that created the elementary matrix. The same idea works for other elementary row operations, such as scalar multiplication

Multiplication by an elementary matrix on the right applies the same operation to the columns of the matrix.

$$\mathbf{E}_{3R_2} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{E}_{3R_2} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 12 & 15 & 18 \\ 7 & 8 & 9 \end{pmatrix}$$

and addition by a scaled row

$$\mathbf{E}_{R_2+2R_3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{E}_{R_2+2R_3} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 18 & 21 & 24 \\ 7 & 8 & 9 \end{pmatrix}$$

4.3 Proof of Existence for the Matrix Inverse

We are now ready to prove the existence of the inverse for any coefficient matrix that is solvable by Gaussian elimination, i.e. any square matrix that can be transformed into the identity matrix with elementary row operations. We will prove existence in three steps.

1. Construct a matrix $\mathbf{P}$ that looks like a left inverse ($\mathbf{PA}= \mathbf{I}$).

2. Show that this left inverse is also a right inverse ($\mathbf{AP}= \mathbf{I}$).

3. Show that the matrix inverse is unique, implying that $\mathbf{P}$ must be the inverse of $\mathbf{A}$.

Theorem. *Suppose matrix* $\mathbf{A}$ *can be reduced to the identity matrix* $\mathbf{I}$ *by elementary row operations. Then there exists a matrix* $\mathbf{P}$ *such that* $\mathbf{PA} = \mathbf{I}$.

Proof. We assume that reducing $\mathbf{A}$ to $\mathbf{I}$ requires k elementary row operations. Let $\mathbf{E}_1, \ldots, \mathbf{E}_k$ be the associated elementary matrices. Then

$$\begin{aligned} \mathbf{E}_k\mathbf{E}_{k-1}\cdots\mathbf{E}_2\mathbf{E}_1\mathbf{A} &= \mathbf{I} \\ (\mathbf{E}_k\mathbf{E}_{k-1}\cdots\mathbf{E}_2\mathbf{E}_1)\mathbf{A} &= \mathbf{I} \\ \mathbf{PA} &= \mathbf{I} \end{aligned}$$

where $\mathbf{P} = \mathbf{E}_k\mathbf{E}_{k-1}\cdots\mathbf{E}_2\mathbf{E}_1$. □

Theorem. *If* $\mathbf{PA} = \mathbf{I}$ *then* $\mathbf{AP} = \mathbf{I}$.

Proof.

$$\begin{aligned} \mathbf{PA} &= \mathbf{I} \\ \mathbf{PAP} &= \mathbf{IP} \\ \mathbf{P}(\mathbf{AP}) &= \mathbf{P} \end{aligned}$$

Since $\mathbf{P}$ multiplied by $\mathbf{AP}$ gives $\mathbf{P}$ back again, $\mathbf{AP}$ must equal the identity matrix ($\mathbf{AP} = \mathbf{I}$). □

Theorem. *The inverse of a matrix is unqiue.*

Proof. Let $\mathbf{A}^{-1}$ be the inverse of $\mathbf{A}$, i.e. $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{AA}^{-1}$. Suppose there exists a matrix $\mathbf{P} \neq \mathbf{A}^{-1}$ such that $\mathbf{PA} = \mathbf{I} = \mathbf{AP}$.

$$\begin{aligned} \mathbf{PA} &= \mathbf{I} \\ \mathbf{PAA}^{-1} &= \mathbf{IA}^{-1} \\ \mathbf{PI} &= \mathbf{A}^{-1} \\ \mathbf{P} &= \mathbf{A}^{-1} \end{aligned}$$

This contradicts our supposition that $\mathbf{P} \neq \mathbf{A}^{-1}$, so $\mathbf{A}^{-1}$ must be unique. □

4.4 Computing the Matrix Inverse

Our proof of existence of the matrix inverse also provided a method of construction. We performed Gaussian elimination on a matrix, constructing an elementary matrix for each step. These elementary matrices were multiplied together to form the matrix inverse. In practice, this method would be wildly inefficient. Transforming an $n \times n$ matrix to reduced row echelon form requires $\mathcal{O}(n^2)$ elementary row operations, so we would need to construct and multiply $\mathcal{O}(n^2)$ elementary matrices. Since naive matrix multiplication requires $\mathcal{O}(n^3)$ operations per matrix, constructing a matrix inverse with this method requires $\mathcal{O}(n^5)$ operations! Since Gaussian elimination is $\mathcal{O}(n^3)$, we would be far better off avoiding the matrix inverse entirely.

The fastest matrix multiplication algorithm is $\mathcal{O}(n^{2.7373})$, although this is not a big help in practice.

Fortunately, there are better methods for constructing matrix inverses. One of the best is called the *side-by-side method*. To see how the side-by-side method works, consider constructing an inverse for the square matrix $\mathbf{A}$. We can use Gaussian elimination to transform $\mathbf{A}$ into the identity matrix, which we can represent with a series of k elementary matrices.

$$\underbrace{\mathbf{E}_k\mathbf{E}_{k-1}\cdots\mathbf{E}_2\mathbf{E}_1}_{\mathbf{A}^{-1}}\mathbf{A} = \mathbf{I}$$

What would happen if we simultaneously apply the same elementary row operations to another identity matrix?

$$\underbrace{\mathbf{E}_k\mathbf{E}_{k-1}\cdots\mathbf{E}_2\mathbf{E}_1}_{\mathbf{A}^{-1}}\mathbf{I} = \mathbf{A}^{-1}\mathbf{I} = \mathbf{A}^{-1}$$

In the side-by-side method, we start with an augmented matrix containing the $n \times n$ matrix $\mathbf{A}$ and an $n \times n$ identity matrix. Then we apply Gaussian elimination to transform $\mathbf{A}$ into $\mathbf{I}$. The augmented matrix ensure that the same elementary row operations will be applied to the identity matrix, yielding the inverse of $\mathbf{A}$:

$$\begin{pmatrix}\mathbf{A} & \mathbf{I}\end{pmatrix} \xrightarrow{\text{EROs}} \begin{pmatrix}\mathbf{I} & \mathbf{A}^{-1}\end{pmatrix}$$

Like the augmented matrix $(\mathbf{A}\,\mathbf{y})$, there is no direct interpretation of $(\mathbf{A}\,\mathbf{I})$. It is simply a convenient way to apply the same EROs to both $\mathbf{A}$ and $\mathbf{I}$.

Let's solve the following system by constructing the matrix inverse.

$$\begin{aligned} 3x_1 + 2x_2 &= 7 \\ x_1 + x_2 &= 4 \end{aligned}$$

In matrix form,

$$\mathbf{A} = \begin{pmatrix} 3 & 2 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 7 \\ 4 \end{pmatrix}$$

We start with the augmented matrix for the side-by-side method.

$$\begin{pmatrix} 3 & 2 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \xrightarrow{R_1 \leftrightarrow R_2} \begin{pmatrix} 1 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{pmatrix}$$
$$\xrightarrow{R_2 - 3R_1} \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & -1 & 1 & -3 \end{pmatrix}$$
$$\xrightarrow{-R_2} \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & -1 & 3 \end{pmatrix}$$
$$\xrightarrow{R_1 - R_2} \begin{pmatrix} 1 & 0 & 1 & -2 \\ 0 & 1 & -1 & 3 \end{pmatrix}$$

Thus, the matrix inverse is

$$\mathbf{A}^{-1} = \begin{pmatrix} 1 & -2 \\ -1 & 3 \end{pmatrix}$$

and we can compute the solution by matrix multiplication.

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y} = \begin{pmatrix} 1 & -2 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} 7 \\ 4 \end{pmatrix} = \begin{pmatrix} -1 \\ 5 \end{pmatrix}$$

We can verify that $\mathbf{A}^{-1}$ is a matrix inverse

$$\mathbf{A}^{-1}\mathbf{A} = \begin{pmatrix} 1 & -2 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} 3 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{I}$$
$$\mathbf{A}\mathbf{A}^{-1} = \begin{pmatrix} 3 & 2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{I}$$

The nice thing about having a matrix inverse is that if only the right hand side of a system of equations change, we do not need to retransform the coefficient matrix. For example, to solve

$$\begin{aligned} 3x_1 + 2x_2 &= 1 \\ x_1 + x_2 &= 3 \end{aligned}$$

we can re-use $\mathbf{A}^{-1}$ since $\mathbf{A}$ is unchanged (only $\mathbf{y}$ is different).

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y} = \begin{pmatrix} 1 & -2 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} -5 \\ 8 \end{pmatrix}$$

4.5 Numerical Issues

Matrix inverses are a powerful method for solving linear systems. However, calculating a matrix inverse should always be your last resort. There are far more efficient and numerically stable methods for solving linear systems. Reasons against using a matrix inverse include:

1. **Computation time.** While the side-by-side method is more efficient for constructing inverses that multiplying elementary matrices, it is still slower than Gaussian elimination for solving linear systems of the form $\mathbf{A}\mathbf{x}= \mathbf{y}$. Both side-by-side and Gaussian elimination reduce an augmented matrix. For an $n \times n$ matrix $\mathbf{A}$, the augmented matrix for Gaussian elimination $(\mathbf{A}\,\mathbf{y})$ is $n\times(n+1)$. The augmented matrix for the side-by-side method $(\mathbf{A}\ \mathbf{I})$ is $n \times 2n$. Solving for the inverse requires nearly twice the computations as solving the linear system directly. Having the inverse allows us to "resolve" the system for a new right hand side ($\mathbf{y}$) for only the cost of a matrix multiplication. However, there are variants of Gaussian elimination – such as LU decomposition – that allow resolving without repeating the entire reduction of the coefficient matrix $\mathbf{A}$.

2. **Memory.** Most large matrices in engineering are *sparse*. Sparse matrices contain very few nonzero entries; matrices with less than 0.01% nonzero entries are not uncommon. Examples of sparse matrices include matrices generated from finite difference approximations or matrices showing connections between nodes in large networks. Computers store sparse matrices by only storing the nonzero entries and their locations. However, there is no guarantee that the inverse of a sparse matrix will also be sparse. Consider the arrow matrix, a matrix with ones along the diagonal and last column and row.

Imagine the connection matrix for Facebook. It would have hundreds of millions of rows and columns, but each person (row) would only have nonzero entries for a few hundred people (columns) that they knew.

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

An $n \times n$ arrow matrix has n^2 entries but only $3n - 2$ nonzeros. However, the inverse of an arrow matrix always has 100% nonzeros. For example, the inverse of the 8×8 matrix above is

A 1000×1000 arrow matrix has less than 0.3% nonzeros.

$$\mathbf{A}^{-1} = \frac{1}{6} \begin{pmatrix} 5 & -1 & -1 & -1 & -1 & -1 & -1 & 1 \\ -1 & 5 & -1 & -1 & -1 & -1 & -1 & 1 \\ -1 & -1 & 5 & -1 & -1 & -1 & -1 & 1 \\ -1 & -1 & -1 & 5 & -1 & -1 & -1 & 1 \\ -1 & -1 & -1 & -1 & 5 & -1 & -1 & 1 \\ -1 & -1 & -1 & -1 & -1 & 5 & -1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & 5 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 \end{pmatrix}$$

Calculating the inverse of a large, sparse matrix could requires orders of magnitude more memory than the original matrix. For some matrices, storing – much less computing – the inverse is impossible.

Despite its disadvantages, the matrix inverse is still a powerful construct. A multiplicative inverse for matrices is necessary for many algebraic manipulations, and the inverse can be used to simply or prove many matrix equations. Just remember to think critically about the need for a matrix inverse before calculating one.

4.6 Inverses of Elementary Matrices

We conclude with another interesting property of elementary matrices. We said before that left multiplication by an elementary matrix performs an elementary row operation (the same ERO that was used to construct the elementary matrix). Left multiplication by the inverse of a elementary matrix "undoes" the operation of the elementary matrix. For example, the elementary matrix $\mathbf{E}_{3R_2}$ scales the second row by two. The inverse $\mathbf{E}_{3R_2}^{-1}$ would scale the second row by 1/2, undoing the scaling by two. Similarly, $\mathbf{E}_{R_2 \leftrightarrow R_3}$ swaps rows two and three, and $\mathbf{E}_{R_2 \leftrightarrow R_3}^{-1}$ swaps them back. The proof of this property is straightforward.

Theorem. *If the elementary matrix $\mathbf{E}_r$ performs the elementary row operation r, then left multiplication by the inverse $\mathbf{E}_r^{-1}$ undoes this operation.*

Proof.

$$\mathbf{E}_r^{-1}(\mathbf{E}_r\mathbf{A}) = (\mathbf{E}_r^{-1}\mathbf{E}_r)\mathbf{A} = (\mathbf{I})\mathbf{A} = \mathbf{A}$$

□

4.7 Rank

Consider the linear system

$$\begin{pmatrix} 1 & 0 & 3 \\ 0 & 2 & -4 \\ -2 & 0 & -6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \\ -4 \end{pmatrix}$$

and the row echelon form of the associated augmented matrix

$$\begin{pmatrix} 1 & 0 & 3 & 2 \\ 0 & 2 & -4 & -2 \\ -2 & 0 & -6 & -4 \end{pmatrix} \xrightarrow{\frac{1}{2}R_2} \xrightarrow{R_3+2R_1} \begin{pmatrix} 1 & 0 & 3 & 2 \\ 0 & 1 & -2 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Notice that the last row is all zeros. We have no information about the last entry (x_3). However, this does not mean we cannot solve the linear system. Since x_3 is unknown, let us assign its value the symbol α. Then, by back substitution

$$\begin{aligned} x_3 &= \alpha \\ x_2 - 2x_3 = -1 \Rightarrow \quad x_2 &= 2\alpha - 1 \\ x_1 + 3x_3 = 2 \Rightarrow \quad x_1 &= 2 - 3\alpha \end{aligned}$$

The above linear system has not one solution, but infinitely many. There is a solution for every value of the parameter α, so we say the system has a *parameterized solution*.

Parameterized solutions are necessary any time row echelon reduction of a matrix leads to one or more rows with all zero entries. The number of nonzero rows in the row echelon form of a matrix is the matrix's *rank*. The rank of a matrix can be calculated by counting the number of nonzero rows after the matrix is transformed into row echelon form by Gaussian elimination. In general, if a matrix with n columns has rank n, it is possible to find a unique solution to the system $\mathbf{Ax} = \mathbf{y}$. If $\text{rank}(\mathbf{A}) < n$, there may be infinitely many solutions. These solutions require that we specify $n - \text{rank}(\mathbf{A})$ parameters.

We denote the rank of a matrix $\mathbf{A}$ as $\text{rank}(\mathbf{A})$.

Matrices has both a *row rank* (the number of nonzero rows in row-reduced echelon form) and a *column rank* (the number of nonzero columns in a column-reduced echelon form). Thus the concept of rank also applies to nonsquare matrices. However, the row and column ranks are always equivalent, even if the matrix is not square:

Theorem. *The row rank of a matrix equals the column rank of the matrix, i.e.* $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^{\text{T}})$.

Proof. We will defer the proof of this theorem until our discussion of the Fundamental Theorem of Linear Algebra. □

The equivalence of the row and column ranks implies an upper bound on the rank of nonsquare matrices.

Theorem. *The rank of a matrix is less than or equal to the smallest dimension of the matrix, i.e.* $\text{rank}(\mathbf{A}) \leq \min(\dim \mathbf{A})$.

Proof. The row rank of $\mathbf{A}$ is the number of nonzero rows in the row-reduced $\mathbf{A}$, so the rank of $\mathbf{A}$ must be less than the number of rows in $\mathbf{A}$. Since the row rank is also equal to the column rank, there must also be $\text{rank}(\mathbf{A})$ nonzero columns in the column-reduced $\mathbf{A}$. So the rank of $\mathbf{A}$ must never be larger than either the number of rows or number of columns in $\mathbf{A}$. □

A matrix that has the maximum possible rank (rank n for an $n \times n$ square matrix or rank $\min(m, n)$ for an $m \times n$ rectangular matrix), we say the matrix is *full rank*. A matrix that is not full rank is *rank deficient*.

Linear Independence

The notion of rank is deeply tied to the concept of *linear independence.* A vector $\mathbf{x}_i$ is linearly dependent on a set of n vectors $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$ if there exits coefficients c_1, c_2, …, c_n such that

$$\mathbf{x}_i = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n$$

Note the difference between $\mathbf{x}_1, \ldots, \mathbf{x}_n$, a set of n vectors; and $x_1, \ldots, x_n$, a set of n scalars that for the elements of a vector $\mathbf{x}$.

A set of vectors are *linearly dependent* if one of the vectors can be expressed as a linear combination of some of the others. This is analogous to saying there exists a set of coefficients $c_1, \ldots, c_n$, not all equal to zero, such that

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n = 0$$

If a matrix with n rows has rank $k < n$, then $n - k$ of the rows are linearly dependent on the other k rows. Going back to our previous example, the matrix

$$\begin{pmatrix} 1 & 0 & 3 \\ 0 & 2 & -4 \\ -2 & 0 & -6 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{pmatrix}$$

has rank 2 since there are two nonzero rows in the row-reduced matrix. Therefore, one of the rows must be linearly dependent on the other rows. Indeed, we see that the last row $\begin{pmatrix} -2 & 0 & -6 \end{pmatrix}$ is -2 times the first row $\begin{pmatrix} 1 & 0 & 3 \end{pmatrix}$. During row reduction by Gaussian elimination, any linearly dependent rows will be completely zeroed out, revealing the rank of the matrix.

Rank and linear dependence tell us about the information content of a coefficient matrix. If some of the rows of the coefficient matrix are linearly dependent, then matrix is rank deficient and no unique solution exists. These matrices are also information deficient – we do not have one independent expression for each variable. Without a separate piece of information for each variable, we cannot unique map between the input $\mathbf{x}$ and the output $\mathbf{y}$. However, if we introduce a separate parameter for each zeroed row, we are artificially providing the missing information. We can find a new solution every time we specify values for the parameters.

Homogeneous Systems ($\mathbf{Ax} = \mathbf{0}$)

A linear systems of equations is *homogeneous* if and only if the right hand side vector ($\mathbf{y}$) is equal to the zero vector ($\mathbf{0}$). Homogeneous systems always have at least one solution, $\mathbf{x} = \mathbf{0}$, since $\mathbf{A0} = \mathbf{0}$. The zero solution to a homogeneous system is called the *trivial solution.*

Some homogeneous systems have a nontrivial solution, i.e. a solution $\mathbf{x} \neq \mathbf{0}$. If a homogeneous system has a nontrivial solution, then it has infinitely many solutions, a result we state as follows

Theorem. *Any linear combination of nontrivial solutions to a homogeneous linear system is also a solution.*

This proof is equivalent to showing that $\mathbf{Ax} = \mathbf{0}$ satisfies our definition of linear systems: $f(k_1 x_1 + k_2 x_2) = k_1 f(x_1) + k_2 f(x_2)$.

Proof. Suppose we had two solutions, $\mathbf{x}$ and $\mathbf{x}'$ to the homogeneous system $\mathbf{Ax} = \mathbf{0}$. Then

$$\begin{aligned}\mathbf{A}(k\mathbf{x} + k'\mathbf{x}') &= \mathbf{A}(k\mathbf{x}) + \mathbf{A}(k'\mathbf{x}') \\ &= k(\mathbf{Ax}) + k'(\mathbf{Ax}') \\ &= k(\mathbf{0}) + k'(\mathbf{0}) \\ &= \mathbf{0}\end{aligned}$$

Since there are infinitely many scalars k and k', we can generate infinitely many solutions to the homogeneous system $\mathbf{Ax} = \mathbf{0}$. □

There is a connection between the solvability of nonhomogeneous systems $\mathbf{Ax} = \mathbf{y}$ and the corresponding homogeneous system $\mathbf{Ax} = \mathbf{0}$. If there exists at least one solution to $\mathbf{Ax} = \mathbf{y}$ and a nontrivial solution to $\mathbf{Ax} = \mathbf{0}$, then there are infinitely many solutions to $\mathbf{Ax} = \mathbf{y}$. To see why, let $\mathbf{x}_{\text{nh}}$ be the solution to the nonhomogeneous system ($\mathbf{Ax}_{\text{nh}} = \mathbf{y}$) and $\mathbf{x}_{\text{h}}$ be a nontrivial solution to the homogeneous system $\mathbf{Ax}_{\text{h}} = \mathbf{0}$. Then any of the infinite linear combinations $\mathbf{x}_{\text{nh}} + k\mathbf{x}_{\text{h}}$ is also a solution to $\mathbf{Ax} = \mathbf{y}$ since

$$\mathbf{A}(\mathbf{x}_{\text{nh}} + k\mathbf{x}_{\text{h}}) = \mathbf{Ax}_{\text{nh}} + k\mathbf{Ax}_{\text{h}} = \mathbf{y} + k\mathbf{0} = \mathbf{y}$$

General Solvability

For any linear system of equations, we can use the rank of the coefficient matrix and the augmented matrix to determine the existence and number of solutions. The relationship between solvability and rank is captured by the Rouché-Capelli theorem:

Theorem. *A linear system* $\mathbf{Ax} = \mathbf{y}$ *where* $\mathbf{x} \in \mathbb{R}^n$ *has a solution if and only if the rank of the coefficient matrix equals the rank of the augmented matrix, i.e.* $\text{rank}(\mathbf{A}) = \text{rank}([\mathbf{A}\,\mathbf{y}])$. *Furthermore, the solution is unique if* $\text{rank}(\mathbf{A}) = n$*; otherwise there are infinitely many solutions.*

Proof. We will sketch several portions of this proof to give intuition about the theorem. A more rigorous proof is beyond the scope of this class.

1. **Homogeneous systems.** For a homogeneous system $\mathbf{Ax} = \mathbf{0}$, we know that $\text{rank}(\mathbf{A}) = \text{rank}([\mathbf{A}\,\mathbf{0}])$. (Since the rank of $\mathbf{A}$ is equal to the number of nonzero columns, adding another column of zeros will never change the rank.) Thus, we know that homogeneous systems are always solvable, at least by the trivial solution $\mathbf{x} = \mathbf{0}$. If $\text{rank}(\mathbf{A}) = n$, then the trivial solution is unique and is the only solution. If $\text{rank}(\mathbf{A}) < n$, there are infinitely many parameterized solutions.

2. **Full rank, nonhomogeneous systems**. For a nonhomogeneous system ($\mathbf{Ax} = \mathbf{y}$, $\mathbf{y} \neq \mathbf{0}$), we expect a unique solution if and only if adding the column $\mathbf{y}$ to the coefficient matrix doesn't change the rank. For this to be true, $\mathbf{y}$ must be linearly dependent on the other columns in $\mathbf{A}$; otherwise, adding a new linearly independent column would increase the rank. If $\mathbf{y}$ is linearly dependent on the n columns of $\mathbf{A}$, it must be true that there exists weights c_1, c_2, …, c_n such that

$$c_1\mathbf{A}(:,1) + c_2\mathbf{A}(:,2) + \cdots + c_n\mathbf{A}(:,n) = \mathbf{y}$$

based on the definition of linear dependence. But the above expression can be rewritten in matrix form as

$$(\mathbf{A}(:,1)\,\mathbf{A}(:,2)\,\cdots\,\mathbf{A}(:,n)) \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} = \mathbf{Ac} = \mathbf{y}$$

which shows that the system has a unique solution $\mathbf{x} = \mathbf{c}$.

3. **Rank deficient, nonhomogeneous systems**. Let $\text{rank}(\mathbf{A}) = k < n$. Then the row-reduced form of $\mathbf{A}$ will have k rows that resemble the identity matrix and $n - k$ rows of all zeros:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2k} & \cdots & a_{2n} \\ & \vdots & & & \vdots & \\ a_{k1} & a_{k2} & \cdots & a_{kk} & \cdots & a_{kn} \\ a_{k+1,1} & a_{k+1,2} & \cdots & a_{k+1,k} & \cdots & a_{k+1,n} \\ & \vdots & & & \vdots & \\ a_{n1} & a_{n2} & \cdots & a_{nk} & \cdots & a_{nn} \end{pmatrix} \xrightarrow{\text{EROs}} \begin{pmatrix} 1 & a'_{12} & \cdots & a'_{1k} & \cdots & a'_{1n} \\ 0 & 1 & \cdots & a'_{2k} & \cdots & a'_{2n} \\ & \vdots & & & \vdots & \\ 0 & 0 & \cdots & 1 & \cdots & a'_{kn} \\ 0 & 0 & \cdots & 0 & \cdots & 0 \\ & \vdots & & & \vdots & \\ 0 & 0 & \cdots & 0 & \cdots & 0 \end{pmatrix}$$

Now imagine we performed the same row reduction on the augmented matrix $(\mathbf{A}\,\mathbf{y})$. We would still end up with $n - k$ rows with zeros in the first n columns (the columns of $\mathbf{A}$):

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} & \cdots & a_{1n} & y_1 \\ a_{21} & a_{22} & \cdots & a_{2k} & \cdots & a_{2n} & y_2 \\ & \vdots & & & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} & \cdots & a_{kn} & y_k \\ a_{k+1,1} & a_{k+1,2} & \cdots & a_{k+1,k} & \cdots & a_{k+1,n} & y_{k+1} \\ & \vdots & & & \vdots & & \\ a_{n1} & a_{n2} & \cdots & a_{nk} & \cdots & a_{nn} & y_n \end{pmatrix} \xrightarrow{\text{EROs}} \begin{pmatrix} 1 & a'_{12} & \cdots & a'_{1k} & \cdots & a'_{1n} & y'_1 \\ 0 & 1 & \cdots & a'_{2k} & \cdots & a'_{2n} & y'_2 \\ & \vdots & & & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 & \cdots & a'_{kn} & y'_k \\ 0 & 0 & \cdots & 0 & \cdots & 0 & y'_{k+1} \\ & \vdots & & & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 & y'_n \end{pmatrix}$$

We know that if $y'_{k+1}, \ldots, y'_n = 0$, we can solve this system by designating $n - k$ parameters for the variables $x_{k+1}, \ldots, x_n$ for which we have no information. However, notice what happens if any of the values $y'_{k+1}, \ldots, y'_n$ are nonzero. Then we have an expression of the form $0 = y'_i \neq 0$, which is nonsensical. Therefore, the only way we can solve this system is by requiring that $y'_{k+1}, \ldots, y'_n = 0$. This is exactly the requirement that the rank of the augmented matrix equal k, the rank of the matrix $\mathbf{A}$ by itself. If any of the $y'_{k+1}, \ldots, y'_n$ are nonzero, then the augmented matrix has one fewer row of zeros, so the rank of the augmented matrix would be greater than the rank of the original matrix. There are two ways to interpret this result. First, we require that the right hand side $\mathbf{y}$ doesn't "mess up" our system by introducing a nonsensical expression. Second, if a row i in the matrix $\mathbf{A}$ is linearly dependent on the other rows in $\mathbf{A}$, the corresponding values y_i must have the same dependency on the other values in $\mathbf{y}$. If so, when the row i is zeroed out during row reduction, the value y_i will also be zeroed out, avoiding any inconsistency.

□

4.8 Rank and Matrix Inverses

For a general nonhomogeneous system $\mathbf{Ax} = \mathbf{y}$ with $\mathbf{A} \in \mathbb{R}^{n \times n}$, we know that a unique solution only exists if $\text{rank}(\mathbf{A}) = \text{rank}([\mathbf{A}\,\mathbf{y}]) = n$. If $\text{rank}(\mathbf{A}) = n$, we know that $\mathbf{A}$ can be transformed into reduced row form without generating any rows with all zero entries. We also know that if an inverse $\mathbf{A}^{-1}$ exists for $\mathbf{A}$, we can use the inverse to uniquely solve for $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$. Putting these facts together, we can now definitely state necessary and sufficient conditions for matrix inversion:

An $n \times n$ matrix $\mathbf{A}$ has an inverse if and only if $\text{rank}(\mathbf{A}) = n$.

4.9 Summary

We've shown in this chapter the tight connections between matrix inversion, solvability, and the rank of a matrix. We will use these concepts many times to understand the solution of linear systems. However, we've also argued that despite their theoretical importance, these concepts have limited practical utility for solving linear systems. For example, computing the rank of a matrix requires transforming the matrix into reduced echelon form. This requires the same computations as solving a linear system involving the matrix, so one would rarely check the rank of a coefficient matrix before attempting to solve a linear system. Instead, we will see rank emerge as a useful tool only when considering matrices by themselves in Part III of this course.