Spectral Methods for Natural Language Processing (Part I of the Dissertation)

Jang Sun Lee (Karl Stratos)

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

Table of Contents

Chapter 1

A Review of Linear Algebra

1.1 Basic Concepts

In this section, we review basic concepts in linear algebra frequently invoked in spectral techniques.

1.1.1 Vector Spaces and Euclidean Space

A **vector space** $\mathcal{V}$ over a field $\mathcal{F}$ of scalars is a set of "vectors", entities with direction, closed under addition and scalar multiplication satisfying certain axioms. It can be endowed with an **inner product** $\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathcal{F}$, which is a quantatative measure of the relationship between a pair of vectors (such as the angle). An inner product also induces a **norm** $||u|| = \sqrt{\langle u, u \rangle}$ which computes the magnitude of u. See Chapter 1.2 of Friedberg *et al.* [2003] for a formal definition of a vector space and Chapter 2 of Prugovečki [1971] for a formal definition of an inner product.

In subsequent sections, we focus on Euclidean space to illustrate key ideas associated with a vector space. The n-dimensional (real-valued) **Euclidean space** $\mathbb{R}^n$ is a vector space over $\mathbb{R}$. The **Euclidean inner product** $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\langle u, v \rangle := [u]_1[v]_1 + \cdots + [u]_n[v]_n \tag{1.1}$$

It is also called the **dot product** and written as $u \cdot v$. The standard vector multiplication notation $u^\top v$ is sometimes used to denote the inner product.

One use of the inner product is calculating the *length* (or *norm*) of a vector. By the Pythagorean theorem, the length of $u \in \mathbb{R}^n$ is given by $||u||_2 := \sqrt{[u]_1^2 + \ldots + [u]_n^2}$ and called the **Euclidean norm** of u. Note that it can be calculated as

$$||u||_2 = \sqrt{\langle u, u \rangle} \tag{1.2}$$

Another use of the inner product is calculating the *angle* θ between two nonzero vectors. This use is based on the following result.

Theorem 1.1.1. *For nonzero $u, v \in \mathbb{R}^n$ with angle θ, $\langle u, v \rangle = ||u||_2 \, ||v||_2 \cos\theta$.*

Proof. Let $w = u - v$ be the opposing side of θ. The law of cosines states that

$$||w||_2^2 = ||u||_2^2 + ||v||_2^2 - 2 \, ||u||_2 \, ||v||_2 \cos\theta$$

But since $||w||_2^2 = ||u||_2^2 + ||v||_2^2 - 2\langle u, v \rangle$, we conclude that $\langle u, v \rangle = ||u||_2 \, ||v||_2 \cos\theta$. □

The following corollaries are immediate from Theorem 1.1.1.

Corollary 1.1.2 (Orthogonality)**.** *Nonzero $u, v \in \mathbb{R}^n$ are orthogonal (i.e., their angle is $\theta = \pi/2$) iff $\langle u, v \rangle = 0$.*

Corollary 1.1.3 (Cauchy–Schwarz inequality)**.** $|\langle u, v \rangle| \leq ||u||_2 \, ||v||_2$ *for all $u, v \in \mathbb{R}^n$.*

1.1.2 Subspaces and Dimensions

A **subspace** S of $\mathbb{R}^n$ is a subset of $\mathbb{R}^n$ which is a vector space over $\mathbb{R}$ itself. A necessary and sufficient condition for $S \subseteq \mathbb{R}^n$ to be a subspace is the following (Theorem 1.3, Friedberg *et al.* [2003]):

1. $0 \in S$
2. $u + v \in S$ whenever $u, v \in S$
3. $au \in S$ whenever $a \in \mathbb{R}$ and $u \in S$

The condition implies that a subspace is always a "flat" (or linear) space passing through the origin, such as infinite lines and planes (or the trivial subspace $\{0\}$).

A set of vectors $u_1 \dots u_m \in \mathbb{R}^n$ are called **linearly dependent** if there exist $a_1 \dots a_m \in \mathbb{R}$ that are not all zero such that $au_1 + \cdots au_m = 0$. They are **linearly independent** if they are not linearly dependent. The **dimension** $\dim(S)$ of a subspace $S \subseteq \mathbb{R}^n$ is the number of linearly independent vectors in S.

The **span** of $u_1 \dots u_m \in \mathbb{R}^n$ is defined to be all their linear combinations:

$$\text{span}\{u_1 \dots u_m\} := \left\{ \sum_{i=1}^{m} a_i u_i \;\middle|\; a_i \in \mathbb{R} \right\} \tag{1.3}$$

which can be shown to be the smallest subspace of $\mathbb{R}^n$ containing $u_1 \dots u_m$ (Theorem 1.5, Friedberg *et al.* [2003]).

The **basis** of a subspace $S \subseteq \mathbb{R}^n$ of dimension m is a set of linearly independent vectors $u_1 \dots u_m \in \mathbb{R}^n$ such that

$$S = \text{span}\{u_1 \dots u_m\} \tag{1.4}$$

In particular, $u_1 \dots u_m$ are called an **orthonormal basis** of S when they are orthogonal and have length $||u_i||_2 = 1$. We frequently parametrize an orthonormal basis as an orthonormal matrix $U = [u_1 \dots u_m] \in \mathbb{R}^{n \times m}$ ($U^\top U = I_{m \times m}$).

Finally, given a subspace $S \subseteq \mathbb{R}^n$ of dimension $m \leq n$, the corresponding **orthogonal complement** $S^\perp \subseteq \mathbb{R}^n$ is defined as

$$S^\perp := \{u \in \mathbb{R}^n : \; u^\top v = 0 \;\; \forall v \in S\}$$

It is easy to verify that the three subspace conditions hold, thus $S^\perp$ is a subspace of $\mathbb{R}^n$. Furthermore, we always have $\dim(S) + \dim(S^\perp) = n$ (see Theorem 1.5, Friedberg *et al.* [2003]), thus $\dim(S^\perp) = n - m$.

1.1.3 Matrices

A matrix $A \in \mathbb{R}^{m \times n}$ defines a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. Given $u \in \mathbb{R}^n$, the transformation $v = Au \in \mathbb{R}^m$ can be thought of as either a linear combination of the columns $c_1 \dots c_n \in \mathbb{R}^m$ of A, or dot products between the rows $r_1 \dots r_m \in \mathbb{R}^n$ of A and u:

$$v = [u]_1 c_1 + \cdots + [u]_n c_n = \begin{bmatrix} r_1^\top u \\ \vdots \\ r_m^\top u \end{bmatrix} \tag{1.5}$$

The **range** (or the **column space**) of A is defined as the span of the columns of A; the **row space** of A is the column space of $A^\top$. The **null space** of A is defined as the set of vectors $u \in \mathbb{R}^n$ such that $Au = 0$; the **left null space** of A is the null space of $A^\top$. We denote them respectively by the following symbols:

$$\text{range}(A) = \text{col}(A) := \{Au : \ u \in \mathbb{R}^n\} \subseteq \mathbb{R}^m \tag{1.6}$$

$$\text{row}(A) := \text{col}(A^\top) \subseteq \mathbb{R}^n \tag{1.7}$$

$$\text{null}(A) := \{u \in \mathbb{R}^n : \ Au = 0\} \subseteq \mathbb{R}^n \tag{1.8}$$

$$\text{left-null}(A) := \text{null}(A^\top) \subseteq \mathbb{R}^m \tag{1.9}$$

It can be shown that they are all subspaces (Theorem 2.1, Friedberg *et al.* [2003]). Observe that $\text{null}(A) = \text{row}(A)^\perp$ and $\text{left-null}(A) = \text{range}(A)^\perp$. In Section 1.3, we show that singular value decomposition can be used to find an orthonormal basis of each of these subspaces.

The **rank** of A is defined as the dimension of the range of A, which is the number of linearly independent columns of A:

$$\text{rank}(A) := \dim(\text{range}(A)) \tag{1.10}$$

An important use of the rank is testing the invertibility of a square matrix: $A \in \mathbb{R}^{n \times n}$ is invertible iff $\text{rank}(A) = n$ (see p. 152 of Friedberg *et al.* [2003]). The **nullity** of A is the dimension of the null space of A, $\text{nullity}(A) := \dim(\text{null}(A))$.

The following theorems are fundamental results in linear algebra:

Theorem 1.1.4 (Rank-nullity theorem)**.** *Let $A \in \mathbb{R}^{m \times n}$. Then*

$$\text{rank}(A) + \text{nullity}(A) = n$$

Proof. See p. 70 of Friedberg *et al.* [2003]. □

Theorem 1.1.5. *Let $A \in \mathbb{R}^{m \times n}$. Then*

$$\dim(\text{col}(A)) = \dim(\text{row}(A))$$

Proof. See p. 158 of Friedberg *et al.* [2003]. □

Theorem 1.1.5 shows that $\text{rank}(A)$ is also the number of linearly independent rows. Furthermore, the rank-nullity theorem implies that if $r = \text{rank}(A)$,

$$\begin{aligned} \text{rank}(A) = \dim(\text{col}(A)) = \dim(\text{row}(A)) = r \\ \dim(\text{null}(A)) = n - r \\ \dim(\text{left-null}(A)) = m - r \end{aligned}$$

We define additional quantities associated with a matrix. The **trace** of a square matrix $A \in \mathbb{R}^{n \times n}$ is defined as the sum of its diagonal entries:

$$\text{Tr}(A) := [A]_{1,1} + \cdots + [A]_{n,n} \tag{1.11}$$

The **Frobenius norm** $||A||_F$ of a matrix $A \in \mathbb{R}^{m \times n}$ is defined as:

$$||A||_F := \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |[A]_{i,j}|^2} = \sqrt{\text{Tr}(A^\top A)} = \sqrt{\text{Tr}(AA^\top)} \tag{1.12}$$

where the trace expression can be easily verified. The relationship between the trace and eigenvalues (1.23) implies that $||A||_F^2$ is the sum of the singular values of A. The **spectral norm** or the **operator norm** $||A||_2$ of a matrix $A \in \mathbb{R}^{m \times n}$ is defined as the maximizer of $||Ax||_2$ over the unit sphere,

$$||A||_2 := \max_{u \in \mathbb{R}^n:\, ||u||_2 = 1} ||Au||_2 = \max_{u \in \mathbb{R}^n:\, u \neq 0} \frac{||Au||_2}{||u||_2} \tag{1.13}$$

The variational characterization of eigenvalues (Theorem 1.2.7) implies that $||A||_2$ is the largest singular value of A. Note that $||Au||_2 \leq ||A||_2 \, ||u||_2$ for any $u \in \mathbb{R}^n$: this matrix-vector inequality is often useful.

An important property of $||\cdot||_F$ and $||\cdot||_2$ is their orthogonal invariance:

Proposition 1.1.1. *Let $A \in \mathbb{R}^{m \times n}$. Then*

$$||A||_F = ||QAR||_F \qquad\qquad ||A||_2 = ||QAR||_2$$

where $Q \in \mathbb{R}^{m \times m}$ and $R \in \mathbb{R}^{n \times n}$ are any orthogonal matrices (see Section 1.1.4).

Proof. Let $A = U\Sigma V^\top$ be an SVD of A. Then $QAR = (QU)\Sigma(R^\top V)^\top$ is an SVD of QAR since QU and $R^\top V$ have orthonormal columns. Thus A and QAR have the same set of singular values. Since $||\cdot||_F$ is the sum of singular values and $||\cdot||_2$ is the maximum singular value, the statement follows. □

1.1.4 Orthogonal Matrices

A square matrix $Q \in \mathbb{R}^{n\times n}$ is an **orthogonal matrix** if $Q^\top Q = I_{n\times n}$. In other words, the columns of Q are an orthonormal basis of $\mathbb{R}^n$; it follows that $QQ^\top = I_{n\times n}$ since $QQ^\top$ is an identity operator over $\mathbb{R}^n$ (see Section 1.1.5). Two important properties of Q are the following:

1. For any $u \in \mathbb{R}^n$, Qu has the same length as u:

$$||Qu||_2 = \sqrt{u^\top Q^\top Qu} = \sqrt{u^\top u} = ||u||_2$$

2. For any nonzero $u, v \in \mathbb{R}^n$, the angle $\theta_1 \in [0, \pi]$ between Qu and Qv and $\theta_2 \in [0, \pi]$ between u and v are the same. To see this, note that

$$||Qu||_2 \, ||Qv||_2 \cos\theta_1 = \langle Qu, Qv\rangle = u^\top Q^\top Qv = \langle u, v\rangle = ||u||_2 \, ||v||_2 \cos(\theta_2)$$

   It follows that $\cos\theta_1 = \cos\theta_2$ and thus $\theta_1 = \theta_2$ (since θ_1, θ_2 are taken in $[0, \pi]$).

Hence an orthogonal matrix $Q \in \mathbb{R}^{n\times n}$ can be seen as a *rotation* of the coordinates in $\mathbb{R}^n$.[1]

1.1.5 Orthogonal Projection onto a Subspace

Theorem 1.1.6. *Let $S \subseteq \mathbb{R}^n$ be a subspace spanned by an orthonormal basis $u_1 \dots u_m \in \mathbb{R}^n$. Let $U := [u_1 \dots u_m] \in \mathbb{R}^{n\times m}$. Pick any $x \in \mathbb{R}^n$ and define*

$$y^* := \arg\min_{y\in S} ||x - y||_2 \tag{1.14}$$

[1] Certain orthogonal matrices also represent *reflection*. For instance, the orthogonal matrix

$$Q = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

is a reflection in $\mathbb{R}^2$ (along the diagonal line that forms an angle of $\pi/4$ with the x-axis).

Then the unique solution is given by $y^* = UU^\top x$.

Proof. Any element $y \in S$ is given by Uv for some $v \in \mathbb{R}^n$, thus $y^* = Uv^*$ where

$$v* = \arg\min_{v \in \mathbb{R}^n} ||x - Uv||_2 = (U^\top U)^{-1} U^\top x = U^\top x$$

is unique, hence y^* is unique. □

In Theorem 1.1.6, $x - y^*$ is orthogonal to the subspace $S = \text{span}\{u_1 \dots u_m\}$ since

$$\langle x - y^*, u_i \rangle = x^\top u_i - x^\top U U^\top u_i = 0 \qquad \forall i \in [m] \tag{1.15}$$

For this reason, the $n \times n$ matrix $\Pi := UU^\top$ is called the **orthogonal projection** onto the subspace $S \subseteq \mathbb{R}^n$. A few remarks on Π:

1. Π is unique. If Π' is another orthogonal projection onto S, then $\Pi x = \Pi' x$ for all $x \in \mathbb{R}^n$ (since this is uniquely given, Theorem 1.1.6). Hence $\Pi = \Pi'$.

2. Π is an identitiy operator for elements in S. This implies that the inherent dimension of $x \in S$ is m (not n) in the sense that the m-dimensional vector

$$\tilde{x} := U^\top x$$

can be restored to $x = U\tilde{x} \in \mathbb{R}^n$ without any loss of accuracy. This idea is used in subspace identification techniques (Section 2.7).

It is often of interest to compute the orthogonal projection $\Pi \in \mathbb{R}^{n \times n}$ onto the range of $A \in \mathbb{R}^{n \times m}$. If A already has orthonormal columns, the projection is given by $\Pi = AA^\top$. Otherwise, a convenient construction is given by

$$\Pi = A(A^\top A)^+ A^\top \tag{1.16}$$

To see this, let $A = U\Sigma V^\top$ be a rank-m SVD of A so that the columns of $U \in \mathbb{R}^{n \times m}$ are an orthonormal basis of range(A). Then

$$A(A^\top A)^+ A^\top = (U\Sigma V^\top)(V\Sigma^{-2} V^\top)(V\Sigma U^\top) = UU^\top \tag{1.17}$$

Gram-Schmidt$(v_1 \dots v_m)$

Input: linearly independent $m \leq n$ vectors $v_1 \dots v_m \in \mathbb{R}^n$

1. Normalize $\bar{v}_1 = v_1 / ||v_1||_2$.
2. For $i = 2 \dots m$,
   (a) Remove the components of v_i lying in the span of $\bar{v}_1 \dots \bar{v}_{i-1}$,
   $$\bar{w}_i = v_i - [\bar{v}_1 \dots \bar{v}_{i-1}][\bar{v}_1 \dots \bar{v}_{i-1}]^\top v_i = v_i - \sum_{j=1}^{i-1} (\bar{v}_j^\top v_i)\bar{v}_j$$
   (b) Normalize $\bar{v}_i = \bar{w}_i / ||\bar{w}_i||_2$.

Output: orthonormal $\bar{v}_1 \dots \bar{v}_m \in \mathbb{R}^n$ such that $\text{span}\{\bar{v}_1 \dots \bar{v}_i\} = \text{span}\{v_1 \dots v_i\}$ for all $i = 1 \dots m$

Figure 1.1: The Gram-Schmidt process.

1.1.6 Gram-Schmidt Process and QR Decomposition

An application of the orthogonal projection yields a very useful technique in linear algebra called the **Gram-Schmidt process** (Figure 1.1).

Theorem 1.1.7. *Let $v_1 \dots v_m \in \mathbb{R}^n$ be linearly independent vectors. The output $\bar{v}_1 \dots \bar{v}_m \in \mathbb{R}^n$ of **Gram-Schmidt**$(v_1 \dots v_m)$ are orthonormal and satisfy*

$$\text{span}\{\bar{v}_1 \dots \bar{v}_i\} = \text{span}\{v_1 \dots v_i\} \qquad \forall 1 \leq i \leq m$$

Proof. The base case $i = 1$ can be trivially verified. Assume $\text{span}\{\bar{v}_1 \dots \bar{v}_{i-1}\}$ equals $\text{span}\{v_1 \dots v_{i-1}\}$ and consider the vector $\bar{v}_i$ computed in the algorithm. It is orthogonal to the subspace $\text{span}\{\bar{v}_1 \dots \bar{v}_{i-1}\}$ by (1.15) and has length 1 by the normalization step, so $\bar{v}_1 \dots \bar{v}_i$ are orthonormal. Furthermore,

$$v_i = (\bar{v}_1^\top v_i)\bar{v}_1 + \dots + (\bar{v}_{i-1}^\top v_i)\bar{v}_{i-1} + ||\bar{w}_i||_2 \, \bar{v}_i$$

is in $\text{span}\{\bar{v}_1 \dots \bar{v}_i\}$, thus $\text{span}\{\bar{v}_1 \dots \bar{v}_i\} = \text{span}\{v_1 \dots v_i\}$. □

QR(A)

Input: $A \in \mathbb{R}^{n\times m}$ with linearly independent columns $a_1 \dots a_m \in \mathbb{R}^n$

1. $Q := [\bar{a}_1 \dots \bar{a}_m] \leftarrow$ **Gram-Schmidt**$(a_1 \dots a_m)$
2. Define an upper triangular matrix $R \in \mathbb{R}^{m\times m}$ by

$$[R]_{i,j} \leftarrow \bar{a}_i^\top a_j \qquad \forall i \in [1,m],\ j \in [i,m]$$

Output: orthonormal matrix $Q \in \mathbb{R}^{n\times m}$ and an upper triangular matrix $R \in \mathbb{R}^{m\times m}$ such that $A = QR$.

Figure 1.2: QR decomposition.

The Gram-Schmidt process yields one of the most elementary matrix decomposition techniques called **QR decomposition**. A simplified version (which assumes only matrices with linearly independent columns) is given in Figure 1.1.

Theorem 1.1.8. *Let $A \in \mathbb{R}^{n\times m}$ be a matrix with linearly independent columns $a_1 \dots a_m \in \mathbb{R}^n$. The output (Q, R) of $\boldsymbol{QR}(A)$ are an orthonormal matrix $Q \in \mathbb{R}^{n\times m}$ and an upper triangular matrix $R \in \mathbb{R}^{m\times m}$ such that $A = QR$.*

Proof. The columns $\bar{a}_1 \dots \bar{a}_m$ of Q are orthonormal by Theorem 1.1.7 and R is upper triangular by construction. The i-th column of QR is given by

$$(\bar{a}_1^\top \bar{a}_1)a_i + \dots + (\bar{a}_i^\top \bar{a}_i)a_i = [\bar{a}_1 \dots \bar{a}_i][\bar{a}_1 \dots \bar{a}_i]^\top a_i = a_i$$

since $a_i \in \text{span}\{\bar{a}_1 \dots \bar{a}_i\}$. □

The Gram-Schmidt process is also used in the non-negative matrix factorization algorithm of Arora *et al.* [2012a].

1.2 Eigendecomposition

In this section, we develop a critical concept associated with a matrix called eigenvectors and eigenvalues. This concept leads to decomposition of a certain class of matrices called *eigen-*

decomposition. All statements (when not proven) can be found in standard introductory textbooks on linear algebra such as Strang [2009].

1.2.1 Square Matrices

Let $A \in \mathbb{R}^{n\times n}$ be a real square matrix. An **eigenvector** v of A is a nonzero vector that preserves its direction in $\mathbb{R}^n$ under the linear transformation defined by A: that is, for some scalar λ,

$$Av = \lambda v \tag{1.18}$$

The scalar λ is called the **eigenvalue** corresponding to v. A useful fact (used in the proof of Theorem 1.2.3) is that eigenvectors corresponding to different eigenvalues are linearly independent.

Lemma 1.2.1. *Eigenvectors (v, v') of $A \in \mathbb{R}^{n\times n}$ corresponding to distinct eigenvalues (λ, λ') are linearly independent.*

Proof. Suppose $v' = cv$ for some scalar c (which must be nonzero). Then the eigen conditions imply that $Av' = A(cv) = c\lambda v$ and also $Av' = \lambda' v' = c\lambda' v$. Hence $\lambda v = \lambda' v$. Since $\lambda \neq \lambda'$, we must have $v = 0$. This contradicts the definition of an eigenvector. □

Theorem 1.2.2. *Let $A \in \mathbb{R}^{n\times n}$. The following statements are equivalent:*

- *λ is an eigenvalue of A.*
- *λ is a scalar that yields $\det(A - \lambda I_{n\times n}) = 0$.*

Proof. λ is an eigenvalue of A iff there is some nonzero vector v such that $Av = \lambda v$, and

$$\begin{aligned}
\exists v \neq 0 :\ (A - \lambda I_{n\times n})v = 0 \quad &\iff \quad \text{nullity}(A - \lambda I_{n\times n}) > 0 \\
&\iff \quad \text{rank}(A - \lambda I_{n\times n}) < n \qquad \text{(by the rank-nullity theorem)} \\
&\iff \quad A - \lambda I_{n\times n} \text{ is not invertible}
\end{aligned}$$

The last statement is equivalent to $\det(A - \lambda I_{n\times n}) = 0$. □

Since $\det(A - \lambda I_{n\times n})$ is a degree n polynomial in λ, it has n roots (counted with multiplicity[2]) by the fundamental theorem of algebra and can be written as

$$\det(A - \lambda I_{n\times n}) = (\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n) \tag{1.19}$$

Let λ be a distinct root of (1.19) and $a(\lambda)$ its multiplicity. Theorem 1.2.2 implies that λ is a distinct eigenvalue of A with a space of corresponding eigenvectors

$$E_{A,\lambda} := \{v : \ Av = \lambda v\} \tag{1.20}$$

(i.e., the null space of $A - \lambda I_{n\times n}$ and hence a subspace) which is called the **eigenspace** of A associated with λ. The dimension of this space is the number of linearly independent eigenvectors corresponding to λ. It can be shown that

$$1 \leq \dim(E_{A,\lambda}) \leq a(\lambda)$$

where the first inequality follows by the definition of λ (i.e., there is a corresponding eigenvector). We omit the proof of the second inequality.

Theorem 1.2.3. *Let $A \in \mathbb{R}^{n\times n}$ be a matrix with eigenvalues $\lambda_1 \ldots \lambda_n$. The following statements are equivalent:*

- *There exist eigenvectors $v_1 \ldots v_n$ corresponding to $\lambda_1 \ldots \lambda_n$ such that*

  $$A = V\Lambda V^{-1} \tag{1.21}$$

  *where $V = [v_1 \ldots v_n]$ and $\Lambda = \operatorname{diag}(\lambda_1 \ldots \lambda_n)$. (1.21) is called an **eigendecomposition** of A.*

- *The eigenspace of A associated with each distinct eigenvalue λ has the maximum dimension, that is, $\dim(E_{A,\lambda}) = a(\lambda)$.*

Proof. For any eigenvectors $v_1 \ldots v_n$ corresponding to $\lambda_1 \ldots \lambda_n$, we have

$$AV = V\Lambda \tag{1.22}$$

[2]Recall that λ is a root of multiplicity k for a polynomial $p(x)$ if $p(x) = (x-\lambda)^k s(x)$ for some polynomial $s(x) \neq 0$.

Thus it is sufficient to show that the existence of an invertible V is equivalent to the second statement. This is achieved by observing that we can find n linearly independent eigenvectors iff we can find $a(\lambda)$ linearly independent eigenvectors for each distinct eigenvalue λ (since eigenvectors corresponding to different eigenvalues are already linearly independent by Lemma 1.2.1). □

Theorem 1.2.3 gives the condition on a square matrix to have an eigendecomposition (i.e., each eigenspace must have the maximum dimension). A simple corollary is the following:

Corollary 1.2.4. *If $A \in \mathbb{R}^{n\times n}$ has n distinct eigenvalues $\lambda_1 \dots \lambda_n$, it has an eigendecomposition.*

Proof. Since $1 \leq \dim(E_{A,\lambda_i}) \leq a(\lambda_i) = 1$ for each (distinct) eigenvalue λ_i, the statement follows from Theorem 1.2.3. □

Since we can write an eigendecomposition of A as

$$V^{-1}AV = \Lambda$$

where Λ is a diagonal matrix, a matrix that has an eigendecomposition is called **diagonalizable**.[3] Lastly, a frequently used fact about eigenvalues $\lambda_1 \dots \lambda_n$ of $A \in \mathbb{R}^{n\times n}$ is the following (proof omitted):

$$\mathrm{Tr}(A) = \lambda_1 + \cdots + \lambda_n \tag{1.23}$$

1.2.2 Symmetric Matrices

A square matrix $A \in \mathbb{R}^{n\times n}$ always has eigenvalues but not necessarily an eigendecomposition. Fortunately, if A is additionally *symmetric*, A is guaranteed to have an eigendecomposition of a convenient form.

Lemma 1.2.5. *Let $A \in \mathbb{R}^{n\times n}$. If A is symmetric, then*

[3] While not every square matrix $A \in \mathbb{R}^{n\times n}$ is diagonalizable, it can be transformed into an *upper triangular form* $T = U^\top AU$ by an orthogonal matrix $U \in \mathbb{R}^{n\times n}$; see Theorem 3.3 of Stewart and Sun [1990]. This implies a decomposition $A = UTU^\top$ known the **Schur decomposition**. A can also always be transformed into a *block diagonal form* called a **Jordan canonical form**; see Theorem 3.7 of Stewart and Sun [1990].

1. *All eigenvalues of A are real.*

2. *A is diagonalizable.*

3. *Eigenvectors corresponding to distinct eigenvalues are orthogonal.*

Proof. For the first and second statements, we refer to Strang [2009]. For the last statement, let (v, v') be eigenvectors of A corresponding to distinct eigenvalues (λ, λ'). Then

$$\lambda v^\top v' = v^\top A^\top v' = v^\top A v' = \lambda' v^\top v'$$

Thus $v^\top v' = 0$ since $\lambda \neq \lambda'$. □

Theorem 1.2.6. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with eigenvalues $\lambda_1 \dots \lambda_n \in \mathbb{R}$. Then there exist orthonormal eigenvectors $v_1 \dots v_n \in \mathbb{R}^d$ of A corresponding to $\lambda_1 \dots \lambda_n$. In particular,*

$$A = V \Lambda V^\top \tag{1.24}$$

for orthogonal matrix $V = [v_1 \dots v_n] \in \mathbb{R}^{n\times n}$ and $\Lambda = \operatorname{diag}(\lambda_1 \dots \lambda_n) \in \mathbb{R}^{n\times n}$.

Proof. Since A is diagonalizable (Lemma 1.2.5), the eigenspace of λ_i has dimension $a(\lambda_i)$ (Theorem 1.2.3). Since this is the null space of a real matrix $A - \lambda_i I_{n\times n}$, it has $a(\lambda_i)$ orthonormal basis vectors in $\mathbb{R}^n$. The claim follows from the fact that the eigenspaces of distinct eigenvalues are orthogonal (Lemma 1.2.5). □

Another useful fact about the eigenvalues of a symmetric matrix is the following.

Proposition 1.2.1. *If $A \in \mathbb{R}^{n\times n}$ is symmetric, the rank of A is the number of nonzero eigenvalues.*

Proof. The dimension of $E_{A,0}$ is the multiplicity of the eigenvalue 0 by Lemma 1.2.5 and Theorem 1.2.3. The rank-nullity theorem gives $\operatorname{rank}(A) = n - \operatorname{nullity}(A) = n - a(0)$. □

Note that (1.24) can be equivalently written as a sum of weighted outer products

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^\top = \sum_{\lambda_i \neq 0} \lambda_i v_i v_i^\top \tag{1.25}$$

1.2.3 Variational Characterization

In Section 1.2.2, we see that any symmetric matrix has an eigendecomposition with real eigenvalues and orthonormal eigenvectors. It is possible to frame this decomposition as a *constrained optimization problem* (i.e., variational characterization).

Theorem 1.2.7. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with orthonormal eigenvectors $v_1 \dots v_n \in \mathbb{R}^n$ corresponding to its eigenvalues $\lambda_1 \geq \dots \geq \lambda_n \in \mathbb{R}$. Let $k \leq n$. Consider maximizing $v^\top Av$ over unit-length vectors $v \in \mathbb{R}^n$ under orthogonality constraints:*

$$v_i^* = \underset{\substack{v\in\mathbb{R}^n: \\ ||v||_2=1 \\ v^\top v_j^*=0\ \forall j<i}}{\arg\max} \quad v^\top Av \qquad\qquad \text{for } i = 1 \dots k$$

Then an optimal solution is given by $v_i^* = v_i$.

Proof. The Lagrangian for the objective for v_1^* is:

$$L(v, \bar{\lambda}) = v^\top Av - \bar{\lambda}(v^\top v - 1)$$

Its stationary conditions $v^\top v = 1$ and $Av = \bar{\lambda}v$ imply that v_1^* is a unit-length eigenvector of A with eigenvalue $\bar{\lambda}$. Pre-multiplying the second condition by $v^\top$ and using the first condition, we have $\bar{\lambda} = v^\top Av$. Since this is the objective to maximize, we must have $\bar{\lambda} = \lambda_1$. Thus any unit-length eigenvector in E_{A,λ_1} is an optimal solution for v_1^*, in particular v_1. The case for $v_2^* \dots v_k^*$ can be proven similarly by induction. □

Note that in Theorem 1.2.7,

$$\lambda_1 = \max_{v:\ ||v||_2=1} v^\top Av = \max_{v\neq 0} \left(\frac{v}{\sqrt{v^\top v}}\right)^\top A \left(\frac{v}{\sqrt{v^\top v}}\right) = \max_{v\neq 0} \frac{v^\top Av}{v^\top v}$$

The quantity in the last expression is called the **Rayleigh quotient**,

$$R(A, v) := \frac{v^\top Av}{v^\top v} \tag{1.26}$$

Thus the optimization problem can be seen as maximizing $R(A, v)$ over $v \neq 0$ (under orthogonality constraints):

$$v_i^* = \underset{\substack{v\in\mathbb{R}^n: \\ v\neq 0 \\ v^\top v_j^*=0\ \forall j<i}}{\arg\max} \quad \frac{v^\top Av}{v^\top v} \qquad\qquad \text{for } i = 1 \dots k$$

Another useful characterization in matrix form is the following:

Theorem 1.2.8. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with orthonormal eigenvectors $v_1 \dots v_n \in \mathbb{R}^d$ corresponding to its eigenvalues $\lambda_1 \geq \dots \geq \lambda_n \in \mathbb{R}$. Let $k \leq n$. Consider maximizing the trace of $\overline{V}^\top A\overline{V} \in \mathbb{R}^{k\times k}$ over orthonormal matrices $\overline{V} \in \mathbb{R}^{n\times k}$:*

$$V^* = \underset{\overline{V}\in\mathbb{R}^{n\times k}:\, \overline{V}^\top\overline{V}=I_{k\times k}}{\arg\max} \operatorname{Tr}(\overline{V}^\top A\overline{V})$$

Then an optimal solution is given by $V^* = [v_1 \dots v_k]$.

Proof. Denote the columns of $\overline{V}$ by $\bar{v}_1 \dots \bar{v}_k \in \mathbb{R}^n$. The Lagrangian for the objective is:

$$L(\{\bar{v}_1 \dots \bar{v}_k\}, \{\bar{\lambda}_i\}_{i=1}^k, \{\gamma_{ij}\}_{i\neq j}) = \sum_{i=1}^k \bar{v}_i^\top A\bar{v}_i - \sum_{i=1}^k \bar{\lambda}_i(\bar{v}_i^\top\bar{v}_i - 1) - \sum_{i\neq j}\gamma_{ij}\bar{v}_i^\top\bar{v}_j$$

It can be verified from stationary conditions that $\bar{v}_i^\top\bar{v}_i = 1$, $\bar{v}_i^\top\bar{v}_j = 0$ (for $i \neq j$), and $A\bar{v}_i = \bar{\lambda}_i\bar{v}_i$. Thus $\bar{v}_1 \dots \bar{v}_k$ are orthonormal eigenvectors of A corresponding to eigenvalues $\bar{\lambda}_1 \dots \bar{\lambda}_k$. Since the objective to maximize is

$$\operatorname{Tr}(\overline{V}^\top A\overline{V}) = \sum_{i=1}^k \bar{v}_i^\top A\bar{v}_i = \sum_{i=1}^k \bar{\lambda}_i$$

any set of orthonormal eigenvectors corresponding to the k largest eigenvalues $\lambda_1 \geq \dots \geq \lambda_k$ are optimal, in particular $V^* = [v_1 \dots v_k]$. □

1.2.4 Semidefinite Matrices

A symmetric matrix $A \in \mathbb{R}^{n\times n}$ always has an eigendecomposition with real eigenvalues. When all the eigenvalues of A are furthermore *non-negative*, A is called **positive semidefinite** or **PSD** and sometimes written as $A \succeq 0$. Equivalently, a symmetric matrix $A \in \mathbb{R}^{n\times n}$ is PSD if $v^\top Av \geq 0$ for all $v \in \mathbb{R}^n$; to see this, let $A = \sum_{i=1}^n \lambda_i v_i v_i^\top$ be an eigendecomposition and note that

$$v^\top Av = \sum_{i=1}^n \lambda_i (v^\top v_i)^2 \geq 0 \;\; \forall v \in \mathbb{R}^n \qquad \iff \qquad \lambda_i \geq 0 \;\; \forall 1 \leq i \leq n$$

A PSD matrix whose eigenvalues are strictly positive is called **positive definite** and written as $A \succ 0$. Similarly as above, $A \succ 0$ iff $v^\top Av > 0$ for all $v \neq 0$. Matrices that are **negative semidefinite** and **negative definite** are symmetrically defined (for non-positive and negative eigenvalues).

These matrices are important because they arise naturally in many settings.

Example 1.2.1 (Covariance matrix)**.** *The covariance matrix of a random variable $X \in \mathbb{R}^n$ is defined as*

$$C_X := \boldsymbol{E}\left[(X - \boldsymbol{E}[X])(X - \boldsymbol{E}[X])^\top\right]$$

which is clearly symmetric. For any $v \in \mathbb{R}^n$, let $Z := v^\top(X - \boldsymbol{E}[X])$ and note that $v^\top C_X v = \boldsymbol{E}[Z^2] \geq 0$, thus $C_X \succeq 0$.

Example 1.2.2 (Hessian)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. The Hessian of f at x is defined as $\nabla^2 f(x) \in \mathbb{R}^{n\times n}$ where*

$$[\nabla^2 f(x)]_{i,j} := \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \qquad \forall i, j \in [n]$$

which is clearly symmetric. If x is stationary, $\nabla f(x) = 0$, then the spectral properties of $\nabla^2 f(x)$ determines the category of x.

- *If $\nabla^2 f(x) \succ 0$, then x is a local minimum. Consider any direction $u \in \mathbb{R}^n$. By Taylor's theorem, for a sufficiently small $\eta > 0$*

$$f(x + \eta u) \approx f(x) + \frac{\eta^2}{2} u^\top \nabla^2 f(x) u > f(x)$$

- *Likewise, if $\nabla^2 f(x) \prec 0$, then x is a local maximum.*

- *If $\nabla^2 f(x)$ has both positive and negative eigenvalues, then x is a saddle point. If $v_+ \in \mathbb{R}^n$ is an eigenvector corresponding to a positive eigenvalue,*

$$f(x + \eta v_+) \approx f(x) + \frac{\eta^2}{2} v_+^\top \nabla^2 f(x) v_+ > f(x)$$

  If $v_- \in \mathbb{R}^n$ is an eigenvector corresponding to a negative eigenvalue,

$$f(x + \eta v_-) \approx f(x) + \frac{\eta^2}{2} v_-^\top \nabla^2 f(x) v_- < f(x)$$

Finally, if $\nabla^2 f(x) \succeq 0$ for all $x \in \mathbb{R}^n$, then f is convex. Given any $x, y \in \mathbb{R}^n$, for some z between x and y,

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 f(z)(y - x) \\ &\geq f(x) + \nabla f(x)^\top (y - x) \end{aligned}$$

Example 1.2.3 (Graph Laplacian)**.** *Consider an undirected weighted graph with n vertices $[n]$ and a (symmetric) adjacency matrix $W \in \mathbb{R}^{n\times n}$. The (i,j)-th entry of W is a non-negative weight $w_{ij} \geq 0$ for edge (i,j) where $w_{ij} = 0$ iff there is no edge (i,j). The degree of vertex $i \in [n]$ is defined as $d_i := \sum_{j=1}^{n} w_{ij}$ and assumed to be positive.*

The (unnormalized) graph Laplacian is a matrix whose spectral properties reveal the connectivity of the graph:

$$L := D - W \tag{1.27}$$

where $D := \mathrm{diag}(d_1, \ldots, d_n)$. Note that L does not depend on self-edges w_{ii} by construction. This matrix has the following properties (proofs can be found in Von Luxburg [2007]):

- *$L \succeq 0$ (and symmetric), so all its eigenvalues are non-negative.*
- *Moreover, the multiplicity of eigenvalue 0 is the number of connected components in the graph (so it is always at least 1).*
- *Suppose there are $m \leq n$ connected components $A_1 \ldots A_m$ (a partition of $[n]$). Represent each component $c \in [m]$ by an indicator vector $\mathbb{1}^c \in \{0,1\}^n$ where*

$$\mathbb{1}^c_i = [\![\text{ vertex } i \text{ belongs to component } c \,]\!] \qquad \forall i \in [n]$$

Then $\{\mathbb{1}^1 \ldots \mathbb{1}^m\}$ is a basis of the zero eigenspace $E_{L,0}$.

1.2.5 Numerical Computation

Numerical computation of eigenvalues and eigenvectors is a deep subject beyond the scope of this thesis. Thorough treatments can be found in standard references such as Golub and Van Loan [2012]. Here, we supply basic results to give insight.

Consider computing eigenvectors (and their eigenvalues) of a diagonalizable matrix $A \in \mathbb{R}^{n\times n}$. A direct approach is to calculate the n roots of the polynomial $\det(A - \lambda I_{n\times n})$ in (1.19) and for each distinct root λ find an orthonormal basis of its eigenspace $E_{A,\lambda} = \{v : (A - \lambda I_{n\times n})v = 0\}$ in (1.20). Unfortunately, finding roots of a high-degree polynomial is a non-trivial problem of its own. But the following results provide more practical approaches to this problem.

1.2.5.1 Power Iteration

Theorem 1.2.9. *Let $A \in \mathbb{R}^{n \times n}$ be a nonzero symmetric matrix with eigenvalues $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$ and corresponding orthonormal eigenvectors $v_1 \dots v_n$. Let $v \in \mathbb{R}^n$ be a vector chosen at random. Then $A^k v$ converges to some multiple of v_1 as k increases.*

Proof. Since $v_1 \dots v_n$ form an orthonormal basis of $\mathbb{R}^n$ and v is randomly chosen from $\mathbb{R}^n$, $v = \sum_{i=1}^{n} c_i v_i$ for some nonzero $c_1 \dots c_n \in \mathbb{R}$. Therefore,

$$A^k v = \lambda_1^k \left(c_1 v_1 + \sum_{i=2}^{n} c_i \left(\frac{\lambda_i}{\lambda_1} \right)^k v_i \right)$$

Since $|\lambda_i / \lambda_1| < 1$ for $i = 2 \dots n$, the second term vanishes as k increases. □

A few remarks on Theorem 1.2.9:

- The proof suggests that the convergence rate depends on $\lambda_2 / \lambda_1 \in [0, 1)$. If this is zero, $k = 1$ yields an exact estimate $Av = \lambda_1 c_1 v_1$. If this is nearly one, it may take a large value of k before $A^k v$ converges.

- The theorem assumes $|\lambda_1| > |\lambda_2|$ for simplicity (this is called a spectral gap condition), but there are more sophisticated analyses that do not depend on this assumption (e.g., Halko *et al.* [2011]).

- Once we have an estimate $\hat{v}_1 = A^k v$ of the dominant eigenvector v_1, we can calculate an estimate $\hat{\lambda}_1$ of the corresponding eigenvalue by solving

$$\hat{\lambda}_1 = \arg\min_{\lambda \in \mathbb{R}} ||A\hat{v}_1 - \lambda \hat{v}_1||_2 \tag{1.28}$$

  whose closed-form solution is given by the Rayleigh quotient:

$$\hat{\lambda}_1 = \frac{\hat{v}_1^\top A \hat{v}_1}{\hat{v}_1^\top \hat{v}_1} \tag{1.29}$$

- Once we have an estimate $(\hat{v}_1, \hat{\lambda}_1)$ of (v_1, λ_1), we can perform a procedure called **deflation**

$$A' := A - \hat{\lambda}_1 \hat{v}_1 \hat{v}_1^\top \approx A - \lambda_1 v_1 v_1^\top = \sum_{i=2}^{n} \lambda_i v_i v_i^\top \tag{1.30}$$

  If $\hat{v}_1 = v_1$, the dominant eigenvector of A' is exactly v_2 which can be estimated in a similar manner.

Input: symmetric $A \in \mathbb{R}^{n \times n}$, number of desired eigen components $m \leq n$

Simplifying Assumption: the m dominant eigenvalues of A are nonzero and distinct, $|\lambda_1| > \cdots > |\lambda_m| > |\lambda_{m+1}| > 0$ ($\lambda_{m+1} = 0$ if $m = n$), with corresponding orthonormal eigenvectors $v_1 \dots v_m \in \mathbb{R}^n$

1. For $i = 1 \dots m$,
   (a) Initialize $\hat{v}_i \in \mathbb{R}^n$ randomly from a unit sphere.
   (b) Loop until convergence:
      i. $\hat{v}_i \leftarrow A\hat{v}_i$
      ii. $\hat{v}_i \leftarrow \hat{v}_i / ||\hat{v}_i||$
   (c) Compute the corresponding eigenvalue $\hat{\lambda}_i \leftarrow \hat{v}_i^\top A \hat{v}_i$.
   (d) Deflate $A \leftarrow A - \hat{\lambda}_i \hat{v}_i \hat{v}_i^\top$.

Output: estimate $(\hat{v}_i, \hat{\lambda}_i)$ of (v_i, λ_i) for $i = 1 \dots m$

Figure 1.3: A basic version of the power iteration method.

Theorem 1.2.9 suggests a scheme for finding eigenvectors of A, one by one, in the order of decreasing eigenvalues. This scheme is called the **power iteration method**; a basic version of the power method is given in Figure 1.3. Note that the eigenvector estimate is normalized in each iteration (Step 1(b)ii); this is a typical practice for numerical stability.

1.2.5.2 Orthogonal Iteration

Since the error introduced in deflation (1.30) propagates to the next iteration, the power method may be unstable for non-dominant eigen components. A natural generalization that remedies this problem is to find eigenvectors of A corresponding to the largest m eigenvalues simultaneously. That is, we start with m linearly independent vectors as columns of $\widetilde{V} = [\hat{v}_1 \dots \hat{v}_m]$ and compute

$$A^k \widetilde{V} = [A^k \hat{v}_1 \dots A^k \hat{v}_m]$$

Input: symmetric $A \in \mathbb{R}^{n \times n}$, number of desired eigen components $m \leq n$

Simplifying Assumption: the m dominant eigenvalues of A are nonzero and distinct, $|\lambda_1| > \cdots > |\lambda_m| > |\lambda_{m+1}| > 0$ ($\lambda_{m+1} = 0$ if $m = n$), with corresponding orthonormal eigenvectors $v_1 \dots v_m \in \mathbb{R}^n$

1. Initialize $\widehat{V} \in \mathbb{R}^{n \times m}$ such that $\widehat{V}^\top \widehat{V} = I_{m \times m}$ randomly.
2. Loop until convergence:
   (a) $\widetilde{V} \leftarrow A\widehat{V}$
   (b) Update $\widehat{V}$ to be an orthonormal basis of range($\widetilde{V}$) (e.g., by computing the QR decomposition: $[\widehat{V}, R] \leftarrow \mathrm{QR}(\widetilde{V})$).
3. Compute the corresponding eigenvalues $\widehat{\Lambda} \leftarrow \widehat{V}^\top A \widehat{V}$.
4. Reorder the columns of $\widehat{V}$ and $\widehat{\Lambda}$ in descending absolute magnitude of eigenvalues.

Output: estimate $(\widehat{V}, \widehat{\Lambda})$ of (V, Λ) where $V = [v_1 \dots v_m]$ and $\Lambda = \mathrm{diag}(\lambda_1 \dots \lambda_m)$.

Figure 1.4: A basic version of the subspace iteration method.

Note that each column of $A^k \widetilde{V}$ converges to the dominant eigen component (the case $m = 1$ degenerates to the power method). As long as the columns remain linearly independent (which we can maintain by orthogonalizing the columns in every multiplication by A), their span converges to the subspace spanned by the eigenvectors of A corresponding to the largest m eigenvalues under certain conditions (Chapter 7, Golub and Van Loan [2012]). Thus the desired eigenvectors can be recovered by finding an orthonormal basis of range($A^k \widetilde{V}$). The resulting algorithm is known as the **orthogonal iteration method** and a basic version of the algorithm is given in Figure 1.4. As in the power iteration method, a typical practice is to compute an orthonormal basis in each iteration rather than in the end to improve numerical stability; in particular, to prevent the estimate vectors from becoming linearly dependent (Step 2b).

1.2.5.3 Lanczos Method

We mention a final algorithm which is particularly effective when the goal is to compute only a small number of dominant eigenvalues of a large, sparse symmetric matrix. The algorithm is known as the **Lanczos method**; details of this method can be found in Chapter 9 of Golub and Van Loan [2012]. We give a sketch of the algorithm to illustrate its mechanics. This is the algorithm we use in in implementing our works. More specifically, we use the SVDLIBC package provided by Rohde [2007] which employs the single-vector Lanczos method.

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. The Lanczos method seeks an orthogonal matrix $Q_n \in \mathbb{R}^{n\times n}$ such that

$$T = Q_n^\top A Q_n = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & \cdots & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \ddots & \vdots \\ 0 & \beta_2 & \ddots & & \\ \vdots & & & \ddots & \beta_{n-1} \\ 0 & \cdots & & \beta_{n-1} & \alpha_n \end{bmatrix} \tag{1.31}$$

is a **tridiagonal** matrix (i.e., $T_{i,j} = 0$ except when $i \in \{j-1, j, j+1\}$). We know that such matrices exist since A is diagonalizable; for instance, Q_n can be orthonormal eigenvectors of A. Note that T is symmetric (since A is). From an eigendecompsition of $T = \widetilde{V}\widetilde{\Lambda}\widetilde{V}^\top$, we can recover an eigendecomposition of the original matrix $A = V\widetilde{\Lambda}V^\top$ where $V = Q_n\widetilde{V}$ since

$$A = Q_n T Q_n^\top = (Q_n\widetilde{V})\widetilde{\Lambda}(Q_n\widetilde{V})^\top$$

is an eigendecomposition. The Lanczos method is an iterative scheme to efficiently calculate Q_n and, in the process, simultaneously compute the tridiagonal entries $\alpha_1 \dots \alpha_n$ and $\beta_1 \dots \beta_{n-1}$.

Lemma 1.2.10. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric and $Q_n = [q_1 \dots q_n]$ be an orthogonal matrix such that $T = Q_n^\top A Q_n$ has the tridiagonal form in* (1.31)*. Define $q_0 = q_{n+1} = 0$, $\beta_0 = 1$, and*

$$r_i := Aq_i - \beta_{i-1}q_{i-1} - \alpha_i q_i \qquad\qquad 1 \le i \le n$$

Then

$$\alpha_i = q_i^\top A q_i \qquad 1 \leq i \leq n \tag{1.32}$$

$$\beta_i = ||r_i||_2 \qquad 1 \leq i \leq n-1 \tag{1.33}$$

$$q_{i+1} = r_i/\beta_i \qquad 1 \leq i \leq n-1 \tag{1.34}$$

Proof. Since $AQ_n = Q_n T$, by the tridiagonal structure of T,

$$Aq_i = \beta_{i-1} q_{i-1} + \alpha_i q_i + \beta_i q_{i+1} \qquad 1 \leq i \leq n$$

Multiplying on the left by q_i and using the orthonormality of $q_1 \dots q_n$, we verify $\alpha_i = q_i^\top A q_i$. Rearranging the expression gives

$$\beta_i q_{i+1} = Aq_i - \beta_{i-1} q_{i-1} - \alpha_i q_i = r_i$$

Since q_{i+1} is a unit vector for $1 \leq i \leq n-1$, we have $\beta_i = ||r_i||$ and $q_{i+1} = r_i/\beta_i$. □

Lemma 1.2.10 suggests that we can seed a random unit vector q_1 and iteratively compute $(\alpha_i, \beta_{i-1}, q_i)$ for all $i \leq n$. Furthermore, it can be shown that if we terminate this iteration early at $i = m \leq n$, the eigenvalues and eigenvectors of the resulting $m \times m$ tridiagonal matrix are a good approximation of the m dominant eigenvalues and eigenvectors of the original matrix A. It can also be shown that the Lanczos method converges faster than the power iteration method [Golub and Van Loan, 2012].

A basic version of the Lanczos method shown in Figure 1.5. The main computation in each iteration is a matrix-vector product $A\hat{q}_i$ which can be made efficient if A is sparse.

1.3 Singular Value Decomposition (SVD)

Singular value decomposition (SVD) is an application of eigendecomposition to factorize *any* matrix $A \in \mathbb{R}^{m \times n}$.

Input: symmetric $A \in \mathbb{R}^{n \times n}$ with dominant eigenvalues $|\lambda_1| \geq \cdots \geq |\lambda_m| > 0$ and corresponding orthonormal eigenvectors $v_1 \ldots v_m \in \mathbb{R}^n$, number of desired eigen components $m \leq n$

1. Initialize $\hat{q}_1 \in \mathbb{R}^n$ randomly from a unit sphere and let $\hat{q}_0 = 0$ and $\hat{\beta}_0 = 1$.

2. For $i = 1 \ldots m$,

   (a) Compute $\hat{\alpha}_i \leftarrow \hat{q}_i^\top A \hat{q}_i$. If $i < m$, compute:

   $$\hat{r}_i \leftarrow A\hat{q}_i - \hat{\beta}_{i-1}\hat{q}_{i-1} - \hat{\alpha}_i \hat{q}_i \qquad \hat{\beta}_i \leftarrow ||\hat{r}_i||_2 \qquad \hat{q}_{i+1} \leftarrow \hat{r}_i / \hat{\beta}_i$$

3. Compute the eigenvalues $|\hat{\lambda}_1| \geq \ldots \geq |\hat{\lambda}_m|$ and the corresponding orthonormal eigenvectors $\hat{w}_1 \ldots \hat{w}_m \in \mathbb{R}^m$ of the $m \times m$ tridiagonal matrix:

   $$\widehat{T} = \begin{bmatrix} \hat{\alpha}_1 & \hat{\beta}_1 & 0 & \cdots & 0 \\ \hat{\beta}_1 & \hat{\alpha}_2 & \hat{\beta}_2 & \ddots & \vdots \\ 0 & \hat{\beta}_2 & \ddots & & \\ \vdots & & & \ddots & \hat{\beta}_{m-1} \\ 0 & \cdots & & \hat{\beta}_{m-1} & \hat{\alpha}_m \end{bmatrix}$$

   (e.g., using the orthogonal iteration method).

4. Let $\hat{v}_i \leftarrow \widehat{Q}_m \hat{w}_i$ where $\widehat{Q}_m := [\hat{q}_1 \ldots \hat{q}_m] \in \mathbb{R}^{n \times m}$.

Output: estimate $(\hat{v}_i, \hat{\lambda}_i)$ of (v_i, λ_i) for $i = 1 \ldots m$

Figure 1.5: A basic version of the Lanczos method.

1.3.1 Derivation from Eigendecomposition

SVD can be derived from an observation that $AA^\top \in \mathbb{R}^{m \times m}$ and $A^\top A \in \mathbb{R}^{n \times n}$ are symmetric and PSD, and have the same number of nonzero (i.e., positive) eigenvalues since $\text{rank}(A^\top A) = \text{rank}(AA^\top) = \text{rank}(A)$.

Theorem 1.3.1. *Let $A \in \mathbb{R}^{m \times n}$. Let $\lambda_1 \geq \ldots \geq \lambda_m \geq 0$ denote the m eigenvalues of $AA^\top$*

and $\lambda'_1 \geq \ldots \geq \lambda'_n \geq 0$ *the* n *eigenvalues of* $A^\top A$*. Then*

$$\lambda_i = \lambda'_i \qquad 1 \leq i \leq \min\{m, n\} \tag{1.35}$$

Moreover, there exist orthonormal eigenvectors $u_1 \ldots u_m \in \mathbb{R}^m$ *of* $AA^\top$ *corresponding to* $\lambda_1 \ldots \lambda_m$ *and orthonormal eigenvectors* $v_1 \ldots v_n \in \mathbb{R}^n$ *of* $A^\top A$ *corresponding to* $\lambda'_1 \ldots \lambda'_n$ *such that*

$$A^\top u_i = \sqrt{\lambda_i} v_i \qquad 1 \leq i \leq \min\{m, n\} \tag{1.36}$$

$$A v_i = \sqrt{\lambda_i} u_i \qquad 1 \leq i \leq \min\{m, n\} \tag{1.37}$$

Proof. Let $u_1 \ldots u_m \in \mathbb{R}^m$ be orthonormal eigenvectors of $AA^\top$ corresponding to eigenvalues $\lambda_1 \geq \ldots \geq \lambda_m \geq 0$. Pre-multiplying $AA^\top u_i = \lambda_i u_i$ by $A^\top$ and $u_i^\top$, we obtain

$$A^\top A(A^\top u_i) = \lambda_i (A^\top u_i) \tag{1.38}$$

$$\lambda_i = \left|\left| A^\top u_i \right|\right|_2^2 \tag{1.39}$$

The first equality shows that $A^\top u_i$ is an eigenvector of $A^\top A$ corresponding to an eigenvalue λ_i. Since this holds for all i and both $AA^\top$ and $A^\top A$ have the same number of nonzero eigenvalues, we have (1.35).

Now, construct $v_1 \ldots v_m$ as follows. Let eigenvectors v_i of $A^\top A$ corresponding to nonzero eigenvalues $\lambda_i > 0$ be:

$$v_i = \frac{A^\top u_i}{\sqrt{\lambda_i}} \tag{1.40}$$

These vectors are unit-length eigenvectors of $A^\top A$ by (1.38) and (1.39). Furthermore, they are orthogonal: if $i \neq j$,

$$v_i^\top v_j = \frac{u_i^\top A A^\top u_j}{\sqrt{\lambda_i \lambda_j}} = \sqrt{\frac{\lambda_j}{\lambda_i}} u_i^\top u_j = 0$$

Let eigenvectors v_i of $A^\top A$ corresponding to zero eigenvalues $\lambda_i = 0$ be any orthonormal basis of $E_{A^\top A, 0}$. Since this subspace is orthogonal to eigenvectors of $A^\top A$ corresponding to nonzero eigenvalues, we conclude that all $v_1 \ldots v_m$ are orthonormal.

It remains to verify (1.36) and (1.37). For $\lambda_i > 0$, they follow immediately from (1.40). For $\lambda_i = 0$, $A^\top u_i$ and Av_i must be zero vectors since $\left|\left|A^\top u_i\right|\right|_2^2 = \lambda_i$ by (1.39) and also $||Av_i||_2^2 = v_i^\top A^\top A v_i = \lambda_i$; thus (1.36) and (1.37) hold trivially. □

The theorem validates the following definition.

Definition 1.3.1. *Let $A \in \mathbb{R}^{m\times n}$. Let $u_1 \dots u_m \in \mathbb{R}^m$ be orthonormal eigenvectors of $AA^\top$ corresponding to eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$, let $v_1 \dots v_n \in \mathbb{R}^n$ be orthonormal eigenvectors of $A^\top A$ corresponding to eigenvalues $\lambda'_1 \geq \cdots \geq \lambda'_n \geq 0$, such that*

$$\begin{aligned} \lambda_i &= \lambda'_i && 1 \leq i \leq \min\{m,n\} \\ A^\top u_i &= \sqrt{\lambda_i} v_i && 1 \leq i \leq \min\{m,n\} \\ Av_i &= \sqrt{\lambda_i} u_i && 1 \leq i \leq \min\{m,n\} \end{aligned}$$

*The **singular values** $\sigma_1 \dots \sigma_{\max\{m,n\}}$ of A are defined as:*

$$\sigma_i := \begin{cases} \sqrt{\lambda_i} & 1 \leq i \leq \min\{m,n\} \\ 0 & \min\{m,n\} < i \leq \max\{m,n\} \end{cases} \tag{1.41}$$

*The vector u_i is called a **left singular vector** of A corresponding to σ_i. The vector v_i is called a **right singular vector** of A corresponding to σ_i. Define*

- *$U \in \mathbb{R}^{m\times m}$ is an orthogonal matrix $U := [u_1 \dots u_m]$.*
- *$\Sigma \in \mathbb{R}^{m\times n}$ is a rectangular diagonal matrix with $\Sigma_{i,i} = \sigma_i$ for $1 \leq i \leq \min\{m,n\}$.*
- *$V \in \mathbb{R}^{n\times n}$ is an orthogonal matrix $V := [v_1 \dots v_n]$.*

*and note that $AV = U\Sigma$. This gives a **singular value decomposition (SVD)** of A:*

$$A = U\Sigma V^\top = \sum_{i=1}^{\min\{m,n\}} \sigma_i u_i v_i^\top \tag{1.42}$$

If A is already symmetric, there is a close relation between an eigendecomposition of A and an SVD of A.

Proposition 1.3.1. *If $A \in \mathbb{R}^{n\times n}$ is symmetric and $A = V\,\mathrm{diag}(\lambda_1 \dots \lambda_n)V^\top$ is an orthonormal eigendecomposition of A with $\lambda_1 \geq \cdots \geq \lambda_n$, then $A = V\,\mathrm{diag}(|\lambda_1| \dots |\lambda_n|)V^\top$ is an SVD of A.*

Proof. Since $V\,\mathrm{diag}(\lambda_1^2 \dots \lambda_n^2)V^\top$ is an eigendecomposition of $AA^\top$ and $A^\top A$, the i-th singular value of A is $\sigma_i = \sqrt{\lambda_i^2} = |\lambda_i|$ and the left and right singular vectors corresponding to σ_i are both the i-th column of V. □

Corollary 1.3.2. *If $A \in \mathbb{R}^{n\times n}$ is symmetric, an eigendecomposition of A and an SVD of A are the same iff $A \succeq 0$.*

As emphasized in Chapter 6.7 of Strang [2009], given a matrix $A \in \mathbb{R}^{m\times n}$ with rank r, an SVD yields an orthonormal basis for each of the four subspaces assocated with A:

$$\begin{aligned}\mathrm{col}(A) &= \mathrm{span}\{u_1 \dots u_r\}\\ \mathrm{row}(A) &= \mathrm{span}\{v_1 \dots v_r\}\\ \mathrm{null}(A) &= \mathrm{span}\{v_{r+1} \dots v_n\}\\ \text{left-null}(A) &= \mathrm{span}\{u_{r+1} \dots u_m\}\end{aligned}$$

A typical practice, however, is to only find singular vectors corresponding to a few dominant singular values (in particular, ignore zero singular values).

Definition 1.3.2 (Low-rank SVD)**.** *Let $A \in \mathbb{R}^{m\times n}$ with rank r. Let $u_1 \dots u_r \in \mathbb{R}^m$ and $v_1 \dots v_r \in \mathbb{R}^n$ be left and right singular vectors of A corresponding to the (only) positive singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Let $k \leq r$. A* ***rank-k SVD*** *of A is*

$$\widehat{A} = U_k \Sigma_k V_k^\top = \sum_{i=1}^{k} \sigma_i u_i v_i^\top \tag{1.43}$$

where $U_k := [u_1 \dots u_k] \in \mathbb{R}^{m\times k}$, $\Sigma_k := \mathrm{diag}(\sigma_1 \dots \sigma_k) \in \mathbb{R}^{k\times k}$, and $V_k := [v_1 \dots v_k] \in \mathbb{R}^{n\times k}$. Note that $\widehat{A} = A$ if $k = r$.

1.3.2 Variational Characterization

Theorem 1.3.3. *Let $A \in \mathbb{R}^{m\times n}$ with left singular vectors $u_1 \dots u_p \in \mathbb{R}^m$ and right singular vectors $v_1 \dots v_p \in \mathbb{R}^n$ corresponding to singular values $\sigma_1 \geq \dots \geq \sigma_p \geq 0$ where $p :=$*

$\min\{m,n\}$. *Let* $k \leq p$. *Consider maximizing* $u^\top Av$ *over unit-length vector pairs* $(u,v) \in \mathbb{R}^m \times \mathbb{R}^n$ *under orthogonality constraints:*

$$(u_i^*, v_i^*) = \underset{\substack{(u,v)\in\mathbb{R}^m\times\mathbb{R}^n: \\ ||u||_2=||v||_2=1 \\ u^\top u_j^*=v^\top v_j^*=0\ \forall j<i}}{\arg\max} u^\top Av \qquad \text{for } i = 1\dots k$$

Then an optimal solution is given by $(u_i^*, v_i^*) = (u_i, v_i)$.

Proof. The proof is similar to the proof of Theorem 1.2.7 and is omitted. □

Theorem 1.3.4. *Let* $A \in \mathbb{R}^{m\times n}$ *with left singular vectors* $u_1 \dots u_p \in \mathbb{R}^m$ *and right singular vectors* $v_1 \dots v_p \in \mathbb{R}^n$ *corresponding to singular values* $\sigma_1 \geq \dots \geq \sigma_p \geq 0$ *where* $p := \min\{m,n\}$. *Let* $k \leq p$. *Consider maximizing the trace of* $\overline{V}^\top A^\top A\overline{V} \in \mathbb{R}^{k\times k}$ *over orthonormal matrices* $\overline{V} \in \mathbb{R}^{n\times k}$*:*

$$\begin{aligned} V^* &= \underset{\overline{V}\in\mathbb{R}^{n\times k}:\ \overline{V}^\top\overline{V}=I_{k\times k}}{\arg\max} \operatorname{Tr}(\overline{V}^\top A^\top A\overline{V}) \\ &= \underset{\overline{V}\in\mathbb{R}^{n\times k}:\ \overline{V}^\top\overline{V}=I_{k\times k}}{\arg\max} \left|\left|A\overline{V}\right|\right|_F^2 \end{aligned}$$

where the second expression is by definition. Then an optimal solution is given by $V^* = [v_1 \dots v_k]$.

Proof. Since this is equivalent to the constrained optimization in Theorem 1.2.8 where the given symmetric matrix is $A^\top A$, the result follows from the definition of right singular vectors. □

1.3.3 Numerical Computation

Numerical computation of SVD is again an involved subject beyond the scope of this thesis. See Cline and Dhillon [2006] for references to a wide class of algorithms. Here, we give a quick remark on the subject to illustrate main ideas.

Let $A \in \mathbb{R}^{m\times n}$ with $m \leq n$ (if not, consider $A^\top$) and consider computing a rank-k SVD $U_k\Sigma_k V_k^\top$ of A. Since the columns of $U_k \in \mathbb{R}^{m\times k}$ are eigenvectors corresponding to the dominant k eigenvalues of $A^\top A \in \mathbb{R}^{m\times m}$ (which are squared singular values of A), we can compute an eigendecomposition of $A^\top A$ to obtain U_k and Σ_k, and finally recover $V_k = \Sigma_k^{-1}U_k^\top A$.

The core of many SVD algorithms is computing an eigendecomposition of $A^\top A$ efficiently without explicitly computing the matrix product. This can be done in various ways. For instance, we can modify the basic Lanczos algorithm in Figure 1.5 as follows: replace the matrix-vector product $A\hat{q}_i$ in Step 2a to $\hat{z}_i := A\hat{q}_i$ followed by $A^\top \hat{z}_i$. As another example, Matlab's sparse SVD (`svds`) computes an eigendecomposition of

$$B := \begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$$

and extracts the singular vectors and values of A from the eigendecomposition of B.

There is also a randomized algorithm for computing an SVD [Halko *et al.*, 2011]. While we do not use it in this thesis since other SVD algorithms are sufficiently scalable and efficient for our purposes, the randomized algorithm can potentially be used for computing an SVD of an extremely large matrix.

1.4 Perturbation Theory

Matrix perturbation theory is concerned with how properties of a matrix change when the matrix is perturbed by some noise. For instance, how "different" are the singular vectors of A from the singular vectors of $\widehat{A} = A + E$ where E is some noise matrix?

In the following, we let $\sigma_i(M) \geq 0$ denote the i-th largest singular value of M. We write $\angle\{u, v\}$ to denote the angle between nonzero vectors u, v taken in $[0, \pi]$.

1.4.1 Perturbation Bounds on Singular Values

Basic bounds on the singular values of a perturbed matrix are given below. They can also be used as bounds on eigenvalues for symmetric matrices.

Theorem 1.4.1 (Weyl [1912])**.** *Let $A, E \in \mathbb{R}^{m\times n}$ and $\widehat{A} = A + E$. Then*

$$\left|\sigma_i(\widehat{A}) - \sigma_i(A)\right| \leq ||E||_2 \qquad \forall i = 1 \ldots \min\{m, n\}$$

Theorem 1.4.2 (Mirsky [1960])**.** *Let $A, E \in \mathbb{R}^{m\times n}$ and $\widehat{A} = A + E$. Then*

$$\sum_{i=1}^{\min\{m,n\}} \left(\sigma_i(\widehat{A}) - \sigma_i(A)\right)^2 \leq ||E||_F^2$$

1.4.2 Canonical Angles Between Subspaces

To measure how "different" the singular vectors of A are from the singular vectors of $\widehat{A} = A + E$, we use the concept of an *angle* between the associated subspaces. This concept can be understood from the one-dimensional case. Suppose $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^n$ are subspaces of $\mathbb{R}^n$ with $\dim(\mathcal{X}) = \dim(\mathcal{Y}) = 1$. An acute angle $\angle\{\mathcal{X}, \mathcal{Y}\}$ between these subspaces can be calculated as:

$$\angle\{\mathcal{X}, \mathcal{Y}\} = \arccos \max_{\substack{x \in \mathcal{X},\, y \in \mathcal{Y}: \\ ||x||_2 = ||y||_2 = 1}} x^\top y$$

This is because $x^\top y = \cos \angle\{x, y\}$ for unit vectors x, y. Maximization ensures that the angle is acute. The definition can be extended as follows:

Definition 1.4.1. *Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^n$ be subspaces of $\mathbb{R}^n$ with $\dim(\mathcal{X}) = d$ and $\dim(\mathcal{Y}) = d'$. Let $m := \min\{d, d'\}$. The **canonical angles** between $\mathcal{X}$ and $\mathcal{Y}$ are defined as*

$$\angle_i\{\mathcal{X}, \mathcal{Y}\} := \arccos \max_{\substack{x \in \mathcal{X},\, y \in \mathcal{Y}: \\ ||x||_2 = ||y||_2 = 1 \\ x^\top x_j = y^\top y_j = 0 \;\; \forall j < i}} x^\top y \qquad \forall i = 1 \ldots m$$

*The **canonical angle matrix** between $\mathcal{X}$ and $\mathcal{Y}$ is defined as*

$$\angle\{\mathcal{X}, \mathcal{Y}\} := \operatorname{diag}(\angle_1\{\mathcal{X}, \mathcal{Y}\} \ldots \angle_m\{\mathcal{X}, \mathcal{Y}\})$$

Canonical angles can be found with SVD:

Theorem 1.4.3. *Let $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^{n \times d'}$ be orthonormal bases for $\mathcal{X} := \operatorname{range}(X)$ and $\mathcal{Y} := \operatorname{range}(Y)$. Let $X^\top Y = U \Sigma V^\top$ be a rank-$(\min\{d, d'\})$ SVD of $X^\top Y$. Then*

$$\angle\{\mathcal{X}, \mathcal{Y}\} = \arccos \Sigma$$

Proof. For all $1 \leq i \leq \min\{d, d'\}$,

$$\cos \angle_i\{\mathcal{X}, \mathcal{Y}\} = \max_{\substack{x \in \operatorname{range}(X) \\ y \in \operatorname{range}(Y): \\ ||x||_2 = ||y||_2 = 1 \\ x^\top x_j = y^\top y_j = 0 \;\; \forall j < i}} x^\top y = \max_{\substack{u \in \mathbb{R}^d,\, v \in \mathbb{R}^{d'}: \\ ||u||_2 = ||v||_2 = 1 \\ u^\top u_j = v^\top v_j = 0 \;\; \forall j < i}} u X^\top Y v = \sigma_i$$

where we solve for u, v in $x = Xu$ and $y = Yv$ under the same constraints (using the orthonormality of X and Y) to obtain the second equality. The final equality follows from a variational characterization of SVD. $\square$

Sine of the canonical angles The sine of the canoincal angles between subspaces is a natural measure of their difference partly because of its connection to the respective orthogonal projections (see Section 1.1.5).

Theorem 1.4.4 (Chapter 2, Stewart and Sun [1990]). *Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^n$ be subspaces of $\mathbb{R}^n$. Let $\Pi_{\mathcal{X}}, \Pi_{\mathcal{Y}} \in \mathbb{R}^{n\times n}$ be the (unique) orthogonal projections onto $\mathcal{X}, \mathcal{Y}$. Then*

$$||\sin \angle \{\mathcal{X}, \mathcal{Y}\}||_F = \frac{1}{\sqrt{2}} ||\Pi_{\mathcal{X}} - \Pi_{\mathcal{Y}}||_F$$

A result that connects the sine of canonical angles to singular values is the following:

Theorem 1.4.5 (Corollary 5.4, p. 43, Stewart and Sun [1990]). *Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^n$ be subspaces of $\mathbb{R}^n$ with the same dimension $\dim(\mathcal{X}) = \dim(\mathcal{Y}) = d$. Let $X, Y \in \mathbb{R}^{n\times d}$ be orthonormal bases of $\mathcal{X}, \mathcal{Y}$. Let $X_\perp, Y_\perp \in \mathbb{R}^{n\times(n-d)}$ be orthonormal bases of $\mathcal{X}^\perp, \mathcal{Y}^\perp$. Then the nonzero singular values of $Y_\perp^\top X$ or $X_\perp^\top Y$ are the sines of the nonzero canonical angles between $\mathcal{X}$ and $\mathcal{Y}$. In particular,*

$$||\sin \angle \{\mathcal{X}, \mathcal{Y}\}|| = \left|\left|Y_\perp^\top X\right|\right| = \left|\left|X_\perp^\top Y\right|\right| \tag{1.44}$$

where the norm can be $||\cdot||_2$ or $||\cdot||_F$.

1.4.3 Perturbation Bounds on Singular Vectors

Given the concept of canonical angles, We are now ready to state important bounds on the top singular vectors of a perturbed matrix attributed to Wedin [1972].

Theorem 1.4.6 (Wedin, spectral norm, p. 262, Theorem 4.4, Stewart and Sun [1990]). *Let $A, E \in \mathbb{R}^{m\times n}$ and $\widehat{A} = A + E$. Assume $m \geq n$. Let $A = U\Sigma V^\top$ and $\widehat{A} = \widehat{U}\widehat{\Sigma}\widehat{V}^\top$ denote SVDs of A and $\widehat{A}$. Choose the number of the top singular components $k \in [n]$ and write*

$$A = [U_1 U_2 U_3] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{bmatrix} [V_1 V_2]^\top \qquad \widehat{A} = \left[\widehat{U}_1 \widehat{U}_2 \widehat{U}_3\right] \begin{bmatrix} \widehat{\Sigma}_1 & 0 \\ 0 & \widehat{\Sigma}_2 \\ 0 & 0 \end{bmatrix} \left[\widehat{V}_1 \widehat{V}_2\right]^\top$$

where the matrices (U_1, Σ_1, V_1) *with* $\Sigma_1 \in \mathbb{R}^{k\times k}$, (U_2, Σ_2, V_2) *with* $\Sigma_2 \in \mathbb{R}^{(n-k)\times(n-k)}$, *and a leftover* $U_3 \in \mathbb{R}^{m\times(m-n)}$ *represent a k-partition of* $U\Sigma V$ *(analogously for* $\widehat{A}$*). Let*

$$\Phi := \angle\left\{\text{range}(U_1), \text{range}(\widehat{U}_1)\right\}$$
$$\Theta := \angle\left\{\text{range}(V_1), \text{range}(\widehat{V}_1)\right\}$$

If there exist $\alpha, \delta > 0$ *such that* $\sigma_k(\widehat{A}) \geq \alpha + \delta$ *and* $\sigma_{k+1}(A) \leq \alpha$, *then*

$$||\sin\Phi||_2 \leq \frac{||E||_2}{\delta} \qquad\qquad ||\sin\Theta||_2 \leq \frac{||E||_2}{\delta}$$

We point out some subtle aspects of Theorem 1.4.6. First, we can choose any matrices to bound $||\sin\Phi||_2$ as long as they have U_1 and $\widehat{U}_1$ as their top k left singular vectors (analogously for Θ). Second, we can flip the ordering of the singular value constraints (i.e., we can choose which matrix to treat as the original). For example, let $\widetilde{A} \in \mathbb{R}^{m\times n}$ be any matrix whose top k left singular vectors are $\widehat{U}_1$ (e.g., $\widetilde{A} = \widehat{U}_1\widehat{\Sigma}_1\widehat{V}_1^\top$). The theorem implies that if there exist $\alpha, \delta > 0$ such that $\sigma_k(A) \geq \alpha + \delta$ and $\sigma_{k+1}(\widetilde{A}) \leq \alpha$, then

$$||\sin\Phi||_2 \leq \frac{\left|\left|\widetilde{A} - A\right|\right|_2}{\delta}$$

There is also a Frobenius norm version of Wedin, which is provided here for completeness:

Theorem 1.4.7 (Wedin, Frobenius norm, p. 260, Theorem 4.1, Stewart and Sun [1990])**.** *Assume the same notations in Theorem 1.4.6. If there exists* $\delta > 0$ *such that* $\sigma_k(\widehat{A}) \geq \delta$ *and* $\min_{i=1\ldots k,\, j=k+1\ldots n}\left|\sigma_i(\widehat{A}) - \sigma_j(A)\right| \geq \delta$, *then*

$$||\sin\Phi||_F^2 + ||\sin\Theta||_F^2 \leq \frac{2\,||E||_F^2}{\delta^2}$$

Applications of Wedin to low-rank matrices Simpler versions of Wedin's theorem can be derived by assuming that A has rank k (i.e., our choice of the number of the top singular components exactly matches the number of nonzero singular values of A). This simplifies the condition in Theorem 1.4.6 because $\sigma_{k+1}(A) = 0$.

Theorem 1.4.8 (Wedin, Corollary 22, Hsu *et al.* [2012])**.** *Assume the same notations in Theorem 1.4.6. Assume* $\text{rank}(A) = k$ *and* $\text{rank}(\widehat{A}) \geq k$. *If* $\left|\left|\widehat{A} - A\right|\right|_2 \leq \epsilon\sigma_k(A)$ *for some*

$\epsilon < 1$, *then*

$$||\sin\Phi||_2 \leq \frac{\epsilon}{1-\epsilon} \qquad\qquad ||\sin\Theta||_2 \leq \frac{\epsilon}{1-\epsilon}$$

Proof. For any value of $\alpha > 0$, define $\delta := \sigma_k(\widehat{A}) - \alpha$. Since $\sigma_k(\widehat{A})$ is positive, we can find a sufficiently small α such that δ is positive, thus the conditions $\sigma_k(\widehat{A}) \geq \alpha + \delta$ and $\sigma_{k+1}(A) = 0 \leq \alpha$ in Theorem 1.4.6 are satisfied. It follows that

$$||\sin\Phi||_2 \leq \frac{\left|\left|\widehat{A} - A\right|\right|_2}{\delta} = \frac{\left|\left|\widehat{A} - A\right|\right|_2}{\sigma_k(\widehat{A}) - \alpha}$$

Since this is true for any $\alpha > 0$, we can take limit $\alpha \to 0$ on both sides to obtain

$$||\sin\Phi||_2 \leq \frac{\left|\left|\widehat{A} - A\right|\right|_2}{\sigma_k(\widehat{A})} \leq \frac{\epsilon\sigma_k(A)}{\sigma_k(\widehat{A})} \leq \frac{\epsilon\sigma_k(A)}{(1-\epsilon)\sigma_k(A)} = \frac{\epsilon}{1-\epsilon}$$

where the last inequality follows from Weyl's inequality: $\sigma_k(\widehat{A}) \geq (1-\epsilon)\sigma_k(A)$. The bound on the right singular vectors can be shown similarly. □

It is also possible to obtain a different bound by using an alternative argument.

Theorem 1.4.9 (Wedin)**.** *Assume the same notations in Theorem 1.4.6. Assume* $\text{rank}(A) = k$ *and* $\text{rank}(\widehat{A}) \geq k$*. If* $\left|\left|\widehat{A} - A\right|\right|_2 \leq \epsilon\sigma_k(A)$ *for some* $\epsilon < 1$*, then*

$$||\sin\Phi||_2 \leq 2\epsilon \qquad\qquad ||\sin\Theta||_2 \leq 2\epsilon$$

Proof. Define $\widetilde{A} := \widehat{U}_1\widehat{\Sigma}_1\widehat{V}_1^\top$. Note that $\left|\left|\widehat{A} - \widetilde{A}\right|\right|_2 \leq \left|\left|\widehat{A} - A\right|\right|_2$ since $\widetilde{A}$ is the optimal rank-k approximation of $\widehat{A}$ in $||\cdot||_2$ (Theorem 2.2.1). Then by the triangle inequality,

$$\left|\left|\widetilde{A} - A\right|\right|_2 \leq \left|\left|\widehat{A} - \widetilde{A}\right|\right|_2 + \left|\left|\widehat{A} - A\right|\right|_2 \leq 2\epsilon\sigma_k(A)$$

We now apply Theorem 1.4.6 with $\widetilde{A}$ as the original matrix and A as a perturbed matrix (see the remark below Theorem 1.4.6). Since $\sigma_k(A) > 0$ and $\sigma_{k+1}(\widetilde{A}) = 0$, we can use the same limit argument in the proof of Theorem 1.4.8 to have

$$||\sin\Phi||_2 \leq \frac{\left|\left|A - \widetilde{A}\right|\right|_2}{\sigma_k(A)} \leq \frac{2\epsilon\sigma_k(A)}{\sigma_k(A)} = 2\epsilon$$

The bound on the right singular vectors can be shown similarly. □

All together, we can state the following convenient corollary.

Corollary 1.4.10 (Wedin)**.** *Let $A \in \mathbb{R}^{m\times n}$ with rank k. Let $E \in \mathbb{R}^{m\times n}$ be a noise matrix and assume that $\widehat{A} := A + E$ has rank at least k. Let $A = U\Sigma V^\top$ and $\widehat{A} = \widehat{U}\widehat{\Sigma}\widehat{V}^\top$ denote rank-k SVDs of A and $\widehat{A}$. If $||E||_2 \leq \epsilon\sigma_k(A)$ for some $\epsilon < 1$, then for any orthonormal bases $U_\perp, \widehat{U}_\perp$ of $\text{range}(U)^\perp, \text{range}(\widehat{U})^\perp$ and $V_\perp, \widehat{V}_\perp$ of $\text{range}(V)^\perp, \text{range}(\widehat{V})^\perp$, we have*

$$\left|\left|\widehat{U}_\perp^\top U\right|\right|_2 = \left|\left|U_\perp^\top \widehat{U}\right|\right|_2 \leq \min\left\{\frac{\epsilon}{1-\epsilon},\ 2\epsilon\right\}$$
$$\left|\left|\widehat{V}_\perp^\top V\right|\right|_2 = \left|\left|V_\perp^\top \widehat{V}\right|\right|_2 \leq \min\left\{\frac{\epsilon}{1-\epsilon},\ 2\epsilon\right\}$$

Note that if $\epsilon < 1/2$ the bound $\epsilon/(1-\epsilon) < 2\epsilon < 1$ is tighter.

Proof. The statement follows from Theorem 1.4.8, 1.4.9, and 1.4.5. □

It is also possible to derive a version of Wedin that does not involve orthogonal complements. The proof illustrates a useful technique: given any orthonormal basis $U \in \mathbb{R}^{m\times k}$,

$$I_{m\times m} = UU^\top + U_\perp U_\perp^\top$$

This allows for a decomposition of any vector in $\mathbb{R}^m$ into $\text{range}(U)$ and $\text{range}(U_\perp)$.

Theorem 1.4.11 (Wedin)**.** *Let $A \in \mathbb{R}^{m\times n}$ with rank k and $\widehat{A} \in \mathbb{R}^{m\times n}$ with rank at least k. Let $U, \widehat{U} \in \mathbb{R}^{m\times k}$ denote the top k left singular vectors of $A, \widehat{A}$. If $\left|\left|\widehat{A} - A\right|\right|_2 \leq \epsilon\sigma_k(A)$ for some $\epsilon < 1/2$, then*

$$\left|\left|\widehat{U}^\top x\right|\right|_2 \geq \sqrt{1-\epsilon_0^2}\ ||x||_2 \qquad \forall x \in \text{range}(U) \tag{1.45}$$

where $\epsilon_0 := \epsilon/(1-\epsilon) < 1$.

Proof. Since we have $||y||_2 = ||Uy||_2 = \left|\left|\widehat{U}y\right|\right|_2$ for any $y \in \mathbb{R}^k$, we can write

$$||y||_2^2 = \left|\left|\widehat{U}\widehat{U}^\top Uy\right|\right|_2^2 + \left|\left|\widehat{U}_\perp\widehat{U}_\perp^\top Uy\right|\right|_2^2 = \left|\left|\widehat{U}^\top Uy\right|\right|_2^2 + \left|\left|\widehat{U}_\perp^\top Uy\right|\right|_2^2$$

By Corollary 1.4.10, we have $\left|\left|\widehat{U}_\perp^\top U\right|\right|_2^2 \leq \epsilon_0^2 < 1$, thus

$$\left|\left|\widehat{U}^\top Uy\right|\right|_2^2 \geq \left(1 - \left|\left|\widehat{U}_\perp^\top U\right|\right|_2^2\right)||y||_2^2 \geq \left(1-\epsilon_0^2\right)||y||_2^2 \qquad \forall y \in \mathbb{R}^k$$

Then the claim follows. □

1.4.3.1 Examples

We now show how Wedin's theorem can be used in practice with some examples. In these examples, we assume a matrix $A \in \mathbb{R}^{m \times n}$ with rank k and an empirical estimate $\widehat{A}$ with rank at least k. Let $U, \widehat{U} \in \mathbb{R}^{m \times k}$ denote the top k left singular vectors of $A, \widehat{A}$.

In order to apply Wedin's theorem, we must establish that the empirical estimate $\widehat{A}$ is sufficiently accurate, so that

$$\left|\left|\widehat{A} - A\right|\right|_2 \leq \epsilon \sigma_k(A) \qquad \epsilon < 1/2 \tag{1.46}$$

Note that the condition depends on the smallest positive singular value of A. Let $\epsilon_0 := \epsilon/(1-\epsilon) < 1$.

Example 1.4.1 (Empirical invertibility, Hsu *et al.* [2008])**.** *Let $O \in \mathbb{R}^{m \times k}$ be a matrix such that* range(O) = range(U)*. Note that $U^\top O \in \mathbb{R}^{k \times k}$ is invertible. We now show that $\widehat{U}^\top O$ is also invertible if* (1.46) *holds. Apply* (1.45) *with $x = Oz$ to obtain:*

$$\left|\left|\widehat{U}^\top O z\right|\right|_2 \geq \sqrt{1-\epsilon_0^2}\ ||Oz||_2 \qquad \forall z \in \mathbb{R}^k$$

Since $\sigma_i(M)$ is the maximum of $||Mz||_2$ over orthonormally constrained z, this implies

$$\sigma_i\left(\widehat{U}^\top O\right) \geq \sqrt{1-\epsilon_0^2}\ \sigma_i(O) \qquad \forall i \in [k]$$

In particular, $\sigma_k\left(\widehat{U}^\top O\right) > 0$ and thus $\widehat{U}^\top O$ is invertible.

Example 1.4.2 (Empirical separability, Chapter 5 of the thesis)**.** *Assume that $Q \in \mathbb{R}^{k \times k}$ is an orthogonal matrix with columns $q_i \in \mathbb{R}^k$. That is,*

$$q_i^\top q_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Let $\widehat{Q} := \widehat{U}^\top U Q$ with columns $\hat{q}_i \in \mathbb{R}^k$. We can bound the separation between the columns $\hat{q}_i$ assuming that (1.46) *holds. By Corollary 1.4.10, we have $\left|\left|\widehat{U}_\perp^\top U q_i\right|\right|_2 \leq \epsilon_0$. Then since*

$$||q_i||^2 = \left|\left|\widehat{U}\widehat{U}^\top U q_i\right|\right|^2 + \left|\left|\widehat{U}_\perp \widehat{U}_\perp^\top U q_i\right|\right|^2 = ||\hat{q}_i||^2 + \left|\left|\widehat{U}_\perp^\top U q_i\right|\right|^2 = 1$$

we have

$$\hat{q}_i^\top \hat{q}_i = 1 - \left|\left|\widehat{U}_\perp^\top U q_i\right|\right|_2^2 \geq 1 - \epsilon_0^2$$

Also, if $i \neq j$,

$$\begin{aligned}
\hat{q}_i^\top \hat{q}_j &= q_i^\top U^\top \widehat{U}\widehat{U}^\top U q_j \\
&= q_i^\top U^\top \left(I_{m\times m} - \widehat{U}_\perp \widehat{U}_\perp^\top \right) U q_j \\
&= q_i^\top q_j - q_i U^\top \widehat{U}_\perp \widehat{U}_\perp^\top U q_j \\
&= -q_i U^\top \widehat{U}_\perp \widehat{U}_\perp^\top U q_j \\
&\leq \left|\left| \widehat{U}_\perp^\top U q_i \right|\right|_2 \left|\left| \widehat{U}_\perp^\top U q_j \right|\right|_2 \\
&\leq \epsilon_0^2
\end{aligned}$$

where the first inequality is the Cauchy-Schwarz inequality.

Chapter 2

Examples of Spectral Techniques

This chapter gives examples of spectral techniques in the literature to demonstrate the range of spectral applications.

2.1 The Moore–Penrose Pseudoinverse

The **Moore–Penrose pseudoinverse** (or just the pseudoinverse) of a matrix $A \in \mathbb{R}^{m\times n}$ is the unique matrix $A^+ \in \mathbb{R}^{n\times m}$ such that[1]

1. $AA^+ \in \mathbb{R}^{n\times n}$ is the orthogonal projection onto range(A), and
2. $A^+A \in \mathbb{R}^{m\times m}$ is the orthogonal projection onto row(A).

A simple construction of the pseudoinverse A^+ is given by an SVD of A.

Proposition 2.1.1. *Let $A \in \mathbb{R}^{m\times n}$ with $r := \operatorname{rank}(A) \leq \min\{m, n\}$. Let $A = U\Sigma V^\top$ denote a rank-r SVD of A. Then $A^+ := V\Sigma^{-1}U^\top \in \mathbb{R}^{m\times n}$ is a matrix such that $AA^+ \in \mathbb{R}^{n\times n}$*

[1]This is a simplified definition sufficient for the purposes of the thesis: see Section 6.7 of Friedberg *et al.* [2003] for a formal treatment. It is defined as the matrix corresponding to a (linear) function $L : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$L(v) = \begin{cases} u & \text{if } \exists u \text{ such that } v = Au \text{ (i.e., } v \in \text{range}(A)) \\ 0 & \text{otherwise} \end{cases}$$

from which the properties in the main text follow.

is the orthogonal projection onto $\text{range}(A)$ *and* $A^+A \in \mathbb{R}^{m\times m}$ *is the orthogonal projection onto* $\text{row}(A)$.

Proof. The orthogonal projections onto $\text{range}(A)$ and $\text{row}(A)$ are respectively given by $UU^\top$ and $VV^\top$, and since $U^\top U = V^\top V = I_{r\times r}$,

$$AA^+ = U\Sigma V^\top V\Sigma^{-1}U^\top = UU^\top$$
$$A^+A = V\Sigma^{-1}U^\top U\Sigma V^\top = VV^\top$$

□

The pseudoinverse A^+ is the unique minimizer of $||AX - I_{m\times m}||_F$ over $X \in \mathbb{R}^{n\times m}$ (p. 257, Golub and Van Loan [2012]) and can be seen as a generalization of matrix inverse:

- If A has linearly independent columns (so $A^\top A$ is invertible),

$$AA^+ = I_{m\times m}$$
$$A^+ = (A^\top A)^{-1}A^\top$$

- If A has linearly independent rows (so $AA^\top$ is invertible),

$$A^+A = I_{n\times n}$$
$$A^+ = A^\top(AA^\top)^{-1}$$

- If A is square and has full rank, then $A^+ = A^{-1}$.

2.2 Low-Rank Matrix Approximation

A celebrated application of SVD is the low-rank matrix approximation problem:

Theorem 2.2.1 (Eckart and Young [1936], Mirsky [1960])**.** *Let* $A \in \mathbb{R}^{m\times n}$. *Let* $k \leq \min\{m, n\}$ *and consider*

$$Z^* = \underset{Z\in\mathbb{R}^{m\times n}:\ \text{rank}(Z)\leq k}{\arg\min} ||A - Z|| \tag{2.1}$$

where $||\cdot||$ *is an orthogonally invariant norm:* $||M|| = ||QMR||$ *for orthogonal* Q *and* R *(e.g., the Frobenius norm* $||\cdot||_F$*, the spectral norm* $||\cdot||_2$*). Then an optimal solution is given by a rank-*k *SVD of* A*,* $Z^* = U_k\Sigma_k V_k^\top$.

Proof. Let $A = U\Sigma V$ be an SVD of A. Then

$$||A - Z||^2 = \left|\left|\Sigma - \overline{Z}\right|\right|^2 = \sum_{i=1}^{r} \left(\sigma_i - \overline{Z}_{i,i}\right)^2 + \sum_{i \neq j} \overline{Z}_{i,j}^2$$

where $\overline{Z} := U^\top Z V \in \mathbb{R}^{m \times n}$ has rank k. This is minimized (uniquely if $\sigma_k > \sigma_{k+1}$) at $\sum_{i=k+1}^{r} \sigma_i^2$ by a rectangular diagonal matrix $\overline{Z}_{i,i} = \sigma_i$ for $1 \leq i \leq k$, which implies $Z = U_k \Sigma_k V_k$. □

It is illuminating to examine a closely related unconstrained problem:

$$\{b_i^*\}_{i=1}^m, \{c_i^*\}_{i=1}^n = \underset{\substack{b_1 \ldots b_m \in \mathbb{R}^k \\ c_1 \ldots c_n \in \mathbb{R}^k}}{\arg\min} \sum_{i,j} \left(A_{i,j} - b_i^\top c_j\right)^2 \tag{2.2}$$

which in matrix form can be written as

$$(B^*, C^*) = \underset{\substack{B \in \mathbb{R}^{k \times m} \\ C \in \mathbb{R}^{k \times n}}}{\arg\min} \left|\left|A - B^\top C\right|\right|_F \tag{2.3}$$

This is equivalent to (2.1) (with the Frobenius norm) since any matrix with rank at most k can be expressed as $B^\top C$ (e.g., by SVD) and $\text{rank}(B^\top C) \leq k$. It has infinite level sets since $\left|\left|A - B^\top C\right|\right|_F = \left|\left|A - \overline{B}^\top \overline{C}\right|\right|_F$ for $\overline{B} = Q^\top B$ and $\overline{C} = Q^{-1} C$ where Q is any $k \times k$ invertible matrix. For convenience, we can fix the form $B = \sqrt{\widetilde{\Sigma}_k}\widetilde{U}_k^\top$ and $C = \sqrt{\widetilde{\Sigma}_k}\widetilde{V}_k^\top$ by a rank-k SVD of $B^\top C = \widetilde{U}_k \widetilde{\Sigma}_k \widetilde{V}_k^\top$. The stationary conditions of (2.3) are then

$$A\widetilde{V}_k = \widetilde{U}_k \widetilde{\Sigma}_k \qquad\qquad A^\top \widetilde{U}_k = \widetilde{V}_k \widetilde{\Sigma}_k$$

which imply that each stationary point is given by some k singular components of A. In particular, the global minima are given by components corresponding to the largest k singular values (Theorem 2.2.1). Surprisingly, all other stationary points are saddle points; a proof can be found on page 29 of Ho [2008]. Thus (2.2) is a (very special type of) non-convex objective for which SVD provides a global minimum.

A slight variant of (2.2) is the following:

$$\{b_i^*\}_{i=1}^m, \{c_i^*\}_{i=1}^n = \underset{\substack{b_1 \ldots b_m \in \mathbb{R}^k \\ c_1 \ldots c_n \in \mathbb{R}^k}}{\arg\min} \sum_{i,j} W_{i,j} \left(A_{i,j} - b_i^\top c_j\right)^2 \tag{2.4}$$

where $W \in \mathbb{R}^{n \times m}$ is a non-negative weight matrix. Unfortunately, there is no SVD-based closed-form solution to this problem [Srebro *et al.*, 2003]. Unlike the unweighted case, the

objective has local optima that are not saddle points and can be shown to be generally NP-hard [Gillis and Glineur, 2011]. Despite the intractability, (2.4) is successfully optimized by iterative methods (e.g., gradient descent) in numerous practical applications such as recommender systems [Koren *et al.*, 2009] and word embeddings [Pennington *et al.*, 2014].

2.3 Finding the Best-Fit Subspace

A very practical interpretation of SVD is that of projecting data points to the "closest" lower-dimensional subspace. Specifically, let $x^{(1)} \dots x^{(M)} \in \mathbb{R}^d$ be M data points in $\mathbb{R}^d$. Given $k \leq d$, we wish to find an orthonormal basis $V^* = [v_1^* \dots v_k^*] \in \mathbb{R}^{d \times k}$ of a k-dimensional subspace such that

$$V^* = \underset{V \in \mathbb{R}^{d \times k} \colon V^\top V = I_{k \times k}}{\arg\min} \sum_{i=1}^{M} \left|\left| x^{(i)} - VV^\top x^{(i)} \right|\right|_2 \tag{2.5}$$

The subspace $\text{span}\{v_1^* \dots v_k^*\}$ is called the **best-fit subspace**. Since $x^{(i)} - VV^\top x^{(i)}$ is orthogonal to $VV^\top x^{(i)}$, by the Pythagorean theorem

$$\left|\left| x^{(i)} - VV^\top x^{(i)} \right|\right|_2^2 = \left|\left| x^{(i)} \right|\right|_2^2 - \left|\left| VV^\top x^{(i)} \right|\right|_2^2 = \left|\left| x^{(i)} \right|\right|_2^2 - \left|\left| V^\top x^{(i)} \right|\right|_2^2$$

Let $X \in \mathbb{R}^{M \times d}$ be a data matrix whose i-th row is given by $x^{(i)}$. Since $\sum_{i=1}^{M} \left|\left| V^\top x^{(i)} \right|\right|_2^2 = \text{Tr}(V^\top X^\top XV) = ||XV||_F^2$, (2.5) is equivalent to

$$V^* = \underset{V \in \mathbb{R}^{d \times k} \colon V^\top V = I_{k \times k}}{\arg\max} ||XV||_F \tag{2.6}$$

An optimal solution is given by $V^* = V_k$ where $U_k \Sigma_k V_k^\top$ is a rank-k SVD of X. The projected data points are given by the rows of $\underline{X} \in \mathbb{R}^{M \times k}$ where

$$\underline{X} = XV_k = U_k \Sigma_k \tag{2.7}$$

2.4 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a classical spectral technique for dimensionality reduction [Pearson, 1901]. A standard formulation of PCA is as follows [Jolliffe, 2002].

Given a random variable $X \in \mathbb{R}^d$, we wish to find $m \leq d$ vectors $a_1 \dots a_m \in \mathbb{R}^d$ such that for each $i = 1 \dots m$:

$$\begin{aligned} a_i = &\arg\max_{a \in \mathbb{R}^d} \quad \mathrm{Var}(a^\top X) \\ &\text{subject to} \quad ||a||_2 = 1, \text{ and} \\ &\qquad\qquad a^\top a_j = 0 \text{ for all } j < i \end{aligned} \tag{2.8}$$

That is, $a_1 \dots a_m$ are orthonormal vectors such that a_i is the direction of the i-th largest variance of X. We express the objective in terms of the covariance matrix:

$$C_X := \boldsymbol{E}\left[(X - \boldsymbol{E}[X])(X - \boldsymbol{E}[X])^\top\right]$$

as $\mathrm{Var}(a^\top X) = a^\top C_X a$. Since $C_X \succeq 0$, it has an eigendecomposition of the form $C_X = U \Lambda U^\top$ where $U = [u_1 \dots u_d]$ is orthonormal and $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$ with $\lambda_1 \geq \dots \geq \lambda_d \geq 0$. Then a solution

$$\begin{aligned} a_i = &\arg\max_{a \in \mathbb{R}^d} \quad a^\top C_X a \\ &\text{subject to} \quad ||a||_2 = 1, \text{ and} \\ &\qquad\qquad a^\top a_j = 0 \text{ for all } j < i \end{aligned} \tag{2.9}$$

is given by $a_i = u_i$ and the value of the maximized variance is the eigenvalue λ_i since $\mathrm{Var}(a_i^\top X) = a_i^\top C_X a_i = \lambda_i$.

2.4.1 Best-Fit Subspace Interpretation

Let $x^{(1)} \dots x^{(M)}$ be M samples of X with the sample mean $\hat{\mu} := \sum_{i=1}^M x^{(i)}/M$. The sample covariance matrix is:

$$\hat{C}_X = \frac{1}{M} \sum_{i=1}^{M} (x^{(i)} - \hat{\mu})(x^{(i)} - \hat{\mu})^\top$$

By pre-processing the data as $\bar{x}^{(i)} := (x^{(i)} - \hat{\mu})/\sqrt{M}$ and organizing it into a matrix $\bar{X} \in \mathbb{R}^{M \times d}$ where $\bar{X}_i = \bar{x}^{(i)}$, we can write:

$$\hat{C}_X = \bar{X}^\top \bar{X}$$

Let $\bar{X} = \hat{U}\hat{\Sigma}\hat{V}^\top$ be an SVD of $\bar{X}$ where $\hat{\Sigma} = \mathrm{diag}(\hat{\sigma}_1 \dots \hat{\sigma}_d)$ is a diagonal matrix of ordered singular values $\hat{\sigma}_1 \geq \dots \geq \hat{\sigma}_d \geq 0$ and $\hat{V} = [\hat{v}_1 \dots \hat{v}_d]$ is the orthonormal matrix of right singular vectors. Since $\hat{C}_X = \hat{V}\hat{\Sigma}^2\hat{V}$ and it is an eigendecomposition in the desired form, the i-th PCA direction is given by $\hat{a}_i = \hat{v}_i$ and the value of the maximized variance is $\hat{\sigma}_i^2$. We make a few observations on this result:

- There is no need to explicitly compute the sample covariance matrix $\hat{C}_X$ and its eigendecomposition. We can directly apply an SVD on the data matrix $\bar{X}$.
- Since $\hat{a_1} \dots \hat{a_m}$ are the right singular vectors of $\sqrt{M}\bar{X}$ corresponding to the largest m singular values, the orthogonal projection $\hat{\Pi} := [\hat{a_1} \dots \hat{a_m}][\hat{a_1} \dots \hat{a_m}]^\top$ minimizes

$$\sum_{i=1}^{M} \left|\left|(x^{(i)} - \hat{\mu}) - \hat{\Pi}(x^{(i)} - \hat{\mu})\right|\right|^2$$

  Hence PCA can be interpreted as finding the best-fit subspace of mean-centered data points.

2.5 Canonical Correlation Analysis (CCA)

Canonical correlation analysis (CCA) is a classical spectral technique for analyzing the correlation between two variables [Hotelling, 1936]. A standard formulation of CCA is as follows [Hardoon *et al.*, 2004]. Given a pair of random variables $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$, we wish to find $m \leq \min(d, d')$ pairs of vectors $(a_1, b_1) \dots (a_m, b_m) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ such that for each $i = 1 \dots m$:

$$\begin{aligned} (a_i, b_i) = \underset{(a,b)\in\mathbb{R}^d\times\mathbb{R}^{d'}}{\arg\max} \quad & \mathrm{Cor}(a^\top X, b^\top Y) \\ \text{subject to} \quad & \mathrm{Cor}(a^\top X, a_j^\top X) = 0 \ \text{ for all } j < i \\ & \mathrm{Cor}(b^\top Y, b_j^\top Y) = 0 \ \text{ for all } j < i \end{aligned} \tag{2.10}$$

That is, (a_i, b_i) projects (X, Y) to 1-dimensional random variables $(a_i^\top X, b_i^\top Y)$ that are maximally correlated, but $a_i^\top X$ is uncorrelated to $a_j^\top X$ for all $j < i$ (respectively for Y). Note that the solution is not unique because the correlation coefficient $\mathrm{Cor}(Y, Z)$ is invariant

under separate linear transformations on $Y, Z \in \mathbb{R}$:

$$\mathrm{Cor}(\alpha Y + \gamma, \beta Z + \lambda) = \mathrm{Cor}(Y, Z)$$

for any constants $\alpha, \beta, \gamma, \lambda \in \mathbb{R}$ where α and β are nonzero.

We express the objective in terms of the cross-covariance and covariance matrices:

$$C_{XY} := \boldsymbol{E}\left[(X - \boldsymbol{E}[X])(Y - \boldsymbol{E}[Y])^\top\right] \qquad C_X := \boldsymbol{E}\left[(X - \boldsymbol{E}[X])(X - \boldsymbol{E}[X])^\top\right]$$
$$C_Y := \boldsymbol{E}\left[(Y - \boldsymbol{E}[Y])(Y - \boldsymbol{E}[Y])^\top\right]$$

Since $\mathrm{Cor}(a^\top X, b^\top Y) = a^\top C_{XY} b / \sqrt{(a^\top C_X a)(b^\top C_Y b)}$, we write:

$$\begin{aligned}(a_i, b_i) = \underset{(a,b)\in\mathbb{R}^d\times\mathbb{R}^{d'}}{\arg\max} \quad & a^\top C_{XY} b \\ \text{subject to} \quad & a^\top C_X a = b^\top C_Y b = 1, \text{ and} \\ & a^\top C_X a_j = b^\top C_Y b_j = 0 \text{ for all } j < i\end{aligned} \tag{2.11}$$

We now consider a change of basis $c = C_X^{1/2} a$ and $d = C_Y^{1/2} b$. Assuming that C_X and C_Y are non-singular, we plug in $a = C_X^{-1/2} c$ and $b = C_Y^{-1/2} d$ above to obtain the auxiliary problem:

$$\begin{aligned}(c_i, d_i) = \underset{(c,d)\in\mathbb{R}^d\times\mathbb{R}^{d'}}{\arg\max} \quad & c^\top C_X^{-1/2} C_{XY} C_Y^{-1/2} d \\ \text{subject to} \quad & c^\top c = d^\top d = 1, \text{ and} \\ & c^\top c_j = d^\top d_j = 0 \text{ for all } j < i\end{aligned} \tag{2.12}$$

whereupon the original solution is given by $a_i = C_X^{-1/2} c_i$ and $b_i = C_Y^{-1/2} d_i$.

A solution of (2.12) is given by $c_i = u_i$ and $d_i = v_i$ where u_i and v_i are the left and right singular vectors of

$$\Omega := C_X^{-1/2} C_{XY} C_Y^{-1/2} \in \mathbb{R}^{d\times d'}$$

corresponding to the i-th largest singular value σ_i. The singular value σ_i is the value of the maximized correlation since $\mathrm{Cor}(a_i^\top X, b_i^\top Y) = a_i^\top C_{XY} b_i = u_i^\top \Omega v_i = \sigma_i$, thus it is bounded as $0 \le \sigma_i \le 1$.

CCA can also be framed as inducing new coordinate systems for the input variables $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$, called the **CCA coordinate systems**, in which they have special covariance structures.

Proposition 2.5.1. *Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$ be random variables with invertible covariance matrices. Let $U\Sigma V^\top$ denote an SVD of $\Omega := C_X^{-1/2} C_{XY} C_Y^{-1/2}$ and let $A := C_X^{-1/2} U$ and $B := C_Y^{-1/2} V$. If $X_{CCA} := A^\top (X - \boldsymbol{E}[X])$ and $Y_{CCA} := B^\top (Y - \boldsymbol{E}[Y])$,*

- *The covariance matrix of X_{CCA} is $I_{d \times d}$.*
- *The covariance matrix of Y_{CCA} is $I_{d' \times d'}$.*
- *The cross-covariance matrix of X_{CCA} and Y_{CCA} is Σ.*

Proof. For the first claim,

$$\boldsymbol{E}[X_{\text{CCA}} X_{\text{CCA}}^\top] = A^\top \boldsymbol{E}[(X - \boldsymbol{E}[X])(X - \boldsymbol{E}[X])^\top] A = U^\top C_X^{-1/2} C_X C_X^{-1/2} U = U^\top U = I_{d \times d}$$

The second claim follows similarly. For the third claim,

$$\boldsymbol{E}[X_{\text{CCA}} Y_{\text{CCA}}^\top] = A^\top \boldsymbol{E}[(X - \boldsymbol{E}[X])(Y - \boldsymbol{E}[Y])^\top] B = U^\top \Omega V = \Sigma$$

□

That is, in the CCA coordinates, the dimensions $i = 1 \ldots \min\{d, d'\}$ of each variable are sorted (in descending order) by the strength of correlation with the corresponding dimensions of the other variable.

2.5.1 Dimensionality Reduction with CCA

A significant part of the recent advance in spectral methods is due to the pioneering theoretical work by Kakade and Foster [2007] and Foster *et al.* [2008] that provides insights into how CCA can be used for dimensionality reduction in certain scenarios. We give a simplified version of these results.

The theory is based on multi-view learning for linear regression. Let $X^{(1)}, X^{(2)} \in \mathbb{R}^d$ be random variables with invertible covariance matrices representing two distinct "views" of another variable (to be specified below). For simplicity, we assume that the two views have the same dimension, but this can be easily relaxed.

CCA coordinate convention Without loss of generality, we assume that $X^{(1)}, X^{(2)}$ are already put in the coordinate systems induced by CCA between $X^{(1)}$ and $X^{(1)}$ (Proposition 2.5.1). Thus they have zero means, identity covariance matrices, and a diagonal cross-covariance matrix $\Sigma = \text{diag}(\sigma_1 \dots \sigma_d)$ where $\sigma_i := \text{Cor}\left(X_i^{(1)}, X_i^{(2)}\right)$ is the i-th maximized correlation. This convention significantly simplifies the notations below. In particular, note that for each $v \in \{1, 2\}$, the top $m \leq d$ most correlated dimensions of $X^{(v)}$ are simply its first m entries which we denote by

$$\underline{X}^{(v)} := (X_1^{(v)} \dots X_m^{(v)})$$

This choice of an m-dimensional representation of the original variable leads to desirable properties under certain assumptions about the relation between $X^{(1)}$ and $X^{(2)}$ with respect to the variable being predicted.

2.5.1.1 Assumption 1: Shared Latent Variable

The **shared latent variable assumption** is that there is some latent variable $H \in \mathbb{R}^m$ where $m \leq d$ such that $X^{(1)}, X^{(2)} \in \mathbb{R}^d$ are (i) conditionally independent given H, and (ii) linear in H in expectation as follows: there exist full-rank matrices $A^{(1)}, A^{(2)} \in \mathbb{R}^{d \times m}$ such that

$$\boldsymbol{E}\left[X^{(1)}|H\right] = A^{(1)}H \qquad\qquad \boldsymbol{E}\left[X^{(2)}|H\right] = A^{(2)}H \tag{2.13}$$

We assume that $\boldsymbol{E}\left[HH^\top\right] = I_{m \times m}$ without loss of generality, since we can always whiten H and the linearity assumption is preserved.

Theorem 2.5.1 (Theorem 3, Foster *et al.* [2008])**.** *We make the shared latent variable assumption defined above. Let $A^{(*|v)} \in \mathbb{R}^{d \times m}$ denote the best linear predictor of $H \in \mathbb{R}^m$ with $X^{(v)} \in \mathbb{R}^d$:*

$$A^{(*|v)} := \underset{A \in \mathbb{R}^{d \times m}}{\arg\min}\ \boldsymbol{E}\left[\left|\left|A^\top X^{(v)} - H\right|\right|_2\right] = \boldsymbol{E}\left[X^{(v)}H^\top\right]$$

Let $\underline{A}^{(v)} \in \mathbb{R}^{m \times m}$ denote the best linear predictor of $H \in \mathbb{R}^m$ with $\underline{X}^{(v)} \in \mathbb{R}^m$:

$$\underline{A}^{(v)} := \underset{A \in \mathbb{R}^{m \times m}}{\arg\min}\ \boldsymbol{E}\left[\left|\left|A^\top \underline{X}^{(v)} - H\right|\right|_2\right] = \boldsymbol{E}\left[\underline{X}^{(v)}H^\top\right]$$

Then the optimal predictor $\underline{A}^{(v)}$ based on the top m most correlated dimensions is precisely as good as the optimal predictor $A^{(*|v)}$ based on all dimensions:

$$\left(A^{(*|v)}\right)^\top X^{(v)} = \left(\underline{A}^{(v)}\right)^\top \underline{X}^{(v)}$$

Proof. With the conditional independence of $X^{(1)}, X^{(2)}$ and the linear relation (2.13),

$$\begin{aligned}
\Sigma = \boldsymbol{E}\left[X^{(1)}\left(X^{(2)}\right)^\top\right] &= \boldsymbol{E}\left[\boldsymbol{E}\left[X^{(1)}\left(X^{(2)}\right)^\top \middle| H\right]\right] \\
&= \boldsymbol{E}\left[\boldsymbol{E}\left[X^{(1)}\middle|H\right]\boldsymbol{E}\left[X^{(2)}\middle|H\right]^\top\right] \\
&= A^{(v)}\,\boldsymbol{E}\left[HH^\top\right]\left(A^{(v)}\right)^\top \\
&= A^{(v)}\left(A^{(v)}\right)^\top
\end{aligned}$$

Thus $\Sigma \in \mathbb{R}^{d\times d}$ has rank m, implying that $\sigma_{m+1} = \cdots = \sigma_d = 0$. Let $\underline{\Sigma} \in \mathbb{R}^{m\times m}$ denote the $(m \times m)$ upper left block of Σ. Next, observe that the best predictor $A^{(*|v)}$ of H based on $X^{(v)}$ is in fact $A^{(v)}$:

$$A^{(v)} = \underset{A\in\mathbb{R}^{d\times m}}{\arg\min}\ \boldsymbol{E}\left[\left|\left|AH - X^{(v)}\right|\right|_2^2\right] = \boldsymbol{E}\left[X^{(v)}H^\top\right] = A^{(*|v)}$$

Together, we have

$$\left(A^{(*|v)}\right)^\top X^{(v)} = \left(A^{(v)}\right)^\top X^{(v)} = \left(A^{(v)}\right)^+ \Sigma X^{(v)} = \left(\underline{A}^{(v)}\right)^+ \underline{\Sigma} X^{(v)} = \left(\underline{A}^{(v)}\right)^\top \underline{X}^{(v)}$$

where we used the fact that $\left(A^{(v)}\right)^\top = \left(A^{(v)}\right)^+ \Sigma$ and that $\underline{A}^{(v)} = \boldsymbol{E}\left[\underline{X}^{(v)}H^\top\right]$ is the first m rows of $A^{(v)} = \boldsymbol{E}\left[X^{(v)}H^\top\right]$. □

2.5.1.2 Assumption 2: Redundancy of the Views

Let $Y \in \mathbb{R}$ and D denote the joint distribution over $(X^{(1)}, X^{(2)}, Y)$ (expectations are with respect to D unless otherwise noted). The **redundancy assumption** is that each individual view $X^{(v)}$ is nearly as (linearly) predictive of the response variable Y as the union of them $X := (X^{(1)}, X^{(2)})$. More precisely, if we denote the best possible predictor $\beta^* \in \mathbb{R}^{2d}$ of Y with both views $X \in \mathbb{R}^{2d}$ by

$$\beta^* := \underset{\beta\in\mathbb{R}^{2d}}{\arg\min}\ \boldsymbol{E}\left[(\beta\cdot X - Y)^2\right]$$

and denote the best possible predictor $\beta^{(v)} \in \mathbb{R}^d$ of Y with only view $X^{(v)} \in \mathbb{R}^d$ by

$$\beta^{(v)} := \underset{\beta \in \mathbb{R}^d}{\arg\min}\ \boldsymbol{E}\left[(\beta \cdot X^{(v)} - Y)^2\right] = \boldsymbol{E}\left[X^{(v)}Y\right] \qquad \forall v \in \{1,2\}$$

then the ϵ-redundancy assumption is that for some ϵ,

$$\boldsymbol{E}\left[\left(\beta^{(v)} \cdot X^{(v)} - Y\right)^2\right] - \boldsymbol{E}\left[(\beta^* \cdot X - Y)^2\right] \leq \epsilon \qquad \forall v \in \{1,2\} \tag{2.14}$$

Lemma 2.5.2 (Lemma 2, Kakade and Foster [2007])**.** *Under the ϵ-redundancy assumption, the optimal predictor $\beta^{(v)}$ of Y with view $X^{(v)} \in \mathbb{R}^d$ cannot have large weights corresponding to weakly correlated dimensions,*

$$\sum_{i=1}^{d} (1-\sigma_i)\left(\beta_i^{(v)}\right)^2 \leq 4\epsilon \tag{2.15}$$

for each view $v \in \{1,2\}$. Note that the bound is independent of d.

Proof. It follows from (2.14) that $\boldsymbol{E}\left[\left(\beta^{(1)} \cdot X^{(1)} - \beta^{(2)} \cdot X^{(2)}\right)^2\right] \leq 4\epsilon$ (Lemma 1, Kakade and Foster [2007]). Furthermore, since $X^{(v)}$ is in the CCA coordinate system,

$$\begin{aligned}
\boldsymbol{E}\left[\left(\beta^{(1)} \cdot X^{(1)} - \beta^{(2)} \cdot X^{(2)}\right)^2\right] &= \sum_{i=1}^{d} \left(\beta_i^{(1)}\right)^2 + \left(\beta_i^{(2)}\right)^2 - 2\sigma_i \beta_i^{(1)} \beta_i^{(2)} \\
&= \sum_{i=1}^{d} (1-\sigma_i)\left(\beta_i^{(1)}\right)^2 + (1-\sigma_i)\left(\beta_i^{(2)}\right)^2 + \sigma_i\left(\beta_i^{(1)} - \beta_i^{(2)}\right)^2 \\
&\geq \sum_{i=1}^{d} (1-\sigma_i)\left(\beta_i^{(v)}\right)^2 \qquad \forall v \in \{1,2\}
\end{aligned}$$

Together, the stated bound is implied. □

Lemma 2.5.2 motivates discarding weakly correlated dimensions. Let

$$m = \left|\left\{i \in [d] :\ \mathrm{Cor}\left(X_i^{(1)}, X_i^{(2)}\right) \geq 1 - \sqrt{\epsilon}\right\}\right|$$

be the number of $(1-\sqrt{\epsilon})$-**strongly correlated dimensions**. Define a thresholded estimator $\beta^{(v)}_{\text{threshold}} \in \mathbb{R}^d$ by

$$[\beta^{(v)}_{\text{threshold}}]_i := \begin{cases} \boldsymbol{E}[X_i^{(v)}Y] & \text{if } \mathrm{Cor}\left(X_i^{(1)}, X_i^{(2)}\right) \geq 1 - \sqrt{\epsilon} \\ 0 & \text{otherwise} \end{cases} \tag{2.16}$$

which can be thought of as a *biased* estimator of $\beta^{(v)}$. Note that $\beta^{(v)}_{\text{threshold}} \cdot X^{(v)} = \underline{\beta}^{(v)} \cdot \underline{X}^{(v)}$, where $\underline{\beta}^{(v)}$ denotes the optimal linear predictor of Y with $\underline{X}^{(v)} \in \mathbb{R}^m$:

$$\underline{\beta}^{(v)} := \arg\min_{\beta \in \mathbb{R}^d} \boldsymbol{E}\left[(\beta \cdot \underline{X}^{(v)} - Y)^2\right] = \boldsymbol{E}\left[\underline{X}^{(v)} Y\right] \qquad \forall v \in \{1, 2\}$$

Sample estimates We assume a set of n samples of $(X^{(1)}, X^{(2)}, Y)$ drawn iid from the distribution D,

$$T := \{(x_1^{(1)}, x_1^{(2)}, y_1) \dots (x_n^{(1)}, x_n^{(2)}, y_n)\}$$

We use the superscript $\wedge$ to denote empirical estimates. For instance, $\hat{\beta}^{(v)} \in \mathbb{R}^d$ is defined as $\hat{\beta}^{(v)} := \frac{1}{n} \sum_{l=1}^n x_l^{(v)} y_l$, and $\underline{\hat{\beta}}^{(v)} \in \mathbb{R}^m$ is defined as $\underline{\hat{\beta}}^{(v)} := \frac{1}{n} \sum_{l=1}^n [x_l^{(v)}]_i y_l$ for $i \in [m]$. Note that the sample estimates are with respect to a fixed T. We use $\boldsymbol{E}_T[\cdot]$ to denote the expected value with respect to T.

Theorem 2.5.3 (Theorem 2, Kakade and Foster [2007])**.** *We make the ϵ-redundancy assumption defined above. Assuming $\boldsymbol{E}[Y^2|X] \leq 1$, the empirical estimate of $\underline{\beta}^{(v)} \in \mathbb{R}^m$ incurs the regret (in expectation)*

$$\boldsymbol{E}_T\left[\text{regret}^*_T\left(\underline{\hat{\beta}}^{(v)}\right)\right] \leq \sqrt{\epsilon}(\sqrt{\epsilon} + 4) + \frac{m}{n}$$

where regret^*_T *is relative to the best possible predictor $\beta^* \in \mathbb{R}^{2d}$ using both views $X \in \mathbb{R}^{2d}$:*

$$\text{regret}^*_T\left(\underline{\hat{\beta}}^{(v)}\right) := \boldsymbol{E}\left[\left(\underline{\hat{\beta}}^{(v)} \cdot \underline{X}^{(v)} - Y\right)^2\right] - \boldsymbol{E}\left[(\beta^* \cdot X - Y)^2\right]$$

A remarkable aspect of this result is that as the number of samples n increases, the empirical estimate of the biased estimator $\beta^{(v)}_{\text{threshold}}$ converges[2] to the optimal estimator β^* with no dependence on the original dimension d; it only depends on the number of $(1-\sqrt{\epsilon})$-strongly correlated dimensions m. Thus if $m \ll d$, then we need much fewer samples to estimate the biased estimator $\beta^{(v)}_{\text{threshold}}$ than to estimate the unbiased estimators $\beta^{(v)}$ or β^* (in which case the regret depends on d) to achieve (nearly) optimal regret.

[2]The suboptimality $\sqrt{\epsilon}(\sqrt{\epsilon} + 4)$ is due to bias and (2.14).

Proof of Theorem 2.5.3. By (2.14), it is sufficient to show that

$$\boldsymbol{E}_T\left[\text{regret}_T\left(\hat{\underline{\beta}}^{(v)}\right)\right] \leq 4\sqrt{\epsilon} + \frac{m}{n}$$

where regret_T is relative to the best possible predictor $\beta^{(v)} \in \mathbb{R}^d$ using view $X^{(v)} \in \mathbb{R}^d$:

$$\text{regret}_T\left(\hat{\underline{\beta}}^{(v)}\right) := \boldsymbol{E}\left[\left(\hat{\underline{\beta}}^{(v)} \cdot \underline{X}^{(v)} - Y\right)^2\right] - \boldsymbol{E}\left[\left(\beta^{(v)} \cdot X^{(v)} - Y\right)^2\right]$$

The regret takes a particularly simple form because of linearity and the choice of coordinates. Given a fixed set of samples T (so that $\hat{\underline{\beta}}^{(v)}$ is not random),

$$\begin{aligned}
\text{regret}_T\left(\hat{\underline{\beta}}^{(v)}\right) &:= \boldsymbol{E}\left[\left(\hat{\underline{\beta}}^{(v)} \cdot \underline{X}^{(v)} - Y\right)^2\right] - \boldsymbol{E}\left[\left(\beta^{(v)} \cdot X^{(v)} - Y\right)^2\right] \\
&= \hat{\underline{\beta}}^{(v)} \cdot \hat{\underline{\beta}}^{(v)} - \beta^{(v)} \cdot \beta^{(v)} - 2\hat{\underline{\beta}}^{(v)} \cdot \boldsymbol{E}[\underline{X}^{(v)}Y] + 2\beta^{(v)} \cdot \boldsymbol{E}[X^{(v)}Y] \\
&= \left|\left|\hat{\beta}^{(v)}_{\text{threshold}}\right|\right|_2^2 - 2\hat{\beta}^{(v)}_{\text{threshold}} \cdot \beta^{(v)} + \left|\left|\beta^{(v)}\right|\right|_2^2 \\
&= \left|\left|\hat{\beta}^{(v)}_{\text{threshold}} - \beta^{(v)}\right|\right|_2^2
\end{aligned}$$

This allows for a *bias-variance decomposition* of the expected regret:

$$\begin{aligned}
\boldsymbol{E}_T\left[\text{regret}_T\left(\hat{\underline{\beta}}^{(v)}\right)\right] &= \boldsymbol{E}_T\left[\left|\left|\hat{\beta}^{(v)}_{\text{threshold}} - \beta^{(v)}\right|\right|_2^2\right] \\
&= \left|\left|\beta^{(v)}_{\text{threshold}} - \beta^{(v)}\right|\right|_2^2 + \boldsymbol{E}_T\left[\left|\left|\hat{\beta}^{(v)}_{\text{threshold}} - \beta^{(v)}\right|\right|_2^2\right] \\
&= \left|\left|\beta^{(v)}_{\text{threshold}} - \beta^{(v)}\right|\right|_2^2 + \sum_{i=1}^{m} \text{Var}\left(\hat{\underline{\beta}}^{(v)}_i\right)
\end{aligned}$$

The first term corresponds to the bias of the estimator, and the second term is the amount of variance with respect to T.

To bound the variance term, note that:

$$\begin{aligned}
\text{Var}\left(\hat{\underline{\beta}}^{(v)}_i\right) &= \frac{1}{n}\text{Var}\left(\underline{X}^{(v)}_i Y\right) \leq \frac{1}{n}\boldsymbol{E}\left[\left(\underline{X}^{(v)}_i Y\right)^2\right] \\
&= \frac{1}{n}\boldsymbol{E}\left[\left(\underline{X}^{(v)}_i\right)^2 \boldsymbol{E}\left[Y^2|X\right]\right] \leq \frac{1}{n}\boldsymbol{E}\left[\left(X^{(v)}_i\right)^2\right] = \frac{1}{n}
\end{aligned}$$

where the second variance is with respect to D. We used the assumption that $\boldsymbol{E}[Y^2|X] \leq 1$. So the variance term is bounded by m/n.

To bound the bias term, it is crucial to exploit the multi-view assumption (2.14). For all $i > m$ we have $\sigma_i < 1 - \sqrt{\epsilon}$ and thus $1 \leq (1 - \sigma_i)/\sqrt{\epsilon}$, so

$$\left|\left|\beta^{(v)}_{\text{threshold}} - \beta^{(v)}\right|\right|_2^2 = \sum_{i>m}\left(\beta^{(v)}_i\right)^2 \leq \sum_{i=1}^{d}\left(\beta^{(v)}_i\right)^2 \leq \sum_{i=1}^{d}\frac{1-\sigma_i}{\sqrt{\epsilon}}\left(\beta^{(v)}_i\right)^2 \leq 4\sqrt{\epsilon}$$

where the last step is by (2.15) and makes the bias term independent of d.

□

Connection to semi-supervised learning The theory suggest a natural way to utilize unlabeled data with CCA to augment supervised training. In a semi-supervised scenario, we assume that the amount of labeled samples is limited: $(x_1^{(1)}, y_1) \dots (x_n^{(1)}, y_n)$ samples of $(X^{(1)}, Y)$ for some small n. But if there is a second view $X^{(2)}$ as predictive of Y as $X^{(1)}$ (i.e., the redundancy assumption) for which it is easy to obtain a large amount of unlabeled samples of $(X^{(1)}, X^{(2)})$,

$$(x_1^{(1)}, x_1^{(2)}) \dots (x_{n'}^{(1)}, x_{n'}^{(2)}) \qquad\qquad n' \gg n$$

then we can leverage these unlabeled samples to accurately estimate CCA projection vectors. These projection vectors are used to eliminate the dimensions of the labeled samples $x_1^{(1)} \dots x_n^{(1)}$ that are not strongly correlated with the other view's. Theorem 2.5.3 implies that the supervised model trained on these low dimensional samples (corresponding to the thresholded estimator) converges to the optimal model at a faster rate.

2.6 Spectral Clustering

Spectral clustering refers to partitioning vertices in an undirected graph by matrix decomposition [Donath and Hoffman, 1973; Fiedler, 1973]. Here, we give one example framed as finding vertex representations suitable for the clustering problem [Shi and Malik, 2000]; this approach is closely relevant to our word clustering method in Chapter 5. For other examples of spectral clustering, see Von Luxburg [2007].

Given an undirected weighted graph described in Example 1.2.3, we wish to find a partition $\mathcal{P} = \{A_1 \dots A_m\}$ of vertices $[n]$ where $m \leq n$. One sensible formulation is to minimize the "flow" $W(A, \overline{A}) := \sum_{i \in A,\, j \in \overline{A}} w_{ij}$ between each cluster $A \in \mathcal{P}$ and its complement $\overline{A} := [n] \backslash A$ to encourage cluster independence, while normalizing by the "volume" $\text{vol}(A) := \sum_{i \in A} d_i$ to discourage an imbalanced partition. This gives the following objective:

$$\mathcal{P}^* = \arg\min_{\mathcal{P}} \sum_{A \in \mathcal{P}} \frac{W(A, \overline{A})}{\text{vol}(A)} \tag{2.17}$$

This problem is NP-hard [Wagner and Wagner, 1993], but there is a spectral method for solving a relaxed version of this problem.

In this method, vertex i is represented as the i-th row of a matrix $X_{\mathcal{P}} \in \mathbb{R}^{n \times m}$ where:

$$[X_{\mathcal{P}}]_{i,c} = \begin{cases} \frac{1}{\sqrt{\text{vol}(A_c)}} & \text{if } i \in A_c \\ 0 & \text{otherwise} \end{cases} \tag{2.18}$$

with respect to a specific partition $\mathcal{P} = \{A_1 \dots A_m\}$ of $[n]$. Note that $X_{\mathcal{P}}^\top D X_{\mathcal{P}} = I_{m \times m}$ (for all $\mathcal{P}$) by design. We invoke the following fact:

$$\sum_{A \in \mathcal{P}} \frac{W(A, \overline{A})}{\text{vol}(A)} = \text{Tr}(X_{\mathcal{P}}^\top L X_{\mathcal{P}})$$

where L denotes the unnormalized graph Laplacian $L := W - D$. This holds by properties of L and the definition of $X_{\mathcal{P}}$; see Von Luxburg [2007] for a proof. Use this fact to rewrite the clustering objective as

$$\begin{aligned} X^* = \underset{X_{\mathcal{P}} \in \mathbb{R}^{n \times m}}{\arg\min} \ & \text{Tr}(X_{\mathcal{P}}^\top L X_{\mathcal{P}}) \\ & X_{\mathcal{P}} \text{ has the form in (2.18) for some } \mathcal{P} \end{aligned} \tag{2.19}$$

whereupon the optimal clusters can be recovered as: $i \in A_c$ iff $X^*_{i,c} > 0$. We obtain a relaxation of (2.19) by weakening the explicit form constraint as:

$$\begin{aligned} \widetilde{X} = \underset{X \in \mathbb{R}^{n \times m}}{\arg\min} \ & \text{Tr}(X^\top L X) \\ & \text{subject to } X^\top D X = I_{m \times m} \end{aligned} \tag{2.20}$$

Using a change of basis $U = D^{1/2} X$ and plugging in $X = D^{-1/2} U$ above, we can solve

$$\begin{aligned} \widetilde{U} = \underset{U \in \mathbb{R}^{n \times m}}{\arg\min} \ & \text{Tr}(U^\top D^{-1/2} L D^{-1/2} U) \\ & \text{subject to } U^\top U = I_{m \times m} \end{aligned} \tag{2.21}$$

and let $\widetilde{X} = D^{-1/2} \widetilde{U}$. It can be verified that the solution of (2.21) is given by the orthonormal eigenvectors of $D^{-1/2} L D^{-1/2}$ (called the normalized graph Laplacian) corresponding to the m smallest eigenvalues $0 \leq \lambda_1 \leq \dots \leq \lambda_m$. More directly, the solution of (2.20) is given

by the eigenvectors of $D^{-1}L$ corresponding to the same eigenvalues $\Lambda := \text{diag}(\lambda_1, \ldots, \lambda_m)$ since

$$(D^{-1/2}LD^{-1/2})\widetilde{U} = \Lambda\widetilde{U} \qquad \Longleftrightarrow \qquad D^{-1}L\widetilde{X} = \Lambda\widetilde{X}$$

This gives the clustering algorithm of Shi and Malik [2000]:

1. Construct the normalized graph Laplacian $\overline{L} = D^{-1}L$.
2. (Rank-m eigendecomposition) Compute the eigenvectors of $\overline{L}$ corresponding to the smallest m eigenvalues as columns of matrix $\widetilde{X} \in \mathbb{R}^{n\times m}$.
3. Cluster the rows of $\widetilde{X}$ into m groups $A_1 \ldots A_m$ (e.g., with k-means).

In summary, the method approximates the idealized vertex representations X^* in (2.19) (which, if given, can be used to trivially recover the optimal clusters) with a surrogate representation $\widetilde{X}$ that is efficiently computable with an eigendecomposition of the normalized graph Laplacian. While this approximation can be arbitrarily suboptimal in the worst case [Guattery and Miller, 1998], it is effective in practice [Shi and Malik, 2000; Ng *et al.*, 2002].

2.7 Subspace Identification

Spectral methods have recently garnered much interest as a promising approach to learning latent-variable models. A pioneering work in this direction is the spectral algorithm of Hsu *et al.* [2008] for estimating distributions under HMMs. The Hsu *et al.* method is an amalgam of many ideas; see the paper for a detailed discussion. A crucial component of the method is the use of SVD to identify a low-dimensional subspace associated with the model. We give a brief, informal review of their algorithm and its extension by Foster *et al.* [2012] from an angle of subspace identification.

Consider an HMM with m hidden states $h \in [m]$ and n observation states $x \in [n]$ where $m \ll n$. This HMM can be parametrized as a matrix-vector tuple (T, O, π) where

$$\begin{aligned}
T \in \mathbb{R}^{m\times m}: &\qquad T_{h',h} = \text{transition probability from state } h \text{ to } h' \\
O \in \mathbb{R}^{n\times m}: &\qquad O_{x,h} = \text{emission probability from state } h \text{ to observation } x \\
\pi \in \mathbb{R}^{m}: &\qquad \pi_h = \text{prior probability of state } h
\end{aligned}$$

It is well-known (and easily checkable) that with the following definition of "observable operators" (A, a_∞, a_1) [Ito *et al.*, 1992; Jaeger, 2000]

$$\begin{aligned} A(x) &:= T \operatorname{diag}(O_{x,1} \dots O_{x,m}) && \forall x \in [n] \\ a_\infty^\top &:= 1_m^\top \\ a_1 &:= \pi \end{aligned}$$

(1_m is a vector of ones in $\mathbb{R}^m$), the probability of any observation sequence $x_1 \dots x_N \in [n]$ under the HMM is given by a product of these operators:

$$p(x_1 \dots x_N) = a_\infty^\top A(x_N) \cdots A(x_1) a_1 \tag{2.22}$$

That is, (2.22) is the matrix form of the forward algorithm [Rabiner, 1989]. The approach pursued by Hsu *et al.* [2008] and Foster *et al.* [2012] is to instead estimate certain linear transformations of the operators:

$$B(x) := GA(x)G^+ \qquad \forall x \in [n] \tag{2.23}$$

$$b_\infty^\top := a_\infty^\top G^+ \tag{2.24}$$

$$b_1 := Ga_1 \tag{2.25}$$

where G is a matrix such that $G^+G = I_{m \times m}$. It is clear that the forward algorithm can be computed by (B, b_∞, b_1), since

$$\begin{aligned} p(x_1 \dots x_N) &= a_\infty^\top A(x_N) \cdots A(x_1) a_1 \\ &= a_\infty^\top G^+ GA(x_N) G^+ G \cdots G^+ GA(x_1) G^+ G a_1 \\ &= b_\infty^\top B(x_N) \cdots B(x_1) b_1 \end{aligned}$$

Let $X_1, X_2 \in [n]$ be random variables corresponding to the first two observations under the HMM (where we assume the usual generative story). A central quantity considered by Hsu *et al.* [2008] is a matrix of bigram probabilities $P_{2,1} \in \mathbb{R}^{n \times n}$ defined as

$$[P_{2,1}]_{x',x} := P(X_1 = x, X_2 = x') \qquad \forall x, x' \in [n] \tag{2.26}$$

The matrix relates the past observation (X_1) to the future observation (X_2). It can be shown that this matrix can be expressed in terms of the HMM parameters as (Lemma 3,

Hsu *et al.* [2008]):

$$P_{2,1} = OT\,\mathrm{diag}(\pi)O^\top \tag{2.27}$$

It follows that $\mathrm{rank}(P_{2,1}) = m$ if $O, T, \mathrm{diag}(\pi)$ have full-rank—even though the dimension of the matrix is $n \times n$.

Hsu *et al.* [2008] apply SVD on $P_{2,1}$ to identify the m-dimensional subspace spanned by the conditional emission distributions: $(O_{1,h}, \ldots, O_{n,h})$ for all $h \in [m]$. Specifically, if $P_{2,1} = U\Sigma V^\top$ is a rank-m SVD, then it can be shown that (Lemma 2, Hsu *et al.* [2008])

$$\mathrm{range}(U) = \mathrm{range}(O) \tag{2.28}$$

This projection matrix $U \in \mathbb{R}^{n\times m}$ is then used to reduce the dimension of observations from $\mathbb{R}^n$ to $\mathbb{R}^m$, whereupon the linearly transformed operators (2.23–2.25) are recovered by the method of moments. Importantly, the spectral dimensionality reduction leads to polynomial sample complexity (Theorem 6, Hsu *et al.* [2008]; Theorem 1, Foster *et al.* [2012]).

Note that the statement is about the *true* probabilities $P_{2,1}$ under the HMM. In order to establish finite sample complexity bounds, we must consider the empirical estimate $\widehat{P}_{2,1}$ of $P_{2,1}$ where each entry

$$[\widehat{P}_{2,1}]_{x',x} := \frac{1}{N}\sum_{i=1}^{N} \left[\left[X_1 = x, X_2 = x'\right]\right] \qquad\qquad \forall x, x' \in [n]$$

is estimated from a finite number of samples N, and examine how a rank-m SVD $\widehat{U}\widehat{\Sigma}\widehat{V}^\top$ of $\widehat{P}_{2,1}$ behaves with respect to a rank-m SVD $U\Sigma V^\top$ of $P_{2,1}$ as a function of N. Deriving such bounds can be quite involved (see Section 1.4) and is a major technical contribution of Hsu *et al.* [2008].

It should be emphasized that the subspace identification component can be disentangled from the method of moments. In particular, it can be verified that removing U in their definitions of $\vec{b}_1$, $\vec{b}_\infty$, and B_x in Hsu *et al.* [2008] still results in a consistent estimator of the distribution in (2.22).

2.8 Alternating Minimization Using SVD

Ando and Zhang [2005] propose learning a shared structure across multiple related classification tasks over a single domain. Specifically, they consider T binary classification tasks

each of which has its own linear classifier $f_t : \mathbb{R}^d \to \mathbb{R}$ mapping a d-dimensional feature vector $x \in \mathbb{R}^d$ to a classification score

$$f_t(x) := (u_t + \Theta v_t)^\top x \tag{2.29}$$

Here, $u_t \in \mathbb{R}^d$ and $v_t \in \mathbb{R}^m$ are task-specific parameters but $\Theta \in \mathbb{R}^{d \times m}$ is a global parameter shared by all classifiers $f_1 \dots f_T$ (we assume $m \leq \min\{d, T\}$). In particular, if Θ is zero then each classifier is an independent linear function $u_t^\top x$. The predicted label is the sign of the classification score $\text{sign}(f_t(x)) \in \{\pm 1\}$.

The parameter sharing makes the estimation problem challenging, but Ando and Zhang [2005] develop an effective alternating loss minimization algorithm using a variational property of SVD. To illustrate their method, let $L : \mathbb{R} \times \{\pm 1\} \to \mathbb{R}$ be a convex loss function for classification, for instance the hinge loss $L(p, y) = \max(0, 1 - py)$.[3] For each task $t \in [T]$, we are given n_t labeled samples $(x^{(1|t)}, y^{(1|t)}) \dots (x^{(n_t|t)}, y^{(n_t|t)}) \in \mathbb{R}^d \times \{\pm 1\}$. A training objective is given by the following empirical loss minimization:

$$\min_{\substack{u_t \in \mathbb{R}^d \ \forall t \in [T] \\ v_t \in \mathbb{R}^m \ \forall t \in [T] \\ \Theta \in \mathbb{R}^{d \times m}}} \sum_{t=1}^{T} \left(r(u_t, v_t) + \frac{1}{n_t} \sum_{i=1}^{n_t} L\left((u_t + \Theta v_t)^\top x^{(i|t)}, y^{(i|t)} \right) \right) + R(\Theta) \tag{2.30}$$

where $r(u_t, v_t)$ and $R(\Theta)$ are appropriate regularizers for the parameters. In other words, we minimize the sum of average losses (averaging is necessary since the amount of labeled data can vary greatly for each task).

Ando and Zhang [2005] choose a particular version of (2.30) to accomodate the use of SVD, given by

$$\min_{\substack{u_t \in \mathbb{R}^d \ \forall t \in [T] \\ v_t \in \mathbb{R}^m \ \forall t \in [T] \\ \Theta \in \mathbb{R}^{d \times m} : \Theta^\top \Theta = I_{m \times m}}} \sum_{t=1}^{T} \left(\lambda_t \left|\left| u_t \right|\right|_2^2 + \frac{1}{n_t} \sum_{i=1}^{n_t} L\left((u_t + \Theta v_t)^\top x^{(i|t)}, y^{(i|t)} \right) \right) \tag{2.31}$$

Note that the parameters v_t are not regularized: $r(u_t, v_t) = \lambda_t \left|\left| u_t \right|\right|_2^2$ for some hyperparameter $\lambda_t \geq 0$. Also, Θ is constrained to be an orthonormal matrix and is thus implicitly

[3] Ando and Zhang [2005] use a quadratically smoothed hinge loss called modified Huber:

$$L(p, y) = \begin{cases} \max(0, 1 - py)^2 & \text{if } py \geq -1 \\ -4py & \text{otherwise} \end{cases}$$

regularized. With an orthonormal $\Theta \in \mathbb{R}^{d \times m}$, the problem can be interpreted as finding an m-dimensional subspace of $\mathbb{R}^d$ which is predictive of labels across all T tasks. If $x_\Theta := \Theta^\top x$ denotes the m-dimensional representation of $x \in \mathbb{R}^d$ projected in the subspace range(Θ), every f_t computes the classification score of x by using this representation:

$$f_t(x) = u_t^\top x + v_t^\top x_\Theta$$

The objective (2.31) can be re-written with the change of variable $w_t := u_t + \Theta v_t$:

$$\min_{\substack{w_t \in \mathbb{R}^d \ \forall t \in [T] \\ v_t \in \mathbb{R}^m \ \forall t \in [T] \\ \Theta \in \mathbb{R}^{d \times m} \colon \Theta^\top \Theta = I_{m \times m}}} \sum_{t=1}^{T} \left(\lambda_t \left|\left| w_t - \Theta v_t \right|\right|_2^2 + \frac{1}{n_t} \sum_{i=1}^{n_t} L\left(w_t^\top x^{(i|t)}, y^{(i|t)} \right) \right) \tag{2.32}$$

Clearly, the original solution can be recovered from the solution of this formulation by $u_t = w_t - \Theta v_t$. The intuition behind considering (2.32) instead of (2.31) is that this allows us to separate the parameters Θ and v_t from the loss function $L(\cdot, \cdot)$ if we fix w_t.

Theorem 2.8.1 (Ando and Zhang [2005])**.** *Assume the parameters $w_t \in \mathbb{R}^d$ are fixed in (2.32) for all $t \in [T]$. Define $A := [\sqrt{\lambda_1} w_1 \dots \sqrt{\lambda_T} w_T] \in \mathbb{R}^{d \times T}$, and let $U = [u_1 \dots u_m] \in \mathbb{R}^{d \times m}$ be the left singular vectors of A corresponding to the largest $m \leq \min\{d, T\}$ singular values. Then the optimal solution for the parameters $\Theta \in \mathbb{R}^{d \times m}$ (under the orthogonality constraint $\Theta^\top \Theta = I_{m \times m}$) and $v_t \in \mathbb{R}^m$ is given by $\Theta^* = U$ and $v_t^* = U^\top w_t$ for all $t \in [T]$.*

Proof. Since w_t's are fixed, the objective (2.32) becomes

$$\min_{\substack{v_t \in \mathbb{R}^m \ \forall t \in [T] \\ \Theta \in \mathbb{R}^{d \times m} \colon \Theta^\top \Theta = I_{m \times m}}} \sum_{t=1}^{T} \lambda_t \left|\left| w_t - \Theta v_t \right|\right|_2^2$$

Note that for any value of orthonormal Θ, the optimal solution for each v_t is given by regression $(\Theta^\top \Theta)^{-1} \Theta^\top w_t = \Theta^\top w_t$. Thus we can plug in $v_t = \Theta^\top w_t$ in the objective to remove dependence on all variables except for Θ,

$$\min_{\Theta \in \mathbb{R}^{d \times m} \colon \Theta^\top \Theta = I_{m \times m}} \sum_{t=1}^{T} \lambda_t \left|\left| w_t - \Theta \Theta^\top w_t \right|\right|_2^2$$

Since $\left|\left| w_t - \Theta \Theta^\top w_t \right|\right|_2^2 = ||w_t||_2^2 - w_t^\top \Theta \Theta^\top w_t$, the objective is equivalent to

$$\max_{\Theta \in \mathbb{R}^{d \times m} \colon \Theta^\top \Theta = I_{m \times m}} \left|\left| A^\top \Theta \right|\right|_F^2$$

Thus the columns of an optimal Θ^* are given by the left singular vectors of A corresponding to the largest m singular values (Theorem 1.3.4). This also gives the claim on v_t^*. □

The theorem yields an alternating minimization strategy for optimizing (2.32). That is, iterate the following two steps until convergence:

- Fix Θ and v_t's: optimize the convex objective (2.32) (convex in w_t).
- Fix w_t's: compute optimal values of Θ and v_t in (2.32) with SVD (Theorem 2.8.1).

Note, however, that in general this does not guarantee the global optimality of the output parameters w_t, v_t, and Θ.

2.9 Non-Negative Matrix Factorization

Non-negative matrix factorization (NMF) is the following problem: given a non-negative matrix $A \in \mathbb{R}^{n \times d}$ (i.e., $A_{i,j} \geq 0$ for all i and j), and also a rank value $m \leq \min\{n, d\}$, find non-negative matrices $B \in \mathbb{R}^{n \times m}$ and $C \in \mathbb{R}^{m \times d}$ such that $A = BC$ (the existence of such B and C is often given by task-specific assumptions). If M_i denotes the i-th row of matrix M, it can be easily verified that

$$A_i = \sum_{j=1}^{m} B_{i,j} \times C_j \tag{2.33}$$

In other words, a row of A is a (non-negative) linear combination of the rows of C. Thus NMF can be seen as finding a set of "dictionary" rows $C_1 \ldots C_m$ that can be non-negatively added to realize all n rows of A. NMF arises naturally in many applications.

Example 2.9.1 (Image analysis [Lee and Seung, 1999])**.** *Suppose that each row of $A \in \mathbb{R}^{n \times d}$ is a facial image represented as a vector of d non-negative pixel values. Let $B \in \mathbb{R}^{n \times m}$ and $C \in \mathbb{R}^{m \times d}$ be non-negative matrices such that $A = BC$. Then each facial image $A_i = B_{i,1}C_1 + \cdots + B_{i,m}C_m$ is a non-negative linear combination of m "basis images" $C_1 \ldots C_m$.*

Example 2.9.2 (Document analysis [Blei *et al.*, 2003; Arora *et al.*, 2012b])**.** *Suppose that each row of $A \in \mathbb{R}^{n \times d}$ is a document represented as the document's distribution over d word*

types (thus non-negative). Let $B \in \mathbb{R}^{n \times m}$ *and* $C \in \mathbb{R}^{m \times d}$ *be non-negative matrices such that* $A = BC$ *and additionally that each row of* B *sums to* 1*. Then the word distribution under the* i*-th document* $A_i = B_{i,1}C_1 + \cdots + B_{i,m}C_m$ *is a convex combination of the word distributions under* m *"topics"* $C_1 \ldots C_m$*.*

Note that while NMF is matrix decomposition, it is somewhat divorced from the theory of eigendecompositon. NMF is often implicit in parameter estimation of probabilistic models; for instance, learning the parameters of latent Dirichlet allocation can be seen as an implicit NMF [Arora *et al.*, 2012b].

Donoho and Stodden [2003] provide an intuitive geometric interpretation of NMF which also leads to an understanding of when an NMF is unique. Since all values involved in the characterization of A's row

$$A_i = \sum_{j=1}^{m} B_{i,j} \times C_j$$

are non-negative, we have that

- A_i is a vector residing in the positive orthant of $\mathbb{R}^d$.
- $C_1 \ldots C_m$ are vectors also in the positive orthant of $\mathbb{R}^d$ such that any A_i can be expressed as their combination (scaled by scalars $B_{i,1} \ldots B_{i,m} \geq 0$).

Hence NMF can be viewed as finding a *conical hull* enclosing all $A_1 \ldots A_n$.[4] If $A_1 \ldots A_n$ do not lie on every axis, there are infinitely many conical hulls that enclose $A_1 \ldots A_n$ and hence NMF does not have a unique solution. Using this intuition, Donoho and Stodden [2003] provide a separability condition for when an NMF is unique.

Vavasis [2009] shows that NMF is NP-hard in general, but Arora *et al.* [2012b; 2012a] develop a provable NMF algorithm by exploiting a natural separability condition. In particular, Arora *et al.* [2012a] derive a purely combinatorial method for extracting dictionary rows $C_1 \ldots C_m$ and successfully apply it to learning topic models. In Chapter 7, we extend this framework to learning hidden Markov models.

[4] When each row of B is contrained to sum to 1 (as in Example 2.9.2), then NMF can be viewed as finding a *convex hull* enclosing all $A_1 \ldots A_n$.

2.10 Tensor Decomposition

(We borrow the tensor notation in previous work [Lim, 2006; Anandkumar *et al.*, 2014].)

A p-th order tensor T is a p-dimensional array with entries $T_{i_1 \dots i_p} \in \mathbb{R}$ (e.g., a matrix is a second-order tensor). For simplicity, we only consider $p \leq 3$. A tensor $T \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ defines a function that maps input matrices V_1, V_2, V_3, where $V_i \in \mathbb{R}^{n_i \times m_i}$, to an output tensor $T(V_1, V_2, V_3) \in \mathbb{R}^{m_1 \times m_2 \times m_3}$ as follows:

$$[T(V_1, V_2, V_3)]_{i,j,k} := \sum_{i',j',k'} T_{i',j',k'} [V_1]_{i',i} [V_2]_{j',j} [V_3]_{k',k} \tag{2.34}$$

This nonlinear function is called **multilinear** since it is linear in V_i if all input matrices are fixed except V_i. A tensor $T \in \mathbb{R}^{n \times n \times n}$ is called **supersymmetric** if its entries are invariant to a permutation on indices, that is, $[T]_{i,j,k} = [T]_{i,k,j} = \cdots$. The **rank** of a supersymmetric T is defined to be the smallest non-negative integer m such that $T = \sum_{i=1}^{m} v_i v_i^\top v_i^\top$ for some vectors $v_1 \dots v_m \in \mathbb{R}^n$. Given vectors $\{u, v, w\}$, the notation $uv^\top w^\top$ denotes a rank-1 tensor with entries $[uv^\top w^\top]_{i,j,k} = [u]_i [v]_j [w]_k$ (analogous to the matrix outer product).

The above terms are similarly defined for the first- and second-order tensors (i.e., vectors and matrices). Note that a supersymmetric second-order tensor $M \in \mathbb{R}^{n_1 \times n_2}$ reduces to the standard definition of a symmetric matrix; the rank of M reduces to the number of nonzero eigenvalues (Proposition 1.2.1); and the tensor product (2.34) reduces to the matrix product $M(V_1, V_2) = V_1^\top M V_2$. The notation (2.34) also accommodates bypassing certain input positions with identity matrices. For example, the matrix-vector product can be expressed as $M(I_{n_1 \times n_1}, v) = Mv \in \mathbb{R}^{n_1}$. For a supersymmetric tensor $T \in \mathbb{R}^{n \times n \times n}$, a unit **eigenvector** $v \in \mathbb{R}^n$ of T is a unit-length vector with a corresponding **eigenvalue** $\lambda \in \mathbb{R}$ such that

$$T(I_{n \times n}, v, v) = \lambda v \tag{2.35}$$

which is a direct analogue of the matrix counterpart (1.18).

Tensor decomposition is often useful (e.g., in the method of moments). Unfortunately, many of the tools developed in conventional linear algebra do not generalize to higher-order tensors. For instance, while a symmetric matrix always has an efficiently computable eigendecomposition (Theorem 1.2.6), it is not the case for a higher-order supersymmetric

tensor T [Qi, 2005]. While the low-rank matrix approximation problem can be solved efficiently using SVD (Section 2.2), computing even a rank-1 approximation of T:

$$\min_{u,v,w} \left|\left|T - uv^\top w^\top\right|\right|_F \tag{2.36}$$

(where $||T||_F = \sqrt{\sum_{i,j,k} T_{i,j,k}^2}$) is NP-hard (Theorem 1.13, Hillar and Lim [2013]).

Anandkumar *et al.* [2014] show that the problem is much more manageable if tensors are *orthogonal* in addition to being supersymmetric. Specifically, they assume a supersymmetric and orthogonal tensor $T \in \mathbb{R}^{n \times n \times n}$ of rank-m, that is,

$$T = \sum_{i=1}^{m} \lambda_i v_i v_i^\top v_i^\top \tag{2.37}$$

where $v_1 \ldots v_m \in \mathbb{R}^n$ are orthonormal and $\lambda_1 \geq \ldots \geq \lambda_m > 0$. Since $T(I_{n\times n}, v_i, v_i) = \lambda_i v_i$, each (v_i, λ_i) is an eigenvector-eigenvalue pair. In this case, a random initial vector $v \in \mathbb{R}^n$ under the **tensor power iterations**:

$$v \mapsto \frac{T(I_{n\times n}, v, v)}{||T(I_{n\times n}, v, v)||_2} \tag{2.38}$$

converges to some v_i (Theorem 4.1, Anandkumar *et al.* [2014]). Thus the eigencomponents of T can be extracted through the power iteration method similar to the matrix case in Figure 1.3. Note a subtle difference: the extracted eigencomponents may not be in a descending order of eigenvalues, since the iteration (2.38) converges to *some* eigenvector v_i, not necessarily v_1.

Another important contribution of Anandkumar *et al.* [2014] is a scheme to *orthogonalize* a rank-m supersymmetric tensor $T = \sum_{i=1}^m w_i u_i u_i^\top u_i^\top$ where $w_1 \geq \ldots \geq w_m > 0$ but $u_1 \ldots u_m$ are not necessarily orthogonal (but assumed to be linearly independent) with a corresponding rank-m symmetric matrix $M = \sum_{i=1}^m w_i u_i u_i^\top$. Let $W \in \mathbb{R}^{n\times m}$ be a whitening matrix for M, that is,

$$M(W, W) = \sum_{i=1}^{m} (\sqrt{w_i} W^\top u_i)(\sqrt{w_i} W^\top u_i)^\top = I_{m\times m}$$

For instance, one can set $W = V\Lambda^{-1/2}$ where $M = V\Lambda V^\top$ is a rank-m SVD of M. This implies that $\sqrt{w_1} W^\top u_1 \ldots \sqrt{w_m} W^\top u_m \in \mathbb{R}^m$ are orthonormal. Then the $(m \times m \times m)$

tensor

$$T(W,W,W) = \sum_{i=1}^{m} \frac{1}{\sqrt{w_i}} (\sqrt{w_i} W^\top u_i)(\sqrt{w_i} W^\top u_i)^\top (\sqrt{w_i} W^\top u_i)^\top$$

is orthogonal and is decomposable by the tensor power iteration method $T(W,W,W) = \sum_{i=1}^{m} \lambda_i v_i v_i^\top v_i^\top$. The original variables can be recovered as $w_i = 1/\lambda_i^2$ and $u_i = \lambda_i (W^\top)^+ v_i$.

In summary, the method of Anandkumar *et al.* [2014] can be used to recover linearly independent $u_1 \dots u_m \in \mathbb{R}^n$ and positive scalars $w_1 \dots w_m \in \mathbb{R}$ from supersymmetric second- and third-order tensors of rank m:

$$M = \sum_{i=1}^{m} w_i u_i u_i^\top \tag{2.39}$$

$$T = \sum_{i=1}^{m} w_i u_i u_i^\top u_i^\top \tag{2.40}$$

Anandkumar *et al.* [2014] show that this can be used as a *learning algorithm* for a variety of latent-variable models. For instance, consider learning a bag-of-words model with n word types and m topic types. The task is to estimate the model parameters

- $w_i \in \mathbb{R}$: probability of topic $i \in [m]$
- $u_i \in \mathbb{R}^n$: conditional distribution over n word types given topic $i \in [m]$

Then it is easily verifiable that the observable quantities $M \in \mathbb{R}^{n \times n}$ and $T \in \mathbb{R}^{n \times n \times n}$ where $M_{i,j}$ is the probability of words $i, j \in [n]$ occurring together in a document (not necessarily consecutively) and $T_{i,j,k}$ is the probability of words $i, j, k \in [n]$ occurring together in a document (not necessarily consecutively) have the form (2.39) and (2.40). Thus the parameters (w_i, u_i) can be estimated by tensor decomposition.

We mention that there is ongoing progress in tensor decomposition. For example, see Kuleshov *et al.* [2015] for a decomposition scheme applicable to a wider class of tensors.

Bibliography

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6:1817–1853, 2005.

Sanjeev Arora, Rong Ge, Yoni Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. *arXiv preprint arXiv:1212.4777*, 2012.

Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 1–10. IEEE, 2012.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Alan Kaylor Cline and Inderjit S Dhillon. Computation of the singular value decomposition. *Handbook of linear algebra*, pages 45–1, 2006.

William E Donath and Alan J Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5):420–425, 1973.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a

correct decomposition into parts? In *Advances in neural information processing systems*, page None, 2003.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973.

Dean P Foster, Sham M Kakade, and Tong Zhang. Multi-view dimensionality reduction via canonical correlation analysis. *Toyota Technological Institute, Chicago, Illinois, Tech. Rep. TTI-TR-2008-4*, 2008.

D. P. Foster, J. Rodu, and L.H. Ungar. Spectral dimensionality reduction for hmms. *Arxiv preprint arXiv:1203.6130*, 2012.

Stephen H. Friedberg, Arnold J. Insel, and Lawrence E. Spence. *Linear Algebra*. Pearson Education, Inc., 4 edition, 2003.

Nicolas Gillis and François Glineur. Low-rank matrix approximation with weights or missing data is np-hard. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1149–1165, 2011.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

Stephen Guattery and Gary L Miller. On the quality of spectral separators. *SIAM Journal on Matrix Analysis and Applications*, 19(3):701–719, 1998.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.

Ngoc-Diep Ho. *Nonnegative matrix factorization algorithms and applications*. PhD thesis, ÉCOLE POLYTECHNIQUE, 2008.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *arXiv preprint arXiv:0811.4413*, 2008.

Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Hisashi Ito, S-I Amari, and Kingo Kobayashi. Identifiability of hidden markov information sources and their minimum degrees of freedom. *Information Theory, IEEE Transactions on*, 38(2):324–333, 1992.

Herbert Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000.

Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

Sham M Kakade and Dean P Foster. Multi-view regression via canonical correlation analysis. In *Learning theory*, pages 82–96. Springer, 2007.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

Volodymyr Kuleshov, Arun Tejasvi Chaganty, and Percy Liang. Tensor factorization via matrix factorization. *arXiv preprint arXiv:1501.07320*, 2015.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Lek-Heng Lim. Singular values and eigenvalues of tensors: a variational approach. *arXiv preprint math/0607648*, 2006.

Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1):50–59, 1960.

Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

K Pearson. On lines and planes of closest fit to system of points in space. philiosophical magazine, 2, 559-572, 1901.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the Empiricial Methods in Natural Language Processing*, volume 12, 2014.

Eduard Prugovečki. *Quantum mechanics in Hilbert space*, volume 41. Academic Press, 1971.

Liqun Qi. Eigenvalues of a real supersymmetric tensor. *Journal of Symbolic Computation*, 40(6):1302–1324, 2005.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Doug Rohde. SVDLIBC (available at `http://tedlab.mit.edu/~dr/SVDLIBC/`), 2007.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

Nathan Srebro, Tommi Jaakkola, et al. Weighted low-rank approximations. In *Proceedings of the International Conference on Machine learning*, volume 3, pages 720–727, 2003.

GW Stewart and Ji-Guang Sun. Matrix perturbation theory (computer science and scientific computing), 1990.

Gilbert Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press Wellesley, MA, 4 edition, 2009.

Stephen A Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2009.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Dorothea Wagner and Frank Wagner. *Between min cut and graph bisection*. Springer, 1993.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.