EE3001 Machine Learning **Fall 2022**

Lecure A2. Basics of Linear Algebra

Lecturer: Jie Wang Date: Sep 15, 2022

Linear algebra is central to almost all areas of mathematics and is also powerful in most sciences and fields of engineering, including machine learning. In this lecture, we review some of the basics of linear algebra. The major reference of this lecture is [1].

1 Linear Space

We start with linear space, which is one of the most important concepts in linear algebra. A critical property of linear spaces is that *each vector can be linearly represented by a finite number of other vectors*. To understand this statement, we need to answer the following questions.

1. What is a linear space?
2. What is the linear combination?
3. What is the basis of a linear space?
4. Do all linear spaces have a basis? Are their bases always finite or countable?

We can find the answers to the first three questions in this lecture. However, to answer the last question, we need **Axiom of Choice** or **Zorn's Lemma**, which are beyond our scope. In this course, we **admit** that every linear space has a set of basis and **know** that some linear spaces have non-countable basis.

Definition 1. Let $\mathcal{V}$ be a nonempty set and F be a number field (e.g., $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$). We say that $\mathcal{V}$ is the **linear space** (or **vector space**) **over** F if the following conditions hold.

1. We have defined two binary operations in $\mathcal{V}$.
   (a) The first operation, called **vector addition** or simply **addition**, assigns to any two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathcal{V}$ a third vector in $\mathcal{V}$ which is commonly written as $\mathbf{u}+\mathbf{v}$, and called the **sum** of these two vectors.
   (b) The second operation, called **scalar multiplication**, assigns to any scalar $a \in F$ and any vector $\mathbf{v} \in \mathcal{V}$ another vector in $\mathcal{V}$, which is denoted $a\mathbf{v}$.
2. The addition and the scalar multiplication defined in $\mathcal{V}$ satisfy the following eight axioms.
   (a) The addition is **commutative**, i.e., $\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}$, $\forall\,\mathbf{u},\mathbf{v}\in\mathcal{V}$.
   (b) The addition is **associative**, i.e., $(\mathbf{u}+\mathbf{v})+\mathbf{w}=\mathbf{u}+(\mathbf{v}+\mathbf{w})$, $\forall\,\mathbf{u},\mathbf{v},\mathbf{w}\in\mathcal{V}$.
   (c) There exists a **zero vector** in $\mathcal{V}$, i.e., there exists $\theta\in\mathcal{V}$ such that $\theta+\mathbf{v}=\mathbf{v}+\theta=\mathbf{v}$, $\forall\,\mathbf{v}\in\mathcal{V}$. The zero vector is also denoted by $\mathbf{0}$.
   (d) There exists an **additive inverse** for each vector in $\mathcal{V}$, i.e., for every $\mathbf{v}\in\mathcal{V}$, there exists a vector $\mathbf{v}'\in\mathcal{V}$ such that $\mathbf{v}+\mathbf{v}'=\mathbf{v}'+\mathbf{v}=\mathbf{0}$. We also denote $\mathbf{v}'$ by $-\mathbf{v}$.
   (e) The scalar multiplication is **compatible** with field multiplication, i.e., $(ab)\mathbf{v}=a(b\mathbf{v})$, $\forall\,\mathbf{v}\in\mathcal{V}$, $a,b\in F$.
   (f) The **multiplicative identity** in F is the **identity element** of scalar multiplication, i.e., $\mathbf{1v}=\mathbf{v}$, $\forall\,\mathbf{v}\in\mathcal{V}$.

(g) The scalar multiplication is **distributive** with respect to vector addition, i.e., $a(\mathbf{u}+\mathbf{v}) = a\mathbf{u} + a\mathbf{v}$, $\forall\, a \in F$, $\mathbf{u}, \mathbf{v} \in \mathcal{V}$.

(h) The scalar multiplication is **distributive** with respect to field addition, i.e., $(a+b)\mathbf{v} = a\mathbf{v} + b\mathbf{v}$, $\forall\, a, b \in F$, $\mathbf{v} \in \mathcal{V}$.

Remark 1. When we talk about a linear space, compared to what the elements in it (i.e., vectors) are, we care more about the operations (i.e., the vector addition and scalar multiplication) defined on it and its linear structure.

Example 1.

1. $\mathbb{R}[x]$ is the linear space consisting of all the polynomials with real coefficients.
2. $C[a, b]$ is the linear space consisting of all the continuous functions on $[a, b]$.

Definition 2. Let $\mathcal{V}$ be a linear space over F and S be a subset of $\mathcal{V}$. For any finite subset $S_1 = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subset S$ and $a_1, \ldots, a_k \in F$, we call $a_1\mathbf{v}_1 + \ldots + a_k\mathbf{v}_k$ a **linear combination** of S. If a vector $\mathbf{u} \in \mathcal{V}$ is a linear combination of S, we say that $\mathbf{u}$ can be **linearly represented** by S. The set of all linear combinations of S is denoted $V(S)$.

Note that for every subset $S \subset \mathcal{V}$, $V(S)$ is a subspace of $\mathcal{V}$. It is easy to show that any subspace $\mathcal{W}$ that contains S also contains $V(S)$. Hence, $V(S)$ is the smallest subspace of $\mathcal{V}$ containing S.

Definition 3. We say that S **spans** or **generates** $V(S)$, $V(S)$ is the **linear span** of S, and S is a **spanning set** or a **generating set** of $V(S)$.

Definition 4. Let $\mathcal{V}$ be a linear space over F and S be a subset of $\mathcal{V}$. If there exists some difnite subset $S_1 = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subset S$ and scalars $a_1, \ldots, a_k$ that are not all 0 such that

$$a_1\mathbf{v}_1 + \ldots + a_k\mathbf{v}_k = \mathbf{0},$$

we say that S is **linearly dependent**. Otherwise, S is **linearly independent**.

Remark 2. Please note the geometric meaning of linear dependencies.

Theorem 1. *Let $\mathcal{V}$ be a linear space over F and S be a subset of $\mathcal{V}$, then S is linearly dependent if and only if there exists some vector $\mathbf{v} \in S$, which is the linear combination of other vectors.*

Definition 5. Let $\mathcal{V}$ be a linear space over F, S be a subset of $\mathcal{V}$, and M be a subset of S. If M is linear independent and $M_{\mathbf{v}} = M \cup \{\mathbf{v}\}$ is linear dependent for any $\mathbf{v} \in S$, we say that M is a **maximal linearly independent system** of S. If one of the maximal linearly independent systems of S, $M = \{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ is finite, the **rank** of S is r, denoted $\operatorname{rank} S = r$.

Remark 3. If every maximal linearly independent system of S is infinite, it is natural to define $\operatorname{rank} S = \infty$. However, we only consider the case where $\operatorname{rank} S < \infty$ in this course.

Definition 6. Let $\mathcal{V}$ be a linear space over F.

1. If $\mathcal{V}$ is the linear span of some finite subset $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, i.e., $\mathcal{V} = V(\{\mathbf{v}_1, \ldots, \mathbf{v}_n\})$, we say that $\mathcal{V}$ is a **finite-dimensional** linear space. Obviously, we have $\operatorname{rank} V \leq n$ in this case.
2. If there exist n vectors in $\mathcal{V}$ that are linearly independent and every $n+1$ vectors in $\mathcal{V}$ are linearly dependent, we say that the **dimension** of $\mathcal{V}$ is n, denoted $\dim \mathcal{V} = n$.

3. If there exists a set of vectors $M = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ such that every vector $v \in \mathcal{V}$ is the linear combination of M, i.e.,

$$\mathbf{v} = a_1\mathbf{v}_1 + \ldots + a_n\mathbf{v}_n,$$

and the coefficients are uniquely determined by $\mathbf{v}$, we say that M is a **basis** of $\mathcal{V}$. The ordered array $(a_1, \ldots, a_n)$ is called the **coordinate** of $\mathbf{v}$ under the basis M.

2 Range and Nullspace

We define the range and the nullsapce of $\mathbf{A} \in \mathbb{R}^{m\times n}$ as follows.

Definition 7. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$. The **range** of $\mathbf{A}$, denoted $\mathcal{R}(\mathbf{A})$, is the set of all vectors in $\mathbb{R}^m$ that can be written as linear combinations of the columns of $\mathbf{A}$, i.e.,

$$\mathcal{R}(\mathbf{A}) = \{\mathbf{A}\mathbf{x} \in \mathbb{R}^m \mid \mathbf{x} \in \mathbb{R}^n\}.$$

The **nullsapce** (or **kernel**) of $\mathbf{A}$, denoted $\mathcal{N}(\mathbf{A})$, is the set of all vectors $\mathbf{x}$ mapped into $\mathbf{0}$ by $\mathbf{A}$, i.e.,

$$\mathcal{N}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{0}\}.$$

Theorem 2. *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$.

1. $\mathcal{R}(\mathbf{A})$ *is a subspace of* $\mathbb{R}^m$ *and* $\mathcal{N}(\mathbf{A})$ *is a subspace of* $\mathbb{R}^n$.
2. *The dimension of* $\mathcal{R}(\mathbf{A})$ *is the rank of* $\mathbf{A}$, *i.e.,*

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

3. *The dimension of* $\mathcal{N}(\mathbf{A})$ *is that of the solution space of* $\mathbf{A}\mathbf{x} = \mathbf{0}$, *i.e.,*

$$\dim \mathcal{N}(\mathbf{A}) = \dim V_{\mathbf{A}}.$$

In fact, we have $\mathcal{N}(\mathbf{A}) = V_{\mathbf{A}}$.

4. *The sum of dimensions of* $\mathcal{R}(\mathbf{A})$ *and* $\mathcal{N}(\mathbf{A})$ *is the dimension of* $\mathbb{R}^n$, *i.e.,*

$$\dim \mathcal{R}(\mathbf{A}) + \dim \mathcal{N}(\mathbf{A}) = n,$$

which is equivalent to what we know about $\mathbf{A}$*:*

$$\operatorname{rank} \mathbf{A} + \dim V_{\mathbf{A}} = n.$$

Question 1. How many ways can you think of to explain the rank of a matrix?

2.1 Orthogonal Decomposition induced by A

For a linear space $\mathcal{V}$, there are some linear spaces where all vectors are orthogonal to $\mathcal{V}$. For example, in $\mathbb{R}^3$, the z-axis is orthogonal (or perpendicular) to the xy-plane.

Definition 8. Let $\mathcal{V}$ be a subspace of $\mathbb{R}^n$, its **orthogonal complement**, denoted $\mathcal{V}^\perp$, is defined as

$$\mathcal{V}^\perp = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{z}^\top \mathbf{x} = 0 \text{ for } \forall\, \mathbf{z} \in \mathcal{V}\}.$$

Theorem 3. *Let $\mathcal{V}$ be a subspace of $\mathbb{R}^n$. The orthogonal complement of the orthogonal complement of $\mathcal{V}$ is $\mathcal{V}$ itself, i.e.,*

$$(\mathcal{V}^\perp)^\perp = \mathcal{V},$$

which matches our expectations for a complement.

Theorem 4. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$.*

1. *The null space of $\mathbf{A}$ is the orthogonal complement of the range of $\mathbf{A}^\top$, and the range of $\mathbf{A}$ is the complement of the nullspace of $\mathbf{A}^\top$, i.e.,*

$$\mathcal{N}(\mathbf{A}) = \mathcal{R}(\mathbf{A}^\top)^\perp,$$
$$\mathcal{R}(\mathbf{A}) = \mathcal{N}(\mathbf{A}^\top)^\perp.$$

2. *We can decompose $\mathbb{R}^n$ as*

$$\mathbb{R}^n = \mathcal{N}(\mathbf{A}) \overset{\perp}{\oplus} \mathcal{R}(\mathbf{A}^\top),$$

*where the symbol $\overset{\perp}{\oplus}$ refers to **orthogonal direct sum**, i.e., the sum of two subspaces that are orthogonal. Such a decomposition of $\mathbb{R}^n$ is called the **orthogonal decomposition induced by** $\mathbf{A}$.*

3 Symmetric Eigenvalue Decomposition

Theorem 5. *Let $\mathbf{A} \in S^n$, i.e., $\mathbf{A}$ is a real symmetric $n \times n$ matrix. Then $\mathbf{A}$ can be factored as*

$$\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top,$$

*where $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is orthogonal and $\boldsymbol{\Lambda} = \mathbf{diag}\,(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$. Such a decomposition is called the **symmetric eigenvalue decomposition** or **spectral decomposition** of $\mathbf{A}$.*

Suppose that $\mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_n)$, then we have

$$\mathbf{A}(\mathbf{q}_1, \ldots, \mathbf{q}_n) = (\mathbf{q}_1, \ldots, \mathbf{q}_n) \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix},$$

which leads to $\mathbf{A}\mathbf{q}_i = \lambda_i \mathbf{q}_i$, $1 \leq i \leq n$. Note that $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$, which means that $\mathbf{q}_i^\top \mathbf{q}_j = 0$, $1 \leq i \neq j \leq n$. Hence $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ is an **orthonormal set** of eigenvectors of $\mathbf{A}$.

We order the eigenvalues as $\lambda_{\max} := \lambda_1 \geq \cdots \geq \lambda_n =: \lambda_{\min}$. Then the determinant, trace, and Frobenius norm can be expressed in terms of eigenvalues, we have

$$\det \mathbf{A} = \prod_{i=1}^{n} \lambda_i,$$
$$\operatorname{tr} \mathbf{A} = \sum_{i=1}^{n} \lambda_i,$$
$$\|\mathbf{A}\|_F := \left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{n} \lambda_i^2\right)^{\frac{1}{2}}.$$

The norms $\|\cdot\|_*$ defined in $\mathbb{R}^m$ and $\mathbb{R}^n$ induce a norm $\|\cdot\|_*$ defined in $\mathbb{R}^{m\times n}$ as

$$\|\mathbf{A}\|_* := \sup_{\mathbf{x}\in\mathbb{R}^n} \frac{\|\mathbf{A}\mathbf{x}\|_*}{\|\mathbf{x}\|_*} = \sup_{\mathbf{x}\in\mathbb{R}^n, \|\mathbf{x}\|_*=1} \|\mathbf{A}\mathbf{x}\|_*.$$

Given the definition of ***p*-norm** of vectors in $\mathbb{R}^n$ and $\mathbb{R}^m$, i.e.,

$$\|\mathbf{x}\|_p = \begin{cases} \left(\sum_{i=1}^n x_i^p\right)^{\frac{1}{p}}, & 1 \leq p < \infty, \\ \max_{1\leq i\leq n} |x_i|, & p = \infty, \end{cases} \quad \text{(here we only give the definition for } \mathbb{R}^n\text{)}$$

we can define the p-norm of matrices by

$$\|\mathbf{A}\|_p := \sup_{\mathbf{x}\in\mathbb{R}^n} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}.$$

We can show that

$$\|\mathbf{A}\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |x_{ij}|,$$
$$\|\mathbf{A}\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |x_{ij}|.$$

Question 2. How to express the 2-norm (also called **spectral norm**) $\|\mathbf{A}\|_2$ of a square matrix $\mathbf{A}$ in terms of eigenvalues? What if $\mathbf{A}$ is not square?

3.1 Definiteness and Matrix Inequalities

First we have a simple observation. Suppose $\mathbf{A} \in S^n$, we let $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ be the maximum and minimum eigenvalue of $\mathbf{A}$, respectively. Then we have

$$\lambda_{\max}(\mathbf{A}) = \sup_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq 0} \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}, \quad \lambda_{\min}(\mathbf{A}) = \inf_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq 0} \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{x}}.$$

To see this, let $\mathbf{A} = \mathbf{Q}^\top\mathbf{\Lambda}\mathbf{Q}$ be the orthogonal decomposition of $\mathbf{A}$, where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{\Lambda} = \mathbf{diag}\,(\lambda_1, \cdots, \lambda_n)$ is a diagonal matrix whose diagonal entries are the eigenvalues of $\mathbf{A}$. Then we have

$$\mathbf{x}^\top\mathbf{A}\mathbf{x} = \mathbf{x}^\top\mathbf{Q}^\top\mathbf{\Lambda}\mathbf{Q}\mathbf{x} \leq \lambda_{\max}(\mathbf{A})\mathbf{x}^\top\mathbf{Q}^\top\mathbf{Q}\mathbf{x} = \lambda_{\max}(\mathbf{A})\mathbf{x}^\top\mathbf{x}.$$

The equality holds if and only if $\mathbf{x}$ is an eigenvector of $\lambda_{\max}(\mathbf{A})$. The equation of $\lambda_{\min}(\mathbf{A})$ holds similarly, and we leave it for an exercise.

In the following context, we are going to investigate how the sign of $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ influence the properties of $\mathbf{A}$. First, we introduce the concept of definiteness.

Definition 9. Let $\mathbf{A} \in S^n$.

1. We say $\mathbf{A}$ is **positive definite** or $\mathbf{A} > 0$, if $\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$. We denote the set of all positive definite matrices by S^n_{++}.
2. We say $\mathbf{A}$ is **positive semidefinite** or $\mathbf{A} \geq 0$, if $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$. We denote the set of all positive definite matrices by S^n_{+}.
3. We say $\mathbf{A}$ is **negative definite** or $\mathbf{A} < 0$, if $\mathbf{x}^\top \mathbf{A}\mathbf{x} < 0$ for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$.
4. We say $\mathbf{A}$ is **negative semidefinite** or $\mathbf{A} \leq 0$, if $\mathbf{x}^\top \mathbf{A}\mathbf{x} \leq 0$ for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$.

Notice that by definition, $\mathbf{A}$ is positive definite is equivalent to $\lambda_{\min}(\mathbf{A}) > 0$; $\mathbf{A}$ is positive semidefinite is equivalent to $\lambda_{\min}(\mathbf{A}) \geq 0$. For negative and negative semidefinite cases, we also have similar results. Another observation is that $\mathbf{A} < 0$ is equivalent to $-\mathbf{A} > 0$, $\mathbf{A} \leq 0$ is equivalent to $-\mathbf{A} \geq 0$.

3.2 Symmetric Squareroot

Suppose $\mathbf{A} \in S^n_+$, $\mathbf{A} = \mathbf{Q}^\top \mathbf{diag}\,(\lambda_1, \cdots, \lambda_n)\mathbf{Q}$ is its orthogonal decomposition, then $\lambda_i \geq 0$ $(i = 1, 2, \cdots, n)$ and we can define the squareroot of $\mathbf{A}$,

$$\mathbf{A}^{\frac{1}{2}} := \mathbf{Q}^\top \mathbf{diag}\,(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_n})\mathbf{Q}.$$

Notice that $\mathbf{A}^{\frac{1}{2}}$ behaves like the arithmetic squareroot of $\mathbf{A}$, where $\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{A}$ and $\mathbf{A}^{\frac{1}{2}} \geq 0$.

4 Singular Value Decomposition (SVD)

Singular value decomposition separates any matrix into simple pieces and is widely used in numerical linear algebra field.

Theorem 6. *Suppose* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *with* $\operatorname{rank}(\mathbf{A}) = r$, *then* $\mathbf{A}$ *can be factorized as*

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top.$$

Here $\mathbf{U} \in \mathbb{R}^{m\times m}$ *and* $\mathbf{V} \in \mathbb{R}^{n\times n}$ *are both orthogonal matrices,* $\mathbf{\Sigma} = \mathbf{diag}\,(\sigma_1, \cdots, \sigma_r, 0, \cdots, 0)$ *with* $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_r > 0$.

We call σ_i $(i = 1, \cdots, r)$ *the singular values of* $\mathbf{A}$, *the columns of* $\mathbf{U}$ *left singular vectors and the columns of* $\mathbf{V}$ *right singular vectors. Then we have*

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top,$$

where $\mathbf{u}_i$ *is the ith column of* $\mathbf{U}$ $(i = 1, \cdots, m)$ *and* $\mathbf{v}_j$ *is the jth column of* $\mathbf{V}$ $(j = 1, \cdots, n)$.

Further, one can show that the set of singular values of $\mathbf{A}$ is equal to the set of the arithmetic square root of non-zero eigenvalues of $\mathbf{A}^\top\mathbf{A}$ or $\mathbf{A}\mathbf{A}^\top$.

We denote the largest singular value of $\mathbf{A}$ by $\sigma_{\max}(\mathbf{A})$. Then we can show that

$$\sigma_{\max}(\mathbf{A}) = \sup_{\mathbf{x}\in\mathbb{R}^m, \mathbf{y}\in\mathbb{R}^n, \mathbf{x},\mathbf{y}\neq 0} \frac{\mathbf{x}^\top \mathbf{A}\mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} = \sup_{\mathbf{y}\in\mathbb{R}^n, \mathbf{y}\neq 0} \frac{\|\mathbf{A}\mathbf{y}\|_2}{\|\mathbf{y}\|_2} = \|\mathbf{A}\|_2.$$

5 Schur Complement

Schur's formula derives from elimination method, and can help us calculate the inverse of block matrices. First we consider a matrix $\mathbf{X} \in S^n$ partitioned as

$$\mathbf{X} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{pmatrix},$$

where $\mathbf{A} \in S^m$ and $m < n$. Suppose that $\det(\mathbf{X}), \det(\mathbf{A}) \neq 0$, we have

$$\begin{pmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{B}^\top\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{pmatrix} \begin{pmatrix} \mathbf{A} & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{C} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B} \end{pmatrix}.$$

This formula is called Schur's formula, and we define the **Schur complement** of $\mathbf{A}$ in $\mathbf{X}$ as $\mathbf{S} = \mathbf{C} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{B}$. We take determinant in both sides of Schur's formula and attain

$$\det(\mathbf{X}) = \det(\mathbf{A})\det(\mathbf{S}).$$

In the following context in this section, we utilize the Schur complement to calculate the inverse of the block matrix $\mathbf{X}$. Let $\mathbf{x}, \mathbf{u} \in \mathbb{R}^n$ and $\mathbf{y}, \mathbf{v} \in \mathbb{R}^{n-m}$. The Schur complement comes up in solving linear equations

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}.$$

Solving this equation we get

$$\begin{cases} \mathbf{B}^\top\mathbf{A}^{-1}(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}) = \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{u} \\ \mathbf{B}^\top\mathbf{x} + \mathbf{C}\mathbf{y} = \mathbf{v}, \end{cases}$$

then we have

$$\begin{cases} \mathbf{x} = (\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1}\mathbf{B}^\top\mathbf{A}^{-1})\mathbf{u} - \mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1}\mathbf{v} \\ \mathbf{y} = \mathbf{S}^{-1}(\mathbf{v} - \mathbf{B}^\top\mathbf{A}^{-1}\mathbf{u}). \end{cases}$$

We first let $\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{I} \\ \mathbf{O} \end{pmatrix}$ and then $\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \mathbf{O} \\ \mathbf{I} \end{pmatrix}$ in the the above solution, attaining $\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{y}_1 \end{pmatrix}$ and $\begin{pmatrix} \mathbf{x}_2 \\ \mathbf{y}_2 \end{pmatrix}$ respectively. Combining the two equations and we have

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 \\ \mathbf{y}_1 & \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{I} \end{pmatrix}.$$

This is sufficient to show that

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1}\mathbf{B}^\top\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{S}^{-1} \\ -\mathbf{S}^{-1}\mathbf{B}^\top\mathbf{A}^{-1} & \mathbf{S}^{-1} \end{pmatrix}.$$

References

[1] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization.* Cambridge university press, 2004.