The Singular Value Decomposition, Applications and Beyond

Zhihua Zhang
Shanghai Jiao Tong University
zhihua@sjtu.edu.cn

Contents

Abstract

The singular value decomposition (SVD) is not only a classical theory in matrix computation and analysis, but also is a powerful tool in machine learning and modern data analysis. In this tutorial we first study the basic notion of SVD and then show the central role of SVD in matrices. Using majorization theory, we consider variational principles of singular values and eigenvalues. Built on SVD and a theory of symmetric gauge functions, we discuss unitarily invariant norms, which are then used to formulate general results for matrix low rank approximation. We study the subdifferentials of unitarily invariant norms. These results would be potentially useful in many machine learning problems such as matrix completion and matrix data classification. Finally, we discuss matrix low rank approximation and its recent developments such as randomized SVD, approximate matrix multiplication, CUR decomposition, and Nyström approximation. Randomized algorithms are important approaches to large scale SVD as well as fast matrix computations.

1

Introduction

The singular value decomposition (SVD) is a classical matrix theory and a key computational technique, and it has also received wide applications in science and engineering. Compared with an eigenvalue decomposition (EVD) which only works on some of square matrices, SVD applies to all matrices. Moreover, many matrix concepts and properties such as matrix pseudoinverses, variational principles and unitarily invariant norms can be induced from SVD. Thus, SVD plays a fundamental role in matrix computation and analysis.

Furthermore, due to recent great developments of machine learning, data mining and theoretical computer science, SVD has been found to be more and more important. It is not only a powerful tool and theory but also an art. SVD makes matrices become a “Language" of data science.

The terminology of *singular values* has been proposed by Horn in 1950 and 1954 [Horn, 1951, 1954]. The first proof of the SVD for general $m \times n$ matrices might be given by Eckart and Young [1939]. But the theory of singular values can date back to the 19th century when it had been studied by the Italian differential geometer E. Beltrami, the French algebraist C. Jordan, the English mathematician J. J. Sylvester,

Table 1.1: Comparison of Matrix Factorization Methods

Matrices	Geometry	Data	Computation
$m \times n$	Polar	CX	QR
$m \times n$	SVD	CUR	QR
SPSD	Spectral	Nyström	(Incomplete) Cholesky

the French mathematician L. Autonne, etc. Please refer to Chapter 3 of Horn and Johnson [1991] in which the authors presented an excellent historical retrospection about the SVD or theory of singular values.

There is a rich literature involving singular values or SVD. Chapter 3 of Horn and Johnson [1991] provides exhaustive studies about inequalities of singular values as well as unitarily invariant norms, and the primary focus is on the matrix theory. The books by Watkins [1991], Demmel [1997], Golub and Van Loan [2012], Trefethen and Bau III [1997] present a detailed introduction to SVD, the primary focus of which is on numerical linear algebra.

This tutorial is motivated by recent successful applications of SVD in machine learning and theoretical computer science [Hastie et al., 2001, Burges, 2010, Halko et al., 2011, Woodruff, 2014b, Mahoney, 2011, Blum et al., 2015]. The primary focus is on a perspective of machine learning. The main purpose of the tutorial includes two aspects. First, it provides a systematic tutorial to the SVD theory and illustrates its functions in matrix and data analysis. Second, it provides an advanced review about recent developments of the SVD theory in applications of machine learning and theoretical computer science.

1.1 Roadmap

The preliminaries about matrices please refer to the book of Horn and Johnson [1985]. This tutorial involves matrix differential calculus, majorization theory, and symmetric gauge functions. For them, the detailed materials can be found in Macnus and Neudecker [2000], Marshal et al. [2010], Schatten [1950], Bhatia [1997]. In Chapter 2 we review some preliminaries such as Kronecker produces and vectorization operators, majorization theory, and derivatives.

In Chapter 3 we introduce the basic notion of SVD, including the existence, construction, and uniqueness. We then rederive some important matrix concepts and properties via SVD. We also study generalized SVD problems, which are concerned with joint decomposition of two matrices. In Chapter 4 we further illustrate the application of SVD in definition of the matrix pseudoinverse and solution of the Procrustes analysis problem. We discuss the role that SVD plays in subspace machine learning methods.

From the viewpoint of computation and modern data analysis, matrix factorization techniques should be the most important issue of matrices. In Table 1.1 we summary matrix factorization methods, which are categorized into three types. In particular, the Polar decomposition, SVD, and spectral decomposition consider geometric representation of a data matrix, whereas the CX, CUR, and Nyström dcompositions consider a compact representation of the data themselves. That is, the latters use a portion of the data to represent the whole data. The primary focus of the QR and Cholesky decomposition is on fast computation. In Chapter 5 we give reviews about the QR and CUR decompositions.

In Chapter 6 we consider variational principles for singular values and eigenvalues. Specifically, we apply matrix differential calculus to rederive the von Neumann theorem [Neumann, 1937] and the Ky Fan theorem [Fan, 1951]. Accordingly, we give some inequalities for singular values and eigenvalues.

Built on the inequalities for singular values, Chapter 7 discusses unitarily invariant norms. Unitarily invariant norms include the nuclear norm, Frobenius norm and spectral norm as their special cases. There is a one-to-one correspondence between a unitarily invariant norm of a matrix and a symmetric gauge function on the singular values of the matrix. This helps us to establish properties of unitarily invariant norms.

In Chapter 8 we study subdifferentials of unitarily invariant norms. We especially present the subdifferentials of the spectral norm and the nuclear norm as well as the applications in matrix low rank approximation. We illustrate several examples in optimization, which are solved via the subdifferentials of the spectral and nuclear norms. The subdifferentials of unitarily invariant norms would have potentially useful in machine learning and optimization.

Matrix low rank approximation is a promising theme in machine learning and theoretical computer science. Chapter 9 gives two important theorems about matrix low rank approximation based on errors of unitarily invariant norms. The first one is an extension of the ordinal least squares estimation problem. The second one was proposed by Mirsky [1960], which is an extension of the novel Eckart Young theorem [Eckart and Young, 1936]. We also discuss approximate matrix multiplication, which can be regarded as an inverse process of matrix low rank approximation.

In Chapter 10 we study randomized SVD, CUR approximation, and Nyström methods to make the applications scalable. The randomized SVD and CUR approximation can be also viewed as matrix low rank approximation techniques. The Nyström approximation is a special case of the CUR decomposition and has been widely used to speed up kernel methods.

1.2 Notation and Definitions

Throughout this tutorial, vectors and matrices are denoted by boldface lowercase letters and boldface uppercase letters, respectively. $\mathbb{R}_{+}^{n} = \{\mathbf{u} = (u_1, \ldots, u_n)^T \in \mathbb{R}^n : u_j \geq 0 \text{ for } j = 1, \ldots, n\}$ and $\mathbb{R}_{++}^{n} = \{\mathbf{u} = (u_1, \ldots, u_n)^T \in \mathbb{R}^n : u_j > 0 \text{ for } j = 1, \ldots, n\}$. Furthermore, if $\mathbf{u} \in \mathbb{R}_{+}^{n}$ (or $\mathbf{u} \in \mathbb{R}_{++}^{n}$), we also denote $\mathbf{u} \geq 0$ (or $\mathbf{u} > 0$).

Given a vector $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$, let $|\mathbf{x}| = (|x_1|, \ldots, |x_n|)^T$, let $\|\mathbf{x}\|_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$ for $p \geq 1$ be the ℓ_p-norm of $\mathbf{x}$, and let $\mathrm{diag}(\mathbf{x})$ be an $n \times n$ diagonal matrix with the ith diagonal element as x_i.

Let $[m] = \{1, 2, \ldots, m\}$, $\mathbf{I}_m$ be the $m \times m$ identity matrix, $\mathbf{1}_m$ be the $m \times 1$ vector of ones, and $\mathbf{0}$ be the zero vector or matrix with appropriate size. Let $\mathbf{A} \oplus \mathbf{B} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}$.

For a matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n] = [a_{ij}] \in \mathbb{R}^{m \times n}$, $\mathbf{A}^T$ denotes the transpose of $\mathbf{A}$, $\mathrm{rank}(\mathbf{A})$ denotes the rank, $\mathrm{range}(\mathbf{A})$ represents the range which is the space spanned by the columns (i.e., $\mathrm{range}(\mathbf{A}) =$

$\{\mathbf{y} \in \mathbb{R}^m : \mathbf{y} = \mathbf{A}\mathbf{x} \text{ for some } \mathbf{x} \in \mathbb{R}^n\} = \text{span}\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$), $\text{null}(\mathbf{A})$ is the null space (i.e., $\text{null}(\mathbf{A}) = \{\mathbf{x} : \mathbf{A}\mathbf{x} = 0\}$), and for $p = \min\{m, n\}$ $\mathsf{dg}(\mathbf{A})$ denotes the p-vector with a_{ii} as the ith element. Sometimes we also use Matlab Colon to represent a submatrix of $\mathbf{A}$. For example, let $I \subset [m]$ and $J \subset [n]$. $\mathbf{A}_{I,J}$ denotes the submatrix of $\mathbf{A}$ with rows indexed by I and columns indexed by J, $\mathbf{A}_{I,:}$ consists of those rows of $\mathbf{A}$ in I, and $\mathbf{A}_{:,J}$ consists of those columns of $\mathbf{A}$ in J.

Let $\|\mathbf{A}\|_F = \sqrt{\sum_{ij} a_{ij}^2}$ denote the Frobenius norm, $\|\mathbf{A}\|_2$ denote the spectral norm, and $\|\mathbf{A}\|_*$ denote the nuclear norm. When $\mathbf{A}$ is square, we let $\mathbf{A}^{-1}$ be the inverse (if exists) of $\mathbf{A}$, $\text{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii}$ be the trace, and $\det(\mathbf{A})$ be the determinant of $\mathbf{A}$.

An $m \times m$ real matrix $\mathbf{U}$ is symmetric if $\mathbf{A}^T = \mathbf{A}$, and skew-symmetric if $\mathbf{A}^T = -\mathbf{A}$, and normal if $\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A}$. Clearly, symmetric and skew-symmetric matrices are normal. An $m \times m$ real matrix $\mathbf{U}$ is said to be orthonormal (or orthogonal) if $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}_m$. An $m \times n$ real matrix $\mathbf{Q}$ for $m > n$ is column orthonormal (or column orthogonal) if $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_n$, and a column orthonormal $\mathbf{Q}$ is always able to be extended to an orthonormal matrix. A matrix $\mathbf{M} \in \mathbb{R}^{m \times m}$ is said to be positive semidefinite (PSD) or positive definite if for any nonzero vector $\mathbf{x} \in \mathbb{R}^m$ $\mathbf{x}^T\mathbf{M}\mathbf{x} \geq 0$ or $\mathbf{x}^T\mathbf{M}\mathbf{x} > 0$.

2

Preliminaries

In this chapter we present some preliminaries, including Kronecker products and vectorization operators, majorization theory, and derivatives. We list some basic results that will be used in this monograph but omit their detailed derivations.

2.1 Kronecker Products and Vectorization Operators

Given two matrices$\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{p\times q}$, the the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ is defined by

$$\mathbf{A} \otimes \mathbf{B} \triangleq \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix},$$

which is $mp \times nq$. The following properties can be found in Muirhead [1982].

Proposition 2.1. The Kronecker product has the following properties.

(a) $(\alpha\mathbf{A}) \otimes (\beta\mathbf{B}) = \alpha\beta(\mathbf{A} \otimes \mathbf{B})$ for any scalars $\alpha, \beta \in \mathbb{R}$.

(b) $(\mathbf{A} \otimes \mathbf{B})^T = \mathbf{A}^T \otimes \mathbf{B}^T$.

(c) $(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C})$.

(d) If $\mathbf{A}$ and $\mathbf{C}$ are both $m \times n$ and $\mathbf{B}$ is $p \times q$, then $(\mathbf{A}+\mathbf{C}) \otimes \mathbf{B} = \mathbf{A} \otimes \mathbf{B}+\mathbf{C} \otimes \mathbf{B}$ and $\mathbf{B} \otimes (\mathbf{A}+\mathbf{C}) = \mathbf{B} \otimes \mathbf{A}+\mathbf{B} \otimes \mathbf{C}$.

(e) If $\mathbf{A}$ is $m \times n$, $\mathbf{B}$ is $p \times q$, $\mathbf{C}$ is $n \times r$, and $\mathbf{D}$ is $q \times s$, then

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD}).$$

(f) If $\mathbf{U}$ and $\mathbf{V}$ are both orthogonal matrices, so is $\mathbf{U} \otimes \mathbf{V}$.

(g) If $\mathbf{A}$ and $\mathbf{B}$ are symmetric positive semidefinite (SPSD), so is $\mathbf{A} \otimes \mathbf{B}$.

Kronecker products often work with vectorization operators together. Let $\mathsf{vec}(\mathbf{A}) = (a_{11}, \ldots, a_{m1}, a_{12}, \ldots, a_{mn})^T \in \mathbb{R}^{mn}$ be vectorization of the matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$. The following lemma gives the connection between Kronecker products and vectorization operators.

Lemma 2.1.

(1) If $\mathbf{B}$ is $p \times m$, $\mathbf{X}$ is $m \times n$, and $\mathbf{C}$ is $n \times q$, then

$$\mathsf{vec}(\mathbf{BXC}) = (\mathbf{C}^T \otimes \mathbf{B})\mathsf{vec}(\mathbf{X}).$$

(2) If $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times p}$, and $\mathbf{C} \in \mathbb{R}^{p \times m}$, then

$$\mathrm{tr}(\mathbf{ABC}) = (\mathsf{vec}(\mathbf{A}^T))^T(\mathbf{I}_m \otimes \mathbf{B})\mathsf{vec}(\mathbf{C}).$$

(3) If $\mathbf{A} \in \mathbb{R}^{m \times p}$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{B} \in \mathbb{R}^{n \times n}$, and $\mathbf{C} \in \mathbb{R}^{p \times m}$, then

$$\begin{aligned}\mathrm{tr}(\mathbf{AX}^T\mathbf{BXC}) &= (\mathsf{vec}(\mathbf{X}))^T((\mathbf{CA})^T \otimes \mathbf{B})\mathsf{vec}(\mathbf{X}) \\ &= (\mathsf{vec}(\mathbf{X}))^T((\mathbf{CA}) \otimes \mathbf{B}^T)\mathsf{vec}(\mathbf{X}).\end{aligned}$$

2.2 Majorization

Given a vector $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$, let $\mathbf{x}^{\downarrow} = (x_1^{\downarrow}, \ldots, x_n^{\downarrow})$ be such a permutation of $\mathbf{x}$ that $x_1^{\downarrow} \geq x_2^{\downarrow} \geq \cdots \geq x_n^{\downarrow}$. Given two vectors $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{x} \geq \mathbf{y}$ means $x_i - y_i \geq 0$ for all $i \in [n]$. We say that $\mathbf{x}$ is majorized by $\mathbf{y}$ (denoted $\mathbf{x} \prec \mathbf{y}$) if $\sum_{i=1}^k x_i^{\downarrow} \leq \sum_{i=1}^k y_i^{\downarrow}$ for $k = 1, \ldots, n-1$ and $\sum_{i=1}^n x_i^{\downarrow} = \sum_{i=1}^n y_i^{\downarrow}$. Similarly, $\mathbf{x} \succ \mathbf{y}$ if $\sum_{i=1}^k x_i^{\downarrow} \geq \sum_{i=1}^k y_i^{\downarrow}$ for $k = 1, \ldots, n-1$ and $\sum_{i=1}^n x_i^{\downarrow} = \sum_{i=1}^n y_i^{\downarrow}$.

We say that $\mathbf{x}$ is weakly submajorized by $\mathbf{y}$ (denoted $\mathbf{x} \prec_w \mathbf{y}$) if $\sum_{i=1}^k x_i^{\downarrow} \leq \sum_{i=1}^k y_i^{\downarrow}$ for $k = 1, \ldots, n$, and $\mathbf{x}$ is weakly superrmajorized by $\mathbf{y}$ (denoted $\mathbf{x} \prec^w \mathbf{y}$) if $\sum_{i=1}^k x_i^{\downarrow} \geq \sum_{i=1}^k y_i^{\downarrow}$ for $k = 1, \ldots, n$,

An $n \times n$ matrix $\mathbf{W} = [w_{ij}]$ is said to be doubly stochastic if the $w_{ij} \geq 0$, $\sum_{j=1}^n w_{ij} = 1$ for all $i \in [n]$, and $\sum_{i=1}^n w_{ij} = 1$ for all $j \in [n]$. Note that if $\mathbf{Q} = [q_{ij}] \in \mathbb{R}^{n \times n}$ is orthonormal, then $\mathbf{W} \triangleq [q_{ij}^2]$ is a doubly stochastic matrix. It is thus called *orthostochastic*.

The following three lemmas are classical results in majorization theory. They will be used in investigating unitarily invariant norms.

Lemma 2.2. [Hardy et al., 1951] Given two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, then $\mathbf{x} \prec \mathbf{y}$ if and only if there exists a doubly stochastic matrix $\mathbf{W}$ such that $\mathbf{x} = \mathbf{W}\mathbf{y}$.

Lemma 2.3 (Birkhoff)**.** Let $\mathbf{W} \in \mathbb{R}^{n \times n}$. Then it is a doubly stochastic matrix if and only if it can be expressed as a convex combination of a set of permutation matrices.

Lemma 2.4. Let $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ be given nonnegative real numbers such that $u_1 \geq \cdots \geq u_n$ and $v_1 \geq \cdots \geq v_n$. If

$$\prod_{i=1}^k u_i \leq \prod_{i=1}^k v_i \quad \text{for } k \in [n],$$

then

$$\sum_{i=1}^k u_i \leq \sum_{i=1}^k v_i \quad \text{for } k \in [n].$$

More generally, assume f is a real-valued function such that $f(\exp(u))$ is increasing and convex. Then

$$\sum_{i=1}^k f(u_i) \leq \sum_{i=1}^k f(v_i) \quad \text{for } k \in [n].$$

2.3 Derivatives and Optimality

First let $f : \mathcal{X} \subset \mathbf{R}^n \to \mathbb{R}$ be a continuous function. The directional derivative of f at $\bar{\mathbf{x}}$ in a direction $\mathbf{u} \in \mathcal{X}$ is defined as

$$f'(\bar{\mathbf{x}}; \mathbf{u}) = \lim_{t \downarrow 0} \frac{f(\bar{\mathbf{x}} + t\mathbf{u}) - f(\bar{\mathbf{x}})}{t},$$

when this limit exists. When the directional derivative $f'(\bar{\mathbf{x}}; \mathbf{u})$ is linear in $\mathbf{u}$ (that is, $f'(\bar{\mathbf{x}}; \mathbf{u}) = \langle \mathbf{a}, \mathbf{u} \rangle$ for some $\mathbf{a} \in \mathcal{X}$) then we say f is (Gâbeaux) differentiable at $\bar{\mathbf{x}}$, with derivative $\nabla f(\bar{\mathbf{x}}) = \mathbf{a}$. If f is differentiable at every point in $\mathcal{X}$ then we say f is differentiable on $\mathcal{X}$.

When f is not differentiable but convex, we consider a notion of subdifferentials. We say $\mathbf{z}$ is the subgradient of f at $\bar{\mathbf{x}}$ if it satisfies

$$f(\bar{\mathbf{x}}) \leq f(\mathbf{x}) + \langle \mathbf{z}, \bar{\mathbf{x}} - \mathbf{x} \rangle \ \text{ for all points } \ \mathbf{z} \in \mathcal{X}.$$

The set of subgradients is called the subdifferential, and denoted by $\partial f(\bar{\mathbf{x}})$. The subdifferential is always a closed convex set. The following result shows a connection between subgradients and directional derivatives.

Lemma 2.5 (Max Formula)**.** If the function $f : \mathcal{X} \to (-\infty, +\infty]$ is convex, then any point $\bar{\mathbf{x}}$ in core(dom$f$) and any direction $\mathbf{u}$ in $\mathcal{X}$ satisfy

$$f'(\bar{\mathbf{x}}; \mathbf{u}) = \max \ \big\{ \langle \mathbf{z}, \mathbf{u} \rangle : \mathbf{z} \in \partial f(\bar{\mathbf{x}}) \big\}.$$

The further details of these results can be found from Borwein and Lewis [2006]. The following lemma then shows the fundamental role of subgradients in optimization.

Lemma 2.6. For any proper convex function $f : \mathcal{X} \to (-\infty, +\infty]$, the point $\bar{\mathbf{x}}$ is a minimizer of f if and only if the condition $\mathbf{0} \in \partial f(\bar{\mathbf{x}})$ holds.

Now let f be a differentiable function from $\mathbb{R}^{m\times n}$ to $\mathbb{R}$. For a matrix $\mathbf{X} = [x_{ij}] \in \mathbf{R}^{m\times n}$, $\frac{df(\mathbf{X})}{d\mathbf{X}} = \left(\frac{df}{dx_{ij}}\right)$ $(m \times n)$ defines the derivative of f w.r.t. $\mathbf{X}$. The Hessian matrix of f w.r..t. $\mathbf{X}$ is defined as $\frac{d^2 f(\mathbf{X})}{d\mathsf{vec}(\mathbf{X}) d\mathsf{vec}(\mathbf{X})^T}$, which is an $mn{\times}mn$ matrix. Let us see an example.

Example 2.1. We define the function f as

$$f(\mathbf{X}) = \mathrm{tr}(\mathbf{X}^T \mathbf{M} \mathbf{X}),$$

where $\mathbf{M} = [m_{ij}] \in \mathbb{R}^{m\times m}$ is a given constant matrix. It is directly computed that $\frac{df}{dx_{ij}} = \sum_{l=1}^{m} (m_{il} + m_{li}) x_{lj}$. This implies that $\frac{df}{d\mathbf{X}} = (\mathbf{M} + \mathbf{M}^T)\mathbf{X}$. In fact, the derivative can be computed as follows. Compute

$$df = \mathrm{tr}(d\mathbf{X}^T \mathbf{M} \mathbf{X} + \mathbf{X}^T \mathbf{M} d\mathbf{X}) = \mathrm{tr}((\mathbf{M} + \mathbf{M}^T) \mathbf{X} d\mathbf{X}^T).$$

We thus have that $\frac{df}{d\mathbf{X}} = (\mathbf{M} + \mathbf{M}^T)\mathbf{X}$.

Additionally, it follows from Lemma 2.1 that $f(\mathbf{X}) = \mathsf{vec}(\mathbf{X})^T(\mathbf{I}_n \otimes \mathbf{M})\mathsf{vec}(\mathbf{X})$. Thus, we have

$$\frac{df}{d\mathsf{vec}(\mathbf{X})} = \mathsf{vec}\Big(\frac{df}{d\mathbf{X}}\Big) = [\mathbf{I}_n \otimes (\mathbf{M} + \mathbf{M}^T)]\mathsf{vec}(\mathbf{X}),$$

and hence,

$$\frac{d^2 f(\mathbf{X})}{d\mathsf{vec}(\mathbf{X}) d\mathsf{vec}(\mathbf{X})^T} = \mathbf{I}_n \otimes (\mathbf{M} + \mathbf{M}^T).$$

3

The Singular Value Decomposition

The singular value decomposition (SVD) is a classical matrix theory and computational tool. In modern data computation and analysis, SVD becomes more and more important. In this chapter we aim to provide a systematical review about the basic principle of SVD.

We will see that there are four approaches to SVD. The first approach is depart from the spectral decomposition of a symmetric positive semidefinite (SPSD) matrix. The second approach gives a construction process via induction. In the third approach the SVD problem is equivalently formulated into an eigenvalue decomposition problem of a symmetric matrix (see Theorem 3.5). The fourth approach is based on the equivalent relationship between the SVD and polar decomposition (see Theorem 3.6).

We also study uniqueness of SVD (see Theorem 3.2 and Corollary 3.3). These results will be used in derivation of subdifferentials of unitarily invariant norms (see Chapter 8). Additionally, we present a generalized SVD (GSVD), which addresses joint decomposition problems of two matrices. When the two matrices form a column orthonormal matrix, the resulting GSVD process is called the CS decomposition.

3.1 Formulations

Given a nonzero SPSD matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$, let γ_i for $i = 1, \ldots, n$ be the eigenvalues of $\mathbf{M}$ and $\mathbf{x}_i$ be the corresponding eigenvectors. That is,

$$\mathbf{M}\mathbf{x}_i = \gamma_i \mathbf{x}_i, \quad i = 1, \ldots, n. \tag{3.1}$$

It is well known that the $\mathbf{x}_i$ can be assumed to be mutually orthonormal. Let $\mathbf{\Gamma} = \mathrm{diag}(\gamma_1, \ldots, \gamma_n)$ and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ such that $\mathbf{X}^T\mathbf{X} = \mathbf{I}_n$. We write (3.1) in matrix form as

$$\mathbf{M}\mathbf{X} = \mathbf{X}\mathbf{\Gamma}.$$

This gives rise to an *eigenvalue decomposition* (EVD) of $\mathbf{M}$:

$$\mathbf{M} = \mathbf{X}\mathbf{\Gamma}\mathbf{X}^T.$$

Since the γ_i are nonnegative, this decomposition is also called a *spectral decomposition* of the SPSD matrix $\mathbf{M}$.

Note that the above EVD always exists when $\mathbf{M}$ is symmetric but not PSD. However, the current eigenvalues γ_i are not necessarily nonnegative. Let $\hat{\mathbf{\Gamma}} = \mathrm{diag}(|\gamma_1|, \ldots, |\gamma_n|)$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$ with $\mathbf{y}_i = \mathrm{sgn}(\gamma_i)\mathbf{x}_i$ where $\mathrm{sgn}(0) = 1$. Then the decomposition is reformulated as

$$\mathbf{M} = \mathbf{X}\hat{\mathbf{\Gamma}}\mathbf{Y},$$

where $\mathbf{X}^T\mathbf{X} = \mathbf{I}_n$, $\mathbf{Y}^T\mathbf{Y} = \mathbf{I}_n$, and $\hat{\mathbf{\Gamma}}$ is a nonnegative diagonal matrix. This new formulation defines a singular value decomposition (SVD) of the symmetric matrix $\mathbf{M}$.

Naturally, a question emerges: does an SVD exist for an arbitrary matrix? Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ of rank r where $r \le \min\{m, n\}$. Without loss of generality, we assume $m \ge n$ for ease of exposition, because we can consider $\mathbf{A}^T$ when $m < n$.

Consider that $\mathbf{A}\mathbf{A}^T$ is SPSD, so it has the spectral decomposition, which is defined as

$$\mathbf{A}\mathbf{A}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T,$$

where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ and $\mathbf{U}^T\mathbf{U} = \mathbf{I}_m$. Since $\mathrm{rank}(\mathbf{A}\mathbf{A}^T) = \mathrm{rank}(\mathbf{A}) = r$, $\mathbf{A}\mathbf{A}^T$ has and only has r positive eigenvalues and the corresponding eigenvectors can form a column orthonormal matrix.

Assume that $\mathbf{\Lambda}_r = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r)$ and $\mathbf{U}_r = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r]$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$ are the positive eigenvalues of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{U}_r$ is the $m \times r$ matrix of the corresponding eigenvectors such that $\mathbf{U}_r^T\mathbf{U}_r = \mathbf{I}_r$. Thus, it follows from the spectral decomposition that

$$\mathbf{U}_r^T\mathbf{A}\mathbf{A}^T\mathbf{U}_r = \mathbf{\Lambda}_r$$

and $\mathbf{U}_{-r}^T\mathbf{A}\mathbf{A}^T\mathbf{U}_{-r} = \mathbf{0}$ where $\mathbf{U}_{-r}$ consists of the last $m{-}r$ columns of $\mathbf{U}$. Thus, we have $\mathbf{A}^T\mathbf{U}_{-r} = \mathbf{0}$. Let $\mathbf{V}_r = [\mathbf{v}_1, \ldots, \mathbf{v}_r] \triangleq \mathbf{A}^T\mathbf{U}_r\mathbf{\Lambda}_r^{-1/2}$. Then it satisfies $\mathbf{V}_r^T\mathbf{V}_r = \mathbf{I}_r$. Note that

$$\mathbf{A}^T\mathbf{U}(\mathbf{\Lambda}_r^{-1/2} \oplus \mathbf{I}_{m-r}) = [\mathbf{V}_r, \mathbf{A}^T\mathbf{U}_{-r}] = [\mathbf{V}_r, \mathbf{0}],$$

which implies that $\mathbf{A}^T = [\mathbf{V}_r, \mathbf{0}](\mathbf{\Lambda}_r^{\frac{1}{2}} \oplus \mathbf{I}_{m-r})\mathbf{U}^T = \mathbf{V}_r\mathbf{\Lambda}_r^{\frac{1}{2}}\mathbf{U}_r^T$. Hence,

$$\mathbf{A} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T, \tag{3.2}$$

where $\mathbf{\Sigma}_r = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$ with $\sigma_i = \lambda_i^{1/2}$ for $i = 1, \ldots, r$. Clearly, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

We refer to (3.2) as the *condensed SVD* of $\mathbf{A}$, where σ_i's are called the singular values, the columns $\mathbf{u}_i$ of $\mathbf{U}_r$ and the columns $\mathbf{v}_i$ of $\mathbf{V}_r$ are called respectively the left and right singular vectors of $\mathbf{A}$.

Recall that we always assume that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. Let $\mathbf{\Sigma}_n = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$ be the $n \times n$ diagonal matrix, and $\mathbf{U}_n$ be an $m \times n$ column-orthonormal matrix consisting of $\mathbf{U}_r$ in the first $m \times r$ block. In this case, we can equivalently write the condensed SVD of $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{U}_n\mathbf{\Sigma}_n\mathbf{V}^T, \tag{3.3}$$

which is called a *thin (or reduced) SVD* of $\mathbf{A}$. Furthermore, we extend $\mathbf{U}_n$ to a square orthonormal matrix (denoted $\mathbf{U}$), and $\mathbf{\Sigma}_n$ to an $m \times n$ matrix $\mathbf{\Sigma}$ by adding $m - n$ rows of zeros below. Then SVD can be also expressed as

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{3.4}$$

which is called a *full SVD* of $\mathbf{A}$.

As we have seen, these three expressions are mutually equivalent. We will sometimes use $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ for the thin SVD for notational simplicity. In a thin SVD version, let us always keep it in mind that

$\mathbf{\Sigma}$ is square and $\mathbf{U}$ or $\mathbf{V}$ is column orthonormal. We now present the formal formation of SVD of an arbitrary $\mathbf{A} \in \mathbb{R}^{m\times n}$ in which $m \geq n$ is not necessarily required.

Theorem 3.1. Given an arbitrary $\mathbf{A} \in \mathbb{R}^{m\times n}$, its full SVD defined in (3.4) always exists. Furthermore, the singular values σ_i are uniquely determined.

Based on the spectral decomposition of $\mathbf{A}\mathbf{A}^T$, we have previously shown the existence proof of the SVD theorem. Here we present a constructive proof, which has been widely given in the literature.

Proof. If $\mathbf{A}$ is zero, the result is trivial. Thus, let $\mathbf{A}$ be a nonzero matrix. Define $\sigma_1 \triangleq \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2$, which exists because $\mathbf{x} \mapsto \|\mathbf{A}\mathbf{x}\|_2$ is continuous and the set $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1\}$ is compact. Moreover, $\sigma_1 > 0$. Let $\mathbf{v}_1 \in \mathbb{R}^n$ be such a vector that $\sigma_1 = \|\mathbf{A}\mathbf{v}_1\|_2$. Define $\mathbf{u}_1 = \mathbf{A}\mathbf{v}_1/\sigma_1$, which satisfies $\|\mathbf{u}_1\|_2 = 1$.

We extend $\mathbf{u}_1$ and $\mathbf{v}_1$ to orthonormal matrices $\mathbf{U} = [\mathbf{u}_1, \mathbf{U}_{-1}]$ and $\mathbf{V} = [\mathbf{v}_1, \mathbf{V}_{-1}]$, respectively. Then

$$\mathbf{U}^T\mathbf{A}\mathbf{V} = \begin{bmatrix} \sigma_1 & \mathbf{u}_1^T\mathbf{A}\mathbf{V}_{-1} \\ \mathbf{0} & \mathbf{U}_{-1}^T\mathbf{A}\mathbf{V}_{-1} \end{bmatrix} \triangleq \mathbf{B},$$

where we use the fact $\mathbf{U}_{-1}^T\mathbf{A}\mathbf{v}_1 = \sigma_1\mathbf{U}_{-1}^T\mathbf{u}_1 = \mathbf{0}$. Note that

$$\max_{\|\mathbf{x}\|_2=1} \|\mathbf{B}\mathbf{x}\|_2^2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{U}^T\mathbf{A}\mathbf{V}\mathbf{x}\|_2^2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2^2 = \sigma_1^2.$$

However,

$$\frac{1}{\sigma_1^2 + \mathbf{z}^T\mathbf{z}} \left\| \mathbf{B} \begin{bmatrix} \sigma_1 \\ \mathbf{z} \end{bmatrix} \right\|_2^2 \geq \sigma_1^2 + \mathbf{z}^T\mathbf{z},$$

where $\mathbf{z} = \mathbf{V}_{-1}^T\mathbf{A}^T\mathbf{u}_1$. This implies that $\mathbf{z}$ must be zero.

The proof is completed by induction. In particular, assume $(m-1) \times (n-1)$ matrix $\mathbf{U}_{-1}^T\mathbf{A}\mathbf{V}_{-1}$ has a full SVD $\mathbf{U}_{-1}^T\mathbf{A}\mathbf{V}_{-1} = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^T$. Then $\mathbf{A}$ has a full SVD:

$$\begin{aligned}\mathbf{A} &= [\mathbf{u}_1, \mathbf{U}_{-1}] \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{U}} \end{bmatrix} \begin{bmatrix} \sigma_1 & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{\Sigma}} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{V}}^T \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{V}_{-1}^T \end{bmatrix} \\ &= [\mathbf{u}_1, \mathbf{U}_{-1}\tilde{\mathbf{U}}] \begin{bmatrix} \sigma_1 & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{\Sigma}} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ (\mathbf{V}_{-1}\tilde{\mathbf{V}})^T \end{bmatrix},\end{aligned}$$

because the matrices $[\mathbf{u}_1, \mathbf{U}_{-1}\tilde{\mathbf{U}}]$ and $[\mathbf{v}_1, \mathbf{V}_{-1}\tilde{\mathbf{V}}]$ are orthonormal. □

As for the uniqueness of the singular values is due to that the σ_i^2 are eigenvalues of $\mathbf{A}\mathbf{A}^T$ which are unique. Unfortunately, the left and right singular matrices $\mathbf{U}_r$ and $\mathbf{V}_r$ are not unique. However, we have the following result.

Theorem 3.2. Let $\mathbf{A} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T$ be a given condensed SVD of $\mathbf{A}$. Assume there are ρ distinct values among the nonzero singular values $\sigma_1, \ldots, \sigma_r$, with respective multiplicities r_i (satisfying $\sum_{i=1}^{\rho} r_i = r$). Then $\mathbf{A} = \tilde{\mathbf{U}}_r\mathbf{\Sigma}_r\tilde{\mathbf{V}}_r^T$ is a condensed SVD if and only if

$$\tilde{\mathbf{U}}_r = \mathbf{U}_r(\mathbf{Q}_1 \oplus \mathbf{Q}_2 \oplus \ldots \oplus \mathbf{Q}_\rho) \text{ and } \tilde{\mathbf{V}}_r = \mathbf{V}_r(\mathbf{Q}_1 \oplus \mathbf{Q}_2 \oplus \ldots \oplus \mathbf{Q}_\rho),$$

where $\mathbf{Q}_i$ is an arbitrary $r_i \times r_i$ orthonormal matrix.

Furthermore, if all the nonzero singular values are distinct, then the $\mathbf{Q}_i$ are either 1 or -1. In other words, the left and right singular vectors are uniquely determined up to signs.

Proof. Let $\delta_1 > \delta_2 > \ldots > \delta_\rho$ be the ρ distinct values among the $\sigma_1, \ldots, \sigma_r$. This implies that

$$\mathbf{\Sigma}_r = \delta_1\mathbf{I}_{r_1} \oplus \delta_2\mathbf{I}_{r_2} \oplus \ldots \oplus \delta_\rho\mathbf{I}_{r_\rho}. \tag{3.5}$$

The sufficiency follows from the fact that

$$(\mathbf{Q}_1 \oplus \ldots \oplus \mathbf{Q}_\rho)(\delta_1\mathbf{I}_{r_1} \oplus \ldots \oplus \delta_\rho\mathbf{I}_{r_\rho})(\mathbf{Q}_1^T \oplus \ldots \oplus \mathbf{Q}_\rho^T) = \mathbf{\Sigma}_r.$$

We now prove the necessary condition. Consider that $\text{range}(\mathbf{U}_r) = \text{range}(\mathbf{A}) = \text{range}(\tilde{\mathbf{U}}_r)$ and $\text{range}(\mathbf{V}_r) = \text{range}(\mathbf{A}^T) = \text{range}(\tilde{\mathbf{V}}_r)$. Thus, we have

$$\tilde{\mathbf{U}}_r = \mathbf{U}_r\mathbf{S} \text{ and } \tilde{\mathbf{V}}_r = \mathbf{V}_r\mathbf{T},$$

where $\mathbf{S}$ and $\mathbf{T}$ are some $r \times r$ orthonormal matrices. Hence, $\mathbf{\Sigma}_r = \mathbf{S}\mathbf{\Sigma}_r\mathbf{T}^T$, or equivalently, $\mathbf{\Sigma}_r\mathbf{T} = \mathbf{S}\mathbf{\Sigma}_r$. As in (3.5) for $\mathbf{\Sigma}$, partition $\mathbf{S}$ and $\mathbf{T}$ into

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \ldots & \mathbf{S}_{1\rho} \\ \vdots & \ddots & \vdots \\ \mathbf{S}_{\rho 1} & \ldots & \mathbf{S}_{\rho\rho} \end{bmatrix} \text{ and } \mathbf{T} = \begin{bmatrix} \mathbf{T}_{11} & \ldots & \mathbf{T}_{1\rho} \\ \vdots & \ddots & \vdots \\ \mathbf{T}_{\rho 1} & \ldots & \mathbf{T}_{\rho\rho} \end{bmatrix},$$

where $\mathbf{S}_{ij}$ and $\mathbf{T}_{ij}$ are $r_i \times r_j$. It follows from $\mathbf{\Sigma}_r \mathbf{T} = \mathbf{S}\mathbf{\Sigma}_r$ that $\delta_i \mathbf{T}_{ii} = \delta_i \mathbf{S}_{ii}$ for $i = 1, \ldots, \rho$ and $\delta_i \mathbf{T}_{ij} = \delta_j \mathbf{S}_{ij}$. As a result, we obtain that $\mathbf{S}_{ii} = \mathbf{T}_{ii}$ for $i = 1, \ldots, \rho$. Since $\mathbf{S}$ and $\mathbf{T}$ are orthonormal, we have

$$\sum_{j=1}^{\rho} \mathbf{S}_{ij}\mathbf{S}_{ij}^T = \mathbf{I}_{r_i} = \sum_{j=1}^{\rho} \mathbf{T}_{ij}\mathbf{T}_{ij}^T.$$

Note that $\sum_{j=1}^{\rho} \mathbf{T}_{\rho j}\mathbf{T}_{\rho j}^T = \sum_{j=1}^{\rho} \frac{\delta_j^2}{\delta_\rho^2} \mathbf{S}_{\rho j}\mathbf{S}_{\rho j}^T$, which implies that

$$\sum_{j<\rho} \Big[1 - \frac{\delta_j^2}{\delta_\rho^2}\Big] \mathbf{S}_{\rho j}\mathbf{S}_{\rho j}^T = \mathbf{0}. \tag{3.6}$$

Since $1 - \frac{\delta_j^2}{\delta_\rho^2} < 0$ for $j < \rho$ and $\mathbf{S}_{\rho j}\mathbf{S}_{\rho j}^T$ is always PSD, we must have $\mathbf{S}_{\rho j} = \mathbf{0}$ for all $j < \rho$, for otherwise, if there were a $k < \rho$ such that $\mathbf{S}_{\rho k} \neq \mathbf{0}$, there would exist a nonzero $\mathbf{x} \in \mathbb{R}^{r_\rho}$ such that $\mathbf{x}^T \mathbf{S}_{\rho k}\mathbf{S}_{\rho k}^T \mathbf{x} > 0$. It would lead to

$$\sum_{j<\rho} \Big[1 - \frac{\delta_j^2}{\delta_\rho^2}\Big] \mathbf{x}^T \mathbf{S}_{\rho j}\mathbf{S}_{\rho j}^T \mathbf{x} < 0,$$

which conflicts with (3.6). Accordingly, $\mathbf{S}_{\rho j} = \mathbf{T}_{\rho j} = \mathbf{0}$ for all $j < \rho$, and hence, $\mathbf{S}_{\rho\rho}\mathbf{S}_{\rho\rho}^T = \mathbf{T}_{\rho\rho}\mathbf{T}_{\rho\rho}^T = \mathbf{I}_{r_\rho}$. It also follows from the orthogonality of $\mathbf{S}$ and of $\mathbf{T}$ that for any $i < \rho$,

$$\mathbf{0} = \sum_{j=1}^{\rho} \mathbf{S}_{ij}\mathbf{S}_{\rho j}^T = \mathbf{S}_{i\rho}\mathbf{S}_{\rho\rho}^T \ \text{ and } \ \mathbf{0} = \sum_{j=1}^{\rho} \mathbf{T}_{ij}\mathbf{T}_{\rho j}^T = \mathbf{T}_{i\rho}\mathbf{T}_{\rho\rho}^T,$$

which leads to $\mathbf{S}_{i\rho} = \mathbf{T}_{i\rho} = \mathbf{0}$ for $i < \rho$.

Similarly, consider the $\rho-1, \rho-2, \ldots, 2$ cases. We have $\mathbf{S}_{ij} = \mathbf{T}_{ij} = \mathbf{0}$ for $i \neq j$, $\mathbf{S}_{ii} = \mathbf{T}_{ii}$ and $\mathbf{S}_{ii}\mathbf{S}_{ii}^T = \mathbf{T}_{ii}\mathbf{T}_{ii}^T = \mathbf{I}_{r_i}$ for $i \in [\rho]$. As a result, setting $\mathbf{Q}_i = \mathbf{S}_{ii}$ completes the proof. □

We now extend the result in Theorem 3.2 to the full SVD and thin SVD of $\mathbf{A}$. The following corollary is immediately obtained.

Corollary 3.3. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be a given full SVD of $\mathbf{A} \in \mathbb{R}^{m \times n}$. Then $\mathbf{A} = \tilde{\mathbf{U}}\mathbf{\Sigma}\tilde{\mathbf{V}}^T$ is a full SVD if and only if $\tilde{\mathbf{U}} = \mathbf{U}\mathbf{Q}$ and $\tilde{\mathbf{V}} = \mathbf{V}\mathbf{P}$ where $\mathbf{Q} = \mathbf{Q}_1 \oplus \cdots \oplus \mathbf{Q}_\rho \oplus \mathbf{Q}_0$ and $\mathbf{P} = \mathbf{Q}_1 \oplus \cdots \oplus \mathbf{Q}_\rho \oplus \mathbf{P}_0$. Here $\mathbf{Q}_1, \ldots, \mathbf{Q}_\rho$ are defined as in Theorem 3.2, and $\mathbf{Q}_0 \in \mathbb{R}^{(m-r)\times(m-r)}$ and

$\mathbf{P}_0 \in \mathbb{R}^{(n-r)\times(n-r)}$ are any orthonormal matrices. Obviously, $\mathbf{Q}\mathbf{\Sigma}\mathbf{P}^T = \mathbf{\Sigma}$ and $\mathbf{Q}^T\mathbf{\Sigma}\mathbf{P} = \mathbf{\Sigma}$ hold.

Assume $m \geq n$ and $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is a given thin SVD of $\mathbf{A} \in \mathbb{R}^{m\times n}$. Then $\mathbf{A} = \tilde{\mathbf{U}}\mathbf{\Sigma}\tilde{\mathbf{V}}^T$ is a thin SVD if and only if $\tilde{\mathbf{U}} = \mathbf{U}\mathbf{Q}$ and $\tilde{\mathbf{V}} = \mathbf{V}\mathbf{P}$ where $\mathbf{Q} = \mathbf{Q}_1 \oplus \cdots \oplus \mathbf{Q}_\rho \oplus \mathbf{Q}_0$ and $\mathbf{P} = \mathbf{Q}_1 \oplus \cdots \oplus \mathbf{Q}_\rho \oplus \mathbf{P}_0$. Currently, $\mathbf{Q}_0 \in \mathbb{R}^{(n-r)\times(n-r)}$ is any orthonormal matrix. Obviously, $\mathbf{Q}\mathbf{\Sigma} = \mathbf{\Sigma}\mathbf{Q}$, $\mathbf{\Sigma}\mathbf{P}^T = \mathbf{P}^T\mathbf{\Sigma}$, and $\mathbf{Q}\mathbf{\Sigma}\mathbf{P}^T = \mathbf{\Sigma}$ hold.

Theorem 3.2 and Corollary 3.3 will be used in derivation of subdifferentials of unitarily invariant norms (see Chapter 8). When the matrix in question is SPSD, the spectral decomposition and SVD are identical. That is, $\mathbf{U} = \mathbf{V}$ in this case. Moreover, the eigenvalues and singular values are identical.

The construction proof of Theorem 3.1 shows that

$\sigma_1(\mathbf{A}) = \max\{\|\mathbf{A}\mathbf{v}\|_2 : \mathbf{v} \in \mathbb{R}^n, \|\mathbf{v}\|_2 = 1\}$, so there exists a unit vector $\mathbf{v}_1 \in \mathbb{R}^n$ such that $\sigma_1(\mathbf{A}) = \|\mathbf{A}\mathbf{v}_1\|_2$;

$\sigma_2(\mathbf{A}) = \max\{\|\mathbf{A}\mathbf{v}\|_2 : \mathbf{v} \in \mathbb{R}^n, \|\mathbf{v}\|_2 = 1, \mathbf{v}^T\mathbf{v}_1 = 0\}$, so there exists a unit vector $\mathbf{v}_2 \in \mathbb{R}^n$ such that $\mathbf{v}_2^T\mathbf{v}_1 = 0$ and $\sigma_2(\mathbf{A}) = \|\mathbf{A}\mathbf{v}_2\|_2$;

$\vdots$

$\sigma_k(\mathbf{A}) = \max\{\|\mathbf{A}\mathbf{v}\|_2 : \mathbf{v} \in \mathbb{R}^n, \|\mathbf{v}\|_2 = 1, \mathbf{v}^T[\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}] = \mathbf{0}\}$, so there exists a unit vector $\mathbf{v}_k \in \mathbb{R}^n$ such that $\mathbf{v}_k^T[\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}] = \mathbf{0}$ and $\sigma_k(\mathbf{A}) = \|\mathbf{A}\mathbf{v}_k\|_2$;

$\vdots$

The following theorem is the generalization of the Courant-Fischer theorem for singular values.

Theorem 3.4. Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$ be the singular values of $\mathbf{A}$ where $p = \min\{m, n\}$. For any $k \in [p]$, then

$$\begin{aligned}\sigma_k &= \min_{\mathbf{v}_1,\ldots,\mathbf{v}_{k-1}\in\mathbb{R}^n} \max_{\substack{\mathbf{v}\in\mathbb{R}^n,\, \|\mathbf{v}\|_2=1\\ \mathbf{v}^T[\mathbf{v}_1,\ldots,\mathbf{v}_{k-1}]=\mathbf{0}}} \|\mathbf{A}\mathbf{v}\|_2 \\ &= \max_{\mathbf{v}_1,\ldots,\mathbf{v}_{n-k}\in\mathbb{R}^n} \min_{\substack{\mathbf{v}\in\mathbb{R}^n,\, \|\mathbf{v}\|_2=1\\ \mathbf{v}^T[\mathbf{v}_1,\ldots,\mathbf{v}_{n-k}]=\mathbf{0}}} \|\mathbf{A}\mathbf{v}\|_2.\end{aligned}$$

3.2 Matrix Properties via SVD

In what follows, we list some matrix properties which can be induced from SVD. These properties show that SVD is fundamental not only in matrix computation but also in matrix analysis.

Proposition 3.1. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be a full SVD of $m \times n$ matrix $\mathbf{A}$, and $\mathbf{A} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T$ be a condensed SVD. Let $p = \min\{m, n\}$. Then

(1) The rank of $\mathbf{A}$ is equal to the number of the nonzero singular values σ_i of $\mathbf{A}$.

(2) $\|\mathbf{A}\|_2 = \sigma_1$ is the spectral norm and $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\sum_{i=1}^p \sigma_i^2}$ is the Frobenius norm.

(3) $\text{range}(\mathbf{A}) = \text{range}(\mathbf{A}\mathbf{A}^T) = \text{range}(\mathbf{U}_r) = \text{span}(\mathbf{u}_1, \ldots, \mathbf{u}_r)$ and $\text{null}(\mathbf{A}) = \text{range}(\mathbf{V}_{-r}) = \text{span}(\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n)$.

(4) $\text{range}(\mathbf{A}^T) = \text{range}(\mathbf{A}^T\mathbf{A}) = \text{range}(\mathbf{V}_r) = \text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_r)$ and $\text{null}(\mathbf{A}^T) = \text{range}(\mathbf{U}_{-r}) = \text{span}(\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m)$.

(5) The eigenvalues of $\mathbf{A}^T\mathbf{A}$ are σ_i^2 for $i = 1, \ldots, r$ and $n-r$ zeros. The right singular vectors $\mathbf{v}_i$ are the corresponding orthonormal eigenvectors.

(6) The eigenvalues of $\mathbf{A}\mathbf{A}^T$ are σ_i^2 for $i = 1, \ldots, r$ and $m - r$ zeros. The left singular vectors $\mathbf{u}_i$ are the corresponding orthonormal eigenvectors.

(7) Let $\mathbf{B} = \mathbf{U}_B\mathbf{\Sigma}_B\mathbf{V}_B$ be the condensed SVD of $\mathbf{B}$. Then $\mathbf{A} \oplus \mathbf{B} = (\mathbf{U} \oplus \mathbf{U}_B)(\mathbf{\Sigma} \oplus \mathbf{\Sigma}_B)(\mathbf{V}^T \oplus \mathbf{V}_B^T)$ is the condensed SVD of $\mathbf{A}\oplus\mathbf{B}$, and $\mathbf{A}\otimes\mathbf{B} = (\mathbf{U}\otimes\mathbf{U}_B)(\mathbf{\Sigma}\otimes\mathbf{\Sigma}_B)(\mathbf{V}^T\otimes\mathbf{V}_B^T)$ is the condensed SVD of $\mathbf{A}\otimes\mathbf{B}$.

(8) If $\mathbf{A}$ is square and invertible, then $\mathbf{A}^{-1} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T$ and $|\det(\mathbf{A})| = \prod_{i=1}^n \sigma_i(\mathbf{A})$.

Theorem 3.5. Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, let $\mathbf{H} = \begin{bmatrix} \mathbf{0} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{bmatrix}$. If $\mathbf{A} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T$ be the condensed SVD, then $\mathbf{H}$ has $2r$ nonzero eigenvalues, which are $\pm\sigma_i$, with the corresponding orthonormal eigenvectors $\frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{v}_i \\ \pm\mathbf{u}_i \end{bmatrix}$, $i = 1, \ldots, r$.

Conversely, if γ_i is the eigenvalue of $\mathbf{H}$, with the corresponding eigenvector $\mathbf{z}_i = \begin{bmatrix} \mathbf{z}_i^{(1)} \\ \mathbf{z}_i^{(2)} \end{bmatrix}$ where $\mathbf{z}_i^{(1)} \in \mathbb{R}^n$ and $\mathbf{z}_i^{(2)} \in \mathbb{R}^m$, then $-\gamma_i$ is the eigenvalue of $\mathbf{H}$, with the corresponding eigenvector $\mathbf{z}_i = \begin{bmatrix} \mathbf{z}_i^{(1)} \\ -\mathbf{z}_i^{(2)} \end{bmatrix}$. Furthermore, let the σ_i denote the r positive values among the $\pm\gamma_i$, and $\frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{v}_i \\ \mathbf{u}_i \end{bmatrix}$ denote the corresponding orthonormal eigenvectors. Then $\mathbf{A} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T$, where $\mathbf{U}_r = [\mathbf{u}_1, \ldots, \mathbf{u}_r]$, $\mathbf{V}_r = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$, and $\mathbf{\Sigma}_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$, is a condensed SVD of $\mathbf{A}$.

Proof. The first part is directly obtained from the fact that

$$\mathbf{H} = \begin{bmatrix} \mathbf{0} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{V}_r \mathbf{\Sigma}_r \mathbf{U}_r^T \\ \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T & \mathbf{0} \end{bmatrix}$$
$$= \frac{1}{2} \begin{bmatrix} \mathbf{V}_r & \mathbf{V}_r \\ \mathbf{U}_r & -\mathbf{U}_r \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_r & \mathbf{0} \\ \mathbf{0} & -\mathbf{\Sigma}_r \end{bmatrix} \begin{bmatrix} \mathbf{V}_r^T & \mathbf{U}_r^T \\ \mathbf{V}_r^T & -\mathbf{U}_r^T \end{bmatrix}.$$

Conversely, consider that

$$\begin{bmatrix} \mathbf{0} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{z}_i^{(1)} \\ -\mathbf{z}_i^{(2)} \end{bmatrix} = \begin{bmatrix} -\mathbf{A}^T \mathbf{z}_i^{(2)} \\ \mathbf{A}\mathbf{z}_i^{(1)} \end{bmatrix} = \begin{bmatrix} -\gamma_i \mathbf{z}_i^{(1)} \\ \gamma_i \mathbf{z}_i^{(2)} \end{bmatrix} = -\gamma_i \begin{bmatrix} \mathbf{z}_i^{(1)} \\ -\mathbf{z}_i^{(2)} \end{bmatrix},$$

which shows that $-\gamma_i$ is the eigenvalue of $\mathbf{H}$, with the corresponding eigenvector $\begin{bmatrix} \mathbf{z}_i^{(1)} \\ -\mathbf{z}_i^{(2)} \end{bmatrix}$. Now using the notation of $\mathbf{\Sigma}_r$, $\mathbf{U}_r$, and $\mathbf{V}_r$, we have the EVD of $\mathbf{H}$:

$$\mathbf{H} = \begin{bmatrix} \mathbf{0} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \mathbf{V}_r & \mathbf{V}_r \\ \mathbf{U}_r & -\mathbf{U}_r \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_r & \mathbf{0} \\ \mathbf{0} & -\mathbf{\Sigma}_r \end{bmatrix} \begin{bmatrix} \mathbf{V}_r^T & \mathbf{U}_r^T \\ \mathbf{V}_r^T & -\mathbf{U}_r^T \end{bmatrix}.$$

It also follows from the orthogonality of the eigenvectors that $\mathbf{U}_r^T \mathbf{U}_r + \mathbf{V}_r^T \mathbf{V}_r = 2\mathbf{I}_r$ and $\mathbf{U}_r^T \mathbf{U}_r - \mathbf{V}_r^T \mathbf{V}_r = \mathbf{0}$. This implies that $\mathbf{U}_r^T \mathbf{U}_r = \mathbf{V}_r^T \mathbf{V}_r = \mathbf{I}_r$. Thus, $\mathbf{A} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T$ is a condensed SVD of $\mathbf{A}$. □

Theorem 3.5 establishes an interesting connection of the SVD of a general matrix with the EVD of a symmetric matrix. This provides an approach to handling an SVD problem of an arbitrary matrix. That is, one transforms the SVD problem into an EVD problem of an associated symmetric matrix. The theorem also gives an alternative proof for the SVD theory.

The following theorem shows that the *Polar Decomposition* of a matrix can be induced from its SVD. Note that SVD can be also derived from the Polar decomposition. Here we do not give the detail of this derivation.

Theorem 3.6 (Polar Decomposition)**.** Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a given matrix where $m \geq n$. Then its polar decomposition exists; that is, there are a column orthonormal matrix $\mathbf{Q}$ and a unique SPSD matrix $\mathbf{S}$ such that $\mathbf{A} = \mathbf{QS}$. Furthermore, if $\mathbf{A}$ is full column rank, then $\mathbf{Q}$ is unique.

Proof. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be a thin SVD of $\mathbf{A}$. Then

$$\mathbf{A} = \mathbf{U}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}\mathbf{V}^T \triangleq \mathbf{QS},$$

where $\mathbf{Q} \triangleq \mathbf{U}\mathbf{V}^T$ is column orthonormal and $\mathbf{S} \triangleq \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T$ is SPSD.

Assume that $\mathbf{A}$ has two Polar decompositions: $\mathbf{A} = \mathbf{Q}_1\mathbf{S}_1$ and $\mathbf{Q}_2\mathbf{S}_2$. Make the full SVDs (spectral decomposition) of $\mathbf{S}_1$ and $\mathbf{S}_2$ as $\mathbf{S}_1 = \mathbf{V}_1\mathbf{\Sigma}_1\mathbf{V}_1^T$ and $\mathbf{S}_2 = \mathbf{V}_2\mathbf{\Sigma}_2\mathbf{V}_2^T$, respectively. Then $\mathbf{A} = (\mathbf{Q}_1\mathbf{V}_1)\mathbf{\Sigma}_1\mathbf{V}_1^T$ and $\mathbf{A} = (\mathbf{Q}_2\mathbf{V}_2)\mathbf{\Sigma}_2\mathbf{V}_2^T$ be two thin SVDs of $\mathbf{A}$. This implies that $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 \triangleq \mathbf{\Sigma}$. Moreover, it follows from Corollary 3.3 that $\mathbf{V}_2 = \mathbf{V}_1\mathbf{P}_1$ and $\mathbf{Q}_2\mathbf{V}_2 = \mathbf{Q}_1\mathbf{V}_1\mathbf{P}_2$ where $\mathbf{P}_1$ and $\mathbf{P}_2$ are orthonormal matrices such that $\mathbf{\Sigma}\mathbf{P}_1^T = \mathbf{P}_1^T\mathbf{\Sigma}$. Thus, $\mathbf{S}_2 = \mathbf{V}_2\mathbf{\Sigma}\mathbf{V}_2^T = \mathbf{V}_1\mathbf{P}_1\mathbf{\Sigma}\mathbf{P}_1^T\mathbf{V}_1^T = \mathbf{V}_1\mathbf{\Sigma}\mathbf{V}_1^T = \mathbf{S}_1$.

If $\mathbf{A}$ is full column rank, then $\mathbf{S}$ is invertible. Hence, $\mathbf{Q}_1 = \mathbf{Q}_2$. □

As we see from the proof, $\mathbf{S} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T = (\mathbf{A}^T\mathbf{A})^{1/2}$; that is, $\mathbf{S}$ is identical to the square root of the matrix $\mathbf{A}^T\mathbf{A}$.

3.3 Matrix Concepts via SVD

All matrices have SVD, so SVD plays a central role in matrix analysis and computation. As we have seen in the previous section, many matrix concepts and properties can be induced from SVD. Here we present other several matrix notions, which are used in modern matrix computations.

Definition 3.1. Assume $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{m \times n}$ are of rank k and rank l, respectively, and $l \geq k$. Let $\mathbf{A} = \mathbf{U}_{A,k} \mathbf{\Sigma}_{A,k} \mathbf{V}_{A,k}^T$ and $\mathbf{B} = \mathbf{U}_{B,l} \mathbf{\Sigma}_{B,l} \mathbf{V}_{B,l}^T$ be the condensed SVDs of $\mathbf{A}$ and $\mathbf{B}$. The cosines of the canonical angles between $\mathbf{A}$ and $\mathbf{B}$ are defined as

$$\cos \theta_i(\mathbf{A}, \mathbf{B}) = \sigma_i(\mathbf{U}_{A,k}^T \mathbf{U}_{B,l}), \ i = 1, \ldots, k.$$

Consider that

$$\sigma^2(\mathbf{U}_{A,k}^T \mathbf{U}_{B,l}) = \lambda(\mathbf{U}_{A,k}^T \mathbf{U}_{B,l} \mathbf{U}_{B,l}^T \mathbf{U}_{A,k})$$

and $\mathbf{U}_{A,k}^T \mathbf{U}_{B,l} \mathbf{U}_{B,l}^T \mathbf{U}_{A,k} + \mathbf{U}_{A,k}^T \mathbf{U}_{B,-l} \mathbf{U}_{B,-l}^T \mathbf{U}_{A,k} = \mathbf{I}_k$, where $\mathbf{U}_{\mathbf{B},-l} \in \mathbb{R}^{m \times n-l}$ is an orthonormal complement of $\mathbf{U}_{B,l}$. Thus, we have that

$$\lambda(\mathbf{U}_{A,k}^T \mathbf{U}_{B,l} \mathbf{U}_{B,l}^T \mathbf{U}_{A,k}) = 1 - \lambda(\mathbf{U}_{A,k}^T \mathbf{U}_{B,-l} \mathbf{U}_{B,-l}^T \mathbf{U}_{A,k}).$$

In other words, $\sigma^2(\mathbf{U}_{A,k}^T \mathbf{U}_{B,l}) = 1 - \sigma^2(\mathbf{U}_{A,k}^T \mathbf{U}_{B,-l})$. Hence,

$$\sin \theta_i(\mathbf{A}, \mathbf{B}) = \sigma_{k+1-i}(\mathbf{U}_{A,k}^T \mathbf{U}_{B,-l}), \ i = 1, \ldots, k.$$

Note that $\sigma_1(\mathbf{U}_{A,k}^T \mathbf{U}_{B,-l}) = \|\mathbf{U}_{A,k}^T \mathbf{U}_{B,-l}\|_2$, which is also cased the distance between two subspaces spanned by $\mathbf{U}_{A,k}$ and $\mathbf{U}_{B,l}$.

Definition 3.2. Given a nonzero matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, let $\sigma_1 \geq \cdots \geq \sigma_p$ where $p = \min\{m, n\}$. The stable rank of $\mathbf{A}$ is defined as $\sum_{i=1}^p \frac{\sigma_i^2}{\sigma_1^2}$, and the nuclear rank is defined as $\sum_{i=1}^p \frac{\sigma_i}{\sigma_1}$.

Clearly, $\sum_{i=1}^p \frac{\sigma_i^2}{\sigma_1^2} \leq \sum_{i=1}^p \frac{\sigma_i}{\sigma_1} \leq \text{rank}(\mathbf{A})$. The concepts have been recently proposed for describing error bounds of matrix multiplication approximation [Magen and Zouzias, 2011, Cohen et al., 2015, Kyrillidis et al., 2014].

Definition 3.3 (Statistical Leverage Score)**.** Given an $m \times n$ matrix $\mathbf{A}$ with $m > n$, let $\mathbf{A}$ have a thin SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, and let $\mathbf{u}^{(i)}$ be the ith row of $\mathbf{U}$. Then the statistical leverage scores of the rows of $\mathbf{A}$ are defined as

$$l_i = \|\mathbf{u}^{(i)}\|_2^2 \ \text{ for } \ i = 1, \ldots, m.$$

The coherence of the rows of $\mathbf{A}$ is defined as

$$\gamma \triangleq \max_{i} \ l_i.$$

The (i, j)-cross leverage scores are defined as

$$c_{ij} = (\mathbf{u}^{(i)})^T \mathbf{u}^{(j)}.$$

The statistical leverage [Hoaglin and Welsch, 1978] measures the extent to which the singular vectors of a matrix are correlated with the standard basis. Recently, it has found usefulness in large-scale data analysis and in the analysis of randomized matrix algorithms [Drineas et al., 2008, Mahoney and Drineas, 2009, Ma et al., 2014]. A related notion is that of matrix coherence, which has been of interest in matrix completion and Nyström-based low rank matrix approximation [Candès and Recht, 2009, Talwalkar and Rostamizadeh, 2010, Wang and Zhang, 2013, Nelson and Nguyên, 2013].

3.4 Generalized Singular Value Decomposition

This section studies simultaneous SVD of two given matrices $\mathbf{A}$ and $\mathbf{B}$. This leads us to a generalized SVD (GSVD) problem.

Theorem 3.7 (GSVD)**.** Suppose two matrices $\mathbf{A} \in \mathbb{R}^{m\times p}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$ with $n \geq p$ are given. Let $q = \min\{m, p\}$. Then there exist two orthonormal matrices $\mathbf{U}_A \in \mathbb{R}^{m\times m}$ and $\mathbf{U}_B \in \mathbb{R}^{n\times n}$, and an invertible matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$ such that

$$\mathbf{U}_A^T \mathbf{A} \mathbf{X} = \mathrm{diag}(\alpha_1, \ldots, \alpha_q) \ \text{ and } \ \mathbf{U}_B^T \mathbf{B} \mathbf{X} = \mathrm{diag}(\beta_1, \ldots, \beta_p),$$

where $\alpha_1 \geq \cdots \alpha_q \geq 0$, and $0 \leq \beta_1 \leq \cdots \beta_p$.

The GSVD theorem was originally proposed by Loan [1976], in which $n \geq p$ (or $m \geq p$) is required. Later on, Paige and Saunders [1981] developed a more general formulation for GSVD in which matrix pencil $\mathbf{A}$ and $\mathbf{B}$ are required only to have the same number of columns. Paige and Saunders [1981] also studied a GSVD of submatrices of a column orthonormal matrix. That is a so-called CS decomposition [Golub et al., 1999] given as follows.

Theorem 3.8 (The CS Decomposition)**.** Let $\mathbf{Q} \in \mathbb{R}^{(m+n)\times p}$ be a column orthonormal matrix. Partition it as $\mathbf{Q}^T = [\mathbf{Q}_1^T, \mathbf{Q}_2^T]$ where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are $m\times p$ and $n\times p$. Then there exist orthonormal matrices $\mathbf{U}_1 \in \mathbb{R}^{m\times m}$, $\mathbf{U}_2 \in \mathbb{R}^{n\times n}$, and $\mathbf{V}_1 \in^{p\times p}$ such that

$$\mathbf{U}_1^T\mathbf{Q}_1\mathbf{V}_1 = \mathbf{C} \ \text{ and } \ \mathbf{U}_2^T\mathbf{Q}_2\mathbf{V}_1 = \mathbf{S},$$

where

$$\mathbf{C} = \begin{array}{c} \\ r \\ s \\ m{-}r{-}s \end{array} \begin{array}{c} \begin{array}{ccc} r & s & p{-}r{-}s \end{array} \\ \begin{pmatrix} \mathbf{I}_r & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \end{array},$$

$$\mathbf{S} = \begin{array}{c} \\ n{+}r{-}p \\ s \\ p{-}r{-}s \end{array} \begin{array}{c} \begin{array}{ccc} r & s & p{-}r{-}s \end{array} \\ \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{p-r-s} \end{pmatrix} \end{array},$$

$\mathbf{C}_1 = \mathrm{diag}(\alpha_1, \ldots, \alpha_s)$ and $\mathbf{S}_1 = \mathrm{diag}(\sqrt{1-\alpha_1^2}, \ldots, \sqrt{1-\alpha_s^2})$, and $1 > \alpha_1 \geq \alpha_2 \geq \cdots \alpha_s > 0$.

Proof. Since $\mathbf{Q}_1^T\mathbf{Q}_1 + \mathbf{Q}_2^T\mathbf{Q}_2 = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_p$, the largest eigenvalue of $\mathbf{Q}_1^T\mathbf{Q}_1$ (reps. $\mathbf{Q}_2^T\mathbf{Q}_2$) is at most 1. This implies $\|\mathbf{Q}_1\|_2 = \sigma_1(\mathbf{Q}_1) \leq 1$ (resp. $\|\mathbf{Q}_2\|_2 \leq 1$). Let $q = \min\{m, p\}$. Make a full SVD of $\mathbf{Q}_1$ as

$$\mathbf{Q}_1 = \mathbf{U}_1\mathbf{C}\mathbf{V}_1^T,$$

where $\mathbf{C} = \mathrm{diag}(c_1, \ldots, c_q)$ is an $m\times p$ diagonal matrix. Assume

$$1 = c_1 = \cdots = c_r > c_{r+1} \geq \cdots \geq c_{r+s} > c_{r+s+1} = \cdots c_p = 0.$$

Let $\mathbf{D} = \mathrm{diag}(c_{r+1}, \ldots, c_{r+s}) \oplus \mathbf{0}$, which is $(m-r) \times (p-r)$, and

$$\mathbf{Q}_2\mathbf{V}_1 = [\underbrace{\mathbf{W}_1}_{r}, \underbrace{\mathbf{W}_2}_{p-r}].$$

Then

$$\begin{bmatrix} \mathbf{U}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix}^T \begin{bmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \end{bmatrix} \mathbf{V}_1 = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \\ \mathbf{W}_1 & \mathbf{W}_2 \end{bmatrix}$$

is column orthonormal. This implies that $\mathbf{W}_1 = \mathbf{0}$ and

$$\mathbf{W}_2^T\mathbf{W}_2 = \mathbf{I}_{p-r} - \mathbf{D}^T\mathbf{D} = \mathrm{diag}(1 - c_{r+1}^2, \ldots, 1 - c_p^2)$$

is nonsingular. Define $s_i = \sqrt{1 - c_i^2}$ for $i \in [p]$. Then

$$\mathbf{Z} \triangleq \mathbf{W}_2\mathrm{diag}(1/s_{r+1}, \ldots, 1/s_p)$$

is column orthonormal. We now extend $\mathbf{Z}$ to an $n \times n$ orthonormal matrix $\mathbf{U}_2$, the last $p - r$ columns of which constitute $\mathbf{Z}$. When setting $\alpha_1 = c_{r+1}, \cdots, \alpha_s = c_{r+s}$, we have

$$\mathbf{U}_2^T\mathbf{Q}_1\mathbf{V}_1 = \mathbf{S}.$$

Thus, the theorem follows. □

Remarks It is worth pointing out that $\mathbf{Q}_1 = \mathbf{U}_2\mathbf{S}\mathbf{V}_1^T$ is not certainly a full SVD of $\mathbf{Q}_1$, because some of the nonzero elements of $\mathbf{S}$ might not lie on the principal diagonal. However, if $n \geq p$, then we can move the first $n - p$ rows of $\mathbf{S}$ to be the last $n - p$ rows by pre-multiplying some permutation matrix $\mathbf{P}$. That is,

$$\mathbf{P}^T\mathbf{U}_2^T\mathbf{Q}_1\mathbf{V}_1 = \begin{array}{c} \\ r \\ s \\ p{-}r{-}s \\ n{-}p \end{array} \begin{array}{c} \begin{array}{ccc} r & s & p{-}r{-}s \end{array} \\ \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{p-r-s} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \end{array}.$$

This is the reason why the restriction $n \geq p$ is required in Theorem 3.7 ($\mathbf{A}$ and $\mathbf{B}$ correspond to $\mathbf{Q}_1$ and $\mathbf{Q}_2$, respectively).

The following theorem gives a more general version of Theorem 3.7 as well as Theorem 3.8. Compared with Theorem 3.7, $m \geq p$ or $n \geq p$ are no longer restricted. Compared with Theorem 3.8, the submatrices in question do not necessarily form a column orthonormal matrix.

Theorem 3.9. Suppose two matrices $\mathbf{A} \in \mathbb{R}^{m\times p}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$ are given. Let $\mathbf{K}^T \triangleq [\mathbf{A}^T, \mathbf{B}^T]$ with the rank t. Then exist orthonormal matrices $\mathbf{U}_A \in \mathbb{R}^{m\times m}$, $\mathbf{U}_B \in \mathbb{R}^{n\times n}$, $\mathbf{W} \in \mathbb{R}^{t\times t}$, and $\mathbf{V} \in \mathbb{R}^{p\times p}$ such that

$$\mathbf{U}_A^T\mathbf{A}\mathbf{V} = \mathbf{\Sigma}_A[\underbrace{\mathbf{W}^T\mathbf{R}}_{t}, \underbrace{\mathbf{0}}_{p-t}] \text{ and } \mathbf{U}_B^T\mathbf{B}\mathbf{V} = \mathbf{\Sigma}_B[\underbrace{\mathbf{W}^T\mathbf{R}}_{t}, \underbrace{\mathbf{0}}_{p-t}],$$

where $\mathbf{R} \in \mathbb{R}^{t\times t}$ is a positive diagonal matrix with its diagonal elements equal to the nonzero of singular values of $\mathbf{K}$,

$$\mathbf{\Sigma}_A = \begin{array}{c} \\ r \\ s \\ m{-}r{-}s \end{array} \!\!\begin{array}{c} \begin{array}{ccc} r & s & t{-}r{-}s \end{array} \\ \begin{pmatrix} \mathbf{I}_r & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \end{array}, \tag{3.7}$$

$$\mathbf{\Sigma}_B = \begin{array}{c} \\ n{+}r{-}t \\ s \\ t{-}r{-}s \end{array} \!\!\begin{array}{c} \begin{array}{ccc} r & s & t{-}r{-}s \end{array} \\ \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_B & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{t-r-s} \end{pmatrix} \end{array}. \tag{3.8}$$

Here r and s depend on the context,

$$\mathbf{D}_A = \mathrm{diag}(\alpha_{r+1}, \ldots, \alpha_{r+s}) \text{ and } \mathbf{D}_B = \mathrm{diag}(\sqrt{1{-}\alpha_{r+1}^2}, \ldots, \sqrt{1{-}\alpha_{r+s}^2}),$$

and $1 > \alpha_{r+1} \geq \cdots \geq \alpha_{r+s} > 0$.

Theorem 3.9 implies that

$$\mathbf{U}_A^T\mathbf{A}\mathbf{X} = [\mathbf{\Sigma}_A, \mathbf{0}] \text{ and } \mathbf{U}_B^T\mathbf{B}\mathbf{X} = [\mathbf{\Sigma}_B, \mathbf{0}],$$

where $\mathbf{X} \triangleq \mathbf{V}(\mathbf{R}^{-1}\mathbf{W} \oplus \mathbf{I}_{p-t})$. With the above remarks, Theorem 3.7 follows. Thus, we now present the proof of Theorem 3.9.

Proof. Since $\mathrm{rank}(\mathbf{K}) = t$, making a full SVD of $\mathbf{K}$ yields

$$\mathbf{P}^T\mathbf{K}\mathbf{V} = \begin{bmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{P} \in \mathbb{R}^{(m+n)\times(m+n)}$ and $\mathbf{V} \in \mathbb{R}^{p\times p}$ are orthonormal matrices, $\mathbf{R}$ is a $t \times t$ diagonal matrix with the diagonal elements as the nonzero singular values of $\mathbf{K}$. Partition $\mathbf{P}$ as

$$\mathbf{P} = [\underbrace{\mathbf{P}_1}_{t}, \underbrace{\mathbf{P}_2}_{m+n-t}] = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{bmatrix} \text{ where } \mathbf{P}_{11} \in \mathbb{R}^{m\times t} \text{ and } \mathbf{P}_{21} \in \mathbb{R}^{n\times t}.$$

Obviously, $\mathbf{P}_1^T\mathbf{P}_1 = \mathbf{P}_{11}^T\mathbf{P}_{11} + \mathbf{P}_{21}^T\mathbf{P}_{21} = \mathbf{I}_t$. Moreover, we have

$$\mathbf{K}\mathbf{V} = [\mathbf{P}_1\mathbf{R}, \mathbf{0}].$$

Applying Theorem 3.8 to $\mathbf{P}_1$ yields that there exist orthonormal matrices $\mathbf{U}_A \in \mathbb{R}^{m\times m}$, $\mathbf{U}_B \in \mathbb{R}^{n\times n}$, and $\mathbf{W} \in \mathbb{R}^{t\times t}$ such that

$$\begin{bmatrix} \mathbf{U}_A^T & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_B^T \end{bmatrix} \begin{bmatrix} \mathbf{P}_{11} \\ \mathbf{P}_{21} \end{bmatrix} \mathbf{W} = \begin{bmatrix} \mathbf{\Sigma}_A \\ \mathbf{\Sigma}_B \end{bmatrix}$$

where $\mathbf{\Sigma}_A$ and $\mathbf{\Sigma}_B$ are defined in (3.7) and (3.8). Hence,

$$\begin{bmatrix} \mathbf{U}_A^T & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_B^T \end{bmatrix} \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \mathbf{V} = \begin{bmatrix} \mathbf{\Sigma}_A \mathbf{W}^T \mathbf{R} & \mathbf{0} \\ \mathbf{\Sigma}_B \mathbf{W}^T \mathbf{R} & \mathbf{0} \end{bmatrix}.$$

That is, $\mathbf{U}_A^T \mathbf{A} \mathbf{V} = \mathbf{\Sigma}_A[\mathbf{W}^T\mathbf{R}, \mathbf{0}]$ and $\mathbf{U}_B^T \mathbf{B} \mathbf{V} = \mathbf{\Sigma}_B[\mathbf{W}^T\mathbf{R}, \mathbf{0}]$. □

In terms of Theorem 3.7, if $\beta_i \neq 0$, then the column $\mathbf{x}_i$ of $\mathbf{X}$ satisfies

$$\mathbf{A}^T\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{B}^T\mathbf{B}\mathbf{x}_i,$$

where $\lambda_i = \frac{\alpha_i^2}{\beta_i^2}$. This implies GSVD can be used to solve generalized eigenvalue problems. Based on this observation, Howland et al. [2003], Park and Park [2005] applied GSVD for solving Fisher linear discriminant analysis (FLDA) and generalized Fisher discriminant analysis [Baudat and Anouar, 2000, Mika et al., 2000].

Recall that the above GSVD procedure requires to implementing an SVD on the $(m+n) \times p$ matrix $\mathbf{K}$. The computational cost is $O((m+n)p * \min\{m+n, p\})$. Thus, when both $m+n$ and p are very large, the GSVD is less efficient. We now consider a special case in which $\mathbf{B} = \mathbf{Z}\mathbf{A}$ where $\mathbf{Z} \in \mathbb{R}^{n\times m}$ is some given matrix. We will see that it is no longer necessary to perform the SVD on $\mathbf{K}$.

Theorem 3.10. Let $\mathbf{A} \in \mathbb{R}^{m\times p}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$ be two given matrices. Assume that $\mathbf{B} = \mathbf{Z}\mathbf{A}$ where $\mathbf{Z} \in \mathbb{R}^{n\times m}$ is some matrix, $\mathrm{rank}(\mathbf{B}) = s$, and $\mathrm{rank}(\mathbf{A}) = t$. Let $\mathbf{A} = \mathbf{U}_t\mathbf{\Sigma}_t\mathbf{V}_t^T$ be a condensed SVD of $\mathbf{A}$, and $\mathbf{Y} = \mathbf{U}_Y\mathbf{\Sigma}_Y\mathbf{V}_Y^T$ be a full SVD of $\mathbf{Y} \triangleq \mathbf{Z}\mathbf{U}_t$. Then

$$(\mathbf{U}_t\mathbf{V}_Y)^T\mathbf{A}\mathbf{V}_t\mathbf{\Sigma}_t^{-1}\mathbf{V}_Y = \mathbf{I}_t \;\text{ and }\; \mathbf{U}_Y^T\mathbf{B}\mathbf{V}_t\mathbf{\Sigma}_t^{-1}\mathbf{V}_Y = \mathbf{\Sigma}_Y.$$

The proof is direct. Assume $\mathbf{U}_t$ and $\mathbf{V}_t$ are extended to orthonormal matrices $\mathbf{U}$ $(m \times m)$ and $\mathbf{V}$ $(p \times p)$. Let

$$\mathbf{X} = \mathbf{V}(\mathbf{\Sigma}_t^{-1}\mathbf{V}_Y \oplus \mathbf{I}_{p-t}).$$

We now have that

$$\mathbf{AX} = \mathbf{UU}^T\mathbf{AV}(\boldsymbol{\Sigma}_t^{-1}\mathbf{V}_Y \oplus \mathbf{I}_{p-t}) = [\mathbf{U}_t\mathbf{V}_Y, \mathbf{0}] = \mathbf{U}(\mathbf{V}_Y \oplus \mathbf{I}_{m-t})(\mathbf{I}_t \oplus \mathbf{0})$$

and

$$\begin{aligned}\mathbf{BX} &= \mathbf{ZUU}^T\mathbf{AV}(\boldsymbol{\Sigma}_t^{-1}\mathbf{V}_Y \oplus \mathbf{I}_{p-t}) = \mathbf{ZU}_t[\mathbf{V}_Y, \mathbf{0}] \\ &= \mathbf{U}_Y\boldsymbol{\Sigma}_Y\mathbf{V}_Y^T[\mathbf{V}_Y, \mathbf{0}] = \mathbf{U}_Y[\boldsymbol{\Sigma}_Y, \mathbf{0}].\end{aligned}$$

Thus,

$$(\mathbf{V}_Y^T \oplus \mathbf{I}_{m-t})\mathbf{U}^T\mathbf{AX} = [\mathbf{I}_t \oplus \mathbf{0}]$$

and

$$\mathbf{U}_Y^T\mathbf{BX} = [\boldsymbol{\Sigma}_Y, \mathbf{0}].$$

In this special case, we only need to implement two SVDs on two matrices with smaller sizes. The diagonal elements of $\boldsymbol{\Sigma}_Y$ and the columns of $\mathbf{V}_t\boldsymbol{\Sigma}_t^{-1}\mathbf{V}_Y$ are the generalized eigenvalues and eigenvectors of the corresponding generalized eigenvalue problem.

Remarks Assume that $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{m\times n}$ have the same size. Gibson [1974] proved that they have joint factorizations $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}_A\mathbf{V}^T$ and $\mathbf{B} = \mathbf{U}\boldsymbol{\Sigma}_B\mathbf{V}^T$ if and only if $\mathbf{AB}^T$ and $\mathbf{B}^T\mathbf{A}$ are both normal. Here $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices, and both $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_B$ are diagonal but their diagonal elements are perhaps complex. These diagonal elements are nonnegative only if both $\mathbf{AB}^T$ and $\mathbf{B}^T\mathbf{A}$ are SPSD.

4

Applications of SVD: Case Studies

In the previous chapter we present the basic notion and some important properties of SVD. Meanwhile, we show that many matrix properties can be rederived via SVD. In this chapter, we further illustrate applications of SVD in matrices, including in the definition of the Moore-Penrose pseudoinverse of an arbitrary matrix and in the analysis of the Procrustes problem.

For any matrix, the Moore-Penrose pseudoinverse exists and is unique. Moreover, it has been found to have many applications. Thus, it is an important matrix notion. In this chapter we exploit the matrix pseudoinverse to solve least squares estimation, giving rise to a more general result. We also show that the matrix pseudoinverse can be used to deal with a class of generalized eigenvalue problems.

In fact, SVD has also wide applications in machine learning and data analysis. For example, SVD is an important tool in spectral analysis [Azar et al., 2001], latent semantic indexing [Papadimitriou et al., 1998], spectral clustering, and projective clustering [Feldman et al., 2013]. We specifically show that SVD plays a fundamental role in subspace methods such as PCA, MDS, FDA and CCA.

4.1 The Matrix MP Pseudoinverse

Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and a vector $\mathbf{b} \in \mathbb{R}^m$, we are concerned with the least squares estimation problem:

$$\hat{\mathbf{x}} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2. \tag{4.1}$$

The minimizer should satisfy the Karush-Kuhn-Tucker (KKT) condition: that is, it is the solution of the following normal equation:

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}. \tag{4.2}$$

Let $\mathbf{A} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T$ be the condensed SVD of $\mathbf{A}$. Then $\mathbf{V}_r\mathbf{\Sigma}_r^2\mathbf{V}_r^T\mathbf{x} = \mathbf{V}_r\mathbf{\Sigma}_r\mathbf{U}_r^T\mathbf{b}$. Define $\mathbf{A}^\dagger = \mathbf{V}_r\mathbf{\Sigma}_r^{-1}\mathbf{U}_r^T \in \mathbb{R}^{n\times m}$. Obviously,

$$\hat{\mathbf{x}} = \mathbf{A}^\dagger\mathbf{b}$$

is a minimizer. It is clear that if $\mathbf{A}$ is invertible, then the minimizer is $\hat{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{b}$. Thus, $\mathbf{A}^\dagger$ is a generalization of $\mathbf{A}^{-1}$ in the case that $\mathbf{A}$ is an arbitrary matrix, i.e., it is not necessarily invertible and even non-square. This leads us to the notion of the matrix Moore-Penrose (MP) pseudoinverse [Ben-Israel and Greville, 2003].

Definition 4.1. Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, a real $n \times m$ matrix $\mathbf{B}$ is called the MP pseudoinverse of $\mathbf{A}$ if it satisfies the following four conditions: (1) $\mathbf{ABA} = \mathbf{A}$, (2) $\mathbf{BAB} = \mathbf{B}$, (3) $(\mathbf{AB})^T = \mathbf{AB}$, and (4) $(\mathbf{BA})^T = \mathbf{BA}$.

It is easily verified that $\mathbf{A}^\dagger = \mathbf{V}_r\mathbf{\Sigma}_r^{-1}\mathbf{U}_r^T$ is a pseudoinverse of $\mathbf{A}$. Moreover, when $\mathbf{A}$ is invertible, $\mathbf{A}^\dagger$ is identical to $\mathbf{A}^{-1}$. The following theorem then shows that $\mathbf{A}^\dagger$ is the unique pseudoinverse of $\mathbf{A}$.

Theorem 4.1. Let $\mathbf{A} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T$ be the condensed SVD of $\mathbf{A} \in \mathbb{R}^{m\times n}$. Then $\mathbf{B}$ is the pseudoinverse of $\mathbf{A}$ if and only if $\mathbf{B} = \mathbf{A}^\dagger \triangleq \mathbf{V}_r\mathbf{\Sigma}_r^{-1}\mathbf{U}_r^T$.

Proof. To complete the proof, it suffices to prove the uniqueness of the pseudoinverse. Assume that $\mathbf{B}$ and $\mathbf{C}$ are two pseudoinverses of $\mathbf{A}$. Then

$$\begin{aligned}\mathbf{AB} &= (\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T = \mathbf{B}^T(\mathbf{ACA})^T = \mathbf{B}^T\mathbf{A}^T\mathbf{C}^T\mathbf{A}^T \\ &= (\mathbf{AB})^T(\mathbf{AC})^T = (\mathbf{ABA})\mathbf{C} = \mathbf{AC}.\end{aligned}$$

Similarly, it also holds that $\mathbf{BA} = \mathbf{CA}$. Thus,

$$\mathbf{B} = \mathbf{BAB} = \mathbf{BAC} = \mathbf{CAC} = \mathbf{C}.$$

□

The matrix pseudoinverse also has wide applications. Let us see its application in solving generalized eigenproblems. Given two matrices $\mathbf{M}$ and $\mathbf{N} \in \mathbb{R}^{m \times m}$, we refer to $(\mathbf{\Lambda}, \mathbf{X})$ where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$ and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_q]$ as q eigenpairs of the matrix pencil $(\mathbf{M}, \mathbf{N})$ if $\mathbf{MX} = \mathbf{NX\Lambda}$; namely,

$$\mathbf{M}\mathbf{x}_i = \lambda_i \mathbf{N}\mathbf{x}_i, \quad \text{for } i = 1, \ldots, q.$$

The problem of finding eigenpairs of $(\mathbf{M}, \mathbf{N})$ is known as a *generalized eigenproblem*. Clearly, when $\mathbf{N} = \mathbf{I}_m$, the problem becomes the conventional eigenvalue problem.

Usually, we are interested in the problem with the nonzero λ_i for $i = 1, \ldots, q$ and refer to $(\mathbf{\Lambda}, \mathbf{X})$ as the nonzero eigenpairs of $(\mathbf{M}, \mathbf{N})$. If $\mathbf{N}$ is nonsingular, $(\mathbf{\Lambda}, \mathbf{X})$ is also referred to as the (nonzero) eigenpairs of $\mathbf{N}^{-1}\mathbf{M}$ because the generalized eigenproblem is equivalent to the eigenproblem:

$$\mathbf{N}^{-1}\mathbf{MX} = \mathbf{X\Lambda}.$$

However, when $\mathbf{N}$ is singular, Zhang et al. [2010] suggested to use a pseudoinverse eigenproblem:

$$\mathbf{N}^{\dagger}\mathbf{MX} = \mathbf{X\Lambda}.$$

Moreover, Zhang et al. [2010] established a connection between the solutions of the generalized eigenproblem and its corresponding pseudoinverse eigenproblem. That is,

Theorem 4.2. Let $\mathbf{M}$ and $\mathbf{N}$ be two matrices in $\mathbb{R}^{m \times m}$. Assume $\mathrm{range}(\mathbf{M}) \subseteq \mathrm{range}(\mathbf{N})$. Then, if $(\mathbf{\Lambda}, \mathbf{X})$ are the nonzero eigenpairs of $\mathbf{N}^{\dagger}\mathbf{M}$, we have that $(\mathbf{\Lambda}, \mathbf{X})$ are the nonzero eigenpairs of the matrix pencil $(\mathbf{M}, \mathbf{N})$. Conversely, if $(\mathbf{\Lambda}, \mathbf{X})$ are the nonzero eigenpairs of the matrix pencil $(\mathbf{M}, \mathbf{N})$, then $(\mathbf{\Lambda}, \mathbf{N}^{\dagger}\mathbf{NX})$ are the nonzero eigenpairs of $\mathbf{N}^{\dagger}\mathbf{M}$.

Proof. Let $\mathbf{M} = \mathbf{U}_1\mathbf{\Gamma}_1\mathbf{V}_1^T$ and $\mathbf{N} = \mathbf{U}_2\mathbf{\Gamma}_2\mathbf{V}_2^T$ be the condensed SVD of $\mathbf{M}$ and $\mathbf{N}$. Thus, we have range($\mathbf{M}$) = range($\mathbf{U}_1$) and range($\mathbf{N}$) = range($\mathbf{U}_2$). Moreover, we have $\mathbf{N}^\dagger = \mathbf{V}_2\mathbf{\Gamma}_2^{-1}\mathbf{U}_2^T$ and $\mathbf{N}\mathbf{N}^\dagger = \mathbf{U}_2\mathbf{U}_2^T$. It follows from range($\mathbf{M}$) $\subseteq$ range($\mathbf{N}$) that range($\mathbf{U}_1$) $\subseteq$ range($\mathbf{U}_2$). This implies that $\mathbf{U}_1$ can be expressed as $\mathbf{U}_1 = \mathbf{U}_2\mathbf{Q}$ where $\mathbf{Q}$ is some matrix of appropriate order. As a result, we have

$$\mathbf{N}\mathbf{N}^\dagger\mathbf{M} = \mathbf{U}_2\mathbf{U}_2^T\mathbf{U}_2\mathbf{Q}\mathbf{\Gamma}_1\mathbf{V}_1^T = \mathbf{M}.$$

It is worth noting that the condition $\mathbf{N}\mathbf{N}^\dagger\mathbf{M} = \mathbf{M}$ is not only necessary but also sufficient for range($\mathbf{M}$) $\subseteq$ range($\mathbf{N}$).

If $(\mathbf{\Lambda}, \mathbf{X})$ are the eigenpairs of $\mathbf{N}^\dagger\mathbf{M}$, then it is easily seen that $(\mathbf{\Lambda}, \mathbf{X})$ are also the eigenpairs of $(\mathbf{M}, \mathbf{N})$ due to $\mathbf{N}\mathbf{N}^\dagger\mathbf{M} = \mathbf{M}$.

Conversely, suppose $(\mathbf{\Lambda}, \mathbf{X})$ are the eigenpairs of $(\mathbf{M}, \mathbf{N})$. Then we have $\mathbf{N}\mathbf{N}^\dagger\mathbf{M}\mathbf{X} = \mathbf{N}\mathbf{X}\mathbf{\Lambda}$. This implies that $(\mathbf{\Lambda}, \mathbf{N}^\dagger\mathbf{N}\mathbf{X})$ are the eigenpairs of $\mathbf{N}^\dagger\mathbf{M}$ due to $\mathbf{N}\mathbf{N}^\dagger\mathbf{M} = \mathbf{M}$ and $\mathbf{N}^\dagger\mathbf{N}\mathbf{N}^\dagger = \mathbf{N}^\dagger$. □

Fisher discriminant analysis (FDA) is a classical method for classification and dimension reduction simultaneously [Mardia et al., 1979]. It is essentially a generalized eigenvalue problem in which the matrices $\mathbf{N}$ and $\mathbf{M}$ correspond to a pooled scatter matrix and a between-class scatter matrix [Ye and Xiong, 2006, Zhang et al., 2010]. Moreover, the condition range($\mathbf{M}$) $\subseteq$ range($\mathbf{N}$) meets. Thus, Theorem 4.2 provides a solution when the pooled scatter matrix is singular or nearly singular. We will present more details about FDA in Section 4.3.

4.2 The Procrustes Problem

Assume that $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{Y} \in \mathbb{R}^{n\times p}$ are two configurations of n data points. The orthogonal Procrustes analysis aims to move $\mathbf{Y}$ relative into $\mathbf{X}$ through rotation [Gower and Dijksterhuis, 2004].

In particular, the Procrustes problem is defined as

$$\min_{\mathbf{Q}\in\mathbb{R}^{p\times p}} \|\mathbf{X} - \mathbf{Y}\mathbf{Q}\|_F^2 \ \text{ s.t. } \ \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_p. \tag{4.3}$$

Theorem 4.3. Let the full SVD of $\mathbf{Y}^T\mathbf{X}$ be $\mathbf{Y}^T\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Then $\mathbf{U}\mathbf{V}^T$ is the minimizer of the Procrustes problem in (4.3).

Proof. Since $\|\mathbf{X} - \mathbf{YQ}\|_F^2 = \mathrm{tr}((\mathbf{X} - \mathbf{YQ})^T(\mathbf{X} - \mathbf{YQ})) = \mathrm{tr}(\mathbf{X}^T\mathbf{X}) + \mathrm{tr}(\mathbf{Y}^T\mathbf{Y}) - 2\mathrm{tr}(\mathbf{Y}^T\mathbf{X}\mathbf{Q}^T)$, the original problem is equivalent to

$$\max\ \mathrm{tr}(\mathbf{Y}^T\mathbf{X}\mathbf{Q}^T) \ \text{ s.t. } \ \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_p.$$

Recall that the constants $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_p$ are equivalent to that $\mathbf{q}_i^T\mathbf{q}_i = 1$ for $i = 1, \ldots, p$, and $\mathbf{q}_i^T\mathbf{q}_j = 0$ for $i \neq j$. Here the $\mathbf{q}_i$ are the columns of $\mathbf{Q}$. Thus, the Lagrangian function is

$$\mathrm{tr}(\mathbf{Y}^T\mathbf{X}\mathbf{Q}^T) - \frac{1}{2}\sum_{i=1}^{p} c_{ii}(\mathbf{q}_i^T\mathbf{q}_i - 1) - \frac{1}{2}\sum_{i>j} c_{ij}(\mathbf{q}_i^T\mathbf{q}_j - 0),$$

which is written in matrix form as

$$L(\mathbf{Q}, \mathbf{C}) = \mathrm{tr}(\mathbf{Y}^T\mathbf{X}\mathbf{Q}^T) - \frac{1}{2}\mathrm{tr}[\mathbf{C}(\mathbf{Q}^T\mathbf{Q} - \mathbf{I}_p)],$$

where $\mathbf{C} = [c_{ij}]$ is a symmetric matrix of the Lagrangian multipliers.

Since

$$dL = \mathrm{tr}(\mathbf{Y}^T\mathbf{X}d\mathbf{Q}^T) - \frac{1}{2}\mathrm{tr}(\mathbf{C}(d\mathbf{Q}^T\mathbf{Q} + \mathbf{Q}^Td\mathbf{Q})),$$

we have $\frac{dL}{d\mathbf{Q}} = \mathbf{Y}^T\mathbf{X} - \mathbf{QC}$. Letting the first-order derivative be zero yields

$$\mathbf{Y}^T\mathbf{X} - \mathbf{QC} = \mathbf{0}.$$

Let $\hat{\mathbf{Q}} = \mathbf{U}\mathbf{V}^T$ and $\hat{\mathbf{C}} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T$, which are obviously the solutions of the above equation systems.

The Hessian matrix of L w.r.t. $\mathbf{Q}$ at $\mathbf{Q} = \hat{\mathbf{Q}}$ and $\mathbf{C} = \hat{\mathbf{C}}$ is $-(\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T) \otimes \mathbf{I}_p$, which is negative definite. Thus, $\mathbf{Q} = \mathbf{U}\mathbf{V}^T$ is the minimizer of the Procrustes problem. □

4.3 Subspace Methods: PCA, MDS, FDA, and CCA

Subspace methods, such as principal component analysis (PCA), multidimensional scaling (MDS), Fisher discriminant analysis (FDA), and canonical correlation analysis (CCA), are a class of important machine learning methods. SVD plays a fundamental role in subspace learning methods.

PCA [Jolliffe, 2002, Kittler and Young, 1973] and MDS [Cox and Cox, 2000] are two classical dimension reduction methods. Let $\mathbf{A} =$

$[\mathbf{a}_1, \ldots, \mathbf{a}_n]^T$ be a given data matrix in which each vector $\mathbf{a}_i$ represents a data instance in $\mathbb{R}^p$. Let $\mathbf{m} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{a}_i = \frac{1}{n}\mathbf{A}^T\mathbf{1}_n$ be the sample mean and $\mathbf{C}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{I}_n^T$ be a so-called centered matrix. The pooled scatter matrix is defined as (a multiplier $1/n$ omitted)

$$\mathbf{S} = \sum_{i=1}^{n}(\mathbf{a}_i - \mathbf{m})(\mathbf{a}_i - \mathbf{m})^T = \mathbf{A}^T\mathbf{C}_n\mathbf{C}_n\mathbf{A} = \mathbf{A}^T\mathbf{C}_n\mathbf{A}.$$

It is well known that PCA computes the spectral decomposition of $\mathbf{S}$, while the classical MDS or principal coordinate analysis (PCO) computes the spectral decomposition of the Gram matrix $\mathbf{C}_n\mathbf{A}\mathbf{A}^T\mathbf{C}_n$. Proposition 3.1-(5)-(6) show that it is equivalent to computing SVD directly on the centerized data matrix $\mathbf{C}_n\mathbf{A}$. Thus, SVD bridges PCA and PCO. That is, there is a duality relationship between PCA and PCO [Mardia et al., 1979]. This relationship has found usefulness in latent semantic analysis, face classification, and microarray data analysis [Deerwester et al., 1990, Turk and Pentland, 1991, Golub et al., 1999, Belhumeur et al., 1997, Muller et al., 2004].

FDA is a joint approach for dimension reduction and classification. Assume that the $\mathbf{a}_i$ are to be grouped into c disjoint classes and that each $\mathbf{a}_i$ belongs to one and only one class. Let $V = \{1, 2, \ldots, n\}$ denote the index set of the data points $\mathbf{a}_i$ and partition V into c disjoint subsets V_j; that is, $V_i \cap V_j = \varnothing$ for $i \neq j$ and $\cup_{j=1}^{c} V_j = V$, where the cardinality of V_j is n_j so that $\sum_{j=1}^{c} n_j = n$. We also make use of a matrix representation for the partitions. In particular, we let $\mathbf{E} = [e_{ij}]$ be an $n \times c$ indicator matrix with $e_{ij} = 1$ if input $\mathbf{a}_i$ is in class j and $e_{ij} = 0$ otherwise.

Let $\mathbf{m}_j = \frac{1}{n_j}\sum_{i \in V_j}\mathbf{a}_i$ be the jth class mean for $j = 1, \ldots, c$. The between-class scatter matrix is defined as $\mathbf{S}_b = \sum_{j=1}^{c} n_j(\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T$. Conventional FDA solves the following generalized eigenproblem:

$$\mathbf{S}_b\mathbf{x}_j = \lambda_j\mathbf{S}\mathbf{x}_j, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q > \lambda_{q+1} = 0,$$

where $q \leq \min\{p,\ c{-}1\}$ and where we refer to $\mathbf{x}_j$ as the jth discriminant direction. The above generalized eigenproblem can can be expressed in matrix form:

$$\mathbf{S}_b\mathbf{X} = \mathbf{S}\mathbf{X}\boldsymbol{\Lambda}, \tag{4.4}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_q]$ $(n \times q)$ and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$ $(q \times q)$.

Let $\mathbf{\Pi} = \mathrm{diag}(n_1, \ldots, n_c)$. Then $\mathbf{S}_b$ can be rewritten as

$$\mathbf{S}_b = \mathbf{A}^T \mathbf{C}_n \mathbf{E} \mathbf{\Pi}^{-1} \mathbf{E}^T \mathbf{C}_n \mathbf{A}.$$

Recall that $\mathbf{S} = \mathbf{A}^T \mathbf{C}_n \mathbf{C}_n \mathbf{A}$. Given these representations of $\mathbf{S}$ and $\mathbf{S}_b$, the problem in (4.4) can be solved by using the GSVD method [Loan, 1976, Paige and Saunders, 1981, Golub and Van Loan, 2012, Howland et al., 2003]. Moreover, it is obvious that $\mathrm{range}(\mathbf{S}_b) \subseteq \mathrm{range}(\mathbf{A}^T \mathbf{C}_n) = \mathrm{range}(\mathbf{S})$. Thus, Theorem 4.2 provides a solution when $\mathbf{S}$ is singular or nearly singular. Moreover, the method given in Theorem 3.10 is appropriate for solving the FDA problem.

CCA is another subspace learning model [Hardoon et al., 2004]. The primary focus is on the relationship between two groups of variables (or features), whereas PCA considers interrelationships within a set of variable. Mathematically, CCA is defined as a generalized eigenvalue problem, so its solution can be borrowed from that of FDA.

4.3.1 Nonlinear Extensions

Reproducing kernel theory [Aronszajn, 1950] provides an approach for nonlinear extensions of subspace methods. For example, kernel PCA [Schölkopf et al., 1998], kernel FDA [Baudat and Anouar, 2000, Mika et al., 2000, Roth and Steinhage, 2000], kernel CCA [Akaho, 2001, Van Gestel et al., 2001, Bach and Jordan, 2002] have been successively proposed and received wide applications in data analysis.

Kernel methods work in a feature space $\mathcal{F}$, which is related to the original input space $\mathcal{X} \subset \mathbb{R}^p$ by a mapping,

$$\boldsymbol{\varphi} : \mathcal{X} \to \mathcal{F}.$$

That is, $\boldsymbol{\varphi}$ is a vector-valued function which gives a vector $\boldsymbol{\varphi}(\mathbf{a})$, called a *feature vector*, corresponding to an input $\mathbf{a} \in \mathcal{X}$. In kernel methods, we are given a reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $K(\mathbf{a}, \mathbf{b}) = \boldsymbol{\varphi}(\mathbf{a})^T \boldsymbol{\varphi}(\mathbf{b})$ for $\mathbf{a}, \mathbf{b} \in \mathcal{X}$. The mapping $\boldsymbol{\varphi}(\cdot)$ itself is typically not given explicitly. Rather, there exist only inner products between feature vectors in $\mathcal{F}$. In order to implement a kernel method without referring to $\boldsymbol{\varphi}(\cdot)$ explicitly, one resorts to the so-called *kernel trick* [Schölkopf and Smola, 2002, Shawe-Taylor and Cristianini, 2004].

Let $L_2(\mathcal{X})$ be the square integrable Hilbert space of functions whose elements are functions defined on $\mathcal{X}$. It is a well-known result that if K is a reproducing kernel for the Hilbert space $L_2(\mathcal{X})$, then $\{K(\cdot, \mathbf{b})\}$ spans $L_2(\mathcal{X})$. Here $K(\cdot, \mathbf{b})$ represents a function that is defined on $\mathcal{X}$ with values at $\mathbf{a} \in \mathcal{X}$ equal to $K(\mathbf{a}, \mathbf{b})$. There are some common kernel functions:

(a) Linear kernel: $K(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T \mathbf{b}$,

(b) Gaussian kernel or radial basis function (RBF): $K(\mathbf{a}, \mathbf{b}) = \exp\big(-\sum_{j=1}^{p} \frac{(a_j - b_j)^2}{\beta_j}\big)$ with $\beta_j > 0$,

(c) Laplacian kernel: $K(\mathbf{a}, \mathbf{b}) = \exp\big(-\sum_{j=1}^{p} \frac{|a_j - b_j|}{\beta_j}\big)$ with $\beta_j > 0$,

(d) Polynomial kernel: $K(\mathbf{a}, \mathbf{b}) = (\mathbf{a}^T \mathbf{b} + 1)^d$ of degree d.

Given a training set of input vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$, the kernel matrix $\mathbf{K} = [K(\mathbf{a}_i, \mathbf{a}_j)]$ is an $n \times n$ SPSD matrix.

5

The QR and CUR Decompositions

The QR factorization and CUR decomposition are the two most important counterparts of SVD. These three factorizations apply to all matrices. In Table 1.1 we have compared their primary focuses. The SVD and QR factorization are two classical matrix theories. The CUR decomposition aims to represent a data matrix in terms of a small number part of the matrix, which makes it easy for us to understand and interpret the data in question. Here we present very brief introductions to the QR factorization and CUR decomposition.

5.1 The QR Factorization

The QR factorization is another decomposition method applicable all matrices. Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the QR factorization is given by

$$\mathbf{A} = \mathbf{Q}\mathbf{R},$$

where $\mathbf{Q} \in \mathbb{R}^{m \times m}$ is orthonormal and $\mathbf{R} \in \mathbb{R}^{m \times n}$ is upper triangular (or low triangular). Let $\mathbf{D}$ be an $m \times m$ diagonal matrix whose diagonal elements are either 1 or -1. Then $\mathbf{A} = (\mathbf{Q}\mathbf{D})(\mathbf{D}\mathbf{R})$ is still a QR factorization of $\mathbf{A}$. Thus, we always assume that $\mathbf{R}$ has nonnegative diagonal elements.

Assume $m \geq n$. The matrix $\mathbf{A}$ also has a thin QR factorization:

$$\mathbf{A} = \mathbf{QR},$$

where $\mathbf{Q} \in \mathbb{R}^{m \times n}$ is currently column orthonormal, and $\mathbf{R} \in \mathbb{R}^{n \times n}$ is upper triangular with nonnegative diagonal elements. If $\mathbf{A}$ is of rank n, $\mathbf{R}$ is uniquely determined. In this case, $\mathbf{Q} = \mathbf{AR}^{-1}$ is also uniquely determined.

Asume $\mathbf{A}$ has rank r ($\leq \min\{m, n\}$). Then there exists an $m \times m$ orthonormal matrix $\mathbf{Q}$ and an $n \times n$ permutation matrix $\mathbf{P}$ such that

$$\mathbf{Q}^T\mathbf{AP} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{R}_{11}$ is an $r \times r$ upper triangular matrix with positive diagonal elements. This is called a *rank revealing QR factorization*.

Computation of the QR factorization can be arranged by the novel Gram-Schmidt orthogonalization process or the modified Gram-Schmidt which is numerically more stable [Trefethen and Bau III, 1997]. Additionally, Gu and Eisenstat [1996] proposed efficient algorithms for computing a rank-revealing QR factorization [Hong and Pan, 1992]. Stewart [1999] devised efficient computational algorithms of truncated pivoted QR approximations to a sparse matrix.

5.2 The CUR Decomposition

As we have see, SVD leads us to a geometrical representation, and the QR factorization facilitates computations. They have little concrete meaning. This makes it difficult for us to understand and interpret the data in question.

Kuruvilla et al. [2002] have claimed: “it would be interesting to try to find basis vectors for all experiment vectors, using actual experiment vectors and not artificial bases that offer little insight.” Therefore, it is of great interest to represent a data matrix in terms of a small number of actual columns and/or actual rows of the matrix. Matrix column selection and CUR matrix decomposition provide such techniques.

Column selection yields a so-called CX decomposition, and the CUR decomposition can be be regarded as a special CX decomposition. The

CUR decomposition problem has been widely discussed in the literature [Goreinov et al., 1997a,b, Stewart, 1999, Tyrtyshnikov, 2000, Berry et al., 2005, Drineas and Mahoney, 2005, Bien et al., 2010], and it has been shown to be very useful in high dimensional data analysis.

The CUR was originally called a skeleton decomposition [Goreinov et al., 1997a]. Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a given matrix of rank r. Then there exists a nonsingular $r \times r$ submatrix in $\mathbf{A}$. Without loss of generality, assume this nonsingular matrix is the first $r \times r$ principal submatrix of $\mathbf{A}$. That is, $\mathbf{A}$ can be partioned into the following form:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

where $\mathbf{A}_{11}$ is a $r \times r$ nonsingular matrix. Consider that $[\mathbf{A}_{21}, \mathbf{A}_{22}] = \mathbf{B}[\mathbf{A}_{11}, \mathbf{A}_{12}]$ for some $\mathbf{B} \in \mathbb{R}^{(m-r)\times r}$. It follows from $\mathbf{A}_{21} = \mathbf{B}\mathbf{A}_{11}$ that $\mathbf{B} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}$. Hence, $\mathbf{A}_{22} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$. So it is obtained that

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} \\ \mathbf{A}_{21} \end{bmatrix} \mathbf{A}_{11}^{-1}[\mathbf{A}_{11}, \mathbf{A}_{12}].$$

In general case, let $\mathbf{A}_{I,J}$ be the nonsingular submatrix where $I = \{i_1, \ldots, i_r\} \subset [m]$ and $J = \{j_1, \ldots, j_r\} \subset [n]$. Then it also hods that

$$\mathbf{A} = \mathbf{C}\mathbf{A}_{I,J}^{-1}\mathbf{R},$$

where $\mathbf{C} = \mathbf{A}_{:,J}$ and $\mathbf{R} = \mathbf{A}_{I,:}$ are respectively a subset of columns and a subset of rows, of $\mathbf{A}$.

In practical applications, however, it is intractable to select $\mathbf{A}_{I,J}$. Alternatively, Stewart [1999] proposed a quasi Gram-Schmidt algorithm, obtaining a sparse column-row (SCA) approximation of the original matrix $\mathbf{A}$ [Berry et al., 2005]. The SCA approximation is of the form $\mathbf{A} \approx \mathbf{XTY}$, where $\mathbf{X}$ and $\mathbf{Y}$ consist of columns and rows of $\mathbf{A}$, and $\mathbf{T}$ minimizes $\|\mathbf{A} - \mathbf{XTY}\|_F^2$. This algorithm is a deterministic peocedure but computationally expensive.

The terminology of the CUR decomposition has been proposed by Drineas and Mahoney [2005], Mahoney et al. [2008]. They reformulated the idea based on random selection. A CUR decomposition algorithm seeks to find a subset of c columns of $\mathbf{A}$ to form a matrix $\mathbf{C} \in \mathbb{R}^{m \times c}$, a subset of r rows to form a matrix $\mathbf{R} \in \mathbb{R}^{r \times n}$, and an intersection matrix $\mathbf{U} \in \mathbb{R}^{c \times r}$ such that $\|\mathbf{A} - \mathbf{CUR}\|_\xi$ is small. Accordingly, $\tilde{\mathbf{A}} = \mathbf{CUR}$ is used to approximate $\mathbf{A}$.

Since there are $\binom{n}{c}$ possible choices of constructing $\mathbf{C}$ and $\binom{m}{r}$ possible choices of constructing $\mathbf{R}$, obtaining the best CUR decomposition is a hard problem. In Chapter 10 we will further study the CUR decomposition problem via random approximation.

The CUR decomposition is also an extension of the novel Nyström approximation to a general matrix. The Nyström method approximates an SPSD matrix only using a subset of its columns, so it can alleviate computation and storage costs when the SPSD matrix in question is large in size. Thus, the Nyström method and its variants [Halko et al., 2011, Gittens and Mahoney, 2013, Kumar et al., 2009, Wang and Zhang, 2013, 2014, Wang et al., 2014b, Si et al., 2014] have been extensively used in the machine learning community. For example, they have been applied to Gaussian processes [Williams and Seeger, 2001], kernel classification [Zhang et al., 2008, Jin et al., 2013], spectral clustering [Fowlkes et al., 2004], kernel PCA and manifold learning [Talwalkar et al., 2008, Zhang et al., 2008, Zhang and Kwok, 2010], determinantal processes [Affandi et al., 2013], etc.

6

Variational Principles

Variational principles correspond to matrix perturbation theory [Stewart and Sun, 1990], which is the theoretical foundation to characterize stability or sensitivity of a matrix computation algorithm. Thus, variational principles are important in analysis for error bounds of matrix approximate algorithms (see Chapters 9 and 10).

In this chapter we specifically study variational properties for eigenvalues of a symmetric matrix as well as for singular values of a general matrix. We will see that these results for eigenvalues and for singular values are almost parallel. The cornerstones are the novel von Neumann theorem [Neumann, 1937] and Ky Fan theorem [Fan, 1951]. We present new proofs for them by using theory of matrix differentials. Additionally, we present some majorization inequalities. They will be used in the latter chapters, especially in investigating unitarily invariant norms (see Chapter 7).

Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we always let $\sigma_1(\mathbf{A}) \geq \cdots \geq \sigma_p(\mathbf{A})$ be the singular values of $\mathbf{A}$ where $p = \min\{m, n\}$. When $\mathbf{A}$ is symmetric, let $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$ be the eigenvalues of $\mathbf{A}$. These eigenvalues or singular values are always arranged in deceasing order. Note that the eigenvalues are real but could be negative. Let

$\boldsymbol{\lambda}(\mathbf{M}) = (\lambda_1(\mathbf{M}), \ldots, \lambda_n(\mathbf{M}))^T$ denote the eigenvalues of an $n \times n$ real square matrix $\mathbf{M}$, and $\boldsymbol{\sigma}(\mathbf{A}) = (\sigma_1(\mathbf{A}), \ldots, \sigma_p(\mathbf{A}))^T$ denote the singular values of an $m \times n$ real matrix $\mathbf{A}$. Sometimes we also write them the σ_i or the λ_i when they are explicit in the context for notational simplicity.

6.1 Variational Properties for Eigenvalues

In this section we consider variational properties for eigenvalues of a real symmetric matrix. It is well known that for an arbitrary symmetric matrix, its eigenvalues are all real. The following cornerstone theorem was originally established by von Neumann [1937].

Theorem 6.1 (von Neumann Theorem). Assume $\mathbf{M} \in \mathbb{R}^{n\times n}$ and $\mathbf{N} \in \mathbb{R}^{n\times n}$ are symmetric. Then

$$\sum_{i=1}^{n} \lambda_i(\mathbf{M})\lambda_i(\mathbf{N}) = \max_{\mathbf{Q}\mathbf{Q}^T=\mathbf{I}_n} \operatorname{tr}(\mathbf{Q}\mathbf{M}\mathbf{Q}^T\mathbf{N}).$$

Moreover,

$$\sum_{i=i}^{n} \lambda_i(\mathbf{M})\lambda_{n-i+1}(\mathbf{N}) = \min_{\mathbf{Q}\mathbf{Q}^T=\mathbf{I}_n} \operatorname{tr}(\mathbf{Q}\mathbf{M}\mathbf{Q}^T\mathbf{N}).$$

Proof. The second part directly follows from the first part because

$$\min_{\mathbf{Q}\mathbf{Q}^T=\mathbf{I}_n} \operatorname{tr}(\mathbf{Q}\mathbf{M}\mathbf{Q}^T\mathbf{N}) = -\max_{\mathbf{Q}\mathbf{Q}^T=\mathbf{I}_n} \operatorname{tr}(\mathbf{Q}\mathbf{M}\mathbf{Q}^T(-\mathbf{N})).$$

We now present the proof of the first part. Make full EVDs of $\mathbf{M}$ and $\mathbf{N}$ as $\mathbf{M} = \mathbf{U}_M\boldsymbol{\Lambda}_M\mathbf{U}_M^T$ and $\mathbf{N} = \mathbf{U}_N\boldsymbol{\Lambda}_N\mathbf{U}_N^T$, where $\boldsymbol{\Lambda}_M = \operatorname{diag}(\lambda_1(\mathbf{M}), \ldots, \lambda_n(\mathbf{M}))$ and $\boldsymbol{\Lambda}_N = \operatorname{diag}(\lambda_1(\mathbf{N}), \ldots, \lambda_n(\mathbf{N}))$, and $\mathbf{U}_M$ and $\mathbf{U}_N$ are orthonormal. It is easily seen that

$$\begin{aligned}\max_{\mathbf{Q}\mathbf{Q}^T=\mathbf{I}_n} \operatorname{tr}(\mathbf{Q}\mathbf{M}\mathbf{Q}^T\mathbf{N}) &= \max_{\mathbf{Q}\mathbf{Q}^T=\mathbf{I}_n} \operatorname{tr}((\mathbf{U}_N^T\mathbf{Q}\mathbf{U}_M)\boldsymbol{\Lambda}_M(\mathbf{U}_N^T\mathbf{Q}\mathbf{U}_M)^T\boldsymbol{\Lambda}_N)\\ &= \max_{\mathbf{Q}\mathbf{Q}^T=\mathbf{I}_n} \operatorname{tr}(\mathbf{Q}\boldsymbol{\Lambda}_M\mathbf{Q}^T\boldsymbol{\Lambda}_N).\end{aligned}$$

Let $\mathbf{Q} = [q_{ij}] = [\mathbf{q}_1, \ldots, \mathbf{q}_n]^T$. We now have

$$\begin{aligned}
&\mathrm{tr}(\mathbf{Q}\mathbf{\Lambda}_M\mathbf{Q}^T\mathbf{\Lambda}_N) \\
&= \sum_{i=1}^{n} \mathbf{q}_i^T \mathbf{\Lambda}_M \mathbf{q}_i \lambda_i(\mathbf{N}) \\
&= \sum_{i=1}^{n-1}\sum_{j=1}^{i} \mathbf{q}_j^T \mathbf{\Lambda}_M \mathbf{q}_j [\lambda_i(\mathbf{N}) - \lambda_{i+1}(\mathbf{N})] + \lambda_n(\mathbf{N}) \sum_{j=1}^{n} \mathbf{q}_j^T \mathbf{\Lambda}_M \mathbf{q}_j \\
&= \sum_{i=1}^{n-1} [\lambda_i(\mathbf{N}) - \lambda_{i+1}(\mathbf{N})] \sum_{j=1}^{i}\sum_{k=1}^{n} q_{jk}^2 \lambda_k(\mathbf{M}) + \lambda_n(\mathbf{N}) \sum_{j=1}^{n} \lambda_j(\mathbf{M}).
\end{aligned}$$

Define $\mathbf{W} \triangleq [q_{ij}^2]$ which is doubly stochastic, and $\mathbf{u} = [u_1, \ldots, u_n]^T$ where $u_j = \sum_{k=1}^n q_{jk}^2 \lambda_k(\mathbf{M})$. That is, $\mathbf{u} = \mathbf{W}\boldsymbol{\lambda}(\mathbf{M})$. By Lemma 2.2, we know that $\mathbf{u} \prec \boldsymbol{\lambda}(\mathbf{M})$. Accordingly,

$$\begin{aligned}
\mathrm{tr}(\mathbf{Q}\mathbf{\Lambda}_M\mathbf{Q}^T\mathbf{\Lambda}_N) &\leq \sum_{i=1}^{n-1} [\lambda_i(\mathbf{N}) - \lambda_{i+1}(\mathbf{N})] \sum_{j=1}^{i} \lambda_j(\mathbf{M}) + \lambda_n(\mathbf{N}) \sum_{j=1}^{n} \lambda_j(\mathbf{M}) \\
&= \sum_{i=1}^{n} \lambda_i(\mathbf{M})\lambda_i(\mathbf{N}).
\end{aligned}$$

When $\mathbf{Q} = \mathbf{I}_n$, the equality holds. That is, $\mathbf{U}_N^T\mathbf{Q}\mathbf{U}_M = \mathbf{I}_n$ in the original problem. The theorem follows. □

The following theorem is a corollary of Theorem 6.1 when taking

$$\mathbf{N} = \begin{bmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Theorem 6.2 (von Neumann Theorem). Assume $\mathbf{M} \in \mathbb{R}^{n\times n}$ is symmetric. Then for $k \in [n]$,

$$\sum_{i=1}^{k} \lambda_i = \max_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_k} \mathrm{tr}(\mathbf{Q}^T\mathbf{M}\mathbf{Q}),$$

which is arrived when $\mathbf{Q}$ is the $n \times k$ matrix of the orthonormal vectors associated with $\lambda_1, \ldots, \lambda_k$. Moreover,

$$\sum_{i=n-k+1}^{n} \lambda_i = \min_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_k} \mathrm{tr}(\mathbf{Q}^T\mathbf{M}\mathbf{Q}).$$

In the appendix we give an other proof based on theory of matrix differentials. The von Neumann theorem describes the variational principle of eigenvalues of a symmetric matrix. Using Theorems 6.2, we have the following variational properties.

Proposition 6.1. Given two $n \times n$ real symmetric matrices $\mathbf{M}$ and $\mathbf{N}$, we have that

(1) $\boldsymbol{\lambda}(\mathbf{M}+\mathbf{N}) \prec \boldsymbol{\lambda}(\mathbf{M}) + \boldsymbol{\lambda}(\mathbf{N})$ and $\boldsymbol{\lambda}(\mathbf{M}) - \boldsymbol{\lambda}(\mathbf{N}) \prec \boldsymbol{\lambda}(\mathbf{M}-\mathbf{N})$.

(2) $\sum_{i=1}^{k} \lambda_i(\mathbf{M}+\mathbf{N}) \geq \sum_{i=1}^{k} \lambda_i(\mathbf{M}) + \sum_{j=n-k+1}^{n} \lambda_j(\mathbf{N})$ for $k \in [n]$.

(3) $(m_{11}, \ldots, m_{nn}) \prec (\lambda_1(\mathbf{M}), \ldots, \lambda_n(\mathbf{M}))$.

Proof. The proof is based on Theorem 6.2. First, for $k \in [n-1]$,

$$\begin{aligned}\sum_{i=1}^{k} \lambda_i(\mathbf{M}+\mathbf{N}) &= \max_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_k} \Big\{ \mathrm{tr}(\mathbf{Q}^T\mathbf{M}\mathbf{Q}) + \mathrm{tr}(\mathbf{Q}^T\mathbf{N}\mathbf{Q}) \Big\} \\ &\leq \max_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_k} \mathrm{tr}(\mathbf{Q}^T\mathbf{M}\mathbf{Q}) + \max_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_k} \mathrm{tr}(\mathbf{Q}^T\mathbf{N}\mathbf{Q}) \\ &= \sum_{i=1}^{k} \lambda_i(\mathbf{M}) + \sum_{i=1}^{k} \lambda_i(\mathbf{N}).\end{aligned}$$

Note that $\mathrm{tr}(\mathbf{M}+\mathbf{N}) = \mathrm{tr}(\mathbf{M}) + \mathrm{tr}(\mathbf{N})$, so $\boldsymbol{\lambda}(\mathbf{M}+\mathbf{N}) \prec \boldsymbol{\lambda}(\mathbf{M}) + \boldsymbol{\lambda}(\mathbf{N})$. Hence, $\boldsymbol{\lambda}(\mathbf{M}) - \boldsymbol{\lambda}(\mathbf{N}) \prec \boldsymbol{\lambda}(\mathbf{M}-\mathbf{N})$. Second,

$$\begin{aligned}\sum_{i=1}^{k} \lambda_i(\mathbf{M}+\mathbf{N}) &= \max_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_k} \Big\{ \mathrm{tr}(\mathbf{Q}^T\mathbf{M}\mathbf{Q}) + \mathrm{tr}(\mathbf{Q}^T\mathbf{N}\mathbf{Q}) \Big\} \\ &\geq \max_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_k} \Big\{ \mathrm{tr}(\mathbf{Q}^T\mathbf{M}\mathbf{Q}) + \min_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_k} \mathrm{tr}(\mathbf{Q}^T\mathbf{N}\mathbf{Q}) \Big\} \\ &= \sum_{i=1}^{k} \lambda_i(\mathbf{M}) + \sum_{j=n-k+1}^{n} \lambda_j(\mathbf{N}).\end{aligned}$$

To prove the third part, we assume that $m_{11} \geq \cdots \geq m_{nn}$ without loss of generality. Now the result is obtained via

$$\sum_{i=1}^{k} \lambda_i(\mathbf{M}) = \max_{\mathbf{Q}^T\mathbf{Q}=\mathbf{I}_k} \mathrm{tr}(\mathbf{Q}^T\mathbf{M}\mathbf{Q}) \geq \mathrm{tr}(\mathbf{H}_k^T\mathbf{M}\mathbf{H}_k) = \sum_{i=1}^{k} m_{ii},$$

where $\mathbf{H}_k$ consists of the first k columns of $\mathbf{I}_n$ for all $k \in [n]$. □

Proposition 6.1-(3) is sometimes referred to as Schur's theorem. The second part of the following proposition is an extension of Schur's theorem.

Proposition 6.2. Let $\mathbf{M} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix}$ be $n \times n$ real symmetric. Here $\mathbf{M}_{11}$ is $k \times k$. Then

$$(1) \quad \lambda_i(\mathbf{M}) \geq \lambda_i(\mathbf{M}_{11}) \geq \lambda_{n-k+i}(\mathbf{M}) \text{ for } \; i = 1, \ldots, k;$$

and (2) $\quad (\boldsymbol{\lambda}(\mathbf{M}_{11}), \boldsymbol{\lambda}(\mathbf{M}_{22})) \prec \boldsymbol{\lambda}(\mathbf{M})$.

Furthermore, for any column-orthonormal matrix $\mathbf{Q} \in \mathbb{R}^{n \times k}$, we have

$$(3) \quad \lambda_i(\mathbf{M}) \geq \lambda_i(\mathbf{Q}^T\mathbf{M}\mathbf{Q}) \geq \lambda_{n-k+i}(\mathbf{M}) \;\text{ for }\; i = 1, \ldots, k.$$

Proof. The first result directly follows from the well known interlacing theorem [Horn and Johnson, 1985]. As for the third part, we can extend $\mathbf{Q}$ to an orthonormal matrix $\tilde{\mathbf{Q}} = [\mathbf{Q}, \mathbf{Q}^{\perp}]$. Consider that

$$\tilde{\mathbf{Q}}^T\mathbf{M}\tilde{\mathbf{Q}} = \begin{bmatrix} \mathbf{Q}^T\mathbf{M}\mathbf{Q} & \mathbf{Q}^T\mathbf{M}\mathbf{Q}^{\perp} \\ (\mathbf{Q}^{\perp})^T\mathbf{M}\mathbf{Q} & (\mathbf{Q}^{\perp})^T\mathbf{M}\mathbf{Q}^{\perp} \end{bmatrix}.$$

Thus,

$$\lambda_i(\mathbf{M}) = \lambda_i(\tilde{\mathbf{Q}}^T\mathbf{M}\tilde{\mathbf{Q}}) \geq \lambda_i(\mathbf{Q}^T\mathbf{M}\mathbf{Q}) \geq \lambda_{n-k+i}(\tilde{\mathbf{Q}}^T\mathbf{M}\tilde{\mathbf{Q}}) = \lambda_{n-k+i}(\mathbf{M}).$$

We now consider the proof of the second part. Let the EVDs of $\mathbf{M}_{11}$ and $\mathbf{M}_{22}$ be $\mathbf{M}_{11} = \mathbf{U}_1\boldsymbol{\Lambda}_1\mathbf{U}_1^T$ and $\mathbf{M}_{22} = \mathbf{U}_2\boldsymbol{\Lambda}_2\mathbf{U}_2^T$. Then

$$\begin{bmatrix} \mathbf{U}_1^T & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_2^T \end{bmatrix} \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{U}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}_1 & \mathbf{U}_1^T\mathbf{M}_{12}\mathbf{U}_2 \\ \mathbf{U}_2^T\mathbf{M}_{21}\mathbf{U}_1 & \boldsymbol{\Lambda}_2 \end{bmatrix}.$$

Since $\mathbf{U}_1$ and $\mathbf{U}_2$ are orthonormal, we have that $\boldsymbol{\lambda}(\mathbf{M}_{11}) = \boldsymbol{\lambda}(\boldsymbol{\Lambda}_1)$, $\boldsymbol{\lambda}(\mathbf{M}_{22}) = \boldsymbol{\lambda}(\boldsymbol{\Lambda}_2)$, and

$$\boldsymbol{\lambda}\left(\begin{bmatrix} \mathbf{U}_1^T & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_2^T \end{bmatrix} \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{U}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_2 \end{bmatrix}\right) = \boldsymbol{\lambda}(\mathbf{M}).$$

Applying Proposition 6.1-(3) completes the proof. □

6.2 Variational Properties for Singular Values

Theorems 6.1 and 6.2 can be extended to a general matrix. In this case, we investigate singular values of the matrix instead. Theorems 6.3 and 6.4 correspond to Theorems 6.1 and 6.2, respectively.

Theorem 6.3 (Ky Fan Theorem)**.** Given two matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{m\times n}$, let $\mathbf{A}$ and $\mathbf{B}$ have full SVDs $\mathbf{A} = \mathbf{U}_A\mathbf{\Sigma}_A\mathbf{V}_A^T$ and $\mathbf{B} = \mathbf{U}_B\mathbf{\Sigma}_B\mathbf{V}_B^T$, respectively. Let $p = \min\{m, n\}$. Then

$$\begin{aligned}\sum_{i=1}^{p} \sigma_i(\mathbf{A})\sigma_i(\mathbf{B}) &= \max_{\mathbf{X}^T\mathbf{X}=\mathbf{I}_m, \mathbf{Y}^T\mathbf{Y}=\mathbf{I}_n} |\mathrm{tr}(\mathbf{X}^T\mathbf{A}\mathbf{Y}\mathbf{B}^T)| \\ &= \max_{\mathbf{X}^T\mathbf{X}=\mathbf{I}_m, \mathbf{Y}^T\mathbf{Y}=\mathbf{I}_n} \mathrm{tr}(\mathbf{X}^T\mathbf{A}\mathbf{Y}\mathbf{B}^T),\end{aligned}$$

which is achieved at $\mathbf{X} = \mathbf{U}_A\mathbf{U}_B^T$ and $\mathbf{Y} = \mathbf{V}_A\mathbf{V}_B^T$.

Proof. Note that

$$\mathrm{tr}(\mathbf{X}^T\mathbf{A}\mathbf{Y}\mathbf{B}^T) = \frac{1}{2}\mathrm{tr}\left(\begin{bmatrix}\mathbf{Y}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{X}^T\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{Y} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0}\end{bmatrix}\right).$$

The theorem is directly obtained from Theorems 6.1 and 3.5. □

Theorem 6.4 (Ky Fan Theorem)**.** Given an $m \times n$ real matrix $\mathbf{A}$, let $p = \min\{m, n\}$, and let the singular values of $\mathbf{A}$ be $\sigma_1, \ldots, \sigma_p$ which are arranged in descending order, with the corresponding left and right singular vectors $\mathbf{u}_i$ and $\mathbf{v}_i$. Then for any $k \in [p]$,

$$\sum_{i=1}^{k} \sigma_i = \max_{\mathbf{X}^T\mathbf{X}=\mathbf{I}_k, \mathbf{Y}^T\mathbf{Y}=\mathbf{I}_k} |\mathrm{tr}(\mathbf{X}^T\mathbf{A}\mathbf{Y})| = \max_{\mathbf{X}^T\mathbf{X}=\mathbf{I}_k, \mathbf{Y}^T\mathbf{Y}=\mathbf{I}_k} \mathrm{tr}(\mathbf{X}^T\mathbf{A}\mathbf{Y}),$$

which is achieved at $\mathbf{X} = [\mathbf{u}_1, \ldots, \mathbf{u}_k]$ and $\mathbf{Y} = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$.

The theorem can be obtained from Theorems 6.2 and 3.5 or from Theorem 6.3. In the appendix we give the third proof.

Proposition 6.3. Given two matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{m\times n}$, let $p = \min\{m, n\}$. Let $\hat{\mathbf{A}}$ be obtained by replacing the last r rows and/or columns of $\mathbf{A}$ by zeros. Then

(1) $\boldsymbol{\sigma}(\mathbf{A}+\mathbf{B}) \prec_w \boldsymbol{\sigma}(\mathbf{A}) + \boldsymbol{\sigma}(\mathbf{B})$.

(2) $\sigma_{i+j-1}(\mathbf{A}+\mathbf{B}) \leq \sigma_i(\mathbf{A}) + \sigma_j(\mathbf{B})$ for $i, j \geq 1$ and $i+j-1 \leq p$.

(3) $\mathbf{a} \prec_w \boldsymbol{\sigma}(\mathbf{A})$ where $\mathbf{a} = (a_{11}, \ldots, a_{pp})^T$.

(4) For $i \in [p-r]$, $\sigma_{r+i}(\mathbf{A}) \leq \sigma_i(\hat{\mathbf{A}}) \leq \sigma_i(\mathbf{A})$.

(5) Let $\mathbf{P} \in \mathbb{R}^{m\times r}$ and $\mathbf{Q} \in \mathbb{R}^{n\times r}$ be column orthonormal matrices where $r \leq p$. Then $\sigma_{r+i}(\mathbf{A}) \leq \sigma_i(\mathbf{P}^T\mathbf{A}) \leq \sigma_i(\mathbf{A})$ and $\sigma_{r+i}(\mathbf{A}) \leq \sigma_i(\mathbf{A}\mathbf{Q}) \leq \sigma_i(\mathbf{A})$ for $i = 1, \ldots, p-r$.

Proof. The proof of Proposition 6.3-(1) and (3) is parallel to that of Proposition 6.1-(1) and (3). Part-(2) is Weyl's monotonicity theorem. It can be proven by the Courant-Fischer theorem (see Theorem 3.4). Consider that $\sigma_i(\mathbf{A}) = \sqrt{\lambda_i(\mathbf{A}^T\mathbf{A})} = \sqrt{\lambda_i(\mathbf{A}\mathbf{A}^T)}$ and $\sigma_i(\hat{\mathbf{A}}) = \sqrt{\lambda_i(\hat{\mathbf{A}}^T\hat{\mathbf{A}})} = \sqrt{\lambda_i(\hat{\mathbf{A}}\hat{\mathbf{A}}^T)}$. Part (4) follows from Proposition 6.2-(1). Part (5) follows then from Proposition 6.2-(3). □

Theorem 6.5. Given two matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{m\times n}$, let $s_i(\mathbf{A}-\mathbf{B}) = |\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})|$ for $i \in [p]$ where $p = \min\{m, n\}$. Then

$$\sum_{i=1}^{k} s_i^{\downarrow}(\mathbf{A}-\mathbf{B}) \leq \sum_{i=1}^{k} \sigma_i(\mathbf{A}-\mathbf{B}) \quad \text{for } k = 1, \ldots, p.$$

Proof. Consider the following two $(m+n)\times(m+n)$ symmetric matrices:

$$\tilde{\mathbf{A}} = \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \tilde{\mathbf{B}} = \begin{bmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{bmatrix}.$$

By Theorem 3.5, the eigenvalues of $\tilde{\mathbf{A}}$ are $\pm\sigma_1(\mathbf{A}), \ldots, \pm\sigma_p(\mathbf{A})$, together with $m+n-2p$ zeros; and similarly for $\tilde{\mathbf{B}}$ as well as for $\tilde{\mathbf{A}}-\tilde{\mathbf{B}}$. Thus, the p largest entries of $\boldsymbol{\lambda}(\tilde{\mathbf{A}}-\tilde{\mathbf{B}})$ are $\sigma_1(\mathbf{A}-\mathbf{B}), \ldots, \sigma_p(\mathbf{A}-\mathbf{B})$. Note that both $\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})$ and $\sigma_i(\mathbf{B}) - \sigma_i(\mathbf{A})$ are the entries of $\boldsymbol{\lambda}(\tilde{\mathbf{A}}) - \boldsymbol{\lambda}(\tilde{\mathbf{B}})$, so the p largest entries of $\boldsymbol{\lambda}(\tilde{\mathbf{A}}) - \boldsymbol{\lambda}(\tilde{\mathbf{B}})$ comprise the set $\{s_1(\mathbf{A}-\mathbf{B}), \ldots, s_p(\mathbf{A}-\mathbf{B})\}$. Proposition 6.1 shows that $\boldsymbol{\lambda}(\tilde{\mathbf{A}}-\tilde{\mathbf{B}}) \prec \boldsymbol{\lambda}(\tilde{\mathbf{A}}) - \boldsymbol{\lambda}(\tilde{\mathbf{B}})$. This implies the result of the theorem. □

Theorem 6.6. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$ be given, and let $q = \min\{m, n, p\}$. Then for $k = 1, \ldots, q$,

$$\prod_{i=1}^{k} \sigma_i(\mathbf{A}\mathbf{B}) \leq \prod_{i=1}^{k} \sigma_i(\mathbf{A})\sigma_i(\mathbf{B}).$$

If $n = p = m$, then equality holds for $k = n$. And

$$\sum_{i=1}^{k} \sigma_i(\mathbf{AB}) \leq \sum_{i=1}^{k} \sigma_i(\mathbf{A})\sigma_i(\mathbf{B}) \leq \Big(\sum_{i=1}^{k} \sigma_i(\mathbf{A})\Big)\Big(\sum_{i=1}^{k} \sigma_i(\mathbf{B})\Big).$$

Proof. Let $\mathbf{AB} = \mathbf{U\Sigma V}^T$ be a full SVD of $\mathbf{AB}$, and for $k \leq q$ let $\mathbf{U}_k$ and $\mathbf{V}_k$ be the first k columns of $\mathbf{U}$ and $\mathbf{V}$, respectively. Now take a polar decomposition of $\mathbf{BV}_k$ as $\mathbf{BV}_k = \mathbf{QS}$. Since $\mathbf{S}^2 = \mathbf{V}_k^T\mathbf{B}^T\mathbf{BV}_k$ and by Proposition 6.3-(4), we obtain

$$\det(\mathbf{S}^2) = \det(\mathbf{V}_k^T\mathbf{B}^T\mathbf{BV}_k) \leq \prod_{i=1}^{k} \sigma_i^2(\mathbf{B})$$

We further have that

$$\begin{aligned}\prod_{i=1}^{k} \sigma_i(\mathbf{AB}) &= |\det(\mathbf{U}_k^T\mathbf{ABV}_k)| = |\det(\mathbf{U}_k^T\mathbf{AQ})\det(\mathbf{S})| \\ &\leq \prod_{i=1}^{k} \sigma_i(\mathbf{A})\sigma_i(\mathbf{B}).\end{aligned}$$

The above inequality again follows from Proposition 6.3-(4), When $n = p = m$, then

$$\prod_{i=1}^{n} \sigma_i(\mathbf{AB}) = |\det(\mathbf{AB})| = |\det(\mathbf{A})| \times |\det(\mathbf{B})| = \prod_{i=1}^{n} \sigma_i(\mathbf{A})\sigma_i(\mathbf{B}).$$

The second part follows from the first part and Lemma 2.4. □

6.3 Appendix: Application of Matrix Differentials

Here we present alternative proofs for Theorem 6.2 and Theorem 6.4, which are based on matrix differentials. It aims at further illustrating how to use matrix differentials.

The Second Proof of Theorem 6.2. To solve the problem, we define the Lagrangian function:

$$L(\mathbf{Q}, \mathbf{C}) = \mathrm{tr}(\mathbf{Q}^T\mathbf{MQ}) - \mathrm{tr}(\mathbf{C}(\mathbf{Q}^T\mathbf{Q} - \mathbf{I}_k)),$$

where $\mathbf{C}$ is a $k \times k$ symmetric matrix of Lagrangian multipliers. Since

$$dL = \mathrm{tr}(d\mathbf{Q}^T\mathbf{M}\mathbf{Q} + \mathbf{Q}^T\mathbf{M}d\mathbf{Q}) - \mathrm{tr}(\mathbf{C}(d\mathbf{Q}^T\mathbf{Q} + \mathbf{Q}^T d\mathbf{Q})),$$

this shows that $\frac{dL}{d\mathbf{Q}} = 2\mathbf{M}\mathbf{Q} - 2\mathbf{Q}\mathbf{C}$. The KKT condition is now

$$\mathbf{M}\mathbf{Q} - \mathbf{Q}\mathbf{C} = \mathbf{0}.$$

Clearly, if $\hat{\mathbf{C}} \triangleq \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ and $\hat{\mathbf{Q}}$ consists of the corresponding orthonormal eigenvectors, they are a solution of the above equation. In this setting, we see that $\mathrm{tr}(\hat{\mathbf{Q}}^T\mathbf{M}\hat{\mathbf{Q}}) = \sum_{i=1}^k \lambda_i$.

Thus, we only need to prove that $\hat{\mathbf{Q}}$ is indeed the maximizer of the original problem. We now compute the Hessian matrix of L w.r.t. $\mathbf{Q}$ at $\mathbf{Q} = \hat{\mathbf{Q}}$ and $\mathbf{C} = \hat{\mathbf{C}}$. Since $\mathsf{vec}(\mathbf{M}\mathbf{Q} - \mathbf{Q}\mathbf{C}) = (\mathbf{I}_k \otimes \mathbf{M} - \mathbf{C} \otimes \mathbf{I}_n)\mathsf{vec}(\mathbf{Q})$, the Hessian matrix is given as

$$\mathbf{H} = 2(\mathbf{I}_k \otimes \mathbf{M} - \hat{\mathbf{C}} \otimes \mathbf{I}_n).$$

For any $\mathbf{X} \in \mathbb{R}^{n \times k}$ such that $\mathbf{X}^T\hat{\mathbf{Q}} = \mathbf{0}$, it suffices for our purpose to prove $\mathbf{x}^T\mathbf{H}\mathbf{x}/2 \leq 0$ where $\mathbf{x} = \mathsf{vec}(\mathbf{X})$. Take the full EVD of $\mathbf{M}$ as $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $\mathbf{U} = [\hat{\mathbf{Q}}, \hat{\mathbf{Q}}^{\perp}]$ such that $\mathbf{U}^T\mathbf{U} = \mathbf{I}_n$. Denote $\mathbf{\Lambda}_2 = \mathrm{diag}(\lambda_{k+1}, \ldots, \lambda_n)$ and $\mathbf{Y} = (\hat{\mathbf{Q}}^{\perp})^T\mathbf{X} = [\mathbf{y}_1, \ldots, \mathbf{y}_k]$. Then,

$$\begin{aligned}
\frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} &= \mathrm{tr}(\mathbf{X}^T\mathbf{M}\mathbf{X}) - \mathrm{tr}(\mathbf{X}\hat{\mathbf{C}}\mathbf{X}^T) \\
&= \mathrm{tr}(\mathbf{X}^T\hat{\mathbf{Q}}^{\perp}\mathbf{\Lambda}_2(\hat{\mathbf{Q}}^{\perp})^T\mathbf{X}) - \mathrm{tr}(\hat{\mathbf{C}}\mathbf{X}^T(\hat{\mathbf{Q}}\hat{\mathbf{Q}}^T + \hat{\mathbf{Q}}^{\perp}(\hat{\mathbf{Q}}^{\perp})^T)\mathbf{X}) \\
&= \mathrm{tr}(\mathbf{Y}^T\mathbf{\Lambda}_2\mathbf{Y}) - \mathrm{tr}(\hat{\mathbf{C}}\mathbf{Y}^T\mathbf{Y}) \\
&= \sum_{i=1}^k \mathbf{y}_i^T\mathbf{\Lambda}_2\mathbf{y}_i - \sum_{i=1}^k \lambda_i\mathbf{y}_i^T\mathbf{y}_i \\
&= \sum_{i=1}^k \mathbf{y}_i^T(\mathbf{\Lambda}_2 - \lambda_i\mathbf{I}_{n-k})\mathbf{y}_i \leq 0.
\end{aligned}$$

□

The Third Proof of Theorem 6.4. To solve the constrained problem in the theorem, we now define the Lagrangian function:

$$L(\mathbf{X}, \mathbf{Y}, \mathbf{C}_1, \mathbf{C}_2) = \mathrm{tr}(\mathbf{X}^T\mathbf{A}\mathbf{Y}) - \frac{1}{2}\mathrm{tr}(\mathbf{C}_1(\mathbf{X}^T\mathbf{X} - \mathbf{I}_k)) - \frac{1}{2}\mathrm{tr}(\mathbf{C}_2(\mathbf{Y}^T\mathbf{Y} - \mathbf{I}_k)),$$

where $\mathbf{C}_1$ and $\mathbf{C}_2$ are two $k \times k$ symmetric matrix of Lagrange multipliers. Since

$$dL = \mathrm{tr}(d\mathbf{X}^T\mathbf{A}\mathbf{Y}) - \frac{1}{2}\mathrm{tr}(\mathbf{C}_1(d\mathbf{X}^T\mathbf{X} + \mathbf{X}^T d\mathbf{X})),$$
$$dL = \mathrm{tr}(\mathbf{X}^T\mathbf{A}d\mathbf{Y}) - \frac{1}{2}\mathrm{tr}(\mathbf{C}_2(d\mathbf{Y}^T\mathbf{Y} + \mathbf{Y}^T d\mathbf{Y})),$$

which yield that $\frac{dL}{d\mathbf{X}} = \mathbf{A}\mathbf{Y} - \mathbf{X}\mathbf{C}_1$ and $\frac{dL}{d\mathbf{Y}} = \mathbf{X}\mathbf{A}^T - \mathbf{Y}\mathbf{C}_2$. The KKT condition is now

$$\mathbf{A}\mathbf{Y} - \mathbf{X}\mathbf{C}_1 = \mathbf{0} \;\text{ and }\; \mathbf{A}^T\mathbf{X} - \mathbf{Y}\mathbf{C}_2 = \mathbf{0}.$$

It then follows from $\mathbf{X}^T\mathbf{X} = \mathbf{I}_k$ and $\mathbf{Y}^T\mathbf{Y} = \mathbf{I}_k$ that $\mathbf{C}_1 = \mathbf{C}_2$. We denote $\mathbf{C} \triangleq \mathbf{C}_1 = \mathbf{C}_2$. So,

$$\mathbf{A}\mathbf{Y} - \mathbf{X}\mathbf{C} = \mathbf{0},$$
$$\mathbf{A}^T\mathbf{X} - \mathbf{Y}\mathbf{C} = \mathbf{0}.$$

That is,

$$\begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \mathbf{C}.$$

Clearly, if $\hat{\mathbf{C}} \triangleq \mathbf{\Sigma}_k = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$, $\hat{\mathbf{X}} \triangleq \mathbf{U}_k = [\mathbf{u}_1, \ldots, \mathbf{u}_k]$, and $\hat{\mathbf{Y}} \triangleq \mathbf{V}_k = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$, then they are a solution of the above equation. In this setting, we see that $\mathrm{tr}(\hat{\mathbf{X}}^T\mathbf{A}\hat{\mathbf{Y}}) = \sum_{i=1}^k \sigma_i$.

Thus, we only need to prove that $(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ is the maximizer of the original problem. We now compute the Hessian matrix of L w.r.t. $(\mathbf{X}, \mathbf{Y})$ at $(\mathbf{X}, \mathbf{Y}) = (\hat{\mathbf{X}}, \hat{\mathbf{Y}})$, and $\mathbf{C} = \hat{\mathbf{C}}$. The Hessian matrix is given as

$$\mathbf{H} \triangleq \begin{bmatrix} \frac{\partial^2 L}{\partial \mathsf{vec}(\hat{\mathbf{X}}) \partial \mathsf{vec}(\hat{\mathbf{X}})^T} & \frac{\partial^2 L}{\partial \mathsf{vec}(\hat{\mathbf{X}}) \partial \mathsf{vec}(\hat{\mathbf{Y}})^T} \\ \frac{\partial^2 L}{\partial \mathsf{vec}(\hat{\mathbf{Y}}) \partial \mathsf{vec}(\hat{\mathbf{X}})^T} & \frac{\partial^2 L}{\partial \mathsf{vec}(\hat{\mathbf{Y}}) \partial \mathsf{vec}(\hat{\mathbf{Y}})^T} \end{bmatrix} = \begin{bmatrix} -\mathbf{\Sigma}_k \otimes \mathbf{I}_m & \mathbf{I}_k \otimes \mathbf{A} \\ \mathbf{I}_k \otimes \mathbf{A}^T & -\mathbf{\Sigma}_k \otimes \mathbf{I}_n \end{bmatrix},$$

because $\mathsf{vec}(\mathbf{A}\mathbf{Y} - \mathbf{X}\mathbf{C}) = (\mathbf{I}_k \otimes \mathbf{A})\mathsf{vec}(\mathbf{Y}) - (\mathbf{C}^T \otimes \mathbf{I}_m)\mathsf{vec}(\mathbf{X})$ and $\mathsf{vec}(\mathbf{A}^T\mathbf{X} - \mathbf{Y}\mathbf{C}) = (\mathbf{I}_k \otimes \mathbf{A}^T)\mathsf{vec}(\mathbf{X}) - (\mathbf{C}^T \otimes \mathbf{I}_n)\mathsf{vec}(\mathbf{Y})$.

Note that

$$\begin{bmatrix} \hat{\mathbf{X}}^T & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{Y}}^T \end{bmatrix} \begin{bmatrix} \hat{\mathbf{X}} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{Y}} \end{bmatrix} = \mathbf{I}_{2k}.$$

Thus, for any $\mathbf{Z}_1 \in \mathbb{R}^{m\times k}$ and $\mathbf{Z}_2 \in \mathbb{R}^{n\times k}$ such that $\mathbf{Z}_1^T\hat{\mathbf{X}} = \mathbf{0}$ and $\mathbf{Z}_2^T\hat{\mathbf{Y}} = \mathbf{0}$, it suffices for our purpose to prove $\mathbf{z}^T\mathbf{H}\mathbf{z} \leq 0$ where $\mathbf{z}^T = (\mathsf{vec}(\mathbf{Z}_1)^T, \mathsf{vec}(\mathbf{Z}_2)^T)$. Compute

$$\begin{aligned}
\mathbf{z}^T\mathbf{H}\mathbf{z} &= [\mathsf{vec}(\mathbf{Z}_1)^T, \mathsf{vec}(\mathbf{Z}_2)^T] \begin{bmatrix} -\boldsymbol{\Sigma}_k \otimes \mathbf{I}_m & \mathbf{I}_k \otimes \mathbf{A} \\ \mathbf{I}_k \otimes \mathbf{A}^T & -\boldsymbol{\Sigma}_k \otimes \mathbf{I}_n \end{bmatrix} \begin{bmatrix} \mathsf{vec}(\mathbf{Z}_1) \\ \mathsf{vec}(\mathbf{Z}_2) \end{bmatrix} \\
&= \mathsf{vec}(\mathbf{Z}_2)^T(\mathbf{I}_k \otimes \mathbf{A}^T)\mathsf{vec}(\mathbf{Z}_1) + \mathsf{vec}(\mathbf{Z}_1)^T(\mathbf{I}_k \otimes \mathbf{A})\mathsf{vec}(\mathbf{Z}_2) \\
&\quad - \mathsf{vec}(\mathbf{Z}_1)^T(\boldsymbol{\Sigma}_k \otimes \mathbf{I}_m)\mathsf{vec}(\mathbf{Z}_1) - \mathsf{vec}(\mathbf{Z}_2)^T(\boldsymbol{\Sigma}_k \otimes \mathbf{I}_n)\mathsf{vec}(\mathbf{Z}_2) \\
&= -\mathrm{tr}(\mathbf{Z}_1^T\mathbf{Z}_1\boldsymbol{\Sigma}_k) - \mathrm{tr}(\mathbf{Z}_2^T\mathbf{Z}_2\boldsymbol{\Sigma}_k) + 2\mathrm{tr}(\mathbf{Z}_1^T\mathbf{A}\mathbf{Z}_2) \triangleq \Delta.
\end{aligned}$$

Take a thin SVD of $\mathbf{A}$ as $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_k \oplus \boldsymbol{\Sigma}_{-k}$, $\mathbf{U} = [\mathbf{U}_k, \mathbf{U}_{-k}]$, and $\mathbf{V} = [\mathbf{V}_k, \mathbf{V}_{-k}]$. Denote $\mathbf{R}_1 = \mathbf{U}_{-k}^T\mathbf{Z}_1$ and and $\mathbf{R}_2 = \mathbf{V}_{-k}^T\mathbf{Z}_2$. Then $\mathrm{tr}(\mathbf{Z}_1^T\mathbf{A}\mathbf{Z}_2) = \mathrm{tr}(\mathbf{Z}_1^T\mathbf{U}_{-k}\boldsymbol{\Sigma}_{-k}\mathbf{V}_{-k}^T)$. And hence,

$$\begin{aligned}
-\Delta &= \mathrm{tr}(\mathbf{Z}_1^T\boldsymbol{\Sigma}_k\mathbf{Z}_1) + \mathrm{tr}(\mathbf{Z}_2^T\boldsymbol{\Sigma}_k\mathbf{Z}_2) - 2\mathrm{tr}(\mathbf{Z}_1^T\mathbf{U}_{-k}\boldsymbol{\Sigma}_{-k}\mathbf{V}_{-k}^T\mathbf{Z}_2) \\
&\geq \mathrm{tr}(\mathbf{Z}_1^T\mathbf{U}_{-k}\mathbf{U}_{-k}^T\mathbf{Z}_1\boldsymbol{\Sigma}_k) + \mathrm{tr}(\mathbf{Z}_2^T\mathbf{U}_{-k}\mathbf{U}_{-k}^T\mathbf{Z}_2\boldsymbol{\Sigma}_k) \\
&\quad - 2\mathrm{tr}(\mathbf{Z}_1^T\mathbf{U}_{-k}\boldsymbol{\Sigma}_{-k}\mathbf{V}_{-k}^T\mathbf{Z}_2) \\
&= \mathrm{tr}(\mathbf{R}_1^T\mathbf{R}_1\boldsymbol{\Sigma}_k) + \mathrm{tr}(\mathbf{R}_2^T\mathbf{R}_2\boldsymbol{\Sigma}_k) - 2\mathrm{tr}(\mathbf{R}_1^T\boldsymbol{\Sigma}_{-k}\mathbf{R}_2) \\
&\geq \mathrm{tr}(\mathbf{R}_1^T\boldsymbol{\Sigma}_{-k}\mathbf{R}_1) + \mathrm{tr}(\mathbf{R}_2^T\boldsymbol{\Sigma}_{-k}\mathbf{R}_2) - 2\mathrm{tr}(\mathbf{R}_1^T\boldsymbol{\Sigma}_{-k}\mathbf{R}_2) \\
&= \mathrm{tr}[(\mathbf{R}_1 - \mathbf{R}_2)^T\boldsymbol{\Sigma}_{-k}(\mathbf{R}_1 - \mathbf{R}_2)] \geq 0.
\end{aligned}$$

The last inequality uses the fact that $\mathrm{tr}(\mathbf{R}_1^T\mathbf{R}_1\boldsymbol{\Sigma}_k) \geq \mathrm{tr}(\mathbf{R}_1^T\boldsymbol{\Sigma}_{-k}\mathbf{R}_1)$ and $\mathrm{tr}(\mathbf{R}_2^T\mathbf{R}_2\boldsymbol{\Sigma}_k) \geq \mathrm{tr}(\mathbf{R}_2^T\boldsymbol{\Sigma}_{-k}\mathbf{R}_2)$. □

7

Unitarily Invariant Norms

In this chapter we study unitarily invariant norms of a matrix, which can be defined via singular values of the matrix. Unitarily invariant norms were contributed by J. von Neumann, Robert Schatten, and Ky Fan. J. von Neumann established an equivalent relationship between unitarily invariant norms and symmetric gauge functions. There are two popular classes of unitarily invariant norms: the Ky Fan norms and Schatten p-norms.

Parallel with the vector p-norms, the Schatten p-norms are defined on singular values of a matrix. Their special cases include the spectral norm, Frobenius norm, and nuclear norm. They have wide applications in modern data analysis and computation. For example, the Frobenius norm is used to measure approximation errors in regression and reconstruction problems because it essentially equivalent to the ℓ_2-norm of a vector. The spectral norm is typically used to describe convergence and convergence rate of an iteration procedure. The nuclear norm provides an effective approach to matrix low rank modeling.

We first briefly review matrix norms, and then present the notion of symmetric gauge functions. Symmetric gauge functions facilitate us to study unitarily invariant norms. First, it transforms a unitarily invariant norm on matrices to a norm on vectors equivalently. Second, it can incorporate majorization theory. Accordingly, we give some important properties of unitarily invariant norms.

7.1 Matrix Norms

A function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is said to be a matrix norm if the following conditions are satisfied:

(1) $f(\mathbf{A}) > 0$ for all nonzero matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$;

(2) $f(\alpha\mathbf{A}) = |\alpha| f(\mathbf{A})$ for any $\alpha \in \mathbb{R}$ and any $\mathbf{A} \in \mathbb{R}^{m\times n}$;

(3) $f(\mathbf{A}+\mathbf{B}) \leq f(\mathbf{A}) + f(\mathbf{B})$ for any $\mathbf{A}$ and $\mathbf{B} \in \mathbb{R}^{m\times n}$.

We denote the norm of a matrix $\mathbf{A}$ by $\|\mathbf{A}\|$. Furthermore, if

(4) $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$ where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$,

the matrix norm is said to be consistent. In some literature, when one refers to a matrix norm on $\mathbf{R}^{n\times n}$. it is required to be consistent. Here we do not make this requirement.

There is an equivalence between any two norms. Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be two norms on $\mathbb{R}^{m\times n}$. Then there exist positive numbers α_1 and α_2 such that for all $\mathbf{A} \in \mathbb{R}^{m\times n}$,

$$\alpha_1\|\mathbf{A}\|_\alpha \leq \|\mathbf{A}\|_\beta \leq \alpha_2\|\mathbf{A}\|_\alpha.$$

Conditions (2) and (3) tell us that the norm is convex. Moreover, it is continuous because

$$\big|\|\mathbf{A}\| - \|\mathbf{B}\|\big| \leq \|\mathbf{A}-\mathbf{B}\| \leq \alpha\|\mathbf{A}-\mathbf{B}\|_F, \quad \text{where } \alpha > 0.$$

A norm always companies with its dual. The dual is a norm. Moreover, the dual of the dual norm is the original norm.

Definition 7.1. Let $\|\cdot\|$ be a given norm on $\mathbb{R}^{m\times n}$. Its dual (denoted $\|\cdot\|^*$) is defined as

$$\|\mathbf{A}\|^* = \max\big\{\mathrm{tr}(\mathbf{A}\mathbf{B}^T) : \ \mathbf{B} \in \mathbb{R}^{m\times n}, \|\mathbf{B}\| = 1\big\}.$$

Proposition 7.1. The dual $\|\cdot\|^*$ has the following properties:

(1) The dual is a norm.

(2) $(\|\mathbf{A}\|^*)^* = \|\mathbf{A}\|$.

(3) $\operatorname{tr}(\mathbf{A}\mathbf{B}^T) \le |\operatorname{tr}(\mathbf{A}^T\mathbf{B})| \le \|\mathbf{A}\|\|\mathbf{B}\|^*$ (or $\|\mathbf{A}\|^*\|\mathbf{B}\|$).

There are two approaches for definition of a matrix norm. In the first approach, the norm of matrix $\mathbf{A}$ is defined via its vectorization $\mathsf{vec}(\mathbf{A})$; that is, $\|\mathbf{A}\| = \|\mathsf{vec}(\mathbf{A})\|$, which obviously satisfies Conditions (1)-(3). We refer to this class of the matrix norms as *matrix vectorization norms* for ease of exposition. Note that the Frobenius norm is a matrix vectorization norm because $\|\mathbf{A}\|_F = \|\mathsf{vec}(\mathbf{A})\|_2$. However, this class of matrix norms are not always consistent. For example, let

$$\mathbf{A} = \mathbf{B} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Since $\mathbf{AB} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$ and

$$2 = \|\mathsf{vec}(\mathbf{AB})\|_\infty > \|\mathsf{vec}(\mathbf{A})\|_\infty \|\mathsf{vec}(\mathbf{B})\|_\infty = 1,$$

this implies that the corresponding matrix norm is not consistent.

In the second approach, the matrix norm is defined by

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|,$$

which is also called the *induced* or *operator* norm.

Theorem 7.1. The operator norm on $\mathbb{R}^{m\times n}$ is a consistent matrix norm.

Proof. Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, the result is trivial If $\mathbf{A} = \mathbf{0}$. Assume that $\mathbf{A} \neq \mathbf{0}$. Then there exists a nonzero vector $\mathbf{z} \in \mathbb{R}^n$ for which $\mathbf{A}\mathbf{z} \neq \mathbf{0}$. So we have $\|\mathbf{A}\mathbf{z}\| > 0$ and $\|\mathbf{z}\| > 0$. Hence,

$$\|\mathbf{A}\| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} \ge \frac{\|\mathbf{A}\mathbf{z}\|}{\|\mathbf{z}\|} > 0.$$

Conditions (2)-(3) are directly obtained from the definition of the vector norm. As for Condition (4), it can be established by

$$\|\mathbf{ABx}\| \leq \|\mathbf{A}\|\|\mathbf{Bx}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|\|\mathbf{x}\|$$

for any $\mathbf{x} \neq \mathbf{0}$. Thus,

$$\|\mathbf{AB}\| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{ABx}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\|\|\mathbf{B}\|.$$

□

As we have shown, $\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2 = \sigma_1(\mathbf{A})$. It is thus called the spectral norm.

Note that $\|\mathbf{UAV}\|_2 = \|\mathbf{A}\|_2$ and $\|\mathbf{UAV}\|_F = \|\mathbf{A}\|_F$ for any $m \times m$ orthonormal matrix $\mathbf{U}$ and any $n \times n$ orthonormal matrix $\mathbf{V}$. In other words, they are unitarily invariant.

Definition 7.2. A matrix norm is said to be *unitarily invariant* if $\|\mathbf{UAV}\| = \|\mathbf{A}\|$ for any unitary matrices $\mathbf{U}$ and $\mathbf{V}$.

In this tutorial, we only consider real matrices. Thus, a unitarily invariant norm should be termed as "orthogonally invariant norm." However, we still follow the term of the unitarily invariant norm and denote it by $|\!|\!| \cdot |\!|\!|$.

Theorem 7.2. Let $\|\cdot\|$ be a given norm on $\mathbb{R}^{m\times n}$. Then it is unitarily invariant if and only if its dual is unitarily invariant.

Proof. Suppose $\|\cdot\|$ is unitarily invariant, and let $\mathbf{U} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ be orthonormal. Then

$$\begin{aligned}\|\mathbf{UAV}\|^* &= \max\left\{\mathrm{tr}(\mathbf{UAVB}^T) : \ \mathbf{B} \in \mathbb{R}^{m\times n}, \|\mathbf{B}\| = 1\right\} \\ &= \max\left\{\mathrm{tr}(\mathbf{A}(\mathbf{U}^T\mathbf{BV}^T)^T) : \ \mathbf{B} \in \mathbb{R}^{m\times n}, \|\mathbf{B}\| = 1\right\} \\ &= \max\left\{\mathrm{tr}(\mathbf{AC}^T) : \ \mathbf{C} \in \mathbb{R}^{m\times n}, \|\mathbf{UCV}\| = 1\right\} \\ &= \max\left\{\mathrm{tr}(\mathbf{AC}^T) : \ \mathbf{C} \in \mathbb{R}^{m\times n}, \|\mathbf{C}\| = 1\right\} = \|\mathbf{A}\|^*.\end{aligned}$$

The converse follows from the fact that $(\|\mathbf{A}\|^*)^* = \|\mathbf{A}\|$. □

We find that $\|\mathbf{A}\|_2 = \|\boldsymbol{\sigma}(\mathbf{A})\|_\infty$ and $\|\mathbf{A}\|_F = \|\boldsymbol{\sigma}(\mathbf{A})\|_2$; that is, they correspond the norms on the vector $\boldsymbol{\sigma}(\mathbf{A})$ of the singular values of $\mathbf{A}$. This sheds light on the relationship of a unitarily invariant norm of a matrix with its singular values.

7.2 Symmetric Gauge Functions

In order to investigate the unitarily invariant norm, we first present the notion of symmetric gauge functions.

Definition 7.3. A real function $\phi : \mathbb{R}^n \to \mathbb{R}$ is called a symmetric gauge function if it satisfies the following four conditions:

(1) $\phi(\mathbf{u}) > 0$ for all nonzero $\mathbf{u} \in \mathbb{R}^n$.

(2) $\phi(\alpha \mathbf{u}) = |\alpha| \phi(\mathbf{u})$ for any constant $\alpha \in \mathbb{R}$.

(3) $\phi(\mathbf{u} + \mathbf{v}) \leq \phi(\mathbf{u}) + \phi(\mathbf{v})$ for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$.

(4) $\phi(\mathbf{D}\mathbf{u}_\pi) = \phi(\mathbf{u})$ where $\mathbf{u}_\pi = (u_{\pi_1}, \ldots, u_{\pi_n})$ with π as a permutation of $[n]$ and $\mathbf{D}$ is an $n \times n$ diagonal matrix with ± 1 diagonal elements.

Furthermore, the gauge function is called normalized if it satisfies the condition:

(5) $\phi(1, 0, \ldots, 0) = 1$.

Conditions (1)-(3) show that that the gauge function is a vector norm. Thus, it is convex and continuous. Condition (4) says that the gauge function is symmetric.

Lemma 7.3. [Schatten, 1950] Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. If $|\mathbf{u}| \leq |\mathbf{v}|$, then $\phi(\mathbf{u}) \leq \phi(\mathbf{v})$ for every symmetric gauge function ϕ.

Proof. In terms of Condition (4), we can directly assume that $\mathbf{u} \geq \mathbf{0}$ and $\mathbf{v} \geq \mathbf{0}$. Currently, the argument is equivalent to

$$\phi(\omega_1 v_1, \ldots, \omega_n v_n) \leq \phi(v_1, \ldots, v_n)$$

for $\omega_i \in [0, 1]$. Thus, by induction, it suffices to prove

$$\phi(v_1, \ldots, v_{n-1}, \omega v_n) \leq \phi(v_1, \ldots, v_n)$$

where $\omega \in [0, 1]$ for every symmetric gauge function ϕ. It follows from the following direct computation:

$$\begin{aligned}
&\phi(v_1, \ldots, v_{n-1}, \omega v_n) \\
&= \phi\Big(\frac{1{+}\omega}{2}v_1{+}\frac{1{-}\omega}{2}v_1, \ldots, \frac{1{+}\omega}{2}v_{n-1}{+}\frac{1{-}\omega}{2}v_{n-1}, \frac{1{+}\omega}{2}v_n - \frac{1{-}\omega}{2}v_n\Big) \\
&\leq \frac{1+\omega}{2}\phi(v_1, \ldots, v_{n-1}, v_n) + \frac{1-\omega}{2}\phi(v_1, \ldots, v_{n-1}, -v_n) \\
&= \phi(v_1, \ldots, v_{n-1}, v_n).
\end{aligned}$$

□

Theorem 7.4. [Fan, 1951] Given two nonnegative vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n_+$, then $\mathbf{u} \prec_w \mathbf{v}$ if and only if $\phi(\mathbf{u}) \leq \phi(\mathbf{v})$ for every symmetric gauge function ϕ.

Proof. The necessity is obtained by setting a set of special symmetric gauge functions ϕ_k for $k \in [n]$. Specifically, they are defined as

$$\phi_k(\mathbf{x}) = \max_{1 \leq i_1 \leq \cdots \leq i_k \leq n} \sum_{l=1}^{k} |x_{i_l}|.$$

where $\mathbf{x} = (x_1, \ldots, x_n)$.

It remains to prove the sufficiency. Without loss of generality, we assume that $u_1 \geq \cdots \geq u_n$ and $v_1 \geq \cdots \geq v_n$. Let $\mathbf{z} = (z_1, \ldots, z_n)^T$ where $z_i = v_i$ for $i \in [n-1]$ and $z_n = v_n - \sum_{i=1}^n (v_i - u_i)$. Obviously, $\mathbf{z} \leq \mathbf{v}$. And it follows from $\mathbf{u} \prec_w \mathbf{v}$ that $\mathbf{u} \prec \mathbf{z}$. In terms of the theorem of Hardy, Littlewood, and Pólya (see Lemma 2.2), there exists a doubly stochastic matrix (say $\mathbf{W}$) such that $\mathbf{u} = \mathbf{W}\mathbf{z}$. Since $\mathbf{W}(\mathbf{v} - \mathbf{z}) \geq \mathbf{0}$, we have $\mathbf{u} \leq \mathbf{W}\mathbf{v}$. Thus, by Lemma 7.3, $\phi(\mathbf{u}) \leq \phi(\mathbf{W}\mathbf{v})$ for every symmetric gauge function. Consider that a doubly stochastic matrix can be expressed a convex combination of a set of permutation matrices (see Lemma 2.3). We write $\mathbf{W} = \sum_{j=1} \alpha_j \mathbf{P}_j$ where $\alpha_j \geq 0$ and $\sum_j \alpha_j = 1$, and the $\mathbf{P}_j$ are permutation matrices. Accordingly,

$$\phi(\mathbf{u}) \leq \phi(\sum_j \alpha_j \mathbf{P}_j \mathbf{v}) \leq \sum_j \alpha_j \phi(\mathbf{P}_j \mathbf{v}) = \sum_j \alpha_j \phi(\mathbf{v}) = \phi(\mathbf{v}).$$

□

It is worth noting that the proof of Theorem 7.4 implies that if $\phi_k(\mathbf{u}) \leq \phi_k(\mathbf{v})$ for $k \in [n]$, then $\phi(\mathbf{u}) \leq \phi(\mathbf{v})$ for every symmetric gauge function ϕ. In other words, an infinite family of norm inequalities follows from a finite one.

Definition 7.4. The dual of a symmetric gauge function ϕ on $\mathbb{R}^n$ is defined as

$$\phi^*(\mathbf{u}) \triangleq \max\left\{\mathbf{u}^T\mathbf{v} : \mathbf{v} \in \mathbb{R}^n, \phi(\mathbf{v}) = 1\right\}.$$

Proposition 7.2. Let ϕ^* be the dual of the symmetric gauge function ϕ. Then ϕ^* is also a symmetric gauge function. Moreover, $(\phi^*)^* = \phi$.

Proof. For a nonzero vector $\mathbf{u} \in \mathbb{R}^n$, then $\phi(\mathbf{u}) > 0$. Hence,

$$\max_{\phi(\mathbf{v})=1} \mathbf{u}^T\mathbf{v} \geq \frac{\mathbf{u}^T\mathbf{u}}{\phi(\mathbf{u})} > 0.$$

It is also seen that

$$\phi^*(\mathbf{u}+\mathbf{v}) = \max_{\phi(\mathbf{z})=1} (\mathbf{u}+\mathbf{v})^T\mathbf{z} \leq \max_{\phi(\mathbf{z})=1} \mathbf{u}^T\mathbf{z} + \max_{\phi(\mathbf{z})=1} \mathbf{v}^T\mathbf{z} \leq \phi^*(\mathbf{u}) + \phi^*(\mathbf{v}).$$

As for the symmetry of ϕ^* can be directly obtained from that of ϕ. Finally, note that ϕ^* is a norm on $\mathbb{R}^n$. Thus, $(\phi^*)^* = \phi$. □

7.3 Unitarily Invariant Norms via SGFs

There is a one-to-one correspondence between a unitarily invariant norm and a symmetric gauge function (SGF).

Theorem 7.5. If $|\!|\!| \cdot |\!|\!|$ is a given unitarily invariant norm on $\mathbb{R}^{m\times n}$, then there is a symmetric gauge function ϕ on $\mathbb{R}^q$ where $q = \min\{m, n\}$ such that $|\!|\!|\mathbf{A}|\!|\!| = \phi(\boldsymbol{\sigma}(\mathbf{A}))$ for all $\mathbf{A} \in \mathbb{R}^{m\times n}$.

Conversely, if ϕ is a symmetric gauge function on $\mathbb{R}^q$, then $|\!|\!|\mathbf{A}|\!|\!| \triangleq \phi(\boldsymbol{\sigma}(\mathbf{A}))$ is a unitarily invariant norm on $\mathbb{R}^{m\times n}$.

Proof. Given a unitarily invariant norm $|\!|\!| \cdot |\!|\!|$ on $\mathbb{R}^{m\times n}$ and a vector $\mathbf{x} \in \mathbb{R}^q$, define $\phi(\mathbf{x}) \triangleq |\!|\!|\mathbf{X}|\!|\!|$ where $\mathbf{X} = [x_{ij}] \in \mathbb{R}^{m\times n}$ satisfying that $x_{ii} = x_i$ for $i \in [q]$ and all other elements are zero. That ϕ is a norm on $\mathbb{R}^q$ follows from the fact that $|\!|\!| \cdot |\!|\!|$ is a norm. The unitary invariance of

$\|\!|\cdot|\!\|$ then implies that ϕ satisfies the symmetry. Now let $\mathbf{A} = \mathbf{U\Sigma V}^T$ be the full SVD of $\mathbf{A}$. Then $\|\!|\mathbf{A}|\!\| = \|\!|\mathbf{U\Sigma V}^T|\!\| = \|\!|\mathbf{\Sigma}|\!\| = \phi(\boldsymbol{\sigma}(\mathbf{A}))$.

Conversely, if ϕ is a symmetric gauge function, for any $\mathbf{A} \in \mathbb{R}^{m\times n}$ define $\|\!|\mathbf{A}|\!\| = \phi(\boldsymbol{\sigma}(\mathbf{A}))$. We now prove that $\|\!|\cdot|\!\|$ is a unitarily invariant norm. First, that $\|\!|\mathbf{A}|\!\| > 0$ for $\mathbf{A} \neq \mathbf{0}$ and $\|\!|\alpha\mathbf{A}|\!\| = |\alpha|\|\!|\mathbf{A}|\!\|$ for any constant α follows the fact that ϕ is a norm. The unitary invariance of $\|\!|\cdot|\!\|$ follows from that for any orthonormal matrices $\mathbf{U}$ $(m\times m)$ and $\mathbf{V}$ $(n \times n)$, $\mathbf{UAV}$ and $\mathbf{A}$ have the same singular values. Finally,

$$\begin{aligned}\|\!|\mathbf{A}+\mathbf{B}|\!\| &= \phi(\boldsymbol{\sigma}(\mathbf{A}+\mathbf{B})) \leq \phi(\boldsymbol{\sigma}(\mathbf{A}) + \boldsymbol{\sigma}(\mathbf{B})) \\ &\leq \phi(\boldsymbol{\sigma}(\mathbf{A})) + \phi(\boldsymbol{\sigma}(\mathbf{B})) \\ &= \|\!|\mathbf{A}|\!\| + \|\!|\mathbf{B}|\!\|.\end{aligned}$$

Here the first inequality follows Proposition 6.3 and Theorem 7.4. □

The following theorem implies that there is also a one-one correspondence between the dual of a symmetric gauge function and a dual unitarily invariant norm.

Theorem 7.6. Let ϕ^* be the dual of symmetric gauge function ϕ. Then $\|\!|\mathbf{A}|\!\| = \phi(\boldsymbol{\sigma}(\mathbf{A}))$ if and only if $\|\!|\mathbf{A}|\!\|^* = \phi^*(\boldsymbol{\sigma}(\mathbf{A}))$.

Proof. Assume that $\|\!|\mathbf{A}|\!\| = \phi(\boldsymbol{\sigma}(\mathbf{A}))$. Then

$$\begin{aligned}\|\!|\mathbf{A}|\!\|^* &= \max\big\{\mathrm{tr}(\mathbf{A}^T\mathbf{B}) : \mathbf{B} \in \mathbb{R}^{m\times n}, \|\!|\mathbf{B}|\!\| = 1\big\} \\ &= \max\big\{\mathrm{tr}(\mathbf{\Sigma}_A^T\mathbf{U}_A^T\mathbf{B}\mathbf{V}_A) : \phi(\boldsymbol{\sigma}(\mathbf{B})) = 1\big\},\end{aligned}$$

where $\mathbf{A} = \mathbf{U}_A\mathbf{\Sigma}_A\mathbf{V}_A^T$ is a full SVD of $\mathbf{A}$. By Theorem 6.3, we have

$$\mathrm{tr}(\mathbf{V}_A^T\mathbf{B}^T\mathbf{U}_A\mathbf{\Sigma}_A) \leq \max_{\mathbf{U}^T\mathbf{U}=\mathbf{I}_m, \mathbf{V}^T\mathbf{V}=\mathbf{I}_n} \mathrm{tr}(\mathbf{V}^T\mathbf{B}^T\mathbf{U}\mathbf{\Sigma}_A) = \sum_{i=1}^{q} \sigma_i(\mathbf{A})\sigma_i(\mathbf{B}).$$

When letting $\mathbf{B} = \mathbf{U}_A\mathbf{\Sigma}_B\mathbf{V}_A^T$ as a full SVD of $\mathbf{B}$, we can obtain that

$$\|\!|\mathbf{A}|\!\|^* = \max\big\{\mathrm{tr}(\mathbf{\Sigma}_A^T\mathbf{\Sigma}_B), \phi(\boldsymbol{\sigma}(\mathbf{B})) = 1\big\} = \phi^*(\boldsymbol{\sigma}(\mathbf{A})).$$

Conversely, the result follows from the fact that $(\phi^*)^* = \phi$. □

Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, let it have a full SVD: $\mathbf{A} = \mathbf{U\Sigma V}^T$. Then $\|\!|\mathbf{A}|\!\| = \|\!|\mathbf{\Sigma}|\!\|$. As we have seen, for $\mathbf{x} \in \mathbb{R}^n$ the function

$$\phi(\mathbf{x}) \triangleq \max_{1\leq i_1 \leq \cdots \leq i_k \leq n} \sum_{l=1}^{k} |x_{i_l}|$$

is a symmetric gauge function. Thus, $\sum_{i=1}^{k} \sigma_i(\mathbf{A})$ defines also a class of unitarily invariant norms which are the so-called *Ky Fan k-norms*.

Clearly, the vector p-norm $\| \cdot \|_p$ for $p \geq 1$ is a symmetric gauge function. Thus, Theorem 7.5 shows that $|\!|\!|\mathbf{A}|\!|\!|_p \triangleq \|\boldsymbol{\sigma}(\mathbf{A})\|_p$ for $p \geq 1$ are a class of unitarily invariant norms. They are well known as the *Schatten p-norms*. Thus, $\|\mathbf{A}\|_F = \|\boldsymbol{\sigma}(\mathbf{A})\|_2 = |\!|\!|\mathbf{A}|\!|\!|_2$ and $\|\mathbf{A}\|_2 = \|\boldsymbol{\sigma}(\mathbf{A})\|_\infty = |\!|\!|\mathbf{A}|\!|\!|_\infty$.

When $p = 1$, $\|\mathbf{A}\|_* \triangleq |\!|\!|\mathbf{A}|\!|\!|_1 = \|\boldsymbol{\sigma}(\mathbf{A})\|_1 = \sum_{i=1}^{\min\{m,n\}} \sigma_i(\mathbf{A})$ is called the *nuclear norm* or *trace norm*, which has been widely used in many machine learning problems such as matrix completion, matrix data classification, multi-task learning, etc. [Srebro et al., 2004, Cai et al., 2010, Mazumder et al., 2010, Liu et al., 2013, Luo et al., 2015, Kang et al., 2011, Pong et al., 2010, Zhou and Li, 2014]. Parallel with the ℓ_1-norm which is used as convex relaxation of the ℓ_0-norm [Tibshirani, 1996], the nuclear norm is a convex alternative of the matrix rank. Since the nuclear norm is the best convex approximation of the matrix rank over the unit ball of matrices, this makes it more tractable to solve the resulting optimization problem (see Example 8.1 below).

7.4 Properties of Unitarily Invariant Norms

Theorem 7.5 opens an approach for exploring unitarily invariant norms by using symmetric gauge functions and majorization theory. We will see that this makes things more tractable.

Theorem 7.7. Let $|\!|\!| \cdot |\!|\!|$ be a unitarily invariant norm on $\mathbb{R}^{n \times n}$. Then it is consistent.

Theorem 7.7 follows immediately from Theorem 6.6. However, when the norm is defined on $\mathbb{R}^{m \times n}$, Theorem 6.6 can not help to establish the consistency of the corresponding unitarily invariant norm.

As an immediate corollary of Theorem 7.5, we have the following result, which shows that unitarily invariant norms are monotone.

Theorem 7.8. Let $|\!|\!| \cdot |\!|\!|$ be a given unitarily invariant norm on $\mathbb{R}^{m \times n}$. Then $|\!|\!|\mathbf{A}|\!|\!| \leq |\!|\!|\mathbf{B}|\!|\!|$ if and only if $\boldsymbol{\sigma}(\mathbf{A}) \prec_w \boldsymbol{\sigma}(\mathbf{B})$.

Proposition 7.3. Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, let $[\mathbf{A}]_r$ be obtained by replacing the last r rows and r columns of $\mathbf{A}$ with zeros, and $\langle \mathbf{A} \rangle_r$ by replacing the last r rows or columns of $\mathbf{A}$ with zeros. Let $q = \min\{m, n\}$. Then for any $r \in [q]$,

$$|\!|\!|[\mathbf{A}]_r|\!|\!| \leq |\!|\!|\langle \mathbf{A} \rangle_r|\!|\!| \leq |\!|\!|\mathbf{A}|\!|\!|.$$

Proof. Part (1) directly follows from Proposition 6.3 which shows that $\boldsymbol{\sigma}([\mathbf{A}]_r) \prec_w \boldsymbol{\sigma}(\langle \mathbf{A} \rangle_r) \prec_w \boldsymbol{\sigma}(\mathbf{A})$. □

Proposition 7.4. Given two matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{m \times n}$, we have that

$$|\!|\!|\mathrm{diag}(\boldsymbol{\sigma}(\mathbf{A}) - \boldsymbol{\sigma}(\mathbf{B}))|\!|\!| \leq |\!|\!|\mathbf{A} - \mathbf{B}|\!|\!|.$$

Furthermore, if both $\mathbf{A}$ and $\mathbf{B}$ are symmetric matrixes in $\mathbb{R}^{m \times m}$, then

$$|\!|\!|\mathrm{diag}(\boldsymbol{\sigma}(\mathbf{A}) - \boldsymbol{\sigma}(\mathbf{B})) \leq |\!|\!|\mathrm{diag}(\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B}))|\!|\!||\!|\!| \leq |\!|\!|\mathbf{A} - \mathbf{B}|\!|\!|.$$

Proof. The first part of the proposition is immediately obtained from Theorem 6.5. As for the second part, Proposition 6.1-(i) says that $\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B}) \prec \boldsymbol{\lambda}(\mathbf{A} - \mathbf{B})$. It then follows from Lemmas 2.2 and 2.3 that $\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B}) = \sum_j \alpha_j \mathbf{P}_j \boldsymbol{\lambda}(\mathbf{A} - \mathbf{B})$ where the $\alpha_j \geq 0$ and $\sum_j \alpha_j = 1$, and the $\mathbf{P}_j$ are some permutation matrices. Accordingly, for every symmetric gauge function ϕ on $\mathbb{R}^m$, we have that

$$\begin{aligned} \phi(\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B})) &= \phi(\sum_j \alpha_j \mathbf{P}_j \boldsymbol{\lambda}(\mathbf{A} - \mathbf{B})) \leq \sum_j \alpha_j \phi(\mathbf{P}_j \boldsymbol{\lambda}(\mathbf{A} - \mathbf{B})) \\ &= \sum_j \alpha_j \phi(\boldsymbol{\lambda}(\mathbf{A} - \mathbf{B})) = \phi(\boldsymbol{\lambda}(\mathbf{A} - \mathbf{B})), \end{aligned}$$

which implies that $|\!|\!|\mathrm{diag}(\boldsymbol{\lambda}(\mathbf{A}) - \boldsymbol{\lambda}(\mathbf{B}))|\!|\!||\!|\!| \leq |\!|\!|\mathbf{A} - \mathbf{B}|\!|\!|$. Additionally, consider that for a symmetric matrix $\mathbf{M}$, it holds that $\sigma_i(\mathbf{M}) = |\lambda_i(\mathbf{M})|$. Hence, we have that

$$|\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})| \geq ||\lambda_i(\mathbf{A})| - |\lambda_i(\mathbf{B})|| = |\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})|.$$

This concludes the proof. □

As a direct corollary of Proposition 6.5, we have that

$$|\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})| \leq \|\mathbf{A} - \mathbf{B}\|_2, \text{ for } i = 1, \ldots, q,$$

where $q = \min\{m, n\}$, and

$$\sqrt{\sum_{i=1}^{q} (\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B}))^2} \leq \|\mathbf{A} - \mathbf{B}\|_F.$$

When $\mathbf{A}$ and $\mathbf{B}$ are both symmetric, we also have that

$$|\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})| \leq \|\mathbf{A} - \mathbf{B}\|_2, \text{ for } \ i = 1, \ldots, m,$$

$$\sqrt{\sum_{i=1}^{m} (\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B}))^2} \leq \|\mathbf{A} - \mathbf{B}\|_F.$$

The latter result is well known as the Hoffman-Wielandt theorem. Note that the Hoffman-Wielandt theorem still hods when $\mathbf{A}$ and $\mathbf{B}$ are normal [Stewart and Sun, 1990].

Theorem 7.9. Let $|\!|\!| \cdot |\!|\!|$ be an arbitrary unitarily invariant norm on $\mathbb{R}^{m\times n}$, and $\mathbf{E}_{11} \in \mathbb{R}^{m\times n}$ have the entry 1 in the $(1, 1)$th position and zeroes elsewhere. Then

(a) $|\!|\!|\mathbf{A}|\!|\!| = |\!|\!|\mathbf{A}^T|\!|\!|$.

(b) $\sigma_1(\mathbf{A})|\!|\!|\mathbf{E}_{11}|\!|\!| \leq |\!|\!|\mathbf{A}|\!|\!| \leq \|\mathbf{A}\|_* |\!|\!|\mathbf{E}_{11}|\!|\!|$.

(c) If the symmetric gauge function ϕ corresponding to the norm $|\!|\!| \cdot |\!|\!|$ is normalized (i.e., $\phi(1, 0, 0, \ldots, 0) = 1$), then

$$\|\mathbf{A}\|_2 \leq |\!|\!|\mathbf{A}|\!|\!| \leq \|\mathbf{A}\|_*.$$

Proof. Part (a) is due to that $\phi(\boldsymbol{\sigma}(\mathbf{A})) = \phi(\boldsymbol{\sigma}(\mathbf{A}^T))$.

If $\phi(1, 0, \ldots, 0) = 1$, then $|\!|\!|\mathbf{E}_{11}|\!|\!| = 1$. Thus, we can have Part (c) from Part (b). Assume $\mathbf{A}$ is nonzero. Otherwise, the result is trivial. Let $q = \min\{m, n\}$. First,

$$\begin{aligned} |\!|\!|\mathbf{A}|\!|\!| &= \phi(\sigma_1(\mathbf{A}), \ldots, \sigma_q(\mathbf{A})) = \sigma_1(\mathbf{A})\phi(1, \sigma_2(\mathbf{A})/\sigma_1(\mathbf{A}), \ldots, \sigma_q(\mathbf{A})/\sigma_1(\mathbf{A})) \\ &\geq \sigma_1(\mathbf{A})\phi(1, 0, \ldots, 0) = \sigma_1(\mathbf{A})|\!|\!|\mathbf{E}_{11}|\!|\!|. \end{aligned}$$

Since $\left(\sigma_1(\mathbf{A})/\sum_{i=1}^{q}\sigma_i(\mathbf{A}), \ldots, \sigma_q(\mathbf{A})/\sum_{i=1}^{q}\sigma_i(\mathbf{A})\right) \prec (1, 0, \ldots, 0)$, we have

$$\begin{aligned} |\!|\!|\mathbf{A}|\!|\!| &= (\sum_{i=1}^{q}\sigma_i(\mathbf{A}))\phi(\sigma_1(\mathbf{A})/\sum_{i=1}^{q}\sigma_i(\mathbf{A}), \ldots, \sigma_q(\mathbf{A})/\sum_{i=1}^{q}\sigma_i(\mathbf{A})) \\ &\leq \|\mathbf{A}\|_*\phi(1, 0, \ldots, 0) = \|\mathbf{A}\|_* |\!|\!|\mathbf{E}_{11}|\!|\!|. \end{aligned}$$

□

Note that a norm $\| \cdot \|$ on $\mathbb{R}^{m \times n}$ is said to be *self adjoint* if $\|\mathbf{A}\| = \|\mathbf{A}^T\|$ for any $\mathbf{A} \in \mathbb{R}^{m \times n}$. Thus, Theorem 7.9-(a) shows that the unitarily invariant norm is self-adjoint.

It is worth mentioning that $\|\!|\mathbf{E}_{ij}|\!\| = \|\!|\mathbf{E}_{11}|\!\|$ where $\mathbf{E}_{ij} \in \mathbb{R}^{m \times n}$ has entry 1 in the (i, j)th position and zeros elsewhere. Moreover, the Schatten p-norms satisfy $\|\mathbf{E}_{11}\|_p = 1$. Theorem 7.9 says that for any unitarily invariant norm $\|\!| \cdot |\!\|$ such that $\|\!|\mathbf{E}_{11}|\!\| = 1$,

$$1 \leq \frac{\|\!|\mathbf{A}|\!\|}{\|\mathbf{A}\|_2} \leq \frac{\|\mathbf{A}\|_*}{\|\mathbf{A}\|_2} \leq \operatorname{rank}(\mathbf{A}).$$

Recall that $\frac{\sum_{i=1}^q \sigma_i^2(\mathbf{A})}{\sigma_1^2(\mathbf{A})} = \frac{\|\mathbf{A}\|_F^2}{\|\mathbf{A}\|_2^2}$ and $\frac{\sum_{i=1}^q \sigma_i(\mathbf{A})}{\sigma_1(\mathbf{A})} = \frac{\|\mathbf{A}\|_*}{\|\mathbf{A}\|_2}$, so called *stable rank* and *nuclear rank* (see Definition 3.2). They have been found usefulness in the analysis of matrix multiplication approximation [Magen and Zouzias, 2011, Cohen et al., 2015, Kyrillidis et al., 2014].

Theorem 7.10. Let $\mathbf{M} \in \mathbb{R}^{m \times m}$, $\mathbf{N} \in \mathbb{R}^{n \times n}$, and $\mathbf{A} \in \mathbb{R}^{m \times n}$ such that the block matrix

$$\begin{bmatrix} \mathbf{M} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{N} \end{bmatrix}$$

is SPSD. Then

$$\|\!|\mathbf{M}|\!\| + \|\!|\mathbf{N}|\!\| \geq 2\|\!|\mathbf{A}|\!\|.$$

Proof. Without loss of generality, we assume $m \geq n$. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be a thin SVD of $\mathbf{A}$. Consider that

$$[\mathbf{U}^T, -\mathbf{V}^T] \begin{bmatrix} \mathbf{M} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{U} \\ -\mathbf{V} \end{bmatrix} = \mathbf{U}^T\mathbf{M}\mathbf{U} + \mathbf{V}^T\mathbf{N}\mathbf{V} - \mathbf{U}^T\mathbf{A}\mathbf{V} - \mathbf{V}^T\mathbf{A}^T\mathbf{U}$$

is PSD. Hence, $\|\!|\mathbf{U}^T\mathbf{M}\mathbf{U} + \mathbf{V}^T\mathbf{N}\mathbf{V}|\!\| \geq 2\|\!|\mathbf{\Sigma}|\!\|$. That is,

$$\|\!|\mathbf{V}^T\mathbf{U}^T\mathbf{M}\mathbf{U}\mathbf{V} + \mathbf{N}|\!\| \geq 2\|\!|\mathbf{A}|\!\|.$$

Note that

$$\|\!|\mathbf{V}^T\mathbf{U}^T\mathbf{M}\mathbf{U}\mathbf{V} + \mathbf{N}|\!\| \leq \|\!|\mathbf{V}^T\mathbf{U}^T\mathbf{M}\mathbf{U}\mathbf{V}|\!\| + \|\!|\mathbf{N}|\!\| \leq \|\!|\mathbf{M}|\!\| + \|\!|\mathbf{N}|\!\|.$$

□

Proposition 7.5. Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, then the following holds

$$\|\mathbf{A}\| = \min_{\mathbf{X},\mathbf{Y}:\mathbf{X}\mathbf{Y}^T=\mathbf{A}} \frac{1}{2}\Big\{\|\mathbf{X}\mathbf{X}^T\| + \|\mathbf{Y}\mathbf{Y}^T\|\Big\}.$$

If $\mathrm{rank}(\mathbf{A}) = r \leq \min\{m, n\}$, then the minimum above is attained at a rank decomposition $\mathbf{A} = \hat{\mathbf{X}}\hat{\mathbf{Y}}^T$ where $\hat{\mathbf{X}} = \mathbf{U}_r\mathbf{\Sigma}_r^{1/2}$ and $\hat{\mathbf{Y}} = \mathbf{V}_r\mathbf{\Sigma}_r^{1/2}$, and $\mathbf{A} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T$ is a condensed SVD of $\mathbf{A}$.

Proof. Let $\mathbf{A} = \mathbf{X}\mathbf{Y}^T$ be any decomposition of $\mathbf{A}$. Then

$$\begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix}[\mathbf{X}^T, \mathbf{X}^T] = \begin{bmatrix}\mathbf{X}\mathbf{X}^T & \mathbf{X}\mathbf{Y}^T\\ \mathbf{Y}\mathbf{X}^T & \mathbf{Y}\mathbf{Y}^T\end{bmatrix}$$

is SPSD. Thus,

$$\frac{1}{2}\Big[\|\mathbf{X}\mathbf{X}^T\| + \|\mathbf{Y}\mathbf{Y}^T\|\Big] \geq \|\mathbf{A}\|.$$

When $\mathbf{X} \triangleq \hat{\mathbf{X}} = \mathbf{U}_r\mathbf{\Sigma}_r^{1/2}$ and $\mathbf{Y} \triangleq \hat{\mathbf{Y}} = \mathbf{V}_r\mathbf{\Sigma}_r^{1/2}$, it holds that $\|\mathbf{A}\| = \frac{1}{2}[\|\hat{\mathbf{X}}\hat{\mathbf{X}}^T\| + \|\hat{\mathbf{Y}}\hat{\mathbf{Y}}^T\|]$. □

Since $\frac{1}{2}\Big[\|\mathbf{X}\mathbf{X}^T\| + \|\mathbf{Y}\mathbf{Y}^T\|\Big] \geq \sqrt{\|\mathbf{X}\mathbf{X}^T\|}\sqrt{\|\mathbf{Y}\mathbf{Y}^T\|}$,

$$\|\mathbf{A}\| \geq \min_{\mathbf{X},\mathbf{Y}:\mathbf{X}\mathbf{Y}^T=\mathbf{A}} \sqrt{\|\mathbf{X}\mathbf{X}^T\|}\sqrt{\|\mathbf{Y}\mathbf{Y}^T\|}.$$

When taking $\hat{\mathbf{X}} = \mathbf{U}_r\mathbf{\Sigma}_r^{1/2}\mathbf{V}_r^T$ and $\hat{\mathbf{Y}} = \mathbf{V}_r\mathbf{\Sigma}_r^{1/2}\mathbf{V}_r^T$, one has

$$\|\mathbf{A}\| = \sqrt{\|\hat{\mathbf{X}}\hat{\mathbf{X}}^T\|}\sqrt{\|\hat{\mathbf{Y}}\hat{\mathbf{Y}}^T\|}.$$

This thus leads us to the following proposition.

Proposition 7.6. Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, then the following holds

$$\|\mathbf{A}\| = \min_{\mathbf{X},\mathbf{Y}:\mathbf{X}\mathbf{Y}^T=\mathbf{A}} \sqrt{\|\mathbf{X}\mathbf{X}^T\|}\sqrt{\|\mathbf{Y}\mathbf{Y}^T\|}.$$

Accordingly, the following inequality hods:

$$\|\mathbf{X}\mathbf{Y}^T\| \leq \|\mathbf{X}\mathbf{X}^T\|^{1/2}\|\mathbf{Y}\mathbf{Y}^T\|^{1/2}. \tag{7.1}$$

This is a form of the Cauchy-Schwarz inequality under the unitarily invariant norms.

As a corollary of Proposition 7.5, the following proposition immediately follows. Moreover, this proposition was widely used in matrix completion problems, because an optimization problem regularized by the Frobenius norm is solved more easily than that regularized by the nuclear norm [Hastie et al., 2014].

Proposition 7.7. [Srebro et al., 2004, Mazumder et al., 2010] Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, then the following holds

$$\|\mathbf{A}\|_* = \min_{\mathbf{X},\mathbf{Y}:\mathbf{X}\mathbf{Y}^T=\mathbf{A}} \frac{1}{2}\Big\{\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2\Big\}.$$

If $\operatorname{rank}(\mathbf{A}) = k \leq \min\{m, n\}$, the minimum above is attained at some rank decomposition.

The following theorem shows that the Frobenius norm has a so-called matrix-Pythagoras' property. However, for other Schatten norms, there needs a strong condition to make the property hold.

Theorem 7.11. Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$. If $\mathbf{A}\mathbf{B}^T = \mathbf{0}$ or $\mathbf{A}^T\mathbf{B} = \mathbf{0}$, then

$$\|\mathbf{A}+\mathbf{B}\|_F^2 = \|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2,$$

$$\max\{\|\mathbf{A}\|_2^2, \|\mathbf{B}\|_2^2\} \leq \|\mathbf{A}+\mathbf{B}\|_2^2 \leq \|\mathbf{A}\|_2^2 + \|\mathbf{B}\|_2^2.$$

If both $\mathbf{A}\mathbf{B}^T = \mathbf{0}$ and $\mathbf{A}^T\mathbf{B} = \mathbf{0}$ are satisfied, then

$$\|\!|\mathbf{A}+\mathbf{B}|\!\|_p^p = \|\!|\mathbf{A}|\!\|_p^p + \|\!|\mathbf{B}|\!\|_p^p$$

for $1 \leq p < \infty$ and $\|\mathbf{A}+\mathbf{B}\|_2 = \max\{\|\mathbf{A}\|_2, \|\mathbf{B}\|_2\}$.

Proof. Since $(\mathbf{A}+\mathbf{B})^T(\mathbf{A}+\mathbf{B}) = \mathbf{A}^T\mathbf{A} + \mathbf{B}^T\mathbf{B}$ when $\mathbf{A}^T\mathbf{B} = \mathbf{0}$, the Pythagorean property for the Frobenius norm is obvious. As for the spectral norm, it is easily seen that

$$\begin{aligned}
\|\mathbf{A}+\mathbf{B}\|_2^2 &= \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^T(\mathbf{A}+\mathbf{B})^T(\mathbf{A}+\mathbf{B})\mathbf{x} \\
&= \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^T(\mathbf{A}^T\mathbf{A}+\mathbf{B}^T\mathbf{B})\mathbf{x} \\
&\leq \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} + \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^T\mathbf{B}^T\mathbf{B}\mathbf{x} \\
&= \|\mathbf{A}\|_2^2 + \|\mathbf{B}\|_2^2.
\end{aligned}$$

Let the condensed SVDs of $\mathbf{A}$ and $\mathbf{B}$ be $\mathbf{A} = \mathbf{U}_A\mathbf{\Sigma}_A\mathbf{V}_A^T$ and $\mathbf{B} = \mathbf{U}_B\mathbf{\Sigma}_B\mathbf{V}_B^T$. If $\mathbf{A}^T\mathbf{B} = \mathbf{0}$ and $\mathbf{A}\mathbf{B}^T = \mathbf{0}$, then $\mathbf{V}_A^T\mathbf{V}_B = \mathbf{0}$ and $\mathbf{U}_A^T\mathbf{U}_B = \mathbf{0}$. Note that

$$\mathbf{A}+\mathbf{B} = [\mathbf{U}_A, \mathbf{U}_B]\begin{bmatrix}\mathbf{\Sigma}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_B\end{bmatrix}\begin{bmatrix}\mathbf{V}_A^T \\ \mathbf{V}_B^T\end{bmatrix}$$

is the condensed SVD of $\mathbf{A}+\mathbf{B}$. So the nonzero singular values of $\mathbf{A}+\mathbf{B}$ consist of those of $\mathbf{A}$ and of $\mathbf{B}$. The theorem accordingly follows. □

Let us end this chapter by showing a relationship among the matrix operator, matrix vectorization, and unitarily invariant norms.

Theorem 7.12. Let f be a matrix norm on $\mathbb{R}^{m\times n}$.

(a) The norm f is both unitarily invariant and operator norm if and only if $f(\mathbf{A}) = \|\mathbf{A}\|_2$ for any $\mathbf{A} \in \mathbb{R}^{m\times n}$. In other words, the spectral norm is only one operator norm that satisfies the self-adjoint property.

(b) Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, $f(\mathbf{A}) \triangleq \|\mathsf{vec}(\mathbf{A})\|$ is unitarily invariant if and only if it is the norm $\gamma\|\mathbf{A}\|_F$ for some $\gamma > 0$.

Proof. The proof of Part (a) can be found in Corollary 5.6.35 of Horn and Johnson [1985]. As for Part (b), it is obvious that the Frobenius norm is both unitarily invariant and vectorization norm. Conversely, given any $\mathbf{A} \in \mathbb{R}^{m\times n}$, the vectorization norm is defined as $\|\mathbf{a}\|$ where $\mathbf{a} = \mathsf{vec}(\mathbf{A})$. Recall that the vector $\mathbf{a}$ can be regarded as an $mn \times 1$ matrix. Let $\mathbf{a} = \mathbf{U}_a\mathbf{\Sigma}_a\mathbf{v}_a^T$ be the full SVD of $\mathbf{a}$. Then it is easily seen that $\mathbf{\Sigma}_a = (\|\mathbf{A}\|_F, 0, \ldots, 0)^T$. Moreover, we can set $\mathbf{v}_a = 1$. For any orthonormal matrices $\mathbf{U} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$, we have that $f(\mathbf{U}\mathbf{A}\mathbf{V}^T) = \|\mathsf{vec}(\mathbf{U}\mathbf{A}\mathbf{V}^T)\| = \|(\mathbf{V}\otimes\mathbf{U})\mathsf{vec}(\mathbf{A})\| = \|\mathbf{a}\|$ due to the unitary invariance. Moreover, we have that $\|\mathbf{a}\| = \|\mathbf{\Sigma}_a\| = \|\mathbf{A}\|_F\|(1, 0, \ldots, 0)\|$. Letting $\gamma = \|(1, 0, \ldots, 0)\| > 0$, we complete the proof. Notice that if the norm is normalized, then $\gamma = 1$. □

8

Subdifferentials of Unitarily Invariant Norms

In the previous chapters, we have used matrix differential calculus. Let $f : \mathbb{R}^{m\times n} \to \mathbf{R}$. We have discussed the gradient and Hessian of f w.r.t. $\mathbf{X} \in \mathbb{R}^{m\times n}$. Especially, the function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is defined as a trace function. Such a function is differentiable. In this chapter we consider f to be a unitarily invariant norm.

Norm functions are not necessarily differentiable. For example, the spectral norm and nuclear norm are not differentiable. But norm functions are convex and continuous, so we can resort to theory of subdifferentials [Rockafellar, 1970, Borwein and Lewis, 2006]. Indeed, the subdifferentials of unitarily invariant norms have been studied by Watson [1992] and Lewis [2003].

Using the properties of unitarily invariant norms and the SVD theory, we present directional derivatives and subdifferentials of unitarily invariant norms. As two special cases, we report the subdifferentials of the spectral norm and nuclear norm. These two norms have been widely used in machine learning such as matrix low rank approximation. We illustrate applications of the subdifferentials in optimization problems regularized by either the spectral norm or the nuclear norm. We also study the use of the subdifferentials of unitarily invariant norms in solving least squares estimation problems, whose loss function is defined as any unitarily invariant norm.

8.1 Subdifferentials

Let $\| \cdot \|$ be a given norm on $\mathbf{R}^{m\times n}$, and $\mathbf{A}$ be a given matrix in $\mathbb{R}^{m\times n}$. The subdifferential, a set of subgradients, of $\|\mathbf{A}\|$ is defined as

$$\Big\{\mathbf{G} \in \mathbb{R}^{m\times n} : \|\mathbf{B}\| \geq \|\mathbf{A}\| + \mathrm{tr}((\mathbf{B}-\mathbf{A})^T\mathbf{G}) \text{ for all } \mathbf{B} \in \mathbb{R}^{m\times n}\Big\},$$

and denoted by $\partial\|\mathbf{A}\|$. When the norm $\|\cdot\|$ is differentiable, the subgradient degenerates to the gradient. That is, the subdifferential is a singleton. For example, when taking the squared Frobenius norm $\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}^T\mathbf{A})$, $\partial\|\mathbf{A}\|_F^2 = \{2\mathbf{A}\}$.

Lemma 8.1. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a given matrix. Then $\mathbf{G} \in \partial\|\mathbf{A}\|$ if and only if $\|\mathbf{A}\| = \mathrm{tr}(\mathbf{G}^T\mathbf{A})$ and $\|\mathbf{G}\|^* \leq 1$.

Proof. The sufficiency is immediate. Now assume that $\mathbf{G} \in \partial\|\mathbf{A}\|$. Then taking $\mathbf{B} = 2\mathbf{A}$ yields $\|\mathbf{G}\| \geq \mathrm{tr}(\mathbf{A}^T\mathbf{G})$ and taking $\mathbf{B} = \frac{1}{2}\mathbf{A}$ yields $\frac{1}{2}\|\mathbf{A}\| \leq \frac{1}{2}\mathrm{tr}(\mathbf{A}^T\mathbf{G})$, which implies that $\|\mathbf{A}\| = \mathrm{tr}(\mathbf{A}^T\mathbf{G})$. Subsequently, $\|\mathbf{B}\| \geq \mathrm{tr}(\mathbf{G}^T\mathbf{B})$ for all matrices $\mathbf{B}$. Thus, the dual norm satisfies

$$\|\mathbf{G}\|^* = \max\{\mathrm{tr}(\mathbf{G}^T\mathbf{B}) : \|\mathbf{B}\| = 1\} \leq 1.$$

□

We especially consider the subdifferential of unitarily invariant norms. Given a unitarily invariant norm $|\!|\!| \cdot |\!|\!|$ on $\mathbb{R}^{m\times n}$, let $p = \min\{m, n\}$. Theorem 7.5 shows there exists a symmetric gauge function $\phi : \mathbb{R}^p \to \mathbb{R}$ associated with the norm $|\!|\!| \cdot |\!|\!|$. Thus, this encourages us to define the subdifferential of unitarily invariant norms via the subdifferential of symmetric gauge functions.

The subdifferential of the symmetric gauge function ϕ at $\mathbf{x} \in \mathbb{R}^p$ is

$$\partial\phi(\mathbf{x}) \triangleq \{\mathbf{z} \in \mathbb{R}^p : \phi(\mathbf{y}) \geq \phi(\mathbf{x}) + (\mathbf{y}-\mathbf{x})^T\mathbf{z} \text{ for all } \mathbf{y} \in \mathbb{R}^p\}.$$

In terms of Lemma 8.1, that $\mathbf{z} \in \partial\phi(\mathbf{x})$ is equivalent to that $\phi(\mathbf{x}) = \mathbf{x}^T\mathbf{z}$ and $\phi^*(\mathbf{z}) \leq 1$. Here ϕ^* is the dual of ϕ (see Definition 7.4) which is a symmetric gauge function for the dual norm $\|\!|\cdot|\!\|^*$. That is, $\phi^*(\boldsymbol{\sigma}(\mathbf{A})) = \|\!|\mathbf{A}|\!\|^*$ (see Theorem 7.6).

Let us return to the subdifferential of unitarily invariant norms. The following lemma gives the directional derivative of $\|\!|\mathbf{A}|\!\|$.

Lemma 8.2. Let $\|\!|\cdot|\!\|$ be a given unitarily invariant norm on $\mathbb{R}^{m\times n}$, and ϕ be the corresponding symmetric gauge function. Then the directional derivative of the norm at $\mathbf{A} \in \mathbb{R}^{m\times n}$ in a direction $\mathbf{R} \in \mathbb{R}^{m\times n}$ is

$$\lim_{t\downarrow 0} \frac{\|\!|\mathbf{A}+t\mathbf{R}|\!\| - \|\!|\mathbf{A}|\!\|}{t} = \max_{\mathbf{d}\in\partial\phi(\boldsymbol{\sigma}(\mathbf{A}))} \sum_{i=1}^{p} d_i \mathbf{u}_i^T \mathbf{R}\mathbf{v}_i = \max_{\mathbf{G}\in\partial\|\!|\mathbf{A}|\!\|} \operatorname{tr}(\mathbf{R}^T\mathbf{G}).$$

Here $p = \min\{m, n\}$, $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$, $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$, $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\sigma}(\mathbf{A}))$, and $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ is a full SVD of $\mathbf{A}$.

Proof. By Lemma 2.5, we immediately have

$$\lim_{t\downarrow 0} \frac{\|\!|\mathbf{A}+t\mathbf{R}|\!\| - \|\!|\mathbf{A}|\!\|}{t} = \max_{\mathbf{G}\in\partial\|\!|\mathbf{A}|\!\|} \operatorname{tr}(\mathbf{R}^T\mathbf{G}).$$

We now prove the first equality. Let $\mathbf{z} = (\mathbf{u}_1^T\mathbf{R}\mathbf{v}_1, \ldots, \mathbf{u}_p^T\mathbf{R}\mathbf{v}_p)^T$. Consider that

$$\|\!|\mathbf{A}+t\mathbf{R}|\!\| = \|\!|\boldsymbol{\Sigma}+t\mathbf{U}^T\mathbf{R}\mathbf{V}|\!\| = \phi(\boldsymbol{\sigma}(\boldsymbol{\Sigma}+t\mathbf{U}^T\mathbf{R}\mathbf{V})) \geq \phi(\boldsymbol{\sigma}(\mathbf{A})+t\mathbf{z})$$

because $\boldsymbol{\sigma}(\mathbf{A})+t\mathbf{z} \prec_w \boldsymbol{\sigma}(\boldsymbol{\Sigma}+t\mathbf{U}^T\mathbf{R}\mathbf{V})$ by Proposition 6.3. Accordingly, we have that

$$\lim_{t\downarrow 0} \frac{\|\!|\mathbf{A}+t\mathbf{R}|\!\| - \|\!|\mathbf{A}|\!\|}{t} \geq \lim_{t\downarrow 0} \frac{\phi(\boldsymbol{\sigma}(\mathbf{A})+t\mathbf{z}) - \phi(\boldsymbol{\sigma}(\mathbf{A}))}{t} = \max_{\mathbf{d}\in\partial\phi(\boldsymbol{\sigma}(\mathbf{A}))} \mathbf{d}^T\mathbf{z}.$$

The above equality follows from Lemma 2.5, when applied to the symmetric gauge function ϕ.

On the other hand, let $\boldsymbol{\sigma}(t) \triangleq \boldsymbol{\sigma}(\mathbf{A}+t\mathbf{R}) = \boldsymbol{\sigma}(\boldsymbol{\Sigma}+t\mathbf{U}^T\mathbf{R}\mathbf{V})$. Now we have

$$\begin{aligned}\frac{\|\!|\mathbf{A}|\!\| - \|\!|\mathbf{A}+t\mathbf{R}|\!\|}{t} &= \frac{\|\!|\mathbf{A}+t\mathbf{R}-t\mathbf{R}|\!\| - \|\!|\mathbf{A}+t\mathbf{R}|\!\|}{t} \\ &= \frac{\phi(\boldsymbol{\sigma}(\boldsymbol{\Sigma}+t\mathbf{U}^T\mathbf{R}\mathbf{V}-t\mathbf{U}^T\mathbf{R}\mathbf{V})) - \phi(\boldsymbol{\sigma}(t))}{t} \\ &\geq \frac{\phi(\boldsymbol{\sigma}(t)-t\mathbf{z}) - \phi(\boldsymbol{\sigma}(t))}{t} \\ &\geq -\mathbf{d}(t)^T\mathbf{z} \ \ [\text{where } \mathbf{d}(t) \in \partial\phi(\boldsymbol{\sigma}(t))].\end{aligned}$$

The above first inequality follows from $\boldsymbol{\sigma}(t) - t\mathbf{z} \prec_w \boldsymbol{\sigma}(\mathbf{A})$. The second inequality is based on the property of the subgradient of ϕ at $\boldsymbol{\sigma}(t)$. Note that ϕ is a continuous function. By the definition of $\partial\phi(\boldsymbol{\sigma}(t))$, it is directly verified that $\lim\limits_{t\to 0+} \mathbf{d}(t) \to \mathbf{d}_0 \in \partial\phi(\boldsymbol{\sigma}(\mathbf{A}))$. Thus,

$$\lim_{t\downarrow 0} \frac{\|\mathbf{A}+t\mathbf{R}\| - \|\mathbf{A}\|}{t} \leq \lim_{t\downarrow 0} \mathbf{d}(t)^T\mathbf{z} = \mathbf{d}_0^T\mathbf{z} \leq \max_{\mathbf{d}\in\partial\phi(\boldsymbol{\sigma}(\mathbf{A}))} \mathbf{d}^T\mathbf{z}.$$

This implies that the first equality also holds. □

Theorem 8.3. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have a full SVD $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, and let $\boldsymbol{\sigma} = \mathsf{dg}(\boldsymbol{\Sigma})$. Then

$$\partial\|\mathbf{A}\| = \mathrm{conv}\Big\{\mathbf{U}\mathbf{D}\mathbf{V}^T : \mathbf{d} \in \partial\phi(\boldsymbol{\sigma}), \mathbf{D} = \mathrm{diag}(\mathbf{d})\Big\}.$$

where ϕ is a symmetric gauge function corresponding to the norm $\|\cdot\|$.

Here the notation "conv$\{\cdot\}$" represents the convex hull of a set, which is closed and convex. If $\mathbf{G} \in \partial\|\mathbf{A}\|$, Theorem 8.3 says that $\mathbf{G}$ can be expressed as

$$\mathbf{G} = \sum_i \alpha_i \mathbf{U}^{(i)}\mathbf{D}^{(i)}(\mathbf{V}^{(i)})^T,$$

where $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$, $\mathbf{A} = \mathbf{U}^{(i)}\boldsymbol{\Sigma}(\mathbf{V}^{(i)})^T$ is a full SVD, $\mathbf{d}_i \in \phi(\boldsymbol{\sigma})$, and $\mathbf{D}^{(i)} = \mathrm{diag}(\mathbf{d}_i)$. According to Corollary 3.3, we can rewrite $\mathbf{G}$ as

$$\mathbf{G} = \sum_i \alpha_i \mathbf{U}\mathbf{Q}^{(i)}\mathbf{D}^{(i)}(\mathbf{P}^{(i)})^T\mathbf{V}^T, \tag{8.1}$$

where $\mathbf{P}^{(i)}$ and $\mathbf{Q}^{(i)}$ are defined as $\mathbf{P}$ and $\mathbf{Q}$ in Corollary 3.3; i.e., they satisfy that $\mathbf{Q}^{(i)}\boldsymbol{\Sigma}(\mathbf{P}^{(i)})^T = \boldsymbol{\Sigma}$ and $(\mathbf{Q}^{(i)})^T\boldsymbol{\Sigma}\mathbf{P}^{(i)} = \boldsymbol{\Sigma}$.

Proof. First of all, we denote the convex hull on the right-hand side by $\mathcal{G}(\mathbf{A})$. Assume that $\mathbf{G} \in \mathcal{G}(\mathbf{A})$. We now prove $\mathbf{G} \in \partial\|\mathbf{A}\|$. Based on Lemma 8.1, we try to show that $\|\mathbf{A}\| = \mathrm{tr}(\mathbf{A}^T\mathbf{G})$ and $\|\mathbf{G}\|^* \leq 1$. In terms of the above discussion, we can express $\mathbf{G}$ as in (8.1). Thus,

$$\begin{aligned}
\mathrm{tr}(\mathbf{A}^T\mathbf{G}) &= \sum_{i=1} \alpha_i \mathrm{tr}(\mathbf{A}^T\mathbf{U}\mathbf{Q}^{(i)}\mathbf{D}^{(i)}(\mathbf{P}^{(i)})^T\mathbf{V}^T) \\
&= \sum_{i=1} \alpha_i \mathrm{tr}((\mathbf{P}^{(i)})^T\boldsymbol{\Sigma}^T\mathbf{Q}^{(i)}\mathbf{D}^{(i)}) = \sum_{i=1} \alpha_i \mathrm{tr}(\boldsymbol{\Sigma}^T\mathbf{D}^{(i)}) \\
&= \sum_{i=1} \alpha_i \mathbf{d}_i^T\boldsymbol{\sigma} = \phi(\boldsymbol{\sigma}) = \|\mathbf{A}\|.
\end{aligned}$$

Additionally,

$$|||\mathbf{G}|||^* = \max_{|||\mathbf{R}|||\le 1} \operatorname{tr}(\mathbf{G}^T\mathbf{R}) = \max_{|||\mathbf{R}|||\le 1} \operatorname{tr}\big(\mathbf{R}^T \sum_{i=1} \alpha_i \mathbf{U}^{(i)}\mathbf{D}^{(i)}(\mathbf{V}^{(i)})^T\big).$$

Since for each i,

$$|||\mathbf{U}^{(i)}\mathbf{D}^{(i)}(\mathbf{V}^{(i)})^T|||^* = |||\mathbf{D}^{(i)}|||^* = \phi^*(\mathbf{d}_i) \le 1,$$

and by Proposition 7.1 we have

$$\operatorname{tr}(\mathbf{R}^T\mathbf{U}^{(i)}\mathbf{D}^{(i)}(\mathbf{V}^{(i)})^T) \le |||\mathbf{R}||| \times |||\mathbf{U}^{(i)}\mathbf{D}^{(i)}(\mathbf{V}^{(i)})^T|||^* \le |||\mathbf{R}|||.$$

Thus, $|||\mathbf{G}|||^* \le 1$. In summary, we have $\mathbf{G} \in \partial|||\mathbf{A}|||$.

Conversely, assume that $\mathbf{G} \in \partial|||\mathbf{A}|||$ but $\mathbf{G} \notin \mathcal{G}(\mathbf{A})$. Then by the well-known separation theorem [Borwein and Lewis, 2006, see Theorem 1.1.1] there exists a matrix $\mathbf{R} \in \mathbb{R}^{m\times n}$ such that

$$\operatorname{tr}(\mathbf{R}^T\mathbf{X}) < \operatorname{tr}(\mathbf{R}^T\mathbf{G}) \;\text{ for all }\; \mathbf{X} \in \mathcal{G}(\mathbf{A}).$$

This implies that

$$\max_{\mathbf{d}\in\partial\phi(\boldsymbol{\sigma})} \sum_{i=1} d_i \mathbf{u}_i^T\mathbf{R}\mathbf{v}_i = \max_{\mathbf{X}\in\mathcal{G}(\mathbf{A})} \operatorname{tr}(\mathbf{R}^T\mathbf{X}) < \max_{\mathbf{G}\in\partial|||\mathbf{A}|||} \operatorname{tr}(\mathbf{R}^T\mathbf{G}).$$

This contradicts with Lemma 8.2. Thus, the theorem follows. □

We are especially interested in the spectral norm $\|\cdot\|_2$ and the nuclear norm $\|\cdot\|_*$. As corollaries of Theorem 8.3, we have the following the results.

Corollary 8.4. Let $\mathbf{A}$ have rank $r \le p = \min\{m, n\}$ and $\mathbf{A} = \mathbf{U}_r\boldsymbol{\Sigma}_r\mathbf{V}_r^T$ be a condensed SVD. Then the subdifferential of $\|\mathbf{A}\|_*$ is give as

$$\partial\|\mathbf{A}\|_* = \Big\{\mathbf{U}_r\mathbf{V}_r^T + \mathbf{W} : \mathbf{W} \in \mathbb{R}^{m\times n} \text{ s.t. } \mathbf{U}_r^T\mathbf{W} = \mathbf{0}, \mathbf{W}\mathbf{V}_r = \mathbf{0}, \|\mathbf{W}\|_2 \le 1\Big\}.$$

Proof. For the nuclear norm, the corresponding symmetric gauge function is $\phi(\boldsymbol{\sigma}) = \|\boldsymbol{\sigma}\|_1 = \sum_{i=1}^p \sigma_i$. Moreover,

$$\partial\|\boldsymbol{\sigma}\|_1 = \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}\|_\infty \le 1 \text{ and } u_i = 1 \text{ for } i = 1, \ldots, r\}.$$

Let $\mathbf{G} \in \partial\|\mathbf{A}\|_*$. By Theorem 8.3 and Corollary 3.3, we have

$$\begin{aligned}\mathbf{G} &= \sum_{i=1} \alpha_i \mathbf{U}\mathbf{Q}^{(i)}\mathbf{D}^{(i)}(\mathbf{P}^{(i)})^T\mathbf{V}^T \\ &= \mathbf{U}_r\mathbf{V}_r^T + \sum_{i=1} \alpha_i \mathbf{U}_{-r}\mathbf{Q}_0^{(i)}\mathbf{D}_{-r}^{(i)}(\mathbf{P}_0^{(i)})^T\mathbf{V}_{-r}^T,\end{aligned}$$

where the $\alpha_i \geq 0$ and $\sum_{i=1} \alpha_i = 1$, $\mathbf{D}^{(i)} = \mathsf{dg}(\mathbf{d}_i)$, $\mathbf{d}_i \in \partial\phi(\boldsymbol{\sigma})$, and $\mathbf{D}_{-r}^{(i)}$ is the last $(m-r) \times (n-r)$ principal submatrix of $\mathbf{D}^{(i)}$. Here $\mathbf{Q}^{(i)} \in \mathbb{R}^{m\times m}$, $\mathbf{P}^{(i)} \in \mathbb{R}^{n\times n}$, $\mathbf{Q}_0^{(i)} \in \mathbb{R}^{(m-r)\times(m-r)}$, and $\mathbf{P}_0^{(i)} \in \mathbb{R}^{(n-r)\times(n-r)}$ are orthonormal matrices, which are defined in Corollary 3.3. Let

$$\mathbf{W} \triangleq \mathbf{U}_{-r}\Big[\sum_{i=1} \alpha_i \mathbf{Q}_0^{(i)} \mathbf{D}_{-r}^{(i)} (\mathbf{P}_0^{(i)})^T\Big] \mathbf{V}_{-r}^T. \tag{8.2}$$

Obviously, $\mathbf{U}_r^T \mathbf{W} = \mathbf{0}$ and $\mathbf{W}\mathbf{V}_r = \mathbf{0}$. Moreover,

$$\|\mathbf{W}\|_2 \leq \sum_{i=1} \alpha_i \|\mathbf{D}_{-r}^{(i)}\|_2 \leq 1.$$

We can also see that any matrix $\mathbf{W}$ satisfying the above three conditions always has an expression as in (8.2). □

Corollary 8.5. Let the largest singular value σ_1 of $\mathbf{A} \in \mathbb{R}^{m\times n}$ have multiplicity t, and $\mathbf{U}_t$ and $\mathbf{V}_t$ consist of the first t columns of $\mathbf{U}$ and $\mathbf{V}$ respectively. Then

$$\partial\|\mathbf{A}\|_2 = \Big\{\mathbf{U}_t \mathbf{H} \mathbf{V}_t^T : \mathbf{H} \in \mathbb{R}^{t\times t} \text{ s.t. } \mathbf{H} \text{ is SPSD}, \operatorname{tr}(\mathbf{H}) = 1\Big\}.$$

Proof. The corresponding symmetric gauge function is $\phi(\boldsymbol{\sigma}) = \|\boldsymbol{\sigma}\|_\infty$, and its subdifferential is

$$\partial\|\boldsymbol{\sigma}\|_\infty = \operatorname{conv}\{\mathbf{e}_i : \ i = 1, \ldots, t\},$$

where $\mathbf{e}_i$ is the ith column of the identity matrix. It then follows from Theorem 8.3 that for any $\mathbf{G} \in \partial\|\mathbf{A}\|_2$, it can be written as

$$\mathbf{G} = \sum_{i=1} \alpha_i \mathbf{U}_t \mathbf{Q}^{(i)} \mathbf{D}_t^{(i)} (\mathbf{Q}^{(i)})^T \mathbf{V}_t^T,$$

where the $\alpha_i \geq 0$ and $\sum_{i=1} \alpha_i = 1$, and $\mathbf{Q}^{(i)}$ is an arbitrary $t \times t$ orthonormal matrix (see Theorem 3.2). Here $\mathbf{D}^i = \mathsf{dg}(\mathbf{d}_i)$, $\mathbf{d}_i \in \partial\phi(\boldsymbol{\sigma})$, and $\mathbf{D}_t^{(i)}$ is the first $t \times t$ principal submatrix of $\mathbf{D}^{(i)}$. Let

$$\mathbf{H} = \sum_{i=1} \alpha_i \mathbf{Q}^{(i)} \mathbf{D}_t^{(i)} (\mathbf{Q}^{(i)})^T, \tag{8.3}$$

which is SPSD and satisfies $\operatorname{tr}(\mathbf{H}) = 1$. Conversely, any SPSD matrix $\mathbf{H}$ satisfying $\operatorname{tr}(\mathbf{H}) = 1$ can be always expressed as the form of (8.3). □

8.2 Applications

In this section we present several examples to illustrate the application of the subdifferential of unitarily invariant norms in solving an optimization problem regularized by a unitarily invariant norm or built on any unitarily invariant norm loss.

Example 8.1. Given a nonzero matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, consider the following optimization problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} \; f(\mathbf{X}) \triangleq \frac{1}{2}\|\mathbf{X}-\mathbf{A}\|_F^2 + \tau\|\mathbf{X}\|_*, \tag{8.4}$$

where $\tau > 0$ is a constant. Clearly, the problem is convex in $\mathbf{X}$. This problem is a steppingstone of matrix completion. Let $\mathbf{A} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T$ be a given condensed SVD of $\mathbf{A}$, and define

$$\hat{\mathbf{X}} = \mathbf{U}_r[\mathbf{\Sigma}_r - \tau\mathbf{I}_r]_+\mathbf{V}_r,$$

where $[\mathbf{\Sigma}_r - \tau\mathbf{I}_r]_+ = \mathrm{diag}([\sigma_1-\tau]_+, \ldots, [\sigma_r-\tau]_+)$ and $[z]_+ = \max(z, 0)$. Now it can be directly checked that

$$\partial f(\hat{\mathbf{X}}) = \hat{\mathbf{X}} - \mathbf{A} + \tau\partial\|\hat{\mathbf{X}}\|.$$

Assume that the first k singular values σ_i are greater than τ. Then,

$$\frac{1}{r}(\mathbf{A}-\hat{\mathbf{X}}) = \mathbf{U}_k\mathbf{V}_k^T + \frac{1}{\tau}\mathbf{U}_{k+1:r}\mathrm{diag}(\sigma_{k+1}, \ldots, \sigma_r)\mathbf{V}_{k+1:r}^T,$$

which belongs to $\partial\|\hat{\mathbf{X}}\|$. In other words, $\mathbf{0} \in \partial f(\hat{\mathbf{X}})$ (see Corollary 8.4). Thus, $\hat{\mathbf{X}}$ is a minimizer of the optimization problem. It is called the singular value thresholding (SVT) operator [Cai et al., 2010]. We can see that the parameter τ controls the rank of the matrix $\hat{\mathbf{X}}$ and the problem is able to yield a low rank solution to the matrix $\mathbf{X}$. That is, $\hat{\mathbf{X}}$ is a low rank approximation to the matrix $\mathbf{A}$.

Example 8.2. Given a nonzero matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, consider the following optimization problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} \; f(\mathbf{X}) \triangleq \frac{1}{2}\|\mathbf{X}-\mathbf{A}\|_F^2 + \tau\|\mathbf{X}\|_2, \tag{8.5}$$

where $\tau > 0$ is a constant. Also, this problem is convex in $\mathbf{X}$. Let $\mathbf{A}$ have the k distinct positive singular values $\delta_1 > \delta_2 > \cdots > \delta_k$ among the σ_i, with respective multiplicities $r_1, \ldots, r_k$. Thus, the rank of $\mathbf{A}$ is $r = \sum_{i=1}^{k} r_i$. Let $m_t = \sum_{i=1}^{t} r_i$ and $\mu_t = \sum_{i=1}^{t} r_i \delta_i$ for $t = 1, \ldots, k$. So $m_k = r$ and $\mu_k = \mathrm{tr}(\mathbf{\Sigma}_r) = \sum_{i=1}^{r} \sigma_i$. Assume that $\tau \leq \mu_k$. We now consider two cases.

In the first case, assume $l \in [k-1]$ is the smallest integer such that

$$\sum_{i=1}^{l} r_i(\delta_i - \delta_{l+1}) = \mu_l - \delta_{l+1} m_l > \tau,$$

and hence, $\delta_l \geq \frac{\mu_l - \tau}{m_l} > \delta_{l+1}$. Note that

$$\begin{aligned}\sum_{i=1}^{l+1} r_i(\delta_i - \delta_{l+2}) &= \sum_{i=1}^{l} r_i(\delta_i - \delta_{l+1}) + \sum_{i=1}^{l+1} r_i(\delta_{l+1} - \delta_{l+2}) \\ &> \sum_{i=1}^{l} r_i(\delta_i - \delta_{l+1}) > \tau.\end{aligned}$$

This implies that l is identifiable. Denoting $\delta = \frac{\mu_l - \tau}{m_l}$, we define $\hat{\mathbf{\Sigma}}$ by replacing the first m_l diagonal elements of $\mathbf{\Sigma}_r$ by δ, and then set $\hat{\mathbf{X}} = \mathbf{U}_r \hat{\mathbf{\Sigma}}_r \mathbf{V}_r^T$. Now note that

$$\frac{1}{\tau}(\mathbf{A} - \hat{\mathbf{X}}) = \mathbf{U}_{m_l} \mathbf{H} \mathbf{V}_{m_l}^T,$$

where $\mathbf{H} = \mathrm{diag}\big((\sigma_1 - \delta)/\tau, \ldots, (\sigma_{m_l} - \delta)/\tau\big)$. Clearly, $\mathbf{H}$ is PSD and $\mathrm{tr}(\mathbf{H}) = \sum_{i=1}^{m_l} \frac{\sigma_i - \delta}{\tau} = \sum_{i=1}^{l} \frac{r_l(\delta_i - \delta)}{\tau} = 1$. It follows from Corollary 8.5 that $\frac{1}{\tau}(\mathbf{A} - \hat{\mathbf{X}}) \in \partial \|\hat{\mathbf{X}}\|_2$. Thus, $\hat{\mathbf{X}}$ is a minimizer.

In the second case, otherwise, $\sum_{i=1}^{k-1} r_i(\delta_i - \delta_k) = \mu_{k-1} - m_{k-1}\delta_k \leq \tau \leq \mu_k$. Let $\delta = \frac{\mu_k - \tau}{m_k}$ such that

$$0 \leq \delta \leq \frac{\mu_k - \mu_{k-1} + \delta_k m_{k-1}}{m_k} = \delta_k.$$

Define $\hat{\mathbf{X}} = \mathbf{U}_r \delta \mathbf{I}_r \mathbf{V}^T$. Then

$$\frac{1}{\tau}(\mathbf{A} - \hat{\mathbf{X}}) = \frac{1}{\tau} \mathbf{U}_r (\mathbf{\Sigma}_r - \delta \mathbf{I}_r) \mathbf{V}_r^T.$$

Since $\frac{1}{\tau}(\mathbf{\Sigma}_r - \delta \mathbf{I}_r)$ is PSD and $\frac{1}{\tau}\mathrm{tr}(\mathbf{\Sigma}_r - \delta \mathbf{I}_r) = 1$, we obtain $\mathbf{0} \in \partial f(\hat{\mathbf{X}})$. This implies that $\hat{\mathbf{X}}$ is a minimizer of the problem.

As we have seen, the minimizer $\hat{\mathbf{X}}$ has the same rank with $\mathbf{A}$. Thus, the problem in (8.5) can not give a low-rank solution. However, this problem makes the singular values of $\hat{\mathbf{X}}$ more well-conditioned because the top singular values decay to δ. Thus, we call it a singular value averaging (SVA) operator.

Example 8.3. Given a nonzero matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, consider the following convex optimization problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} f(\mathbf{X}) \triangleq \|\mathbf{X}-\mathbf{A}\|_2 + \tau\|\mathbf{X}\|_*, \tag{8.6}$$

where $\tau > 0$ is a constant. In the above model the loss function and regularization term are respectively defined as the spectral norm and the nuclear norma, which are mutually dual. Moreover, this model can be regarded as a parallel version of the Dantzig selector [Candès and Tao, 2007]. Thus, this model might be potentially interesting.

Let $\mathbf{A} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T$ be a condensed SVD. Assume that $r\tau > 1$. Assume there are the k distinct positive singular values $\delta_1 > \delta_2 > \cdots > \delta_k$ among the σ_i, with respective multiplicities $r_1, \ldots, r_k$. Let $m_t = \sum_{i=1}^t r_i$ for $t = 1, \ldots, k$.

Let $l \in [k]$ be the smallest integer such that $m_l\tau \geq 1 > m_{l-1}\tau$. Define $\hat{\mathbf{X}} = \mathbf{U}_r[\mathbf{\Sigma}_r - \delta_l\mathbf{I}_r]_+\mathbf{V}_r^T = \mathbf{U}_{m_{l-1}}\mathrm{diag}(\sigma_1-\delta_l, \ldots, \sigma_{m_{l-1}}-\delta_l)\mathbf{V}_{m_{l-1}}^T$. Then $\mathbf{A}-\hat{\mathbf{X}}$ has the maximum singular value δ_l with multiplicity m_l. It follows from Corollaries 8.4 and 8.5 that

$$\partial\|\hat{\mathbf{X}}\|_* = \Big\{\mathbf{U}_{m_{l-1}}\mathbf{V}_{m_{l-1}}^T + \mathbf{W} : \mathbf{W}^T\mathbf{U}_{m_{l-1}} = \mathbf{0}, \mathbf{W}\mathbf{V}_{m_{l-1}} = \mathbf{0}, \|\mathbf{W}\|_2 \leq 1\Big\}$$

and

$$\partial\|\mathbf{A}-\hat{\mathbf{X}}\|_2 = \Big\{-\mathbf{U}_{m_l}\mathbf{H}\mathbf{V}_{m_l}^T : \mathbf{H} \text{ is PSD}, \mathrm{tr}(\mathbf{H}) = 1\Big\}.$$

Take $\mathbf{W}_0 = \mathbf{U}_{[m_{l-1}+1:m_l]}\frac{(1-m_{l-1}\tau)}{r_l\tau}\mathbf{I}_{r_l}\mathbf{V}_{[m_{l-1}+1:m_l]}^T$. Note that $\mathbf{W}_0\mathbf{V}_{m_{l-1}} = 0$, $\mathbf{W}_0^T\mathbf{U}_{m_{l-1}} = \mathbf{0}$, and $\|\mathbf{W}_0\|_2 = \frac{(1-m_{l-1}\tau)}{r_l\tau} \leq 1$ due to $m_{l-1}\tau + r_l\tau = m_l\tau \geq 1$ and $m_{l-1}\tau < 1$. Hence,

$$\tau\partial\|\hat{\mathbf{X}}\|_* \ni \tau(\mathbf{U}_{m_{l-1}}\mathbf{V}_{m_{l-1}}^T + \mathbf{W}_0) = \mathbf{U}_{m_l}\mathbf{H}_0\mathbf{V}_{m_l}^T,$$

where $\mathbf{H}_0 = \tau(\mathbf{I}_{m_{l-1}} \oplus \frac{(1-m_{l-1}\tau)}{r_l\tau}\mathbf{I}_{r_l})$. Clearly, $\mathbf{H}_0$ is PSD and $\mathrm{tr}(\mathbf{H}_0) = 1$. Thus,

$$-\mathbf{U}_{m_l}\mathbf{H}_0\mathbf{V}_{m_l}^T \in \partial\|\mathbf{A}-\hat{\mathbf{X}}\|_2.$$

As a result, $\mathbf{0} \in \partial\|\mathbf{A} - \hat{\mathbf{X}}\|_2 + \tau\partial\|\hat{\mathbf{X}}\|_*$. Consequently, $\hat{\mathbf{X}}$ is a minimizer of the problem in (8.6). Compared with SVT in the model (8.4) which uses the tuning parameter τ as the thresholding value, the current model uses δ_l as the thresholding value.

We also consider the following convex optimization problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} \ f(\mathbf{X}) \triangleq \|\mathbf{X} - \mathbf{A}\|_* + \frac{1}{\tau}\|\mathbf{X}\|_2. \tag{8.7}$$

Clearly, the minimizer of the problem is $\mathbf{A} - \hat{\mathbf{X}}$ where $\hat{\mathbf{X}}$ is the minimizer of the problem (8.6).

Example 8.4. Finally, we consider the following optimization problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times p}} \ f(\mathbf{X}) \triangleq |\!|\!|\mathbf{AX} - \mathbf{B}|\!|\!|,$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{m\times p}$ are two given matrices. This is a novel matrix low rank approximation problem. We will further discuss this problem in Theorem 9.1 of Chapter 9. Here we are concerned with the use of Theorem 8.3 in solving the problem based on unitarily invariant norm loss functions.

Let $\mathbf{A} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T$ be a condensed SVD of $\mathbf{A}$, and $\mathbf{U}_{-r}$ and $\mathbf{V}_{-r}$ be respective orthonormal complements of $\mathbf{U}_r$ and $\mathbf{V}_r$. Now $\mathbf{B} - \mathbf{AA}^\dagger\mathbf{B} = \mathbf{U}_{-r}\mathbf{U}_{-r}^T\mathbf{B}$. Thus, when taking $\hat{\mathbf{X}} = \mathbf{A}^\dagger\mathbf{B}$, one has

$$\partial f(\hat{\mathbf{X}}) = \mathbf{A}^T\partial|\!|\!|\mathbf{U}_{-r}\mathbf{U}_{-r}^T\mathbf{B}|\!|\!|.$$

Let $\mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0^T = \mathbf{U}_{-r}^T\mathbf{B}$ be a thin SVD of $\mathbf{U}_{-r}^T\mathbf{B}$, $\mathbf{D}$ be a diagonal matrix, and ϕ be a symmetric gauge function associated with the norm $|\!|\!|\cdot|\!|\!|$. It follows from Theorem 8.3 that

$$\partial|\!|\!|\mathbf{U}_{-r}\mathbf{U}_{-r}^T\mathbf{B}|\!|\!| = \mathrm{conv}\{\mathbf{U}_{-r}\mathbf{U}_0\mathbf{D}\mathbf{V}_0^T : \mathbf{U}_0, \mathbf{V}_0, \mathsf{dg}(\mathbf{D}) \in \phi(\mathsf{dg}(\mathbf{\Sigma}_0))\}.$$

Thus, for any $\mathbf{G} \in \partial|\!|\!|\mathbf{U}_{-r}\mathbf{U}_{-r}^T\mathbf{B}|\!|\!|$, it holds that $\mathbf{A}^T\mathbf{G} = \mathbf{0}$. This implies that $\partial f(\hat{\mathbf{X}}) = \{\mathbf{0}\}$. Hence, $\mathbf{0} \in \partial f(\hat{\mathbf{X}})$. This implies that $\hat{\mathbf{X}}$ is a minimizer of the problem. In other words,

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times p}} |\!|\!|\mathbf{AX} - \mathbf{B}|\!|\!| = |\!|\!|\mathbf{AA}^\dagger\mathbf{B} - \mathbf{B}|\!|\!|.$$

9

Matrix Low Rank Approximation

Matrix low rank approximation is very important, because it has received wide applications in machine learning and data mining. On the one hand, many machine learning methods involve computing linear equation systems, matrix decomposition, matrix determinants, matrix inverses, etc. How to compute them efficiently is challenging in big data scenarios. Matrix low rank approximation is a potentially powerful approach for addressing computational challenge. On the other hand, many machine learning tasks can be modeled as matrix low rank approximation problems such as matrix completion, spectral clustering, and multi-task learning.

Approximate matrix multiplication is an inverse process of the matrix low rank approximation problem. Recently, many approaches to approximate matrix multiplication [Drineas et al., 2006a, Sarlos, 2006, Cohen and Lewis, 1999, Magen and Zouzias, 2011, Kyrillidis et al., 2014, Kane and Nelson, 2014] have been developed. Meanwhile, they are used to obtain fast solutions for the ℓ_2 regression and SVD problems [Drineas et al., 2006b, 2011b, Nelson and Nguyên, 2013, Halko et al., 2011, Clarkson and Woodruff, 2013, Martinsson et al., 2011, Woolfe et al., 2008]. This makes matrix low rank approximation also become increasingly popular in the theoretical computer science community [Sarlos, 2006, Drineas et al., 2006a].

In this chapter we first present some important theoretical results in matrix low rank approximation. We then discuss approximate matrix multiplication. In the following chapter we are concerned with large scale matrix approximation. We will study randomized SVD and CUR approximation. They can be also cast into the matrix low rank approximation framework.

9.1 Basic Results

Usually, matrix low rank approximation is formulated as a least squares estimation problem based on the Frobenius norm loss. However, Tropp [2015] pointed out that Frobenius-norm error bounds are not acceptable in most cases of practical interest. He even said "Frobenius-norm error bounds are typically vacuous." Thus, spectral norm as a loss function is also employed. In this chapter, we present several basic results, some of which hold even for every unitarily invariant norm.

Theorem 9.1. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{C} \in \mathbb{R}^{m\times c}$. Then for any $\mathbf{X} \in \mathbb{R}^{c\times n}$ and any unitarily invariant norm $|\!|\!| \cdot |\!|\!|$,

$$|\!|\!|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger\mathbf{A}|\!|\!| \leq |\!|\!|\mathbf{A} - \mathbf{C}\mathbf{X}|\!|\!|.$$

In other words,

$$\mathbf{C}^\dagger\mathbf{A} = \operatorname*{argmin}_{\mathbf{X}\in\mathbb{R}^{c\times n}} |\!|\!|\mathbf{C}\mathbf{X} - \mathbf{A}|\!|\!|. \tag{9.1}$$

As we have seen, Theorem 9.1 was discussed in Example 8.4, where the problem is solved via the subdifferentials of unitarily invariant norms given in Theorem 8.3. Here, we present an alternative proof.

Proof. Let $\mathbf{E}_1 = \mathbf{A} - \mathbf{C}\mathbf{C}^\dagger\mathbf{A}$, $\mathbf{E}_2 = \mathbf{C}\mathbf{C}^\dagger\mathbf{A} - \mathbf{C}\mathbf{X}$, and $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 = \mathbf{A} - \mathbf{C}\mathbf{X}$. Since

$$\mathbf{E}_1^T\mathbf{E}_2 = \mathbf{A}^T(\mathbf{I} - \mathbf{C}\mathbf{C}^\dagger)\mathbf{C}(\mathbf{C}^\dagger\mathbf{A} - \mathbf{X}) = \mathbf{A}^T\mathbf{0}(\mathbf{C}^\dagger\mathbf{A} - \mathbf{X}) = \mathbf{0},$$

we have $\mathbf{E}^T\mathbf{E} = \mathbf{E}_1^T\mathbf{E}_1 + \mathbf{E}_2^T\mathbf{E}_2$, and thus $\lambda_i(\mathbf{E}_1) \leq \lambda_i(\mathbf{E})$. It then follows that $\sigma_i(\mathbf{E}_1) \leq \sigma_i(\mathbf{E})$, and thereby $\boldsymbol{\sigma}(\mathbf{E}_1) \prec_w \boldsymbol{\sigma}(\mathbf{E})$. It then follows from

Theorems 7.4 and 7.5 that

$$\|\mathbf{E}_1\| \;\leq\; \|\mathbf{E}\|$$

for any unitarily invariant norm $\|\cdot\|$. □

Recall that Problem (9.1) gives an extension to the least squares problem (4.1) in Section 4.1. Theorem 9.1 shows that there is an identical solution w.r.t. all unitarily invariant norm errors. The following theorem shows the solution of a more complicated problem. However, the theorem holds only for the Frobenius norm loss.

Theorem 9.2. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{C} \in \mathbb{R}^{m\times c}$, and $\mathbf{R} \in \mathbb{R}^{r\times n}$. Then for all $\mathbf{X} \in \mathbb{R}^{c\times r}$,

$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}\mathbf{R}^{\dagger}\mathbf{R}\|_F \leq \|\mathbf{A} - \mathbf{C}\mathbf{X}\mathbf{R}\|_F.$$

Equivalently, $\mathbf{X}^{\star} = \mathbf{C}^{\dagger}\mathbf{A}\mathbf{R}^{\dagger}$ minimizes the following problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{c\times n}} \|\mathbf{C}\mathbf{X}\mathbf{R} - \mathbf{A}\|_F^2. \tag{9.2}$$

Proof. Let $\mathbf{E}_1 = (\mathbf{I}_m - \mathbf{C}\mathbf{C}^{\dagger})\mathbf{A}$, $\mathbf{E}_2 = \mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}(\mathbf{I}_n - \mathbf{R}^{\dagger}\mathbf{R})$, $\mathbf{E}_3 = \mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}\mathbf{R}^{\dagger}\mathbf{R} - \mathbf{C}\mathbf{X}\mathbf{R}$, and $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3$. Then $\mathbf{E}_1 + \mathbf{E}_2 = \mathbf{A} - \mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}\mathbf{R}^{\dagger}\mathbf{R}$ and $\mathbf{E} = \mathbf{A} - \mathbf{C}\mathbf{X}\mathbf{R}$. Since $\mathbf{E}_1^T\mathbf{E}_2 = \mathbf{0}$, $\mathbf{E}_3\mathbf{E}_2^T = \mathbf{0}$, $\mathbf{E}_1^T\mathbf{E}_3 = \mathbf{0}$, it follows from the matrix Pythagorean theorem that

$$\|\mathbf{E}\|_F^2 = \|\mathbf{E}_1\|_F^2 + \|\mathbf{E}_2\|_F^2 + \|\mathbf{E}_3\|_F^2 = \|\mathbf{E}_1 + \mathbf{E}_2\|_F^2 + \|\mathbf{E}_3\|_F^2.$$

Thus, $\|\mathbf{E}_1 + \mathbf{E}_2\|_F^2 \leq \|\mathbf{E}\|_F^2$. □

Theorem 9.3. [Eckart and Young, 1936, Mirsky, 1960] Given an $m\times n$ real matrix $\mathbf{A}$ of rank r ($\leq \min\{m, n\}$), let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the full SVD of $\mathbf{A}$. Define $\mathbf{A}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T$, where $\mathbf{U}_k$ and $\mathbf{V}_k$ consist of the first k columns of $\mathbf{U}$ and $\mathbf{V}$ respectively, and $\mathbf{\Sigma}_k$ is the first $k\times k$ principal submatrix of $\mathbf{\Sigma}$. Then for all $m\times n$ real matrices $\mathbf{B}$ of rank at most k,

$$\|\mathbf{A} - \mathbf{A}_k\| \leq \|\mathbf{A} - \mathbf{B}\|$$

holds for all unitarily invariant norm $\|\cdot\|$. In other words,

$$\mathbf{A}_k = \operatorname*{argmin}_{\mathbf{B}\in\mathbb{R}^{m\times n},\mathrm{rank}(\mathbf{B})\leq k} \|\mathbf{A} - \mathbf{B}\|. \tag{9.3}$$

Theorem 9.3 shows that the rank k truncated SVD produces the best rank k approximation. The theorem was originally proposed by Eckart and Young [1936] under the setting of the Frobenius norm, and generalized to any unitarily invariant norms by Mirsky [1960].

Proof. For any $m \times n$ real matrix $\mathbf{B}$ of rank at most k, we can write it as $\mathbf{B} = \mathbf{QC}$ where $\mathbf{Q}$ is an $m \times k$ column orthonormal matrix and $\mathbf{C}$ is some $k \times n$ matrix. Thus,

$$|||\mathbf{A} - \mathbf{B}||| = |||\mathbf{A} - \mathbf{QC}||| \geq |||\mathbf{A} - \mathbf{QQ}^T\mathbf{A}||| = |||\mathbf{Q}^{\perp}(\mathbf{Q}^{\perp})^T\mathbf{A}|||,$$

where $\mathbf{Q}^{\perp}$ $(m \times (m{-}k))$ is the orthogonal complement of $\mathbf{Q}$. By Proposition 6.3, we have $\sigma_i(\mathbf{Q}^{\perp}(\mathbf{Q}^{\perp})^T\mathbf{A}) = \sigma_i((\mathbf{Q}^{\perp})^T\mathbf{A}) \geq \sigma_{k+i}$ for $i = 1, \ldots, p-k$. This implies that

$$\boldsymbol{\sigma}(\mathbf{A} - \mathbf{A}_k) = (\sigma_{k+i}, \sigma_p, 0, \ldots, 0)^T \prec_w \boldsymbol{\sigma}(\mathbf{Q}^{\perp}(\mathbf{Q}^{\perp})^T\mathbf{A}).$$

Hence, $|||\mathbf{A} - \mathbf{B}||| \geq |||\mathbf{A} - \mathbf{A}_k|||$. □

The above proof procedure also implies that for all $m \times k$ column orthonormal matrices $\mathbf{Q}$,

$$|||\mathbf{A} - \mathbf{U}_k\mathbf{U}_k^T\mathbf{A}||| \leq |||\mathbf{A} - \mathbf{QQ}^T\mathbf{A}|||$$

holds for every unitarily invariant norm $|||\cdot|||$.

When $k < r$, $\mathbf{A}_k$ is called a truncated SVD of $\mathbf{A}$ and the closest rank-k approximation of $\mathbf{A}$. Note that when the Frobenius norm is used, $\mathbf{A}_k$ is the unique minimizer of the problem in (9.3). However, when other unitarily invariant norms are used, the case does not always hold. For example, let us take the spectral norm. Clearly, if

$$\tilde{\boldsymbol{\Sigma}} = \mathrm{diag}(\sigma_1 - \omega\sigma_{k+1}, \sigma_2 - \omega\sigma_{k+1}, \ldots, \sigma_k - \omega\sigma_{k+1}, 0, \ldots, 0)$$

for any $\omega \in [0, 1]$, then $\mathbf{U}\tilde{\boldsymbol{\Sigma}}\mathbf{V}^T$ is also a minimizer of the corresponding problem.

Theorem 9.4. Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and a column orthonormal matrix $\mathbf{Q} \in \mathbb{R}^{m\times p}$, let $\mathbf{B}_k$ be the rank-k truncated SVD of $\mathbf{Q}^T\mathbf{A}$ for $1 \leq k \leq p$. Then $\mathbf{B}_k$ is an optimal solution of the following problem:

$$\min_{\mathbf{B}\in\mathbb{R}^{l\times n}, \mathrm{rank}(\mathbf{B})\leq k} \|\mathbf{A} - \mathbf{QB}\|_F^2 = \|\mathbf{A} - \mathbf{QB}_k\|_F^2. \tag{9.4}$$

Proof. Note that $(\mathbf{A} - \mathbf{Q}\mathbf{Q}^T\mathbf{A})^T(\mathbf{Q}\mathbf{B} - \mathbf{Q}\mathbf{Q}^T\mathbf{A}) = \mathbf{0}$, so

$$\begin{aligned}\|\mathbf{A} - \mathbf{Q}\mathbf{B}\|_F^2 &= \|\mathbf{A} - \mathbf{Q}\mathbf{Q}^T\mathbf{A}\|_F^2 + \|\mathbf{Q}\mathbf{B} - \mathbf{Q}\mathbf{Q}^T\mathbf{A}|_F^2 \\ &= \|\mathbf{A} - \mathbf{Q}\mathbf{Q}^T\mathbf{A}\|_F^2 + \|\mathbf{B} - \mathbf{Q}^T\mathbf{A}|_F^2.\end{aligned}$$

The result of the theorem follows from Theorem 9.3. □

Theorem 9.4 is a variant of Theorem 9.3 and of Theorem 9.1. Unfortunately, $\mathbf{B}_k$ might not be the solution to the above problem in every unitarily invariant norm, even in the spectral norm error. The reason is that the matrix Pythagorean identity hods only for the Frobenius norm (see Theorem 7.11).

However, Tropp [2015] pointed out that Frobenius-norm error bounds are not acceptable in most cases of practical interest. He even said “Frobenius-norm error bounds are typically vacuous” [Tropp, 2015]. The following theorem was proposed by Gu [2015], which relates the approximation error in the Frobenius norm to that in the spectral norm.

Theorem 9.5. [Gu, 2015] Given any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, let $p = \min\{m, n\}$ and $\mathbf{B}$ be a matrix with rank at most k such that

$$\|\mathbf{A} - \mathbf{B}\|_F \leq \sqrt{\eta^2 + \sum_{j=k+1}^{p} \sigma_j^2(\mathbf{A})}$$

for some $\eta \geq 0$. Then we must have $\sqrt{\sum_{j=1}^{k}(\sigma_j(\mathbf{A}) - \sigma_j(\mathbf{B}))^2} \leq \eta$ and

$$\|\mathbf{A} - \mathbf{B}\|_2 \leq \sqrt{\eta^2 + \sigma_{k+1}^2(\mathbf{A})}.$$

Proof. By Proposition 6.3-(2), we have

$$\sigma_{i+k}(\mathbf{A}) \leq \sigma_i(\mathbf{A} - \mathbf{B}) + \sigma_{k+1}(\mathbf{B}) = \sigma_i(\mathbf{A} - \mathbf{B}) \;\text{ for }\; i \in [p-k]$$

due to $\mathrm{rank}(\mathbf{B}) \leq k$. It then follows that

$$\begin{aligned}\|\mathbf{A} - \mathbf{B}\|_F^2 = \sum_{i=1}^{p} \sigma_i^2(\mathbf{A} - \mathbf{B}) &\geq \sigma_1^2(\mathbf{A} - \mathbf{B}) + \sum_{i=2}^{p-k} \sigma_i^2(\mathbf{A} - \mathbf{B}) \\ &\geq \sigma_1^2(\mathbf{A} - \mathbf{B}) + \sum_{i=2}^{p-k} \sigma_{i+k}^2(\mathbf{A}).\end{aligned}$$

We thus obtain

$$\|\mathbf{A}-\mathbf{B}\|_2^2 = \sigma_1^2(\mathbf{A}-\mathbf{B}) \le \eta^2 + \sigma_{k+1}^2(\mathbf{A}).$$

Additionally, it follows from Theorem 6.5 that

$$\sum_{i=1}^{k}(\sigma_i(\mathbf{A})-\sigma_i(\mathbf{B}))^2 + \sum_{j=k+1}^{p}\sigma_j^2(\mathbf{B}) \le \|\mathbf{A}-\mathbf{B}\|_F^2 \le \eta^2 + \sum_{j=k+1}^{p}\sigma_j^2(\mathbf{A}),$$

which leads to the result. □

Let us apply Theorem 9.5 to Theorem 9.4 to establish a spectral norm error bound. It follows from Theorem 9.4 that

$$\|\mathbf{A}-\mathbf{A}_k\|_F \le \|\mathbf{A}-\mathbf{Q}\mathbf{B}_k\|_F \le \|\mathbf{A}-\mathbf{Q}\mathbf{Q}^T\mathbf{A}_k\|_F.$$

Consider that

$$\begin{aligned}\|\mathbf{A}-\mathbf{Q}\mathbf{Q}^T\mathbf{A}_k\|_F^2 &= \|\mathbf{A}-\mathbf{A}_k+\mathbf{A}_k-\mathbf{Q}\mathbf{Q}^T\mathbf{A}_k\|_F^2\\ &= \|(\mathbf{I}_m-\mathbf{Q}\mathbf{Q}^T)\mathbf{A}_k\|_F^2 + \|\mathbf{A}-\mathbf{A}_k\|_F^2\end{aligned}$$

due to $(\mathbf{A}-\mathbf{A}_k)\mathbf{A}_k^T(\mathbf{I}_m-\mathbf{Q}\mathbf{Q}^T)=\mathbf{0}$. Thus,

$$\|\mathbf{A}-\mathbf{Q}\mathbf{B}_k\|_F^2 \le \|(\mathbf{I}_m-\mathbf{Q}\mathbf{Q}^T)\mathbf{A}_k\|_F^2 + \sum_{i=k+1}^{n}\sigma_i^2(\mathbf{A}).$$

By Theorem 9.5, we have that

$$\|\mathbf{A}-\mathbf{Q}\mathbf{B}_k\|_2^2 \le \|(\mathbf{I}_m-\mathbf{Q}\mathbf{Q}^T)\mathbf{A}_k\|_F^2 + \sigma_{k+1}^2(\mathbf{A}),$$

which can give an error bound in the spectral norm.

9.2 Approximate Matrix Multiplication

Given matrices $\mathbf{A}\in\mathbb{R}^{n\times d}$ and $\mathbf{B}\in\mathbb{R}^{n\times p}$, it is well known that the complexity of computing $\mathbf{A}^T\mathbf{B}$ is $O(dnp)$. Approximate matrix multiplication aims to obtain a matrix $\mathbf{C}\in\mathbb{R}^{d\times p}$ with $o(dnp)$ time complexity such that for a small $\varepsilon>0$,

$$\|\mathbf{A}^T\mathbf{B}-\mathbf{C}\| \le \varepsilon\|\mathbf{A}\|\|\mathbf{B}\|.$$

This shows that approximate matrix multiplication can be viewed as an inverse process of the conventional matrix low rank approximation problem.

Approximate matrix multiplication is a potentially important approach for fast matrix multiplication [Drineas et al., 2006a, Clarkson and Woodruff, 2009, Cohen and Lewis, 1999, Kane and Nelson, 2014, Drineas et al., 2011b, Nelson and Nguyên, 2013, Clarkson and Woodruff, 2013]. It is the foundation of approximate least square methods and matrix low rank approximation methods [Sarlos, 2006, Halko et al., 2011, Kyrillidis et al., 2014, Martinsson et al., 2011, Woolfe et al., 2008, Magdon-Ismail, 2011, Magen and Zouzias, 2011, Cohen and Lewis, 1999, Kane and Nelson, 2014, Drineas et al., 2011b, Nelson and Nguyên, 2013, Clarkson and Woodruff, 2013]. Moreover, it can be also used in large scalable k-means clustering [Cohen et al., 2014], approximate leverage scores [Drineas et al., 2011a], etc.

Most of work for matrix approximations is based on error bounds w.r.t. the Frobenius norm [Drineas et al., 2006a, Sarlos, 2006, Cohen and Lewis, 1999, Kane and Nelson, 2014, Drineas et al., 2011b, Nelson and Nguyên, 2013, Clarkson and Woodruff, 2013]. In contrast, there is a few work based on spectral-norm error bounds [Halko et al., 2011, Kyrillidis et al., 2014, Martinsson et al., 2011, Woolfe et al., 2008, Magdon-Ismail, 2011, Magen and Zouzias, 2011]. As we have mentioned earlier, spectral-norm error bounds are also of great interest.

In approximate matrix multiplication, oblivious subspace embedding matrix is a key ingredient. For example, gaussian matrix and random sign matrix are oblivious matrix. However, leverage score sketching matrix depends on data matrix, hence, it is not an oblivious subspace embedding matrix.

Definition 9.1. [Woodruff, 2014b] Given $\varepsilon > 0$ and $\delta > 0$, let Π be a distribution on $l \times n$ matrices, where l relies on n, d, ε and δ. Suppose that with probability at lest $1 - \delta$, for any fixed $n \times d$ matrix $\mathbf{A}$, a matrix $\mathbf{S}$ drawn from distribution Π is a $(1+\varepsilon)$ ℓ_2-subspace embedding for $\mathbf{A}$, that is, for all $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{SAx}\|_2^2 = (1 \pm \varepsilon)\|\mathbf{Ax}\|_2^2$ with probability $1 - \delta$. Then we call Π an (ε, δ)-oblivious ℓ_2-subspace embedding,

Recently, Cohen et al. [2015] proved optimal approximate matrix multiplication in terms of stable rank by using subspace embedding [Batson et al., 2014].

Theorem 9.6. [Cohen et al., 2015] Given ε, $\delta \in (0, 1/2)$, let $\mathbf{A}$ and $\mathbf{B}$ be two conforming matrices, and Π be a (ε, δ) subspace embedding for the $2\tilde{r}$-dimensional subspace, where $\tilde{r}$ is the maximum of the stable ranks of $\mathbf{A}$ and $\mathbf{B}$. Then,

$$||(\Pi\mathbf{A})^T(\Pi\mathbf{B}) - \mathbf{A}^T\mathbf{B}|| \leq \varepsilon||\mathbf{A}||||\mathbf{B}||$$

holds with at least $1 - \delta$.

To analyze approximate matrix multiplication with the Frobenius error, Kane and Nelson [2014] introduced the JL-moment property.

Definition 9.2. A distribution $\mathcal{D}$ over $\mathbf{R}^{n\times d}$ has the $(\varepsilon, \delta, \ell)$-JL moment property if for all $\mathbf{x} \in \mathbf{R}^d$ with $\|\mathbf{x}\|_2 = 1$,

$$\mathbb{E}_{\Pi\sim\mathcal{D}}\left|\|\Pi\mathbf{x}\|_2^2 - 1\right|^\ell \leq \varepsilon^\ell \cdot \delta$$

Based on the JL-moment property, these is an approximate matrix multiplication method with the Frobenius error.

Theorem 9.7. Given ε, $\delta \in (0, 1/2)$, let $\mathbf{A}$ and $\mathbf{B}$ be two conforming matrices, and Π be a matrix satisfying the $(\varepsilon, \delta, \ell)$-JL moment property for some $\ell \geq 2$. Then,

$$||(\Pi\mathbf{A})^T(\Pi\mathbf{B}) - \mathbf{A}^T\mathbf{B}||_F \leq \varepsilon||\mathbf{A}||_F||\mathbf{B}||_F$$

holds with at least $1 - \delta$.

Note that both the subspace embedding property and the JL moment property have close relationships. More specifically, they can be converted into each other [Kane and Nelson, 2014].

There are other methods, which do not use subspace embedding matrices, in the literature. Magen and Zouzias [2011] gave a method based on columns selection. Bhojanapalli et al. [2015] proposed a new method with sampling and alternating minimization to directly compute a low-rank approximation to the product of two given matrices.

For low-rank matrix approximation in the streaming model, Clarkson and Woodruff [2009] gave the near-optimal space bounds by the sketches. Liberty [2013] came up with a deterministic streaming algorithm, with an improved analysis studied by Ghashami and Phillips [2014] and space lower bound obtained by Woodruff [2014a].

10

Large-Scale Matrix Approximation

In this chapter we discuss fast computational methods of the SVD, kernel methods, and CUR decomposition via randomized approximation. The goal is to make the matrix factorizations fill the use on large scale data matrices.

It is notoriously difficult to compute SVD because the exact SVD of an $m \times n$ matrix takes $\mathcal{O}(mn \min\{m, n\})$ time. Fortunately, many machine learning methods such as latent semantic indexing [Deerwester et al., 1990], spectral clustering [Shi and Malik, 2000], manifold learning [Tenenbaum et al., 2000, Belkin and Niyogi, 2003] are interested in only the top singular value triples. The Krylov subspace method computes the top k singular value triples in $\tilde{\mathcal{O}}(mnk)$ time [Saad, 2011, Musco and Musco, 2015], where the $\tilde{\mathcal{O}}$ notation hides the logarithm factors and the data dependent condition number. If a low precision solution suffices, the time complexity can be even lower. Here we will make main attention on randomized approximate algorithms that demonstrate high scalability. Randomized algorithms are a feasible approach for large scale machine learning models [Rokhlin et al., 2009, Mahoney, 2011, Tu et al., 2014]. In particular, we will consider randomized SVD methods [Halko et al., 2011].

In contrast to the randomized SVD which is based on random projection, the CUR approximation mainly employs column selection. Column selection has been extensively studied in the theoretical computer science (TCS) and numerical linear algebra (NLA) communities. The work in TCS mainly focuses on choosing good columns by randomized algorithms with provable error bounds [Frieze et al., 2004, Deshpande et al., 2006, Drineas et al., 2008, Deshpande and Rademacher, 2010, Boutsidis et al., 2014, Guruswami and Sinop, 2012]. The focus in NLA is then on deterministic algorithms, especially the rank-revealing QR factorizations, that select columns by pivoting rules [Foster, 1986, Chan, 1987, Stewart, 1999, Bischof and Hansen, 1991, Hong and Pan, 1992, Chandrasekaran and Ipsen, 1994, Gu and Eisenstat, 1996, Berry et al., 2005].

10.1 Randomized SVD

All the randomized SVD algorithms essentially have the same idea: first draw a random projection matrix $\mathbf{\Omega} \in \mathbb{R}^{n \times c}$, then form the sketch $\mathbf{C} = \mathbf{A}\mathbf{\Omega} \in \mathbb{R}^{m \times c}$ and compute its orthonormal bases $\mathbf{Q} \in \mathbb{R}^{m \times c}$, and finally compute a rank k matrix $\mathbf{X} \in \mathbb{R}^{c \times n}$ such that $\|\mathbf{A} - \mathbf{Q}\mathbf{X}\|_\xi^2$ is small compared to $\|\mathbf{A} - \mathbf{A}_k\|_\xi^2$. Here $\| \cdot \|_\xi$ denotes either the Frobenius norm or the spectral norm.

The following lemma is the foundation in theoretical analysis of the randomized SVD [Halko et al., 2011, Gu, 2015].

Lemma 10.1. Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a given matrix, and $\mathbf{Z} \in \mathbb{R}^{n \times k}$ be column orthonormal. Let $\mathbf{\Omega} \in \mathbb{R}^{n \times c}$ be any matrix such that $\mathrm{rank}(\mathbf{Z}^T\mathbf{\Omega}) = \mathrm{rank}(\mathbf{Z}) = k$, and define $\mathbf{C} = \mathbf{A}\mathbf{\Omega} \in \mathbb{R}^{m \times c}$. Then

$$\|\mathbf{A} - \Pi_{\mathbf{C},k}^{\xi}(\mathbf{A})\|_\xi^2 \leq \|\mathbf{E}\|_\xi^2 + \|\mathbf{E}\mathbf{\Omega}(\mathbf{Z}^T\mathbf{\Omega})^\dagger\|_\xi^2,$$

where $\mathbf{E} = \mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^T$, and $\Pi_{\mathbf{C},k}^{\xi}(\mathbf{A}) \in \mathbb{R}^{m \times n}$ denotes the best approximation to $\mathbf{A}$ within the column space of $\mathbf{C}$ that has rank at most k w.r.t. the norm $\| \cdot \|_\xi$ loss.

Proof. In terms of definition of $\Pi_{\mathbf{C},k}^{\xi}(\mathbf{A})$, we have

$$\|\mathbf{A} - \Pi_{\mathbf{C},k}^{\xi}(\mathbf{A})\|_\xi^2 \leq \|\mathbf{A} - \mathbf{X}\|_\xi^2$$

for all matrices $\mathbf{X} \in \mathbb{R}^{m\times n}$ of rank at most k in the column space of $\mathbf{C}$. Obviously, $\mathbf{C}(\mathbf{Z}^T\mathbf{\Omega})^\dagger\mathbf{Z}^T$ is such a matrix. Thus,

$$\begin{aligned}\|\mathbf{A} - \Pi^{\xi}_{\mathbf{C},k}(\mathbf{A})\|_\xi^2 &\leq \|\mathbf{A} - \mathbf{C}(\mathbf{Z}^T\mathbf{\Omega})^\dagger\mathbf{Z}^T\|_\xi^2 \\ &= \|\mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^T + \mathbf{A}\mathbf{Z}\mathbf{Z}^T - \mathbf{C}(\mathbf{Z}^T\mathbf{\Omega})^\dagger\mathbf{Z}^T\|_\xi^2 \\ &= \|\mathbf{E} + (\mathbf{A}\mathbf{Z}\mathbf{Z}^T - \mathbf{A})\mathbf{\Omega}(\mathbf{Z}^T\mathbf{\Omega})^\dagger\mathbf{Z}^T\|_\xi^2 \\ &= \|\mathbf{E} + \mathbf{E}\mathbf{\Omega}(\mathbf{Z}^T\mathbf{\Omega})^\dagger\mathbf{Z}^T\|_\xi^2.\end{aligned}$$

Here we use the fact that $\mathbf{Z}^T\mathbf{\Omega}(\mathbf{Z}^T\mathbf{\Omega})^\dagger = \mathbf{I}_k$ because $\mathrm{rank}(\mathbf{Z}^T\mathbf{\Omega}) = k$. Consider that

$$\mathbf{E}\mathbf{\Omega}(\mathbf{Z}^T\mathbf{\Omega})^\dagger\mathbf{Z}^T\mathbf{E}^T = \mathbf{E}\mathbf{\Omega}(\mathbf{Z}^T\mathbf{\Omega})^\dagger\mathbf{Z}^T(\mathbf{A}^T - \mathbf{Z}\mathbf{Z}^T\mathbf{A}^T) = \mathbf{0}.$$

The theorem follows from Theorem 7.11. □

Consider the rank-k truncated SVD $\mathbf{A}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T$. Then we can write $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{A}\mathbf{V}_k\mathbf{V}_k^T + (\mathbf{A} - \mathbf{A}_k).$$

Let $\mathbf{Z} = \mathbf{V}_k$ and $\mathbf{E} = \mathbf{A} - \mathbf{A}_k$ in Lemma 10.1. Then the following theorem is an immediate corollary of Lemma 10.1.

Theorem 10.2. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the full SVD of $\mathbf{A} \in \mathbb{R}^{m\times n}$, fix $k \geq 0$, and let $\mathbf{A}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T$ be the best at most rank k approximation of $\mathbf{A}$. Choose a test matrix $\mathbf{\Omega}$ and construct the sketch $\mathbf{C} = \mathbf{A}\mathbf{\Omega}$. Partition $\mathbf{\Sigma} = \begin{bmatrix}\mathbf{\Sigma}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{-k}\end{bmatrix}$ and $\mathbf{V} = [\mathbf{V}_k, \mathbf{V}_{-k}]$. Define $\mathbf{\Omega}_1 = \mathbf{V}_k^T\mathbf{\Omega}$ and $\mathbf{\Omega}_2 = \mathbf{V}_{-k}^T\mathbf{\Omega}$. Assume that $\mathbf{\Omega}_1$ has full row rank. Then

$$\|(\mathbf{I}_m - \mathbf{C}\mathbf{C}^\dagger)\mathbf{A}\|_\xi^2 \leq \|\mathbf{A} - \Pi^{\xi}_{\mathbf{C},k}(\mathbf{A})\|_\xi^2 \leq \|\mathbf{\Sigma}_{-k}\|_\xi^2 + \|\mathbf{\Sigma}_{-k}\mathbf{\Omega}_2\mathbf{\Omega}_1^\dagger\|_\xi^2.$$

In Lemma 10.1 and Theorem 10.2, the condition $\mathrm{rank}(\mathbf{V}_k^T\mathbf{\Omega}) = \mathrm{rank}(\mathbf{V}_k) = k$ is essential for an effective randomized SVD algorithm. An idealized case for meeting this condition is that $\mathrm{range}(\mathbf{V}_k) \subset \mathrm{range}(\mathbf{\Omega})$. In this case, the randomized SVD degenerates an exact truncated SVD procedure. Thus, the above condition aims to relax this idealized case. Moreover, the key for an effective randomized SVD is to select a test matrix $\mathbf{\Omega}$ such that the condition $\mathrm{rank}(\mathbf{V}_k^T\mathbf{\Omega}) = \mathrm{rank}(\mathbf{V}_k) = k$ holds as much as possible. Lemma 10.1 and Theorem 10.2 are also fundamental in random column selection [Boutsidis et al., 2014].

10.1.1 Randomized SVD: Frobenius Norm Bounds

In this subsection, we describe two randomized SVD algorithms which have $(1+\epsilon)$ relative-error bound.

Random Projection. In order to reduce computational expenses, randomized algorithms [Frieze et al., 2004, Vempala, 2000] have been introduced to truncated SVD and low-rank approximation. The Johnson & Lindenstrauss (JL) transform [Johnson and Lindenstrauss, 1984, Dasgupta and Gupta, 2003] is known to keep isometry in expectation or with high probability. Halko et al. [2011], Boutsidis et al. [2014] used the JL transform for sketching and showed relative-error bounds. However, the Gaussian test matrix is dense and cannot efficiently apply to matrices. Several improvements have been proposed to make the sketching matrix sparser; see the review [Woodruff, 2014b] for the complete list of the literature. In particular, the count sketch [Clarkson and Woodruff, 2013] applies to $\mathbf{A}$ in only $\mathcal{O}(\mathsf{nnz}(\mathbf{A}))$ time and exhibits very similar properties as the JL transform. Specifically, Woodruff [2014b] showed that an $m \times \mathcal{O}(k/\epsilon)$ sketch $\mathbf{C} = \mathbf{A}\mathbf{\Omega}$ can be obtained in $\mathcal{O}(\mathsf{nnz}(\mathbf{A}))$ time and

$$\min_{\mathrm{rank}(\mathbf{X})\leq k} \left\|\mathbf{A} - \mathbf{Q}\mathbf{X}\right\|_F^2 \;\leq\; (1+\epsilon)\left\|\mathbf{A} - \mathbf{A}_k\right\|_F^2 \tag{10.1}$$

holds with high probability.

The Prototype Algorithm. Halko et al. [2011] proposed to directly solve the left-hand side of (10.1), which has closed-form solution $\mathbf{X}^\star = (\mathbf{Q}^T\mathbf{A})_k$. This leads to the prototype algorithm shown in Algorithm 1. The optimality of $\mathbf{X}^\star$ is given in Theorem 9.4.

The prototype algorithm is not time efficient because the matrix product $\mathbf{Q}^T\mathbf{A}$ costs $\mathcal{O}(mnc)$ time, which is not lower than the exact solutions. Nevertheless, the prototype algorithm is still useful in large-scale applications because it is pass-efficient—it goes only two passes through $\mathbf{A}$.

Faster Randomized SVD. The bottleneck of the prototype algorithm is the matrix product in computing $\mathbf{X}^\star$. Notice that (9.4) is a strongly over-determined system, so it can be approximately solved by once more random projection. Let $\mathbf{P} = \mathbf{P}_1\mathbf{P}_2 \in \mathbb{R}^{m\times p}$ be another random projection matrix, where $\mathbf{P}_1$ is a count sketch and $\mathbf{P}_2$ is a JL

Algorithm 1 Randomized SVD: The Prototype Algorithm.

```
Input: a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $m \geq n$, target rank $k$, the size of sketch $c$ where $0 < k \leq c < n$;
Draw a sketching matrix $\mathbf{\Omega} \in \mathbb{R}^{n \times c}$, e.g. a Gaussian test matrix or a count sketch
Compute $\mathbf{C} = \mathbf{A}\mathbf{\Omega} \in \mathbb{R}^{m \times c}$ and its orthonormal bases $\mathbf{Q} \in \mathbb{R}^{m \times c}$;
Compute the rank $k$ truncated SVD: $\mathbf{Q}^T\mathbf{A} \approx \bar{\mathbf{U}}_k \tilde{\mathbf{\Sigma}}_k \tilde{\mathbf{V}}_k^T$;
return $\tilde{\mathbf{U}}_k = \mathbf{Q}\bar{\mathbf{U}}_k$, $\tilde{\mathbf{\Sigma}}_k$, $\tilde{\mathbf{V}}_k$—an approximate rank-$k$ truncated SVD of $\mathbf{A}$.
```

transform matrix. Then we solve

$$\tilde{\mathbf{X}} \; = \; \min_{\mathrm{rank}(\mathbf{X}) \leq k} \; \|\mathbf{P}^T(\mathbf{A} - \mathbf{Q}\mathbf{X})\|_F^2$$

instead of (9.4), and $\tilde{\mathbf{X}}$ has closed-form solution

$$\tilde{\mathbf{X}} = \tilde{\mathbf{R}}^{\dagger}(\tilde{\mathbf{Q}}^T\mathbf{P}^T\mathbf{A})_k,$$

where $\tilde{\mathbf{Q}}\tilde{\mathbf{R}}$ be the economy size QR decomposition of $(\mathbf{P}^T\mathbf{Q}) \in \mathbb{R}^{p \times c}$. Finally, the rank k matrix $\mathbf{Q}\tilde{\mathbf{X}}$ is the obtained approximation to $\mathbf{A}$, and its SVD can be very efficiently computed. Clarkson and Woodruff [2013], Woodruff [2014b] showed that

$$\left\|\mathbf{A} - \mathbf{Q}\tilde{\mathbf{R}}^{\dagger}(\tilde{\mathbf{Q}}^T\mathbf{P}^T\mathbf{A})_k\right\|_F^2 \; \leq \; (1+\epsilon)\, \|\mathbf{A} - \mathbf{A}_k\|_F^2$$

for a large enough p, and the overall time cost is $\mathcal{O}(\mathsf{nnz}(\mathbf{A}) + (m + n)\mathsf{poly}(k/\epsilon))$.

10.1.2 Randomized SVD: Spectral Norm Bounds

The previous section shows that the approximate truncated SVD can be computed highly efficiently, with the $(1+\epsilon)$ Frobenius relative-error guaranteed. The Frobenius norm bound tells that the total elementwise distance is small, but it does not inform us the closeness of their singular vectors. Therefore, we need spectral norm bounds or even stronger principal angle bounds; here we only consider the former. We seek to find an $m \times k$ column orthogonal matrix $\tilde{\mathbf{U}}$ such that

$$\left\|\mathbf{A} - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^T\mathbf{A}\right\|_2^2 \; \leq \; \eta\|\mathbf{A} - \mathbf{A}_k\|_2^2,$$

where η will be specified later.

The Prototype Algorithm. Unlike the Frobenius norm bound, the prototype algorithm is unlikely to attain a constant factor bound (i.e., η is independent of m, n), letting alone the $1+\epsilon$ bound. It is because the lower bounds [Witten and Candès, 2013, Boutsidis et al., 2014] showed that if $\mathbf{\Omega} \in \mathbb{R}^{n\times c}$ in Algorithm 1 is the Gaussian test matrix or any column selection matrix, the order of η must be at least n/c. We apply Gu's theorem [Gu, 2015] (Theorem 9.5) to obtain an $\mathcal{O}(n)$-factor spectral norm bound, and then introduce iterative algorithms with the $(1+\epsilon)$ spectral norm bound.

Let $\tilde{\mathbf{U}}_k$, $\tilde{\mathbf{\Sigma}}_k$, and $\tilde{\mathbf{V}}_k$ be the outputs of Algorithm 1. We have that

$$\begin{aligned} \big\|\mathbf{A} - \tilde{\mathbf{U}}_k\tilde{\mathbf{U}}_k^T\mathbf{A}\big\|_F^2 &\leq \big\|\mathbf{A} - \tilde{\mathbf{U}}_k\tilde{\mathbf{\Sigma}}_k\tilde{\mathbf{V}}_k^T\big\|_F^2 \\ &= \big\|\mathbf{A} - \mathbf{Q}\mathbf{X}^\star\big\|_F^2 \;\leq\; (1+\epsilon)\,\|\mathbf{A}-\mathbf{A}_k\|_F^2, \end{aligned}$$

where the first inequality follows from Theorem 9.1, the equality follows from the definitions, and the second inequality follows from (10.1) provided that $c = \mathcal{O}(k/\epsilon)$ and $\mathbf{\Omega}$ is the Gaussian test matrix or the count sketch. We let $\epsilon = 1$ and $c = \mathcal{O}(k)$ and apply Theorem 9.5 to obtain

$$\begin{aligned} \big\|\mathbf{A} - \tilde{\mathbf{U}}_k\tilde{\mathbf{U}}_k^T\mathbf{A}\big\|_2^2 &\leq \|\mathbf{A}-\mathbf{A}_k\|_2^2 + \|\mathbf{A}-\mathbf{A}_k\|_F^2 \\ &\leq (n-k+1)\|\mathbf{A}-\mathbf{A}_k\|_2^2. \end{aligned} \tag{10.2}$$

Here the second inequality follows from that $\|\mathbf{A}-\mathbf{A}_k\|_F^2 = \sum_{i=k+1}^n \sigma_i^2 \leq (n-k)\sigma_{k+1}^2 = (n-k)\|\mathbf{A}-\mathbf{A}_k\|_2^2$. To this end, we have shown that the prototype algorithm 1 satisfies $\mathcal{O}(n)$-factor spectral norm bound. However, the result itself has little meaning.

The Simultaneous Power Iteration can be used to refine the sketch [Halko et al., 2011, Gu, 2015]. The algorithm is described in Algorithm 2 and analyzed in the following. Let $\mathbf{\Omega} \in \mathbb{R}^{n\times c}$ be a Gaussian test matrix or count sketch and $\mathbf{B} = (\mathbf{A}\mathbf{A}^T)^t\mathbf{A}$. Let us take $\mathbf{B}$ instead of $\mathbf{A}$ as the input of the prototype algorithm 1 and obtain the approximate left singular vectors $\tilde{\mathbf{U}}_k$. It is easy to verify that $\tilde{\mathbf{U}}_k$ is the same to the output of Algorithm 2. We will show that when $t = \mathcal{O}(\frac{\log n}{\epsilon})$,

$$\big\|\mathbf{A} - \tilde{\mathbf{U}}_k\tilde{\mathbf{U}}_k^T\mathbf{A}\big\|_2^2 \;\leq\; (1+\epsilon)\|\mathbf{A}-\mathbf{A}_k\|_2^2. \tag{10.3}$$

To show this result, we need the lemma of Halko et al. [2011].

Algorithm 2 Subspace Iteration Methods.

```
Input: any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the target rank $k$, the size of sketch $c$ where $0 < k \leq c < n$;
Generate an $n \times c$ Gaussian test matrix $\mathbf{\Omega}$ and perform sketching $\mathbf{C}^{(0)} = \mathbf{A\Omega}$;
for $i = 1$ to $t$ do
    Optional: orthogonalize $\mathbf{C}^{(i-1)}$;
    Compute $\mathbf{C}^{(i)} = \mathbf{A}\mathbf{A}^T\mathbf{C}^{(i-1)}$;
end for
The Power Method: orthonalize $\mathbf{C}^{(t)}$ to obtain $\mathbf{Q} \in \mathbb{R}^{m \times c}$;
The Krylov Subspace Method: orthonalize $\mathbf{K} = [\mathbf{C}^{(0)}, \cdots, \mathbf{C}^{(t)}]$ to obtain $\mathbf{Q} \in \mathbb{R}^{m \times (t+1)c}$;
Compute the rank $k$ truncated SVD: $\mathbf{Q}^T\mathbf{A} \approx \bar{\mathbf{U}}_k \tilde{\mathbf{\Sigma}}_k \tilde{\mathbf{V}}_k^T$;
return $\tilde{\mathbf{U}}_k = \mathbf{Q}\bar{\mathbf{U}}_k$, $\tilde{\mathbf{\Sigma}}_k$, $\tilde{\mathbf{V}}_k$—an approximate rank-$k$ truncated SVD of $\mathbf{A}$.
```

Lemma 10.3 (Halko, Martinsson, & Tropp)**.** Let $\mathbf{A}$ be any matrix and $\mathbf{U}$ have orthonormal columns. Then for any positive integer t,

$$\big\|(\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{A}\big\|_2 \;\leq\; \big\|(\mathbf{I} - \mathbf{U}\mathbf{U}^T)(\mathbf{A}\mathbf{A}^T)^t\mathbf{A}\big\|_2^{1/(2t+1)}.$$

By Lemma 10.3, we have that

$$\begin{aligned}\big\|(\mathbf{I} - \tilde{\mathbf{U}}_k\tilde{\mathbf{U}}_k^T)\mathbf{A}\big\|_2^2 &\leq \big\|(\mathbf{I} - \tilde{\mathbf{U}}_k\tilde{\mathbf{U}}_k^T)\mathbf{B}\big\|_2^{2/(2t+1)} \\ &\leq (n-k+1)^{1/(2t+1)}\sigma_{k+1}^{2/(2t+1)}(\mathbf{B}) \\ &= (1+\epsilon)\sigma_{k+1}^2(\mathbf{A}).\end{aligned}$$

Here the second inequality follows from (10.2) and the definitions of $\mathbf{B}$ and $\tilde{\mathbf{U}}_k$, and we show the equality in the following. Let $2t + 1 = \frac{\log(n-k+1)}{0.5\epsilon}$. We have that $\frac{1}{2t+1}\log(n-k+1) = 0.5\epsilon \leq \log(1+\epsilon)$, where the inequality holds for all for all $\epsilon \in [0, 1]$. Taking the exponential of both sides, we have $(n-k+1)^{1/(2t+1)} \leq 1 + \epsilon$. Finally, (10.3) follows from that $\sigma_{k+1}^2(\mathbf{A}) = \|\mathbf{A} - \mathbf{A}_k\|_2^2$.

The Krylov Subspace Method. From Algorithm 2 we can see that the power iteration repeats t times, but only the output of the last iteration $\mathbf{C}^{(t)}$ is used. In fact, the intermediate results $\mathbf{C}^{(0)}, \cdots, \mathbf{C}^{(t)}$ are also useful. The matrix $\mathbf{K} = [\mathbf{C}^{(0)}, \cdots, \mathbf{C}^{(t)}] \in \mathbb{R}^{m \times (t+1)c}$ is well known as the Krylov matrix, and range($\mathbf{K}$) is called the Krylov subspace. We show the Krylov subspace method in Algorithm 2, which differs from simultaneous power iteration in only one line. It turns out that the Krylov subspace method converges much faster than the power iteration [Saad, 2011]. Very recently, Musco and Musco [2015] showed that with $t = \frac{\log n}{\sqrt{\epsilon}}$ power iteration, the $1+\epsilon$ spectral norm bound (10.3) holds with high probability. This result is evidently stronger than the simultaneous power iteration.

It is worth mentioning that the Krylov subspace method described in Algorithm 2 is a simplified version, and it may be instable when t is large. This is because the columns of $\mathbf{C}^{(0)}, \cdots, \mathbf{C}^{(t)}$ tend to be linearly dependent as t grows. In practice, re-orthogonalization or partial re-orthogonalization are employed to prevent the instability from happening [Saad, 2011].

10.2 Kernel Approximation

Kernel methods are important tools in machine learning, computer vision, and data mining [Schölkopf and Smola, 2002, Shawe-Taylor and Cristianini, 2004, Vapnik, 1998, Rasmussen and Williams, 2006]. For example, kernel ridge regression (KRR), Gaussian processes, kernel support vector machine (KSVM), spectral clustering, and kernel principal component analysis (KPCA) are classical nonlinear models for regression, classification, clustering, and dimensionality regression. Unfortunately, the lack of scalability has always been the major drawback of kernel methods. The three steps of most kernel methods—forming the kernel matrix, training, generalization—can all be prohibitive in big-data applications.

Specifically, suppose we are given n training data and m test data, all of d dimension. Firstly, it takes $\mathcal{O}(n^2 d)$ time to form an $n \times n$ kernel matrix $\mathbf{K}$, e.g., the Gaussian RBF kernel matrix. Secondly, the training requires either SVD or matrix inversion of the kernel matrix. For example, spectral clustering, KPCA, Isomap [Tenenbaum et al., 2000], and Laplacian eigenmaps [Belkin and Niyogi, 2003] compute the top k singular vectors of the (normalized) kernel matrix, where k is the number of classes or the target dimensionality. This costs $\mathcal{O}(n^2k)$ time and $\mathcal{O}(n^3)$ memory. Thirdly, to generalize the trained model to the test data, kernel methods such as KRR, KSVM, KPCA cost $\mathcal{O}(nmd)$ time to form an $n \times m$ cross kernel matrix between the training and test data. If m is as large as n, generalization is as challenging as training.

Low rank approximation is the most popular approach to scalable kernel approximation. If we have the low rank approximation $\mathbf{K} \approx \mathbf{C}\mathbf{X}\mathbf{C}^T$, then the approximate eigenvalue decomposition can be immediately obtained by

$$\mathbf{K} \approx \mathbf{C}\mathbf{X}\mathbf{C}^T = \mathbf{U}_C \underbrace{(\mathbf{\Sigma}_C \mathbf{V}_C^T \mathbf{X} \mathbf{V}_C \mathbf{\Sigma}_C)}_{=\mathbf{Z}} \mathbf{U}_C^T = (\mathbf{U}_C\mathbf{U}_Z)\mathbf{\Lambda}_Z(\mathbf{U}_C\mathbf{U}_Z)^T.$$

Here $\mathbf{C} = \mathbf{U}_{\mathbf{C}}\mathbf{\Sigma}_{\mathbf{C}}\mathbf{V}_{\mathbf{C}}^T$ is the SVD and $\mathbf{Z} = \mathbf{U}_{\mathbf{Z}}\mathbf{\Lambda}_{\mathbf{Z}}\mathbf{U}_{\mathbf{Z}}^T$ is the spectral decomposition. Since the tall-and-skinny matrix $\mathbf{U}_C\mathbf{U}_Z$ has orthonormal columns and the diagonal entries of $\mathbf{\Lambda}_Z$ are in the descending order, the leftmost columns of $\mathbf{U}_C\mathbf{U}_Z$ are approximately the top singular vectors of $\mathbf{K}$. This approach only costs $\mathcal{O}(nc^2)$ time, where c is the number of columns of $\mathbf{C}$. Our objective is thereby to find such a low rank approximation.

Difference from Randomized SVD. Why cannot we directly use the randomized SVD to approximate the kernel matrix? The randomized SVD assumes that the matrix is fully observed; unfortunately, this is not true for kernel methods. When the number of data samples is million scale, even forming the kernel matrix is impossible. Therefore, the primary objective of kernel approximation is to avoid forming the whole kernel matrix. The existing random projection methods all require the full observation of the matrix, so random projection is not a feasible option. We must use column selection in the kernel approximation problem.

The Prototype Algorithm. Let $\mathbf{S}$ be an $n \times c$ sketching matrix and let $\mathbf{C} = \mathbf{K}\mathbf{S}$. It remains to find the $c \times c$ intersection matrix $\mathbf{X}$. The most intuitive approach is to minimize the approximation error by

$$\mathbf{X}^\star \ = \ \underset{\mathbf{X}}{\operatorname{argmin}} \ \big\|\mathbf{K} - \mathbf{C}\mathbf{X}\mathbf{C}^T\big\|_F^2 \ = \ \mathbf{C}^\dagger\mathbf{K}(\mathbf{C}^\dagger)^T, \tag{10.4}$$

where the second equality follows from Theorem 9.2. This method was proposed by Halko et al. [2011] for approximating symmetric matrix. Wang et al. [2014a] showed that by randomly sampling $\mathcal{O}(k/\epsilon)$ columns of $\mathbf{K}$ to form $\mathbf{C}$ by a certain algorithm, the approximation is high accurate:

$$\big\|\mathbf{K} - \mathbf{C}\mathbf{X}^{\star}\mathbf{C}^T\big\|_F^2 \;\leq\; (1+\epsilon)\big\|\mathbf{K} - \mathbf{K}_k\big\|_F^2.$$

This upper bound matches the lower bound $c \geq 2k/\epsilon$ up to a constant factor [Wang et al., 2014a]. Unfortunately, the prototype algorithm has two obvious drawbacks. Firstly, to compute the intersection matrix $\mathbf{X}^{\star}$, every entry of $\mathbf{K}$ must be known. As is discussed, it takes $\mathcal{O}(n^2 d)$ time to form the kernel matrix $\mathbf{K}$. Secondly, the matrix multiplication $\mathbf{C}^{\dagger}\mathbf{K}$ costs $\mathcal{O}(n^2 c)$ time. In sum, the prototype algorithm costs $\mathcal{O}(n^2 c + n^2 d)$ time. Although it is substantially faster than the exact solution, the prototype algorithm has the same time complexity as the exact solution.

Faster SPSD Matrix Sketching. Since $\mathbf{C} = \mathbf{K}\mathbf{S}$ has much more rows than columns, the optimization problem (10.4) is strongly overdetermined. Wang et al. [2015b] proposed to use sketching to approximately solve (10.4). Specifically, let $\mathbf{P}$ be a certain $n \times p$ column selection matrix with $p \geq c$ and compute

$$\tilde{\mathbf{X}} \;=\; \operatorname*{argmin}_{\mathbf{X}} \big\|\mathbf{P}^T(\mathbf{K} - \mathbf{C}\mathbf{X}\mathbf{C}^T)\mathbf{P}\big\|_F^2 \;=\; (\mathbf{P}^T\mathbf{C})^{\dagger}(\mathbf{P}^T\mathbf{K}\mathbf{P})(\mathbf{C}^T\mathbf{P})^{\dagger}.$$

In this way, we need only $nc + p^2$ entries of $\mathbf{K}$ to form the approximation $\mathbf{K} \approx \mathbf{C}\tilde{\mathbf{X}}\mathbf{C}^T$. The intersection matrix $\tilde{\mathbf{X}}$ can be computed in $\mathcal{O}(ncd + p^2 d + p^2 c)$ time, given $\mathbf{S}$ and n data points of d dimension. Wang et al. [2015b] devised an algorithm that sets $p = \sqrt{n}c/\sqrt{\epsilon}$ and very efficiently forms the column selection matrix $\mathbf{P}$; and the following error bound holds with high probability:

$$\big\|\mathbf{K} - \mathbf{C}\tilde{\mathbf{X}}\mathbf{C}^T\big\|_F^2 \;\leq\; (1+\epsilon)\,\min_{\mathbf{X}} \big\|\mathbf{K} - \mathbf{C}\mathbf{X}\mathbf{C}^T\big\|_F^2.$$

By this choice of p, the overall time cost is linear in n.

Motivated by the matrix ridge approximation of Zhang [2014], Wang et al. [2014b] proposed a spectral shifting kernel approximation method. When the spectrum of $\mathbf{K}$ decays slowly, the shifting term helps to improve the approximation accuracy and numerical stability.

Wang et al. [2014a] also showed that the spectral shifting approach can be used to improve other kernel approximation models such as the memory efficient kernel approximation (MEKA) model [Si et al., 2014].

The Nyström Method is the most popular kernel approximation approach. It is named after its inventor Nyström [1930] and gained its popularity in the machine learning society after its application in Gaussian procession regression [Williams and Seeger, 2001]. Let $\mathbf{S}$ be a column selection matrix, $\mathbf{C} = \mathbf{KS}$, and $\mathbf{W} = \mathbf{S}^T\mathbf{KS}$. The Nyström method approximates $\mathbf{K}$ by $\mathbf{CW}^\dagger\mathbf{C}^T$. In fact, the Nyström method is a special case of the faster SPSD matrix sketching where $\mathbf{P}$ and $\mathbf{S}$ are equal. This also indicates that the Nyström method is an approximate solution to (10.4). Gittens and Mahoney [2013] offered comprehensive error analysis of the Nyström method. The Nyström method has been applied to solve million scale kernel methods [Talwalkar et al., 2013]. But unlike the faster SPSD matrix sketching, the Nyström method cannot generate high quality approximation. The lower bound [Wang and Zhang, 2013] indicates that the Nyström method cannot attain $(1+\epsilon)$ relative-error bound unless it is willing to spend $\Omega(n^2k/\epsilon)$ time.

To this end, we have shown how to efficiently approximate any kernel matrix and use the obtained low rank approximation to speed up training. We will introduce efficient generalization using the CUR matrix decomposition in the next section.

10.3 The CUR Approximation

Let $\mathbf{A}$ by any $m \times n$ matrix. The CUR matrix decomposition is formed by selecting c columns of $\mathbf{A}$ to form $\mathbf{C} \in \mathbb{R}^{m\times c}$, r rows to form $\mathbf{R} \in \mathbb{R}^{r\times n}$, and computing an intersection matrix $\mathbf{U} \in \mathbb{R}^{c\times r}$ such that $\mathbf{CUR} \approx \mathbf{A}$. In this section, we first discussion the motivations and then describe algorithms and error analyses.

Motivations. Firstly, let us continue the generalization problem of kernel methods which remains unsolved in the previous section. Suppose we are given n training data and m test data, all of d dimension. To generalize the trained model to the test data, supervised kernel methods such as Gaussian processes and KRR require evaluating the kernel function of every train and test data pair—that is to form an $m \times n$ cross kernel matrix $\mathbf{K}_*$—which costs $\mathcal{O}(mnd)$ time. By the fast CUR algorithm described later in this section, the approximation $\mathbf{K}_* \approx \mathbf{CUR}$ can be obtained in time linear in $d(m+n)$. With such a decomposition at hand, the matrix product $\mathbf{K}_*\mathbf{M} \approx \mathbf{CURM}$ can be computed in $\mathcal{O}(nrk + mck)$ time. In this way, the overall time cost of generalization is linear in $m + n$.

Secondly, CUR forms a compressed representation of the data matrix, as well as the truncated SVD, and it can be very efficiently converted to the SVD-like form:

$$\mathbf{A} \;\approx\; \mathbf{CUR} \;=\; \mathbf{U}_C \underbrace{\mathbf{\Sigma}_C \mathbf{V}_C^T \mathbf{U} \mathbf{U}_R \mathbf{\Sigma}_R}_{=\mathbf{B}} \mathbf{V}_R^T \;=\; (\mathbf{U}_C \mathbf{U}_B) \mathbf{\Sigma}_B (\mathbf{V}_R \mathbf{V}_B)^T.$$

Here $\mathbf{C} = \mathbf{U}_C \mathbf{\Sigma}_C \mathbf{V}_C^T$, $\mathbf{R} = \mathbf{U}_R \mathbf{\Sigma}_R \mathbf{V}_R^T$, $\mathbf{B} = \mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_R$ are the SVD. Since CUR is formed by sampling columns and rows, it preserves the sparsity and nonnegativity of the original data matrix. The sparsity makes CUR cheaper to store than SVD, and the nonnegativity makes CUR a nonnegative matrix factorization.

Thirdly, CUR consists of the actual columns and rows, and thus it enables human to to understand and interpret the data. In comparison, the basis vectors of SVD has little concrete meaning. An example of Drineas et al. [2008] and Mahoney and Drineas [2009] has well shown this viewpoint; that is, the vector $[(1/2)\text{age} - (1/\sqrt{2})\text{height} + (1/2)\text{income}]$, the sum of the significant uncorrelated features from a data set of people's features, is not particularly informative. Therefore, it is of great interest to represent a data matrix in terms of a small number of actual columns and/or actual rows of the matrix.

Column Selection. Several different column selection strategies have been devised, among which the leverage score sampling [Drineas et al., 2008] and the adaptive sampling [Wang and Zhang, 2013, Boutsidis and Woodruff, 2014] attain relative error bounds. In particular, Boutsidis and Woodruff [2014] showed that with $c = \mathcal{O}(k/\epsilon)$ columns and $r = \mathcal{O}(k/\epsilon)$ rows selected by adaptive sampling to form $\mathbf{C}$ and $\mathbf{R}$,

$$\min_{\mathbf{X}} \|\mathbf{A} - \mathbf{CXR}\|_F^2 \;\leq\; (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

holds in expectation. A further refinement was developed by Woodruff

[2014b]. We will not go to the details of the leverage score sampling or adaptive sampling. The users only need to know that such algorithms randomly sample columns/rows according to some non-uniform distributions. Unfortunately, it requires observing the whole matrix $\mathbf{A}$ to compute such non-uniform distributions, thus such column selection algorithms cannot be applied to speed up computation. It remains an open problem whether there is a relative-error sampling algorithm that needs not observing the whole of $\mathbf{A}$. In practice, the users can simply sample columns/rows uniformly without replacement, which usually has acceptable empirical performance.

The Intersection Matrix. With the selected columns $\mathbf{C}$ and rows $\mathbf{R}$ at hand, we can simply compute the intersection matrix by

$$\mathbf{U}^\star \;=\; \underset{\mathbf{U}}{\operatorname{argmin}} \; \big\|\mathbf{A} - \mathbf{CUR}\big\|_F^2 \;=\; \mathbf{C}^\dagger \mathbf{A} \mathbf{R}^\dagger. \tag{10.5}$$

Here the second equality follows from Theorem 9.2. This approach has been used by Stewart [1999], Wang and Zhang [2013], Boutsidis and Woodruff [2014]. This approach is very similar to the prototype SPSD matrix approximation method in the previous section, and it costs at least $\mathcal{O}(mn \cdot \min\{c, r\})$ time and requires observing every entry of $\mathbf{A}$. Apparently, it cannot help speed up matrix computation.

Wang et al. [2015a] proposed a more practical CUR decomposition method which solves (10.5) approximately. The method first draws two column selection matrices $\mathbf{P}_C \in \mathbb{R}^{m \times p_c}$ and $\mathbf{P}_R \in \mathbb{R}^{n \times p_r}$ ($p_c, p_r \geq c, r$), which costs $\mathcal{O}(mc^2 + nr^2)$ time. It then computes the intersection matrix by

$$\tilde{\mathbf{U}} = \underset{\mathbf{U}}{\operatorname{argmin}} \; \big\|\mathbf{P}_C^T(\mathbf{A} - \mathbf{CUR})\mathbf{P}_R\big\|_F^2 = (\mathbf{P}_C^T \mathbf{C})^\dagger (\mathbf{P}_C^T \mathbf{A} \mathbf{P}_R)(\mathbf{R}\mathbf{P}_R)^\dagger.$$

This method needs observing only $p_c \times p_r$ entries of $\mathbf{A}$, and the overall time cost is $\mathcal{O}(p_c p_r \cdot \min\{c, r\} + mc^2 + nr^2)$. When

$$p_c \geq \mathcal{O}\Big(c\sqrt{\min\{m,n\}/\epsilon}\Big) \quad \text{and} \quad p_r \geq \mathcal{O}\Big(r\sqrt{\min\{m,n\}/\epsilon}\Big),$$

the following inequality holds with high probability:

$$\big\|\mathbf{A} - \mathbf{C}\tilde{\mathbf{U}}\mathbf{R}\big\|_F^2 \;\leq\; (1+\epsilon) \min_{\mathbf{U}} \big\|\mathbf{A} - \mathbf{CUR}\big\|_F^2.$$

In sum, a high quality CUR decomposition can be computed in time linear in $\min\{m, n\}$.

Acknowledgements

I would like to thank my graduate students Cheng Chen, Luo Luo, Shusen Wang, Haishan Ye, and Qiaomin Ye. Specifically, Cheng Chen, Luo Luo and Qiaomin Ye helped to proofread the whole manuscript. Haishan Ye helped to revise Chapter 9.2, and Shusen Wang helped to revise Chapter 10. I would also like to thank other students who took my course “ Matrix Methods in Massive Data Analysis” in the summer term 2015. They helped to improve the lecture notes, which provide the main materials for this tutorial.

References

Raja Hafiz Affandi, Alex Kulesza, Emily B. Fox, and Ben Taskar. Nyström approximation for large-scale determinantal processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013.

S. Akaho. A kernel method for canonical correlation analysis. In *International Meeting of Psychometric Society*, 2001.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

Yossi Azar, Amos Fiat, Anna Karlin, Frank McSherry, and Jared Saia. Spectral analysis of data. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 619–626. ACM, 2001.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

J. Batson, D. Spielman, and N. Srivastave. Twice-Ramanujan sparsifiers. *SIAM Review*, 56(2):315–334, 2014.

G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12:2385–2404, 2000.

P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. PAMI*, 19 (7):711–720, 1997.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications. Second Edition.* Springer, 2003.

M. W. Berry, S. A. Pulatova, and G. W. Stewart. Algorithm 844: computing sparse reduced-rank approximations to sparse matrices. *ACM Transactions on Mathematical Software*, 31(2):252–269, 2005.

Rajendra Bhatia. *Matrix Analysis.* Springer, 1997.

Srinadh Bhojanapalli, Prateek Jain, and Sujay Sanghavi. Tighter low-rank approximation via sampling the leveraged element. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 902–920. SIAM, 2015.

J. Bien, Y. Xu, and M. W. Mahoney. CUR from a sparse optimization viewpoint. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

C. H. Bischof and P. C. Hansen. Structure-preserving and rank-revealing QR-factorizations. *SIAM Journal on Scientific and Statistical Computing*, 12 (6):1332–1350, 1991.

Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science.* 2015.

Jonathan M. Borwein and Adrian S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples.* Springer, second edition, 2006.

Christos Boutsidis and David P. Woodruff. Optimal CUR matrix decompositions. *STOC*, pages 353–362, 2014.

Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal column-based matrix reconstruction. *SIAM Journal on Computing*, 43(2): 687–717, 2014.

Christopher J. C. Burges. Dimension reduction: A guided tour. *Foundations and Trends in Machine Learning*, 2:275–365, 2010.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Emmanuel J Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Emmanuel J Candès and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6): 2313–2351, 2007.

T. F. Chan. Rank revealing QR factorizations. *Linear Algebra and Its Applications*, 88:67–82, 1987.

S. Chandrasekaran and I. C. F. Ipsen. On rank-revealing factorisations. *SIAM Journal on Matrix Analysis and Applications*, 15(2):592–622, 1994.

Kenneth L Clarkson and David P Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 205–214. ACM, 2009.

Kenneth L Clarkson and David P Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 81–90. ACM, 2013.

Edith Cohen and David D Lewis. Approximating matrix multiplication for pattern recognition tasks. *Journal of Algorithms*, 30(2):211–252, 1999.

Michael Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. *arXiv preprint arXiv:1410.6801*, 2014.

Michael B. Cohen, Jelani Nelson, and David P. Woodruff. Optimal approximate matrix product in terms of stable rank. *arXiv preprint arXiv:1507.02268*, 2015.

T. F. Cox and M. A. A. Cox. *Multidimensional Scaling.* Chapman & Hall/CRC, second edition, 2000.

S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structure & Algorithms*, 22(1):60–65, 2003.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of The American Society for Information Science*, 41(6):391–407, 1990.

J. Demmel. *Applied Numerical Linear Algebra.* SIAM, Philadelphia, 1997.

A. Deshpande and L. Rademacher. Efficient volume sampling for row/column subset selection. In *Proceedings of the 51st IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 329–338, 2010.

A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. *Theory of Computing*, 2(2006):225–247, 2006.

P. Drineas and M. W. Mahoney. On the Nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

P. Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30:844–881, 2008.

Petros Drineas, Ravi Kannan, and Michael W Mahoney. Fast monte carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1):132–157, 2006a.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Sampling algorithms for l_2 regression and applications. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, SODA '06, pages 1127–1136, Philadelphia, PA, USA, 2006b. Society for Industrial and Applied Mathematics. ISBN 0-89871-605-5.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(1):3475–3506, 2011a.

Petros Drineas, Michael W Mahoney, S Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011b.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.

C. Eckart and G. Young. A principal axis transformation for non-Hermitian matrices. *Bulletin of the American Mathematical Society*, 45(2):118–121, 1939.

Ky Fan. Maximum properties and inequalities for the eigenvalues of completely continuous operators. *Proc. Nat. Acad. Sci. USA*, 37:760–766, 1951.

Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1434–1453. SIAM, 2013.

L. V. Foster. Rank and null space calculations using matrix decomposition without column interchanges. *Linear Algebra and its Applications*, 74:47–71, 1986.

C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.

A. Frieze, K. Kannan, and Rademacher S. Vempala. Fast Monte Carlo algorithms for finding low-rank approximation. *Journal of the ACM*, 51(6): 1025–1041, 2004.

Mina Ghashami and Jeff M Phillips. Relative errors for deterministic low-rank matrix approximations. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 707–717. SIAM, 2014.

P. M. Gibson. Simultaneous diagonalization of rectangular complex matrices. *Linear Algebra and Its Applications*, 9:45–53, 1974.

A. Gittens and M. W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. In *International Conference on Machine Learning (ICML)*, 2013.

Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU Press, 3rd edition, 2012.

T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, and M. Caligiuri. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–536, 1999.

S. A. Goreinov, E. E. Tyrtyshnikov, and N. L. Zamarashkin. A theory of pseudoskeleton approximations. *Linear Algebra and Its Applications*, 261: 1–21, 1997a.

S. A. Goreinov, N. L. Zamarashkin, and E. E. Tyrtyshnikov. Pseudo-skeleton approximations by matrices of maximal volume. *Mathematical Notes*, 62 (4):619–623, 1997b.

J. C. Gower and G. B. Dijksterhuis. *Procrustes Problems*. Oxford University Press, 2004.

Ming Gu. Subspace iteration randomization and singular value problems. *SIAM Journal on Scientific Computing*, 37(3):1139–1173, 2015.

Ming Gu and S. C. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

V. Guruswami and A. K. Sinop. Optimal column-based low-rank matrix reconstruction. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2012.

N Halko, P G Martinsson, and J A Tropp. Finding Structure with Randomness : Probabilistic Algorithms for Matrix Decompositions. *SIAM Review*, 53(2):217–288, 2011.

D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16:2639–2664, 2004.

G. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, second edition, 1951.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

Trevor Hastie, Rahul Mazumder, Jason Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *arXiv preprint arXiv:1410.2596*, 2014.

D. C. Hoaglin and R. E. Welsch. The hat matrix in regression and ANOVA. *The American Statistician*, 32(1):17–22, 1978.

Y. P. Hong and C. T. Pan. Rank-revealing QR factorizations and the singular value decomposition. *Mathematics of Computation*, 58(197):213–232, 1992.

A. Horn. On the singular values of a product of completely continuous operators. *Proc. Nat. Acad. Sci. USA*, 36:374–375, 1951.

A. Horn. On the eigenvalues of a matrix with prescribed singular values. *Proc. Amer. Math. Soc.*, 5:4–7, 1954.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge, UK, 1985.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, second edition, 1991.

P. Howland, M. Jeon, and H. Park. Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 25(1):165–179, 2003.

R. Jin, T. Yang, M. Mahdavi, Y. F. Li, and Z. H. Zhou. Improved bound for the Nyström method and its application to kernel classification. *IEEE Transactions on Information Theory*, 59(10):6939–6949, 2013.

W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mapping into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

I.T. Jolliffe. *Principal component analysis.* Springer, New York, second edition edition, 2002.

Daniel M Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):4, 2014.

Zhuoliang Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 521–528, 2011.

J. Kittler and P. C. Young. A new approach to feature selection based on the Karhunen-Loève expansion. *Pattern Recognition*, 5:335–352, 1973.

S. Kumar, M. Mohri, and A. Talwalkar. Ensemble Nyström method. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

F. G. Kuruvilla, P. J. Park, and S. L. Schreiber. Vector algebra in the analysis of genome-wide expression data. *Genome Biology*, 3, 2002.

Anastasios Kyrillidis, Michail Vlachos, and Anastasios Zouzias. Approximate matrix multiplication with application to linear embeddings. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 2182–2186. IEEE, 2014.

Adrian S Lewis. The mathematics of eigenvalue optimization. *Mathematical Programming*, 97(1-2):155–176, 2003.

Edo Liberty. Simple and deterministic matrix sketching. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 581–588. ACM, 2013.

Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. In *Pattern Analysis and Machine Intelligence*, volume 35, pages 208–220. IEEE, 2013.

C. F. Van Loan. Generalizing the singular value decomposition. *SIAM Journal on numerical Analysis*, 13:76–83, 1976.

Luo Luo, Yubo Xie, Zhihua Zhang, and Wu-Jun Li. Support matrix machines. In *The International Conference on Machine Learning (ICML)*, 2015.

Ping Ma, Michael Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. In *International Conference on Machine Learning (ICML)*, 2014.

Jan R. Macnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, second edition, 2000.

Malik Magdon-Ismail. Using a non-commutative Bernstein bound to approximate some matrix algorithms in the spectral norm. *arXiv preprint arXiv:1103.5453*, 2011.

Avner Magen and Anastasios Zouzias. Low rank matrix-valued chernoff bounds and approximate matrix multiplication. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 1422–1436. SIAM, 2011.

M. W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3): 697–702, 2009.

M. W. Mahoney, M. Maggioni, and P. Drineas. Tensor-CUR decompositions for tensor-based data. *SIAM Journal on Matrix Analysis and Applications*, 30(3):957–987, 2008.

Michael W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3:123–224, 2011.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, New York, 1979.

Albert W. Marshal, Ingram Olkin, and Barry C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, second edition, 2010.

Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis*, 30(1):47–68, 2011.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11:2287–2322, 2010.

S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, and K. R. Müller. Invariant feature extraction and classification in kernel space. In *Advances in Neural Information Processing Systems 12*, volume 12, pages 526–532, 2000.

L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *Quarterly Journal of Mathemathics*, 11:50–59, 1960.

R. J. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley and Sons, New York, 1982.

N. Muller, L. Magaia, and B. M. Herbst. Singular value decomposition, eigenfaces, and 3 D reconstruction. *SIAM Review*, 46:518–545, 2004.

Cameron Musco and Christopher Musco. Stronger approximate singular value decomposition via the block lanczos and power methods. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Jelani Nelson and Huy L Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 117–126. IEEE, 2013.

J. von Neumann. Some matrix-inequalities and metrication of matrix-space. *Tomsk University Review*, 1:286–300, 1937.

Evert J. Nyström. Über die praktische auflösung von integralgleichungen mit anwendungen auf randwertaufgaben. *Acta Mathematica*, 54(1):185–204, 1930.

C. C. Paige and M. A. Saunders. Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis*, 18(3):398–405, 1981.

Christos H Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 159–168. ACM, 1998.

C. H. Park and H. Park. Nonlinear discriminant analysis using kernel functions and the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 27(1):87–102, 2005.

Ting Kei Pong, Paul Tseng, Shuiwang Ji, and Jieping Ye. Trace norm regularization: reformulations, algorithms, and multi-task learning. *SIAM Journal on Optimization*, 20(6):3465–3489, 2010.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* The MIT Press, Cambridge, MA, 2006.

T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, New Jersey, 1970.

V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31:1100–1124, 2009.

V. Roth and V. Steinhage. Nonlinear discriminant analysis using kernel functions. In *Advances in Neural Information Processing Systems 12*, volume 12, pages 568–574, 2000.

Yousef Saad. Numerical methods for large eigenvalue problems. *preparation. Available from: http://www-users. cs. umn. edu/saad/books. html*, 2011.

Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 143–152. IEEE, 2006.

Robert Schatten. *A Theory of Cross-Space.* Princeton University Press, 1950.

B. Schölkopf and A. Smola. *Learning with Kernels.* The MIT Press, 2002.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, Cambridge, UK, 2004.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8): 888–905, 2000.

Si Si, Cho-Jui Hsieh, and Inderjit Dhillon. Memory efficient kernel approximation. In *International Conference on Machine Learning (ICML)*, pages 701–709, 2014.

Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336, 2004.

G. W. Stewart. Four algorithms for the efficient computation of truncated pivoted QR approximations to a sparse matrix. *Numerische Mathematik*, 83(2):313–323, 1999.

G. W. Stewart and J. G. Sun. *Matrix Perturbation Theory.* Academic Press, New York, 1990.

A. Talwalkar and A. Rostamizadeh. Matrix coherence and the Nyström method. In *In Proceedings of the 26th Conference in Uncertainty in Artificial Intelligence*, 2010.

A. Talwalkar, S. Kumar, and H. Rowley. Large-scale manifold learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

Ameet Talwalkar, Sanjiv Kumar, Mehryar Mohri, and Henry Rowley. Large-scale SVD and manifold learning. *Journal of Machine Learning Research*, 14:3129–3152, 2013.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500): 2319–2323, 2000.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

Lloyd N. Trefethen and David Bau III. *Numerical Linear Algebra.* SIAM, 1997.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

Bojun Tu, Zhihua Zhang, Shusen Wang, and Hui Qiani. Making fisher discriminant analysis scalable. In *Proceedings of the 31th International Conference on Machine Learning (ICML'14)*, 2014.

M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 586–591, 1991.

E. E. Tyrtyshnikov. Incomplete cross approximation in the mosaic-skeleton method. *Computing*, 64:367–380, 2000.

T. Van Gestel, J. A. K. Suykens, J. De Brabanter, B. De Moor, and J. Vandewalle. Kernel canonical correlation analysis and least squares support vector machines. In *The International Conference on Artificial Neural Networks (ICANN)*, pages 381–386, 2001.

V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

Santosh S. Vempala. *The Random Projection Method*. American Mathematical Society, 2000.

Shusen Wang and Zhihua Zhang. Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling. *Journal of Machine Learning Research*, 14:2729–2769, 2013.

Shusen Wang and Zhihua Zhang. Efficient algorithms and error analysis for the modified nyström method. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

Shusen Wang, Luo Luo, and Zhihua Zhang. The modified Nyström method: Theories, algorithms, and extension. *CoRR, abs/1406.5675*, 2014a. URL `http://arxiv.org/abs/1406.5675`.

Shusen Wang, Chao Zhang, Hui Qian, and Zhihua Zhang. Improving the modified nyström method using spectral shifting. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2014b.

Shusen Wang, Zhihua Zhang, and Tong Zhang. Improved analyses of the randomized power method and block Lanczos method. *arXiv:1508.06429*, 2015a. URL `http://arxiv.org/abs/1508.0642`.

Shusen Wang, Zhihua Zhang, and Tong Zhang. Towards more efficient symmetric matrix sketching and cur matrix decomposition. *arXiv preprint arXiv:1503.08395*, 2015b.

D. S. Watkins. *Fundamentals of Matrix Computations*. John Wiley and Sons, New York, 1991.

G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and Its Applications*, 170:33–45, 1992.

C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2001.

Rafi Witten and Emmanuel Candès. Randomized algorithms for low-rank matrix factorizations: sharp performance bounds. *Algorithmica*, 72(1):264–281, 2013.

David Woodruff. Low rank approximation lower bounds in row-update streams. In *Advances in Neural Information Processing Systems*, pages 1781–1789, 2014a.

David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014b.

Franco Woolfe, Edo Liberty, Vladimir Rokhlin, and Mark Tygert. A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25(3):335–366, 2008.

J. Ye and T. Xiong. Computational and theoretical analysis of null space and orthogonal linear discriminant analysis. *Journal of Machine Learning Research*, 7:1183–1204, 2006.

K. Zhang and J. T. Kwok. Clustered Nyström method for large scale manifold learning and dimension reduction. *IEEE Transactions on Neural Networks*, 21(10):1576–1587, 2010.

K. Zhang, I. W. Tsang, and J. T. Kwok. Improved Nyström low-rank approximation and error analysis. In *International Conference on Machine Learning (ICML)*, 2008.

Zhihua Zhang. The matrix ridge approximation: algorithms and applications. *Machine Learning*, 97:227–258, 2014.

Zhihua Zhang, Guang Dai, Congfu Xu, and Michael I. Jordan. Regularized discriminant analysis, ridge regression and beyond. *Journal of Machine Learning Research*, 11:2199–2228, 2010.

Hua Zhou and Lexin Li. Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):463–483, 2014.