EE2211 : Introduction to Machine Learning (Fall 2020)
Solutions of Linear Equations

Vincent Y. F. Tan

In this brief document, we will summarize some key concepts involving the solution set of a system of linear equations. We will first review some basics of linear algebra. Most of the material here can be found in standard linear algebra texts such as Strang [Str16].

1 Review of Linear Algebra

Here we review some linear algebra which you should have seen before.

Definition 1. *A* vector space *over the reals consists of a set $\mathcal{V}$, a vector sum operation $+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$ and a scalar multiplication operation $\cdot : \mathbb{R} \times \mathcal{V} \to \mathcal{V}$ satisfying the following properties.*

- *Commutativity:* $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ *for all* $\mathbf{x}, \mathbf{y} \in \mathcal{V}$*;*
- *Associativity:* $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$ *for all* $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{V}$*;*
- *Identity element of addition:* $\mathbf{x} + \mathbf{0} = \mathbf{x}$ *for all* $\mathbf{x} \in \mathcal{V}$*;*
- *Inverse element of addition: For every* $\mathbf{x} \in \mathcal{V}$*, there exists an element* $-\mathbf{x} \in \mathcal{V}$ *such that* $\mathbf{x} + (-\mathbf{x}) = 0$*;*
- *Associativity of scalar multiplication: For all* $a, b \in \mathbb{R}$ *and* $\mathbf{x} \in \mathcal{V}$*,* $a(b\mathbf{x}) = (ab)\mathbf{x}$*;*
- *Identity element of scalar multiplication:* $1\mathbf{x} = \mathbf{x}$ *for* $\mathbf{x} \in \mathcal{V}$*.*
- *Distributivity of scalar multiplication w.r.t. vector addition:* $a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$ *for all* $a \in \mathbb{R}$ *and* $\mathbf{x}, \mathbf{y} \in \mathcal{V}$*;*
- *Distributivity of scalar multiplication w.r.t. addition in* $\mathbb{R}$*:* $(a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$ *for all* $a, b \in \mathbb{R}$ *and* $\mathbf{x} \in \mathcal{V}$*.*

Definition 2. *A set of vectors* $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ *from a vector space* $\mathcal{V}$ *is* linearly independent *if*

$$\beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \ldots + \beta_k\mathbf{x}_k = \mathbf{0} \tag{1}$$

implies that $\beta_1 = \beta_2 = \ldots = \beta_k = 0$*.*

Check that the condition that $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is linearly independent is equivalent to the fact that no vector $\mathbf{x}_i$ can be expressed as a linear combination of the other vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_k$.

Definition 3. *A set of vectors* $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ *is a* basis *for a vector space* $\mathcal{V}$ *if*

- $\mathcal{V} = \mathrm{span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$*;*
- $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ *is linearly independent.*

Equivalently, every $\mathbf{x} \in \mathcal{V}$ can be *uniquely* written as $\sum_{i=1}^k \beta_i\mathbf{x}_i$ for some $\{\beta_i\}_{i=1}^k \subset \mathbb{R}$. The number of vectors in any basis of $\mathcal{V}$ is called the *dimension* of $\mathcal{V}$, written as $\dim(\mathcal{V})$.

Definition 4. *The* nullspace *of a matrix* $\mathbf{A} \in \mathbb{R}^{m\times d}$ *is defined as*

$$\mathcal{N}(\mathbf{A}) := \{\mathbf{x} \in \mathbb{R}^d : \mathbf{A}\mathbf{x} = \mathbf{0}\}. \tag{2}$$

The range *or* column space *of* $\mathbf{A}$ *is defined as*

$$\mathcal{R}(\mathbf{A}) := \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathbb{R}^d\} \subset \mathbb{R}^m. \tag{3}$$

Definition 5. *The* column rank *or* rank *of a matrix* $\mathbf{A} \in \mathbb{R}^{m\times d}$ *is defined as*

$$\operatorname{rank}(\mathbf{A}) = \dim(\mathcal{R}(\mathbf{A})). \tag{4}$$

In other words, the column rank of $\mathbf{A}$ *is the dimension of the column space of* $\mathbf{A}$*.*

The *rank-nullity theorem* says that

$$\operatorname{rank}(\mathbf{A}) + \dim(\mathcal{N}(\mathbf{A})) = d. \tag{5}$$

It is always true that $\operatorname{rank}(\mathbf{A}) \le \min\{m, d\}$. We say that a matrix is *full rank* if $\operatorname{rank}(\mathbf{A}) = \min\{m, d\}$. A matrix is *full column rank* (resp. *full row rank*) if the set of columns (resp. rows) of the matrix is linearly independent. If the matrix $\mathbf{A}$ is square (i.e., $m = d$) and it is full rank, then the inverse $\mathbf{A}^{-1}$ exists.

2 Nature of Solutions to Linear Systems

Often in engineering, we would like to "solve" systems of equations of the form

$$\mathbf{X}\mathbf{w} = \mathbf{y} \qquad \text{or} \qquad \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_d \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \tag{6}$$

where $\mathbf{X} \in \mathbb{R}^{m\times d}$ and $\mathbf{y} \in \mathbb{R}^m$ are given and $\mathbf{w} \in \mathbb{R}^d$ is to be found. As mentioned, if the matrix $\mathbf{X}$ is square and full rank, $\mathbf{X}^{-1}$ exists and so we can solve for $\mathbf{w}$ by simple matrix inversion and multiplication $\mathbf{w} = \mathbf{X}^{-1}\mathbf{y}$. However, most of the time in engineering, $m \ne d$ and more care is needed to discuss the existence and uniqueness of solutions to the linear system in (6). For this, we appeal to the Rouché-Capelli Theorem. We need the notion of the *augmented matrix*

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} & \mathbf{y} \end{bmatrix} \in \mathbb{R}^{m\times(d+1)}. \tag{7}$$

Note that the augmented matrix $\tilde{\mathbf{X}}$ has rank at least as large as that of $\mathbf{X}$, i.e., $\operatorname{rank}(\mathbf{X}) \le \operatorname{rank}(\tilde{\mathbf{X}})$. This is because $\tilde{\mathbf{X}}$ has more columns than $\mathbf{X}$ so the dimension of its column space must be as large as that of $\mathbf{X}$.

Theorem 1 (Rouché-Capelli Theorem)**.** *For the linear system in* (6)*, the following hold:*

(i) The system in (6) *admits a* <u>*unique*</u> *solution if and only if* $\operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\tilde{\mathbf{X}}) = d$*;*

(ii) The system in (6) *has* <u>*no solution*</u> *if and only if* $\operatorname{rank}(\mathbf{X}) < \operatorname{rank}(\tilde{\mathbf{X}})$*;*

(iii) The system in (6) *has* <u>*infinitely many*</u> *solutions if and only if* $\operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\tilde{\mathbf{X}}) < d$*.*

Proof sketch (Only the $\Longleftarrow$ *directions).* For part (i), we note that the condition that $\operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\tilde{\mathbf{X}})$ means that $\mathbf{y}$ is in the column space of $\mathbf{X}$. This means that there exists $\{w_i\}_{i=1}^d \subset \mathbb{R}$ such that $\sum_{i=1}^d w_i\mathbf{x}_i = \mathbf{y}$ where the $\mathbf{x}_i$'s are the d columns of $\mathbf{X}$. Since $\operatorname{rank}(\mathbf{X}) = d$, $\{\mathbf{x}_i\}_{i=1}^d$ span $\mathbb{R}^d$ and the representation $\sum_{i=1}^d w_i\mathbf{x}_i = \mathbf{y}$ is unique (see discussion after Definition 3), which means there is a unique solution.

For part (ii), the condition that $\operatorname{rank}(\mathbf{X}) < \operatorname{rank}(\tilde{\mathbf{X}})$ means that $\mathbf{y}$ is not in the column space of $\mathbf{X}$ so there is no solution.

For part (iii), since $\operatorname{rank}(\mathbf{X}) < d$, $\{\mathbf{x}_i\}_{i=1}^d$ do not span $\mathbb{R}^d$ and the dimension of the nullspace of $\mathbf{X}$ is non-zero. This means that if $\mathbf{w}_{\mathrm{p}}$ is a particular solution so is $\mathbf{w}_{\mathrm{p}} + \mathbf{w}_0$ where $\mathbf{w}_0 \in \mathcal{N}(\mathbf{X})$. Hence, are infinitely many solutions. $\square$

Let us consider a few examples.

- Consider the following over-determined system in which $m = 3$ and $d = 2$:

$$\mathbf{X} = \begin{bmatrix} 2 & 1 \\ 4 & 3 \\ 5 & 6 \end{bmatrix}, \qquad \text{and} \qquad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}. \tag{8}$$

  In this case $\text{rank}(\mathbf{X}) = 2$ and $\text{rank}(\tilde{\mathbf{X}}) = 3$. This is case (ii) of the Rouché-Capelli Theorem and there is no solution. This is the usual case for over-determined systems.

- Consider the following over-determined system in which $m = 3$ and $d = 2$:

$$\mathbf{X} = \begin{bmatrix} 2 & 1 \\ 4 & 3 \\ 5 & 6 \end{bmatrix}, \qquad \text{and} \qquad \mathbf{y} = \begin{bmatrix} 4 \\ 10 \\ 17 \end{bmatrix}. \tag{9}$$

  In this case $\text{rank}(\mathbf{X}) = 2$ and $\text{rank}(\tilde{\mathbf{X}}) = 2$. This is case (i) of the Rouché-Capelli Theorem and there is a unique solution even though the system is over-determined. Note that $\mathbf{y}$ is one times the first column of $\mathbf{X}$ plus two times the second column of $\mathbf{X}$, so it is in the linear span of the columns of $\mathbf{X}$.

- Consider the following over-determined system in which $m = 3$ and $d = 2$:

$$\mathbf{X} = \begin{bmatrix} 2 & 1 \\ 4 & 2 \\ 6 & 3 \end{bmatrix}, \qquad \text{and} \qquad \mathbf{y} = \begin{bmatrix} 8 \\ 16 \\ 24 \end{bmatrix}. \tag{10}$$

  In this case $\text{rank}(\mathbf{X}) = 1$ and $\text{rank}(\tilde{\mathbf{X}}) = 1$ and both these ranks are $< d = 2$. This is case (iii) of the Rouché-Capelli Theorem and there are infinitely many solution even though the system is over-determined. Note that the three columns of $\tilde{\mathbf{X}}$ are collinear.

- Consider the following under-determined system in which $m = 2$ and $d = 3$:

$$\mathbf{X} = \begin{bmatrix} 2 & 1 & 3 \\ 4 & 2 & 5 \end{bmatrix}, \qquad \text{and} \qquad \mathbf{y} = \begin{bmatrix} 10 \\ 7 \end{bmatrix}. \tag{11}$$

  In this case, $\text{rank}(\mathbf{X}) = 2$ and $\text{rank}(\tilde{\mathbf{X}}) = 2$ but $d = 3$. This is case (iii) of the Rouché-Capelli Theorem and there are infinitely many solutions. This is the usual case for under-determined systems.

- Consider the following under-determined system in which $m = 2$ and $d = 3$:

$$\mathbf{X} = \begin{bmatrix} 2 & 1 & 3 \\ 4 & 2 & 6 \end{bmatrix}, \qquad \text{and} \qquad \mathbf{y} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}. \tag{12}$$

  In this case, $\text{rank}(\mathbf{X}) = 1$ and $\text{rank}(\tilde{\mathbf{X}}) = 2$ because $\mathbf{y} \notin \mathcal{R}(\mathbf{X})$. This is case (ii) of the Rouché-Capelli Theorem and there is no solution. Note that $\mathbf{y}$ boosts the rank of $\mathbf{X}$ by 1 in the augmented matrix $\tilde{\mathbf{X}}$, i.e., $\mathbf{y}$ is not in the column space of $\mathbf{X}$, which is the ray $\{[t, 2t]^\top : t \in \mathbb{R}\}$.

- For under-determined systems ($m < d$), can we have case (i)?

3 Least Squares Estimation for $m > d$

We consider the case in which $\mathbf{X}$ is tall ($m > d$) and full rank. This means that $\text{rank}(\mathbf{X}) = d$; equivalently, all columns are linearly independent. This *over-determined* scenario happens a lot in engineering. For example, this happens in estimation problems, where one tries to estimate a small number d of parameters given a

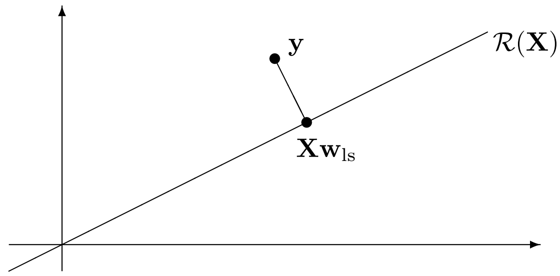

Figure 1: The least squares problem consists in finding $\mathbf{w}_{\mathrm{ls}}$, the point such that $\mathbf{X}\mathbf{w}_{\mathrm{ls}} \in \mathcal{R}(\mathbf{X})$ is closest to a given point $\mathbf{y}$.

lot of (noisy) experimental measurements, say $m > d$. As seen from the usual case of the Rouché-Capelli Theorem (case (ii) in which $\mathbf{y}$ is not in the linear span of the columns of $\mathbf{X}$), there is no solution. Hence, one way to find "the best" solution is to minimize the sum of squares of the errors

$$\text{minimize} \quad \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2. \tag{13}$$

The optimal $\mathbf{w}$ is the *least squares estimate* (why can we use the article "the" here?). We can solve this problem by means of calculus (a more elegant way is through the projection theorem). Let $f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$. Then

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \nabla_{\mathbf{w}} \left(\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - 2\mathbf{y}^\top \mathbf{X}\mathbf{w} + \mathbf{y}^\top \mathbf{y} \right) = 2\mathbf{X}^\top \mathbf{X} - 2\mathbf{X}^\top \mathbf{y}. \tag{14}$$

Setting this to zero, we see that the optimal $\mathbf{w}$ is the least squares estimate

$$\mathbf{w}_{\mathrm{ls}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{15}$$

The existence of $(\mathbf{X}^\top \mathbf{X})^{-1}$ is guaranteed by the fact that $\mathbf{X}$ has full column rank. See the geometry of the problem in Fig. 2.

A few words about the matrix $\mathbf{X}^\dagger := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$. This matrix is called the *pseudo-inverse* of the full rank tall matrix $\mathbf{X}$. It is also called the *left-inverse* of $\mathbf{X}$ because if we multiply $\mathbf{X}$ on the left with $\mathbf{X}^\dagger$, we obtain $\mathbf{X}^\dagger \mathbf{X} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} = \mathbf{I}$.

4 Least Norm Solution for $m < d$

Now we consider the case in which $\mathbf{X}$ is wide ($m < d$) and full rank. This means that $\mathrm{rank}(\mathbf{X}) = m$; equivalently, all rows are linearly independent (full row rank). This *under-determined* situation also occurs a lot in engineering.

Example 1. *For example, in control engineering (a field of study within mechanical and electrical engineering), one often considers the following discrete-time state-space system (e.g., describing the dynamics of a robot operating over a quantized time interval):*

$$v_{i+1} = a v_i + b w_i, \quad i = 0, 1, \ldots, d-1, \tag{16}$$

where v_i is the state *of the system at time i and w_i is our* control. *We assume the system starts at the origin $v_0 = 0$. We desire to design the w_i's such that the terminal state $v_d = y$ (for some given y) while minimizing the* cost *of the control $\sum_{i=0}^{d-1} w_i^2$. After some algebra, this can be rewritten as*

$$\textit{minimize} \quad \|\mathbf{w}\|^2 \quad \textit{subject to} \quad y = \begin{bmatrix} b & ab & a^2 b & \cdots & a^{d-1} b \end{bmatrix} \begin{bmatrix} w_{d-1} \\ w_{d-2} \\ \vdots \\ w_0 \end{bmatrix}. \tag{17}$$

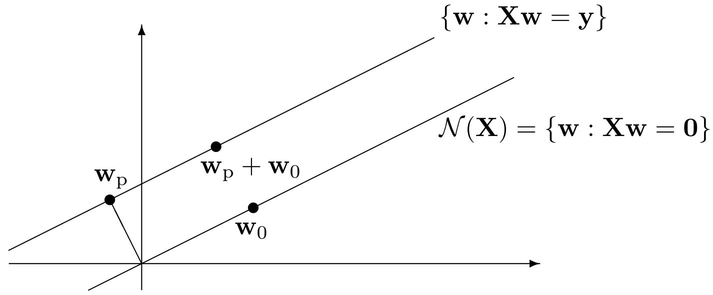

Figure 2: The least norm problem consists of finding the particular solution $\mathbf{w}_{\mathrm{p}}$ that minimizes the norm. All other solutions $\mathbf{w}_{\mathrm{p}} + \mathbf{w}_0$ where $\mathbf{w}_0 \in \mathcal{N}(\mathbf{X})$ have larger norms.

This is exactly an under-determined problem if we make the identifications $\mathbf{X} = \begin{bmatrix} b & ab & a^2b & \cdots & a^{d-1}b \end{bmatrix}$ *(for obvious reasons, this is called the* d-step reachability matrix*) and* $\mathbf{w} = \begin{bmatrix} w_{d-1} & w_{d-2} & \cdots & w_0 \end{bmatrix}^\top$ *and* y *is scalar (i.e.,* $m = 1$*).*

We return to the equation $\mathbf{X}\mathbf{w} = \mathbf{y}$ in which $m < d$ and the matrix $\mathbf{X}$ has full row rank. It is clear that $(\mathbf{X}\mathbf{X}^\top)^{-1}$ exists and

$$\mathbf{w}_{\mathrm{p}} = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} \tag{18}$$

is a solution to the equation (check!). The subscript p indicates that this is a *particular* solution to the linear system. From the usual case of the Rouché-Capelli Theorem (case (iii)), we know that there are infinitely many solutions. Where are these infinitely many solutions? Let $\mathbf{w}_0 \in \mathcal{N}(\mathbf{X})$ be any vector in the nullspace of $\mathbf{X}$. Note that the nullspace has positive dimension because $\mathrm{rank}(\mathbf{X}) = m < d$ so $\mathbf{w}_0$ can be chosen to be a non-zero vector. Then $\mathbf{w}_{\mathrm{p}} + \mathbf{w}_0$ is also a solution to (6) (check!). In the following, we argue that among all the solutions, $\mathbf{w}_{\mathrm{p}}$ has a special place in our hearts because is the *least norm solution* to (6).

Suppose that $\mathbf{w}$ is any solution to $\mathbf{X}\mathbf{w} = \mathbf{y}$. Then since $\mathbf{w}_{\mathrm{p}}$ is also a solution, we have $\mathbf{X}(\mathbf{w} - \mathbf{w}_{\mathrm{p}}) = \mathbf{0}$ and

$$(\mathbf{w} - \mathbf{w}_{\mathrm{p}})^\top \mathbf{w}_{\mathrm{p}} \overset{(18)}{=} (\mathbf{w} - \mathbf{w}_{\mathrm{p}})^\top \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} = \big(\mathbf{X}(\mathbf{w} - \mathbf{w}_{\mathrm{p}})\big)^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} = \mathbf{0}. \tag{19}$$

This means that $\mathbf{w} - \mathbf{w}_{\mathrm{p}}$ is orthogonal to $\mathbf{w}_{\mathrm{p}}$. By the Pythagorean theorem,

$$\|\mathbf{w}\|^2 = \|(\mathbf{w} - \mathbf{w}_{\mathrm{p}}) + \mathbf{w}_{\mathrm{p}}\|^2 = \|\mathbf{w} - \mathbf{w}_{\mathrm{p}}\|^2 + \|\mathbf{w}_{\mathrm{p}}\|^2 \geq \|\mathbf{w}_{\mathrm{p}}\|^2, \tag{20}$$

which shows that $\mathbf{w}_{\mathrm{p}}$ is the least norm solution to $\mathbf{X}\mathbf{w} = \mathbf{y}$.

Finally, we say a few words about the matrix $\mathbf{X}^\dagger = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1}$. This matrix is called the *pseudo-inverse* of the full-rank wide matrix $\mathbf{X}$. It is also know as the *right-inverse* of $\mathbf{X}$ (why?).

References

[Str16] G. Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 5th edition, 2016.