

Data Science for Engineers
Prof. Raghunathan Rangaswamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 14
Solving Linear Equations

We will continue the lecture on solving linear equations. In the last lecture I discussed the case of many more equations and variables, where we might not have a solution and how we can use an optimization perspective to find a solution. In this lecture I am going to give you some examples for that case show you what happens when we apply the solution that we derived last time, and then after that I will go on to look at the case of more variables than equations.

(Refer Slide Time: 00:49)

The slide, titled "Case 2: Example - I", displays a linear system $\begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -0.5 \\ 5 \end{bmatrix}$. Below this, it notes $m = 3, n = 2$ and states "Using the optimization concept,". The least squares solution is given as $x = (A^T A)^{-1} A^T b$. This is followed by the explicit calculation: $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \left(\begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -0.5 \\ 5 \end{bmatrix}$. Red circles highlight the coefficient matrix and the $(A^T A)^{-1}$ term in the formula.

So, let us look at an $Ax = b$ example system as shown in the screen. Here we have a matrix with 3 rows and 2 columns, which basically means that there are 3 equations in 2 variables number of equations more than number of variables. And we have to read these equations as $x_1 = 1$, $2x_1 = -0.5$ and $3x_1 + x_2 = 5$.

So, if you notice these equations you would realize that the first 2 equations are inconsistent. For example, if we were to take the first equation is true then $x_1 = 1$ and if we substitute that value into the second equation you will get $2 = -0.05$. If you were to take the second equation as true then $2x_1$ is -0.5 . So, x_1 will be -0.25 and that would

not solve the first equation. So, these 2 equations are inconsistent. The third equation since it is $3x_1 + x_2$ irrespective of whatever value you get for x_1 you can always use this equation to calculate the value for x_2 ; however, we cannot solve this set of equations.

Now, let us see what is the solution that we get, by using the optimization concept that we described in the last lecture. We said $x = A^T A \text{ inverse } A^T b$. the A matrix is 1 0 2 0 3 1. So, A^T matrix is 1 2 3 0 0 1. Simply plugging in the matrices here.

(Refer Slide Time: 02:30)

Data science for Engineers

Case 2: Example continued

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.2 & -0.6 \\ -0.6 & 2.8 \end{bmatrix} \begin{bmatrix} 15 \\ 5 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 5 \end{bmatrix}$$

- Thus, the solution for the given example is $(x_1, x_2) = (0, 5)$
- Substituting in the equation shows

$$\begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 5 \end{bmatrix} \neq \begin{bmatrix} 1 \\ -0.5 \\ 5 \end{bmatrix}$$

Linear Algebra 3

And then doing the calculation gives us this equation which says x_1 x_2 is a matrix times 15 5. This is an intermediate step for the calculation. And when you further simplify it you get a solution $x_1 = 0$, $x_2 = 5$. Notice that the optimum solution here that is chosen does not have either one of the 2 cases that we talked about in the last slide, which is $x_1 = 1$ and $x_1 = -0.25$ the optimization approach chooses $x_1 = 0$ and $x_2 = 5$ and when you substitute it back into the equation you get b as 0 0 5 whereas, the actual b that we are interested in is 1 - 0.55.

So, you can see that while the third equation is being solved exactly the first take both the first 2 equations are not solve for; however, as we described before this is the best solution in a collective minimization of error sense, which is what we defined as minimizing sum of squared of errors. We will now move on to the next example.

(Refer Slide Time: 03:47)


Data science for Engineers

Case 2: Example

$$\begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$$

- $m = 3, n = 2$
- Using the optimization concept,
$$\mathbf{x} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}$$
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \left(\begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$$

Linear Algebra



Let us consider another example for us to illustrate something different here. We have taken the same left hand side we have the same a matrix; however, the right-hand side has been modified to be 1 2 5. we have done this for a specific reason which we will see presently. So, when you look at this equation; if you take the first equation it reads as $x_1 = 1$. If you look at the second equation it reads as $2x_1 = 2$. The third equation reads as $3x_1 + x_2 = 5$.

So, from the first equation you can get a solution for $x_1 = 1$ and the second equation since it reads as $2x_1 = 2$, we have to simply substitute the solution that we get from the first equation and see whether the second equation is also satisfied since $x_1 = 1$ 2 times x_1 2 times 1 is 2 the second equation is also satisfied.

Now, let us see what happens to the third equation. The third equation reads as $3x_1 + x_2 = 5$, we already know $x_1 = 1$ satisfies the first 2 equations. So, $3x_1 + x_2 = 5$ would give you $x_2 = 2$. Now you notice that if I get a solution 1 and 2 for x_1 and x_2 ; though the number of equations are more than the variables, the equations are in such a way that I can get a solution for x_1 and x_2 that satisfies all the 3 equations.

Now let us see whether the expression that we had for this case actually uncovers this solution. So, we said $\mathbf{x} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}$ and we do the same manipulation as the last example except that this \mathbf{b} has become 1 2 5 now.

(Refer Slide Time: 05:40)

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.2 & -0.6 \\ -0.6 & 2.8 \end{bmatrix} \begin{bmatrix} 20 \\ 5 \end{bmatrix}$$
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$


- Thus, the solution for the given example is $(x_1, x_2) = (1, 2)$
- Substituting in the equation shows
$$\begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$$

R Code

```
A=matrix(c(1,2,3,0,0,1),ncol=2, byrow=F)
b=matrix(c(1,2.5),ncol=2, byrow=F)
x=inv(t(A)%*%A)%*%t(A)%*%b
x
```

Console output

```
> x=inv(t(A)%*%A)%*%t(A)%*%b
> x
[1]
[1,] 1
[2,] 2
>
```



Linear Algebra

After some more calculations you will see that $x_1 = 1$ $x_2 = 2$. Thus, the solution is 1 2 and we had already verified that this would solve the equation and we had verified that 1 2 is a solution that we can directly get by observation from the previous slide.


So, the important point here is that if we have more equations than variables then you can always use this least square solution which is $(A^T A)^{-1} A^T b$. The only thing to keep in mind is that $(A^T A)^{-1}$ exists if the columns of A are linearly independent. If the columns of A are not linearly independent, then we have to do something else which you will see as we go through this lecture.

(Refer Slide Time: 06:32)

Data science for Engineers

Case 3: $m < n$

- This case addresses the problem of more attributes or variables than equations
- Since the number of attributes is greater than the number of equations, one can obtain multiple solutions for the attributes
- This is termed as an infinite-solution case
- How does one choose a single solution from the set of infinite possible solutions?



Linear Algebra

So, that finishes the case where the number of equations are more than the number of variables. Now let us address the last case where the number of equations are less than the number of variables, which would be m less than n in this case we address the problem of more attributes or variables than equations.

Now since I have many more variables and equations I would have infinite number of solutions the way to think about this is the following. If I had, let us say, 2 equations and 3 variables. You can think of this situation as one where you could choose any value for x_3 and then simply put it into the 2 equations. And whatever are the terms with respect to x_3 you collect them and take them to the right-hand side; that would leave you with 2 equations and 2 variables and once we solve for that 2 equations and 2 variables we will get values for x_1 and x_2 .

So, basically what this means is that, I can choose any value for x_3 and then corresponding to that I will get values for x_1 and x_2 . So, I will get infinite number of solutions. Since I have infinite number of solutions then the question that I ask is how do I find one single solution from the set of infinite possible solutions? Clearly if you are looking at only solvability of the equation, there is no way to distinguish between this infinite possible solutions. So, we need to bring some other metric that we could possibly use, which would have some value for us to pick one solution that we can say is a solution to this case.

(Refer Slide Time: 08:12)


Data science for Engineers
Case 3: An optimization perspective

- Pose the following optimization problem

$$\min \left(\frac{1}{2} x^T x \right) \text{ s.t. } Ax = b$$
- Define a Lagrangian function $f(x, \lambda)$

$$\min \left[f(x, \lambda) = \frac{1}{2} x^T x + \lambda^T (Ax - b) \right]$$
- Differentiating the Lagrangian with respect to x , and setting to zero

$$x + A^T \lambda = 0$$



Linear Algebra

Similar to the previous example we are going to take an optimization view here, what we are going to do is we are going to minimize $x^T x$, this half is just to make sure the solution comes out in a

nice form. And notice here something that is important we also have a constraint for this optimization problem $s \cdot t$ means subject to.

So, I want to minimize this half; $x^T x$ subject to the constraint $Ax = b$. So, in other words what we are saying is whatever solution we get for x that has to necessarily satisfy this equation. And this is not a problem we can find infinite number of solutions x which will satisfy these equations. So, what this objective does is of all of those solutions how do I pick, that one solution which will minimize this $x^T x$. We have to think about a rationale for, why we would choose $x^T x$ as an objective.

This basically says that of all the solutions I want the solution which is closest to the origin is what this is saying in terms of $x^T x$. From an engineering viewpoint one could justify this as the following; if you have lots of design parameters that you are trying to optimize and so on, you would like to keep the sizes small for example, so you might want small numbers. So, you want to be as close to origin as possible this is just one justification for doing something like this nonetheless this is one way of picking one solution from this infinite number of solutions.

Now, in the previous example and in this example, we are solving these optimization problems; however, we have not taught in this course how to solve optimization problems. For people who already know how to solve optimization problems this would be obvious. For other participants who do not know how to solve optimization problems, I would encourage you to just bear with me and then go through this solution and see what the solution form is and once this module on linear algebra is finished we will have a couple of modules on optimization from the viewpoint of data science.

So, when we do that, you will see how we solve these kinds of optimization problems. The optimization problem that we solved for the last case is what is called an unconstrained optimization problem because there are no constraints to that problem whereas, this problem that we are solving is called a constrained optimization problem because while we have an objective we also have a set of constraints that we need to solve.

So, you will have to bear with us till you go through the optimization module to understand this. Interestingly it is generally a good idea to teach linear algebra on optimization, but interestingly, some of the linear algebra concepts you can view as optimization problems and solving optimization problems requires lots of linear algebra concepts. So, in that sense they are both coupled. In any case to solve optimization problems of this form we can define what is called a Lagrangian function $f(x)$ comma λ , λ are extra parameters that we introduce into this optimization formulation. And what you do is you

minimize this Lagrangian with respect to x to get a set of equations. And you also minimize this with respect to Lagrangian which will back out the constraint. So, whatever solution you have, has to solve both the differentiation with respect to x which should give you $x + A^T \lambda = 0$ and also differentiation with λ which will simply give you $Ax - b = 0$. That would basically say that whatever solution you get, that has to satisfy the equation $Ax = b$ we will see how this is useful in identifying a solution.

(Refer Slide Time: 12:35)

Data science for Engineers

Case 3: An optimization perspective

$$x = -A^T \lambda$$

Pre-multiplying by A
 $Ax = b = -AA^T \lambda$

Thus we obtain $\lambda = -(AA^T)^{-1}b$ assuming that all the rows are linearly independent

$$x = -A^T \lambda = A^T (AA^T)^{-1} b$$

Linear Algebra

So, let us look at this equation $x + A^T \lambda = 0$. So, from this we can get a solution for x which is $-A^T \lambda$. Now what you could do is; you do not know x and you do not know λ also. So, there has to be some way of finding out both of them. So, what we are going to do is we are going to use the knowledge that any solution that we get has to satisfy the equation $Ax = b$.

So, what we are going to do is we are going to pre-multiply this x by A . So, we pre-multiply on both sides so, we get $Ax = -AA^T \lambda$ by pre-multiplying this equation by A . Now since any solution x satisfies $Ax = b$, I can replace this Ax by b and I get this equation $b = -AA^T \lambda$ and from this equation we can get λ to be $-(AA^T)^{-1}b$. And this is possible and this inverse exists only if all the rows are linearly independent.

Now, since we have an expression for λ we can substitute that expression here and we will get $x = -A^T \lambda$ and your λ is this expression which is from here. So, this solves for x in the equation $Ax = b$. And since we use this idea here the x that we get is such that $Ax = b$ that is satisfies the original equation.

(Refer Slide Time: 14:17)

Data science for Engineers

Case 3: Example


$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

- $m = 2, n = 3$
- Using the optimization concept,

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

$$x = \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Linear Algebra



Now, let us take an example to understand this. I have an $A x = b$ here, I have a as 1 2 3 0 0 1 and b as 2 1. So, again notice here since there are 2 equations, I have 2 rows and 3 columns, I have 3 variables, these equations are read as $x_1 + 2x_2 + 3x_3 = 2$ and $x_3 = 1$. Now clearly when you look at this equation you will notice that $x_3 = 1$ has to be a solution. So, the question is how do I choose x_1 and x_2 , nonetheless we will use the optimization solution to actually see what happens here.

So, the optimization solution from the previous slide is the following $x = A^T (A A^T)^{-1} b$. Now A^T is 1 2 3 0 0 1 here. And this is my A and A^T again I take an inverse of this and b now is 2 1.

(Refer Slide Time: 15:28)

Data science for Engineers

Case 3: Example

$$x = \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 14 & 3 \\ 3 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$x = \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} -0.2 \\ 1.6 \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -0.2 \\ -0.4 \\ 1 \end{bmatrix}$$

- The solution for the given example is $(x_1, x_2, x_3) = (-0.2, -0.4, 1)$


Linear Algebra

R Code

```
A=matrix(c(1,0,2,0,3,1),ncol=3)
b=c(2,1)
library(MASS)
x=t(A)%*%inv(A%*%t(A))%*%b
x
```

Console output

```
A=matrix(c(1,0,2,0,3,1),ncol=3, byrow=F)
b=c(2,1)
x=t(A)%*%inv(A%*%t(A))%*%b
x
```




And when I do some more algebra I finally get a solution to x_1, x_2, x_3 which is the following; And we had already seen that $x_3 = 1$ has to be a solution because the last equation basically said $x_3 = 1$. Now x_1 and x_2 you could have found several numbers to satisfy the first equation after you choose $x_3 = 1$ of all of these this solution says this - 0.2 - 0.4 is the minimum norm solution or this vector is the closest vector from the origin; that satisfies my equation $Ax = b$. So, I can finally, say my solution x_1, x_2, x_3 is - 0.2 - 0.41.

(Refer Slide Time: 16:14)

Data science for Engineers
Case 3: Example

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

- The solution for the given example is $(x_1, x_2, x_3) = (-0.2, -0.4, 1)$
- Verify this is a solution that satisfies the original equation
- This also turns out to be minimum norm solution



Linear Algebra

And you can easily verify that this satisfies the original equation since x_3 is 1, the second equation is $x_3 = 1$.

So, that gets satisfied when you look at the other equation you have one times - 0.2 + 2 times - 0.4. That will be - 0.2 - 0.81 + 3 times 1 will give you 3 - 1 = 2 which is this. So, the solution that we found satisfies the original equation and this also turns out to be the minimum norm solution as we discussed.

(Refer Slide Time: 16:56)

Data science for Engineers

Generalization

- The described cases cover all the scenarios one might encounter while solving linear equations
- Is there any form in which the results obtained for cases 1, 2 and 3 can be generalized ?
- The concept we used to generalize the solutions is called as Moore-Penrose pseudo-inverse of a matrix
- The pseudo inverse is used as follows

$Ax = b$ $x = A^{-1}b$

The solution becomes

$x = A^+b$

- Singular Value Decomposition can be used to calculate the pseudo inverse or the generalized inverse (A^+)

Linear Algebra 12

So, when we have a set of linear equations we basically said that there are 3 cases that one needs look at, one case is where number of equations and variables are the same $m = n$. The second case is where the number of equations are lot more than the number of variables m greater than n . And the third case was when number of equations less than number of variables m less than n . And we saw that one case is an exact solution if it is a full rank matrix.

And if it is not a full rank matrix then you could have infinite solutions or no solutions, and interestingly the next 2 cases covers these 2 aspects when I have lot more equations than variables I have a no solution case, and when I have lot more variables than equations I have infinite solution case and since we are able to solve all the 3 we should be able to use the solution to the case 2 and 3 for the case one where the rank is not full. And depending on whether it is a consistent set of equation or inconsistent set of equation you should be able to use the corresponding infinite number of solutions or no solutions result right?

So, in some sense we understand that there should be some generalization of all of these results. So, that we can write one equation which solves all of these cases square rectangular cases and so on. So, that is a question that we are asking, is there any form in which the results obtained from cases 1, 2 and 3 can be generalized. It turns out that there is a concept that we can use to generalize all of these, this is what is called the Moore Penrose pseudo inverse of a matrix.

So, when we typically have equations of the form $Ax = b$, we write $x = A^{-1}b$ as a solution. The generalization of this is to write x as A^+b where A^+ have used this term to denote the pseudo inverse of A . And as long as we can calculate the pseudo inverse in a fashion that is irrespective of the size of A , irrespective of whether the columns and rows are dependent or independent.

If I can write one general solution like this which will reduce to the cases that we discussed in this lecture, then that is a very convenient way of representing all kinds of solutions instead of looking at whether the number of rows are more, number of columns are more, is rank full and so on. All of them if they can be subsumed in one expression like this it would be very nice and it turns out that there is an expression like that and that expression is called the pseudo inverse.

Now, the pseudo inverse of A for A can be calculated using a singular value decomposition as one technique. There are many other ways of computing this, but singular value decomposition is one way of computing this. And as far as this course is concerned you just need to know that we can compute this. We do not have to really worry about how singular value decomposition is done.

(Refer Slide Time: 20:17)

Data science for Engineers

Two examples revisited

Example 2

R Code

```
A=matrix(c(1,2,3,0,0,1),ncol=2, byrow=F)
b=matrix(c(1,2,5),ncol=1, byrow=F)
library(MASS)
x=ginv(A)%*%b
```

Solution

```
> x
[1] -1
[2] 1
[3] 2
```

Example 3

R Code

```
A=matrix(c(1,0,2,0,3,1),ncol=3, byrow=F)
b=c(2,1)
library(MASS)
x=ginv(A)%*%b
```

Solution

```
> x
[1] -0.2
[2] -0.4
[3] 1.0
```

Linear Algebra

So, how do I get this in R? So the way you do this in R is you use this library and the pseudo inverse is usually calculated using this `ginv` function. Here `g` stands for generalized. So, what R does is whatever size of the problem you give here we have given 2 different examples,

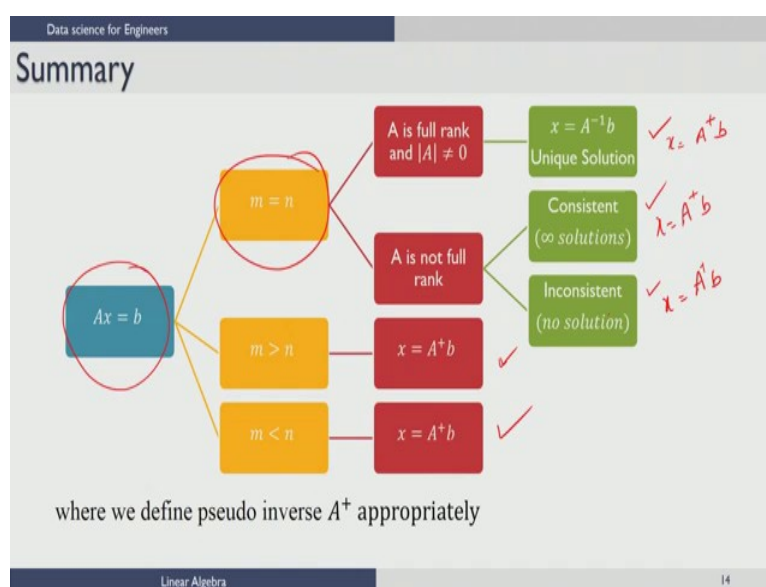
where one example has more equations than variables the second example has more variables than equation.

These are the examples that were picked from this lecture itself and we show that irrespective of whatever be the sizes of this matrices a and b , we use the same equation g inverse A and the solution $1\ 2$ that we got in one example and the solution $-0.2\ -0.4$ and one we got in the other case come out of this g inverse.

Now, the key point to understand is you simply use g inverse in R to get these solutions, but the interpretation of these solutions is what we have taught in this class. So, interpretation for this solution here is that; this is the least square solution or this is the solution that will minimize the errors collectively or this is a solution that will minimize $e_1^2 + e_2^2$ and so on.

This is what is called the minimum norm solution. While there are infinite number of solutions this is a solution that is the closest to origin. So, that is the interpretation for these 2 solutions that that we want to keep in mind as far as solving linear equations is concerned, nonetheless the operationalization for how to use R is very simple you simply use g inverse as a function.

(Refer Slide Time: 22:21)



So, let me summarize this lecture, we said we are interested in solving equations of the form $Ax = b$. We talked about 3 cases $m = n$ and $m \neq n$ if A is full rank unique solution $A^{-1}b$. If A is not full rank there are 2 possibilities either the equations are consistent or inconsistent. And if m is greater than n we look at a least square solution and if m is less than n then we look at a least norm solution.

We can write this as $A^{-1}b$ or I could also write this as pseudo inverse b . In this case the pseudo inverse and A inverse will be exactly the same and as I mentioned before since these 2 cases are covered by these 2. I should be able to use the same a pseudo inverse b for both these cases also without worrying about whether they are consistent inconsistent and so on. In all of these cases I will get a solution by using the idea of generalized inverse.

So, this concludes the section on solving linear equations irrespective of whether it is a square or a rectangular system or not, worrying about really whether the columns are dependent independent and so on. You can use generalized inverse as one unifying concept to find a solution to all these cases.

Thank you and in the next lecture we will take a geometric view of the same equations and variables that is useful in data science.